electronics

MDPI

Article

DuSiamIE: A Lightweight Multidimensional Infrared-Enhanced RGBT Tracking Algorithm for Edge Device Deployment

Jiao Li, Haochen Wu, Yuzhou Gu, Junyu Lu and Xuecheng Sun *

School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China; lijiaoshu@shu.edu.cn (J.L.); wuhaochen@shu.edu.cn (H.W.); xaviergu@shu.edu.cn (Y.G.); ljy284089474@shu.edu.cn (J.L.)
* Correspondence: sunxc@shu.edu.cn

Abstract: Advancements in deep learning and infrared sensors have facilitated the integration of RGB-thermal (RGBT) tracking technology in computer vision. However, contemporary RGBT tracking methods handle complex image data, resulting in inference procedures with a large number of floating-point operations and parameters, which limits their performance on general-purpose processors. We present a lightweight Siamese dual-stream infrared-enhanced RGBT tracking algorithm, called DuSiamIE.It is implemented on the low-power NVIDIA Jetson Nano to assess its practicality for edge-device applications in resource-limited settings. Our algorithm replaces the conventional backbone network with a modified MobileNetV3 and incorporates light-aware and infrared feature enhancement modules to extract and integrate multimodal information. Finally, NVIDIA TensorRT is used to improve the inference speed of the algorithm on edge devices. We validated our algorithm on two public RGBT tracking datasets. On the GTOT dataset, DuSiamIE achieved a precision (PR) of 83.4% and a success rate (SR) of 66.8%, with a tracking speed of 40.3 frames per second (FPS). On the RGBT234 dataset, the algorithm achieved a PR of 75.3% and an SR of 52.6%, with a tracking speed of 34.7 FPS. Compared with other algorithms, DuSiamIE exhibits a slight loss in accuracy but significantly outperforms them in speed on resource-constrained edge devices. It is the only algorithm among those tested that can perform real-time tracking on such devices.

Keywords: RGBT tracking; Siamese; infrared-enhanced; NVIDIA Jetson Nano; NVIDIA TensorRT

check for updates

Citation: Li, J.; Wu, H.; Gu, Y.; Lu, J.; Sun, X. DuSiamIE: A Lightweight Multidimensional Infrared-Enhanced RGBT Tracking Algorithm for Edge Device Deployment. *Electronics* **2024**, *13*, 4721. https://doi.org/10.3390/electronics13234721

Academic Editor: George A. Tsihrintzis

Received: 23 October 2024
Revised: 21 November 2024
Accepted: 27 November 2024
Published: 29 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The demands for computer vision are continuously increasing due to the advancements in video surveillance, autonomous driving, and human–computer interaction technologies. Target tracking is a fundamental research area in this field. It entails continuously monitoring an object's location across consecutive frames by utilizing data from the initial frame. This approach is widely used in various domains, such as security surveillance and autonomous driving [1]. However, tracking systems frequently encounter challenging scenarios such as dynamic changes in background illumination, target occlusions, rapid target movements, and deformations, all of which pose significant difficulties for accurate tracking. Specifically, dynamic changes in background illumination can lead to substantial variations in image brightness and contrast, making it difficult to consistently identify the target under different lighting conditions. Target occlusions can partially or completely hide the target, preventing tracking algorithms from accurately detecting the target's position. Rapid target movements can cause image blurring or large frame-to-frame position changes, rendering traditional tracking methods ineffective. These factors collectively have a severe impact on the accuracy and reliability of tracking systems. Although target tracking methods have been developed to address these challenges, conventional RGB tracking, owing to its dependence on visible-light imaging, is prone to failure in scenarios characterized by pronounced variations in illumination, occlusions, or rapid target dynamics, thereby posing significant challenges to the maintenance of robustness and real-time performance

in practical applications [2]. Therefore, improving robustness and dynamic performance in real-world scenarios has become a primary focus of research in target tracking [3].

The traditional approach for target tracking typically uses cameras and other equipment to capture visual data of objects. However, relying on a single modality presents challenges in acquiring sufficient information about the object. Specifically, visible light imaging is susceptible to environmental factors such as fluctuations in illumination, rain, smoke, and haze, which can significantly impair image quality. In contrast, infrared imaging is less impacted by these conditions but suffers from lower resolution and reduced textural details [4]. Furthermore, during thermal crossover, as shown in Figure 1, distinguishing between the target and the background within infrared images becomes difficult. To overcome these limitations, an integrated vision tracker that combines both visible and thermal infrared modalities can exploit their inherent correlation and complementarity. This multimodal approach alleviates the limitations and uncertainties associated with single-modal information, thereby enhancing the reliability of the tracking system [5].



Figure 1. Examples of poor lighting conditions (**left**) and exposure to shade (**right**).

The integration of multiple modalities necessitates the management of a substantially larger volume of feature information, thereby demanding increased storage capacity [6], enhanced computational power, and extended computation times [7]. Accurate target tracking demands superior real-time performance compared to other visual tasks. Edge devices, such as drones [8], autonomous vehicles [9], and wearable gadgets with limited processing capabilities, require both high precision and real-time frame rates [10]. A significant challenge in the practical implementation of RGBT tracking is ensuring both tracking precision and real-time performance on resource-constrained edge devices [11].

Recently, the increasing prevalence of edge computing devices, such as the NVIDIA Jetson Nano, has facilitated the advancement of effective target-tracking algorithms in resource-limited environments. However, there remains a paucity of research dedicated to reducing the computational complexity and resource consumption of these models. Achieving accurate real-time target tracking on edge devices continues to pose significant challenges [12]. Consequently, developing efficient and lightweight multimodal target-tracking models for edge devices has become a primary focus of contemporary research [13].

This study presents a novel and efficient dual-stream RGBT tracking network specifically designed for deployment on edge devices. The contributions of this work are primarily as follows:

1. **Dual-stream Siamese network:** We present a specialized module designed to enhance infrared features and address the problem of feature loss during extraction, particularly the under-utilization of infrared features commonly encountered in light sensing. The incorporation of this module enhances the PR and SR by 2.7% and 3.3%, respectively.
2. **Development of an infrared feature enhancement module:** We introduce a novel dual-stream Siamese network with optical sensing architecture for edge devices, which demonstrates significant improvements in the PR and SR of 10.5% and 12.4%, respectively, compared to a single-stream Siamese network.
3. **Real-time tracking optimization using NVIDIA TensorRT:** To facilitate real-time tracking on resource-constrained edge devices, we utilize NVIDIA TensorRT. The results demonstrate that the operating speed increased from 18.4 frames per second (FPS) to 40.3 FPS, thereby fully meeting the real-time requirements for target tracking.

The subsequent sections of the paper are organized as follows: Section 2 presents a review of pertinent research on RGBT target tracking. In Section 3, a comprehensive discussion is provided regarding the three fundamental components of the proposed tracking methodology. Section 4 offers a summary of the experimental results and analyses conducted on publicly available RGBT datasets. Finally, Section 5 concludes the paper.

2. Related Work

2.1. RGB Tracking

RGB target tracking methodologies utilize visible light images to monitor and analyze the positional dynamics of targets across consecutive frames. These methodologies are broadly categorized into two primary paradigms: correlation filter-based methods and deep learning-based approaches. Correlation filter-based RGB tracking systems, such as the minimum output sum of squared error (MOSSE) algorithm [14], operate by matching target templates in the frequency domain [15], achieving computational efficiency under simplistic environmental conditions. However, these methods exhibit significant limitations in complex scenarios, including low illumination, background clutter, and target occlusions, where their robustness and tracking accuracy degrade considerably. Advanced techniques, such as multi-task deep dual correlation filters (MDDCF) [16], have been proposed to address these challenges by integrating alternating optimization and dual correlation mechanisms. Nevertheless, the inherent dependency of these methods on visible-light imaging restricts their adaptability in adverse and dynamic environmental conditions.

