MaeFuse: Transferring Omni Features with Pretrained Masked Autoencoders for Infrared and Visible Image Fusion via Guided Training

Jiayang Li[1] , **Junjun Jiang**[1*] , **Pengwei Liang**[1] , **Jiayi Ma**[2]
[1]Harbin Institute of Technology
[2]Wuhan University
jiangjunjun@hit.edu.cn

Abstract

In this research, we introduce MaeFuse, a novel autoencoder model designed for infrared and visible image fusion (IVIF). The existing approaches for image fusion often rely on training combined with downstream tasks to obtain high-level visual information, which is effective in emphasizing target objects and delivering impressive results in visual quality and task-specific applications. MaeFuse, however, deviates from the norm. Instead of being driven by downstream tasks, our model utilizes a pretrained encoder from Masked Autoencoders (MAE), which facilities the omni features extraction for low-level reconstruction and high-level vision tasks, to obtain perception friendly features with a low cost. In order to eliminate the domain gap of different modal features and the block effect caused by the MAE encoder, we further develop a guided training strategy. This strategy is meticulously crafted to ensure that the fusion layer seamlessly adjusts to the feature space of the encoder, gradually enhancing the fusion effect. It facilitates the comprehensive integration of feature vectors from both infrared and visible modalities, preserving the rich details inherent in each. MaeFuse not only introduces a novel perspective in the realm of fusion techniques but also stands out with impressive performance across various public datasets.

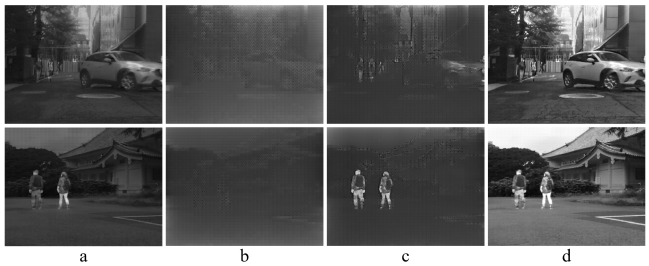
1 Introduction

Multimodal sensing technology has significantly contributed to the widespread use of multimodal imaging in various fields, and the fusion of infrared and visible modalities is the most common application technique. Infrared and visible image fusion (IVIF) can merge the complementary information of these two modalities. Infrared images display thermal contours and are unaffected by lighting effects, but lack texture and color accuracy. Visible images can capture texture but depend on lighting. Fused images can combine their strengths, and thus enhancing visual clarity and the effectiveness of downstream visual tasks [Sun *et al.*, 2019; Zhang *et al.*, 2020] and surpassing the results achievable by each modality when used separately.

To effectively fuse the useful elements of both modalities, diverse fusion techniques and training approaches have been developed. In end-to-end networks, there are implementations utilizing residual connections [Li *et al.*, 2021] and adversarial learning techniques [Ma *et al.*, 2019; Ma *et al.*, 2020]. In autoencoder networks, some directly employ pretrained convolutional encoders [Li and Wu, 2018], others integrate fusion in different frequency domains [Zhao *et al.*, 2021], and there are also methods combining transformers with convolutional architectures [Zhao *et al.*, 2023b]. These methods have preserved visual texture information to some extent. However, the resulting fused images may struggle to retain high-level semantic visual information adequately, which can lead to the obscurity of high-level visual targets and potentially suboptimal performance in downstream tasks.

In an effort to learn high-level semantic information, researchers have employed downstream tasks to drive fusion networks to better learn the high-level semantic information of target objects. Some have utilized semantic segmentation tasks [Tang *et al.*, 2022b], while others have employed object detection [Liu *et al.*, 2022]. However, using separate modules for image fusion and downstream tasks, and training directly with limited annotated data, makes a challenge in integrating high-level visual information into the fusion module through downstream tasks. Consequently, researchers proposed joint training of the encoder with downstream segmentation tasks [Tang *et al.*, 2023], enhancing the encoder's ability to extract high-level visual information under limited data. There are also proposals to use meta-learning training methods [Zhao *et al.*, 2023a], allowing the model to progressively learn relevant object information and fully utilize the information brought by downstream tasks. However, these methods also introduce complexities in model architecture and training approaches, and still face the issue of insufficient labeled data, leading to a risk of overfitting on training data.

