Optimizing Object Detection with Multispectral RGB/IR Fusion

Sofia Maria Palomino Chamizo, Daeyoung Kim
School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
{sofia, kimd}@kaist.ac.kr

***Abstract*—This paper addresses the challenge of effectively fusing visible and infrared images for robust object detection under varying lighting conditions. Due to their distinct characteristics, existing methods often struggle to integrate the complementary information from these modalities fully. To overcome these limitations, this work introduces the Multi-Scale Attention Fusion Transformer (MAFT), which improves feature integration at the mid-level feature stage (P4/16). Unlike previous transformer-based approaches that apply uniform fusion across multiple levels, MAFT focuses on P4/16, incorporating multi-scale attention mechanisms to refine feature extraction and strengthen RGB-IR fusion. This approach results in more precise object detection, particularly for small and low-contrast targets. Evaluations on the VEDAI, FLIR-Aligned, and KAIST datasets show that MAFT achieves state-of-the-art mean Average Precision (mAP) performance while maintaining competitive inference speed, making it suitable for real-time applications.**

***Index Terms*—Multispectral fusion, object detection, RGB, Infrared, Deep Learning, Transformers, Multi-Scale attention**

I. Introduction

Multispectral object detection, particularly the fusion of visible (RGB) and infrared (IR) images has gained increasing attention due to its robustness in challenging lighting conditions. Applications such as autonomous driving, surveillance, and aerial reconnaissance benefit from combining the complementary information from both modalities. RGB images offer high spatial detail but are highly sensitive to lighting conditions. At the same time, IR captures thermal variations, making objects visible in low-light environments. However, effectively integrating these modalities remains a challenge due to differences in spatial resolution, noise characteristics, and the need for efficient real-time processing.

Fusion approaches can be classified into early, middle, and late fusion [1]. Early fusion combines raw data from both modalities at the input level, potentially preserving rich information but introducing difficulties in handling varying spatial resolutions and noise characteristics. Late fusion aggregates features at higher network stages, after modality-specific feature extraction, and may lose some synergistic benefits of the complementary information. In contrast, middle fusion, which combines features at intermediate stages of the network, allows for integrating complementary strengths while maintaining the distinct characteristics of each modality. This approach is efficient for multispectral object detection as it ensures that both modalities contribute meaningfully to the final output without overwhelming the network with unprocessed raw data or losing vital information.

Deep learning-based methods have made significant progress in multispectral fusion. CNN-based models such as DenseFuse [2] and INSANet [3] integrate RGB and IR features but often lack global contextual understanding. More recent transformer-based architectures, such as the Cross-Modality Fusion Transformer (CFT) [4], enhance feature interaction by utilizing self-attention mechanisms and hierarchical fusion at P3/8, P4/16, and P5/32. However, the effectiveness of fusion at each level varies, and current methods do not fully exploit the most informative stages.

This paper introduces a Multi-Scale Attention Fusion Transformer (MAFT), specifically designed to improve feature integration at P4/16, where crucial mid-level features are extracted. Unlike CFT, which applies uniform fusion across multiple levels, our approach refines feature extraction at P4/16 by incorporating multi-scale attention mechanisms, strengthening the fusion of RGB and IR data. This modification leads to more precise object detection, particularly for small and low-contrast targets, which are often missed in conventional fusion methods.

The key contributions of this work are as follows:

- Improved Fusion at P4/16: We introduce MAFT to refine feature extraction and fusion at the mid-level feature stage.
- State-of-the-Art Performance on Multispectral Datasets: Our model is evaluated on the VEDAI [5], FLIR-Aligned [6], [7], and KAIST [8] datasets, showing improved accuracy over existing approaches, especially for detecting small objects.
- Efficient Computation for Real-Time Applications: The method maintains high detection accuracy while keeping inference time low, making it suitable for military, automotive, and surveillance applications.

II. Related Work

Multispectral object detection has undergone significant advancements with the development of deep learning. Traditional approaches relied on handcrafted fusion techniques such as Simple Averaging and Weighted Summation [1],

