

Structured Probabilistic Latent Modeling for Representation Learning and 3D Perception

Jiaxun Guo

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Information and Systems Engineering)

at Concordia University

Montréal, Québec, Canada

April 2026

© Jiaxun Guo, 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Jiaxun Guo**

Entitled: **Structured Probabilistic Latent Modeling for Representation Learning
and 3D Perception**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

Dr. Honghao Fu Chair

Dr. Guillaume-Alexandre Bilodeau External Examiner

Dr. Yassine Yaakoubi Examiner

Dr. Abdessamad Ben Hamza Examiner

Dr. Arash Mohammadi Examiner

Dr. Nizar Bouguila Supervisor

Dr. Wentao Fan Co-supervisor

Dr. Manar Amayri Co-supervisor

Approved by

Dr. Andrea Schiffauerova, GPD
Department of Concordia Institute for Information Systems Engineering

2026

Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Structured Probabilistic Latent Modeling for Representation Learning and 3D Perception

Jiaxun Guo, Ph.D.

Concordia University, 2026

Traditional representation learning and 3D perception methods commonly rely on Gaussian priors in latent spaces, which are inadequate for modeling complex structural asymmetries, directional uncertainty, and rotational symmetries. This thesis introduces structured probabilistic latent models with non-Gaussian distributions to address these challenges, enabling uncertainty-aware representation learning and robust 3D geometric perception. The research first investigates foundational probabilistic modeling for latent representations. We propose GamMM-VAE, a deep generative clustering framework that employs an asymmetric Gamma mixture prior together with a novel reparameterization to learn high-quality latent embeddings for non-negative vectors. To further model directional and axially symmetric data, we develop VIBinMM, a scalable variational inference framework for Bingham mixture models. By exploiting the mathematical relationship between Bingham and Gaussian distributions, the proposed approach bypasses intractable normalization constants and enables efficient inference in high-dimensional directional spaces. Building upon these foundations, the thesis extends structured latent priors to rotation-invariant 3D perception. We introduce the Shadow-informed Pose Feature (SiPF) and the Ga4DPF framework, which leverage Bingham distributions over unit quaternions to dynamically model global pose references. These mechanisms preserve strict rotation invariance while resolving local geometric ambiguities and mitigating feature collapse in symmetric structures. In addition, we present BRVSNet, a self-supervised framework for probabilistic rotation estimation and consistent multi-view generation. Extensive experiments on multiple benchmarks demonstrate that incorporating structured, non-Gaussian priors, particularly Gamma and Bingham distributions, leads to more expressive latent representations and robust, uncertainty-aware perception. The proposed models consistently outperform state-of-the-art methods across tasks ranging from unsupervised learning to fine-grained 3D spatial discrimination.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Nizar Bouguila, and my two co-supervisors, Dr. Wentao Fan and Dr. Manar Amayri. Over the past four years, they provided me with unwavering support, guidance, and encouragement. Whenever I encountered failures, challenges, or moments of self-doubt, they were always there with motivation and inspiration. Their expertise and insightful feedback were crucial in shaping both the direction and the quality of this research, and in helping me grow into a qualified doctoral researcher. Their invaluable lessons and experience will benefit me throughout my lifelong research career.

I would also like to thank my colleagues and friends in the laboratory for the wonderful time we spent together and for the supportive and collaborative environment we shared.

I would like to extend my special thanks to my close friend, Dr. Yunke Chen, for his emotional support during the most difficult moments of my journey.

Finally, I am deeply grateful to my family. I would not have achieved what I have today without their unwavering support and love. To my parents and my beloved Xin, your constant encouragement, understanding, and unconditional love have always been my source of strength and motivation. Your belief in me has given me the confidence to pursue my dreams and overcome every challenge along the way.

Contribution of Authors

This Ph.D. thesis comprises five manuscripts, with the two chapters already published, and the remaining chapters submitted for publication in reputable academic journals and conferences. Each chapter represents the contents of a manuscript, which has been carefully reformatted and reorganized to adhere to the guidelines provided by the School of Graduate Studies.

- **Manuscript 1 (Chapter 2):** Guo, J., Fan, W., Amayri, M., & Bouguila, N. (2025). Deep clustering analysis via variational autoencoder with Gamma mixture latent embeddings. *Neural Networks*, 183, 106979.
- **Manuscript 2 (Chapter 3):** Guo, J., Fan, W., Amayri, M., & Bouguila, N. (2025). Variational Inference for the Bingham Mixture Model and Applications. Manuscript submitted for publication to the *Pattern Recognition* journal.
- **Manuscript 3 (Chapter 4):** Guo, J., Fan, W., Amayri, M., & Bouguila, N. (2025). BRVSNet: A Self-Supervised Framework for Probabilistic Rotation Estimation and Consistent Multi-View Generation. Manuscript submitted for publication to the *IEEE Transactions on Pattern Analysis and Machine Intelligence* journal.
- **Manuscript 4 (Chapter 5):** Guo, J., Amayri, M., Bouguila, N., Liu, X., & Fan, W. (2026). Enhancing Rotation-Invariant 3D Learning with Global Pose Awareness and Attention Mechanisms. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*.
- **Manuscript 5 (Chapter 6):** Guo, J., Fan, W., Amayri, M., & Bouguila, N. (2026). 4D Local Modeling Toward Dynamic Global Perception for Ambiguity-free Rotation-Invariant Point Cloud Analysis. Manuscript submitted for publication to the *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Contents

List of Figures	x
List of Tables	xv
1 Introduction	1
1.1 Motivation and Challenges	1
1.2 Background and Related Works	3
1.2.1 Variational Inference and Probabilistic Latent Modeling	4
1.2.2 Structured Priors for Complex Data Modeling	6
1.2.3 Advancing Representation Learning toward 3D Geometric Perception	9
1.3 Thesis Contributions	12
1.4 Thesis Overview	15
2 Deep Clustering Analysis via Variational Autoencoder	17
2.1 Introduction	17
2.2 Related Works	20
2.3 Proposed Model	22
2.3.1 Objective Function	22
2.3.2 Reparameterization trick and approximation on Gamma distribution	25
2.3.3 Hyper-parameter learning and choosing strategy of GamMM-VAE	27
2.4 Experimental Results	28
2.4.1 Datasets and Experimental Settings	29
2.4.2 Clustering Results	31

2.4.3	Visualization of Learned Embeddings	33
2.4.4	Selecting the Number of Latent Clusters	37
2.4.5	Ablation Study	37
2.4.6	Comparison of the Convergence Speed	39
2.5	Conclusion	41
3	Variational Inference for the Bingham Mixture Model and Applications	42
3.1	Introduction	42
3.2	The Proposed Model	44
3.2.1	Bingham Distribution	45
3.2.2	Variational Inference for Bingham Mixture Model	46
3.3	Experiments	54
3.3.1	Synthetic Data	54
3.3.2	Earthquake Data Analysis	57
3.3.3	Gene Expression Data Clustering	60
3.4	Conclusion	64
4	BRVSNet: A Self-Supervised Framework for Probabilistic Rotation Estimation and Consistent Multi-View Generation	65
4.1	Introduction	65
4.2	Related Work	69
4.3	The Proposed Method	70
4.3.1	Revisit of DBN	70
4.3.2	Probabilistic Modeling of Rotation	71
4.3.3	Bingham-Based Latent Space	72
4.3.4	RWTA-ELBO loss	74
4.3.5	Weight Loss	77
4.3.6	Training Procedure	78
4.4	Experiments	79
4.4.1	Experimental Setup	79
4.4.2	Rotation Regression	81

4.4.3	Multi-View Generation	82
4.4.4	Hyperparameter Analysis	84
4.4.5	Ablation Study	86
4.5	Conclusion and Future Works	88
5	Enhancing Rotation-Invariant 3D Learning with Global Pose Awareness and Attention Mechanisms	90
5.1	Introduction	90
5.2	Related Works	93
5.2.1	Rotation-Sensitive 3D Point Cloud Analysis	93
5.2.2	Rotation-Robust Methods	93
5.3	Problem Definition and Background	94
5.4	The Proposed Method	96
5.4.1	Shadow-Informed Pose Feature	96
5.4.2	The RIAttnConv Operator	99
5.4.3	Network Architecture	100
5.5	Experiments	101
5.5.1	Implementation Details	101
5.5.2	3D Shape Classification	101
5.5.3	Real-World Shape Classification	102
5.5.4	Shape Part Segmentation	103
5.5.5	Ablation Studies	104
5.6	Conclusion	106
6	4D Local Modeling Toward Dynamic Global Perception for Ambiguity-free Rotation-Invariant Point Cloud Analysis	107
6.1	Introduction	107
6.2	Related Works	109
6.2.1	Rotation-Sensitive 3D Point Cloud Analysis	109
6.2.2	Rotation-Aware Methods	110
6.3	Preliminaries	111

6.3.1	Equivariance and Invariance	111
6.3.2	Receptive Fields in Point Cloud Convolutions	111
6.4	The Proposed Method	112
6.4.1	4D Robust Modeling for Local Representation	113
6.4.2	Dynamic Perception for Global Pose	116
6.4.3	Network Architecture	117
6.5	Experiments	118
6.5.1	Implementation Details	118
6.5.2	3D Shape Classification	118
6.5.3	Real-World Shape Classification	119
6.5.4	Shape Part Segmentation	121
6.5.5	Model Complexity	121
6.5.6	Ablation Studies	122
6.5.7	Visualization of Learned RI Features	123
6.6	Conclusion	124
7	Conclusions and Future Works	125
7.1	Conclusions	125
7.2	Future Works	126
	Bibliography	128

List of Figures

Figure 1.1	The limitation of Gaussian isotropy in modeling skewed data. The figure demonstrates the approximation of a Gamma distribution using a 3-component GMM. While the combined fit captures the general profile, the individual symmetric components struggle to align with the strictly non-negative and asymmetric nature of the target feature.	7
Figure 1.2	Visual comparison of distribution fitting on $\mathbb{S}^2$. The proposed Bingham distribution (blue) successfully models both the antipodal symmetry and anisotropy (elliptical shape) of the data. In contrast, the vMF distribution (red) fails due to its unimodal assumption, while the Watson distribution (green) provides a suboptimal fit due to its restriction to isotropic (circular) contours.	8
Figure 1.3	Comparison between Local and Global Pose-aware Descriptors. Strictly local invariant descriptors like PPF (left) suffer from feature collapse. In contrast, the proposed SiPF (right) utilizes a global reference R_g , derived from directional probabilistic modeling, to introduce global directional cues. This allows the model to distinguish between symmetric or distant components while maintaining rotation invariance.	11
Figure 2.1	The Flowchart of GamMM-VAE. The data generative process is as follows: 1) Encoder $g(\mathbf{x}; \phi)$ gets the original data $\mathbf{x}$ and outputs the parameters $\tilde{\mu}$ and $\tilde{\sigma}$ of the adjusted Gaussian mixture model. 2) Latent embedding $\mathbf{z}$ is obtained by reparameterization trick and indirectly affected by the parameter π, α, β during backpropagation. 3) Decoder $f(\mathbf{z}; \theta)$ is responsible for the mapping of latent embedding $\mathbf{z}$ to generated data $\mu_{\mathbf{x}}$. 4) The difference between $\mathbf{x}$ and $\mu_{\mathbf{x}}$, the parameters $\tilde{\alpha}$ and $\tilde{\beta}$ of the Gamma mixture model, and the initial parameters π, α , and β jointly construct the evidence lower bound as a loss to continuously optimize the model and update π, α , and β during training.	18

Figure 2.2	Approximation from Gamma to Gaussian.	26
Figure 2.3	Visualization of the embeddings learned by GamMM-VAE on MNIST. All points in each sub-figure are predictions for the clusters of digits given by the subtitle. The different marks in sub-figures are the ground-truth labels.	34
Figure 2.4	t-SNE visualization of obtained latent embeddings by GamMM-VAE for the MNIST dataset.	35
Figure 2.5	The generated samples of MNIST and F-MNIST by ClusterGAN and GamMM-VAE. The digits and clothes in the same row come from the same cluster.	36
Figure 2.6	The generated samples of MNIST with different numbers of clusters. (Each row represents a cluster.)	37
Figure 2.7	Gradient descent of loss in pretrain with respect to the adjustment coefficients on MNIST dataset.	38
Figure 2.8	Accuracy with different adjustment coefficients on MNIST dataset.	39
Figure 2.9	Comparison of the convergence speed by different models on MNIST dataset.	40
Figure 3.1	Visualizations of 3D Bingham distributions with varying Ω and fixed orientation matrix $\mathbf{U} = [[0, 0, 1], [0, 1, 0], [1, 0, 0]]$. The shape transitions from uniform to highly concentrated and annular as dispersion parameters vary.	46
Figure 3.2	Visualization of the relationship between the Bingham and zero-mean Gaussian distributions. The parameters Ω_k and $\mathbf{U}_k$ of the Bingham distribution are obtained as the eigenvalues and eigenvectors of $-\frac{1}{2}\mathbf{\Lambda}_k$ from the corresponding Gaussian component.	47
Figure 3.3	Directed acyclic graph (DAG) representation of the Bingham mixture model via Gaussian transformation. Nodes represent random variables, edges indicate dependencies, and plates denote replication across data points.	49
Figure 3.4	The visualization of the synthetic data sets.	55
Figure 3.5	Comparison of original and estimated distributions of earthquake data by regions.	58
Figure 3.6	Top 4 clustering results obtained from the YMG, YS, HFS and YCC data sets by the proposed model.	62
Figure 3.7	Performance of the proposed model on YMG, YS, HFS and YCC data sets in terms of Homogeneity and Separation with different values of K	64

Figure 4.1	Illustration of the Rotation-Awareness in the Proposed Framework. The framework builds upon a Bingham-based latent space to enable rotation-aware learning and downstream tasks. (Top) In the rotation-aware building and view synthesis module, training data are embedded into the latent space, and the generator leverages its rotation-awareness to synthesize novel views from arbitrary latent directions. (Bottom) In the rotation uncertainty estimation module, test data are projected into the latent space, where the Bingham distribution models rotation hypotheses and quantifies uncertainty for robust rotation regression.	66
Figure 4.2	Framework Pipeline. The framework consists of four main components: (1) the DBN block processes images or point clouds to generate mixture parameters; (2) the Bingham-based space embedding, together with the rotation modeling block, transforms mixture embeddings into latent orientation distributions and aligns them with ground-truth quaternion constraints; (3) the supervision function block refines estimation via RWTA-ELBO loss and weight loss; and (4) the discriminator block produces pseudo ground truths to enhance latent representation, while only the mode latent z_{mode} is used as input to the Multi-View Generator, which, after training, develops strong rotation-awareness on the Bingham distribution and synthesizes novel views (images or point clouds) that are geometrically consistent with the estimated rotations. During inference, it is sufficient to provide only the corresponding quaternion to synthesize the desired multi-view outputs.	68
Figure 4.3	Visual Comparison of Rotation Parameters to Point Clouds on the ModelNet10 Dataset. The inputs of Ours: unseen angles. The inputs of DiffusionPM and WarpingGAN: ground truth point clouds.	83
Figure 4.4	Visual comparison of rotation parameter predictions on images from the AmReLoc dataset. Each category includes two comparison sets: the first and third rows show the ground truth, and the second and fourth rows show the corresponding generated images.	83
Figure 4.5	The accuracy comparison of NofM on AmReLoc dataset by category and threshold.	85
Figure 4.6	The CD comparison of NofM on ModelNet10 dataset by category. The red circles indicate the minimum CD for each category, corresponding to the optimal NofM value.	86

Figure 5.1 Illustration of pose ambiguity caused by symmetric local structures. Traditional Point Pair Features (PPF) capture only local geometric relationships and thus yield identical representations for symmetric regions with similar local geometry (e.g., left and right wing tips). In contrast, the proposed SiPF incorporates global pose awareness by introducing a ‘shadow’ reference point derived from a shared rotation matrix R_g to incorporate pose-aware context, enabling consistent and discriminative RI representations across arbitrary orientations. . . . 91

Figure 5.2 Overview of the proposed SiPF and RIAttnConv pipeline. Our framework consists of three main components: (1) Task-adaptive Shadow Locating estimates a globally consistent reference rotation R_g by modeling unit quaternions with a Bingham distribution, generating ‘shadow’ points for each input; (2) SiPF and Input Feature Extraction constructs LRFs $\{\mathcal{L}_j\}_{j=1}^k$ and computes SiPFs $\mathbf{P}_r$ by combining local geometric descriptors with global shadow-based pose differences; (3) RIAttnConv embeds SiPFs to produce adaptive kernel weights $\mathbf{W}_r$, which are used in a RI attention block to guide feature aggregation. A reversed EdgeConv module then fuses neighbourhood and center features x_r to produce the final output x'_r 97

Figure 5.3 Visualization of part segmentation results on ShapeNetPart Dataset under z/SO(3) setting. The leftmost column is the ground truth. The testing results under arbitrary rotations are on the rest columns. 105

Figure 6.1 Traditional local pose-aware descriptors (e.g., PPF [1]) exhibit limited expressiveness and are prone to ambiguity in symmetric structures. In contrast, our approach enhances local geometric representation by lifting point-pair features into a higher-dimensional 4D space, enabling more robust and distinctive feature encoding. By further incorporating a global anchor derived from a dynamic global rotation R_g , Ga4DPF achieves more discriminative and rotation-consistent representations. 108

Figure 6.2	The pipeline of the proposed method. Our framework consists of two key components: (1) Dynamic Perception for Global Pose, which addresses the inherent <i>ambiguity of symmetric local structures</i> by estimating a globally consistent reference rotation R_g through Bingham distribution modeling, generating adaptive anchor points that provide additional global cues; (2) 4D Robust Modeling for Local Representation, where constructs LRFs $\{\mathcal{L}_j\}_{j=1}^k$ and a learnable steerable transform (ST) lifts the point cloud into 4D space, enabling fine-grained geometric encoding and enhancing the robustness of local representations against <i>point-pair ambiguity</i> . Finally, a Rotation-Invariant Convolution module embeds and aggregates the resulting Ga4DPF features, producing discriminative representations suitable for downstream tasks such as shape classification and part segmentation. . . .	113
Figure 6.3	Visualization of part segmentation results on ShapeNetPart Dataset under $z/SO(3)$ setting. The leftmost column is the ground truth. The testing results under arbitrary rotations are on the rest columns.	121
Figure 6.4	Visualization of learned dense features for shape part segmentation under different rotations. Our method (bottom row) yields more consistent and discriminative representations, especially for symmetric parts, while preserving rotation invariance.	124

List of Tables

Table 2.1	Summary of the Benchmark Datasets	31
Table 2.2	Comparison of Clustering Performance on Benchmark Datasets in terms of ACC and NMI.	32
Table 2.3	Comparison of Clustering Performance on Benchmark Datasets in terms of ARI. . . .	32
Table 3.1	True and Estimated Parameters for the First Synthetic Data Set.	56
Table 3.2	True and Estimated Parameters for the Second Synthetic Data Set.	57
Table 3.3	Estimated parameters of Bingham distribution for the earthquake data.	59
Table 3.4	Wasserstein distance from the original earthquake data by Bingham and Watson distribution. (<i>Average results of 10 experiments</i>)	59
Table 3.5	H_{ave} and S_{ave} of gene expression data in terms of different methods. (<i>Average results of 10 experiments</i>)	61
Table 3.6	Computational run time (in seconds) of different methods on clustering gene expression data.	63
Table 4.1	Rotation Estimation Accuracy on the AmReLoc Dataset. DBN results are reproduced under our experimental settings. The best results are shown in bold, and the second-best results are marked with a dagger ($\dagger$).	81
Table 4.2	Rotation Estimation on ModelNet10. DBN results are reproduced under our experimental settings. CD values are scaled by 10^2 . The best results are shown in bold, and the second-best results are marked with a dagger ($\dagger$).	81
Table 4.3	Generation performance on the AmReLoc dataset. PSNR: 0– ∞ ; SSIM: 0–1; LPIPS: 0–1.	82

Table 4.4	Generation Performance on ModelNet10. CD and EMD values are scaled by 10^2 . The best results are shown in bold.	82
Table 4.5	Ablation study of rotation regression with varying numbers of mixtures (NofM). For AmReLoc, the metric is accuracy at different thresholds; for ModelNet10, the metric is CD, scaled by 10^2 . The results are the average value across categories on each dataset. The best results for each method are in bold, and the best overall results are marked with a star (*). . . .	85
Table 4.6	Ablation study of rotation regression with and without a discriminator (Disc.). For AmReLoc, the metric is accuracy at different thresholds; for ModelNet10, the metric is CD, scaled by 10^2 . The best results for each method are in bold, and the best overall results are marked with a star (*).	87
Table 5.1	Shape classification accuracy (%) on ModelNet40 dataset. 'pc' and 'n' represent the 3D coordinates and normal vectors of input data, respectively.	102
Table 5.2	Real-world classification accuracy (%) on ScanObjectNN OBJ_BG dataset.	103
Table 5.3	Part segmentation results on ShapeNetPart dataset. 'C. mIoU' stands for mIoU (%) over all classes while 'I. mIoU' denotes mIoU over all instances.	104
Table 5.4	Ablation study on the relative pose feature $\mathcal{P}_r^j$	104
Table 5.5	Ablation study on the RI convolution.	105
Table 5.6	Comparison of FLOPs, parameter count, and mIoU among different models with point coordinates only. (* indicates the results are estimated based on the classification architecture.)	106
Table 6.1	Shape classification accuracy (%) on ModelNet40 dataset. 'pc' and 'n' represent the 3D coordinates and normal vectors of input data, respectively.	119
Table 6.2	Real-world classification accuracy (%) on ScanObjectNN OBJ_BG dataset.	120
Table 6.3	Part segmentation results on ShapeNetPart dataset. 'C. mIoU' stands for mIoU (%) over all classes while 'I. mIoU' denotes mIoU over all instances.	120
Table 6.4	Comparison of FLOPs, parameter count, and mIoU on ShapeNetPart (z/SO(3)) with point coordinates only. (* indicates the results are estimated based on the classification architecture.)	122
Table 6.5	Ablation study on the relative pose feature. Reported results are classification accuracy under the z/SO(3) setting.	122

Table 6.6 Ablation study on the sampling number N_s for geometric mean estimation. Reported results are classification accuracy under the z/SO(3) setting. 123

Chapter 1

Introduction

In this chapter, we present an overview of the motivations, fundamental challenges, methodological foundations, and main contributions of this thesis. Section 1.1 introduces the motivation of the research, highlighting the limitations of conventional Gaussian-based latent modeling in representation learning and rotation-invariant 3D perception. Section 1.2 reviews the background and related works on probabilistic latent modeling, structured priors, and representation learning for 3D geometric perception, providing the theoretical context for the proposed methods. Section 1.3 summarizes the key contributions of this thesis, emphasizing the role of structured non-Gaussian probabilistic modeling in addressing asymmetry, uncertainty, and rotational ambiguity. Finally, Section 1.4 outlines the organization of the thesis and guides the reader through the subsequent chapters.

1.1 Motivation and Challenges

The rapid progress of deep representation learning has fundamentally transformed the processing of high-dimensional data, shifting the paradigm from manual feature engineering to autonomous, data-driven feature extraction. By mapping raw observations into lower-dimensional latent spaces, neural networks can disentangle complex dependencies and uncover latent generative factors [2, 3, 4, 5]. This capacity underpins a wide spectrum of tasks, from unsupervised clustering and pattern recognition [6, 7] to advanced 3D geometric perception and robotic vision [8, 9, 10, 11]. However, as deep learning systems transition from controlled lab settings to complex real-world environments, the limitations of conventional latent modeling approaches have become increasingly pronounced.

At the core of most deep generative models, such as Variational Autoencoders (VAEs) [12] and Generative Adversarial Networks (GANs) [13], lies the assumption of Gaussian latent priors. Despite their analytical tractability and closed-form solutions, multivariate Gaussians are isotropic and Euclidean by nature. This “one-size-fits-all” prior often fails to reflect the intrinsic geometric and statistical structures of real-world data.

First, the inherent symmetry of Gaussian distributions is inappropriate for modeling skewed, strictly positive, or directional features. Imposing symmetric priors on asymmetric or non-negative data often demands excessive mixture components to approximate distribution tails, leading to overparameterization and reduced interpretability [14, 15, 16, 17]. Second, most latent representations are deterministic, collapsing uncertainty into point estimates. This simplification disregards inherent uncertainties arising from sensor noise, occlusions, and viewpoint variability. These factors are critical to safety-critical applications such as autonomous driving [18, 19]. Third, existing 3D perception methods often employ local receptive fields to attain rotation invariance [20, 21, 22], but such locality comes at the cost of discarding global structure, resulting in limited context and the inability to distinguish symmetric or repetitive patterns [23, 24].

Although unsupervised learning and 3D perception are distinct subfields, they share a foundational challenge: modeling data structure while respecting its inherent asymmetries, symmetries, and uncertainties.

- **Asymmetry in Latent Embeddings:** In clustering, real-world data rarely conforms to symmetric distributions. Capturing skewness or non-negativity is essential for learning compact, discriminative cluster-specific representations beyond the capacity of Gaussian Mixture Models [14, 15].
- **Symmetry and Ambiguity in 3D Space:** In 3D perception, rotational symmetry is both necessary and problematic. While rotation invariance is desirable to handle arbitrary viewpoints, physical symmetries introduce ambiguity [25, 26, 27]. Without explicit modeling, symmetric structures can cause feature collapse, where distinct but locally similar parts become indistinguishable.
- **Uncertainty as a Unifying Challenge:** Both domains require reasoning under multimodal uncertainty. Whether to consider the uncertainty of a data point belonging to multiple clusters or the ambiguity of objects pose when viewed from a symmetric angle, the point estimate is insufficient [28, 29]. Capturing this uncertainty requires a probabilistic framework that can represent distributions over non-Euclidean spaces, such as unit hyperspheres where directional data naturally reside [30, 31, 32].

These challenges motivate a shift toward Structured Probabilistic Latent Modeling. Unlike conventional black-box embeddings, this approach integrates domain-specific inductive biases directly into the latent space via tailored probabilistic priors. Here, “structure” refers to the principled design of latent distributions that align with the topological or algebraic properties of the data, such as non-negativity for physical measurements or axial symmetry for 3D rotations [33, 34]. By moving beyond Gaussian-centric and deterministic assumptions, structured probabilistic models provide three key advantages:

- (1) **Expressiveness:** Structured models can capture skewed, directional, or periodic features using fewer parameters and with greater fidelity than traditional isotropic priors.
- (2) **Disambiguation:** Incorporating global structural priors enables the resolution of local geometric ambiguities, a common limitation in existing rotation-invariant pipelines.
- (3) **Robustness:** Probabilistic reasoning facilitates uncertainty quantification, providing confidence-aware predictions for perception tasks such as multi-view synthesis and spatial reasoning.

This thesis investigates how structured, non-Gaussian latent priors, particularly Gamma and Bingham distributions, can be rigorously derived, efficiently optimized, and seamlessly integrated into deep learning pipelines. It advances the state of probabilistic representation learning and enables robust, uncertainty-aware 3D geometric perception.

1.2 Background and Related Works

This section establishes the theoretical and methodological foundations of the proposed research. We first review variational inference and probabilistic latent variable modeling, emphasizing their role in scalable posterior approximation and deep generative learning. Then, we examine structured probabilistic priors beyond conventional Gaussian assumptions, highlighting their importance for modeling asymmetry, constraints, and directional uncertainty in complex data. Finally, these foundations are situated within the context of 3D representation learning, where the limitations of local rotation-invariant descriptors are analyzed, motivating structured directional probabilistic modeling for global, pose-aware representations.

1.2.1 Variational Inference and Probabilistic Latent Modeling

The core challenge in probabilistic latent variable modeling lies in the exact inference of the posterior distribution $p(\mathbf{z} \mid \mathbf{x})$ over latent variables $\mathbf{z}$ given observations $\mathbf{x}$. By Bayes' rule, the posterior is defined as

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})}, \quad (1)$$

where the marginal likelihood (or evidence)

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \quad (2)$$

requires integrating over the latent space. In deep generative models with high-dimensional latent variables and nonlinear parameterizations, this integral is generally intractable, making exact posterior inference computationally prohibitive [12].

Variational Inference (VI) addresses this challenge by reformulating posterior inference as a functional optimization problem [14, 15]. The central idea is to approximate the intractable posterior $p_{\theta}(\mathbf{z} \mid \mathbf{x})$ with a tractable variational distribution $q_{\phi}(\mathbf{z})$ drawn from a restricted family $\mathcal{Q}$. The approximation quality is typically measured by the Kullback–Leibler (KL) divergence [15, 16]:

$$KL(q_{\phi}(\mathbf{z}) \parallel p_{\theta}(\mathbf{z} \mid \mathbf{x})) = \mathbb{E}_{q_{\phi}(\mathbf{z})} \left[\log \frac{q_{\phi}(\mathbf{z})}{p_{\theta}(\mathbf{z} \mid \mathbf{x})} \right]. \quad (3)$$

By optimizing the variational parameters ϕ , VI enables scalable and efficient posterior approximation in high-dimensional latent variable models [14].

Since the true posterior $p_{\theta}(\mathbf{z} \mid \mathbf{x})$ is intractable, direct minimization of the KL divergence in Equation 3 is not feasible, as it depends on the marginal likelihood $\log p_{\theta}(\mathbf{x})$. Variational methods circumvent this issue by introducing a tractable lower bound on the marginal log-likelihood via Jensen's inequality [12, 35]:

$$\log p_{\theta}(\mathbf{x}) = KL(q_{\phi}(\mathbf{z}) \parallel p_{\theta}(\mathbf{z} \mid \mathbf{x})) + \mathcal{L}(\theta, \phi; \mathbf{x}), \quad (4)$$

where $\mathcal{L}(\theta, \phi; \mathbf{x})$ denotes the Evidence Lower Bound (ELBO) [12]. Maximizing the ELBO is therefore equivalent to minimizing the KL divergence between the variational distribution and the true posterior, while simultaneously providing a lower bound on the data evidence [35]. Within the Variational Autoencoder

(VAE) framework, the ELBO can be expressed as:

$$\mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\boldsymbol{\theta}}(\mathbf{x} | \mathbf{z})] - KL(q_{\phi}(\mathbf{z} | \mathbf{x}) \| p(\mathbf{z})), \quad (5)$$

where the first term encourages accurate reconstruction of the observations from the latent variables, and the second term regularizes the latent space by aligning the prior $p(\mathbf{z})$ with the variational posterior $q_{\phi}(\mathbf{z} | \mathbf{x})$. This regularization not only mitigates overfitting but also enforces coherent generative structure in the learned latent space.

The ELBO provides a unified objective for the joint optimization of the model parameters $\boldsymbol{\theta}$ and the variational parameters ϕ . To enable efficient learning in high-dimensional latent variable models, the Stochastic Gradient Variational Bayes (SGVB) estimator was introduced. SGVB allows the ELBO to be optimized using stochastic gradient descent with Monte Carlo estimates of the expectations, thereby avoiding expensive iterative inference procedures such as MCMC [12]. By shifting the burden of inference to the optimization of tractable variational distributions, SGVB substantially improves the scalability of variational models to large-scale datasets and complex deep architectures [3, 5].

A major challenge in VAE training arises from the need to backpropagate gradients through stochastic latent variables sampled as $\mathbf{z} \sim q_{\phi}(\mathbf{z} | \mathbf{x})$. The reparameterization trick resolves this issue by expressing the latent variable as a deterministic function of the variational parameters and an auxiliary noise variable, i.e., $\mathbf{z} = g_{\phi}(\boldsymbol{\epsilon})$ with $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$. For example, in the Gaussian case, this takes the form $\mathbf{z} = \boldsymbol{\mu}_{\phi}(\mathbf{x}) + \boldsymbol{\sigma}_{\phi}(\mathbf{x}) \odot \boldsymbol{\epsilon}$. This formulation enables gradients to propagate through the sampling process, thereby allowing stable and efficient training of deep generative models via standard backpropagation [12].

The choice of latent space distributions fundamentally determines the flexibility and robustness of representation learning. Foundational probabilistic models commonly adopt Gaussian priors due to their mathematical convenience and analytical tractability [12]. However, such assumptions are often inadequate for data exhibiting intrinsic asymmetry, non-negativity, or geometric constraints [14, 32]. Building upon the variational inference framework and its associated optimization principles, this thesis develops novel reparameterization techniques for structured, non-Gaussian latent distributions. In particular, we introduce Gamma distribution reparameterizations tailored to non-negative latent variables, and propose efficient variational inference frameworks for axially symmetric directional latent spaces, including mixture models based on the Bingham distribution. Collectively, these contributions enable more expressive probabilistic

latent models that explicitly encode structural constraints inherent to the data.

1.2.2 Structured Priors for Complex Data Modeling

In deep probabilistic graphical models, the choice of the prior distribution fundamentally determines the capacity of the latent space to capture the underlying structure of the data. This section examines the advantages of structured probabilistic priors for modeling asymmetric and constrained latent variables, in contrast to conventional Gaussian-based assumptions.

Traditional Variational Autoencoders predominantly adopt a multivariate Gaussian distribution as the latent prior due to its analytical tractability [12]. The probability density function (PDF) of a D -dimensional Gaussian distribution is given by:

$$\mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp \left(-\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right). \quad (6)$$

Despite its widespread use, the symmetry and unbounded support of the Gaussian distribution impose intrinsic limitations when modeling real-world latent features. First, Gaussian priors are defined over the entire real space $(-\infty, \infty)$, whereas many physical measurements and feature magnitudes are strictly non-negative [14, 15]. Second, Gaussian distributions offer limited flexibility in capturing skewed or heavy-tailed behavior commonly observed in practical data [16]. As a result, Gaussian Mixture Models (GMMs) often require a large number of components to approximate asymmetric structures, leading to redundant parameterization and diminished interpretability [36].

For data that is strictly non-negative and exhibits skewed characteristics, the Gamma distribution provides a principled structured alternative to Gaussian priors as Figure 1.1. The probability density function of a univariate Gamma distribution is defined as:

$$p(z \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}, \quad z > 0, \quad (7)$$

where $\alpha > 0$ and $\beta > 0$ denote the shape and rate parameters, respectively. In contrast to symmetric Gaussian distributions with unbounded support, the Gamma distribution naturally enforces non-negativity and flexibly models a wide range of distributional shapes. By varying the shape parameter α , the distribution can transition from highly skewed regimes to approximately symmetric forms, making it well suited for

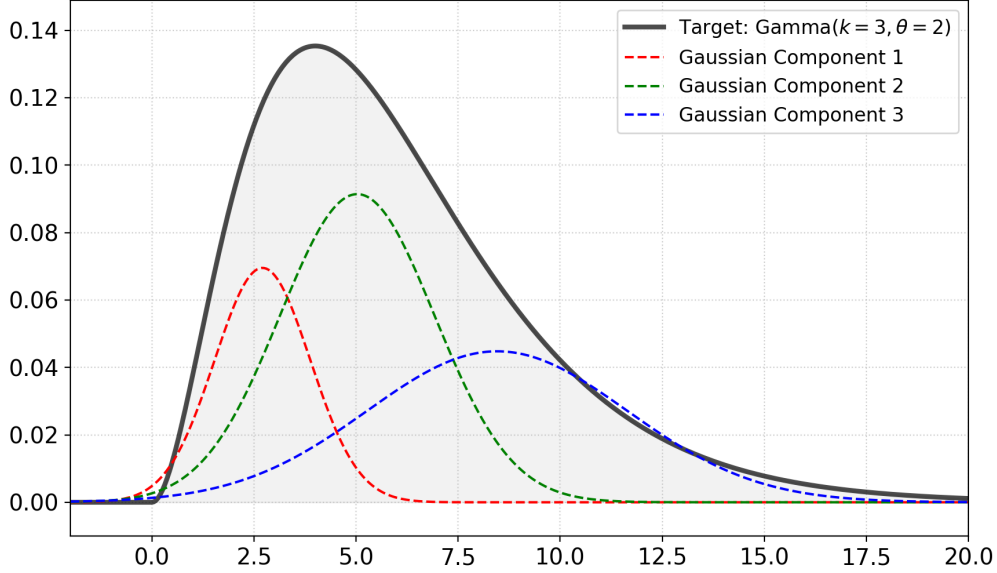


Figure 1.1: The limitation of Gaussian isotropy in modeling skewed data. The figure demonstrates the approximation of a Gamma distribution using a 3-component GMM. While the combined fit captures the general profile, the individual symmetric components struggle to align with the strictly non-negative and asymmetric nature of the target feature.

asymmetric latent representations [36].

To integrate the Gamma distribution into the VAE framework while avoiding the instability of direct Gamma reparameterization, a novel “bridge” is constructed between the normal and Gamma distributions. This design is motivated by the asymptotic convergence of the Gamma distribution to a normal distribution as the shape parameter α increases [37]. Specifically, the mean and variance of a Gamma distribution satisfy:

$$\mu = \frac{\alpha}{\beta}, \quad \sigma^2 = \frac{\alpha}{\beta^2}, \quad (8)$$

which yield the parameter mapping

$$\alpha = \frac{\mu^2}{\sigma^2}, \quad \beta = \frac{\mu}{\sigma^2}. \quad (9)$$

Here, μ and σ^2 correspond to the mean and variance of an associated Gaussian approximation. In practice, the encoder network predicts μ and σ , and latent variables are sampled using the standard Gaussian reparameterization trick, $\mathbf{z}^{(l)} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\sigma}} \odot \boldsymbol{\epsilon}^{(l)}$. This construction enables stable gradient-based optimization while approximating Gamma-distributed latent embeddings. To encourage positivity and control the scale of the latent variables, the predicted mean μ is further constrained to a bounded positive range via a normalization function, thereby reducing the probability mass assigned to invalid negative regions and improving

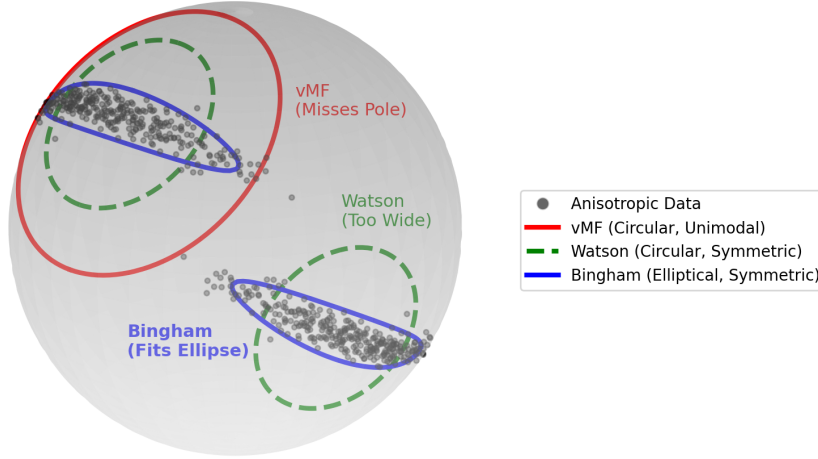


Figure 1.2: Visual comparison of distribution fitting on $\mathbb{S}^2$. The proposed Bingham distribution (blue) successfully models both the antipodal symmetry and anisotropy (elliptical shape) of the data. In contrast, the vMF distribution (red) fails due to its unimodal assumption, while the Watson distribution (green) provides a suboptimal fit due to its restriction to isotropic (circular) contours.

generalization for non-negative latent representations.

When data is embedded on the unit hypersphere $\mathbb{S}^{D-1}$, as is common for L_2 -normalized feature representations or unit quaternions encoding 3D rotations, conventional Euclidean distributions become inappropriate. In such directional representations, the orientation of a vector conveys substantially more information than its magnitude [34]. The Bingham distribution is a fundamental model in directional statistics for axially symmetric data. Its probability density function is defined as [32]

$$f(\mathbf{x} \mid \mathbf{V}, \mathbf{\Lambda}) = \frac{1}{N(\mathbf{\Lambda})} \exp(\mathbf{x}^T \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \mathbf{x}), \quad (10)$$

where $\mathbf{V}$ is an orthogonal matrix whose columns specify principal directions, $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_D)$ contains the concentration parameters with $\lambda_1 \leq \dots \leq \lambda_D = 0$, and $N(\mathbf{\Lambda})$ denotes the intractable normalization constant [38]. The matrix $\mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$ thus defines a symmetric parameter matrix governing the concentration of the distribution on the hypersphere.

In contrast to the von Mises–Fisher distribution [39], which models unimodal directional data, and the Watson distribution [40], which assumes axisymmetric concentration around a principal direction, the Bingham distribution explicitly exhibits antipodal symmetry, satisfying $f(\mathbf{x}) = f(-\mathbf{x})$ [32]. This property is critical for 3D rotation modeling, as unit quaternions $\mathbf{q}$ and $-\mathbf{q}$ correspond to the same physical rotation.

Crucially, there exists a deep mathematical connection between the Bingham distribution and zero-mean

Gaussian distributions [32]. In particular, a Bingham distribution can be characterized through an angular Gaussian construction, whereby samples from a zero-mean Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ are normalized to lie on the unit hypersphere. By exploiting this relationship under an appropriate parameterization, the variational inference framework (VIBinMM) developed in this thesis effectively circumvents the intractable normalization constant $N(\Lambda)$, enabling scalable and efficient inference for high-dimensional directional latent variables. This principled relationship between Euclidean Gaussian modeling and structured directional distributions forms the theoretical foundation for robust, uncertainty-aware rotation estimation and consistent multi-view generation, as demonstrated by the proposed BRVSNet framework.

1.2.3 Advancing Representation Learning toward 3D Geometric Perception

This section examines how principles of representation learning extend from conventional visual understanding to three-dimensional geometric perception. We first formalize the notion of rotation invariance (RI) for 3D point cloud data and then analyze why widely adopted local invariant representations, when considered in isolation, are fundamentally insufficient for capturing global geometric structure [41, 42, 25]. These observations naturally motivate the need for structured probabilistic modeling that can bridge local geometric reasoning with global pose awareness.

The central objective of representation learning is to learn a mapping function $f : \mathcal{X} \rightarrow \mathcal{Z}$ that transforms high-dimensional raw observations into compact latent representations $\mathbf{z}$, thereby facilitating downstream inference tasks [3, 5]. Importantly, representation learning is not merely concerned with dimensionality reduction. It aims to uncover latent structures that encode meaningful semantic and geometric information. In the context of 3D perception, such representations must further exhibit robustness to spatial transformations, notably rigid rotations, so that semantic consistency is preserved across arbitrary view-points [42, 26].

To formalize rotation invariance, a 3D object is represented as a point cloud $\mathcal{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_N\} \in \mathbb{R}^{N \times 3}$, where each point $\mathbf{p}_i$ denotes a spatial coordinate on the object surface. Let

$$SO(3) = \{\mathbf{R} \in \mathbb{R}^{3 \times 3} \mid \mathbf{R}^T \mathbf{R} = \mathbf{I}, \det(\mathbf{R}) = 1\} \quad (11)$$

denote the group of 3D rotations. A mapping function $f : \mathbb{R}^{N \times 3} \rightarrow \mathbb{R}^D$ is said to be rotation invariant if its

output remains unchanged under any global rigid rotation applied uniformly to all points in the input [41]:

$$f(\mathcal{P}\mathbf{R}) = f(\{\mathbf{p}_1\mathbf{R}, \dots, \mathbf{p}_N\mathbf{R}\}) = f(\mathcal{P}), \quad \forall \mathbf{R} \in SO(3). \quad (12)$$

This condition ensures that the extracted latent representation $\mathbf{z}$ depends only on the intrinsic geometry of the object, and is independent of its absolute orientation in space [43].

While rotation-equivariant representations preserve orientation information by transforming synchronously with the input, i.e., $f(\mathcal{P}\mathbf{R}) = f(\mathcal{P})\mathbf{R}$ when the latent features reside in a rotation-aware space [44], this work focuses on constructing strictly rotation-invariant representations through structured modeling. A common strategy for achieving invariance is to eliminate dependence on a global reference frame by encoding relative geometric relationships between points [1]. A representative example is the Point Pair Feature (PPF) formulation [1], where for two points $(\mathbf{p}_i, \mathbf{p}_j)$ and their associated normal vectors $(\mathbf{n}_i, \mathbf{n}_j)$, a rotation-invariant descriptor is defined as

$$\psi_{ij} = (\|\mathbf{d}\|_2, \angle(\mathbf{n}_i, \mathbf{d}), \angle(\mathbf{n}_j, \mathbf{d}), \angle(\mathbf{n}_i, \mathbf{n}_j)), \quad (13)$$

where $\mathbf{d} = \mathbf{p}_j - \mathbf{p}_i$ denotes the relative displacement vector. Since distances and angles are preserved under $SO(3)$ transformations, such descriptors are mathematically rotation invariant. However, this invariance is achieved at the expense of discarding absolute pose information, which fundamentally limits the model’s ability to reason about global structure and to disambiguate geometrically symmetric configurations [43].

This limitation becomes particularly evident in the dominant paradigm of rotation-invariant point cloud analysis [25, 26], which follows a “local description followed by global aggregation” strategy. Specifically, local neighbourhoods are constructed using K-Nearest Neighbors (KNN), defining a support region $\mathcal{N}(p_i) = \{p_{j_1}, \dots, p_{j_k}\}$ for each query point. Invariant features are then extracted within these neighbourhoods via local reference frames or pairwise geometric descriptors [1]. From an information-theoretic and geometric perspective, this process effectively restricts feature learning to a small region in the Euclidean space, approximately characterized by a ball $B(p_i, r)$, thereby capturing only local tangent-space geometry while neglecting long-range structural dependencies [43, 45].

More critically, strictly local invariant descriptors induce an inherent many-to-one mapping [43, 45].