Inspired by the observation that MAE encoder [He *et al.*, 2022], trained self-supervisedly on a large image dataset, has demonstrated proficiency in assimilating both low-level and high-level visual information [Li *et al.*, 2022; Liu *et al.*, 2023a], making it an ideal choice for extracting robust features for better fusion. In this paper, we propose to leverage pretrained MAE encoder to extract the features of infrared and visible images for better fusion. By applying the pretrained model with strong representation ability, we



a b c d

Figure 1: The fusion results with different fusion strategies: (a) mean fusion for two features, (b) cross-attention based fusion without alignment in the feature domain, (c) cross-attention based fusion with alignment in the feature domain, and (d) our MaeFuse. Here we only show the grayscale images for better comparison.

can expect to reduce the dependence on data of IVIF task, where the paired samples are usually scarce. However, as shown in Figure 1, fusing infrared and visible light features using conventional cross-attention [Li and Wu, 2024; Liu *et al.*, 2023b] may converge to the local optima, where fused image have obvious block effects and inharmonious regions. To this end, in this paper we further introduce a guided training strategy to mitigate this issue. This strategy is employed to expedite the alignment of the fusion layer output with the encoder feature domain, thereby avoiding the potential risk of converging to local optima. By adopting this approach, a simple yet effective fusion network was developed, which successfully extracts and retains both low-level image reconstruction and high-level visual features. Our contributions can be distilled into two main aspects:

- Employing a pretrained encoder, *i.e.*, MAE, as the encoder for fusion tasks enables the fusion network to acquire comprehensive low-level and high-level visual information. This approach addresses the issue of lacking high-level visual information in fusion features and also simplifies the overall network structure. Additionally, it resolves the problem of insufficient labeled data during training.
- A guided training strategy is proposed, aiding the fusion layer in rapidly adapting to and aligning with the encoder's feature space. This effectively resolves the issue of fusion training in vision transformer (ViT) architectures becoming trapped in local optima.

2 Related Works

In this section, we introduce some typical works related to our method. Firstly, we examine notable deep learning approaches in infrared and visible image fusion. Subsequently, we discuss research focusing on downstream task-driven image fusion. Lastly, we explore various pretrained models utilized for feature representation.

2.1 Deep Learning-based IVIF

The development of deep learning has brought many novel methods to image fusion. The architecture of infrared and visible image fusion networks based on deep learning can mainly be divided into autoencoder networks and end-to-end networks. Autoencoder networks can be further divided into three parts: the encoder, fusion layer, and decoder. Li et al. proposed the first pretrained fusion model [Li and Wu, 2018], introducing dense connections in the encoder. However, it adopted manually designed fusion strategies, which limit its flexibility and fusion performance. Later, Xu et al. used a classification saliency-based rule for image fusion [Xu *et al.*, 2021], allowing the fusion strategy to be learned as well. End-to-end fusion frameworks eliminate the need for manually designed fusion strategies. In this context, Xu et al. proposed a method [Xu *et al.*, 2020a] for general fusion tasks, which can automatically estimate the importance of source images. In addition, there are some methods that use multi-scale fusion, which also achieves very good performance [Li *et al.*, 2021]. Since the image fusion problem usually does not have ground truth, some researchers have turned to exploit new training strategies, such as generative adversarial methods [Ma *et al.*, 2019; Ma *et al.*, 2020; Li *et al.*, 2020] and unsupervised methods [Jung *et al.*, 2020; Cheng *et al.*, 2023]. All the methods above mainly focus on integrating the complementary information of the two modalities and enhancing the information in the fused images. A good visual result is their common pursuit. However, they generally cannot preserve the high-level visual information of the fused images to help us highlight the interested objects.

2.2 Downstream Task-Driven IVIF

To enhance the acquisition of high-level visual information for improved image fusion, Tang et al. initially employed semantic segmentation tasks as drivers for image fusion [Tang *et al.*, 2022b]. In a similar vein, Liu et al. utilized object detection for the same purpose [Liu *et al.*, 2022]. However, the limitation of labeled fusion datasets hampers the ability to provide semantic feedback to the fusion network through backpropagation. Addressing this, the work of [Tang *et al.*, 2023] introduces a technique to infuse high-level semantic information at the feature level within the fusion network. This is executed by co-training the encoder with both the downstream and fusion tasks. On the other hand, Zhao et al. recommended the adoption of meta-learning techniques [Zhao *et al.*, 2023a], allowing the fusion network to progressively assimilate both low-level and high-level visual information. These methods have brought fresh air to the field of image fusion, and have recently become one of the research hotspots. Nonetheless, these methods have led to increased complexity in both the fusion network and its training strategy, thereby complicating the design and implementation of fusion methods.

