Registration of Infrared and Visible Images Using Style Transfer-Based Semantic Segmentation

Si-Ting Lin, Chih-Hung Han, Chieh-Ling Lee, Po-Chyi Su*, Feng-Tsun Chien† and Min-Kuan Chang‡
* National Central University, Taoyuan, Taiwan
E-mail: pochyisu@csie.ncu.edu.tw Tel/Fax: +886-3-4227151/+886-3-4222681
† National Yang Ming Chiao Tung University, Hsinchu, Taiwan
‡ National Chung-Hsing University, Taichung,Taiwan

***Abstract*—Infrared and visible image fusion combines complementary information from both modalities to produce a single image that enhances human perception and supports high-level tasks such as semantic segmentation and object detection. Most existing fusion methods assume perfectly aligned infrared and visible image pairs, an assumption that rarely holds in real-world scenarios due to spatial misalignment, frame drops, and differences in resolution or field of view. While recent approaches can handle minor misalignments under equal resolution, more robust alignment strategies are needed for significant discrepancies. Furthermore, current fusion datasets lack semantic and object annotations, limiting the development of task-driven models. To address these challenges, we propose a new method for constructing a fusion dataset enriched with semantic segmentation labels. We apply style transfer to existing annotated datasets to generate synthetic infrared-visible image pairs tailored to target applications. These images are used to retrain segmentation models, creating datasets with needed annotations. For spatial alignment, we estimate scaling and translation using logarithmic polar transformations and the Fourier Transform. To handle temporal misalignment, we integrate spatial alignment with mask matching to maximize object overlap between paired frames. Experimental results demonstrate the effectiveness of our method in aligning infrared and visible image pairs.**

I. Introduction

Due to constraints of filming conditions or hardware limitations, a single-type sensor often captures only limited information, leading to gaps in real-world scene depiction and an incomplete representation of the observed environment. The aim of image fusion is to extract complementary information from multiple images of the same scene and generate a fused image that combines the distinctive features to fulfill various application requirements. Among different image fusion tasks, the fusion of infrared and visible images has attracted significant attention. Visible sensors capture reflected light, providing rich texture details that closely resemble human visual perception. However, in complex environments such as nighttime, obstructions, or smoke, visible images cannot effectively capture critical target objects like pedestrians and vehicles. In contrast, infrared sensors collect thermal radiation information, which allows them to effectively highlight target objects even under extreme environmental interference, compensating for the shortcomings of visible images. For example, in insufficient light conditions, infrared sensors can detect the presence and movement of objects by capturing the heat emitted from human bodies or vehicle surfaces, though they often lack detailed texture. By leveraging the complementary nature of visible and infrared images, integrating the valuable information from both into a fused image contributes to advanced visual tasks such as object detection, object tracking, and semantic segmentation. In recent years, numerous deep-learning-based methods for the fusion of infrared and visible images have been proposed, including techniques like Autoencoders (AE), Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and others. Despite the breakthroughs in algorithmic performance achieved by existing methods, most of them have not fully addressed potential issues encountered in practical applications. In real-world applications, due to the differences in imaging principles between infrared and visible images, various degrees of misalignment often occur, resulting in distortions in the fusion outcomes. Moreover, issues such as frame drops during the capture process or temporal misalignment caused by discrepancies in sensor settings may arise. If the alignment issues are not resolved, even the most advanced fusion methods will struggle to achieve satisfactory results.

To address this problem, recent research such as UMF-CMGR [1] and SuperFusion [2] has proposed aligning the input images before fusion to mitigate the effects of slight shifts and deformations on the fusion outcomes. However, these methods require the input images to have the same resolution, while the actually captured infrared and visible images may have different resolutions and capture range disparities. For example, in Fig. 1, the left image is the visible image having a resolution of 1800×1600, while the center infrared image has a resolution of 640×512. In the visible image, the "STOP" sign on the road is fully visible, while the sign is incomplete in the infrared image. The size of the vehicle in the visible image is also slightly smaller than in the infrared image, indicating that the two sensors capture scenes with different fields of view. For input images with such significant positional discrepancies, even after being adjusted to the same resolution, the effectiveness of existing methods remains limited. The right image in Fig. 1 shows the fused image by SuperFusion, and we can see noticeable artifacts in both the "STOP" marking on the road and the "STOP" sign in the upper right corner.

