Enhanced Video Tracking Based on Fusion of Visible and Infrared Images

Mohamed Elsayed Awad Mohamed

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Electrical and Computer Engineering) at
Concordia University
Montréal, Québec, Canada

July 2025

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Mohamed Elsayed Awad Mohamed**

Entitled: **Enhanced Video Tracking Based on Fusion of Visible and Infrared Images**

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

Doctor of Philosophy (Electrical and Computer 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:

______________________________ Chair
Dr. Caroline Hachem-Vermette

______________________________ External Examiner
Dr. Q.M. Jonathan Wu

______________________________ Arms-length Examiner
Dr. Chun-Yi Su

______________________________ Examiner
Dr. Chunyan Wang

______________________________ Examiner
Dr. Wei-Ping Zhu

______________________________ Thesis Supervisor
Dr. M. Omair Ahmad

______________________________ Thesis Supervisor
Dr. M.N.S. Swamy

Approved by ______________________________
Dr. Jun Cai, Graduate Program Director

September 8, 2025 ______________________________
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Enhanced Video Tracking Based on Fusion of Visible and
Infrared Images

Mohamed Elsayed Awad Mohamed, Ph.D.

Concordia University, 2025

Video tracking is the process of automated identification, localization, and continuous monitoring of objects of interest throughout consecutive video frames. Video tracking is core of many cutting-edge vision applications such as surveillance systems, autonomous vehicles, augmented reality, robotics, and human-computer interaction. However, reliance solely on visible (RGB) imagery introduces significant challenges, including poor visibility, low illumination, occlusion, and appearance variations. To overcome these challenges, fusion of RGB with thermal infrared (TIR) data has been explored to leverage complementary modality information for improved tracking performance under challenging conditions.

Many existing RGB-Thermal (RGB-T) trackers use deep learning (DL) methods for strong object feature representation. However, despite their superior performance, these tracking methods mostly rely on dual-branch architectures, complex fusion modules, or external teacher-student frameworks, leading to increased model size and significant training overhead. To address this problem, this thesis proposes unified RGB-T tracking schemes that enhance conventional RGB trackers without altering their network architectures or significantly increasing the computational complexity.

In the first part of the thesis, a novel pixel-level fusion network, symmetric bidirectional dynamic fusion (SBiDF), is introduced. SBiDF enhances RGB inputs by dynamically integrating TIR data at the pixel level prior to tracking, utilizing modality-specific autoencoders, dynamic convolutional filtering (DCF) blocks, and an output fusion module. The DCF blocks perform adaptive, bidirectional, content-aware enhancement, enabling balanced cross-modal refinement. Importantly, SBiDF generalizes effectively beyond TIR to additional modalities such as depth and event data, providing superior tracking accuracy and broad applicability without modifying the tracker architecture.

The second part of the thesis presents a novel learning-based framework, multi-level self-distillation (MSD), adapting a single-stream RGB tracker to the RGB-T setting through advanced training strategies rather than architectural changes. MSD integrates RGB and TIR data via a shared backbone guided by self-supervised contrastive and modality-gap alignment losses alongside supervised focal and modality-specific losses.

Extensive evaluations are performed to demonstrate that SBiDF and MSD provide performance superior to that of state-of-the-art tracking methods in terms of robust accuracy, simplified implementation, and enhanced computational efficiency, making them highly practical for real-world applications.

Acknowledgments

First and foremost, I am honored to express my deepest gratitude to my supervisors, Professor M. Omair Ahmad and Professor M.N.S. Swamy, for their invaluable supervision, guidance, generous advice, constructive criticism, and continuous encouragement throughout this research. Their doors were always open whenever I encountered difficulties or had questions about my research or writing. They consistently allowed this thesis to be my own work, yet gently steered me in the right direction whenever they thought I needed it.

My sincere appreciation also goes to Dr. Ahmed Elliethy for his continuous and invaluable guidance, encouragement, and insightful discussions, particularly regarding methodological aspects of my research. Without his passionate participation, this work would not have been successfully completed.

Last but certainly not least, I convey my warmest thanks to my parents for their sacrifices, unconditional love, and limitless support in helping me achieve my goals. I am especially grateful to my loving and supportive wife, Eman, who provided endless inspiration and stood by me through the highs and lows of my Ph.D. journey. I also extend heartfelt thanks to my children, Zain and Yahya, who patiently endured my busy schedule that often prevented us from spending many weekends together.

This thesis would not have been possible without the support and encouragement of these wonderful people. Thank you all.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

CLE	Center location error
CM	Contrastive module
CNN	Convolutional neural network
DAPFNet	Dual adversarial pixel fusion network
DCF	Dynamic convolutional filtering
DL	Deep learning
FPS	Frames per second
GT	Ground-truth
IoU	Intersection over union
KD	Knowledge distillation
LN	Layer normalization
LoRA	Low-rank adaptation
MACs	Multiply–accumulate operations
MLP	Multi-layer perceptron
MSD	Multi-level self-distillation
PL	Prompt learning
PR	Precision rate
Pr	Precision
Re	Recall

ReLU	Rectified linear unit
RGB	Visible red-green-blue
RGB-D	Visible-depth
RGB-E	Visible-event
RGB-T	Visible-thermal
SBiDF	Symmetric bidirectional dynamic fusion
SD	Self-distillation
SR	Success rate
TBSI	Template-bridged search interaction
TIR	Thermal infrared
ViT	Vision transformer
X-modality	Auxiliary modality

Symbols

α	Focusing parameter
β	Focusing parameter
$\hat{\mathbf{F}}^{(l)}$	Intermediate normalized fused features from layer l
$\hat{\mathbf{S}}$	Predicted response map
$\hat{F}$	Enhanced feature representation
$\hat{I}^m$	Enhanced m-modality image
λ	Controlling hyperparameter
$\mathbb{B}$	Batch size
$\mathbb{D}$	Projection dimension
$\mathcal{D}_1$	Detail-1 decoder
$\mathcal{D}_2$	Detail-2 decoder
$\mathcal{D}_{base}$	Base decoder
$\mathcal{E}_l$	Encoder layer l
$\mathcal{L}_{\text{con}}$	Contrastive loss

$\mathcal{L}_{\text{foc}}$	Focal loss
$\mathcal{L}_{\text{gap}}$	Modality-gap loss
$\mathcal{L}_{\text{total}}$	Total loss
$\mathcal{L}_{\text{track}}$	Tracking loss
$\mathcal{L}_{cls}$	Classification loss
$\mathcal{L}_{iou}$	IoU regression loss
$\mathcal{L}_{l_1}$	l_1 bounding-box regression loss
μ	Mean
Ω	Convolution window
ϕ	Dynamic kernel generation network
ψ	Upsampling operation
σ	Standard deviation
τ	Temperature hyper-parameter
$\mathbf{C}$	Cosine similarity matrix
$\mathbf{F}^{(l)}$	Intermediate fused features from layer l
$\mathbf{S}^{\text{GT}}$	Ground-truth heatmap
$\mathbf{x}$	Search features
$\mathbf{x}_p$	Projected search features
$\mathbf{z}$	Template features
$\mathbf{z}_p$	Projected template features
Proj	Projection head
B	Reconstructed base image
b	Convolutional bias
B^{RGB}	Base RGB image
B^{TIR}	Base TIR image
d^{TIR}	TIR feature difference
D_1	Reconstructed medium-scale details

D_2	Reconstructed high-scale details
F_l	Features extracted at layer l
f_l	Feature maps at layer l
f_{base}	Base image features
G	Guide modality features
H^{TIR}	Higher-level TIR image
I^m	Input m-modality image
I^{Fused}	Fused RGB image
I^{RGB}	Input RGB image
I^{TIR}	input TIR image
I^{fused}	Output fused image
K	Convolutional kernels
N^+	Number of positive samples
N^-	Number of negative samples
S	Source modality features
W	Convolutional weight

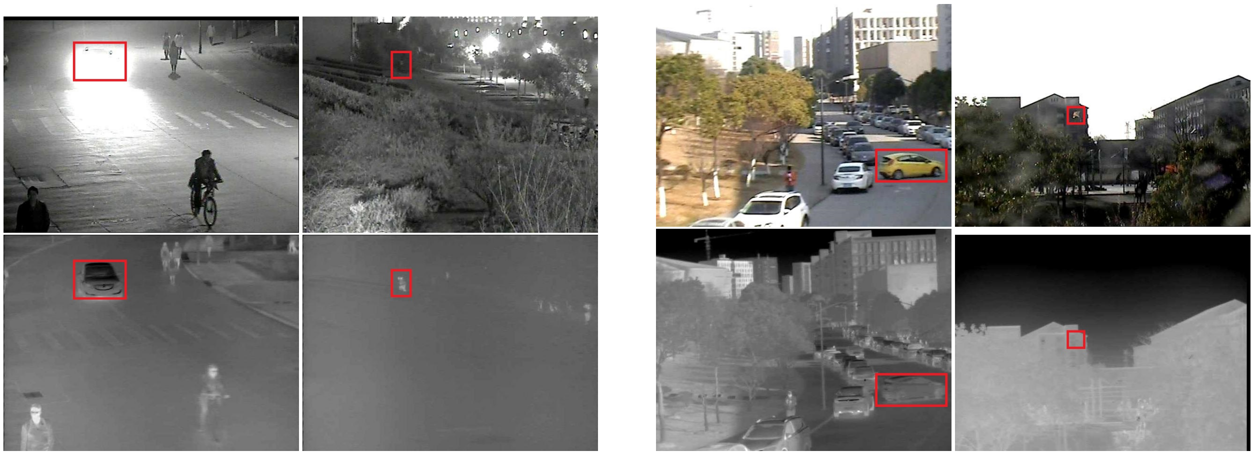
Chapter 1

Introduction

1.1 Background

Video tracking is the process of automatically detecting, locating, and following objects of interest over time across consecutive frames of a video sequence. Nowadays, video tracking is becoming the core process of many vision applications such as surveillance systems, autonomous vehicles, augmented reality, robotic systems, and human-computer interaction. Developing such applications faces many challenges in video tracking such as object variety, pose variation, occlusion, and motion blur. Several tracking approaches were proposed based on visible (RGB) band video sequences trying to overcome the aforementioned challenges. Visible video sequences capture the rich visual information about objects in a scene such as colors, textures, and fine-grained details. This information is suitable for a wide range of tracking scenarios which makes visual object tracking readily available in many surveillance and monitoring systems. However, video sequences captured by visible band sensors may lack some scene details under poor visibility conditions such as haze, fog, mist, and intense or poor illumination. Under such conditions, the effectiveness of video tracking approaches based solely on visible imagery may be questionable.

On the contrary, thermal infrared (TIR) band sensors can capture video sequences with better scene details under low-light or nighttime conditions, where visible light may be insufficient for



(a) (b)

Figure 1.1: Examples of visible (first row) and infrared (second row) image pairs captured for the same scenes (column-wise) and selected from the RGBT234 dataset [1].

effective tracking, as shown in the example images in Fig. 1.1 (a). Therefore, many tracking approaches were recently proposed based on infrared band video sequences. Infrared video sequences capture the heat emitted by objects which allows infrared object tracking for improved detection and tracking of living beings and heat-emitting objects. Furthermore, infrared sensors are less susceptible to visual noise and clutter caused by changing lighting conditions or shadows which provides more stable tracking in challenging environments. However, the infrared video sequences lack some scene details in cases of thermal crossover and also lack the scene's color and texture details which are well provided in the visible video sequences, as shown in Fig. 1.1 (b). In such cases, the video tracking approaches based only on infrared imagery may be unreliable.

To address these limitations, RGB-Thermal (RGB-T) tracking has emerged as a promising solution by integrating complementary RGB and TIR modalities. While RGB modality provides rich color and texture information useful under normal lighting, TIR modality offers valuable information about objects in low lighting, complete darkness, or under adverse weather conditions. By effectively combining the strengths of both modalities, RGB-T tracking significantly enhances tracking robustness, ensuring reliable performance despite environmental variances such as illumination changes, shadows, thermal crossover, and partial occlusions [2–4]. The RGB-T fusion tracking approaches differ mainly based on the adopted theory in their methodologies. Among these approaches, the ones that adopt deep learning (DL) techniques usually exhibit superior performance compared to other ones thanks to the powerful feature representation capabilities of deep neural networks and their ability to learn adaptive fusion rules that can predict how much a modality is reliable to contribute to the tracking process.

Recent DL-based approaches dominate the field, typically extending pre-trained RGB trackers with modality-specific branches and fusion modules. The RGB-T fusion process mostly happens at either the feature level [5–9] or decision level [10–13]. In feature-level fusion, features from RGB and TIR images are extracted and merged into a joint representation to guide the tracking process. For example, [14] employs a generative network trained with tracking and reconstruction losses to produce refined fused features. In contrast, decision-level fusion applies a tracking-before-fusion strategy, where tracking is performed separately on RGB and TIR images, and the results are later combined using ensemble methods or confidence-based weighting. For instance, a dual-stream Siamese tracker in [10] generates separate response maps per modality and aggregates them to localize the target. Although being effective, these paradigms often introduce substantial computational overhead, increase model complexity, and require nontrivial architectural changes, limiting their practicality for resource-constrained applications.

To mitigate the computational overhead and complexity of RGB-T tracking systems, recent studies have explored efficiency-oriented strategies such as prompt learning [15–19] and knowledge distillation [20–24]. These methods aim to preserve multi-modal benefits while reducing the computational and architectural burden. Prompt-based approaches insert lightweight adapters into frozen backbones to guide modality interaction, aiming to preserve efficiency while still enabling multi-modal reasoning. For example, a lightweight prompt-branch is used to extract TIR features guided by RGB features extracted using a pre-trained backbone in [16]. The interaction between the two modalities' features is used then in obtaining the tracking results. Despite reduced complexity of such approaches, their performance is often limited by the capacity of fixed prompts. On the other hand, distillation-based methods transfer knowledge from a fusion-heavy teacher to a compact student, but they require careful tuning of distillation objectives. For instance, Lu et al. [22]

employs coupled knowledge distillation to bridge the modality gap by aligning feature distributions between modalities, while using modality-specific teachers to guide a dual-branch student model. While more efficient than traditional multi-branch architectures, these approaches often require two-stage training pipeline and careful design of distillation losses. The need for a separate teacher network also increases memory usage and training time, which can limit scalability and deployment.

1.2 Motivation and Objective

Despite extensive research in feature-level and decision-level fusion, pixel-level fusion remains relatively under-explored. Pixel-level fusion integrates complementary modalities at the earliest stage of the tracking pipeline, potentially yielding richer representations and simplifying integration with existing tracking frameworks. Such a fusion-before-tracking approach eliminates the need for architectural modifications such as additional modality-specific branches or complex fusion modules. For instance, RGB and TIR images are concatenated along the channel dimension in [25], demonstrating enhanced accuracy through end-to-end fusion and training. However, existing pixel-level methods often rely on static fusion strategies, lacking flexibility to adapt dynamically to content variations across diverse tracking scenarios. Therefore, developing a dynamic pixel-level fusion method that effectively leverages complementary modality information in a content-aware manner, without increasing architectural complexity, remains crucial for enhancing deployment feasibility and performance robustness in practical scenarios.

Apart from pixel-level fusion, recent efficiency-oriented techniques such as prompt learning and knowledge distillation attempt to mitigate model complexity, but often encounter practical limitations. Prompt-based methods typically restrict model adaptability due to the limited flexibility of fixed prompts, whereas knowledge distillation approaches involve intricate two-stage training processes and dependencies on external teacher networks, complicating deployment. Recently, self-distillation has emerged as a promising streamlined alternative to address these limitations.

Unlike traditional knowledge distillation, self-distillation leverages internal representations from the network itself as both teacher and student to guide learning [26]. Nevertheless, self-distillation in the context of RGB-T tracking remains underexplored. For example, Hou et al. [23] propose a self-distillation-based method enhancing robustness by guiding masked modality features using unmasked counterparts. Combined with prompt-based fine-tuning, this approach adapts a frozen RGB backbone to the RGB-T setting. While reducing architectural complexity compared to traditional distillation, these methods still rely on modality-specific components or additional parameters. Thus, achieving a balance between performance and efficiency to facilitate real-world deployment remains an ongoing challenge, motivating exploration of novel learning-based fusion strategies.

In light of the aforementioned limitations, the primary objective of this thesis is to propose unified RGB-T tracking frameworks capable of enhancing conventional RGB trackers without architectural modifications or significant computational overhead. Specifically, in our study two directions are taken to achieve this goal.

In the first part of the thesis, we develop a novel pixel-level fusion framework, named symmetric bidirectional dynamic fusion (SBiDF), designed to enrich the RGB modality *before* employing it as input for tracking. Unlike generic pixel-level fusion approaches designed for image enhancement or reconstruction, SBiDF is explicitly tailored for object tracking, adaptively enhancing RGB content by dynamically incorporating complementary cues from the TIR modality. It generates fused RGB-T images that are optimized for tracking, with features emphasized according to their relevance to the target object. During training, the fused RGB-T image serves as the sole input to the tracking network, and the tracker loss provides the only supervision. Thus, SBiDF avoids relying on modality-specific priors or auxiliary objectives; instead, it learns a fusion strategy specifically oriented toward tracking performance. The design of SBiDF are driven by a key idea: allowing both RGB and the auxiliary modality to influence each other dynamically rather than relying solely on a fixed fusion rule. Consequently, the proposed SBiDF generalizes effectively across multiple modalities (thermal, depth, event) through dynamic pixel-level fusion, enabling enhanced tracking accuracy and broader applicability.

In the second part of the thesis, we introduce a novel multi-level self-distillation framework (MSD) to transform a standard RGB tracker into a high-performance RGB-T solution without architectural modifications, additional parameters, or external teachers. The core insight behind MSD is that a shared backbone can learn robust cross-modal representations through carefully designed self-supervision and multi-level training signals. MSD integrates internal self-distillation losses and supervised learning signals, including contrastive feature distillation, modality-gap alignment, intermediate focal loss, modality-specific losses, and a primary tracking loss, enhancing cross-modal feature representation and achieving robust RGB-T tracking accuracy. By unifying these learning objectives, MSD enables a single backbone to dynamically adapt to RGB, TIR, or fused inputs without modality-specific branches, maintaining the computational efficiency, parameter count, and inference speed of the original RGB tracker.

By conducting investigations in these two directions, in this thesis we aim to bridge gaps between the performance and efficiency in multi-modal tracking.

1.3 Organization of the Thesis

The thesis is organized as follows. In Chapter 2, a comprehensive review of state-of-the-art RGB-T and multi-modal tracking approaches is introduced. Various DL-based fusion strategies including pixel-level, feature-level, and decision-level fusion are discussed. Additionally, recent trends in efficiency-oriented approaches such as prompt learning and knowledge distillation for multi-modal tracking are presented, further positioning this thesis within the existing literature. Chapter 3 provides a detailed description of the proposed pixel-level fusion framework, termed symmetric bidirectional dynamic fusion (SBiDF). The chapter thoroughly presents the methodology, highlighting the architecture components, including modality-specific autoencoders, dynamic convolutional filtering (DCF) blocks, and the output fusion module. Furthermore, comprehensive experimental validation of SBiDF is provided, demonstrating its generalization ability across various modalities (thermal, depth, and event data) and its ability to improve tracking performance compared to state-of-the-art unified multi-modal trackers as well as modality-specific trackers. In Chapter 4, details and formulation of the proposed multi-level self-distillation framework (MSD) are described. The chapter details the proposed objectives leveraged to adapt a standard RGB tracker for robust RGB-T tracking, including contrastive loss, modality-gap alignment loss, intermediate focal loss, modality-specific losses, and the final tracking loss. Moreover, extensive experimental results are provided, demonstrating that MSD achieves state-of-the-art tracking accuracy while maintaining the parameter efficiency and inference speed of the original RGB tracker. Finally, Chapter 5 concludes the thesis by summarizing the proposed SBiDF and MSD frameworks along with their key contributions. Additionally, this chapter provides a concise discussion of potential directions for future research and opportunities for further investigation building upon the work presented in this thesis.

Chapter 2

Literature Review

2.1 Introduction

To position our work within the broader landscape of multi-modal single object tracking, we review three key areas most relevant to the proposed approaches: general multi-modal tracking frameworks, pixel-level fusion methods, and efficiency-driven approaches. The first area, multi-modal tracking, encompasses researches integrating complementary sensing modalities mostly RGB-Thermal (RGB-T), but also RGB-Depth (RGB-D) and RGB-Event (RGB-E), to enhance tracking robustness under diverse and challenging conditions. Such methods typically rely on modality-specific architectures or fusion mechanisms at the feature or decision levels. The second area, pixel-level fusion, merges multi-modal data directly at the image level. The third area, efficiency-driven approaches, explores how learning-based schemes such as knowledge distillation and prompt learning offer lightweight and efficient alternatives to re-training or expanding existing models.

The proposed SBiDF approach lies at the intersection of the multi-modal tracking and pixel-level fusion domains, introducing an adaptive pixel-level fusion explicitly optimized for tracking performance, thereby achieving robust multi-modal tracking without architectural modifications or additional computational overhead. Meanwhile, the proposed MSD approach lies at the intersection of the multi-modal tracking and efficiency-driven directions, eliminating explicit fusion modules while achieving efficiency through internal supervision and self-distillation. In the following sections, we review recent representative works in each area and highlight their connections and distinctions relative to our proposed frameworks.

2.2 Related Work on Multi-modal Tracking Schemes

Multi-modal object tracking has gained considerable attention due to its robustness in challenging environments where single-modality trackers (particularly RGB-only trackers) typically struggle. By leveraging complementary information from additional modalities, multi-modal trackers achieve enhanced performance under adverse conditions such as low visibility, occlusions, and background clutter.