Similarly, deep learning-based RGB tracking approaches, exemplified by the Siamese fully convolutional (SiamFC) neural network [17], frame the tracking task as a template matching problem, thereby enhancing computational efficiency. However, these methods often exhibit a lack of robustness when confronted with variations in target size or environmental disturbances. For instance, the Siamese box adaptive network (SiamBAN), which introduces an anchor-free tracking framework capable of directly regressing target position and size, circumvents complexities associated with predefined anchor boxes [18]. Despite this innovation, SiamBAN demonstrates suboptimal performance in scenarios involving significant size variations or severe occlusions. Additionally, while the discriminative tracker multi-domain network (MDNet) [19] achieves superior accuracy by training multi-domain classifiers to distinguish targets from background datasets, its real-time applicability is hindered by the computational overhead associated with online training, limiting its robustness in dynamic environments [20].

Although RGB tracking methodologies have achieved considerable advancements, they still encounter significant challenges in maintaining robust performance under adverse conditions such as low illumination, occlusion, and dynamically changing environments. These inherent limitations highlight the necessity for multi-modal tracking approaches that

can leverage complementary information from diverse sensor modalities. By integrating data from multiple sources, multi-modal tracking systems can enhance both the accuracy and robustness of target tracking in complex and challenging scenarios.

2.2. RGBT Tracking

To address the limitations inherent in RGB tracking, researchers have increasingly focused on multi-modal approaches, such as RGBT tracking, which integrates information from both visible and infrared sensors. The fusion of these complementary modalities helps mitigate issues related to occlusion and poor visibility, thereby ensuring more reliable performance across a broader range of environmental conditions.

The integration of sensing devices and the advancement of deep learning has spurred a growing interest in RGBT target tracking [21]. SiamFT [22] and dynamic Siamese networks with multi-layer fusion DSiamMFT [23] expand the RGB Siamese network architecture by incorporating a two-stream network designed to address the challenge of RGBT fusion tracking. Despite the Siamese-based network's simplicity, enabling faster computation, its inability to fully utilize the capabilities of infrared and visible light modalities leads to reduced accuracy [24]. Furthermore, the multi-adapter convolutional network (MANet) [25] employs a multi-adapter network that performs modality-sharing, modality-specific, and instance-aware feature learning within a unified deep architecture for RGBT tracking. Additionally, the inclusion of a hierarchical multimodal fusion method in the hierarchical multi-modal fusion tracker (HMFT) [26] greatly enhances accuracy. This approach not only extracts shared features but also focuses on learning modality-specific features.

Recent studies have introduced sophisticated fusion strategies to enhance tracking performance. The decision fusion adaptive tracker (DFAT) [27] addresses biases in RGB and infrared modalities through dynamic weighting, achieving robust performance across diverse conditions. It also optimizes template updates for long-term accuracy. The attribute-based progressive fusion network (APFNet) [28] uses five attribute-specific branches to tackle challenges like thermal crossover, illumination variation, and occlusion. By employing lightweight models, APFNet reduces dependency on large-scale training data while maintaining accuracy. Its aggregation fusion module and enhancement fusion transformer effectively combine and strengthen features, demonstrating superior performance on GTOT and RGBT234 datasets. However, its relatively high computational complexity remains a limitation for edge device applications [29,30].

3. Methodology

This paper presents the architecture of a lightweight Siamese dual-stream infrared-enhanced RGBT tracking algorithm (DuSiamIE). Unlike conventional RGB tracking systems, this method integrates complementary features from both infrared and visible-light imaging, enhancing robustness against environmental challenges such as low illumination, occlusions, and rapid motion, as shown in Figure 2.

The algorithm integrates a light-aware module into a lightweight Siamese network tailored for edge devices. We introduce the multidimensional infrared feature enhancement (MIFE) module, which enhances infrared features and improves modal fusion details, addressing the issue of inadequate feature extraction and missing infrared modalities in light-perception-guided modal fusion processes. Furthermore, to enable real-time inference on edge devices, we optimize the algorithm's feature extraction and classification-regression layers to further enhance the performance of real-time tracking, as detailed in Section 3.1.

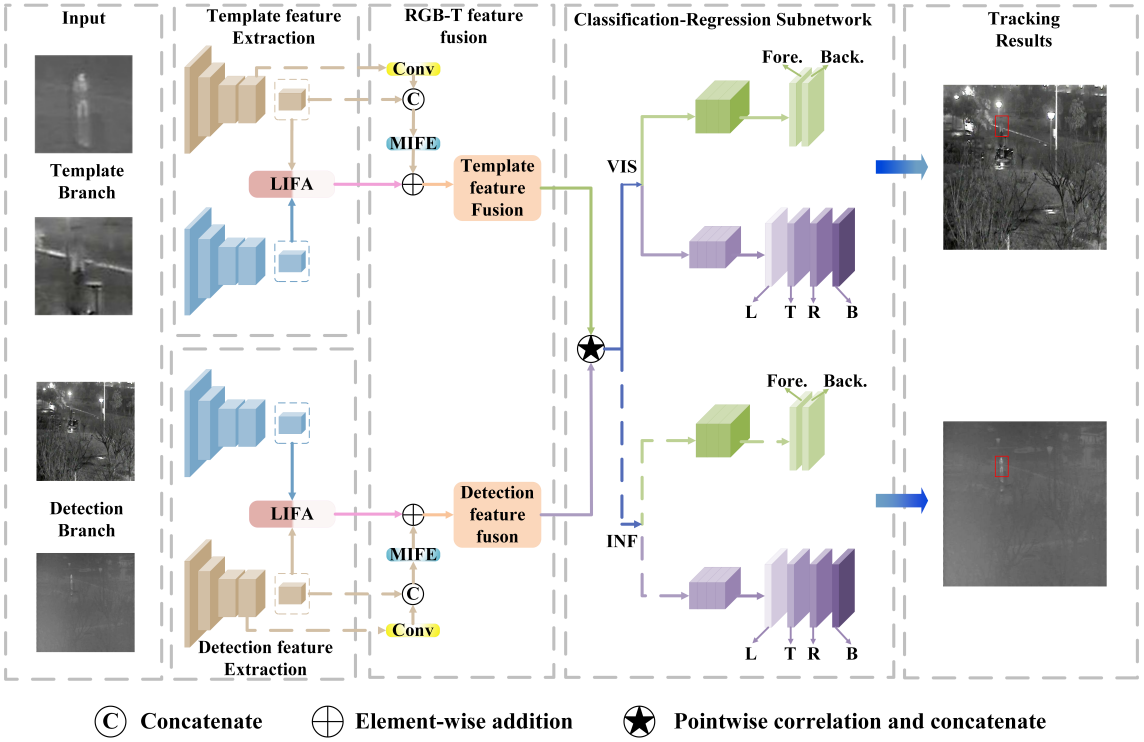


Figure 2. The DuSiamIE network: feature extraction via a Siamese network, modal fusion with the LIFA module, infrared enhancement using the MIFE module, and tracking output after classification and regression.

3.1. Lightweight Dual Stream Siamese Network Optimized for Edge Deployment

Target tracking technology has witnessed significant advancements in recent years. The limitations of RGB target tracking have spurred a shift towards the development of tracking algorithms that utilize multimodal fusion. The integration of multimodal information can offer a more comprehensive understanding of target characteristics and enhance the robustness and accuracy of tracking algorithms. However, the processing of multimodal data leads to higher computational requirements and reduced real-time performance, particularly when deployed on edge devices. To mitigate this problem, we expand the RGB tracking network to ensure that the algorithm remains lightweight and efficient while achieving high-precision tracking.

