Application-Guided Image Fusion: A Path to Improve Results in High-Level Vision Tasks

Gisel Bastidas-Guacho[1,2][a], Patricio Moreno-Vallejo[2,1][b], Boris Vintimilla[1][c] and Angel D. Sappa[1,3][d]

[1]*ESPOL Polytechnic University, Campus Gustavo Galindo, Km. 30.5 Vía Perimetral, Guayaquil, Ecuador*

[2]*ESPOCH Polytechnic University, South Pan-American Highway km 1 1/2, Riobamba, Ecuador*

[3]*Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain*

Keywords: Multimodal, Fusion, Semantic Segmentation, Application-Driven.

Abstract: This paper proposes an enhanced application-driven image fusion framework to improve final application results. This framework is based on a deep learning architecture that generates fused images to better align with the requirements of applications such as semantic segmentation and object detection. The color-based and edge-weighted correlation loss functions are introduced to ensure consistency in the YCbCr space and emphasize structural integrity in high-gradient regions, respectively. Together, these loss components allow the fused image to retain more features from the source images by producing an application-ready fused image. Experiments conducted on two public datasets demonstrate a significant improvement in mIoU achieved by the proposed approach compared to state-of-the-art methods.

1 INTRODUCTION

In recent years, the proliferation of imaging devices capable of capturing data across a range of spectral bands has unlocked new possibilities for computer vision applications beyond the visible spectrum. These multimodal devices are increasingly used in fields such as surveillance, driver assistance, mobile technology, and industrial monitoring, just to mention a few (Sun et al., 2021; Karim et al., 2023). Effectively processing the diverse information from these sensors requires advanced multimodal image fusion techniques, which combine data from different spectral bands to provide a comprehensive, enhanced view of the scene. It is important to note that for optimal fusion results, the images captured by different modalities must be properly registered to ensure accurate alignment and integration of information (Velesaca et al., 2024).

Multimodal image fusion techniques aim to extract and integrate complementary information from images captured in different modalities. As an example of these multimodal image fusion techniques, we can mention approaches based on the usage of visible (VIS) and thermal infrared (IR). In these cases VIS and IR are commonly fused because each provides complementary information. For instance, visible images offer texture details of the image, but low-light conditions and occlusions can hinder it. In contrast, infrared images capture thermal radiation, allowing them to provide temperature information independently of lighting conditions (Liu et al., 2024b). Therefore, the resulting fused image contains valuable enriched information for high-level vision tasks such as semantic segmentation, object detection, and classification.

Several multimodal image fusion approaches exist; however, most do not consider application. In other words, the fused image is not optimized for a high-level task (Tang et al., 2022c; Le et al., 2022; Karim et al., 2023; Liang et al., 2024; Yang et al., 2024). For instance, PIAFusion is a fusion framework that integrates a cross-modality differential fusion module and an illumination-aware loss function to adaptively combine salient features from both modalities according to the lighting context. However, this framework does not consider specific high-level vision application contexts that could be useful

[a] https://orcid.org/0000-0002-6070-7193
[b] https://orcid.org/0000-0002-9317-9884
[c] https://orcid.org/0000-0001-8904-0209
[d] https://orcid.org/0000-0003-2468-0031

Bastidas-Guacho, G., Moreno-Vallejo, P., Vintimilla, B. and Sappa, A. D.
Application-Guided Image Fusion: A Path to Improve Results in High-Level Vision Tasks.
DOI: 10.5220/0013307300003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 178-187
ISBN: 978-989-758-728-3; ISSN: 2184-4321

to decide which salient feature is more important.

In order to enhance high-level vision tasks with multimodal images, several studies propose end-to-end frameworks where the fusion result is not represented as an image (Zhang et al., 2021; Sun et al., 2021; Chu and Lee, 2023; Xiao et al., 2024; Liu et al., 2024c). However, these frameworks often need more flexibility and adaptability to different applications. Therefore, recent deep learning approaches integrate the fusion and application process in a unified pipeline, using the fused image as an intermediate output (Tang et al., 2022b; Tang et al., 2022a; Sun et al., 2022; Liu et al., 2024b). These application-driven multimodal image fusion methods connect the fusion and application networks through a feedback mechanism during learning, enabling mutual enhancement. That is, the high-level vision application benefits from improved fusion, while the fusion process is simultaneously optimized to more effectively support the application (Bastidas-Guacho et al., 2023). Figure 1 illustrates the pipeline of the end-to-end application framework and the multimodal application-driven image fusion framework.

Among the different applications that take advantage of multimodal source of information, we can find the multimodal semantic segmentation (MSS). MSS is a high-level vision task that has gained significant attention due to its ability to capture comprehensive information from various modalities, such as infrared and visible images, to enhance segmentation accuracy. This approach has proven invaluable in fields such as autonomous driving (Feng et al., 2020) and surveillance (Wang et al., 2021), where single-modal data often need more detail to ensure reliable decision-making (Dong et al., 2024). Application-driven frameworks have been developed for multimodal semantic segmentation. For example, Tang et al. introduce a semantic-aware framework for infrared and visible image fusion (Tang et al., 2022b). This framework employs the loss function as a feedback mechanism between fusion and application to optimize both the quality of the fused image and the effectiveness of high-level vision tasks. Another approach is proposed in (Liu et al., 2023), which employs a hierarchical interactive attention block to enhance feature exchange between fusion and segmentation tasks.