2.2.1 RGB-T Tracking

Early research predominantly focused on combining RGB images with thermal infrared (TIR) modality (RGB-T). RGB-T tracking has progressed rapidly with the rise of deep learning (DL), which provides more robust and expressive feature representations than traditional hand-crafted approaches. Modern DL-based RGB-T trackers typically employ convolutional neural networks (CNNs), Siamese architectures that formulate tracking as a similarity matching task between template and search regions, or more recently, transformer-based designs capable of modeling long-range dependencies. Several DL-based RGB-T tracking frameworks emerged, generally extending powerful RGB trackers by incorporating specialized branches or fusion modules for TIR data.

Feature-level fusion, in particular, became widely adopted, extracting modality-specific features separately and merging them to form a unified representation suitable for tracking. For instance, Li et al. [27] introduced a challenge-aware framework with separate modality-shared and modality-specific branches to dynamically fuse thermal and RGB features based on environmental conditions. In [6], a Siamese-based tracker that combines hierarchical feature aggregation and bidirectional modality fusion is introduced. RGB and TIR features are extracted using a CNN backbone, aggregated across multiple levels, and collaboratively fused before being processed by a region proposal network for target localization. To adapt with object's appearance variations, a challenge-aware framework is introduced in [7] by modeling both modality-shared challenges (e.g., occlusion, low resolution) and modality-specific challenges (e.g., illumination changes, thermal crossover). Specialized feature representations are extracted through dedicated challenge branches and then fused via an interaction module to obtain the tracking result.

Decision-level fusion has also been explored, wherein tracking is performed separately on each modality and predictions are later combined based on confidence measures or ensemble methods. For example, a dual-branch Siamese tracker is introduced in [11] to produce individual response maps for the input RGB and thermal streams, then apply a joint modal channel attention module to adaptively weight and merge these response maps into a single one for the final bounding-box prediction. Similarly, Feng et al. [10] use a Siamese network with hierarchical channel attention and depth-wise correlation to generate RGB and TIR response maps, which are then adaptively fused using a contribution-aware aggregation module before classification and regression. In [13], a three-branch network is proposed by combining complementary and discriminative feature fusion with adaptive decision fusion to leverage diverse modality cues. It learns shared and distinct representations across modalities, weights them channel-wise, and dynamically fuses decision maps using confidence-aware attention.

Building on the success of transformers in vision tasks, recent RGB-T trackers have incorporated transformer-based modules to better capture complex cross-modal interactions and long-range dependencies. For instance, Shi et al. [9] utilize a CNN backbone followed by a transformer-based encoder-decoder structure. RGB and TIR features are combined using an adaptive feature-mixing mechanism and decoded via a corner-distribution-based prediction head. Similarly, an encoder-decoder network is employed in [28] to enhance global context of features extracted by a shared

CNN backbone, while a reliability estimator updates a dynamic memory online to adapt to temporal appearance changes. Feng et al. [8] propose a trifurcate tree backbone with sparse transformers to capture both shared and modality-specific representations, followed by a confidence-aware aggregation module and dual-head prediction for classification and localization. In [29], a spatio-temporal module is combined with the transformer backbone to perform bidirectional modality enhancement via cross-attention between RGB and TIR features, while integrating dynamic tokens from previous frames to inject temporal awareness. By jointly refining search and template tokens with both spatial and temporal cues, the tracker can captures target variations and modality interactions across time. Similarly, Hui et al [5] proposed the template-bridged search interaction (TBSI) module, using the initial template as a medium to gather and distribute target-relevant features between RGB and TIR search regions to enhance tracking accuracy.

2.2.2 RGB-D Tracking

More recently, the success of RGB-T tracking inspired similar multi-modal approaches that combine RGB with other modalities, notably depth (RGB-D) and event data (RGB-E). Depth sensors provide critical spatial and geometric cues, enhancing tracking stability under cluttered scenes and background ambiguity. RGB-D trackers typically integrate depth information using dedicated fusion modules, or via spatial attention mechanisms emphasizing object contours and depth discontinuities. For instance, Yan et al. [30] introduced the DepthTrack benchmark as well as an RGB-D tracking baseline, highlighting the need for deeper integration of depth data rather than simply using it as auxiliary cues. Li et al. [31] presented a 3D prompt-based learning framework, using geometric cues from point-cloud representations to enhance the discriminative capability of RGB trackers. Additionally, hierarchical modality aggregation methods such as the HMAD tracker [32] have shown that selectively incorporating depth features at different network stages greatly improves tracking robustness, especially in cluttered and ambiguous scenes.

2.2.3 RGB-E Tracking

Event-based cameras, offering low latency and high temporal resolution, have also begun to be explored for tracking rapidly moving objects. RGB-E methods typically exploit temporal cues provided by asynchronous events to improve tracking in highly dynamic scenes. For example, Wang et al. [33] introduced the VisEvent benchmark as well as an RGB-E tracking baseline, using a cross-modality transformer that fuses visible and event streams, significantly improving tracking under low light and fast motion. Recent advancements include the frame-event alignment method [34], which introduced a two-part network for aligning asynchronous event data with traditional RGB frames, significantly improving high-speed tracking performance. Zhang et al. [35] proposed an RGB-E tracker that treats single-object tracking as a multi-object problem, using appearance and motion embeddings from RGB and event modalities within a transformer framework.

Despite the effectiveness of these multi-modal tracking methods, they often require substantial architectural modifications, such as modality-specific branches, multi-stage fusion mechanisms, or carefully designed interaction modules. These additions significantly increase computational costs, making them less suitable for real-time or resource-constrained applications. As a result, a clear need remains for fusion methods that combine efficiency with simplicity, scalability, and minimal architectural overhead. Pixel-level fusion has the potential to address precisely these needs, aiming to deliver effective multi-modal tracking through a fusion-before-tracking strategy, which is largely under-explored. On the other hand, there has been growing interest in more efficient strategies, such as knowledge distillation and prompt learning, which aim to enhance multi-modal tracking performance without modifying the base architecture or introducing large numbers of additional parameters. We discuss these strategies in the following sections.

2.3 Related Work on Pixel-level Fusion Schemes

Pixel-level fusion represents an alternative fusion paradigm that merges multi-modal information directly at the image level, prior to feature extraction or tracking. Unlike feature-level or decision-level approaches, pixel-level fusion produces a unified multi-modal representation as an enhanced input image suitable for direct use by conventional RGB trackers while retaining low-level visual and structural information from both modalities.

Early pixel-level fusion methods primarily employed simple concatenation or weighted averaging strategies. For instance, Chen et al. [36] introduced a fully 3D CNN-based encoder-decoder framework that concatenates each RGB image and its depth map as two "time" slices in a 4-D input, then uses inflated 3D convolutions in the encoder to jointly learn from both modalities at the pixel level. A mirrored 3D decoder with rich back-projection paths and channel-modality attention refines these fused representations before prediction. Similarly, a four-channel RGB-T representation is employed in [37] to generate fused histograms for correlation-filter-based tracking, enhancing robustness through improved foreground-background segmentation. Zhang et al. [38] experimented with straightforward fusion techniques, such as averaging RGB and TIR images or replacing an RGB channel with thermal data in conjunction with Siamese tracking frameworks [39]. mfDiMP [25] concatenated RGB and TIR images along the channel dimension, achieving superior tracking accuracy through end-to-end fusion and training. In [40], RGB and depth patches are stacked as multi-channel inputs into a spatial-attention module to select RGB-D fused glimpses. A parallel CNN and dynamic-filter streams are employed to extract appearance and spatial cues which are then fed into an LSTM-based state estimator and an MLP to iteratively refine the target bounding box. Such methods frequently suffer from image misregistration, where subtle pixel-level shifts between RGB and auxiliary modality (X-modality) introduce noise, significantly impairing tracking performance.

To address misregistration issues, adaptive pixel-level fusion techniques have been proposed. Notably, the dual adversarial pixel fusion network (DAPFNet) [41] adaptively combines RGB and thermal modalities within a generator-discriminator framework, allowing effective fusion of non-aligned images. Although computationally efficient, DAPFNet depends on explicit assumptions about texture and structural characteristics, such as presuming RGB imagery provides richer texture details and thermal images primarily contain structural information, which may not consistently hold across varying scenarios. The reliance on handcrafted loss functions enforcing these assumptions can hinder adaptability and overall fusion effectiveness. Similarly, DFAT [42] employs a modality decomposition approach (MDLatLRR [43]), decomposing RGB and TIR images separately into base and detail components before fusion. Although DFAT produces visually appealing fused images, its fusion approach is completely detached from the tracking task, using predefined rules aiming for improved visual quality rather than tracking-specific performance. Recently, Zheng et al. [44] introduced a comprehensive multi-level fusion method, explicitly combining RGB and thermal modalities at pixel, feature, and decision levels, demonstrating improved robustness under challenging visual conditions. Nevertheless, its pixel-level fusion remains fundamentally guided by general-purpose rules rather than task-specific optimization for tracking performance.

Building on this area of research, our proposed SBiDF framework takes a fundamentally different approach by explicitly learning fusion strategies tailored to tracking across multiple modalities. Instead of relying on fixed assumptions, handcrafted rules, or visually driven fusion objectives, SBiDF introduces a symmetric, bidirectional dynamic convolutional filtering mechanism that adaptively modulates both RGB and X-modality features at the pixel level. The fusion is directly supervised by the tracking loss, ensuring that the learned representations are optimized for tracking accuracy rather than human perception. The proposed SBiDF establishes deep, bidirectional feature interactions, allowing RGB and X-modality features to guide each other dynamically. This enables a more expressive and generalizable fusion mechanism, effective across diverse modalities such as thermal, depth, and event data. **To the best of our knowledge, SBiDF is the first unified multi-modal tracker explicitly built on a pixel-level fusion strategy.**

2.4 Related Work on Efficiency-driven Schemes

2.4.1 Prompt Learning

Prompt learning-based methods have recently emerged as lightweight alternatives for multi-modal tracking. Prompt-based methods typically insert compact adapters into pre-trained and frozen RGB tracking backbones, enabling efficient interaction between modalities at a minimal computational cost. For example, Cao et al. [17] propose a universal bi-directional adapter that enables cross-prompting between modalities within frozen transformer layers, facilitating adaptive feature exchange without assuming a fixed dominant modality. Similarly, Zhu et al. [15] inject modality-specific prompts into intermediate layers of a frozen transformer using lightweight complementary blocks. Building upon this, Wang et al. [18] extend the idea into a two-stream framework that incorporates spatio-temporal interactions to aggregate features over time. In [45], a unified multi-modal tracker is introduced by injecting learnable cross-modal prompts into a frozen RGB transformer to enrich its feature tokens with auxiliary modality cues, and by efficiently fine-tuning this backbone through low-rank adaptation (LoRA) for end-to-end tracking. While these prompt-based approaches offer high efficiency and minimal parameter overhead, their representational capacity may be limited by fixed prompt dimensions, making them less adaptable in dynamic or complex tracking scenarios.

2.4.2 Knowledge Distillation

In parallel, knowledge distillation (KD) has gained traction as a complementary strategy for improving efficiency, wherein a compact student network learns to replicate the behavior of a more powerful teacher model [46]. KD has been widely explored across domains such as machine translation [47], image classification [48], pose estimation [49], and person re-identification [50]. In RGB-T tracking, KD is increasingly employed to transfer the rich multi-modal representations of fusion-heavy teacher networks into lightweight student trackers. Zhang et al. [20] pioneer this idea by distilling semantic and spatial knowledge from a sophisticated dual-stream teacher into a simpler student via multi-path distillation. More recently, Hong et al. [24] propose an online degradation distillation strategy, where degraded inputs are fed to the student branch while the teacher processes clean inputs. This enables the student to recover missing modality information through supervised alignment with the teacher's outputs. While effective, KD-based methods often require two-stage training and careful design of distillation losses. The need for a separate teacher network also increases memory usage and training time, which can limit scalability and deployment.

Self-distillation (SD) addresses these limitations by using a single network as both teacher and student, leveraging internal representations for self-guided learning. This paradigm has shown particular promise for improving model robustness and feature discriminability [26, 51]. However, in RGB-T tracking, SD remains relatively unexplored but has demonstrated potential in recent works. For instance, Hong et al. [21] introduce a dual-branch transformer with shared parameters and concatenated RGB-TIR tokens. Their self-distillation framework dynamically identifies the stronger modality and distills feature- and attention-level knowledge from strong to weak branches during training. Although this SD framework improves modality interaction and robustness, it relies on a dual-branch transformer design, which increases both parameter count and computational overhead.

While prompting- and distillation-based approaches offer various trade-offs in accuracy, speed, and complexity, achieving a unified solution that delivers robust tracking, architectural simplicity, and training efficiency remains an open challenge. Our proposed MSD approach addresses this gap by introducing multi-level self-distillation framework that maintains the structure and efficiency of the original RGB tracker. Building on the aforementioned efficiency-driven schemes, MSD takes a fundamentally different approach by essentially shifting the focus from *how to fuse* (the dominant question in RGB-T tracking) to *how to train* effectively with multi-modal data.

Generally speaking, An RGB-T tracker receives as input two synchronized video streams: RGB frames I^{RGB} capturing color and texture cues, and TIR frames I^{TIR} capturing heat signatures robust to illumination changes. Given an initial target bounding box in the first frame, the goal of the

RGB-T tracker is to estimate the target's state for each subsequent frame. The output of the tracker is a predicted bounding box specifying the top-left location of the tracked object along with the object's size (width and height).

A generic RGB-T tracker performs three main stages: (1) feature extraction of the input modalities, (2) cross-modal fusion to combine complementary cues, and (3) target state estimation through similarity matching or regression. The cross-modal fusion stage can occur at different levels, pixel, feature, or decision, depending on the tracker design.

Performance is commonly evaluated by precision and success rates (measuring localization accuracy and overlap with ground truth), as well as efficiency metrics such as inference speed (frames per second) and parameter count. Typical challenges include large appearance variations, illumination changes, occlusion, and thermal crossover. This formal description defines the system studied in this thesis and provides the foundation for the SBiDF and MSD methods proposed in the following chapters.

2.5 Summary

In this chapter, we have reviewed the existing literature on multi-modal object tracking, with a focus on RGB-Thermal (RGB-T) fusion approaches, pixel-level fusion methods, and efficiency-driven strategies. We have first discussed how early and recent RGB-T tracking methods have leveraged deep learning architectures to integrate complementary modality information, often employing dual-stream networks, hierarchical fusion modules, or transformer-based encoders to achieve state-of-the-art performance. While effective, these solutions frequently incur significant computational overhead, architectural complexity, and increased parameter counts that limit their practicality in resource-constrained or real-time scenarios. Moreover, we have examined pixel-level fusion techniques, which integrate modalities at the earliest stage of the tracking pipeline. While early pixel-level fusion methods relied on static concatenation or averaging strategies, more recent work has introduced adaptive fusion networks and modality decomposition approaches. However, most of these methods either depend on general-purpose visual quality objectives, predefined fusion rules, or modality-specific designs, limiting their scalability and effectiveness for tracking tasks. To address efficiency concerns, recent research has explored prompt learning and knowledge distillation as promising alternatives to complex fusion-heavy designs. Despite their benefits, these approaches still face challenges such as limited adaptability, reliance on multi-stage training, or dependence on external teachers. In light of these observations, there remains a clear need for unified frameworks that combine the simplicity and efficiency of pixel-level fusion with the learning capabilities of modern self-supervised strategies, while avoiding architectural modifications and excessive training complexity. The proposed SBiDF and MSD frameworks presented in Chapter 3 and 4, respectively, directly address these gaps by introducing dynamic pixel-level fusion and multi-level self-distillation schemes explicitly optimized for robust, efficient multi-modal tracking. The following chapters describe these contributions in detail.

Chapter 3

A Pixel-level Fusion Framework for Unified Multi-modal Object Tracking

3.1 Introduction

In RGB-X tracking systems, where X denotes an auxiliary modality such as thermal (T), depth (D), or event (E), fusion is typically performed at either the feature level or the decision level. In this work, we take a fundamentally different approach by proposing a unified framework that transforms a standard RGB tracker into a multi-modal one without architectural modifications, auxiliary branches, or external teacher networks. Our key idea is to perform pixel-level fusion, generating a fused RGB-X image that can be directly processed by any off-the-shelf RGB tracker. This fusion-before-tracking strategy preserves the simplicity and training dynamics of the original tracker while leveraging complementary information from the auxiliary modality. Although largely underexplored in the deep learning-based multi-modal tracking literature, pixel-level fusion offers key advantages: it retains fine-grained spatial details across modalities and enables seamless integration with a wide range of pre-trained RGB trackers, without requiring backbone changes or modality-specific components. Moreover, it allows for reusing existing RGB models via fine-tuning on the fused image, reducing training cost while preserving the benefits of large-scale RGB pretraining.

In this chapter, we introduce the symmetric bidirectional dynamic fusion (SBiDF), a novel pixel-level fusion network designed to enrich the RGB modality before employing it as input for tracking [52, 53]. The network architecture comprises four key components: an RGB autoencoder, an X-modality autoencoder, dynamic convolutional filtering (DCF) blocks, and an output fusion module. Each modality-specific autoencoder contains an encoder responsible for extracting multi-scale hierarchical features, followed by three specialized decoders that reconstruct base and detailed representations for both RGB and X-modality. Central to SBiDF are four bidirectional DCF blocks, which enable content-aware cross-modal enhancement. Specifically, two X-guided RGB fusion blocks dynamically modulate hierarchical RGB detail maps conditioned upon the corresponding X-modality detail maps, while two RGB-guided X-modality fusion blocks conversely modulate X-modality features using RGB details. These DCF blocks generate pixel-wise kernels adaptively based on the guiding modality, enabling soft, content-aware enhancement of one modality by the other. Subsequently, the dynamically enriched RGB and X-modality features are independently aggregated and then averaged to obtain the final fused output. This bidirectional refinement promotes more effective inter-modality interactions and balanced modality contributions, resulting in more informative fused inputs for the tracker that improve its robustness across diverse conditions.

Unlike generic pixel-level fusion approaches designed for image enhancement or reconstruction, SBiDF is tailored explicitly for object tracking, adaptively enhancing RGB content by dynamically incorporating complementary cues from the X-modality. It generates fused RGB-X images that are optimized for tracking, with features emphasized according to their relevance to the target object. During training, the fused RGB-X image serves as the sole input to the tracking network, and the tracker loss provides the only supervision. Thus, SBiDF avoids relying on modality-specific priors or auxiliary objectives; instead, it learns a fusion strategy specifically oriented toward tracking performance.

The effectiveness of the proposed approach is validated on four diverse RGB-X tracking benchmarks: LasHeR [54], RGBT234 [1], DepthTrack [30], and VisEvent [33], covering RGB-T, RGB-D, and RGB-E tracking scenarios. Our method is evaluated against several recent state-of-the-art trackers, including unified multi-modal (RGB-X) trackers and modality-specific (RGB-T, RGB-D, and RGB-E) trackers. The compared trackers cover a wide range of adopted fusion strategies, including feature-, decision-, and pixel-level fusion, as well as efficiency-oriented methods relying on prompt learning and knowledge distillation. Across all benchmarks, SBiDF achieves strong performance, demonstrating its generality, robustness, and adaptability under a wide range of conditions.

This chapter is organized as follows. Section 3.2 provides a detailed description of the proposed SBiDF approach. Experimental results are presented in Section 3.3 followed by a conclusion in Section 3.4.

3.2 Methodology

This section presents our proposed *symmetric bidirectional dynamic fusion* (SBiDF) network. The development of SBiDF was motivated by our initial work, the *hierarchical feature difference autoencoder* (HFDAE) [52], which lays the foundation for task-driven pixel-level RGB-T fusion. The HFDAE approach is designed to perform pixel-level fusion between RGB and TIR images to enhance the quality of fused inputs for robust RGB-T tracking. HFDAE utilizes an asymmetric encoder-decoder architecture, in which a shallow RGB encoder captures coarse-level features, and a deeper TIR encoder extracts hierarchical details at multiple scales. Fusion occurs through a combination of a base RGB image and TIR-derived detail differences, explicitly designed to preserve thermal structural cues and RGB textural information within a unified fused representation. This simple asymmetric encoder-decoder architecture, primarily designed to preserve thermal structural cues, limits the approach to RGB-T fusion only. In contrast, the SBiDF approach proposes a significantly improved network, which employs deeper, symmetric, and bidirectional feature interactions between RGB and auxiliary modalities. The enhancements are driven by a key idea: allow both RGB and X-modality to influence each other dynamically rather than relying solely on a fixed difference-based fusion. Consequently, the proposed SBiDF generalizes effectively across multiple modalities (thermal, depth, event) through dynamic pixel-level fusion, enabling enhanced tracking accuracy and broader applicability.

In the following subsections, we first outline the preliminary HFDAE framework [52]. Then, we provide a detailed description of the proposed architecture of SBiDF [53]. Then, we provide an explanation of the loss function used for optimization, and finally, we present the training procedure.

3.2.1 HFDAE: Preliminary Pixel-level Fusion Framework

The proposed HFDAE network [52] is designed to enhance the RGB modality *prior* to its use as input for a tracker. HFDAE adaptively refines RGB content for tracking tasks by leveraging complementary cues from the TIR modality. This adaptive mechanism enables dynamic, pixel-level fusion tailored to the unique characteristics of both the object and its environment, thereby improving tracking performance.

The overall architecture of the proposed HFDAE network is illustrated in Fig. 3.1. HFDAE consists of a shallow RGB autoencoder, a TIR encoder with three variable-depth decoders, and a fusion module that integrates multi-scale differential features from TIR into RGB. The TIR encoder generates feature maps at three different depths, each processed by a corresponding decoder. The outputs of these decoders, along with the output from the RGB autoencoder, are fed into the fusion module to generate the final fused image. The following subsections elaborate on each component.