This study introduces a dual-stream Siamese network that comprises two branches, each dedicated to processing its distinct input modality. To facilitate implementation on edge devices, an optimized version of the lightweight network architecture, MobileNetV3, is employed. It is made to reduce computational overhead, thereby enabling efficient deployment and acceleration on these devices. During the feature extraction phase, only the output features from the fourth-layer inverted residual block are used for subsequent processing.

Under optimal illumination, visible light can capture features such as color, shape, and texture. Conversely, in low-light or night-time conditions, infrared images can highlight the thermal features of a target object by capturing its heat distribution and temperature differences. To address this, we propose the light-aware feature aggregation (LIFA) module, which optimizes the fusion process by considering illumination conditions. This allows the module to adapt to the illumination of the input video and to identify and exploit complementary and common features between different modalities. The structure of the LIFA module is shown in Figure 3. The initial visible light frame is input into the network with illumination perception capabilities. Features are further extracted through convolution, and the model is stabilized by a batch normalization layer to accelerate the training process. To achieve feature abstraction, fusion, and dimensionality reduction, two fully connected layers are utilized. Consequently, two weights denoted as ω_{vi} and

ω_{ir}, which represent the illumination information, are obtained. The weights are derived using the softmax function. Despite being remarkably lightweight, this perceptual network possesses sufficient sensitivity to precisely detect illumination information and allocate weights for fusion.

Infrared features are often underutilized in contemporary optical sensing networks. This study introduces an approach that tackles this problem by enhancing infrared features across multiple dimensions using the MIFE module. The combined and enhanced template and search feature data are then fed into a high-level neural network for further classification and regression tasks. This study employs anchor-free tracking in the final head network designed for classification and regression tasks. The proposed methodology deviates from earlier techniques that employed predefined anchor boxes for tracking. Instead, the algorithm directly predicts the target's centroid and size, thereby reducing the parameter count [31]. To generate the regression and classification feature maps, the template and search features are concatenated at the pixel level. Furthermore, dimensionality reduction techniques are implemented to decrease computational load, thereby improving processing efficiency. This method is further refined to achieve a lightweight design. To complete the target tracking process, the classification score map and regression offset are derived from the classification and regression prediction layers, respectively.

During tracking, the object may exhibit localized movement near the boundary or may move beyond the frame limits. To mitigate frame drift, this work proposes the use of reflective padding. This technique involves mirroring pixels at the image's edges into the convolutional padding zone [32]. We specify the edge coordinates as (i, j), where i represents the horizontal axis and j represents the vertical axis. Given a border pixel with a value of p, reflective padding transfers this value to the adjacent padding location.

Horizontal direction:

$$X_{padded}(i, j + P) = X(i, j). \tag{1}$$

Vertical direction:

$$X_{padded}(i + P, P) = X(i, j). \tag{2}$$

This form of padding is generally considered to have a more natural visual impact compared to zero padding. This is because it preserves the uninterrupted structure of the image edges and reduces edge effects, thereby enhancing the authenticity of edge processing. The method successfully mitigates the loss of information in the fused image when objects are located near the boundary. Moreover, the padding procedure efficiently reduces the influence of similarly shaped objects at the boundary. Subsequently, convolution is used to extract relevant illumination properties, and adaptive average pooling (AAP) establishes the dimensions of the output feature maps, ensuring the efficient preservation of spatially essential information for subsequent tracking tasks.

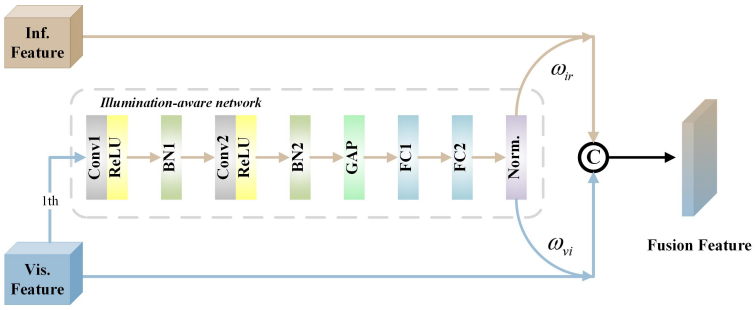


Figure 3. LIFA module.

3.2. Multidimensional Infrared Feature Enhancement Module

The assessment of the DuSiamIE algorithm indicates that in environments with rapid fluctuating illumination and significant target-background contrast, the algorithm frequently encounters difficulties in differentiating between tracked objects, resulting in

tracking failures. To resolve this issue, we employ mathematical modeling and provide the LIFA-guided fusion method outlined in the preceding sections.

The variable L denotes light information, encompassing aspects such as spectral distribution, geographical distribution, and reflectance. It mostly denotes the intensity of light. The aim is to formulate a mathematical equation that accurately delineates the entire process and extracts and integrates the feature information of L. Equation (3) delineates the complete feature information $F(L)$, comprising the visible light information $f_{VI}(L)$ and the infrared feature information $f_{IR}(L)$, amalgamated by a weighting factor.

$$F(L) = \omega_{vi}(L) \cdot f_{VI}(L) + \omega_{ir}(L) \cdot f_{IR}(L). \tag{3}$$

The weighting coefficients ω_{ir} and ω_{vi}, obtained from light perception, are associated with light intensity and meet the equation $\omega_{ir} + \omega_{vi} = 1$. In the model,

$$\omega_{vi}(L) = \frac{L}{L+M}, \tag{4}$$

$$\omega_{ir}(L) = \frac{M}{L+M}. \tag{5}$$

In this context, M denotes an environmental constant that is separate from light. When L is minimal, ω_{vi} approaches 0, and ω_{ir} approaches 1; conversely, when L is maximal, the reverse occurs.

The infrared feature information primarily relies on the thermal radiation of the object and is typically considered a constant function:

$$f_{IR}(L) = c. \tag{6}$$

The modeling of the visible light feature information $f_{VI}(L)$ is a more complex operation. The feature information of visible light is inversely related to its intensity; hence, the maximum value can only be attained under optimal lighting circumstances. Consequently, an entropy model similar to the Gaussian distribution is utilized, expressed by the following equation:

$$f_{vi}(L) = A \cdot \exp(-\frac{(L-L_0)^2}{2\sigma^2}). \tag{7}$$

In this application, L_0 is the illumination level at which the optimal quantity of feature information is accessible. A represents the maximal quantity of information, whereas σ indicates the breadth of the control curve.

Substituting the aforementioned function into Equation (3) yields the following result:

$$F(L) = \frac{L}{L+M} \cdot A \cdot \exp(-\frac{(L-L_0)^2}{2\sigma^2}) + \frac{M}{L+M} \cdot c. \tag{8}$$

In the limit of high light intensities ($L \to \infty$), the total feature information is predominantly determined by the visible light information, with the infrared information being effectively disregarded. Conversely, at low light intensities, infrared feature information should, in theory, be the primary source of information. However, it was discovered that under specific low-light conditions, infrared features were unable to exert a dominant influence, with the majority of feature information still derived from visible light.

This is due to the fact that as L decreases, while the IR weight $\omega_{ir}(L)$ increases, the IR feature information c remains constant and relatively insignificant. The visible light feature information, $f_{VI}(L)$, although decreasing with L, is likely to remain high during the rapid transition from high to low light. This is due to the smoothness of its exponential decay, which results in a continuous dominance of the visible mode contribution.

