

DEEP LEARNING-BASED OBJECT TRACKING SYSTEM BY USING VISUAL AND
THERMAL INFRARED FUSION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ABBAS TÜRKOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF MODELING AND SIMULATION

AUGUST 2023

Deep Learning-Based Object Tracking System By Using Visual and Thermal Infrared Fusion

submitted by **ABBAS TÜRKOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Modeling and Simulation Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Elif Sürer
Head of Department, **Modeling and Simulation**

Assoc. Prof. Dr. Elif Sürer
Supervisor, **Modeling and Simulation**

Assoc. Prof. Dr. Erdem Akagündüz
Co-supervisor, **Modeling and Simulation**

Examining Committee Members:

Prof. Dr. Alptekin Temizel
Modeling and Simulation Department, METU

Assoc. Prof. Dr. Elif Sürer
Modeling and Simulation Department, METU

Assist. Prof. Dr. İrem Ülkü
Computer Engineering Department, Ankara University

Date: 28.08.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Abbas Türkođlu

Signature :

ABSTRACT

DEEP LEARNING-BASED OBJECT TRACKING SYSTEM BY USING VISUAL AND THERMAL INFRARED FUSION

Türkoğlu, Abbas

M.S., Department of Modeling and Simulation

Supervisor: Assoc. Prof. Dr. Elif Sürer

Co-Supervisor: Assoc. Prof. Dr. Erdem Akagündüz

August 2023, 51 pages

Object tracking in computer vision presents a formidable challenge, particularly when faced with adverse conditions like occlusion, variations in illumination, and motion blur. In recent years, deep learning has shown great promise for object tracking. However, the vast majority of deep learning-based object trackers use only visible band images. This limits their performance in challenging conditions, as thermal infrared electromagnetic waves can provide additional information about the object, such as its temperature. This thesis proposes a deep learning-based object tracking system that uses visual band (RGB) and thermal infrared (RGBT) fused images. The system consists of two main components: a feature extractor and a tracker. The feature extractor extracts features from both RGB and thermal infrared (TIR) images. The tracker then uses these features to track the object in the next frame. The proposed system is evaluated on the RGBT234 and LasHeR datasets, which are the mostly used RGBT object tracking datasets in the literature. The results show that the proposed system outperforms state-of-the-art RGB object trackers on the RGBT234 and LasHeR datasets.

Keywords: Deep learning, object tracking, RGBT object tracking, computer vision, thermal and visual fusion

ÖZ

TERMAL KIZILÖTESİ VE GÖRÜNÜR BANT KAYNAŞTIRMA KULLANARAK DERİN ÖĞRENME TABANLI NESNE TAKİP SİSTEMİ

Türkođlu, Abbas

Yüksek Lisans, Modelleme ve Simülasyon Bölümü

Tez Yöneticisi: Doç. Dr. Elif Sürer

Ortak Tez Yöneticisi: Doç. Dr. Erdem Akagündüz

Ağustos 2023, 51 sayfa

Nesne takibi, özellikle engeller (oklüzyon), aydınlatma deđişiklikleri ve hareket bulanıklığı gibi zorlu koşullarda bilgisayarla görünüm zorlu bir problem olarak karşımıza çıkmaktadır. Son yıllarda, derin öğrenme, nesne takibi için büyük umut vaat etmektedir. Ancak, çođu derin öğrenme tabanlı nesne takip modeli yalnızca görünür bant (RGB) görüntüleri kullanır. Bu nedenle zorlu koşullarda sınırlı performans gösterirler. Fakat termal kızılötesi elektromanyetik dalgalar (TIR) nesne sıcaklığı gibi ek bilgiler sağlayabilmektedirler. Bu tez, kaynaştırılmış RGBT görüntülerini kullanan derin öğrenme tabanlı bir nesne izleme sistemi önermektedir. Sistem iki ana bileşenden oluşur: öznitelik çıkarıcı ve takip edici. Öznitelik çıkarıcı, hem RGB hem de TIR görüntülerinden özellikler ve nirengi noktaları çıkarır. Takip edici daha sonra bir sonraki karedeki nesneyi izlemek için bu özellikleri kullanır. Önerilen sistem, yaygın olarak kullanılan RGBT nesne izleme veri kümeleri olan RGBT234 ve LasHeR veri kümeleri üzerinde değerlendirilmiştir. Sonuçlar, önerilen sistemin RGBT234 ve LasHeR veri setlerinde son teknoloji ürünü RGB nesne izleyicilerinden daha iyi performans sergilediđini göstermektedir.

Anahtar Kelimeler: Derin öğrenme, nesne takibi, RGBT nesne takibi, bilgisayarlı görü, termal görünür band kaynaştırma

To my family

ACKNOWLEDGMENTS

This thesis is partially supported by Middle East Technical University Scientific Research Projects Coordination Unit (METU-BAP), under the project title "Real-Time Visual Tracking System Based on Deep Learning Using Infrared and Visible Band Fusion" (Kızılötesi ve Görünür Bant Kaynaştırma Kullanarak Derin Öğrenme Tabanlı ve Gerçek Zamanlı Görsel Takip Sistemi - AGEP-704-2022-11000).

TABLE OF CONTENTS

ABSTRACT.....	iv
ÖZ.....	v
DEDICATION.....	vi
ACKNOWLEDGMENTS.....	vii
TABLE OF CONTENTS.....	viii
LIST OF TABLES.....	xi
LIST OF FIGURES.....	xii
LIST OF ABBREVIATIONS.....	xiv
CHAPTERS	
1 INTRODUCTION.....	1
1.1 RGBT Tracking Challenges.....	1
1.2 Problem Definition.....	2
1.3 Research Questions.....	3
1.4 Objectives of the Thesis.....	4
1.5 Contributions of the Study.....	4
1.6 Organization of the Thesis.....	5
2 RELATED WORK.....	7

2.1	RGBT Fusion Tracking In pre-Deep Learning Era	7
2.2	Deep Learning-based RGBT Fusion Object Tracking Methods	8
2.2.1	Convolutional Neural Network (CNN) Based RGBT Fusion Object Trackers	8
2.2.2	Siamese-based RGBT Trackers	11
2.2.3	Other Deep RGBT Fusion Object Trackers	13
2.3	Different Fusion Levels	13
2.3.1	Pixel-level Fusion	15
2.3.2	Feature-level Fusion	15
2.3.3	Decision-level Fusion	16
3	EXPERIMENTAL SETUP	17
3.1	Datasets and Benchmarks	17
3.1.1	VOT Challenges	17
3.1.2	GTOT	17
3.1.3	RGBT210	18
3.1.4	RGBT234	18
3.1.5	LasHeR	19
3.2	Evaluation Metrics	20
3.2.1	Precision Rate	20
3.2.2	Success Rate	20
3.3	Experimental Settings	21
4	METHODOLOGY	23
4.1	Network Architecture	23
4.2	Training	24

4.3	Online Tracking	25
5	RESULTS AND DISCUSSION.....	27
5.1	Evaluation on RGBT234 Dataset	27
5.2	Evaluation on LasHeR Dataset	36
5.3	Computational Efficiency	38
5.4	Experiments and Ablation Study	39
5.4.1	Experiments on Robustness of Our Model	42
5.5	Experiments on Other RGBT Tracking Models	42
6	CONCLUSION	45
	REFERENCES	47

LIST OF TABLES

Table 1	Results of Fusion Object Trackers on Different Benchmarks	14
Table 2	RGBT234 Dataset - List Of Attributes	18
Table 3	RGBT Fusion Object Tracking Benchmarks/Datasets and Their Contents	19
Table 4	RGBT234 dataset attribute-based PR and SR scores. The highest scores are shown in red color.	32
Table 5	Comparison of FPS Rates for RGB-T Tracking Models	39
Table 6	Results of Ablation Study Evaluation on RGBT234 Dataset	40
Table 7	Results of Experiments on RGBT234 Dataset	42
Table 8	Results of Experiments on LasHeR Dataset	42
Table 9	Results of Experiments on Other RGBT Tracking Models Tested on RGBT234 Dataset	43

LIST OF FIGURES

Figure 1	Pipeline of CNN-based RGBT Trackers	8
Figure 2	Pipeline of Siamese-based RGBT Trackers	12
Figure 3	Pixel-level Fusion Tracking	15
Figure 4	Feature-level Fusion Tracking	16
Figure 5	Decision-level Fusion Tracking	16
Figure 6	The architecture of the Enhanced Attribute-based Network.	24
Figure 7	The structure of the proposed “Enhanced Fusion Branch Module”.	25
Figure 8	The structure of the proposed “Enhanced Aggregation Module”	26
Figure 9	Precision Rate evaluation curve on the RGBT234 dataset.	28
Figure 10	Success Rate evaluation curve on the RGBT234 dataset.	28
Figure 11	Comparison of EANet to three state-of-the-art trackers on different sequences. (a) The basketballwalking sequence, which has challenge of heavy occlusion; (b) The man4 sequence, which also has challenge of heavy occlusion (c) The soccerinhand sequence with the challenges of heavy occlusion, scale variation, thermal crossover, and fast motion; and (d) the redbag sequence with partial occlusion, deformation, and scale variation. The top row for each sequence displays the frames in RGB, while the bottom row displays the frames in thermal. Different colored rectangles are used to show the results of various trackers.	29
Figure 12	PR score of the BC challenge on the RGBT234 dataset.	33
Figure 13	SR score of the BC challenge on the RGBT234 dataset.	33
Figure 14	PR score of the Camera Moving challenge on the RGBT234 dataset.	33
Figure 15	SR score of the Camera Moving challenge on the RGBT234 dataset.	33
Figure 16	PR score of the Deformation challenge on the RGBT234 dataset.	33
Figure 17	SR score of the Deformation challenge on the RGBT234 dataset.	33
Figure 18	PR score of the Fast Motion challenge on the RGBT234 dataset	34
Figure 19	SR score of the Fast Motion challenge on the RGBT234 dataset	34

Figure 20	PR score of the Heavy Occlusion challenge on the RGBT234 dataset	34
Figure 21	SR score of the Heavy Occlusion challenge on the RGBT234 dataset	34
Figure 22	PR score of the Low Illumination challenge on the RGBT234 dataset	34
Figure 23	SR score of the Low Illumination challenge on the RGBT234 dataset	34
Figure 24	PR score of the Low Resolution challenge on the RGBT234 dataset	35
Figure 25	SR score of the Low Resolution challenge on the RGBT234 dataset	35
Figure 26	PR score of the Motion Blur challenge on the RGBT234 dataset	35
Figure 27	SR score of the Motion Blur challenge on the RGBT234 dataset	35
Figure 28	PR score of the No Occlusion challenge on the RGBT234 dataset	35
Figure 29	SR score of the No Occlusion challenge on the RGBT234 dataset	35
Figure 30	PR score of the Partial Occlusion challenge on the RGBT234 dataset	36
Figure 31	SR score of the Partial Occlusion challenge on the RGBT234 dataset	36
Figure 32	PR score of the Scale Variation challenge on the RGBT234 dataset	36
Figure 33	SR score of the Scale Variation challenge on the RGBT234 dataset	36
Figure 34	PR score of the Thermal Crossover challenge on the RGBT234 dataset	36
Figure 35	SR score of the Thermal Crossover challenge on the RGBT234 dataset	36
Figure 36	Precision Rate evaluation curve on the LasHeR dataset.	37
Figure 37	Success Rate evaluation curve on the LasHeR dataset.	38
Figure 38	RGBT Tracking Model With Attentional Feature Fusion Module	40
Figure 39	The Structure of "ABr+AFF" Variation	41
Figure 40	The Structure of "ESKBr+AFF" Variation	41

LIST OF ABBREVIATIONS

BC	Background Clutter
CM	Camera Moving
CME	Camera Motion Estimation
DEF	Deformation
ESK	Enhanced Selective Kernel
FC	Fully-connected
FM	Fast Motion
HOG	Histogram Of Oriented Gradient
HO	Heavy Occlusion
IV	Illumination Variation
LI	Low Illumination
LR	Low Resolution
LBP	Local Binary Patterns
MB	Motion Blur
NO	No Occlusion
OCC	Occlusion
OPE	One Pass Evaluation
PO	Partial Occlusion
PR	Precision Rate
RGB	Red, Green, and Blue
RGBT	Red, Green, Blue and Thermal
SIFT	Scale Invariant Feature Transform

SOT	Single Object Tracking
SR	Success Rate
SV	Scale Variation
TC	Thermal Crossover
TMP	Target Motion Prediction

CHAPTER 1

INTRODUCTION

In computer vision, object tracking is a significant issue with wide-ranging applications, including surveillance, autonomous vehicles, and human-computer interaction. The ability to accurately and robustly track objects in complex and dynamic scenes is crucial for tasks such as object recognition, activity analysis, and behavior understanding. Traditional tracking methods often rely on single-modal data, such as RGB imagery, which may suffer from limitations such as poor visibility under challenging lighting conditions. In recent years, there has been a growing interest in fusing thermal images and visual RGB images to leverage the complementary strengths of both modalities, enabling more reliable and accurate object tracking. This thesis focuses on exploring and developing advanced fusion techniques using deep learning approaches to enhance object tracking performance by leveraging the benefits of RGB and TIR fusion.

Due to the intrinsic disparities between the two modalities, typical tracking approaches that seek to combine thermal and visual pictures into a unified stream might not be the best option for RGBT tracking in difficult conditions. RGB images collect visible light and provide precise color information, whereas thermal images catch heat radiated by objects and provide vital temperature information. To properly monitor objects under a variety of environmental situations, the fusion of various modalities necessitates a careful evaluation of their individual properties.

Utilizing the complimentary properties of thermal and visual pictures to address the difficulties of RGBT tracking has been made possible by recent developments in deep learning. Deep learning-based approaches have the potential to learn and extract significant characteristics from both modalities, allowing for more robust and accurate object tracking. We can improve object tracking performance in difficult settings by leveraging the capabilities of deep learning and the rich information offered by RGB and thermal images.

1.1 RGBT Tracking Challenges

To achieve accurate and dependable tracking of RGBT objects, a number of challenges must be solved. One of the main challenges lies in the intrinsic contrast between RGB and thermal modalities. RGB images record visual appearance and color information, whereas thermal images record temperature. These modalities have diverse properties such as varying spatial resolutions, dynamic ranges, and environmental sensitivity. Integrating these modalities efficiently necessitates resolving modality differences, such as aligning spatial and temporal information, calibrating intensity values, and dealing with variations in texture and appearance between RGB and thermal images. Robust fusion techniques

are needed to capitalize on the complementary strengths of RGB and thermal modalities, allowing for effective object tracking in a variety of settings.

Handling occlusion is another key problem in RGBT tracking. When objects of interest are partially or completely obscured by other objects in the scene, tracking problems occur. Because of the disparities in appearance and temperature information, occlusion is especially difficult in RGBT settings. When an object is occluded in one modality, the other modality may nevertheless give useful tracking information. Appropriate occlusion management techniques must be employed in order to precisely estimate the object's position and follow it consistently even in obstructed parts. This can include strategies such as using context information from both modalities, including motion cues, adopting appearance modeling techniques, and recovering from occlusion occurrences using temporal coherence.

Variations in illumination conditions also present a substantial barrier to RGBT tracking. Significant alterations in the appearance and temperature patterns of objects can result from changes in illumination. The result is a warped view of the world that makes it difficult to predict the future. Addressing illumination variations necessitates adaptive algorithms capable of handling changes in lighting conditions while maintaining accurate object tracking. Techniques like adaptive feature selection, robust fusion algorithms capable of handling variations in illumination, and learning-based systems capable of adapting to changes in the appearance and temperature patterns of the tracked object are examples of such techniques.

Scale and resolution variations might also provide challenges for RGBT tracking. Misalignment and difficulty identifying comparable object regions between the two modalities may occur. To efficiently handle size and resolution variations, strategies such as scale estimates, region alignment techniques, and adaptive fusion processes must be employed.

Addressing these challenges in RGBT tracking is critical for improving the accuracy, resilience, and efficiency of object tracking systems in a variety of contexts such as autonomous driving, surveillance, and robotics.

There are datasets available in the literature to develop appropriate models for real-world challenges. These datasets contain sequences with different challenge attributes. Information about these attributes is given in Chapter 3. This study has been developed with these challenges in mind. While presenting the results of the study, the achievements against the challenges are also discussed separately.

