Lightweight RGB-T Object Tracking with Mobile Vision Transformers

Mahdi Falaki

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

August 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mahdi Falaki**

Entitled: **Lightweight RGB-T Object Tracking with Mobile Vision Transformers**

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

Master of Applied Science (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. Krzysztof Skonieczny

_____ External Examiner
Dr. Charalambos Poullis

_____ Examiner
Dr. Krzysztof Skonieczny

_____ Supervisor
Dr. Maria Amer

Approved by _____
Dr. Jun Cai, Chair
Department of Electrical and Computer Engineering

_____ 2025

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Lightweight RGB-T Object Tracking with Mobile Vision Transformers

Mahdi Falaki

Single-modality object tracking (e.g., RGB-only) encounters difficulties in challenging imaging conditions, such as low illumination and adverse weather conditions. To solve this, multimodal tracking (e.g., RGB-T models) aims to leverage complementary data such as thermal infrared features. While recent Vision Transformer-based multimodal trackers achieve strong performance, they are often computationally expensive due to large model sizes. In this work, we propose a novel lightweight RGB-T tracking algorithm based on Mobile Vision Transformers (MobileViT). Our tracker introduces a progressive fusion framework that jointly learns intra-modal and inter-modal interactions between the template and search regions using separable attention. This design produces effective feature representations that support more accurate target localization while achieving a small model size and fast inference speed. Compared to state-of-the-art efficient multimodal trackers, our model achieves comparable accuracy while offering significantly lower parameter counts (less than 4 million) and the fastest GPU inference speed of 122 frames per second. This thesis is the first to propose a tracker using Mobile Vision Transformers for RGB-T tracking and multimodal tracking at large.

Acknowledgments

I would like to express my deepest gratitude to Professor Amer for his invaluable guidance and support throughout the course of this research.

I am also thankful to my previous labmates, Milad and Goutam, for their support. In particular, I would like to thank Goutam for his outstanding contributions to the field. His work on SMAT has been a major source of inspiration and a foundation for much of what I have learned.

Finally, I owe my deepest thanks to my family, whose unconditional love, patience, and encouragement have always been with me. Their unwavering support has been the foundation that made this work possible.

Contents

List of Figures viii

List of Tables x

1 Introduction 1

1.1 Background and Motivation . . . 1

1.2 Limitations of Existing Multimodal Models . . . 2

1.3 Proposed Solution Overview . . . 2

1.4 Datasets and Evaluation Protocol . . . 3

1.4.1 Datasets . . . 3

1.4.2 Evaluation Metrics . . . 4

1.5 Experimental Scope and Architectural Variants . . . 5

1.6 Thesis Statement . . . 5

1.7 Contributions . . . 5

1.8 Thesis Organization . . . 6

2 Literature Review 8

2.1 From RGB-only to Multimodal Tracking . . . 8

2.2 Design Choices in Multimodal Trackers . . . 9

2.3 Gaps in RGB-T Efficiency and Insights from RGB-only Solutions . . . 11

2.4 Recent Efficient RGB-T Trackers . . . 13

3 Methodology **15**

3.1 Overview . . . 15

3.2 MobileViTv2-based Multimodal Backbone . . . 16

3.3 Neck Module . . . 21

3.4 Proposed Cross-Modal Fusion Transformer . . . 22

3.5 Prediction Head . . . 23

3.6 Training Loss Function . . . 24

4 Experimental Results **26**

4.1 Results . . . 26

4.1.1 Implementation Details . . . 26

4.1.2 Evaluation Scope and Benchmarks . . . 27

4.1.3 Datasets . . . 28

4.1.4 Evaluation Metrics . . . 29

4.1.5 Comparison to Related Work . . . 30

4.1.6 Comparison with Baseline Architecture . . . 32

4.2 Ablation Studies . . . 33

4.2.1 Visual Results . . . 33

4.2.2 Effect of Progressive Intra- and Inter-Modal Fusion . . . 36

4.2.3 Attribute-Based Performance Analysis . . . 38

4.2.4 Ablation Study on Final Fusion Strategies . . . 40

5 Analysis **42**

5.1 Overview . . . 42

5.2 Prompt Learning for RGB-T Tracking . . . 42

5.2.1 Motivation for Prompt Learning . . . 43

5.2.2 Pretrained Backbone and Prompt Learning Setup . . . 43

5.2.3 Prompt Integration into Separable Attention . . . 43

5.2.4 Iterative Prompt Refinement Across Layers . . . 45

5.2.5 Prompt Learning Results on LasHeR . . . 46

5.3 Siamese-Based Tracker Architecture 47
5.3.1 Analysis of Siamese Tracker Results 49

6 Conclusion and Future Work **52**

6.1 Conclusion 52
6.2 Future Work 53

7 Appendix A: Experiment on RGB-D **55**

Bibliography **57**

List of Figures

Figure 3.1 The pipeline of proposed RGB-T tracker. MV2 stands for MobileNetV2 (Inverted Residual blocks) and mmMobileViT for multimodal MobileViT (see Figure 3.2). $\downarrow 2$ indicates spatial downsampling by 2. $\{X^{\mathrm{IR}}, Z^{\mathrm{IR}}\}$ show the input search and template frames of Thermal Infrared Modality (IR). $\times 3$ shows the number of subsequent MV2 blocks in layer_2. 16

Figure 3.2 The detailed Architecture of the proposed mmMobileViT with demonstration of intra-modal and inter-modal separable attention in layer_3 and layer_4, respectively. L are the number of transformer layers at mmMobileViT blocks in layers 3 and 4. 17

Figure 3.3 Visual comparison between standard multi-head self-attention (MHA) and separable self-attention (SSA). While MHA performs dense pairwise interactions between all tokens, SSA avoids this by using a softmax-normalized query and efficient per-token weighting, leading to significantly reduced computational complexity. 20

Figure 4.1 Trade-off between parameter count and inference speed (FPS) for lightweight RGB-T trackers. Our tracker (top-left) achieves the best combination of high FPS and low parameter count. 31

Figure 4.2 Architecture of SMAT (Gopal and Amer, 2024), a strong RGB-only baseline. Our RGB-T tracker extends this design with progressive modality interaction in the backbone and a dedicated cross-fusion transformer after the neck, as described in Sections 3.2 and 3.4. 33

Figure 4.3 Evolution of architectures from MobileViTv2 (top) for classification/detection, to SMAT (middle) for RGB tracking with search/template inputs, and finally to our RGB-T tracker (bottom) with progressive fusion and cross-fusion transformer. . . . 34

Figure 4.4 Tracking results on two GTOT (Li et al., 2016) sequences comparing RGB-only (upper frames) and our RGB-T method (lower frames), as in Table 4.2. Red boxes are predictions; green boxes are ground truth. Frame indices appear above and below the frames. ***RainyCar2*** illustrates rainy conditions; ***WalkingOcc*** shows partial occlusion. 35

Figure 4.5 Attribute-based evaluation on RGBT234. (a) Max Precision Rate (MPR). (b) Max Success Rate (MSR). 39

Figure 5.1 Illustration of the prompt-adapted **mmMobileViT** block used for RGB-T tracking. During each transformer stage, RGB and IR features are unfolded and used to generate modality-aware prompts via convolutional Prompt Blocks (Zhu et al., 2023a), which are added to the frozen RGB token stream before each Separable Mixed-Attention layer. The blocks highlighted in orange indicate the only learnable modules during the prompt learning stage, while all other components remain frozen. This design enables efficient cross-modal adaptation with minimal parameter overhead. 44

Figure 5.2 Overview of the Siamese-based RGB-T tracker architecture. Each modality (RGB and IR) is processed by its own backbone and neck module. Each backbone branch uses the same MobileViTv2 architecture described in Section 3.2 and visualized in Figure 3.2, but without inter-modal interaction in Layer 4. Instead of fusing features within the backbone, this variant performs cross-modal integration using a single cross-modal fusion transformer, as described in Section 3.4, before passing the output to the prediction head. 48

List of Tables

Table 2.1 Summary of recent RGB-T and RGB-X trackers and their key design features. 13

Table 4.1 Comparison with state-of-the-art lightweight RGB-T trackers on LasHeR (Li et al., 2021), RGBT234 (Li et al., 2019), and GTOT (Li et al., 2016). The best, second-best, and third-best results are highlighted in red, green, and blue, respectively. Our model achieves the best trade-off between speed, accuracy, and computational cost. **SMAT*** (Gopal and Amer, 2024) is the baseline model that only has the RGB pipeline trained on LasHeR. Parameters and MAC counts are reported in millions (M) and gigas (G), respectively. 31

Table 4.2 Ablation results on RGBT234 (Li et al., 2019) and GTOT (Li et al., 2016) comparing *base_rgb* (RGB-only), dual-modality without cross-fusion transformer, and the full model. Integrating IR features substantially improves performance, and our transformer-based fusion module yields additional gains while keeping the model lightweight and efficient. 35

Table 4.3 Ablation of fusion strategies in backbone Layers 3 and 4. The proposed progressive fusion strategy achieves the best balance of accuracy and efficiency across all three benchmarks. 37

Table 4.4 Comparison of final-stage fusion strategies after cross-modal transformer. The proposed sigmoid-based channel-wise weighting achieves the highest accuracy across benchmarks with minimal parameters and real-time speed. 40

Table 5.1 Performance comparison of prompt learning on the LasHeR dataset. Our model adds 1.5M trainable prompt parameters on top of a 3.926M frozen RGB backbone. ViPT results are from their original paper. 46

Table 5.2 Performance comparison of the Siamese-based RGB-T tracker on the LasHeR dataset. The Siamese model uses independent backbones and necks for each modality and fuses features at the final stage. Despite having more parameters, its GMACs are comparable due to the use of lightweight separable attention and similar backbone structure in both models (see Eq. 4 in Section 3.2). 50

Table 7.1 Comparison of our RGB-D extension with representative multimodal trackers on DepthTrack. *SUTrack* and *EMTrack* also appear in the main results table and are discussed in Related Works. 56

Chapter 1

Introduction

1.1 Background and Motivation

Visual object tracking is a fundamental task in computer vision, aiming to continuously localize a target object across a video sequence given only its initial bounding box (Ye et al., 2022). While significant progress has been made in RGB-based tracking through deep convolutional and transformer-based models (Ye et al., 2022; Bai et al., 2024), these approaches often degrade under challenging imaging conditions such as low illumination, motion blur, or partial occlusion. In such scenarios, relying solely on RGB information can lead to visual ambiguities, background clutter confusion, or complete target loss.

To overcome the limitations of RGB-only tracking, recent research has focused on multimodal tracking, where RGB data is combined with auxiliary signals such as thermal infrared (T), depth (D), event-based data (E), or natural language (L) (Chen et al., 2025). These additional modalities provide complementary cues that are less invariant to illumination changes, motion blur, or occlusion, thereby improving robustness and continuity in challenging scenarios. Among these, RGB-Thermal (RGB-T) tracking has emerged as a widely studied subclass, owing to the passive nature and ambient-light independence of thermal imaging. Thermal data is especially effective in adverse conditions such as fog, heavy rain, poor lighting, or partial occlusion, where RGB features tend to degrade significantly (Lu et al., 2025; Zhang et al., 2023a).

1.2 Limitations of Existing Multimodal Models

While multimodal tracking has advanced significantly with the introduction of transformer-based architectures, these gains often come at the expense of computational efficiency. State-of-the-art models such as SUTrack (Chen et al., 2025), STTrack (Hu et al., 2025a), and TBSI (Hui et al., 2023) leverage Vision Transformer (ViT) (Dosovitskiy et al., 2021) or Swin Transformer (Liu et al., 2021) backbones to perform dense global reasoning across modalities, resulting in strong tracking accuracy. However, this performance is typically accompanied by high model complexity and latency. For instance, the SUTrack-L384 variant (Chen et al., 2025) contains 247 million parameters and operates at only 12 frames per second (FPS) on standard GPU hardware, rendering it unsuitable for real-time deployment. These computational demands pose significant barriers to adoption in latency-sensitive or resource-constrained environments.

In response to the high computational cost of standard Vision Transformers, the RGB-only tracking community has increasingly adopted efficient backbone architectures to reduce complexity and improve real-time performance. Backbone models such as LeViT (Graham et al., 2021), MobileViT (Mehta and Rastegari, 2022), and MobileViTv2 (Mehta and Rastegari, 2023) have been employed in lightweight tracking frameworks to mitigate the quadratic complexity of self-attention and reduce parameter overhead. These architectures offer strong trade-offs between speed and accuracy by incorporating structural optimizations for a more efficient attention mechanism (Gopal and Amer, 2024; Blatter et al., 2023; Zhai et al., 2024). Despite their demonstrated success in RGB-only tracking, these compact backbone designs remain largely absent from RGB-T and broader multimodal tracking literature.

1.3 Proposed Solution Overview

To address the computational inefficiencies of existing RGB-T models, we propose the first RGB-T tracker based on the MobileViTv2 backbone (Mehta and Rastegari, 2023), designed to balance performance with speed and model compactness. A key component of our model is Separable Attention (Mehta and Rastegari, 2023), which replaces the dense token-to-token operations of standard Multi-Head Attention (MHA) with a lightweight formulation based on softmax-weighted query

broadcasting. While MHA exhibits quadratic complexity $\mathcal{O}(k^2d)$ in sequence length k, Seaprable achieves a linear complexity of $\mathcal{O}(kd)$, enabling efficient global reasoning without incurring high memory or compute costs (Mehta and Rastegari, 2023), as shown in Figure 3.3.

To integrate complementary cues across modalities, we adopt a progressive fusion strategy that first processes RGB and thermal inputs independently and then gradually merges them through separable mixed-attention based attention layers. A final transformer-based cross-fusion module combines the feature streams before prediction, enabling both early and late fusion without introducing redundant backbone components. This architecture results in a lightweight RGB-T tracker with only 3.93 million parameters and an inference speed of 122 FPS on standard GPU hardware, while offering comparable accuracy to state-of-the-art efficient multimodal trackers.

Our model design builds directly on the theoretical foundation of the MobileViTv2 backbone and the separable-mixed attention from SMAT (Gopal and Amer, 2024) which is built upon SSA mechanism in MobileViTv2 (Mehta and Rastegari, 2023). In turn, this enables efficient global reasoning with linear complexity. For implementation, we leverage the publicly available SMAT framework (Gopal and Amer, 2024), which applies MobileViTv2 to RGB-only tracking and achieves state-of-the-art performance in this category. We extend SMAT by introducing an RGB-T data pipeline, modifying it to handle both RGB and thermal infrared inputs. Our key architectural contributions, including progressive fusion across backbone layers and cross-modal fusion transformer, are implemented within this foundation, resulting in a lightweight and effective RGB-T tracker built on a well-established architecture with proven performance.

1.4 Datasets and Evaluation Protocol

1.4.1 Datasets

To evaluate the effectiveness of our proposed RGB-T tracker, we conduct experiments on three standard benchmarks: LasHeR (Li et al., 2021), RGBT234 (Li et al., 2019), and GTOT (Li et al., 2016). These datasets span a variety of real-world challenges including low illumination, occlusion, and thermal crossover, providing a comprehensive basis for multimodal tracking evaluation.

LasHeR is the largest RGB-T dataset to date, with 1224 annotated sequences covering 32 object

categories and 19 challenging attributes. It provides frame-aligned RGB and thermal pairs and reports metrics such as Success Rate (SR), Precision Rate (PR), and Normalized PR (NPR).

RGBT234 consists of 234 sequences, capturing diverse scenarios with partial RGB-T misalignment. To account for this, it uses Maximum PR (MPR) and Maximum SR (MSR), which select the better modality per frame for evaluation.

GTOT includes 50 sequences, mainly focused on pedestrian tracking under occlusion and small object size. It uses SR and a stricter PR metric, evaluating predictions within a 5-pixel threshold.

1.4.2 Evaluation Metrics

To quantitatively assess tracking performance, we adopt standard metrics used across RGB-T benchmarks:

Precision Rate (PR) measures the percentage of predicted bounding box centers that fall within a predefined pixel distance from the ground truth, typically 20 pixels for LasHeR and RGBT234, and 5 pixels for GTOT due to smaller object sizes.

Success Rate (SR) computes the proportion of frames where the Intersection over Union (IoU) between prediction and ground truth exceeds a threshold. In most cases, SR is reported as the area under the success curve (AUC).

Normalized PR (NPR) used in LasHeR accounts for target scale by normalizing the precision threshold based on object size.

Maximum Precision Rate (MPR) and **Maximum Success Rate (MSR)** are adopted by RGBT234 to compensate for minor misalignment between RGB and thermal frames. These metrics select the better performing modality at each frame, reflecting the best-case performance achievable with cross-modal inputs.

