A Comparative Analysis of IR-VIS Image Fusion Methods: Object Detection with YOLO Architectures and Fusion Quality Evaluation

Büşra ÇELİK[1,2], İbrahim Halil GİDEN[1,2], Kubilay Ayturan[2], Fırat HARDALAÇ[2]
[1]*ASELSAN Inc.*
[2]*Department of Electrical Electronics Engineering, Gazi University*
e-mail: bcolak@aselsan.com, busra.colak1@gazi.edu.tr, ihgiden@aselsan.com, kubilay.ayturan@gazi.edu.tr, firat@gazi.edu.tr
Ankara, Türkiye

2025 Fourteenth International Conference on Image Processing, Theory, Tools & Applications (IPTA) | 978-1-6654-5739-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/IPTA66025.2025.11222000

***Abstract*— Image fusion is a technique that integrates information from multiple sources into a single image to produce richer and more meaningful visual content. This study applied three classical fusion methods—Multi-Scale Transform (MST), Saliency-Based fusion, and Anisotropic Diffusion with KL Transform—to RGB and thermal image pairs from the LLVIP dataset. CLAHE preprocessing was applied to enhance contrast, and the resulting fusion images were evaluated in terms of object detection performance and image quality metrics. For object detection, YOLOv4, YOLOv4-Tiny, and YOLOv7 models were used. According to YOLOv4 results, MST fusion applied to CLAHE-preprocessed RGB images achieved 0.18% higher mAP@0.5 accuracy compared to thermal images alone and 9.53% higher accuracy compared to RGB images alone. Regarding the authors' knowledge, this is the first time the study presents the comparative analysis of these models on the LLVIP dataset. The fusion outputs were also evaluated using SSIM, Mutual Information, and $Q_{AB/F}$ metrics. The results indicate that the Saliency-Based fusion method yielded the best performance, particularly regarding visual quality.**

Keywords—IR-VIS image fusion, object detection, LLVIP dataset, low-light conditions, fusion quality metrics, YOLOv4, YOLOv7, CLAHE.

I. Introduction

With the advancement of imaging technologies, combining images obtained from different spectral bands has enabled more reliable analyses, particularly under challenging environmental conditions. In scenarios where RGB cameras are insufficient—such as in dark, low-contrast, or shadowy environments—thermal images offer advantages due to their sensitivity to the heat emitted by objects. Conversely, RGB images provide greater situational awareness by clearly capturing environmental details, colors, and sharp edges. The complementary nature of these two modalities enriches the information content of the scene, and combining thermal and visible band data significantly improves object detection performance.

Image fusion, combining data from multiple sensors into a single image, offers more comprehensive information than any individual source. Especially in low-light and shadow-dominant conditions, neither visible nor thermal images alone provide sufficient visual detail, which, in turn, necessitates the use of fusion techniques. Thermal images highlight temperature variations, while RGB images contribute to high-frequency details such as edges and color. Merging these complementary characteristics improves object detection accuracy and overall image quality, enhancing situational awareness for infrared search and track (IRST) systems. Prior studies have demonstrated that the fusion of visible and infrared images under low-light and complex scene conditions significantly benefits detection performance and perceptual image quality [1] [2].

In this context, as noted in the literature: "The images acquired by a single visible light sensor are very susceptible to light conditions, weather changes, and other factors, while the images acquired by a single infrared light sensor generally have poor resolution, low contrast, low signal-to-noise ratio, and blurred visual effects." [3]. Hence, combining visible and infrared light mitigates the limitations of individual sensors, merging their strengths to enhance the overall quality of the resulting fused images substantially.

This study applied three image fusion methods—Multi-Scale Transform (MST), Saliency-Based fusion, and KL Transform with Anisotropic Diffusion—to RGB and thermal image pairs from the LLVIP dataset [4]. The effects of contrast enhancement (via CLAHE) and varying contribution ratios of source images were examined about object detection performance and image quality metrics. The proposed methods are evaluated for their potential usability in real-time object detection and tracking, particularly in defense and surveillance applications.