3.2.1.1 RGB Autoencoder

The input RGB image I^{RGB} is processed by a shallow autoencoder to extract a base RGB image B^{RGB}. This base image serves as the fundamental representation of the scene, capturing primary structural and color information. The autoencoder consists of a single-layer convolutional encoder

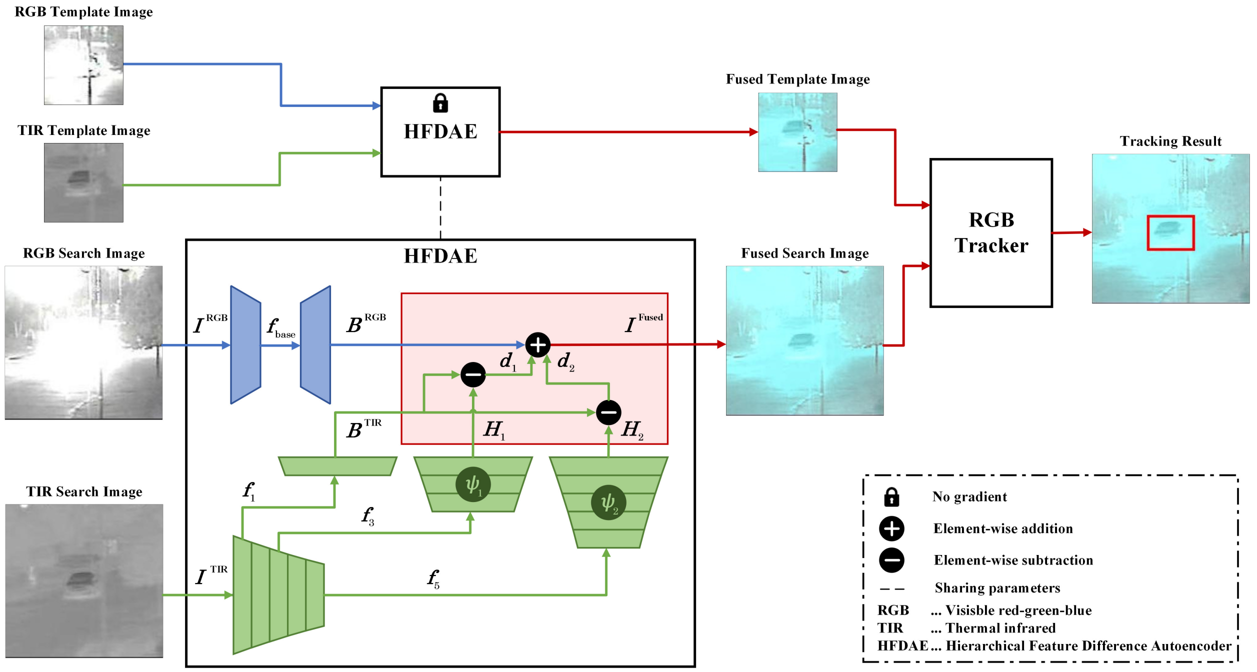


Figure 3.1: Block diagram of the proposed hierarchical feature difference auto-encoder (HFDAE) network. The network consists of three main components: a shallow RGB autoencoder, a TIR encoder with three variable-depth decoders, and a fusion module. The shallow RGB autoencoder processes the RGB modality to extract a base RGB image, capturing primary structural and color information. The TIR modality is progressively encoded, and the outputs from different encoder layers are fed into separate decoders to produce base and hierarchical TIR representations. By subtracting the higher-level TIR images (which contain fewer details) from the base TIR image, salient thermal features are extracted. These features are then added to the base RGB image to generate the final fused image.

and decoder with ReLU activation:

$$f_{\text{base}} = \sigma(W_e * I^{\text{RGB}} + b_e), \tag{3.1}$$

where W_e and b_e represent the encoder's convolutional weight and bias, respectively, $\sigma(\cdot)$ is the ReLU activation function, and $*$ denotes the convolution operation. The decoder reconstructs B^{RGB} as

$$B^{\text{RGB}} = W_d * f_{\text{base}} + b_d, \tag{3.2}$$

where W_d and b_d are the decoder's convolutional weight and bias. This base image serves as a reference onto which TIR-derived details will be applied. In other words, this base image preserves the essential RGB spatial and textural details to ensure that the additional details from the TIR modality would enhance the RGB image meaningfully for tracking, rather than distorting its original features.

3.2.1.2 TIR Autoencoder

The input TIR image I^{TIR} is processed by a five-layer convolutional encoder, where each layer progressively refines the extracted features:

$$\begin{aligned} f_l =& \sigma(W_l * f_{l-1} + b_l), \quad l = 1, 2, ..., 5, \\ f_0 =& I^{\text{TIR}}, \end{aligned} \tag{3.3}$$

where f_l, W_l, and b_l denote the feature maps, convolutional weights, and biases at layer l, respectively. Features from the first, third, and fifth layers (f_1, f_3, and f_5) are decoded to reconstruct images at different detail levels:

$$B^{\text{TIR}} = W_b * f_1 + b_b, \tag{3.4}$$

$$H_1^{\text{TIR}} = \psi_1(W_{d_1} * f_3 + b_{d_1}), \tag{3.5}$$

$$H_2^{\text{TIR}} = \psi_2(W_{d_2} * f_5 + b_{d_2}), \tag{3.6}$$

where $W_b, b_b, W_{d_1}, W_{d_2}, b_{d_1}, b_{d_2}$ are learnable parameters, and $\psi(\cdot)$ represents an upsampling operation followed by convolution.

3.2.1.3 Fusion

After obtaining the base RGB image B^{RGB}, the base thermal image, B^{TIR}, and the higher-level thermal images H_1^{TIR} and H_2^{TIR}, we employ them for effectively integrating the RGB and TIR information. Since higher-level images contain fewer details, subtracting them from this base image yields two levels of useful details, which can then be integrated into the base RGB image. Thus, these differences highlight salient thermal features.

The differences between the higher-level thermal images and the base thermal image are computed as:

$$d_1^{\text{TIR}} = H_1^{\text{TIR}} - B^{\text{TIR}}, \tag{3.7}$$

$$d_2^{\text{TIR}} = H_2^{\text{TIR}} - B^{\text{TIR}}. \tag{3.8}$$

Then, the final fused RGB image is obtained as:

$$I^{\text{Fused}} = B^{\text{RGB}} + \lambda_1 d_1^{\text{TIR}} + \lambda_2 d_2^{\text{TIR}}, \tag{3.9}$$

where λ_1 and λ_2 are fixed weights controlling the contribution of TIR perturbations. This structured fusion approach ensures that thermal information enhances object visibility and tracking robustness without distorting the original RGB features.

3.2.2 SBiDF: Pixel-level Bidirectional Fusion Framework

Building on the preliminary HFDAE framework that rely on modality-specific encoders and unidirectional enhancement, the proposed SBiDF [53] establishes deep, bidirectional feature interactions, allowing RGB and X-modality features to guide each other dynamically. This enables a more expressive and generalizable fusion mechanism, effective across diverse modalities such as thermal, depth, and event data.

As illustrated in Fig. 3.2, SBiDF comprises four primary components:

- Modality-specific (RGB and X-modality) Autoencoders: Each autoencoder contains an encoder responsible for extracting individual RGB and X multi-scale hierarchical features, followed by specialized decoders that reconstruct base and detailed representations to capture rich spatial and structural information.

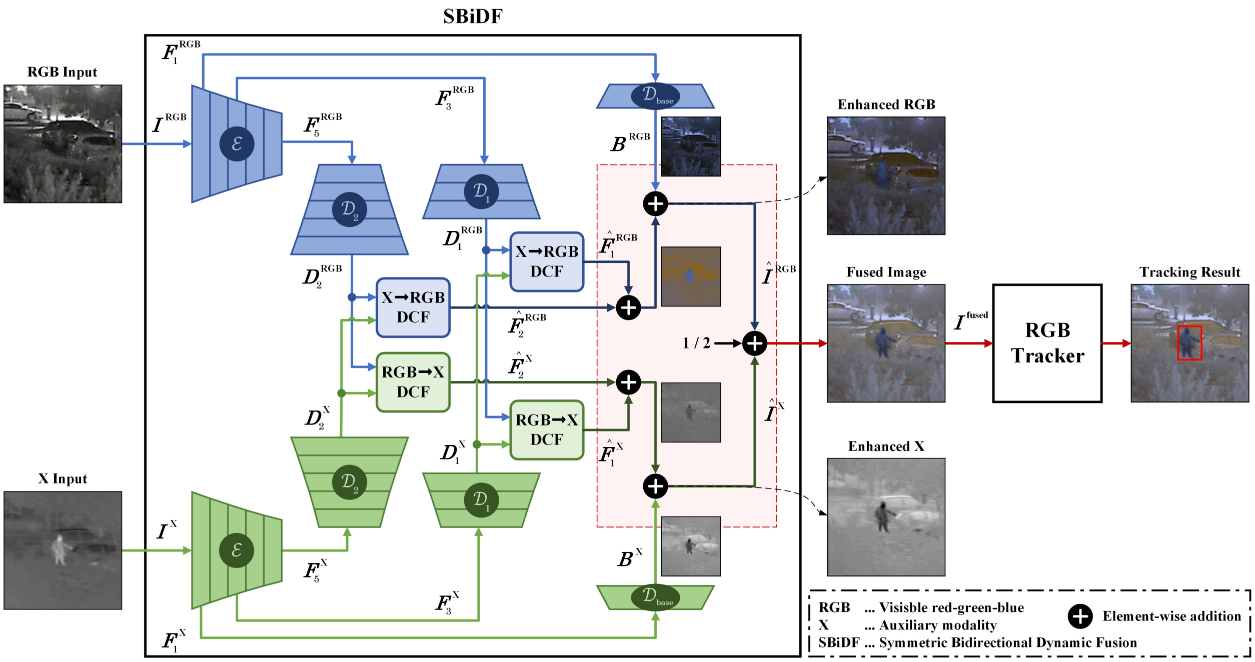


Figure 3.2: Block diagram of the proposed symmetric bidirectional dynamic fusion (SBiDF) network. The architecture consists of four main components: an RGB autoencoder, an X-modality autoencoder, dynamic convolutional filtering (DCF) blocks, and an output fusion module. Each autoencoder comprises a single encoder and three variable-depth decoders, generating base and hierarchical detail maps from the input RGB and auxiliary modality (X) images. The DCF blocks perform adaptive bidirectional fusion to produce enhanced representations for the final fused output optimized for tracking.

- Dynamic Convolutional Filtering (DCF) Blocks: Implements novel adaptive convolutional operations that dynamically modulate RGB and X-modality detail maps bidirectionally, using pixel-wise kernels conditioned on complementary modalities.

- Final Fusion: Aggregates the dynamically enhanced features from both modalities, combining them into a single fused image optimized explicitly for object tracking tasks.

The proposed network processes both RGB and X-modality inputs through identical five-layer encoders. This symmetry allows both streams to contribute richer semantic and structural information across scales. Then, four dynamic convolutional fusion blocks are employed to modulate the detail features in one modality using spatially-varying kernels generated from the other modality. These blocks perform guided enhancement, where the structure in one modality is used to refine feature content in the other. This bidirectional guidance ensures stronger cross-modal interaction, improving the quality of the fused image. The fusion process is completed by independently enhancing both RGB and X reconstructions and then aggregating the two into a single output. This final image integrates RGB and X-modality information in a spatially-aware and adaptive manner, significantly improving robustness to modality-specific artifacts, noise, and occlusions. The resulting fused image serves directly as input to an RGB tracker, without requiring architectural modifications or additional training stages. In the following, we detail each architectural component of the proposed network.

3.2.2.1 Modality-specific Auto-encoders

The auto-encoder architecture in SBiDF is specifically designed to extract, reconstruct, and represent rich hierarchical features from each input modality (RGB and X modalities). This structure comprises modality-specific encoders and multiple decoders, to capture both base-level content and fine-grained details at multiple scales within each modality.

Encoder Modules Each modality employs a dedicated encoder composed of five convolutional layers. Given an input modality image I^m, where $m \in \{RGB, X\}$, the encoder progressively extracts hierarchical features at multiple scales. Formally, the features extracted at each encoder layer l can be described as

$$\begin{aligned} F_l^m =& \mathcal{E}_l^m(F_{l-1}^m), \quad l \in \{1,2,3,4,5\}, \\ F_0^m =& I^m, \end{aligned} \tag{3.10}$$

where $\mathcal{E}_l^m(\cdot)$ denotes the convolutional and activation operations at layer l. Each encoder layer downsamples the spatial dimensions while increasing the channel depth, capturing progressively abstract features suitable for reconstructing diverse levels of image details.

The hierarchical structure of these encoders enable rich multi-scale feature extraction, allowing the network to effectively capture high-level semantic context alongside low-level spatial details, both crucial for robust fusion.

Multi-Scale Decoders To reconstruct meaningful representations from encoded features, each modality uses three specialized decoders: a Base decoder and two hierarchical Detail decoders (Detail-1 and Detail-2).

- Base Decoder: Reconstructs the fundamental modality-specific appearance from shallow encoder features (F_1^m) at the highest spatial resolution. The base reconstruction is defined as

  $$B^m = \mathcal{D}_{base}^m(F_1^m), \tag{3.11}$$

  where $\mathcal{D}_{base}^m$ represents a convolutional operation.

- Detail Decoders: Generate finer-grained detail maps from deeper features, capturing mid- and high-level information. Detail-1 and Detail-2 decoders utilize the intermediate-level encoder features (F_3^m) and the deepest features (F_5^m), respectively, to reconstruct medium-scale details D_1^m and high-scale fine-grained details D_2^m as

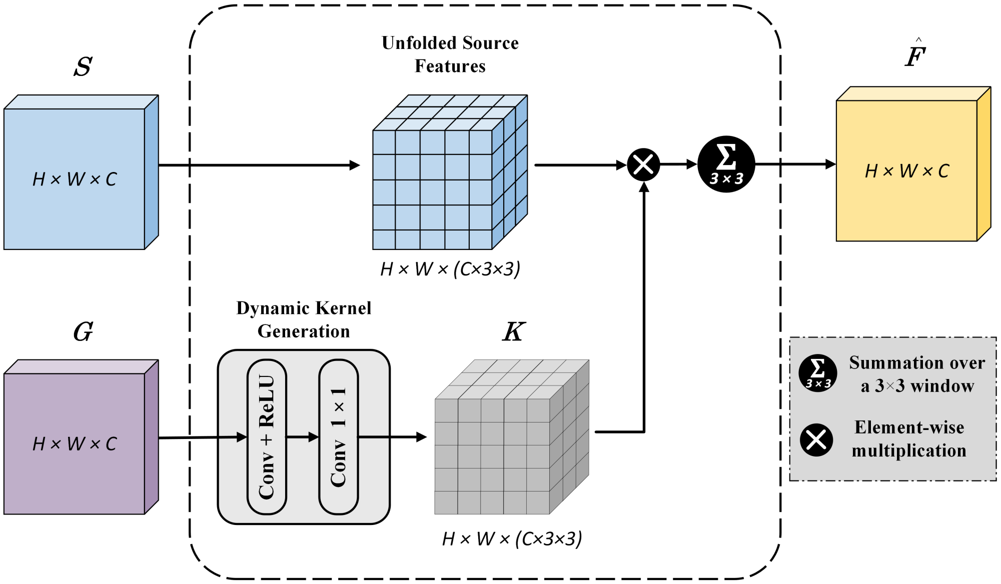
  $$\begin{aligned} D_1^m =& \mathcal{D}_1^m(F_3^m), \\ D_2^m =& \mathcal{D}_2^m(F_5^m), \end{aligned} \tag{3.12}$$

  where $\mathcal{D}_1^m$ and $\mathcal{D}_2^m$ includes multiple transpose convolutional layers followed by a single convolutional layer to progressively upsample features to the original image resolution.

The hierarchical reconstruction facilitated by these decoders ensures accurate and discriminative representations at multiple spatial scales. These rich, scale-aware representations form the basis for subsequent dynamic cross-modality interactions.

3.2.2.2 Dynamic Convolutional Filtering (DCF) Blocks

To effectively fuse complementary details across modalities, SBiDF employs novel dynamic convolutional filtering (DCF) blocks. Unlike standard fusion methods that use fixed convolutional kernels, our DCF blocks adaptively generate convolutional kernels conditioned on complementary



S
$H \times W \times C$
Unfolded Source Features
$H \times W \times (C{\times}3{\times}3)$
$\hat{F}$
$H \times W \times C$
Σ 3×3
G
$H \times W \times C$
Dynamic Kernel Generation
Conv + ReLU
Conv 1×1
K
$H \times W \times (C{\times}3{\times}3)$
Summation over a 3×3 window
Element-wise multiplication

Figure 3.3: Detailed block diagram of the dynamic convolutional filtering (DCF) block. The DCF block adaptively generates spatially varying convolutional kernels K conditioned on the guide modality features G. These kernels are applied pixel-wise to the unfolded patches of the source modality S, producing dynamically enhanced representations $\hat{F}$.

modality features. This mechanism allows pixel-wise, context-aware fusion, leading to enhanced representation quality tailored specifically for object tracking tasks. Our design of DCF blocks is inspired by recent advances in dynamic convolution [55, 56]. In particular, MFGNet [55] uses a cross-modal attention mechanism to generate convolutional filters tailored to each input modality. Differently, DFNet [56] dynamic fuses fixed modality-shared and non-shared convolutional kernels in the kernel-space to capture modality-specific and common features. Building upon these pioneering ideas, our proposed DCF blocks leverage dynamic convolution principles but in a different manner introducing a more symmetric, bidirectional approach.

Specifically, we introduce two types of DCF to facilitate bidirectional cross-modal interaction: RGB fusion guided by X-modality (X→RGB) and X-modality fusion guided by RGB (RGB→X). The general block diagram of the DCF block is shown in Fig. 3.3. Formally, let S and G represent the source and guiding modality feature maps, respectively. Each ($G \rightarrow S$) DCF block dynamically generates spatially varying kernels from the guiding modality G, which are then applied to the source modality S. The dynamic kernel generation process can be expressed as

$$K_{x,y} = \phi(G)_{x,y}, \tag{3.13}$$

where $\phi(\cdot)$ is a small neural network (comprising convolutional and nonlinear activation layers) that outputs adaptive convolutional 3×3 kernels at each spatial location (x, y). To perform adaptive convolution on the source feature S, each pixel location is processed as follows

$$\hat{F}_{x,y} = \sum_{(i,j)\in\Omega} K_{x,y}(i, j) \cdot S_{x+i,y+j}, \tag{3.14}$$

where $\hat{F}_{x,y}$ is the dynamically fused feature at position (x, y), Ω is the convolution window corresponding to the 3×3 local neighborhood of (x, y) in S, and $K_{x,y}(i, j)$ is the dynamically generated kernel weight for spatial offsets (i, j) at pixel (x, y). In practice, the convolution operation defined in (3.14) is implemented using unfold process as illustrated in Fig. 3.3. Specifically, the local neighborhoods (patches) of size 3×3 around each spatial location in S are first extracted into an unfolded representation. Subsequently, adaptive kernels K are multiplied element-wise with these unfolded patches, and the results are summed along the patch dimension to produce the dynamically convolved output $\hat{F}$.

In SBiDF, four DCF blocks are used:

- Two X-modality-guided RGB fusion blocks (X→RGB): These blocks enhance RGB detail maps by dynamically modulating RGB features with kernels conditioned on corresponding X-modality detail maps as

$$\hat{F}_d^{RGB} = DCF(S = D_d^{RGB}, G = D_d^{X}), d \in \{1, 2\}, \tag{3.15}$$

- Two RGB-guided X-modality fusion blocks (RGB→X): Similarly, these blocks enrich X-modality details by applying dynamically generated kernels from RGB detail maps as

$$\hat{F}_d^X = DCF(S = D_d^X, G = D_d^{RGB}), d \in \{1, 2\}, \tag{3.16}$$

This symmetric bidirectional approach ensures balanced interaction, allowing each modality to effectively enhance the other based on local visual context. Such adaptive fusion significantly improves the overall tracking representation by emphasizing task-specific features critical for accurate tracking.

3.2.2.3 Final Fusion

After obtaining dynamically enhanced modality-specific representations through the proposed DCF blocks, SBiDF combines these enhanced features to generate the final fused image, optimized explicitly for object tracking. Formally, let

$$\begin{aligned} \hat{I}^{RGB} =& B^{RGB} + \lambda_1^{RGB} \hat{F}_1^{RGB} + \lambda_2^{RGB} \hat{F}_2^{RGB}, \\ \hat{I}^X =& B^X + \lambda_1^X \hat{F}_1^X + \lambda_2^X \hat{F}_2^X, \end{aligned} \tag{3.17}$$

be the dynamically enhanced representation obtained through bidirectional dynamic convolutional fusion for RGB and X-modality, respectively, where B^{RGB} and B^X are the reconstructed base images, $\hat{F}_d^{RGB}$ and $\hat{F}_d^X$ ($d \in \{1, 2\}$) represent dynamically fused detail maps, and λ parameters control their relative contributions. Then, the final RGB-X fused output I^{fused} is computed through a straightforward aggregation of both enhanced representations as

$$I^{fused} = \frac{\hat{I}^{RGB} + \hat{I}^X}{2}. \tag{3.18}$$

3.2.3 Loss Function and Optimization

The proposed SBiDF network is designed explicitly for improving object tracking performance by producing fused RGB-X images optimized for the tracking task, rather than general-purpose visual fusion. To achieve this, the network is trained end-to-end using a tracking-specific loss function that directly measures and optimizes tracking accuracy.

Specifically, we follow the tracking loss formulation introduced in [57], which effectively balances object localization accuracy and foreground-background classification robustness. The tracking loss $\mathcal{L}_{\text{track}}$ is composed of three complementary terms: classification loss ($\mathcal{L}_{cls}$), Intersection-over-Union (IoU) regression loss ($\mathcal{L}_{iou}$), and l_1 bounding-box regression loss ($\mathcal{L}_{l_1}$). Formally, the overall tracking loss function is expressed as:

$$\mathcal{L}_{\text{track}} = \mathcal{L}_{cls} + \lambda_{iou}\mathcal{L}_{iou} + \lambda_{l_1}\mathcal{L}_{l_1}, \tag{3.19}$$

where λ_{iou} and λ_{l_1} are hyperparameters controlling the relative importance of each regression loss component.