Collectively, these metrics offer a balanced view of tracking robustness (SR), localization accuracy (PR), and modality adaptability (MPR/MSR).

1.5 Experimental Scope and Architectural Variants

We benchmark our model against state-of-the-art lightweight RGB-T trackers including SUTrack-Tiny (Chen et al., 2025), EMTrack (Liu et al., 2024a), CMD (Zhang et al., 2023a), and TBSI (Hui et al., 2023). Our tracker achieves the lowest parameter count (3.93M), a fair amount of compute (4.35 GMACs), and highest speed (122 FPS) among all compared models. While being significantly more efficient, it delivers competitive accuracy across LasHeR (Li et al., 2021), RGBT234 (Li et al., 2019), and GTOT (Li et al., 2017).

To further investigate the adaptability and modularity of our approach, we explore two prevalent architectural extensions. First, we integrate a prompt learning mechanism in which thermal-aware prompts are injected into the frozen RGB backbone to adapt it for RGB-T tracking without full model re-training. Second, we implement a Siamese variant with modality-specific MobileViTv2 backbones and defer fusion to a final transformer layer. These extensions allow us to analyze different fusion strategies, adaptation mechanisms, and training cost trade-offs in multimodal settings.

1.6 Thesis Statement

This thesis addresses the challenge of building lightweight RGB-T object trackers capable of real-time inference on resource-constrained devices. While transformer-based models offer strong tracking performance, their high computational cost limits deployment in practical settings. To overcome this, we propose a lightweight RGB-T tracker built upon the MobileViTv2 backbone, incorporating separable mixed-attention and progressive fusion strategies.

1.7 Contributions

The main contributions of this thesis are summarized as follows:

- We propose the first RGB-T tracker based on the MobileViTv2 architecture, introducing separable mixed-attention for efficient multimodal modeling and the novel progressive modality fusion protocol.

- Our model achieves significant reductions in parameter count and inference latency while maintaining competitive accuracy across three challenging RGB-T benchmarks.
- We investigate two architectural extensions to evaluate adaptability and fusion strategies: (1) a prompt learning mechanism that adapts a frozen RGB-pretrained backbone using thermal-aware prompts with minimal training cost; and (2) a Siamese variant with modality-specific backbones and late fusion, highlighting trade-offs in modularity and fusion timing.
- We conduct comprehensive empirical studies and ablation experiments to benchmark our model against state-of-the-art RGB-T trackers and analyze two major design extensions.
- The implementation will be available at: code

1.8 Thesis Organization

The remainder of this thesis is structured as follows:

- **Chapter 2 – Literature Review:** Reviews prior literature on lightweight RGB-only and RGB-T tracking, emphasizing transformer-based models, lightweight designs, and fusion strategies.
- **Chapter 3 – Methodology:** Presents the proposed RGB-T tracker based on MobileViTv2, including architectural details, attention mechanisms, and fusion design.
- **Chapter 4 – Experimental Results:** Reports quantitative comparisons against existing lightweight RGB-T trackers on three benchmarks (LasHeR, RGBT234, and GTOT), along with ablation studies evaluating the contributions of thermal input, fusion modules, and other architectural components.
- **Chapter 5 – Analysis of Variants:** Introduces two architectural extensions, prompt learning and Siamese fusion, and analyzes their theoretical foundations, implementation, and experimental outcomes. Although neither method outperforms our proposed design in Chapter 3, they offer valuable insights into design trade-offs and serve as baselines for future improvement.

- **Chapter 6 – Conclusion and Future Work:** Summarizes key findings and discusses future directions for research in efficient multimodal tracking.

Chapter 2

Literature Review

2.1 From RGB-only to Multimodal Tracking

Object tracking has traditionally relied on RGB images as the sole input modality. With the progression from handcrafted features in correlation filter-based methods to convolutional neural networks (CNNs), RGB-based tracking has become increasingly robust and accurate. Models such as SiamFC (Bertinetto et al., 2016) utilized deep CNNs to extract discriminative features for target representation, enabling improved performance across various benchmarks. More recently, transformer-based models such as OSTrack (Ye et al., 2022) have further enhanced tracking accuracy by incorporating global attention mechanisms that better capture context and spatial dependencies in the input frames (Bertinetto et al., 2016; Ye et al., 2022). Despite these advancements, RGB-only trackers face persistent limitations, particularly in challenging scenarios involving low illumination, partial occlusion, or background–foreground similarity. In such cases, relying on RGB input alone often results in degraded performance due to insufficient discriminative information.

To enhance tracking robustness under these conditions, the community has increasingly explored the integration of additional information sources. These complementary signals may originate from either low-level sensor data or high-level semantic cues. Low-level modalities such as thermal infrared and depth provide invariant information under appearance shifts or poor lighting, making them suitable for scenarios where RGB fails. High-level auxiliary data, such as natural language prompts or event-based representations, offer task-specific or temporally precise information

that can refine the tracking objective (Chen et al., 2025). The move toward multimodal tracking architectures reflects the recognition that richer input representations can better address the diverse challenges encountered in real-world tracking applications.

Building on this shift toward richer representations, the field of object tracking has increasingly adopted multimodal architectures that integrate RGB with other complementary modalities. Among these, RGB-X tracking refers to the broad family of methods where the auxiliary modality X can include thermal infrared (T), depth (D), event-based signals (E), or even natural language guidance (L). These approaches aim to enhance robustness by leveraging the complementary strengths of each modality in varied conditions. Within this landscape, RGB-T tracking has gained particular prominence due to the practical accessibility of thermal cameras and its natural alignment with critical application areas such as night-time surveillance and autonomous navigation in low-visibility environments. RGB-T specialized trackers (Wang et al., 2024a; Lu et al., 2025; Sun et al., 2024; Wang et al., 2024b; Cao et al., 2024; Hui et al., 2023; Lu et al., 2024; Xiao et al., 2025; Chen et al., 2024) are explicitly designed to model interactions between RGB and thermal inputs. In contrast, more general RGB-X trackers (Chen et al., 2025; Zhu et al., 2023a; Hou et al., 2024; Hong et al., 2024; Wu et al., 2024; Hu et al., 2025a,b; Liu et al., 2024a) target broader cross-modal compatibility by adopting unified frameworks that handle multiple modality types.

2.2 Design Choices in Multimodal Trackers

Multimodal trackers can also be distinguished by their architectural design, which typically falls into two broad paradigms: transformer-based and Siamese-based frameworks. Transformer-based trackers, such as TBSI (Hui et al., 2023), AINet (Lu et al., 2025), and CAFormer (Xiao et al., 2025), adopt a unified architecture where feature extraction and cross-modal fusion occur within shared attention layers. These models are often built on Vision Transformers (ViT) (Dosovitskiy et al., 2021) or their variants (Zhu et al., 2023a; Hou et al., 2024; Hong et al., 2024; Wu et al., 2024; Hu et al., 2025a,b; Liu et al., 2024a; Wang et al., 2024a; Sun et al., 2024; Wang et al., 2024b; Cao et al., 2024; Chen et al., 2024; Lu et al., 2024), allowing them to model long-range dependencies and modality interactions through global attention. This design enables strong contextual

reasoning across modalities, contributing to superior tracking accuracy. However, these advantages come at the cost of high computational overhead due to dense attention operations and large model sizes. In contrast, Siamese-based trackers like SiamTFA (Zhang et al., 2024) and SiamTDR (Wang et al., 2023) utilize separate network branches to process each modality independently. While this approach reduces architectural complexity and offers improved modularity and interpretability, it often lacks deep cross-modal interaction, which can limit tracking performance.

Based on the different architectural paradigms, multimodal trackers adopt varying fusion strategies depending on architectural design, typically categorized into early, mid-level, and late fusion. While early (pixel-level) and late (decision-level) fusion are conceptually simple, they often suffer from limited cross-modal interaction. Recent RGB-T trackers predominantly use mid-level fusion, which combines features at intermediate stages for richer multimodal representation. Among these, dense all-layer fusion is adopted by AINet (Lu et al., 2025) and M3PT (Wang et al., 2024b), where features from both modalities are fused at multiple transformer layers to promote comprehensive modality interaction. Models such as CAFormer (Xiao et al., 2025) and TGTrack (Chen et al., 2024) embed fusion within the attention mechanism through cross-modulated attention, generating modality-adaptive representations. TBSI (Hui et al., 2023) introduces a unique bridging strategy, in which a fused template feature bridges the RGB and thermal search branches to guide interaction. On the other hand, CMD (Zhang et al., 2023a) employs cross-modal knowledge distillation, using a pre-trained RGB teacher to supervise the thermal student stream during training, thereby eliminating fusion at inference time.

Another emerging design choice in multimodal tracking is the use of prompt learning, a technique originally developed in NLP and recently adapted for vision tasks. Prompt learning enables models to conditionally adapt to different input modalities with minimal changes to the core architecture, facilitating parameter-efficient training. This is especially relevant in multimodal tracking, where large modality-specific branches can increase model complexity and training cost. ViPT (Zhu et al., 2023a) pioneered this approach by introducing visual prompts for each modality, enabling flexible fusion with minimal overhead. Building on this idea, EMTrack (Liu et al., 2024a) adopts a lightweight temporal prompting mechanism to enhance feature representation across frames. Both

models demonstrate that prompt-based methods not only improve adaptability to multiple modalities but also serve as a practical solution for limited training data and model generalization.

2.3 Gaps in RGB-T Efficiency and Insights from RGB-only Solutions

Despite the effectiveness of transformer-based architectures in improving multimodal tracking performance, their widespread adoption introduces notable computational burdens. Vision Transformer (ViT)–based and Swin Transformer-based backbones, employed in models such as SUTrack (Chen et al., 2025), AINet (Lu et al., 2025), and CAFormer (Xiao et al., 2025), offer strong global context modeling and seamless modality interaction through dense self-attention. However, the self-attention mechanism in standard ViT has a computational complexity of $\mathcal{O}(N^2 \cdot d)$, where N is the number of tokens and d is the embedding dimension (Dosovitskiy et al., 2021). In multimodal settings, N increases with the addition of modality-specific tokens, exacerbating the quadratic scaling and inflating both memory usage and inference time. For instance, SiamTFA employs a Swin Transformer backbone with over 192 million parameters (Zhang et al., 2024), while SUTrack, based on HiViT (Zhang et al., 2023b), introduces hierarchical token reduction to alleviate the cost, yet still suffers from latency limitations due to global attention layers. These constraints render many state-of-the-art RGB-T models impractical for real-time or edge deployment, where computational resources and latency budgets are limited.

In response to these challenges, the RGB-only tracking community has explored using a range of lightweight transformer models that optimize the trade-off between performance and efficiency. Backbone designs such as LeViT (Graham et al., 2021), MobileViT (Mehta and Rastegari, 2022), and MobileViTv2 (Mehta and Rastegari, 2023) significantly reduce computational overhead through architectural innovations. These include convolutional tokenization for spatially efficient embeddings, inverted bottlenecks to limit parameter growth, and streamlined attention mechanisms designed to reduce redundancy. Among these, MobileViTv2 introduces separable, which reduces the spatial attention complexity from $\mathcal{O}(N^2)$ to near-linear $\mathcal{O}(N)$ in practice, offering efficient modeling of long-range dependencies with substantially lower computational cost. Beyond the above, several notable transformer variants also directly address the quadratic complexity of self-attention.

Models such as Reformer (Kitaev et al., 2020) reduce attention complexity to $\mathcal{O}(L \log L)$ using locality-sensitive hashing, while Longformer (Beltagy et al., 2020) and Linformer (Wang et al., 2020) achieve $O(L)$ complexity via windowed sparse attention and low-rank factorization, respectively. Sparse Transformer models (Child et al., 2019) attain a complexity of $\mathcal{O}(L\sqrt{L})$ through sparse attention patterns. Additionally, VMamba (Liu et al., 2024b) introduces a 2D visual State Space attention (SS2D) mechanism tailored to vision inputs, also achieving linear-time complexity in the visual domain. These works collectively reinforce the broader trend toward efficient attention mechanisms, complementing the separable attention achieved in MobileViTv2.

Building on Mobile Vision Transformer backbones or alternative lightweight principles, several RGB-only trackers have been proposed to deliver efficient performance. Trackers such as SMAT (Gopal and Amer, 2024), ETTrack (Blatter et al., 2023), MVT (Gopal and Amer, 2023), HiT (Kang et al., 2023), and MobileTrack (Zhai et al., 2024) leverage the aforementioned efficient backbones to achieve real-time operation with competitive accuracy. Notably, HiT, based on LeViT (Graham et al., 2021) backbones, comes in multiple variants with varying resolution and parameter budgets, allowing for flexible deployment across devices with different resource constraints. In parallel, CNN-based trackers like LightTrack (Yan et al., 2021a) and LightFC (Li et al., 2024) demonstrate that fully-convolutional architectures can also provide strong efficiency–accuracy trade-offs. LightTrack employs a differentiable one-shot neural architecture search (NAS) framework to automatically discover optimal lightweight architectures for tracking, achieving decent FPS on standard benchmarks. LightFC uses knowledge distillation from a strong teacher to train a compact fully-convolutional Siamese model, maintaining competitive performance with minimal parameters. These lightweight RGB-only architectures have demonstrated that compact designs can deliver competitive tracking performance with significantly fewer multiply–accumulate (MAC) operations, reduced latency, and lower memory consumption. For instance, MobileViT-based trackers (Gopal and Amer, 2024, 2023) operate at real-time speeds while maintaining accuracy comparable to larger ViT-based designs (Ye et al., 2022; Bai et al., 2024), as shown in benchmarks across RGB-only tracking. However, these advances have not yet been widely adopted in the multimodal domain. RGB-T trackers still rely heavily on large token sets and fusion modules with dense attention, amplifying resource consumption. While existing methods often emphasize accuracy, they

Table 2.1: Summary of recent RGB-T and RGB-X trackers and their key design features.

Model	Architecture	Fusion Strategy	Prompt Learning	Backbone	Venue/Year
CMD (Zhang et al., 2023a)	Siamese	Distillation	No	ResNet-18	CVPR 2023
SiamTFA (Zhang et al., 2024)	Siamese (Triple)	Late	No	Swin Transformer	TCSVT 2024
SiamTDR (Wang et al., 2023)	Siamese	Late	No	AlexNet	TMM 2023
LightFC-X (Li et al., 2025)	Siamese	Distillation	No	CNN	ArXiv 2025
TBSI (Hui et al., 2023)	Transformer	Mid	No	ViT	TCSVT 2023
EMTrack (Liu et al., 2024a)	Transformer	Early (Addition)	Yes (Temporal)	D-MAE-Tiny	ICCV 2024
ViPT (Zhu et al., 2023a)	Transformer	Mid	Yes (Modality Prompts)	ViT	ECCV 2022
CAFormer (Xiao et al., 2025)	Transformer	Mid (Cross-Modulated)	No	ViT	TMM 2025
AINet (Lu et al., 2025)	Transformer	Mid (Dense All-layer)	No	Mamba	TIP 2025
M3PT (Wang et al., 2024b)	Transformer	Mid (Dense)	No	ViT	TCSVT 2024
TGTrack (Chen et al., 2024)	Transformer	Mid (Cross-Modulated)	No	ViT	TIP 2024
SUTrack (Chen et al., 2025)	Transformer	Mid	No	HiViT	CVPR 2025
Ours	Hybrid CNN-Transformer	Mid(SSA) + Late(3.4)	No	MobileViTv2	–

frequently neglect deployment feasibility. This leaves a compelling opportunity for RGB-T tracking architectures that incorporate lightweight design principles, such as separable attention and efficient tokenization, to close the gap between high performance and practical deployability on real-world devices.

2.4 Recent Efficient RGB-T Trackers

Several recent works have attempted to improve the efficiency of RGB-T tracking by either simplifying architectural components or optimizing training strategies. The CMD tracker (Zhang et al., 2023a) exemplifies this trend by combining a lightweight backbone with a cross-modal knowledge distillation strategy. Rather than relying on a heavy transformer, CMD uses a small CNN-based architecture trained to imitate a stronger teacher network. Similarly, SiamTDR (Wang et al., 2023) adopts a streamlined Siamese-based architecture with shallow encoders and minimal fusion, aiming to minimize computational burden. While these models demonstrate promising efficiency, their simplified designs lead to limited representational capacity. CMD requires a strong teacher model and staged training, complicating reproducibility, and SiamTDR often underperforms due to its limited depth and weak modality interaction. On the other end of the spectrum, SiamTFA (Zhang et al., 2024) utilizes a Swin Transformer-based triple-stream network that independently encodes the template, RGB search region, and thermal search region. While it achieves strong performance, the architecture is over-engineered for practical deployment, its 192M+ parameter count makes it one of the heaviest RGB-T trackers to date.