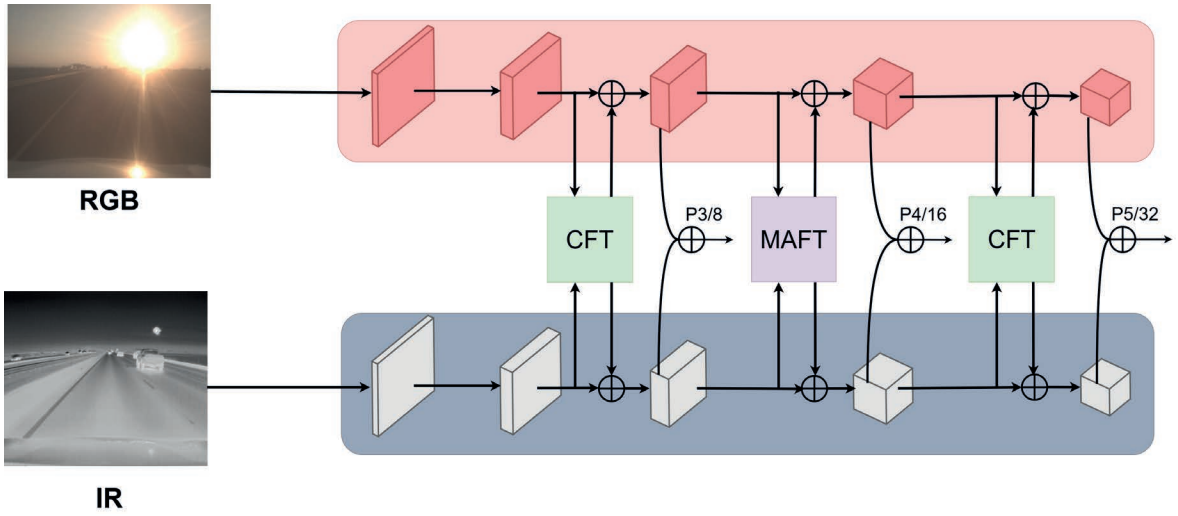

Fig. 1. The architecture of the proposed method. RGB and infrared features are independently extracted through convolutional layers and fused at P3/8, P4/16, and P5/32 using Multi-Scale Attention and Cross-Modality Fusion Transformers.

which often failed to exploit the complementary nature of RGB and IR data fully. More recent methods integrate deep learning-based fusion strategies, which can be broadly categorized into CNN-based fusion and transformer-based fusion.

A. CNN-Based Fusion Approaches

Early deep learning models primarily relied on convolutional neural networks (CNNs) to extract and merge modality-specific features. DenseFuse [2] introduced an encoder-decoder architecture that combined RGB and IR data through element-wise addition or L1-norm fusion. Another approach, Halfway Fusion [3], proposed an intermediate-stage feature fusion mechanism within a CNN-based object detection model, allowing the network to retain modality-specific information before merging them into a unified representation. While these methods improved multispectral fusion, CNN-based architectures struggle with long-range dependencies and fail to capture the global interactions between RGB and IR features fully.

B. Transformer-Based Fusion Approaches

With the success of transformers in vision tasks, recent multispectral detection models incorporate self-attention mechanisms to capture global dependencies. The Multi-Modal Feature Pyramid Transformer [9] improves RGB-IR fusion by applying cross-modal attention at multiple feature scales. CAFF-DINO [10] strengthens multispectral object detection by introducing cross-attention modules, which facilitate better modality interaction in challenging conditions. The Cross-Modality Fusion Transformer (CFT) [4] has emerged as a state-of-the-art approach, combining RGB and IR features at P3/8, P4/16, and P5/32 using transformer-based fusion.

Despite these advancements, existing transformer-based fusion methods typically apply uniform self-attention mechanisms without explicitly adapting to variations in object size and modality characteristics.

III. Methodology

This work addresses the limitations of existing fusion methods by introducing the Multi-Scale Attention Fusion Transformer, which enhances feature extraction and integration through the use of both spatial and channel attention mechanisms embedded within the transformer block. MAFT specifically enhances feature fusion at the P4/16 level, enabling more accurate detection of small and low-contrast objects while maintaining computational efficiency.

A. Overall Architecture

The proposed model follows a two-stream feature extraction pipeline, where RGB and IR images are processed independently through a YOLOv5 [10] backbone before fusion. This backbone extracts hierarchical features at different scales, ensuring each modality retains its unique characteristics while enabling effective integration. Features are extracted and fused at three key stages of the feature hierarchy:

- P3/8: Captures fine-grained spatial details helpful in detecting large objects.
- P4/16: A mid-level stage where features contain local structure and contextual information.
- P5/32: Captures high-level semantic information needed for robust object classification.

At each stage, transformer-based fusion mechanisms integrate the extracted features from both modalities. Figure 1 illustrates the two-stream architecture, highlighting the distinction between CFT and MAFT, where the latter refines feature extraction through multi-scale attention.

