Abstract—Infrared and visible image fusion plays a vital role in the field of computer vision. Previous approaches make efforts to design various fusion rules in the loss functions. However, these experimental designed fusion rules make the methods more and more complex. Besides, most of them only focus on boosting the visual effects, thus showing unsatisfactory performance for the follow-up high-level vision tasks. To address these challenges, in this letter, we develop a semantic-level fusion network to sufficiently utilize the semantic guidance, emancipating the experimental designed fusion rules. In addition, to achieve a better semantic understanding of the feature fusion process, a fusion block based on the transformer is presented in a multi-scale manner. Moreover, we devise a regularization loss function, together with a training strategy, to fully use semantic guidance from the high-level vision tasks. Compared with state-of-the-art methods, our method does not depend on the hand-crafted fusion loss function. Still, it achieves superior performance on visual quality along with the follow-up high-level vision tasks.

Index Terms—Image fusion, semantic-level fusion network, semantic-driven fusion training strategy.

I. INTRODUCTION

Infrared and visible image fusion (IVIF), referring to providing the typical characteristics from source images, has witnessed rapid development in recent years. IVIF can effectively break the limitation of information loss from a single sensor and plays an important role for the follow-up high-level vision tasks, e.g., object detection [1], semantic segmentation [2] and so on. Although dual-branch feature-level fusion methods [3], [4] have achieved promising performance, pixel-level IVIF methods can serve as versatile preprocessing to facilitate various high-level vision tasks and provide a visually pleasing result for human vision.

Manuscript received 22 November 2022; revised 15 February 2023; accepted 5 April 2023. Date of publication 13 April 2023; date of current version 19 April 2023. This work was supported in part by the National Key R&D Program of China under Grant 2020YFB1313503, in part by the National Natural Science Foundation of China under Grant 1222B0252, and in part by the Fundamental Research Funds for the Central Universities and the Major Key Project of PCL under Grant PCL2021A12. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yuichi Tanaka. (Corresponding author: Risheng Liu.)

Yuhui Wu and Zhu Liu are with the School of Software Technology, Dalian University of Technology, Dalian 116024, China (e-mail: wuyuhui63@hotmail.com; liuzhu_dlut@mail.dlut.edu.cn).

Jinyuan Liu is with the School of Mechanical of Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: atlantis918@hotmail.com).

Xin Fan and Risheng Liu are with the DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Dalian 116024, China, and also with Peng Cheng Laboratory, Shenzhen 518052, China (e-mail: xin.fan@dlut.edu.cn; rlshu@dlut.edu.cn).

Digital Object Identifier 10.1109/LSP.2023.3269880

In the past, traditional fusion methods prevailed, including [5], [6], [7], [8], [9], [10], [11], [12]. Recently, deep learning-based methods have shown great potential in various image-processing fields [13], [14], [15] by their powerful nonlinear feature extraction capabilities. Learning-based IVIF approaches achieve promising performance based on the design of various fusion rules and loss functions. We can roughly divide current learning schemes into two categories: fusion-rule based methods [16], [17], [18], [19] and end-to-end learning schemes [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30].

In specific, the first kind of fusion method relies on manually designed rules to aggregate the modal feature approximately. These methods first utilize the auto-encoder mechanisms to extract and reconstruct multimodal features to sufficiently learn the significant feature extraction. Then they develop various fusion rules for feature fusion, e.g., weighted average, summation, maximum selection, and \( \ell_1 \) norm. For instance, Li et al. [16] pioneer the dense blocks as the learnable auto-encoder and utilized the weighted-average strategy to fuse modal features. After that, Li et al. [17] also provide spatial/channel attention mechanisms as the fusion strategies to fuse features with nested connections. Subsequently, Liu et al. [31] introduce the edge attention-guided auto-encoder to extract the feature and adopts the simple fusion rules.