The classification loss term $\mathcal{L}_{cls}$ is implemented using a weighted focal loss [58], effectively addressing the class imbalance typically observed between foreground objects and background regions in tracking tasks. By emphasizing challenging examples (e.g., partially occluded objects or visually ambiguous backgrounds), this focal loss significantly enhances object discrimination accuracy. On the other hand, bounding-box regression accuracy is ensured through two complementary loss terms. The generalized IoU loss $\mathcal{L}_{iou}$ [59] measures spatial overlap between the predicted and ground-truth bounding boxes, ensuring robust localization and scale alignment of tracked objects. Meanwhile, the l_1 loss term ($\mathcal{L}_{l_1}$) promotes precise regression by directly penalizing deviations from the ground-truth bounding-box coordinates, enhancing stability during tracking and reducing jitter.

Through direct optimization of this comprehensive tracking objective, SBiDF learns fusion strategies explicitly tuned for robust tracking performance. Importantly, no additional fusion-specific or modality-specific reconstruction losses are required, simplifying training and ensuring the learned fused representations directly support improved tracking outcomes.

3.2.4 Training Procedure

The proposed SBiDF network is trained end-to-end using paired modality data (RGB-TIR, RGB-D, or RGB-E images). Prior to initiating the joint training process, we first initialize the RGB tracking backbone with pretrained parameters obtained from large-scale RGB datasets. Leveraging this pretrained knowledge provides the tracker with strong baseline performance and stability, enabling efficient adaptation to fused multi-modal images during subsequent training phases.

During training, the fusion network (SBiDF) is trained from scratch, allowing modality-specific feature extraction and fusion processes to be optimized explicitly for tracking. In contrast, the RGB tracker is finetuned based on the fused images generated by SBiDF. Thus, the tracker adapts its pre-trained visual representations specifically toward multi-modal fused images, improving robustness across diverse tracking conditions.

Each training sample consists of paired modality frames (e.g., RGB-TIR, RGB-D, or RGB-E), cropped into two sets: template and search images. While both template and search image pairs pass through the SBiDF fusion network, gradient backpropagation is only permitted through the fusion of the search images. Consequently, SBiDF explicitly learns an effective pixel-level fusion strategy that minimizes the tracking loss specifically on search images. This learned fusion strategy is then applied consistently to template images during inference, ensuring compatibility and robust visual correspondence between template and search pairs.

Training optimization employs the AdamW optimizer, which combines adaptive gradient-based optimization with weight decay regularization to improve convergence stability and reduce over-fitting. The network is trained for a total of 30 epochs, with each epoch consisting of 60,000 randomly sampled training pairs. At the conclusion of every epoch, model validation is performed on an independent set comprising 10,000 validation samples. The validation performance is evaluated using the mean IoU metric, providing an objective measure of tracking localization accuracy. To select the optimal SBiDF model, we retain the checkpoint yielding the highest validation mean IoU across all epochs. This rigorous validation and checkpointing approach ensures the final fusion network generalizes effectively across a variety of unseen scenarios and maximizes practical tracking performance.

3.3 Experimental Results

To validate the effectiveness and generalization of the proposed SBiDF framework, we conduct extensive experiments across multiple RGB-T, RGB-D, and RGB-E tracking benchmarks. In these experiments, we compare SBiDF with unified multi-modal trackers and modality-specific trackers, covering a wide range of fusion strategies, including feature-, decision-, and pixel-level fusion, as well as efficiency-oriented methods relying on prompt learning and knowledge distillation. This section describes the evaluation setup, implementation details, and a comprehensive comparison against recent state-of-the-art methods. We further present visual evaluation and ablation studies to highlight the contributions of each component within the proposed architecture.

3.3.1 Datasets and Evaluation Metrics

We evaluate SBiDF across several large-scale benchmarks spanning RGB-T, RGB-D, and RGB-E modalities:

LasHeR [54] A comprehensive benchmark for RGB-T tracking comprising 1224 paired RGB and thermal infrared video sequences. Of these, 979 sequences are designated for training, and 245 sequences form the test set. The approximate average number of frames per video sequence pair is 596 frames. Moreover, LasHeR includes 19 carefully annotated challenge attributes such as occlusion, deformation, low illumination, and scale variation, facilitating detailed analysis of tracker performance under diverse and realistic conditions.

RGBT234 [1] A widely-used RGB-T tracking dataset containing 234 sequences with precisely aligned RGB and thermal frames. The average number of frames per sequence is approximately 1000 frames. Each sequence is annotated with one or more of 12 challenge attributes, including heavy occlusion, motion blur, thermal crossover, and low resolution, enabling evaluation under various environmental and visual challenges.

DepthTrack [30] Consists of 202 video sequences for RGB-D tracking, partitioned into 152 training sequences and 50 testing sequences. Each sequence contains densely annotated RGB frames and aligned depth maps, with an average of 1473 frames per sequence. This dataset emphasizes scenarios requiring geometric reasoning and depth-aware discrimination.

VisEvent [33] A widely-used RGB-E tracking dataset, providing 820 sequences in total, with an average of 450 frames for each sequence. The dataset includes 500 sequences for training and 320 sequences for testing. VisEvent focuses on high-speed motion and low-light conditions, leveraging event data's high temporal resolution to complement RGB information.

To ensure fair and comprehensive comparisons across datasets, we adopt standard evaluation metrics specific to each benchmark. For the LasHeR, RGBT234, and VisEvent datasets, we report Precision Rate (PR) and Success Rate (SR), which are widely used in object tracking benchmarks. The PR measures the percentage of frames where the predicted target center falls within a predefined distance threshold from the ground truth. While, the Success Rate evaluates the overlap between the predicted bounding box and the ground truth by computing the Intersection over Union (IoU). For the DepthTrack dataset, we adopt the F-score, Precision (Pr), and Recall (Re) evaluation metrics. Precision quantifies the proportion of correctly predicted bounding boxes relative to all predictions, while Recall measures the proportion of correctly predicted bounding boxes relative to all ground truth annotations. The F-score provides a balanced measure of tracking accuracy, combining Precision and Recall into a single metric.

3.3.2 Implementation Details

The proposed SBiDF framework is designed as a general pixel-level fusion module applicable to a wide range of trackers and modalities, rather than being tied to a specific tracking architecture. In all experiments, we integrate SBiDF by feeding the fused RGB-X images directly as input to an off-the-shelf RGB tracker. For consistency and reproducibility, we adopt OSTrack [57] as the base tracker across all modalities. Specifically, the tracker is initialized using the pretrained weights from DropTrack [60], which provides a strong starting point for fine-tuning on fused multi-modal inputs. For each modality configuration (RGB-T, RGB-D, RGB-E), the SBiDF fusion module and OSTrack backbone are trained together end-to-end on the corresponding training subsets of LasHeR, DepthTrack, and VisEvent, respectively. The entire framework is implemented in Python using the PyTorch deep learning library. In the model development and training phase, we used a computing platform with 32GB RAM and an NVIDIA V100 GPU [61]. Inference and evaluation are performed on a another workstation equipped with an Intel Core i9-9900K CPU @3.6 GHz, 64GB of RAM, and a single NVIDIA GeForce RTX 2080Ti GPU with 11GB of memory. Under these settings, the proposed method achieves an average tracking speed of **45 FPS**, demonstrating practical runtime performance suitable for real-time deployment across RGB-T, RGB-D, and RGB-E tracking tasks. Additionally, the total number of added parameters for the SBiDF fusion network is **11.7M**.

In our implementation, we set $\lambda_1^X = 1$ and $\lambda_2^X = 1$. Also, we set $\lambda_1^{RGB} = 1$ and $\lambda_2^{RGB} = 1$ for the RGB-T tracking task, while $\lambda_1^{RGB} = 0.5$ and $\lambda_2^{RGB} = 0.5$ for the RGB-D and RGB-E ones. These hyperparameters were selected based on empirical evaluations to strike a balance between tracking accuracy and training efficiency. For the loss weights in (3.19), we set $\lambda_{iou} = 2$ and $\lambda_{l_1} = 5$ following the empirical recommendations in [57]. The complete training and inference settings for the proposed SBiDF framework are summarized in Table 3.1.

Table 3.1: Training and inference settings of the proposed SBiDF framework.

Component	Setting
Template size	128×128
Search size	256×256
Encoder depth	5 layers per modality
Detail feature-scales	F_3, F_5
DCF window (Ω)	3×3
Loss weights	$\lambda_{iou} = 2$, $\lambda_{l_1} = 5$
Optimizer	AdamW
Learning rate	5×10^{-5}
Batch size	10
Training epochs	30
Inference GPU	NVIDIA RTX 2080Ti
Inference speed	45 FPS
Baseline parameters	92.5M
Added parameters	11.7M

3.3.3 Quantitative Results

To fairly evaluate the proposed tracker, we compare its performance with several state-of-the-art trackers. We organize our comparisons into four groups, each summarized in a separate table, to facilitate a structured and comprehensive evaluation. The first group is the most related to our work by including unified multi-modal trackers, which are designed to operate across diverse modalities such as RGB-T, RGB-D, and RGB-E. The second group focuses specifically on thermal-aware (RGB-T) trackers, reflecting the maturity and variety of methods developed for thermal fusion. The third group includes RGB-D trackers that leverage depth information, and the fourth group comprises RGB-E trackers utilizing event-based data. For each group, we selected representative methods covering a wide range of fusion strategies, including feature-level fusion, decision-level fusion, prompt learning, knowledge distillation, and pixel-level fusion. This selection ensures a fair and rigorous comparison with state-of-the-art approaches across different design paradigms and modality combinations.

Table 3.2: Evaluation scores (%) of the compared multi-modal (RGB-T/D/E) trackers as well as the baseline tracker. Best, second best, and third best scores are represented in **red**, **blue**, and **green**, respectively.

Category	Tracker	Year	RGBT234		LasHeR		DepthTrack			VisEvent	
			PR	SR	PR	SR	F-score	Pr	Re	PR	SR
Baseline	OSTrack [57]	2022	72.9	54.9	51.5	41.2	52.9	53.6	52.2	69.5	53.4
Prompt learning	ProTrack [19]	2022	78.6	58.7	50.9	42.1	57.8	58.3	57.3	63.2	47.1
	ViPT [15]	2023	83.5	61.7	65.1	52.5	59.4	59.2	59.6	**75.8**	59.2
	Un-Track [45]	2024	83.7	61.8	66.7	53.6	**61.2**	**61.3**	**61.0**	**76.3**	**59.7**
	OneTracker [62]	2024	**85.7**	**64.2**	**67.2**	**53.8**	60.9	**60.7**	60.4	**76.7**	**60.8**
	EMTrack [64]	2025	81.8	60.1	65.9	53.3	58.3	58.0	58.5	72.4	58.4
	M3Track [65]	2025	84.5	**63.0**	65.8	52.5	57.7	56.6	58.8	73.5	59.2
Knowledge distillation	SDSTrack [23]	2024	84.8	62.5	66.5	53.1	**61.4**	**61.9**	**60.9**	**76.7**	**59.7**
	CMDTrack [63]	2025	**85.9**	61.8	**68.8**	**56.6**	59.8	59.1	60.7	**75.8**	**61.3**
Pixel-level	SBiDF [53] (Proposed)	-	**88.3**	**65.3**	**71.4**	**57.2**	**61.0**	59.8	**62.3**	75.1	**61.3**

3.3.3.1 Comparison with Unified (Multi-modal) Trackers

Table 3.2 presents a comprehensive comparison of the proposed SBiDF framework against recent unified multi-modal trackers evaluated across RGBT234, LasHeR, DepthTrack, and VisEvent datasets. The baseline OSTrack [57], which does not incorporate any auxiliary modalities, achieves modest results on all benchmarks, reflecting the limitations of single-modality tracking in the challenging conditions of these multi-modal benchmarks. Among prompt learning-based approaches, OneTracker [62] and Un-Track [45] achieve strong performance, with OneTracker reaching the best PR (76.7%) on VisEvent. CMDTrack [63], a recent distillation-based tracker, achieves the highest SR (61.3%) on VisEvent, while its counterpart SDSTrack [23] performs competitively, with the top F-score (61.4%) and Precision (61.9%) on DepthTrack.

Compared to these strong trackers, the proposed SBiDF tracker delivers consistently superior or comparable results across all benchmarks. Specifically, SBiDF achieves the highest PR (88.3%) and SR (65.3%) on RGBT234, outperforming the closest competitor CMDTrack by 2.4% and 3.5%, respectively. On LasHeR, SBiDF also ranks first, with 71.4% PR and 57.2% SR, demonstrating its robustness in diverse thermal tracking scenarios. On DepthTrack, SBiDF achieves an F-score of 61.0%, which is only 0.4% lower than the top-performing SDSTrack (61.4%) and higher than all prompt-based approaches. Notably, SBiDF obtains the highest Recall (62.3%), indicating strong ability to maintain target localization across long sequences. On VisEvent, SBiDF achieves 75.1% PR and 61.3% SR, matching or exceeding most recent methods. Although SDSTrack and OneTracker marginally surpass SBiDF in PR on VisEvent, SBiDF demonstrates balanced performance without relying on multi-stage distillation or modality-specific prompt tuning. These results confirm that SBiDF delivers competitive or superior accuracy across all tested modalities and datasets, outperforming fusion-heavy, prompt learning, and distillation frameworks while maintaining a simple, tracker-agnostic pixel-level fusion strategy.

3.3.3.2 Comparison with Thermal-specific (RGB-T) Trackers

Table 3.3 reports results comparing SBiDF with a broad range of state-of-the-art RGB-T trackers evaluated on the RGBT234 and LasHeR datasets. The compared methods span four main categories: feature-level and decision-level fusion, prompt learning, knowledge distillation, and pixel-level fusion. As shown in the table, the proposed SBiDF tracker outperforms all compared RGB-T trackers across all categories on both datasets, except for DKDTrack [21] but only with a 0.2% lower SR on RGBT234. Specifically, SBiDF reaches the highest PR (88.3%) and the second-best SR (65.3%) on RGBT234, while achieving the best PR (71.4%) and SR (57.2%) on LasHeR. Moreover, SBiDF achieves significant improvements compared to all prior pixel-level fusion approaches namely, mfDiMP [25], CSR-DCF [37], DAPFNet [41], DFAT [42], and the proposed HFDAE [52]. These results demonstrate that SBiDF not only matches or exceeds leading prompt learning and distillation methods but also sets a new benchmark among pixel-level fusion trackers on RGB-T datasets. The consistent gains over existing methods validate the effectiveness of SBiDF's adaptive, tracking-oriented fusion design.

3.3.3.3 Comparison with Depth-specific (RGB-D) Trackers

Table 3.4 summarizes the performance of SBiDF compared to specialized RGB-D trackers on the DepthTrack dataset. The recent prompt-based tracker VADT [68] attains the highest F-score

Table 3.3: PR and SR scores (%) of the compared RGB-T trackers on the RGBT234 and LasHeR datasets. Best, second best, and third best scores are represented in **red**, **blue**, and **green**, respectively.

Category	Tracker	Year	RGBT234		LasHeR	
			PR	SR	PR	SR
Feature-level and/or Decision-level	TBSI [5]	2023	87.1	63.7	69.2	55.6
	MACFT [66]	2023	85.7	62.2	65.3	51.4
	MLAAS [10]	2024	79.5	58.4	53.8	43.1
	MELT [12]	2024	75.3	54.6	57.3	45.0
	TransAgg [8]	2024	85.5	63.2	66.7	53.4
	SiamSCR [6]	2024	79.1	57.5	52.2	40.1
	CAT++ [7]	2024	84.0	59.2	50.9	35.6
	STMT [29]	2024	86.5	63.8	67.4	53.7
	LMINet [67]	2024	83.8	59.2	49.0	34.8
Prompt learning	BAT [17]	2024	86.8	64.1	70.2	56.3
	TATrack [18]	2024	87.2	**64.4**	70.2	56.1
	TUFNet [16]	2025	**88.2**	64.1	70.8	55.7
Knowledge distillation	CMD [20]	2023	82.4	58.4	59.0	46.4
	DKDTrack [21]	2025	**88.0**	**65.5**	**70.9**	**56.5**
Pixel-level	mfDiMP* [25]	2019	82.4	58.3	58.3	45.6
	CSR-DCF [37]	2023	70.7	51.4	-	-
	DAPFNet [41]	2023	86.0	62.0	**71.1**	51.3
	DFAT [42]	2023	75.8	55.2	44.6	33.9
	HFDAE [52] (Proposed)	2025	87.4	63.6	**71.1**	**57.0**
	SBiDF [53] (Proposed)	-	**88.3**	**65.3**	**71.4**	**57.2**

(61.0%) and Precision (60.6%), highlighting the benefits of efficient modality conditioning. The proposed SBiDF tracker achieves an F-score of 61.0%, matching the top-performing VADT and outperforming all feature-level fusion-based trackers. SBiDF also reaches 62.3% Recall, the highest among all compared trackers, as well as the second-best Precision (59.8%), indicating superior capability in maintaining accurate localization over long sequences with depth data. Notably, SBiDF matches the overall performance of VADT, despite being designed as a unified, modality-agnostic fusion framework rather than a depth-specific RGB-D tracker. These results confirm that the adaptive pixel-level fusion of SBiDF generalizes effectively beyond RGB-T scenarios and achieves state-of-the-art performance in RGB-D tracking without requiring modality-specific architecture modifications or prompt tuning.

Table 3.4: F-score, Precision and Recall scores (%) of the compared RGB-D trackers on the DepthTrack dataset. Best, second best, and third best scores are represented in **red**, **blue**, and **green**, respectively.

Category	Tracker	Year	DepthTrack		
			F-score	Pr	Re
Prompt learning	VADT [68]	2024	**61.0**	**60.6**	**60.3**
Feature-level	DDiMP [69]	2020	48.5	50.3	46.9
	ATCAIS [69]	2020	47.6	45.5	50.0
	CLGS_D [69]	2020	45.3	**58.4**	36.9
	DeT [30]	2021	53.2	56.0	50.6
	SPT [70]	2023	**53.8**	52.7	**54.9**
	DepthRefiner [71]	2024	51.0	51.3	50.7
	SSLTrack [72]	2024	52.5	56.5	49.1
	CDAAT [73]	2025	**59.0**	57.8	**60.3**
Pixel-level	SBiDF [53] (Proposed)	-	**61.0**	**59.8**	**62.3**

Table 3.5: PR and SR scores (%) of the compared RGB-E trackers on the VisEvent dataset. Best, second best, and third best scores are represented in **red**, **blue**, and **green**, respectively.

Category	Tracker	Year	VisEvent	
			PR	SR
Prompt learning	eMoE-Tracker [74]	2025	**76.4**	**61.3**
Feature-level	MDNet(EF) [75]	2016	66.1	42.6
	ATOM(EF) [76]	2019	60.8	41.2
	SiamRCNN(EF) [77]	2020	65.9	49.9
	LTMU(EF) [78]	2020	65.5	45.9
	PrDiMP(EF) [79]	2020	64.4	45.3
	SiamCAR(EF) [80]	2020	59.9	42.0
	SiamMask(EF) [81]	2023	56.2	36.9
	MMHT [82]	2024	73.4	**55.3**
	TENet [83]	2025	**76.5**	**60.1**
Pixel-level	SBiDF [53] (Proposed)	-	**75.1**	**61.3**

3.3.3.4 Comparison with Event-specific (RGB-E) Trackers

Table 3.5 compares SBiDF against recent RGB-E trackers evaluated on the VisEvent dataset. From the table, the proposed SBiDF tracker achieves the best SR score (61.3%) along with the recent prompt-based eMoE-Tracker [74], outperforming all the other compared trackers with a margin of

1.2% higher than the second-best SR score. Moreover, SBiDF achieves a 75.1% PR score ranking the third-best with a 1.4% lower than the highest PR score achieved by TENet [83]. Notably, SBiDF accomplishes these results without relying on any prompt-specific modules or event-conditioned feature encoders, confirming the strength of its general pixel-level fusion strategy. Such results validate the versatility and robustness of SBiDF in RGB-E tracking, demonstrating that an adaptive, tracker-agnostic fusion framework can achieve competitive performance even in event-based scenarios where temporal resolution and motion cues are critical.