Unlike prior methods that apply uniform self-attention across all feature extraction stages, MAFT introduces a multi-scale attention mechanism at the P4/16 stage. Applying this mechanism at every stage would significantly increase

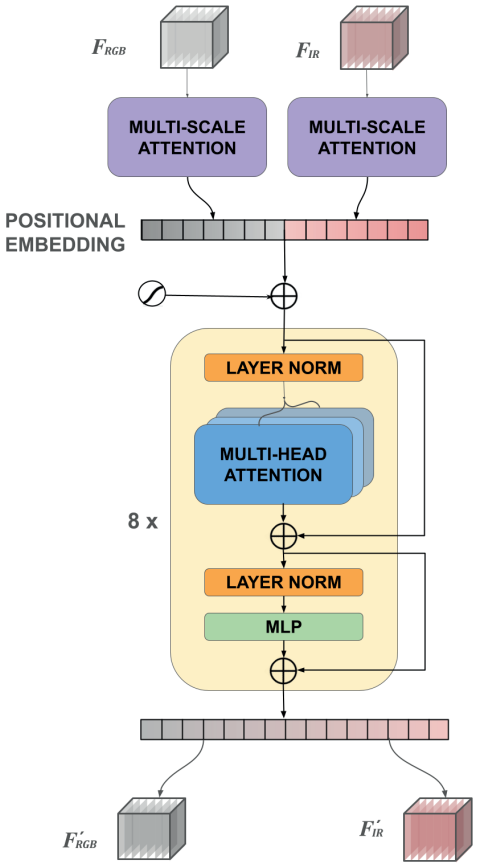

Fig. 2. The architecture of the proposed transformer. The key components include separate multi-scale attention modules applied to the RGB and IR input features, as well as standard transformer blocks that process the fused representation. The multi-scale attention enables the model to capture relevant features at different spatial resolutions, thereby improving its ability to combine information from the two modalities effectively.

computational complexity, making real-time applications impractical. Instead, P4/16 serves as an optimal point for enhancement, as it captures both low-level spatial details from P3/8 and high-level semantic features from P5/32. This stage effectively acts as a bridge, ensuring a smooth transition across feature scales, which is crucial for detecting objects of varying sizes and enhancing modality integration.

B. Multi-Scale Attention Mechanism

To improve feature representation before fusion, we introduce a dual attention mechanism that combines spatial and channel attention.

1) Channel Attention: Given an input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, we compute channel-wise importance using a squeeze-and-excitation structure. The global descriptor is obtained via global average pooling:

$$\mathbf{z} = \text{GAP}(\mathbf{F}) \in \mathbb{R}^C \tag{1}$$

This descriptor passes through two fully connected layers with a nonlinearity to yield channel attention weights:

$$\mathbf{z}_c = \sigma\left(W_2 \cdot \delta(W_1 \cdot \mathbf{z})\right) \tag{2}$$

where W_1, W_2 are learnable weights, $\delta(\cdot)$ denotes ReLU, and $\sigma(\cdot)$ denotes the sigmoid activation. These weights are applied to the original feature map to modulate channel importance.

2) Spatial Attention: Spatial attention identifies salient regions within a feature map using average and max pooling, followed by convolution:

$$\mathbf{z}_s = \sigma\left(f^{7\times7}\left(\text{AvgPool}(\mathbf{F}) \parallel \text{MaxPool}(\mathbf{F})\right)\right) \tag{3}$$

where $\parallel$ denotes channel-wise concatenation and $f^{7\times7}(\cdot)$ is a convolution with a 7×7 kernel.

The refined feature map after applying both attentions becomes:

$$\mathbf{F}' = \mathbf{F} \cdot \mathbf{z}_c \cdot \mathbf{z}_s \tag{4}$$

This dual attention mechanism ensures that the model focuses on the most informative channels and regions in both RGB and IR streams.

C. Transformer Fusion at P4/16

The core innovation in MAFT lies in its transformer-based fusion block, applied selectively at the P4/16 feature level. This stage balances low-level spatial details and high-level semantics, making it a prime candidate for enhanced fusion.

After applying the multi-scale attention mechanism to the modality-specific feature maps $\mathbf{F}_{\text{RGB}}$ and $\mathbf{F}_{\text{IR}}$, we flatten and concatenate them to form a joint sequence representation:

$$\mathbf{I} = \text{Concat}(\text{Flatten}(\mathbf{F}_{\text{RGB}}), \text{Flatten}(\mathbf{F}_{\text{IR}})) \in \mathbb{R}^{2HW\times C} \tag{5}$$

A learnable positional embedding is added to $\mathbf{I}$ to retain spatial context, and the sequence is passed through a transformer block comprising multi-head self-attention and a feed-forward network. The attention mechanism models both intra- and inter-modality dependencies.