1.2 Problem Definition

In computer vision, object tracking is a significant issue that has significant applications in a wide range of disciplines, including autonomous driving, security surveillance, and robotics. Traditional object tracking methods frequently rely exclusively on visible band (RGB) images, which can be limiting in difficult circumstances. In addition, RGB-based tracking methods may struggle to maintain precision in a number of environmental conditions. To overcome these limitations, the integration of thermal infrared data, which provides additional information about an object's temperature, has shown promise in improving object tracking performance in challenging scenarios.

RGBT tracking is difficult since RGB and thermal images are including different modalities. Traditional tracking methods typically fuse RGB and thermal images in a single stream, which can be suboptimal in challenging scenarios.

The aim of this thesis is to find an effective method to fuse RGB and thermal modalities, enhancing RGBT tracking by incorporating deep architectures that combine high-level features with precision and efficiency, ultimately providing a viable solution for achieving precise and robust object tracking in challenging scenarios while minimizing parameter requirements and reducing reliance on extensive training data.

This thesis investigates how to effectively perform RGBT object tracking by combining information from thermal and RGB pictures. The objective is to develop a deep learning-based RGBT method for object tracking that exploits the advantages of both modalities while overcoming their distinctions. This requires effectively managing the differences between RGB and thermal images, such as varying illumination conditions, scale and resolution differences. The goal is to accomplish accurate and consistent object tracking despite occlusions and cluttered backgrounds, enabling reliable tracking performance in real-world scenarios.

Finding an efficient method to combine the RGB and thermal modalities is an additional crucial aspect of the issue. The challenge is to effectively integrate high-level features from both modalities in order to increase tracking precision and efficiency. By utilizing fewer parameters, the suggested approach aims to accomplish strong feature fusion and lessens the dependence on extensive training datasets. This thesis aims to resolve these problems and develop RGBT object tracking methods by offering insightful analysis and possible solutions for tracking objects in difficult situations.

We assess the efficacy of our method by comparing its results with those of other cutting-edge techniques using benchmark datasets like RGBT234 [1] and LasHeR [2]. The ultimate goal is to demonstrate that our work offers a promising solution for improving the accuracy and efficiency of RGBT tracking systems in various applications, including autonomous driving, surveillance security, and robotics.

1.3 Research Questions

In this section, we present a series of research questions that serve as the foundation of our investigation, guiding our exploration into the depths of the subject matter. The following questions will be meticulously examined to achieve a comprehensive understanding of the topic at hand:

- How can the integration of RGB and thermal modalities be used for object tracking in challenging scenarios?
- What are the limits of current tracking methods for RGBT and how can they be overcome?
- How can the distinctions between the RGB and thermal modalities be efficiently handled to produce precise and reliable object tracking?
- What methods can be used to deal with occlusion and keep precise tracking even when objects are partially or completely blocked?

- How can illumination variations be handled in an efficient manner to guarantee reliable object tracking under varying lighting conditions?
- What methods can be used to deal with scale and resolution differences between thermal and RGB images in RGBT tracking?
- How does the suggested RGBT tracking technique compared to current methods?
- How can we create a feature fusion technique that effectively fuses features from the thermal and RGB modalities to improve overall tracking performance under challenging circumstances?

By addressing these research questions, our goals are to further RGBT tracking methods and provide insights into the effectiveness and limitations of existing techniques.

1.4 Objectives of the Thesis

Within the framework of this thesis, we outline the core objectives that lay the groundwork for our scholarly pursuit. The following articulates the primary objectives that will steer our investigation and contribute to the broader understanding of the subject matter:

- Create an RGBT object tracking technique based on deep learning that effectively combines thermal and visual data to enhance object tracking performance in difficult scenarios including motion blur, occlusion, and lighting variations.
- Investigate and analyze the limitations of current RGBT tracking methods, highlighting the major issues and drawbacks, and suggesting fresh approaches to address these issues.
- To ensure that the tracking system keeps accurate localization and tracking of objects even when they are partially or entirely occluded from view, develop and test enhanced methods to deal with occlusion.
- Assess the accuracy, robustness, and computational efficiency of the proposed RGBT tracking technique by conducting thorough evaluations and comparisons with state-of-the-art tracking methods.

By achieving these objectives, the goal of this thesis is to improve RGBT object tracking, offer insightful information about how to overcome challenges and offer practical solutions to enhance object tracking precision and reliability across a wide range of applications.

1.5 Contributions of the Study

With due consideration to the comprehensive literature survey presented in the previous subsections, this section expounds on the substantial contributions of the thesis, accentuating their pertinence in filling the existing research gaps and extending the boundaries of knowledge in the field. Each technical contribution is meticulously examined below, revealing their novel insights in the context of the literature reviewed earlier:

- **Enhanced RGBT Object Tracking Approach:** This study proposes an enhanced deep learning-based RGBT object tracking approach that efficiently fuses information from both RGB and thermal modalities. By integrating the strengths of these two imaging sources, the proposed method enhances tracking performance, especially in challenging scenarios characterized by occlusion, changes in lighting, and motion blur.
- **Handling Modalities Differences:** A significant contribution lies in the investigation and implementation of techniques to address the distinctions between RGB and thermal modalities. The study ensures that the tracking system can reliably and accurately track objects across diverse imaging sources.
- **Robust Occlusion Handling:** This study introduces enhanced methods to effectively handle occlusion, ensuring that the tracking system can maintain accurate localization and tracking of objects even when they are partially or entirely obscured from view.
- **Scale and Resolution Adaptation:** A crucial contribution is the development and integration of techniques to handle scale and resolution discrepancies in thermal and RGB images. This adaptive approach ensures the tracking system's ability to handle targets of different sizes and resolutions.
- **Thorough Evaluation and Comparison:** The study evaluates the precision and success rate of the proposed RGBT tracking technique through thorough evaluations and comparisons with state-of-the-art tracking methods. This provides valuable insights into the performance and effectiveness of the proposed approach.

Overall, the contributions of this study advance the field of RGBT object tracking, offering enhanced solutions to overcome challenges and improve tracking precision and success rate.

On the other hand, the results obtained with the RGBT object tracking model developed with this thesis is also used in the preparation of the paper titled "EANet: Enhanced Attribute-based RGBT Tracker Network" [3]. This paper is accepted at The 16th International Conference on Machine Vision (ICMV-2023).

1.6 Organization of the Thesis

This subsection elucidates the organizational structure of the thesis, providing an overview of how the various components are interconnected to present a unified and comprehensive narrative. In the first chapter, which is an introduction to the thesis's subject matter, the problem is defined. The literature related to the subject of the thesis is discussed in Chapter 2, along with the deep learning-based RGBT fusion object tracking methods. The experimental design for the thesis is presented in Chapter 3. Datasets, metrics for evaluation, and experimental setups are described in further depth. The technique that is being proposed in this thesis is broken down in great length in Chapter 4, which is devoted to elaborating on the methodology. Chapter 4 begins with an explanation of Network Architecture then moves on to discuss Training. The tracking approach that is proposed in this thesis is then broken down and explained. The results of the tests are presented in Chapter 5, together with an analysis of the numerical results obtained from those experiments. This chapter also addresses the failure cases, explains and investigates the consequences of the findings, and talks about the probable causes and

remedies. In the last chapter, discussion is provided regarding the results that were reported in Chapter 5.

CHAPTER 2

RELATED WORK

The incorporation of deep learning techniques into computer vision has resulted in significant advancements in RGB thermal fused tracking, particularly through the utilization of RGB imagery. RGB thermal fused tracking combines thermal and RGB data to improve object detection and tracking in challenging environments. By utilizing the complementary strengths of both modalities, this method enhances tracking performance in scenarios with low visibility or poor lighting.

This chapter provides a comprehensive literature review on RGB thermal fused tracking, ranging from pre-deep learning methods to the current era of deep learning. Beginning with traditional tracking methods that rely on handcrafted features and simple algorithms, we examine the evolution of techniques. Then, we examine the emergence of deep learning and its influence on the field, highlighting early successes in object tracking.

The discussion then shifts to recent advancements in RGB thermal fused tracking based on deep learning. We investigate state-of-the-art architectures, network designs, and training methodologies, as well as commonly employed benchmark datasets and evaluation metrics. This chapter seeks to establish a foundation for our proposed approach by synthesizing existing knowledge, identifying research gaps, and setting the stage for subsequent chapters.

2.1 RGBT Fusion Tracking In pre-Deep Learning Era

Handcrafted features, including Histogram Of Oriented Gradient (HOG), Scale Invariant Feature Transform (SIFT), and Local Binary Patterns (LBP) have been used in previous RGB-infrared tracking fusion methods [4, 5]. Mean shift and filters like Kalman or particle were utilized by several of the methods as well [4, 5, 6, 7]. In contrast to Kalman filtering, which is only applicable to linear dynamics and Gaussian noise, particle filters can handle nonlinear and non-Gaussian visual tracking [7]. When it comes to recovering from a tracking failure, the mean shift approach can get stuck in a local minimum [8] and is therefore not ideal for use in RGB-infrared fusion tracking.

To characterize the human visual system, researchers have used sparse representation-based methods in RGB-infrared fusion tracking [9]. Several approaches have been suggested for addressing this matter, including [10] concatenating image patches and sparsely representing them; in order to create similarity and probability functions, [11] uses a joint sparse representation as primary method of data representation; [12] inducing generative multimodal feature models, and [13] designing modality-correlation-aware sparse representation models [14].

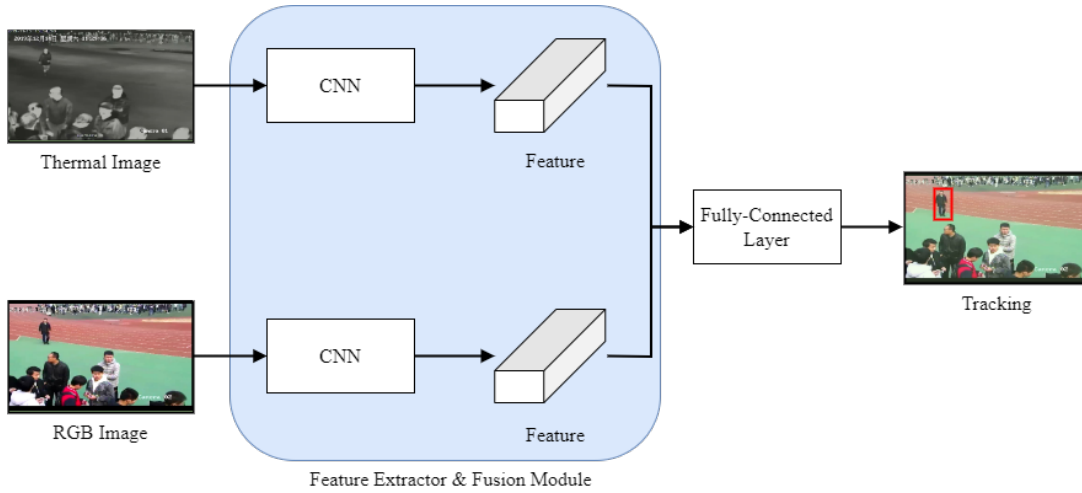


Figure 1: Pipeline of CNN-based RGBT Trackers

In recent times, there has been a utilization of graph-based techniques in addressing the challenge of RGB-infrared fusion tracking. With picture patches serving as nodes and weights indicating foreground or background information, these methods describe the intended bounding box as a graph [14]. During tracking, a dynamic learning process determines the weights and how they relate to individual patches. Background noise is reduced and feature representation is enhanced with the help of graph-based fusion trackers. By improving graph affinity, propagating patch weights, including cross-modal ranking, and multi-graph descriptors, several research have demonstrated the efficacy of graph-based methods [15, 16, 17]. Modality weights are also taken into account by some of these methods [1] for adaptive fusion of several modalities.

2.2 Deep Learning-based RGBT Fusion Object Tracking Methods

In this section, information about deep learning-based RGBT object tracking in the literature will be given. We analyzed the models in three subsections. These are Convolutional Neural Network (CNN) based models, Siamese-based models, and other models. The results of the models in the literature for different benchmarks can be seen in Table 1.

2.2.1 Convolutional Neural Network (CNN) Based RGBT Fusion Object Trackers

The common pipeline of CNN-based RGBT Fusion Object Trackers can be seen in Figure 1. A well-known visual tracking method built on the representations from discriminatively trained CNN is called Multi-Domain Network (**MDNet**) [18]. The proposed algorithm has shown superior performance compared to existing tracking benchmarks. A distinct training sequence is associated with each domain, and MDNet is formed from a combination of common layers and many branches of domain-specific layers. Each video is treated as a separate domain, and at the network's end, domain-specific layers are used. The layers before these for generic representation learning share data from all the sequences. To distinguish between domain-specific and domain-independent information, the algo-

rithm also incorporates a multi-domain learning framework. The method has been successfully used for visual tracking, where the CNN is updated online within the context of a new sequence after being pre-trained using multi-domain learning. In experiments on the VOT2014 [19], the algorithm outperformed state-of-the-art trackers. The vast majority of CNN-based RGBT object tracking algorithms have used MDNet as a baseline.

FANet [20] explores the use of visual and thermal infrared data to conduct stable visual tracking in difficult and unfavorable situations. As an alternative to traditional methods of RGBT fusion object tracking, it offers a quality-aware Feature Aggregation Net (FANet) based system. To overcome the difficulty of drastic visual shifts aggregates hierarchical deep features within each modality using FANet. It offers a productive method for using every layer feature in a model that has been trained and for picking up reliable target representations. Learns a nonlinear interaction among channels with different modalities using a Fully Connected layer (FC). Subsequently, employing a secondary fully connected layer (FC) alongside a SoftMax activation function, the model makes predictions regarding the modality weights responsible for regulating the transmission of information between modalities throughout the process of adaptive aggregation. In order to enhance the efficiency of feature aggregation, FANet employs the methodologies of max pooling and 1x1 convolution to compress feature dimensions. Adaptive aggregation additionally uses the activation functions SoftMax, fully connected layer, and global average pooling to predict the weights of the modalities.

To produce more precise outcomes, **DAPNet** [21] fully leverages shallow-to-deep spatial and semantic information and utilizes max pooling to normalize feature maps of varying widths and compress feature channels to cut down on unnecessary data. DAPNet uses shared parameters for the RGB and thermal backbone networks in order to minimize network parameters and capture common characteristics across several modalities. Weighted random sampling and global average pooling techniques are combined to achieve this. The DAPNet framework introduces a novel method for aggregating and pruning features, which involves recursively combining deep features from all layers and compressing feature channels. Similar to MDNet [18], as the backbone network, VGG-M [22] is chosen.

Deep Adaptive Fusion Network for High Performance RGBT Tracking (**DAFNet**) [23] is a real-time RGB and thermal tracking framework. It utilizes a recursive fusion chain to adaptively integrate features from different modalities. The fusion of multi-layer features in DAFNet allows for the aggregation of information at different scales and the integration of the complementary advantages of different levels of features. The adaptive weighting operations further improve performance.

MANet [24] is another deep neural network for a robust RGBT tracking. MANet consists of three different network blocks that play crucial roles in the MANet network's architecture. Remarkably, the authors present convincing data to highlight the critical role played by the instance adapter and the modality adapter components in the improvement of overall performance. Their research highlights the importance of these particular adapters to the network's performance.

Yang et al. [25] proposed a deep RGBT tracking method that is guided by dual visual attention. This method makes use of two crucial attentional mechanisms to efficiently train deep classifiers. The approach uses local attention first, which is based on the shared visual attention present in both RGB and thermal data. This allows the network to recognize important features and patterns for precise classification. Second, the method makes use of target-driven global attention, which enhances local attention by providing the classifier with global proposals. These global proposals work in conjunction with localized proposals derived from prior tracking outcomes, fostering a comprehensive understanding of

the tracked object's context and ensuring superior tracking performance. The researchers conducted two different experiments on RGBT benchmarks and found that their approach achieved comparable or better tracking performance compared to other methods, particularly when utilizing the local attention mechanism. Global attention was also found to improve the results of visual trackers.

CMPP [26] framework facilitates the diffusion of instance patterns in both the spatial and temporal domains by integrating past contexts over extended periods of time into the current framework, hence improving information inheritance. The CMPP framework is also robust to modality loss, with reduced performance degradation when a modality is lost. In practice, the CMPP framework has a frame rate of 1.3 FPS on the RGBT234 [1] dataset, making it suitable for real-time tracking applications.

Duality-gated mutual condition network (**DMCNet**) [27] is a neural network that can effectively suppress the effects of data noise and improve the discriminative abilities of target features. DMCNet is a model that uses discriminative information from one modality to guide feature learning in another modality. The researchers integrated a duality-gated mechanism to address the issue of noise stemming from single and multiple modalities during information propagation.