2.3 Pretrained Models for Feature Representation

A large pre-trained feature model achieves strong generalization and performance for downstream tasks, benefiting low-sample scenarios. For pretrained models, there are mainly two types of architectures: based on convolution and based on transformers. Within the realm of traditional convolution, VGG [Simonyan and Zisserman, 2015] initially served as the foundational backbone for encoder architectures. Subsequently, the advent of ResNet [He *et al.*, 2016] addressed the issue of gradient vanishing, enabling deeper network structures for enhanced feature extraction. However, it's important

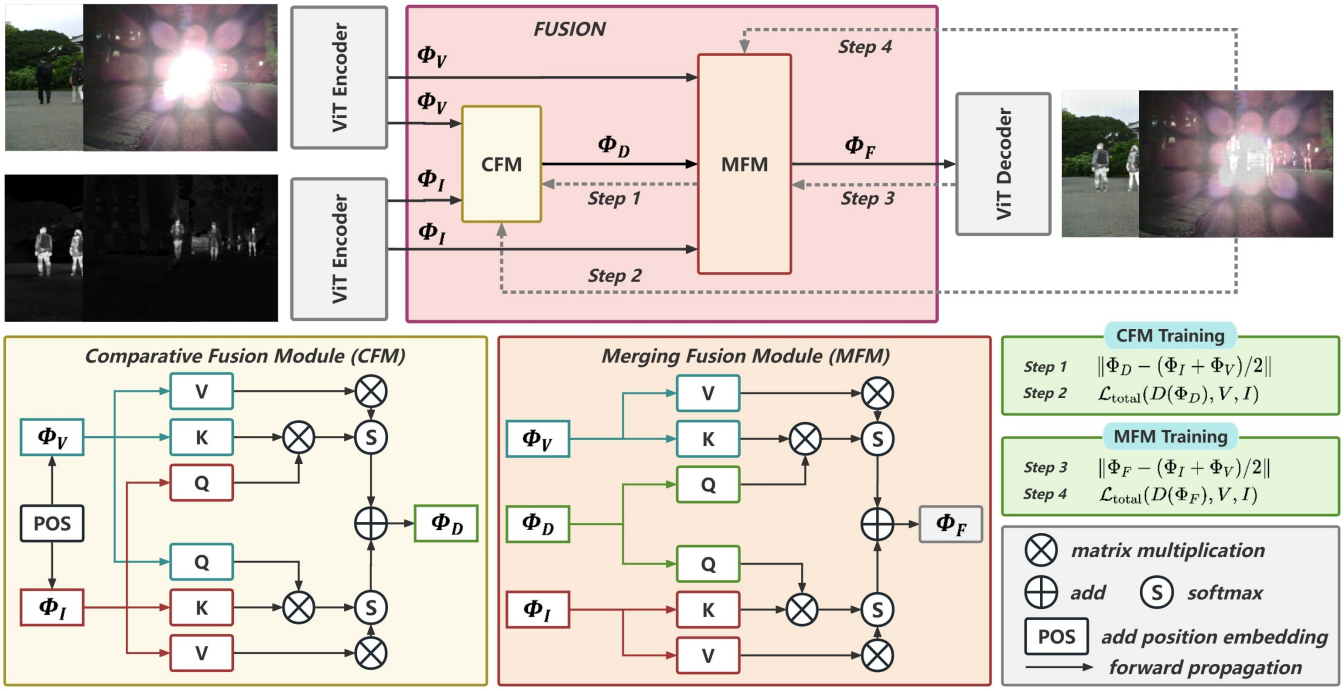


Figure 2: Workflow of our proposed MaeFuse. The upper part describes the overall architecture of the network, while the lower part elaborates on the content of the CFM and MFM structures. CFM cross-learning retains useful information, and MFM fuses enriched detail information based on the output content of CFM as a reference.

to note that both VGG and ResNet depend heavily on labeled data. To improve feature extraction capability and reduce dependence on labels, SimCLR [Chen *et al.*, 2020] uses contrastive learning to help the model obtain useful image representations. Regarding ViT, two primary architectures stand out: SimMIM [Xie *et al.*, 2022] and MAE [He *et al.*, 2022]. Both employ a masking strategy for self-supervised training during their pretraining phase. Notably, the MAE pretrained encoder excels in feature extraction, effectively bridging low-level and high-level visual information. Consequently, in our MaeFuse framework, we leverage the MAE pretrained encoder to enrich and enhance the comprehensiveness of feature information in our fusion tasks.

3 Method

In this section, we first clarify how to utilize the MAE's pretrained encoder for feature fusion, and then provide a detailed exposition of the guided training strategy, focusing on two aspects: the loss function and the training process.

3.1 Network Architecture

The proposed MaeFuse primarily consists of two architectures: 1) A pretrained encoder and decoder based on MAE, utilizing the encoder weights from MAE to directly obtain our feature vectors. 2) A learnable two-layer fusion network. We illustrate the proposed method in Figure 2.