3.3.4 Visual Results

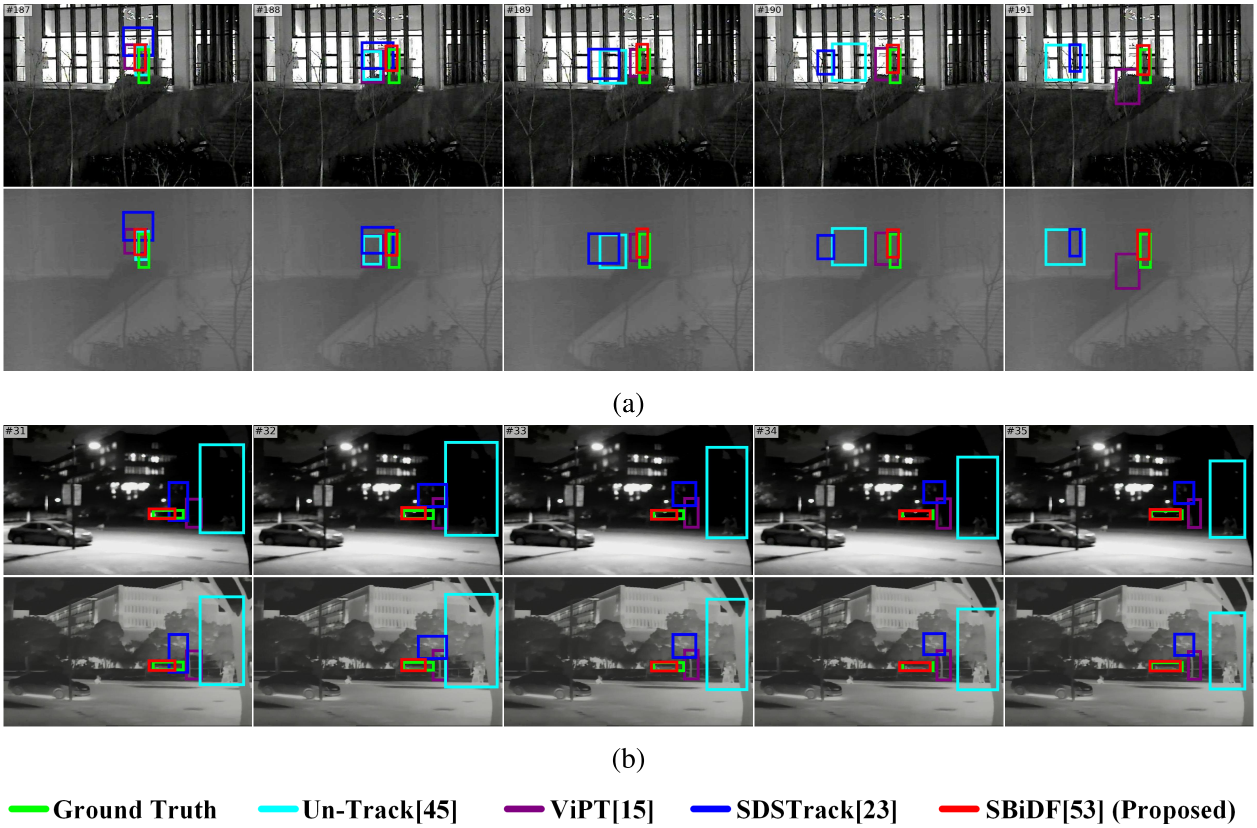
For visual evaluation, Fig. 3.4 and 3.5 present representative examples of tracking results obtained using the proposed tracker and three recent unified multi-modal trackers: ViPT [15], Un-Track [45], and SDSTrack [23]. These examples cover diverse and challenging cases across RGB-T, RGB-D, and RGB-E tracking scenarios. Specifically, in case Fig. 3.4 (a), object details are missing in the TIR image due to reflective surfaces but are captured clearly in the RGB frame. In contrast, cases Fig. 3.4 (b) and Fig. 3.5 (a) through (c) illustrate situations of overexposure (intense light) or underexposure (low light), where RGB images fail to capture critical object information, while the auxiliary modality (thermal, depth, or event) preserves these details.

As shown in the figures, the proposed SBiDF tracker maintains precise target localization closely aligned with the ground truth, even in scenarios where the compared trackers exhibit visible drift or tracking failures. Additionally, SBiDF preserves essential RGB spatial and textural details, effectively preventing misleading cues from the auxiliary modality from degrading the tracking performance.

3.3.5 Ablation Study

Architectural Components Analysis

To assess the contribution of architectural depth and the selection of feature scales in the proposed fusion framework, we conducted an ablation study comparing three SBiDF variants against the

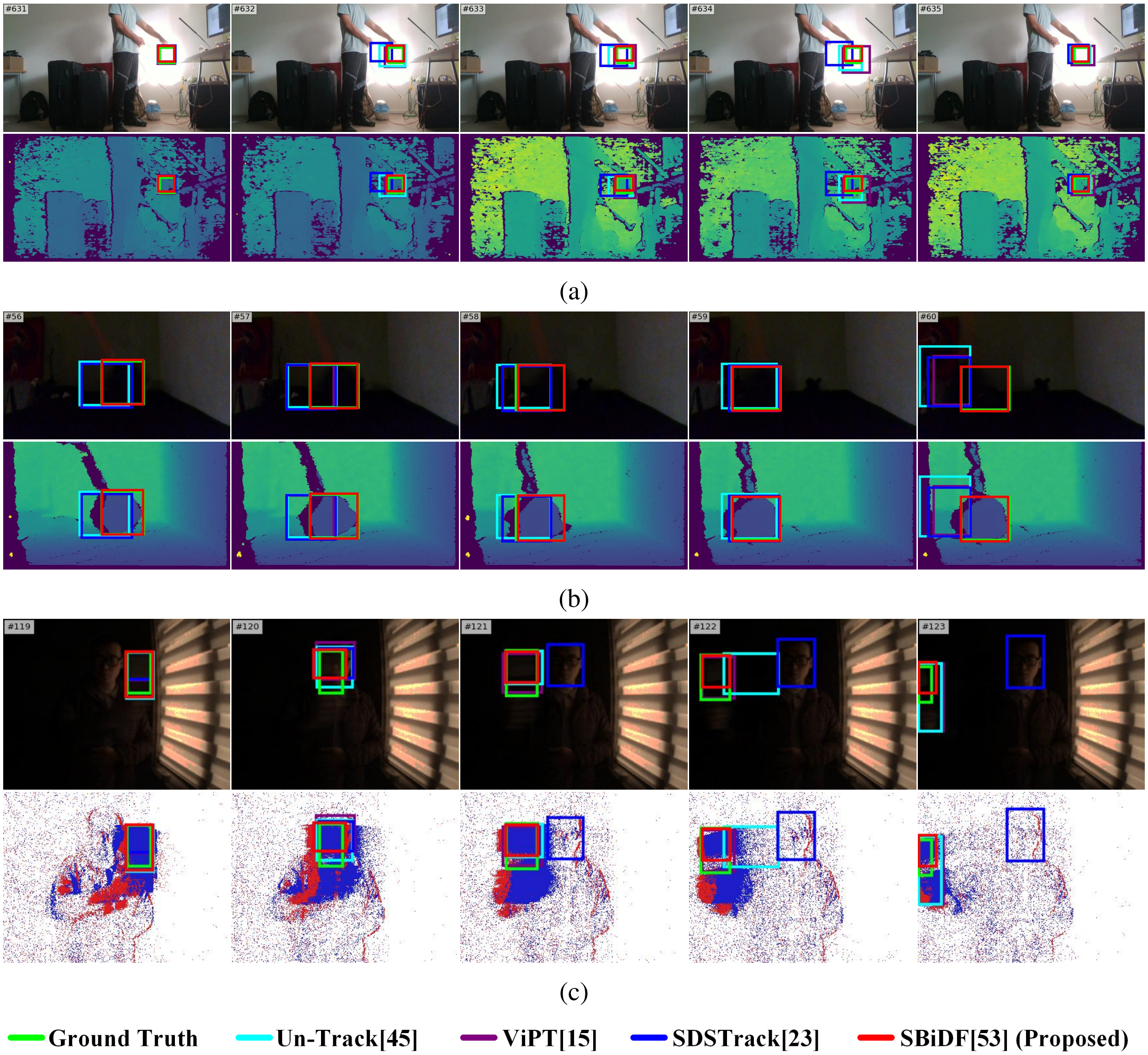


(a)

(b)

Figure 3.4: Examples of the RGB-T tracking results obtained using the compared trackers.

proposed configuration. Table 3.6 summarizes the results on the RGBT234 and LasHeR datasets. The All-Scales variant, which uses all four feature scales (2–5) as detail components, achieves strong results but slightly lower performance compared to the final design, suggesting that including all scales may introduce redundant or less discriminative information that does not consistently benefit tracking. The Shallow-Scale3 configuration, with a three-layer encoder and a single detail scale, demonstrates competitive performance (e.g., 64.0% SR on RGBT234), confirming that even reduced representations can provide useful complementary cues. However, the results remain consistently below the proposed version, indicating that deeper feature hierarchies improve robustness. The Deep-Scale357 variant extends the encoder to seven layers and fuses three detail scales (3, 5, and 7). While this setup achieves comparable results to the All-Scales variant, it does not yield additional gains and slightly reduces PR, potentially due to overfitting or the inclusion of less relevant high-level features. Overall, the proposed configuration, using a five-layer encoder with



(a)

(b)

(c)

Figure 3.5: Examples of the RGB-D and RGB-E tracking results obtained using the compared trackers. Cases in (a) and (b) are for RGB-D tracking, and (c) is for RGB-E tracking.

scales 3 and 5, achieves the best performance across both benchmarks. These results validate our design choice, demonstrating that a balanced selection of encoder depth and intermediate detail scales provides optimal fusion capacity without unnecessary complexity.

Impact of Base Tracker Initialization and Fusion Methods

To gain deeper insights into the robustness of our fusion framework in enhancing tracking performance, we conduct a comprehensive ablation study. Specifically, we analyze the impact of the base

Table 3.6: Ablation study evaluating different encoder depths and detail-scale configurations in the proposed SBiDF framework. The best PR and SR scores (%) for each dataset are highlighted in **bold**.

Variants	RGBT234		LasHeR		Encoder	Detail
	PR	SR	PR	SR	#layers	feature-scales
All-Scales	86.6	63.1	70.5	56.6	5	F_2, F_3, F_4, F_5
Shallow-Scale3	86.1	64.0	70.3	56.4	3	F_3
Deep-Scale357	86.4	63.1	70.0	56.2	7	F_3, F_5, F_7
SBiDF (Proposed)	**88.3**	**65.3**	**71.4**	**57.2**	5	F_3, F_5

tracker's initial performance on the proposed SBiDF tracker by initializing it with the pretrained weights of: (1) OSTrack [57] and (2) DropTrack [60]. Additionally, we compare the SBiDF tracker in each initialization scenario against other pixel-level fusion approaches that are not optimized for tracking. These include weighted sum fusion (WtdSum) and semantic-aware image fusion (SeAFusion) [84]. SeAFusion utilizes segmentation networks to guide fusion through semantic loss. For fair evaluation, we apply the same end-to-end training procedure (detailed in Section 3.2.4) to train WtdSum, SeAFusion, and SBiDF alongside the RGB tracker. Fig. 3.6 presents a histogram of PR scores on RGBT234, comparing the pretrained trackers (OSTrack and DropTrack) with their respective WtdSum-, SeAFusion-, and SBiDF-based variants. As shown, all three fusion methods improve the performance of both pretrained RGB trackers. However, the proposed SBiDF fusion consistently achieves the highest improvement in PR scores (about **11%**), demonstrating its superior ability to leverage complementary modality information. Unlike WtdSum, which blindly combines RGB and TIR data, and SeAFusion, which prioritizes semantic classification over object localization, SBiDF effectively integrates multi-modal features to enhance tracking accuracy.

3.4 Summary

In this chapter, we have introduced a novel symmetric bidirectional dynamic fusion (SBiDF) framework, designed to perform adaptive pixel-level fusion of RGB and auxiliary modalities (thermal, depth, and event) for robust multi-modal object tracking. In this regard, we have first outlined the

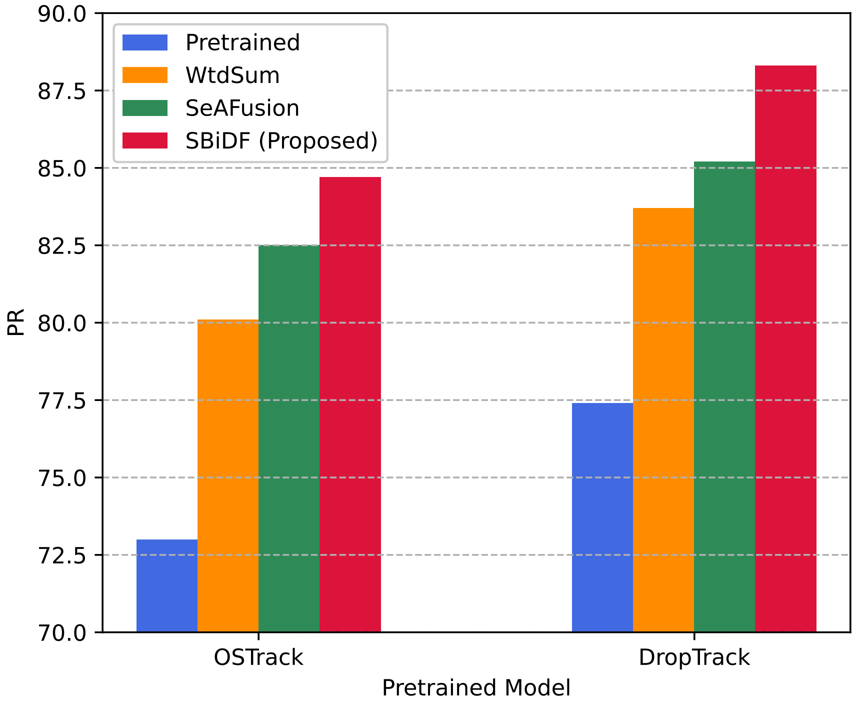


Figure 3.6: Histogram of PR scores calculated over the RGBT234 dataset using the proposed, pretrained, WtdSum-based and SeAFusion [84]-based trackers.

preliminary hierarchical feature difference autoencoder (HFDAE) framework that formed the basis for a more advanced tracker, SBiDF. A detailed description of SBiDF has then been presented. Unlike previous fusion methods that typically require modality-specific adaptations or complex fusion modules, SBiDF directly generates a fused input image optimized explicitly for tracking tasks. The proposed network employs symmetric encoder-decoder structures coupled with novel dynamic convolutional filtering blocks, facilitating effective bi-directional interaction between RGB and auxiliary modality features. Ablation studies have been carried out to validate the effectiveness of our architectural choices. Extensive evaluations on diverse multi-modal tracking benchmarks have demonstrated that SBiDF consistently outperforms or matches the performance of recent state-of-the-art unified trackers as well as the ones that are specific to a single auxiliary modality, including fusion-heavy, prompt learning and knowledge distillation-based trackers, while maintaining a simple and computationally efficient design.

Chapter 4

A Multi-level Self-distillation Framework for Robust RGB-T Object Tracking

4.1 Introduction

In RGB-T tracking systems, most existing trackers rely on extending pretrained RGB trackers using complex fusion modules or modality-specific architectures, sacrificing efficiency for performance. In Chapter 3, we have introduced an alternative solution through a pixel-level fusion framework (SBiDF) that enhances RGB inputs with TIR information prior to tracking, without modifying the underlying tracker. This method demonstrated strong tracking performance and generalization while preserving architectural simplicity, as it required no modifications to the backbone of the RGB tracker. However, it relies on explicit image-level fusion and operates independently of the tracker's internal learning process, limiting its adaptability to changing tracking dynamics and modality-specific uncertainties. Efficiency-driven alternatives have explored knowledge distillation and prompt learning to reduce architectural complexity or added parameters. However, they often require multi-stage training pipeline and external teacher networks, and suffer from the limited capacity of fixed prompts. In this work, we argue that the key to efficient and robust RGB-T tracking lies not in designing increasingly complex fusion mechanisms, but in rethinking how to train a single backbone to natively harness multi-modal data.

In this chapter, we propose the multi-level self-distillation framework (MSD) that transforms a standard RGB tracker into a high-performance RGB-T solution without architectural modifications, added parameters, or external teachers [85]. The key idea behind MSD is that a shared backbone can learn robust cross-modal representations through carefully designed self-supervision and multi-level training signals. Specifically, MSD integrates two novel self-distillation losses: (1) Contrastive feature distillation that aligns semantically similar regions across template-search pairs while repulsing dissimilar ones, enhancing discriminability across the feature hierarchy, and (2) Modality-gap alignment that minimizes the divergence between RGB and TIR features in latent space, ensuring consistent representations under modality-specific perturbations. These losses complement a suite of supervised tracking objectives, including: (1) An intermediate focal loss to strengthen shallow and mid-level features for early target localization, (2) Modality-specific losses applied to lightweight auxiliary heads, preserving unique cues when one modality degrades, and (3) A primary tracking loss for final bounding box prediction from fused features. By unifying these objectives, MSD enables a single backbone to dynamically adapt to RGB, TIR, or fused inputs without modality-specific branches. Additionally, it maintains efficiency by preserving the original model's parameter count and inference speed. Finally, it generalizes across challenges, from thermal crossover to partial occlusions.

MSD is validated on three benchmark datasets, LasHeR [54], RGBT234 [1], and GTOT [86], where it consistently outperforms state-of-the-art methods in accuracy while maintaining the efficiency of a standard RGB tracker. In addition to the main evaluation, we conduct extensive ablation studies to assess the contribution of each loss component in the proposed framework. Furthermore, we examine the tracker's performance under modality-missing scenarios, where only one modality (RGB or TIR) is available. This setup reflects practical challenges in real-world applications such as autonomous driving and surveillance, where sensor failures or adverse environmental conditions may result in missing modality data [87, 88]. In these challenging settings, the proposed MSD tracker exhibits great robustness and maintains the tracking performance, demonstrating its ability to leverage rich internal supervision for resilient cross-modal representation learning.

This chapter is organized as follows. Section 4.2 describes the details and formulation of the proposed MSD framework. Experimental results are presented in Section 4.3 followed by a conclusion in Section 4.4.

4.2 Methodology

This section presents the proposed *multi-level self-distillation* (MSD) unified tracker for RGB-T tracking. The tracker employs a shared backbone to independently extract features from both RGB and TIR inputs. These features are then fused through element-wise summation to form unified feature representations. To enhance learning, the architecture is trained with a set of auxiliary self-distillation losses that provide rich supervision across multiple layers, modalities, and feature distributions. These losses provide internal supervision that improves representation consistency without the need for an external teacher model. In addition to self-distillation, the training is guided by a set of supervised tracking losses. These include an intermediate focal loss that emphasizes shallow and mid-level features to improve early localization, modality-specific losses that retain distinctive information under partial modality degradation, and a fused tracking loss that refines the final bounding box prediction. Together, these losses encourage the model to learn modality-specific cues, develop modality-invariant representations, and improve localization robustness.

This section is organized as follows: We begin with an overview of the network architecture, followed by a detailed description of the multi-level self-distillation components, and conclude with the training methodology.

4.2.1 Network Architecture

The proposed RGB-T tracker adopts a unified architecture built upon a shared transformer backbone to process both RGB and TIR inputs. In contrast to conventional dual-branch designs that

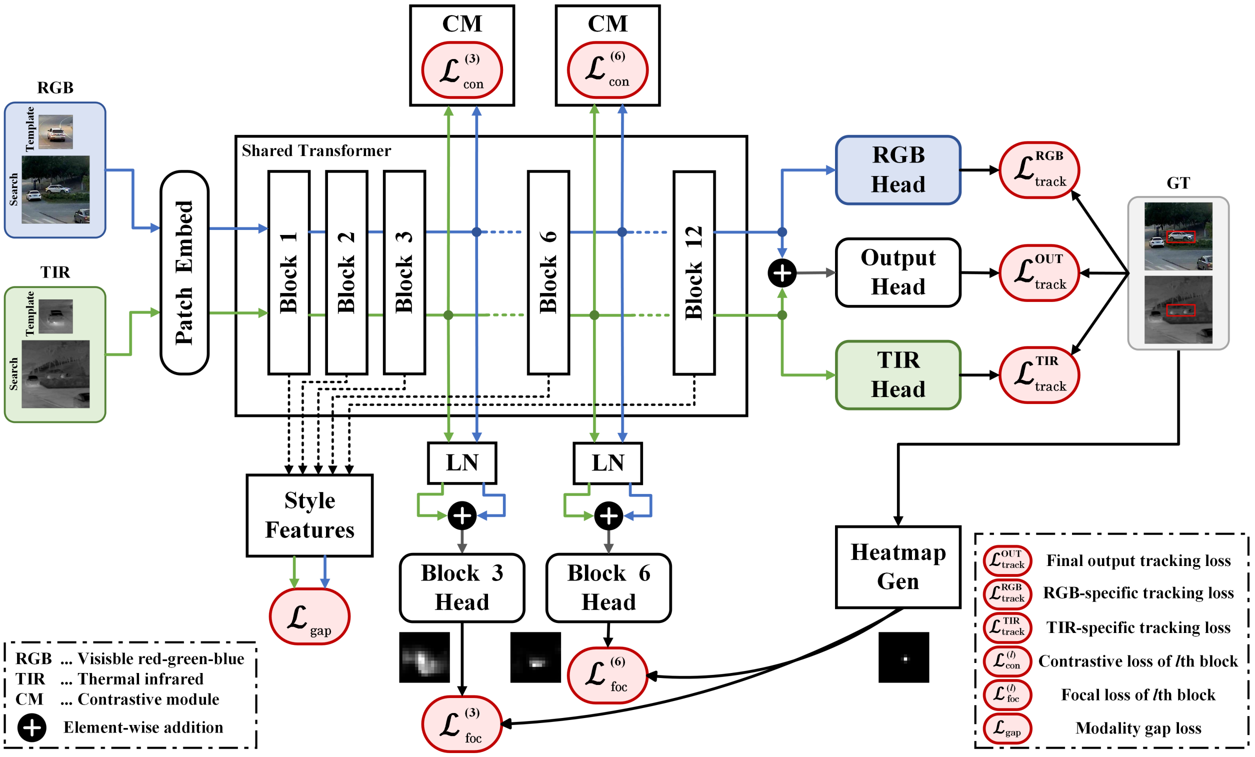

RGB
Template
Search
TIR
Template
Search
Patch Embed
Shared Transformer
Block 1
Block 2
Block 3
Block 6
Block 12
CM
$\mathcal{L}_{\text{con}}^{(3)}$
CM
$\mathcal{L}_{\text{con}}^{(6)}$
LN
LN
Style Features
$\mathcal{L}_{\text{gap}}$
Block 3 Head
Block 6 Head
$\mathcal{L}_{\text{foc}}^{(3)}$
$\mathcal{L}_{\text{foc}}^{(6)}$
RGB Head
Output Head
TIR Head
$\mathcal{L}_{\text{track}}^{\text{RGB}}$
$\mathcal{L}_{\text{track}}^{\text{OUT}}$
$\mathcal{L}_{\text{track}}^{\text{TIR}}$
GT
Heatmap Gen
RGB ... Visisble red-green-blue
TIR ... Thermal infrared
CM ... Contrastive module
Element-wise addition
$\mathcal{L}_{\text{track}}^{\text{OUT}}$ Final output tracking loss
$\mathcal{L}_{\text{track}}^{\text{RGB}}$ RGB-specific tracking loss
$\mathcal{L}_{\text{track}}^{\text{TIR}}$ TIR-specific tracking loss
$\mathcal{L}_{\text{con}}^{(l)}$ Contrastive loss of lth block
$\mathcal{L}_{\text{foc}}^{(l)}$ Focal loss of lth block
$\mathcal{L}_{\text{gap}}$ Modality gap loss

Figure 4.1: Block diagram of the proposed Multi-level Self-Distillation (MSD) network. The architecture comprises a single shared transformer and a tracking head. RGB and TIR template-search pairs are first processed independently through the transformer, and their features are subsequently fused via element-wise summation before being passed to the tracking head for bounding box prediction. To enhance learning, the network incorporates auxiliary modules and complementary loss functions that enable multi-level self-distillation and supervised tracking, guiding feature learning across layers and modalities.

maintain separate pathways for each modality, our approach reuses a single pre-trained RGB backbone to extract features from both RGB and TIR images. This design significantly reduces model complexity and computational overhead while maintaining strong representational capacity.