MANet++ [28] concurrently engages in target representation learning that encompasses "modality-shared", "modality-specific", and "instance-aware" aspects, all tailored for the demands of RGBT tracking. Its purpose is to harness the distinct strengths of both visible and thermal data, resulting in a resilient approach to object tracking. MANet++ is based on MANet and includes the previously expanded functionality within each modality prior to fusion.

CBPNet [29], like other RGBT Trackers, includes a feature extractor. It is supported with attention to increase the accuracy of object tracking. The channel attention mechanism selectively enhances informative features and suppresses less relevant ones in the channel dimension, contributing to improved performance. The cross-layer bilinear pooling module facilitates hierarchical feature interaction and effective information integration, capturing complex patterns and relationships. A robust representation is produced by the quality-aware fusion module, which integrates deep semantic information and low textural information from several modalities. Leveraging these multimodal representations, the three fully connected layers conduct accurate tracking decisions, making the model well-suited for RGBT tracking tasks.

M⁵L [30] is based on deep metric learning, in which a loss function is used to optimize the distance between samples in feature space. M⁵L employs a unique loss function known as the "Multi-modal Multi-margin Structured Loss". The employed loss function successfully preserves the structured information derived from samples acquired from both thermal and RGB images. Additionally, it enhances the discrimination between perplexing positive signals and regular ones by incorporating various margins. The system additionally incorporates a module that integrates features adaptively to enable end-to-end training using a CNN.

After the cross-modal interactions, each of the trackers indicated above maintains a fused feature representation. However, thermal and visual features are also kept in some approaches to maintain the discriminative nature of the modality-specific properties.

To get beyond the limitations of single modality tracking, **MaCNet** [31] makes use of competitive learning strategies and a modal-aware attention network. The technique consists of three different modules for effective cross-modality attention and a classification network that is used to classify the extracted features as either the target object or the background.

Trident Fusion Network (**TFNet**) [32] uses both thermal and RGB data to track certain instances across a series of frames. For classification and regression, TFNet uses aggregated and modality-specific features to thoroughly mine the data from both modalities. The network consists of an aggregation module, feature pruning, and fusion module and is based on the VGG-M [22] network. The utilization of the feature pruning technique aims to mitigate the issue of overfitting and improve the acquisition of valuable feature representation.

Challenge-Aware RGBT Tracking (**CAT**) [33] comprises a CNN backbone, challenge branches that share modalities, a module to aggregate adaptively for all branches, and architecture in layers. The shared challenge branches in modality collaboration are employed to capture the target's appearance consistently across all modalities for cooperation, while the modality-specific branches are separate branches for each modality that have different levels of granularity and complexity. The module for aggregation adaptively combines challenge-aware target representations to form discriminative target representations. The adaptive fusion module of the hierarchical design adaptively fuses various modality features, while the discriminative feature transfer module is responsible for transferring features from one modality to another.

ADRNet [34] is another RGBT Tracker with high-speed computational efficiency with 25 fps. The ADRNet is specifically built to simulate the target appearance in various characteristics by utilizing the attribute-driven branch. This branch comprises a 3x3 convolutional layer, which is subsequently followed by a ReLU [35] activation function. The attribute-driven branches are then adaptively aggregated via a summation and selection operations module. The ADRNet also includes a spatial-wise ensemble network (SENet) to enhance feature representation and avoid drifting to similar surroundings.

Five difficulty attributes ("thermal crossover, illumination variation, scale variation, occlusion, and fast motion") are being identified by **APFNet** [36]. Each attribute is trained one by one with a small subset of training data with the corresponding attribute. This allows APFNet to use small-size data for each branch, focusing on attribute-specific feature fusion.

A three-stage approach that is intended to be both efficient and effective is used to train APFNet. APFNet includes an improvement transformer with two decoders and three encoders. APFNet can be trained well with small data and can tackle a wide range of challenges.

2.2.2 Siamese-based RGBT Trackers

The common pipeline of Siamese-based RGBT Fusion Object Trackers can be seen in Figure 2. Because of the end-to-end training approach's efficiency, the Siamese network is continued by the ground-breaking work, **SiamFC** [37] in visual object tracking. Overall, it seeks to learn a broad similarity evaluation measure. According to the figure, the RGB and TIR feature extractors used by the current Siamese-based trackers are not equivalent. The published Siamese-based RGBT trackers' feature aggregation is accomplished through the usage of the multi-modal fusion module. Then, before each of their separate heads, the same method for assessing similarity is applied to both regression and classification.

Zhang et al. [38] is a tracking method at the pixel-level using Siamese Networks, which has been demonstrated to be effective in RGB object tracking. The proposed Siamese network is used to perform

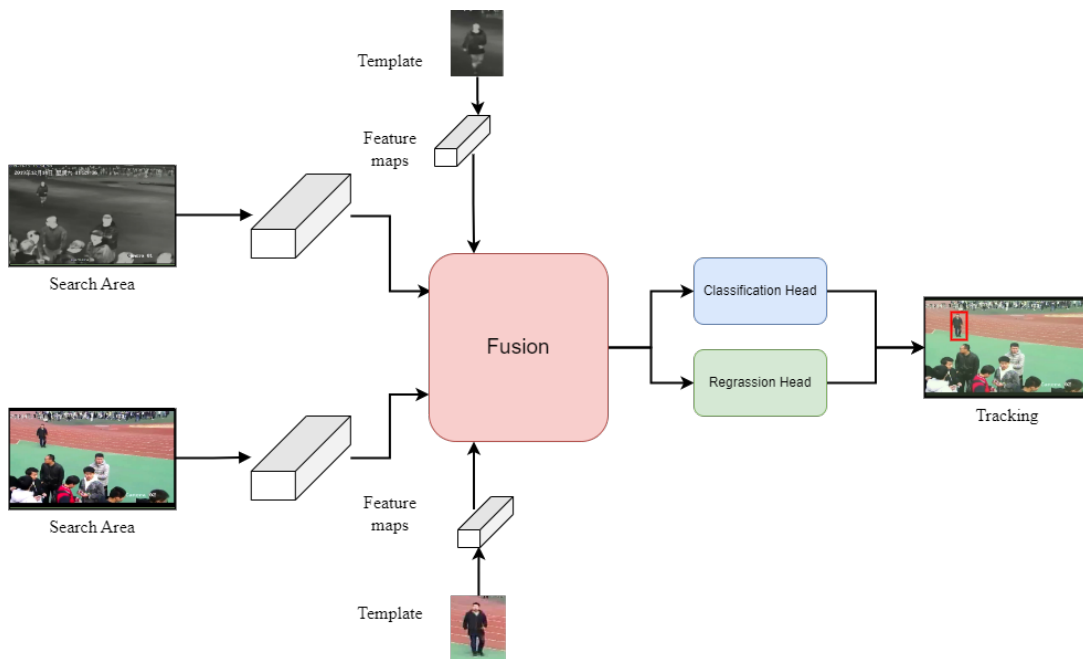


Figure 2: Pipeline of Siamese-based RGBT Trackers

tracking after the fusion of visible and infrared images. This is the first work to use Siamese networks for fusion tracking.

SiamFT [39] is also based on Siamese Networks. The method processes thermal and visual images separately using two Siameses. Predicting image dependability using modality weight computation is also proposed. The SiamFC [37] was used as the backbone of SiamFT. The framework of SiamFT is generic and can be used with other Siamese network-based tracking methods. The modality weight computation method adaptively predicts the reliability of thermal and RGB images in different exposure conditions.

The dynamic Siamese networks and multi-layer feature fusion that form the foundation of **DSiamMFT** [40] are responsible for the adaptive integration of multi-level deep features between the two networks. The final response map, which is utilized to find the target, is formed by combining the response maps that were produced from the various fused layer features using an elementwise fusion approach. When compared to systems that are just based on images from a single modality, the tracking performance of the proposed methodology is greatly improved.

DuSiamRT [41] is accomplished by utilizing a dual Siamese network to combine two modal projected position maps. It is predicated on the notion that under challenging settings, such as poor light or rainy weather, the target may be more clearly visible in thermal infrared photographs than in RGB images. The information that is provided by both modalities is utilized by DuSiamRT through the usage of handcrafted features and an attention mechanism. Additionally, the bounding box prediction is improved through the utilization of a region proposal network.

SiamCDA [42] another Siamese-based RGBT Tracker, aims to enhance tracking performance while maintaining computational efficiency as its primary objective. To accomplish this objective, it in-

tegrates a multi-modal feature fusion module that is attuned to capturing complementarity between modalities, alongside a region proposal selection module that is sensitive to the presence of distractors [42]. The authors conducted extensive experiments on large-scale datasets consisting of visible and infrared images.

2.2.3 Other Deep RGBT Fusion Object Trackers

Li et al. [43] make use of a two-stream architecture for building an adaptive representation of features by fusing thermal and visual streams, which extract general information about objects. This is accomplished through the process of learning an adaptive feature representation. After that, a multi-channel correlation filter is applied in order to provide accurate predictions regarding the locations of the objects using the fused feature maps. A fusion sub-network is one of the components of the ConvNet design, which also includes generic sub-networks. ReLU [35] serves as the non-linearity in the fusion sub-networks, two independent convolutional layers that make up its architecture. The proposed method incorporates features into a filter in order to locate the object and is able to manage scale fluctuations and partial occlusions in an efficient manner.

mfDiMP [44] is a method using an end-to-end tracking mechanism. At several framework levels, such as the pixel and feature levels, the authors take into account a number of fusion processes. The proposed method is based on the Discriminative Model for Prediction (DiMP) [45] tracker.

(**JMMAC**) [46] is suggested as a joint model for motion and appearance cues. "Multimodal fusion" and "motion mining" are included in the framework. Target and camera motions are included in the motion cue, which is used to reinforce the tracker when the appearance cue is faulty. A tracker switcher is also suggested so that the motion and appearance trackers can be switched around in a flexible way. The framework uses the ECO [47] method as the base tracker and a novel late fusion method called MFNet [48]. The target motion prediction module in the motion mining component allows it to dynamically choose which cue to utilize for the target position. The CME scheme is used to estimate camera motion, and the TMP scheme is used to switch between the appearance and motion trackers.

2.3 Different Fusion Levels

Experimental findings of [50] demonstrate that combining thermal and RGB images can effectively address a number of issues that become intractable when only one source image modality is used. The outcomes of fusion tracking vary at different fusion levels due to the various methods of using usable information. Pixel-level fusion is a simple way to introduce duplicate data, which could reduce the tracking accuracy. Feature-level fusion is a technique that combines the distinctive characteristics of images in order to improve the visibility and dependability of the target. Results gathered from diverse modalities are tracked at the decision level, comprising two or more instances. Therefore, when the accuracy of a single method is low, the benefit of decision-level fusion to monitoring may be overlooked.

Tracker	GTOT [49]		RGBT210 [15]		RGBT234 [1]		LasHeR [2]	
	PR	SR	PR	SR	PR	SR	PR	SR
FANet [20]	0.891	0.728	-	-	0.787	0.553	0.442	0.309
DAPNet [21]	0.882	0.707	-	-	0.766	0.537	0.431	0.314
DAFNet [23]	0.891	0.712	-	-	0.796	0.544	0.449	0.311
MANet [24]	0.894	0.724	-	-	0.777	0.539	0.457	0.330
[25]	0.843	0.677	-	-	0.787	0.545	-	-
CMPP [26]	0.926	0.738	-	-	0.823	0.575	-	-
DMCNet [27]	0.909	0.733	0.797	0.555	0.839	0.593	0.491	0.357
MANet++ [28]	0.901	0.723	-	-	0.800	0.554	0.467	0.317
CBPNet [29]	0.885	0.716	-	-	0.794	0.541	-	-
M ⁵ L [30]	0.896	0.710	-	-	0.795	0.542	-	-
MaCNet [31]	0.880	0.714	-	-	0.790	0.554	0.483	0.352
TFNet [32]	0.886	0.729	0.777	0.529	0.806	0.560	-	-
CAT [33]	0.889	0.717	0.792	0.533	0.804	0.561	0.451	0.317
ADRNet [34]	0.904	0.739	-	-	0.809	0.571	-	-
[38]	-	-	-	-	0.610	0.428	-	-
SiamFT [39]	0.826	0.700	-	-	0.688	0.486	-	-
DSiamMFT [40]	-	-	0.642	0.432	-	-	-	-
DuSiamRT [41]	0.766	0.628	-	-	0.567	0.384	-	-
SiamCDA [42]	0.877	0.732	-	-	0.760	0.569	-	-
[43]	0.852	0.626	-	-	-	-	-	-
mfDiMP [44]	-	-	0.786	0.555	0.785	0.559	0.447	0.344
JMMAC [46]	0.902	0.732	-	-	0.790	0.573	-	-
APFNet [36]	0.905	0.739	-	-	0.827	0.579	0.500	0.362

Table 1: Results of Fusion Object Trackers on Different Benchmarks

2.3.1 Pixel-level Fusion

The fusion-before-tracking strategy, also known as pixel-level fusion tracking, involves fusing images from various modalities to create more informative images before conducting object tracking, as illustrated in Figure 3. However, pixel-level picture fusion retains the majority of the information from the original photos, making it very computationally expensive and may slow down the entire process. Some RGBT tracking models [51, 52, 53, 54] used pixel-level fusion.

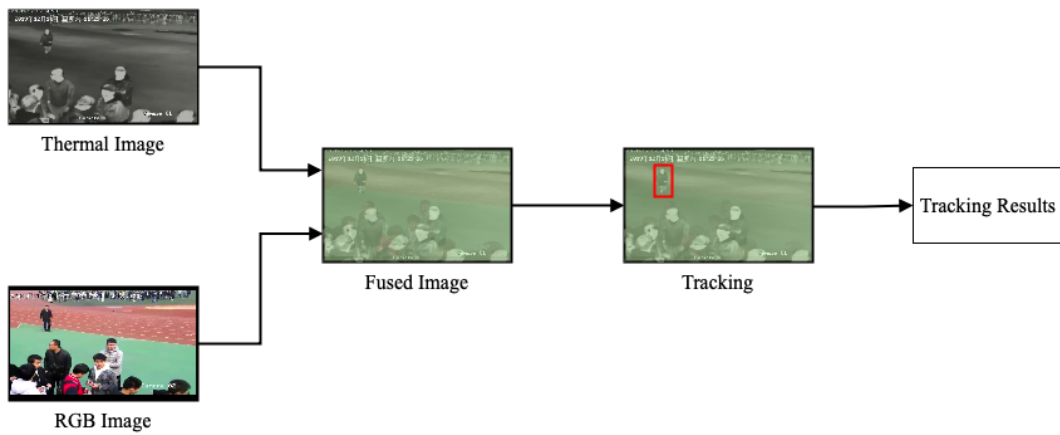


Figure 3: Pixel-level Fusion Tracking

2.3.2 Feature-level Fusion

The first step in feature-level fusion for tracking is feature extraction from the RGB and thermal pictures. Then, a pre-established fusion rule is applied to combine these features into a fused feature. This fused characteristic is then used for tracking. The performance can be enhanced since fused features typically provide additional information. The simple architecture of feature-level fusion is shown in Figure 4. The feature-level fusion method is simpler than the pixel-level method since it directly creates multi-modal features. The extraction of RGB and thermal pictures, as well as their efficient fusion, are the important elements of feature-level fusion. Feature-level fusion is used by the majority of deep RGBT object trackers.

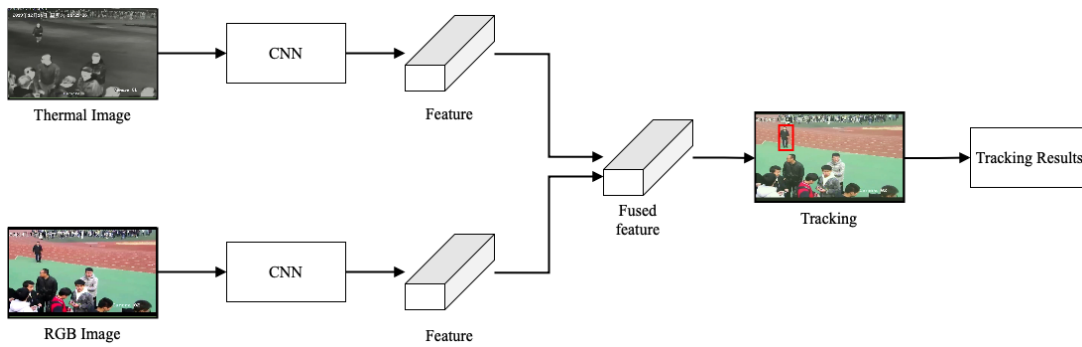


Figure 4: Feature-level Fusion Tracking

2.3.3 Decision-level Fusion

The decision-level fusion, sometimes referred to as the tracking-before-fusion technique, is the highest degree of fusion tracking. Tracking is carried out using different modalities, as indicated in Figure 5. Fusion techniques at the decision-level have certain benefits. First, based on RGB and thermal images, several trackers can be selected to do tracking. The bounding box surrounding the target is the sole thing that the majority of decision-level fusion tracking techniques require. Second, compared to other fusion tracking approaches, the computational cost is typically lower. The tracking speed might be quicker than approaches for the fusion at other levels. Additionally, there are fewer criteria for the alignment of thermal and RGB pictures in decision-level fusion tracking. Some tracking methods focused directly on decision-level fusion to design a robust RGBT fusion object tracking system, such as [55] and [56].