Instead of using handcrafted fusion rules, we adopt self-attention to compute interactions:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{6}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are learned projections of $\mathbf{I}$, and d_k is a scaling factor.

The output of the transformer block is reshaped and merged back with the original feature streams using residual cross-modality fusion. This helps retain modality-specific information while reinforcing shared representations:

$$\mathbf{F}_{\text{fused}} = \mathbf{F}_{\text{RGB}} + \text{CrossAtt}_{\text{RGB}\to\text{IR}} + \text{CrossAtt}_{\text{IR}\to\text{RGB}} \tag{7}$$

By limiting this enhanced fusion to P4/16, MAFT achieves improved accuracy with minimal impact on inference speed, striking a practical balance for real-time applications.

IV. Evaluation

To assess the effectiveness of the proposed method, we evaluate its performance using mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds. The primary evaluation metrics include mAP@50, mAP@75, and mAP@50:95, which measure detection accuracy at varying levels of localization precision.

The Average Precision (AP) for a given class is defined as:

$$\text{AP} = \int_0^1 P(r)\, dr \tag{8}$$

where $P(r)$ is the precision-recall curve. The mean Average Precision (mAP) across all classes is then given by:

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^{n} \text{AP}_i \tag{9}$$

Where N is the total number of classes.

For a more comprehensive evaluation, we compute mAP@50, which considers predictions correct if the IoU is at least 0.50, and mAP@75, which applies a stricter threshold by requiring an IoU of at least 0.75. Additionally, mAP@50:95 averages AP across IoU thresholds ranging from 0.50 to 0.95 in increments of 0.05, providing a more detailed performance measure.

A. Experimental Setup

The experiments are conducted on the VEDAI [4], FLIR-Aligned [5], and KAIST [6] datasets, each offering different challenges in multispectral object detection. VEDAI consists of aerial images containing small vehicles captured in both RGB and infrared. FLIR-Aligned provides thermal and visible imagery for detecting pedestrians and vehicles in real-world driving scenarios. KAIST is a multispectral pedestrian detection dataset that includes both day and night scenes, making it particularly useful for evaluating robustness in low-light conditions. All datasets are preprocessed to ensure alignment between RGB and IR modalities, maintaining consistent image resolution and aspect ratios. Standard data augmentation techniques, including random cropping, horizontal flipping, and brightness adjustments, are applied to enhance generalization.

The model is trained using stochastic gradient descent (SGD) with a learning rate of 1×10^{-2}. The batch size is set to 16, and training is conducted for 600 epochs on an NVIDIA GeForce RTX 4090 Ti GPU. The number of epochs was determined empirically, ensuring that training converges optimally without overshooting.

B. Experimental Results

Table I presents the quantitative results comparing MAFT with baseline models, including CFT and standard YOLO-based approaches. The results demonstrate that MAFT consistently outperforms previous methods across all datasets, achieving the highest mAP@50, mAP@75, and mAP@50:95 in most cases.

TABLE I
Comparison of mAP results on VEDAI, FLIR-Aligned, and KAIST datasets. Best results are marked in bold, second-best results are underlined.

Dataset	Method	mAP@50	mAP@75	mAP
VEDAI	YOLOv5 (RGB)	0.535	0.287	0.299
	YOLOv5 (Thermal)	0.444	0.225	0.243
	CFT (RGB+T)	<u>0.610</u>	<u>0.313</u>	<u>0.331</u>
	MAFT (RGB+T)	**0.719**	**0.484**	**0.437**
FLIR	YOLOv5 (RGB)	0.927	0.706	0.642
	YOLOv5 (Thermal)	**0.954**	0.728	**0.695**
	CFT (RGB+T)	0.949	<u>0.755</u>	0.679
	MAFT (RGB+T)	<u>0.950</u>	**0.762**	<u>0.679</u>
KAIST	YOLOv5 (RGB)	0.973	0.931	0.776
	YOLOv5 (Thermal)	0.954	0.906	0.695
	CFT (RGB+T)	<u>0.974</u>	<u>0.934</u>	<u>0.814</u>
	MAFT (RGB+T)	**0.975**	**0.935**	**0.815**

On the VEDAI dataset, which contains small objects in aerial images, MAFT achieves a mAP@50 of 0.719, significantly outperforming CFT (0.610) and both single-modality YOLO baselines. The improvement is even more pronounced in mAP@75, where MAFT achieves 0.484, compared to 0.313 for CFT. This highlights the effectiveness of multi-scale attention at P4/16, which enhances small object detection by improving feature extraction and modality fusion.