As illustrated in Fig. 4.1, the network takes as input a pair of template and search images from both the RGB and TIR modalities. Each input image is first embedded using a shared patch embedding layer, and then passed through a unified Vision Transformer (ViT) backbone composed of 12 transformer blocks. Although the RGB and TIR inputs are processed independently, they share weights throughout the backbone, encouraging the learning of modality-common feature representations.

To generate the final fused representation, the features extracted from the last transformer block for both modalities are combined via element-wise summation. These fused features are then fed directly into the original tracking head, which remains unchanged from the base RGB tracker, to produce the final output: the predicted bounding box of the tracked object.

This lightweight and modular design eliminates the need for additional fusion modules or modality-specific adapters. Instead, it leverages a rich set of auxiliary self-distillation and supervised tracking losses (detailed in the next section) to guide the learning process. These losses guide the network to generate aligned, discriminative, and modality-robust feature representations, enabling effective tracking in challenging multi-modal scenarios.

4.2.2 Loss Functions and Multi-level Self-distillation

To enable effective learning within the shared backbone and fully harness the complementary information of RGB and TIR modalities, we propose a comprehensive framework that integrates multi-level self-distillation with supervised tracking losses.

As illustrated in Fig. 4.1, the proposed tracker incorporates five complementary loss components, categorized into two groups: *multi-level self-distillation losses* and *supervised tracking losses*. The multi-level self-distillation losses provide internal supervision to guide the backbone's feature learning by exploiting relationships within the model itself. On the other hand, the supervised tracking losses utilize ground-truth bounding box annotations and serve as key drivers for training the tracker. Specifically,

- Multi-level self-distillation losses include:
  - Contrastive Loss: Encourages early-stage alignment between template and search features. The loss is applied to intermediate layers of the shared transformer backbone. This loss operates by pulling corresponding regions (i.e., the template-search pairs of the same target) closer in the embedding space while pushing apart non-corresponding regions (i.e., different targets or samples). By enforcing this alignment at earlier stages,

the network learns to project semantically consistent representations across views, improving discriminability in the feature hierarchy.

  - Modality Gap Alignment Loss: Minimizes the distributional discrepancy between RGB and TIR features. Thus, it reduces modality-induced variance and facilitates effective feature fusion within the shared backbone.

- Supervised tracking losses include:

  - Intermediate Focal Loss: Applied to shallow and mid-level features to encourage early target localization. This loss helps the network focus on spatially informative regions early in the hierarchy, improving the discriminative quality of intermediate representations.
  - Modality-Specific Tracking Losses: Applied to lightweight tracking heads connected to the final RGB and TIR features. These supervised losses help preserve modality-specific cues and improve robustness under specific modality degradation.
  - Main Tracking Loss: The primary loss used to supervise the final fused RGB-TIR representation. It is computed from the output of the tracking head and directly drives the prediction of the target's bounding box.

Together, these loss components act as complementary supervision signals, helping the network to learn robust, aligned, and generalizable multi-modal features within a unified architecture. Each loss component is described in detail in the following subsections.

4.2.2.1 Multi-level Self-distillation Losses

Contrastive Loss

This loss aligns template and search features at intermediate transformer layers by attracting corresponding template-search regions of the same target and repelling non-corresponding ones, fostering semantic consistency across views.

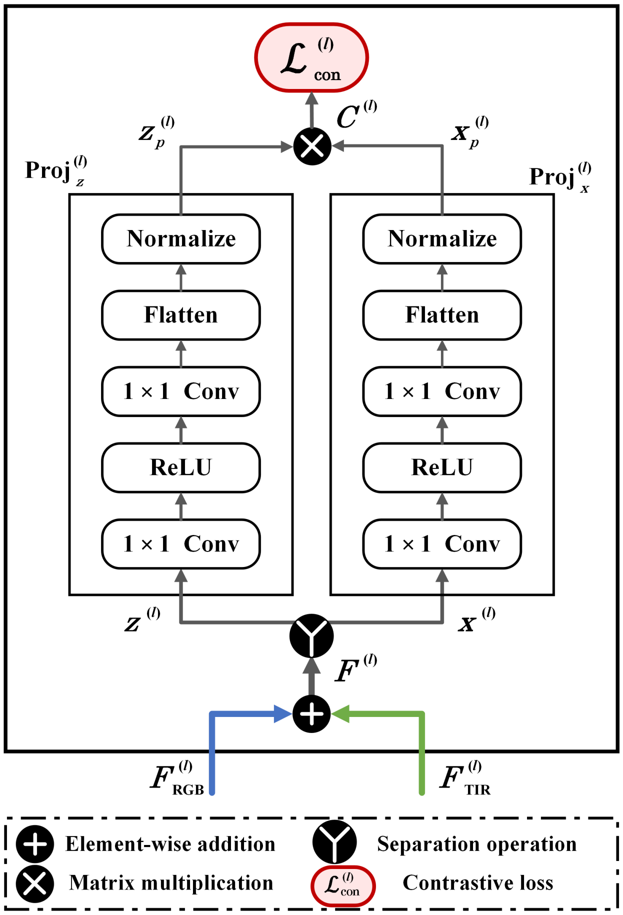


Figure 4.2: Block diagram of the contrastive module (CM). Intermediate features from both RGB and TIR modalities are first fused via element-wise summation, and then separated into template and search branches. Each branch is passed through its own projection head to generate embeddings that are then multiplied to form a similarity matrix, where diagonal elements represent positive pairs and off-diagonal elements represent negative ones.

The process of calculating the contrastive loss is illustrated in Fig. 4.2. Let $\mathbf{F}^{(l)}$ denote the intermediate features from layer $l \in \{3, 6\}$, where $\mathbf{F}^{(l)} = \mathbf{F}_{\text{RGB}}^{(l)} + \mathbf{F}_{\text{TIR}}^{(l)}$. We first split $\mathbf{F}^{(l)}$ into the template and search features, $\mathbf{z}^{(l)}$ and $\mathbf{x}^{(l)}$, respectively. These features are then passed through different projection heads to map them to a lower-dimensional embedding space as

$$\mathbf{z}_p^{(l)} = \text{Proj}_z^{(l)}\left(\mathbf{z}^{(l)}\right), \quad \mathbf{x}_p^{(l)} = \text{Proj}_x^{(l)}\left(\mathbf{x}^{(l)}\right), \tag{4.1}$$

where $\mathbf{z}_p$,$\mathbf{x}_p \in \mathbb{R}^{\mathbb{B}\times\mathbb{D}}$, with $\mathbb{B}$ and $\mathbb{D}$ being the batch size and projection dimension, respectively. Proj denotes a 1×1 convolutional layer followed by ReLU activation and feature normalization, serving as the projection head. The projected features are then used to compute cosine similarities between every template-search pair of samples across the batch to form a similarity matrix as

$$\mathbf{C}^{(l)} = \mathbf{z}_p^{(l)} \left(\mathbf{x}_p^{(l)}\right)^T \in \mathbb{R}^{\mathbb{B}\times\mathbb{B}}. \tag{4.2}$$

As the diagonal elements of $\mathbf{C}$ represent the similarity between corresponding template-search pairs (positive pairs) and the remaining elements are considered as the in-batch negative pairs, the contrastive loss is computed using the InfoNCE [89] formulation as

$$\mathcal{L}_{\text{con}}^{(l)} = -\frac{1}{\mathbb{B}} \sum_{i=1}^{\mathbb{B}} \log \left(\frac{\exp\left(\mathbf{C}_{ii}^{(l)}/\tau\right)}{\sum_{j=1}^{B} \exp\left(\mathbf{C}_{ij}^{(l)}/\tau\right)} \right), \tag{4.3}$$

where τ is a temperature hyper-parameter (typically 0.07). This formulation encourages each template to be most similar to its corresponding search feature, while discouraging similarity to features of other samples within the batch. The contrastive loss is applied at both layers (3 and 6) using separate projection heads for each layer and each pathway (template/search). The final contrastive loss is the sum of both

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{con}}^{(3)} + \mathcal{L}_{\text{con}}^{(6)}. \tag{4.4}$$

Modality-gap Loss

In RGB-T tracking, a fundamental challenge lies in the inherent domain gap between RGB and TIR modalities due to their distinct visual characteristics. Inspired by the recent work in [22], we introduce a modality-gap loss that aligns the “styles” of features extracted from both modalities, thereby promoting feature-level consistency and reducing cross-modal discrepancy.

Specifically, we measure and minimize the discrepancy in first-order (mean) and second-order (standard deviation) statistics of the RGB and TIR features extracted from multiple depths of the shared transformer backbone. The modality-gap loss at layer l is defined as:

$$\mathcal{L}_{\text{gap}}^{(l)} = \frac{1}{\mathbb{B}} \sum_{i=1}^{\mathbb{B}} \left\| \mu_{\text{RGB},i}^{(l)} - \mu_{\text{TIR},i}^{(l)} \right\|_2^2 + \left\| \sigma_{\text{RGB},i}^{(l)} - \sigma_{\text{TIR},i}^{(l)} \right\|_2^2, \tag{4.5}$$

where μ and σ denote the channel-wise mean and standard deviation, respectively, and i indexes the samples in the batch.

This loss is computed across all layers of the transformer backbone (from 1 to 12), and the total modality-gap loss is aggregated as:

$$\mathcal{L}_{\text{gap}} = \sum_{l=1}^{12} \mathcal{L}_{\text{gap}}^{(l)}. \tag{4.6}$$

By minimizing $\mathcal{L}_{\text{gap}}$, the network is encouraged to extract modality-invariant features with similar distributional characteristics across RGB and TIR streams. This facilitates more effective cross-modal fusion and contributes to robust tracking under modality variation.

Please note that, unlike the contrastive loss that promotes instance-level semantic alignment across samples using cross-view contrastive supervision, the modality-gap loss focuses on intra-sample distributional alignment between modalities. While the contrastive loss encourages the network to project semantically similar regions into a shared embedding space, the modality-gap loss ensures that the underlying feature distributions (style) of RGB and TIR are statistically aligned. Together, these complementary objectives enable both robust semantic matching and effective cross-modal fusion.

Combining the contrastive and the modality gap alignment objectives in this multi-level self-distillation framework helps the network to learn early shared semantic representations across modalities and frames, making the features more invariant to modality discrepancies and appearance changes. It also effectively guides shallow and mid-layer features to encode both target identity and cross-modal consistency.

4.2.2.2 Supervised Tracking Losses

Intermediate Focal Loss

To encourage early spatial localization and enhance discriminative feature learning in the shared backbone, we introduce an intermediate focal loss at two layers of the transformer encoder, specifically after the 3rd and 6th blocks. This mechanism supervises intermediate feature representations, enabling the backbone to localize the target before reaching the final prediction stage.

At each selected layer $l \in \{3, 6\}$, the RGB and TIR features are first normalized using LayerNorm and then fused via element-wise summation to form a joint representation:

$$\hat{\mathbf{F}}^{(l)} = \text{LayerNorm}(\mathbf{F}^{(l)}_{\text{RGB}}) + \text{LayerNorm}(\mathbf{F}^{(l)}_{\text{TIR}}). \tag{4.7}$$

This fused feature $\hat{\mathbf{F}}^{(l)}$ is passed through a lightweight tracking head to generate a predicted response map $\hat{\mathbf{S}}^{(l)}$, where each element of the map indicates the confidence of the target's presence at that spatial location. To supervise $\hat{\mathbf{S}}^{(l)}$, we generate a ground-truth heatmap $\mathbf{S}^{\text{GT}}$ centered at the target location, using a Gaussian distribution whose radius is computed as in [58], based on the size of the ground-truth bounding box. The heatmap values decay radially, providing soft supervision around the target center.

The focal loss is computed between $\hat{\mathbf{S}}^{(l)}$ and $\mathbf{S}^{\text{GT}}$ as follows:

$$\mathcal{L}^{(l)+}_{\text{foc}} = \sum_{i=1}^{N^+} \left(1 - \hat{S}^{(l)}_i\right)^{\alpha} \log\left(\hat{S}^{(l)}_i\right), \tag{4.8}$$

$$\mathcal{L}^{(l)-}_{\text{foc}} = \sum_{i=1}^{N^-} \left(\hat{S}^{(l)}_i\right)^{\alpha} \log\left(1 - \hat{S}^{(l)}_i\right) \left(1 - S^{\text{GT}}_i\right)^{\beta}, \tag{4.9}$$

$$\mathcal{L}^{(l)}_{\text{foc}} = -\frac{1}{N^+}\left(\mathcal{L}^{(l)+}_{\text{foc}} + \mathcal{L}^{(l)-}_{\text{foc}}\right), \tag{4.10}$$

where α and β are focusing parameters, N^+ is the number of positive samples (locations where $S^{\text{GT}}_i = 1$), and N^- is the number of negative samples (locations where $S^{\text{GT}}_i < 1$). This formulation prioritizes hard examples among both positives and negatives, improving spatial precision under challenging conditions.

The total intermediate focal loss is defined as:

$$\mathcal{L}_{\text{foc}} = \lambda_{\text{foc}}^{(3)} \mathcal{L}_{\text{foc}}^{(3)} + \lambda_{\text{foc}}^{(6)} \mathcal{L}_{\text{foc}}^{(6)}, \tag{4.11}$$

where $\lambda_{\text{foc}}^{(3)}$ and $\lambda_{\text{foc}}^{(6)}$ are weighting coefficients that balance the contribution of each layer's loss.

By enforcing $\mathcal{L}_{\text{foc}}$ alongside the main tracking task loss (discussed next), we encourage the backbone to progressively refine the target's location across the network depth, resulting in improved stability and robustness in tracking performance.

Modality-specific Tracking

To encourage the network to retain modality-specific information, we introduce two auxiliary tracking heads: an RGB-specific head and a TIR-specific head, as illustrated in Fig. 4.1. Both heads receive the final-layer features from the shared transformer before feature fusion, i.e., $\mathbf{F}_{\text{RGB}}^{(12)}$ and $\mathbf{F}_{\text{TIR}}^{(12)}$, respectively. Each head is responsible for independently predicting the target from its corresponding modality. These predictions provide additional supervision that reinforces modality-awareness and strengthens feature representations in both branches during training.

Main Tracking Loss

The final tracking result is generated from a shared-output head, which receives the summed features $\mathbf{F}_{\text{RGB}}^{(12)} + \mathbf{F}_{\text{TIR}}^{(12)}$ as input. This fused representation captures complementary cues from both modalities, enabling robust and accurate tracking under challenging conditions.

Please note that, the tracking heads (including the fused head and the auxiliary RGB/TIR heads) share the same loss formulation adopted from [57]. It consists of a classification loss ($\mathcal{L}_{cls}$) for object-background discrimination and two regression losses: l_1 loss ($\mathcal{L}_{l_1}$) and generalized IoU loss ($\mathcal{L}_{iou}$) [59] for bounding box prediction. The total tracking loss is given by:

$$\mathcal{L}_{\text{track}} = \mathcal{L}_{cls} + \lambda_{iou} \mathcal{L}_{iou} + \lambda_{l_1} \mathcal{L}_{l_1}, \tag{4.12}$$

where $\lambda_{iou} = 2$ and $\lambda_{l_1} = 5$. The loss is applied to each head independently, and the total tracking supervision during training is the sum of the losses from all three heads.

The proposed supervised tracking losses encourage the backbone to progressively refine the target's location across the network depth. Additionally, the three-head tracking strategy (the fused head and the auxiliary RGB/TIR heads) ensures that the network not only learns to track from the fused representation but also preserves discriminative capabilities for each individual modality. This design regularizes the shared backbone to encode both modality-specific and modality-common cues, enhancing robustness across diverse scenarios. It is important to emphasize that only the shared-output tracking head is used during inference, keeping the test-time pipeline efficient.

4.2.2.3 Total Loss

The final training objective integrates the main tracking loss with all self-distillation objectives. The total loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{track}}^{\text{OUT}} &+ \lambda_{\text{track}} \left(\mathcal{L}_{\text{track}}^{\text{RGB}} + \mathcal{L}_{\text{track}}^{\text{TIR}} \right) \\ &+ \mathcal{L}_{\text{foc}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \lambda_{\text{gap}} \mathcal{L}_{\text{gap}}, \end{aligned} \tag{4.13}$$

where λ_{track}, λ_{con}, and λ_{gap} are balancing coefficients. Here, $\mathcal{L}_{\text{track}}^{\text{OUT}}$ denotes the loss from the final fused-output tracking head, while $\mathcal{L}_{\text{track}}^{\text{RGB}}$ and $\mathcal{L}_{\text{track}}^{\text{TIR}}$ correspond to the modality-specific tracking heads. This comprehensive loss formulation encourages the model to effectively balance modality-specific specialization with modality-invariant representation learning.

4.2.3 Training Procedure

The proposed tracking framework is trained in an end-to-end manner using paired RGB and TIR frames, allowing the model to effectively learn multi-modal representations jointly. Prior to training, the backbone of the RGB tracker is initialized with pre-trained weights obtained from large-scale RGB datasets, ensuring strong visual feature extraction from the outset. This pre-initialization helps preserve the tracker's original capacity while enabling efficient adaptation to multi-modal inputs.

During training, pairs of RGB and TIR frames are cropped into fixed-size template and search regions using ground-truth bounding boxes. The cropped regions from both modalities are then processed sequentially and independently within a single forward pass. Specifically, in the first pass, the RGB template and search regions are split, flattened, and linearly projected before being concatenated and fed into the shared backbone. The TIR template and search regions undergo the same preprocessing steps in a subsequent pass. Alternatively, an equivalent result can be obtained by concatenating all RGB and TIR images along the batch dimension and processing them in a single pass through the shared backbone. In practice, we adopt this latter strategy. The resulting features are then fused via simple element-wise summation, followed by the application of multi-level self-distillation and supervised tracking losses, as shown in Fig. 4.1. All loss components are jointly optimized within a unified training loop. This design maintains full compatibility with the pretrained RGB tracker and avoids any architectural modifications or additional fusion modules.

The model is trained using the AdamW optimizer for 20 epochs. Each epoch consists of 60,000 randomly sampled training pairs, drawn from the RGB-T training-set. To monitor the learning process, validation is performed after each epoch using 10,000 held-out samples, and the mean Intersection over Union (IoU) is computed as the validation metric. The checkpoint yielding the highest validation IoU across all epochs is selected as the final model. Notably, the training pipeline is fully end-to-end and does not require any separate pretraining or teacher networks, ensuring simplicity and efficiency.

During inference, the proposed tracker operates using only the shared backbone and final tracking head, ensuring a streamlined and efficient pipeline. The auxiliary heads and losses modules introduced during training are completely removed such as those used for self-distillation (contrastive and style modules), intermediate supervision (LN, block tracking heads and heatmap generator), and modality-specific learning (RGB- and TIR-specific heads). Only the fused RGB-TIR features (obtained through element-wise summation) are forwarded to the main head for prediction. This design ensures that the additional supervision mechanisms used during training do not introduce any inference-time overhead, allowing the model to maintain its original architecture and tracking speed at test time.

4.3 Experimental Results

This section presents a comprehensive evaluation of the proposed MSD tracker on widely used RGB-T tracking benchmarks. We first introduce the experimental settings, including datasets, evaluation metrics, and implementation details. Then, we report quantitative comparisons with state-of-the-art trackers, followed by ablation studies and efficiency analysis.

4.3.1 Datasets and Evaluation Metrics

To thoroughly evaluate the effectiveness of the proposed MSD framework, we conduct extensive experiments on three widely adopted RGB-T tracking benchmarks: GTOT [86], RGBT234 [1], and LasHeR [54]. The LasHeR dataset provides a large-scale benchmark for RGB-T tracking, comprising a total of 1224 paired RGB and TIR video sequences, which are split into 979 sequences for training and 245 sequences for testing. On the other hand, GTOT is one of the earliest RGB-T tracking datasets and contains 50 video sequences, each with aligned RGB and TIR frames, while RGBT234 is a larger dataset containing 234 video sequences covering more tracking challenges. To facilitate detailed performance analysis, each sequence in RGBT234 is annotated with one or more of 12 challenge attributes that reflect the specific environmental or visual conditions under which the videos were captured such as heavy occlusion, low resolution, thermal crossover, motion blur, and deformation. Table 4.1 lists the attributes annotated in RGBT234. This diverse set of attributes enables a comprehensive evaluation of trackers across a wide range of real-world tracking difficulties.

To evaluate the tracker performance, we follow the official evaluation protocols for each dataset

Table 4.1: List of attributes annotated in the RGBT234 dataset.

Attribute Notation	Attribute Description	No. of sequences
NO	No occlusion	41
PO	Partial occlusion	96
HO	Heavy occlusion (over 80% percentage)	96
LI	Low illumination	63
LR	Low resolution	50
TC	Thermal crossover	28
DEF	Deformation	76
FM	Fast motion (larger than 20 pixels between two adjacent frames)	32
SC	Scale change (scale ratio is out of the range [0.5,1])	120
MB	Motion blur	55
CM	Camera moving	89
BC	Background clutter	54

and report precision rate (PR) and success rate (SR) as the main metrics. PR measures the proportion of frames in which the predicted bounding box is sufficiently close to the ground truth, specifically when the center location error (CLE) falls below a predefined threshold. On the other hand, SR evaluates the tracker's accuracy based on the IoU metric, reporting the percentage of frames where the predicted and ground truth bounding boxes overlap beyond a set threshold. Together, PR and SR offer a balanced assessment of a tracker's accuracy and robustness in predicting object locations and scales.