Encoder. Here we utilize the MAE (large) architecture, characterized by 24 layers of ViT blocks. To process both infrared and visible images, we employ a unified encoder. The visible images are initially transformed into YCrCb format, with only the Y channel being used for input. This approach enables us to project information from both modalities into a singular, coherent feature space.

A brief notation is introduced for clarity. The input images, namely the visible image and the infrared image, are denoted as $V \in \mathbb{R}^{H\times W}$ and $I \in \mathbb{R}^{H\times W}$, respectively. Given that a singular encoder is utilized for both, it is referred to as $\mathbb{E}(\cdot)$ in our discussion.

Our encoding process can be described as converting the inputs $\{V, I\}$ from visible and infrared images into corresponding image features $\{\Phi_V, \Phi_I\}$

$$\Phi_V = \mathbb{E}(V), \quad \Phi_I = \mathbb{E}(I). \tag{1}$$

Fusion Layer. In the fusion process, our objective is for the network to selectively integrate modal information in various areas, drawing on the feature information from both modalities. Essentially, the aim is to ensure that the fusion outcome in each region is enriched with a higher density of information. The initial layer of the fusion layer, known as the Comparative Fusion Module (CFM), primarily enables the cross-learning of feature information from modalities Φ_V and Φ_I. The output of CFM, represented by Φ_D, predominantly captures essential contour information from both modalities. However, it is important to note that during the cross-learning process, this module often loses a significant amount of detail information.

To mitigate the problem of the Comparative Fusion Module (CFM)'s output lacking fine details, we incorporated the Merging Fusion Module (MFM). Within this framework, the feature vector Φ_D, produced by the CFM, acts as a guide for contour features, aiding in the re-fusion of the initial encoded features. The original encoded features Φ_V and Φ_I from the two modalities are then compared with Φ_D. In regions where

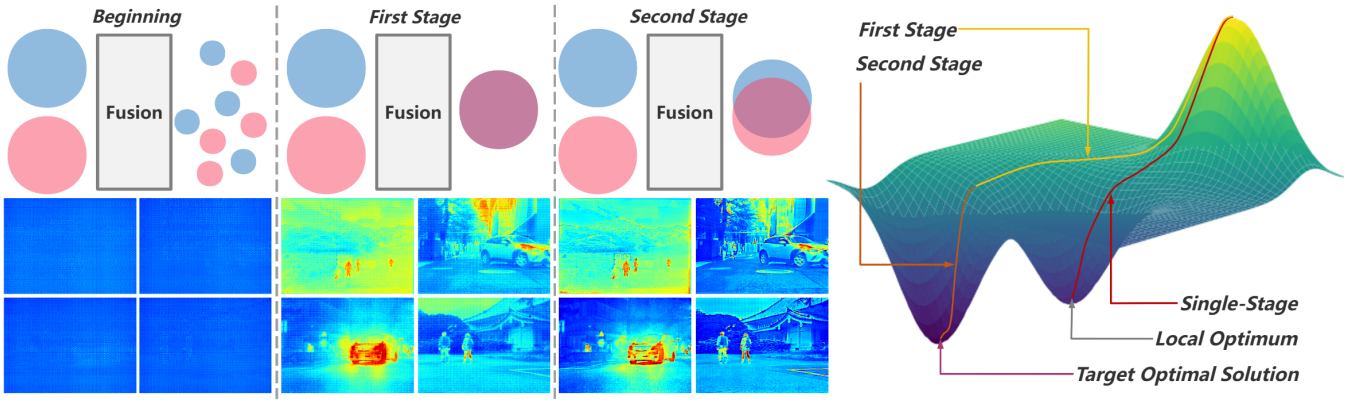


Figure 3: The schematic diagram illustrates our two-stage training process. The left side of the fusion module displays the original features outputted by the encoder, whereas the right side shows the features outputted by the fusion layer. The first stage involves aligning the feature domains of the fusion layer and the encoder. The second stage progresses with training using a fusion loss function Eq. (4). This two-stage training strategy is designed to effectively circumvent the issue of becoming trapped in local optima.

the original encoding closely aligns with Φ_D, the information is retained, while dissimilar regions are de-emphasized. This method ensures that the resulting fusion feature vector not only retains the contour information of Φ_D but also enriches it with more detailed information.