For the FLIR-Aligned dataset, MAFT achieves the best mAP@75 (0.762), surpassing both CFT (0.755) and the single-modality thermal YOLO baseline. However, mAP@50, the thermal-only YOLO model, reaches 0.954, slightly outperforming MAFT (0.950). This suggests that single-modality models may retain an advantage in high-contrast thermal imagery in certain conditions. However, MAFT still provides a more balanced fusion strategy across multiple IoU thresholds.

On the KAIST dataset, which includes pedestrian detection in both day and night conditions, MAFT achieves the highest mAP scores across all metrics, with a mAP@50 of 0.975 and a mAP@75 of 0.935, slightly improving over CFT (0.974 and 0.934, respectively). The relatively small difference suggests that both models perform strongly on this dataset, but the additional refinement from multi-scale attention at P4/16 provides a slight edge.

In addition to accuracy improvements, MAFT maintains competitive computational efficiency, as shown in Table **??**. The inference speed results indicate that CFT remains slightly faster across all datasets, with MAFT introducing a small computational overhead due to the added attention mechanisms. For instance, in VEDAI, MAFT's inference time is 11.0 ms, compared to 9.2 ms for CFT, resulting in a drop in FPS from 108.7 to 90.9. A similar pattern is observed in FLIR and KAIST, where MAFT trades off a modest decrease in speed for improved detection accuracy.

Despite this, the model remains within an acceptable range for real-time applications, demonstrating that the selective use

of multi-scale attention at P4/16 effectively enhances accuracy without significantly compromising computational efficiency.

TABLE II
INFERENCE SPEED COMPARISON ACROSS DIFFERENT METHODS

Dataset	Method	Time (ms)	FPS
VEDAI	CFT (RGB+T)	**9.2**	**108.7**
	MAFT (RGB+T)	11.0	90.9
FLIR	CFT (RGB+T)	**5.1**	**196.1**
	MAFT (RGB+T)	5.8	172.4
KAIST	CFT (RGB+T)	**4.3**	**232.6**
	MAFT (RGB+T)	5.3	188.7

V. CONCLUSION

In conclusion, this paper has tackled the critical challenge of effectively integrating RGB and IR images for robust object detection. To this end, we introduced the MAFT, a transformer architecture specifically designed to strenghen feature extraction and fusion at the crucial mid-level feature stage, P4/16. Unlike prior transformer-based methods that apply uniform fusion across all feature scales, MAFT incorporates multi-scale attention mechanisms at P4/16 to refine both spatial and channel interactions between RGB and IR features before fusion. The experimental evaluation of MAFT on the VEDAI, FLIR-Aligned, and KAIST datasets has demonstrated state-of-the-art performance across key evaluation metrics, including mAP@50, mAP@75, and mAP@50:95. Notably, MAFT showed significant improvements in detecting small and low-contrast objects, as evidenced by the substantial gains on the VEDAI dataset. While achieving these accuracy improvements, MAFT maintains competitive computational efficiency, demonstrating its potential for real-time applications in domains such as autonomous driving, surveillance, and aerial reconnaissance. Future work could explore extending the multi-scale attention mechanism or investigating its application to other fusion stages to enhance performance and efficiency further.

ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (RS-2025-02222776) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).

REFERENCES

[1] F. Farahnakian and J. Heikkonen, "Deep learning based multi-modal fusion architectures for maritime vessel detection," *Remote Sensing*, vol. 12, no. 2509, 2020.

[2] H. Li and X. J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, 2019, doi: 10.1109/TIP.2018.2887342.

[3] S. Lee, T. Kim, J. Shin, N. Kim, and Y. Choi, "INSANet: INtra-INter spectral attention network for effective feature fusion of multispectral pedestrian detection," *Sensors*, vol. 24, no. 4, p. 1168, 2024.

[4] Q. Fu, D. Hu, and Z. Wang, "Cross-Modality Fusion Transformer for multispectral object detection," *arXiv:2111.00273*, 2022.

[5] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, 2015.

[6] FLIR Systems, "Free FLIR thermal dataset for algorithm training," Teledyne FLIR. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[7] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, UAE, Oct. 2020, pp. 276–280.

[8] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1037–1045.

[9] Y. Zhu, X. Sun, M. Wang, and H. Huang, "Multi-modal feature pyramid transformer for RGB-infrared object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9984–9995, 2023, doi: 10.1109/TITS.2023.3266487.

[10] K. Helvig, B. Abeloos, and P. Trouvé-Peloux, "CAFF-DINO: Multispectral object detection transformers with cross-attention features fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 3037–3046.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv:1506.02640*, 2016.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.