However, these fusion strategies are not sensitive to diverse data distribution and are easy to induce visual artifacts and blurs. More importantly, the manual design of fusion strategies is too fragile to preserve suitable modal characteristics for supporting follow-up high-level vision tasks.

Instead of manually designed fusion rules, end-to-end learning methods are proposed to establish the connection between source and fused images directly. Specifically, architectures and loss functions are two challenging stumbling stones for these methods. Existing methods concentrate on designing architectures based on the current effective practices, rather than considering the particular properties of fusion tasks. For example, dense blocks [21] and residual blocks [32] are widely utilized for IVIF. Besides, diverse loss functions are proposed to enforce different principles for IVIF. Typically, Ma et al. [24] introduce the dual generative adversarial criterion to push the generated images as similar to the source images. Xu et al. [21] propose the feature-level measurement to endow the richness of information into fused images.

Though visually appealing results which can acquire remarkable statistical metrics are obtained. These loss functions coupled with training strategies make the methods more complex. Furthermore, these architectures cannot effectively extract the modal characteristics, limited by the local awareness of convolution networks. We argue that both categories of methods neglect the requirement of follow-up semantic tasks.
A multi-scale self-attention-based image fusion network is proposed to address the limitations of existing methods. Liu et al. [20] calculate saliency masks to construct fusion loss, neglecting that the visual saliency cannot reflect semantic richness. Tang et al. [33] creatively utilize semantic information to promote both registration and fusion. Despite the promising results, the intensity maximization loss limits the performance on extreme scenes.

To partially mitigate these issues, rather than explicitly constructing fusion rules, we propose a generic semantic-driven learning paradigm to investigate task-specific image fusion. To be concrete, we first propose a multi-scale fusion network with a self-attention mechanism to sufficiently aggregate the modal features. Multi-scale extraction can effectively combine features in a coarse-to-fine manner from the scene structure to context details. Self-attention mechanism is to establish the long-range dependency of multi-modal features, and better depict the global representation of salient targets. Then we introduce a correlated regularization to describe the relationships between source images and fused images. Based on this, we only utilize the criterion of high-level vision tasks to train both the fusion and high-level networks. Thus, this strategy emancipates the experimental design of fusion rules, discards the restriction of modal statistic metrics, and drastically improves the performance of high-level vision tasks. We summarize the core contribution as follows:

- Imposing a correlated regularization, a fully semantic-driven training strategy is introduced to break free from handcrafted fusion rules.
- A multi-scale self-attention-based image fusion network is proposed to effectively represent the global structures in a coarse-to-fine manner.

II. THE PROPOSED METHOD

A. Network Architecture

While some previous works focus on pursuing high efficiency at the expense of the fitting ability of the fusion network, it does not work well to utilize the semantic guidance from the follow-up high-level vision tasks. In contrast, we hypothesize that the fusion network should be capable of semantic understanding to achieve flexible fusion effects among various semantic classes.

To this end, we employ the multi-scale mechanism from [17] to deal with the textual details and semantic information, respectively. As shown in Fig. 1, we use the downsampling (i.e. max-pooling) operators to obtain feature maps of different resolutions. Among them, the shallow layer feature maps contain more textures information, while the deep layer feature maps contain more semantic information. As for segmentation, we directly adopt SegFormer [34].

On the other hand, to fuse the extracted cross-modal feature maps, we devise a generic fusion block based on efficient self-attention [35]. As shown in the bottom of Fig. 1, the fusion block consists of two self-attention modules, which can capture and reinforce the useful components in the global receptive field. In the self-attention module, we reshape the feature maps of $R^{H \times W \times C}$ to vectors of $R^{N \times C}$, where $N = H \times W$. Then, we use the linear layers to encode the vectors into query $Q$, key $K$, and value $V$. We get the attention map by a matrix multiplication $K^TV$, then we acquire the final attention result $Q_{\text{Softmax}}(K^TV)$. The strengthened component is obtained by element-wise multiplying the attention result and the input feature maps. To preserve the detailed information, we further introduce a residual connection.