In application-driven approaches, the loss function is the feedback mechanism, integrating fusion and application during learning. The current work proposes an adaptation of an existing application-driven fusion approach to improve MSS results. A new loss function is proposed to simultaneously optimize both fusion and final application, incorporating additional semantic information into the learning process. The network architecture is based on the framework presented in (Tang et al., 2022b). We propose the use of color and correlation between the source multimodal images and fused images in order to maximizes the preservation of the salient features from visible and infrared modalities. The correlation is computed in a weighted manner to measure the similarity of intensity variations between the fused and source images. This approach aims to preserve more detail in the fused image to improve high-level vision tasks results. The approach proposed in this paper is compared with the state-of-the-art application-driven image fusion frameworks.

This paper is organized as follows. Section 2 briefly introduces existing application-driven multimodal image fusion approaches for semantic segmentation. In Section 3, the adaptation proposed in the current work is presented. Section 4 presents the experimental validations of the proposed method. Finally, Section 5 provides conclusions and discusses future directions for deep learning-based multimodal image fusion in semantic segmentation.

2 RELATED WORK

This section details deep learning-based multimodal image fusion methods for semantic segmentation. Firstly, standalone approaches are summarized, then end-to-end frameworks, and finally application-driven approaches are reviewed.

2.1 Standalone Fusion Frameworks

Traditional multimodal image fusion methods are limited by fixed fusion rules and high computational demands (Li et al., 2023). Recent deep learning-based frameworks automate feature extraction, integration, and image reconstruction, addressing limitations in traditional methods. These frameworks dynamically balance the contributions of each modality. (Zhang and Demiris, 2023) categorize these approaches as CNN-based (Mustafa et al., 2020; Li et al., 2020; Tang et al., 2022c; Xu et al., 2022), autoencoder-based (Li and Wu, 2018; Wang et al., 2022; Yang et al., 2024; Liu et al., 2024a), GAN-based (Ma et al., 2020; Fu et al., 2021; Zhang et al., 2024a), and transformer-based.

CNN-based frameworks first automate feature extraction and fusion with enhancements such as residual and dense connections to retain good detail and attention mechanisms for feature emphasis. For instance, in (Li et al., 2020) the author propose to use

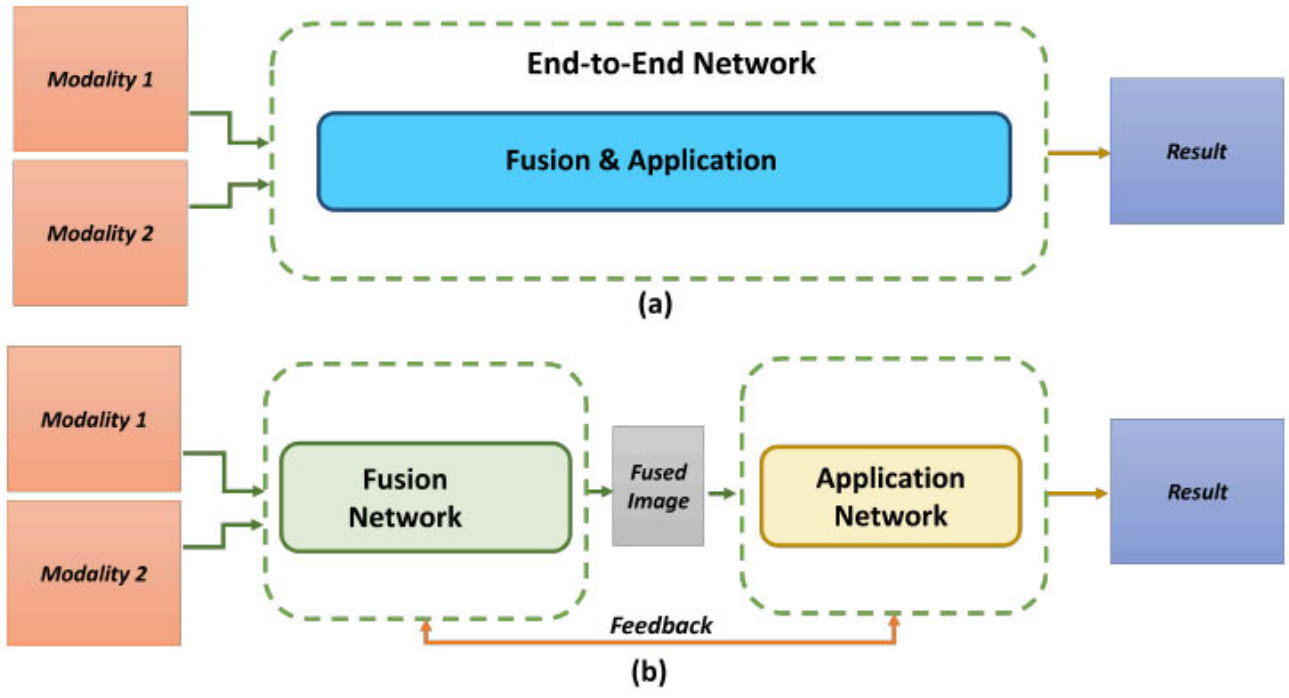

Figure 1: (a) End-to-end approaches of multimodal image-based application pipeline. (b) Application-driven image fusion approaches pipeline.

a dense connection structure combined with an attention mechanism to enhance fusion performance. They introduce cross-dimensional weighting and aggregation to calculate an attention map tailored for infrared and visible image fusion. Another standalone fusion framework has been proposed in (Tang et al., 2022c), referred to as PIAFusion; this architecture is a CNN-based framework incorporating an illumination estimation module that detects the lighting conditions of the scene (day or night). This module adapts the fusion process by focusing on visible details in well-lit conditions and prioritizing infrared information in low-light scenarios.