Mathmatically, CFM is denoted as $\mathbb{C}(\cdot)$, and MFM as $\mathbb{M}(\cdot)$. Φ_D represents the output content of CFM, while Φ_F is the final output result. Thus, we have

$$\begin{aligned} \Phi_D &= \mathbb{C}(\Phi_I, \Phi_V), \\ \Phi_F &= \mathbb{M}(\Phi_I, \Phi_V, \Phi_D). \end{aligned} \tag{2}$$

Decoder. For the ViT architecture, a simple few-layer decoder can achieve very good reconstruction results, so we choose 4-layer ViT blocks as the decoder. Here, $\mathbb{D}(\cdot)$ represents the decoder and we have

$$F = \mathbb{D}(\Phi_F). \tag{3}$$

3.2 Loss Function

For the purpose of maintaining image information integrity, our approach predominantly incorporates the $\mathcal{L}_1$ loss function along with the gradient loss:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{int}} + \alpha \mathcal{L}_{\text{grad}}, \\ \mathcal{L}_{\text{int}} &= \frac{1}{HW} \left\| F - \max(V, I) \right\|, \\ \mathcal{L}_{\text{grad}} &= \frac{1}{HW} \left\| |\nabla F| - \max\left(|\nabla V|, |\nabla I|\right) \right\|. \end{aligned} \tag{4}$$

Certainly, these loss functions are designed to globally optimize the image. Since we have already included high-level semantic information, we do not need related loss functions. Before calculating the fusion loss, we first need to align the output feature domain of the fusion layer with the feature domain of the encoder. Here, we use the mean of the two modalities as the baseline for training

$$\begin{aligned} \mathcal{L}_{\text{CFM-align}} &= \left(\frac{(\Phi_I + \Phi_V)}{2} - \Phi_D \right)^2, \\ \mathcal{L}_{\text{MFM-align}} &= \left(\frac{(\Phi_I + \Phi_V)}{2} - \Phi_F \right)^2. \end{aligned} \tag{5}$$

Algorithm 1 Two-Stage Training

Input: Φ_V and Φ_I
Output: Φ_F

```
Let t = 0.
Initialize Φ_F.
while t < 100 do
    if t < 20 then
        Compute loss with Eq. (5) between Φ_F and (Φ_V + Φ_I)/2.
    else
        Compute loss of D(Φ_F) with Eq. (4).
    end if
    Update MFM's weight based on the computed loss.
    t = t + 1.
end while
return Φ_F
```

3.3 Guided Training Strategy

Guided learning is an approach that focuses on achieving specific objectives to enhance learning efficiency. In our fusion task, we faced two primary challenges: limited data availability and the emergence of block effects with discordant image information when directly fusing features from the MAE encoder. To tackle these issues, we implemented guided training. This method sets predetermined objectives to steer the training process. Employing this strategy allows us to effectively address both challenges. It ensures the model avoids settling at local optima and achieves more accurate results.

Two-Stage Training

The MAE encoder can represent image data in a feature domain, where we need to fuse features of two modalities. Therefore, we first need to align the feature domain output by the fusion layer with the feature domain of the encoder. This alignment is our goal for guided learning. Once our feature domains are aligned, we can use a fusion loss function to promote the integration of information from the two modalities. Ignoring this phased training approach may lead to getting trapped in local optima. Our training strategy schematic is shown in Figure 3.

Figure 4: Visual comparisons utilizing high-exposure images on the MSRS dataset. Compare these with both state-of-the-art deep learning-based fusion methods and downstream task-driven fusion methods

We first define our guiding objective, which is the mean of the feature vectors of the two modalities. We calculate the least squares of the fusion network's output features with this target. This process helps to quickly align the feature domain output by the fusion layer with the feature domain of the encoder. Afterwards, we use an image texture loss function to guide the fusion effect towards a direction with richer texture. See Algorithm 1 for the pseudocode.

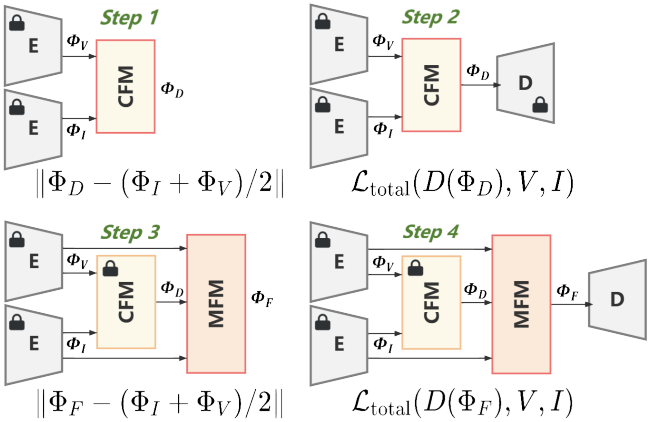


Figure 5: Visualization results of the training process. The first and second rows represent the training process of CFM and MFM, respectively. When conducting training for each module, there are two steps involved: modality alignment training based on the average of encoder and training based on the fusion loss. These two steps are known as two-stage training.