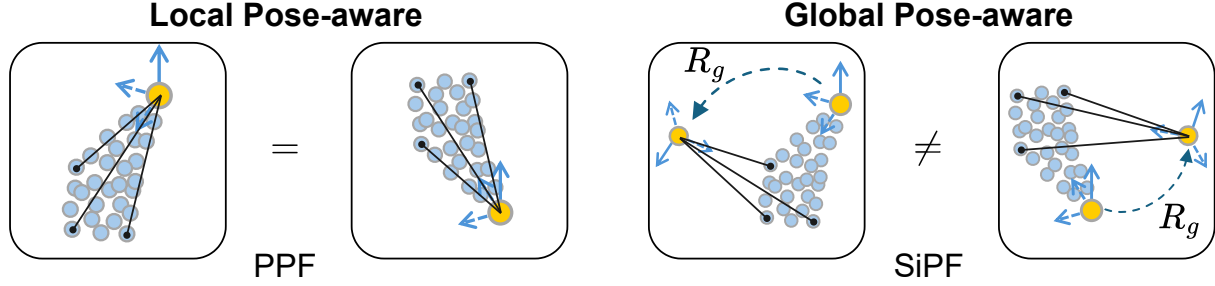


Figure 1.3: Comparison between Local and Global Pose-aware Descriptors. Strictly local invariant descriptors like PPF (left) suffer from feature collapse. In contrast, the proposed SiPF (right) utilizes a global reference R_g , derived from directional probabilistic modeling, to introduce global directional cues. This allows the model to distinguish between symmetric or distant components while maintaining rotation invariance.

For objects exhibiting geometric symmetries, spatially distant regions may share nearly identical local configurations [46]. As a consequence, local invariant mappings satisfy

$$\psi(\mathcal{N}(p_{\text{left}})) \approx \psi(\mathcal{N}(p_{\text{right}})), \quad (14)$$

rendering symmetric components indistinguishable in the latent space. This phenomenon, which we refer to as feature collapse, arises when spatially distinct yet locally similar regions are mapped to overlapping representations. Because absolute pose information is explicitly removed, the model loses access to global directional cues, making it impossible to consistently define relational concepts such as “left” and “right” [41]. This collapse fundamentally limits performance in fine-grained tasks such as part segmentation and correspondence estimation [23, 24].

These observations indicate that achieving rotation invariance through local geometric normalization alone is insufficient for global 3D perception. To overcome this limitation, this thesis adopts directional probabilistic modeling as a principled mechanism for capturing global pose information in a distributional manner, while preserving strict invariance of the learned representations. In particular, rotational uncertainty is modeled directly on the unit quaternion hypersphere S^3 using the Bingham distribution [32],

$$f(\mathbf{q} \mid \mathbf{V}, \mathbf{\Lambda}) \propto \exp(\mathbf{q}^T \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \mathbf{q}), \quad (15)$$

which naturally encodes antipodal symmetry and supports rotational ambiguity.

By learning a probabilistic global reference frame rather than a deterministic pose estimate, the model

introduces global directional structure without sacrificing rotation invariance. This formulation forms the theoretical foundation for the Shadow-informed Pose Feature (SiPF) and the Ga4DPF framework developed in subsequent chapters.

1.3 Thesis Contributions

This thesis addresses fundamental limitations in traditional latent space modeling, with a particular focus on the restrictive assumptions imposed by isotropic Gaussian priors. To this end, it develops a series of structured probabilistic latent models designed to capture complex structural asymmetries, directional uncertainties, and rotational symmetries that naturally arise in real-world data distributions. The key contributions of this research are summarized as follows:

- **Structured Asymmetric Latent Modeling for Deep Generative Clustering**

We propose a novel variational autoencoder (VAE)-based deep clustering model called GamMM-VAE, which is able to learn latent representations of the training data for clustering. First, most existing VAE-based deep clustering methods use the Gaussian mixture model (GMM) as a prior on the latent space, and we employ a more flexible asymmetric Gamma mixture model instead to perform a higher quality embedding of the data latent space. Second, since the Gamma is strictly positive, in order to exploit the reparameterization trick of VAE, we propose a transformation method from Gaussian distribution to Gamma distribution. This method can also be considered as a Gamma distribution reparameterization trick. Finally, we derive the evidence lower bound (ELBO) based on the Gamma mixture model in an effective way for the stochastic gradient variational Bayesian (SGVB) estimator to optimize the proposed model. The effectiveness of our model is validated by quantitative comparisons with other state-of-the-art deep clustering models on five benchmark datasets. Moreover, due to the generative nature of VAE, the proposed model can generate highly realistic samples of specific classes without supervised information. This work is published in the journal Neural Networks [47].

- **Scalable Directional Latent Modeling for Variational Inference**

The Bingham distribution has gained increasing attention in recent years for modeling axially symmetric directional data. However, its practical use is hindered by the intractable normalization constant, which makes exact inference computationally expensive and limits the scalability of existing methods,

particularly those relying on numerical approximations or Markov Chain Monte Carlo (MCMC) techniques. To address these limitations, this work makes two key contributions. First, we propose a novel formulation that exploits the mathematical relationship between the Bingham distribution and zero-mean Gaussian distributions, allowing us to bypass the need to compute the normalization constant. This reformulation preserves the generality of the Bingham mixture model while significantly enhancing computational efficiency. Second, we introduce the first variational inference framework tailored specifically for Bingham mixture models, offering a scalable and flexible solution for modeling axially symmetric data in high dimensions. We demonstrate the effectiveness of the proposed method through extensive experiments on synthetic datasets, as well as two real-world applications: earthquake focal mechanism analysis and gene expression clustering. In both domains, our model consistently achieves superior performance, highlighting its practical utility in capturing the complex structure of axially symmetric data. The source code is publicly available at: <https://github.com/jiaxunguo/VIBinMM>. This manuscript is submitted to the Pattern Recognition journal [48].

- **Probabilistic Representation Learning for Uncertainty-Aware 3D Perception**

Rotation estimation is a fundamental yet inherently challenging task in computer vision and robotics, particularly under conditions of uncertainty introduced by object symmetries, occlusions, and ambiguous viewpoints. To tackle these challenges, we introduce BRVSNet (Bingham Rotation and View Synthesis Network), a unified self-supervised framework that jointly performs rotation regression and multi-view generation. BRVSNet leverages a Bingham mixture model over quaternions to represent rotations probabilistically, effectively capturing directional uncertainty in a compact and differentiable manner. At the heart of the framework is a Bingham mixture latent space, which enables two core functionalities: (1) a discriminator for self-supervised rotation learning, and (2) a multi-view generator guided by the highest-weight Bingham component, enabling the synthesis of consistent metadata conditioned solely on rotation quaternions. To enable joint training, we introduce a novel Relaxed Winner-Take-All Evidence Lower Bound (RWT-ELBO) loss that balances rotation regression with self-supervised multi-view consistency. Additionally, we propose two auxiliary weight regularization losses to enhance mixture component selection and improve representation quality. Extensive experiments across multiple datasets demonstrate that BRVSNet significantly outperforms existing baselines in both rotation estimation accuracy and multi-view generation consistency, highlighting the benefits

of integrating probabilistic rotation modeling within a generative, self-supervised learning paradigm. The source code of BRVSNet is publicly available at: <https://github.com/jiaxunguo/BRVSNet>. This work is submitted to the IEEE Transactions on Pattern Analysis and Machine Intelligence [49].

- **Global Pose-Aware Attention for Rotation-Invariant 3D Learning**

Recent advances in rotation-invariant (RI) learning for 3D point clouds typically replace raw coordinates with handcrafted RI features to ensure robustness under arbitrary rotations. However, these approaches often suffer from the loss of global pose information, making them incapable of distinguishing geometrically similar but spatially distinct structures. We identify that this limitation stems from the restricted receptive field in existing RI methods, leading to *Wing-tip feature collapse*, a failure to differentiate symmetric components (e.g., left and right airplane wings) due to indistinguishable local geometries. To overcome this challenge, we introduce the Shadow-informed Pose Feature (SiPF), which augments local RI descriptors with a globally consistent reference point (referred to as the “shadow”) derived from a learned shared rotation. This mechanism enables the model to preserve global pose awareness while maintaining rotation invariance. We further propose Rotation-invariant Attention Convolution (RIAttnConv), an attention-based operator that integrates SiPFs into the feature aggregation process, thereby enhancing the model’s capacity to distinguish structurally similar components. Additionally, we design a task-adaptive shadow locating module based on the Bingham distribution over unit quaternions, which dynamically learns the optimal global rotation for constructing consistent shadows. Extensive experiments on 3D classification and part segmentation benchmarks demonstrate that our approach substantially outperforms existing RI methods, particularly in tasks requiring fine-grained spatial discrimination under arbitrary rotations. The source code of BRVSNet is publicly available at: <https://github.com/jiaxunguo/EnRI-GAM>. This work is published in the Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence [50].

- **Ambiguity-Free Dynamic Global Perception via 4D Local Modeling**

Rotation invariance remains a core challenge in point cloud analysis, where existing methods often struggle with structural ambiguities and insufficient global context. Most rotation-invariant (RI) representations are derived from local coordinate systems, which inherently suffer from point-pair ambiguities and fail to capture discriminative features in symmetric or repetitive structures, while discarding

informative global pose cues. To overcome these limitations, we propose Ga4DPF, a novel framework that offers a robust, global-aware RI representation by converting rotation-equivariant geometric representations into invariant ones, while concurrently integrating global pose awareness. Specifically, Ga4DPF introduces a learnable steerable transform that equivariantly lifts point clouds into 4D space, facilitating robust local feature construction and mitigating point-pair ambiguities. Concurrently, we model a dynamic global pose reference using the Bingham distribution, which adaptively estimates a consistent global rotation and enhances global feature discriminability. Extensive experiments on multiple benchmark datasets demonstrate that Ga4DPF achieves state-of-the-art performance with high computational efficiency, offering a new paradigm for rotation-invariant point cloud analysis. This work has been accepted by the IEEE/CVF Conference on Computer Vision and Pattern Recognition [51].

1.4 Thesis Overview

This section outlines the organization of the thesis and provides a brief overview of each chapter.

- In Chapter 2, we propose GamMM-VAE, a novel deep generative clustering framework that replaces traditional symmetric Gaussian priors with an asymmetric Gamma mixture prior. To enable stable end-to-end training, we develop a new reparameterization technique that establishes a mathematical "bridge" between Normal and Gamma distributions, effectively handling the skewness and non-negativity of physical data features while mitigating gradient vanishing issues.
- In Chapter 3, to model directional and axially symmetric data, we introduce VIBinMM, a scalable variational inference framework for Bingham mixture models. By exploiting the mathematical relationship between Bingham and zero-mean Gaussian distributions, this framework bypasses the computation of intractable normalization constants, enabling efficient posterior approximation in high-dimensional directional spaces.
- In Chapter 4, we develop BRVSNet, a self-supervised framework that utilizes structured directional features for robust 3D perception. By integrating the Bingham distribution as a structured prior, the model achieves consistent multi-view generation and quantifies rotational uncertainty, offering a principled alternative to deterministic point estimates.

- In Chapter 5, we address the "feature collapse" phenomenon in 3D perception by proposing mechanisms that preserve strict rotation invariance while incorporating global pose awareness. Through the integration of attention mechanisms and global spatial context, the proposed model resolves geometric ambiguities in symmetric structures that local invariant descriptors cannot distinguish.
- In Chapter 6, We propose Ga4DPF, a novel framework that achieves robust, globally-aware rotation-invariant (RI) representations by equivariantly lifting point clouds into 4D space. This framework introduces a learnable steerable transform to facilitate robust local feature construction, effectively mitigating the point-pair ambiguities inherent in 3D local coordinate systems. Concurrently, we leverage a sampling-based approximation and geometric mean optimization of the Bingham distribution to model a dynamic global pose reference, which adaptively estimates a consistent global rotation and enhances feature discriminability in symmetric or repetitive structures.
- In Chapter 7, we conclude this thesis, summarizing the main findings and implications of our research. Additionally, we identify and discuss open research directions that can serve as valuable avenues for future work in the field.

Chapter 2

Deep Clustering Analysis via Variational Autoencoder

2.1 Introduction

Clustering is an unsupervised machine learning method that groups similar data points into clusters. It has been widely used in pattern recognition, data mining, and other fields [6, 7]. In the past decades, various clustering methods have been proposed, including K-means [52] and a class of clustering methods based on mixture models [53, 54, 55, 14, 56, 57, 58]. Gaussian mixture model (GMM) is the most well known among these mixture model-based clustering methods. Although the mentioned methods have proven to be effective in various fields, they only establish a direct relationship with the original data when describing the data and lack the ability to automatically learn more representative features of the data on the latent space.

With the development of neural networks and deep learning, data representation learning techniques have matured. The unique structure of neural networks enables deep learning models to effectively decompose the direct dependencies of the original data and map its features to a low-dimensional space for representation learning [2, 3, 4, 5]. In recent years, some deep generative clustering models have been proposed and proven to be effective, such as generative adversarial networks (GANs) [13, 59, 60] and variational autoencoders [12, 61, 62, 63]. They are effective in generating corresponding fake data for different clusters and also have the potential ability to identify similar points and aggregate them into clusters. GAN

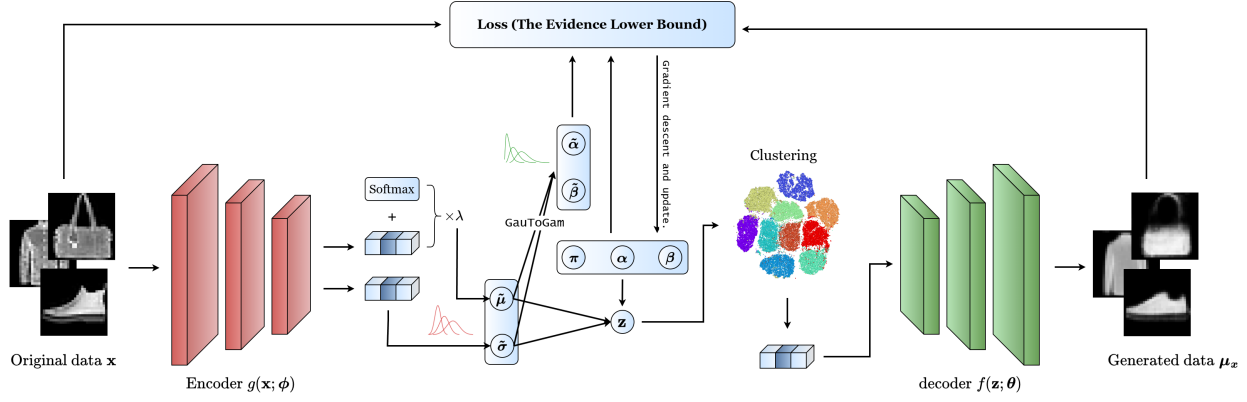


Figure 2.1: The Flowchart of GamMM-VAE. The data generative process is as follows: 1) Encoder $g(\mathbf{x}; \phi)$ gets the original data $\mathbf{x}$ and outputs the parameters $\tilde{\mu}$ and $\tilde{\sigma}$ of the adjusted Gaussian mixture model. 2) Latent embedding $\mathbf{z}$ is obtained by reparameterization trick and indirectly affected by the parameter π, α, β during backpropagation. 3) Decoder $f(\mathbf{z}; \theta)$ is responsible for the mapping of latent embedding $\mathbf{z}$ to generated data μ_x . 4) The difference between $\mathbf{x}$ and μ_x , the parameters $\tilde{\alpha}$ and $\tilde{\beta}$ of the Gamma mixture model, and the initial parameters π, α , and β jointly construct the evidence lower bound as a loss to continuously optimize the model and update π, α , and β during training.

achieves generative capabilities by training the generators and discriminators to continuously adversely improve the authenticity of the fake data generated by the generators and the ability of the discriminators to distinguish between real and fake data. VAE is a type of autoencoder (AE) neural network that consists of an encoder, which is responsible for properly downscaling the original data into the latent space and extracting the representation, and a decoder, which is responsible for reconstructing the original input data. But one thing that distinguishes VAE from AE is the regularized coding distribution to ensure that a good characterization of the latent space is used to generate new data. The specification of the latent space is usually to encode the reduced-dimensional data as some probability distribution so that the regularized latent space can also mitigate overfitting.

Many VAE-based deep generative clustering models assume that the hidden embedding follows a Gaussian distribution or a Gaussian Mixture Model (GMM) due to its mathematical tractability. However, the symmetry of the Gaussian probability density function requires GMMs to use more components to fit asymmetric distributions and the tails of skewed distributions adequately. This necessity leads to a more complex structure in GMMs and poses significant challenges in interpreting asymmetric data. To address these challenges, recent research has explored non-Gaussian mixture models, including the Beta mixture model [14], the Dirichlet mixture model [15, 16, 17], and the Student's t mixture model [64, 65]. Among these, the

Dirichlet distribution has been successfully integrated into VAEs with some modifications, yielding impressive experimental results [35]. Clearly, the choice of the latent space mixture model’s density distribution plays a crucial role in determining the flexibility and robustness of the VAE.

Despite these advancements, a significant research gap remains in effectively modelling univariate and multivariate positive vectors within VAEs. The Gamma Mixture Model (GamMM) [66] has emerged as a promising non-Gaussian mixture model capable of addressing this gap. GamMMs have shown effectiveness in various applications, such as analyzing capacity and error probabilities [36] and object shape clustering [67]. However, their integration into VAE frameworks for clustering tasks has not been thoroughly explored. This chapter proposes a novel deep generative clustering model based on VAE and GamMM latent embeddings to fill this gap. The proposed model offers several advantages. Firstly, utilizing GamMM as a prior for the latent space distribution provides a more flexible and efficient embedding, enhancing clustering performance compared to GMM-based VAEs [68, 69]. Secondly, the Gamma distribution’s property of strictly positive values allows the use of the Softmax function within the network structure to regulate intermediate values, thereby mitigating overfitting. Detailed design ideas for this approach are discussed in Section 2.3.2. In summary, this research aims to address the limitations of existing VAE-based clustering models by integrating GamMMs, thereby improving clustering performance and providing a robust framework for modelling positive vectors in latent spaces.

The main contributions of this work can be summarized below.

- (1) We propose a novel deep generative clustering model based on VAE with GamMM latent embedding called GamMM-VAE.
- (2) We construct GamMM-based evidence lower bound (ELBO) using two effective approximations of the integral forms.
- (3) We introduce a reparameterization trick of Gamma distribution based on the approximation of Gaussian distribution to Gamma distribution. It is shown to be reliable and valid in VAE.

The flowchart of GamMM-VAE framework is illustrated in Figure 2.1. The rest of this chapter is organized as follows. Some related works are presented in Section 2.2. Section 2.3 describes the proposed deep generative model. In Section 2.4, we provide a detailed description and discussion of experimental results. Finally, we conclude this chapter in Section 2.5.

2.2 Related Works

The development of neural networks and representations learning algorithms has made it possible for many clustering and classification models to give convincing results based on the learned feature information without specialized knowledge. For example, Xie et al. [70] proposed a classical deep clustering model called deep embedding clustering (DEC), which first iteratively learns how to reduce the observation space to a low-dimensional hidden space, after which it continuously optimizes the soft assignment between the embedded points and the cluster centroid as well as refines the cluster centroids using KL divergence as the convergence criterion. The process is repeated until the criterion is satisfied. Since the clustering loss defined in DEC may corrupt the feature space, resulting in non-representative meaningless features, Guo et al. [71] proposed an improved version of DEC called IDEC, which uses clustering loss to scatter the feature space data points and maintains local structure of data generating distribution using an under-complete autoencoder. DEC and IDEC achieve good performance on clustering tasks, however, their feature embeddings are designed for clustering and lack the ability to extend the models to generative samples.

Deep generative models have seen impressive developments in recent years. VAE, GAN, and Diffusion models are three of the most successful deep generative frameworks, and all are based on unsupervised generative processes. Diffusion models are powerful in their generative ability but face several challenges in clustering. These models struggle to identify patterns and regularities in low-quality data, which prevents their ability from generalizing to new datasets or scenarios [72]. When working with large-scale datasets, they face significant computational challenges, including long training times, high memory usage, and difficulties in achieving convergence. Additionally, uneven or biased data sampling can limit their effectiveness in producing adaptable and diverse outputs [72]. These challenges collectively impact the effectiveness of diffusion models in clustering tasks. On the other hand, VAE and GAN are able to capture data features for clustering and generate unseen data through neural networks. They are currently being continuously researched for extensions and enhancements, and have achieved impressive experimental results in a wide range of fields [73, 74, 75, 76, 77, 78].

Chen et al. [79] extended the information theory of generative adversarial networks to propose InfoGAN, which is able to learn disentangled representations in a completely unsupervised approach. It maximizes the mutual information between a small fraction of latent embeddings and observations, and continuously enhances the model by optimizing a lower bound on the mutual information objective. InfoGAN can find

latent information such as writing style in handwritten data and hairstyle and emotion in face data. ClusterGAN [80] was proposed by Mukherjee et al. to tackle the clustering task, where they solve the problem by sampling latent embeddings from a mixture of hot-coded variables and continuous latent embeddings, coupled with an inverse network that projects the data into the latent space to achieve clustering. The results show that ClusterGAN can maintain potential space interpolation across categories and demonstrate excellent performance in comparison with various clustering baselines.

Jiang et al. [81] proposed a variational depth embedding (VaDE) technique, which is an extension of VAE that uses GMM as a prior to clustering the latent space and uses the SGVB estimator and reparameterization trick to optimize the ELBO. VaDE demonstrated excellent clustering accuracy and the ability to generate highly realistic samples on six benchmark datasets. However, since the GMM in VaDE fits individual dimensions separately using independent Gaussian distributions, it cannot be considered as a true GMM for data generation. Dilokthanakul et al. [82] proposed a Gaussian mixture VAE (GMVAE) with a prior distribution of the GMM depending on two latent spaces. This approach allows GMM to capture a clustering representation of non-single-peaked data, which has advantages over conventional VAE. Ojo and Bouguila [77] present a novel generalized Dirichlet variational Autoencoder (GD-VAE) for topic modelling and image classification. The deployed generalized Dirichlet distribution allows VAE for the capture of both positively and negatively correlated topics. An elastic deep autoencoder for text embedding clustering (EDA-TEC) is proposed by Daneshfar et al. [78]. It combines Frobenius and $L_{2,1}$ -norms into an elastic loss function to handle Gaussian and Laplacian noise effectively. Additionally, it employs an improved graph regularization term based on weighted cosine similarity for better geometric data structure maintenance and a sparsity regularization term to enhance clustering results.

While these models have shown promising results, they primarily rely on Gaussian distributions or GMMs, which can struggle with asymmetric data distributions. This study builds on this body of work by introducing a Gamma Mixture Model (GamMM) as the prior for the latent space in a VAE framework. By leveraging the strictly positive values and flexibility of the Gamma distribution, the proposed model aims to overcome the limitations of existing VAE-based clustering models, offering improved performance and robustness in clustering tasks. Despite the use of non-Gaussian and asymmetric distribution methods in Ojo and Bouguila’s Generalized Dirichlet VAE (GD-VAE) [77] and Daneshfar et al.’s EDA-TEC [78], our approach provides a different yet complementary direction. By introducing GamMM as the prior in a VAE framework, we leverage the strictly positive values and flexibility of the Gamma distribution to enhance

clustering performance and robustness, particularly in handling positive vectors in latent spaces.

We will compare the proposed model with most of the models mentioned in this section to demonstrate the advanced clustering and generation capabilities of our model in the experimental results section.

2.3 Proposed Model

In this section, we propose a novel deep generative clustering model based on VAE in which a GamMM is the prior of latent embeddings. We also introduce a reparameterization trick that approximates the Gamma distribution by Gaussian distribution.

2.3.1 Objective Function

Due to the requirement for appropriate loss functions in deep neural network models for optimization through backpropagation, this section begins with the Gamma mixture model, in conjunction with the network structure shown in Figure 2.1, progressively derives the objective optimization function.

Gamma Mixture Model

Given a H-dimensional data set $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with N samples, assume that the generation of $\mathcal{X}$ follows some stochastic process influenced by a peer D-dimensional set of variables $\mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ where $\mathbf{z}_i = (z_{i1}, \dots, z_{iD})$. And this set of random variables $\mathcal{Z}$ acts as a latent embedding of $\mathcal{X}$ and follows a mixture of K multivariate Gamma (mGam) distributions. The probability density function of the GamMM is

$$p(\mathbf{z}_i | \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{k=1}^K \pi_k \mathbf{Gam}(\mathbf{z}_i | \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \quad (16)$$

where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K\}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$, $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are the parameters of the mGam distribution representing component k , where $\boldsymbol{\alpha}_k = (\alpha_{k1}, \dots, \alpha_{kD})$, $\boldsymbol{\beta}_k = (\beta_{k1}, \dots, \beta_{kD})$. $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ provides the mixing coefficients with the constraints that they are not negative and sum to one. $\mathbf{Gam}(\mathbf{z}_i | \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ is a mGam distribution representing component k with parameters $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ and is defined using the following structure with conditionally independent features

$$\begin{aligned}
\mathbf{Gam}(\mathbf{z}_i|\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) &= \prod_{d=1}^D \mathbf{Gam}(z_{id}|\alpha_{kd}, \beta_{kd}) \\
&= \prod_{d=1}^D \frac{\beta_{kd}^{\alpha_{kd}}}{\Gamma(\alpha_{kd})} z_{id}^{\alpha_{kd}-1} \exp(-\beta_{kd} z_{id}),
\end{aligned} \tag{17}$$

where $z_{id} \in (0, \infty)$, α_{kd} and β_{kd} are the shape and rate parameters of Gamma distribution and are strictly positive.

The Generative Process

The probability distribution of the clustering indicator k is generated based on a categorical distribution parameterized by $\boldsymbol{\pi}$:

$$p(k) = \mathbf{Cat}(k|\boldsymbol{\pi}). \tag{18}$$

The latent embedding $\mathbf{z}$ corresponding to the cluster k can be sampled from the mGam distribution using the known $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and the corresponding conditional probability is

$$p(\mathbf{z}|k) = \mathbf{Gam}(\mathbf{z}|\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k). \tag{19}$$

A sample $\mathbf{x}$ from the latent embedding $\mathbf{z}$ is generated according to $f(\mathbf{z}; \boldsymbol{\theta})$ which can be considered as a decoder neural network whose input is $\mathbf{z}$ and parameterized by $\boldsymbol{\theta}$. The probability density function is

$$p(\mathbf{x}|\mathbf{z}) = \mathbf{Gam}(\mathbf{x}|\boldsymbol{\alpha}_x, \boldsymbol{\beta}_x), \tag{20}$$

where $[\log \boldsymbol{\alpha}_x, \log \boldsymbol{\beta}_x] = f(\mathbf{z}; \boldsymbol{\theta})$. In the implementation, $\boldsymbol{\alpha}_x$ and $\boldsymbol{\beta}_x$ will not be output as middle parameters, and the decoder $f(\mathbf{z}; \boldsymbol{\theta})$ will directly output $\mathbf{x}$ that obeys $\mathbf{Gam}(\mathbf{x}|\boldsymbol{\alpha}_x, \boldsymbol{\beta}_x)$, which is denoted by $\boldsymbol{\mu}_x^{(l)}$ in the ELBO expression below.

Based on the above generative process, the joint probability $p(\mathbf{x}, \mathbf{z}, k)$ can be given as:

$$p(\mathbf{x}, \mathbf{z}, k) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|k)p(k), \tag{21}$$

where $\mathbf{x}$ and k are independent conditioned on $\mathbf{z}$.

Variational Lower Bound

In order to generate more reliable data using the model, we need to maximize the likelihood for the given point set to optimize the parameters. According to Jensen's inequality, the log-likelihood equation for the above joint probability can be written as

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int_{\mathbf{z}} \sum_k p(\mathbf{x}, \mathbf{z}, k) d\mathbf{z} \\ &\geq \left\langle \log \frac{p(\mathbf{x}, \mathbf{z}, k)}{q(\mathbf{z}, k | \mathbf{x})} \right\rangle_{q(\mathbf{z}, k | \mathbf{x})} = \mathcal{L}_{\text{ELBO}}(\mathbf{x}),\end{aligned}\tag{22}$$

where $\langle \cdot \rangle$ is the calculation of expectation, $\mathcal{L}_{\text{ELBO}}(\mathbf{x})$ denotes the ELBO of the marginal log likelihood. $q(\mathbf{z}, k | \mathbf{x})$ represents the variational posterior distribution as an approximation to the true posterior. Because of the common use of the mean-field assumption in variational Bayesian inference [83], $q(\mathbf{z}, k | \mathbf{x})$ can be decomposed as

$$q(\mathbf{z}, k | \mathbf{x}) = q(\mathbf{z} | \mathbf{x})q(k | \mathbf{x}),\tag{23}$$

where $q(\mathbf{z} | \mathbf{x})$ can be considered as a mGam distribution:

$$q(\mathbf{z} | \mathbf{x}) = \mathbf{Gam}(\mathbf{z} | \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}),\tag{24}$$

where $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ are obtained from an encoder neural network $g(\mathbf{x}; \phi)$ with input $\mathbf{x}$ and the parameters ϕ as

$$[\log \tilde{\boldsymbol{\alpha}}, \log \tilde{\boldsymbol{\beta}}] = g(\mathbf{x}; \phi).\tag{25}$$

Substituting Equation 21 and Equation 23 into the $\mathcal{L}_{\text{ELBO}}(\mathbf{x})$ of Equation 22, we have

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{x}) &= \left\langle \log \frac{p(\mathbf{x}, \mathbf{z}, k)}{q(\mathbf{z}, k | \mathbf{x})} \right\rangle_{q(\mathbf{z}, k | \mathbf{x})} \\ &= \langle \log p(\mathbf{x}, \mathbf{z}, k) - \log q(\mathbf{z}, k | \mathbf{x}) \rangle_{q(\mathbf{z}, k | \mathbf{x})} \\ &= \langle \log p(\mathbf{x} | \mathbf{z}) + \log p(\mathbf{z} | k) + \log p(k) \\ &\quad - \log q(\mathbf{z} | \mathbf{x}) - \log q(k | \mathbf{x}) \rangle_{q(\mathbf{z}, k | \mathbf{x})}.\end{aligned}\tag{26}$$

For flexibility and convenience in the ELBO optimization process, $\langle \log p(\mathbf{x} | \mathbf{z}) \rangle_{q(\mathbf{z}, k | \mathbf{x})}$ will use the SGVB estimator to make approximations. This helps the neural network to update parameters through gradient descent. For ease of expression, we assume that $\mathbf{x}$ follows Bernoulli distribution. $\langle \log p(\mathbf{x} | \mathbf{z}) \rangle_{q(\mathbf{z}, k | \mathbf{x})}$

approximated by the SGVB estimator can be written as

$$\begin{aligned}\langle \log p(\mathbf{x}|\mathbf{z}) \rangle_{q(\mathbf{z},k|\mathbf{x})} &= \frac{1}{L} \sum_{l=1}^L \langle \log p(\mathbf{x}|\mathbf{z}^{(l)}) \rangle \\ &= \frac{1}{L} \sum_{l=1}^L \sum_{h=1}^H x_h \log \mu_{x_h}^{(l)} + (1 - x_h) \log (1 - \mu_{x_h}^{(l)}),\end{aligned}\tag{27}$$

where L is the number of Monte Carlo samples in the SGVB estimator. H is the dimensionality of $\mathbf{x}$. $\mu_x^{(l)}$ can be the output of decoder neural network $f(\mathbf{z}^{(l)}; \theta)$ after using the SGVB estimator. By substituting the terms in Equation 26 with Equation 18, Equation 19, Equation 20 and Equation 24, the $\mathcal{L}_{\text{ELBO}}(\mathbf{x})$ can be given by

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{x}) &= \frac{1}{L} \sum_{l=1}^L \sum_{h=1}^H x_h \log \mu_{x_h}^{(l)} + (1 - x_h) \log (1 - \mu_{x_h}^{(l)}) \\ &\quad + \sum_{k=1}^K \gamma_k \sum_{d=1}^D \left[\log \left(\frac{\beta_{kd}^{\alpha_{kd}}}{\Gamma(\alpha_{kd})} \right) - \frac{\Gamma(\tilde{\alpha}_d + 1)}{\Gamma(\tilde{\alpha}_d)} \frac{\beta_{kd}}{\tilde{\beta}_d} \right. \\ &\quad \quad \left. + (\alpha_{kd} - 1) \left(\psi(\tilde{\alpha}_d) - \log(\tilde{\beta}_d) \right) \right] \\ &\quad - \sum_{d=1}^D \left[\log \left(\frac{\tilde{\beta}_d^{\tilde{\alpha}_d}}{\Gamma(\tilde{\alpha}_d)} \right) - \frac{\Gamma(\tilde{\alpha}_d + 1)}{\Gamma(\tilde{\alpha}_d)} \right. \\ &\quad \quad \left. + (\tilde{\alpha}_d - 1) \left(\psi(\tilde{\alpha}_d) - \log(\tilde{\beta}_d) \right) \right] \\ &\quad + \sum_{k=1}^K \gamma_k \log \frac{\pi_k}{\gamma_k},\end{aligned}\tag{28}$$

where D is the dimensionality of $\mathbf{z}$ and the corresponding $\tilde{\alpha}$ and $\tilde{\beta}$. π_k is the prior probability of cluster k representing $p(k)$. γ_k denotes $q(k|\mathbf{x})$ for simplicity and can be approximated by the SGVB estimator and reparameterization trick as follows:

$$q(k|\mathbf{x}) = \langle \log q(k|\mathbf{z}) \rangle_{q(\mathbf{z}|\mathbf{x})} = \frac{1}{L} \sum_{l=1}^L \frac{p(k)p(\mathbf{z}^{(l)}|k)}{\sum_{k'=1}^K p(k')p(\mathbf{z}^{(l)}|k')}.\tag{29}$$

2.3.2 Reparameterization trick and approximation on Gamma distribution

The reparameterization trick is an important step in VAE. In the previous subsection, we mentioned that $\mu_x^{(l)}$ in Equation 27 can be directly obtained from decoder neural network $f(\mathbf{z}^{(l)}; \theta)$ after using the SGVB estimator. While $\mathbf{z}^{(l)}$ of $f(\mathbf{z}^{(l)}; \theta)$ need to be sampled from the mGam distribution in Equation 24,

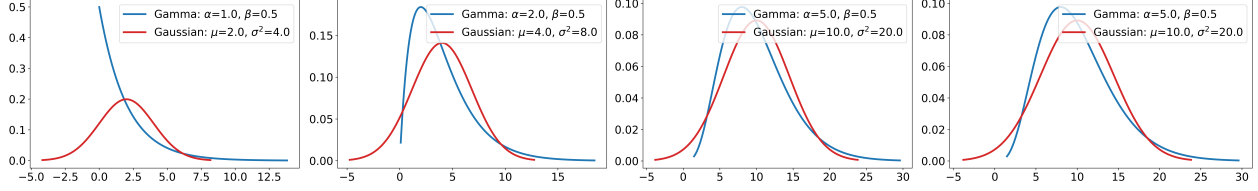


Figure 2.2: Approximation from Gamma to Gaussian.

as $\mathbf{z}^{(l)} \sim \mathbf{Gam}(\tilde{\alpha}^{(l)}, \tilde{\beta}^{(l)})$. If using the traditional sampling methods, this sampling operation will be non-derivable in the gradient descent process of the neural network, which fails to update the gradient and parameters. The solution to this sampling problem in VAE is the reparameterization trick. As an example, assume that the latent embedding $\mathbf{z}$ is sampled from a normal distribution, i.e., $\mathbf{z}^{(l)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I})$. According to the reparameterization trick, $\mathbf{z}^{(l)}$ is obtained by

$$\mathbf{z}^{(l)} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\sigma}} \circ \boldsymbol{\epsilon}^{(l)}, \quad (30)$$

where $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbf{I})$, $\circ$ is the operator of element-wise multiplication, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}$ are the mean and variance derived by a corresponding encoder like Equation 25.

The nature of the reparameterization trick can be understood as making the randomness of the parameter ϕ in $g(\mathbf{x}; \phi)$ transfer to the meta-variables of the variational posterior distribution, and the model needs to learn how to adjust the meta variables. A reparameterization trick [84] based on rejection sampling of the Gamma distribution was proposed for the gradient reparameterization. However, the latent embeddings that are reparameterized using the parameters from the encoder output are ineffective in contributing to the model clustering after using the Gamma reparameterization trick in our model.

The Gamma distribution has two main parameters, the shape parameter α and the rate parameter β , which are also the outputs of the encoder. Based on the analysis of the Gamma distribution reparameterization trick, we finally attribute the reason for the failure to a division operation of the reparameterization. Depending on the method, the shape of the distribution is firstly respecified based on the parameter α . Then, the distribution needs to be scaled to the appropriate size by a division step based on the rate parameter. However, the division operation results in a vanishing gradient problem because the numerator and denominator are both outputs of the same encoder, and thus the division erased the gradient of the encoder. Finally, the model cannot optimize the parameters based on gradient descent.

The reparameterization trick (Equation 30) for the normal distribution has been used in many models

and shown to be reliable. So we still rely on Equation 30 for the reparameterization problem in this work and find a bridge to connect the normal and Gamma distributions. The relationship between the normal and Gamma distributions has been presented in [37] as follows

$$\mu = \frac{\alpha}{\beta}, \sigma^2 = \frac{\alpha}{\beta^2}, \alpha \rightarrow \infty, \quad (31)$$

where μ and σ^2 are the mean and variance of normal distribution, α and β are the shape and rate of Gamma distribution. Because the support of normal distribution is the set of real numbers, while Gamma distribution is greater than 0. So strictly speaking, this transformation can only be performed from Gamma to normal distribution. Moreover, as shown in Figure 2.2, this approximation becomes more accurate as α tends to infinity.

Because of the use of the normal distribution reparameterization trick, our model has to approximate Gamma distribution from normal distribution. This requires some restrictions on the proposed approximation method (Equation 31). We propose the following improvement methods:

- (1) The μ of the encoder output is normalized to between [0,1] using the *softmax* function. This ensures that the support of the resulting normal distribution is a subset of the Gamma distribution, satisfying the prerequisites for approximation from normal to Gamma distribution.
- (2) With σ^2 unchanged, the μ normalized by the *softmax* function is appropriately scaled using λ . According to Equation 31, this step magnifies the α to some extent, leading to a more accurate approximation. However, a larger λ doesn't represent better results because it will erase the influence of variance σ^2 , which is an important output of the encoder and the reparameterization trick. Hence, λ needs to be chosen carefully as an initialization parameter.

Finally, our proposed approximation from normal to Gamma distribution is as follows:

$$\begin{aligned} \alpha &= \frac{(\text{softmax}(\mu) * \lambda)^2}{\sigma^2}, \\ \beta &= \frac{\text{softmax}(\mu) * \lambda}{\sigma^2}. \end{aligned} \quad (32)$$

2.3.3 Hyper-parameter learning and choosing strategy of GamMM-VAE

During the hyperparameter learning, some of the processes in the first subsection need to be changed since we use the reparameterization trick of normal distribution to approximate Gamma distribution.

- (1) First of all, after simplifying Equation 32, the learning of Gamma distribution hyper-parameters should be divided into the following two parts:

$$\begin{aligned} [\tilde{\boldsymbol{\mu}}, \log \tilde{\boldsymbol{\sigma}}^2] &= g(\mathbf{x}; \boldsymbol{\phi}), \\ [\log \tilde{\boldsymbol{\alpha}}, \log \tilde{\boldsymbol{\beta}}] &= \mathbf{GauToGam}(\tilde{\boldsymbol{\mu}}, \log \tilde{\boldsymbol{\sigma}}^2, \lambda), \end{aligned} \quad (33)$$

where $\tilde{\boldsymbol{\mu}}$ and $\log \tilde{\boldsymbol{\sigma}}^2$ are the mean and log-variance of normal distribution directly obtained from the encoder neural network. λ is the parameter which needs to be chosen at the beginning. $\log \tilde{\boldsymbol{\alpha}}$ and $\log \tilde{\boldsymbol{\beta}}$ are the log-shape and log-rate of Gamma distribution.

- (2) $q(\mathbf{z}|\mathbf{x})$ in Equation 24 needs to be replaced using the normal distribution reparameterization trick (Equation 30), and the input parameters will be the normal distribution parameters obtained in the first step of Equation 33. The normal distribution reparameterization is a bridge to obtain the corresponding Gamma parameters. The latent representation is a normal distribution in the forward of the model. However, when the approximation is sufficient, the latent representation can be optimized to a Gamma distribution during back-propagation because $\mathcal{L}_{\text{ELBO}}(\mathbf{x})$ is built based on Gamma distribution.

The hyper-parameter adjustment coefficient λ needs to be chosen in a way that the $\tilde{\boldsymbol{\alpha}}$ of Equation 32 is as large as possible to improve the accuracy of the approximation. However, a too large $\tilde{\boldsymbol{\alpha}}$ will also lead to a large exponential calculation in $\mathcal{L}_{\text{ELBO}}(\mathbf{x})$ (Equation 28), which in turn makes NaN in the gradient calculation. Therefore, the proper selection of lambda is an important step in the model. In the experiments shown of this work, λ was selected empirically. Usually, the λ that leads to the best results appears in [20, 40].

The training steps of the proposed deep generative model are shown in Algorithm 1.

2.4 Experimental Results

In this section, the proposed model will be compared with other state-of-the-art models using six publicly available benchmark datasets. The performance of the model is shown in terms of clustering ability, generative ability, and several other aspects with figures and tables on different metrics.

Algorithm 1 Training Steps of Proposed Model

Input:

- $\mathcal{X}$ - training data.
- H - the dimension of $\mathbf{x}$.
- D - the dimension of $\mathbf{z}$.
- lr - learning rate.
- λ - Adjustment coefficient.

Output:

- ϕ - parameters of Encoder.
 - θ - parameters of Decoder.
 - π, α, β - parameters of the multivariate Gamma mixture model.
 - 1: Initialization ϕ and θ using a pretrained stacked autoencoder.
 - 2: Initialization π, α and β using a pretrained GamMM.
 - 3: **while** not converged **do**
 - 4: Sample a data $\mathbf{x}$ from $\mathcal{X}$.
 - 5: $[\tilde{\mu}, \log \tilde{\sigma}^2] \leftarrow g(\mathbf{x}; \phi)$
 - 6: $\tilde{\mu} = \tilde{\mu} \times \lambda$
 - 7: $[\log \tilde{\alpha}, \log \tilde{\beta}] \leftarrow \text{GauToGam}(\tilde{\mu}, \log \tilde{\sigma}^2)$
 - 8: $\mathbf{z} \leftarrow \text{reparameterization}(\tilde{\mu}, \tilde{\sigma})$
 - 9: Generated data $\mu_x \leftarrow f(\mathbf{z}; \theta)$
 - 10: $\mathcal{L}_{\text{ELBO}}(\mathbf{x}) \leftarrow \left\langle \log \frac{p(\mathbf{x}, \mathbf{z}, k)}{q(\mathbf{z}, k | \mathbf{x})} \right\rangle_{q(\mathbf{z}, k | \mathbf{x})}$ (Equation 28)
 - 11: Update the parameters of the decoder, encoder and GamMM.
 - 12: $\phi, \theta, \pi, \alpha, \beta \leftarrow \text{Adam}(\mathcal{L}_{\text{ELBO}}(\mathbf{x}), \text{lr})$
 - 13: **end while**
-

2.4.1 Datasets and Experimental Settings

Six benchmark datasets were selected to examine the performance of the proposed models. The datasets were chosen considering their popularity and the datasets used in the experiments of the articles proposing the comparison models. The structures of the encoder ($g(\mathbf{x}; \phi)$) and decoder ($f(\mathbf{z}; \theta)$) neural networks were fixed as input-500-500-2000-hidden(10) and hidden(10)-500-500-2000-output in experiments. All layers are fully connected with ReLU activation function. Adam optimizer is used to maximize the $\mathcal{L}_{\text{ELBO}}(\mathbf{x})$ in Equation 28. Hyperparameters such as batch size and learning rate varied depending on the dataset. The descriptions and experimental settings of the selected datasets MNIST [85], FATHION-MNIST (F-MNIST) [86], Heterogeneity Human Activity Recognition (HHAR) [87], REUTERS-10K [88], REUTERS [88] and WEBKB [89] are as follows:

- (1) **MNIST:** MNIST is a famous image classification and clustering dataset with a collection of 70,000 handwritten digital 28*28 pixel images for ten categories from 0 to 9. Since we use a fully connected network to train the model, each image is flattened and reshaped to 784 dimensions before training.

During training, the mini-batch size is chosen to be 800 and the learning rate is kept at the default 0.001.

- (2) **F-MNIST:** F-MNIST is a dataset of Zalando’s article images and is similar to MNIST. There are 70000 images in total from 10 different categories, each in 28*28 grayscale. However, F-MNIST is dominated by shoes, backpacks and clothing. This dataset is preprocessed in the same way as MNIST and needs to be reshaped to 784 dimensions. The mini-batch size is also 800, but the learning rate needs to be adjusted to 0.00001.
- (3) **HHAR:** There are 10299 sensor records about human activity in this dataset collected from smart-phones and smartwatches. All data are 561-dimensional and classified into six categories ‘Biking’, ‘Sitting’, ‘Standing’, ‘Walking’, ‘Stair Up’ and ‘Stair down’. The mini-batch size and learning rate are 100 and 0.0001, respectively.
- (4) **REUTERS:** REUTERS is a dataset of English news stories. For comparison, we used the same method as [90] to retain only the four root categories corporate/industrial, government/social, markets, and economics, finally left with a dataset of 685071 articles. Also, the 2000 most frequent words in each article were used to compute tf-idf (term frequency–inverse document frequency) features instead of whole articles. The mini-batch size selection for this dataset is 800 and the learning rate is 0.0005.
- (5) **REUTERS-10K:** This dataset was obtained by randomly picking up 10,000 stories from REUTERS. Since some models cannot be run on the huge REUTERS full dataset, a subset is taken for model performance testing. On this dataset, the mini-batch size is also kept at 800 and the learning rate needs to be turned up to 0.001.
- (6) **WEBKB:** The WEBKB dataset consists of 8334 web pages collected by CMU groups in 1977. There are seven classes in the dataset: Course, Faculty, Student, Staff, Department, Project, and Other. However, we only include the clusters of Course, Faculty, Student, Project due to adequate sample sizes. As a result, the dataset consists of 4199 samples with 2000 selected features. The mini-batch size and learning rate for this dataset are 1000 and 0.002.