B. Training Strategy

Existing end-to-end deep-learning methods focus on devising fusion rules to acquire visually appealing results. Unfortunately, hand-crafted fusion rules are heavily limited to the scene and cannot meet the essential requirement of follow-up semantic tasks. To address this issue, we develop a semantic-driven training strategy to emancipate the manual design.

1) Warm-Start Phase: Jointly training both the fusion and segmentation networks is an intuitive strategy. However, at the beginning of training, the parameters of the fusion model are randomly initialized and thus cannot provide meaningful fused images for the segmentation network to handle. Consequently, the training process deviates from our expectations. To address this issue, we use an average strategy to pre-train the fusion model to obtain a malleable initialization. This learning procedure can be formulated as

$$
\min_{\theta} \mathcal{L}_{WS} (\mathcal{N}_F (I_{vis}, I_{ir}; \theta)),
$$

where $\mathcal{N}_F$ is the fusion network with learnable parameters $\theta$, $I_{vis}$ and $I_{vis}$ denote infrared image and visible image, respectively. After this phase, we obtain $\theta'$, which can fuse the source images into substantially average and meaningful results for the next training procedure.

2) Semantic Training Phase: In this phase, we fine-tune the fusion network by jointly training with the segmentation network, which can be formulated as

$$
\min_{\theta', \omega} \mathcal{L}_{ST} (\mathcal{N}_S (\mathcal{N}_F (I_{vis}, I_{ir}; \theta'); \omega)),
$$

where $\mathcal{N}_S$ is the semantic segmentation network with learnable parameters $\omega$. The semantic segmentation task will learn to adjust the proportion of infrared and visible components away from the average fusion state.

It is instructive to note that the fusion model without additional constraints is unstable, resulting in degraded segmentation performance. To mitigate this, we design an auxiliary regularization loss function to constrain the fusion model for utilizing far-reaching semantic guidance.
C. Loss Function

1) Warm-Start Loss Function: The warm-start fusion loss function can be formulated as:
\[ L_{WS} = \frac{1}{HW} \| I_{f} - \frac{I_{ir} + I_{vis}}{2} \|_1, \]
where \( H \) and \( W \) denote the height and width of the source image, respectively. \( I_{f} \) is the fused image. \( \| \cdot \|_1 \) represents the calculation of the \( \ell_1 \) norm.

2) Semantic Training Loss Function: During the semantic training phase, we use the semantic training loss function as follows:
\[ L_{ST} = L_{sem} + \lambda L_{reg}, \]
where the \( L_{sem} \) is a commonly used cross-entropy loss function, while the \( L_{reg} \) is a regularization term. The \( \lambda \) is a hyper-parameter to strike a balance between the two terms. We define the regularization loss function as:
\[ L_{reg} = \frac{1}{Corr(I_{ir}, I_{f}) + Corr(I_{vis}, I_{f})}, \]
where \( Corr(\cdot) \) denotes the calculation of the correlation of two image tensors.

III. EXPERIMENTS

In this section, we first conduct qualitative and quantitative comparisons on three datasets: MFNet [2], TNO [36], and RoadScene [21]. The competitors include MST-SR [37], DenseFuse [16], RFN-Nest [17], SMoA [18], GANMcC [38], U2Fusion [21], MFEIF [31], and SeAFusion [32]. We also conducted ablation studies to validate the proposed architecture and semantic-driven training strategy.

A. Fusion Results

As shown in Fig. 2, we can see that our method can flexibly preserve abundant and useful textural details, successfully highlighting the important targets in diverse harsh environments. For instance, persons in our method have sharper edges, and thus stand out from the background, which will be helpful for follow-up segmentation. As for quantitative comparisons, our method outperformed others in terms of commonly used statistical evaluation metrics spatial frequency (SF) [39] and average gradient (AG) [40], as shown in Fig. 3. This indicates our results have more information richness, containing more details and high contrast.