The derivation of Equation (8) is as follows:

$$F'(L) = \frac{L}{(L+M)^2} \cdot f_{vi}(L) + \frac{L}{L+M} \cdot f'_{vi}(L) - \frac{M}{(L+M)^2} \cdot c. \tag{9}$$

In which,

$$f'_{vi}(L) = \frac{A}{-\sigma^2} \cdot (L - L_0) \cdot exp(-\frac{(L-L_0)^2}{2\sigma^2}).$$

In instances of rapid alterations in illumination and elevated target-background contrast, the parameters σ and A exhibit a notable increase, resulting in a considerable rise in $f_{VI}(L)$ and $f'_{vi}(L)$. These remain the dominant components of the total information $F(L)$. Consequently, the potential of infrared features is not fully exploited across a broad spectrum, and multimodal fusion is insufficient to attain optimal tracking outcomes in certain instances.

To address this issue, this paper proposes enhancements to the infrared features. However, this paper finds that merely adjusting the fusion weights is insufficient to fully utilise the infrared features, and this approach also affects the ability to perceive light normally. Accordingly, this paper exploits the distinctive visual characteristics of infrared images, which predominantly document temperature distribution and exhibit minimal color and texture information. Furthermore, such feature information is typically distributed in feature layers that are shallower than those containing visible light features.

In light of the aforementioned considerations, we propose the design of the MIFE module. This method combines shallow infrared features with the current layer of infrared features and performs multi-scale fusion enhancement, facilitating cross-scale information interaction and capturing more valuable infrared feature information. The MIFE module facilitates the acquisition of additional shallow infrared feature information and introduces self-attention, which captures global information across different dimensions and enhances the representation of infrared features. The incorporation of MIFE significantly enhances tracking performance in scenarios characterized by rapid illumination changes and difficulty in distinguishing between the foreground and background.

In Figure 4a, the architectural structure of MIFE is shown to consist of three parallel branches. The cross-interactions between the C, H, and W dimensions are captured by the latter two branches of the input feature map $X_{MSI} \in R^{C \times H \times W}$ by rotating along the H and W axes. After the multi-dimensional fusion processes, the three branches implement an attention mechanism, producing intermediate features across multiple dimensions. As illustrated in Equations (9) and (10), the final output is achieved via weighted averaging $X'_{MSI} \in R^{C \times H \times W}$, where C, H, and W represent the processes in three distinct dimensions:

$$X'_{MSI_i} = Dim_i(X_{MSI}) + X_{MSI}, i = C, H, W, \tag{10}$$

$$X'_{MSI} = \frac{X'_{MSI_C} + X'_{MSI_H} + X'_{MSI_W}}{3}. \tag{11}$$

In this section, we describe three different operations, each representing a distinct dimension. Specifically, in the first branch, the operation along the C dimension (channel dimension operation) aims to eliminate the influence of the channel dimension while constructing the dependency between height and width dimensions, denoted as H:W. As illustrated in the green section of Figure 4b, assume the input feature map is given by $X_{HW} \in R^{C \times H \times W}$. The input feature map X_{HW} is first adjusted into three distinct forms as inputs for the self-attention mechanism: $X^Q_{HW} \in R^{N \times C}$, $X^K_{HW} \in R^{C \times N}$, and $X^V_{HW} \in R^{C \times N}$, where Q, K, and V represent the query, key, and value matrices, respectively, that are used in the self-attention operation to capture the internal dependencies. Here, $N = H \times W$ represents the number of spatial locations.

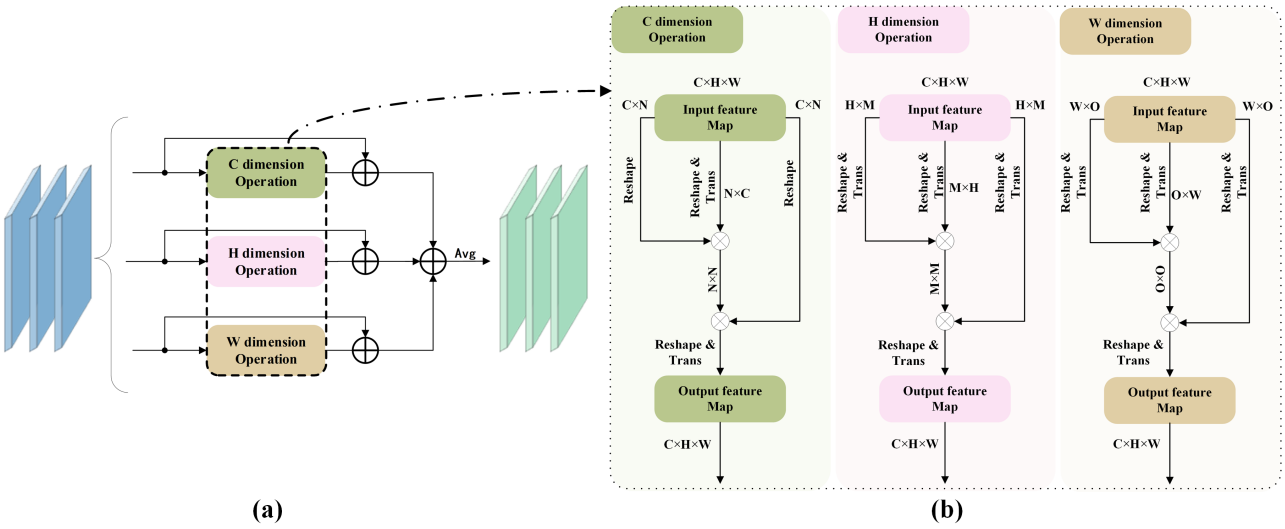


Figure 4. MIFE module: (**a**) MIFE's network structure. (**b**) Structure of the mechanism of self-attention in different dimensions.

Next, we perform matrix multiplication between X_{HW}^{Q} and X_{HW}^{K}, yielding a feature vector $X_{HW}^{QK} \in R^{N\times N}$. This operation is intended to retain rich details while squeezing the depth of the channel dimension. Subsequently, the optimized weights are obtained by applying the Sigmoid function to $X_{HW}^{QK} \in R^{N\times N}$. Finally, a matrix multiplication between X_{HW}^{V} and X_{HW}^{QK} is performed, followed by a weighted summation with the original input X_{MSI} to obtain the final attention-fused output $X_{MSI_C}^{'}$.

In the second branch, the operation along the H dimension (height dimension operation) aims to eliminate the influence of the height dimension while establishing dependencies between the channel and width dimensions, denoted as *C*:*W*. As illustrated in the pink section of Figure 4b, the input feature map $X_{MSI} \in R^{C\times H\times W}$ is rotated counterclockwise by 90 degrees along the H axis, yielding the feature map $X_{CW} \in R^{W\times H\times C}$. To obtain the output, we use the same self-attention mechanism and Sigmoid activation function as in the first branch. Finally, the output is rotated clockwise by 90 degrees along the H axis to maintain the original shape of the features, followed by a weighted summation with the input to obtain the output feature map $X_{MSI_H}^{'}$.

In the third branch, interactions are conducted along the W axis, which involves the operation along the W dimension (width dimension operation) to eliminate the influence of the width dimension and establish dependencies between the channel and height dimensions, denoted as *C*:*H*. As illustrated in the brown section of Figure 4b, the input feature map $X_{MSI} \in R^{C\times H\times W}$ is rotated counterclockwise by 90 degrees along the W axis, yielding the feature map $X_{CH} \in R^{H\times C\times W}$. Similarly, we employ the same self-attention mechanism and Sigmoid activation function as in the first branch. Finally, the output is rotated clockwise by 90 degrees along the W axis to maintain the original feature shape, followed by a weighted summation with the input to obtain the output feature map $X_{MSI_W}^{'}$.