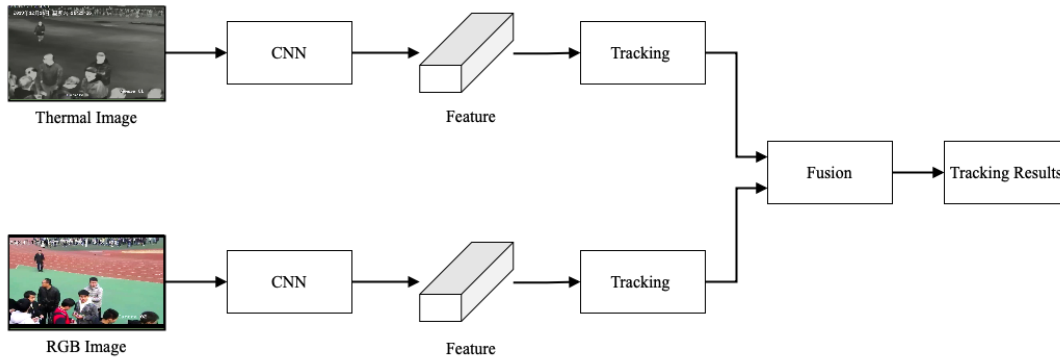


Figure 5: Decision-level Fusion Tracking

CHAPTER 3

EXPERIMENTAL SETUP

In this section, we describe the datasets used for experiments and the evaluation metrics.

3.1 Datasets and Benchmarks

This section provides detailed information about the benchmarks/datasets currently used in the literature. A summary of RGBT Fusion object tracking benchmarks/datasets and their contents can be seen in Table 3.

3.1.1 VOT Challenges

VOT-TIR (Visual Object Tracking with Thermal Infrared) [57] is a benchmark specifically designed for evaluating the performance of tracking algorithms for objects in videos with thermal infrared (TIR) channels. It includes a collection of video sequences with ground truth annotations and is designed to test the performance of tracking algorithms in situations where the objects being tracked have TIR information.

TIR imagery is often used in tracking applications because it can provide additional information about the temperature and surface properties of objects, which can be useful for distinguishing between different objects and improving tracking accuracy. The VOT-TIR benchmark is designed to test the performance of tracking algorithms in these types of situations.

The VOT-TIR benchmark is commonly used in the research and development of tracking algorithms and systems and can help researchers understand the strengths and limitations of different approaches and identify areas for improvement.

Since this dataset is generally used for VOT Challenges, it is not used in this study, but it is included in this section because it is widely used and has an important role in the literature.

3.1.2 GTOT

The GTOT [49] evaluation standard comprises 50 video sequences, which were taken using thermal and grayscale cameras. These sequences encompass a range of challenging scenarios like low light,

Attr	Description
NO	No Occlusion.
PO	Partial Occlusion - there is some obstruction in the path to the objective object.
HO	Heavy Occlusion - more than 80% of object visibility is blocked.
LI	Low Illumination - the amount of light that is present in the area being targeted is inadequate.
LR	Low Resolution - the target region has a low resolution.
TC	Thermal Crossover - indicates that the temperature of the target is comparable to that of other objects or the background surroundings.
DEF	Deformation - a deformation of an object that is not rigid.
FM	Fast Motion - the ground truth moves more than 20 pixels in either direction between the two frames that are immediately adjacent to one another.
SV	Scale Variation - it has been determined that the ratio of the original bounding box to the current bounding box does not fall within the range of [0.5,1].
MB	Motion Blur - the blurred image data results from the moving target item.
CM	Camera Moving - a moving camera catches the target item.
BC	Background Clutter - the object is included in the disorganized background information.

Table 2: RGBT234 Dataset - List Of Attributes

swift movement, partial obstruction, and cluttered backdrops. In addition to the video sequences, the GTOT [49] benchmark includes evaluation metrics where researchers can compare the performance of their tracking algorithms. The GTOT [49] benchmark has been used a lot to measure how well visual tracking algorithms work that uses information from both grayscale and thermal modalities. It has been shown to be a useful resource for researchers working on improving the robustness and adaptiveness of object tracking algorithms in challenging scenarios.

3.1.3 RGBT210

RGBT210 [15] has become a large RGBT dataset with 210 video pairs. RGBT210 [15] includes aligned video pairs for object tracking in both night and day. It has the advantages of being large and having a diverse set of challenging scenarios, such as deformity, partial occlusion, and scale estimation. This dataset was later expanded to be named RGBT234 [1].

3.1.4 RGBT234

The RGBT234 [1] dataset is an improvement upon previous RGBT datasets, such as RGBT210 [15], because it contains a larger number of video pairs and includes videos captured on hot days, reducing bias due to the temperature sensitivity of thermal sensors.

RGBT234 [1] has 234 RGBT videos, each of which has a color video (RGB) and a thermal video (T). It has about 234K frames in total, and the largest video pair has an 8K frame. This is a benchmark dataset that shows how to track objects well in difficult situations by using information from RGBT videos in a flexible way.

Benchmark	Year	Videos	Frames	Aligned	Category	Attributes	Reference
GTOT	2016	50	15.8K	N	9	7	[49]
RGBT210	2017	210	210K	Y	22	12	[15]
RGBT234	2018	234	233.8K	N	22	12	[1]
VOT-RGBT2020	2019	60		-	13	12	[57]
LasHeR	2021	245+979	730K	Y	32	19	[2]

Table 3: RGBT Fusion Object Tracking Benchmarks/Datasets and Their Contents

Imaging equipment for RGBT234 [1] includes a platform that can be turned, a thermal infrared camera, and a charge-coupled device (CCD) camera. The collimator lines up the optical axes of two cameras so that they point in the same direction.

The RGBT Fusion Object Tracking benchmark includes datasets that capture a variety of real-world scenarios, such as:

- Indoor and outdoor environments
- Different lighting conditions (e.g., day, night, low light)
- Occlusions and clutter
- Multiple objects moving simultaneously

The challenging parts of object tracking are represented by the 12 attributes used by RGBT234 [1] to annotate the sequences. List of attributes can be seen in Table 2. RGBT234 [1] benchmark is intended to provide a comprehensive and realistic evaluation of object tracking algorithms, and the authors hope that it will become a widely used standard for evaluating the performance of object tracking systems.

3.1.5 LasHeR

LasHeR [2] is made up of 1224 pairs, or over 730,000 frame pairs, of videos taken in both the visible and thermal infrared spectrums (245 for testing and 979 for training). Each frame has been carefully aligned and given a bounding box by hand. This dataset will have a big effect on how advanced RGBT fusion object trackers are trained and how thoroughly RGBT fusion object tracking methods are tested.

LasHeR [2] captures numerous item categories, camera angles, complicated scenes, and environmental variables that span day and night, as well as seasons and weather. In data creation, various new issues are taken into account as a result of real-world applications. It will encourage the development of practical tracking algorithms.

In addition, an unaligned version of LasHeR [2] is released to encourage research on alignment-free RGBT fusion object tracking, which is more practical for real-world applications.

3.2 Evaluation Metrics

The RGBT Tracking evaluation metrics show variation depending on the selected benchmark. Precision rate (PR) and success rate (SR) are two metrics that are the same for the following benchmarks: GTOT [49], RGBT210 [15], RGBT234 [1], and LasHeR [2]. The difference between the projected and ground-truth bounding boxes is measured by the Precision Rate (PR). On the other hand, the Success Rate (SR) measures the proportion of tracking failures when the Intersection over Union (IoU) value is less than a predefined threshold. Within the field of quantitative performance evaluation, the one-pass evaluation (OPE) yields PR and SR values. Using the area under the curve, a representative SR score is obtained, which represents the number of correctly tracked frames with overlaps exceeding predetermined thresholds.

The metrics that are used in VOT-RGBT2019 [58] and VOT-RGBT2020 [57] include Accuracy, Robustness, and Excepted Average Overlap (EAO). Accuracy can be seen as the degree to which the ground truth and the prediction are consistent with one another. The concept of robustness refers to the ability to measure the proportion of unsuccessful tracking attempts relative to the total number of image frames. EAO is widely regarded as the most crucial and provides an all-encompassing indication of the tracker’s superiority. Since VOT-RGBT2019 [58] and VOT-RGBT2020 [57] benchmarks are not used in this study, these metrics are not explained in detail.

3.2.1 Precision Rate

Precision rate (PR) is a metric used to evaluate the performance of a deep object tracking algorithm. It can be defined as the proportion of accurate detections to all detections the algorithm made. What constitutes an acceptable PR is the percentage of frames in which the output location falls between a predefined threshold and the ground truth. It calculates the average Euclidean distance between the center coordinates of the tracked target and the manually annotated ground-truth positions of each frame. The threshold is 20 pixels in order to determine the representative PR score.

In other words, it measures the proportion of detections that are correct. A high precision rate indicates that the algorithm is able to accurately identify objects in the scene and makes correct detections, while a low precision rate indicates that the algorithm is prone to making incorrect detections.

3.2.2 Success Rate

The success rate (SR) is a measure of the performance of a tracking algorithm. It is commonly expressed as the proportion of frames in a particular video sequence in which the algorithm properly locates the tracked item. The formula can be expressed as:

$$\text{Successrate} = \frac{\text{Number of Successful Outcomes}}{\text{Total Number of Trials}} \quad (1)$$

To compute the success rate, one needs to define a criterion for determining whether the object has been correctly located in a given frame. This criterion could be based on the overlap between the bounding box around the tracked object and the ground truth bounding box.

One such criterion for evaluation could involve determining if the degree of overlap between the two bounding boxes exceeds a specified threshold, such as 50%. In this scenario, the measure of success would be determined by computing the ratio of frames in which the degree of overlap surpasses the predetermined threshold to the overall number of frames encompassing the video sequence.

The success rate is an important metric for evaluating the performance of object tracking algorithms, as it allows researchers and practitioners to compare the performance of different algorithms on the same dataset.

3.3 Experimental Settings

The code for this thesis was written in Python, which has become the standard for deep learning in recent years because of its higher-level syntax and interpretation, but especially because of the GPU acceleration tools that it supports. PyTorch was used in this thesis because it is one of the most famous deep learning frameworks that GPUs can speed up. Also, the tests were done on a computer system including GPU (NVIDIA A4000), with 16 gigabytes of RAM.

Conducting a comprehensive assessment, our approach is thoroughly evaluated using the RGBT234 [1] benchmark, renowned as a premier benchmark for RGB and thermal object tracking. The primary objective behind evaluating our methodology on the RGBT234 [1] benchmark is to showcase its exceptional precision and efficiency in tracking targets under challenging conditions.

In this study, we use GTOT [49] to train our proposed model, and RGBT234 [1] and LasHeR [2] to test our proposed model, the details of which are given in the 3.1. The details of the training are explained in detail in 4.2. Success rate and precision rate, details of which are given in 3.2, are used as evaluation metrics.

CHAPTER 4

METHODOLOGY

4.1 Network Architecture

Our research focuses on merging an aggregation module proposed by [59] with the attribute-specific fusion influenced by [36]. For fusion, APFNet uses the so-called "Attribute-based Progressive Fusion Module". "Attribute-Specific Fusion Branch", "Attribute-based Aggregation", and "Attribute-based Enhanced Fusion" [36] are the three primary parts of this module. This architecture shows to be an effective RGBT tracker network, but in this study, we investigate methods for replacing the "Attribute-based Enhanced Fusion Module" [36] to simplify the system.

We adopt our idea from a different architecture, the ESKNet [59], which adds spatial attention techniques to the SKNet [60] model. The network achieves precise calibration of the importance of different regions by employing spatial attention, which dynamically assigns more weights or emphasis to specific spatial parts. Less informative regions are ignored or suppressed by the spatial attention mechanism, which gives higher weights to significant regions and prioritizes them for tracking. The model has the capability to enhance its depiction of the item's visual attributes and effectively accommodate variations in its spatial orientation or visual characteristics by prioritizing the most significant spatial regions. By calibrating the spatial dimension features through the integration of spatial attention, we hope to improve the model's ability to pay attention to and follow the object accurately, even in difficult environments. To enhance the model's capacity to track the object precisely, particularly in challenging circumstances, we calibrate the spatial dimension features by integrating spatial attention.

To extract features from thermal and visual pictures, we use VGG-M [22] as the backbone. As the network architecture, we used the MDNet [18] structure like most of deep learning-based object tracking models.

We use a parallel network as our network's backbone [43] to account for the variations between the different imaging modalities and extract features separately from RGB and thermal infrared images. The initial three convolutional layers of VGG-M, featuring kernel dimensions of 7x7, 5x5, and 3x3, serve as the fundamental framework of our model. The convolution kernel parameters are initialized using a pre-trained model on ImageNet-vid [18]. Through the integration of the improved attribute-based module at every tier of the backbone, we establish a hierarchical architecture that enables efficient fusion of various modalities. Three fully connected layers are added after the final convolutional layer. The FC layer at the end is modified for various domains, similar to the methodology used in [18].

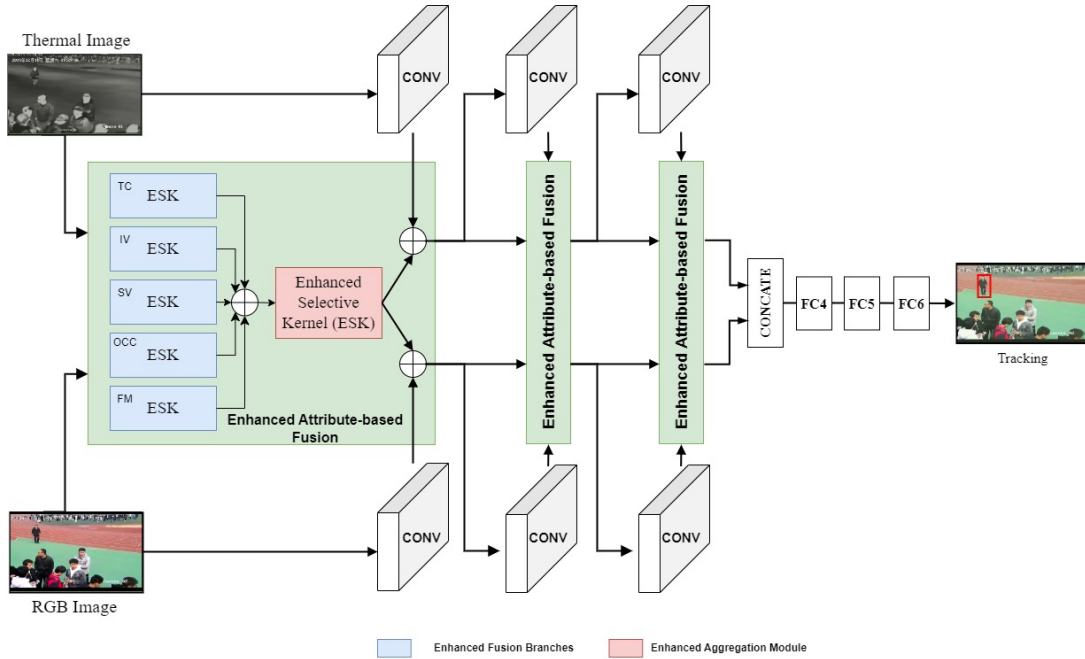


Figure 6: The architecture of the Enhanced Attribute-based Network.

As shown in Figure 6, while RGB and Thermal data are subjected to convolution as in MDNet [18], a fusion process is applied to these data at the same time. The data first passes through enhanced fusion branches that are specially trained for different attributes. Similar to [59], we employ Enhanced Selective Kernel (ESK) for both branch data aggregation and fusion. The structure of the proposed "Enhanced Fusion Branches" and "Enhanced Aggregation Module" can be seen in Figure 7 and 8, respectively. Following the aggregation process, element-wise addition is applied to both the convolutional and aggregation-derived data. At the end of the third layer, this process—which runs concurrently for RGB and Thermal data—is concatenated.