B. Segmentation Results

We also evaluate the fusion quality from the perspective of segmentation. We trained the same segmentation network (SegFormer-b0) from scratch with all comparative methods. As shown in Fig. 4, the segmentation model with our method can provide more accurate results, e.g., bikes. In contrast, other methods cannot estimate the shapes of the car, interfered with by the strong glare. As shown in Table I, our method got the top scores on mAcc and mIoU, indicating our method can intelligently reserve useful information for different classes.
C. Ablation Studies

1) Analyzing the Fusion Module: We conducted experiments including removing the self-attention module (w/o SLA), replacing it with channel attention (CHA) [41], and spatial attention (SPA) [42]. As shown in Fig. 5, persons of our fused result are more natural and conspicuous, which is also demonstrated by segmentation results, as reported in Table II.

2) Analyzing the Warm-Start and Regularization Loss: We further conducted a set of experiments including removing the warm-start phase (w/o WS) and removing the regularization term (w/o $L_{reg}$). As shown in Table II, the semantic segmentation performance of “w/o WS” decreases significantly. The blurred visual effects of “w/o WS” also validate the necessity of the warm-start phase, as shown in Fig. 6. “w/o $L_{reg}$” leads to slight performance degradation as shown in Table II. But the difference in visual effects is hard to recognize in Fig. 6. We also provide an intensity maximization warm-start strategy, i.e., $L_{WS} = \frac{1}{|W|} [I_f - \max(I_{ir}, I_{vis})]$, where $\max(\cdot)$ is denoted as element-wise maximum selection. The fused results are denoted as “Max-WS”. The corresponding results of the proposed scheme are denoted as “Ave-WS”. It is hard to judge which is better from Fig. 6. But from Table II, we can see that the maximum selection fusion rule provides better initialization to preserve those classes that contain larger intensity, e.g., pedestrians and cars. But it fails to deal with other classes, such as car stops. It reveals that the average warm-start strategy is a more malleable choice.

3) Analyzing the Effect of Semantic Loss: We conducted experiments removing some classes, e.g., the car class (w/o Car), the person class (w/o Person), and both (w/o Car&Person). We also conducted an experiment removing the semantic loss (w/o $L_{sem}$). As shown in Fig. 7, the semantic loss makes the pedestrians and the cars stand out from the background as shown in (c). The impact of the person class is more significant than the car class. It is worth noting that (b) is substantially the same as (d), reflecting that the other classes (apart from Car and Person) have a minor influence. This class-imbalance issue implies the full utilization of semantic loss, while other semantic-driven methods heavily depend on fusion loss.

IV. Conclusion

In this letter, to break free from manually designing fusion rules, we develop a semantic-level fusion network with an adaptive semantic-driven training strategy to take full advantage of the guidance from the follow-up semantic tasks. Experimental results reveal that semantic loss can not only replace the manual design of fusion rules but also provide a flexible and robust semantic-level fusion that satisfies both human vision and high-level vision tasks.
REFERENCES

[1] T. T. Zin, H. Takahashi, T. Toriu, and H. Hama, “Fusion of infrared and visible images for robust person detection,” Image Fusion, pp. 239–264, 2011.

[2] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, “MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes,” in Proc. IEEE/RJS Int. Conf. Intell. Robots Syst., 2017, pp. 5108–5115.

[3] Y. Sun, W. Zuo, and M. Liu, “RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes,” IEEE Robot. Automat. Lett., vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[4] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, “Revisiting feature fusion for RGB-T salient object detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 5, pp. 1804–1818, May 2021.

[5] Y. Liu, X. Chen, R. K. Ward, and J. Wang, “Image fusion with convolutional sparse representation,” IEEE Signal Process. Lett., vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[6] H. Li and X.-J. Wu, “Densefuse: A fusion approach to infrared and visible image fusion,” Proc. 3rd Int. Conf. Inf. Fusion, 2021.