Although various object detection studies using different YOLO architectures have been conducted on the LLVIP dataset image fusion [5], [6], no previous study has been found up to now in which the performances of YOLOv4, YOLOv4-Tiny, and YOLOv7 were directly compared. To the authors' knowledge, this study presents the first comparative analysis of these three models in conjunction with image fusion methods. This study may offer an in-depth assessment of the regarding YOLO algorithms' respective detection capabilities.

II. Related Work

The field of image fusion, particularly the fusion of thermal and visible band images, offers significant advantages in both perceptual image quality and deep learning-based object detection applications. In the literature, image fusion approaches have been developed using traditional and CNN-based deep learning techniques. [7]

Numerous methods have been proposed to integrate information from different spectral bands. Among these, multi-scale transforms (MST), saliency-based approaches, and differential diffusion techniques are prominent.

Bavirisetti and Dhuli [8] proposed a Saliency-Based two-scale image fusion method for visible and infrared images. This approach decomposes the source images into base and

978-1-6654-5739-2/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on November 02,2026 at 09:40:05 UTC from IEEE Xplore. Restrictions apply.

detail layers using a mean filter. Then, saliency maps are generated by calculating the difference between mean and median filtering outputs. These saliency maps are normalized on a per-pixel basis to construct adaptive weight maps, which are used to fuse the detail layers. This method allows for preserving fine details and generating content-aware fusion results that are well-suited to human visual perception.

Na et al. [9] proposed a wavelet transform (WT)-based multi-scale fusion approach. This method decomposes source images into low- and high-frequency components using WT. The low-frequency components are fused using weighted average rules, while the high-frequency components are combined based on absolute coefficient values. For color images, each color channel is processed separately during fusion, resulting in combined images with improved contrast and higher information content. This method has been proven effective in preserving structural details and enhancing contrast when fusing thermal and visible images.

Indira et al. [10] fused PET/CT images by applying discrete wavelet transform (DWT) to decompose images into sub-bands. Low-frequency components were fused using average rules, while high-frequency components were combined using maximum or contrast-based selection. The final fused image was reconstructed using inverse DWT (IDWT).

Another approach published on the MathWorks platform combines anisotropic diffusion with KL transform to fuse thermal and visible images [11]. This method aims to preserve structural and edge details by harmoniously merging the base and detail components while reducing noise to improve fusion quality.

This study focuses on image fusion methods that are suitable for real-time applications with limited computational resources and power consumption. While recent advancements in deep learning have led to powerful AI-based fusion architectures, they often require significant processing power, which can be a major constraint in real-time defense and surveillance systems. For this reason, this work is centered on evaluating classical fusion methods, which offer a more favorable balance between performance and computational efficiency, making them more practical for resource-constrained environments.

The LLVIP dataset used in this study is widely adopted in the image fusion domain due to its aligned thermal-visible image pairs captured under low-light conditions and the availability of labeled pedestrian objects. In some previous studies, only visual quality metrics such as SSIM and MI were evaluated after fusion; in others, object detection was performed using architectures like YOLOv5, YOLOv7, and CNN [12], [13].

In the existing literature, it is seen that classical image fusion methods are primarily evaluated only by visual quality metrics, and limited analysis is performed in terms of object detection performance. In this context, this current study could be considered as one of the first studies that perform comparative performance analysis of YOLOv4, YOLOv4-Tiny, and YOLOv7 architectures on images generated by MST, Saliency, and Anisotropic Diffusion and KL Transform based fusion methods on LLVIP dataset.