The main information of the six datasets has been summarized in Table 2.1. All experiments do not distinguish between training and testing sets, which are based on the experiments in [90] in order to directly

compare the experimental results. In addition, due to the variability of each dataset, the model needs to gradually decay the learning rate during training according to the actual situation, and the specific learning rate decay strategy can be referred to the provided code. A stacked autoencoder will pretrain the neural networks $g(\mathbf{x}; \phi)$ and $f(\mathbf{z}; \theta)$ at the beginning to provide better initialization. Moreover, at the end of pretraining, all training data are projected into the latent space $\mathbf{z}$, which is then fed to GamMM for training to obtain π, α, β .

Table 2.1: Summary of the Benchmark Datasets

Dataset	#Samples	#Clusters	Dimensions
MNIST	70000	10	1*28*28
F-MNIST	70000	10	1*28*28
HHAR	10299	6	561
REUTERS	685071	4	2000
REUTERS-10k	10000	4	2000
WEBKB	4199	4	2000

2.4.2 Clustering Results

First, we will analyze the clustering ability of the models based on their clustering results for the six benchmark datasets mentioned above. Three metrics, Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), are used to present the clustering results. While NMI and ARI are the classical clustering evaluation metrics, ACC can also demonstrate the classification ability of the proposed clustering model. We will also present the results of other clustering models on these three metrics to highlight the performance of our model. These clustering models include two classical clustering models, K-means [52] and GMM [53], and seven other advanced deep generative clustering methods, including DEC [70], IDEC [71], InfoGAN [79], ClusterGAN [80], VaDE [90], EDA-TEC [78] and GD-VAE [77].

The experimental results of all mentioned models on different datasets have been presented in Table 2.2 (ACC and NMI) and Table 2.3 (ARI) where the best results of different models on the same dataset and with the same metrics are marked in bold, the second and third best results are marked with '*' and '†', respectively. The clustering results of other models are mostly collected from published articles or obtained based on open-source code with the hyper-parameters mentioned in the articles. All the models are pre-trained if required. Those that could not get better results based on open source code and hyper-parameters

Table 2.2: Comparison of Clustering Performance on Benchmark Datasets in terms of ACC and NMI.

Model	MNIST		F-MNIST		HHAR		REUTERS		REUTERS-10k		WEBKB	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.532	0.500	0.474	0.512	0.600	0.589	0.515	0.309	0.538	0.405	0.567	0.358
GMM	0.540	0.509	0.523	0.409	0.659	0.659	0.535	0.337	0.547	0.410	0.589	0.356
DEC	0.863	0.834	0.595	0.510	0.820	0.734	0.756	0.481	0.743	0.497	0.623	0.362
IDEC	0.874	0.859	0.609	0.513	0.849	0.762	0.530	0.209	0.756	0.498	0.631	0.366
InfoGAN	0.870	0.840	0.610	0.590	0.816	0.690	0.766	0.518	0.753	0.500	0.633	0.295
ClusterGAN	0.950*	0.890*	0.630	0.640*	0.828	0.772 [†]	0.810 [†]	0.577*	0.795	0.553*	0.651	0.314
VaDE	0.944 [†]	0.891	0.629	0.611	0.845 [†]	-	0.794	0.541	0.798 [†]	0.529	0.648	0.376 [†]
EDA-TEC	0.952	0.871	0.644 [†]	0.659	0.789	0.820	-	-	0.741	0.558	0.653 [†]	0.400
GD-VAE	0.893	0.779	0.702*	0.599	0.851*	0.755	0.840	0.575 [†]	0.801*	0.503	0.662*	0.391*
GamMM-VAE	0.952	0.888 [†]	0.708	0.628 [†]	0.863	0.780*	0.828*	0.600	0.805	0.538 [†]	0.689	0.400

Table 2.3: Comparison of Clustering Performance on Benchmark Datasets in terms of ARI.

Model	MNIST	F-MNIST	HHAR	REUTERS	REUTERS-10k	WEBKB
	ARI	ARI	ARI	ARI	ARI	ARI
K-means	0.372	0.370	0.461	0.299	0.361	0.274
GMM	0.386	0.353	0.580	0.308	0.372	0.282
DEC	0.751	0.403	0.673	0.480	0.468	0.328
IDEC	0.773	0.410	0.710 [†]	-	0.466	0.336
InfoGAN	0.838	0.519	0.641	0.503	0.479	0.293
ClusterGAN	0.889 [†]	0.566	0.710 [†]	0.537 [†]	0.513	0.315
VaDE	0.899	0.533*	-	0.503	0.490	0.358 [†]
EDA-TEC	0.883	0.518	0.728*	-	0.562*	0.373
GD-VAE	0.783	0.503	0.709	0.629*	0.556 [†]	0.348
GamMM-VAE	0.898*	0.530 [†]	0.737	0.657	0.605	0.369*

are marked with '-', but the published articles of the corresponding models have given other results that can be used as references.

It can be seen from Tables 2.2 and 2.3 that the deep generative clustering model relies on the hierarchical features learned by the neural network to show significant improvements in each metric compared to the traditional clustering models namely K-means and GMM. Notably, there is a minimum improvement of 0.3 in all three metrics on the MNIST dataset. Moreover, our proposed model performs well compared to different deep generative clustering models, owing to its prominent embedding approach. The accuracy of the proposed model on the MNIST dataset is 0.952, which is the best ACC, while NMI and ARI are 0.888 and 0.898 and have only around 0.002 difference from the best value. The proposed model achieves a best clustering accuracy (0.708) on F-MNIST. However, EDA-TEC and ClusterGAN perform better on NMI and ARI of F-MNIST. For the HHAR dataset, the GamMM-VAE achieved the best results in terms of ACC and ARI, but ranked second in the NMI metric. As for the REUTERS, REUTERS-10k and WEBKB datasets,

the proposed model still significantly outperformed other models in total.

These results highlight the effectiveness of the proposed model in leveraging the Gamma mixture model for more flexible and higher-quality embedding in the latent space. Our findings confirm the initial hypothesis that employing a non-symmetric distribution like the Gamma mixture model can enhance clustering performance, particularly in handling asymmetric data distributions. Furthermore, the superior performance across diverse datasets underscores the robustness and generalizability of the proposed model.

The significant improvements observed in the clustering metrics align with our research objectives, demonstrating that the integration of GamMM into the VAE framework can effectively address the limitations of existing models that primarily rely on Gaussian distributions. This study provides strong evidence that our model not only excels in clustering tasks but also offers a promising direction for future research in deep generative clustering methods.

Overall, the clustering results of the proposed model on the six datasets are significantly outstanding compared to the other listed models, and most of the experimental results are bolded and marked as the best, or at least as the third best by ‘†’. This aligns with our hypothesis that deep generative models with more flexible latent space distributions can achieve superior clustering performance.

2.4.3 Visualization of Learned Embeddings

In order to show the separation ability of the proposed model for each class during clustering, we use t-SNE (t-distributed Stochastic Neighbor Embedding), a technique for visualizing high-dimensional data in two or three dimensions [91], to transfer the latent embeddings from the encoder output to the 2D space. Taking the MNIST dataset as an example, the clustering of the latent embedding $\mathbf{z}$ in training at Epoch 1, Epoch 50, Epoch 150 and the final Epoch has been given in Figure 2.4. At Epoch 1, the model has been able to find the clustering centers of each class well with the help of pre-train, despite the poor clustering of the data. As the number of epochs rises, the points of each class are gradually clustered correctly. The latent embeddings have been well separated from each other in the final epoch.

To observe the correct assignment of each class prediction, we classify the predicted results and transfer them to the 2D space with t-SNE in the final epoch, and label each predicted class with its true label as shown in Figure 2.3. In conjunction with Figure 2.4, we can see in Figure 2.3 that the error points in the prediction are mostly adjacent clusters. For example, the digit 9 is adjacent to the digits 3, 4, and 7 in the latent space, so some of the predicted digits 3, 4, and 7 have a true label of 9, while some of the predicted

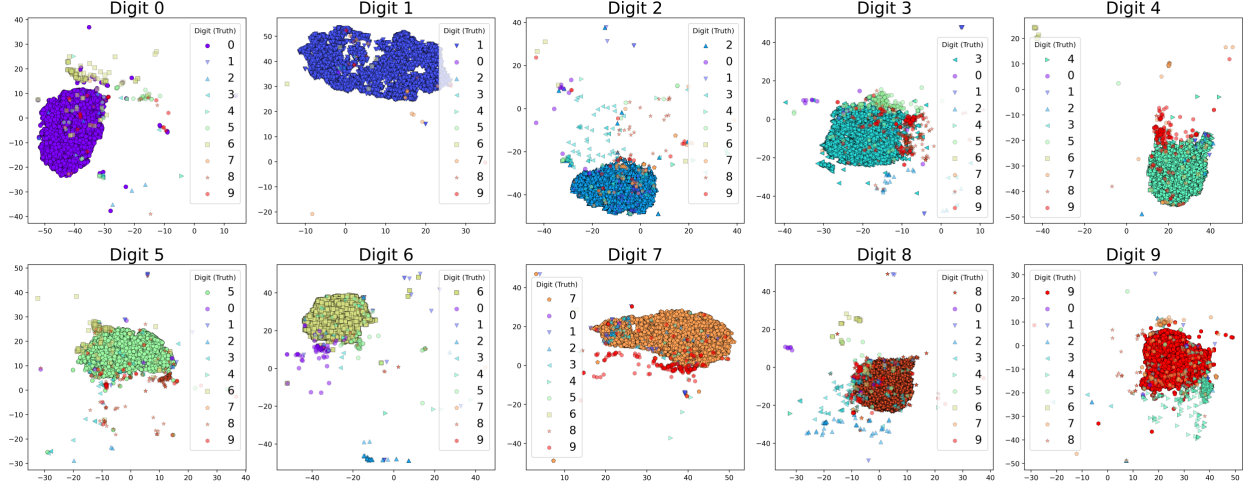
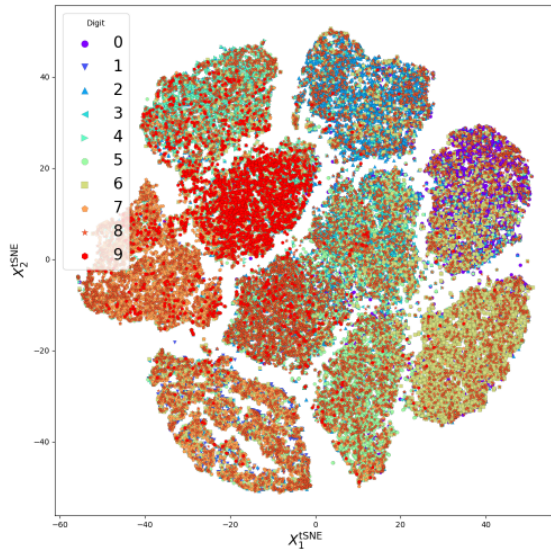


Figure 2.3: Visualization of the embeddings learned by GamMM-VAE on MNIST. All points in each sub-figure are predictions for the clusters of digits given by the subtitle. The different marks in sub-figures are the ground-truth labels.

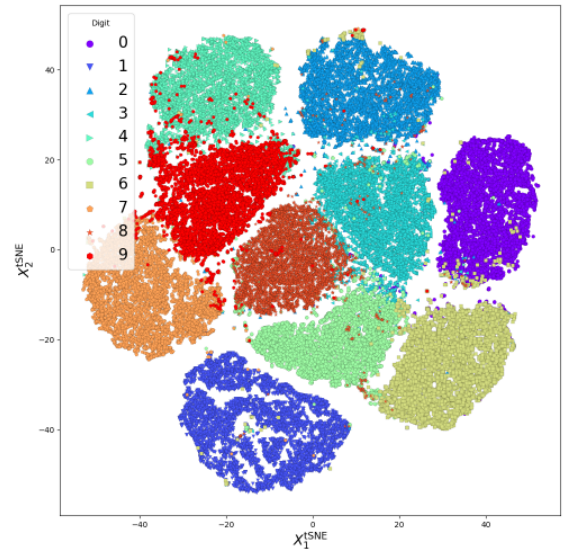
digit 9 is obviously actually the digit 4. Such prediction error points are often written with some ambiguity due to the similar shapes, which makes it difficult for the model to classify them. Overall, the majority of handwritten digits on MNIST were correctly clustered by GamMM-VAE.

Another important feature of the proposed model, in addition to its clustering capabilities, is its ability to generate high-quality images of specific classes due to the generative nature of the VAE model. Among the nine models compared, ClusterGAN [80] also demonstrates strong generative capabilities inherent to GANs. In this subsection, we compare the generated images produced by ClusterGAN and GamMM-VAE.

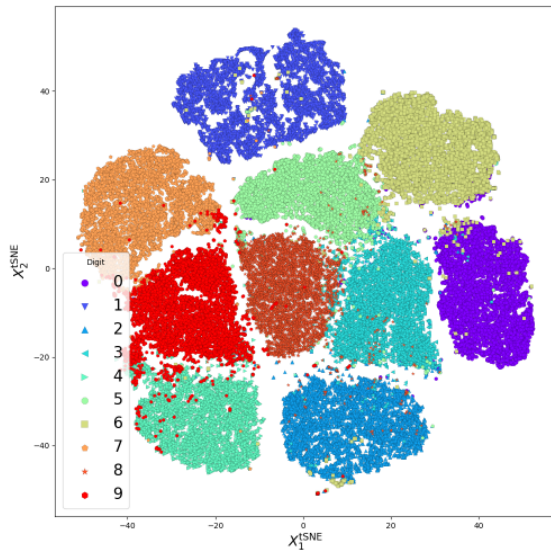
The generated images using ClusterGAN and GamMM-VAE for MNIST and F-MNIST datasets are given in Figure 2.5. The handwritten digits generated by GamMM-VAE exhibit good diversity, smoothness, and recognizability compared to those generated by ClusterGAN. For the F-MNIST dataset, the images generated by GamMM-VAE are also smoother than those produced by ClusterGAN. As for F-MNIST, the generated images of GamMM-VAE are smoother than those of ClusterGAN. Moreover, since GamMM-VAE has a significant advantage in the clustering accuracy of the four classes of Bag, Sandal, Sneaker and Ankle boot, the recognizability and distinguishability of the corresponding generated images are also much better than the F-MNIST images generated by ClusterGAN.



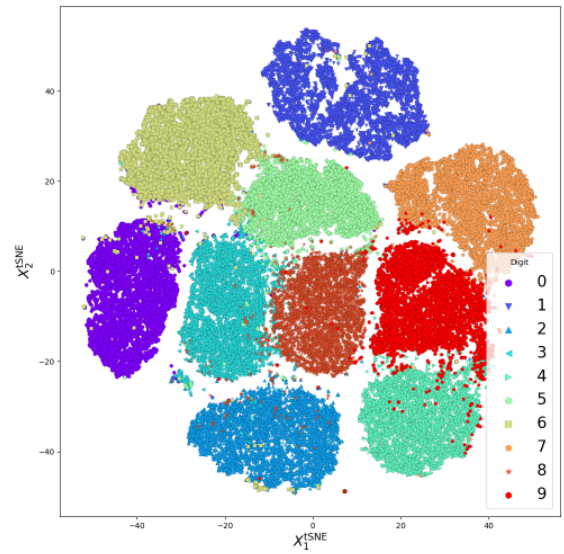
(a) Epoch 1 (46.35%)



(b) Epoch 50 (93.83%)



(c) Epoch 150 (94.63%)



(d) Epoch End (95.17%)

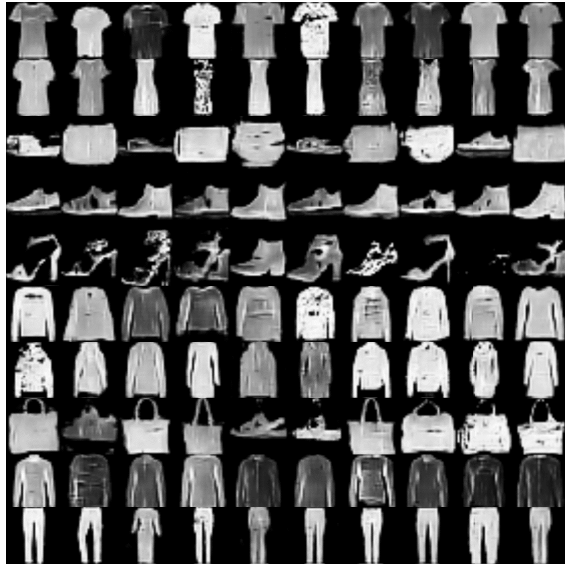
Figure 2.4: t-SNE visualization of obtained latent embeddings by GamMM-VAE for the MNIST dataset.



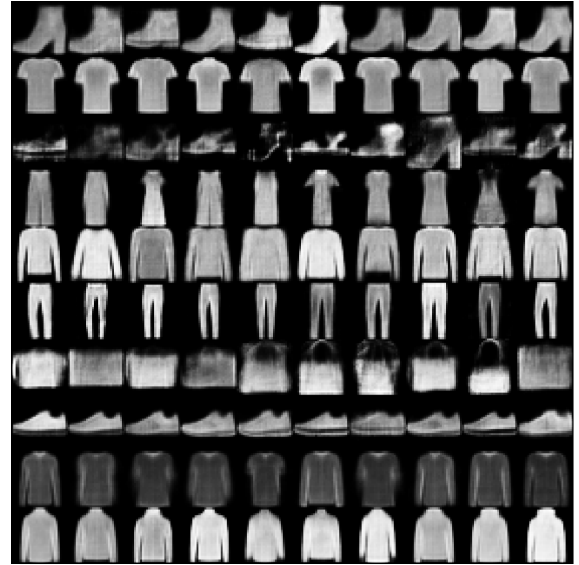
(a) MNIST ClusterGAN



(b) MNIST GamMM-VAE

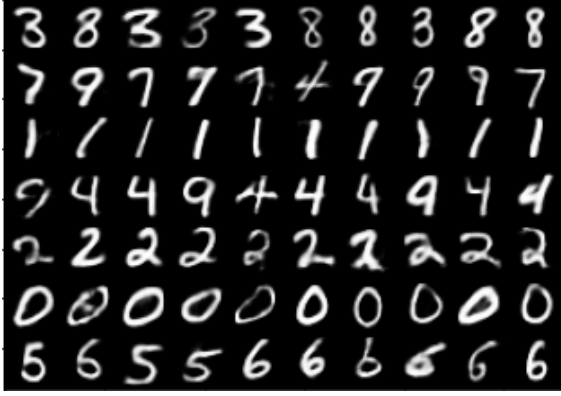


(c) F-MNIST ClusterGAN



(d) F-MNIST GamMM-VAE

Figure 2.5: The generated samples of MNIST and F-MNIST by ClusterGAN and GamMM-VAE. The digits and clothes in the same row come from the same cluster.



(a) 7 clusters



(b) 14 clusters

Figure 2.6: The generated samples of MNIST with different numbers of clusters. (Each row represents a cluster.)

2.4.4 Selecting the Number of Latent Clusters

The number of latent space clusters must be predetermined for the proposed GamMM-VAE model. Once the correct number of clusters is selected, GamMM-VAE can effectively perform both clustering and generation tasks. In Figure 2.6, we show the generation results obtained by specifying a particular number of clusters when the MNIST dataset is used as training data.

When the selected number of clusters is fewer than the true number, similar categories are clustered together, as shown in Figure 2.6a. For instance, digits '3' and '8', '5' and '6', '4' and '9', and '7' and '9' are grouped together due to their similar appearances. Conversely, when more clusters are selected than the true number, different shapes and variations of the same digits in the MNIST dataset are distinguished. In Figure 2.6b, orthogonal and skewed '1' and '8' are distinguished, and different writings of '2' and '6' are distinguished.

2.4.5 Ablation Study

To examine the impact of deep neural networks on experimental results, VaDE integrates an autoencoder network structure on top of GMM for clustering tasks. The experimental results in Tables 2.2 and 2.3 indicate that VaDE significantly surpasses GMM in clustering ability. Moreover, while maintaining the framework of a variational autoencoder, the proposed GamMM-VAE surpasses VaDE in most comparison metrics. This superiority showcases the benefits of employing a non-symmetric Gamma mixture model for embedding in the latent space. Furthermore, the proposed model demonstrates certain advantages

even when juxtaposed with the most recent models like EDA-TEC and GD-VAE. Nonetheless, EDA-TEC shows commendable performance on the NMI and ARI metrics, which are crucial for evaluating clustering performance. This performance is attributed to its variational autoencoder foundation, augmented with an improved graph regularization term based on weighted cosine similarity and sparsity regularization to better preserve the geometric data structure and enhance clustering outcomes. Contrarily, GD-VAE, primarily optimized for classification rather than clustering, does not showcase as pronounced performance on NMI and ARI as EDA-TEC and GamMM-VAE. The overall advantage demonstrated by the proposed model over other baseline models underscores the significance of each model component.

To address the stability of our model, we introduced an approximation of the Gamma distribution from the normal distribution in Subsection 2.3.2, incorporating a hyperparameter termed the adjustment coefficient. We explored its impact on model performance through experiments, where a larger adjustment coefficient theoretically yields a closer approximation of the Gamma to the normal distribution. Experimentation with λ on the MNIST dataset in increments of 5, across a spectrum from $[5,50]$, was conducted to ascertain its influence on model efficacy.

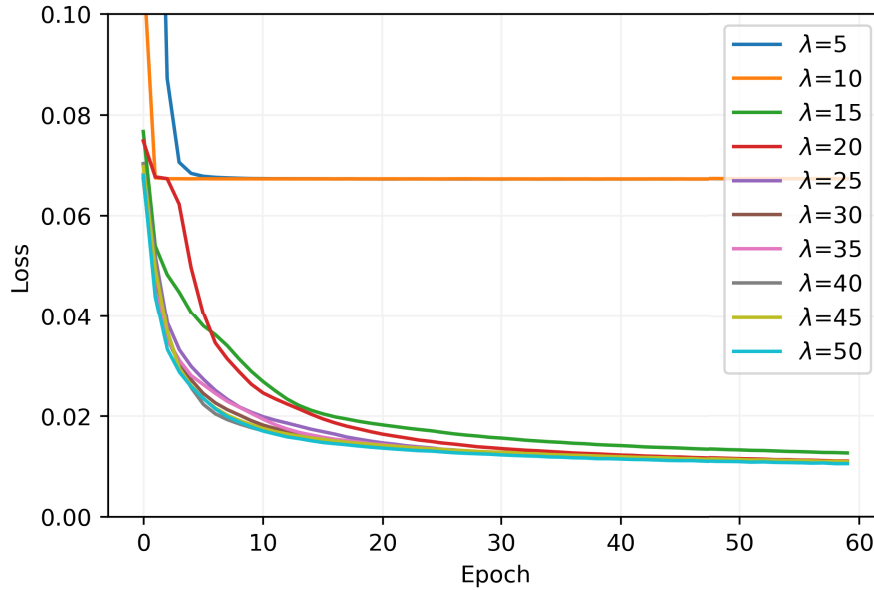


Figure 2.7: Gradient descent of loss in pretrain with respect to the adjustment coefficients on MNIST dataset.

Despite not utilizing our complete reparameterization trick in the pretraining phase, the output μ from $g(\mathbf{x}; \phi)$ requires scaling as per Equation 34 to ensure network weight stability throughout training. Figure

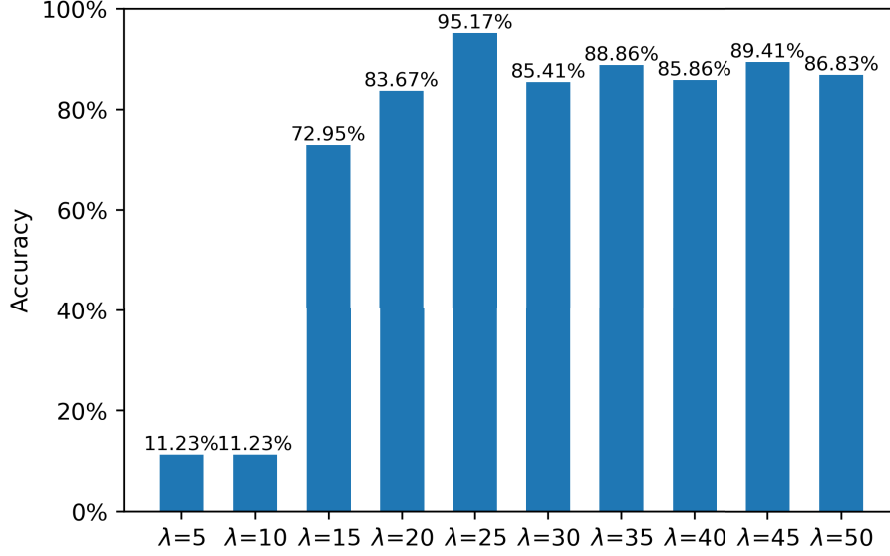


Figure 2.8: Accuracy with different adjustment coefficients on MNIST dataset.

2.7 depicts the pretraining loss decrease across different λ values. Notably, at $\lambda = 5$ and $\lambda = 10$, the models exhibit poor gradient descent, mirrored by an accuracy of 11.23% in Figure 2.8, suggesting negligible clustering capability. An improvement in gradient descent and corresponding accuracy was observed at $\lambda = 15$ and $\lambda = 20$, indicating an enhanced fit of the Gamma distribution to the normal distribution.

$$\mu^* = softmax(\mu) * \lambda \quad (34)$$

For $\lambda \geq 20$, gradient descent and accuracy levels out, suggesting an optimal range for λ . However, excessively high λ values, such as $\lambda = 45$ and $\lambda = 50$, can destabilize and crash the model training due to computational limitations in handling operations within $\mathcal{L}_{ELBO}(\mathbf{x})$ involving β as the base and α as the exponent. Similar outcomes across other datasets recommend a λ range of $[20, 40]$, balancing approximation accuracy and computational stability.

2.4.6 Comparison of the Convergence Speed

Our proposed model has the property of fast convergence after pretraining stacked autoencoders. Figure 2.9 compares our model with the other seven models in terms of the change in accuracy over 200 epochs on the MNIST dataset. Our proposed model and EDA-TEC can rapidly reach an accuracy of about 90%

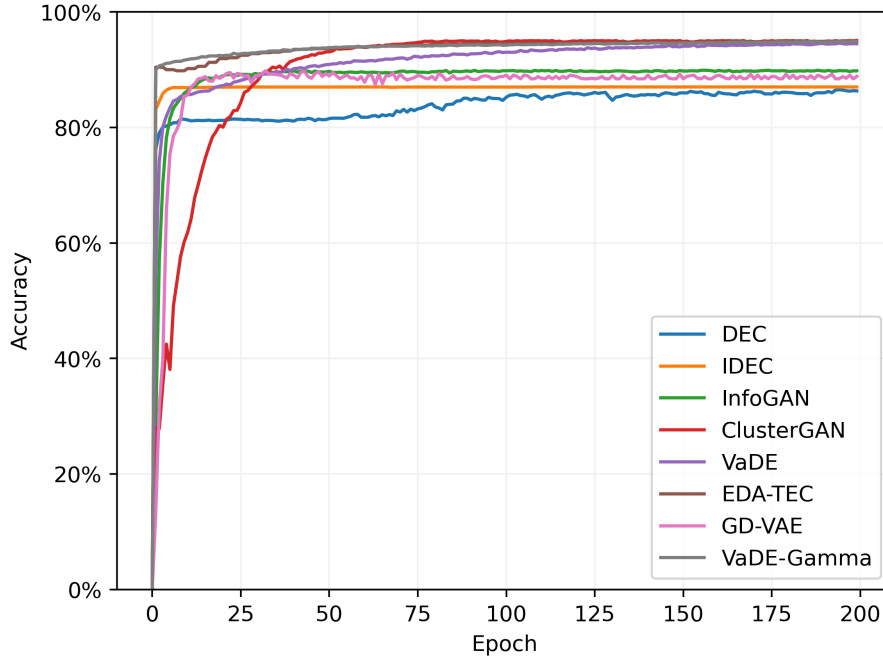


Figure 2.9: Comparison of the convergence speed by different models on MNIST dataset.

once training is started, and eventually reaches 95% at epoch 200. DEC, IDEC, InfoGAN and GD-VAE not only do not converge as fast as our model, but also end up with accuracy rates below 90%. The final accuracy of VaDE is similar to that of our model, but the convergence speed is much less than ours. It is worth noting that although the convergence speed of ClusterGAN is not as fast as that of GamMM-VAE at the beginning of the experiment, it outperformed GamMM-VAE around epoch 60. Overall, the fast convergence of our model outperforms most compared models.

Our proposed model demonstrates fast convergence after pretraining stacked autoencoders. Figure 2.9 compares the convergence speed of our model with seven other models by observing the change in accuracy over 200 epochs on the MNIST dataset. Both our model and EDA-TEC quickly achieve an accuracy of approximately 90% soon after training begins, ultimately reaching 95% by epoch 200. In contrast, DEC, IDEC, InfoGAN, and GD-VAE do not converge as quickly and have final accuracy rates below 90%. While VaDE’s final accuracy is similar to ours, its convergence speed is significantly slower. Notably, although ClusterGAN’s initial convergence speed is slower than GamMM-VAE, it surpasses GamMM-VAE’s accuracy around epoch 60. Overall, our model’s rapid convergence outperforms most of the compared models.

2.5 Conclusion

In this chapter, we propose a deep clustering approach based on a Gamma mixture model integrated with a variational autoencoder (GamMM-VAE). This model not only demonstrates excellent clustering performance but also possesses the capability to generate original data for specified classes. By adjusting the Gamma mixture model as the prior distribution in the VAE’s latent space, we enhance the model’s clustering capabilities through improved latent embeddings. We derived an Evidence Lower Bound (ELBO) based on the Gamma mixture to optimize the model during training. Additionally, we introduced a method to approximate the Gamma distribution using a Gaussian reparameterization trick. Testing on six benchmark datasets and comparing with other state-of-the-art models, our proposed model exhibited significant clustering and generative capabilities. These datasets cover three application areas: image clustering, human activity recognition, and text clustering, highlighting the practical applicability of our model in real-world scenarios.

Despite its strengths, a limitation of our proposed model is the need for manual selection of the number of clusters. This requirement stems from the necessity to predetermine and train the distribution parameters of the mixture model components embedded in the latent space. Additionally, the current model cannot gradually eliminate the weights of non-significant components, which may impact its efficiency. Future work may address these limitations by employing methods such as Bayesian variational inference, which could allow for automatic determination of the number of clusters and gradual elimination of non-significant components.

Further research will focus on addressing these limitations and exploring additional directions for improvement. Given that GANs can continuously enhance their generative capabilities through adversarial training between generators and discriminators, they offer a promising avenue for enhancing our model. Combining VAE and GAN, as seen in models like VAE-GAN [92], could further improve the characterization of latent embeddings. Additionally, exploring different non-Gaussian priors and applying the model to diverse real-world datasets could provide deeper insights and enhance its practical applicability.

Overall, our proposed GamMM-VAE model represents a significant advancement in deep clustering methods, offering robust performance and practical utility across various domains. The promising results and identified areas for improvement pave the way for future research to build upon this foundation.

Chapter 3

Variational Inference for the Bingham Mixture Model and Applications

3.1 Introduction

Foundational probabilistic models play a pivotal role in enhancing the accuracy and interpretability of data-driven tasks. In recent years, directional data analysis, which focuses on multivariate data constrained to a unit hypersphere, has attracted increasing interest due to its relevance in a wide range of real-world applications. Unlike conventional data types, directional data emphasize orientation rather than magnitude, making them particularly suitable for domains where the direction of vectors is more informative than their length. These include bioinformatics, collaborative filtering, wind direction modeling, text categorization, and speaker recognition [28, 29, 93, 94].

In machine learning pipelines, data normalization, particularly using the L_2 norm, is a common preprocessing step. This normalization projects data onto the unit hypersphere, suggesting that directional probability distributions are more appropriate for modeling than traditional Euclidean-based distributions (such as the multivariate Gaussian) for capturing the geometry of such data. Among these, the von Mises–Fisher (vMF) [30], Watson [31], and Bingham [32] distributions are widely studied. The vMF distribution, often viewed as the spherical analogue of the Gaussian, is characterized by a mean direction and a concentration parameter. It has been successfully applied in tasks such as topic modeling, fMRI analysis, and deep generative learning [95, 96, 97, 61]. However, the vMF distribution is limited in its ability to model axially

symmetric data, where the probability density remains unchanged when a vector is flipped (i.e., $\mathbf{x}$ and $-\mathbf{x}$ are equally likely). Such symmetry arises in many real-world scenarios, including diffusion tensor imaging, structural geology, and molecular orientation data. To address these limitations, the Watson and Bingham distributions offer more flexible alternatives. The Watson distribution captures antipodal symmetry and has been employed in clustering gene expression data [98], axial image modeling [99, 100], and speech signal separation [101]. The Bingham distribution generalizes the Watson distribution, enabling modeling of arbitrary axial symmetric patterns and has been used in diverse applications such as earthquake focal mechanisms [102, 103], rock fracture analysis [104, 105], hydrodynamics [106, 107], and shape analysis [108].

Despite its expressive power, the practical adoption of the Bingham distribution has been limited, primarily due to the computational intractability of its normalization constant. This constant depends on the eigenvalues of the parameter matrix and lacks a closed-form solution, posing a major challenge for likelihood-based inference and parameter estimation. Traditional solutions include Markov chain Monte Carlo (MCMC) sampling [109], precomputed lookup tables [110], saddlepoint approximations [111], series expansions [33], and holonomic gradient descent (HGD) [34]. While these approaches offer various trade-offs between accuracy and computational efficiency, they often suffer from scalability limitations, poor generalization to higher dimensions, or lack of adaptability in dynamic inference settings.

To address these challenges, we propose a variational Bayesian framework for inference in Bingham mixture models (BMMs). The key motivation is to circumvent the need to compute the intractable normalization constants by leveraging a zero-mean Gaussian reparameterization and directional constraints. Our approach provides a scalable and flexible alternative to MCMC and other approximation-based methods, while still preserving uncertainty quantification. In contrast to previous efforts that either rely heavily on sampling or precomputation, our method transfers the computational burden to the posterior approximation of a tractable distribution (i.e., a zero-mean Gaussian) thus enabling efficient inference via variational techniques.

Moreover, existing Bayesian treatments of the Bingham distribution remain relatively underexplored. For instance, [103] proposed an approximate maximum likelihood estimator using MCMC for Bayesian computation. While effective in low dimensions, the method incurs substantial computational overhead. In [102], rejection sampling was employed to achieve exact Bayesian inference under the angular central Gaussian formulation, but evaluation was again limited to low-dimensional data. Additionally, [112] introduced

a recursive Bingham filter for directional data, but its reliance on precomputed tables for normalization constants limits adaptability. In contrast, our variational inference framework is conceptually simple, computationally efficient, and broadly applicable. It eliminates the need for sampling-based or lookup-based approximations and shifts the technical complexity to the estimation of a zero-mean Gaussian mixture. Our method offers improved scalability, enables closed-form updates, and provides uncertainty estimates over model parameters. Extensive experiments on both synthetic and real-world datasets demonstrate that the proposed approach achieves competitive or superior performance compared to existing methods, while significantly reducing computational cost. In summary, this work makes the following contributions:

- We propose a novel variational inference framework for Bingham mixture models that avoids computing the intractable normalization constants.
- We establish a theoretically grounded connection between the Bingham distribution and the zero-mean Gaussian, enabling a tractable inference mechanism.
- We provide a scalable implementation and demonstrate its effectiveness on synthetic and real directional data, highlighting the advantages in both accuracy and efficiency.

The rest of this chapter is structured as follows: Section 3.2 reviews the Bingham distribution and presents the proposed variational inference framework. Section 3.3 reports the experimental setup and results. Section 3.4 concludes the chapter with a discussion of implications and future directions.

3.2 The Proposed Model

In this section, we first introduce the formulation of the Bingham distribution and its mixture model. We then analyze the effects of the distribution’s parameters on its shape using both equations and illustrative visualizations. Furthermore, we discuss the theoretical connection and transformation between the Bingham mixture model and the Gaussian mixture model. Finally, we present our variational inference framework for the Bingham mixture model.

3.2.1 Bingham Distribution

Let $\mathbf{x} \in \mathbb{S}^{D-1}$ denote a D -dimensional random vector constrained to the unit hypersphere, where $\mathbb{S}^{D-1} = \{\mathbf{x} \in \mathbb{R}^D \mid \|\mathbf{x}\| = 1\}$. The Bingham distribution models axially symmetric data on this hypersphere. Its probability density function is given by:

$$\mathcal{B}(\mathbf{x} \mid \mathbf{U}, \boldsymbol{\Omega}) = {}_1F_1\left(\frac{1}{2}, \frac{D}{2}, \boldsymbol{\Omega}\right)^{-1} \exp(\mathbf{x}^T \mathbf{U} \boldsymbol{\Omega} \mathbf{U}^T \mathbf{x}), \quad (35)$$

where $\mathbf{U}$ is an orthonormal matrix $[\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_D]$, and $\boldsymbol{\Omega}$ is a diagonal matrix with negative parameters $(\omega_1 \leq \omega_2 \leq \dots \leq \omega_{D-1} < 0)$. The function ${}_1F_1\left(\frac{1}{2}, \frac{D}{2}, \boldsymbol{\Omega}\right)$ is the confluent hypergeometric function of matrix argument, which serves as the normalization constant:

$${}_1F_1\left(\frac{1}{2}, \frac{D}{2}, \boldsymbol{\Omega}\right) = \sum_{t=0}^{\infty} \sum_{\tau \vdash t} \frac{\left(\frac{1}{2}\right)_{\tau}^{(2)}}{\left(\frac{D}{2}\right)_{\tau}^{(2)}} \frac{C_{\tau}^{(2)}(\boldsymbol{\Omega})}{t!}, \quad (36)$$

where $\tau \vdash t$ denotes that τ is a partition of integer t , $C_{\tau}^{(2)}$ represents the zonal polynomials, and $(\cdot)_{\tau}^{(2)}$ is the generalized Pochhammer symbol.

The Bingham distribution is rotationally invariant and particularly well-suited for modeling data with antipodal symmetry. To facilitate intuition, we present visualizations of several 3-dimensional Bingham distributions with a fixed orientation matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$ and varying dispersion parameters $\boldsymbol{\Omega} = (\omega_1, \omega_2)$, as shown in Figure 3.1.

When $\omega_1 = \omega_2 = 0$, the distribution becomes uniform over the hypersphere. As the values of ω_1 and ω_2 decrease uniformly, the distribution increasingly concentrates around two antipodal points, forming a symmetric pattern. When $\omega_1 \neq \omega_2$, the distribution stretches from a symmetric circular band into elliptical shapes, eventually forming a ring-like structure. The difference between ω_1 and ω_2 governs the eccentricity of the distribution on the sphere.

The orientation matrix $\mathbf{U}$ determines the principal axes of concentration. In particular, $\mathbf{u}_3$ denotes the direction of maximum accumulation. When the difference between ω_1 and ω_2 is large, $\mathbf{u}_3$ and the eigenvector associated with the more negative eigenvalue (e.g., $\mathbf{u}_d$) define the plane in which the distribution compresses into a ring.

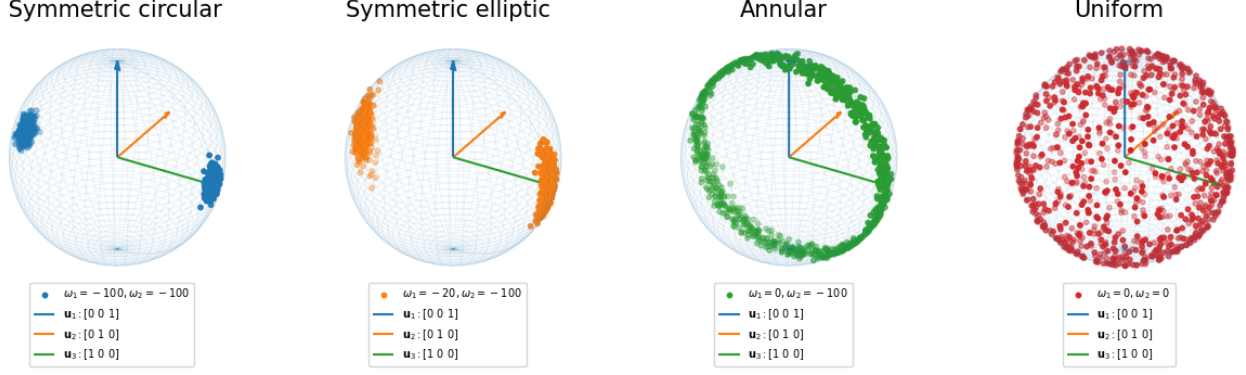


Figure 3.1: Visualizations of 3D Bingham distributions with varying Ω and fixed orientation matrix $\mathbf{U} = [[0, 0, 1], [0, 1, 0], [1, 0, 0]]$. The shape transitions from uniform to highly concentrated and annular as dispersion parameters vary.

3.2.2 Variational Inference for Bingham Mixture Model

This section presents a detailed formulation of the variational inference framework for the Bingham mixture model (BinMM). We begin with a formal model description, followed by the derivation of the evidence lower bound (ELBO), and conclude with the parameter estimation process.

Model Description

Let $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ denote a set of N i.i.d observations on the unit hypersphere, and assume that the data are generated from a mixture model with K Bingham components. The probability density function (PDF) of the BinMM is defined as:

$$p(\mathbf{X} \mid \underline{\mathbf{U}}, \underline{\Omega}, \underline{\pi}) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \mathcal{B}(\mathbf{x}_n \mid \mathbf{U}_k, \Omega_k), \quad (37)$$

where $\underline{\pi} = \{\pi_k\}_{k=1}^K$ are the mixture weights satisfying $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$, $\underline{\mathbf{U}} = \{\mathbf{U}_k\}_{k=1}^K$ are the orthonormal direction matrices, and $\underline{\Omega} = \{\Omega_k\}_{k=1}^K$ are the corresponding shape matrices. The posterior responsibility, i.e., the probability that $\mathbf{x}_n$ belongs to the k -th component, is given by Bayes' rule:

$$p(k \mid \mathbf{x}_n) \propto \pi_k \mathcal{B}(\mathbf{x}_n \mid \mathbf{U}_k, \Omega_k). \quad (38)$$

In our work, we exploit a well-established mathematical equivalence between the Bingham distribution and a zero-mean Gaussian distribution with a precision matrix Λ_k , as described in [32, 29]. Using this, the

responsibilities can be redefined as:

$$p(k | \mathbf{x}_n) \propto \pi_k \mathcal{N}(\mathbf{x}_n | \mathbf{0}, \mathbf{\Lambda}_k^{-1}), \quad (39)$$

where $\mathcal{N}(\mathbf{x}_n | \mathbf{0}, \mathbf{\Lambda}_k^{-1})$ denotes a D -dimensional Gaussian distribution with zero mean and precision matrix $\mathbf{\Lambda}_k$. The relationship between the Gaussian and Bingham parameters is given by the eigen-decomposition of $-\frac{1}{2}\mathbf{\Lambda}_k$, whose eigenvalues and eigenvectors correspond to $\mathbf{\Omega}_k$ and $\mathbf{U}_k$, respectively. This transformation is further discussed in Section 3.2.2.

Consequently, the BinMM can be equivalently represented as a zero-mean Gaussian mixture model (GMM):

$$p(\mathbf{X} | \mathbf{\underline{\Lambda}}, \pi) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mathbf{0}, \mathbf{\Lambda}_k^{-1}), \quad (40)$$

where $\mathbf{\underline{\Lambda}} = \{\mathbf{\Lambda}_k\}_{k=1}^K$ denotes the set of precision matrices. Figure 3.2 illustrates the relationship between the Bingham and zero-mean Gaussian distributions.

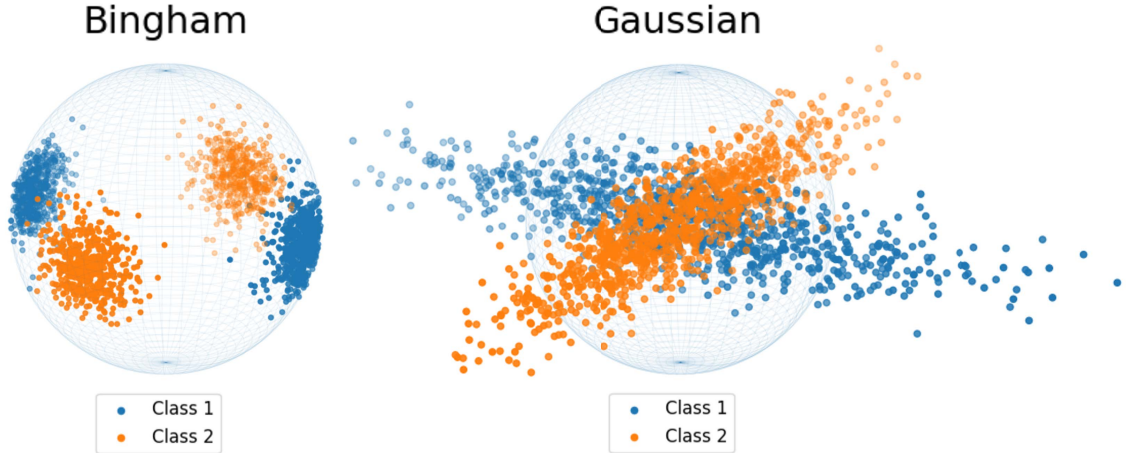


Figure 3.2: Visualization of the relationship between the Bingham and zero-mean Gaussian distributions. The parameters $\mathbf{\Omega}_k$ and $\mathbf{U}_k$ of the Bingham distribution are obtained as the eigenvalues and eigenvectors of $-\frac{1}{2}\mathbf{\Lambda}_k$ from the corresponding Gaussian component.