Autoencoder-based frameworks use the encoder for feature extraction and the decoder for image reconstruction. Recently, in (Liu et al., 2024a), the DS-Fusion model has been proposed. It consists of an encoder, a local attention module, an attention mechanism for feature extraction and fusion, and a decoder to generate the fused image. Similarly, the CEFusion framework has been proposed in (Yang et al., 2024). It includes a multigranularity feature separation encoder, a triple-branch scene fidelity module, a progressive cross-granularity interaction feature enhancement module, and an image reconstruction decoder.

GAN-based frameworks have been also proposed for image fusion. For instance, FusionGAN (Ma et al., 2019) use adversarial networks to integrate visible and infrared features while retaining target clarity. Recently, (Zhang et al., 2024a) introduces a Guided Restoration Module, which uses a Conditional GAN to recover relevant information hidden in the darkness in the visible modality. GAN-based models incorporate multiple discriminators to effectively maintain thermal and visible information. Due to transformers' ability to handle long-range dependencies, they have recently been applied to feature extraction in image fusion. Frameworks like DNDT (Zhao and Nie, 2021) and SwinFusion (Ma et al., 2022), utilize self-attention to enhance multi-scale feature extraction. Other architectures combine transformers with GANs and CNNs, integrating spatial and channel-based attention such as CGTF (Li et al., 2022).

Although standalone approaches are designed to obtain the best representation based on the input data, these approaches do not consider the specific needs or constraints of the application where the fused image will be used. For this reason, when attempting to preserve the input information in a balanced way, the best result may not be optimized for a given application. Thus, if the method does not adapt the fusion to the needs of applications, the intended multimodal preservation balance can degrade the performance of the application. It highlights the need for application-driven image fusion strategies.

2.2 End-to-End Fusion and Application Frameworks

End-to-end frameworks for multimodal fusion and applications consist of a unified pipeline to streamline workflow, from data preprocessing to feature extraction, fusion and prediction. Most of these approaches use a two-branched encoder-decoder backbone network. In (Sun et al., 2021) a backbone network based on residual FCN architecture is used. It introduces two types of attention fusion modules (in-

tramodal and intermodal) to integrate features within and across modalities effectively. The network also employs a multiscale supervision training strategy to optimize performance across different resolutions of the feature.

On the other hand, (Wang et al., 2023) proposes SGFNet. This network uses an asymmetric encoder and a decoder. The encoder processes RGB and IR branches. It uses semantic guidance head for semantic information extraction from IR images. Besides, the multimodal coordination and distillation unit and the cross-level and semantic-guided enhancement unit are designed to enhance feature fusion while minimizing interference from lighting noise in RGB inputs. Moreover, the edge-aware Lawin ASPP decoder (Yan et al., 2022) refines segmentation boundaries by incorporating edge information from low-level features. Recently, in (Xiao et al., 2024) the authors use the encoder and transformer decoder as the backbone of the framework and propose GM-DETR, which employs a two-stage training strategy that involves isolated training on individual RGB and IR data sets followed by a fusion stage with aligned multimodal data.

End-to-end frameworks for multimodal fusion and application often do not allow visualizing intermediate fusion results, which may limit the ability to assess how well information from each modality is integrated. This limitation highlights the need for approaches that offer modular fusion and application stages to allow more control and interpretation of the fusion process, as well as the use of different application frameworks.

2.3 Application-Driven Fusion Frameworks

Application-driven multimodal image fusion frameworks optimize the fusion process to improve performance in a given high-level vision task—in the current work semantic segmentation will be considered. These approaches leverage the complementary information of each modality by aligning the fused output with the specific needs of high-level vision tasks. In (Tang et al., 2022b) the SeAFusion is proposed; it performs cascade optimization between the fusion and application networks using semantic loss to guide high-level vision task information back into the image fusion module. The infrared and visible image fusion network consists of a feature extractor and an image reconstructor. The feature extractor is based on gradient residual dense blocks, while the image reconstructor utilizes a convolutional layer. The training strategy iteratively alternates between the fusion network and the segmentation network. As segmentation network the authors propose to use the architecture presented in (Peng et al., 2021), but other networks could be also considered.