The uniform geometry of the MIFE module's input and output feature maps allows its application to any layer, thus helping to improve feature representation. The MIFE module is added to DuSiamIE before the fourth feature extraction layer to incorporate shallow infrared features. This integration uses a sophisticated separable convolutional architecture to capture cross-dimensional interactions and improve the representation of infrared information, effectively resolving the problem of limited infrared feature contribution. This work empirically confirms its efficacy in the ablation experiment described in Section 4.4.

3.3. Optimization of Inference Speed for Embedded Devices

Efficient online processing is fundamental for practical applications of multimodal target tracking systems on edge devices. Enhancing the computational efficiency and real-time performance of the model while maintaining tracking accuracy is of great research importance, especially for embedded devices with constrained resources.

The present study investigates the implementation of lightweight models on edge devices and proposes optimization strategies to enhance operational efficiency. We implement the algorithms on edge devices to assess their real-time inference performance. This enables the complete process, including acquisition, processing, and result storage, to be executed entirely on the edge device. An NVIDIA Jetson Nano with 4 GB of memory was chosen as the edge device for this research. This development board has a processing capacity of 4.7 trillion operations per second (TOPS) at 10 W, which is adequate for supporting deep learning model inference [33].

We further optimized the model using the NVIDIA TensorRT deep learning inference optimizer to accommodate current software and hardware constraints. This optimization streamlined the network architecture and enhanced the efficiency of resource utilization of the underlying hardware components. In deep learning model inference, the convolutional, activation, and normalization layers explicitly invoke the appropriate CUDA version 12.2 interfaces. The integration of layers and tensors using NVIDIA TensorRT enables the consolidation of the initial network layers into a unified module, thereby reducing GPU utilization, enhancing computational efficiency and preserving the fundamental operations of the network.

The precise operating process of NVIDIA TensorRT is depicted in Figure 5. Initially, the network architecture is vertically fused by integrating the convolution, bias, and rectified linear unit (ReLU) layers into a composite "CBR" layer. The network architecture is subsequently consolidated horizontally, whereby layers in the neural network that receive identical input tensors and execute identical operations are combined into a single layer. Ultimately, NVIDIA TensorRT facilitates the direct linkage of multiple output matrices and buffers, thus eliminating the need for a separate concatenation operation. This enables the output of the present layer to be immediately connected to the subsequent layer [34].

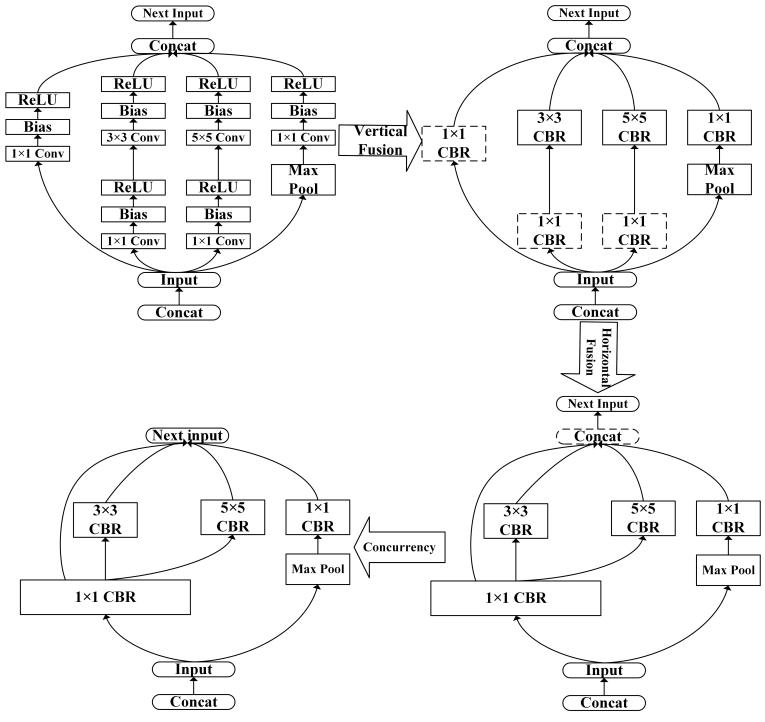


Figure 5. NVIDIA TensorRT optimization process.

The majority of deep learning frameworks employ 32-bit floating-point (FP32) precision for tensors during the training process [35]. Although higher-precision data types can improve accuracy, they also require more memory and computational resources. However, in real-world inference scenarios, since the process does not involve backpropagation, it is possible to reduce computational precision without significantly affecting accuracy. NVIDIA TensorRT's precision calibration technology allows the conversion of FP32 models into lower-precision FP16 or INT8 formats. These optimizations reduce memory usage, latency, and model size. By employing calibration, NVIDIA TensorRT effectively quantizes FP32 models to INT8 precision while preserving performance, thus preventing degradation due to reduced precision. This methodology aims to maintain accuracy while decreasing resource consumption, thereby enabling more convenient deployment in resource-constrained environments. Hence, the lightweight network described in this study can be implemented on an embedded edge computing development board, facilitating effective real-time processing for tracking visible light and infrared targets.

An analysis of the model's performance on the NVIDIA Jetson Nano embedded platform reveals that, after NVIDIA TensorRT acceleration, the real-time tracking frame rate increases dramatically from 18.4 fps to 40.3 fps. This significant enhancement in frame rate adequately satisfies the real-time tracking requirements of edge devices.

4. Experiments and Discussions

4.1. Datasets and Evaluation Indicators

4.1.1. GTOT Dataset and Metrics

The GTOT [36] dataset contains 50 video pairs from different scenarios and conditions, and each pair consists of visible and thermal infrared videos. It consists of frames with manually labelled real-world scenarios, and the challenge attributes are classified into seven groups based on the state of the target. Two widely used evaluation metrics, the PR and SR from the one-time evaluation (OPE), are used as evaluation metrics for the tracker. For GTOT, the target object is usually small, and in this paper, the threshold is set to 5 pixels following previous work.

4.1.2. RGBT234 Dataset and Metrics

The RGBT234 [37] dataset is a large-scale RGBT tracking dataset with 234 sequences and 12 challenge attributes. The acquisition devices for this dataset are a thermal infrared camera and a CCD camera, and the imaging parameters of the two cameras are identical, which ensures a high degree of accuracy in the alignment between visible and thermal infrared sequence pairs. Following previous work, the evaluation metrics in this paper are based on the maximum precision rate (MPR) and maximum success rate (MSR), both with a threshold of 20 pixels.

4.2. Experimental Environment

This study conducted the training of the proposed algorithm on a Windows 10 operating system, utilizing an AMD Ryzen 7 4800H CPU (4.20 GHz) and an NVIDIA GeForce RTX 2060 GPU with 6 GB of dedicated RAM. The software environment comprised Python 3.8.18, CUDA 11.6, and PyTorch 1.13.0. Finally, to demonstrate the lightweight nature of the algorithm during inference, it was deployed on an NVIDIA Jetson Nano edge device equipped with a quad-core ARM Cortex-A57 CPU, 4 GB of LPDDR4 RAM, and a 128-core Maxwell GPU. The NVIDIA Jetson Nano, with its balance of computational power and low power consumption, is well-suited for assessing the efficiency of the proposed model in real-time edge computing environments.

During the initial training phase, the backbone network of the tracker was trained using multiple RGB datasets, including COCO [38], TrackingNet [39], LaSOT [40], and DET [41], aiming to achieve optimal tracking performance. The backbone architecture adopted a pre-trained MobileNetV3-Small model, which provides a good trade-off between efficiency and accuracy. To effectively transfer learning from the pre-trained model, the

backbone parameters were frozen during the first 10 epochs to train only the head network. Subsequently, selective fine-tuning was performed on the feature extraction layers of the backbone, allowing it to adapt specifically to the tracking task while retaining the knowledge from the original pre-training.