Other efforts aim to strike a more practical trade-off between efficiency and accuracy by using unified architectures and streamlined fusion. TBSI (Hui et al., 2023) introduces a ViT-based framework where a novel Template-Bridged Search region Interaction module enables effective modality fusion via shared tokens. While still moderately heavy, it avoids redundant computation by integrating cross-modal interactions within the transformer layers, offering a balance between interpretability and cost. EMTrack (Liu et al., 2024a) pushes efficiency further by adopting a unified framework with modality-specific patch embedding and shared attention layers. Its lightweight D-MAE-Tiny backbone, combined with simple addition-based fusion and light knowledge distillation from ViPT, enables real-time CPU inference at 29.1 FPS with only 16M parameters. However, despite its efficient runtime design, EMTrack requires a three-stage training pipeline involving supervised learning, temporal prompt tuning, and cross-modal knowledge distillation, which adds significant overhead and complexity to the training process. LightFC-X (Li et al., 2025) proposed a multimodal extension of the RGB-only LightFC (Li et al., 2024) tracker, following a fully-convolutional Siamese design with late fusion and cross-modal knowledge distillation. By leveraging a compact CNN backbone and efficient training strategy, it achieves strong performance with minimal computational overhead, demonstrating that unimodal lightweight architectures can be effectively adapted to the multimodal tracking setting.

Chapter 3

Methodology

3.1 Overview

This chapter presents our lightweight RGB-T tracking model based on MobileViTv2 (Mehta and Rastegari, 2023), designed for efficient multimodal object tracking. We first describe the multimodal backbone and its use of separable mixed-attention (Gopal and Amer, 2024) in Section 3.2, followed by the neck module in Section 3.3, the cross-modal fusion transformer in Section 3.4, the prediction head in Section 3.5, and the training strategy and loss functions in Section 3.6. Throughout the chapter, we highlight key architectural decisions and theoretical comparisons to multi-head attention (Dosovitskiy et al., 2021).

To clearly distinguish the architectural components employed in our backbone, we first clarify the related terminologies. **Separable self-attention (SSA)** refers to the lightweight attention formulation introduced in MobileViTv2 (Mehta and Rastegari, 2023), reducing the quadratic complexity of conventional multi-head attention to a linear form with respect to the token dimension. Building upon this principle, SMAT (Gopal and Amer, 2024) defined **separable mixed-attention** as the case where tokens from both the template and search frames are concatenated and jointly processed in the backbone. Finally, the general term **separable attention** is used to denote the conceptual class of attention mechanisms derived from SSA that retain linear complexity, regardless of whether the input tokens originate from a single input or from mixed template–search pairs (or even from multimodal token mixing).

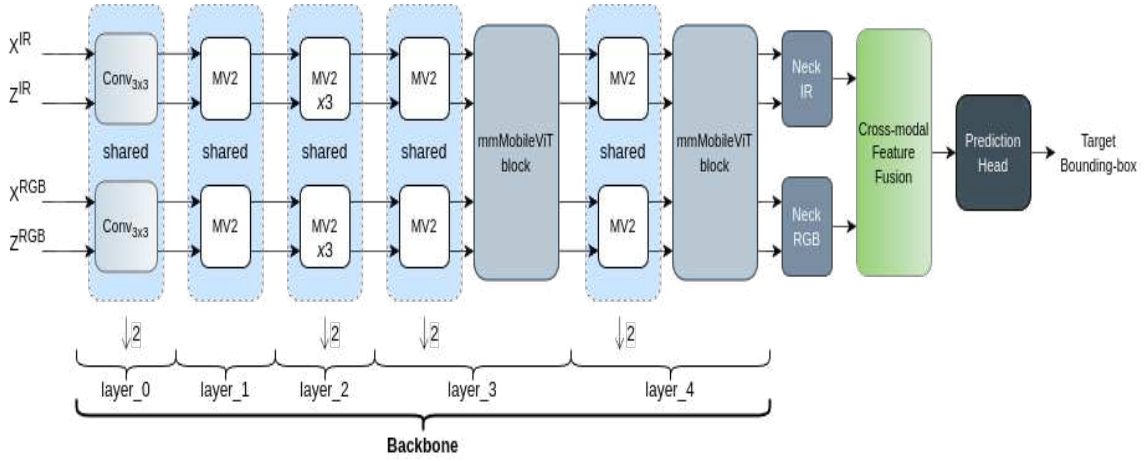


Figure 3.1: The pipeline of proposed RGB-T tracker. MV2 stands for MobileNetV2 (Inverted Residual blocks) and mmMobileViT for multimodal MobileViT (see Figure 3.2). $\downarrow 2$ indicates spatial downsampling by 2. $\{X^{\mathrm{IR}}, Z^{\mathrm{IR}}\}$ show the input search and template frames of Thermal Infrared Modality (IR). $\times 3$ shows the number of subsequent MV2 blocks in layer_2.

3.2 MobileViTv2-based Multimodal Backbone

The proposed RGB-T tracking model is built on top of a lightweight and unified feature extraction backbone based on MobileViTv2 (Mehta and Rastegari, 2023), a hybrid architecture that integrates convolutional and transformer-based operations. The backbone receives template and search images from both RGB and infrared (IR) modalities (see Figure 3.1), denoted as $X_{\mathrm{in}}^{\mathrm{RGB}}, X_{\mathrm{in}}^{\mathrm{IR}} \in \mathbb{R}^{W_x \times H_x \times 3}$ and $Z_{\mathrm{in}}^{\mathrm{RGB}}, Z_{\mathrm{in}}^{\mathrm{IR}} \in \mathbb{R}^{W_z \times H_z \times 3}$, where H_x, W_x and H_z, W_z represent the spatial dimensions of the search and template images, respectively. In our implementation, input search images from both modalities have a spatial size of 256×256, while template images are sized at 128×128. Each input is first passed through a shared depth-wise 3×3 convolution followed by a point-wise 1×1 convolution, which increases the channel dimension to 32 and reduces the spatial resolution by a factor of two. As a result, the output feature maps after this stage have a size of 32×128×128 for the search branch and 32×64×64 for the template branch. This convolutional stem operates identically across all modalities and inputs, enabling a unified representation space early in the network.

These feature maps are then processed by two layers composed of MobileNetV2 (Sandler et al., 2018) inverted residual blocks. In Layer 1, a single block is used to double the number of channels to 64 while preserving spatial resolution. Layer 2 consists of two consecutive inverted residual blocks:

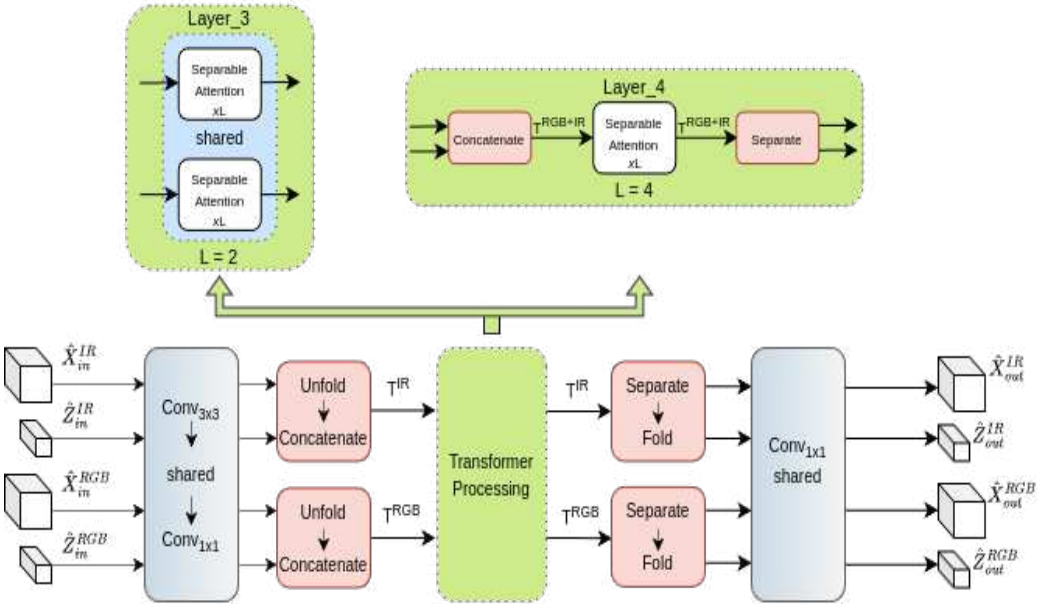


Figure 3.2: The detailed Architecture of the proposed mmMobileViT with demonstration of intra-modal and inter-modal separable attention in layer_3 and layer_4, respectively. L are the number of transformer layers at mmMobileViT blocks in layers 3 and 4.

the first performs downsampling, reducing the resolution to 64×64 for search features and 32×32 for template features, while the second maintains these spatial sizes and increases the channel dimension to 128. To ensure both efficiency and consistency across modalities, the same MobileNetV2 blocks are shared between the RGB and IR branches. This design not only avoids parameter duplication and reduces computation but also promotes the learning of modality-invariant features early in the backbone. The inverted residual structure itself provides an efficient alternative to standard convolutional layers (Sandler et al., 2018), leveraging depthwise separable convolutions and linear bottlenecks to reduce the number of operations and parameters. By expanding the feature dimension before applying non-linear transformations, and by decoupling spatial filtering from channel mixing through depthwise convolutions, these blocks preserve representational power while significantly lowering the computational cost.

Following the convolutional feature extraction stages, the backbone transitions to two transformer-based layers (Layer 3 and Layer 4), each implemented using our proposed mmMobileViTv2 block (see Figure 3.2). These blocks extend the original MobileViTv2 (Mehta and Rastegari, 2023) design by incorporating progressive multimodal interactions specifically tailored for RGB-T tracking.

As illustrated in Figure 3.2, each mmMobileViT block takes modality-specific feature maps from both the search and template branches, denoted as $\hat{X}_{\text{in}}^{\text{RGB}}, \hat{Z}_{\text{in}}^{\text{RGB}}, \hat{X}_{\text{in}}^{\text{IR}}, \hat{Z}_{\text{in}}^{\text{IR}}$, and enhances their contextual representation through separable mixed-attention mechanisms (Gopal and Amer, 2024). Layer 3 focuses on intra-modal modeling, where RGB and IR streams are processed independently to preserve modality-specific structures while capturing global relations across template and search regions. By postponing cross-modal interaction until Layer 4, the architecture avoids early fusion, which may compromise the unique spatial or thermal features of each modality. This two-stage progressive design enables the backbone to learn strong contextualized features within each modality before performing joint reasoning and fusion in the final stage.

In Layer 3, the feature maps of each modality are first processed by a shared 3×3 depth-wise convolution followed by a 1×1 point-wise projection. This operation preserves the channel dimension at 128 while reducing the spatial resolution, resulting in feature maps of size 32×32 for search inputs and 16×16 for template inputs. For each modality (RGB and IR) and input type (template and search), the feature maps are then partitioned into non-overlapping patches of size $p_1 \times p_1$ ($p_1 = 2$). We select a small patch size of 2×2 to preserve fine-grained spatial detail for high-precision tracking. Each patch is flattened into a token vector, forming a sequence of $N = \frac{H \cdot W}{p_1^2}$ tokens of dimension d. The patch tokens from the template (Z) and search (X) branches are concatenated along the sequence dimension for each modality:

$$\mathbf{T}^{\text{RGB}} = [\mathbf{Z}_{\text{patch}}^{\text{RGB}} \parallel \mathbf{X}_{\text{patch}}^{\text{RGB}}] \in \mathbb{R}^{C \times N \times d}, \quad \mathbf{T}^{\text{IR}} = [\mathbf{Z}_{\text{patch}}^{\text{IR}} \parallel \mathbf{X}_{\text{patch}}^{\text{IR}}] \in \mathbb{R}^{C \times N \times d} \tag{1}$$

These modality-specific token sequences are then independently processed by L ($L = 2$ for Layer_3) shared transformer layers equipped with separable mixed-attention, which captures global intra-modal dependencies across the template and search regions.

To model global dependencies between template and search regions efficiently, we employ Separable Mixed Attention within each transformer layer of the mmMobileViT block. In contrast to standard multi-head self-attention (MHA), which performs dense token-to-token interactions, separable attention adopts a lightweight formulation that achieves linear complexity with respect to sequence length—making it well-suited for mobile or real-time tracking. Given a token sequence

$\mathbf{T} \in \mathbb{R}^{k \times d}$ formed by concatenating patch tokens from the search and template inputs, separable attention first generates the query (Q), key (K), and value (V) matrices using shared 1×1 convolutional projections:

$$\mathrm{Q} \in \mathbb{R}^{k \times 1}, \quad \mathrm{K} \in \mathbb{R}^{k \times d}, \quad \mathrm{V} \in \mathbb{R}^{k \times d}. \tag{2}$$

The query is then normalized using softmax and broadcast across the feature dimension:

$$\tilde{\mathrm{Q}} = \mathrm{Softmax}(\mathrm{Q}) \in \mathbb{R}^{k \times 1} \quad \Rightarrow \quad \tilde{\mathrm{Q}} \in \mathbb{R}^{k \times d}. \tag{3}$$

This enables attention computation through efficient per-token weighting:

$$\mathrm{A} = \sum_{k} \left(\tilde{\mathrm{Q}} \odot \mathrm{K} \right), \quad \mathrm{M} = \tilde{\mathrm{A}} \odot \mathrm{ReLU}(\mathrm{V}), \tag{4}$$

where $\odot$ denotes element-wise multiplication and $\mathrm{M} \in \mathbb{R}^{k \times d}$ represents the final contextualized output. This SSA formulation has a linear complexity of $\mathcal{O}(kd)$, making it computationally attractive for sequences with large k. As shown on the right of Figure 3.3, this formulation avoids computing dense attention matrices by broadcasting relevance scores derived from the softmax-normalized query vector.

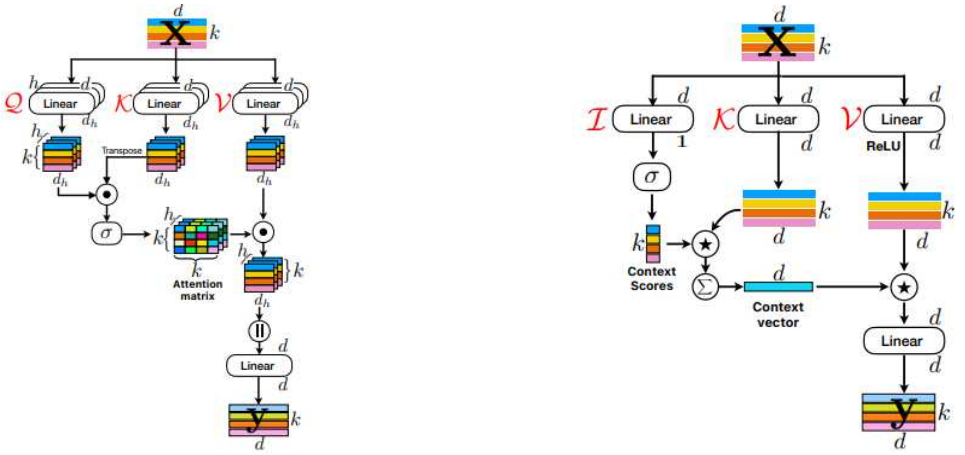
By comparison, standard multi-head attention (MHA) in Vision Transformers (Dosovitskiy et al., 2021) uses full pairwise interactions. Given the same token input $\mathbf{T} \in \mathbb{R}^{k \times d}$, it computes queries, keys, and values via linear projections:

$$\mathrm{Q} = \mathbf{T}W_Q, \quad \mathrm{K} = \mathbf{T}W_K, \quad \mathrm{V} = \mathbf{T}W_V, \quad W_Q, W_K, W_V \in \mathbb{R}^{d \times d}, \tag{5}$$

yielding $\mathrm{Q}, \mathrm{K}, \mathrm{V} \in \mathbb{R}^{k \times d}$. Attention is then computed using the standard scaled dot-product:

$$\mathrm{A} = \mathrm{Softmax}\left(\frac{\mathrm{QK}^\top}{\sqrt{d}} \right) \mathrm{V}, \tag{6}$$

which requires evaluating a full similarity matrix $\mathrm{QK}^\top \in \mathbb{R}^{k \times k}$. As a result, the overall complexity scales quadratically with sequence length, $\mathcal{O}(k^2 d)$, which can be prohibitive for long token sequences or high-resolution inputs. In contrast, the left side of Figure 3.3 illustrates how



(a) MHA in Transformers (Dosovitskiy et al., 2021)

(b) Separable Self-Attention (Mehta and Rastegari, 2023)

Figure 3.3: Visual comparison between standard multi-head self-attention (MHA) and separable self-attention (SSA). While MHA performs dense pairwise interactions between all tokens, SSA avoids this by using a softmax-normalized query and efficient per-token weighting, leading to significantly reduced computational complexity.