4.3.2 Implementation Details

In our experiments, we adopt OSTrack [57] as the baseline tracker, utilizing ViT as its backbone. For simplification, we refer to the combination of the proposed MSD framework and OSTrack as the "MSD tracker". We initialized the tracker with the pretrained weights of DropTrack [60], then trained it using the training subset of LasHeR. The proposed MSD tracker is implemented in Python using the PyTorch deep learning library and trained on a computing platform with 32GB of RAM and a NVIDIA V100 GPU [61]. On the other hand, all evaluation experiments are conducted on a system equipped with an Intel Core i9-9900K CPU @ 3.6GHz, 64GB RAM, and a single NVIDIA GeForce RTX 2080Ti GPU with 11GB of memory. Under these settings, our method achieves an

Table 4.2: Settings of the proposed MSD framework.

Component	Setting
Template size	128×128
Search size	256×256
Contrastive loss weights	$\lambda_{\text{con}} = 0.5$
Modality-gap loss weights	$\lambda_{\text{gap}} = 0.01$
Focal loss weights	$\lambda_{\text{foc}}^{(3)} = 0.2$, $\lambda_{\text{foc}}^{(6)} = 0.5$
Tracking loss weights	$\lambda_{\text{track}} = 0.5$, $\lambda_{iou} = 2$, $\lambda_{l_1} = 5$
Optimizer	AdamW
Learning rate	5×10^{-5}
Batch size	10
Training epochs	20
Inference GPU	NVIDIA RTX 2080Ti
Inference speed	52 FPS
Baseline parameters	92.5M
Added parameters	0

average tracking speed of 52 FPS, demonstrating high computational efficiency suitable for real-time applications.

During training, we employ a combination of multiple self-distillation and supervised-tracking loss functions, each weighted to control its contribution to the total loss. The associated loss weights are empirically set as follows: $\lambda_{\text{foc}}^{(3)} = 0.2$, $\lambda_{\text{foc}}^{(6)} = 0.5$, $\lambda_{\text{track}} = 0.5$, $\lambda_{\text{con}} = 0.5$, and $\lambda_{\text{gap}} = 0.01$. These hyperparameters are chosen to strike a balance between supervision strength at different depths and maintaining efficient training dynamics. The settings details of the proposed MSD tracker are summarized in Table 4.2.

4.3.3 Quantitative Comparison

We compare the performance of the proposed RGB-T tracker with several state-of-the-art trackers selected based on three categories: (1) **fusion-based architectures** that extend traditional RGB trackers by incorporating modality-specific pathways, cross-modal attention modules, and fusion mechanisms at the feature or decision levels, (2) **prompt learning** (PL)-based approaches that

Table 4.3: PR and SR scores (%) of the compared RGB-T trackers. Best and second best scores are represented in **red** and **blue**, respectively.

Category	Tracker	Year	Backbone	GTOT		RGBT234		LasHeR	
				PR	SR	PR	SR	PR	SR
Fusion-based architecture	HMFT [13]	2022	ResNet-50	91.2	74.9	78.8	56.8	-	-
	DFMTNet [28]	2023	ResNet-50	-	-	86.2	63.6	65.1	52.0
	TBSI [5]	2023	ViT	-	-	87.1	63.7	69.2	55.6
	MLAAS [10]	2024	ResNet-50	91.3	75.1	79.5	58.4	53.8	43.1
	MELT [12]	2024	ResNet-50	-	-	75.3	54.6	57.3	45.0
	MMFT [9]	2024	ResNet-50	-	-	-	-	69.8	55.9
	TransAgg [8]	2024	Swin-T	92.3	75.1	85.5	63.2	66.7	53.4
	SiamSCR [6]	2024	ResNet-50	91.7	73.7	79.1	57.5	52.2	40.1
	CAT++ [7]	2024	VGG-M	91.5	73.3	84.0	59.2	50.9	35.6
	STMT [29]	2024	ViT	-	-	86.5	63.8	67.4	53.7
Prompt learning	ProTrack [19]	2022	ViT	-	-	78.6	58.7	50.9	42.1
	ViPT [15]	2023	ViT	-	-	83.5	61.7	65.1	52.5
	BAT [17]	2024	ViT	-	-	86.8	64.1	70.2	56.3
	TATrack [18]	2024	ViT	-	-	87.2	64.4	70.2	56.1
	TUFNet [16]	2025	ViT	**93.8**	76.4	88.2	64.1	70.8	55.7
Knowledge distillation	CMD [20]	2023	ResNet-18	89.2	73.4	82.4	58.4	59.0	46.4
	CKD [22]	2024	ViT	93.2	**77.2**	**90.0**	**67.4**	**73.2**	**58.1**
	SDSTrack [23]	2024	ViT	-	-	84.8	62.5	66.5	53.1
	DKDTrack [21]	2025	ViT	-	-	88.0	65.5	70.9	56.5
	AF-tbsi [24]	2025	ViT	-	-	87.3	65.4	69.5	55.9
	AF-bat [24]	2025	ViT	-	-	87.9	64.9	70.6	56.6
	MSD [85] (Proposed)	-	ViT	**94.2**	**78.8**	**90.7**	**67.6**	**72.7**	**58.2**

insert lightweight adapters or prompts into frozen backbones to guide modality interaction without extensive architectural changes, and (3) **knowledge distillation** (KD)-based trackers that rely on external teacher models or self-guided mechanisms for supervising their networks. The selected trackers include MELT [12], MMFT [9], TBSI [5], TransAgg [8], SiamSCR [6], CAT++ [7], CMD [20], DKDTrack [21], CKD [22], SDSTrack [23], AF-tbsi [24], AF-bat [24], ViPT [15], TUFNet [16], BAT [17], TATrack [18], ProTrack [19], HMFT [13], DFMTNet [28], MLAAS [10], and STMT [29]. A summary of the comparison results is provided in Table 4.3.

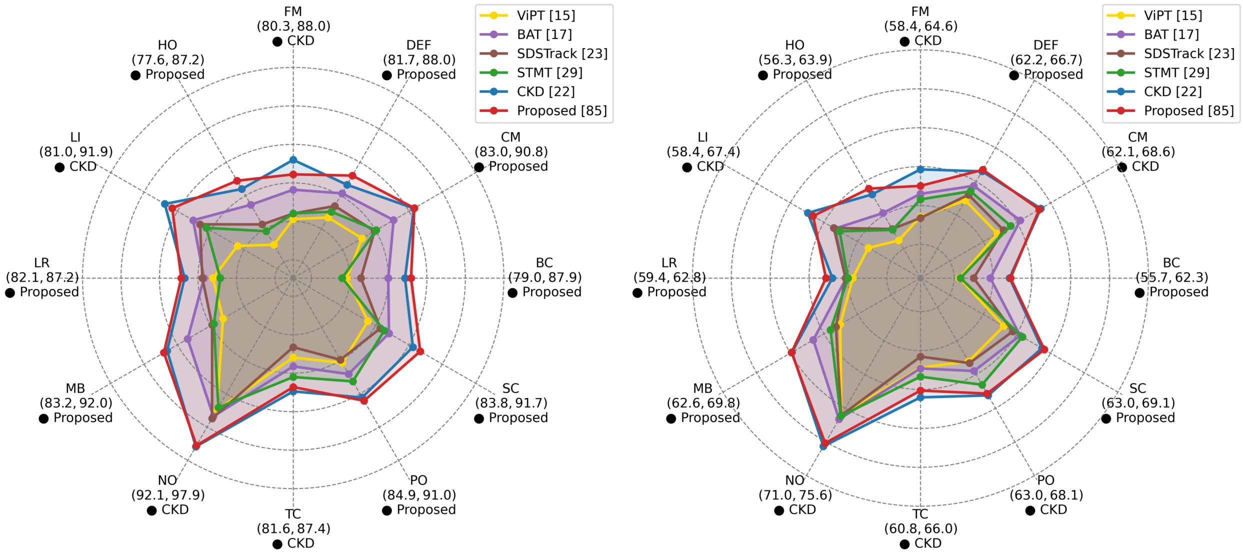
4.3.3.1 Evaluation on GTOT Dataset

The GTOT dataset is a widely used benchmark for RGB-T tracking, known for its relatively balanced modality contribution and moderate-level challenges. As shown in Table 4.3, our proposed tracker achieves a PR of 94.2% and an SR of 78.8%, outperforming all other methods. Compared to the best-performing KD-based tracker CKD (93.2/77.2%) and the best-performing PL-based tracker TUFNet (93.8/76.4%), our tracker demonstrates a significant boost in both precision and overlap. This improvement is primarily attributed to our lightweight and unified architecture, which avoids fusion-heavy branches while benefiting from rich internal supervision through multi-level self-distillation.

4.3.3.2 Evaluation on RGBT234 Dataset

The RGBT234 dataset introduces more diverse object appearances, illumination variations, and background clutter, making it a rigorous benchmark for evaluating cross-modal robustness. Our tracker achieves 90.7% PR and 67.6% SR, slightly outperforming the best KD-based trackers: CKD (90.0/67.4%) and DKDTrack (88.0/65.5%). While several KD-based and transformer-heavy methods like AF-tbsi, AF-bat, BAT, TATrack, and TUFNet also show competitive performance, their architectures rely on modality-specific supervision or cross-prompt modules. In contrast, our model maintains a simpler design without relying on external supervision, yet achieves the best results.

To gain deeper insight into the performance of the proposed tracker under different tracking challenges, we further conduct an attribute-based comparison on the RGBT234 dataset. As illustrated in the radar charts (Fig. 4.3), the proposed tracker consistently ranks among the top across most of the 12 annotated attributes (listed in Table 4.1), particularly excelling in challenging scenarios such as BC (background clutter), MB (motion blur), LR (low resolution), and HO (heavy occlusion). Specifically, our tracker achieves the highest PR on 8 out of 12 attributes and the best SR on 6 attributes, including SC (scale change), PO (partial occlusion), and DEF (deformation),



(a) PR

(b) SR

Figure 4.3: PR and SR scores of the compared RGB-T trackers computed for the different attributes of the RGBT234 dataset. For each attribute, the minimum and maximum scores of all compared trackers are presented, in parenthesis, along with the name of the tracker that achieves the best score.

which are known to be particularly difficult in RGB-T tracking. Compared to other state-of-the-art KD-based trackers like CKD and SDSTrack, the proposed tracker demonstrates more stable and balanced performance, highlighting the effectiveness of the proposed framework in improving target discriminability under varying environmental conditions.

4.3.3.3 Evaluation on LasHeR Dataset

LasHeR is the most challenging benchmark among the three, featuring longer sequences, frequent occlusions, thermal crossover, and low-visibility conditions. On this benchmark, our tracker achieves the highest SR score of 58.2% and the second-best PR score of 72.7%, just behind CKD (73.2/58.1%). Notably, despite CKD being a two-stage KD-based framework with separate teacher-student optimization, our method attains nearly identical accuracy using only a single backbone and joint supervision. Compared to transformer-based fusion networks like TBSI and MMFT, our tracker performs better in overlap metrics, indicating superior localization accuracy in different situations.

These consistent results across all datasets highlight the robustness and efficiency of the proposed design. By unifying modality processing within a shared backbone and integrating a multi-level self-distillation framework alongside supervised tracking objectives, the network effectively enhances cross-modal feature representation. This enables the proposed approach to achieve state-of-the-art performance without relying on complex fusion architectures or external teacher networks.

4.3.4 Ablation Study

Component Analysis

To better understand the contribution of each component in our proposed MSD framework, we conduct a comprehensive ablation study by progressively disabling specific loss terms and evaluating their impact on performance. The ablation variants include:

- **wo/ all**: represents the baseline tracker by removing all self-distillation and tracking supervision losses, retaining only the final output tracking loss ($\mathcal{L}_{\text{track}}^{\text{OUT}}$).
- **wo/ trk**: removes the modality-specific tracking losses ($\mathcal{L}_{\text{track}}^{\text{RGB}}$ and $\mathcal{L}_{\text{track}}^{\text{TIR}}$) while keeping all other losses.
- **wo/ gap**: removes only the modality gap loss ($\mathcal{L}_{\text{gap}}$).
- **wo/ con**: removes only the contrastive loss ($\mathcal{L}_{\text{con}}$).
- **wo/ foc**: removes only the intermediate focal loss ($\mathcal{L}_{\text{foc}}$).

Table 4.4 presents the performance of each variant across the three evaluation datasets. As shown in the table, removing all losses (wo/ all) results in the lowest performance across all datasets, highlighting the importance of the proposed combination of losses. Furthermore, the removal of each individual component leads to a measurable drop in performance. Specifically, removing the modality-specific tracking heads (wo/ trk) slightly degrades performance, particularly

Table 4.4: Ablation study evaluating the impact of removing individual loss components on the performance of the proposed MSD tracker. The best PR and SR scores (%) for each dataset are highlighted in **bold**.

Variants	GTOT		RGBT234		LasHeR	
	PR	SR	PR	SR	PR	SR
wo/ all	92.2	77.6	88.8	65.5	71.4	57.3
wo/ trk	92.3	77.5	90.3	67.1	72.0	57.8
wo/ gap	93.6	78.5	90.5	67.1	72.1	57.8
wo/ con	93.5	**79.1**	89.1	66.5	72.1	57.8
wo/ foc	93.3	78.3	89.8	67.1	72.0	57.7
MSD (Proposed)	**94.2**	78.8	**90.7**	**67.6**	**72.7**	**58.2**

on GTOT and RGBT234, suggesting their role in providing specialized modality supervision. Excluding the modality gap loss (wo/ gap) causes a small but consistent drop in PR and SR, especially on LasHeR, emphasizing its role in aligning the modality distributions.

The contrastive loss (wo/ con) plays a crucial role in localization precision, as its removal results in the largest drop in SR on RGBT234 and LasHeR. Interestingly, the SR on GTOT improves slightly, though PR drops, suggesting that this loss contributes more to fine-grained spatial alignment than to coarse detection, since GTOT contains the images with lowest spatial resolution among the three datasets. Finally, excluding the intermediate focal loss (wo/ foc) also impacts the model's ability to maintain consistent performance across layers. Altogether, the full MSD tracker achieves the best overall performance across all datasets, validating the complementary nature of the proposed losses and the design choice to self-distill information and employ supervised tracking tasks at multiple levels.

In summary, each component of the MSD framework contributes to different aspects of tracking robustness, including modality alignment, feature discrimination, and localization precision. When combined, these components enable the framework to achieve strong generalization across datasets.

Table 4.5: Ablation study evaluating the impact of missing modality data on the performance of the proposed MSD tracker. The best PR and SR scores (%) for each dataset are highlighted in **bold**.

Input modality	RGB				TIR			
Tracker/variant	RGBT234		LasHeR		RGBT234		LasHeR	
	PR	SR	PR	SR	PR	SR	PR	SR
baseline	76.8	37.5	57.4	26.3	74.1	36.9	49.0	25.2
wo/ trk	80.3	58.5	60.5	48.7	70.7	48.4	49.2	39.0
MSD (Proposed)	**83.7**	**61.5**	**64.1**	**51.4**	**78.8**	**54.7**	**58.5**	**45.9**

Modality-missing Analysis

We evaluate the performance of the proposed tracker in scenarios where only a single modality (either RGB or TIR) is available. This setting simulates practical challenges in real-world applications, such as autonomous driving or surveillance, where sensor failures or adverse environmental conditions can result in missing modality data. The evaluation is conducted on the RGBT234 and LasHeR datasets, and both PR and SR scores are reported in Table 4.5. As shown in the table, the baseline tracker (OSTrack [57]), which uses pretrained weights without any modality-specific tracking losses, performs poorly in unimodal settings with SR scores dropping to 37.5/26.3% for RGB and 36.9/25.2% for TIR in RGBT234/LasHeR, respectively. Despite being pretrained on large-scale RGB datasets, the baseline tracker struggles even with RGB-only inputs. This is likely because the RGBT234 and LasHeR datasets are designed such that each modality complements the other, and RGB alone does not offer the same reliability or strength as in conventional RGB-only benchmarks.

In contrast, the proposed tracker achieves significantly higher performance in unimodal settings: 83.7/64.1% PR and 61.5/51.4% SR for RGB-only input, and 78.8/58.5% PR and 54.7/45.9% SR for TIR-only input in RGBT234/LasHeR, respectively. These results highlight the effectiveness of the introduced modality-specific tracking losses, which not only help the model specialize per modality but also enhance the quality of unimodal feature representations. The consistent improvement over the “wo/ trk” variant further emphasizes the contribution of these losses in learning discriminative, modality-aware features even in the absence of cross-modal fusion.

Table 4.6: Total number of parameters, MACs, and inference time per frame of the compared RGB-T trackers.

Tracker	Params (M)	MACs (G)	Time (ms)
TBSI [5]	202.4	82.6	45.6
ViPT [15]	93.4	21.8	26.2
BAT [17]	92.8	56.7	35.9
CMD [20]	19.9	-	-
CKD [22]	183.9	57.9	33.5
SDSTrack [23]	107.8	108.5	80.1
MSD [85] (Proposed)	92.5	56.5	30.3

These findings underscore the proposed tracker's resilience to missing modalities. Unlike traditional dual-branch or fusion-heavy designs, which often assume the presence of both modalities and may require fallback mechanisms or degrade significantly when one stream is lost, the proposed unified-backbone design inherently supports modality dropouts. Thanks to its integrated modality-specific supervision, the tracker can enter a "safe mode" that maintains strong tracking capabilities using only a single modality. This robustness makes the proposed tracker well-suited for deployment in safety-critical environments where sensor reliability cannot be guaranteed.

4.3.5 Efficiency Analysis

Table 4.6 presents a comparative overview of the efficiency of the proposed MSD tracker compared to several state-of-the-art RGB-T trackers. The table particularly reports the total number of parameters, multiply–accumulate operations (MACs), and inference time per frame for each tracker. The inference times of all the compared trackers are measured using the same machine to ensure fair comparison. To further evaluate the performance-efficiency trade-off, we compare our tracker with state-of-the-art RGB-T trackers in terms of accuracy and parameter count. Specifically, Fig. 4.4 plots the average of PR and SR scores computed on the LasHeR dataset versus the number of model parameters (in millions). As shown in the figure, the proposed tracker achieves the second highest overall accuracy (65.45%) among all the compared trackers with only 92.5M parameters and 56.5G MACs with a very slight difference of 0.2% compared to the highest overall accuracy

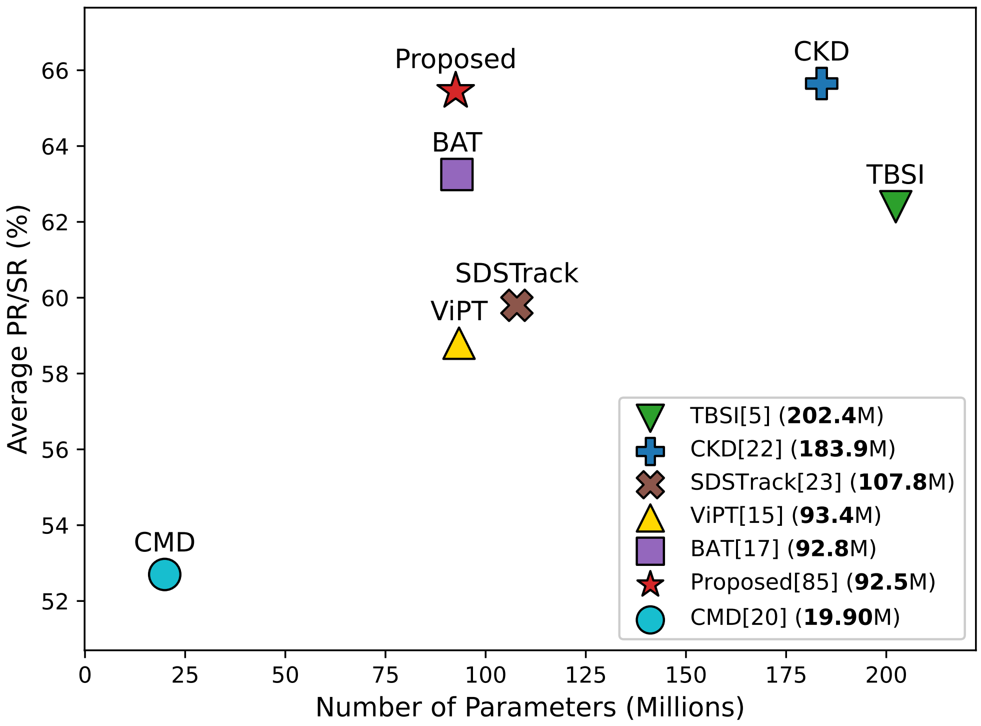


Figure 4.4: Trade-off between model size represented as the number of network parameters and tracking performance computed on the LasHeR dataset. The x-axis indicates the number of parameters (in millions), while the y-axis shows the average of PR and SR scores (%). Each tracker is represented with a distinct marker, while its number of parameters is presented in the legend. The proposed tracker (red star) achieves the best balance between tracking performance and the number of network parameters.

achieved by CKD (65.65%). However, the proposed tracker requires significantly fewer parameters and MACs than CKD (183.9M, 57.9G), whose size is approximately twice as large as the proposed tracker. Moreover, the proposed tracker outperforms other trackers that require more parameters and MACs, such as TBSI (202.4M, 82.6G) and SDSTrack (107.8M, 108.5G), for instance. While CMD is the lightest model (19.9M), it sacrifices a substantial drop in accuracy, demonstrating the limitations of aggressive model compression even in teacher-student KD frameworks. These observations highlight the strong ability of the proposed tracker in balancing between performance and efficiency, thanks to the proposed multi-level self-distillation and tracking supervision framework.