The training process employed the stochastic gradient descent (SGD) optimizer with a weight decay of 1×10^{-4} and a momentum of 0.9. The model was trained for 200 epochs with a batch size of 64. The learning rate was initialized at 0.005 and decayed logarithmically to 0.0005, providing a gradual reduction that helps maintain stability as the model converges. A warm-up strategy was implemented during the first 5 epochs, with the learning rate starting from 0.001 and increasing incrementally by 0.001 per epoch. This strategy ensured a stable initialization phase, thereby preventing gradient-related issues during the early stages of training.

After the initial training on RGB datasets, the dual-stream Siamese network was further refined using the RGBT234 dataset. This retraining aimed to improve the model's robustness to variations in illumination by leveraging both visible (RGB) and thermal information. During this phase, the SGD optimizer with identical hyperparameters was used, and training was conducted for an additional 20 epochs to adjust both the LIFA and MIFE modules. This retraining process leveraged the previously learned features while adapting the model to the additional thermal domain.

Following the training phase, the model was evaluated using the GTOT and RGBT234 datasets to assess its tracking accuracy and robustness under different environmental conditions. The learning rate was adjusted logarithmically throughout the training process to facilitate rapid and stable convergence, thus eliminating the need for continuous or dynamic tuning and reducing computational overhead compared to conventional learning rate adjustment techniques.

To enhance the generalization capabilities of the model, several data augmentation techniques were employed during training, including translation, scaling, blurring, flipping, and color perturbation applied to both the template and search regions. These augmentation strategies were essential for mitigating overfitting and improving model robustness across diverse scenarios. Moreover, a fixed random seed was used during training to ensure reproducibility of the experimental results, allowing for consistent evaluation and comparison.

4.3. Comparison with Other Methods

4.3.1. Comparison of Results on GTOT

To evaluate the overall performance, the algorithms were deployed on an edge device, the NVIDIA Jetson Nano, and compared with state-of-the-art trackers and lightweight networks using Siamese networks, as shown in Table 1. In this paper, the unaccelerated DuSiamIE and the TensorRT-accelerated DuSiamIE are compared with five other methods, including SiamRPN++ MobileNet [42], SiamCSR [7], DFAT [27], APFNet [28], and HMFT [26]. As shown in the table, DuSiamIE's PR reaches 83.4%, SR reaches 66.8% and FPS reaches 18.4. However, the real-time performance of the most precise algorithms, HMFT and APFNet, for inference on the NVIDIA Jetson Nano is currently limited to 0.02 FPS, respectively. These speeds are insufficient to fulfill the real-time demands. After optimizing the network layers and floating-point precision using NVIDIA TensorRT, the method described in this paper achieves a real-time running speed of 40.3 FPS. This speed is approximately 2000 times faster than the highest-accuracy algorithm, indicating strong performance and fully satisfying real-time requirements [43]. A more intuitive comparison of accuracy and reasoning tracking speed is shown in the bubble chart Figure 6.

The following section presents a visual comparison of six algorithms, focusing on the performance of DuSiamIE in relation to five other trackers across three sequences. Figure 7 illustrates that DuSiamIE precisely monitors the location of the target. Figure 7c shows a sequence that includes all problematic characteristics from the dataset. The scene depicts a fast-moving car in low illumination, partially occluded by dense foliage, overtaking

another vehicle with similar thermal characteristics. The target's dimensions undergo substantial changes relative to the initial tracking frame. Despite these challenges, the proposed method effectively tracks the target. However, as seen in Figure 7b, the method described in this work repositions to the tracked target more slowly when faced with occlusion compared to existing state-of-the-art techniques. This limitation arises from the insufficient exploitation of deep features and the absence of an online template update technique in our algorithm.

Table 1. Comparison of PR/SR scores (%) and FPS with other trackers on the GTOT dataset.

	SiamRPNppMobile	SiamCSR	DFAT	APFNet	HMFT	DuSiamIE	DuSiamIE + TensorRT
PR	78.3	89.0	89.3	90.5	91.2	83.4	83.4
SR	65.8	70.7	72.3	73.7	74.9	66.8	66.8
FPS	5.3	4.8	0.69	0.02	0.02	18.4	40.3

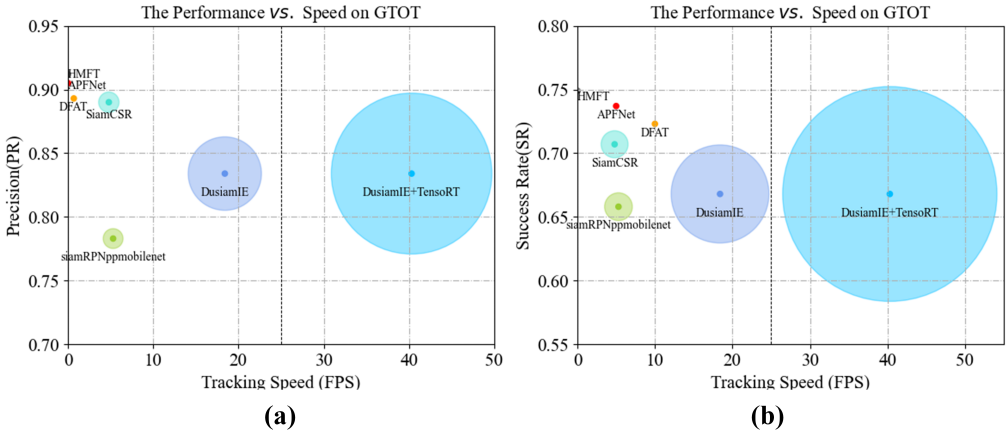


Figure 6. Comparison of speeds of various tracking methods. (**a**) PR and speed based on GTOT. (**b**) SR and speed based on GTOT.

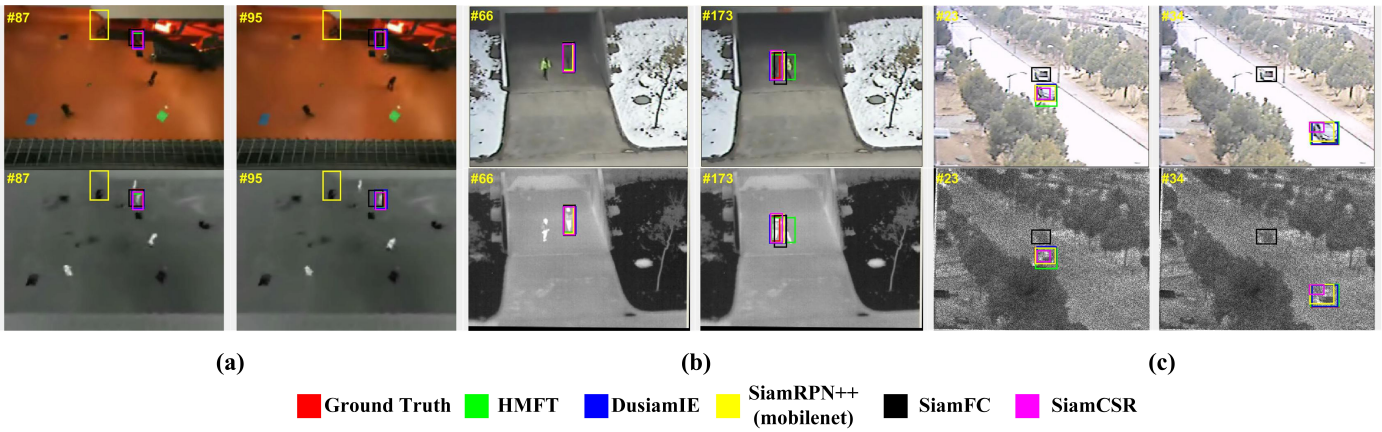


Figure 7. Visual comparison of the tracker proposed in this article with other trackers on three video sequences of GTOT: (**a**) Torabi1, (**b**) Quarrying, and (**c**) Fastcar2.