Another application-driven fusion framework is the SegMiF (Liu et al., 2023); it is a multi-interactive feature learning framework for joint multimodal image fusion and segmentation. This framework utilizes a cascade structure combining a fusion sub-network and a segmentation sub-network, where semantic information is exchanged through a hierarchical interactive attention block to enhance feature integration. In order to balance the importance of each task, a dynamic weighting factor is proposed, which automatically optimizes parameters. Recently, the mutually reinforcing image fusion and segmentation framework has been proposed in (Zhang et al., 2024b). It leverages a coupled learning approach that uses the interactive gated mix attention module, which refines visual features and addresses issues like mismatched feature scores, and the progressive cycle attention module, which enhances semantic understanding by enabling both self-reinforcement within a single modality and cross-modal complementarity. HitFusion (Chen et al., 2024) is another example of a application-driven framework. It is based on transformers using a three-stage training strategy. A cross-feature transformer module is introduced to enhance feature extraction by capturing correlations between visible texture and infrared contrast features. The architecture also uses a dual-branch network design with contrast residual and texture enhancement modules to allow deep feature extraction from source images.

The aforementioned application-driven frameworks have progressively refined loss functions to enhance the extraction and integration of critical information across modalities. The enhancement in the loss function design helps preserve salient features from infrared images and detailed textures from visible images. As a result, the fused images are more valuable and practical for high-level vision applications. Therefore, it demonstrates that a well-structured loss function is necessary for effective feature integration and task performance in multimodal image fusion for a given application.

3 METHODOLOGY

This section presents in detail the proposed framework introducing the designed loss function.

3.1 Problem Definition

The infrared and visible fusion problem in this paper aims to combine complementary information from an infrared image I_{ir} and a visible image I_{vi} into a single fused image I_{f}. Formally, given two registered input images I_{ir} and I_{vi}, the objective is to construct a fused image I_{f}, which is defined as:

$$I_{\text{f}} = F(I_{\text{ir}}, I_{\text{vi}}), \tag{1}$$

where $F(\cdot)$ represents the fusion network. The network $F(\cdot)$ must learn to capture and retain the contrast information from I_{ir} (such as thermal radiation and object saliency) and the fine-grained texture details from I_{vi}, optimizing I_{f} in order to improve the high-level vision application performance. Since ground truth fused images are unavailable, the fusion problem is approached as an unsupervised learning task. Thus, a loss function that guides the training is defined.

3.2 Framework

In the current paper the SeAFusion framework (Tang et al., 2022b) is used as the basis for the application-driven multimodal image fusion approach. Through SeAFusion's application-driven structure, we can optimize the fusion output for visual fidelity and support high-level vision tasks such as semantic segmentation; other tasks, or segmentation architectures, could be also considered. In that way, we ensure that the fused images are informative and applicable to real-world scenarios. In this framework we introduce a new loss function to address certain limitations in SeAFusion and preserve more structural and texture details in the fused image. Thus, the new loss function generates a more balanced fusion output, which helps to enhance the performance of application-driven image fusion approaches and makes them more robust for real-world applications.

3.3 Loss Function

The fused image quality relies on the loss function because it is the primary guide for the fusion process. The loss function ensures that the fused image retains features and details from the source images. A well-designed loss function can balance preserving structural information, enhancing feature contrast, and reducing artifacts. In order to reinforce semantic information in the fused image, we propose a loss function that maximizes the preservation of the content and structural integrity of I_{f} with respect to both, I_{ir} and I_{vi} to facilitate high-level vision tasks such as semantic segmentation, which is the case of the current study. The loss function includes both fusion loss ($\mathcal{L}_{\text{fusion}}$) and segmentation loss ($\mathcal{L}_{\text{seg}}$), formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fusion}}(I_{\text{f}}, I_{\text{ir}}, I_{\text{vi}}) + \lambda \mathcal{L}_{\text{seg}}(I_{\text{f}}), \tag{2}$$

where λ is a hyperparameter to trade off the content and segmentation losses.

3.3.1 Fusion Loss

In order to achieve a more accurate fusion output that preserves structural and texture details to align with the requirements of the given high-level vision tasks, we propose $\mathcal{L}_{\text{fusion}}$, which consists of four terms defined as follows:

$$\mathcal{L}_{\text{fusion}} = \mathcal{L}_{\text{int}} + \alpha \mathcal{L}_{\text{text}} + \beta \mathcal{L}_{\text{color}} + \gamma \mathcal{L}_{\text{wcorr}}, \tag{3}$$

where $\mathcal{L}_{\text{int}}$, $\mathcal{L}_{\text{text}}$, $\mathcal{L}_{\text{color}}$, and $\mathcal{L}_{\text{wcorr}}$ denote the intensity loss, the texture loss, the color loss, and the edge-weighted correlation loss, respectively. α, β, and γ are hyperparameters that control the influence of each component on the total fusion loss.

The intensity loss aims to preserve the intensity patterns in the fused image by measuring pixel-by-pixel intensity difference between fused images and source image. This loss is represented as:

$$\mathcal{L}_{\text{int}} = \frac{1}{HW} \left\| I_f - \max(I_{\text{ir}}, I_{\text{vi}}) \right\|_1, \tag{4}$$

where H and W represent the height and width of the images, respectively. $\|\cdot\|_1$ denotes ℓ_1-norm. $\max(\cdot)$ stands for the element-wise maximum selection.