Most existing image fusion methods focus on improving visual quality and optimizing evaluation metrics, often overlooking the practicality of fusion results for subsequent higher-level visual tasks. Some research has recognized this problem

979-8-3315-7206-8/25/$31.00

Fig. 1. Misaligned infrared and visible images from the RoadScene dataset. From left to right are an visible image, an infrared image, and a fused image by SuperFusion.

and proposed corresponding solutions, such as TarDAL [3] and SeAFusion [4]. These methods introduce loss functions from higher-level visual tasks into the training process of fusion networks, guiding the network to generate images that are beneficial for those tasks. These methods require the fusion network and the higher-level visual task to be trained simultaneously, implying that the training data should include precise annotation. However, in existing infrared and visible image fusion datasets, only a small portion of data includes object annotations or semantic segmentation labels, leading to the fusion network's training being limited to these sparse data instances. Besides, we've noticed that images from different datasets are heavily influenced by the environment and capture conditions, often exhibiting varying structural and textural features. Most existing high-performance fusion models rely on large deep learning architectures. However, in specific applications such as military and surveillance scenarios, models may need to be deployed on edge devices. In such cases, it is crucial to consider lightweight image fusion networks with fewer parameters, aiming to operate effectively in environments with limited computational resources, and ensure real-time processing. Exploring how to achieve infrared and visible image fusion with reduced parameters without compromising performance is a valuable area of research, especially for complex operational environments. This paper proposes a method for creating a semantic segmentation dataset for infrared and visible image fusion. Using existing semantic segmentation datasets, we generate infrared and visible images that match specific scene visuals through style transfer techniques. With the additional semantic segmentation details, we introduce spatial and temporal alignment methods to effectively address issues of significant spatial misalignment and temporal desynchronization issues in input images. To enhance the feasibility of fusion models in real-world scenarios, we propose a low-parameter fusion network, which not only reduces computational requirements but also maintains fusion performance and operational efficiency. The rest of the paper is organized as follows: Sect. II discusses related research and datasets for infrared and visible image fusion. Sect. III presents the proposed methodology, including the establishment of the dataset, image alignment methods, and image fusion methods. Sect. IV shows the results. Finally, Sect. V concludes the paper and outlines future work.

II. Related Work

Common methods for infrared and visible image fusion include AE, CNN, and GAN. These methods typically involve three steps: feature extraction, feature fusion, and image reconstruction. Features are first extracted from both images, then fused into a unified representation, and finally reconstructed into a fused image, enhancing image quality through cross-modal fusion. The image fusion method based on AE typically involve pre-training the Autoencoder on public datasets. The encoder extracts features from the input images, while the decoder reconstructs the input images based on these encoded features. The trained autoencoder can subsequently be applied to tackle two subtasks in image fusion: feature extraction and image reconstruction. However, the feature fusion in this approach usually follows traditional fusion rules and lacks the ability to learn. For example, DenseFuse [5] and RFN-Nest [6] are two AE-based image fusion methods. Both of them train encoders and decoders on the MS-COCO dataset. DenseFuse introduces dense blocks into the encoder, connecting the output of each layer with the outputs of all other layers. This architecture enables the encoding process to capture more useful features from input images. Furthermore, DenseFuse designs two fusion strategies: add and L1-norm, to fuse these features. The fused features are then reconstructed into an image by the decoder. RFN-Nest introduces a Residual Fusion Network (RFN), which replaces traditional fusion methods with a residual structure. Since RFN is a learnable architecture, it effectively enhances the encoder's and decoder's abilities in feature extraction and reconstruction. CNN-based image fusion methods are divided into two different forms. Firstly, there's an end-to-end approach that encompasses feature extraction, feature fusion, and image reconstruction, using carefully designed loss functions and network architectures. The other approach involves using pre-trained CNN models to formulate fusion rules and then performing feature extraction and image reconstruction using conventional methods. For instance, PIAFusion [7] designs an illumination-aware sub-network to estimate the illumination distribution and calculate the illumination probability. Subsequently, it utilizes the illumination probability to construct an illumination-aware loss to guide the training of the fusion network. GAN-based image fusion methods can implicitly accomplish feature extraction, feature fusion, and image reconstruction. These methods mainly rely on two types of loss functions: content loss and adversarial loss. Content loss ensures that the generated image closely resembles the target image in terms of content and structure, meaning that the fused image retains the essential features of the input images. Adversarial loss, on the other hand, enhances the realism of the generated image, helping the generator produce more convincing fused images. However, a single discriminator may cause an imbalance between infrared and visible image information. To address this, DDcGAN [8] proposes using two discriminators to separately assess the structural differences between the fused image and each of the input images. The first discriminator focus on retaining of thermal features, the second