III. Methodology

A. Image Fusion Methods

1) MST (Multi-Scale Transform) Fusion: In wavelet-based multi-scale image fusion, the source images are decomposed using wavelet transform. Both thermal and visible images are separated into low- and high-frequency components. The low-frequency components are fused using weighted averaging, with 70% weight assigned to the thermal image and 30% to the RGB image. The high-frequency components are fused by calculating their average. The combined low- and high-frequency components are reconstructed into a fused image using inverse discrete wavelet transform (IDWT). This approach preserves both the low-frequency information from the thermal image (e.g., temperature differences) and the high-frequency information from the RGB image (e.g., edges and fine details). Since the fusion operates on frequency components rather than direct pixel values, it yields more efficient results. This study applies color MST fusion (by processing RGB channels individually) and grayscale MST fusion (by using single-channel inputs).

2) Saliency-Based Fusion: Images are first decomposed into base layers using a mean filter. The detail layers are then obtained by subtracting the base layers from the original images. A saliency map is generated using the difference between mean and median filtering. These saliency maps are normalized individually for each source image. The weight maps are computed based on the normalized saliency values at each pixel, as shown in Equation (1).

$$w_{\text{RGB}} = \frac{0.5 \cdot S_{\text{RGB}}}{0.5 \cdot S_{\text{RGB}} + 0.5 \cdot S_{\text{Thermal}}} \quad ; \quad w_{\text{Thermal}} = \frac{0.5 \cdot S_{\text{Thermal}}}{0.5 \cdot S_{\text{RGB}} + 0.5 \cdot S_{\text{Thermal}}} \tag{1}$$

The detail layers are then fused using these weight maps. Base layers are combined using simple pixel-wise averaging. The final fused image is obtained as shown in Equation (2).

$$I_{\text{fused}} = \text{Base}_{\text{fused}} + \text{Detail}_{\text{fused}} \tag{2}$$

3) Anisotropic Diffusion and KL Transform Fusion: This method applies anisotropic diffusion to suppress noise and preserve significant edges in the images. The flowchart of the regarding algorithm is schematically drawn in Fig. 1. The diffusion layers are subtracted from the original images to obtain the detail layers for both thermal and RGB images. Principal Component Analysis (PCA) is performed on the detail layers. The eigenvectors obtained from PCA are used to select the principal component with the highest eigenvalue. This selected PCA component is used to fuse the detail layers. The diffusion layers are combined with equal weights. Finally, the fused image combines the fused detail and diffusion layers. This approach preserves key structural information from both images and reduces noise, resulting in a more refined fusion output [14].

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on November 02,2026 at 09:40:05 UTC from IEEE Xplore. Restrictions apply.

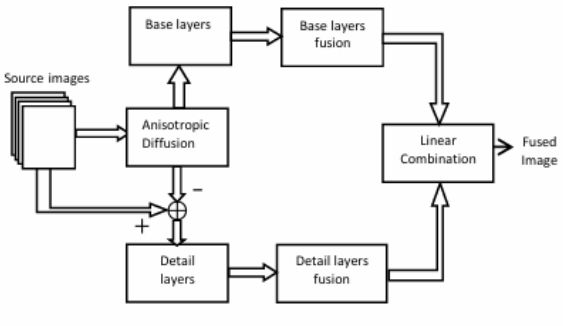


Fig. 1. Anisotropic Diffusion and KL Transform Fusion Method [14]

B. Dataset and Preprocessing

1) LLVIP Dataset: The LLVIP (Low-Light Visible and Infrared Paired) dataset is widely used for image fusion and object detection tasks under low-light conditions. It consists of 15,488 synchronized visible-infrared image pairs (30,976 images) with a resolution of 1280×1024. The dataset provides bounding box annotations for pedestrian objects. Pedestrian annotations are provided for both modalities. The images are carefully aligned in time and space, enabling pixel-level comparison and fusion. LLVIP was designed to evaluate the complementary strengths of visible and infrared imagery in scenarios where RGB cameras struggle due to poor illumination [4].

2) Preprocessing: CLAHE (Contrast Limited Adaptive Histogram Equalization) was applied to the input images prior to the fusion process. The CLAHE parameters were set as follows: Tile Size: 8×8, Clip Limit: 0.01, and Distribution: Rayleigh. The fusion results were evaluated using deep learning-based object detection algorithms and fusion quality metrics. CLAHE has shown a positive impact, especially on visible images captured in low-light conditions.