It is noteworthy that this transformation is critical as it enables efficient inference by circumventing the intractable normalization constant of the Bingham distribution. The parameter estimation of BinMM can thus be formulated through variational inference over the corresponding zero-mean GMM.

For each observation $\mathbf{x}_n$, we introduce a latent variable $\mathbf{z}_n = \{z_{n1}, \dots, z_{nK}\}$ such that $z_{nk} \in \{0, 1\}$, $\sum_{k=1}^K z_{nk} = 1$ and $z_{nk} = 1$ if $\mathbf{x}_n$ belongs to component k and 0, otherwise. Let $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$ denote the

full set of latent variables. The conditional distribution of $\underline{\mathbf{Z}}$ given the mixture weights $\underline{\pi}$ is:

$$p(\underline{\mathbf{Z}} \mid \underline{\pi}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}}. \quad (41)$$

Given the assignments $\underline{\mathbf{Z}}$ and precision matrices $\underline{\Lambda}$, the conditional distribution of the data $\underline{\mathbf{X}}$ is:

$$p(\underline{\mathbf{X}} \mid \underline{\mathbf{Z}}, \underline{\Lambda}) = \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n \mid \mathbf{0}, \Lambda_k^{-1})^{z_{nk}}. \quad (42)$$

To complete the Bayesian model, we place a Dirichlet prior on the mixture weights $\underline{\pi}$:

$$\begin{aligned} p(\underline{\pi}) &= \text{Dir}(\underline{\pi} \mid \underline{\alpha}_0) \\ &= \frac{\Gamma\left(\sum_{k=1}^K \alpha_{0,k}\right)}{\prod_{k=1}^K \Gamma(\alpha_{0,k})} \prod_{k=1}^K \pi_k^{\alpha_{0,k}-1}, \end{aligned} \quad (43)$$

where $\underline{\alpha}_0 = \{\alpha_{0,k}\}_{k=1}^K$ are hyperparameters. Furthermore, we assign a Wishart prior to each precision matrix Λ_k :

$$p(\underline{\Lambda}) = \prod_{k=1}^K \mathcal{W}(\Lambda_k \mid \mathbf{W}_0, \nu_0). \quad (44)$$

where the Wishart distribution $\mathcal{W}(\Lambda_k \mid \mathbf{W}_0, \nu_0)$ is defined as:

$$\begin{aligned} \mathcal{W}(\Lambda_0 \mid \mathbf{W}_0, \nu_0) &= \\ B(\mathbf{W}_0, \nu_0) |\Lambda_k|^{(\nu_0-D-1)/2} \exp\left(-\frac{1}{2} \text{Tr}(\mathbf{W}_0^{-1} \Lambda_k)\right), \end{aligned} \quad (45)$$

and the normalizing constant $B(\mathbf{W}_0, \nu_0)$ is given by:

$$\begin{aligned} B(\mathbf{W}_0, \nu_0) &\equiv \\ |\mathbf{W}_0|^{-\nu_0/2} \left(2^{\nu_0 D/2} \pi^{D(D-1)/4} \prod_{i=1}^D \Gamma\left(\frac{\nu_0 + 1 - i}{2}\right) \right)^{-1}, \end{aligned} \quad (46)$$

where $\mathbf{W}_0$ is a symmetric, positive-definite $D \times D$ matrix, and ν_0 is the degrees of freedom, which must satisfy $\nu_0 > D - 1$ to ensure the validity of the Gamma function.

Figure 3.3 presents the graphical model of the proposed variational inference framework for the BinMM, illustrating the hierarchical structure and dependencies among latent variables and parameters.

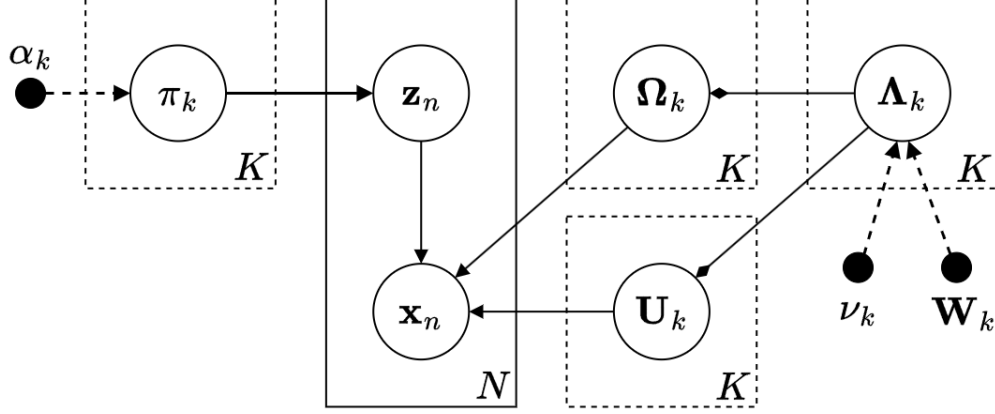


Figure 3.3: Directed acyclic graph (DAG) representation of the Bingham mixture model via Gaussian transformation. Nodes represent random variables, edges indicate dependencies, and plates denote replication across data points.

Variational Inference

We begin by formulating the joint distribution over all random variables involved in the model to facilitate the derivation of the variational framework. The joint distribution is expressed as:

$$p(\underline{\mathbf{X}}, \underline{\mathbf{Z}}, \underline{\pi}, \underline{\Lambda}) = p(\underline{\mathbf{X}} \mid \underline{\mathbf{Z}}, \underline{\Lambda})p(\underline{\mathbf{Z}} \mid \underline{\pi})p(\underline{\pi})p(\underline{\Lambda}), \quad (47)$$

where all terms have been previously defined. Note that only the data $\underline{\mathbf{X}} = \{\mathbf{x}_n\}_{n=1}^N$ are observed.

To perform inference, we employ a mean-field variational approximation, wherein the joint variational distribution factorizes as:

$$q(\underline{\mathbf{Z}}, \underline{\pi}, \underline{\Lambda}) = q(\underline{\mathbf{Z}})q(\underline{\pi})q(\underline{\Lambda}). \quad (48)$$

This factorization not only allows for efficient inference but also enables uncertainty to be captured through distributions over parameters, thereby facilitating the propagation of information and improving convergence during optimization.

Variational inference seeks to approximate the true posterior distribution $p(\underline{\mathbf{Z}}, \underline{\pi}, \underline{\Lambda} \mid \underline{\mathbf{X}})$ by minimizing the Kullback–Leibler (KL) divergence with respect to the variational distribution $q(\underline{\mathbf{Z}}, \underline{\pi}, \underline{\Lambda})$. This is equivalent to maximizing the evidence lower bound (ELBO) on the marginal log-likelihood:

$$\log p(\underline{\mathbf{X}}) = \mathcal{L}(q) + \text{KL}[p(\underline{\mathbf{Z}}, \underline{\pi}, \underline{\Lambda} \mid \underline{\mathbf{X}}) \parallel q(\underline{\mathbf{Z}}, \underline{\pi}, \underline{\Lambda})], \quad (49)$$

where the ELBO is defined as:

$$\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\mathbf{X} | \mathbf{Z}, \mathbf{\Lambda}) p(\mathbf{Z} | \pi) p(\pi) p(\mathbf{\Lambda})}{q(\mathbf{Z}) q(\pi) q(\mathbf{\Lambda})} \right]. \quad (50)$$

Maximization of the ELBO is carried out with respect to the variational factors $q(\mathbf{Z})$, $q(\pi)$, and $q(\mathbf{\Lambda})$, where the expectation $\mathbb{E}_q[\cdot]$ is taken over the variational distributions.

Variational Expectation Step

To derive the optimal distribution over the latent assignments $q(\mathbf{Z})$, we isolate the terms in Equation 50 that depend on $\mathbf{Z}$, absorbing all other terms into a constant:

$$\begin{aligned} \mathcal{L}(q) &= \text{cst} + \mathbb{E}_q \left[\log \frac{p(\mathbf{X} | \mathbf{Z}, \mathbf{\Lambda}) p(\mathbf{Z} | \pi)}{q(\mathbf{Z})} \right] \\ &= \text{cst} + \sum_{k=1}^K \sum_{n=1}^N \log \rho_{nk} \mathbb{E}_q [z_{nk}] - \mathbb{E}_q [\log q(\mathbf{Z})], \end{aligned} \quad (51)$$

where $\log \rho_{nk}$ is defined as

$$\begin{aligned} \log \rho_{nk} &= \mathbb{E} [\log \pi_k] + \frac{1}{2} \mathbb{E} [\log |\mathbf{\Lambda}_k|] - \frac{D}{2} \log(2\pi) \\ &\quad - \frac{1}{2} \mathbb{E}_{\mathbf{\Lambda}_k} [\mathbf{x}_n^T \mathbf{\Lambda}_k \mathbf{x}_n]. \end{aligned} \quad (52)$$

Each term is computed as follows:

$$\begin{aligned} \mathbb{E}_{\mathbf{\Lambda}_k} [\mathbf{x}_n^T \mathbf{\Lambda}_k \mathbf{x}_n] &= D + \nu_k (\mathbf{x}_n^T \mathbf{W}_k \mathbf{x}_n), \\ \mathbb{E} [\ln |\mathbf{\Lambda}_k|] &= \sum_{i=1}^D \psi \left(\frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |\mathbf{W}_k|, \\ \mathbb{E} [\ln \pi_k] &= \psi(\alpha_k) - \psi(\hat{\alpha}), \end{aligned} \quad (53)$$

where $\psi(\cdot)$ is the digamma function, and $\hat{\alpha} = \sum_{k=1}^K \alpha_k$.

Using these quantities, the optimal variational distribution over the latent variables is:

$$\log q^*(\mathbf{Z}) = \text{const} + \sum_{n=1}^N \sum_{k=1}^K z_{nk} \log \rho_{nk}, \quad (54)$$

which implies:

$$q^*(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^K \rho_{nj}}, \quad (55)$$

where r_{nk} represents the posterior responsibility of cluster k for data point $\mathbf{x}_n$, and satisfies $\sum_k r_{nk} = 1$.

Therefore, the expected value of the assignment variable is:

$$\mathbb{E}_q [z_{nk}] = r_{nk}. \quad (56)$$

Variational Maximization Step

We now derive updates for the variational distributions over the parameters $\underline{\pi}$ and $\underline{\Lambda}$.

Update for $q(\underline{\pi})$ Isolating the terms in Equation 50 that involve $\underline{\pi}$, yielding:

$$\begin{aligned} \mathcal{L}(q) &= \text{cst} + \mathbb{E}_q \left[\log \frac{p(\mathbf{Z} | \underline{\pi}) p(\underline{\pi})}{q(\underline{\pi})} \right] \\ &= \text{cst} + \sum_{k=1}^K \mathbb{E}_q [f(\pi_k)] - \mathbb{E}_q [\log q(\underline{\pi})], \end{aligned} \quad (57)$$

where $f(\pi_k) = ((\alpha_{0,k} - 1) + \sum_{n=1}^N r_{nk}) \log \pi_k$. Thus, the log of the optimized $q(\underline{\pi})$ is given by

$$\log q^*(\underline{\pi}) = \text{cst} + \sum_{k=1}^K f(\pi_k). \quad (58)$$

Taking the exponential of both sides of this expression, $q^*(\underline{\pi})$ is recognized to have a Dirichlet distribution,

$$q^*(\underline{\pi}) = \text{Dir}(\underline{\pi} | \underline{\alpha}), \quad \alpha_k = \alpha_{0,k} + \sum_{n=1}^N r_{nk}, \quad (59)$$

where $\underline{\alpha} = \{\alpha_k\}_{k=1}^K$ and $\alpha_{0,k} = 1/K$.

Update for $q(\underline{\Lambda})$ The optimization of $q(\underline{\Lambda})$ is performed by rewriting Equation 50 by including all terms except $\underline{\Lambda}$ involved into a constant term, which is defined as:

$$\begin{aligned}\mathcal{L}(q) &= \text{cst} + \mathbb{E}_q \left[\log \frac{p(\underline{\mathbf{X}} | \underline{\mathbf{Z}}, \underline{\Lambda}) p(\underline{\Lambda})}{q(\underline{\Lambda})} \right] \\ &= \text{cst} + \sum_{k=1}^K \mathbb{E}_q [f(\mathbf{\Lambda}_k)] - \mathbb{E}_q [\log q(\mathbf{\Lambda}_k)],\end{aligned}\tag{60}$$

where

$$\begin{aligned}f(\mathbf{\Lambda}_k) &= \text{cst} + \sum_{n=1}^N \mathbb{E} [z_{nk}] \log \mathcal{N}(\mathbf{x}_n | \mathbf{0}, \mathbf{\Lambda}_k) \\ &\quad + \log \mathcal{W}(\mathbf{\Lambda}_k | \mathbf{W}_0, \nu_0).\end{aligned}\tag{61}$$

Then, the log of the optimized $q(\mathbf{\Lambda}_k)$ is given by

$$\log q^*(\mathbf{\Lambda}_k) = \log \mathcal{W}(\mathbf{\Lambda}_k | \mathbf{W}_k, \nu_k),\tag{62}$$

where the components $\mathbf{W}_k^{-1}, \nu_k$ are given by

$$\begin{aligned}\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k \mathbf{S}_k, \\ \nu_k &= \nu_0 + N_k,\end{aligned}\tag{63}$$

where $\nu_0 = D$, and $\mathbf{W}_0$ is initialized as the covariance matrix of the given data $\underline{\mathbf{X}}$. Here, the intermediate statistics are:

$$\begin{aligned}N_k &= \sum_{n=1}^N r_{nk}, \\ \mathbf{S}_k &= \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\mathbf{x}_n^T \mathbf{x}_n).\end{aligned}\tag{64}$$

Posterior Expectations

The final posterior expectations for the mixture weights and precision matrices are computed as:

$$\begin{aligned}\mathbb{E}[\pi] &= \frac{\alpha}{\sum_{k=1}^K \alpha_k} \\ \mathbb{E}[\underline{\Lambda}] &= \underline{\nu} \cdot \underline{\mathbf{W}},\end{aligned}\tag{65}$$

where $\underline{\nu} = \{\nu_k\}_{k=1}^K$ denotes a set of the number of degrees of freedom, $\underline{\mathbf{W}} = \{\mathbf{W}_k\}_{k=1}^K$ denotes a set of positive definite metrics. These expectations are used in subsequent iterations of the variational inference algorithm.

Bridging GMM and BinMM through Eigendecomposition of Precision Matrices

A key insight in the construction of the BinMM lies in leveraging the structural correspondence between Gaussian and Bingham distributions. As established in the foundational work of [32], a zero-mean Gaussian Mixture Model (GMM) can be utilized to parameterize a BinMM through eigendecomposition of the precision matrices. This connection enables the formulation of a variational inference framework for directional data, facilitating the practical deployment of BinMM in hyperspherical data modeling scenarios.

In this transformation, the mixture weights π_k remain invariant under the mapping and can be directly reused from the GMM to the BinMM. The primary challenge, however, is to derive the directional parameters $\mathbf{U}_k$ (orthonormal axes) and the concentration matrices $\mathbf{\Omega}_k$ (diagonal structure) of the Bingham distributions from the precision matrices $\mathbf{\Lambda}_k$ of the GMM. This is accomplished through eigendecomposition of the inverse precision matrices, yielding the following transformation:

$$\begin{aligned} \pi_k &= \pi_k \\ \mathbf{\Omega}_k, \mathbf{U}_k^{-1} &= \text{Eig} \left(-\frac{1}{2} \mathbf{\Lambda}_k^{-1} \right) \end{aligned} \tag{66}$$

where $\text{Eig}(\cdot)$ denotes the eigendecomposition operator, which returns the eigenvalues and eigenvectors of the matrix argument.

This transformation effectively decouples the covariance structure of the Gaussian components into directional statistics, yielding parameters that align with the Bingham distribution on the hypersphere. The eigenvectors define the orientation of the distribution, while the negative eigenvalues control the concentration along each axis. As a result, the BinMM inherits a coherent and interpretable geometric structure from the GMM representation. The complete variational inference procedure for the BinMM, including this transformation step, is summarized in Algorithm 2.

Algorithm 2 Variational Inference for Bingham Mixture Model

Input:

- $\underline{\mathbf{X}}$ on the $\mathbb{S}^{D-1}$ - Training data.
- K-means or Random data points - The initial method.
- n_components(K) - The initial number of components.
- max_iter - The maximum number of EM iterations.
- tol - Convergence threshold.

Output:

- $\underline{\pi}$ - The weights of each mixture components.
 - $\underline{\mathbf{U}}$ - The direction matrices of BinMM.
 - $\underline{\mathbf{\Omega}}$ - The shape matrices of BinMM.
 - 1: Set the prior parameters: $\alpha_{0,k}$, $\mathbf{W}_{0,k}$ and $\nu_{0,k}$.
 - 2: Initialization r_{nk} by the initial method.
 - 3: **while** not converged **and** not reach max_iter **do**
 - 4: Evaluate the responsibilities (Equation 55) by computing Equation 52.
 - 5: Update posterior parameters α_k , $\mathbf{W}_k$, ν_k by Equation 59 and Equation 63.
 - 6: **end while**
 - 7: Calculate $\mathbb{E}[\underline{\pi}]$ and $\mathbb{E}[\underline{\mathbf{A}}]$ by Equation 65.
 - 8: Based on Equation 66, extracting BinMM parameters $\underline{\mathbf{U}}$, $\underline{\mathbf{\Omega}}$ by eigendecomposing $-\frac{1}{2}\underline{\mathbf{A}}$ obtained from zero-mean GMM, with mixture weights $\underline{\pi}$ preserved.
-

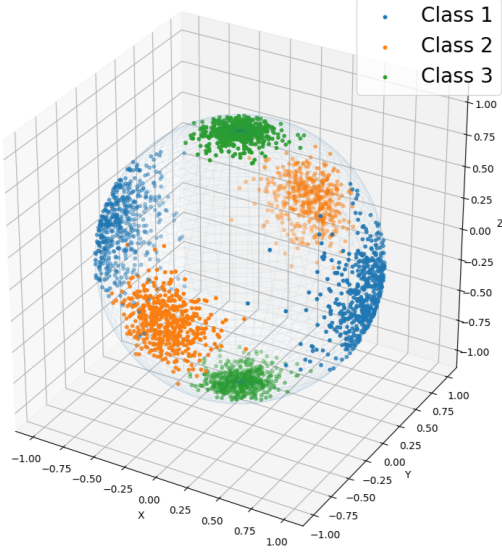
3.3 Experiments

This section evaluates the effectiveness of the proposed BinMM using both synthetic and real-world datasets. The synthetic datasets are generated using a Bayesian sampling-based method for the Bingham distribution [102] and a simulation procedure proposed by Kent et al. [113]. These experiments serve as a theoretical validation of the model. To highlight practical applicability, two downstream tasks are considered: (1) modeling symmetry-rich earthquake source mechanism data from New Zealand, and (2) clustering of gene expression profiles. The first experiment evaluates the proposed method’s capability to fit a single Bingham distribution, while the second experiment extends this evaluation to mixture models by benchmarking BinMM against competitive baselines in a real-world clustering scenario.

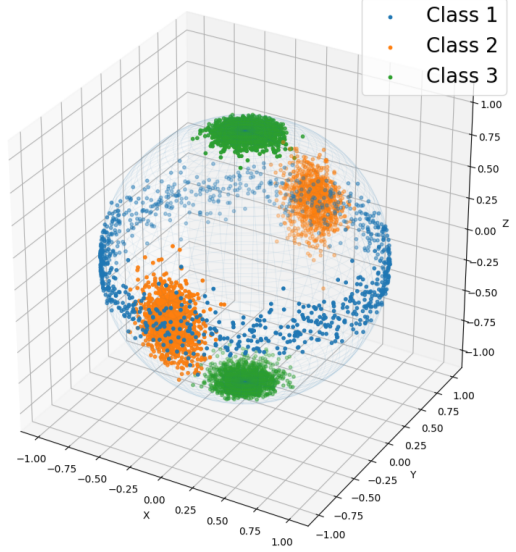
3.3.1 Synthetic Data

Synthetic experiments provide a controlled setting to verify the correctness of the proposed model, as the underlying ground truth parameters are known. This enables direct assessment of parameter recovery and clustering performance.

The synthetic data are composed of 3D directional samples drawn from multiple Bingham distributions,



(a) The first synthetic data set.



(b) The second synthetic data set.

Figure 3.4: The visualization of the synthetic data sets.

visualized on the unit sphere. In all experiments, the number of components K in the BinMM is initialized to match the true number of distributions. Initialization of means and responsibilities in the underlying GMM is performed using randomly selected data points. The EM optimization is configured with a maximum of 1500 iterations and a convergence threshold (tolerance) of 10^{-3} .

We design two synthetic datasets to evaluate the performance under varied distributional settings:

- Dataset 1: Contains three Bingham mixtures aligned along orthogonal axes on the unit sphere S^2 , with an equal number of samples per class.
- Dataset 2: Contains three Bingham mixtures exhibiting different geometric shapes (circular, elliptical, and annular) with varying sample sizes.

To facilitate interpretation of the results, we first provide an explanation of the relationship between Bingham shape types and their estimated parameters. Tables 3.1 and 3.2 report estimated parameters $\hat{\pi}_k, \hat{\omega}_1, \hat{\omega}_2, \hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \hat{\mathbf{u}}_3$, along with the corresponding ground truth values. Then:

- For a symmetric circular Bingham distribution with $\omega_1 = \omega_2$, the condition $|\hat{\mathbf{u}}_3| = |\mathbf{u}_3|$ is both necessary and sufficient for equivalence.
- For a symmetric elliptical distribution, equivalence requires $|\hat{\mathbf{U}}| = |\mathbf{U}|$.

Table 3.1: True and Estimated Parameters for the First Synthetic Data Set.

N_k	k	π_k	ω_1	ω_2	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\hat{\pi}_k$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\mathbf{u}}_1$	$\hat{\mathbf{u}}_2$	$\hat{\mathbf{u}}_3$
1000	1	0.333	-10	-10	[0, 0, 1]	[0, 1, 0]	[1, 0, 0]	0.33	-9.94	-9.27	[0, -0.19, 0.98]	[-0.02, -0.98, -0.19]	[1, -0.02, 0]
1000	2	0.333	-20	-20	[1, 0, 0]	[0, 0, 1]	[0, 1, 0]	0.33	-22.27	-19.44	[-0.35, 0, -0.94]	[0.94, 0.01, -0.35]	[0.01, -1, 0]
1000	3	0.333	-30	-30	[1, 0, 0]	[0, 1, 0]	[0, 0, 1]	0.33	-32.64	-28.20	[0.34, -0.94, 0]	[0.94, 0.34, 0]	[0, 0, 1]

- For an annular distribution with a unique non-zero eigenvalue ω_d , the condition $|\hat{\mathbf{u}}_d| = |\mathbf{u}_d|$ is sufficient for equivalence.

Results on Dataset 1

Figure 3.4a visualizes the first synthetic dataset. Each Bingham component is symmetrically distributed along a different axis of the unit sphere, with similar dispersion structures. The concentration parameters increase progressively from Class 1 to Class 3.

Table 3.1 presents both the ground-truth and estimated parameters. Since all three classes contain the same number of samples, the true mixing proportions are $\pi_k = 1/3$ for all k . The estimated proportions $\hat{\pi}_k$ closely match these values, indicating accurate component assignment. Additionally, the estimated concentration parameters $\hat{\omega}_1$ and $\hat{\omega}_2$ are consistent with the ground truth. The alignment between estimated and true eigenvectors, particularly $\hat{\mathbf{u}}_3 \approx \mathbf{u}_3$, further supports the validity of our model for symmetric circular Bingham distributions.

Results on Dataset 2

Figure 3.4b shows the second synthetic dataset, which features diverse and overlapping Bingham components. Notably, Class 2 (elliptical) and Class 3 (annular) partially intersect, posing a greater challenge for clustering.

Table 3.2 presents the estimated parameters. The estimated eigenvectors $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \hat{\mathbf{u}}_3$ are closely aligned with their true counterparts, demonstrating successful localization of the distribution centers. Despite partial overlap between Class 1 and Class 2, the model accurately distinguishes Class 3, as evidenced by the high precision of $\hat{\pi}_3, \hat{\omega}_1$, and $\hat{\omega}_2$. Minor deviations in $\hat{\pi}_1$ and $\hat{\pi}_2$ are attributable to sample overlap, but overall, the results confirm the model’s ability to handle heterogeneous and intersecting Bingham components.

Table 3.2: True and Estimated Parameters for the Second Synthetic Data Set.

N_k	k	π_k	ω_1	ω_2	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\hat{\pi}_k$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\mathbf{u}}_1$	$\hat{\mathbf{u}}_2$	$\hat{\mathbf{u}}_3$
1000	1	0.170	-40	0	[0, 0, 1]	[0, 1, 0]	[1, 0, 0]	0.20	-35.42	-2.09	[0, 0.01, 1]	[1, -0.02, 0]	[-0.02, -1, 0]
2000	2	0.330	-40	-20	[1, 0, 0]	[0, 0, 1]	[0, 1, 0]	0.30	-46.70	-19.55	[-1, 0, -0.02]	[-0.02, -0.01, 1]	[0, -1, -0.01]
3000	3	0.500	-40	-40	[1, 0, 0]	[0, 1, 0]	[0, 0, 1]	0.50	-41.64	-40.44	[-1, -0.01, 0]	[-0.01, 1, 0]	[0, 0, 1]

3.3.2 Earthquake Data Analysis

The Bingham distribution has been widely used in earthquake source mechanism analysis due to its capacity to model axial symmetry on the spherical manifold. In this section, we apply the proposed BinMM framework to real-world earthquake focal mechanism data from different regions of New Zealand [114], demonstrating the model’s effectiveness in capturing both the geometrical and statistical characteristics of seismic data.

Experimental Settings

The GeoNet project partitions New Zealand into ten regions: Wellington & Marlborough, Hawke’s Bay, Nelson & West Coast, Fiordland, Taranaki, Auckland & Northland, Canterbury, Otago & Southland, Tongariro & Bay of Plenty, and Gisborne. The dataset comprises 2,945 focal mechanism records for earthquakes of magnitude ≥ 4 , occurring between August 21, 2003, and June 20, 2023. To mitigate imbalances in sample size across the ten regions, we group neighboring areas into five clusters, ensuring a more balanced distribution of data while preserving regional characteristics. The resulting composition is summarized in Table 3.3. Each earthquake record was originally represented using plunge and azimuth angles, which describe orientation in a projected hemispherical coordinate system. To convert this into unit vector data on the full 3D sphere, we first map each sample to spherical coordinates and then apply symmetry doubling, a necessary transformation for modeling axially symmetric distributions.

Model initialization involves selecting random data points as initial component means (as opposed to k-means or k-means++), based on empirical performance. Other initialization settings follow the conventions of the proposed framework. The maximum number of EM iterations is set to 1500, and the convergence threshold is fixed at tolerance = 10^{-3} . In line with prior studies on Bingham-based seismic analysis [102, 103, 115], we fit a single Bingham distribution to each region for interpretability and comparison.

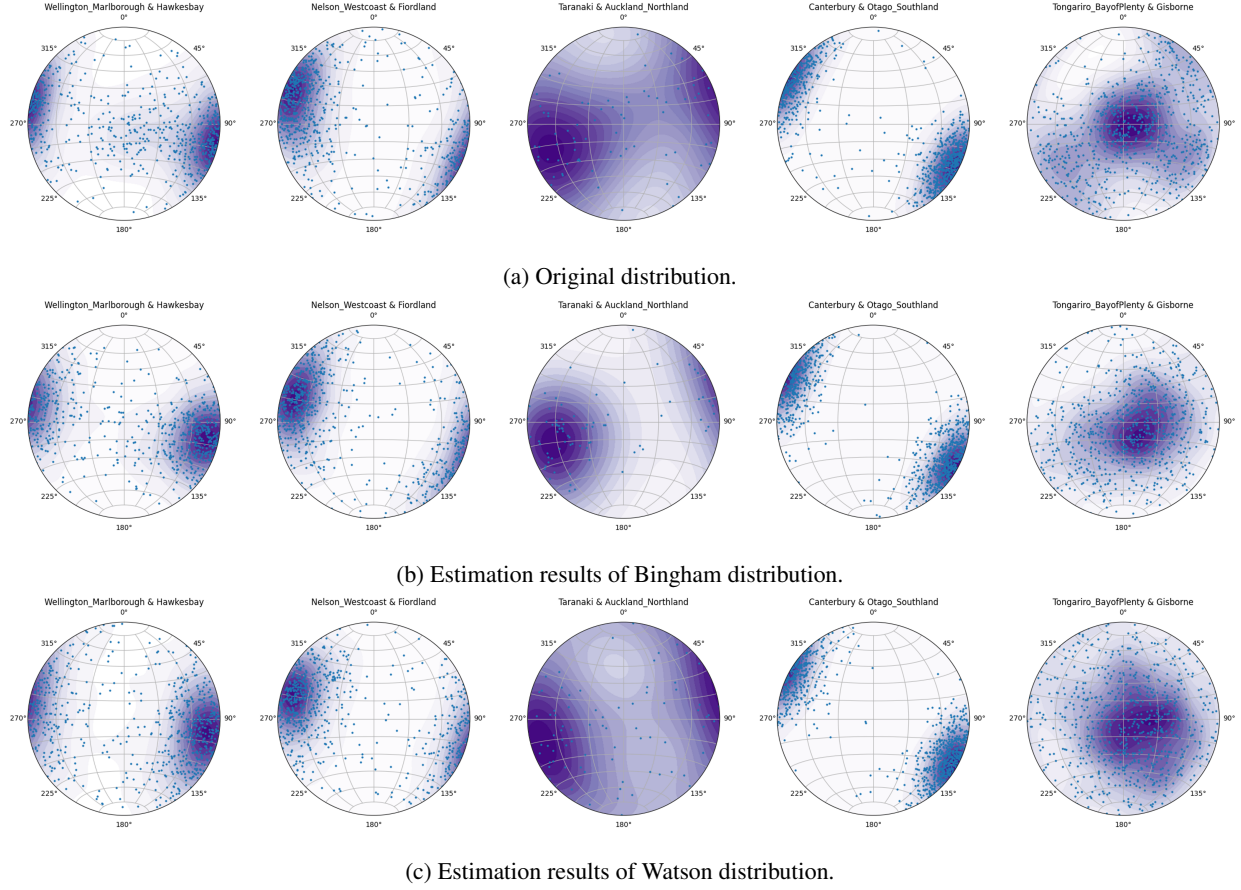


Figure 3.5: Comparison of original and estimated distributions of earthquake data by regions.

Experimental Results

Figure 3.5a visualizes the raw focal mechanism data for each of the five aggregated regions. The spatial distribution suggests that each region’s data approximately follows one or two axially symmetric clusters, making them amenable to Bingham modeling. We begin by fitting a single Bingham distribution to each region and present the estimated parameters and synthesized samples in Table 3.3 and Figure 3.5b, respectively. These results reveal several key insights:

- The Wellington–Marlborough & Hawke’s Bay and Nelson–West Coast & Fiordland regions exhibit similar dispersion patterns, as reflected in the comparable values of $\hat{\omega}_1$ and $\hat{\omega}_2$, and shared directional structure in the eigenvectors $\hat{\mathbf{u}}_1$, $\hat{\mathbf{u}}_2$, and $\hat{\mathbf{u}}_3$.
- In contrast, the remaining three regions display more distinct parameterizations, indicating region-specific seismotectonic behavior.

Table 3.3: Estimated parameters of Bingham distribution for the earthquake data.

Regions	N_k	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\mathbf{u}}_1$	$\hat{\mathbf{u}}_2$	$\hat{\mathbf{u}}_3$
Wellington_Marlborough & Hawkesbay	622	-5.28	-2.84	[-0.183, 0.107, -0.977]	[-0.189, -0.979, -0.072]	[-0.964, 0.172, 0.201]
Nelson_Westcoast & Fiordland	662	-5.318	-3.794	[0.296, -0.907, 0.300]	[0.191, 0.364, 0.912]	[0.936, 0.213, -0.281]
Taranaki & Auckland_Northland	74	-3.84	-2.60	[0.306, -0.048, -0.951]	[0.341, -0.927, 0.157]	[-0.889, -0.372, -0.268]
Canterbury & Otago_Southland	857	-8.308	-6.820	[0.145, -0.895, 0.422]	[-0.455, -0.439, -0.775]	[0.878, -0.079, -0.471]
Tongariro_BayofPlenty & Gisborne	730	-3.586	-2.584	[-0.422, -0.212, 0.881]	[-0.873, -0.166, 0.458]	[-0.243, 0.963, 0.115]

Table 3.4: Wasserstein distance from the original earthquake data by Bingham and Watson distribution. (Average results of 10 experiments)

Regions	Bingham distribution	Watson distribution
Wellington_Marlborough & Hawkesbay	0.0234	0.0439
Nelson_Westcoast & Fiordland	0.0130	0.0188
Taranaki & Auckland_Northland	0.0659	0.0841
Canterbury & Otago_Southland	0.0122	0.0201
Tongariro_BayofPlenty & Gisborne	0.0219	0.0307

- These differences are clearly observable in the synthetic visualizations (Figure 3.5b), where each fitted Bingham distribution captures unique regional dispersion characteristics.

From a geophysical perspective, the estimated Bingham parameters offer valuable insights into local tectonic stress orientations. Because the Bingham distribution captures both direction and symmetry, geoscientists can utilize the learned parameters to monitor temporal or spatial variations in regional stress fields, aiding in the assessment of seismic hazards.

To quantitatively evaluate model fit, we compute the p -Wasserstein distance between the synthetic distribution (sampled from the fitted Bingham model) and the empirical data distribution. The Wasserstein distance between two probability measures μ and ν on $\mathbb{R}^d$ is defined as [116]:

$$W_p(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} (\mathbb{E} \|X - Y\|^p)^{1/p}, \quad p \geq 1, \quad (67)$$

where the inf is taken over all pairs of d -dimensional random vectors X and Y marginally distributed as μ and ν , respectively. A lower Wasserstein distance indicates a closer match between the two distributions.

To further contextualize the performance of the Bingham model, we conduct a comparative experiment using the Watson distribution, a commonly used axial symmetric distribution on the sphere. While the Watson distribution captures circular symmetry, it is unable to model elliptical structures, thus making it a

strict special case of the Bingham distribution.

Figure 3.5c shows samples generated from the fitted Watson distributions, and Table 3.4 reports the average Wasserstein distances for both models. As expected, the Bingham distribution consistently achieves lower Wasserstein distances across all five regions, confirming its superior capacity to model complex elliptical and annular patterns in focal mechanism data.

From the visual comparisons in Figures 3.5a, 3.5b, and 3.5c, the Bingham distribution demonstrates superior fitting, especially in regions with elliptical symmetry, such as Taranaki & Auckland–Northland. Even in less structured regions, such as Tongariro–Bay of Plenty & Gisborne, the Bingham model yields a closer fit than the Watson model. These results confirm the proposed model’s effectiveness in modeling real-world axially symmetric data, particularly when the underlying structure deviates from strict circularity.

3.3.3 Gene Expression Data Clustering

Gene expression data often exhibits diverse structural patterns and can be meaningfully grouped into functionally similar categories. However, manual labeling of such data is costly and time-consuming due to the high dimensionality and complexity of biological systems. As such, unsupervised clustering of gene expression data provides a practical and valuable alternative for exploratory analysis.

Directional distributions, such as the vMF mixture model, have been shown to be effective for clustering gene expression data because spatial features often carry more biological significance than absolute expression magnitudes [117, 39]. Moreover, gene expression vectors often exhibit axial symmetry, where both positive and negative correlations are meaningful [118]. This justifies the use of axial symmetric distributions such as the Watson mixture model (WMM) [57, 39]. As a generalization of the Watson distribution, the Bingham distribution offers greater flexibility by capturing both elliptical and circular symmetry [32].

Accordingly, our proposed variational inference framework for the BinMM provides a principled and expressive tool for modeling gene expression data. It not only captures richer spatial structures but also supports automatic inference of the number of clusters in a fully unsupervised manner.

Experimental Settings

We evaluate the proposed BinMM model on four publicly available gene expression datasets obtained from the National Center for Biotechnology Information (NCBI)¹:

¹<https://www.ncbi.nlm.nih.gov>

Table 3.5: H_{ave} and S_{ave} of gene expression data in terms of different methods. (Average results of 10 experiments)

Model	YMG		YS		HFS		YCC	
	H_{ave}	S_{ave}	H_{ave}	S_{ave}	H_{ave}	S_{ave}	H_{ave}	S_{ave}
K-means	+0.74	-0.71	+0.73	-0.60	+0.46	-0.30	+0.15	-0.12
GMM[83]	+0.73	-0.70	+0.67	-0.58	+0.44	-0.28	+0.09	-0.06
vMFMM[39]	+0.79	-0.76	+0.82	-0.66	+0.60	-0.43	+0.29	-0.13
WMM[40]	+0.80	-0.76	+0.81	-0.64	+0.62	-0.46	+0.31	-0.10
DP-Wat[57]	+0.81	-0.75	+0.83	-0.66	+0.64	-0.45	+0.30	-0.15
HDP-Gau[123]	+0.83	-0.77	+0.83	-0.69	+0.67	-0.45	+0.40	-0.19
HDP-Wat[124]	+0.84	-0.70	+0.86	-0.68	+0.71	-0.43	+0.44	-0.23
HPYP-Wat[124]	+0.85	-0.79	+0.87	-0.66	+0.72	-0.43	+0.47	-0.26
BinMM	+0.84	-0.85	+0.83	-0.87	+0.75	-0.50	+0.44	-0.42

- Yeast Metabolic and Genetic Control (YMG) [119]: 6,400 genes at 7 time points; 614 genes retained after preprocessing.
- Yeast Sporulation (YS) [120]: 6,384 genes at 7 time points; 1,156 genes retained.
- Human Fibroblast Serum (HFS) [121]: 8,613 genes at 18 time points; 349 genes retained.
- Yeast Cell Cycle (YCC) [122]: 4,382 genes at 23 time points; after removing the first 5 low-variance time points, 666 genes at 18 time points remained.

Preprocessing includes removal of missing values and filtering out genes with low spatial variance. All vectors are L_2 normalized to lie on the unit hypersphere. We initialize BinMM using K-means clustering with 10 clusters. The maximum number of EM iterations is set to 1500, and the convergence tolerance is set to $1e^{-3}$.

To assess clustering performance, we compare BinMM with the classical K-means algorithm and five other mixture model-based clustering methods, Gaussian mixture model (GMM) [83], von Mises-Fisher mixture model (vMFMM)[39], Watson mixture model (WMM) [40], Dirichlet process mixture model with Watson distributions (DP-Wat)[57], HDP mixture model with Gaussian distributions (HDP-Gau)[123], HDP and HPYP mixture model with Watson distributions (HDP-Wat and HPYP-wat)[124]. The initialization parameters of these methods are set as follows: K in the K-means algorithm is set to 4. α_0 and β_0 in the GMM are initialized to 0.01 and 0.1. The hyperparameters $\alpha_{0,k}$, $\beta_{0,k}$, $a_{0,k}$, and $b_{0,k}$ of the vMFMM are set to 0.5, 0.1, 1, and 0.01. The hyperparameters $\alpha_{0,k}$, $a_{0,k}$, and $b_{0,k}$ in the WMM are initialized to 0.1, 1, and 0.01. a_k , b_k , c_k , and d_k of DP-Wat are initialized to 0.1, 0.1, 0.01, and 0.1. α_0 and α_1 in HDP-Gau are

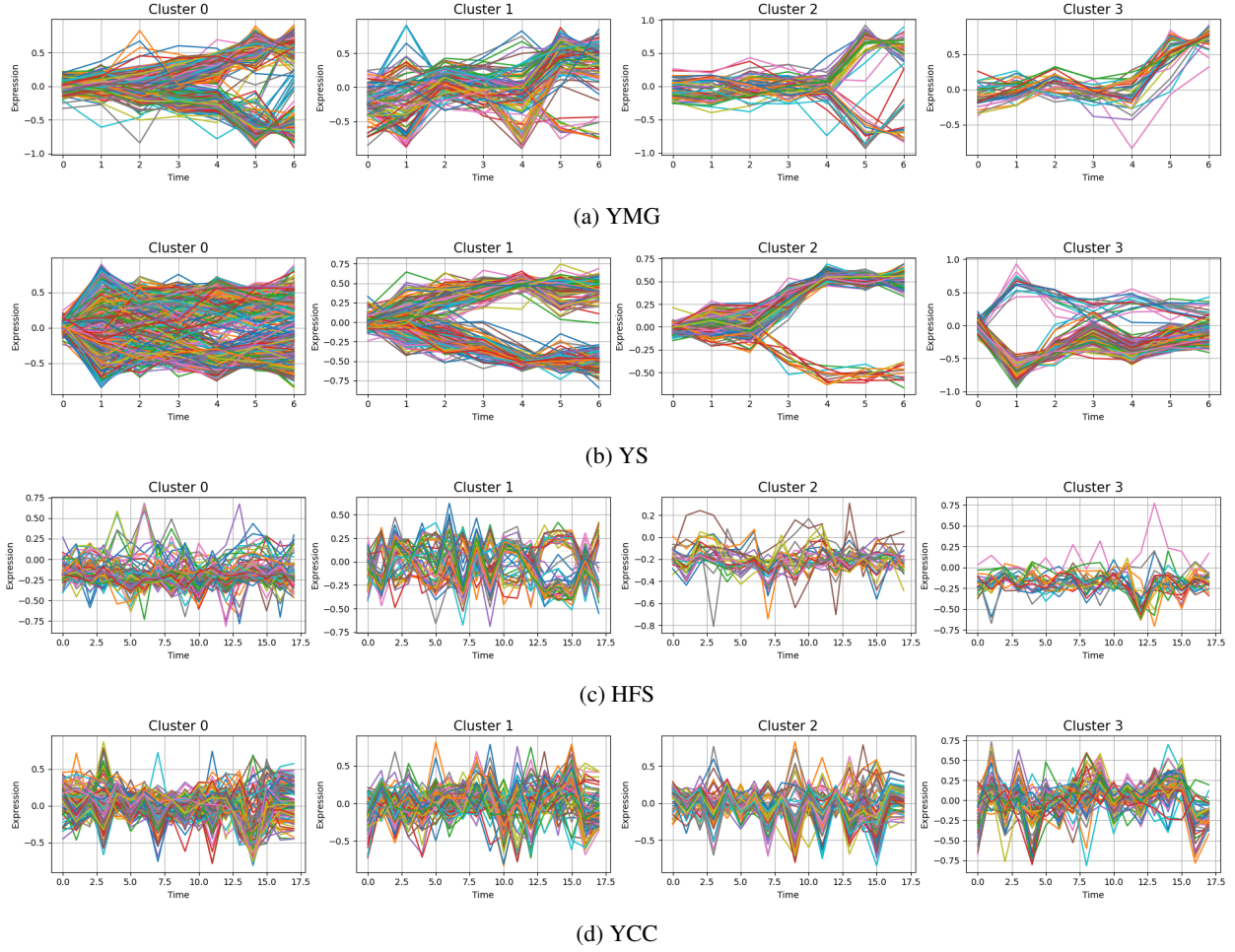


Figure 3.6: Top 4 clustering results obtained from the YMG, YS, HFS and YCC data sets by the proposed model.

initialized to 15 and 0.01. The hyperparameters α , β , ϕ , μ and ν of HDP-Wat are respectively initialized to 1, 0.01, 1, 1 and 0.01. The hyperparameters α , β , η , ω , ϕ , μ and ν of HPYP-Wat are set to 1, 0.01, 0.5, 0.1, 1, 1 and 0.001. The initial number of components for GMM, vMFMM, WMM and DP-Wat is set to 10. For HDP-Gau, 30 and 15 are given to the number of base-level and group-level components. For HDP-Wat and HPYP-Wat, the number of base-level and group-level components are both set to 10.

Table 3.6: Computational run time (in seconds) of different methods on clustering gene expression data.

Model	YMG	YS	HFS	YCC
K-means	0.45	0.46	0.46	0.47
GMM[83]	0.15	0.20	0.13	0.16
vMFMM[39]	2.20	2.33	1.21	2.00
WMM[40]	2.29	2.37	1.19	2.21
DP-Wat[57]	3.14	3.41	1.59	2.40
HDP-Gau[123]	12.61	15.83	10.43	11.79
HDP-Wat[124]	6.16	7.30	4.96	5.43
HPYP-Wat[124]	6.20	7.49	4.83	5.48
BinMM	0.42	0.71	0.14	0.40

Experimental Results

Since ground-truth labels are not available for these datasets, we evaluate clustering performance using two widely adopted internal metrics [98, 40]: *Homogeneity* H_{ave} and *Separation* S_{ave} , which are given by

$$H_{ave} = \frac{1}{N} \sum_{k=1}^K \sum_{\mathbf{x}_n \in \underline{\mathbf{X}}_k} \left(\mathbf{x}_n^\top \boldsymbol{\mu}_k \right)^2, \quad (68)$$

and

$$S_{ave} = \frac{1}{\sum_{k \neq s} |\underline{\mathbf{X}}_k| |\underline{\mathbf{X}}_s|} \sum_{k \neq s} |\underline{\mathbf{X}}_k| |\underline{\mathbf{X}}_s| \min \left(\boldsymbol{\mu}_k^\top \boldsymbol{\mu}_s, -\boldsymbol{\mu}_k^\top \boldsymbol{\mu}_s \right), \quad (69)$$

where $\underline{\mathbf{X}}_k$ is the set of samples in the k -th cluster, and $\boldsymbol{\mu}_k$ is the cluster mean direction. Higher H_{ave} and lower S_{ave} indicate better clustering performance.