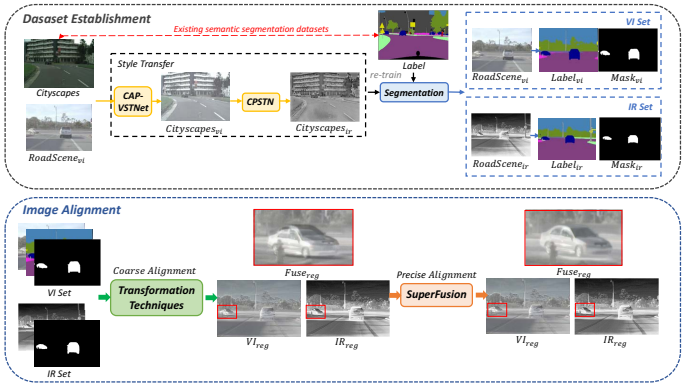

Fig. 2. The flowchart of the proposed method.

emphasizes the detail and texture preservation. This approach balances infrared and visible characteristics, ensuring effective feature preservation in the fused image. While existing fusion methods can achieve excellent results, a critical requirement for effective image fusion in practical applications is ensuring proper image alignment. Misaligned images often lead to artifacts, impacting the final fusion quality. However, due to significant differences between visible and infrared features, traditional image feature extraction methods are not suitable. To meet the requirements of image alignment, UMF-CMGR [1] and SuperFusion [2] propose methods for image alignment before fusion to ensure the quality of the final fused image. The quantity and quality of datasets are also crucial for model training. Several public datasets provide valuable resources for researchers and cover a variety of scenarios. Common datasets for infrared and visible image fusion include M3FD [3], MSRS [7], TNO [9], RoadScene [10], and LLVIP [11].

III. Proposed Method

Our framework is shown in Fig. 2, including three main parts: dataset establishment, image alignment and fusion.

A. Dataset Establishment

Since there are significant differences in both infrared and visible images in different fusion datasets, we apply style transfer to images from an existing semantic segmentation dataset Cityscapes [12], transforming their visual style to match that of the target fusion dataset. The images of Cityscapes dataset are widely used for developing visual deep models for tasks such as semantic segmentation, object detection, and instance segmentation in urban scenes. For visible images, we used an existing style transfer model to generate stylized images $Cityscapes_{vi}$. Although Cityscapes images are also visible images, using style transfer to create images that more closely resemble those of the target dataset is still a crucial step for subsequent processing. Training a style transfer model from scratch is unnecessary because existing general models already achieve the desired results. Training a custom model might instead lead to concerns about insufficient training images for the specific target scenes. For infrared images, we convert the stylized visible images $Cityscapes_{vi}$ into pseudo-infrared images $Cityscapes$ using a cross-modal model. Finally, we retrain the segmentation model using $Cityscapes_{vi}$,$Cityscapes_{ir}$ and the segmentation annotations from Cityscapes to meet the specific requirements. For visible images, we selected the Content Affinity Preserved Versatile Style Transfer network (CAP-VSTNet [13]), which combines a reversible residual network with an unbiased linear transformation module and introduces a Matting Laplacian training loss. The Matting Laplacian, based on the Laplacian matrix, utilizes image edge information to maintain the clarity of the foreground and background boundaries through an appropriate loss function, enhancing the quality of the generated images. The reversible residual network in CAP-VSTNet can transfer the details of the original image to the output image, preventing information loss.