C. Image Fusion Quality Evaluation Metrics

The following parameters are investigated as evaluation metrics for the quality of the fused images:

1) Structural Similarity Index Measure (SSIM): Measures the structural similarity between the fused image and a reference image on a scale from 0 to 1. It considers luminance, contrast, and structure components to generate a similarity score consistent with human visual perception. A score closer to 1 indicates higher structural similarity.

2) Mutual Information (MI): Quantifies the shared information between the fused image and each source images. Typically, the mutual information values with both source images are summed. A higher MI indicates that a larger portion of the information from both sources is retained in the fused image.

3) $Q_{AB/F}$: This metric quantitatively evaluates how well the edge information from source images A and B is preserved in the fused image F. It measures the similarity of local gradient magnitudes and orientations between the source and fused images using Sobel-based edge extraction. The output score ranges from 0 to 1, with higher values indicating better retention of structural detail in the fusion result [15].

D. YOLO-Based Object Detection and Hardware Settings

Model training was performed on an NVIDIA GeForce GTX 1050 GPU system. The input image size was set to 640×480, and all models were trained for 6000 epochs. In all tests, 4000 images were used for training, 500 for validation, and 500 for testing. After training, object detection performance was compared using thermal only, RGB only, and fused images (Thermal+RGB).

In this study, YOLOv4, YOLOv4-Tiny, and YOLOv7 were selected to represent a trade-off spectrum between inference speed and detection accuracy, as reported in recent benchmarks. YOLOv4 offers a balanced compromise between accuracy and computational cost, achieving high mAP while maintaining reasonable inference time on standard GPUs (mAP@0.5 = 65.7%, ~65 FPS) [16]. YOLOv4-Tiny, as a lightweight variant, significantly reduces model size and inference time achieving extremely high frame rates (mAP@0.5 = 42.0%, ~443 FPS on RTX 2080 Ti) making it suitable for embedded or resource-limited devices, albeit at the cost of reduced detection accuracy [17]. YOLOv7, the most recent and advanced among the three, achieves state-of-the-art accuracy (mAP@0.5 = 69.7%, 161 FPS on Tesla V100) [18] in various object detection benchmarks. Although its architecture involves higher computational complexity compared to YOLOv4 and YOLOv4-Tiny, modern GPUs enable it to deliver real-time inference at high frame rates. This high accuracy is particularly beneficial in challenging low-light scenarios.

Table I summarizes the benchmark performance of these models on the COCO dataset, demonstrating the trade-off between speed and accuracy that motivated their selection for this study. This approach enables a comparative evaluation of fusion performance across different points of the accuracy–speed spectrum, providing insight into model suitability for both high-performance and real-time embedded applications.

TABLE I. YOLOv4 – YOLOv4-TINY – YOLOv7 COMPARISON ON COCO DATASET. DATA COMPILED FROM [16], [17], [18].

Model	mAP@0.5	FPS	GPU
YOLOv4	65.7%	~65	Tesla V100
YOLOv4-Tiny	42.0%	~443	RTX 2080 Ti
YOLOv7	69.7%	161	Tesla V100

Evaluating these three models on the LLVIP dataset allows a comprehensive performance comparison under low-light thermal–visible fusion conditions, covering deployment settings ranging from high-performance servers to low-power embedded surveillance systems.

IV. EXPERIMENTAL RESULTS AND EVALUATION

A. Object Detection Results

Object detection performance was evaluated using the mAP@0.5 (mean Average Precision at IoU=0.5) metric, which is commonly used in YOLO-based object detection models. This metric represents the average precision value calculated using an Intersection over Union (IoU) threshold of 0.5 and is widely used in the literature.

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on November 02,2026 at 09:40:05 UTC from IEEE Xplore. Restrictions apply.