4.4 Summary

In this chapter, we have presented a new perspective on RGB-T tracking by shifting the focus from architectural fusion design to effective training through multi-level self-distillation. Our approach uses a shared transformer backbone and injects diverse self-supervision signals across the network to guide feature learning at multiple depths and modalities. By introducing intermediate focal losses, contrastive alignment, modality-specific tracking supervision, and a style-based modality-gap loss, the network becomes more robust to modality variations and appearance changes without relying on additional branches or external teachers. The proposed framework preserves the structural simplicity of the base tracker while achieving strong gains in tracking accuracy. Quantitative results on multiple RGB-T tracking benchmarks have confirmed that our method achieves a strong performance in both accuracy and efficiency.

Chapter 5

Conclusion

5.1 Concluding Remarks

This thesis has been concerned with the problem of enhancing video tracking of single objects by leveraging the complementary strengths of visible (RGB) and thermal infrared (TIR) imagery. Video tracking is the automated process of detecting, localizing, and continuously monitoring objects of interest throughout consecutive video frames. It serves as a crucial component in advanced vision applications such as surveillance, autonomous driving, augmented reality, robotics, and human-computer interaction. Such applications face significant challenges when relying solely on visible imagery, including poor visibility, illumination variations, occlusions, and appearance changes. Recent approaches employing deep learning typically transform pretrained RGB-based trackers into RGB-Thermal (RGB-T) tracking by introducing specialized modality-specific branches or by employing intricate fusion modules. Such methods often require substantial architectural modifications, significantly increasing the complexity of the trackers, sacrificing efficiency for performance. Alternative approaches introduce lightweight adapters into frozen backbone networks to facilitate modality interaction, which limits their performance due to the constrained adaptability of static prompts. Other approaches employ teacher-student frameworks to distill knowledge from complex, fusion-intensive trackers into more compact models, which demands additional computational resources due to the necessity of maintaining external teacher networks and careful design of distillation objectives, further complicating the overall training process. Consequently, such approaches pose practical challenges for deployment in resource-constrained environments or real-time applications, motivating the need for more streamlined and efficient unified RGB-T tracking frameworks.

In this thesis, the overall objective has been to develop unified RGB-T tracking frameworks to enhance conventional RGB trackers without altering their core network architectures or substantially increasing computational demands. To this end, the proposed research has investigated two distinct solutions for achieving efficient RGB-T tracking: a pixel-level fusion method that enriches the input representation, and a learning-based strategy that adapts the training process through multi-level self-distillation.

In the first part of the thesis, the primary objective was to enhance the quality of RGB images used in visual object tracking by leveraging TIR information in a task-driven, adaptive manner, prior to the tracking process. To address this objective, we have proposed a novel symmetric bidirectional dynamic fusion (SBiDF) framework. Specifically, SBiDF employs modality-specific autoencoders to extract hierarchical representations from both RGB and TIR modalities. These autoencoders consist of encoders that extract multi-scale features and specialized decoders that reconstruct base-level and detailed information at various resolutions. Then, dynamic convolutional filtering (DCF) blocks are utilized to perform adaptive, content-aware, bidirectional enhancement of the extracted representations. The DCF blocks dynamically modulate RGB and TIR detail representations, enhancing each modality based on the features of the complementary modality, thereby enabling balanced refinement across modalities. Extensive experimental evaluations conducted on multiple multi-modal benchmark datasets have demonstrated that the proposed SBiDF tracker significantly improves tracking accuracy without modifying the tracker architecture, surpassing several state-of-the-art RGB-T trackers. Importantly, the generalization capability of SBiDF has been also validated on other auxiliary modalities such as depth and event data, highlighting its versatility and broad applicability in diverse multi-modal tracking scenarios.

The second part of the thesis has aimed to adapt standard RGB trackers for efficient and robust RGB-T tracking through learning-based strategies without introducing additional parameters or architectural modifications. To achieve this, a novel multi-level self-distillation (MSD) framework has been proposed. MSD integrates RGB and TIR modalities within a shared backbone architecture, employing advanced training strategies rather than architectural modifications. Specifically, it leverages self-supervised contrastive learning to enhance the discriminability of cross-modal features, ensuring effective alignment and representation across modalities. Additionally, modality-gap alignment losses have been introduced to reduce discrepancies between RGB and TIR feature distributions. Furthermore, supervised focal losses have been employed to handle foreground-background imbalance effectively, along with modality-specific tracking losses to ensure optimal feature representation for each modality individually. Extensive evaluations conducted on several benchmark datasets has clearly demonstrated that MSD achieves a tracking performance that is superior to that of the existing state-of-the-art trackers, while maintaining the simplicity and computational efficiency of the standard RGB tracker. Therefore, the proposed MSD framework effectively bridges the existing gap between tracking accuracy and computational practicality, making it highly suitable for practical deployment.

In conclusion, this thesis has advanced the current knowledge of RGB-T object tracking by introducing a design philosophy that emphasizes task-driven fusion and architectural simplicity. Unlike existing approaches that often trade efficiency for performance through architectural complexity, in this work high-accuracy RGB-T tracking has been achieved through pre-tracking pixel-level enhancement or through multi-level training strategies without altering the backbone of existing RGB trackers. By proposing SBiDF and MSD as two effective and generalizable paradigms, this thesis has established a foundation for a unified, efficient, and scalable multi-modal tracking.

5.2 Scope for Future Work

While the proposed SBiDF and MSD frameworks have demonstrated promising results for RGB-T tracking, several routes remain open for further exploration and enhancement. Future work could extend the proposed SBiDF approach by investigating strategies utilizing attention mechanisms, potentially enhancing the fusion capability in dynamically changing scenarios. Additionally, the generalization capability of SBiDF to other modalities suggests exploring its applicability to other multi-modal tasks beyond tracking, such as semantic segmentation or action recognition, where early-stage fusion could yield significant benefits. Moreover, the proposed MSD approach can be further extended by incorporating more sophisticated learning strategies to support auxiliary modalities beyond thermal (RGB-T), such as depth (RGB-D) and event data (RGB-E), significantly enhancing its applicability and versatility. Investigating lightweight transformer-based architectures combined with MSD could also be beneficial, aiming at achieving even higher tracking accuracy with reduced computational overhead. Finally, deploying and validating the proposed frameworks in realistic, resource-constrained environments such as embedded platforms or edge devices will be a valuable direction. Evaluating performance metrics such as latency, power consumption, and memory efficiency will ensure the practical feasibility and broader applicability of the developed tracking solutions in real-world multi-modal vision systems.

An especially promising future direction would be to integrate the proposed SBiDF and MSD strategies into a single unified framework. Such an approach could leverage the early-stage cross-modal enhancement capabilities of SBiDF with the advanced training supervision and modality alignment offered by MSD. By combining these two paradigms, it may be possible to construct a robust and highly adaptable RGB-T tracking system that benefits from both enhanced input representations and effective learning dynamics, achieving superior performance without compromising efficiency or architectural simplicity.

References

[1] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019.

[2] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Information Fusion*, vol. 63, pp. 166–187, 2020.

[3] H. Zhang, D. Yuan, X. Shu, Z. Li, Q. Liu, X. Chang, Z. He, and G. Shi, "A comprehensive review of RGBT tracking," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–23, 2024.

[4] M. Feng and J. Su, "RGBT tracking: A comprehensive review," *Information Fusion*, vol. 110, p. 102492, 2024.

[5] T. Hui, Z. Xun, F. Peng, J. Huang, X. Wei, X. Wei, J. Dai, J. Han, and S. Liu, "Bridging search region interaction with template for RGB-T tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 13 630–13 639.

[6] Y. Liu, D. Zhou, J. Cao, K. Yan, and L. Geng, "Specific and collaborative representations Siamese network for RGBT tracking," *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18 520–18 534, 2024.

[7] L. Liu, C. Li, Y. Xiao, R. Ruan, and M. Fan, "RGBT tracking via challenge-based appearance disentanglement and interaction," *IEEE Transactions on Image Processing*, vol. 33, pp. 1753–1767, 2024.

[8] M. Feng and J. Su, "RGBT image fusion tracking via sparse trifurcate transformer aggregation network," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

[9] H. Shi, X. Mu, D. Shen, and C. Zhong, "Learning a multimodal feature transformer for RGBT tracking," *Signal, Image and Video Processing*, pp. 1–12, 2024.

[10] M. Feng and J. Su, "Learning multi-layer attention aggregation siamese network for robust RGBT tracking," *IEEE Transactions on Multimedia*, vol. 26, pp. 3378–3391, 2024.

[11] C. Guo, D. Yang, C. Li, and P. Song, "Dual siamese network for RGBT tracking via fusing predicted position maps," *The Visual Computer*, vol. 38, no. 7, pp. 2555–2567, 2022.

[12] Z. Tang, T. Xu, X.J. Wu, and J. Kittler, “Multi-level fusion for robust RGBT tracking via enhanced thermal representation,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 10, Oct. 2024. [Online]. Available: https://doi.org/10.1145/3678176

[13] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, “Visible-thermal UAV tracking: A large-scale benchmark and new baseline,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 8886–8895.

[14] Z. Tang, T. Xu, X. Wu, X.F. Zhu, and J. Kittler, “Generative-based fusion mechanism for multi-modal tracking,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5189–5197, Mar. 2024.

[15] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, “Visual prompt multi-modal tracking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9516–9526.

[16] Y. Liu, Z. Gao, Y. Cao, and D. Zhou, “Two-stage unidirectional fusion network for RGBT tracking,” *Knowledge-Based Systems*, vol. 310, p. 112983, 2025.

[17] B. Cao, J. Guo, P. Zhu, and Q. Hu, “Bi-directional adapter for multimodal tracking,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 927–935.

[18] H. Wang, X. Liu, Y. Li, M. Sun, D. Yuan, and J. Liu, “Temporal adaptive RGBT tracking with modality prompt,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5436–5444.

[19] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, “Prompting for multi-modal tracking,” in *Proceedings of the 30th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2022, p. 3492–3500. [Online]. Available: https://doi.org/10.1145/3503161.3547851

[20] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, “Efficient RGB-T tracking via cross-modality distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8 2023, pp. 5404–5413.

[21] F. Hong, M. Wen, A. Lu, and Q. Wang, “DKDTrack: dual-granularity knowledge distillation for RGBT tracking,” in *Sixteenth International Conference on Graphics and Image Processing (ICGIP 2024)*, vol. 13539. SPIE, 2025, pp. 837–844.

[22] A. Lu, J. Zhao, C. Li, Y. Xiao, and B. Luo, “Breaking modality gap in RGBT tracking: Coupled knowledge distillation,” in *Proceedings of the 32nd ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2024, p. 9291–9300. [Online]. Available: https://doi.org/10.1145/3664647.3680878

[23] X. Hou, J. Xing, Y. Qian, Y. Guo, S. Xin, J. Chen, K. Tang, M. Wang, Z. Jiang, L. Liu *et al.*, “SDSTrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 551–26 561.

[24] F. Hong, J. Wang, A. Lu, and Q. Wang, “Augmentative fusion network for robust RGBT tracking,” in *Sixteenth International Conference on Graphics and Image Processing (ICGIP 2024)*, vol. 13539. SPIE, 2025, pp. 828–836.

[25] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. Shahbaz Khan, “Multimodal fusion for end-to-end RGB-T tracking,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[26] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, “Be your own teacher: Improve the performance of convolutional neural networks via self distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[27] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, “Challenge-aware RGBT tracking,” in *European conference on computer vision*, 2020, pp. 222–237.

[28] M. Li, P. Zhang, M. Yan, H. Chen, and C. Wu, “Dynamic feature-memory transformer network for RGBT tracking,” *IEEE Sensors Journal*, vol. 23, no. 17, pp. 19 692–19 703, 2023.

[29] D. Sun, Y. Pan, A. Lu, C. Li, and B. Luo, “Transformer RGBT tracking with spatio-temporal multimodal tokens,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 12 059–12 072, 2024.

[30] S. Yan, J. Yang, J. Käpylä, F. Zheng, A. Leonardis, and J.K. Kämäräinen, “DepthTrack: Unveiling the power of RGBD tracking,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 725–10 733.

[31] B. Li, Y. Zhuge, S. Jiang, L. Wang, Y. Wang, and H. Lu, “3D prompt learning for RGB-D tracking,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2024, pp. 2527–2544.

[32] B. Xu, Y. Xu, R. Hou, J. Bei, T. Ren, and G. Wu, “RGB-D tracking via hierarchical modality aggregation and distribution network,” in *Proceedings of the 5th ACM International Conference on Multimedia in Asia*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–7.

[33] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, “VisEvent: Reliable object tracking via collaboration of frame and event flows,” *IEEE Transactions on Cybernetics*, vol. 54, no. 3, pp. 1997–2010, 2024.

[34] J. Zhang, Y. Wang, W. Liu, M. Li, J. Bai, B. Yin, and X. Yang, “Frame-event alignment and fusion network for high frame rate tracking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9781–9790.

[35] T. Zhang, K. Debattista, Q. Zhang, G. Ding, and J. Han, “Revisiting motion information for RGB-Event tracking with MOT philosophy,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 89 346–89 372.

[36] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, “RGB-D salient object detection via 3D convolutional neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1063–1071.

[37] W. Liu, W. Liu, and Y. Sun, “Visible–infrared dual-sensor fusion for single-object tracking,” *IEEE Sensors Journal*, vol. 23, no. 4, pp. 4118–4128, 2023.

[38] X. Zhang, P. Ye, D. Qiao, J. Zhao, S. Peng, and G. Xiao, “Object fusion tracking based on visible and infrared images using fully convolutional siamese networks,” in *2019 22th International Conference on Information Fusion (FUSION)*, 2019, pp. 1–8.

[39] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.H.S. Torr, “Fully-convolutional siamese networks for object tracking,” in *Computer Vision – ECCV 2016 Workshops*, 2016, pp. 850–865.

[40] M. Rasoulidanesh, S. Yadav, S. Herath, Y. Vaghei, and S. Payandeh, “Deep attention models for human tracking using RGBD,” *Sensors*, vol. 19, no. 4, 2019.

[41] H. Zheng, N. Yuan, H. Ding, P. Hu, and Z. Yang, “Thermal infrared and visible sequences tracking via dual adversarial pixel fusion,” *Multimedia Tools and Applications*, pp. 1–20, 2023.

[42] Z. Tang, T. Xu, H. Li, X.J. Wu, X. Zhu, and J. Kittler, “Exploring fusion strategies for accurate RGBT visual object tracking,” *Information Fusion*, vol. 99, p. 101881, 2023.

[43] H. Li, X.J. Wu, and J. Kittler, “MDLatLRR: A novel decomposition method for infrared and visible image fusion,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4733–4746, 2020.

[44] J. Zheng, B. Lin, C. Xue, and Q. Feng, “Robust RGB-T object tracking via multilevel fusion and motion-cue-based correction,” *Signal, Image and Video Processing*, vol. 19, no. 9, pp. 1–8, 2025.

[45] Z. Wu, J. Zheng, X. Ren, F.A. Vasluianu, C. Ma, D.P. Paudel, L. Van Gool, and R. Timofte, “Single-model and any-modality for video object tracking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 19 156–19 166.

[46] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[47] Y. Wan, W. Zhang, Z. Li, H. Zhang, and Y. Li, “Dual knowledge distillation for neural machine translation,” *Computer Speech & Language*, vol. 84, p. 101583, 2024.

[48] Y. Zhou, C. Yang, Y. Li, L. Huang, Z. An, and Y. Xu, “Online relational knowledge distillation for image classification,” in *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2024, pp. 365–370.

[49] Z. Zhao, Z. Yan, M. She, and G. Chen, “Process knowledge distillation for multi-person pose estimation,” in *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023)*, vol. 12707. SPIE, 2023, pp. 659–664.

[50] W. Zhu, B. Peng, and W.Q. Yan, “Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification,” *IEEE Transactions on Multimedia*, vol. 26, pp. 7359–7371, 2024.

[51] T.B. Xu and C.L. Liu, “Data-distortion guided self-distillation for deep neural networks,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 5565–5572.

[52] M. Awad, A. Elliethy, M.O. Ahmad, and M.N.S. Swamy, “Adaptive hierarchical feature difference auto-encoder for robust RGB-T object tracking,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, September 2025.

[53] M. Awad, A. Elliethy, M.O. Ahmad, and M.N.S. Swamy, “SBiDF: A pixel-level bidirectional fusion framework for unified multi-modal object tracking,” *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, 2025, under review.

[54] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun, “LasHeR: A large-scale high-diversity benchmark for RGBT tracking,” *IEEE Transactions on Image Processing*, vol. 31, pp. 392–404, 2022.

[55] X. Wang, X. Shu, S. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu, “MFGNet: Dynamic modality-aware filter generation for RGB-T tracking,” *IEEE Transactions on Multimedia*, vol. 25, pp. 4335–4348, 2023.

[56] J. Peng, H. Zhao, and Z. Hu, “Dynamic fusion network for RGBT tracking,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 3822–3832, 2023.

[57] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, “Joint feature learning and relation modeling for tracking: A one-stream framework,” in *Computer Vision – ECCV 2022*, 2022, pp. 341–357.

[58] H. Law and J. Deng, “CornerNet: Detecting objects as paired keypoints,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[59] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[60] Q. Wu, T. Yang, Z. Liu, B. Wu, Y. Shan, and A.B. Chan, “DropMAE: Masked autoencoders with spatial-attention dropout for tracking tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 561–14 571.

[61] S. Mokhov, G. Roper, C.A. Meza, F. Salhany, and et al,, “Speed: Gina Cody School HPC Facility: Scripts, Tools, and Refs,” accessed on 2025-07-22. [Online]. Available: https://github.com/NAG-DevOps/speed-hpc

[62] L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen, and W. Zhang, “OneTracker: Unifying visual object tracking with foundation models and efficient tuning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 19 079–19 091.

[63] T. Zhang, Q. Zhang, K. Debattista, and J. Han, “Cross-modality distillation for multi-modal tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 7, pp. 5847–5865, 2025.

[64] C. Liu, Z. Guan, S. Lai, Y. Liu, H. Lu, and D. Wang, "EMTrack: Efficient multimodal object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 3, pp. 2202–2214, 2025.

[65] Z. Tang, T. Xu, X.J. Wu, and J. Kittler, "M3Track: Meta-prompt for multi-modal tracking," *IEEE Signal Processing Letters*, vol. 32, pp. 1705–1709, 2025.

[66] Y. Luo, X. Guo, M. Dong, and J. Yu, "Learning modality complementary features with mixed attention mechanism for RGB-T tracking," *Sensors*, vol. 23, no. 14, 2023.

[67] J. Mei, J. Zhou, J. Wang, J. Hao, D. Zhou, and J. Cao, "Learning multifrequency integration network for RGBT tracking," *IEEE Sensors Journal*, vol. 24, no. 9, pp. 15 517–15 530, 2024.

[68] G. Zhang, Q. Liang, Z. Mo, N. Li, and B. Zhong, "Visual adapt for RGBD tracking," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 9391–9395.

[69] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.K. Kämäräinen, M. Danelljan, L.Č. Zajc, A. Lukežič, O. Drbohlav *et al.*, "The eighth visual object tracking VOT2020 challenge results," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 547–601.

[70] X.F. Zhu, T. Xu, Z. Tang, Z. Wu, H. Liu, X. Yang, X.J. Wu, and J. Kittler, "RGBD1K: A large-scale dataset and benchmark for RGB-D object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3870–3878.

[71] S. Lai, D. Wang, and H. Lu, "DepthRefiner: Adapting RGB trackers to RGBD scenes via depth-fused refinement," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[72] X.F. Zhu, T. Xu, S. Atito, M. Awais, X.J. Wu, Z. Feng, and J. Kittler, "Self-supervised learning for RGB-D object tracking," *Pattern Recognition*, vol. 155, p. 110543, 2024.

[73] X.F. Zhu, T. Xu, X.J. Wu, Z. Feng, and J. Kittler, "Adaptive colour-depth aware attention for RGB-D object tracking," *IEEE Signal Processing Letters*, vol. 32, pp. 1680–1684, 2025.

[74] Y. Chen and L. Wang, "eMoE-Tracker: Environmental MoE-based transformer for robust event-guided object tracking," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1393–1400, 2025.

[75] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4293–4302.

[76] M. Danelljan, G. Bhat, F.S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 4660–4669.

[77] P. Voigtlaender, J. Luiten, P.H. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 6578–6588.

[78] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, "High-performance long-term tracking with meta-updater," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 6298–6307.

[79] M. Danelljan, L.V. Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 7183–7192.

[80] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese fully convolutional classification and regression for visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 6269–6277.

[81] W. Hu, Q. Wang, L. Zhang, L. Bertinetto, and P.H. Torr, "SiamMask: A framework for fast online object tracking and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3072–3089, 2023.

[82] H. Sun, R. Liu, W. Cai, J. Wang, Y. Wang, H. Tang, Y. Cui, D. Yao, and D. Guo, "Reliable object tracking by multimodal hybrid feature extraction and transformer-based fusion," *Neural Networks*, vol. 178, p. 106493, 2024.

[83] P. Shao, T. Xu, Z. Tang, L. Li, X.J. Wu, and J. Kittler, "TENet: targetness entanglement incorporating with multi-scale pooling and mutually-guided fusion for RGB-E object tracking," *Neural Networks*, vol. 183, p. 106948, 2025.

[84] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28–42, 2022.

[85] M. Awad, A. Elliethy, M.O. Ahmad, and M.N.S. Swamy, "A multi-level self-distillation-based unified tracker for efficient RGBT tracking," *IEEE Transactions on Image Processing (TIP)*, 2025, under review.

[86] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.

[87] A. Lu, C. Li, J. Zhao, J. Tang, and B. Luo, "Modality-missing RGBT tracking: Invertible prompt learning and high-quality benchmarks," *International Journal of Computer Vision*, pp. 1–21, 2024.

[88] A. Rathinam, L. Pauly, A.E.R. Shabayek, W. Rharbaoui, A. Kacem, V. Gaudillière, and D. Aouada, "Hybrid attention for robust RGB-T pedestrian detection in real-world conditions," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 319–326, 2025.

[89] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019. [Online]. Available: https://arxiv.org/abs/1807.03748