For infrared images, we adopted the Cross-modality Perceptual Style Transfer Network (CPSTN) in UMF-CMGR. This approach utilizes a style transfer network to convert visible images into corresponding pseudo-infrared images. The goal is to reduce the cross-modality differences between visible and infrared images, facilitating the subsequent image alignment steps. Fig. 3 illustrates the processing methods for visible and infrared images using different style transfer networks in this study. Given the feature differences across various datasets and their potential impact on subsequent alignment and fusion results, we train the networks individually for each corresponding dataset. This approach allows the network to better understand and reproduce the inherent characteristics of each dataset, thereby improving the quality and realism of the generated images. Fig. 3 uses the RoadScene dataset as an example, treating it as the target for alignment processing. We first use the pre-trained CAP-VSTNet style transfer model, taking the Cityscapes images with segmentation labels as the content image and the target RoadScene visible images as the style image. CAP-VSTNet extracts content features from the content image and style features from the style image, generating the stylized image $Cityscapes_{vi}$. This step not only enhances the subsequent model's segmentation performance on visible images but also provides more suitable inputs for the CPSTN in the next step. The stylized image $Cityscapes_{vi}$ is closer to the characteristics of the RoadScene visible images, allowing CPSTN to more effectively apply style transfer and improve the performance. For the infrared part, we retrain the CPSTN model to adapt to the RoadScene dataset. Specifically, we used the visible and infrared image pairs from the RoadScene dataset to generate the stylized pseudo-infrared image $Cityscapes_{ir}$ from $Cityscapes_{vi}$. We generate a set of visible and infrared images with RoadScene-style characteristics from the original semantic segmentation dataset, Cityscapes. We then use InternImage [14] as the semantic segmentation network and retrain it using the Cityscapes style-transferred visible and pseudo-infrared images generated from different fusion dataset styles. By utilizing the stylized image pairs we generated, the segmentation network can adapt to

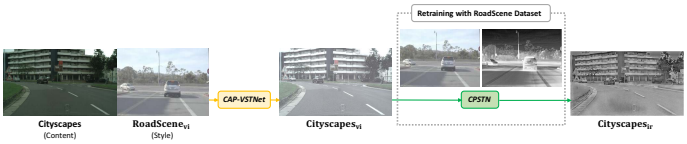

Fig. 3. Style transfer processing in this study.

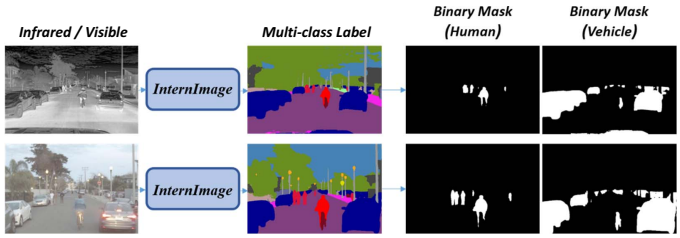

Fig. 4. Semantic segmentation processing.

the characteristics of visible and infrared images in the target dataset, thereby creating more representative training data and improving the accuracy of target image classification. After performing segmentation on the two types of images, we also convert the multi-class semantic segmentation labels into two binary masks, "person" and "car," as shown in Fig. 4.

B. Image Alignment

The image alignment method based on semantic segmentation labels is outlined as follows. The steps are shown in Fig. 5. We choose to use semantic segmentation labels or only important object masks (such as pedestrians or vehicles) for alignment based on whether the background contains common label categories.

1) Scaling and Translation Alignment Based on Classical Image Processing Techniques: Due to the potential differences in Field of View (FOV) between the two input images, we attempt to calculate the scaling factor and translation required for aligning the two images. To determine the scaling factor, we used the Log-Polar Mapping [15] to convert the Cartesian coordinates into logarithmic polar coordinates. Due to the significant differences in the imaging principles of infrared and visible images, direct comparison using the original images may be challenging. Our image fusion dataset with semantic segmentation labels effectively addresses this issue. The original images, processed with semantic segmentation to generate labels and masks, simplify the image content and focus on key information. Each object in the semantic segmentation label maps is marked with a corresponding color, accurately reflecting the shape and boundaries of the objects. Using these labeled maps in the Log-polar Mapping and phase correction processes is feasible and helps avoid alignment errors caused by content differences. Additionally, since semantic segmentation is widely used in applications such as autonomous driving and smart city surveillance, the model is primarily trained on datasets that include common elements like pedestrians, vehicles, vegetation, and sidewalks. However,