TABLE II. OBJECT DETECTION RESULTS ON THE LLVIP DATASET

Image Set Description	YOLOv7 mAP@0.5	YOLOv4 mAP@0.5	YOLOv4-Tiny mAP@0.5
Thermal Only	**0.957**	0.904	**0.916**
RGB Only	0.792	0.811	0.808
Anisotropic Diffusion + KL Fusion	0.868	0.894	0.891
CLAHE-Thermal	0.952	0.900	0.912
CLAHE-RGB	0.770	0.773	0.807
CLAHE-Colored RGB	0.791	0.799	0.813
Anisotropic Diffusion + KL Fusion: CLAHE Thermal (0.5) + CLAHE RGB (0.5)	0.884	0.877	0.874
Anisotropic Diffusion + KL Fusion: Raw Thermal (0.6) + Raw RGB (0.4)	0.928	0.901	0.898
Anisotropic Diffusion + KL Fusion: Raw Thermal(0.5) + CLAHE-RGB (0.5)	0.913	0.873	0.882
Anisotropic Diffusion + KL Fusion: Raw Thermal(0.4) + Raw RGB(0.6)	0.912	0.886	0.882
Anisotropic Diffusion + KL Fusion: CLAHE-Thermal (0.5) + CLAHE-Colored RGB (0.5)	0.872	0.873	0.876
Anisotropic Diffusion + KL Fusion: CLAHE-Thermal (0.6) + CLAHE-RGB (0.4)	0.889	0.882	0.890
Saliency-Based Fusion	0.912	0.892	0.893
Saliency-Based Fusion: Raw Thermal + CLAHE-RGB	0.924	0.886	0.894
MST Fusion (Colored)	0.947	0.905	0.912
MST Fusion (Grayscale)	0.852	0.898	0.894
MST Fusion (Grayscale): CLAHE-Thermal + CLAHE-RGB	0.931	0.894	0.905
MST Fusion(Grayscale): Raw Thermal + CLAHE-RGB	0.949	**0.906**	0.909

As shown in Table II, using the YOLOv4 model, the fused images generated by combining CLAHE-enhanced RGB images with thermal images via the colored Multi-Scale Fusion method achieved the highest mAP@0.5 score—0.18% higher than thermal images alone and 9.53% higher than RGB images alone. This result highlights the positive impact of CLAHE preprocessing on RGB inputs in low-light object detection.

To further examine the influence of thermal and visible image contributions on detection performance, selected fusion results were generated by varying their weight ratios (e.g., 0.5/0.5 vs. 0.6/0.4). This partial ablation analysis, summarized in Table II, provides insight into how thermal-RGB dominance can affect accuracy under low-light conditions. The best mAP@0.5 result for each YOLO model is highlighted in bold for clarity. While YOLOv7 achieves the highest mAP@0.5 in most cases, consistent with its generally higher accuracy in standard benchmarks, YOLOv4 outperforms YOLOv7 in some fusion configurations. For example, in Anisotropic Diffusion + KL Fusion, CLAHE-RGB, and MST Fusion (Grayscale) configurations, YOLOv4's mAP@0.5 scores are higher than YOLOv7's. This can be explained by the YOLOv4 architecture being less prone to overfitting on small or homogeneous datasets of pedestrian images taken at close range at night, particularly in low-light conditions such as LLVIP. In such conditions, thermal imaging plays a dominant role in target visibility, and YOLOv4 can utilize this information more efficiently than YOLOv7 at certain fusion ratios. However, it's not surprising that YOLOv7 achieves the best results in most configurations; this success is supported by its higher capacity and advanced optimization strategies. The difference in the best YOLOv4 case (0.906 vs. 0.904 for thermal only) also indicates that in scenarios where thermal-only input is already highly informative, the relative gain from fusion can be small, and differences between architectures may be subtle. These results suggest that while YOLOv7 is the most robust choice overall, YOLOv4 remains competitive in specific fusion scenarios — particularly those where the dataset characteristics align with its architecture's strengths and where computational efficiency is a consideration.