MHA computes full pairwise token interactions using scaled dot-product attention, which leads to quadratic complexity. Comparing Equations (4) and (6), separable attention avoids computing expensive pairwise token similarities by leveraging broadcasted softmax weights (Equation (3)) to aggregate features more efficiently. This dramatically reduces memory and computational demands, enabling separable mixed-attention to preserve global contextual reasoning while scaling gracefully with input size, ideal for efficient RGB-T tracking in constrained or real-time environments. A visual comparison between standard MHA and separable attention is illustrated in Figure 3.3.

After separable mixed-attention is applied in Layer 3, the output token sequences are folded back into their original spatial configuration. A 1×1 convolutional projection restores the original channel dimension to $C = 128$, producing spatial feature maps for each input type and modality: $\hat{X}_{\text{out}}^{\text{RGB}}, \hat{Z}_{\text{out}}^{\text{RGB}}, \hat{X}_{\text{out}}^{\text{IR}}, \hat{Z}_{\text{out}}^{\text{IR}}$. These features remain modality-specific but now encode globally contextualized representations, forming the input to the final transformer stage.

In mmMobileViT block (see Figure 3.2) of Layer 4, each modality again passes through a shared depth-wise 3×3 convolution and a 1×1 point-wise projection. This stage reduces the spatial resolution of the search features to 16×16 and template features to 8×8, while increasing the channel dimension to 192. The resulting feature maps are tokenized into patch sequences as

before, but now the RGB and IR token sequences are concatenated to form a unified multimodal representation:

$$\mathbf{T}^{\text{RGB+IR}} = [\mathbf{T}^{\text{RGB}} \parallel \mathbf{T}^{\text{IR}}] \in \mathbb{R}^{k \times 2d}, \tag{7}$$

where each modality contributes its contextualized template and search tokens. This fused sequence is then processed using $L = 4$ layers of separable mixed-attention with an attention dimension of 192, enabling the model to capture inter-modal dependencies across spatial and temporal contexts.

By deferring fusion to Layer 4, the network ensures that modality-specific features are first well-developed and globally contextualized, avoiding premature mixing that could distort thermal or visual cues. The resulting fused token sequence is folded back into spatial maps and passed through a concluding 1×1 convolution, which projects the channel dimension to $C_f = 384$ for all search and template feature maps. This yields globally-aware RGB-T feature maps that capture both intra-modal and inter-modal relations, which are then passed to the neck module for further alignment and refinement.

3.3 Neck Module

Following backbone processing, the modality-specific feature maps of the template and search branches, denoted as $\hat{Z}^{\text{RGB}}, \hat{X}^{\text{RGB}}, \hat{Z}^{\text{IR}}, \hat{X}^{\text{IR}}$, are transformed into a comparison space via pixel-wise cross-correlation (PW-XCorr), a common operation in tracking frameworks (Yan et al., 2021b; Gopal and Amer, 2024). This operation computes localized similarity between the template and search features, effectively encoding their spatial alignment and correspondence without introducing additional parameters.

At this stage, each modality retains its own feature maps from Layer 4 of the backbone. The input template and search maps for RGB and IR branches have dimensions $384 \times 8 \times 8$ and $384 \times 16 \times 16$, respectively. These are processed independently by the PW-XCorr operation.

For modality $m \in \{\text{RGB}, \text{IR}\}$, the fused feature representation is computed as:

$$\mathbf{F}^m = \text{PWCorr}(\hat{X}^m, \hat{Z}^m), \tag{8}$$

where $\mathbf{F}^m \in \mathbb{R}^{128\times16\times16}$ retains the spatial layout of the search region while integrating information from the corresponding template. This operation is applied independently for RGB and IR branches, allowing each modality to preserve its unique visual or thermal characteristics during similarity computation (no weight sharing). The use of PW-XCorr here offers a lightweight and interpretable way to focus the model's attention on spatially consistent regions, making it particularly well-suited for object tracking, where temporal coherence and position similarity are crucial. Importantly, since the operation is performed per modality, the resulting features remain separated in representation space, enabling effective and interpretable fusion in the next stage.

These modality-specific correlation features serve as the input to the Cross-Modal Fusion Transformer, where inter-modal reasoning and joint refinement are performed.

3.4 Proposed Cross-Modal Fusion Transformer

After intra-modal feature extraction in Layer 3 and inter-modal fusion in Layer 4 of the backbone, as well as pixel-wise alignment via the Neck Module, the RGB and IR branches yield two modality-specific feature maps: $\mathbf{F}^{\mathrm{RGB}}$ and $\mathbf{F}^{\mathrm{IR}}$, each of size $128 \times 16 \times 16$. These are then passed into the Cross-Modal Fusion Transformer for final joint reasoning and integration before prediction.

It is important to note that, although Layer 4 in the backbone also applied a form of multimodal attention, the focus there was on fusing RGB and IR search/template sequences jointly, leveraging spatial and temporal correspondence between modalities. In contrast, the attention here operates on the modality-level fused outputs from the Neck Module, treating each as a compressed, task-aligned representation. This separation ensures that fusion happens progressively: from intra-modal contextualization to spatial alignment, and finally to semantic-level cross-modal integration.

To prepare for fusion, the feature maps $\mathbf{F}^{\mathrm{RGB}}$ and $\mathbf{F}^{\mathrm{IR}}$ are first divided into non-overlapping patches of size $p_2 \times p_2$, where $p_2 = 8$. Compared to the earlier patching in the backbone (e.g., $p_1 = 2$ in Layers 3 and 4), this step acts on already downsampled features and serves to reduce the sequence length to just a few tokens per modality, while preserving coarse spatial structure. Each patch is flattened into a token vector, and the resulting RGB and IR token sequences are concatenated along the token dimension to form a single multimodal input sequence.

This sequence is then processed by $L = 1$ transformer layer equipped with Separable-Mixed Attention, consistent with the design in the backbone. The use of separable mixed-attention again emphasizes efficiency: by avoiding full token-to-token attention and instead applying softmax-normalized relevance scores across channels, the model maintains low complexity while still capturing important cross-modal dependencies. This step refines the fused representations while keeping the computational budget small, critical for real-time or resource-constrained tracking systems.

Once processed, the tokens are reshaped back into spatial feature maps. These are combined using a learnable channel-wise fusion module, which uses sigmoid-normalized weights to adaptively balance RGB and IR contributions:

$$\mathbf{F}_{\text{fused}} = \sigma(\mathbf{W}^{\text{RGB}}) \odot \mathbf{F}^{\text{RGB}} + \sigma(\mathbf{W}^{\text{IR}}) \odot \mathbf{F}^{\text{IR}}, \tag{9}$$

where $\mathbf{W}^{\text{RGB}}, \mathbf{W}^{\text{IR}} \in \mathbb{R}^{1\times C\times 1\times 1}$ are learnable parameters ($C = 128$), $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication.

This fusion transformer thus completes the multimodal feature pipeline: starting from intra-modal context modeling in Layer 3, inter-modal token fusion in Layer 4, per-modality alignment via PW-XCorr, and finally semantic-level integration through attention-guided reweighting. The resulting fused feature map $\mathbf{F}_{\text{fused}}$ serves as the input to the prediction head for final bounding box estimation and classification.

3.5 Prediction Head

The prediction head serves as the final module of the proposed RGB-T tracker, transforming the fused representation $\mathbf{F}_{\text{fused}}$ into tracking outputs. Structurally, the head is composed of two parallel branches: a classification head for foreground-background discrimination, and a regression head for bounding box prediction. We adopt the design from SMAT (Gopal and Amer, 2024), which integrates lightweight convolutional and attention-based modules to balance accuracy and efficiency.

Each branch first processes the input $\mathbf{F}_{\text{fused}}$ using a 3×3 convolution, followed by a small number of transformer layers based on Separable Mixed-Attention. These layers maintain consistency

with the rest of the architecture and further refine the features while keeping the computational burden low. A final 3×3 convolution projects the output to the required number of channels: one for classification logits, and four for bounding box offsets.

Importantly, inspired by the SMAT architecture, the regression branch uses double the number of transformer layers compared to the classification branch. This design accounts for the greater difficulty of accurately regressing bounding box coordinates, which typically requires a deeper feature refinement process (Gopal and Amer, 2024). In our implementation, the classification head uses two separable mixed-attention layers, while the regression head employs 4.

The classification branch outputs a spatial heatmap indicating the likelihood of target presence at each location, while the regression branch predicts the bounding box parameters (l, t, r, b) relative to each spatial location. These outputs are decoded with respect to the receptive field of the feature map, which retains the spatial resolution of 16×16 from the fusion transformer. The resulting predictions form the final tracking output for each frame.

This head design ensures a lightweight yet expressive prediction mechanism, fully aligned with the rest of the efficient architecture. By reusing separable attention throughout the network, from backbone to head, the model preserves architectural consistency while enabling real-time tracking performance.

3.6 Training Loss Function

To train the proposed RGB-T tracker, we adopt a multi-task loss function that jointly optimizes classification and regression objectives. Each output from the prediction head (Section 3.5) is supervised with a modality-agnostic loss term designed to balance detection accuracy and localization quality.

The overall training loss is formulated as:

$$L_{\text{total}} = L_{\text{cls}} + \lambda_1 \cdot L_1 + \lambda_2 \cdot L_{\text{giou}}, \tag{10}$$

where L_{cls} is the classification loss, and L_1 and L_{giou} represent the regression losses for bounding box prediction. The weights λ_1 and λ_2 control the balance between localization precision and shape

alignment.

Classification Loss: We employ the Focal Loss (Law and Deng, 2018), which addresses class imbalance by down-weighting easy negatives and focusing the model on hard foreground-background examples:

$$L_{\text{cls}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{11}$$

where p_t is the predicted probability for the correct class, and α_t, γ are hyperparameters controlling the loss shape.

Regression Losses: Localization supervision consists of two terms:

- L_1: the smooth ℓ_1 loss encourages precise coordinate matching.
- L_{giou}: the Generalized Intersection-over-Union (GIoU) loss (Rezatofighi et al., 2019) captures shape misalignment and provides informative gradients even when boxes do not overlap.

Both losses are applied at all valid spatial locations that correspond to the ground-truth object center. We use fixed weights $\lambda_1 = 5.0$, $\lambda_2 = 2.0$, following prior work in SMAT (Gopal and Amer, 2024).

Chapter 4

Experimental Results

This chapter presents a comprehensive evaluation of our proposed RGB-T tracker across standard benchmarks. We begin by detailing implementation and evaluation protocols, followed by comparisons with state-of-the-art lightweight trackers in Section 4.1. We then conduct four ablation studies to assess: (1) the effect of thermal modality and the transformer-based fusion module (see Section 4.2.1); (2) the role of progressive intra- and inter-modal fusion in backbone Layers 3 and 4 (see Section 4.2.2); (3) the impact of different final fusion strategies after the transformer (see Section 4.2.4); and (4) the model's attribute-wise robustness across 12 visual challenges (see Section 4.2.3).

4.1 Results

4.1.1 Implementation Details

Model: We use a template-search architecture where RGB and IR frames are cropped and resized to 128×128 and 256×256 for template and search branches, respectively. The backbone consists of an input convolutional block, two MobileNetV2 stages, and two transformer-based mm-MobileViT layers as shown in Figure 3.2. Channel dimensions increase as $\{3 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 384\}$, with four downsampling operations resulting in output sizes of 8×8 (template) and 16×16 (search). Layer_3 and Layer_4 contain 2 and 4 transformer blocks, respectively. We use a patch size of 2×2 in the backbone (Section 3.2) and 8×8 in the cross-fusion transformer

(Section 3.4). The cross-modal fusion transformer uses a hidden dimension $C_f = 128$ and consists of a single transformer block.

Training: We train only on the LasHeR training set (Li et al., 2021), using 60,000 image pairs per epoch for 60 epochs. The optimizer is AdamW (Loshchilov and Hutter, 2017) with learning rate 4×10^{-4}, weight decay 10^{-4}, and gradient clipping at 0.1. The backbone learning rate is scaled by 0.1. A step decay (factor 0.1) is applied at epoch 40. The loss follows Equation (10), with weights $\lambda_1 = 5$ and $\lambda_2 = 2$ as in (Gopal and Amer, 2024). We use pretrained MobileViTv2 weights (Mehta and Rastegari, 2023) for initialization and remove positional embeddings from all transformer blocks. Augmentations include horizontal flipping and brightness jitter. All training is conducted on an NVIDIA Tesla V100 GPU (32GB) with batch size 128.

Hyperparameters: We do not perform any dedicated hyperparameter tuning. Due to the large search space and limited computational resources, we adopt the same training settings used in SMAT (Gopal and Amer, 2024), including optimizer type, learning rate, weight decay, and loss weights. This is a common practice in the object tracking community; for example, SMAT itself inherits its hyperparameters directly from OSTrack (Ye et al., 2022). The only exceptions are the number of training epochs and the learning rate decay point, which we slightly adjust to better match the LasHeR dataset size and the reduced parameter count of our lightweight backbone. All other hyperparameters are preserved to ensure comparability and reproducibility.

Inference: The template is fixed from the first frame. At each time step t, a $4\times$ region around the last predicted box is extracted, resized to 256×256, and passed through the search branch. A Hanning window is applied to the classification map $\mathcal{R}$, and the peak is selected as the new target center. Inference results are generated using an Intel Core i9-12900KF CPU and an NVIDIA RTX 3090 GPU.

4.1.2 Evaluation Scope and Benchmarks

To rigorously assess the performance of our proposed RGB-T tracker, we conduct experiments on three publicly available RGB-T tracking benchmarks: LasHeR (Li et al., 2021), RGBT234 (Li et al., 2019), and GTOT (Li et al., 2016). These datasets are selected for their diversity, annotation quality, and wide adoption in the multimodal tracking literature. All evaluations follow a standard

protocol, where the model is trained only on the LasHeR training set and tested directly on all three benchmarks without fine-tuning. Performance is measured using well-established metrics to capture both overlap-based and location-based accuracy under varying tracking challenges.

4.1.3 Datasets

LasHeR (Li et al., 2021) is currently the largest RGB-T tracking benchmark, comprising a total of 1224 sequences, 979 for training and 245 for testing. It covers 32 object categories and 19 diverse tracking attributes such as fast motion, occlusion, deformation, and scale variation. All sequences are frame-wise aligned across RGB and thermal modalities. LasHeR reports three metrics: Success Rate (SR), Precision Rate (PR), and Normalized PR (NPR), making it a comprehensive testbed for performance evaluation in complex multimodal scenarios.

RGBT234 (Li et al., 2019) includes 234 video sequences with 12 challenging attributes such as low illumination, thermal crossover, background clutter, and camera motion. Unlike LasHeR, slight misalignment exists between RGB and thermal frames. To mitigate this, the benchmark adopts Maximum Precision Rate (MPR) and Maximum Success Rate (MSR), which compute performance using the better modality per frame, offering a robust estimate of tracking capability in the presence of modality inconsistency.

GTOT (Li et al., 2016) is a widely used benchmark focusing on RGB-T pedestrian tracking. It contains 50 short sequences featuring relatively small targets and frequent occlusions. GTOT uses SR and a stricter PR metric defined at a 5-pixel threshold, in contrast to the 20-pixel threshold used in LasHeR and RGBT234. This makes GTOT a challenging dataset for fine-grained localization and suitable for evaluating high-precision tracking.

Real-world RGB-T tracking often suffers from spatial and temporal desynchronization between modalities due to sensor misalignment, varying fields of view, or unsynchronized frame rates. These

factors can degrade fusion quality and tracking robustness, especially for methods that assume perfect pixel-level alignment. Recent efforts have addressed these challenges by proposing alignment-aware strategies such as learnable feature warping or quality-aware weighting that adaptively downplay unreliable modalities. One study (Zhou et al., 2023) tried to explore handling spatial misalignment by learning flexible cross-modal associations without relying on aligned inputs. However, another study (Zhu et al., 2023b) highlight that widely used datasets like GTOT and RGBT234 exhibit non-negligible spatial misalignment, challenging the assumption of perfect correspondence across modalities. Despite these developments, most prior RGB-T tracking studies, including those using LasHeR, GTOT, and RGBT234, continue to evaluate models on spatially and temporally aligned inputs. To stay consistent with the established literature and ensure fair comparison, our evaluations are also conducted on the aligned versions of these standard benchmarks.