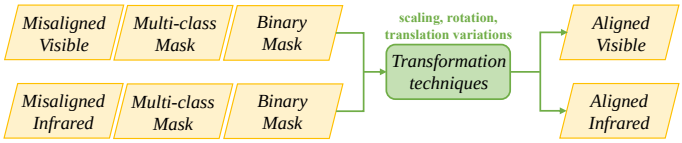

Fig. 5. Flowchart of alignment method based on semantic segmentation.

some scenes may lack common segmentation categories in the background. The background of the image primarily composed of walls and windows, which makes it difficult for the semantic segmentation model to accurately classify them. In such cases, we choose to use masks to ensure alignment accuracy and reduce background interference. This approach helps achieve acceptable alignment results even when the background lacks semantic segmentation information, compensating for the deficiencies in annotations under specific circumstances. Convert the semantic segmentation labels or masks generated from the infrared and visible images to log-polar coordinates, and then use phase correlation to calculate the scaling factor between the images. Phase correlation is a Fourier-transform-based technique that determines the displacement between two images by calculating the phase difference in the frequency domain. This method effectively removes scale differences between the images, allowing them to be aligned on a unified scale. It is noteworthy that when the scale factor between the images to be aligned is too large, the accuracy of this method may be insufficient, or it may require using a larger coordinate grid. In our implementation, we first compare the sizes of the objects in the masks of the two images. After calculating their differences, we crop the image with the larger field of view to ensure accurate scale factor computation. After calculating the scale factor and resizing the original images to the same size, we used the segmentation labels or masks, which are generated through semantic segmentation, to compute the image translation using Fourier transform. Since the label maps are unaffected by variations in brightness or differences in imaging principles, they ensure alignment accuracy. Fourier transform converts images from the spatial domain to the frequency domain. In the frequency domain, phase changes reveal the translation between images. This method allows us to determine the displacement between two segmented images.

2) Alignment of Minor Displacements and Deformations Based on Deep Learning: We used Log-polar Mapping and Fourier transform to eliminate significant differences between the two input images, aligning the infrared and visible images on a unified scale. Frequency domain alignment methods effectively handle changes in scaling, rotation, and displacement in images. For further correction of minor displacements or deformations in the images, deep learning techniques can be employed for local adjustments. SuperFusion captures and corrects these subtle displacements and deformations by performing feature extraction and matching at different scales, achieving more precise local alignment of images.

Regarding temporal alignment, we first select the frames

from the visible dataset that contain the largest object regions using object masks. This selection helps ensure that important features in the images are fully utilized, reducing alignment errors caused by a lack of object content. We select this frame as the reference image n. Next, we align the images from the infrared dataset within the specified range with this reference image in the spatial domain and calculate the IoU of the objects. A visible image is aligned with five infrared images in the spatial domain. The number of infrared images to compare with the reference image depends on the number of images in each dataset. After the comparison, the infrared image $n+1$ with the maximum object overlap is selected, and its geometric transformation parameters are used as the fixed adjustment values for this dataset. Subsequently, all images will be aligned in the spatial domain according to these parameters.

C. Image Fusion

This study modifies the DenseFuse [5] fusion network architecture. DenseFuse employs a dense block structure in the encoder, where each layer's output is connected to outputs of all previous layers, effectively enhancing the capability to extract features from input images. The decoder leverages the features to reconstruct the final fused image. By adjusting the number of convolutional filters in the feature extraction and feature reconstruction sub-networks, we reduced the model's parameters from $147,985$ to $9,417$. The network architecture is shown in Fig. 6. We improved the fusion strategy for infrared and visible images by replacing the original L1-norm and addition fusion strategy with one that uses the ReLU activation function. This change is primarily motivated by its suitability for real-time image processing on edge computing devices. Subsequent experiments have confirmed that even with reduced computational complexity, the model maintains a certain level of performance in image fusion. We used the fusion loss and gradient loss as the loss functions for the fusion network, i.e.,