B. Image Fusion Quality Metrics

Each image set was evaluated using three different quality metrics: $Q_{AB/F}$, Mutual Information (MI), and SSIM. A ranking score of 3 (highest), 2 (second), or 1 (third) was assigned to each image set for each metric, and the total score was computed accordingly. The Saliency-Based Fusion method achieved the highest cumulative score based on this scoring system. Table III highlights the top-ranked values in green, second in blue, and third in red to facilitate comparison across methods.

TABLE III. IMAGE FUSION QUALITY METRICS RESULTS

Image Set Description	**$Q_{AB/F}$**	**Mutual Information**	**SSIM**	**Total Score (Σ)**
Anisotropic Diffusion + KL Fusion	0.255	2.447	0.624 (1)	1
Anisotropic Diffusion + KL Fusion: CLAHE Thermal (0.5) + CLAHE RGB (0.5)	0.349	1.393	0.499	
Anisotropic Diffusion + KL Fusion: Raw Thermal (0.6) + Raw RGB (0.4)	0.265	2.638 (2)	0.620	2
Anisotropic Diffusion + KL Fusion: Raw Thermal (0.5) + CLAHE-RGB (0.5)	0.313	1.885	0.587	
Anisotropic Diffusion + KL Fusion: Raw Thermal (0.4) + Raw RGB (0.6)	0.244	2.308	0.624	
Anisotropic Diffusion + KL Fusion: CLAHE-Thermal (0.5) + CLAHE-Colored RGB (0.5)	0.349	1.396	0.498	
Anisotropic Diffusion + KL Fusion: CLAHE-Thermal (0.6) + CLAHE-RGB (0.4)	0.355	1.432	0.499	
Saliency-Based Fusion	0.620 (3)	2.238	0.642 (2)	5
Saliency-Based Fusion: Raw Thermal + CLAHE-RGB	0.569	1.693	0.591	
MST Fusion (Colored)	0.490	3.120 (3)	0.604	3
MST Fusion (Grayscale)	0.548	2.571 (1)	0.646 (3)	4
MST Fusion (Grayscale): CLAHE-Thermal + CLAHE-RGB	0.577 (1)	1.530	0.494	1
MST Fusion (Grayscale): Raw Thermal + CLAHE RGB	0.613 (2)	2.514	0.580	2

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on November 02,2026 at 09:40:05 UTC from IEEE Xplore. Restrictions apply.

Obtained results as in Table III support the idea that the Saliency-Based fusion method is an effective approach in terms of visual quality for IR-VIS image fusion applications.

A detailed analysis of the metrics reveals that Saliency-Based Fusion achieves the highest cumulative score of 5, a result primarily driven by its top-ranked performance in the $Q_{AB/F}$ metric. The $Q_{AB/F}$ score, which measures the preservation of edge information, highlights the method's effectiveness in retaining critical structural details from the source images during the fusion process. This observation aligns with the visual comparisons, where the Saliency-Based method exhibits superior preservation of scene contrast and edge details, particularly in low-light conditions.

While Saliency-Based Fusion demonstrates overall superiority, it is also important to note the strengths of other methods on specific metrics. For instance, the MST Fusion (Colored) method achieved the highest Mutual Information score of 3.120, suggesting that it is highly effective at retaining the shared information from both the visible and infrared source images. Similarly, the MST Fusion (Grayscale) method, with a score of 0.646, demonstrates the highest structural similarity (SSIM) to the original images, a key factor for maintaining visual fidelity. These findings collectively indicate that while Saliency-Based Fusion is the most robust choice in terms of overall visual quality metrics, the other fusion methods also offer unique advantages that could be considered depending on the specific visual quality requirements of an application

C. Visual Comparisons

Visual object detection comparisons were conducted using YOLOv4 across different image types, see Fig. 2 and Fig. 3. Visual comparisons reveal differences in edge clarity, object boundary preservation, and contrast. Fused images, particularly from Saliency-Based and MST methods, display enhanced structural detail compared to individual thermal or RGB inputs, which could be comparatively noticed from Fig. 2 and Fig. 3.