4.3.2. Comparison of Results on RGBT234

This paper compares the proposed tracker with five existing trackers across 12 RGBT234 attributes: no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossing (TC), distortion (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera movement (CM), and background clutter (BC). The results are summarized in Table 2, where the top three trackers for each

challenge are highlighted in red, blue, and green, respectively. A dash ('-') indicates the absence of available data or code, preventing the reproduction of the corresponding results. The proposed tracker achieves PR and SR of 75.3% and 52.6%, respectively, with performance comparable to the state-of-the-art trackers and notable robustness across multiple attribute tests.

Table 2. Comparison of PR/SR scores (%) and FPS with other trackers on the RGBT234 dataset.

	SiamRPNppMobile	SiamFT	DFAT	HMFT	APFNet	DuSiamIE + TensorRT
NO	83.8 /64.2	84.8/62.0	80.9/55.7	90.9/67.4	80.8/54.1	82.2/61.2
PO	73.5/54.3	72.7/50.6	68.4/50.3	85.7/62.1	81.8/58.2	81.7/58.2
HO	59.4/43.4	57.6/40.7	75.4/55.6	66.4/46.9	84.7/59.6	67.5/45.1
LI	59.3/42.4	68.8/47.4	72.9/52.5	83.3/59.1	83.7/58.8	61.4/42.4
LR	66.4/46.5	69.6/46.5	74.1/53.3	76.3/57.1	83.4/58.8	74.1/46.4
TC	70.6/53.0	70.7/50.3	80.2/56.9	72.2/50.4	84.6/64.2	74.9/54.8
DEF	69.5/53.2	69.7/51.8	71.6/52.3	77.6/57.9	86.6/60.7	75.3/55.0
FM	65.3/46.9	62.0/42.2	71.4/52.2	65.9/46.9	79.9/59.1	75.8/51.2
SV	72.7/55.5	71.1/50.5	73.9/53.6	80.0/59.2	85.6/59.4	78.7/56.2
MB	64.5/48.7	59.0/43.3	76.9/53.1	70.6/50.9	89.0/62.7	66.3/47.3
CM	66.4/49.9	63.2/45.7	70.9/49.5	77.9/56.2	85.8/60.9	69.1/48.3
BC	57.8/39.3	60.5/40.3	79.5/57.7	73.8/49.8	84.5/61.2	64.8/41.9
ALL	69.7/51.7	68.8/48.6	75.8/55.2	78.8/56.8	82.7/57.9	75.3/52.6
FPS	5.3	-	0.7	0.32	0.5	34.7

All trackers were tested on the NVIDIA Jetson NANO, revealing a significant advantage for the proposed DuSiamIE tracker. Despite a marginal decrease in accuracy, DuSiamIE, when deployed with NVIDIA TensorRT, attained a frame rate of 34.7 FPS, which is over 60 times faster than the highest accuracy algorithm, demonstrating superior performance among all evaluated trackers. This substantial speed gain aligns well with the requirements for real-time inference in tracking applications. These findings, combined with earlier evaluations, demonstrate that DuSiamIE offers not only high accuracy but also a lightweight structure, which is uncommon among contemporary trackers, enabling efficient real-time tracking on edge devices.

We compared DuSiamIE against five other state-of-the-art trackers, including ADRNet [44], HMFT, MDNet, DFAT, and APFNet, on three different sequences. As shown in Figure 8, DuSiamIE accurately tracks the target. When the target moves or deforms rapidly, DuSiamIE effectively handles this challenge due to its lightweight network structure. In cases of varying degrees of occlusion, the LIFA module in DuSiamIE effectively manages a wide variety of occlusion situations by extracting and fusing different modal features. However, in low-illumination and non-occlusion environments, the tracking accuracy is slightly reduced due to limitations in the feature extraction network and insufficient information from the fused features. Nevertheless, the introduction of MIFE significantly improves tracking performance compared to the standard Siamese network.



Figure 8. Visual comparison of the tracker proposed in this article with other trackers on three video sequences of RGBT234: (**a**) dog11, (**b**) electric bicycle in front car, and (**c**) flower2.

4.4. Ablation Study

An ablation study using the GTOT dataset is used to verify the essential components of DuSiamIE in this work. We employed two reduced versions of DuSiamIE: one without the LIFA and MIFE modules (DuSiamIE) and another with only the MIFE module removed (DuSiamIE + LIFA). The data presented in Table 3 indicate that the baseline model scores 10.5% and 12.4% lower in PR and SR, respectively, compared to DuSiamIE + LIFA. These findings suggest that LIFA effectively utilizes multimodal information to improve tracking accuracy by considering illumination conditions. The full DuSiamIE model (with both LIFA and MIFE) achieves scores that are 2.7% and 3.3% higher in PR and SR, respectively, than DuSiamIE + LIFA. This indicates that MIFE effectively addresses the insufficient utilization of infrared features in modal fusion. This enhancement improves the exploitation

of infrared image data. These results validate the effectiveness of the primary modules of DuSiamIE. Furthermore, we compare the performance of the DuSiamIE algorithm before and after NVIDIA TensorRT optimization, demonstrating significant speed enhancements in inference without a loss in accuracy. NVIDIA TensorRT optimizes the network architecture and employs precision calibration techniques, which reduce computational overhead and improve real-time performance on edge devices.

Table 3. Ablation studies.

DuSiamIE	LIFA	MIFE	TensorRT	PR(%)	SR(%)	FPS
✓				70.2	51.1	
✓	✓			80.7	63.5	
✓	✓	✓		83.4	66.8	18.3
✓	✓	✓	✓	83.4	66.8	40.3

4.5. Discussions

The results demonstrate that LIFA greatly enhances the accuracy of RGBT tracking by using illumination information for modal fusion. This enables a more precise selection of fusion feature proportions and focuses on a greater number of image features. The MIFE module improves the identification of infrared characteristics that may not be readily detected during the modal fusion procedure. The MIFE module further augments the representation of infrared features by incorporating additional feature information. The use of a lightweight network architecture and NVIDIA TensorRT optimization significantly improves model performance. Within edge device testing, the real-time execution speed significantly exceeds that of other algorithms. The configuration described here fully meets the criteria for robust real-time performance and convenient implementation.

5. Conclusions

This paper presents DuSiamIE, a lightweight Siamese dual-stream network with infrared feature enhancement specifically designed for edge devices. By utilizing the LIFA and MIFE modules, DuSiamIE achieves state-of-the-art performance in real-time tracking under various challenging conditions, such as low illumination and occlusion. Experimental evaluations on the GTOT and RGBT234 datasets demonstrate that DuSiamIE ranks among the top performers in terms of PR and SR scores. Furthermore, DuSiamIE achieves real-time tracking speeds of 40.3 FPS and 34.7 FPS on the NVIDIA Jetson Nano, significantly outperforming other state-of-the-art trackers in terms of speed. Ablation studies validate the effectiveness of the LIFA and MIFE modules, while optimizations using NVIDIA TensorRT significantly enhance inference speed without compromising accuracy. In conclusion, DuSiamIE offers a robust and efficient tracking solution suitable for deployment on resource-constrained platforms.