$$L_{total} = L_{int} + \alpha L_{grad}. \tag{1}$$

The fusion loss combines L1 loss and structural similarity loss (SSIM). To enable the fusion network to better integrate meaningful information, this paper designs a fusion loss function L_{int}. The specific definition of this loss function is as follows:

$$\begin{aligned} L_{int} = &\frac{1}{2}(L_{SSIM}(fusion_i, vi_i) + \beta\|fusion_i - vi_i\|_1) \\ &+ \frac{1}{2}(L_{SSIM}(fusion_i, ir_i) + \beta\|fusion_i - ir_i\|_1) \end{aligned} \tag{2}$$

The L1 loss measures the absolute differences between the fused image and the visible and infrared images, while the structural similarity loss L_{SSIM} is used to preserve the structural similarity of the images. The constant β is used to balance the L1 loss and the structural similarity loss. To emphasize the edges and details in the image, we introduce the gradient loss L_{grad}. This loss function calculates the L1 loss between the gradients of the fused image and the maximum gradient image, which is defined as follows:

$$L_{grad} = \|fusion_{grad} - max(vi_{grad}, ir_{grad})\|_1, \tag{3}$$

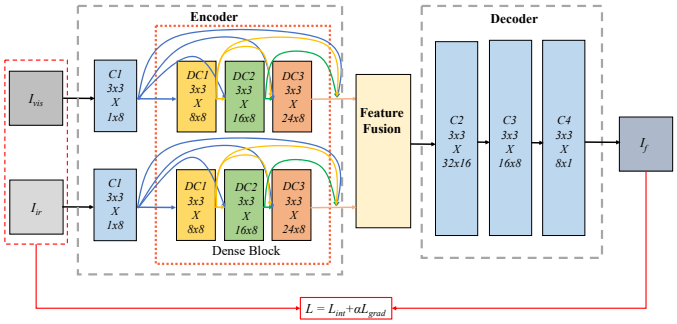

Fig. 6. The architecture for the low-parameter fusion design.

where $fusion_{grad}$, vi_{grad} and ir_{grad} represent the gradients of the fused image, visible image, and infrared image, respectively, extracted using the Sobel operator. The Sobel operation is a commonly used image processing technique for calculating image gradients, which helps to emphasize the edges and details within the image.

IV. Experimental Results

This paper utilized Python 3.7, PyTorch 1.10.1 and torchvision 0.11.2 as the deep learning framework. The operating system was Ubuntu 18.04 LTS. In terms of hardware configuration, the CPU was an Intel Core i9-12900K at 3.2 GHz with 64GB of memory, and the GPU was an NVIDIA GeForce RTX 3090. The versions of CUDA and cuDNN were 11.3 and 8.1, respectively. The RoadScene and FLIR-ADAS-v2 datasets were employed for experimental evaluation. The RoadScene dataset comprises 221 pairs of unaligned and aligned infrared and visible images. The FLIR-ADAS-v2 dataset includes over ten thousand infrared and visible images, though the number of images for each modality differs. It provides several sequences of continuous images suitable for alignment testing. The pretrained CAP-VSTNet model was employed to perform the style transfer on existing semantic segmentation datasets. CP-STN was retrained for each dataset to ensure that the generated pseudo-infrared images closely resembled the original infrared images in that dataset. We also retrained InternImage using the stylized visible iamges and the pseudo-infrared images to improve the segmentation performance of the semantic segmentation model across different datasets.