The texture loss is used to preserve the texture details of the source images, which is defined as:

$$\mathcal{L}_{\text{texture}} = \frac{1}{HW} \left\| |\nabla I_f| - \max(|\nabla I_{\text{ir}}|, |\nabla I_{\text{vi}}|) \right\|_1, \tag{5}$$

where ∇ refers to Sobel gradient operator. $|\cdot|$ indicates the absolute operation.

The color loss aims to maintain color consistency in YCbCr space as proposed in (Zhang et al., 2024b), which is defined as:

$$\mathcal{L}_{\text{color}} = \left\| \text{Cb}^f - \text{Cb}^{\tilde{\text{vi}}} \right\|_1 + \left\| \text{Cr}^f - \text{Cr}^{\tilde{\text{vi}}} \right\|_1, \tag{6}$$

where Cb^f and Cr^f represent the Cb and Cr channels of I_f, respectively. $\tilde{I}_{\text{vi}}$ denotes the visible image that is transformed using gamma correction, and $\text{Cb}^{\tilde{\text{vi}}}$ and $\text{Cr}^{\tilde{\text{vi}}}$ denote the Cb and Cr channels of $\tilde{I}_{\text{vi}}$.

Finally, the edge-weighted correlation loss ($\mathcal{L}_{\text{wcorr}}$) measures the Pearson correlation coefficient in a weighted manner. Traditional correlation measures treat all regions of an image equally, which can lead to suppression of structural features, especially in areas such as edges. These features convey critical

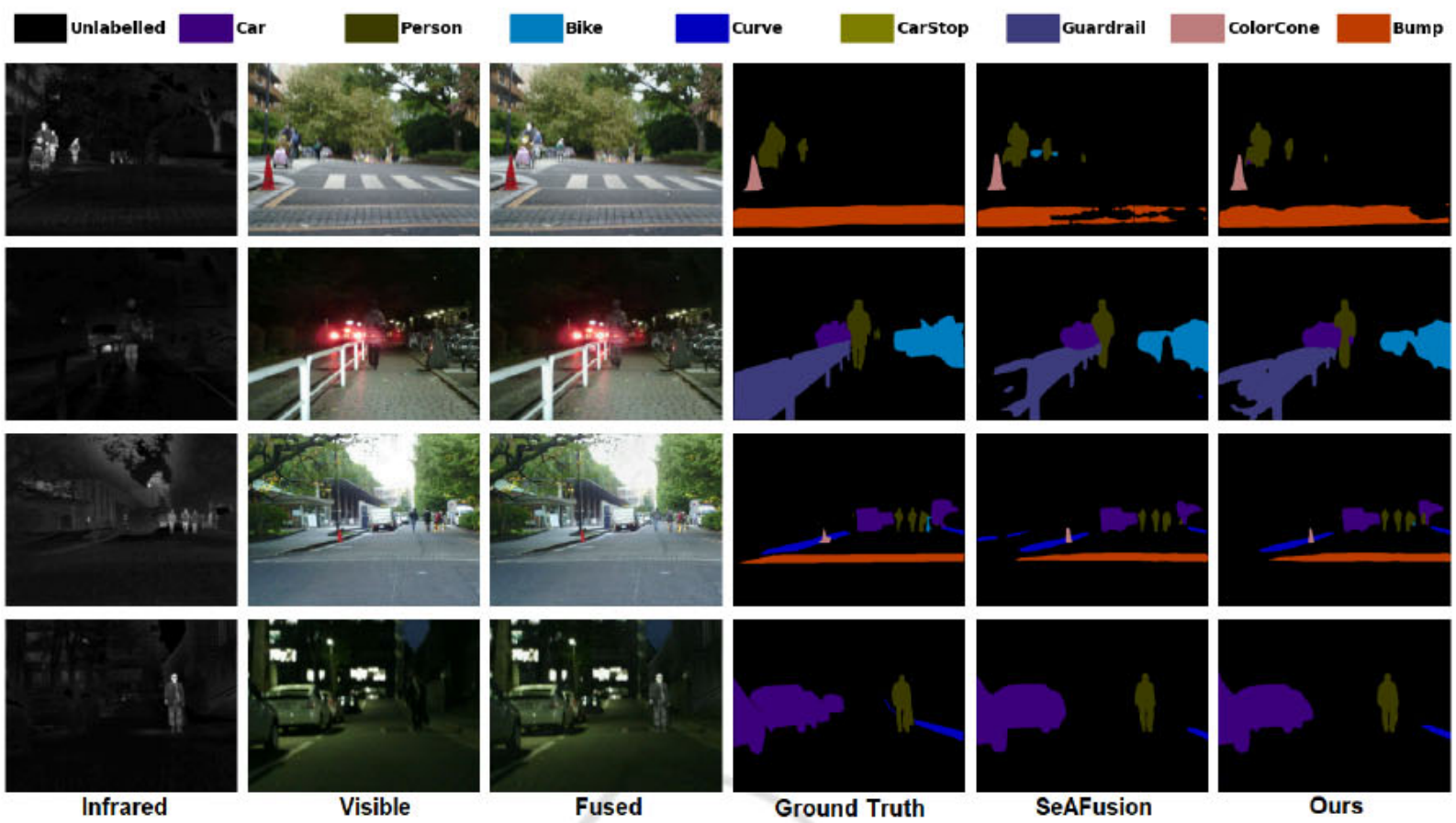

Figure 2: Results on MSRS dataset.