Among the visual results shown in Fig. 2 and Fig. 3, the Saliency-Based Fusion method exhibits superior preservation of edge details and scene contrast, particularly under low-light conditions. This observation aligns with its top ranking in fusion quality metrics, as reported in Table III.

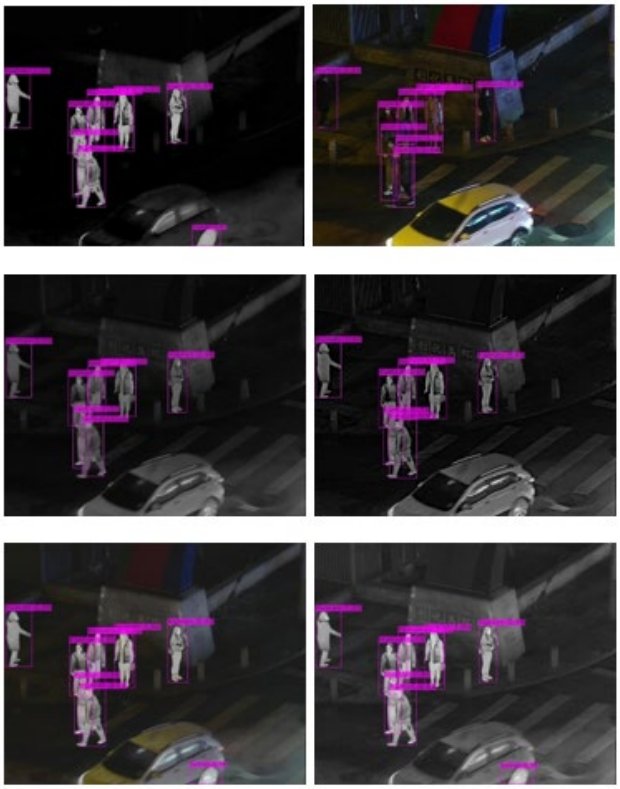


Fig. 2. Visual object detection results obtained using YOLOv4 across different image types. From top-left to bottom-right: (1) Thermal image, (2) RGB image, (3) Anisotropic Diffusion + KL Transform Fusion, (4) Saliency-Based Fusion, (5) MST Fusion (Colored), and (6) MST Fusion (Grayscale) using Raw Thermal + CLAHE-RGB



Fig. 3. Additional visual object detection results obtained using YOLOv4 across different image types. From top-left to bottom-right: (1) Thermal image, (2) RGB image, (3) Anisotropic Diffusion + KL Transform Fusion, (4) Saliency-Based Fusion, (5) MST Fusion (Colored), and (6) MST Fusion (Grayscale) using Raw Thermal + CLAHE-RGB

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on November 02,2026 at 09:40:05 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion

This research investigated how various image fusion techniques influence object detection performance and image quality metrics when applied to the LLVIP dataset. MST, Saliency, and Anisotropic Diffusion + KL methods were compared after CLAHE preprocessing and evaluated with YOLOv7, YOLOv4, and YOLOv4-Tiny models. For YOLOv4, the colored MST fusion generated from CLAHE-preprocessed RGB images yielded the highest mAP@0.5, achieving 0.18% higher accuracy than thermal images alone and 9.53% higher accuracy than RGB images alone. In addition, according to the image-quality-metric evaluation, the Saliency-Based fusion method delivered the best overall performance.

These findings demonstrate that classical image fusion methods can significantly improve object detection performance under low-light conditions and have strong potential for integration into real-time defense and surveillance systems. Although real-time deployment was not evaluated in this study, the relatively low computational complexity of the proposed fusion strategies—compared to deep learning-based approaches—further supports their suitability for real-time use on embedded platforms. Future work will focus on implementing and validating these methods in real-time applications.

Acknowledgment

We thank Kursad Ozcelik (ASELSAN) for his fruitful discussions about image fusion and the YOLO algorithms.