To compare SuperFusion with the proposed methods, we conducted evaluations using the RoadScene dataset. In Fig. 7, the SuperFusion result shows noticeable artifacts in the front vehicle, and significant distortions in the rear vehicle and the road surface. In contrast, our method demonstrates much better aligned content. By further combining the spatial domain alignment and the mask comparison approach, we identify the corresponding images with the most object overlap, as shown in Fig. 8. The top row of the figure displays a continuous image sequence where the contents are not aligned spatially and temporally. After processing with the proposed method, the bottom row shows that the vehicles are effectively aligned

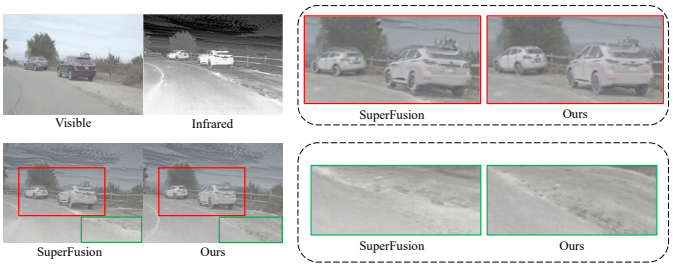

Fig. 7. Image alignment results - RoadScene.

Fig. 8. Image alignment results in the spatial and temporal domain.

in the images without noticeable artifacts. Moreover, while SuperFusion can improve the accuracy of aligned images in a single frame, we found that using SuperFusion for local adjustments on each frame may cause distortion in the background due to local adjustments, making it difficult to maintain consistent backgrounds across the sequence of images.

We compared our method with recent fusion approaches–TarDAL[3], MFEIF[16] and DenseFuse(DF) [5] on the LLVIP dataset. We evaluated using nine commonly used metrics in image fusion tasks: Entropy(EN), Mutual Information(MI), Spatial Frequency(SF), Visual Information Fidelity(VIF), Average Gradient(AG), Sum of the Correlations of Differences(SCD), and Edge-based similarity measure (Q). Table I shows that the proposed light-weight model demonstrates promising results.

V. Conclusion

This paper proposed a method for creating semantic segmentation datasets for infrared and visible image fusion using style transfer and existing datasets, eliminating the need for additional manual labeling. We introduced a semantic segmentation-based image alignment approach that improves label accuracy and efficiently resolves spatial and temporal alignment issues, reducing preprocessing time and resources. We also presented a low-parameter design for infrared and visible fusion that enhances performance while reducing network parameters. Since the image alignment methods rely on semantic segmentation and could be affected by challenging conditions such as strong lighting or smoke, future work could explore more effective data augmentation to overcome these limitations.

TABLE I
Evaluation of fusion on LLVIP [11]

Method	Metric							
	Para.	EN	MI	SF	AG	Q	SCD	VIF
TarDAL	296K	6.27	2.75	0.039	2.98	0.28	1.04	0.79
MFEIF	158K	7.03	3.31	0.054	3.95	0.53	1.56	0.81
DF	74K	6.05	2.48	0.045	3.29	0.38	1.38	0.66
Ours	**9K**	**7.25**	**3.44**	**0.076**	**5.39**	**0.61**	**1.57**	**0.82**

Acknowledgment

This research is supported by National Science and Technology Council, Taiwan, under the grant sNSTC 113-2221-E-008-093 and 114-2221-E-008-028.

References

[1] D Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration," *in arXiv preprint, arXiv:2205.11876*, 2022.

[2] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, 2022.

[3] J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," *in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5811, 2022.

[4] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28-42, 2022.

[5] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[6] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72-86, 2021.

[7] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Information Fusion*, vol. 83, pp. 79-92, 2022.

[8] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4980-4995, 2020.

[9] A. Toet, "The TNO multiband image data collection," *Data in brief*, vol. 15, pp. 249-251, 2017.

[10] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502-518, 2020.

[11] 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*, pp. 3496-3504, 2021.

[12] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," *in Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213-3223, 2016.

[13] L. Wen, C. Gao, and C. Zou, "CAP-VSTNet: content affinity preserved versatile style transfer," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18300-18309, 2023.

[14] W. Wang et al., "Internimage: Exploring large-scale vision foundation models with deformable convolutions," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14408-14419, 2023.

[15] J. N. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration using log-polar transform and phase correlation," *in TENCON, IEEE region 10 conference*, pp. 1-5, 2009.

[16] J. Liu, X. Fan, J. Jiang, R. Liu and Z. Luo, "Learning a Deep Multi-Scale Feature Ensemble and an Edge-Attention Guidance for Image Fusion," in IEEE Trans. on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 105-119, Jan. 2022.