Hierarchical Training of Network Structures

For the CFM and MFM, we adopt a sequential guided training approach. Initially, we focus on the CFM, using a two-stage training strategy. This involves directly connecting the CFM's output to the decoder while keeping the MFM inactive. This step prevents the direct connection of the MFM with the encoder from diminishing the cross-learning effect in CFM. Once CFM training is complete, we lock its weights. Subsequently, we apply a similar two-stage strategy for the MFM, where its input is the output of the now-trained CFM, and its output is linked to the decoder. This approach ensures that the MFM effectively learns the fusion features from the CFM, thereby maximizing the retention of information acquired by the CFM.

4 Experiments

4.1 Implementation Details

We conducted experiments on four representative datasets (TNO [Toet and Hogervorst, 2012], LLVIP [Jia *et al.*, 2021], RoadScene [Xu *et al.*, 2020b] and MSRS [Tang *et al.*, 2022c]). The AdamW optimizer was employed to update parameters of different modules of the model. We utilized PyTorch's automatic mixed precision (AMP) for computation. And we trained the network structure using a fixed learning rate of $1e^{-4}$. All experiments were implemented under the PyTorch framework.

4.2 Evaluation in Multi-modality Image Fusion

We conducted qualitative and quantitative analyses with eight state-of-the-art competitors, including ReCoNet [Huang *et al.*, 2022], SuperFusion [Tang *et al.*, 2022a], DeFusion [Liang *et al.*, 2022], SwinFusion [Ma *et al.*, 2022], CDDFuse [Zhao *et al.*, 2023b], SeAFusion [Tang *et al.*, 2022b], MetaFusion [Zhao *et al.*, 2023a], and BDLFusion [Liu *et al.*, 2023c].

Qualitative comparisons. To demonstrate that the MAE pretrained encoder can effectively extract high-level visual information, we categorize the comparative methods into two

Method	MSRS			LLVIP			RoadScene			TNO		
	EN	CC	SCD	EN	CC	SCD	EN	CC	SCD	EN	CC	SCD
DeFusion	6.342	0.596	1.291	7.141	0.645	1.299	6.857	0.645	1.27	6.683	0.553	1.491
ReCoNet	5.092	0.556	1.255	6.186	0.662	1.472	7.086	0.65	1.636	6.711	0.543	1.681
SuperFusion	6.587	0.601	1.657	7.272	0.662	1.533	7.000	0.621	1.375	6.886	0.527	1.579
MetaFusion	6.365	0.623	1.51	7.101	0.638	1.379	7.407	0.647	1.595	7.237	0.542	1.712
BDLFusion	6.093	0.653	1.563	7.165	0.644	1.371	7.01	0.646	1.339	6.968	0.549	1.632
MaeFuse	6.575	0.628	1.674	7.362	0.672	1.59	7.081	0.672	1.644	6.978	0.576	1.730

Table 1: Quantitative results of the IVF task on four datasets. Red and blue represent the best and second-best values, respectively.

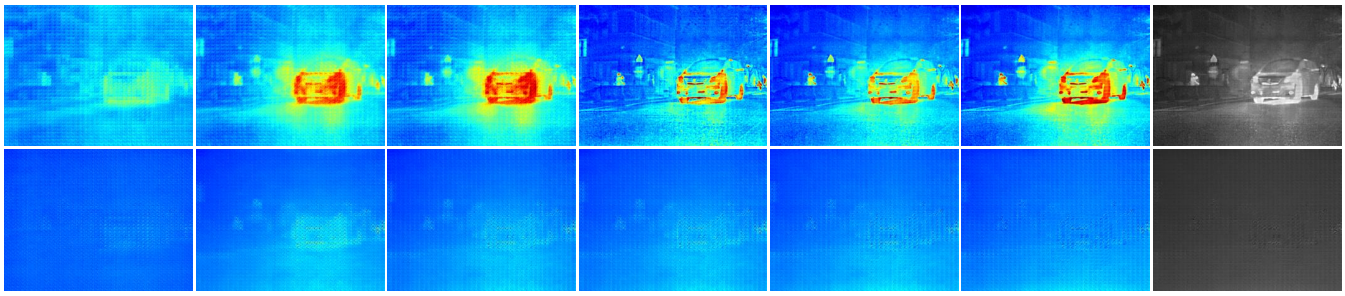


Figure 6: Visualization results of two-stage training ablation studies. From left to right, the number of training iterations gradually increases. The first row shows the results using two-stage training; the first three images are from the first stage of training, and the next three images are from the second stage. The second row shows the results without using two-stage training, where each image corresponds to the same number of iterations as the images in the first row.