[7] H. Li and X.-J. Wu, “Overwater image color correction using exposure-bracketing imaging,” IEEE Signal Process. Lett., vol. 25, no. 6, pp. 893–897, Jun. 2018.

[8] H. Li and X.-J. Wu, “Densefuse: A fusion approach to infrared and visible image fusion,” IEEE Signal Process. Lett., vol. 26, no. 3, pp. 485–489, Mar. 2019.

[9] R. Gao, S. A. Vorobyov, and H. Zhao, “Image fusion with cosparse analysis operator,” IEEE Signal Process. Lett., vol. 24, no. 7, pp. 943–947, Jul. 2017.

[10] R. Gao, S. A. Vorobyov, and H. Zhao, “Image fusion with cosparse analysis operator,” IEEE Signal Process. Lett., vol. 24, no. 7, pp. 943–947, Jul. 2017.

[11] J. Ma, C. Chen, C. Li, and J. Huang, “Infrared and visible image fusion via multi-objective optimization,” IEEE Trans. Image Process., vol. 25, no. 12, pp. 885–889, Dec. 2018.

[12] J. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, “Toward all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark,” in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 3783–3791.

[13] L. Komura, D. Sugimura, and T. Hamamoto, “Underwater image color correction using exposure-bracketing imaging,” IEEE Signal Process. Lett., vol. 25, no. 6, pp. 893–897, Jun. 2018.

[14] L. Komura, D. Sugimura, and T. Hamamoto, “Underwater image color correction using exposure-bracketing imaging,” IEEE Signal Process. Lett., vol. 25, no. 6, pp. 893–897, Jun. 2018.

[15] 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,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 1, pp. 105–119, Jan. 2022.

[16] J. Liu, X. Fan, and R. Liu, “Unsupervised misaligned infrared and visible image fusion via cross-modality image registration and generation,” in Proc. Int. Joint Conf. Artif. Intell., 2022.

[17] J. Liu, X. Fan, and R. Liu, “Unsupervised misaligned infrared and visible image fusion via cross-modality image registration and generation,” in Proc. Int. Joint Conf. Artif. Intell., 2022.

[18] N. G. Zhang, Y. Xiao, X. Guo, and J. Ma, “Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity,” in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 12797–12804.

[19] J. Liu, Y. Deng, Y. Ma, J. Huang, and J. Ma, “Superfusion: A versatile image registration and fusion network with semantic awareness,” IEEE J. Automat. Sci. Syst., vol. 9, no. 12, pp. 2112–2117, Dec. 2022.

[20] E. Xie, W. Z. Wang, A. Y. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” in Proc. Adv. Neural Inf. Process. Syst., 2021, vol. 34, pp. 12077–12090.

[21] Z. Shen, M. Zhang, H. Zhao, S. Y. Li, and H. Li, “Efficient attention: Attention with linear complexities,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2021, pp. 3531–3539.

[22] A. Toet, “TNO image fusion dataset,” Apr. 2014. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

[23] Y. Liu, S. Liu, and Z. Wang, “A general framework for image fusion based on multi-scale transform and sparse representation,” Inf. Fusion, vol. 24, pp. 147–164, 2015.

[24] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, “GANMcA: A generative adversarial network with multiclassification constraints for infrared and visible image fusion,” IEEE Trans. Instrum. Meas., vol. 70, no. 12, pp. 1–14, Dec. 2021.

[25] A. M. Eskicioglu and P. S. Fisher, “Image quality measures and their performance,” IEEE Trans. Commun., vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[26] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, “Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition,” Opt. Commun., vol. 341, pp. 199–209, 2015.

[27] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.

[28] L. Chen et al., “SCA-CNN: Spatial and channel-wise in convolutional networks for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6298–6306.