Table 3.5 presents the results for all methods across the four datasets. The proposed BinMM consistently outperforms all other baseline models. It especially excels when compared to non-directional models (e.g., GMM), reaffirming the advantage of using directional distributions for clustering spatially embedded data. While the HPYP-Wat model achieves slightly better scores on certain metrics due to its hierarchical structure, BinMM offers a strong trade-off between performance and computational efficiency. In particular, as shown in Table 3.6, BinMM is significantly faster than hierarchical models and remains competitive even compared to simpler models such as K-means and GMM.

To assess model robustness, we conducted experiments across varying numbers of initial components $K \in 10, 20, 30, 40, 50$. The results, visualized in Figure 3.7, demonstrate the stability of BinMM across a wide range of configurations, particularly on the YMG and YS datasets.

Finally, we visualize the top 4 discovered clusters for each dataset in Figure 3.6. On the YMG and

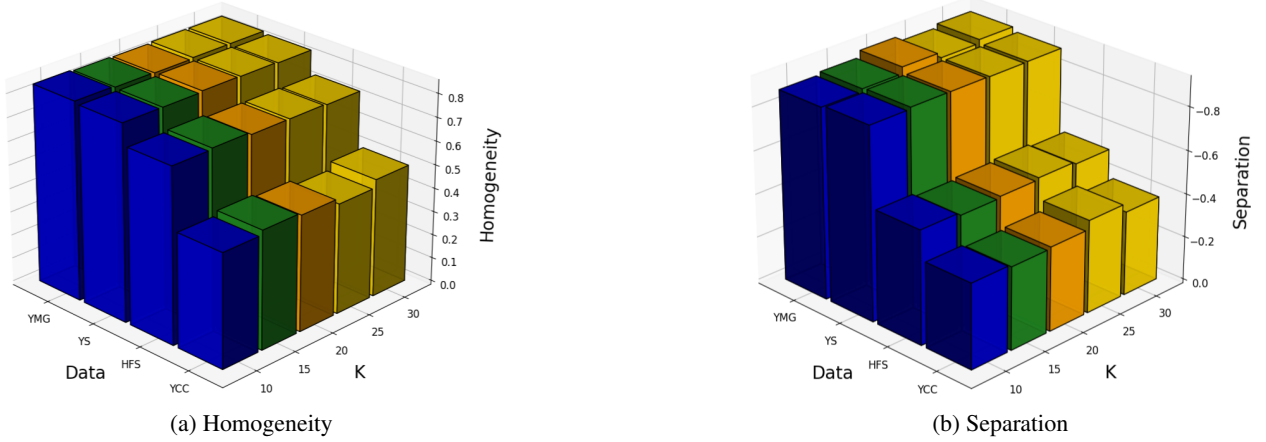


Figure 3.7: Performance of the proposed model on YM, YS, HFS and YCC data sets in terms of Homogeneity and Separation with different values of K .

YS datasets, the clusters exhibit strong symmetric profiles with distinct positive and negative correlation patterns. While the HFS and YCC datasets show less pronounced fluctuations, BinMM still captures clear shape-based structures. These results confirm that the proposed model effectively captures both positive and negative correlation patterns and groups gene profiles with similar regulatory behaviour.

3.4 Conclusion

In this chapter, we propose a novel and principled variational inference framework for the BinMM. In contrast to previous approaches that treat directional data modeling in isolation, our method establishes a systematic connection between classical Gaussian representations and hyperspherical distributions. By leveraging this relationship, we reformulate the Bingham mixture model to enable efficient variational inference without the need to compute the intractable normalization constant, thereby improving both scalability and tractability. We validated the effectiveness of our method through comprehensive experiments on both synthetic and real-world datasets. On synthetic data, the model accurately recovered known Bingham mixture structures, demonstrating its theoretical soundness. In the application to earthquake focal mechanism data, the model effectively captured axial symmetries and provided physically meaningful interpretations. In the gene expression clustering task, our approach outperformed existing directional and hierarchical mixture models in terms of clustering quality and runtime efficiency. Future directions include extending the model to incorporate hierarchical structures for richer clustering semantics and exploring reparameterization techniques for integration into variational autoencoder frameworks.

Chapter 4

BRVSNet: A Self-Supervised Framework for Probabilistic Rotation Estimation and Consistent Multi-View Generation

4.1 Introduction

The increasing expressiveness of deep neural networks has significantly improved their ability to quantify uncertainty and generalize to unseen scenarios, thereby offering valuable insights for downstream tasks in computer vision and robotics [18, 19, 125]. Among these, rotation-related tasks, such as 3D object pose estimation and multi-view image generation, are particularly crucial, given their foundational role in understanding spatial transformations and enabling robust perception in real-world environments. Recent progress in rotation estimation [126, 127, 128] and multi-view generation [129, 130, 131] has expanded the applicability of these tasks to a wide range of domains, including robotics, autonomous navigation, augmented reality, and 3D reconstruction.

Despite the progress in deterministic models, modeling rotation remains a challenging problem due to its inherent geometric complexity and multimodal uncertainty. Probabilistic modeling of rotation in the $SO(3)$ space has emerged as a promising solution, enabling the representation of uncertainty in rotation predictions and improving interpretability. Several approaches have been proposed: normalizing flows on $SO(3)$ [128], von Mises distributions over Euler angles [132], matrix Fisher distributions for rotation matrices [133, 134],

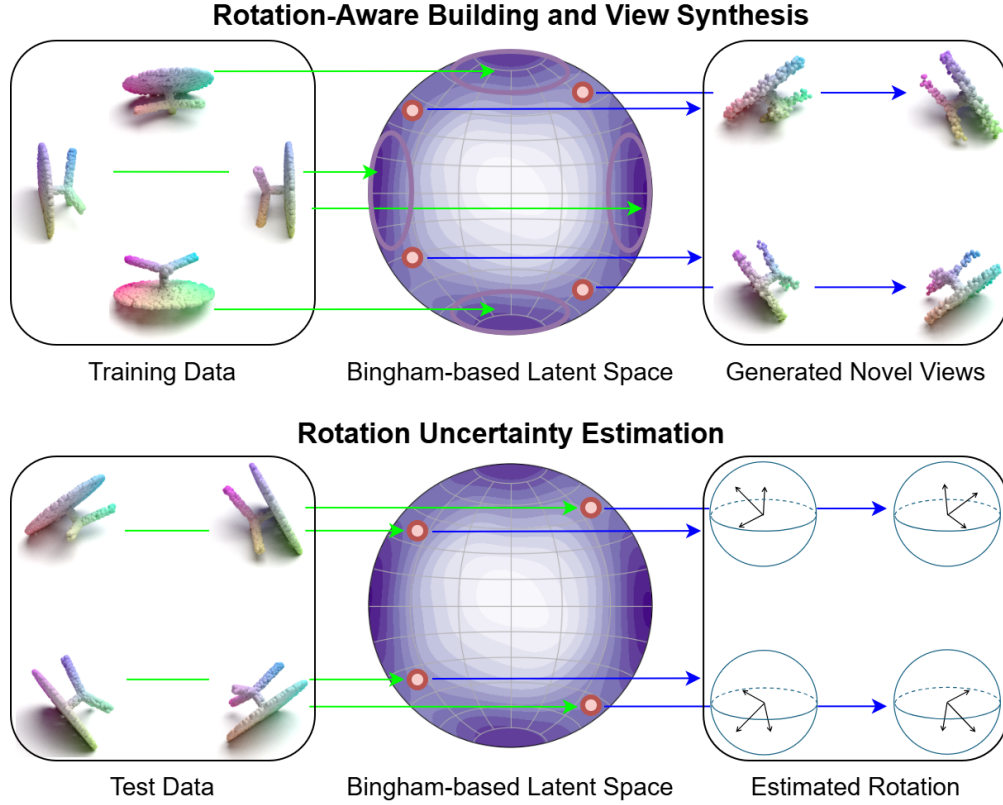


Figure 4.1: **Illustration of the Rotation-Awareness in the Proposed Framework.** The framework builds upon a Bingham-based latent space to enable rotation-aware learning and downstream tasks. (Top) In the rotation-aware building and view synthesis module, training data are embedded into the latent space, and the generator leverages its rotation-awareness to synthesize novel views from arbitrary latent directions. (Bottom) In the rotation uncertainty estimation module, test data are projected into the latent space, where the Bingham distribution models rotation hypotheses and quantifies uncertainty for robust rotation regression.

and Bingham distributions for quaternions [38, 135]. Each of these methods offers unique advantages in capturing rotational uncertainty while preserving manifold geometry. Simultaneously, multi-view generation, particularly from a single image input, has seen notable advances with the rise of powerful generative models. Methods such as Zero-1-to-3 [136], One-2-3-45++ [130], and SyncDreamer [137] leverage large-scale 2D pre-trained models (e.g., CLIP [138], Stable Diffusion [139]) to synthesize consistent multi-view representations. However, when these models are extended to camera relocalization or point cloud pose estimation, they encounter limitations. These tasks involve different input modalities and require highly accurate rotational representations to ensure consistency across views [136, 137]. In such contexts, traditional generative models often struggle to capture the intricate uncertainty and symmetry properties inherent in rotational data. Moreover, the lack of fine-grained rotational supervision hampers the performance of

multi-view generation systems in real-world 6D pose estimation tasks.

This gap reveals an important opportunity: by integrating principled probabilistic models of rotation into deep generative frameworks, it is possible to enhance the rotational accuracy of multi-view generation and simultaneously benefit rotation estimation through auxiliary supervision. Probabilistic rotation models offer inherent advantages (such as rotational invariance, geometric consistency, and smooth latent space optimization) that can act as strong inductive biases for generative tasks [132, 134, 38]. Furthermore, if view-specific representations are generated with awareness of rotational characteristics, these representations can provide a form of pseudo-labeling or self-supervised signals for refining rotation regression [140, 141]. Despite these advantages, the integration of rotation-aware probabilistic modeling into multi-view generation pipelines remains underexplored.

Motivated by the complementary strengths of rotation regression and view synthesis, we propose a unified probabilistic framework that jointly addresses rotation estimation and multi-view generation. Figure 4.1 summarizes the problems we address in this work. Specifically, we explore how incorporating traditional rotational probabilistic models (particularly the Bingham mixture model for quaternions) can enhance the expressiveness and reliability of generative systems for tasks involving pose understanding. We adopt quaternions as our rotation representation due to their compactness and absence of singularities such as gimbal lock, which commonly afflict Euler angle-based models [142]. Compared to rotation matrices or $SO(3)$ -based methods, quaternion-based approaches are more amenable to neural parameterization and backpropagation, making them well-suited for integration into deep learning pipelines. Moreover, the Bingham mixture model provides a flexible mechanism for capturing multi-modal directional uncertainty and representing ambiguous rotations, which are crucial for real-world applications involving occlusions or symmetries.

To instantiate these ideas, we build upon Deep Bingham Networks (DBNs) [38], leveraging them to extract quaternion-based rotational features. These features are then used to construct a latent space parameterized by Bingham noise, enabling uncertainty-aware rotation modeling and guiding the generation of multi-view representations. By leveraging both the probabilistic structure of the Bingham distribution and the generative modeling capabilities of deep generative models, our method simultaneously supports accurate pose estimation and view synthesis, while enabling a self-supervised feedback loop.

Our contributions are summarized as follows:

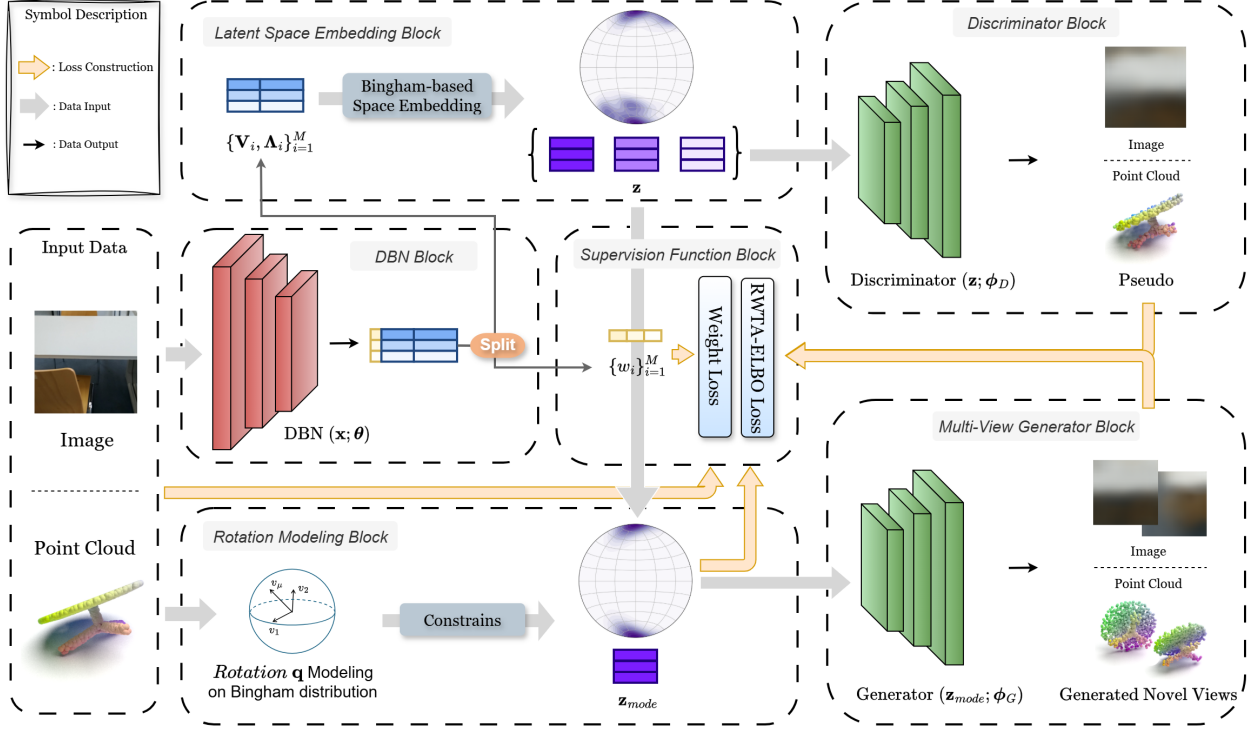


Figure 4.2: **Framework Pipeline.** The framework consists of four main components: (1) the DBN block processes images or point clouds to generate mixture parameters; (2) the Bingham-based space embedding, together with the rotation modeling block, transforms mixture embeddings into latent orientation distributions and aligns them with ground-truth quaternion constraints; (3) the supervision function block refines estimation via RWTA-ELBO loss and weight loss; and (4) the discriminator block produces pseudo ground truths to enhance latent representation, while only the mode latent z_{mode} is used as input to the Multi-View Generator, which, after training, develops strong rotation-awareness on the Bingham distribution and synthesizes novel views (images or point clouds) that are geometrically consistent with the estimated rotations. During inference, it is sufficient to provide only the corresponding quaternion to synthesize the desired multi-view outputs.

- We propose a unified framework that integrates Bingham mixture modeling into both rotation regression and multi-view generation. This framework enables accurate uncertainty modeling in quaternion space and facilitates differentiable optimization via deep neural networks.
- We introduce a novel Bingham noise-based latent module to capture uncertain rotational features, which supports effective gradient flow and underpins the construction of a generator-discriminator architecture used for self-supervision.
- We develop a custom Relaxed Winner-Take-All Evidence Lower Bound (RWTA-ELBO) loss that enables joint training of the DBN and discriminator modules, balancing rotation regression with consistency-driven supervision from generated views.

- We design two new weight regularization losses tailored to optimize the mixture model, allowing the system to select the most representative rotation based on learned mixture weights.

The overall architecture of our proposed framework is illustrated in Figure 4.2.

4.2 Related Work

The core challenge in rotation regression lies in the selection of an appropriate probabilistic distribution for modeling rotations, as different distributions encode rotations in distinct forms. Prokudin *et al.* [132] utilized a mixture of von Mises distributions to model rotations via Euler angles. Deng *et al.* [38] proposed modeling quaternion-based rotations using a mixture of Bingham distributions, which are well-suited for unit hyperspherical data. Mohlin *et al.* [133] adopted the matrix Fisher distribution to directly model rotation matrices, offering a mathematically tractable representation in $SO(3)$. To enable greater expressiveness and flexibility in modeling arbitrary rotation distributions, Liu *et al.* [128] introduced normalizing flows on the $SO(3)$ manifold, combining the power of flow-based models with geometric constraints for probabilistic rotation learning.

In parallel, significant progress has been made in the field of multi-view generation, largely driven by advances in 2D prior models and their generative capabilities. For example, Zero-1-to-3 by Liu *et al.* [136] incorporated rotation and translation embeddings into a conditional diffusion framework to synthesize novel views from a single image. One-2-3-45++ [130] enhanced generation consistency by fixing camera poses during training. SyncDreamer [137] further improved geometry and texture alignment through a 3D-aware feature attention mechanism. Beyond image-based generation, methods such as DiffusionPM [143] and WarpingGAN [144] extended generative modeling to point clouds using diffusion and GAN-based approaches, respectively, showing promising potential for geometric view synthesis.

While prior work in multi-view generation has largely focused on realism and perceptual quality using generative models such as diffusion and GANs, our framework emphasizes the generation of view-specific metadata that aligns accurately with the underlying rotational input. Our work is inspired by and integrates the strengths of the aforementioned rotation modeling and view generation paradigms. Specifically, we aim to bridge the gap between these two areas by introducing a framework that leverages uncertainty-aware quaternion modeling to guide multi-view synthesis, while simultaneously using generated views to improve the accuracy and robustness of rotation regression. This bidirectional synergy highlights the unexplored

potential of combining geometric learning and generative modeling under a principled probabilistic framework.

4.3 The Proposed Method

In this section, we present a general framework for joint rotation regression and multi-view generation, trained using two types of labeled datasets: (i) RGB images with associated camera locations, and (ii) point clouds with annotated object poses. Once trained, the model predicts the rotation of an input instance and generates multiple novel views parameterized by quaternions.

4.3.1 Revisit of DBN

Our framework is motivated by the DBN method proposed in [38] for rotation extraction, and we briefly review its core principles here. DBN provides a generic solution for handling pose-related uncertainties in 3D data applications, comprising two main modules.

The first module is a backbone network, selected according to the input modality. For RGB image inputs, DBN employs ResNet [145], whereas for point cloud inputs, it utilizes a tailored variant of PointNet [42]. This module encodes the input into a 7-dimensional feature vector.

The second module maps the 7-dimensional vector into the parameters of a Bingham distribution and its mixture model. A key contribution of DBN lies in demonstrating that a 7-dimensional vector suffices to parameterize a Bingham distribution, despite its inherent complexity. Assuming the obtained 7-d vector $(\mathbf{z}_1, \mathbf{z}_2)$, where the first 4-d vector $\mathbf{z}_1$ are normalized as a quaternion first and then following *Birdal Strategy* [146] to construct the parameter metric $\mathbf{V}$ as:

$$\mathbf{V}(\mathbf{z}_1) \triangleq \begin{bmatrix} z_{11} & -z_{12} & -z_{13} & z_{14} \\ z_{12} & z_{11} & z_{14} & z_{13} \\ z_{13} & -z_{14} & z_{11} & -z_{12} \\ z_{14} & z_{13} & -z_{12} & -z_{11} \end{bmatrix}. \quad (70)$$

The last 3-d vector $\mathbf{z}_2$ is first activated by the *softplus* function ϕ , and then transformed into a parameter

matrix $\mathbf{\Lambda}$ with decreasing diagonal values obtained through the negative accumulative sum:

$$\begin{aligned}\lambda_1 &= -\phi(z_{21}) - \phi(z_{22}) - \phi(z_{23}) \\ \lambda_2 &= -\phi(z_{21}) - \phi(z_{22}) \\ \lambda_3 &= -\phi(z_{21}).\end{aligned}\tag{71}$$

While the Bingham distribution is well known for its expressiveness in representing rotations, the parameter constraints make direct neural parameterization challenging. DBN effectively overcomes these constraints, pioneering the integration of Bingham distributions into neural network-based rotation estimation.

Motivated by DBN, we extend its principles to jointly address the challenges of rotation regression and multi-view generation. Specifically, we propose BRVSNet (Bingham Rotation and View Synthesis Network), a novel architecture that combines a regressor with a discriminator to enable self-supervised rotation regression, while simultaneously facilitating multi-view generation from quaternion representations.

4.3.2 Probabilistic Modeling of Rotation

To capture the uncertainty inherent in rotation estimation, we adopt the Bingham distribution to construct a probabilistic model of quaternion-based rotation prediction, following [38].

Let $\mathcal{Q} = \mathbf{q}_n$ denote a set of unit quaternions on $\mathbb{S}^3$. The probability density function (PDF) of the Bingham distribution with parameter matrices $\mathbf{V}$ and $\mathbf{\Lambda}$ is given by:

$$\mathcal{B}(\mathbf{q} \mid \mathbf{V}, \mathbf{\Lambda}) = \frac{1}{F(\mathbf{\Lambda})} \exp(\mathbf{q}^T \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \mathbf{q}),\tag{72}$$

where $\mathbf{V} \in \mathbb{R}^{4 \times 4}$ is an orthogonal matrix, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3, 0)$ is a diagonal matrix with ordered negative entries ($\lambda_1 \leq \lambda_2 \leq \lambda_3 < 0$). The term $F(\mathbf{\Lambda})$ is the normalization constant. The last column of $\mathbf{V}$ specifies the mode of the distribution, while the first three columns define the dispersion orientation, with λ_d ($d \in 1, 2, 3$) controlling the dispersion magnitude along each axis.

A Bingham mixture model extends this formulation by combining multiple Bingham components, each weighted by a mixture coefficient. Given N observations $\mathcal{X} = \{\mathbf{x}_n\}$, where each $\mathbf{x}_n$ is associated with M

potential poses, the mixture PDF is:

$$\mathcal{B}_{\text{mix}}(\mathbf{q} \mid \mathbf{x}_{\mathbf{n}}, \boldsymbol{\theta}) = \sum_{i=1}^M w_i(\mathbf{x}_{\mathbf{n}}, \boldsymbol{\theta}) \cdot p_i(\mathbf{q} \mid \mathbf{x}_{\mathbf{n}}, \boldsymbol{\theta}), \quad (73)$$

where $p_i(\mathbf{q} \mid \mathbf{x}_{\mathbf{n}}, \boldsymbol{\theta}) = \mathcal{B}(\mathbf{q} \mid \mathbf{V}_i, \boldsymbol{\Lambda}_i)$, the mixture weights $\{w_i\}_{i=1}^M$ satisfy $w_i > 0$ and $\sum_{i=1}^M w_i = 1$, and $\boldsymbol{\theta}$ denotes all learnable parameters, including both shared and component-specific terms.

There exist multiple approaches for representing probability distributions of rotation [147], including geometric methods [148] and measure-theoretic approaches [149]. Quaternions offer a continuous, minimally redundant, and singularity-free parameterization, making them an appropriate choice for direct regression tasks [150, 151]. The antipodal symmetry of the Bingham distribution [32] aligns naturally with quaternion topology, enabling robust modeling of directional uncertainties. Previous works have successfully applied the Bingham distribution to quaternion-based rotation modeling [152, 146, 153].

In our framework, quaternions serve as the primary rotation representation, with the Bingham distribution forming the probabilistic foundation for our model. Compared to the matrix Fisher distribution for rotation matrices, the Bingham–quaternion formulation offers greater mathematical convenience, particularly when constructing a structured latent space for downstream tasks.

4.3.3 Bingham-Based Latent Space

The Bingham-based latent space serves as the key interface linking the DBN, discriminator, and generator. Once DBN extracts the rotation features of an input and maps them to the parameters of a Bingham mixture model, it is necessary to account for the stochasticity in these learned parameters. This treatment ensures that gradients can propagate through the random variables, enabling the parameters to be updated via backpropagation in the subsequent discriminator and generator modules [12].

Constructing the Latent Space with ACG Sampling

In conventional sampling methods, the random sampling process is decoupled from the model parameters, making it non-differentiable and thus preventing parameter updates through standard backpropagation. The key to addressing this limitation lies in the reparameterization trick, which introduces an independent noise variable such that the randomness is isolated in the noise. This allows the parameters predicted by the neural network to remain differentiable and optimizable [12, 154].

However, the Bingham distribution poses a challenge: its normalizing constant is analytically intractable. To circumvent this, we incorporate the Angular Central Gaussian (ACG) distribution as a surrogate for constructing the Bingham-based latent space. This choice is motivated by two key factors: (i) the ACG distribution is well-suited for modeling antipodally symmetric directional data [155]; and (ii) it has a known, closed-form normalizing constant, facilitating efficient computation. Leveraging these properties, we adopt the acceptance–rejection sampling strategy [113], which employs the ACG distribution as an envelope density to sample from the Bingham distribution, thereby enabling a differentiable latent space construction.

Angular Central Gaussian (ACG) Distribution

Let $\mathbf{y} \in \mathbb{R}^4$ follow a zero-mean multivariate normal distribution with covariance matrix Ψ . The normalized quaternion $\mathbf{q} = \mathbf{y}/\|\mathbf{y}\|$ then follows an ACG distribution on the unit 3-sphere $\mathbb{S}^3$ [156]. Its probability density function is:

$$g(\mathbf{q}; \Psi) = |\Psi|^{-1/2} (\mathbf{q}^T \Psi^{-1} \mathbf{q})^{-2} = c_{\text{ACG}}(\Psi) g^*(\mathbf{q}; \Psi), \quad (74)$$

where $c_{\text{ACG}}(\Psi) = |\Psi|^{-1/2}$ is the normalizing constant, Ψ is a 4×4 symmetric positive-definite parameter matrix.

Acceptance–Rejection Sampling for the Bingham-Based Latent Space

Motivated by [113], samples from a Bingham distribution with parameter matrix $\mathbf{A} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$ can be generated using the ACG distribution as an envelope density. The procedure is as follows:

(1) Initialization:

- Set an acceptance list $\mathbf{L}$.
- Compute $\underline{\Psi}^{-1} = \mathbf{I} + \frac{2}{b} \mathbf{A}$, where b is a scaling parameter.
- Choose a constant $\mathbf{M}$ such that $\mathbf{M}^* \geq \text{upper}\{\frac{f^*(\mathbf{q}_n)}{g^*(\mathbf{q}_n)}\}$.

(2) Sampling:

- Draw $\mathbf{u} \sim \text{Uniform}(0, 1)$.
- Generate candidate quaternions $\mathcal{Q}$ from the ACG distribution with parameter matrix $\underline{\Psi}$.

(3) Acceptance Test:

- Accept $\mathbf{q}_n \in \mathcal{Q}$ if:

$$\mathbf{u} < \frac{f^*(\mathbf{q}_n; \mathbf{A}_n)}{\mathbf{M}^* g^*(\mathbf{q}_n; \mathbf{\Psi}_n)}, \quad (75)$$

where $f^*(\mathbf{q}_n; \mathbf{A}_n) = \exp(-\mathbf{q}_n^T \mathbf{A}_n \mathbf{q}_n)$, and $g^*(\mathbf{q}_n; \mathbf{\Psi}_n) = (\mathbf{q}_n^T \mathbf{\Psi}_n^{-1} \mathbf{q}_n)^{-2}$.

- Otherwise, return to Step 2. The process terminates when $\mathbf{L}$ contains only accepted samples.

Here, $\mathbf{\Psi}$ and $\mathbf{A}$ denote the parameter matrices of the ACG and Bingham distributions, respectively. In practice, setting $b = 1$ works well for most cases. Introducing the acceptance list $\mathbf{L}$ significantly improves parallel sampling efficiency and accelerates optimization across the network.

4.3.4 RWTA-ELBO loss

Since BRVSNet involves both latent space encoding and data reconstruction, its loss design naturally draws inspiration from the Evidence Lower Bound (ELBO) [12] used in variational autoencoders (VAEs). However, unlike conventional VAEs, BRVSNet emphasizes parameter regression, with reconstruction acting as a self-supervised mechanism that accounts for the complexity of the regression task. Motivated by the Relaxed Winner Takes All (RWTA) strategy [38], we introduce the RWTA-ELBO loss, which combines ELBO-based self-supervision with RWTA’s multi-modal learning capability. This formulation not only mitigates mode collapse and enables modeling of multiple distinct rotational modes [157, 158], but also enforces reconstruction consistency to guide rotation regression.

ELBO Formulation

In generative models, maximizing the marginal likelihood of the observed data is generally intractable due to integration over latent variables. The ELBO provides a tractable lower bound on the log-likelihood by introducing an approximate posterior in place of the true posterior:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_n) = & \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_n)} [\log p_{\boldsymbol{\theta}}(\mathbf{x}_n|\mathbf{z})]}_{\text{Reconstruction term (Discrimination)}} - \underbrace{D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_n) \| p_{\boldsymbol{\theta}}(\mathbf{z}))}_{\text{KL divergence (Regression)}}, \end{aligned} \quad (76)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are the parameters of the prior distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$ and the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_n)$, respectively, and $\mathbf{z}$ denotes the latent Bingham variables. The KL divergence term regularizes the approximate posterior towards the prior for rotation regression, while the reconstruction term enforces consistency

between the predicted pseudo ground truths and the original inputs, enabling self-supervised refinement of the regression process.

Discrimination Term

Given a set of N observations $\mathcal{X} = \{\mathbf{x}_n\}$, each $\mathbf{x}_n \in \mathbb{R}^D$, the reconstruction term is estimated using the Stochastic Gradient Variational Bayes (SGVB) estimator [12]:

$$\mathbb{E}_{q_{\underline{\phi}}(\mathbf{z}|\mathbf{x}_n)} [\log p_{\underline{\theta}}(\mathbf{x}_n|\mathbf{z})] = \frac{1}{L} \sum_{l=1}^L (\log p_{\underline{\theta}}(\mathbf{x}_n|\mathbf{z}_l)), \quad (77)$$

where L is the number of Monte Carlo samples, and $\mathbf{z}_l$ is drawn via the acceptance–rejection sampling scheme described in Section 4.3.3.

The choice of reconstruction loss depends on the data modality:

- For image data (AmReLoc dataset), we use Binary Cross Entropy (BCE):

$$\begin{aligned} \mathcal{L}_{\text{BCE}}(\hat{\mathbf{x}}_n, \mathbf{x}_n) = & - \sum_{d=1}^D [x_{n,d} \log(\hat{x}_{n,d}) \\ & + (1 - x_{n,d}) \log(1 - \hat{x}_{n,d})], \end{aligned} \quad (78)$$

- For 3D point clouds (ModelNet10 dataset), we adopt the Chamfer Distance (CD):

$$\begin{aligned} \mathcal{L}_{\text{CD}}(\hat{\mathbf{x}}_n, \mathbf{x}_n) = & \frac{1}{|\hat{\mathbf{x}}_n|} \sum_{\hat{x} \in \hat{\mathbf{x}}_n} \min_{x \in \mathbf{x}_n} \|\hat{x} - x\|_2^2 \\ & + \frac{1}{|\mathbf{x}_n|} \sum_{x \in \mathbf{x}_n} \min_{\hat{x} \in \hat{\mathbf{x}}_n} \|x - \hat{x}\|_2^2, \end{aligned} \quad (79)$$

where $\hat{\mathbf{x}}_n$ and $\mathbf{x}_n$ denote predicted and target point sets, $|\cdot|$ denotes the number of points, and $\|\cdot\|_2^2$ is the squared Euclidean distance.

Regression Term

RWTA focuses learning on the mixture component that yields the most accurate prediction, determined by the distance between the predicted quaternion $\hat{\mathbf{q}}$ and the ground-truth quaternion $\mathbf{q}$. In our formulation,

RWTA is implemented via KL divergence:

$$\begin{aligned}\mathcal{L}_{\text{RWTA-KL}}(\underline{\theta}, \underline{\phi}; \mathbf{x}_n) &= \sum_{i=1}^M \pi_i D_{\text{KL}}(q_{\phi}^{(i)} \| p_{\theta}^{(i)}), \\ &= \sum_{i=1}^M \pi_i (h(q_{\phi}^{(i)}, p_{\theta}^{(i)}) - h(q_{\phi}^{(i)})),\end{aligned}\tag{80}$$

where π_i is the mixture attention weight:

$$\pi_i = \begin{cases} \epsilon, & \text{if } i\text{-th component is selected} \\ \frac{1-\epsilon}{M-1}, & \text{otherwise,} \end{cases}\tag{81}$$

where ϵ is the attention factor, $h(\cdot, \cdot)$ denotes cross entropy, and $h(\cdot)$ denotes entropy.

RWTA-ELBO Formulation

Combining Equation 76 and Equation 80, the RWTA-ELBO loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{RWTA-ELBO}}(\underline{\theta}, \underline{\phi}; \mathbf{x}_n) &= \\ &\underbrace{\mathbb{E}_{q_{\underline{\phi}}(\mathbf{z}|\mathbf{x}_n)} [\log p_{\underline{\theta}}(\mathbf{x}_n|\mathbf{z})]}_{\text{Discrimination term}} - \underbrace{\mathcal{L}_{\text{RWTA-KL}}(\underline{\phi}, \underline{\theta}; \mathbf{x}_n)}_{\text{Regression term}}.\end{aligned}\tag{82}$$

The specific form of the discrimination term depends on the dataset type (BCE for AmReLoc, CD for ModelNet10). The complete derivation is shown in Equation 83.

$$\begin{aligned}
\mathcal{L}_{\text{RWT A-ELBO}}(\underline{\theta}, \underline{\phi}; \mathbf{x}_n) &= \mathbb{E}_{q_{\underline{\phi}}(\mathbf{z}|\mathbf{x}_n)} [\log p_{\underline{\theta}}(\mathbf{x}_n|\mathbf{z})] - \mathcal{L}_{\text{RWT A-KL}}(\underline{\phi}, \underline{\theta}; \mathbf{x}_n) \\
&= \frac{1}{L} \sum_{l=1}^L (\log p_{\underline{\theta}}(\mathbf{x}_n|\mathbf{z}_{n,l})) - \sum_{i=1}^M \pi_i D_{\text{KL}}(q_{\underline{\phi}}^{(i)} \| p_{\underline{\theta}}^{(i)}) \\
&= \mathcal{L}_{\text{BCE}}(\hat{\mathbf{x}}_n, \mathbf{x}_n) \text{ or } \mathcal{L}_{\text{CD}}(\hat{\mathbf{x}}_n, \mathbf{x}_n) \\
&\quad - \sum_{i=1}^M \pi_i (h(q_{\underline{\phi}}^{(i)}, p_{\underline{\theta}}^{(i)}) - h(q_{\underline{\phi}}^{(i)})) \\
&= - \sum_{d=1}^D [x_{n,d} \log(\hat{x}_{n,d}) + (1 - x_{n,d}) \log(1 - \hat{x}_{n,d})] \\
&\text{or } \frac{1}{|\hat{\mathcal{X}}|} \sum_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} \min_{\mathbf{x} \in \mathcal{X}} \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 + \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \min_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \\
&\quad - \sum_{i=1}^M \pi_i (h(q_{\underline{\phi}}^{(i)}, p_{\underline{\theta}}^{(i)}) - h(q_{\underline{\phi}}^{(i)})).
\end{aligned} \tag{83}$$

4.3.5 Weight Loss

The posterior $q_{\underline{\phi}}$ and prior $p_{\underline{\theta}}$ in Equation 82 do not inherently account for the mixture weights w_i . To address this, we design two complementary weight-regularization strategies: Divergence Mixture Entropy (DME) and Aggressive Cross Entropy (ACE).

(1) Divergence Mixture Entropy (DME): Analogous to Equation 80, DME is defined as:

$$\begin{aligned}
\mathcal{L}_{\text{DME}}(w_{\mathcal{B}_{\text{mix}}}, \underline{\theta}, \underline{\phi}) &= \\
&= - \log \left(\sum_{i=1}^M w_i \exp \left(-D_{\text{KL}}(q_{\underline{\phi}}^{(i)} \| p_{\underline{\theta}}^{(i)}) \right) \right),
\end{aligned} \tag{84}$$

where the transformation yields a differentiable entropy-like form, enabling joint updates of mixture weights and distribution parameters during training.

(2) Aggressive Cross Entropy (ACE): An alternative approach treats weight learning as a classification task [38], but with aggressive rescaling:

$$\mathcal{L}_{\text{ACE}}(w_{\mathcal{B}_{\text{mix}}}) = - \sum_{i=1}^M \delta_i \log(w_i) \cdot y_i, \tag{85}$$

where $\{y_i\}_{i=1}^M$ represents the one-hot selection of the best component, and

$$\delta_i = \begin{cases} 1/w_i, & \text{if } w_i > \xi, \\ w_i, & \text{otherwise,} \end{cases} \quad (86)$$

where δ_i is a scaling factor that suppresses dominant components when $w_i > \xi$ (set to 0.01 in our experiments). ACE accelerates convergence to optimal components while mitigating mode collapse by reducing the dominance of high-probability modes. Empirically, both DME and ACE enhance model convergence, offering complementary strategies for optimizing mixture weights within RWTA-ELBO.

4.3.6 Training Procedure

The training process is organized into three sequential stages, where each stage builds upon the previous one to progressively refine different aspects of the model. Each stage employs a tailored combination of loss functions designed to enhance specific capabilities.

Stage 1: Alignment Training In the first stage, the model is optimized using a combination of the regression term and the discrimination term.

- The regression term ensures accurate estimation of rotation parameters.
- The discrimination term enforces consistency between generated and real data, encouraging the model to capture real-world rotational distributions.

This stage establishes the foundation for precise rotation estimation and enables stable progression to later phases. The loss function is:

$$\mathcal{L}_{S1} = \mathcal{L}_{\text{RWTA-ELBO}}(\underline{\theta}, \underline{\phi}; \mathbf{x}_n). \quad (87)$$

Stage 2: Weight Training The second stage focuses on improving the model’s ability to select the most relevant component from the Bingham mixture model. The weight regularization terms in Equation 84 and Equation 85 rescale mixture components and guide the model to prioritize components most relevant to the current rotation estimation task, while suppressing less informative ones. Consequently, the loss function is extended to include the weight loss:

$$\mathcal{L}_{S2} = \mathcal{L}_{S1} + \mathcal{L}_{\text{Weight}}. \quad (88)$$

Stage 3: Generation Training In the final stage, an additional reconstruction term is incorporated. Unlike the discrimination term, this reconstruction is computed only for the optimal mixture component, using its associated likelihood and approximate posterior. This targeted reconstruction step enables the generator to synthesize the corresponding metadata directly from the learned rotation parameters, ensuring that the generated outputs are both rotation-consistent and semantically coherent. The loss function for this stage is:

$$\mathcal{L}_{S3} = \mathcal{L}_{S2} + \underbrace{\mathbb{E}_{q_{\hat{\phi}}(\mathbf{z}|\mathbf{x}_n)} [\log p_{\hat{\theta}}(\mathbf{x}_n|\mathbf{z})]}_{\text{Reconstruction term}}, \quad (89)$$

where $p_{\hat{\theta}}(\mathbf{x}_n|\mathbf{z})$ and $q_{\hat{\phi}}(\mathbf{z}|\mathbf{x}_n)$ denote the likelihood and approximate posterior associated with the selected mixture component.

4.4 Experiments

To comprehensively evaluate our general framework for rotation regression and multi-view generation, we selected datasets from two representative domains: camera relocalization (image-based) and point cloud pose estimation. Each dataset is used for both tasks to ensure consistent cross-domain evaluation.

4.4.1 Experimental Setup

Datasets

We first employ the AmReLoc dataset [38], which contains photographs captured along circular trajectories around objects such as dining tables, round tables, and staircases, as well as real camera motion sequences. This dataset is specifically designed with repetitive structures, making it well-suited for assessing performance in complex and ambiguous environments. The training and test sets consist of 2,414 and 1,326 frames, respectively. During training, camera poses (including both rotation and translation) are provided as labels to supervise the learning process. For multi-view generation, the trained model predicts novel viewpoints for unseen rotations, while the corresponding translations are directly taken from the test labels to preserve geometric consistency in the synthesized views.

For point cloud evaluation, we adopt the ModelNet10 dataset [159], which comprises 10 distinct geometric categories. Pose estimation is performed on objects from each category using the original train/test split. For multi-view generation with point clouds, we select one representative object from each category,

fix its geometry, and assign multiple poses. This procedure produces a synthetic dataset containing 200 shapes per category for both training and testing. All point clouds are uniformly downsampled to 2,048 points per sample before training.

During training, the model is provided with either images or point clouds paired with their corresponding camera or object pose parameters, enabling it to learn a direct mapping from observations to rotations. In the testing phase, for rotation regression, the model takes an image or point cloud from the test set as input and predicts its rotation parameters. For multi-view generation, the model uses ground-truth pose parameters from the test set to generate the corresponding novel-view image or rotated point cloud, ensuring geometric consistency across viewpoints.

Evaluation Metrics

For rotation regression on AmReLoc, following [38], we evaluate using:

- Angular error:

$$d_q(\mathbf{q}, \hat{\mathbf{q}}) = 2 \arccos(|\mathbf{q} \circ \hat{\mathbf{q}}|), \quad (90)$$

- Translational error:

$$d_t(\mathbf{t}, \hat{\mathbf{t}}) = \|\mathbf{t} - \hat{\mathbf{t}}\|_2. \quad (91)$$

A pose is considered correct if both errors are below predefined thresholds. Accuracy is reported under multiple threshold settings.

For point cloud pose estimation, where distinct poses may yield visually similar alignments, we measure the Chamfer Distance (CD) between point clouds transformed using the ground-truth and predicted poses.

In the comparison with other state-of-the-art models for multi-view generation, in addition to visual comparisons, we provide Chamfer Distance (CD) and Earth Mover’s Distance (EMD) as quantitative metrics on the point cloud datasets [143]. For the image dataset AmReLoc, where no suitable baselines are available, we report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) as quantitative references, in addition to visual comparisons with ground-truth images.

Table 4.1: Rotation Estimation Accuracy on the AmReLoc Dataset. DBN results are reproduced under our experimental settings. The best results are shown in bold, and the second-best results are marked with a dagger ($\dagger$).

Method	Blue chairs			Meeting table			Staircase			Staircase_ext			Seminar		
	10/0.1	15/0.2	20/0.3	10/0.1	15/0.2	20/0.3	10/0.1	15/0.2	20/0.3	10/0.1	15/0.2	20/0.3	10/0.1	15/0.2	20/0.3
PoseNet	0.19	0.69	0.90 $\dagger$	0.01	0.05	0.10	0.14	0.45	0.60	0.07	0.31	0.49	0.37 $\dagger$	0.81 $\dagger$	0.90
BayesPoseNet	0.39	0.78	0.88	0.04	0.13	0.22	0.13	0.32	0.49	0.02	0.14	0.31	0.18	0.58	0.78
FisherMatch	0.58	0.85 $\dagger$	0.87	0.08	0.29 $\dagger$	0.35	0.20	0.48	0.65	0.10 $\dagger$	0.40	0.58	0.37 $\dagger$	0.77	0.92 $\dagger$
RotationNormFlow	0.35	0.73	0.79	0.05	0.23	0.29	0.19	0.42	0.55	0.10 $\dagger$	0.39	0.57	0.37 $\dagger$	0.77	0.92 $\dagger$
DBN	0.50	0.84	0.88	0.12 $\dagger$	0.35	0.43 $\dagger$	0.21 $\dagger$	0.49 $\dagger$	0.66 $\dagger$	0.12	0.42 $\dagger$	0.59 $\dagger$	0.37 $\dagger$	0.79	0.92 $\dagger$
Ours	0.55 $\dagger$	0.90	0.94	0.14	0.35	0.48	0.22	0.53	0.70	0.12	0.48	0.67	0.42	0.84	0.94

Table 4.2: Rotation Estimation on ModelNet10. DBN results are reproduced under our experimental settings. CD values are scaled by 10^2 . The best results are shown in bold, and the second-best results are marked with a dagger ($\dagger$).

	Bathtub	Bed	Chair	Desk	Dresser	Monitor	Nightstand	Sofa	Table	Toilet
PointNetLK	10.110	12.330	8.280	10.730	7.260	13.230	5.700	12.460	16.550	7.430
BayesPoseNet	5.994	1.907	0.866	6.068	5.124	2.980	3.535	0.820	7.063	0.730 $\dagger$
FisherMatch	4.594	1.677	1.048	4.751	3.228	3.345	2.427	0.644	7.746	0.989
RotationNormFlow	1.213	0.731	0.669 $\dagger$	1.755	1.114	1.352	1.176	0.299 $\dagger$	0.660	0.765
DBN	0.754 $\dagger$	0.413 $\dagger$	0.708	2.638	1.694	2.138	1.510	0.405	0.587 $\dagger$	0.828
Ours	0.734	0.223	0.529	2.123 $\dagger$	1.238 $\dagger$	1.406 $\dagger$	1.212 $\dagger$	0.218	0.520	0.608

4.4.2 Rotation Regression

We quantitatively compare our method with several state-of-the-art rotation regression models: PoseNet [160], PointNetLK [161], BayesPoseNet [162], FisherMatch [134], RotationNormFlow [128], and DBN [38], evaluated on the AmReLoc and ModelNet10 datasets. Among these, PoseNet is applied only to image-based datasets, whereas PointNetLK is used exclusively for point cloud datasets.