information about the underlying content, which can be helpful for computer vision applications. Therefore, $\mathcal{L}_{\text{wcorr}}$ uses the Sobel gradient value as the mask for prioritizing regions with high structural importance to make the correlation measure more sensitive to edge preservation in the fusion process. High-weighted correlations in edge regions indicate effective retention of essential features from the source images. On the other hand, lower correlations in homogeneous regions imply less emphasis on these areas. It is expressed as:

$$\mathcal{L}_{\text{wcorr}} = \frac{1}{\text{WCorr}_{\text{ir}}(I_f, I_{\text{ir}}, S_{\text{ir}}) + \text{WCorr}_{\text{vi}}(I_f, I_{\text{vi}}, S_{\text{vi}})}, \quad (7)$$

where WCorr $(\cdot)$ denotes the weighted correlation function, S_{ir} and S_{vi} refer to Sobel gradient mask for infrared and visible modalities, respectively. WCorr_x is computed as follows:

$$\text{WCorr}_x = \frac{\sum_{i=1}^{n} S_{x_i} \cdot (I_{f_i} - \bar{I}_f)(I_{x_i} - \bar{I}_x)}{\sqrt{\sum_{i=1}^{n} S_{x_i} \cdot (I_{f_i} - \bar{I}_f)^2} \cdot \sqrt{\sum_{i=1}^{n} S_{x_i} \cdot (I_{x_i} - \bar{I}_x)^2}}, \quad (8)$$

where x denotes the infrared or visible modalities. The weighted correlation function ensures that the correlation calculation focuses on high-gradient regions (edges) while reducing the influence of homogeneous areas (non-edge regions).

In conclusion, the proposed fusion loss components together balance preservation and adaptation among multimodality information. In that way, they allow the production of a fusion result that is both visually consistent and application-ready.

3.3.2 Segmentation Loss

In this work, a semantic segmentation task is used as a case study for application-driven fusion; specifically, the real-time semantic segmentation model (Peng et al., 2021) is employed to segment the fused images. Therefore, the semantic loss ($\mathcal{L}_{\text{seg}}$) includes a main semantic loss and an auxiliary semantic loss, defined as:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{main}} + \lambda_1 \mathcal{L}_{\text{aux}}, \quad (9)$$

where $\lambda_1 = 0.1$ balances the main and auxiliary losses. The main semantic loss and the auxiliary semantic loss are expressed as:

$$\mathcal{L}_{\text{main}} = \frac{-1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} L_{\text{so}}^{(\text{h, w, c})} \log(I_s^{(\text{h, w, c})}), \quad (10)$$

$$\mathcal{L}_{\text{aux}} = \frac{-1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} L_{\text{so}}^{(\text{h, w, c})} \log(I_{sa}(\text{h, w, c})), \quad (11)$$

where $L_{\text{so}} \in \mathbb{R}^{H \times W \times C}$ is a one-hot vector transformed from the segmentation label $L_s \in (1, C)^{H \times W}$. For more comprehensive details on the segmentation loss function definitions and network architecture, refer to (Peng et al., 2021).

4 EXPERIMENTAL RESULTS

This section details the experimental settings and datasets used to evaluate the proposed framework.

Table 1: Comparison of Segmentation Results on MFNet Dataset.

Framework	Car	Person	Bike	Curve	Carstop	Color Tone	Bump	mIoU
SegMiF	87.80	71.40	63.20	47.60	31.10	48.90	50.30	57.19
PSFusion	88.62	75.00	65.53	47.34	42.59	52.96	58.68	61.53
MRFS	89.40	75.40	65.00	49.00	37.20	53.10	58.80	61.13
Seafusion	92.95	82.69	81.25	88.52	43.87	54.90	72.32	73.79
Ours	**93.36**	**84.16**	**81.54**	**90.29**	**46.49**	**55.69**	**73.13**	**74.95**

Table 2: Performance comparison of the proposed approach on MFNet and MSRS datasets.

	Car	Person	Bike	Curve	Carstop	Color Tone	Bump	mIoU
MFNet	93.36	84.16	81.54	90.29	46.49	55.69	73.13	74.95
MSRS	97.90	89.08	90.30	97.72	92.05	83.24	99.05	**92.76**

Then, comparison with recent application-driven image fusion approaches are presented to demonstrate the advantages of the approach proposed for semantic segmentation.

In order to evaluate the proposed method, the MFNet (Ha et al., 2017) and MSRS (Tang et al., 2022c) datasets are selected. The MFNet consists of 1569 pairs of infrared and visible images, while MSRS consists of 1083. Additionally, the proposed approach is compared with recent application-driven image fusion frameworks for semantic segmentation, including **Seafusion** (Tang et al., 2022b), **SegMiF** (Liu et al., 2023), **PSFusion**(Tang et al., 2023), **MRFS** (Zhang et al., 2024b). The fusion network is iteratively trained with the segmentation network based on the joint adaptive learning strategy proposed in (Tang et al., 2022b). The hyperparameters for the fusion loss are set as $\alpha = 10$, $\beta = 10$, and $\gamma = 0.01$. The visible images are converted to the YCbCr color space, where the Y channel of visible images and infrared images are fused. The fused images are then converted to the RGB color space using the Cb and Cr channels from the given visible images. Gamma correction in the color loss is set to 0.1.