The fusion branches that map to distinct features of the same structure are mapped in an attempt to simplify the situation as much as possible. More precisely, we start by using a rectified linear unit (ReLU) [35], a convolutional layer to extract features from two modalities for each branch. Every branch undergoes a repetition of this process. Next, we utilize the ESK [59] to choose the features in an adaptable manner. Figure 6 illustrates the hierarchical structure of our suggested network.

4.2 Training

During the training and the test phase, we adopt a similar strategy as defined in [36]. As part of the trials, we extract attribute-based training data from the GTOT [49] dataset using challenge labels in order to prepare our attribute-specific fusion branches for testing on the RGBT234 [1] and LasHeR [2] datasets. After that, we train an attribute-based aggregation network and an enhancement fusion transformer module using the full GTOT [49] dataset.

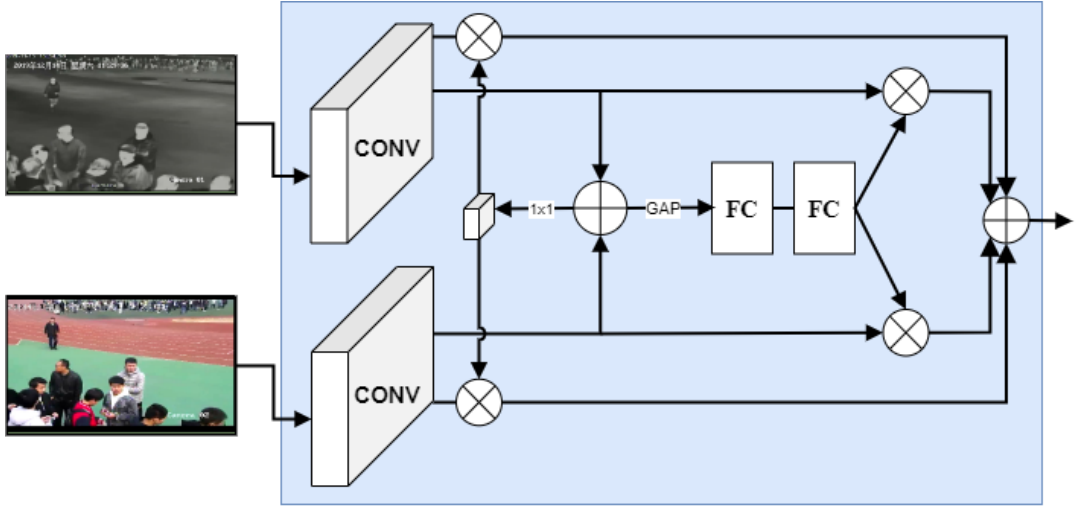


Figure 7: The structure of the proposed “Enhanced Fusion Branch Module”.

A two-phase strategy is used for the training. Every branch is trained separately during the initial stage. The VGG-M [22] parameters that are pre-trained on the ImageNet dataset are used to initialize the dual-stream CNN, which consists of three convolutional layers and FC4 and FC5 layers (see Figure 6). Next, we add the new FC6 classification branches and set their initial values for all fusion branches. The learning rates for FC6 and fusion branches are 0.001 and 0.0005, respectively. There will be 200 total training epochs. To remove the FC layer’s impact, we have only been saving the parameters of the attribute-specific fusion branches up to this point.

In step two, we correct the trained branches and use all available training data to improve the aggregation fusion modules. There are 500 training epochs in this process. Everything else is configured exactly as it was in the first stage. The aggregation fusion modules FC4 and FC5 parameters are stored. The stochastic gradient descent (SGD) is employed for the purpose of optimizing the network with a momentum of 0.9 and a weight attenuation of 0.0005. The learning rates for the convolutional and fully-connected layers are set at 0.0001 and 0.001, respectively.

We randomly select eight image frames and the matching tracking object location from a video clip for each training iteration. Gaussian sampling is then used to extract 256 positive samples and 768 negative samples from the previously mentioned eight image frames, where 32 positive samples and 96 negative samples are needed for each frame. A candidate box obtained using Gaussian sampling is deemed as a positive sample if the overlap rate with the ground truth value is within the range of [0.7, 1]. When the range of values falls within the interval [0, 0.5], it is commonly interpreted as indicating a negative outcome for the sample. Notably, we train our tracker on the GTOT [49] dataset to evaluate it on the RGBT234 [1] and LasHeR [2] datasets.

4.3 Online Tracking

The tracking phase begins with the tracker establishing itself using the position of the target and the first frame of the series, as most trackers do. Using Gaussian sampling in the first frame, 500 positive

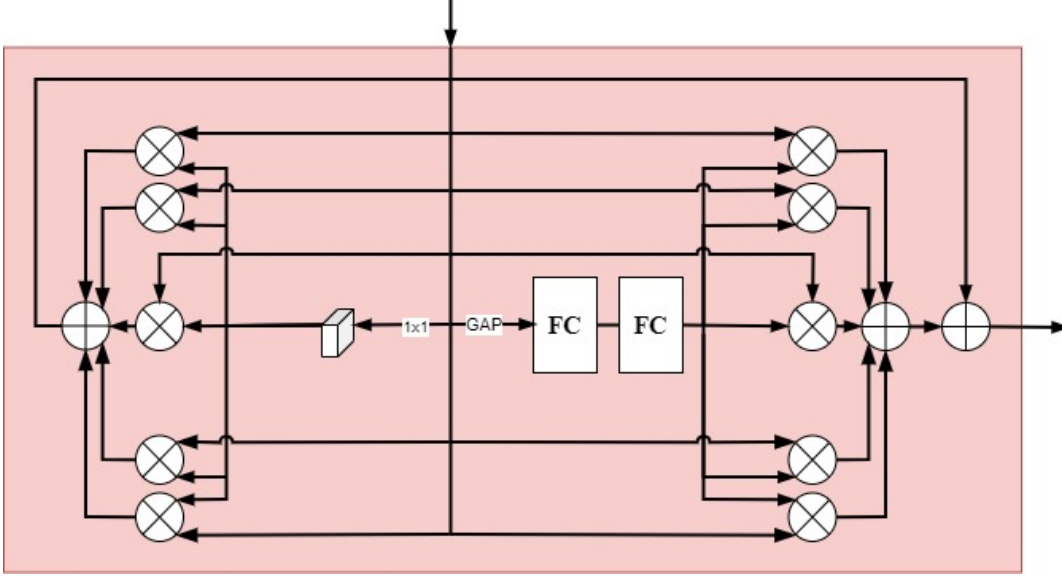


Figure 8: The structure of the proposed “Enhanced Aggregation Module”

samples of different scales around the object are collected. In order to train the regressor, 1000 samples are randomly collected. The coordinates of the tracking findings are changed by the utilization of a regressor, hence enhancing the accuracy of the subsequent tracking process. Importantly, we only modify the fully-connected layers’ parameters during this phase. This approach we use is the same approach that [18] use. The target is monitored in the t -th frame, and a collection of 256 samples is obtained for the current frame using Gaussian sampling. These samples are based on the tracking outcome from the $t - 1$ -th frame. Initially, the mean value is calculated for the top five samples with the highest scores at the given point in time. This process is accomplished by employing the trained model to calculate the scores for a total of 256 samples. The learned regressor is then used to modify the target location. In line with [18], we ensure the robustness of our approach by tracking faults using both short- and long-term update settings. $f^-(x_t^i)$ represents the negative scores, and $f^+(x_t^i)$ represents the positive scores for each sample. At time t , we choose the candidate region sample that has the greatest score to represent the tracking result X_t^* , and the equation for the formula is as follows:

$$X_t^* = \arg \max f^+(x_t^i) \quad i = 1, 2, \dots, N \quad (2)$$

CHAPTER 5

RESULTS AND DISCUSSION

This section presents the experimental results and analyze the performance of the different RGB-TIR fusion techniques and deep learning models and also compare the results with other state-of-the-art methods in object tracking.

5.1 Evaluation on RGBT234 Dataset

On the RGBT234 dataset, our method performed admirably in terms of overall performance. The precision rate (attained 83.5%, demonstrating the level of target tracking accuracy. Furthermore, the success rate (SR) is 58.4%, suggesting the efficiency of our method in effectively tracking targets. Our model is compared with with state-of-art models such as [24, 26, 28, 30, 33, 34, 36, 46]. A visual comparison of EANet to three of these state-of-art trackers [24, 34, 36] on various sequences can be seen in Figure 11. Figure 9 and 10 show the Precision Rate and Success Rate evaluation curves, respectively.

We further analyzed our method’s effectiveness based on several dataset attributes to give a full analysis of its performance. We are able to evaluate how well our method dealt with common object tracking issues like occlusion, light changes, motion blur, and others thanks to our attribute-based study.

In terms of precision rate and success rate, our method displayed strong performance across the majority of attributes. We evaluate our attribute-based performance against the most advanced trackers available. Table 4 shows our attribute-based performance findings.

Here are the attribute-based results:

Background clutter (BC): Our method obtained a success rate of 54.6% and a precision rate of 83.2% in sequences with high background clutter. This demonstrates that it is efficient in precisely tracking targets in crowded background environments. The evaluation curves of PR and SR curve for the background clutter challenge on the RGBT234 dataset can be seen in Figure 12 and 13, respectively.

Camera moving (CM): Our method performed well in sequences where the camera is in motion, achieving a precision score of 77.0% and a success rate of 54.9%. This demonstrates its ability to handle camera motion and maintain accurate tracking. The evaluation curve of PR and SR curve for the camera moving challenge on the RGBT234 dataset can be seen in Figure 12 and 13, respectively.

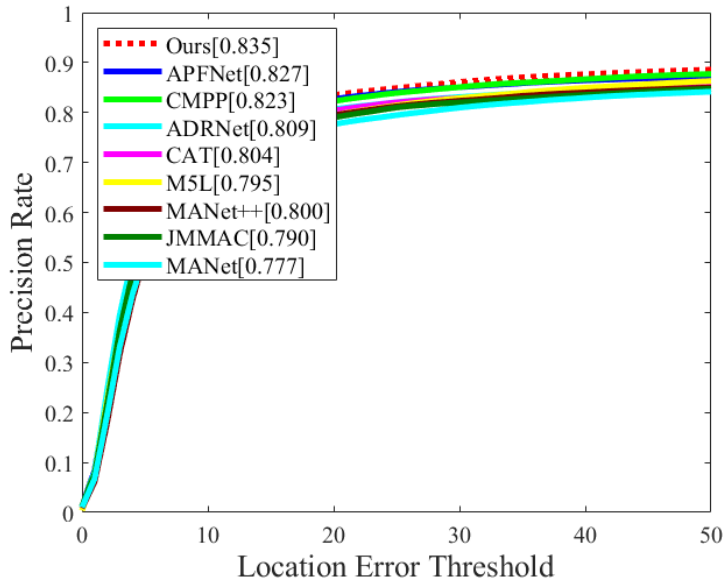


Figure 9: Precision Rate evaluation curve on the RGBT234 dataset.

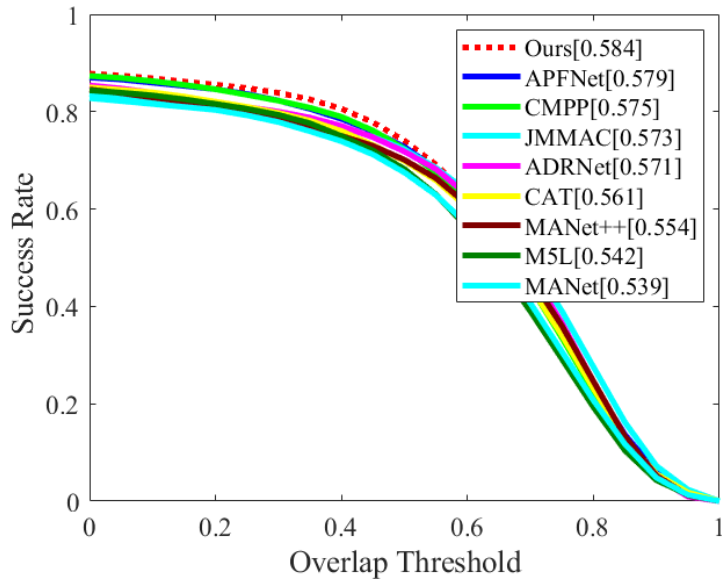


Figure 10: Success Rate evaluation curve on the RGBT234 dataset.

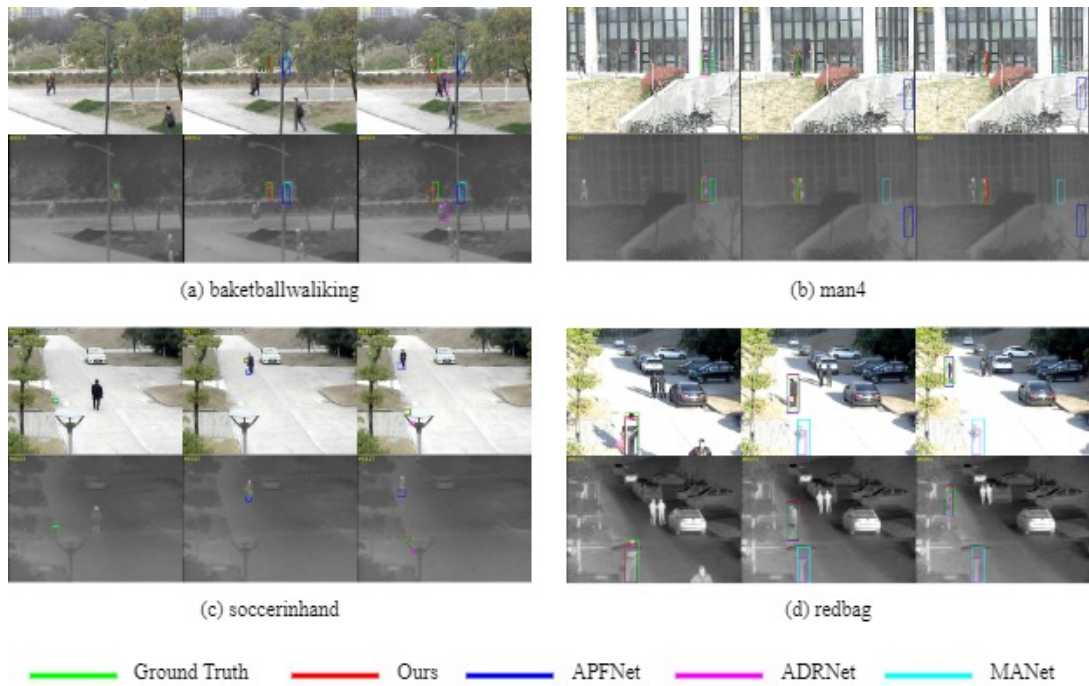


Figure 11: Comparison of EANet to three state-of-the-art trackers on different sequences. (a) The basketballwalking sequence, which has challenge of heavy occlusion; (b) The man4 sequence, which also has challenge of heavy occlusion (c) The soccerinhand sequence with the challenges of heavy occlusion, scale variation, thermal crossover, and fast motion; and (d) the redbag sequence with partial occlusion, deformation, and scale variation. The top row for each sequence displays the frames in RGB, while the bottom row displays the frames in thermal. Different colored rectangles are used to show the results of various trackers.

Deformation (DEF): Our method performed well in sequences with target deformation, achieving a precision score of 77.9% and a success rate of 55.4%. It demonstrates its capability to handle deformed target shapes and maintain accurate tracking. However, additional enhancements are needed in other aspects. The evaluation curve of PR and SR curve for the deformation challenge on the RGBT234 dataset can be seen in Figure 16 and 17, respectively.

Fast motion (FM): When faced with fast motion, our method achieved a precision score of 76.3% and a success rate of 49.2%. Although the success rate is slightly lower, improvements can be made in other areas to enhance overall performance. The evaluation curve of PR and SR curve for the fast motion challenge on the RGBT234 dataset can be seen in Figure 18 and 19, respectively.

Heavy occlusion (HO): The performance of our method was satisfactory even in situations when there was significant occlusion, as evidenced by a precision rate of 76.0% and a success rate of 52.2%. The findings of this study demonstrate that our methodology exhibits proficiency in handling complex scenarios characterized by heavy occlusion of the target. The evaluation curve of PR and SR curve for the heavy occlusion challenge on the RGBT234 dataset can be seen in Figure 20 and 21, respectively.