For the AmReLoc dataset, we ensure a fair comparison by strictly following the same architectural design in both training and inference for the camera translation module across all competing methods. Results are summarized in Tables 4.1 and 4.2. For each class in the AmReLoc dataset, rotation accuracy is evaluated under three thresholds: $10^\circ/0.1\text{m}$, $15^\circ/0.2\text{m}$, and $20^\circ/0.3\text{m}$. On ModelNet10, rotation performance is assessed using CD for each class. DBN results are reproduced under our experimental settings.

As shown in Table 4.1, our method achieves the highest accuracy across most thresholds for all AmReLoc categories. DBN, which also adopts the *RWTA* strategy, ranks second overall, highlighting the effectiveness of *RWTA* and related approaches such as *DME* and *ACE* for rotation regression in complex, ambiguous

Table 4.3: Generation performance on the AmReLoc dataset. PSNR: 0– ∞ ; SSIM: 0–1; LPIPS: 0–1.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Blue chairs	15.39	0.67	0.71
Meeting table	11.39	0.30	0.73
Staircase	12.88	0.57	0.69
Staircase_ext	15.39	0.62	0.70
Seminar	12.95	0.60	0.72

Table 4.4: Generation Performance on ModelNet10. CD and EMD values are scaled by 10^2 . The best results are shown in bold.

	Metric	Bathtub	Bed	Chair	Desk	Dresser	Monitor	Nightstand	Sofa	Table	Toilet
DiffusionPM	CD	10.464	8.066	7.115	9.495	6.300	5.574	17.671	5.694	9.812	6.477
	EMD	7.899	7.840	5.630	7.830	4.396	4.319	15.637	4.410	8.019	4.423
WarpingGAN	CD	5.559	3.978	5.227	4.460	4.375	8.588	8.371	8.892	6.139	2.470
	EMD	4.505	3.062	3.760	4.150	3.249	5.018	5.383	5.316	3.403	2.374
Ours	CD	0.371	0.379	0.388	0.335	0.435	0.291	0.509	0.333	0.436	0.392
	EMD	0.771	0.627	0.621	0.560	0.481	0.562	0.759	0.799	0.676	0.593

environments. A similar trend is observed in ModelNet10 results (Table 4.2), except that RotationNormFlow slightly outperforms our method for certain categories. Overall, our method consistently demonstrates superior performance in rotation regression.

4.4.3 Multi-View Generation

In our multi-view generation experiments, the rotation parameters from the test set are fed into the trained model to assess generative performance at unseen viewpoints. For other methods compared in this section, differences in framework design require that we input ground-truth data from the test set during evaluation. Despite this disadvantage in testing conditions, our approach still demonstrates superior generative performance.

Quantitative Comparison

For the AmReLoc dataset, there are no directly comparable baselines for image-based multi-view generation. To provide quantitative reference, we report PSNR, SSIM, and LPIPS scores (Table 4.3), which measure reconstruction fidelity, structural similarity, and perceptual similarity, respectively, between the generated and ground-truth images. The results indicate that our method effectively captures global scene

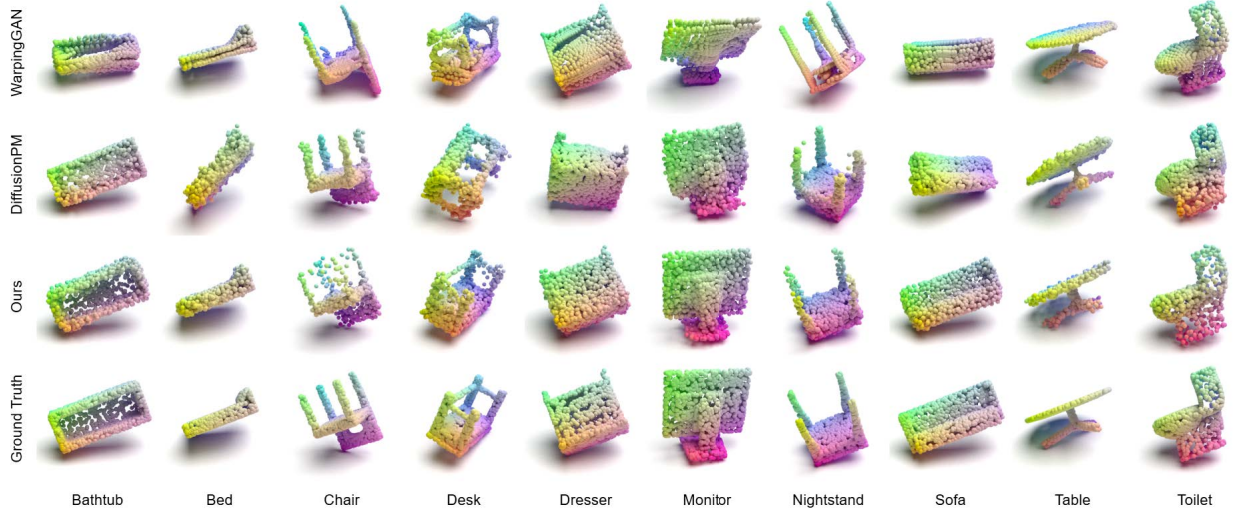


Figure 4.3: Visual Comparison of Rotation Parameters to Point Clouds on the ModelNet10 Dataset. The inputs of Ours: unseen angles. The inputs of DiffusionPM and WarpingGAN: ground truth point clouds.

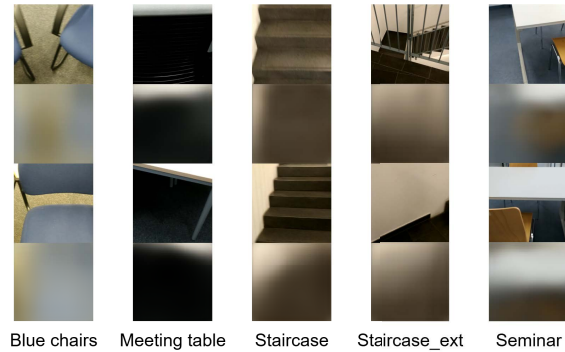


Figure 4.4: Visual comparison of rotation parameter predictions on images from the AmReLoc dataset. Each category includes two comparison sets: the first and third rows show the ground truth, and the second and fourth rows show the corresponding generated images.

structures and generates visually coherent views, though fine-grained textures and lighting variations remain challenging in certain categories.

In contrast, the ModelNet10 dataset allows direct quantitative comparison in the point cloud domain. We compare against two advanced point cloud generation models (DiffusionPM [143] and WarpingGAN [144]), both built on widely adopted generative frameworks. Table 4.4 reports CD and EMD scores, demonstrating that our generated point clouds align more closely with ground truth than those from competing methods. Notably, the performance gap is primarily due to the inability of baseline models to preserve rotational consistency, often leading to mode collapse. By leveraging the Bingham mixture model, our method generates structurally consistent point clouds directly from rotational parameters.

Visual Comparison

Figure 4.3 presents a visual comparison between our method, DiffusionPM, and WarpingGAN on ModelNet10. Our generated point clouds align more closely with the ground truth in terms of both geometry and orientation. While competing methods sometimes produce more visually detailed shapes, they frequently exhibit large angular misalignments. Figure 4.4 shows a qualitative comparison between our generated images and ground truth on AmReLoc. Despite minor blurriness, our results preserve clear object contours (e.g., chairs, tables, staircases) and maintain consistency with the real scenes.

4.4.4 Hyperparameter Analysis

In this section, we analyze the impact of the Number of Mixtures (NofM) in the proposed method. As a key hyperparameter of the Bingham mixture model, NofM is initialized during training and controls the model’s representational capacity. It determines how finely the directional distribution can be approximated, directly influencing both rotation regression accuracy and multi-view generation quality. Therefore, identifying its optimal setting for different datasets is crucial for maximizing performance while maintaining generalization.

Analysis on the AmReLoc dataset

For the AmReLoc dataset, we evaluate the model’s performance under DME and ACE losses with different NofM (Figure 4.5). With DME loss, the best accuracy for most categories is achieved at NofM = 10, though certain categories benefit slightly from NofM = 25. This trend suggests that DME is capable of capturing fine-grained structural variations even as the model capacity increases moderately. Under ACE loss, the optimal performance for the majority of categories is also achieved at NofM = 10, but accuracy decreases noticeably at NofM = 25 and NofM = 50. This indicates that ACE maintains stability only when NofM remains small, and larger mixture counts may reduce its representational focus.

However, when comparing the quantitative results in Table 4.5, ACE outperforms DME across the most NofM settings, with the largest margins at the optimal NofM = 10. These improvements suggest that despite DME’s advantage in modeling fine structural details for specific categories, ACE provides a more balanced representation in AmReLoc. Therefore, from a practical perspective, ACE emerges as the preferred choice for the image dataset, with NofM = 10 offering the best trade-off between capacity and performance.

Table 4.5: Ablation study of rotation regression with varying numbers of mixtures (NofM). For AmReLoc, the metric is accuracy at different thresholds; for ModelNet10, the metric is CD, scaled by 10^2 . The results are the average value across categories on each dataset. The best results for each method are in bold, and the best overall results are marked with a star (*).

Method	NofM	AmReLoc			ModelNet10
		10/0.1 $\uparrow$	15/0.2 $\uparrow$	20/0.3 $\uparrow$	CD $\downarrow$
BRVSNNet+DME	5	0.201	0.497	0.623	1.493
	10	0.203	0.528	0.637	1.181
	25	0.192	0.521	0.660	1.015
	50	0.179	0.485	0.632	0.923*
BRVSNNet+ACE	5	0.187	0.460	0.620	1.591
	10	0.269*	0.597*	0.716*	1.349
	25	0.234	0.579	0.681	1.332
	50	0.236	0.566	0.694	1.288

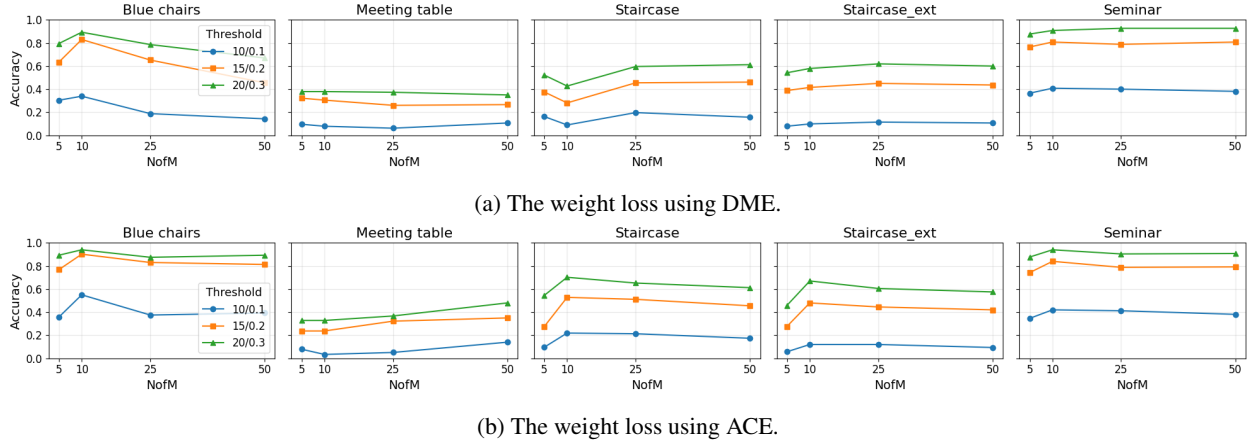
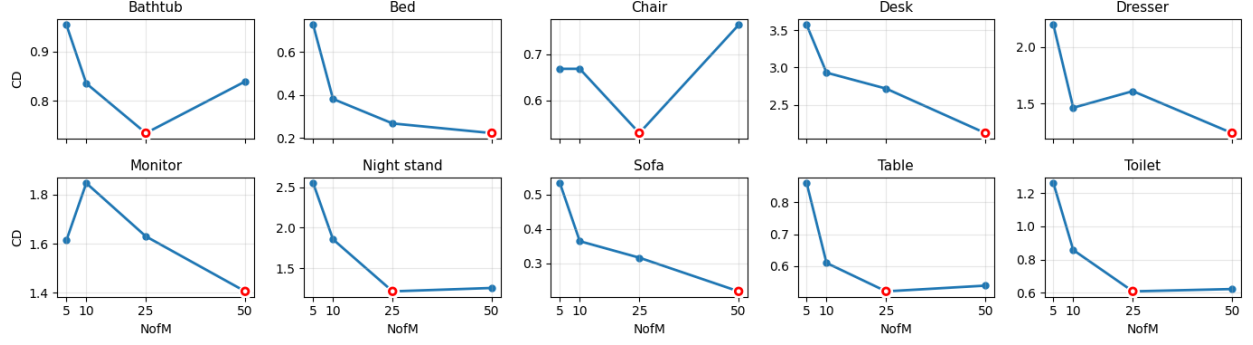


Figure 4.5: The accuracy comparison of NofM on AmReLoc dataset by category and threshold.

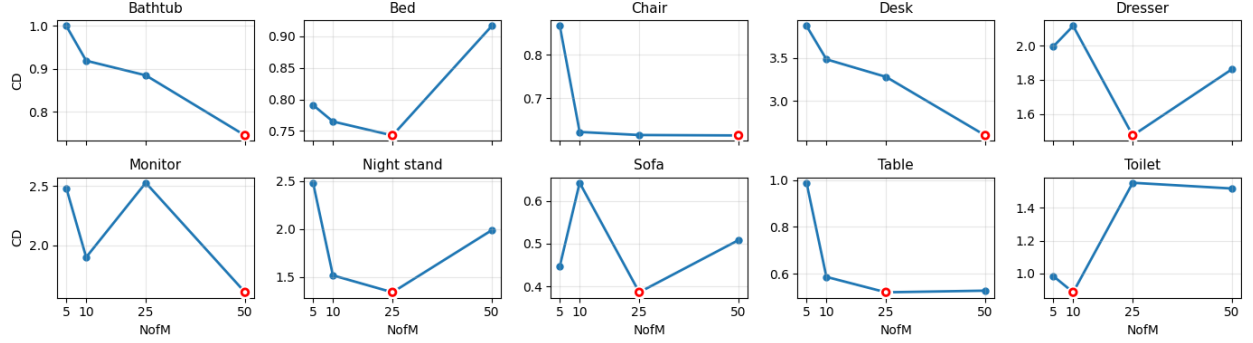
Analysis on the ModelNet10 dataset

For the ModelNet10 dataset, we use Chamfer Distance (CD) as the evaluation metric to study the effect of NofM on generation quality (Figure 4.6). With DME loss, the influence of NofM varies across categories: many categories achieve the minimum CD at NofM = 50, while some reach optimal performance at NofM = 25. In contrast, ACE loss typically yields the best results around NofM = 25 and exhibits relatively lower sensitivity to changes in NofM.

Although Table 4.5 shows that the best mean CD values for both DME and ACE occur at NofM = 50, a deeper analysis reveals that for DME, the performance gap between NofM = 25 and NofM = 50 is marginal, and for ACE, the results for NofM = 10, 25, and 50 are similarly close. This pattern aligns with the trends



(a) The weight loss using DME.



(b) The weight loss using ACE.

Figure 4.6: The CD comparison of NofM on ModelNet10 dataset by category. The red circles indicate the minimum CD for each category, corresponding to the optimal NofM value.

observed in Figure 4.6. Importantly, DME achieves lower CD values than ACE across most NofM settings, indicating that its ability to capture fine-grained structural details is particularly beneficial for point cloud data. This advantage allows DME to produce more precise reconstructions for 3D geometry, making it the preferred choice for tasks where high-fidelity shape recovery is critical.

4.4.5 Ablation Study

In this section, we investigate the contribution of key components in the proposed method through ablation experiments. Specifically, we evaluate three aspects: (i) the supervision of the discriminator on self-supervised rotation regression, (ii) the comparison of the two weight losses (*DME* and *ACE*) across different datasets, and (iii) the importance of Mixtures in the Bingham mixture model. These analyses provide deeper insights into how each factor impacts performance, and guide optimal configurations for different tasks

Table 4.6: Ablation study of rotation regression with and without a discriminator (Disc.). For AmReLoc, the metric is accuracy at different thresholds; for ModelNet10, the metric is CD, scaled by 10^2 . The best results for each method are in bold, and the best overall results are marked with a star (*).

Method	Disc.	AmReLoc			ModelNet10
		10/0.1 $\uparrow$	15/0.2 $\uparrow$	20/0.3 $\uparrow$	CD $\downarrow$
BRVSNet+ <i>DME</i>	✓	0.233	0.574	0.686	0.881*
	✗	0.200	0.554	0.654	1.168
BRVSNet+ <i>ACE</i>	✓	0.290*	0.620*	0.746*	1.091
	✗	0.263	0.577	0.699	1.210

The Discriminator for Supervision

We evaluate the contribution of the discriminator to self-supervised rotation regression by comparing four experimental variants: with and without the discriminator, under both *DME* and *ACE* losses. As shown in Table 4.6, incorporating the discriminator consistently improves rotation regression accuracy across all evaluation metrics and datasets.

On the AmReLoc dataset, the accuracy increases are evident for all thresholds, with the largest relative improvements observed at the threshold (10/0.1), where the discriminator helps the model better distinguish subtle orientation differences. For example, under *ACE* loss, accuracy improves from 0.263 to 0.290, while under *DME* loss, it rises from 0.200 to 0.233. On ModelNet10, the discriminator reduces Chamfer Distance in both loss settings, with *DME* benefiting more significantly (from 1.168 to 0.851), indicating enhanced ability to capture 3D geometry.

The Difference of Weight Losses

We further compare the two proposed weight loss functions, *DME* and *ACE*, to understand their relative strengths across different datasets. As shown in Tables 4.5 and 4.6, *ACE* generally performs better than *DME* on the image-based AmReLoc dataset under the same discriminator setting, achieving the highest accuracy across all thresholds and NofM values. The results suggest that *ACE* provides a more balanced representation for image data, effectively capturing structural consistency while maintaining robustness against variations in NofM.

In contrast, *DME* demonstrates superior performance on the point cloud-based ModelNet10 dataset, achieving lower Chamfer Distance (CD) in most NofM settings. This trend aligns with the qualitative observations that *DME* excels at capturing fine-grained geometric details, which are critical for precise 3D

shape representation. These complementary characteristics enable the selection of *ACE* for image-based tasks prioritizing stability and *DME* for point cloud tasks requiring detailed geometric capture.

The Importance of the Mixture Model

Table 4.5 analyzes the effect of the NofM in the Bingham mixture model. For both datasets, the performance with $\text{NofM} = 5$ is consistently the lowest, regardless of the weight loss used. This clear and repeated underperformance highlights the critical role of having multiple mixture components to approximate the directional distribution. With too few mixtures, the model’s representational capacity is insufficient, leading to degraded rotation regression accuracy on AmReLoc and higher Chamfer Distance on ModelNet10.

Beyond the importance of mixtures, the results also show dataset-specific optimal values: for AmReLoc, peak performance is reached at $\text{NofM} = 10$, while for ModelNet10, the best results emerge at $\text{NofM} = 50$. These differences reflect the varying complexity of directional distributions in image-based versus point cloud-based tasks. In particular, the more complex geometric variations in 3D shapes appear to benefit from a larger number of mixtures, whereas image-based data can achieve strong generalization with fewer components. Overall, these findings demonstrate that the mixture model is a key factor of the proposed method’s performance, and that tailoring NofM to the target data modality is essential for optimal results.

4.5 Conclusion and Future Works

This chapter presents BRVSNet, a unified self-supervised framework for joint rotation estimation and multi-view generation. By leveraging a Bingham mixture model over quaternions, BRVSNet captures directional uncertainty in a compact and differentiable latent space. This latent space supports two key functionalities: a self-supervised discriminator for rotation regression and a multi-view generator conditioned on the dominant mixture component, enabling high-fidelity synthesis from novel viewpoints. To facilitate joint optimization, we introduced the RWTA-ELBO along with two auxiliary weight regularization losses, collectively improving rotation accuracy and latent representation quality. Extensive experiments on the AmReLoc and ModelNet10 datasets demonstrate that BRVSNet consistently outperforms state-of-the-art baselines in both rotation estimation accuracy and multi-view generation fidelity across image and point cloud modalities, confirming the effectiveness of integrating probabilistic modeling within a generative self-supervised learning paradigm. Despite these promising results, our framework exhibits limitations in textural fidelity

compared to state-of-the-art generative models. Future research will explore integrating BRVSNet with more expressive generative paradigms, such as denoising diffusion models, to further enhance rendering quality and adaptability to more complex visual domains.

Chapter 5

Enhancing Rotation-Invariant 3D Learning with Global Pose Awareness and Attention Mechanisms

5.1 Introduction

The rapid advancement of 3D sensing technologies has led to the widespread adoption of point cloud data in various 3D computer vision tasks, including shape classification, part segmentation, and object recognition [8, 9, 10, 11]. While recent learning-based models [25, 26, 27] have achieved notable success on pre-aligned datasets, their generalization to real-world scenarios remains limited. A common and often unrealistic assumption in these models is that input objects are consistently aligned in a canonical orientation. However, in practical applications such as autonomous driving, robotics, and augmented reality, 3D objects naturally appear under arbitrary rotations. Models that lack true *rotation-invariant* (RI) capabilities experience substantial performance degradation when confronted with such orientation variations, even when extensively trained with rotation-augmented data. This limitation fundamentally restricts the deployment of point cloud models in unconstrained environments.

To overcome this issue, a growing body of research has focused on designing RI models by replacing raw 3D coordinates with handcrafted or learned geometric features that are invariant to global rotations [20, 21, 22]. These approaches demonstrate improved robustness to orientation changes by leveraging local

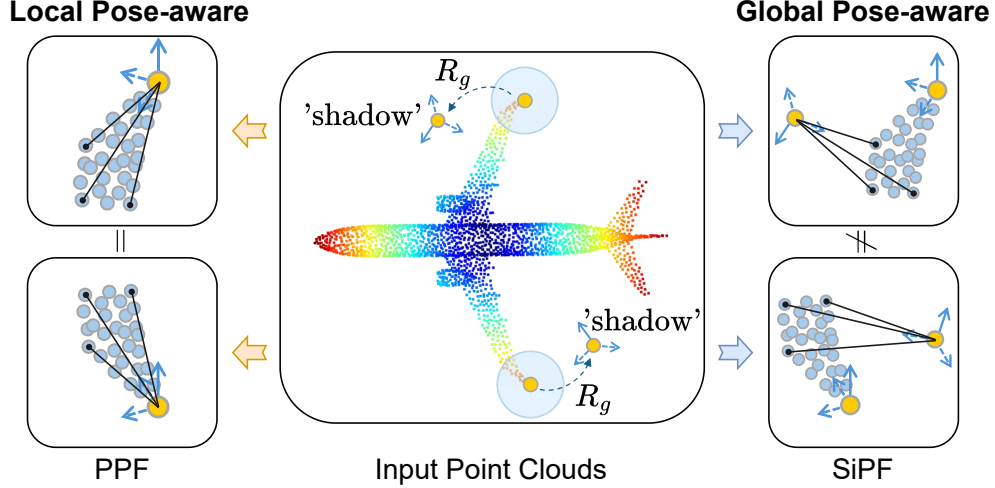


Figure 5.1: Illustration of pose ambiguity caused by symmetric local structures. Traditional Point Pair Features (PPF) capture only local geometric relationships and thus yield identical representations for symmetric regions with similar local geometry (e.g., left and right wing tips). In contrast, the proposed SiPF incorporates global pose awareness by introducing a ‘shadow’ reference point derived from a shared rotation matrix R_g to incorporate pose-aware context, enabling consistent and discriminative RI representations across arbitrary orientations.

descriptors such as point pair features or RI tensors. However, a critical drawback remains: by discarding absolute coordinate information, these methods also lose global pose context, which is essential for differentiating spatially distinct but geometrically similar regions. For instance, in the case of an airplane model, symmetric components such as the left and right wings may be indistinguishable when only local geometry is considered—regardless of rotation invariance. This leads to what we refer to as the *wing-tip feature collapse* phenomenon. Recent attempts to mitigate this issue involve expanding the receptive field and incorporating pairwise relationships within local neighbourhoods to capture broader context while maintaining RI properties [23, 24]. Despite these advances, existing methods still struggle with structural ambiguity in repetitive or symmetric components, primarily due to their limited capacity to encode spatial distinctiveness without access to a consistent global reference.

We argue that the inherent limitation of the receptive field in current RI methods is a fundamental bottleneck. Specifically, for geometrically similar regions, such as opposing airplane wings, the local features (even aggregated from larger neighbourhoods) lack the discriminative information necessary to encode their distinct global poses. As shown in Figure 5.1, conventional RI descriptors fail to preserve relative global positioning, resulting in ambiguous representations that impede fine-grained recognition tasks like part segmentation. Furthermore, these limitations explain the suboptimal performance of many RI pipelines, which

focus on preserving local geometric consistency but neglect global relational cues critical for downstream tasks.

To address these challenges, in this work we propose a novel feature representation termed *Shadow-informed Pose Feature* (SiPF) that augments local RI features with global pose cues while preserving rotation invariance. Inspired by the classical point pair feature (PPF) [1], SiPF introduces a globally consistent shadow reference for each point, derived via a learned shared rotation. This “shadow” serves as an anchor point for encoding relative position in a globally consistent manner. Specifically, we introduce a task-adaptive shadow locating module that estimates a shared optimal rotation via the Bingham distribution over unit quaternions. Using the learned rotation, each point is paired with its shadow to construct SiPF descriptors that retain local geometric relations while incorporating global spatial awareness. We further design a novel attention-based operator, termed *Rotation-invariant Attention Convolution* (RIAttnConv), that integrates SiPF into a cross-attention mechanism fused with convolutional operations. This design enables dynamic aggregation of spatially-aware features, thereby improving the model’s ability to distinguish structurally similar but spatially distinct components. We validate our method on multiple benchmark datasets across classification and part segmentation tasks. Experimental results demonstrate that our framework consistently outperforms state-of-the-art RI methods, particularly in challenging settings that require fine-grained spatial discrimination under arbitrary orientations.

Our main contributions are summarized as follows:

- We identify the restricted receptive field as a key limitation in current RI methods and address it by introducing the SiPF, which incorporates global pose context while maintaining rotation invariance.
- We develop RIAttnConv, a novel convolutional module that integrates SiPF through an attention-augmented feature aggregation strategy, significantly enhancing the discriminative capacity of RI representations.
- We design a task-adaptive shadow locating module that learns a globally consistent reference rotation using the Bingham distribution over unit quaternions, enabling dynamic and robust shadow construction.

5.2 Related Works

5.2.1 Rotation-Sensitive 3D Point Cloud Analysis

PointNet [42] is a pioneering work in point cloud analysis, introducing a straightforward yet effective framework for directly processing raw 3D point sets in tasks such as classification and segmentation. To address its limitations in capturing local geometric structures, PointNet++ [25] extends this architecture with hierarchical feature learning over local neighbourhoods. Subsequent works, such as PointCNN [163] and DGCNN [26], further enhance local context modeling through specialized convolutional and dynamic graph-based operations, respectively. Despite these advances, a shared limitation among these approaches is their sensitivity to input orientation. Because they operate directly on absolute coordinates, their performance tends to degrade significantly when encountering arbitrarily rotated inputs, which limits their applicability in real-world scenarios.

5.2.2 Rotation-Robust Methods

To mitigate the sensitivity to rotations, early approaches utilize principal component analysis (PCA) to align input point clouds to a canonical orientation. For instance, Li et al. [164] address pose ambiguity through hybrid canonicalization strategies. However, such methods often rely heavily on data augmentation or heuristic approximations, failing to achieve true rotation invariance. An alternative direction involves equivariant neural networks that encode predictable transformations of features under rotations via structured linear mappings [165, 166, 167]. These models preserve pose information while improving rotation robustness. Nonetheless, maintaining equivariance typically imposes strict architectural constraints, such as enforcing linear compositions in encoders [165, 168], which can limit model expressiveness and hinder learning capacity. Recently, constructing RI features within local neighbourhoods has emerged as a prevalent strategy. Most of these methods employ handcrafted features derived from geometric priors such as pairwise distances and inter-point angles [22, 20]. For instance, RI-CNN [46] and PaRI-Conv [23] design 4D and 8D local descriptors, respectively, to enhance geometric encoding. However, these designs often sacrifice global contextual information, thereby constraining their ability to capture complex spatial semantics [46]. To alleviate this drawback, several methods have attempted to expand the receptive field or increase connectivity. ClusterNet [169], for example, constructs a densely connected k NN graph (with $k = 80$) to retain richer relational information. LocoTrans [170] enhances local RI features by leveraging

representations derived from an equivariant backbone.

Recent advances such as RIGA [171], and rotation-invariant transformers [172] show that incorporating global context or attention is beneficial for rotation-robust point matching. Attention-based sampling strategies [173, 174] further highlight the importance of globally informed feature construction. Building on these insights, our approach explicitly integrates global pose information with local geometric encoding, yielding more expressive RI representations and consistently strong performance across rotation-robust benchmarks.

5.3 Problem Definition and Background

Rotation-Invariant Functions. Let a point cloud be denoted by $P = \{p_1, \dots, p_N\} \in \mathbb{R}^{N \times 3}$ and let $R \in SO(3)$ represents an arbitrary 3D rotation matrix. The rotated point cloud is thus $P_R = P R$. A function Φ is RI if and only if:

$$\Phi(P_R) = \Phi(P), \quad \forall R \in SO(3). \quad (92)$$

Receptive Field in 3D Point Cloud Convolutions. Given a point cloud $P = [p_1, \dots, p_N] \in \mathbb{R}^{N \times 3}$ with associated input features $X = [x_1, \dots, x_N] \in \mathbb{R}^{N \times c_{in}}$, consider a reference point p_r with its local neighbourhood denoted by $\mathcal{N}(p_r)$. The output feature at p_r , written as part of $X' = [x'_1, \dots, x'_N] \in \mathbb{R}^{N \times c_{out}}$, is typically obtained by aggregating transformed features of its neighbours:

$$f(p_r) = \bigwedge_{j \in \mathcal{N}(p_r)} W_j^T \cdot h(p_j). \quad (93)$$

where $h : \mathbb{R}^3 \rightarrow \mathbb{R}^c$ is a nonlinear point-encoding function, W_j^T denotes the learnable dynamic kernel weight associated with the pair (p_r, p_j) , and $\bigwedge$ represents an aggregation operator such as MAX, SUM, or AVG.

Let the local patch around point p_j be defined as $\Omega(p_j) = \{p_k \mid k \in \mathcal{N}(p_j)\}$. Since h may also depend on the neighbourhood of p_j , the effective receptive field at p_r becomes the union of its neighbours' patches:

$$\Gamma(p_r) = \bigcup_{j \in \mathcal{N}(p_r)} \Omega(p_j). \quad (94)$$

To make the dependence on the receptive field explicit, Equation 93 can be reformulated as:

$$f(p_r | \Gamma(p_r)) = \bigwedge_{j \in \mathcal{N}(p_r)} W_j^r \cdot h(p_j | \Omega(p_j)). \quad (95)$$

Limitations of the Receptive Field in Existing RI Methods. Consider an airplane point cloud and let $p_{\text{left}}, p_{\text{right}} \in P$ be the two wing-tip points. Due to their geometric symmetry, there exists a rotation $R_{\text{sym}} \in SO(3)$ such that:

$$\Omega(p_{\text{right}}) = \Omega(p_{\text{left}}) R_{\text{sym}}. \quad (96)$$

Patch-Swapping Transformation. For any reference point p_r with neighbourhood $\mathcal{N}(p_r)$, we define a transformation on the receptive field as:

$$T(\Gamma(p_r)) = \bigcup_{j \in \mathcal{N}(p_r)} \Omega(p_j) R_j, \quad (97)$$

where each $R_j \in SO(3)$ maps the local patch $\Omega(p_j)$ to its geometrically equivalent counterpart (e.g. R_{sym} swaps the left and right wing-tip patches). Applying Equation 95 to the transformed data yields:

$$f(p_r | T(\Gamma(p_r))) = \bigwedge_{j \in \mathcal{N}(p_r)} W_j^r \cdot h(p_j | \Omega(p_j) R_j). \quad (98)$$

Since h is rotation-invariant by definition (Equation 92), we have $h(p_j | \Omega(p_j) R_j) = h(p_j | \Omega(p_j))$. Therefore, Equations 95 and 98 are equivalent, leading to:

$$f(p_r | \Gamma(p_r)) = f(p_r | T(\Gamma(p_r))). \quad (99)$$

Wing-Tip Feature Collapse. By setting $p_r = p_{\text{left}}$ and choosing T to swap the left and right wing-tip patches, Equation (99) simplifies to: $f(p_{\text{left}}) = f(p_{\text{right}})$. This result implies that the network assigns identical features to symmetric yet globally opposite regions. As a consequence, crucial pose information (e.g. distinction between left and right) is lost, introducing ambiguity in downstream tasks and limiting the discriminative power of the learned representations.

5.4 The Proposed Method

To address the global pose ambiguity that arises among geometrically equivalent local patches, we propose a novel RI point descriptor, termed SiPF. It augments each point with a global pose-aware reference, known as the *shadow*, obtained by applying a shared rotation to the entire point cloud. This augmentation allows local patches to encode relative pose information without compromising rotation invariance, thereby enabling the network to distinguish globally distinct yet locally similar structures.

Building upon SiPF, we further introduce RIAttnConv, a rotation-invariant convolutional operator that incorporates shadow-informed relative pose information into the feature aggregation process. Specifically, for a given reference point p_r , the output feature is computed as:

$$x'_r = \bigwedge_{j \in \mathcal{N}(p_r)} \text{Attn} \left(\{ \mathcal{W}(\mathcal{P}_r^j) \cdot x_j \}_{j=1}^k \right), \quad (100)$$

where $\mathcal{P}_r^j$ encodes the relative pose among the reference point p_r , its shadow p'_r , and neighbour p_j . The kernel weight W_j^r adapts dynamically to the global context introduced by p'_r , allowing the model to distinguish symmetric structures based on their global orientation. As a result, W_j^r in Equations (95) and (98) varies under local patch rotations, enabling the attention module $\text{Attn}(\cdot)$ to differentiate identical local geometries in distinct global poses while preserving overall rotation invariance. The full architectural pipeline of our method is illustrated in Figure 5.2.

5.4.1 Shadow-Informed Pose Feature

LRF Construction. Local Reference Frame (LRF) is a fundamental component for constructing RI representations. While the encoding on the relative pose $\mathcal{P}_r^j$ between point pairs, the LRF $\mathcal{L}_r$ serves as a local coordinate system independently constructed at each point $p_r \in P$. Formally, $\mathcal{L}_r = [\partial_r^1, \partial_r^2, \partial_r^3] \in \mathbb{R}^{3 \times 3}$ consisting of three orthonormal basis vectors, typically derived from two direction vectors e_r^1, e_r^2 via Gram–Schmidt orthonormalization:

$$\partial_r^1 = e_r^1, \quad \partial_r^3 = \frac{\partial_r^1 \times e_r^2}{\|\partial_r^1 \times e_r^2\|}, \quad \partial_r^2 = \partial_r^3 \times \partial_r^1. \quad (101)$$

Recent works [23, 24] adopted a robust and purely local LRF construction strategy, wherein the vector

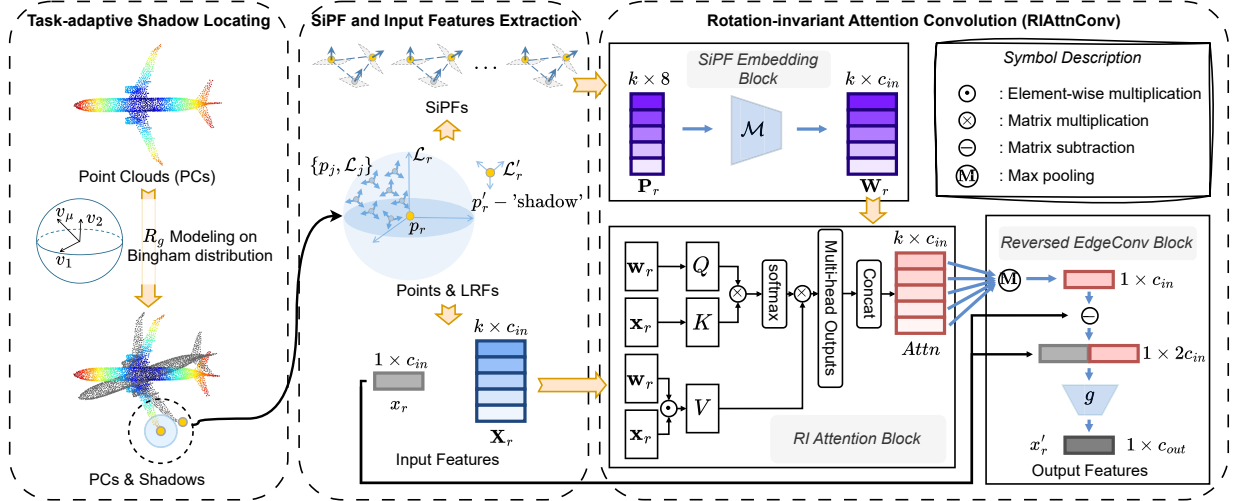


Figure 5.2: Overview of the proposed SiPF and RIAttnConv pipeline. Our framework consists of three main components: (1) Task-adaptive Shadow Locating estimates a globally consistent reference rotation R_g by modeling unit quaternions with a Bingham distribution, generating ‘shadow’ points for each input; (2) SiPF and Input Feature Extraction constructs LRFs $\{\mathcal{L}_j\}_{j=1}^k$ and computes SiPFs $\mathbf{P}_r$ by combining local geometric descriptors with global shadow-based pose differences; (3) RIAttnConv embeds SiPFs to produce adaptive kernel weights $\mathbf{W}_r$, which are used in a RI attention block to guide feature aggregation. A reversed EdgeConv module then fuses neighbourhood and center features x_r to produce the final output x'_r .

from the barycenter of point neighbours to the point itself defines one axis. Specifically, the barycenter is computed as $m_r = \frac{1}{k} \sum_{j=1}^k p_j^r$, where p_j^r are the k -nearest neighbours of p_r , and the axis e_r^2 is defined as $\overrightarrow{p_r m_r} = m_r - p_r$. In this work, we adopt this LRF construction strategy based on local geometry ($e_r^1 = n_r, e_r^2 = \overrightarrow{p_r m_r}$) as input features X to ensure essential robustness and consistency in real-world point clouds.

SiPF Extraction. To describe the relative pose between the LRFs of a reference point p_r and its neighbour p_j , we design our RI feature based on the standard Point Pair Feature (PPF) [1], a 4D vector constructed from the relative position and the primary LRF axes $\partial_r^1, \partial_j^1$ of the two points:

$$\begin{aligned} \text{PPF}(p_r, p_j) &= (\|d\|_2, \cos(\alpha_1), \cos(\alpha_2), \cos(\alpha_3)), \\ \alpha_1 &= \angle(\partial_r^1, d), \quad \alpha_2 = \angle(\partial_j^1, d), \quad \alpha_3 = \angle(\partial_r^1, \partial_j^1), \end{aligned} \tag{102}$$

where $d = p_j - p_r$ and $\angle(\cdot, \cdot)$ denotes the angle between two vectors. While PPF is sufficient to represent the pairwise pose, it becomes ambiguous in the context of multiple neighbours. As shown in prior works [175, 23], for two neighbours p_j and p'_j with identical normals and equal distances to p_r , if both lie on a circle around the axis ∂_r^1 , their PPF values with respect to p_r will be identical, i.e., $\text{PPF}(p_r, p_j) = \text{PPF}(p_r, p'_j)$,

despite having distinct spatial arrangements.

To resolve this ambiguity and incorporate global pose awareness, we introduce the *shadow* point p'_r of p_r , derived from a shared global rotation. We then define the *Shadow-informed Point Pair Feature* (SiPPF) as:

$$\text{SiPPF}(p_r, p'_r, p_j) = \frac{\text{PPF}(p_r, p'_r) - \text{PPF}(p_j, p'_r)}{\|\text{PPF}(p_r, p'_r) - \text{PPF}(p_j, p'_r)\|_2}, \quad (103)$$

where $\text{PPF}(p_r, p'_r)$ and $\text{PPF}(p_j, p'_r)$ encode the relative geometry between each point and the shadow. We normalize the difference using the ℓ_2 -norm to emphasize the pose difference of p_j under the shared global reference.

Finally, we concatenate the local geometric encoding with the global-aware difference to obtain our 8D *Shadow-informed Pose Feature* (SiPF):

$$\mathcal{P}_r^j = (\text{PPF}(p_r, p_j), \text{SiPPF}(p_r, p'_r, p_j)) \in \mathbb{R}^8. \quad (104)$$

Task-Adaptive Shadow Locating. As previously discussed, a shared global rotation is applied to generate the shadow point p'_r , which serves as a consistent reference for global-relative feature extraction. However, in certain geometric configurations, the shadow may fail to enhance the discriminability of local features due to inherent ambiguities. Two representative failure cases are outlined below:

- (1) When the shadow p'_r lies along the primary LRF axis ∂_r^1 , and the primary axes ∂_r^1 and $\partial_{r'}^1$ coincide, the SiPF degenerates to standard PPF due to unresolved degrees of freedom of p_j .
- (2) If the global rotation coincides with the local rotation path R_j used in Equation 97, the resulting SiPF only cancels the internal degrees of freedom of p_j , but fails to distinguish the receptive field $\Omega(p_j)$ from its rotated variant $\Omega(p_j)R_j$.

These cases highlight the necessity of a task-adaptive, ambiguity-resilient global rotation strategy. To account for the ambiguity in selecting $\mathbf{R}_g$ over the entire dataset, we model the uncertainty using a *Bingham distribution* on the hypersphere S^3 of unit quaternions [32]. Let q represent candidate quaternions corresponding to potential global rotations. The Bingham probability density function is defined as:

$$\mathcal{B}(q \mid \mathbf{V}, \mathbf{\Lambda}) = \frac{1}{F(\mathbf{\Lambda})} \exp \left(q^\top \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top q \right), \quad (105)$$

where $\mathbf{V} \in \mathbb{R}^{4 \times 4}$ is an orthogonal matrix, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3, 0)$ is a diagonal matrix with $\lambda_1 \leq \lambda_2 \leq \lambda_3 < 0$. The matrix $\mathbf{V}$ defines the mode and dispersion axes, while $\mathbf{\Lambda}$ controls the strength of dispersion along these axes. $F(\mathbf{\Lambda})$ is the normalization constant. To connect $\mathbf{R}_g$ with the learned mode of the Bingham distribution, we extract the mode vector in $\mathbf{V}$ as the optimal global rotation candidate for generating $p'_r = p_r \mathbf{R}_g$ at each iteration.

To ensure that the selected global rotation aligns with the task-specific objectives, we embed the optimization of $\mathbf{R}_g$ into the end-to-end training. Specifically, we jointly optimize the rotation with network parameters by minimizing a composite loss that integrates both the task objective and the probabilistic alignment with the Bingham distribution:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \delta \cdot \sqrt{(\mathcal{L}_{\text{bingham}} - 0.1 \cdot \mathcal{L}_{\text{task}})^2}, \quad (106)$$

where $\mathcal{L}_{\text{bingham}}$ denotes the negative log-likelihood of the Bingham distribution evaluated at the quaternion $q = \text{Quat}(\mathbf{R}_g)$, and δ is a balancing hyperparameter, empirically set to 0.8. The subtraction term encourages to maintain consistency between the uncertainty captured by the Bingham model and the alignment needs of the downstream task.

5.4.2 The RIAttnConv Operator

To enhance RI feature learning, we propose the RIAttnConv operator. Unlike prior approaches that rely only on relative pose for determining kernel weights, RIAttnConv employs an attention mechanism that adaptively weights neighbour features based on both geometric cues and feature similarity. This design enables the model to capture broader contextual dependencies while facilitating effective aggregation.

RI Attention. Given a reference feature $x_r \in \mathbb{R}^{c_{\text{in}}}$ and its k -nearest neighbouring features $\{x_j\}_{j=1}^k$, we define the stacked neighbour feature matrix as $\mathbf{X}_r = [x_1, \dots, x_k]^\top \in \mathbb{R}^{k \times c_{\text{in}}}$. Similarly, the corresponding SiPF descriptors are stacked as $\mathbf{P}_r = [\mathcal{P}_r^1, \dots, \mathcal{P}_r^k]^\top \in \mathbb{R}^{k \times 8}$. RIAttnConv constructs a scaled dot-product attention mechanism using both $\mathbf{X}_r$ and $\mathbf{P}_r$:

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{Softmax} \left(\frac{QK^\top}{\sqrt{c_{\text{in}}}} \right) V, \\ Q &= \mathbf{W}_r, \quad K = \mathbf{X}_r, \quad V = \mathbf{W}_r \cdot \mathbf{X}_r, \end{aligned} \quad (107)$$

where $\mathbf{W}_r = [W_1^r, \dots, W_k^r]^\top \in \mathbb{R}^{k \times c_{in}}$, and $W_j^r = \mathcal{M}(\mathcal{P}_r^j)$ as $\mathcal{W}(\mathcal{P}_r^j)$ in Equation (100). $\mathcal{M}$ is a Multi-layer Perceptron (MLP) that maps the relative pose $\mathcal{P}_r^j$ into the respective dimension.

This design enables the $\text{Softmax}(\cdot)$ in Equation (107) to produce a $k \times k$ attention matrix over the k -nearest neighbours, enhancing contextual sensitivity. Compared to prior methods [23, 176], this formulation offers a favourable trade-off between computational complexity and expressive power. A detailed efficiency analysis is provided in ablation study (see Table 5.5).