For fair comparisons with the baseline framework of our approach, we retrain SeAFusion using both the original loss function and our proposed loss function. The results presented for the other application-driven image fusion methods are the reported results in the original papers. The semantic segmentation performance is measured by pixel intersection-over-union (*IoU*). The results on the MFNet dataset are shown in Table 1. The MFNet dataset includes nine classes: car, person, bike, curve, carstop, guardrail, color tone, bump, and background. The results reported here are based on the classes available across all application-driven approaches, meaning that the guardrail and background classes are excluded from the results on the MFNet dataset.

The proposed method achieves the highest mean Intersection over Union (*mIoU*) score, outperforming others in all categories. These scores reflect our approach's ability to retain fine details and semantic information in complex image regions. Additionally, our approach performs well in less frequent categories, such as Car stop and Color Tone. These results demonstrate that our proposed method consistently orients the fused image toward the segmentation task, improving the segmentation quality across various classes.

In addition, the proposed framework is also evaluated on the MSRS dataset. This dataset is generated from MFNet dataset by removing 125 misaligned image pairs. Furthermore, an image enhancement algorithm based on the dark channel prior is applied to optimize the contrast and signal-to-noise ratio of the infrared images. On this dataset, the proposed approach achieves a *mIoU* of $92,76$. Figure 2 shows some results for a qualitative comparison between SeAFusion and the proposed approach across different scenes on the MSRS dataset. The results show a more precise scene segmentation by enhancing the delineation of object boundaries, especially obstacles such as bumps, and cones. This advantage highlights our approach's robustness in both day and night scenes by contributing better segmentation accuracy.

Table 2 compares the performance of the proposed method on the MFNet and MSRS datasets. The results show that the proposed approach achieves high *mIoU* scores in the MSRS dataset, with a score of 92.76. Additionally, there are performance differences between datasets for specific classes, such as carstop, colortone, and bump, where the method performs significantly better in MSRS than in MFNet.

An ablation study is performed to demonstrate the effectiveness of the proposed framework. Specifically, the impact of each loss component on the performance of the proposed framework is evaluated.

Table 3: Ablation Study on Different Loss Configurations using MFNet dataset.

$\mathcal{L}_{\text{int}}$	$\mathcal{L}_{\text{text}}$	$\mathcal{L}_{\text{color}}$	$\mathcal{L}_{\text{wcorr}}$	**mIoU**
✓				66.65
	✓			68,99
		✓		55.19
			✓	69.13
✓	✓			73.79
✓		✓		68.09
✓			✓	66.32
	✓	✓		67.27
	✓	✓	✓	66.55
	✓		✓	66.41
		✓	✓	68.92
✓	✓	✓		74.13
✓	✓	✓	✓	**74.95**

The experiments are conducted on the MFNet dataset

The results in Table 3 indicate that the configuration that includes all loss components yields the highest mIoU score. These results show that the proposed loss function improves overall segmentation performance by balancing intensity preservation, texture details, color fidelity, and structural coherence.

5 CONCLUSIONS

This paper presents an application-driven image fusion framework that introduces a new loss function that addresses the challenges of maintaining semantic and structural integrity in fused images to improve high-level vision tasks. The experimental results on the MFNet and MSRS datasets highlight that the framework outperforms recent application-driven fusion approaches. In particular, the proposed framework achieves a mIoU of 92.76% on the MSRS dataset. This performance is evident in accurately segmenting challenging objects, such as bumps, guardrails, and cones. The ablation study further validates the contribution of the color-based and edge-weighted correlation loss to enhancing the fusion quality for high-level vision tasks. Therefore, this framework generates fused images that benefit the performance of high-level vision applications. Future work will focus on integrating this framework with additional high-level vision, such as object detection.

ACKNOWLEDGEMENTS

This work was supported in part by Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206 and in part by the ESPOL project CIDIS-003-2024. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities, and the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR01499.

REFERENCES

Bastidas-Guacho, G., Moreno-Vallejo, P., Vintimilla, B., and Sappa, A. D. (2023). Application on the loop of multimodal image fusion: Trends on deep-learning based approaches. In *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–8.

Chen, J., Ding, J., and Ma, J. (2024). Hitfusion: Infrared and visible image fusion for high-level vision tasks using transformer. *IEEE Transactions on Multimedia*, 26:10145–10159.

Chu, S.-Y. and Lee, M.-S. (2023). Mt-detr: Robust end-to-end multimodal detection with confidence fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5252–5261.

Dong, S., Feng, Y., Yang, Q., Huang, Y., Liu, D., and Fan, H. (2024). Efficient multimodal semantic segmentation via dual-prompt learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 14196–14203. IEEE.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., and Dietmayer, K. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360.

Fu, Y., Wu, X.-J., and Durrani, T. (2021). Image fusion based on generative adversarial network consistent with perception. *Information Fusion*, 72:110–125.

Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., and Harada, T. (2017). Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE.

Karim, S., Tong, G., Li, J., Qadir, A., Farooq, U., and Yu, Y. (2023). Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion*, 90:185–217.