Low illumination (LI): Under low illumination conditions, our method achieved a precision score of 84.1% and a success rate of 56.6%. This highlights the strength of our approach in such situations. Nonetheless, there is room for improvement in other areas. The evaluation curve of PR and SR curve for the low illumination challenge on the RGBT234 dataset can be seen in Figure 22 and 23, respectively.

Low resolution (LR): With a success rate of 56.2% and a precision rate of 86.0%, our technique performed well in low-resolution sequences. This suggests that our method can work in scenarios when the target appears at a lesser resolution. However, additional improvements are required in other aspects. The evaluation curve of PR and SR curve for the low resolution challenge on the RGBT234 dataset can be seen in Figure 24 and 25, respectively.

Motion blur (MB): Our approach achieved 55.7% success rate and 76.9% precision score under motion blur circumstances. This demonstrates how well it tracks targets even when motion blur issues are present. The evaluation curve of PR and SR curve for the motion blur challenge on the RGBT234 dataset can be seen in Figure 26 and 27, respectively.

No occlusion (NO): Our method achieved a precision score of 93.4% and a success rate of 67.2% in sequences without occlusion. This demonstrates the resilience of our approach in properly tracking targets in the absence of occlusion. However, improvements are needed in other aspects as well. The evaluation curve of PR and SR curve for the no occlusion challenge on the RGBT234 dataset can be seen in Figure 28 and 29, respectively.

Partial occlusion (PO): Our technique achieved 60.6% success rate and 86.6% precision score when faced with partial occlusion. This demonstrates how our approach can handle targets that are partially occluded while maintaining precise tracking. The evaluation curve of PR and SR curve for the partial occlusion challenge on the RGBT234 dataset can be seen in Figure 30 and 31, respectively.

Scale variation (SV): With a success rate of 58.5% and a precision score of 83.1%, our approach showed good performance in sequences with scale variation. This demonstrates how well it can adapt to variations in target size and continue to track accurately. However, improvements are needed in

other aspects to further enhance overall performance. The evaluation curve of PR and SR curve for the scale variation challenge on the RGBT234 dataset can be seen in Figure 32 and 33, respectively.

Thermal crossover (TC): The precision rate of 82.8% and success rate of 59.2% is attained by our approach in sequences that had thermal crossover. This demonstrates how well our method tracks targets when the spatial modalities of RGB and thermal overlap. The evaluation curve of PR and SR curve for the thermal crossover challenge on the RGBT234 dataset can be seen in Figure 34 and 35, respectively.

The performance results based on attributes offer a thorough understanding of our method's effectiveness in addressing various difficulties within the RGBT234 dataset. The model exhibits a notable improvement of nearly five highest scores across twelve attributes. Our methodology exhibited good performance compared to the comparative trackers in terms of precision rate (PR) and success rate (SR), hence showcasing its efficacy across diverse and demanding settings.

	JMMAC	M5L	MANet++	MANet	CAT	ADRNNet	CMPP	APFNNet	Ours
BC	0.687/0.485	0.750/0.477	0.767/0.491	0.739/0.486	0.811/0.519	0.789/0.527	0.832/0.538	0.813/0.545	0.832/0.546
CM	0.762/0.556	0.752/0.529	0.747/0.523	0.719/0.508	0.752/0.527	0.757/0.535	0.756/0.541	0.779/0.563	0.770/0.549
DEF	0.706/0.529	0.736/0.511	0.753/0.535	0.720/0.524	0.762/0.541	0.743/0.529	0.750/0.541	0.785/0.564	0.779/0.554
FM	0.610/0.417	0.728/0.465	0.700/0.453	0.694/0.449	0.731/0.470	0.776/0.503	0.786/0.508	0.791/0.511	0.763/0.492
HO	0.677/0.483	0.665/0.450	0.704/0.471	0.689/0.465	0.700/0.480	0.708/0.491	0.732/0.503	0.738/0.507	0.760/0.522
LI	0.840/0.588	0.821/0.547	0.811/0.551	0.769/0.513	0.810/0.547	0.802/0.551	0.862/0.584	0.843/0.569	0.841/0.566
LR	0.771/0.517	0.823/0.535	0.823/0.545	0.757/0.515	0.820/0.539	0.831/0.556	0.865/0.571	0.844/0.565	0.860/0.562
MB	0.751/0.549	0.738/0.528	0.720/0.511	0.726/0.516	0.683/0.490	0.727/0.530	0.754/0.541	0.745/0.545	0.768/0.557
NO	0.932/0.694	0.931/0.646	0.898/0.654	0.887/0.646	0.932/0.668	0.917/0.658	0.956/0.678	0.948/0.680	0.934/0.672
PO	0.841/0.611	0.863/0.589	0.852/0.593	0.816/0.566	0.851/0.593	0.863/0.612	0.855/0.601	0.863/0.606	0.866/0.606
SV	0.837/0.616	0.796/0.542	0.789/0.554	0.777/0.542	0.797/0.566	0.790/0.562	0.815/0.572	0.831/0.579	0.831/0.585
TC	0.749/0.526	0.821/0.564	0.803/0.576	0.754/0.543	0.803/0.577	0.789/0.589	0.835/0.583	0.822/0.581	0.828/0.592
ALL	0.790/0.573	0.795/0.542	0.800/0.554	0.777/0.539	0.804/0.561	0.809/0.571	0.823/0.575	0.827/0.579	0.835/0.584

Table 4: RGBT234 dataset attribute-based PR and SR scores. The highest scores are shown in red color.

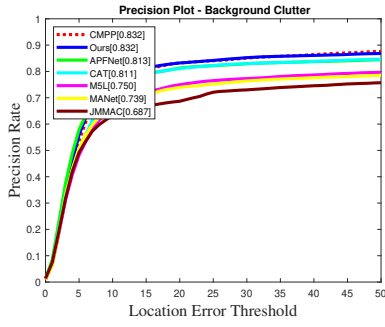


Figure 12: PR score of the BC challenge on the RGBT234 dataset

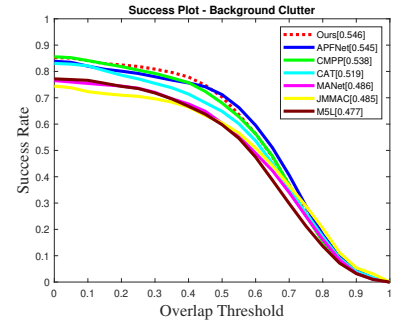


Figure 13: SR score of the BC challenge on the RGBT234 dataset

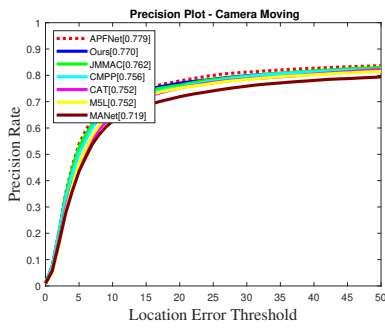


Figure 14: PR score of the Camera Moving challenge on the RGBT234 dataset

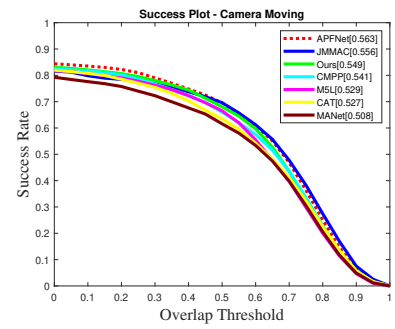


Figure 15: SR score of the Camera Moving challenge on the RGBT234 dataset

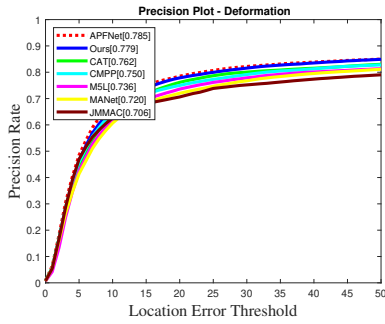


Figure 16: PR score of the Deformation challenge on the RGBT234 dataset

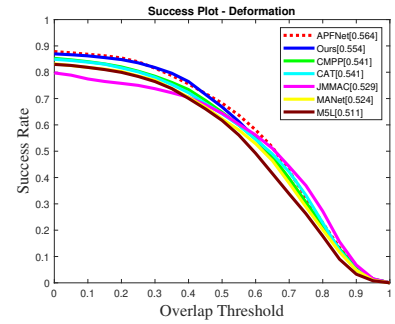


Figure 17: SR score of the Deformation challenge on the RGBT234 dataset

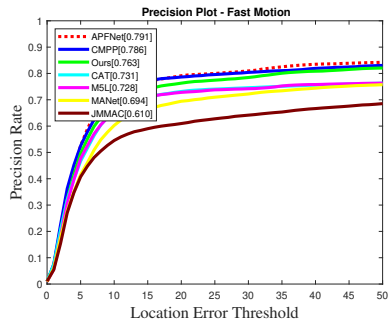


Figure 18: PR score of the Fast Motion challenge on the RGBT234 dataset

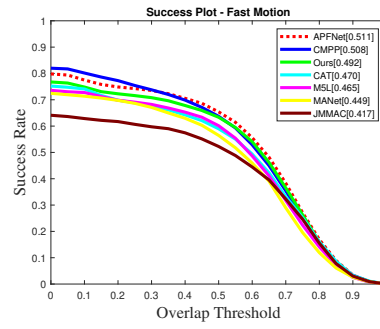


Figure 19: SR score of the Fast Motion challenge on the RGBT234 dataset

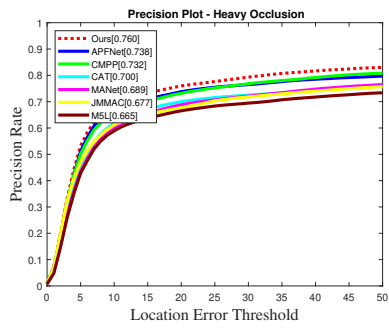


Figure 20: PR score of the Heavy Occlusion challenge on the RGBT234 dataset

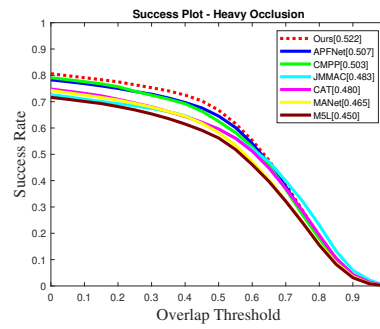


Figure 21: SR score of the Heavy Occlusion challenge on the RGBT234 dataset

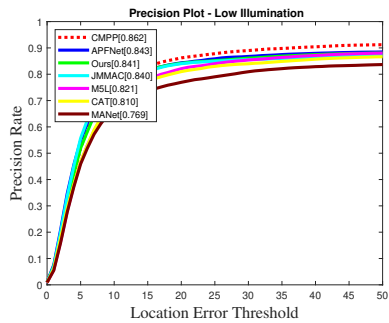


Figure 22: PR score of the Low Illumination challenge on the RGBT234 dataset

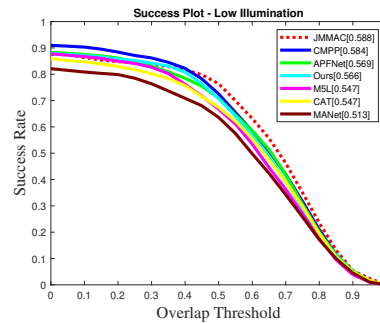


Figure 23: SR score of the Low Illumination challenge on the RGBT234 dataset

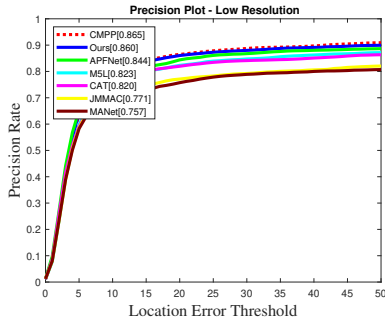


Figure 24: PR score of the Low Resolution challenge on the RGBT234 dataset

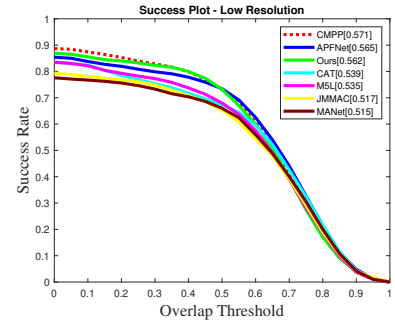


Figure 25: SR score of the Low Resolution challenge on the RGBT234 dataset

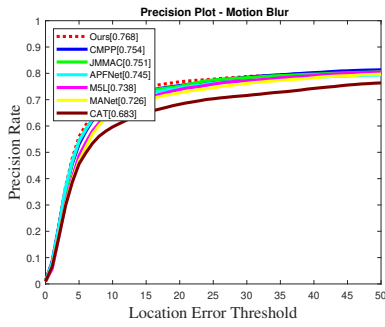


Figure 26: PR score of the Motion Blur challenge on the RGBT234 dataset

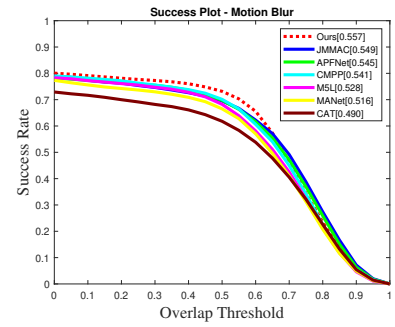


Figure 27: SR score of the Motion Blur challenge on the RGBT234 dataset

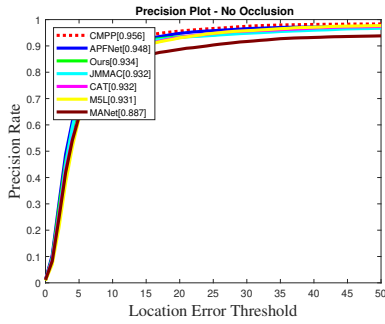


Figure 28: PR score of the No Occlusion challenge on the RGBT234 dataset

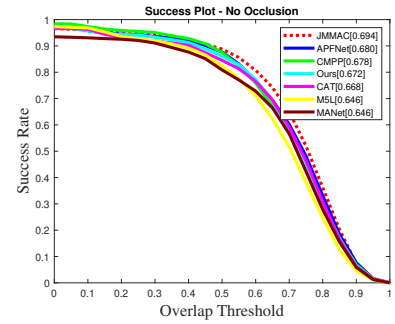


Figure 29: SR score of the No Occlusion challenge on the RGBT234 dataset

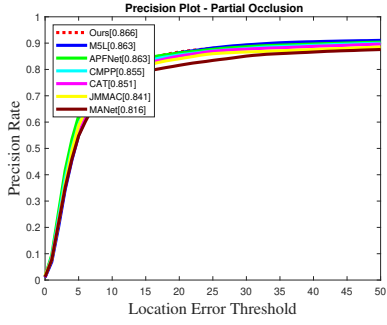


Figure 30: PR score of the Partial Occlusion challenge on the RGBT234 dataset

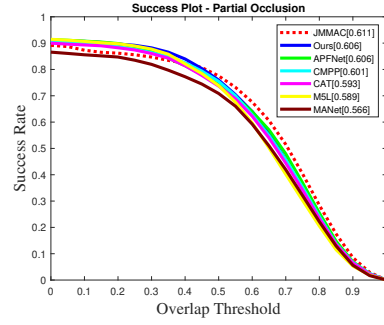


Figure 31: SR score of the Partial Occlusion challenge on the RGBT234 dataset

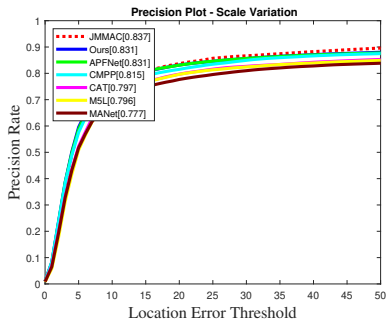


Figure 32: PR score of the Scale Variation challenge on the RGBT234 dataset

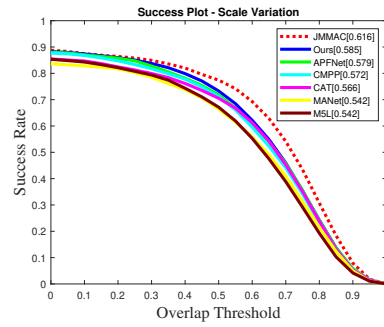


Figure 33: SR score of the Scale Variation challenge on the RGBT234 dataset

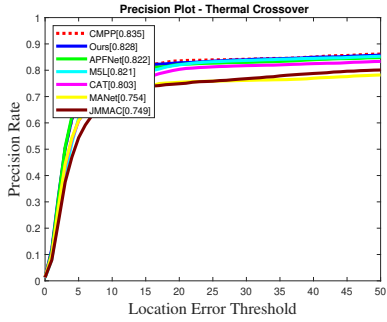


Figure 34: PR score of the Thermal Crossover challenge on the RGBT234 dataset

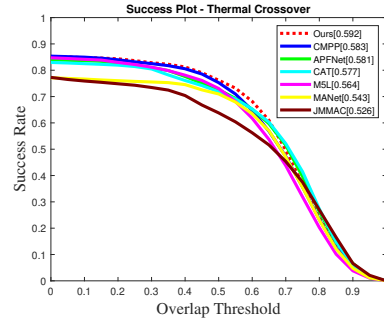


Figure 35: SR score of the Thermal Crossover challenge on the RGBT234 dataset

5.2 Evaluation on LasHeR Dataset

Furthermore, our technique is assessed on the LasHeR dataset in addition to the RGBT234 dataset, in order to further analyze its performance.