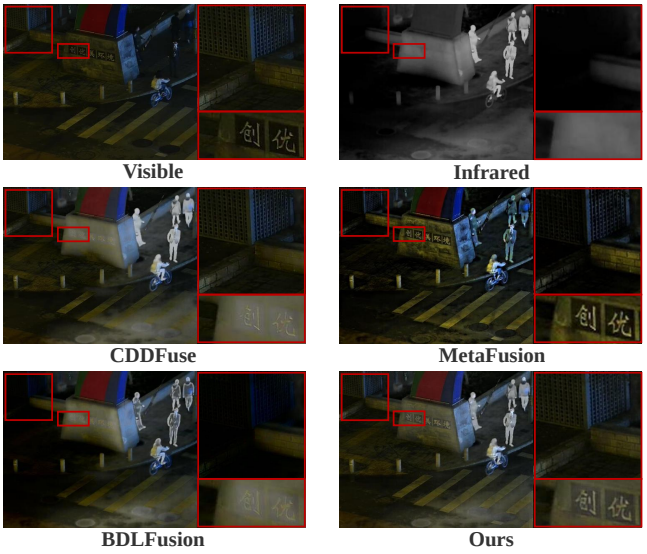


Figure 7: Visual comparison for '010418' in LLVIP IVF dataset.

groups: those without fusion training driven by downstream tasks and those with fusion training driven by downstream tasks. In the exposed images of the MSRS dataset, we find that methods not driven by downstream tasks fail to effectively represent object information and are easily influenced by high exposure. However, methods driven by downstream tasks can to some extent reveal object information. Our method, not driven by downstream tasks, also shows equally good or even better results, indicating that our encoder can extract and fuse high-level visual information. This is exemplified in Figure 4.

Since our encoder can establish a unity between low-level and high-level visual information, the fused images by MaeFuse also maintain texture details very well. In the data of LLVIP, our fused images fully preserve the texture information of the images, such as the grids in the picture. Moreover, our fused images are not influenced by infrared images, thus avoiding artifacts. For example, we can still clearly see the text on buildings. Our fused images only highlight significant infrared information on people, which demonstrates our model's ability to extract high-level visual information. The content is shown in Figure 7.

Quantitative comparison. We also report in Table 1 the numerical comparison results of our method against five other fusion competitors on four datasets: MSRS, LLVIP, RoadScene, and TNO. We use three objective metrics, EN, CC, and SCD, for analysis, which means that the fused images are compared in terms of the amount of graphic information, correlation with the input images, color distribution, and multi-scale structural similarity. It is clearly observed that our method demonstrates superiority in these statistical metrics.

4.3 Ablation studies

To demonstrate the effectiveness of our guided training strategy, we conducted ablation experiments on both two-stage training and hierarchical training strategies. First, for the two-stage training, we trained a group of fusion layers directly with the fusion loss function for 50 iterations, while another group of fusion layers underwent domain alignment training for 25 iterations, followed by another 25 iterations of training with the same fusion loss function. From the qualitative results, we observed that the data not undergoing two-stage training gradually failed to express image information, falling into local traps. In contrast, data with two-stage training rapidly aligned feature domains in the first 25 iterations, then refined image details with the fusion loss function in the latter 25 iterations, thus progressively eliminating visual issues caused by the ViT block effect. From a quantitative standpoint, the loss results without two-stage training gradually stabilized, indicating a local optimum trap. Meanwhile, data with two-stage training initially did not move towards

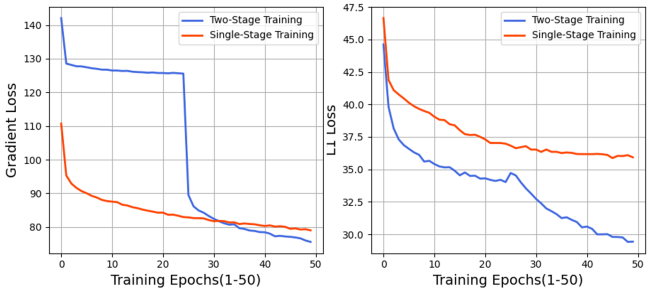


Figure 8: The training process for two-stage training. Blue represents two-stage training, with the first 25 iterations as the first stage and the latter 25 iterations as the second stage. Orange represents training without using the two-stage approach. The left graph shows the gradient loss in fusion, and the right graph shows the $\mathcal{L}_1$ loss.

Figure 9: Visualization results of hierarchical training ablation experiment. The first row shows the results with activated CFM weights, and the second row shows the results with locked CFM weights. The number of iterations corresponding to each image increases from left to right.

the best fusion loss result but rapidly converged after domain alignment and achieved better outcomes, also showing a continuous downward trend. These points prove the effectiveness of our training method. Qualitative results are shown in Figure 6, and quantitative results are presented in Figure 8.