Author Contributions: Conceptualization, H.W. and J.L. (Jiao Li); methodology, H.W.; software, H.W. and Y.G.; validation, H.W. and J.L. (Jiao Li); formal analysis, H.W. and J.L. (Junyu Lu); investigation, X.S.; resources, J.L. (Junyu Lu); data curation, H.W.; writing—original draft preparation, H.W.; writing—review and editing, J.L. (Jiao Li) and X.S.; visualization, H.W.; supervision, J.L. (Jiao Li) and X.S.; project administration, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be made available upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. You, S.; Zhu, H.; Li, M.; Li, Y. A review of visual trackers and analysis of its application to mobile robot. *arXiv* **2019**, arXiv:1910.09761.
2. Wan, M.; Gu, G.; Qian, W.; Ren, K.; Maldague, X.; Chen, Q. Unmanned aerial vehicle video-based target tracking algorithm using sparse representation. *IEEE Internet Things J.* **2019**, *6*, 9689–9706. [CrossRef]
3. Sun, Q.; Wang, Y.; Yang, Y.; Xu, P. Research on target tracking problem of fixed scene video surveillance based on unlabeled data. In Proceedings of the 2021 3rd World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, 18–20 June 2021; pp. 29–33.
4. Zhang, T.; Liu, X.; Zhang, Q.; Han, J. SiamCDA: Complementarity-and distractor-aware RGB-T tracking based on Siamese network. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1403–1417. [CrossRef]
5. Guo, C.; Yang, D.; Li, C.; Song, P. Dual Siamese network for RGBT tracking via fusing predicted position maps. *Vis. Comput.* **2022**, *38*, 2555–2567. [CrossRef]
6. Bayoudh, K.; Knani, R.; Hamdaoui, F.; Mtibaa, A. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.* **2022**, *38*, 2939–2970. [CrossRef] [PubMed]
7. Guo, C.; Xiao, L. High speed and robust RGB-thermal tracking via dual attentive stream siamese network. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 803–806.
8. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Hift, Y.L. Hierarchical feature transformer for aerial tracking. In Proceedings of the CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15457–15466.
9. Zhang, W. A robust lateral tracking control strategy for autonomous driving vehicles. *Mech. Syst. Signal Process.* **2021**, *150*, 107238. [CrossRef]
10. Shao, J.; Du, B.; Wu, C.; Zhang, L. Tracking objects from satellite videos: A velocity feature based correlation filter. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7860–7871. [CrossRef]
11. Shao, J.; Du, B.; Wu, C.; Zhang, L. Can we track targets from space? A hybrid kernel correlation filter tracker for satellite video. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8719–8731. [CrossRef]
12. Deng, X.; Li, J.; Guan, P.; Zhang, L. Energy-efficient UAV-aided target tracking systems based on edge computing. *IEEE Internet Things J.* **2021**, *9*, 2207–2214. [CrossRef]
13. Sun, C.; Wang, X.; Liu, Z.; Wan, Y.; Zhang, L.; Zhong, Y. Siamohot: A lightweight dual siamese network for onboard hyperspectral object tracking via joint spatial-spectral knowledge distillation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5521112. [CrossRef]
14. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
15. Kuai, Y.; Wen, G.; Li, D.; Xiao, J. Target-aware correlation filter tracking in RGBD videos. *IEEE Sens. J.* **2019**, *19*, 9522–9531. [CrossRef]
16. Zheng, Y.; Liu, X.; Cheng, X.; Zhang, K.; Wu, Y.; Chen, S. Multi-task deep dual correlation filters for visual tracking. *IEEE Trans. Image Process.* **2020**, *29*, 9614–9626. [CrossRef] [PubMed]
17. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
18. Guo, Q.; Feng, W.; Zhou, C.; Huang, R.; Wan, L.; Wang, S. Learning dynamic siamese network for visual object tracking. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1763–1771.
19. Nam, H.; Han, B. Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4293–4302.
20. Li, H.; Li, Y.; Porikli, F. Deeptrack: Learning discriminative feature representations online for robust visual tracking. *IEEE Trans. Image Process.* **2015**, *25*, 1834–1848. [CrossRef] [PubMed]
21. Peng, J.; Zhao, H.; Hu, Z.; Zhuang, Y.; Wang, B. Siamese infrared and visible light fusion network for RGB-T tracking. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 3281–3293. [CrossRef]
22. Zhang, X.; Ye, P.; Peng, S.; Liu, J.; Gong, K.; Xiao, G. SiamFT: An RGB-infrared fusion tracking method via fully convolutional Siamese networks. *IEEE Access* **2019**, *7*, 122122–122133. [CrossRef]
23. Zhang, X.; Ye, P.; Peng, S.; Liu, J.; Xiao, G. DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion. *Signal Process. Image Commun.* **2020**, *84*, 115756. [CrossRef]
24. Xue, Y.; Zhang, J.; Lin, Z.; Li, C.; Huo, B.; Zhang, Y. SiamCAF: Complementary attention fusion-based Siamese network for RGBT tracking. *Remote Sens.* **2023**, *15*, 3252. [CrossRef]
25. Long Li, C.; Lu, A.; Hua Zheng, A.; Tu, Z.; Tang, J. Multi-adapter RGBT tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
26. Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; Ruan, X. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8886–8895.

27. Tang, Z.; Xu, T.; Li, H.; Wu, X.J.; Zhu, X.; Kittler, J. Exploring fusion strategies for accurate RGBT visual object tracking. *Inf. Fusion* **2023**, *99*, 101881. [CrossRef]
28. Xiao, Y.; Yang, M.; Li, C.; Liu, L.; Tang, J. Attribute-based progressive fusion network for rgbt tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 2831–2838.
29. Zhang, T.; Guo, H.; Jiao, Q.; Zhang, Q.; Han, J. Efficient rgb-t tracking via cross-modality distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5404–5413.
30. Yu, Z.; Fan, H.; Wang, Q.; Li, Z.; Tang, Y. Region selective fusion network for robust rgb-t tracking. *IEEE Signal Process. Lett.* **2023**, *30*, 1357–1361. [CrossRef]
31. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese box adaptive network for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6668–6677.
32. Feng, Z.; Yan, L.; Xia, Y.; Xiao, B. An adaptive padding correlation filter with group feature fusion for robust visual tracking. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1845–1860. [CrossRef]
33. Süzen, A.A.; Duman, B.; Şen, B. Benchmark analysis of jetson tx2, jetson nano and raspberry pi using deep-cnn. In Proceedings of the 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 26–28 June 2020; pp. 1–5.
34. Liu, L.; Blancaflor, E.B.; Abisado, M. A lightweight multi-person pose estimation scheme based on Jetson Nano. *Appl. Comput. Sci.* **2023**, *19*, 1–14. [CrossRef]
35. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. Mixed precision training. *arXiv* **2017**, arXiv:1710.03740.
36. Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; Lin, L. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Trans. Image Process.* **2016**, *25*, 5743–5756. [CrossRef]
37. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [CrossRef]
38. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
39. Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; Ghanem, B. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 300–317.
40. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5374–5383.
41. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
42. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
43. Madeo, S.; Pelliccia, R.; Salvadori, C.; del Rincon, J.M.; Nebel, J.C. An optimized stereo vision implementation for embedded systems: Application to RGB and infra-red images. *J.-Real-Time Image Process.* **2016**, *12*, 725–746. [CrossRef]
44. Zhang, P.; Wang, D.; Lu, H.; Yang, X. Learning adaptive attribute-driven representation for real-time RGB-T tracking. *Int. J. Comput. Vis.* **2021**, *129*, 2714–2729. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.