The evaluation on the LasHeR dataset verifies the efficacy of our model while exhibiting marginally inferior performance in comparison to the RGBT234 dataset. The performance outcomes on the Lasher dataset can be summarized as follows:

Precision (PR): The precision rate attained by our approach is 50.6%. While the rate achieved by our methodology is comparatively lower than that of the RGBT234 dataset, it outperforms other trackers, hence showcasing the efficacy of our method in precisely localizing and tracking targets. Figure 36 shows the Precision Rate evaluation curve.

Success rate (SR): Our technique achieves a 36.7% success rate. Although the obtained rate may appear relatively lower, it is indicative of the difficulties inherent in the Lasher dataset and the considerable capability of our approach to effectively address these obstacles. The success rate evaluation curve is shown in Figure 37. The SR score of our model exhibits the best value in comparison to other trackers.

The evaluation conducted on the LasHeR dataset provides more evidence to substantiate the resilience and adaptability of our approach in effectively addressing various tracking circumstances and obstacles. Although the performance of our technique was significantly inferior in comparison to the RGBT234 dataset, it demonstrated its efficacy in tracking targets across various conditions.

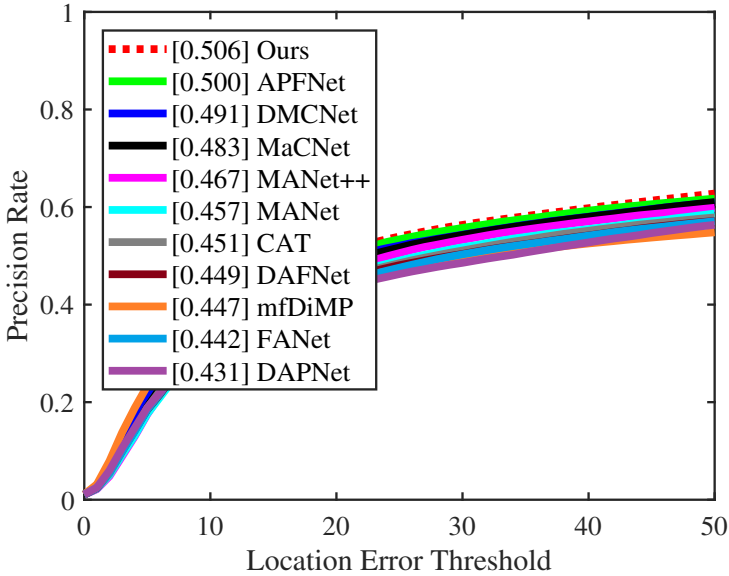


Figure 36: Precision Rate evaluation curve on the LasHeR dataset.

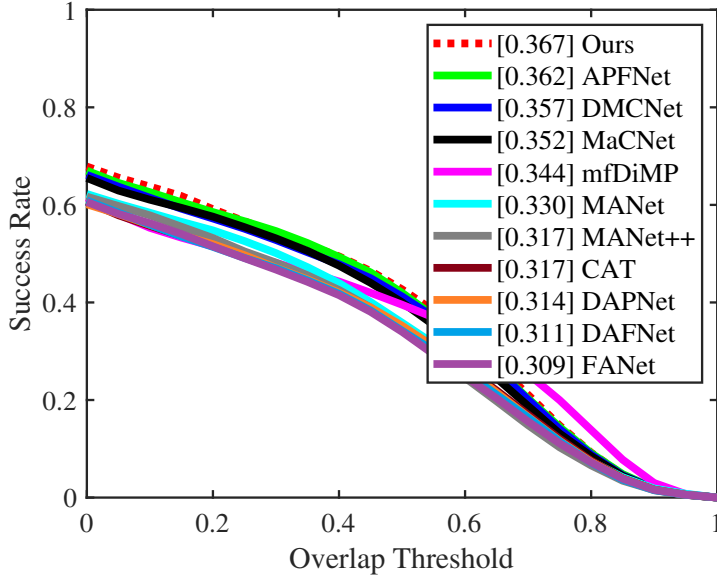


Figure 37: Success Rate evaluation curve on the LasHeR dataset..

5.3 Computational Efficiency

The computational efficiency of an object tracking model is a critical factor for its real-time applicability. In this section, we compare our suggested model to various state-of-the-art RGB-T tracking models in order to assess its computing efficiency.

The frame per second (fps) rate of our model is 1.5, which is far from real-time. This is because our model does not use any ROI layers [61, 62]. ROI layers are a common way to accelerate the speed of object tracking models by extracting features from a small region of interest (ROI) around the target object. The crucial task of directly collecting features from feature maps for the area of interest of each tracked object is performed by ROI layers. This method greatly lessens the computational load, which quickens the model’s execution speed. ROI layers are known to increase frame rates by reducing the number of unnecessary calculations made across the entire image and by concentrating exclusively on pertinent object sections. Although our proposed model does not have enough fps rate for real-time object tracking, it has a higher fps rate than other models that also do not use ROI layer, as can be seen in Table 5. By not using ROI layers, our model has to extract features from the entire image, which is more computationally expensive.

Some other RGB-T tracking models that use ROI layers have achieved higher fps rates. For example, the DAFNet [23] achieves a fps rate of 20, and the ADRNet [34] achieves a fps rate of 25. However, these models also use more complex features and algorithms, which may sacrifice accuracy for speed. Comparison of FPS Rates for state-of-art RGB-T Tracking Models in the literature can be seen in Table 5.

Our proposed model is not as computationally efficient as some other RGB-T tracking models. However, it achieves good accuracy without using RoI layers, which makes it a promising approach for real-time object tracking.

Table 5: Comparison of FPS Rates for RGB-T Tracking Models

Tracker	Speed (fps)	RoI Layer
MANet++	25.4	RoI Align [61]
ADNet	25	RoI Pooling [62]
CAT	20	RoI Align [61]
DAFNet	20	RoI Align [61]
FANet	19	RoI Align [61]
mfDiMP	10.3	RoI Pooling [62]
M5L	9.75	RoI Align [61]
Ours	1.5	-
CMPP	1.3	-
APFNet	1.3	-
MANet	1.11	-
MaCNet	0.8	-

5.4 Experiments and Ablation Study

When we started working on this thesis, we chose APFNet as the baseline. We focused on discovering what is open for improvement or what is missing from this model. In all experiments, we used GTOT as the training dataset and RGBT234 as the test dataset. We trained the system in three steps with the GTOT dataset as described in the paper. We also tested it on the RGBT234 dataset. APFNet performs fusion with the "Attribute-Based Progressive Fusion Module". As mentioned in the paper, this module has three primary components: "Attribute-Specific Fusion Branch", "Attribute-Based Aggregation Fusion", and "Attribute-Based Enhanced Fusion" [36]. "Attribute-Based Enhanced Fusion" [36] is a transformer structure consisting of three encoders and two decoders. Encoders and decoders contain self single-head attention and cross single-head attention respectively, to keep the system simple [36]. We first tried to change the attention mechanism within the encoders and decoders from single-head to multi-head. As can be seen in the "multi-head" column in Table 6, our PR score was 81.3% and our SR score was 56.0%. Since this PR/SR value is lower than our baseline model with 82.7% and 57.9%, we conclude that changing the single-head attention mechanism to multi-head does not contribute to the model.

Our second experiment was based on a new fusion strategy different from the baseline model. As seen in Figure 38, we tried Attentional Feature Fusion (AFF) [63] to fuse the features, where the data from RGB and Thermal images were fused with AFF on each layer, and the data obtained from the fusion on each layer was subjected to element-wise addition. With this model, we obtained PR and SR values of 78.9% and 54.5% respectively. We assessed that this fusion strategy alone is not enough.

We then used this fusion strategy to aggregate data from Attribute-based Feature Branches. We called the variation of the model "ABr+AFF" and the structure of this variation can be seen in Figure 39. We obtained PR and SR values of 79.8% and 55.3% respectively. We found that AFF contributed to the

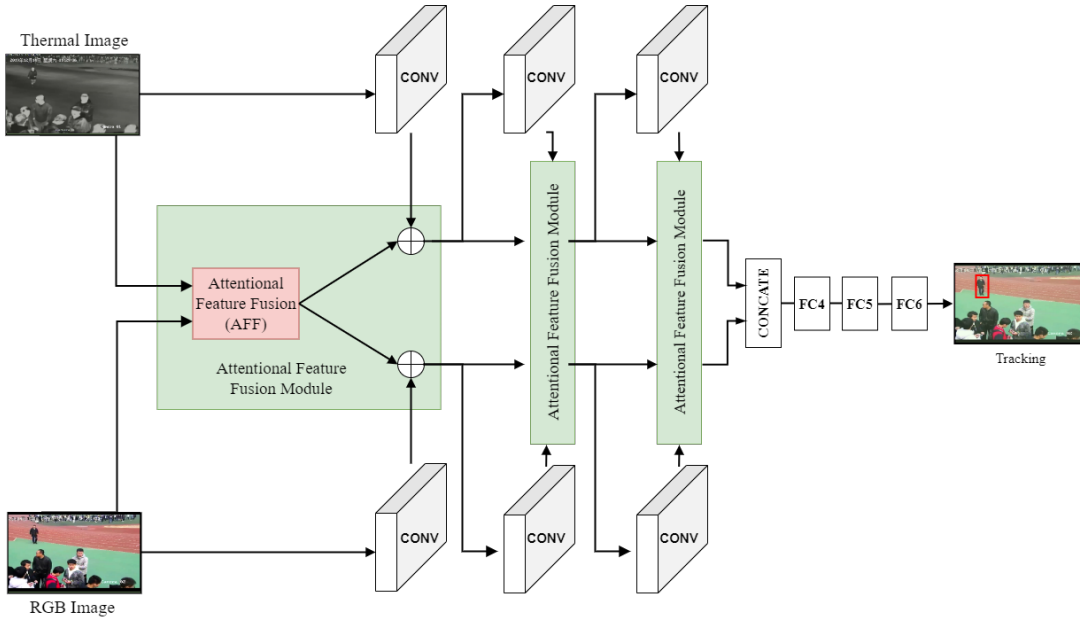


Figure 38: RGBT Tracking Model With Attentional Feature Fusion Module

success with the aggregate process, but this variation of the model was not sufficiently successful. We used AFF again for aggregate operation and enhanced the attribute-based branches with ESK. With this variation called "ESKBr+AFF", we obtained PR and SR values of 81.6% and 56.2%. The Structure of "ESKBr+AFF" Variation can be seen in Figure 40. Since we could not get enough efficiency from our experiments with AFF, we focused on ESK. We had already used ESK in Attribute-based branches, we thought we could also use it for Aggregate, so we used ESK. In this way, we reached the final form of our model.

On the other hand, in order to evaluate how well the individual components of our model work together, we carried out an ablation study. We removed the Enhanced Aggregation Module from the model in order to determine how much of a role it played in the overall success of the model. To do this, we aggregated the data coming from the Enhanced Fusion Branch Modules using element-wise addition. In Table 6, we referred to this particular model iteration as "Var-AggESK." The 'Var-AggESK' version obtained a PR score of 81.2% and an SR score of 56.4%. When we compare "Var-AggESK" with the suggested model, as can be seen in Table 6 we concluded that the Enhanced Aggregation Module is a crucial part of our approach and contributes to its success. In other words, the Enhanced Aggregation Module is required for our model.

Table 6: Results of Ablation Study Evaluation on RGBT234 Dataset

	AFF	ABr+AFF	Multi-head	ESKBr+AFF	Var-AggESK	Ours
PR	0.789	0.798	0.813	0.816	0.812	0.835
SR	0.545	0.553	0.560	0.562	0.564	0.584

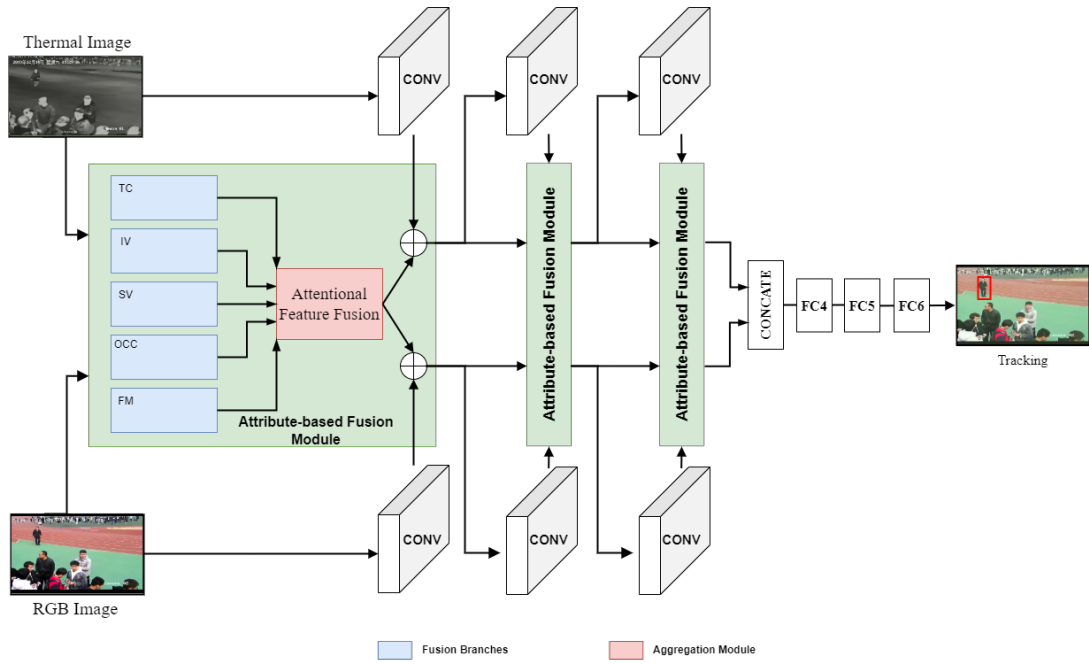


Figure 39: The Structure of "ABr+AFF" Variation

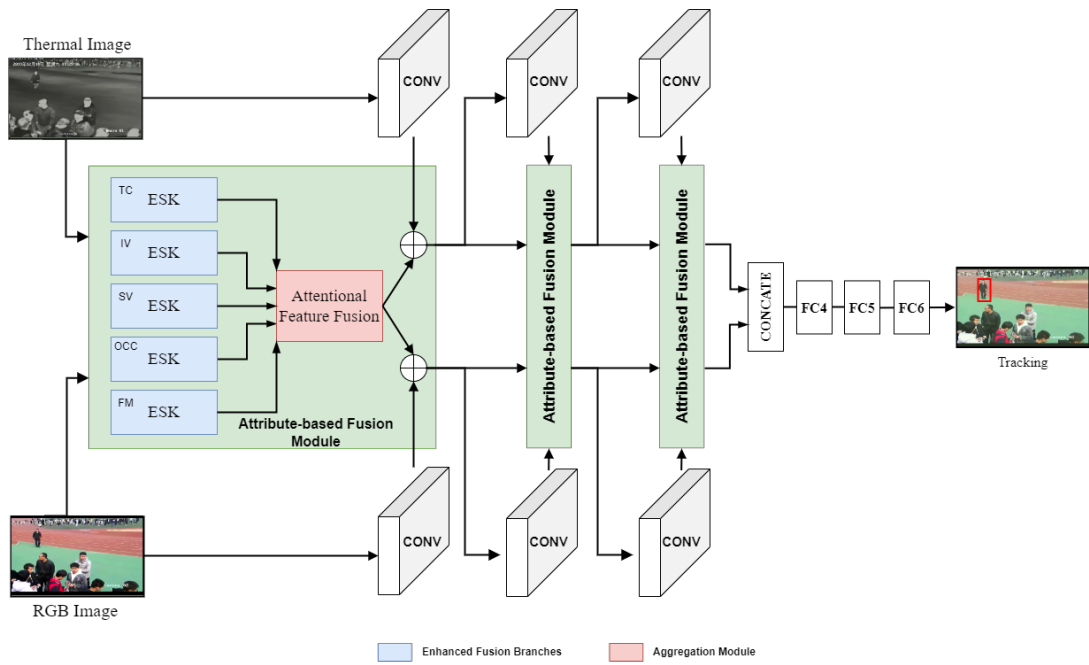


Figure 40: The Structure of "ESKBr+AFF" Variation

5.4.1 Experiments on Robustness of Our Model

In this section, we evaluate the robustness of our deep learning-based object tracking system. We train the model four times independently to evaluate its stability and consistency, and we observe variations in the precision rates and success rates throughout these experiments. We conduct four separate training sessions, and the precision and success rates achieved in each experiment on RGBT234 and LasHeR datasets can be seen in Table 7 and 8.