VOT-RGBT2019 (Kristan et al., 2019) and **VOT-RGBT2020** (Kristan et al., 2020) are two additional RGB-T benchmarks that focus on tracking under specific attributes such as occlusion and camera motion. However, both are subsets of the RGBT234 dataset, with overlapping sequences and the same attribute coverage. Accordingly, we do not include them as separate test sets in our experiments.

4.1.4 Evaluation Metrics

All trackers are evaluated using the one-pass evaluation (OPE) protocol, where tracking is initialized in the first frame and runs continuously without access to future frames. Performance is measured using both overlap-based and location-based metrics:

- **Success Rate (SR)**: Measures the area under the curve (AUC) of the success plot. A frame is considered successful if the Intersection over Union (IoU) between the predicted box B_t and ground truth G_t exceeds a threshold τ:

$$\text{SR} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(\frac{|B_t \cap G_t|}{|B_t \cup G_t|} > \tau\right), \quad \tau \in [0, 1]$$

- **Precision Rate (PR)**: Computes the percentage of frames in which the Euclidean distance between the predicted center (x_t, y_t) and the ground truth (x_t^*, y_t^*) is less than a fixed threshold

θ:

$$\text{CLE}_t = \sqrt{(x_t - x_t^*)^2 + (y_t - y_t^*)^2}$$

LasHeR and RGBT234 use $\theta = 20$ pixels; GTOT uses a stricter threshold of 5 pixels.

- **Normalized PR (NPR)**: A scale-invariant version of PR introduced by LasHeR, where the distance threshold is normalized with respect to the object size, ensuring fairness across varying object scales.

- **Maximum PR and SR (MPR/MSR)**: Used in RGBT234 to account for minor modality misalignments. These metrics select the better of RGB or thermal predictions at each frame to represent the upper bound of modality performance.

This combination of metrics allows us to quantify both coarse and fine-grained accuracy across a diverse set of benchmarks, enabling fair comparison with prior RGB-T tracking methods.

4.1.5 Comparison to Related Work

Table 4.1 compares our model to state-of-the-art lightweight multimodal trackers including SUTrack-Tiny (Chen et al., 2025), EMTrack (Liu et al., 2024a), CMD (Zhang et al., 2023a), and TBSI-Tiny (Hui et al., 2023). We report accuracy (PR, SR, NPR, MPR, MSR), efficiency (FPS), and computational cost (GMACs). Our tracker achieves the fewest parameters (3.93M), fastest inference (121.9 FPS), and competitive or best accuracy on all benchmarks. For instance, on GTOT, our model ranks first in SR (0.741) and second in PR (0.884), indicating superior overlap consistency. On RGBT234, our method achieves the third-best MPR and MSR, despite using over 5× fewer parameters than SUTrack.

While SUTrack-Tiny obtains slightly higher accuracy on LasHeR and RGBT234, it uses 22M parameters and is 20% slower (100 FPS). CMD and EMTrack perform reasonably well, but are heavier and slower. TBSI-Tiny is closer in scale but still lags in speed and accuracy. To further contextualize the importance of thermal information and modality fusion, we include SMAT* as a baseline RGB-only tracker trained on LasHeR (Li et al., 2021). SMAT (Gopal and Amer, 2024) is a state-of-the-art model for unimodal RGB tracking, built on the same MobileViTv2 backbone as our

Tracker	#Params (M)	MACs (G)	FPS (GPU)	LasHeR (Li et al., 2021) PR	SR	NPR	RGBT234 (Li et al., 2019) MPR	MSR	GTOT (Li et al., 2016) PR	SR
SUTrack_Tiny (Chen et al., 2025)	22	3	100	0.667	0.539	–	0.859	0.638	0.853	0.726
EMTrack (Liu et al., 2024a)	16	2	83.8	0.659	0.533	–	0.838	0.601	–	–
CMD (Zhang et al., 2023a)	19.9	–	30	0.590	0.464	0.546	0.824	0.584	0.892	0.734
TBSI_Tiny (Hui et al., 2023)	14.9	–	40	0.617	0.489	0.578	0.794	0.555	0.881	0.706
Ours	3.93	4.35	121.9	0.603	0.473	0.567	0.806	0.589	0.895	0.7467
SMAT*	3.76	–	154.6	0.549	0.438	0.512	0.737	0.536	0.690	0.578

Table 4.1: Comparison with state-of-the-art lightweight RGB-T trackers on LasHeR (Li et al., 2021), RGBT234 (Li et al., 2019), and GTOT (Li et al., 2016). The best, second-best, and third-best results are highlighted in red, green, and blue, respectively. Our model achieves the best trade-off between speed, accuracy, and computational cost. **SMAT*** (Gopal and Amer, 2024) is the baseline model that only has the RGB pipeline trained on LasHeR. Parameters and MAC counts are reported in millions (M) and gigas (G), respectively.

method. However, when evaluated on RGB-T benchmarks under challenging conditions, such as low illumination, occlusion, or thermal crossover, its performance degrades noticeably (see SMAT* results in 4.1). This comparison illustrates that even strong RGB-only models benefit significantly from the inclusion of thermal cues and modality-aware fusion. Our results demonstrate that incorporating a dedicated thermal pipeline and progressive fusion strategy is essential to achieving robust and reliable performance in such scenarios.

Compared to these, our method strikes a good balance between speed, size, and accuracy. Notably, while our model has the lowest parameter count (3.93M), its MACs are higher (4.35G) than those of SUTrack (3G) and EMTrack (2G). This is primarily due to the use of global attention operations in our backbone and the presence of transformer layers throughout the network. However,

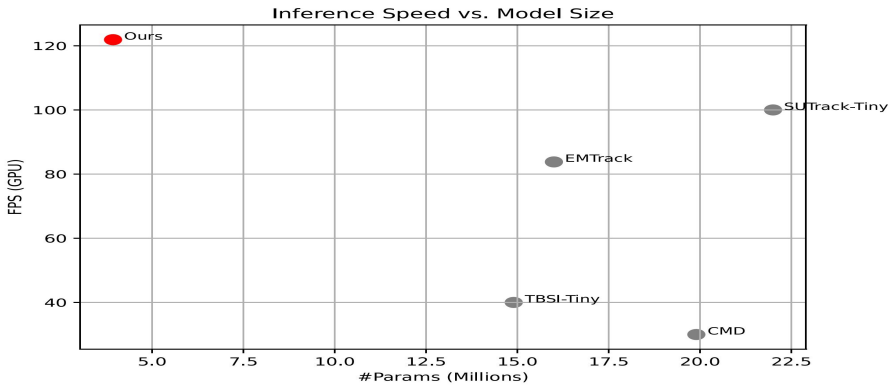


Figure 4.1: Trade-off between parameter count and inference speed (FPS) for lightweight RGB-T trackers. Our tracker (top-left) achieves the best combination of high FPS and low parameter count.

these operations are implemented using separable mixed-attention, which maintains linear complexity and enables real-time performance. As a result, our model still achieves the highest FPS (121.9), demonstrating that slightly higher compute can be efficiently amortized by careful architectural optimization. It is also important to clarify that all inference speed comparisons are reasonably fair despite minor differences in evaluation hardware across methods. Our model and EMTrack (Liu et al., 2024a) are both evaluated on an NVIDIA RTX 3090 GPU. In contrast, SUTrack (Chen et al., 2025) reports FPS using an RTX 2080Ti, CMD (Zhang et al., 2023a) uses an older RTX 1080Ti, and TBSI (Hui et al., 2023) uses an RTX 3080Ti. While the RTX 3090 and 3080Ti offer slightly better performance than the 2080Ti and 1080Ti, the gap is not large enough to affect the FPS of lightweight models with fewer parameters and low computational cost. A FPS vs. #Params scatter plot (Figure 4.1) further visualizes the inference speed vs. model size trade-off across models.

To ensure a fair comparison, we examined the training configurations of all four RGB-T trackers in Table 4.1. SUTrack (Chen et al., 2025), EMTrack (Liu et al., 2024a), and TBSI-Tiny (Hui et al., 2023) adopt the loss as used in OSTrackYe et al. (2022), with identical loss weights ($\lambda_1 = 5$, $\lambda_2 = 2$). CMD (Zhang et al., 2023a), on the other hand, employs a different distillation-based loss function. All models share same hyperparameters, including a resolution for search and target, a drop path rate of 0.1, a gradient clipping norm of 0.1, and an AdamW optimizer.

4.1.6 Comparison with Baseline Architecture

To highlight the architectural contributions of our RGB-T tracker, we compare its structure to SMAT (Gopal and Amer, 2024), a strong RGB-only baseline. The architecture of SMAT is shown in Figure 4.2, while our full model is detailed in Figure 3.1. We summarize the key differences below:

- **Modality Support:** SMAT is designed exclusively for RGB tracking. In contrast, our model operates on RGB-T input, requiring dedicated mechanisms for cross-modal interaction.
- **Backbone Fusion Strategy:** SMAT uses a lightweight MobileViT-based backbone with no modality-specific reasoning. Our model adapts the MobileViTv2 backbone and introduces progressive modality interaction using separable mixed-attention: intra-modal attention in

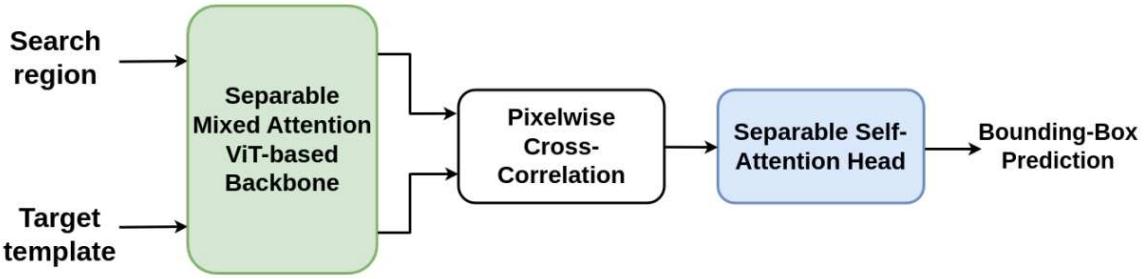


Figure 4.2: Architecture of SMAT (Gopal and Amer, 2024), a strong RGB-only baseline. Our RGB-T tracker extends this design with progressive modality interaction in the backbone and a dedicated cross-fusion transformer after the neck, as described in Sections 3.2 and 3.4.

Layer 3 and inter-modal fusion in Layer 4 (see Section 3.2).

- **Neck and Fusion Modules:** Both models use a neck module adapted from DiMP (Bhat et al., 2019). However, SMAT feeds the neck output directly into the prediction head. In our model, we introduce a cross-fusion transformer between the neck and head to enable late-stage modality fusion (Section 3.4).
- **Prediction Head:** Our prediction head (Section 3.5) follows the same design as SMAT.

To further clarify the interactions in our backbone model, we illustrate in Figure 4.3 how the MobileViTv2 backbone has evolved across three contexts. Originally, MobileViTv2 was introduced for classification and detection tasks with separable self-attention as the feature extractor. This backbone was later adapted in SMAT for tracking, where search and template frames are processed using separable mixed-attention. Our RGB-T tracker builds upon this foundation by introducing dual RGB and thermal streams, progressive intra- and inter-modal fusion within backbone layers. This visualization highlights how our contributions extend the baseline pipeline with modality-aware reasoning while preserving the lightweight efficiency of MobileViTv2.

4.2 Ablation Studies

4.2.1 Visual Results

To assess the contributions of each major component in our architecture, namely, the thermal modality and the cross-fusion transformer, we conduct a series of ablation experiments. These are

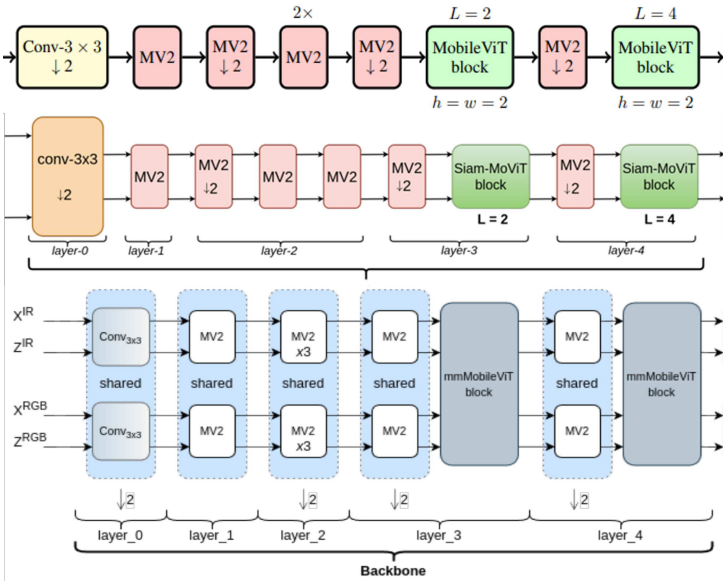


Figure 4.3: Evolution of architectures from MobileViTv2 (top) for classification/detection, to SMAT (middle) for RGB tracking with search/template inputs, and finally to our RGB-T tracker (bottom) with progressive fusion and cross-fusion transformer.

evaluated on the RGBT234 (Li et al., 2019) and GTOT (Li et al., 2016) benchmarks, with results reported in Table 4.2.

(1) *base_rgb*. This variant removes the infrared (IR) stream entirely and operates using only RGB input. It also excludes the fusion module described in Section 3.4, resulting in a streamlined RGB-only tracker. This model achieves the fastest inference speed at 154.64 FPS and the smallest parameter size (3.767M), as shown in Table 4.2. However, the absence of thermal cues leads to significant performance degradation: on RGBT234, it records 0.7378 MPR and 0.5364 MSR, drops of 6.8% and 5.1%, respectively, compared to the full model. On GTOT, the decline is even more pronounced, with 0.6904 PR and 0.5785 SR (–19.3% PR and –16.5% SR), confirming that thermal input is particularly beneficial under adverse conditions such as poor illumination and occlusion.

(2) *w/o Cross-Fusion Transformer*. This variant retains both RGB and IR inputs but removes the transformer-based fusion mechanism introduced in Section 3.4, replacing it with a simple weighted addition. As reported in Table 4.2, this modification increases the model size only slightly to 3.786M and improves FPS to 124.00, but tracking performance also drops: MPR and MSR decrease by 2.6% and 2.0% on RGBT234, and by 5.2% PR and 5.1% SR on GTOT compared to the full model. This demonstrates that while simple fusion retains modality information, it fails to

Model Variant	#Params (in millions)	FPS (GPU)	RGB	IR	Fusion Transformer	RGBT234 (Li et al., 2019) MPR	RGBT234 (Li et al., 2019) MSR	GTOT (Li et al., 2016) PR	GTOT (Li et al., 2016) SR
base_rgb (SMAT)	3.767	154.64	✓	✗	✗	0.7378	0.5364	0.6904	0.5785
w/o Cross-Fusion Transformer	3.786	124.00	✓	✓	✗	0.7860	0.5704	0.8318	0.6902
Proposed Model	3.926	121.92	✓	✓	✓	**0.8063**	**0.5890**	**0.8949**	**0.7467**

Table 4.2: Ablation results on RGBT234 (Li et al., 2019) and GTOT (Li et al., 2016) comparing *base_rgb* (RGB-only), dual-modality without cross-fusion transformer, and the full model. Integrating IR features substantially improves performance, and our transformer-based fusion module yields additional gains while keeping the model lightweight and efficient.