Le, Z., Huang, J., Xu, H., Fan, F., Ma, Y., Mei, X., and Ma, J. (2022). Uifgan: An unsupervised continual-learning generative adversarial network for unified image fusion. *Information Fusion*, 88:305–318.

Li, G., Qian, X., and Qu, X. (2023). Sosmaskfuse: An infrared and visible image fusion architecture based on salient object segmentation mask. *IEEE Transactions*

on Intelligent Transportation Systems, 24(9):10118–10137.

Li, H. and Wu, X.-J. (2018). Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623.

Li, J., Zhu, J., Li, C., Chen, X., and Yang, B. (2022). Cgtf: Convolution-guided transformer for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14.

Li, Y., Wang, J., Miao, Z., and Wang, J. (2020). Unsupervised dense attention network for infrared and visible image fusion. *Multimedia Tools and Applications*, 79(45):34685–34696.

Liang, L., Shen, X., and Gao, Z. (2024). Ifici: Infrared and visible image fusion based on interactive compensation illumination. *Infrared Physics & Technology*, 136:105078.

Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z., and Fan, X. (2023). Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8115–8124.

Liu, K., Li, M., Chen, C., Rao, C., Zuo, E., Wang, Y., Yan, Z., Wang, B., Chen, C., and Lv, X. (2024a). Dsfusion: Infrared and visible image fusion method combining detail and scene information. *Pattern Recognition*, page 110633.

Liu, X., Huo, H., Li, J., Pang, S., and Zheng, B. (2024b). A semantic-driven coupled network for infrared and visible image fusion. *Inf. Fusion*, 108(C).

Liu, X., Xu, X., Xie, J., Li, P., Wei, J., and Sang, Y. (2024c). Fdenet: Fusion depth semantics and edge-attention information for multispectral pedestrian detection. *IEEE Robotics and Automation Letters*, 9(6):5441–5448.

Ma, J., Liang, P., Yu, W., Chen, C., Guo, X., Wu, J., and Jiang, J. (2020). Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98.

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., and Ma, Y. (2022). Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217.

Ma, J., Yu, W., Liang, P., Li, C., and Jiang, J. (2019). Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48:11–26.

Mustafa, H. T., Yang, J., Mustafa, H., and Zareapoor, M. (2020). Infrared and visible image fusion based on dilated residual attention network. *Optik*, 224:165409.

Peng, C., Tian, T., Chen, C., Guo, X., and Ma, J. (2021). Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation. *Neural Networks*, 137:188–199.

Sun, Y., Cao, B., Zhu, P., and Hu, Q. (2022). Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4003–4011.

Sun, Y., Fu, Z., Sun, C., Hu, Y., and Zhang, S. (2021). Deep multimodal fusion network for semantic segmentation using remote sensing image and lidar data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18.

Tang, L., Deng, Y., Ma, Y., Huang, J., and Ma, J. (2022a). Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137.

Tang, L., Yuan, J., and Ma, J. (2022b). Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42.

Tang, L., Yuan, J., Zhang, H., Jiang, X., and Ma, J. (2022c). Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92.

Tang, L., Zhang, H., Xu, H., and Ma, J. (2023). Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, 99:101870.

Velesaca, H., Bastidas, G., Rouhani, M., and Sappa, A. (2024). Multimodal image registration techniques: a comprehensive survey. *Multimedia Tools and Applications*, 83:1–29.

Wang, C., Yang, G., Sun, D., Zuo, J., Wang, E., and Wang, L. (2021). Frequency domain fusion algorithm of infrared and visible image based on compressed sensing for video surveillance forensics. In *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 832–839. IEEE.

Wang, Y., Li, G., and Liu, Z. (2023). Sgfnet: semantic-guided fusion network for rgb-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7737–7748.

Wang, Z., Wu, Y., Wang, J., Xu, J., and Shao, W. (2022). Res2fusion: Infrared and visible image fusion based on dense res2net and double nonlocal attention models. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12.

Xiao, Y., Meng, F., Wu, Q., Xu, L., He, M., and Li, H. (2024). Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5541–5549.

Xu, D., Zhang, N., Zhang, Y., Li, Z., Zhao, Z., and Wang, Y. (2022). Multi-scale unsupervised network for infrared and visible image fusion based on joint attention mechanism. *Infrared Physics & Technology*, 125:104242.

Yan, H., Zhang, C., and Wu, M. (2022). Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *arXiv preprint arXiv:2201.01615*.

Yang, B., Hu, Y., Liu, X., and Li, J. (2024). Cefusion: An infrared and visible image fusion network based on cross-modal multi-granularity information interaction

and edge guidance. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):17794–17809.

Zhang, H., Tang, L., Xiang, X., Zuo, X., and Ma, J. (2024a). Dispel darkness for better fusion: A controllable visual enhancer based on cross-modal conditional adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26487–26496.

Zhang, H., Zuo, X., Jiang, J., Guo, C., and Ma, J. (2024b). Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26974–26983.

Zhang, X. and Demiris, Y. (2023). Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554.

Zhang, Y., Sidibé, D., Morel, O., and Mériaudeau, F. (2021). Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042.

Zhao, H. and Nie, R. (2021). Dndt: Infrared and visible image fusion via densenet and dual-transformer. In *2021 International Conference on Information Technology and Biomedical Engineering (ICITBE)*, pages 71–75. IEEE.