Reversed EdgeConv. To retain the feature information of the reference point x_r , we propose a reversed variant of EdgeConv [26], where neighbour features are first aggregated and then fused with the reference feature. Specifically, given the attention mechanism, we compute:

$$\begin{aligned} \mathcal{Attn} \left(\{\mathcal{W}(\mathcal{P}_r^j) \cdot x_j\}_{j=1}^k \right) &= \text{Attention}(Q, K, V), \\ \hat{x}_r &= \text{MAX}_{j \in \mathcal{N}(p_r)} \mathcal{Attn} \left(\{\mathcal{W}(\mathcal{P}_r^j) \cdot x_j\}_{j=1}^k \right), \end{aligned} \quad (108)$$

where $\mathcal{Attn}$ is the output from the RI attention module defined in Equation (107). Then, the updated feature x'_r is obtained by fusing $\hat{x}_r$ and the original reference feature x_r as:

$$x'_r = g \left((\hat{x}_r - x_r) \oplus x_r \right), \quad (109)$$

where $g(\cdot)$ denotes a one-layer MLP and $\oplus$ represents feature concatenation.

5.4.3 Network Architecture

To validate the generalizability and effectiveness of the proposed RIAttnConv operator, we integrate it into two representative backbone architectures widely adopted in point cloud learning: DGCNN [26] for classification, and AdaptConv [176] for segmentation.

Shape Classification. We adopt the DGCNN backbone and replace all EdgeConv layers with RIAttnConv. The k -nearest neighbour graph is constructed in Euclidean space at each feature layer, with $k = 20$ as the neighbourhood size.

Shape Part Segmentation. For part segmentation, we follow the AdaptConv-based design. Specifically, we stack four RIAttnConv layers followed by a graph convolution layer. Three intermediate pooling stages are inserted to progressively enlarge the receptive field. The k -nearest neighbour graph is constructed with

$k = 40$ for all layers, using Euclidean distance in feature space.

5.5 Experiments

We evaluate our method on three benchmarks: ModelNet40 [159] for shape classification, ShapeNet-Part [177] for part segmentation, and ScanObjectNN [178] for real-world classification. To assess rotation invariance, we adopt three standard train/test protocols: z/z, z/SO(3), and SO(3)/SO(3), where z and SO(3) represent vertical and arbitrary rotations. For fair comparison, RIAttnConv is integrated into standard backbones without additional modifications.

5.5.1 Implementation Details

We use stochastic gradient descent (SGD) with an initial learning rate of 0.1, decayed to 0.001 via cosine annealing [179]. Models are trained for 300 epochs with a batch size of 32 for classification and 16 for segmentation. A dropout rate of 0.5 is applied to fully connected layers.

To ensure rotation-invariant input representations, we initialize a point-wise descriptor for each point p_i based on its relative position to the global centroid O [180], including $\|\overrightarrow{Op_i}\|_2$, $\sin(\angle(\partial_i^1, \overrightarrow{Op_i}))$, and $\cos(\angle(\partial_i^1, \overrightarrow{Op_i}))$, where ∂_i^1 is the local reference axis. By default, we use surface normals as ∂_i^1 (denoted as pc+normal). For methods using only coordinates, we replace ∂_i^1 with the vector $\overrightarrow{p_r m_r}$, from the barycenter to the point.

5.5.2 3D Shape Classification

Dataset. We evaluate our method on the widely used ModelNet40 [159] benchmark, which contains 12,311 CAD models from 40 categories, with 9,843 samples for training and 2,468 for testing. Following [42], we randomly sample 1,024 points for each object and augment the training data with random scaling and translation.

Results. Table 5.1 reports classification accuracy under three rotation protocols: z/z, z/SO(3), and SO(3)/SO(3). While rotation-sensitive methods (e.g., PointCNN, DGCNN) perform well under aligned conditions, their accuracy drops sharply (up to 70%) under unseen rotations. In contrast, RI methods exhibit stable performance across settings but often suffer from limited expressiveness without normals; for instance, RIConv and ClusterNet achieve only 86–87% accuracy in such cases.

Methods	Inputs	z/z	z/SO(3)	SO(3)/SO(3)
Rotation-sensitive				
PointNet [42]	pc	89.2	16.2	75.5
PointNet++ [25]	pc+n	89.3	28.6	85.0
PointCNN [163]	pc	92.2	41.2	84.5
DGCNN [26]	pc	92.2	20.6	81.1
Rotation-robust				
RIConv [46]	pc	86.5	86.4	86.4
ClusterNet [169]	pc	87.1	87.1	87.1
RI-GCN [181]	pc+n	91.0	91.0	91.0
VN-DGCNN [165]	pc	89.5	89.5	90.2
SGMNet [20]	pc	90.0	90.0	90.0
OrientedMP [182]	pc+n	88.4	88.4	88.4
ELGANet [21]	pc	90.3	90.3	90.3
PaRot [183]	pc	90.9	91.0	90.8
TetraSphere [184]	pc	90.5	90.5	90.3
Ours	pc	91.8	91.8	91.8
Ours	pc+n	92.6	92.6	92.6

Table 5.1: Shape classification accuracy (%) on ModelNet40 dataset. ‘pc’ and ‘n’ represent the 3D coordinates and normal vectors of input data, respectively.

Our proposed RIAttnConv achieves 91.8% accuracy using only point coordinates, surpassing several normal-dependent methods such as OrientedMP (88.4%) and RI-GCN (91.0%). With surface normals, performance further improves to 92.6%, yielding a 1.6% gain over RI-GCN and outperforming recent approaches including TetraSphere and PaRot. These results highlight the effectiveness of our attention-based design in learning pose-aware yet rotation-invariant features.

5.5.3 Real-World Shape Classification

Dataset. We further evaluate our method on the ScanObjectNN dataset [178], a real-world benchmark comprising 2,902 objects from 15 categories captured in cluttered and noisy environments. Following standard protocol, we use the OBJ_BG subset, consisting of 2,890 samples with 2,312 for training and 578 for testing. As surface normals are unavailable, only raw point coordinates are used.

Results. As shown in Table 5.2, our method achieves the highest accuracy under both z/SO(3) and SO(3)/SO(3) settings, despite using only point coordinates. It outperforms PaRI-Conv by 0.7% in both cases, highlighting the effectiveness of our global pose-aware attention in capturing robust local geometry under real-world noise and occlusion.

Methods	z/SO(3)	SO(3)/SO(3)
Input pc only		
PointNet [42]	16.7	54.7
PointNet++ [25]	15.0	47.4
PointCNN [163]	14.6	63.7
DGCNN [26]	17.7	71.8
RIConv [46]	75.3	75.5
RI-GCN [181]	80.5	80.6
VN-DGCNN [165]	79.8	80.3
OrientedMP [182]	76.7	77.2
LGR-Net [185]	81.2	81.4
PaRI-Conv [23]	83.3	83.3
PaRot [183]	82.1	82.6
Ours	84.0	84.0

Table 5.2: Real-world classification accuracy (%) on ScanObjectNN OBJ_BG dataset.

5.5.4 Shape Part Segmentation

Dataset. We evaluate our method on the widely adopted ShapeNetPart dataset [177], which contains 16,881 3D models across 16 object categories. Each object is annotated with 2 to 6 parts, yielding a total of 50 distinct part labels. Following the standard protocol in [25], we use the official train/test split and uniformly sample 2,048 points from each object for input.

Results. As shown in Table 5.3, our method achieves state-of-the-art performance under both z/SO(3) and SO(3)/SO(3) settings. Under the more challenging z/SO(3) setup, it achieves a class mean Intersection-over-Union (C. mIoU) of 82.9% and an instance mean IoU (I. mIoU) of 85.0% with normals, outperforming recent methods. Without normals, it still achieves 84.4% and 85.1% instance mIoU, surpassing existing RI baselines such as LocoTrans and TetraSphere, and rivaling normal-dependent approaches like PaRI-Conv and RISurConv. Furthermore, our method consistently achieves strong class-level and instance-level segmentation performance under unseen rotations. This improvement stems from the proposed SiPF and global pose-aware reference, which enable the model to disambiguate locally similar parts by leveraging global spatial context. Figure 5.3 visualizes representative segmentation outputs under the z/SO(3) setting, further illustrating the model’s robustness to arbitrary spatial rotations.

Method	Input	z/SO(3)		SO(3)/SO(3)	
		C. mIoU	I. mIoU	C. mIoU	I. mIoU
PointNet [42]	pc	37.8	-	74.4	-
PointNet++ [25]	pc	48.2	-	76.7	-
DGCNN [26]	pc	37.4	-	73.3	-
RI-CNN [46]	pc	75.3	-	75.5	-
RI-GCN [181]	pc+n	77.2	-	77.3	-
VN-DGCNN [165]	pc	-	81.4	-	81.4
LGR-Net [185]	pc+n	-	-	-	82.8
PaRI-Conv [23]	pc+n	-	84.6	-	84.6
CRIN [186]	pc	80.5	-	80.5	-
PaRot [183]	pc	79.2	-	79.5	-
RI-PCR [187]	pc	-	80.3	-	80.4
LocoTrans [170]	pc	80.1	84.0	80.0	83.8
TetraSphere [184]	pc	-	82.3	-	82.3
RISurConv [24]	pc+n	81.5	-	81.5	-
Ours	pc	81.7	84.4	82.4	85.1
Ours	pc+n	82.9	85.0	83.8	85.5

Table 5.3: Part segmentation results on ShapeNetPart dataset. ‘C. mIoU’ stands for mIoU (%) over all classes while ‘I. mIoU’ denotes mIoU over all instances.

No.	RI Representation	Dim	z/SO(3)	
			C. mIoU	I. mIoU
1	PPF	4	81.1	84.1
2	Aug. PPF	8	81.8	84.2
3	SiPF-w/oDirection	5	82.4	84.5
4	SiPF	8	82.9	85.0

Table 5.4: Ablation study on the relative pose feature $\mathcal{P}_r^j$.

5.5.5 Ablation Studies

In this section, we present a series of ablation experiments to assess the effectiveness of the key components in our proposed framework. Unless otherwise stated, all experiments are performed on ShapeNetPart dataset with normal vector under the z/SO(3) setting and reported with C. mIoU (%) and I. mIoU (%).

Effectiveness of SiPF. To assess the effectiveness of SiPF in capturing both rotation invariance and global pose cues, we compare it with the standard Point Pair Feature (PPF) and the augmented PPF (Aug. PPF) from [23], which introduces an extra local axis to resolve PPF ambiguity. As shown in Table 5.4, SiPF clearly outperforms both variants, underscoring the importance of incorporating a global pose-aware reference. We

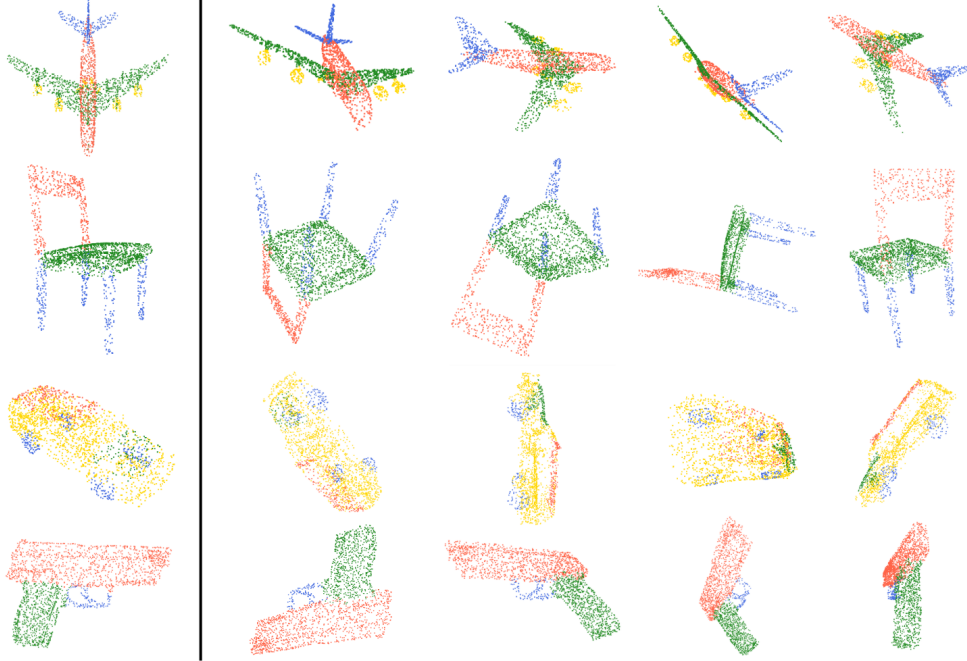


Figure 5.3: Visualization of part segmentation results on ShapeNetPart Dataset under $z/SO(3)$ setting. The leftmost column is the ground truth. The testing results under arbitrary rotations are on the rest columns.

RI Convolution	Params	FLOPs	$z/SO(3)$	
			C. mIoU	I. mIoU
Concat.	3.00M	4895M	82.1	84.2
AdaptConv	3.91M	18896M	82.0	84.2
PaRI-Conv	2.95M	4574M	82.4	84.5
RIAttnConv	3.01M	4795M	82.9	85.0

Table 5.5: Ablation study on the RI convolution.

also report ablations on SiPF’s directional components, confirming their contribution to enhanced robustness and discriminability.

Effectiveness of RIAttnConv. We evaluate the effectiveness of RIAttnConv by comparing it with alternative convolutional strategies, as shown in Table 5.5. All variants employ SiPF as the RI descriptor for fair comparison. In addition to two recent convolutional baselines [176, 23], we include a variant that directly concatenates SiPF with latent features without structural modification. RIAttnConv achieves the best performance while maintaining comparable or lower complexity on parameters and FLOPs, highlighting its efficiency in utilizing SiPF.

Computational Efficiency. Table 5.6 reports the computational cost of our method compared to recent

	LocoTrans	TetraSphere	RISurConv	Ours
Params	6.72M*	1.31M	4.06M	3.01M
FLOPs	7998M*	4092M	9726M	4795M
C. mIoU	80.1	-	81.3	81.7
I. mIoU	84.0	82.3	-	84.4

Table 5.6: Comparison of FLOPs, parameter count, and mIoU among different models with point coordinates only. (* indicates the results are estimated based on the classification architecture.)

state-of-the-art approaches. While some methods improve accuracy by increasing parameters and FLOPs, our model achieves the best segmentation performance with substantially lower complexity, demonstrating superior efficiency and scalability.

5.6 Conclusion

In this chapter, we propose the Shadow-informed Pose Feature (SiPF), a novel rotation-invariant descriptor that enriches local geometric features with global pose context via a learned shadow reference. Together with the introduced RIAttnConv operator, our framework effectively distinguishes symmetric or spatially ambiguous structures and achieves strong performance across classification and segmentation benchmarks under arbitrary rotations. A current limitation is that our evaluation is conducted on object-level datasets. Scene-level assessment remains an appealing future direction, which may further highlight pose ambiguity and validate the broader applicability of our framework.

Chapter 6

4D Local Modeling Toward Dynamic Global Perception for Ambiguity-free Rotation-Invariant Point Cloud Analysis

6.1 Introduction

The rapid advancement of 3D sensing technologies (such as LiDAR, stereo, and time-of-flight sensors) has led to the widespread adoption of point clouds in a variety of 3D vision tasks, including shape classification, segmentation, and object recognition [11, 10, 9]. As a compact geometric representation, point clouds are especially favoured in safety-critical domains such as autonomous driving and robotics [188, 189]. However, most existing deep learning models implicitly assume that input point clouds are consistently aligned to a canonical orientation [25, 26, 27]. In real-world settings, objects are frequently observed under arbitrary and unpredictable rotations, and even extensive data augmentation with random rotations cannot guarantee true robustness. This lack of inherent rotation invariance significantly degrades model performance and limits their deployment in unconstrained environments.

To address this issue, recent studies have focused on constructing rotation-invariant representations that eliminate the dependency on canonical alignment. These approaches often utilize auxiliary modules to transform coordinate frames [190, 23] or encode pairwise geometric relationships within enlarged local neighborhoods [22, 20]. Although effective at achieving rotation invariance, these methods typically incur increased

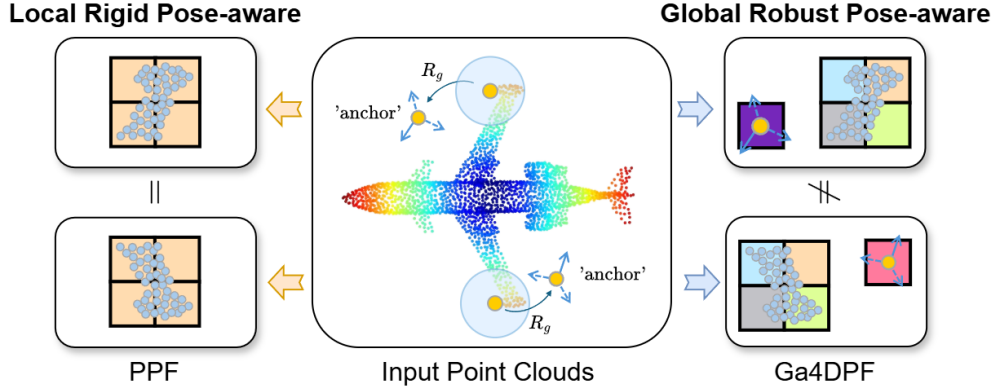


Figure 6.1: Traditional local pose-aware descriptors (e.g., PPF [1]) exhibit limited expressiveness and are prone to ambiguity in symmetric structures. In contrast, our approach enhances local geometric representation by lifting point-pair features into a higher-dimensional 4D space, enabling more robust and distinctive feature encoding. By further incorporating a global anchor derived from a dynamic global rotation R_g , Ga4DPF achieves more discriminative and rotation-consistent representations.

computational overhead and reduced performance on already aligned data. A fundamental challenge that persists is the issue of *structural ambiguity* in RI feature construction. First, rigid local RI encoding schemes lack robustness. The different spatial arrangements may yield nearly identical RI descriptors, which motivates continued refinement of local representations [170, 24]. Second, the use of local reference frames inherently discards global pose information, which is critical for disambiguating structures that are locally similar but globally distinct. While advances in expressive equivariant architectures [184, 165], existing RI frameworks remain limited by unstable local features and global ambiguity.

Our method is an equivariance-to-invariance framework that directly addresses the structural ambiguity of existing RI methods. Drawing inspiration from the observation that equivariant representations retain more structural information than invariant ones [191, 41], we design a representation pipeline that lifts point clouds into a 4D space using steerable 3D spherical neurons [192]. This lifting operation retains critical geometric structures while offering flexibility across varying viewpoints. Crucially, the lifting is fully learnable, enabling adaptive and elastic local feature construction beyond rigid coordinate dependence, thus enhancing the stability of RI representations. To further resolve global ambiguity, we introduce a novel mechanism that assigns each point a learnable global *anchor*, derived from a statistically consistent global rotation modeled by the Bingham distribution. As illustrated in Figure 6.1, this anchor injects global context into local neighborhoods, enabling the model to disambiguate symmetric or locally similar but globally distinct structures. By integrating robust 4D local representations with dynamic global pose awareness, our framework

produces ambiguity-free RI features while remaining compatible with standard RI learning pipelines. We validate the effectiveness of our method across multiple benchmark datasets on both classification and part segmentation tasks. Experimental results demonstrate that our approach significantly outperforms existing state-of-the-art RI methods in both performance and generalization capability.

Our main contributions are summarized as follows:

- We identify and formalize the intrinsic ambiguities in existing RI representations, both at local and global levels, revealing their limitations in learning robust and discriminative features under arbitrary rotations.
- We propose a novel 4D statistical framework that leverages steerable spherical neurons to construct rotation-equivariant local features in a lifted space, offering a principled solution for point-pair encoding in 4D and significantly reducing the ambiguity of spatial arrangements.
- We introduce a learnable global pose reference based on the Bingham distribution, jointly optimized with task-specific objectives, to dynamically capture and utilize global pose information.

6.2 Related Works

6.2.1 Rotation-Sensitive 3D Point Cloud Analysis

Early point cloud processing methods, such as PointNet [42] and its hierarchical extension PointNet++ [25], operate directly on raw 3D coordinates but are inherently sensitive to rotations due to their reliance on absolute spatial positions. Subsequent models like PointCNN [163] and DGCNN [26] introduced learnable operators and graph-based neighborhood aggregation to capture local geometry. More recent approaches, such as CurveNet [193] and methods grounded in conformal geometry [192], integrate stronger geometric priors. However, they still process in the original coordinate space, limiting their robustness to arbitrary orientation changes.

A complementary line of research attempts to address rotation sensitivity via canonical pose alignment. These methods predict a transformation that aligns each input to a consistent canonical orientation, often using modules like Spatial Transformer Networks (STNs) [194], self-consistent regressors [195], or PCA-based normalization [190, 164, 22]. While effective to some extent, these methods typically assume

category-level consistency and rely on extensive rotation augmentation during training, which limits their generalizability to unseen categories or partial observations.

Although these rotation-sensitive models perform well under aligned conditions, their performance deteriorates significantly under arbitrary rotations. This motivates the development of rotation-aware approaches that are inherently robust to orientation changes.

6.2.2 Rotation-Aware Methods

To overcome the challenges posed by arbitrary rotations in 3D point clouds, a range of rotation-aware approaches have been proposed. These techniques eliminate the assumption of pre-aligned input and can be broadly categorized into rotation-invariant and rotation-equivariant methods.

Rotation-invariant methods aim to produce features that remain unchanged under any 3D rotation. A dominant paradigm in RI research is local patch modeling, which constructs invariant features by encoding pairwise or patch-level geometric relationships. One common approach defines Local Reference Frames (LRFs) for each point and computes descriptors based on relative geometry [22, 20]. Classical examples include Point Pair Features (PPF) [1] and their variants, which encode distances and angles between point pairs in a rotation-independent manner. These methods eliminate the need for alignment and provide finer local geometric characterization. However, the construction of rigid local frames is sensitive to noise and perturbations, often leading to structural ambiguity in symmetric or repetitive patterns and reduced robustness of RI representations.

Rotation-equivariant methods, on the other hand, preserve rotation-dependent information by ensuring that features transform consistently with the input under rotations. Tensor field networks [167, 166] achieve strict equivariance through spherical harmonics, while recent vector-based approaches [196, 197, 198] offer more flexible and lightweight alternatives. These methods maintain local structural information and exhibit strong expressive power. However, they often lack explicit mechanisms for enforcing full rotation invariance, which may limit their effectiveness in applications requiring orientation-independent predictions.

Despite the complementary strengths of RI and equivariant methods, few works have successfully integrated both perspectives. RI approaches typically discard orientation cues to achieve invariance, while equivariant models retain such cues but do not guarantee strict invariance. Moreover, enforcing full invariance directly in 3D often comes at the cost of geometric fidelity [44]. A more effective strategy, motivated by recent theoretical insights [191, 41], is to first build a rotation-equivariant representation that preserves

geometric structure and then derive an invariant representation from it. Following this principle, we propose a unified framework that bridges the gap between equivariant and RI learning. Specifically, we embed point clouds into a higher-dimensional 4D space using steerable 3D spherical neurons [192], which linearize $\text{SO}(3)$ group actions and form a principled equivariant backbone. On top of this, we enhance local features with 4D equivariant modeling and introduce a dynamically learned global anchor, enabling the construction of ambiguity-free RI representations.

6.3 Preliminaries

This section introduces essential concepts of equivariance and invariance, along with prior theoretical foundations. These preliminaries provide insight into point cloud transformations and the inherent limitations of local receptive fields in point cloud convolutional architectures.

Let a point cloud be denoted as $P = \{\mathbf{p}_1, \dots, \mathbf{p}_N\} \in \mathbb{R}^{N \times 3}$, and let $\mathbf{R} \in \text{SO}(3)$ represent an arbitrary 3D rotation matrix. The rotated point cloud is defined as $P_R = P\mathbf{R}$.

6.3.1 Equivariance and Invariance

Let G be a group, and let $\{T_g : P \rightarrow P\}_{g \in G}$ be a family of transformations acting on the point cloud. A mapping $f : P \rightarrow P'$ is said to be *G-equivariant* if, for every $g \in G$, there exists a transformation $V_g : P' \rightarrow P'$ such that:

$$V_g[f(\mathbf{p})] = f(T_g[\mathbf{p}]) \quad \text{for all } g \in G, \mathbf{p} \in P. \quad (110)$$

G-invariance is the special case of equivariance in which V_g is the identity map for all $g \in G$, i.e.,

$$f(\mathbf{p}) = f(T_g[\mathbf{p}]) \quad \text{for all } g \in G, \mathbf{p} \in P. \quad (111)$$

To apply transformation T_g on a 3D point $\mathbf{p} \in \mathbb{R}^3$, the group elements $g \in \text{SO}(3)$ are typically represented by 3×3 rotation matrices $\mathbf{R}$ [199].

6.3.2 Receptive Fields in Point Cloud Convolutions

Local *receptive fields* are commonly used in point cloud learning to capture geometric context. Let $X = \{x_1, \dots, x_N\} \in \mathbb{R}^{c_{\text{in}}}$ denote the input feature set associated with point cloud P . For a reference

point $\mathbf{p}_r$, its local neighborhood is denoted by $\mathcal{N}(\mathbf{p}_r)$, and the stacked neighbor feature matrix is defined as $\mathbf{X}_r = [x_1, \dots, x_k]^\top \in \mathbb{R}^{k \times c_{\text{in}}}$.

Point cloud convolutions compute output features $X' = \{x'_1, \dots, x'_N\} \in \mathbb{R}^{c_{\text{out}}}$ at each point $\mathbf{p}_r$ by aggregating transformed features from its local neighborhood:

$$f(\mathbf{p}_r) = \bigwedge_{j \in \mathcal{N}(\mathbf{p}_r)} g(W_j^T \cdot h(\mathbf{p}_j \mid \Omega(\mathbf{p}_j)), x_r), \quad (112)$$

where h is a nonlinear function that maps the $\mathbf{p}_j$ to feature x_j , $\Omega(\mathbf{p}_j)$ denote the local patch centered at point $\mathbf{p}_j$, g is a one-layer MLP for aggregation [26], W_j^T denotes a learnable or dynamic kernel weight for the pair $(\mathbf{p}_r, \mathbf{p}_j)$, and $\wedge$ represents a permutation-invariant aggregation operator such as MAX, SUM, or AVG.

Since the function h also depend on the neighbourhood of $\mathbf{p}_j$, the effective receptive field of $\mathbf{p}_r$ becomes the union of its neighbors' patches:

$$\Gamma(\mathbf{p}_r) = \bigcup_{j \in \mathcal{N}(\mathbf{p}_r)} \Omega(\mathbf{p}_j). \quad (113)$$

While local receptive fields capture rich geometric structure, they are inherently limited in distinguishing globally identical configurations. For instance, consider two symmetric points $\mathbf{p}_{\text{left}}, \mathbf{p}_{\text{right}} \in P$ located on opposite sides of an airplane wing. There exists a rotation $\mathbf{R}_{\text{sym}} \in \text{SO}(3)$ such that their local receptive fields are rotationally equivalent:

$$\Gamma(\mathbf{p}_{\text{right}}) = \Gamma(\mathbf{p}_{\text{left}}) \mathbf{R}_{\text{sym}}. \quad (114)$$

As a result, the network produces identical features for both points, i.e., $f(\mathbf{p}_{\text{left}}) = f(\mathbf{p}_{\text{right}})$, leading to *global ambiguity* and the loss of critical global pose information.

6.4 The Proposed Method

In this section, we propose an ambiguity-free framework for RI point cloud analysis that explicitly addresses both local and global symmetry challenges. Our method consists of two key components: (1) A lifting mechanism that projects 3D point clouds into 4D via steerable spherical neurons, enabling robust and expressive point-pair representations; and (2) A global-aware module based on a Bingham-inspired formulation that complements local receptive fields by resolving global symmetry ambiguity while preserving rotation invariance. An overview of the full architectural pipeline is illustrated in Figure 6.2.

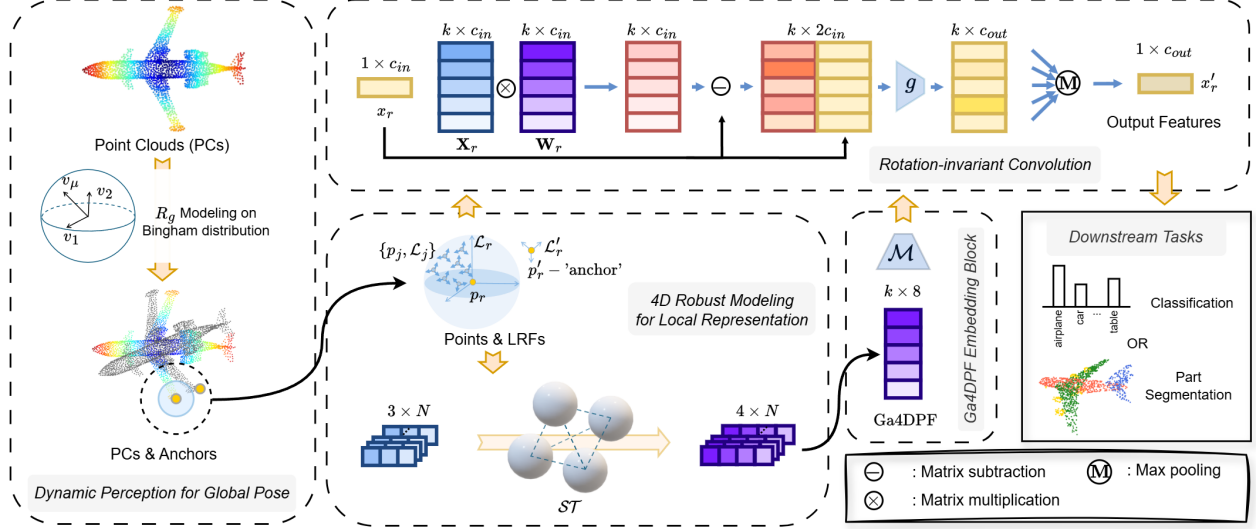


Figure 6.2: The pipeline of the proposed method. Our framework consists of two key components: (1) Dynamic Perception for Global Pose, which addresses the inherent *ambiguity of symmetric local structures* by estimating a globally consistent reference rotation R_g through Bingham distribution modeling, generating adaptive anchor points that provide additional global cues; (2) 4D Robust Modeling for Local Representation, where constructs LRFs $\{\mathcal{L}_j\}_{j=1}^k$ and a learnable steerable transform (ST) lifts the point cloud into 4D space, enabling fine-grained geometric encoding and enhancing the robustness of local representations against *point-pair ambiguity*. Finally, a Rotation-Invariant Convolution module embeds and aggregates the resulting Ga4DPF features, producing discriminative representations suitable for downstream tasks such as shape classification and part segmentation.

6.4.1 4D Robust Modeling for Local Representation

LRF Construction. LRFs are fundamental to RI feature extraction. For each point $\mathbf{p}_r \in P$, we construct an LRF $\mathcal{L}_r = \{\partial_r^1, \partial_r^2, \partial_r^3\} \in \mathbb{R}^{3 \times 3}$ via Gram–Schmidt orthonormalization using two directional vectors e_r^1 and e_r^2 :

$$\partial_r^1 = e_r^1, \quad \partial_r^3 = \frac{\partial_r^1 \times e_r^2}{\|\partial_r^1 \times e_r^2\|}, \quad \partial_r^2 = \partial_r^3 \times \partial_r^1. \quad (115)$$

Following prior works [23, 24], we adopt a purely local construction where e_r^2 is defined as the vector from the barycenter of the k -nearest neighbours $\mathbf{m}_r = \frac{1}{k} \sum_{j=1}^k \mathbf{p}_j^r$ to the reference point $\mathbf{p}_r$, i.e., $e_r^2 = \mathbf{m}_r - \mathbf{p}_r$, and $e_r^1 = n_r$ is the surface normal. This construction ensures geometric consistency and forms the input feature X in our method.

Steerable 3D Spherical Neurons. A steerable spherical neuron [192] consists of a learnable spherical decision surface $\mathbf{S} \in \mathbb{R}^5$ [200, 201] and three rotated variants located at the vertices of a regular tetrahedron. Let $\mathbf{R}_O$ denote the geodesic rotation that moves the learned center $\mathbf{c}_0$ to $\frac{\|\mathbf{c}_0\|}{\sqrt{3}} (1, 1, 1)$, and let $\{\mathbf{R}_{T_i}\}_{i=0}^3$ denote the rotation isomorphisms in $\mathbb{R}^5$ that rotate from $(1, 1, 1)$ to each tetrahedral vertex (with $\mathbf{R}_{T_0} = \mathbf{I}_5$).

The resulting 4×5 filter bank is defined as:

$$B(\mathbf{S}) = [(\mathbf{R}_O^\top \mathbf{R}_{T_i} \mathbf{R}_O \mathbf{S})^\top]_{i=0\dots 3}, \quad (116)$$

which can be viewed as a mapping $f_{4S}(\cdot; \mathbf{S}) : \mathbb{R}^5 \rightarrow \mathbb{R}^4$, parameterized by five learnable weights. Given an embedded input point $\mathbf{P} \in \mathbb{R}^5$ and a 3D rotation representation $\mathbf{R} \in \mathbb{R}^5$, the filter bank satisfies the equivariance condition [200]:

$$V_{\mathbf{R}} B(\mathbf{S}) \mathbf{P} = B(\mathbf{S}) \mathbf{R} \mathbf{P}, \quad (117)$$

where the induced 4×4 action on the output is:

$$V_{\mathbf{R}} = \mathbf{M}^\top \mathbf{R}_O \mathbf{R} \mathbf{R}_O^\top \mathbf{M},$$

$$\mathbf{M} = \frac{1}{2} \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \quad (118)$$

Thus, $V_{\mathbf{R}} \in G \subset \text{SO}(4)$ represents the rotation group acting in the output space defined by the tetrahedral basis. We use this equivariant filter output as a 4D replacement for the original 3D points.

4D Point Pair Features (4DPPF). We define the 4D lifting operation as a steerable transform $\mathcal{ST}(\cdot; \mathbf{S}) : \mathbb{R}^{N \times 3} \rightarrow \mathbb{R}^{N \times 4}$. Following standard point cloud processing pipelines [42, 26], the operation $B(\mathbf{S}) \mathbf{P}$ is applied uniformly across points, yielding a rotation-equivariant output $\mathbf{P}' = \mathcal{ST}(\mathbf{P}; \mathbf{S}) \in \mathbb{R}^{N \times 4}$.

To encode local geometric relationships in this 4D space, we define the 4D point pair feature between a reference point $\mathbf{p}'_r$ and its neighbour $\mathbf{p}'_j$ as:

$$4\text{DPPF}(\mathbf{p}'_r, \mathbf{p}'_j) = (\|d\|_2, \cos(\alpha_1), \cos(\alpha_2), \cos(\alpha_3)),$$

$$\alpha_1 = \angle(\Delta_r^1, d), \quad \alpha_2 = \angle(\Delta_j^1, d), \quad \alpha_3 = \angle(\Delta_r^1, \Delta_j^1), \quad (119)$$

where $d = \mathbf{p}'_j - \mathbf{p}'_r$, $\angle(\cdot, \cdot)$ denotes the angle between two 4D vectors, and $\Delta = \mathcal{ST}(\partial; \mathbf{S})$ represents the orthonormal basis vectors lifted to 4D space.

While Equation 119 resembles the classic 3D PPF formulation [1], it fundamentally enhances discriminability by leveraging 4D geometry. Unlike 3D PPFs, which often yield identical descriptors for symmetric point pairs [175, 23], our 4DPPF naturally resolves these ambiguities due to its richer representational capacity.

Theorem 1 (Disambiguation Guarantee of 4DPPF.). *Let $\mathbf{p}_r$ be the reference point with local axis ∂_r^1 . Consider a set of neighbors $\mathbf{p}_j(\theta)$ obtained by rotating a fixed point $\mathbf{p}_j(0)$ around ∂_r^1 by angle θ , the lifted 4D point pair feature satisfies:*

$$4DPPF(\mathbf{p}'_r, \mathbf{p}'_j) \neq 4DPPF(\mathbf{p}'_r, \mathbf{p}'_j(\theta)), \quad \theta \neq 0, \quad (120)$$

where $\mathbf{p}' = \mathcal{ST}(\mathbf{p})$ and $V_R(\theta) = M^\top R_O R(\theta) R_O^\top M$ is the induced 4D rotation.

Proof. For points $\mathbf{p}_r, \mathbf{p}_j \in \mathbb{R}^3$, if $\mathbf{p}_r$ lies on the axis of rotation, then:

$$\langle R(\theta)\mathbf{p}_j, \mathbf{p}_r \rangle = \langle \mathbf{p}_j, R(\theta)^\top \mathbf{p}_r \rangle = \langle \mathbf{p}_r, \mathbf{p}_j \rangle, \quad (121)$$

implying that $\|d(\theta)\|_2$ and the 3D PPF remain unchanged with θ , and $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors. In contrast, for lifted 4D points $\mathbf{p}'_r, \mathbf{p}'_j \in \mathbb{R}^4$, we have:

$$\langle V_R(\theta)\mathbf{p}'_j, \mathbf{p}'_r \rangle = \langle \hat{\mathbf{p}}_j, R_O R(-\theta) R_O^\top \hat{\mathbf{p}}_r \rangle, \quad (122)$$

where $\hat{\mathbf{p}} = M\mathbf{p}'$. Unlike the 3D case, $\mathbf{p}'_r$ is generally *not* an eigenvector of $V_R(\theta)$, hence $R_O R(-\theta) R_O^\top \hat{\mathbf{p}}_r \neq \hat{\mathbf{p}}_r$ for $\theta \neq 0$, and the inner product varies with θ , leading to:

$$\|d(\theta)\|_2 = \sqrt{\|\mathbf{p}'_j\|_2^2 + \|\mathbf{p}'_r\|_2^2 - 2\langle \mathbf{p}'_j, V_R(\theta)^\top \mathbf{p}'_r \rangle}. \quad (123)$$

Likewise, the basis vectors $\Delta_j^1(\theta) = V_R(\theta)\Delta_j^1(0)$ and $\hat{\Delta}_r^1$ satisfy $R_O R(-\theta) R_O^\top \hat{\Delta}_r^1 \neq \hat{\Delta}_r^1$, so the angles $\cos(\alpha_1), \cos(\alpha_2), \cos(\alpha_3)$ also vary with θ . Therefore,

$$4DPPF(\mathbf{p}'_r, \mathbf{p}'_j) \neq 4DPPF(\mathbf{p}'_r, \mathbf{p}'_j(\theta)) \quad (124)$$

for $\theta \neq 0$, except on a degenerate set of measure zero. $\square$

6.4.2 Dynamic Perception for Global Pose

Global-Aware Pose Feature. While local 4DPPF features effectively capture fine-grained geometric relationships, they are inherently constrained by the local receptive fields. As a result, they may fail to distinguish symmetric configurations due to a lack of global context, an issue we refer to as *global ambiguity*. To overcome this limitation, we propose the *Global-aware 4D Pose Feature* (Ga4DPF), which augments local descriptors with a global reference. Formally, Ga4DPF is defined as:

$$\text{Ga4DPF}(\mathbf{p}_r) = \left(4\text{DPPF}(\mathbf{p}'_r, \mathbf{p}'_j), 4\text{DPPF}(\mathbf{p}_r^\dagger, \mathbf{p}'_j) \right) \in \mathbb{R}^8, \quad (125)$$

where the *anchor point* is defined as $\mathbf{p}_r^\dagger = \mathcal{ST}(\mathbf{p}_r \mathbf{R}_g; \mathbf{S})$, with $\mathbf{R}_g \in \text{SO}(3)$ a consistent global rotation defined later.

By incorporating the anchor $\mathbf{p}_r^\dagger$, the receptive field $\Gamma(\mathbf{p}_r)$ is no longer restricted to the local neighbourhood, but aligned with a global reference. This breaks the local equivariance described in Equation 114, i.e., $\Gamma(\mathbf{p}_r, \mathbf{p}_r^\dagger) \neq \Gamma(\mathbf{p}_r, \mathbf{p}_r^\dagger) \mathbf{R}_{\text{sym}}$, and injects global pose awareness into the representation:

$$\text{Ga4DPF}(\mathbf{p}_r) \neq \text{Ga4DPF}(\mathbf{p}_r \mathbf{R}_{\text{sym}}). \quad (126)$$

Dynamic Global Rotation. The anchor point $\mathbf{p}_r^\dagger$ is constructed through a global rotation $\mathbf{R}_g$, which is applied consistently across the dataset. However, using a fixed $\mathbf{R}_g$ may not improve discriminability, particularly when $\mathbf{R}_g$ coincides with a scene symmetry transformation $\mathbf{R}_{\text{sym}}$ as shown in Equation 114, in which case Ga4DPF cannot effectively disambiguate symmetric local regions.

To overcome this, we propose a *dynamic strategy* that models uncertainty in choosing $\mathbf{R}_g$ by employing the *Bingham distribution* $\mathcal{B}(q \mid \mathbf{V}, \mathbf{\Lambda})$ over the unit quaternion hypersphere $\mathbb{S}^3$, where q denotes a global rotation encoded as a unit quaternion. The Bingham distribution is defined as:

$$\mathcal{B}(q \mid \mathbf{V}, \mathbf{\Lambda}) = \frac{1}{F(\mathbf{\Lambda})} \exp \left(q^\top \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top q \right), \quad (127)$$

where $F(\mathbf{\Lambda})$ is the normalization constant, $\mathbf{V} \in \mathbb{R}^{4 \times 4}$ is an orthogonal matrix that specifies the mode and dispersion axes, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3, 0)$ controls the concentration with $\lambda_1 < \lambda_2 \leq \lambda_3 < 0$.

To couple global pose learning with downstream task performance, we jointly optimize the Bingham

parameters and network using the following objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \delta \cdot \sqrt{(\mathcal{L}_{\text{bingham}} - 0.1 \cdot \mathcal{L}_{\text{task}})^2}, \quad (128)$$

where $\mathcal{L}_{\text{bingham}}$ is the negative log-likelihood of sampled quaternions, and δ is a balancing hyperparameter (empirically set to 0.8). The subtraction term encourages consistency between global pose adaptivity and task performance.

To enable gradient-based learning, we adopt a sampling-based approximation of the Bingham distribution. Given learned $\mathbf{V}$ and $\mathbf{\Lambda}$, we sample N_s quaternions $\{q_i\}_{i=1}^{N_s} \sim \mathcal{B}(q \mid \mathbf{V}, \mathbf{\Lambda})$, and compute the global rotation $\mathbf{R}_g = \text{Rot}(q_g)$, where the representative quaternion q_g is estimated via the geometric mean:

$$q_g = \arg \max_{q \in \mathbb{S}^3} \sum_{i=1}^{N_s} \langle q, q_i \rangle^2, \quad (129)$$

which corresponds to the dominant eigenvector of the matrix $\sum_i q_i q_i^\top$. In practice, we use $N_s = 10$.

6.4.3 Network Architecture

To validate the robustness and effectiveness of the proposed 4DPPF and dynamic global rotation, we integrate them into the PaRI-Conv framework [23] and evaluate them using two widely adopted point cloud backbones for fair evaluation.

Shape Classification. For classification, we employ DGCNN [26] as the backbone and replace all Edge-Conv layers with PaRI-Conv modules enhanced by Ga4DPF. At each layer, we construct a k -nearest neighbor graph in feature space using Euclidean distance, with $k = 20$.

Shape Part Segmentation. For part segmentation, we follow the AdaptConv design [176]. Specifically, we stack five Ga4DPF-augmented PaRI-Conv layers, interleaved with three intermediate pooling layers to progressively expand the receptive field. Across all layers, the k -nearest neighbor graph is constructed in Euclidean space with $k = 40$.

6.5 Experiments

We evaluate the proposed method on three widely used benchmarks: ModelNet40 [159] for shape classification, ShapeNetPart [177] for part segmentation, and ScanObjectNN [178] for real-world object classification. To systematically assess rotation robustness, we adopt three standard evaluation protocols: z/z, z/SO(3), and SO(3)/SO(3), where z and SO(3) denote upright and arbitrary 3D rotations, respectively. In addition, we perform comprehensive ablation studies to investigate the contribution of each key component in our framework.

6.5.1 Implementation Details

We adopt stochastic gradient descent (SGD) for optimization, with an initial learning rate of 0.1, decayed to 0.001 via cosine annealing [179]. All models are trained for 300 epochs. The batch size is set to 32 for classification and 16 for segmentation, with a dropout rate of 0.5 applied to fully connected layers for regularization.

To ensure RI initialization, we define a point-wise descriptor for each point $\mathbf{p}_i$ based on its relative geometry to the global centroid O [180], comprising three elements: $\|\overrightarrow{O\mathbf{p}_i}\|_2$, $\sin(\angle(\partial_i^1, \overrightarrow{O\mathbf{p}_i}))$, and $\cos(\angle(\partial_i^1, \overrightarrow{O\mathbf{p}_i}))$. Here, ∂_i^1 denotes the local axis, typically the surface normal under pc+n setting. For coordinate-only settings, we replace ∂_i^1 with $\overrightarrow{\mathbf{p}_r\mathbf{m}_r}$, the direction from the barycenter to the point.

6.5.2 3D Shape Classification

Dataset. We first evaluate our approach on the widely adopted ModelNet40 benchmark [159], which comprises 12,311 synthetic CAD models across 40 object categories, with 9,843 samples for training and 2,468 for testing. Following the standard protocol of [25], we uniformly sample 1,024 points from each shape. Random scaling and translation are applied to improve generalization during training.