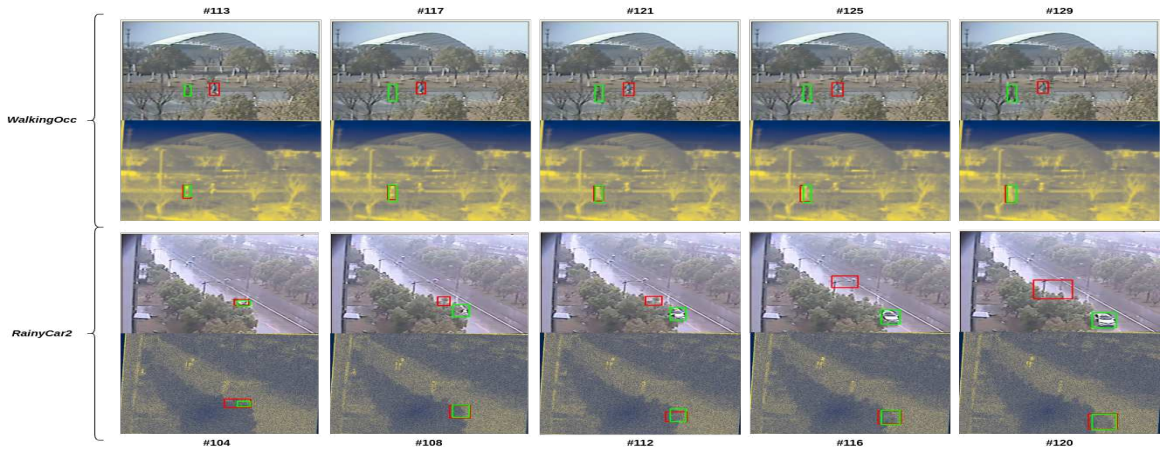


Figure 4.4: Tracking results on two GTOT (Li et al., 2016) sequences comparing RGB-only (upper frames) and our RGB-T method (lower frames), as in Table 4.2. Red boxes are predictions; green boxes are ground truth. Frame indices appear above and below the frames. ***RainyCar2*** illustrates rainy conditions; ***WalkingOcc*** shows partial occlusion.

exploit their cross-modal relationships as effectively as our transformer-based design.

(3) Proposed Model. The full model combines RGB and thermal streams using the cross-fusion transformer module, balancing accuracy and efficiency. As shown in Table 4.2, it achieves the highest accuracy across all metrics: 0.8063 MPR and 0.589 MSR on RGBT234, and 0.8838 PR and 0.7409 SR on GTOT. These results validate the effectiveness of the transformer-based fusion module for learning spatial and semantic interactions between modalities, with only marginal increases in parameters (3.926M) and minimal runtime overhead (121.92 FPS).

Qualitative Results. Figure 4.4 illustrates visual comparisons between the full and the RGB-only models. In challenging scenarios such as *RainyCar2* (low visibility) and *WalkingOcc* (partial occlusion), the RGB-only tracker drifts significantly, while the RGB-T model remains fairly aligned with the ground truth, thanks to complementary thermal features and cross-modal integration.

Summary. These findings confirm that (1) thermal data significantly boosts robustness and accuracy, and (2) the proposed transformer-based fusion outperforms simple alternatives with minimal computational tradeoff.

4.2.2 Effect of Progressive Intra- and Inter-Modal Fusion

To evaluate the empirical benefits of progressive fusion in the mmMobileViTv2 backbone, we perform an ablation study that isolates the contributions of intra-modal attention in Layer 3 and inter-modal attention in Layer 4. This study aims to justify the architectural decision to delay cross-modal fusion until the final backbone layer. Specifically, we compare the proposed progressive design to two alternative configurations:

- **No-Fusion Backbone:** Disables all multimodal interaction in the backbone. Both Layer 3 and Layer 4 apply intra-modal separable mixed-attention independently to RGB and IR streams.
- **All-Fusion Backbone:** Applies inter-modal separable mixed-attention in both Layer 3 and Layer 4, allowing early and repeated cross-modal mixing.
- **Proposed:** Performs intra-modal attention in Layer 3 and inter-modal attention in Layer 4, supporting a two-stage progression from modality-specific reasoning to joint fusion.

All three variants retain the same input configuration, tokenization scheme, and backbone architecture (with $L = 2$ transformer blocks in Layer 3 and $L = 4$ in Layer 4), and use separable mixed-attention for global context modeling as defined in Equation (4). Since separable mixed-attention exhibits linear complexity with respect to input length and channel dimension ($\mathcal{O}(kd)$), and the number of transformer layers remains constant across variants, the GMACs for all models are effectively identical. This ensures that observed performance differences are attributable solely to the fusion strategy, not computational budget.

Results are reported in Table 4.3 across three benchmarks: LasHeR (Li et al., 2021), RGBT234 (Li et al., 2019), and GTOT (Li et al., 2016). Accuracy metrics include Precision Rate (PR), Normalized PR (nPR), and Success Rate (SR) for LasHeR, Maximum PR and SR (MPR, MSR) for RGBT234, and PR/SR for GTOT. We also report GMACs, parameter count, and inference speed (FPS).

Model Variant	Params (M)	GMACs	FPS (GPU)	LasHeR			RGBT234		GTOT	
				PR	nPR	SR	MPR	MSR	PR	SR
Proposed (Progressive)	3.926	4.35	121.92	**0.6026**	**0.5674**	**0.4729**	**0.8063**	**0.5890**	**0.8949**	**0.7467**
No-Fusion Backbone	3.926	4.35	110.00	0.5889	0.5563	0.4667	0.7824	0.5652	0.8612	0.7171
All-Fusion Backbone	3.926	4.35	129.00	0.5839	0.5427	0.4503	0.7812	0.5714	0.8639	0.7077

Table 4.3: Ablation of fusion strategies in backbone Layers 3 and 4. The proposed progressive fusion strategy achieves the best balance of accuracy and efficiency across all three benchmarks.

As shown in Table 4.3, the proposed model achieves the highest accuracy across all benchmarks and metrics. This confirms that delaying inter-modal fusion to Layer 4, after each modality has developed strong contextual representations, leads to more effective integration. In contrast, the all-fusion variant, which performs early cross-modal mixing in Layer 3, consistently underperforms. Early fusion introduces information from the external modality before features are semantically mature, often injecting noise that disrupts the attention distribution. This premature entanglement distorts modality-specific cues before sufficient intra-modal reasoning occurs. The no-fusion variant also falls short, confirming the necessity of explicit multimodal interaction for effective RGB-T tracking. In particular, without any modality fusion, the model is unable to learn joint long-range representations between the RGB and thermal views, which are essential for capturing complementary cues under challenging conditions.

Interestingly, all models maintain identical GMACs (4.35G) and parameter counts (3.926M), validating the linear complexity of separable attention (Eq. 4) and its ability to support flexible fusion strategies without increasing theoretical compute. However, the observed FPS varies due to architectural differences in how the attention layers process modality inputs. The all-fusion variant is the fastest (129 FPS), likely because it concatenates RGB and IR tokens into a single sequence, allowing the attention blocks to operate on uniform input shapes using optimized dense matrix operations. In contrast, the no-fusion and the proposed model use shared separable mixed-attention layers that process each modality stream independently within the same block. This introduces conditional logic, fragmented memory access, and reduced tensor core utilization, factors that hinder GPU efficiency. The no-fusion model, which fully duplicates the separable attention path for both modalities, incurs the most latency (110 FPS). The proposed model, while slightly slower than the all-fusion version (121.92 FPS), avoids full duplication and still benefits from late joint attention in Layer 4. These differences underscore that runtime efficiency is not only governed by GMACs, but

also by how attention is structured and parallelized in practice. More broadly, these results highlight that while multimodal fusion is necessary, it must be carefully timed, naive or premature fusion can degrade attention quality and harm performance, rather than help it.

In conclusion, this ablation confirms the value of the proposed progressive fusion of the backbone. The proposed design, which emphasizes intra-modal reasoning before cross-modal integration, yields the best trade-off between speed, compute, and accuracy, supporting the architectural rationale introduced in Section 3.2.

4.2.3 Attribute-Based Performance Analysis

To provide deeper insight into the strengths and limitations of our model, we perform an attribute-wise analysis on the RGBT234 benchmark. This evaluation assesses how well the proposed tracker handles specific visual challenges by partitioning sequences according to predefined attributes (Li et al., 2019). Each attribute reflects a common difficulty in RGB-T tracking, such as occlusion, thermal crossover, or motion blur.

Figure 4.5 presents radar plots showing the model's Max Precision Rate (MPR) and Max Success Rate (MSR) across 12 key attributes. These results reveal several consistent trends. The model performs best under the *No Occlusion* (NO), *Thermal Crossover* (TC), and *Partial Occlusion* (PO) conditions, highlighting the effectiveness of the cross-modal fusion mechanism in disentangling modality-specific features even when thermal and RGB cues are ambiguous or redundant. Performance is also robust under fast motion (FM), motion blur (MB), and scale variation (SV), likely due to the global context modeling provided by transformer layers in both backbone and fusion stages.

By contrast, the most challenging scenarios are *Background Clutter* (BC), *Hyaline Occlusion* (HO), and *Low Resolution* (LR), where both modalities tend to provide weak or noisy supervision. These cases emphasize the importance of temporal consistency and appearance modeling, which may be explored in future extensions via memory mechanisms or trajectory modeling.

Overall, this analysis validates the versatility of the proposed tracker under a wide range of conditions, and suggests specific avenues, such as clutter robustness and resolution enhancement, for future improvement.

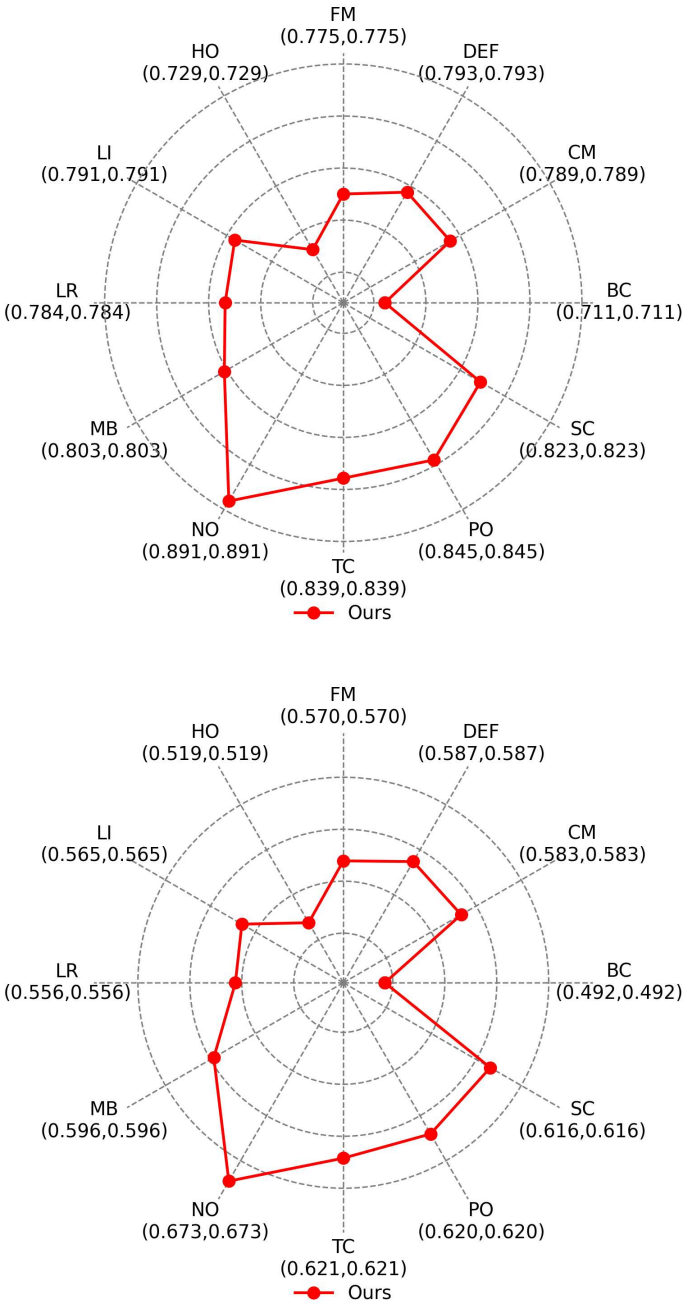


Figure 4.5: Attribute-based evaluation on RGBT234. (a) Max Precision Rate (MPR). (b) Max Success Rate (MSR).

Fusion Method	Params (M)	+MACS (G) (Fusion Block)	FPS (GPU)	RGBT234		GTOT	
				MPR	MSR	MPR	MSR
Sigmoid (Proposed)	3.926	–	122	**0.8063**	**0.5890**	**0.8949**	**0.7467**
Addition	3.926	+0.060	128	0.7741	0.5583	0.8614	0.7139
Softmax	3.926	+0.130	126	0.7860	0.5737	0.8798	0.7265
Concat + Conv	3.926	+0.100	125	0.7809	0.5642	0.8411	0.6863
Input Attention	3.926	+0.100	126	0.7810	0.5660	0.8801	0.7312

Table 4.4: Comparison of final-stage fusion strategies after cross-modal transformer. The proposed sigmoid-based channel-wise weighting achieves the highest accuracy across benchmarks with minimal parameters and real-time speed.

4.2.4 Ablation Study on Final Fusion Strategies

To provide more insights into the final channel-wise weighted fusion mechanism in Eq. 9, we conducted an ablation study to evaluate alternative strategies. As noted, our proposed design applies fixed per-channel weights followed by a sigmoid activation. While efficient, this approach may be limited in adaptability. To test its impact and explore potential improvements, we replaced the fusion block with four variants: element-wise addition, softmax-weighted fusion, 1×1 convolution over concatenated features, and an input-conditioned attention mechanism.

Each method was implemented as a lightweight plug-in to the final fusion stage of our model, following the transformer-based cross-modal reasoning described in Section 3.4. Crucially, all variants introduce negligible computational cost and parameter overhead, less than 130K additional parameters in all cases, making them suitable for real-time operation. Since the fusion block comes after transformer-based integration, it is intentionally kept lightweight to preserve efficiency.

Table 4.4 summarizes the results. Notably, all variants maintain real-time speed above 120 FPS and introduce no meaningful difference in GMACs. Accuracy varies modestly across fusion types. The proposed sigmoid-based fusion achieves the highest performance overall, with 0.8063 MPR and 0.589 MSR on RGBT234, and 0.8838 MPR and 0.7409 MSR on GTOT. Input-attention fusion ranks second, with slightly lower RGBT234 scores but comparable performance on GTOT (0.8801 MPR, 0.7312 MSR). Other strategies, including addition and softmax weighting, also perform competitively but fall short in one or more metrics.

We attribute the superior performance of the sigmoid-based fusion to its ability to softly gate the contribution of each channel without introducing inter-channel dependencies. Empirically, we

found that applying a per-channel sigmoid allows the model to suppress noisy or less informative features while preserving dominant activations from each modality. Unlike softmax, which forces a competition across channels, or addition, which lacks adaptive weighting, the sigmoid function enables independent, bounded modulation in a stable and efficient manner.

These results support the continued use of sigmoid-weighted fusion as an effective and efficient design choice. While slightly more adaptive alternatives like input-conditioned attention can approach similar performance, the gains are marginal. Given the role of this layer as a lightweight summarizer after transformer-based fusion, additional complexity appears unnecessary. This validates our decision to favor simplicity and efficiency at the final stage of the architecture.

Chapter 5

Analysis

5.1 Overview

Building on the lightweight and competitive performance of the proposed MobileViTv2-based RGB-T tracker, in this chapter, we explored how popular paradigms in multimodal tracking literature behave when ported to this efficient backbone. Specifically, we implemented two widely studied methods, prompt learning and Siamese modeling, within the same MobileVision framework. These variants were designed to test the adaptability and modularity of multimodal tracking under constrained compute budgets. Although neither method outperforms our proposed design in Chapter 3, they offer valuable insights into design trade-offs and serve as baselines for future improvement.

Section 5.2 presents a prompt-based adaptation mechanism for frozen RGB backbones, while Section 5.3 introduces a dual-branch Siamese architecture with late fusion. Their corresponding experimental results are analyzed in subsections 5.2.5 and 5.3.1, respectively.

5.2 Prompt Learning for RGB-T Tracking

To investigate the adaptability of pretrained RGB trackers to the RGB-T setting, we integrate a prompt learning mechanism into our architecture. This approach allows us to reuse a strong RGB-trained backbone while introducing minimal trainable parameters for modality adaptation, aligning

with both parameter efficiency and interpretability goals.

5.2.1 Motivation for Prompt Learning

Prompt learning originates in Natural Language Processing (NLP), where small sets of learnable tokens, called prompts, are prepended to input sequences of frozen language models to condition their behavior, primarily used for solving downstream tasks (Liu et al., 2023). In the context of visual object tracking, ViPT (Zhu et al., 2023a) extended this concept to the RGB-T domain by training prompt tokens to inject thermal-aware information into frozen RGB-only trackers. Inspired by this idea, we incorporate prompt-based multimodal adaptation into our MobileViTv2-based architecture to evaluate its capacity for RGB-T generalization with minimal supervision.