To demonstrate the effectiveness of hierarchical training, we conducted an experiment using the same fusion training function. In this experiment, one group had the weights of the CFM activated, while the other group had the CFM weights locked. As shown in the results, the training with locked weights better maintained the original input features. In contrast, training with activated weights tended to cause the loss of original learning parameters, leading to overfitting of the loss function in the image results. Therefore, hierarchical training is beneficial for learning features content layer by layer. Qualitative results are shown in Figure 9.

4.4 Discovery

Infrared and visible image fusion tasks often operate under the assumption that infrared images lack detail. Contrary to this belief, our experiments reveal that infrared images actually possess significant detail. To uncover these hidden details, we employ Gamma correction to the infrared image, some results are shown in Figure 10.

We can clearly see that infrared images have relatively complete image contour information. At this point, we

Figure 10: Lighting the infrared image by Gamma correction. With the decrease of γ, more and more details are being revealed.

Figure 11: The fused results by replacing the CFM with infrared features under normal (the first two rows) and overexposure (the last two rows) light conditions.

wanted to know if the information from the infrared images has a strong effect. We used the feature vector of the infrared image as the input for the MFM module, replacing the CFM module. Figure 11 shows the results and demonstrate outstanding performance. When the lighting conditions are normal, the performance does not decrease, and it can improve performance in extreme overexposure. Therefore, how to better utilize the information from infrared images can be a consideration of our future work!

5 Conclusion

This paper presents relevant experiments demonstrating the effectiveness of using a pretrained encoder with Masked Autoencoders (MAE). This approach enables the encoded features to capture both high-level and low-level visual information efficiently. Additionally, our guided training strategy significantly accelerates the network's convergence towards the optimal solution, while minimizing the risk of entrapment in local optima. A notable discovery is the abundance of contour information in infrared images. Identifying strategies for the effective utilization of this information represents a promising avenue for future fusion research.

References

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[Cheng *et al.*, 2023] Chunyang Cheng, Tianyang Xu, and Xiao-Jun Wu. Mufusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92:80–92, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[Huang *et al.*, 2022] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European Conference on Computer Vision*, pages 539–555. Springer, 2022.

[Jia *et al.*, 2021] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[Jung *et al.*, 2020] Hyungjoo Jung, Youngjung Kim, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Unsupervised deep image fusion with structure tensor representations. *IEEE Transactions on Image Processing*, 29:3845–3858, 2020.

[Li and Wu, 2018] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

[Li and Wu, 2024] Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

[Li *et al.*, 2020] Jing Li, Hongtao Huo, Chang Li, Renhua Wang, and Qi Feng. Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Transactions on Multimedia*, 23:1383–1396, 2020.

[Li *et al.*, 2021] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.

[Li *et al.*, 2022] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[Liang *et al.*, 2022] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022.

[Liu *et al.*, 2022] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 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*, pages 5802–5811, 2022.

[Liu *et al.*, 2023a] Feng Liu, Xiaosong Zhang, Zhiliang Peng, Zonghao Guo, Fang Wan, Xiangyang Ji, and Qixiang Ye. Integrally migrating pre-trained transformer encoder-decoders for visual object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6825–6834, 2023.

[Liu *et al.*, 2023b] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. 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, 2023.

[Liu *et al.*, 2023c] Zhu Liu, Jinyuan Liu, Guanyao Wu, Long Ma, Xin Fan, and Risheng Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1240–1248. International Joint Conferences on Artificial Intelligence Organization, 2023.

[Ma *et al.*, 2019] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[Ma *et al.*, 2020] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[Ma *et al.*, 2022] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[Sun *et al.*, 2019] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

[Tang *et al.*, 2022a] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image

registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

[Tang *et al.*, 2022b] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. 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, 2022.

[Tang *et al.*, 2022c] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[Tang *et al.*, 2023] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. 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*, page 101870, 2023.

[Toet and Hogervorst, 2012] Alexander Toet and Maarten A Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):010901–010901, 2012.

[Xie *et al.*, 2022] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[Xu *et al.*, 2020a] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020.

[Xu *et al.*, 2020b] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12484–12491, 2020.

[Xu *et al.*, 2021] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7:824–836, 2021.

[Zhang *et al.*, 2020] Xingchen Zhang, Ping Ye, and Gang Xiao. Vifb: A visible and infrared image fusion benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 104–105, 2020.

[Zhao *et al.*, 2021] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: deep image decomposition for infrared and visible image fusion. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 976–976, 2021.

[Zhao *et al.*, 2023a] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023.

[Zhao *et al.*, 2023b] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023.