References

[1] Z. Hu, Y. Jing, and G. Wu, "Decision-level fusion detection method of visible and infrared images under low light conditions," EURASIP J. Adv. Signal Process., vol. 2023, no. 38, pp. 1–13, 2023, doi: 10.1186/s13634-023-01002-5

[2] J. Zhang, H. Zhu, Y. Gao, and B. Li, "GIVFuse: Global infrared-visible fusion method based on L1 distance," in Proc. China Autom. Congr. (CAC), 2023, pp. 2219–2224, doi: 10.1109/CAC59555.2023.10451430

[3] W. Ma, K. Wang, J. Li, S. X. Yang, J. Li, L. Song, and Q. Li, "Infrared and visible image fusion technology and application: A review," Sensors, vol. 23, no. 2, p. 599, Jan. 2023, doi: 10.3390/s23020599

[4] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, QC, Canada, Oct. 2021, pp. 3496–3504, doi: 10.1109/ICCVW54120.2021.00389

[5] Z. Hong, J. Li, W. Hu, and J. Wang, "YOLOv8-FA: Multimodal fusion image object detection algorithm," in Proc. 3rd IEEE Ind. Electron. Soc. Annu. Online Conf. (ONCON), Dec. 2024, Art. no. 10931300, doi: 10.1109/ONCON62778.2024.10931300.

[6] S. Meng and Y. Liu, "Multimodal feature fusion YOLOv5 for RGB-T object detection," in Proc. China Autom. Congr. (CAC), Xiamen, China, Nov. 2022, pp. 2333–2338, doi: 10.1109/CAC57257.2022.10055263

[7] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Vancouver, BC, Canada, Jun. 2023, pp. 403–411, doi: 10.1109/CVPRW59228.2023.00046

[8] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," Infrared Phys. Technol., vol. 76, pp. 52–64, 2016, doi: 10.1016/j.infrared.2016.01.009

[9] S. Na, J. Xue, Y. Liu, and Y. Zhang, "An image fusion method based on wavelet transform," in Proc. AEST 2016, Atlantis Press, 2016, doi: 10.2991/aest-16.2016.26

[10] K. P. Indira, "Image fusion for PET CT images using average maximum and average contrast rules," Int. J. Appl. Eng. Res., vol. 10, no. 1, pp. 673–680, 2015

[11] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and KL transform," MATLAB Central File Exchange, 2017. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/63591. [Accessed: Aug. 18, 2025]

[12] G. Lv, X. Wang, Z. Wei, J. Cheng, G. Ma, and H. Bao, "IDFusion: An infrared and visible image fusion network for illuminating darkness," in Proc. 27th Int. Conf. Comput. Supported Cooperative Work Design (CSCWD), Hangzhou, China, May 2024, pp. 3140–3145, doi: 10.1109/CSCWD61410.2024.10580330.

[13] J.-S. Yun, S.-H. Park, and S. B. Yoo, "Infusion-Net: Inter- and intra-weighted cross-fusion network for multispectral object detection," Mathematics, vol. 10, no. 21, Art. no. 3966, Oct. 2022, doi: 10.3390/math10213966

[14] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen–Loeve transform," IEEE Sensors Journal, vol. 16, no. 1, pp. 498–506, Jan. 2016, doi: 10.1109/JSEN.2015.2478655

[15] C. Xydeas and V. Petrovic, "Objective image fusion performance measure," Electronics Letters, vol. 36, no. 4, pp. 308–309, Mar. 2000, doi: 10.1049/el:20000291

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, Apr. 2020.

[17] C.-Y. Wang, A. Bochkovskiy and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling Cross Stage Partial Network," arXiv preprint arXiv:2011.08036v2, Feb. 2021.

[18] C.-Y. Wang, A. Bochkovskiy and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696v1, Jul. 2022.

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on November 02,2026 at 09:40:05 UTC from IEEE Xplore. Restrictions apply.