5.2.2 Pretrained Backbone and Prompt Learning Setup

We initialize our model using the pretrained RGB weights from SMAT (Gopal and Amer, 2024), a lightweight transformer-based tracker trained on large-scale RGB datasets such as GOT-10k (Huang et al., 2021), LaSOT (Fan et al., 2019), and TrackingNet (Muller et al., 2018). To enable multimodal adaptation without altering the learned RGB representations, we freeze all pretrained weights and insert learnable prompt modules into the separable mixed-attention layers of the mmMobileViT blocks at Layer 3 and Layer 4. The complete architecture of this prompt-integrated backbone is shown in Figure 5.1.

5.2.3 Prompt Integration into Separable Attention

Let $\mathbf{F}^{\mathrm{RGB}}, \mathbf{F}^{\mathrm{IR}} \in \mathbb{R}^{C\times H\times W}$ denote the feature maps from the RGB and IR branches of either the template or search input. These feature maps are first processed by the local representation module (a series of convolutional layers), and then unfolded into non-overlapping patch tokens of size $p \times p$. This yields a sequence of patch embeddings:

$$\mathbf{T} \in \mathbb{R}^{B\times N\times d}, \quad \text{where } N = \frac{H \cdot W}{p^2}, \quad d = \text{embedding dimension.} \tag{12}$$

This token sequence $\mathbf{T}$ is subsequently passed through the Separable Mixed-Attention layers

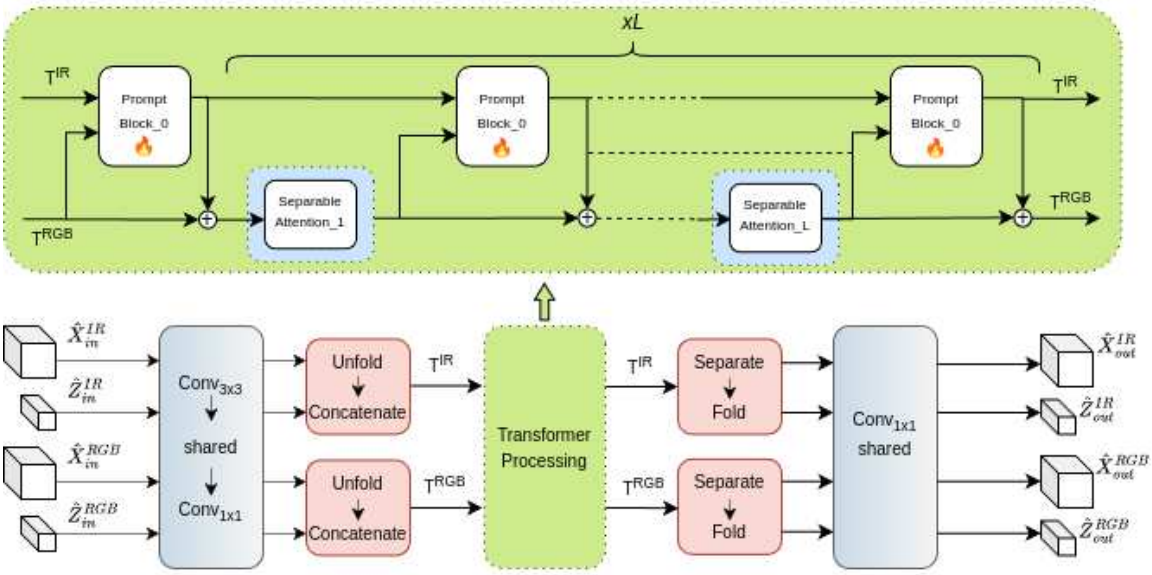


Figure 5.1: Illustration of the prompt-adapted **mmMobileViT** block used for RGB-T tracking. During each transformer stage, RGB and IR features are unfolded and used to generate modality-aware prompts via convolutional Prompt Blocks (Zhu et al., 2023a), which are added to the frozen RGB token stream before each Separable Mixed-Attention layer. The blocks highlighted in orange indicate the only learnable modules during the prompt learning stage, while all other components remain frozen. This design enables efficient cross-modal adaptation with minimal parameter overhead.

described by equation 4, which perform efficient inter-token interactions with linear complexity. Prior to each transformer block, we inject modality-aware prompts as detailed below.

Before applying separable attention, we generate a prompt tensor $\mathbf{P} \in \mathbb{R}^{N \times d}$ using a convolutional Prompt Block. This prompt is computed from the concatenated RGB and IR feature maps as:

$$\mathbf{F}_{\text{concat}} = \text{Concat}(\mathbf{F}^{\text{RGB}}, \mathbf{F}^{\text{IR}}) \in \mathbb{R}^{2C \times H \times W}, \tag{13}$$

$$\mathbf{P} = \text{PromptBlock}(\mathbf{F}_{\text{concat}}), \tag{14}$$

followed by unfolding and normalization. The prompt is then added to the RGB token sequence:

$$\mathbf{T}_{\text{aug}} = \mathbf{T} + \mathbf{P}. \tag{15}$$

This augmented sequence is passed to the transformer for global reasoning guided by the prompt, allowing modality-aware adaptation without modifying the frozen RGB backbone.

This additive prompting scheme is depicted in Figure 5.1, which shows the position and recurrence of Prompt Blocks alongside transformer layers in Layer 3 and Layer 4.

Each Prompt Block follows a lightweight dual-branch design, composed of three 1×1 convolution layers and a soft attention mechanism called *Fovea*. Given the concatenated feature map $\mathbf{F}_{\text{concat}} \in \mathbb{R}^{2C \times H \times W}$, we first split it along the channel dimension:

$$\mathbf{F}_{\text{IR}} = \mathbf{F}_{\text{concat}}[: \tfrac{C}{2}, :], \tag{16}$$

$$\mathbf{F}_{\text{RGB}} = \mathbf{F}_{\text{concat}}[\tfrac{C}{2} :, :], \tag{17}$$

where $\mathbf{F}_{\text{IR}}$ and $\mathbf{F}_{\text{RGB}}$ represent the modality-specific inputs. These are projected to a hidden space via separate convolutions:

$$\mathbf{H}_{\text{IR}} = \text{Conv}_{\text{IR}}(\mathbf{F}_{\text{IR}}), \tag{18}$$

$$\mathbf{H}_{\text{RGB}} = \text{Conv}_{\text{RGB}}(\mathbf{F}_{\text{RGB}}). \tag{19}$$

The RGB branch is then weighted using soft spatial attention:

$$\hat{\mathbf{H}}_{\text{RGB}} = \text{Softmax}(\mathbf{H}_{\text{RGB}}) \odot \mathbf{H}_{\text{RGB}}. \tag{20}$$

The final prompt feature is computed via addition and projected back:

$$\mathbf{P} = \text{Conv}_{1\times1}(\hat{\mathbf{H}}_{\text{RGB}} + \mathbf{H}_{\text{IR}}). \tag{21}$$

This process allows the prompt to encode cross-modal dependencies in a lightweight, interpretable manner.

5.2.4 Iterative Prompt Refinement Across Layers

Our design supports multiple separable mixed-attention layers per mmMobileViT block. After each separable mixed-attention layer, the updated RGB tokens are folded back to spatial maps, and new prompts are generated based on the intermediate outputs. These refined prompts are unfolded

again and added to the tokens before the next transformer block. This recursive mechanism allows prompts to adapt progressively to deeper semantic features:

$$\mathbf{F}_t = \text{Folding}(\mathbf{T}_t), \tag{22}$$

$$\mathbf{P}_{t+1} = \text{PromptBlock}(\text{Concat}(\mathbf{F}_t, \mathbf{F}^{\text{IR}})), \tag{23}$$

$$\mathbf{T}_{t+1} = \mathbf{T}_t + \text{Unfolding}(\mathbf{P}_{t+1}). \tag{24}$$

This iterative prompting continues across all transformer layers in the mmMobileViT block, enabling a dynamic interplay between RGB tokens and modality-aware prompts. Since only the Prompt Blocks are trained, the overhead is minimal and the pretrained RGB weights remain unchanged.

It is important to note that our prompt-learning model preserves the same architecture and layer-sharing strategy as the original mmMobileViTv2 backbone. The only change lies in the inputs to transformer blocks, which now receive adaptively generated prompts derived from both RGB and thermal features. This design maintains structural consistency while leveraging thermal signals to guide the transformer's attention through prompt injection.

5.2.5 Prompt Learning Results on LasHeR

To evaluate the effectiveness of prompt-based modality adaptation, we compare our prompt-integrated model against both the original RGB-only tracker and the ViPT (Zhu et al., 2023a) method on the LasHeR (Li et al., 2021) benchmark. The results are summarized in Table 5.1.

Model	**#Params**	**GMACs**	**PR**	**nPR**	**SR**	**FPS (GPU)**
base_rgb (SMAT)	3.767 M	–	0.5498	0.5124	0.3996	148.8
Proposed (original)	3.926 M	4.35 G	**0.6026**	**0.5674**	**0.4729**	121.92
Prompt-adapted (ours)	5.5 M (1.5 M)	4.83 G	0.5610	0.5233	0.4130	83.47
ViPT (Zhu et al., 2023a)	93 M	22.95 G	0.651	–	0.525	39.5

Table 5.1: Performance comparison of prompt learning on the LasHeR dataset. Our model adds 1.5M trainable prompt parameters on top of a 3.926M frozen RGB backbone. ViPT results are from their original paper.

Our results show that prompt learning yields a modest performance gain over the RGB-only

baseline, improving PR from 0.5498 to 0.5610 and SR from 0.3996 to 0.4130. However, this improvement is less substantial than that reported by ViPT (Zhu et al., 2023a), which benefits from a deeper transformer backbone (ViT-Base (Dosovitskiy et al., 2021), 12 encoder blocks) and a much larger model capacity. ViPT inserts prompts into every transformer layer, enabling fine-grained multimodal conditioning throughout the network.

In contrast, our model uses a hybrid convolution-transformer backbone (mmMobileViTv2), where only six Separable Attention layers are available for prompt injection—two in Layer3 and four in Layer4 (see Figure 5.1). This limited injection depth inherently restricts the influence of the prompts.

Moreover, our Prompt Blocks operate on spatial feature maps, not token sequences, and thus require repeated folding and unfolding operations across transformer depths. This increases computational overhead and explains the drop in speed, from 121.92 FPS in the original model to 83.47 FPS in the prompt-adapted version, despite adding 1.5M learnable parameters.

In summary, our prompt learning design achieves a favorable trade-off: it allows multimodal adaptation while keeping the original RGB tracker completely frozen. The relatively small gains in accuracy highlight the architectural constraints of applying prompting to lightweight backbones, motivating future exploration of more efficient prompt injection mechanisms for shallow transformer hierarchies.

5.3 Siamese-Based Tracker Architecture

Siamese networks are a widely adopted design paradigm in multimodal object tracking (Zhang et al., 2024; Wang et al., 2023), offering simplicity, modularity, and interpretability. As discussed in Chapter 2, Siamese-based RGB-T trackers typically process each modality independently and perform fusion at a later stage, in contrast to transformer-based models that embed fusion within the backbone. While this modular approach avoids the complexity of dense cross-modal attention, it also limits the extent of interaction between modalities during early feature learning. To investigate this trade-off and to isolate the role of cross-modal fusion, we implement a Siamese variant of our RGB-T tracker in which RGB and infrared (IR) modalities are processed by separate backbone

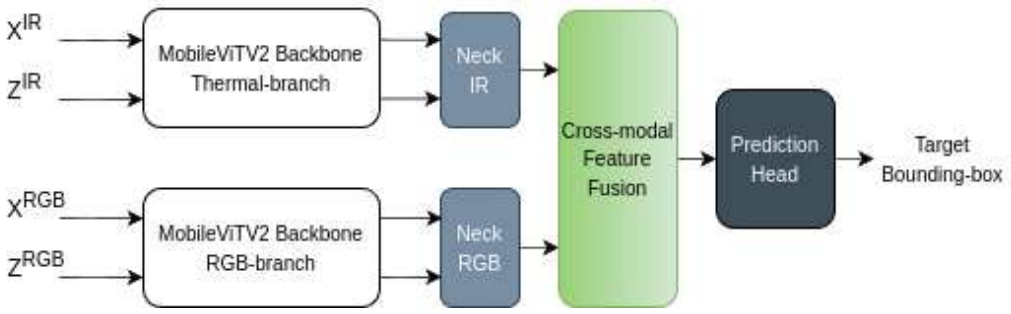


Figure 5.2: Overview of the Siamese-based RGB-T tracker architecture. Each modality (RGB and IR) is processed by its own backbone and neck module. Each backbone branch uses the same MobileViTv2 architecture described in Section 3.2 and visualized in Figure 3.2, but without inter-modal interaction in Layer 4. Instead of fusing features within the backbone, this variant performs cross-modal integration using a single cross-modal fusion transformer, as described in Section 3.4, before passing the output to the prediction head.

branches, with all interaction deferred to a dedicated cross-modal fusion transformer.

This design differs significantly from the MobileViTv2-based backbone described in Section 3.2. In the original architecture, modality-specific features from the RGB and IR branches are processed by shared mmMobileViT blocks. Layer 3 performs intra-modal reasoning, while Layer 4 explicitly fuses RGB and IR token sequences through inter-modal separable mixed-attention (see Figure 3.2). This enables joint context modeling across modalities within the backbone itself. By contrast, the Siamese variant disables this early fusion mechanism: Layer 4 in each branch now mirrors Layer 3 in performing only intra-modal attention, meaning that RGB and IR features are completely disentangled throughout the entire backbone.

Each modality has its own dedicated backbone pipeline, composed of convolutional stem layers, MobileNetV2 inverted residual blocks, and mmMobileViT blocks for Layer 3 and Layer 4. These branches process the template and search regions independently and apply separable mixed-attention within each modality, capturing global dependencies between the template and search inputs without any inter-modal token mixing. As a result, the Siamese backbone produces two modality-specific feature maps: $\hat{X}^{\mathrm{RGB}}, \hat{Z}^{\mathrm{RGB}}$ for RGB and $\hat{X}^{\mathrm{IR}}, \hat{Z}^{\mathrm{IR}}$ for IR.

The outputs from each branch are passed to their respective neck modules, as described in Section 3.3. Each neck module performs pixel-wise cross-correlation (PW-XCorr) independently per modality, yielding $\mathbf{F}^{\mathrm{RGB}}$ and $\mathbf{F}^{\mathrm{IR}}$, two spatially aligned but modality-specific feature maps.

These serve as the input to the Cross-Modal Fusion Transformer, which is responsible for all cross-modal reasoning. Importantly, this transformer is not shared across branches or applied at multiple levels; rather, it operates as a single joint module that fuses the final RGB and IR representations produced by the Siamese necks.

Its structure and operation are identical to the one detailed in Section 3.4, with tokenization, separable mixed-attention based fusion, and learnable adaptive weighting for channel-wise integration (see Figure 3.3). Notably, since the inputs from both modalities are independently tokenized and passed through attention and feedforward sub-blocks, the transformer internally applies double the number of attention layers, one for each modality, before fusion. This expanded structure allows symmetric global reasoning on both streams prior to joint interaction, at the cost of increased parameter count and inference latency.

By deferring all cross-modal interaction to this final transformer stage, the Siamese design allows us to directly analyze the importance of cross-modal attention for RGB-T tracking. Specifically, it enables a clean architectural ablation: by comparing with the unified mmMobileViTv2-based model, which includes inter-modal attention in Layer 4, we can evaluate how the absence of early fusion affects final tracking accuracy. Experimental results in Chapter 4 confirm that while the Siamese model retains strong intra-modal modeling capacity, early inter-modal fusion in Layer 4 leads to higher performance, highlighting the critical role of cross-modal interaction within the backbone.

A summary of the Siamese variant architecture is illustrated in Figure 5.2, which shows the parallel RGB and IR backbone pipelines, independent neck modules, and the shared fusion and prediction head modules. This design offers a modular and interpretable alternative to the original unified backbone, while also serving as an effective baseline for analyzing the contributions of progressive fusion.

5.3.1 Analysis of Siamese Tracker Results

Table 5.2 presents the performance of the Siamese-based RGB-T tracker on the LasHeR dataset. Compared to the RGB-only baseline (*base_rgb*), the Siamese architecture achieves noticeable gains

Model	#Params (M)	MACs (G)	LasHeR PR	LasHeR nPR	LasHeR SR	FPS (GPU)
Proposed	3.926M	4.35G	**0.6026**	**0.5674**	**0.4729**	**121.92**
Siamese Variant	6.100M	4.35G	0.5716	0.5410	0.4586	109.22

Table 5.2: Performance comparison of the Siamese-based RGB-T tracker on the LasHeR dataset. The Siamese model uses independent backbones and necks for each modality and fuses features at the final stage. Despite having more parameters, its GMACs are comparable due to the use of lightweight separable attention and similar backbone structure in both models (see Eq. 4 in Section 3.2).

in tracking accuracy (PR: +0.0218, nPR: +0.0184, SR: +0.0134), highlighting the benefit of incorporating thermal information, even without inter-modal fusion in the backbone.