The variations in precision rates and success rates observed in these experiments offer valuable insights into the robustness of our model. We notice a relatively steady range of precision rates across the four experiments: values for the RGBT234 dataset range from 0.825 to 0.835, while for the LasHeR dataset, values range from 0.497 to 0.506. This consistency suggests that, even under varying training sessions. Like precision rates, success rates exhibit only minor variations, ranging from 0.577 to 0.584 for the RGBT234 dataset and 0.363 to 0.369 for the LasHeR dataset. These minor variations imply that our model consistently achieves a high level of tracking object success across various training sessions. Our model is resilient to modifications in the training procedure, as seen by the minimal variations in both success rates and precision rates. In conclusion, our experiments on the robustness of the deep learning-based object tracking system show that it keeps up its performance consistently across numerous training sessions.

Table 7: Results of Experiments on RGBT234 Dataset

Experiment	Precision Rate	Mean	Variance	Success Rate	Mean	Variance
Experiment-1	0.835	0.830	0.0000193	0.584	0.580	0,000009
Experiment-2	0.828			0.578		
Experiment-3	0.825			0.577		
Experiment-4	0.832			0.581		

Table 8: Results of Experiments on LasHeR Dataset

Experiment	Precision Rate	Mean	Variance	Success Rate	Mean	Variance
Experiment-1	0.506	0.503	0.000018	0.367	0.366	0.000007
Experiment-2	0.506			0.369		
Experiment-3	0.504			0.365		
Experiment-4	0.497			0.363		

5.5 Experiments on Other RGBT Tracking Models

The experimental findings from training and evaluating five different RGBT object tracking models are presented in this section. The RGBT234 dataset is used to test these models after they are trained on the GTOT dataset. The aim of the study is to evaluate these models' performance in relation to the findings presented in the corresponding papers.

Table 9: Results of Experiments on Other RGBT Tracking Models Tested on RGBT234 Dataset

Model	Experimental Precision Rate	Precision Rate on Paper	Experimental Success Rate	Success Rate on Paper
APFNet [36]	0.817	0.827	0.569	0.579
ADRNet [34]	0.807	0.809	0.570	0.571
MANet [24]	0.798	0.777	0.560	0.539
MANet++ [28]	0.795	0.800	0.559	0.554
DAFNet [23]	0.794	0.796	0.540	0.544

The precision and success rates obtained from our experiments are compared to those reported in the papers for each model. The results of the experiments are shown in Table 9.

Our experiments show that, on the RGBT234 dataset, the chosen RGBT tracking models' performance is largely in line with the findings published in the papers. Although there are small changes in success rates and precision, these variations can be ascribed to a number of things, such as differences in the preparation of the dataset, the optimization settings, and the tracking environment.

The resilience of the APFNet and ADRNet models in RGBT tracking was demonstrated by their precision rates, which were generally extremely close to the original research findings. Although their success rates were slightly lower than those of the original study, MANet and MANet++ performed admirably. The MANet model is considered to be developed after the first version of the article is published. With a slightly lower success rate than the original article, DAFNet demonstrated competitive performance.

CHAPTER 6

CONCLUSION

In conclusion, our approach demonstrates exceptional efficacy in object tracking tasks, as evidenced by its impressive performance on the RGBT234 and LasHeR datasets. The assessment conducted on the RGBT234 dataset revealed the method's notable accuracy and efficacy, yielding remarkable outcomes in terms of precision and success rate. The robustness of the attribute-based analysis is further demonstrated through its successful mitigation of many challenges, including but not limited to motion blur, fluctuations in illumination, occlusion, and other factors.

While the performance scores obtained from evaluating the LasHeR dataset were marginally lower compared to our results on the RGBT234 dataset, our model outperformed other trackers and reaffirmed the effectiveness of our model in accurately tracking targets across diverse scenarios. Taking into account the difficulties that are unique to the LasHeR dataset, the success rate and precision rate that we got suggest that our method is effective.

Overall, our approach demonstrates considerable promise in the domain of RGBT fusion object tracking. The accuracy, versatility, and capability of handling challenging scenarios of the model are supported by its performance on the RGBT234 and Lasher datasets. The findings of our assessment make a valuable contribution to the progression of object tracking methodologies and establish a solid groundwork for subsequent investigations and innovations in this domain.

Moving forward, we plan to continue refining our method and exploring ways to further enhance its performance. We think that with continuous improvement, our approach can contribute to the development of more accurate and robust RGBT object tracking systems, opening up new possibilities for applications in various domains.

REFERENCES

- [1] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, “Rgb-t object tracking: Benchmark and baseline,” *Pattern Recognition*, vol. 96, p. 106977, 2019.
- [2] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun, “Lasher: A large-scale high-diversity benchmark for rgbt tracking,” *IEEE Transactions on Image Processing*, vol. 31, pp. 392–404, 2022.
- [3] A. Türkoğlu and E. Akagündüz, “Eanet: Enhanced attribute-based rgbt tracker network,” 2023.
- [4] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, “Geodesic active contour based fusion of visible and infrared video for persistent object tracking,” in *2007 IEEE Workshop on Applications of Computer Vision (WACV '07)*, pp. 35–35, 2007.
- [5] N. Cvejic, S. G. Nikolov, H. D. Knowles, A. Loza, A. Achim, D. R. Bull, and C. N. Canagarajah, “The effect of pixel-level fusion on object tracking in multi-sensor surveillance video,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2007.
- [6] C. Beyan and A. Temizel, “Mean-shift tracking for surveillance applications using thermal and visible band data fusion,” in *Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications VIII* (D. J. Henry, B. T. Cheng, D. C. L. von Berg, and D. L. Young, eds.), vol. 8020, p. 802010, International Society for Optics and Photonics, SPIE, 2011.
- [7] M. Isard and A. Blake, “Condensation - conditional density propagation for visual tracking,” *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.
- [8] C. Conaire, N. O’Connor, and A. Smeaton, “Thermo-visual feature fusion for object tracking using multiple spatiogram trackers,” *Machine Vision and Applications*, vol. 19, pp. 483–494, 2008.
- [9] X. Mei and H. Ling, “Robust visual tracking using $l(1)$ minimization,” in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1436–1443, 2009.
- [10] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, “Multiple source data fusion via sparse representation for robust visual tracking,” in *14th International Conference on Information Fusion*, pp. 1–8, 2011.
- [11] H. Liu and F. Sun, “Fusion tracking in color and infrared images using joint sparse representation,” *Science China Information Sciences*, vol. 55, pp. 590–599, 2012.
- [12] L. Li, C. Li, Z. Tu, and J. Tang, “A fusion approach to grayscale-thermal tracking with cross-modal sparse representation,” in *Image and Graphics Technologies and Applications* (Y. Wang, Z. Jiang, and Y. Peng, eds.), (Singapore), pp. 494–505, Springer Singapore, 2018.

- [13] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, “Modality-correlation-aware sparse representation for rgb-infrared object tracking,” *Pattern Recognition Letters*, vol. 130, pp. 12–20, 2020. Image/Video Understanding and Analysis (IUVA).
- [14] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, “Object fusion tracking based on visible and infrared images: A comprehensive review,” *Information Fusion*, vol. 63, pp. 166–187, 11 2020.
- [15] 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, (New York, NY, USA), p. 1856–1864, Association for Computing Machinery, 2017.
- [16] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, “Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking,” Jan 2018.
- [17] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, and J. Tang, “Learning local-global multi-graph descriptors for rgb-t object tracking,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 2913–2926, 2019.
- [18] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4293–4302, 2016.
- [19] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, ..., and Z. H. Niu, “The visual object tracking vot2014 challenge results,” in *Computer Vision - ECCV 2014 Workshops* (L. Agapito, M. M. Bronstein, and C. Rother, eds.), (Cham), pp. 191–217, Springer International Publishing, 2015.
- [20] Y. Zhu, C. Li, J. Tang, and B. Luo, “Quality-aware feature aggregation network for robust rgbt tracking,” *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 121–130, 2021.
- [21] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, “Dense feature aggregation and pruning for RGBT tracking,” in *Proceedings of the 27th ACM International Conference on Multimedia*, ACM, oct 2019.
- [22] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.
- [23] Y. Gao, C. Li, Y. Zhu, J. Tang, T. He, and F. Wang, “Deep adaptive fusion network for high performance rgbt tracking,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 91–99, 2019.
- [24] C. L. Li, A. Lu, A. H. Zheng, Z. Tu, and J. Tang, “Multi-adapter rgbt tracking,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2262–2270, 2019.
- [25] R. Yang, Y. Zhu, X. Wang, C. Li, and J. Tang, “Learning target-oriented dual attention for robust rgb-t tracking,” in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3975–3979, 2019.

- [26] C. Wang, C. Xu, Z. Cui, L. Zhou, T. Zhang, X. Zhang, and J. Yang, “Cross-modal pattern-propagation for rgb-t tracking,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7062–7071, 2020.
- [27] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang, “Duality-gated mutual condition network for rgbt tracking,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.
- [28] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, “Rgbt tracking via multi-adapter network with hierarchical divergence loss,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5613–5625, 2021.
- [29] Q. Xu, Y. Mei, J. Liu, and C. Li, “Multimodal cross-layer bilinear pooling for rgbt tracking,” *IEEE Transactions on Multimedia*, vol. 24, pp. 567–580, 2022.
- [30] Z. Tu, C. Lin, W. Zhao, C. Li, and J. Tang, “M5l: Multi-modal multi-margin metric learning for rgbt tracking,” *IEEE Transactions on Image Processing*, vol. 31, pp. 85–98, 2022.
- [31] H. Zhang, L. Zhang, L. Zhuo, and J. Zhang, “Object tracking in rgb-t videos using modal-aware attention network and competitive learning,” *Sensors (Switzerland)*, vol. 20, 1 2020.
- [32] Y. Zhu, C. Li, J. Tang, B. Luo, and L. Wang, “Rgbt tracking by trident fusion network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 579–592, 2 2022.
- [33] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, “Challenge-aware rgbt tracking,” in *Computer Vision – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 222–237, Springer International Publishing, 2020.
- [34] P. Zhang, D. Wang, H. Lu, and X. Yang, “Learning adaptive attribute-driven representation for real-time rgb-t tracking,” *International Journal of Computer Vision*, vol. 129, pp. 2714–2729, 9 2021.
- [35] Y. B. Xavier Glorot, Antoine Bordes, “Deep sparse rectifier neural networks,” in *International Conference on Artificial Intelligence and Statistics*, 2011.
- [36] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, “Attribute-based progressive fusion network for rgbt tracking,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 2831–2838, 6 2022.
- [37] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-convolutional siamese networks for object tracking,” in *Computer Vision – ECCV 2016 Workshops* (G. Hua and H. Jégou, eds.), (Cham), pp. 850–865, Springer International Publishing, 2016.
- [38] X. Zhang, P. Ye, D. Qiao, J. Zhao, S. Peng, and G. Xiao, “Object fusion tracking based on visible and infrared images using fully convolutional siamese networks,” in *2019 22th International Conference on Information Fusion (FUSION)*, pp. 1–8, 2019.
- [39] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, and G. Xiao, “Siamft: An rgb-infrared fusion tracking method via fully convolutional siamese networks,” *IEEE Access*, vol. 7, pp. 122122–122133, 2019.
- [40] X. Zhang, P. Ye, S. Peng, J. Liu, and G. Xiao, “Dsiammft: An rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion,” *Signal Processing: Image Communication*, vol. 84, 5 2020.

- [41] C. Guo, D. Yang, C. Li, and P. Song, “Dual siamese network for rgb-t tracking via fusing predicted position maps,” *The Visual Computer*, vol. 38, pp. 2555–2567, 7 2022.
- [42] T. Zhang, X. Liu, Q. Zhang, and J. Han, “Siamcda: Complementarity- and distractor-aware rgb-t tracking based on siamese network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 1403–1417, 3 2022.
- [43] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, “Fusing two-stream convolutional neural networks for rgb-t object tracking,” *Neurocomputing*, vol. 281, pp. 78–85, 2018.
- [44] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. Shahbaz Khan, “Multi-modal fusion for end-to-end rgb-t tracking,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2252–2261, 2019.
- [45] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, “Learning discriminative model prediction for tracking,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6181–6190, 2019.
- [46] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, and X. Yang, “Jointly modeling motion and appearance cues for robust rgb-t tracking,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3335–3347, 2021.
- [47] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “Eco: Efficient convolution operators for tracking,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6931–6939, 2017.
- [48] J. Zhu, G. He, and P. Zhou, “Mfnet: A novel multilevel feature fusion network with multibranch structure for surface defect detection,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.
- [49] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, “Learning collaborative sparse representation for grayscale-thermal tracking,” *IEEE Transactions on Image Processing*, vol. 25, pp. 5743–5756, 2016.
- [50] C. Luo, B. Sun, Q. Deng, Z. Wang, and D. Wang, “Comparison of different level fusion schemes for infrared-visible object tracking: An experimental survey,” in *2018 2nd International Conference on Robotics and Automation Sciences (ICRAS)*, pp. 1–5, 2018.
- [51] S. R. Schnelle and A. L. Chan, “Enhanced target tracking through infrared-visible image fusion,” in *14th International Conference on Information Fusion*, pp. 1–8, 2011.
- [52] C. Conaire, N. O’Connor, E. Cooke, and A. Smeaton, “Comparison of fusion methods for thermo-visual surveillance tracking,” in *2006 9th International Conference on Information Fusion*, pp. 1–7, 2006.
- [53] A. L. Chan and S. R. Schnelle, “Target tracking using concurrent visible and infrared imageries,” in *Signal Processing, Sensor Fusion, and Target Recognition XXI* (I. Kadar, ed.), vol. 8392, p. 83920P, International Society for Optics and Photonics, SPIE, 2012.
- [54] A. Chan and S. Schnelle, “Fusing concurrent visible and infrared videos for improved tracking performance,” *Optical Engineering*, vol. 52, 2013.

- [55] C. Tang, Y. Ling, H. Yang, X. Yang, and W. Tong, “Decision-level fusion tracking for infrared and visible spectra based on deep learning,” *Laser and Optoelectronics Progress*, vol. 56, 2019.
- [56] Z. Tang, T. Xu, H. Li, X.-J. Wu, X. Zhu, and J. Kittler, “Exploring fusion strategies for accurate rgbt visual object tracking,” *Information Fusion*, vol. 99, p. 101881, 2023.
- [57] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, ..., and Z. Ma, “The eighth visual object tracking vot2020 challenge results,” in *Computer Vision – ECCV 2020 Workshops* (A. Bartoli and A. Fusiello, eds.), (Cham), pp. 547–601, Springer International Publishing, 2020.
- [58] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, ..., and Z. Ni, “The seventh visual object tracking vot2019 challenge results,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2206–2241, 2019.
- [59] G. Chen, J. Zhang, Y. Liu, J. Yin, X. Yin, L. Cui, and Y. Dai, “Esknet-an enhanced adaptive selection kernel convolution for breast tumors segmentation,” 2022.
- [60] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” 2019.
- [61] I. Jung, J. Son, M. Baek, and B. Han, “Real-time mdnet,” 8 2018.
- [62] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, “Acquisition of localization confidence for accurate object detection,” in *Computer Vision – ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), (Cham), pp. 816–832, Springer International Publishing, 2018.
- [63] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, “Attentional feature fusion,” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3559–3568, 2021.