Results. Table 6.1 summarizes the classification accuracy on ModelNet40 under three protocols: z/z, z/SO(3), and SO(3)/SO(3), reflecting increasing rotational difficulty. Rotation-sensitive methods (e.g., PointNet++ [25], DGCNN [26]) perform well under z/z but degrade sharply under unseen rotations (e.g., PointNet++ drops from 89.3% to 28.6%). In contrast, RI-based methods such as RIConv [46] and ClusterNet [169] maintain stable accuracy (86%–87%) but are limited in capturing rich geometry using coordinates

Methods	Inputs	z/z	z/SO(3)	SO(3)/SO(3)
Rotation-sensitive				
PointNet [42]	pc	89.2	16.2	75.5
PointNet++ [25]	pc+n	89.3	28.6	85.0
DGCNN [26]	pc	92.2	20.6	81.1
Rotation-aware				
RIConv [46]	pc	86.5	86.4	86.4
ClusterNet [169]	pc	87.1	87.1	87.1
RI-GCN [181]	pc+n	91.0	91.0	91.0
OrientedMP [182]	pc+n	88.4	88.4	88.4
ELGANet [21]	pc	90.3	90.3	90.3
PaRot [183]	pc	90.9	91.0	90.8
TetraSphere [184]	pc	90.5	90.5	90.3
Ours	pc	91.9	91.9	91.9
Ours	pc+n	92.8	92.8	92.8

Table 6.1: Shape classification accuracy (%) on ModelNet40 dataset. ‘pc’ and ‘n’ represent the 3D coordinates and normal vectors of input data, respectively.

alone. More recent methods like ELGANet [21], PaRot [183], and TetraSphere [184] further improve performance (90.3%–90.5%) without requiring normals. Our method achieves 91.9% using only coordinates, outperforming several normal-dependent methods such as OrientedMP [182] (88.4%) and RI-GCN [181] (91.0%). With normals, performance increases to 92.8%, surpassing PaRot and RI-GCN by up to 1.8%. These results highlight the effectiveness of our 4D representation and global pose integration in learning expressive RI representations.

6.5.3 Real-World Shape Classification

Dataset. To evaluate performance under real-world conditions, we conduct experiments on the ScanObjectNN dataset [178], which contains over 15,000 point clouds scanned from 2,902 real-world indoor objects. We adopt the challenging OBJ_BG subset, consisting of 2,890 samples from 15 categories, split into 2,312 for training and 578 for testing. This subset introduces real-world challenges such as noise, occlusion, and background clutter, which significantly impact rotation stability. As normals are unavailable in the processed data, only raw point coordinates are used.

Results. Table 6.2 presents classification accuracy under z/SO(3) and SO(3)/SO(3) protocols. Our method achieves the highest accuracy across both settings, outperforming the previous best model, LocoTrans [170],

Methods	z/SO(3)	SO(3)/SO(3)
Input pc only		
PointNet [42]	16.7	54.7
PointNet++ [25]	15.0	47.4
DGCNN [26]	17.7	71.8
RIConv [46]	75.3	75.5
RI-GCN [181]	80.5	80.6
OrientedMP [182]	76.7	77.2
PaRI-Conv [23]	83.3	83.3
LocoTrans [170]	85.0	84.5
Ours	87.4	87.3

Table 6.2: Real-world classification accuracy (%) on ScanObjectNN OBJ_BG dataset.

Method	Input	z/SO(3)		SO(3)/SO(3)	
		C. mIoU	I. mIoU	C. mIoU	I. mIoU
PointNet [42]	pc	37.8	-	74.4	-
PointNet++ [25]	pc	48.2	-	76.7	-
DGCNN [26]	pc	37.4	-	73.3	-
RI-GCN [181]	pc+n	77.2	-	77.3	-
PaRI-Conv [23]	pc+n	-	84.6	-	84.6
CRIN [186]	pc	80.5	-	80.5	-
PaRot [183]	pc	79.2	-	79.5	-
LocoTrans [170]	pc	80.1	84.0	80.0	83.8
TetraSphere [184]	pc	-	82.3	-	82.3
RISurConv [24]	pc+n	81.5	-	81.5	-
Ours	pc	81.3	84.0	81.2	83.9
Ours	pc+n	82.2	85.1	82.6	85.0

Table 6.3: Part segmentation results on ShapeNetPart dataset. ‘C. mIoU’ stands for mIoU (%) over all classes while ‘I. mIoU’ denotes mIoU over all instances.

by 2.4% and 2.8%, respectively. This demonstrates strong generalization and robustness in real-world conditions. Compared to PaRI-Conv [23], which relies on rigid, handcrafted local reference frames and suffers under noise and occlusion (83.3%), our method dynamically lifts point-normal pairs into a 4D space and learns features using steerable spherical neurons. This flexible and learnable design enhances both local stability and geometric expressiveness. Moreover, the minimal performance gap (0.1%) between z/SO(3) and SO(3)/SO(3) further confirms the rotation invariance of our model.

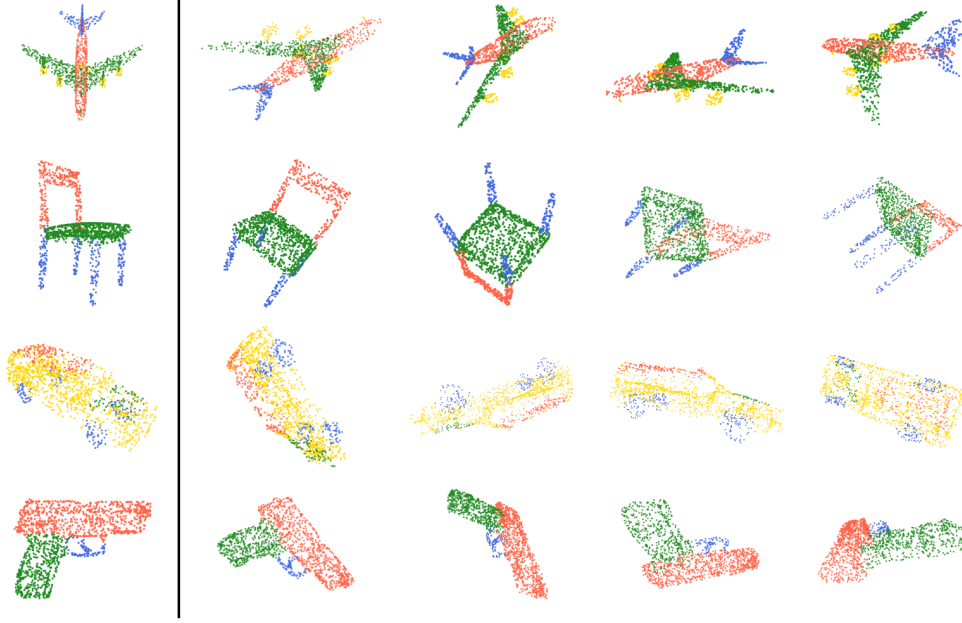


Figure 6.3: Visualization of part segmentation results on ShapeNetPart Dataset under $z/SO(3)$ setting. The leftmost column is the ground truth. The testing results under arbitrary rotations are on the rest columns.

6.5.4 Shape Part Segmentation

Dataset. We further evaluate our method on the ShapeNetPart dataset [177], a large-scale benchmark for fine-grained 3D part segmentation. It contains 16,881 models across 16 object categories, each annotated with 2 to 6 semantic parts, resulting in a total of 50 unique part labels. Following the standard protocol of [25], we adopt the official train/test split and uniformly sample 2,048 points per shape as input.

Results. As shown in Table 6.3, our method achieves the highest mean Intersection-over-Union (mIoU) across all classes and all instances. Under the $pc+n$ and $z/SO(3)$ setting, our model achieves a class mIoU of 82.2%, surpassing RISurConv [24] (81.5%), demonstrating superior segmentation quality with normal vectors. Under the pc and $z/SO(3)$ setting, our method attains a class mIoU of 81.3%, outperforming LocoTrans [170] by 1.2% (80.1%), highlighting the robustness of our learned rotation-invariant features even without normal input. Figure 6.3 visualizes representative segmentation results under the $z/SO(3)$ protocol, further illustrating the model’s robustness to arbitrary rotations.

6.5.5 Model Complexity

We compare the computational efficiency of our model against recent RI methods in Table 6.4. Our method achieves competitive segmentation performance while maintaining a favorable balance between

	LocoTrans	TetraSphere	RISurConv	Ours
Params	6.72M*	1.31M	4.06M	2.25M
FLOPs	7998M*	7996M	12120M	3841M
C. mIoU	80.1	-	81.3	81.3
I. mIoU	84.0	82.3	-	84.0

Table 6.4: Comparison of FLOPs, parameter count, and mIoU on ShapeNetPart (z/SO(3)) with point coordinates only. (* indicates the results are estimated based on the classification architecture.)

No.	RI Representation	Dim	MN40.	ScanObj.
1	PPF	4	91.8	82.8
2	Aug.PPF [23]	8	92.4	83.3
3	4DPPF($\mathbf{p}'_r, \mathbf{p}'_j$)	4	92.4	84.4
4	(PPF, 4DPPF($\mathbf{p}'_r, \mathbf{p}'_j$))	8	92.5	86.7
5	(4DPPF($\mathbf{p}'_r, \mathbf{p}'_j$), 4DPPF($\mathbf{p}'_r, \mathbf{p}'_j$))	8	92.8	87.4

Table 6.5: Ablation study on the relative pose feature. Reported results are classification accuracy under the z/SO(3) setting.

accuracy and complexity. Specifically, the proposed model has only 2.25M parameters and 3841M FLOPs, significantly lower than LocoTrans [170] and RISurConv [24]. Compared to TetraSphere [184], which has a similar parameter count, our method improves I. mIoU by 1.7% while using fewer FLOPs. These results highlight the efficiency and scalability of our model in resource-constrained scenarios.

6.5.6 Ablation Studies

In this section, we conduct ablation analysis to assess the effectiveness of key components in our framework. Specifically, we analyze the influence of (i) the relative pose representation and (ii) the number of samples N_s used in geometric mean estimation for global pose aggregation. All experiments are conducted under the z/SO(3) protocol for consistency and fairness.

Effect of Relative Pose Representation. Table 6.5 presents a detailed comparison of different RI representations. The baseline PPF uses the classic point pair feature formulation [1]. Aug.PPF [23] incorporates handcrafted enhancements to mitigate ambiguity from symmetric point pairs (Equation 121). In contrast, our learned formulation 4DPPF($\mathbf{p}'_r, \mathbf{p}'_j$), derived from lifted point pairs using steerable spherical neurons, replaces handcrafted descriptors and achieves superior performance on ScanObjectNN, despite operating in a lower-dimensional feature space. The hybrid design (PPF, 4DPPF($\mathbf{p}'_r, \mathbf{p}'_j$)) combines the classic PPF

N_s	1	5	10	15	20
Acc.(ModelNet40)	92.3	92.5	92.8	92.8	92.8
Acc.(ScanObjectNN)	85.5	85.9	87.4	87.2	86.9

Table 6.6: Ablation study on the sampling number N_s for geometric mean estimation. Reported results are classification accuracy under the z/SO(3) setting.

with our global-aware variant and significantly outperforms both PPF and Aug.PPF alone. Finally, our full representation Ga4DPF($\mathbf{p}_r$), defined as $(4DPPF(\mathbf{p}'_r, \mathbf{p}'_j), 4DPPF(\mathbf{p}^\dagger_r, \mathbf{p}'_j))$, integrates both local and global-aware 4DPPFs. It consistently achieves the highest accuracy across both ModelNet40 and ScanObjectNN, validating the effectiveness of unifying local and global relational cues.

Effect of Sampling Number N_s . Table 6.6 investigates the impact of the sampling number N_s used for estimating the geometric mean. We report classification accuracy on both ModelNet40 and ScanObjectNN under the z/SO(3) setting. Increasing N_s from 1 to 10 yields consistent performance improvements, with the highest accuracy achieved at $N_s = 10$, with 92.8% on ModelNet40 and 87.4% on ScanObjectNN. However, further increasing N_s yields no improvement or slight degradation, indicating that $N_s = 10$ offers an effective balance between accuracy and efficiency.

6.5.7 Visualization of Learned RI Features

We visualize the dense point-wise features learned under various rotations in Figure 6.4. Compared to DGCNN [26], whose features exhibit significant variation under different poses, and PaRI-Conv [23], which shows improved consistency, our method produces the most rotation-invariant and semantically discriminative representations. Notably, our features remain stable across arbitrary rotations and demonstrate fine-grained distinctions between symmetric structures (such as airplane wings and engines) indicating enhanced local expressiveness. While PaRI-Conv also leverages LRFs to achieve rotation invariance, its reliance on handcrafted frame construction restricts its capacity to resolve structural ambiguities in symmetric geometries. In contrast, our model integrates a global pose anchor within the lifted 4D feature space, allowing more consistent feature alignment across poses and enabling compact yet expressive representation learning.

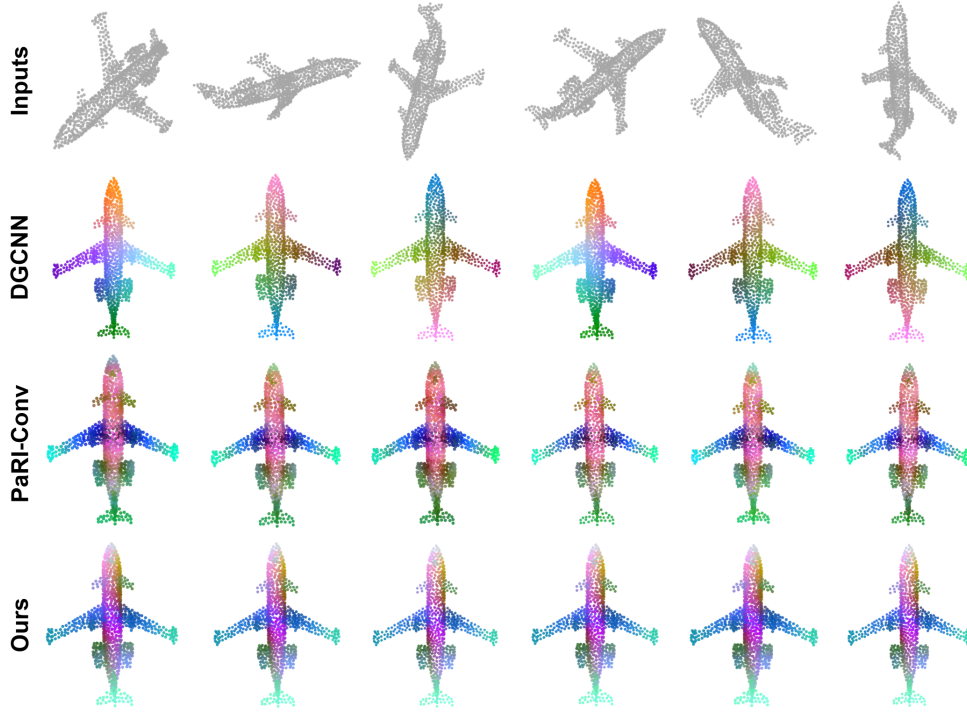


Figure 6.4: Visualization of learned dense features for shape part segmentation under different rotations. Our method (bottom row) yields more consistent and discriminative representations, especially for symmetric parts, while preserving rotation invariance.

6.6 Conclusion

In this work, we presented Ga4DPF, a robust and global-aware RI representation that directly addresses the long-standing challenge of structural ambiguity in RI point cloud analysis. Our key insight lies in bridging rotation equivariance and invariance through joint local–global modeling: a learnable steerable transform enables 4D lifting for robust local feature construction against point-pair ambiguity, while a dynamic global pose reference modeled by the Bingham distribution provides consistent orientation cues. By jointly eliminating both local and global ambiguities, Ga4DPF yields a compact, discriminative, and truly RI representation. Extensive experiments on multiple benchmarks validate the effectiveness and efficiency of our method, significantly outperforming existing RI methods and highlighting its potential as a principled direction for future RI representation learning. In future work, we plan to extend this representation to more complex scenes, explore its integration with generative models, and investigate its application in multimodal and spatiotemporal frameworks.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

This thesis investigated structured probabilistic latent modeling as a principled foundation for representation learning and rotation-invariant 3D perception. Departing from conventional Gaussian-centric and deterministic latent assumptions, the thesis demonstrated that many longstanding limitations in unsupervised learning and 3D geometric understanding originate not from network architectures, but from a mismatch between latent space structure and the intrinsic statistical and geometric properties of data.

At the level of probabilistic representation learning, the thesis first addressed the problem of asymmetric and constrained latent variables. By integrating Gamma mixture priors into variational autoencoders, the proposed GamMM-VAE established an effective framework for modeling non-negative and skewed latent embeddings, leading to improved clustering quality and generative capability. This work highlighted the importance of aligning latent priors with data constraints, and demonstrated that structured non-Gaussian priors can significantly enhance both expressiveness and interpretability in unsupervised learning.

The thesis then extended structured latent modeling to directional and axially symmetric data through a novel variational inference framework for Bingham mixture models. By exploiting the mathematical relationship between the Bingham distribution and zero-mean Gaussian distributions, the proposed approach enabled scalable and tractable inference without reliance on intractable normalization constants. This contribution provided a general and efficient tool for modeling uncertainty on hyperspherical manifolds, bridging classical Euclidean probabilistic modeling and modern directional statistics.

Building upon these probabilistic foundations, the thesis advanced representation learning toward uncertainty-aware 3D perception. Through BRVSNet, rotational uncertainty was modeled directly on the unit quaternion hypersphere using Bingham mixtures, enabling joint self-supervised rotation estimation and consistent multi-view generation within a unified probabilistic latent space. This framework demonstrated that probabilistic pose modeling not only improves robustness under symmetry and occlusion, but also provides a coherent mechanism for uncertainty-aware generative reasoning.

Beyond rotation estimation, the thesis addressed a fundamental limitation of rotation-invariant 3D learning: the inherent ambiguity induced by strictly local invariant descriptors. By introducing global pose awareness through probabilistic modeling, the proposed Shadow-informed Pose Feature (SiPF) enriched local invariant representations with global directional context while preserving strict rotation invariance. Further extending this idea, the Ga4DPF framework unified robust local modeling via 4D lifting and dynamic global pose estimation via Bingham distributions, effectively resolving both local point-pair ambiguity and global symmetry-induced feature collapse. These results demonstrated that rotation invariance and discriminative power are not mutually exclusive when global pose information is modeled probabilistically.

In conclusion, the contributions of this thesis establish a coherent framework in which structured, non-Gaussian probabilistic latent models serve as a unifying mechanism for expressiveness, disambiguation, and uncertainty quantification. By integrating domain-aligned probabilistic structure directly into latent spaces, the proposed methods advance the state of representation learning from heuristic modeling toward principled, uncertainty-aware geometric reasoning.

7.2 Future Works

The research directions explored in this thesis open several promising avenues for future investigation.

First, adaptive and nonparametric latent modeling remains an important extension. While the proposed mixture-based frameworks demonstrate strong performance, incorporating Bayesian nonparametric techniques could enable automatic determination of latent complexity, such as the number of mixture components, and further improve scalability and flexibility in unsupervised settings.

Second, integrating structured probabilistic latent models with more expressive generative paradigms presents a natural next step. In particular, combining directional latent distributions with diffusion-based or

autoregressive generative models may significantly enhance synthesis fidelity while preserving uncertainty-aware geometric reasoning. Such integration could extend the applicability of the proposed frameworks to complex, high-resolution visual and 3D domains.

Third, extending global pose-aware rotation-invariant representations to scene-level and spatiotemporal settings constitutes a compelling direction. Real-world environments introduce hierarchical structure, dynamic interactions, and multi-object relationships, which may further amplify the role of probabilistic global pose modeling in resolving ambiguity and ensuring consistency across time and space.

Finally, structured probabilistic modeling offers opportunities for multimodal representation learning, where geometric, visual, and semantic cues coexist. Exploring how structured latent priors can unify heterogeneous modalities under shared probabilistic representations may lead to more robust and interpretable perception systems in robotics, embodied AI, and autonomous sensing.

In summary, this thesis demonstrates that embedding principled probabilistic structure into latent spaces is a powerful and general strategy for advancing representation learning. By moving beyond isotropic and deterministic assumptions, structured probabilistic latent modeling provides a solid foundation for uncertainty-aware, geometry-consistent, and semantically meaningful learning, laying the groundwork for future progress in both theoretical modeling and real-world perception systems.

Bibliography

- [1] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 998–1005. Ieee, 2010.
- [2] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- [4] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, et al. Deep learning. *nature*, 521 (7553), 436-444. *Google Scholar Google Scholar Cross Ref Cross Ref*, 2015.
- [5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [6] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
- [7] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.
- [8] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.
- [9] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for

- 3d point cloud understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9902–9912, 2022.
- [10] Lunhao Duan, Shanshan Zhao, Nan Xue, Mingming Gong, Gui-Song Xia, and Dacheng Tao. Condaformer: Disassembled transformer with local structure enhancement for 3d point cloud understanding. *Advances in Neural Information Processing Systems*, 36:23886–23901, 2023.
- [11] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023.
- [12] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [14] Zhanyu Ma and Arne Leijon. Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2160–2173, 2011.
- [15] Zhanyu Ma, Pravin Kumar Rana, Jalil Taghia, Markus Flierl, and Arne Leijon. Bayesian estimation of dirichlet mixture model with variational inference. *Pattern Recognition*, 47(9):3143–3157, 2014.
- [16] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543, 2004.
- [17] Wentao Fan, Nizar Bouguila, and Djemel Ziou. Variational learning for finite dirichlet mixture models and applications. *IEEE transactions on neural networks and learning systems*, 23(5):762–774, 2012.
- [18] Yukun Ding, Jinglan Liu, Xiaowei Xu, Meiping Huang, Jian Zhuang, Jinjun Xiong, and Yiyu Shi. Uncertainty-aware training of neural networks for selective medical image segmentation. In Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Hervé Lombaert, and Christopher

- Pal, editors, *International Conference on Medical Imaging with Deep Learning, MIDL 2020, 6-8 July 2020, Montréal, QC, Canada*, volume 121 of *Proceedings of Machine Learning Research*, pages 156–173. PMLR, 2020.
- [19] Xiaolin Tang, Kai Yang, Hong Wang, Jiahang Wu, Yechen Qin, Wenhao Yu, and Dongpu Cao. Prediction-uncertainty-aware decision-making for autonomous vehicles. *IEEE Trans. Intell. Veh.*, 7(4):849–862, 2022.
- [20] Jianyun Xu, Xin Tang, Yushi Zhu, Jie Sun, and Shiliang Pu. Sgmnet: Learning rotation-invariant point cloud representations via sorted gram matrix. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10468–10477, 2021.
- [21] Ruibin Gu, Qiuxia Wu, Yuqiong Li, Wenxiong Kang, Wing WY Ng, and Zhiyong Wang. Enhanced local and global learning for rotation-invariant point cloud representation. *IEEE MultiMedia*, 29(4):24–37, 2022.
- [22] Xianzhi Li, Ruihui Li, Guangyong Chen, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. A rotation-invariant framework for deep point cloud analysis. *IEEE transactions on visualization and computer graphics*, 28(12):4503–4514, 2021.
- [23] Ronghan Chen and Yang Cong. The devil is in the pose: Ambiguity-free 3d rotation-invariant learning via pose-aware convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2022.
- [24] Zhiyuan Zhang, Licheng Yang, and Zhiyu Xiang. Risurconv: Rotation invariant surface attention-augmented convolutions for 3d point cloud classification and segmentation. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024.
- [25] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.
- [26] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

- [27] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8895–8904, 2019.
- [28] Christophe Ley and Thomas Verdebout. *Applied directional statistics: modern methods and case studies*. CRC Press, 2018.
- [29] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000.
- [30] Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.
- [31] Geoffrey S Watson. Equatorial distributions on a sphere. *Biometrika*, 52(1/2):193–201, 1965.
- [32] Christopher Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974.
- [33] Plamen Koev and Alan Edelman. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, 75(254):833–846, 2006.
- [34] Tamio Koyama, Hiromasa Nakayama, Kenta Nishiyama, and Nobuki Takayama. Holonomic gradient descent for the fisher–bingham distribution on the d-dimensional sphere. *Computational Statistics*, 29(3-4):661–683, 2014.
- [35] Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514, 2020.
- [36] Jaehoon Jung, Sang-Rim Lee, Haewook Park, Sunho Lee, and Inkyu Lee. Capacity and error probability analysis of diversity reception schemes over generalized- k fading channels using a mixture gamma distribution. *IEEE Transactions on Wireless Communications*, 13(9):4721–4730, 2014.
- [37] Lawrence M Leemis and Jacquelyn T McQueston. Univariate distribution relationships. *The American Statistician*, 62(1):45–53, 2008.

- [38] Haowen Deng, Mai Bui, Nassir Navab, Leonidas J. Guibas, Slobodan Ilic, and Tolga Birdal. Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation. *Int. J. Comput. Vis.*, 130(7):1627–1654, 2022.
- [39] Jalil Taghia, Zhanyu Ma, and Arne Leijon. Bayesian estimation of the von-mises fisher mixture model with variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 36(9):1701–1715, 2014.
- [40] Jalil Taghia and Arne Leijon. Variational inference for watson mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1886–1900, 2015.
- [41] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the european conference on computer vision (ECCV)*, pages 52–68, 2018.
- [42] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [43] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
- [44] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.
- [45] Carlos Esteves. Theoretical aspects of group equivariant neural networks. *CoRR*, abs/2004.05154, 2020.
- [46] Zhiyuan Zhang, Binh-Son Hua, David W Rosen, and Sai-Kit Yeung. Rotation invariant convolutions for 3d point clouds deep learning. In *2019 International conference on 3d vision (3DV)*, pages 204–213. IEEE, 2019.
- [47] Jiaxun Guo, Wentao Fan, Manar Amayri, and Nizar Bouguila. Deep clustering analysis via variational autoencoder with gamma mixture latent embeddings. *Neural Networks*, 183:106979, 2025.

- [48] Jiaxun Guo, Wentao Fan, Manar Amayri, and Nizar Bouguila. Variational inference for the bingham mixture model and applications. *Pattern Recognition (Submitted)*, 2025.
- [49] Jiaxun Guo, Wentao Fan, Manar Amayri, and Nizar Bouguila. Brvsnet: A self-supervised framework for probabilistic rotation estimation and consistent multi-view generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (Submitted)*, 2025.
- [50] Jiaxun Guo, Manar Amayri, Nizar Bouguila, Xin Liu, and Wentao Fan. Enhancing rotation-invariant 3d learning with global pose awareness and attention mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.
- [51] Jiaxun Guo, Wentao Fan, Manar Amayri, and Nizar Bouguila. 4d local modeling toward dynamic global perception for ambiguity-free rotation-invariant point cloud analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (Submitted)*, 2026.
- [52] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [53] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.
- [54] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.
- [55] Thanh Minh Nguyen and QM Jonathan Wu. Bounded asymmetrical student’s-t mixture model. *IEEE transactions on cybernetics*, 44(6):857–869, 2013.
- [56] Nizar Bouguila and Djemel Ziou. High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1716–1731, 2007.
- [57] Wentao Fan, Nizar Bouguila, Ji-Xiang Du, and Xin Liu. Axially symmetric data clustering through dirichlet process mixture models of watson distributions. *IEEE transactions on neural networks and learning systems*, 30(6):1683–1694, 2018.

- [58] Wentao Fan, Lin Yang, and Nizar Bouguila. Unsupervised grouped axial data modeling via hierarchical Bayesian nonparametric models with Watson distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9654–9668, 2022.
- [59] Xiaodong Luo, Xiaohai He, Xiang Chen, Linbo Qing, and Jin Zhang. Dualg-gan, a dual-channel generator based generative adversarial network for text-to-face synthesis. *Neural Networks*, 155:155–167, 2022.
- [60] Yongqi Tian, Xueyuan Gong, Jialin Tang, Binghua Su, Xiaoxiang Liu, and Xinyuan Zhang. Giu-gans: Global information utilization for generative adversarial networks. *Neural Networks*, 152:487–498, 2022.
- [61] Lin Yang, Wentao Fan, and Nizar Bouguila. Deep clustering analysis via dual variational autoencoder with spherical latent embeddings. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6303–6312, 2023.
- [62] Samir Sadok, Simon Leglaive, Laurent Girin, Xavier Alameda-Pineda, and Renaud Séguier. A multimodal dynamical variational autoencoder for audiovisual speech representation learning. *Neural Networks*, 172:106120, 2024.
- [63] Zedong Bi, Haoran Li, and Liang Tian. Top-down generation of low-resolution representations improves visual perception and imagination. *Neural Networks*, 171:440–456, 2024.
- [64] Thanh Minh Nguyen and QM Jonathan Wu. Robust student’s-t mixture model with spatial constraints and its application in medical image segmentation. *IEEE Transactions on Medical Imaging*, 31(1):103–116, 2011.
- [65] Yong Yang, Shoudong Han, Tianjiang Wang, Wenbing Tao, and Xue-Cheng Tai. Multilayer graph cuts based unsupervised color–texture image segmentation using multivariate mixed student’s t-distribution and regional credibility merging. *Pattern recognition*, 46(4):1101–1124, 2013.
- [66] Kostas N Oikonomou. Prediction with the dynamic bayesian gamma mixture model. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(4):529–542, 1997.

- [67] Nizar Bouguila, Khaled Almakadmeh, and Sabri Boutemedjet. A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection. *Expert Systems with Applications*, 39(7):6641–6656, 2012.
- [68] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised adversarial invariance. *Advances in Neural Information Processing Systems*, 31, 2018.
- [69] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4066–4075, 2019.
- [70] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.
- [71] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Ijcai*, pages 1753–1759, 2017.
- [72] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- [73] Chunyan Li, Junfeng Yao, Wei Wei, Zhangming Niu, Xiangxiang Zeng, Jin Li, and Jianmin Wang. Geometry-based molecular generation with deep constrained variational autoencoder. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [74] Pengfei Jiao, Xuan Guo, Xin Jing, Dongxiao He, Huaming Wu, Shirui Pan, Maoguo Gong, and Wenjun Wang. Temporal network embedding for link prediction via vae joint attention mechanism. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7400–7413, 2021.
- [75] Wen Yu, Baiying Lei, Michael K Ng, Albert C Cheung, Yanyan Shen, and Shuqiang Wang. Tensorizing gan with high-order pooling for alzheimer’s disease assessment. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4945–4959, 2021.
- [76] Wenxiu Diao, Feng Zhang, Jiande Sun, Yinghui Xing, Kai Zhang, and Lorenzo Bruzzone. Zergan: Zero-reference gan for fusion of multispectral and panchromatic images. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

- [77] Akinlolu Oluwabusayo Ojo and Nizar Bouguila. A topic modeling and image classification framework: The generalized dirichlet variational autoencoder. *Pattern Recognition*, 146:110037, 2024.
- [78] Fatemeh Daneshfar, Sayvan Soleymanbaigi, Ali Nafisi, and Pedram Yamini. Elastic deep autoencoder for text embedding clustering by an improved graph regularization. *Expert Systems with Applications*, 238:121780, 2024.
- [79] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.
- [80] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4610–4617, 2019.
- [81] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.
- [82] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.
- [83] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [84] Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR, 2017.
- [85] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [86] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

- [87] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pages 127–140, 2015.
- [88] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.
- [89] Peter W Eklund and Philippe Martin. Www indexation and document navigation using conceptual structures. In *2nd IEEE Conference on Intelligent Information Processing Systems (ICIPS’98)*, IEEE Press, pages 217–221, 1998.
- [90] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1965–1972, 2017.
- [91] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [92] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.
- [93] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.
- [94] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.
- [95] Qi He, Kuiyu Chang, Ee-Peng Lim, and Arindam Banerjee. Keep it simple with time: A reexamination of probabilistic topic detection models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1795–1808, 2010.

- [96] Kenji Hara, Ko Nishino, and Katsushi Ikeuchi. Mixture of spherical distributions for single-view relighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):25–35, 2008.
- [97] Wentao Fan, Lin Yang, and Nizar Bouguila. Grouped spherical data modeling through hierarchical nonparametric bayesian models and its application to fmri data analysis. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2022.
- [98] Suvrit Sra and Dmitrii Karp. The multivariate watson distribution: Maximum-likelihood estimation and other aspects. *Journal of Multivariate Analysis*, 114:256–269, 2013.
- [99] Md Abul Hasnat, Olivier Alata, and Alain Trémeau. Unsupervised clustering of depth images using watson mixture model. In *2014 22nd International Conference on Pattern Recognition*, pages 214–219. IEEE, 2014.
- [100] Lin Yang, Yuhang Liu, and Wentao Fan. Axial data modeling with collapsed nonparametric watson mixture models and its application to depth image analysis. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 17–28. Springer, 2020.
- [101] Dang Hai Tran Vu and Reinhold Haeb-Umbach. Blind speech separation employing directional statistics in an expectation maximization framework. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 241–244. IEEE, 2010.
- [102] Christopher J Fallaize and Theodore Kypraios. Exact bayesian inference for the bingham distribution. *Statistics and Computing*, 26:349–360, 2016.
- [103] Marco Bee, Roberto Benedetti, and Giuseppe Espa. Approximate maximum likelihood estimation of the bingham distribution. *Computational Statistics & Data Analysis*, 108:84–96, 2017.
- [104] Atsushi Yamaji and Katsushi Sato. Clustering of fracture orientations using a mixed bingham distribution and its application to paleostress analysis from dike or vein orientations. *Journal of Structural Geology*, 33(7):1148–1157, 2011.
- [105] Atsushi Yamaji. Genetic algorithm for fitting a mixed bingham distribution to 3d orientations: a tool for the statistical and paleostress analyses of fracture orientations. *Island Arc*, 25(1):72–83, 2016.

- [106] Kevin Breuer, Markus Stommel, and Wolfgang Korte. Analysis and evaluation of fiber orientation reconstruction methods. *Journal of Composites Science*, 3(3):67, 2019.
- [107] Tong Gao, Meredith D Betterton, An-Sheng Jhang, and Michael J Shelley. Analytical structure, dynamics, and coarse graining of a kinetic model of an active fluid. *Physical Review Fluids*, 2(9):093302, 2017.
- [108] Onur C. Hamsici and Aleix M. Martinez. Rotation invariant kernels and their application to shape analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1985–1999, 2009.
- [109] Jesper Møller, Anthony N Pettitt, Robert Reeves, and Kasper K Berthelsen. An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.
- [110] J Glover. libbingham bingham statistics library, 2013.
- [111] Alfred Kume and Andrew TA Wood. Saddlepoint approximations for the bingham and fisher–bingham normalising constants. *Biometrika*, 92(2):465–476, 2005.
- [112] Igor Gilitschenski, Gerhard Kurz, Simon J Julier, and Uwe D Hanebeck. Unscented orientation estimation based on the bingham distribution. *IEEE Transactions on Automatic Control*, 61(1):172–177, 2015.
- [113] John T Kent, Asaad M Ganeiber, and Kanti V Mardia. A new method to simulate the bingham and related distributions in directional data analysis with applications. *arXiv preprint arXiv:1310.8110*, 2013.
- [114] GNS Science. Geonet aotearoa new zealand earthquake moment tensor solutions, 2006.
- [115] Alfred Kume and Tomonari Sei. On the exact maximum likelihood inference of fisher–bingham distributions using an adjusted holonomic gradient method. *Statistics and Computing*, 28:835–847, 2018.
- [116] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

- [117] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.
- [118] Hagit Shatkay, Stephen Edwards, W John Wilbur, and Mark Boguski. Genes, themes and microarrays. In *Proc. Int. Conf. Intell. Syst. Mol. Biol*, volume 8, pages 317–327, 2000.
- [119] Joseph L DeRisi, Vishwanath R Iyer, and Patrick O Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.
- [120] Shelley Chu, Joe DeRisi, Michael Eisen, Jon Mulholland, David Botstein, Patrick O Brown, and Ira Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699–705, 1998.
- [121] Vishwanath R Iyer, Michael B Eisen, Douglas T Ross, Greg Schuler, Troy Moore, Jeffrey CF Lee, Jeffrey M Trent, Louis M Staudt, James Hudson Jr, Mark S Boguski, et al. The transcriptional program in the response of human fibroblasts to serum. *science*, 283(5398):83–87, 1999.
- [122] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.
- [123] Liming Wang and Xiaodong Wang. Hierarchical dirichlet process model for gene expression clustering. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013:1–14, 2013.
- [124] Wentao Fan, Lin Yang, and Nizar Bouguila. Unsupervised grouped axial data modeling via hierarchical bayesian nonparametric models with watson distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9654–9668, 2021.
- [125] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

- [126] Fabien Delattre, David Dirnfeld, Phat Nguyen, Stephen Scarano, Michael J. Jones, Pedro Miraldo, and Erik G. Learned-Miller. Robust frame-to-frame camera rotation estimation in crowded scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9718–9728. IEEE, 2023.
- [127] Yingda Yin, Yang Wang, He Wang, and Baoquan Chen. A laplace-inspired distribution on $SO(3)$ for probabilistic rotation estimation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [128] Yulin Liu, Haoran Liu, Yingda Yin, Yang Wang, Baoquan Chen, and He Wang. Delving into discrete normalizing flows on $SO(3)$ manifold for probabilistic rotation modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21264–21273. IEEE, 2023.
- [129] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [130] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 10072–10083. IEEE, 2024.
- [131] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9698–9707. IEEE, 2024.
- [132] Sergey Prokudin, Peter V. Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, volume 11213 of *Lecture Notes in Computer Science*, pages 542–559. Springer, 2018.

- [133] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [134] Yingda Yin, Yingcheng Cai, He Wang, and Baoquan Chen. Fishermatch: Semi-supervised rotation regression via entropy-based filtering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11154–11163. IEEE, 2022.
- [135] Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a bingham loss. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [136] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9264–9275. IEEE, 2023.
- [137] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [138] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
- [139] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer*

- Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.
- [140] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5242–5251. Computer Vision Foundation / IEEE, 2020.
- [141] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1077–1086. Computer Vision Foundation / IEEE, 2019.
- [142] Evan G Hemingway and Oliver M O’Reilly. Perspectives on euler angle singularities, gimbal lock, and the orthogonality of applied forces and applied moments. *Multibody system dynamics*, 44:31–56, 2018.
- [143] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2837–2845. Computer Vision Foundation / IEEE, 2021.
- [144] Yingzhi Tang, Yue Qian, Qijian Zhang, Yiming Zeng, Junhui Hou, and Xuefei Zhe. Warpinggan: Warping multiple uniform priors for adversarial 3d point cloud generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6387–6395. IEEE, 2022.
- [145] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.
- [146] Tolga Birdal, Umut Simsekli, Mustafa Onur Eken, and Slobodan Ilic. Bayesian pose graph optimization via bingham distributions and tempered geodesic MCMC. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in*

- Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 306–317, 2018.
- [147] Adrian Haarbach, Tolga Birdal, and Slobodan Ilic. Survey of higher order rigid body motion interpolation methods for keyframe animation and continuous-time trajectory estimation. In *2018 International Conference on 3D Vision, 3DV 2018, Verona, Italy, September 5-8, 2018*, pages 381–389. IEEE Computer Society, 2018.
 - [148] Timothy D. Barfoot and Paul Timothy Furgale. Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Trans. Robotics*, 30(3):679–693, 2014.
 - [149] Luca Falorsi, Pim de Haan, Tim R. Davidson, and Patrick Forré. Reparameterizing distributions on lie groups. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 3244–3253. PMLR, 2019.
 - [150] F. Sebastian Grassia. Practical parameterization of rotations using the exponential map. *J. Graphics, GPU, & Game Tools*, 3(3):29–48, 1998.
 - [151] Benjamin Busam, Tolga Birdal, and Nassir Navab. Camera pose filtering with local regression geodesics on the riemannian manifold of dual quaternions. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 2436–2445. IEEE Computer Society, 2017.
 - [152] Jared Glover and Leslie Pack Kaelbling. Tracking the spin on a ping pong ball with the quaternion bingham filter. In *2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, May 31 - June 7, 2014*, pages 4133–4140. IEEE, 2014.
 - [153] Mai Bui, Tolga Birdal, Haowen Deng, Shadi Albarqouni, Leonidas J. Guibas, Slobodan Ilic, and Nassir Navab. 6d camera relocalization in ambiguous scenes via continuous multimodal inference. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, volume 12363 of *Lecture Notes in Computer Science*, pages 139–157. Springer, 2020.

- [154] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org, 2014.
- [155] Brian D. Ripley. *Stochastic simulation*. Wiley series in probability and mathematical statistics : applied probability and statistics. Wiley, 1987.
- [156] K.V. Mardia and P.E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009.
- [157] Christian Rupprecht, Iro Laina, Robert S. DiPietro, and Maximilian Baust. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3611–3620. IEEE Computer Society, 2017.
- [158] Fabian Manhardt, Diego Martín Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6840–6849. IEEE, 2019.
- [159] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.
- [160] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6555–6564. IEEE Computer Society, 2017.
- [161] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7163–7172. Computer Vision Foundation / IEEE, 2019.

- [162] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In Danica Kragic, Antonio Bicchi, and Alessandro De Luca, editors, *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 4762–4769. IEEE, 2016.
- [163] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.
- [164] Feiran Li, Kent Fujiwara, Fumio Okura, and Yasuyuki Matsushita. A closer look at rotation-invariant deep point cloud analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16218–16227, 2021.
- [165] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021.
- [166] Adrien Poulenard and Leonidas J Guibas. A functional approach to rotation equivariant nonlinearities for tensor field networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13174–13183, 2021.
- [167] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d rotation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.
- [168] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.
- [169] Chao Chen, Guanbin Li, Ruijia Xu, Tianshui Chen, Meng Wang, and Liang Lin. Clusternet: Deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4994–5002, 2019.
- [170] Yiyang Chen, Lunhao Duan, Shanshan Zhao, Changxing Ding, and Dacheng Tao. Local-consistent transformation learning for rotation-invariant point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5418–5427, 2024.

- [171] Hao Yu, Ji Hou, Zheng Qin, Mahdi Saleh, Ivan Shugurov, Kai Wang, Benjamin Busam, and Slobodan Ilic. Riga: Rotation-invariant and globally-aware descriptors for point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3796–3812, 2024.
- [172] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant transformer for point cloud matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5384–5393, 2023.
- [173] Chengzhi Wu, Junwei Zheng, Julius Pfommer, and Jürgen Beyerer. Attention-based point cloud edge sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2023.
- [174] Chengzhi Wu, Yuxin Wan, Hao Fu, Julius Pfommer, Zeyun Zhong, Junwei Zheng, Jiaming Zhang, and Jürgen Beyerer. Samble: Shape-specific point cloud sampling for an optimal trade-off between local detail and global uniformity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1342–1352, 2025.
- [175] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European conference on computer vision (ECCV)*, pages 602–618, 2018.
- [176] Haoran Zhou, Yidan Feng, Mingsheng Fang, Mingqiang Wei, Jing Qin, and Tong Lu. Adaptive graph convolution for point cloud analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4965–4974, 2021.
- [177] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.
- [178] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Re-visiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019.

- [179] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [180] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, 2018.
- [181] Seohyun Kim, Jaeyoo Park, and Bohyung Han. Rotation-invariant local-to-global representation learning for 3d point cloud. *Advances in Neural Information Processing Systems*, 33:8174–8185, 2020.
- [182] Shitong Luo, Jiahan Li, Jiaqi Guan, Yufeng Su, Chaoran Cheng, Jian Peng, and Jianzhu Ma. Equivariant point cloud analysis via learning orientations for message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18932–18941, 2022.
- [183] Dingxin Zhang, Jianhui Yu, Chaoyi Zhang, and Weidong Cai. Parot: Patch-wise rotation-invariant network via feature disentanglement and pose restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3418–3426, 2023.
- [184] Pavlo Melnyk, Andreas Robinson, Michael Felsberg, and Mårten Wadenbäck. Tetrasphere: A neural descriptor for o(3)-invariant point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5620–5630, 2024.
- [185] Chen Zhao, Jiaqi Yang, Xin Xiong, Angfan Zhu, Zhiguo Cao, and Xin Li. Rotation invariant point cloud analysis: Where local geometry meets global topology. *Pattern Recognition*, 127:108626, 2022.
- [186] Yujing Lou, Zelin Ye, Yang You, Nianjuan Jiang, Jiangbo Lu, Weiming Wang, Lizhuang Ma, and Cewu Lu. Crin: rotation-invariant point cloud analysis and rotation estimation via centrifugal reference frame. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1817–1825, 2023.
- [187] Jianhui Yu, Chaoyi Zhang, and Weidong Cai. Rethinking rotation invariance with point cloud registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3313–3321, 2023.

- [188] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022.
- [189] Hanjiang Hu, Zuxin Liu, Sharad Chitlangia, Akhil Agnihotri, and Ding Zhao. Investigating the impact of multi-lidar placement on object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2550–2559, 2022.
- [190] Ruixuan Yu, Xin Wei, Federico Tombari, and Jian Sun. Deep positional and relational feature learning for rotation-invariant point cloud analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 217–233. Springer, 2020.
- [191] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International conference on machine learning*, pages 2747–2755. PMLR, 2018.
- [192] Pavlo Melnyk, Michael Felsberg, and Mårten Wadenbäck. Steerable 3d spherical neurons. In *International Conference on Machine Learning*, pages 15330–15339. PMLR, 2022.
- [193] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 915–924, 2021.
- [194] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.
- [195] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.
- [196] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pages 9377–9388. PMLR, 2021.

- [197] Victor Garcia Satorras, Emiel Hoogetboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.
- [198] Zhuo Su, Max Welling, Matti Pietikäinen, and Li Liu. Svnet: Where so (3) equivariance meets binarization on point cloud representation. In *2022 International Conference on 3D Vision (3DV)*, pages 547–556. IEEE, 2022.
- [199] Gregory S Chirikjian and Alexander B Kyatkin. *Engineering applications of noncommutative harmonic analysis: with emphasis on rotation and motion groups*. CRC press, 2000.
- [200] Pavlo Melnyk, Michael Felsberg, and Mårten Wadenbäck. Embed me if you can: A geometric perceptron. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1276–1284, 2021.
- [201] Christian Perwass, Vladimir Banarier, and Gerald Sommer. Spherical decision surfaces using conformal modelling. In *Joint Pattern Recognition Symposium*, pages 9–16. Springer, 2003.