However, relative to the unified mmMobileViTv2 model, the Siamese variant consistently underperforms. Precision drops by 0.0310, normalized precision by 0.0264, and success rate by 0.0233. These results confirm the advantage of early cross-modal interaction: as discussed in Section 3.2, inter-modal attention in Layer 4 of the mmMobileViTv2 backbone allows the model to capture shared contextual dependencies between RGB and IR modalities before fusion, leading to richer representations.

Interestingly, the Siamese variant has significantly more parameters (6.1M vs. 3.926M) due to the use of duplicated backbone modules for each modality. However, its GMACs are approximately the same (4.35G), which may initially seem surprising given the added modules. This can be explained by the use of separable attention, which has linear complexity in the sequence length and embedding dimension, as shown in equation 4. In the mmMobileViTv2 model, the RGB and IR token sequences are concatenated in Layer 4, doubling the input feature dimension from d to $2d$, which increases compute per attention layer to $\mathcal{O}(k \cdot 2d)$. In contrast, the Siamese model processes RGB and IR streams independently using attention with input dimension d, but does so twice, once per modality, leading to a similar overall compute cost.

Thus, although both models exhibit comparable GMACs due to the linear scaling of separable attention (Mehta and Rastegari, 2023), the Siamese model incurs a higher parameter count from backbone duplication. Additionally, its inference speed (109.22 FPS vs. 121.92 FPS) is lower, indicating increased latency caused by more parameters in the non-shared architecture.

In summary, the Siamese variant offers a modular and interpretable baseline for studying RGB-T

fusion. However, its lack of progressive inter-modal reasoning results in reduced accuracy, despite a similar GMACs profile. These results reinforce the importance of fusing multimodal cues throughout the network and demonstrate that backbone-level fusion, even when modestly more complex, leads to stronger performance under comparable conditions.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis introduced a lightweight RGB-T tracking framework that balances performance and efficiency by integrating MobileViTv2 with progressive cross-modal fusion. Motivated by the growing need for real-time and resource-constrained multimodal applications, our work contributes to the RGB-T tracking literature by prioritizing compact design, modularity, and adaptability.

We addressed the inherent challenges of RGB-T tracking, such as illumination variation, partial occlusion, and modality misalignment, by leveraging separable attention in place of standard multi-head attention, significantly reducing computational overhead. The proposed model combines intra-modal reasoning and inter-modal fusion within the same architecture, resulting in a unified yet efficient transformer backbone. With only 3.93 million parameters, our tracker achieves 122 FPS and demonstrates competitive accuracy on three challenging benchmarks: LasHeR, RGBT234, and GTOT.

To further investigate fusion timing and modality adaptation, we explored two architectural extensions within the same lightweight framework. The first integrates prompt learning into separable mixed-attention layers to enable RGB-T adaptation using a frozen RGB-only tracker. The second adopts a Siamese architecture with separate modality-specific backbones and defers fusion to a final cross-modal transformer. These two variants were designed to evaluate the trade-offs between modularity, fusion timing, and parameter sharing, and their comparative analysis offers insights to

our model design.

Our results validate that a compact architecture can deliver strong performance in both speed and accuracy, reinforcing the feasibility of real-time RGB-T tracking on edge platforms. Moreover, the thesis offers a structured framework for evaluating design choices, shared vs. separate backbones, early vs. late fusion, frozen vs. trainable adaptation, which can serve as a guideline for future multimodal tracking research.

6.2 Future Work

While this thesis lays the groundwork for lightweight RGB-T object tracking using Mobile Vision Transformers, several directions remain open for exploration:

- **Reducing Fusion Overhead:** Despite the linear complexity of separable attention, inter-modal fusion through token concatenation increases sequence length and leads to high MACs. Future research could explore low-rank token interactions, sparse fusion modules, or learned token selection to minimize redundancy while preserving cross-modal context.

- **Generalization to Other Lightweight Backbones:** Our framework can be extended to other compact transformers such as LeViT (Graham et al., 2021), EfficientFormer (Li et al., 2022). Studying how separable mixed-attention and progressive fusion behave under different backbones can broaden the applicability of our design principles.

- **Edge Deployment and Latency Profiling:** While our model achieves high FPS on high-end GPUs, practical deployment requires evaluation on mobile platforms (e.g., Jetson Nano, ARM SoCs). Future work will include latency profiling, memory benchmarking, and hardware-aware pruning for embedded use.

- **Low-Rank Thermal Fine-Tuning:** Instead of training full multimodal models from scratch, future work can explore fine-tuning RGB-pretrained trackers using only a small amount of thermal infrared data. Leveraging low-rank adaptation techniques (e.g., LoRA or adapter layers) from NLP (Hu et al., 2022) can enable efficient modality transfer with minimal parameter updates, offering a lightweight and scalable alternative to prompt learning for RGB-T

adaptation.

In conclusion, this thesis contributes a practical and extensible foundation for real-time RGB-T tracking. The proposed architecture advances the state-of-the-art in terms of inference speed with compact multimodal design, but also opens multiple paths for future innovation at the intersection of vision transformers, modality fusion, and real-world deployment.

Chapter 7

Appendix A: Experiment on RGB-D

A.1 Background on RGB-D Tracking

While our primary focus is on RGB-T tracking, extending models to other modality pairs such as RGB-D has been an active area of research. Several recent works have aimed to design *efficient multimodal trackers* that can generalize across different sensory inputs.

SUTrack-T224 (AAAI 2025) (Chen et al., 2025): Built on the HiViT backbone, SUTrack adopts a unified framework that supports multiple modality pairs (e.g., RGB-T, RGB-D, RGB-Event), enabling broad applicability across multimodal tracking tasks. Its hierarchical vision transformer and efficient fusion strategy make it competitive in both RGB-T and RGB-D (see Related Works and the main results table for more detail).

EMTrack (TCSVT 2025) (Liu et al., 2024a): Based on a ViT backbone, EMTrack emphasizes modality-invariant embeddings within an efficient multimodal design. Although primarily optimized for RGB-T, its architectural flexibility allows adaptation to RGB-D and other modality combinations (see Related Works and the main table).

These methods indicate that while specialized fusion designs for RGB-T can be highly effective, generalizable multimodal frameworks are increasingly promising for handling RGB-D as well.

Model	#Params (M)	DepthTrack Pr	Re	F	FPS (GPU)
Our model	3.926M	47.5	49.7	48.6	**120**
SUTrack-T224 (AAAI 2025, HiViT)	22M	61.2	62.1	61.7	100
EMTrack (TCSVT 2025, ViT)	16M	58.0	58.5	61.4	83.8

Table 7.1: Comparison of our RGB-D extension with representative multimodal trackers on DepthTrack. *SUTrack* and *EMTrack* also appear in the main results table and are discussed in Related Works.

A.2 Experiment with Our Model on RGB-D

To assess the generalization capability of our framework, we replaced the thermal stream with depth input and trained the model on the **DepthTrack** benchmark (Yan et al., 2021c), which contains 200 sequences across diverse indoor and outdoor scenarios. Using the same training pipeline as in RGB-T, we directly trained our MobileViTv2-based multimodal backbone on RGB-D pairs and evaluated on the DepthTrack test split.

Our model achieved a **precision of 47.5**, **recall of 49.7**, and **F-score of 48.6** at **120 FPS**. Compared with our strong performance on RGB-T datasets, these results are weaker, which we attribute to: (i) **dataset scale** (DepthTrack is much smaller than LasHeR, 200 vs. 1224 sequences), limiting robust cross-modal learning; (ii) **depth modality limitations** (noise, missing values, and low texture, especially outdoors), reducing complementarity with RGB; and (iii) a **fusion design mismatch**, since our strategy was tailored to RGB-T where thermal provides illumination-robust appearance cues, while depth primarily encodes geometry.

Nevertheless, the experiment shows that our lightweight design adapts to RGB-D and maintains real-time efficiency. Fully exploiting depth likely requires modality-specific architectural choices, as seen in specialized RGB-D trackers such as SUTrack and EMTrack.

A.3 Quantitative Comparison on DepthTrack

Since our main thesis focus is RGB-T, we leave a fuller exploration of RGB-D adaptations and fusion designs to future work.

Bibliography

Y. Bai, Z. Zhao, Y. Gong, and X. Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19048–19057, June 2024.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019.

P. Blatter, M. Kanakis, M. Danelljan, and L. Van Gool. Efficient visual tracking with exemplar transformers. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 1571–1581, 2023.

B. Cao, J. Guo, P. Zhu, and Q. Hu. Bi-directional adapter for multimodal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 927–935, 2024.

L. Chen, B. Zhong, Q. Liang, Y. Zheng, Z. Mo, and S. Song. Top-down cross-modal guidance for robust rgb-t tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

X. Chen, B. Kang, W. Geng, J. Zhu, Y. Liu, D. Wang, and H. Lu. Sutrack: Towards simple and

unified single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2239–2247, 2025.

R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5369–5378, 2019. doi: 10.1109/CVPR.2019.00552.

G. Y. Gopal and M. Amer. Mobile vision transformer-based visual object tracking. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. URL https://papers.bmvc2023.org/0800.pdf.

G. Y. Gopal and M. A. Amer. Separable self and mixed attention transformers for efficient object tracking. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6708–6717, 2024.

B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021.

L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19079–19091, 2024.

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*, pages 26551–26561, 2024.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

X. Hu, Y. Tai, X. Zhao, C. Zhao, Z. Zhang, J. Li, B. Zhong, and J. Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3581–3589, 2025a.

X. Hu, B. Zhong, Q. Liang, L. Shi, Z. Mo, Y. Tai, and J. Yang. Adaptive perception for unified visual multi-modal object tracking. *IEEE Transactions on Artificial Intelligence*, 2025b.

L. Huang, X. Zhao, and K. Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1562–1577, 2021. doi: 10.1109/TPAMI.2019.2957464.

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*, pages 13630–13639, 2023.

B. Kang, X. Chen, D. Wang, H. Peng, and H. Lu. Exploring lightweight hierarchical vision transformers for efficient visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9612–9621, 2023.

N. Kitaev, Ł. Kaiser, and A. Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. Cehovin Zajc, O. Drbohlav, A. Lukezic, A. Berg, A. Eldesokey, J. Kapyla, G. Fernandez, A. Gonzalez-Garcia, A. Memarmoghadam, A. Lu, A. He, A. Varfolomieiev, A. Chan, A. Shekhar Tripathi, A. Smeulders, B. Suraj Pedasingu, B. Xin Chen, B. Zhang, B. Wu, B. Li, B. He, B. Yan, B. Bai, B. Li,

B. Li, B. Hak Kim, and B. Hak Ki. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L. Č. Zajc, A. Lukežič, O. Drbohlav, L. He, Y. Zhang, S. Yan, J. Yang, G. Fernández, A. Hauptmann, A. Memarmoghadam, Á. García-Martín, A. Robinson, A. Varfolomieiev, A. H. Gebrehiwot, B. Uzun, B. Yan, B. Li, C. Qian, C.-Y. Tsai, C. Micheloni, D. Wang, F. Wang, F. Xie, F. J. Lawin, F. Gustafsson, G. L. Foresti, G. Bhat, G. Chen, H. Ling, H. Zhang, H. Cevikalp, H. Zhao, H. Bai, H. C. Kuchibhotla, H. Saribas, H. Fan, H. Ghanei-Yakhdan, H. Li, H. Peng, H. Lu, H. Li, J. Khaghani, J. Bescos, J. Li, J. Fu, J. Yu, J. Xu, J. Kittler, J. Yin, J. Lee, K. Yu, K. Liu, K. Yang, K. Dai, L. Cheng, L. Zhang, L. Wang, L. Wang, L. Van Gool, L. Bertinetto, M. Dunnhofer, M. Cheng, M. M. Dasari, N. Wang, N. Wang, P. Zhang, P. H. S. Torr, Q. Wang, R. Timofte, R. K. S. Gorthi, S. Choi, S. M. Marvasti-Zadeh, S. Zhao, S. Kasaei, S. Qiu, S. Chen, T. B. Schön, T. Xu, W. Lu, W. Hu, W. Zhou, X. Qiu, X. Ke, X.-J. Wu, X. Zhang, X. Yang, X. Zhu, Y. Jiang, Y. Wang, Y. Chen, Y. Ye, Y. Li, Y. Yao, Y. Lee, Y. Gu, Z. Wang, Z. Tang, Z.-H. Feng, Z. Mai, Z. Zhang, Z. Wu, and Z. Ma. The eighth visual object tracking vot2020 challenge results. In A. Bartoli and A. Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 547–601, Cham, 2020. Springer International Publishing. ISBN 978-3-030-68238-5.

H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

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*, 25(12):5743–5756, 2016. doi: 10.1109/TIP.2016.2614135.

C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1856–1864, New York, NY, USA, 2017. Association for Computing

Machinery. ISBN 9781450349062. doi: 10.1145/3123266.3123289. URL https://doi.org/10.1145/3123266.3123289.

C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2019.106977. URL https://www.sciencedirect.com/science/article/pii/S0031320319302808.

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*, 31:392–404, 2021.

Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren. Efficientformer: Vision transformers at mobilenet speed. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 12934–12949. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5452ad8ee6ea6e7dc41db1cbd31ba0b8-Paper-Conference.pdf.

Y. Li, B. Wang, X. Wu, Z. Liu, and Y. Li. Lightweight full-convolutional siamese tracker. *Knowledge-Based Systems*, 286:111439, 2024.

Y. Li, B. Wang, and Y. Li. Lightfc-x: Lightweight convolutional tracker for rgb-x tracking, 2025. URL https://arxiv.org/abs/2502.18143.

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*, 2024a.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), Jan. 2023. ISSN 0360-0300. doi: 10.1145/3560815. URL https://doi.org/10.1145/3560815.

Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024b.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

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*, pages 9291–9300, 2024.

A. Lu, W. Wang, C. Li, J. Tang, and B. Luo. Rgbt tracking via all-layer multimodal interactions with progressive fusion mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5793–5801, 2025.

S. Mehta and M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vh-0sUt8HlG.

S. Mehta and M. Rastegari. Separable self-attention for mobile vision transformers. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=tBl4yBEjKi.

M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018.

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.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals

and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

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*, 2024.

G. Wang, Q. Jiang, X. Jin, Y. Lin, Y. Wang, and W. Zhou. Siamtdr: Time-efficient rgbt tracking via disentangled representations. *IEEE Transactions on Industrial Cyber-Physical Systems*, 1: 167–181, 2023.

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*, volume 38, pages 5436–5444, 2024a.

Q. Wang, Y. Bai, and H. Song. Middle fusion and multi-stage, multi-form prompts for robust rgb-t tracking. *Neurocomputing*, 596:127959, 2024b.

S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

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*, pages 19156–19166, 2024.

Y. Xiao, J. Zhao, A. Lu, C. Li, B. Yin, Y. Lin, and C. Liu. Cross-modulated attention transformer for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8682–8690, 2025.

B. Yan, H. Peng, K. Wu, D. Wang, J. Fu, and H. Lu. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15189, 2021a.

B. Yan, X. Zhang, D. Wang, H. Lu, and X. Yang. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5289–5298, June 2021b.

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*, pages 10725–10733, 2021c.

B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 341–357, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20047-2.

J. Zhai, Z. Cheng, W. Zhang, D. Zhu, and W. Yang. Efficient object tracking on edge devices with mobiletrack. *Journal of Visual Communication and Image Representation*, 100:104126, 2024.

J. Zhang, Y. Qin, S. Fan, Z. Xiao, and J. Zhang. Siamtfa: Siamese triple-stream feature aggregation network for efficient rgbt tracking. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

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*, pages 5404–5413, 2023a.

X. Zhang, Y. Tian, L. Xie, W. Huang, Q. Dai, Q. Ye, and Q. Tian. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=3F6I-0-57SC.

X. Zhou, Q. Wang, H. Wu, Z. Zhang, X. Li, and T. Xu. Robust rgb-t tracking with cross-modal alignment transformer. In *Image and Graphics (ICIG)*, pages 20–32. Springer, 2023. doi: 10.1007/978-3-031-57037-7_2.

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*, pages 9516–9526, 2023a.

S. Zhu, Y. Song, C. Liang, C. Li, X. Yu, Y. Zhang, and Z. Wang. Towards alignment-robust rgb-t tracking: Benchmark and baseline. *IEEE Transactions on Instrumentation and Measurement*, 2023b. doi: 10.1109/TIM.2023.3282644.