Multi-modal Gated Mixture of Local-to-Global Experts for Dynamic Image Fusion

Bing Cao*, Yiming Sun*, Pengfei Zhu†, Qinghua Hu
Tianjin Key Lab of Machine Learning, College of Intelligence and Computing, Tianjin University, China
Haihe Laboratory of Information Technology Application Innovation, China
{caobing,sunyiming1895,zhupengfei,huqinghua}@tju.edu.cn

Abstract

Infrared and visible image fusion aims to integrate comprehensive information from multiple sources to achieve superior performances on various practical tasks, such as detection, over that of a single modality. However, most existing methods directly combined the texture details and object contrast of different modalities, ignoring the dynamic changes in reality, which diminishes the visible texture in good lighting conditions and the infrared contrast in low lighting conditions. To fill this gap, we propose a dynamic image fusion framework with a multi-modal gated mixture of local-to-global experts, termed MoE-Fusion, to dynamically extract effective and comprehensive information from the respective modalities. Our model consists of a Mixture of Local Experts (MoLE) and a Mixture of Global Experts (MoGE) guided by a multi-modal gate. The MoLE performs specialized learning of multi-modal local features, prompting the fused images to retain the local information in a sample-adaptive manner, while the MoGE focuses on the global information that complements the fused image with overall texture detail and contrast. Extensive experiments show that our MoE-Fusion outperforms state-of-the-art methods in preserving multi-modal image texture and contrast through the local-to-global dynamic learning paradigm, and also achieves superior performance on detection tasks. Our code is available: https://github.com/SunYM2020/MoE-Fusion.

1. Introduction

Infrared and visible image fusion focus on generating appealing and informative fused images that enable superior performance in practical downstream tasks over that of using single modality alone [24, 45, 41, 21]. In recent years, infrared-visible image fusion has been widely used in many applications, such as autonomous vehicles [9] and unmanned aerial vehicles [34]. According to the thermal infrared imaging mechanism, infrared images can be adapted to various lighting conditions but has the disadvantage of few texture details [25, 50, 19]. By contrast, visible images contain rich texture detail information, but cannot provide clear information in low light conditions. Therefore, how to design advanced fusion methods such that the fused images preserve sufficient texture details and valuable thermal information has attracted a lot of research attention.

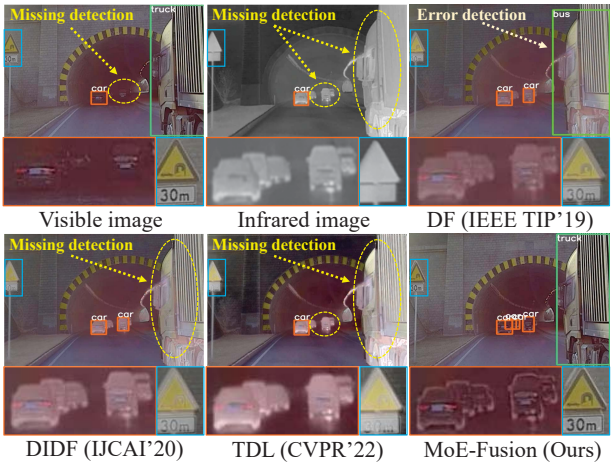

Figure 1. The importance of dynamic image fusion. In the SOTA methods (DF [16], DIDF [49], and TDL [18]), the texture details of objects (*e.g.*, car, truck, and traffic sign) in the fused image are suppressed by the contrast of infrared image, leading to terrible detection results. Benefiting from the dynamic fusion, our method can preserve clear texture details without being interrupted by unsuitable contrast, achieving the best performance.

Existing infrared-visible fusion methods [46, 38, 33, 42] can be mainly categorized to traditional approaches (image decomposition [15], sparse representation [47], *etc.*) and deep learning-based approaches (autoencoder based methods [16, 49, 17], generative adversarial network based

*Equal contribution
†Corresponding author

approaches [18, 26], transformer based approaches [40, 36], *etc*.). However, most of these methods directly combined the texture details and object contrast of different modalities, ignoring the dynamic changes in reality, leading to poor fusion results and even weaker downstream task performance than that of a single modality. As shown in Fig. 1, the infrared image should adaptively enhance *cars* in dim light while avoiding compromising the textural detail of *truck* in bright light. However, the object textures in these state-of-the-art fusion methods are significantly disturbed by the infrared thermal information due to the lack of dynamic learning of multi-modal local and global information, resulting in terrible object detection performance.

In complex scenes, different modalities have different characteristics: under good lighting conditions, the texture of an object should not be disturbed by thermal infrared information; under low lighting conditions, the contrast of an object also should not be suppressed by the darkness of the visible image. Most existing methods perform image fusion in a fixed correlation paradigm, ignoring the dominant modality changes dynamically in reality, and often fall into domain bias. To break the traditional fixed fusion paradigm, we pioneered sample-adaptive local-to-global experts to dynamically enhance the dominant modality for image fusion. Fig. 1 show that the proposed method not only eliminates domain bias but also achieves sample-adaptive dynamic fusion, yielding the best detection results. Specifically, we propose a dynamic image fusion framework with a multi-modal gated mixture of local-to-global experts, termed MoE-Fusion, which consists of a Mixture of Local Experts (MoLE) and a Mixture of Global Experts (MoGE) guided by a multi-modal gate. In MoLE, we introduce the attention map generated by an auxiliary network to construct multi-modal local priors and perform dynamic learning of multi-modal local features guided by multi-modal gating, achieving sample-adaptive multi-modal local feature fusion. Moreover, MoGE performs dynamic learning of multi-modal global features to achieve a balance of texture details and contrasts globally in the fused images. With the proposed dynamic fusion paradigm from local to global, our model is capable of performing a reliable fusion of different modal images.

We summarize our main contributions as follows:

- We propose a dynamic image fusion model, providing a new multi-modal gated mixture of local-to-global experts for reliable infrared and visible image fusion (benefiting from the dynamically integrating effective information from the respective modalities).
- The proposed model is an effective and robust framework for sample-adaptive infrared-visible fusion from local to global. Further, it prompts the fused images to dynamically balance the texture details and contrasts.
- We conduct extensive experiments on multiple infrared-visible datasets, which clearly validate our superiority, quantitatively and qualitatively. Moreover, we also demonstrate our effectiveness in object detection.

2. Related Works

2.1. Infrared and Visible Image Fusion

The infrared and visible image fusion task focuses on generating a fused image containing sufficient information through the learning of multi-modal features [24, 45, 20, 22, 12, 8]. Ma *et al*. [23] defined the goal of image fusion as preserving more intensity information in infrared images as well as gradient information in visible images. Li *et al*. [16] use the autoencoder to extract multi-modal overall features and fuse them by designed fusion rules, which inspired a series of subsequent works [14, 17]. Zhao *et al*. [49, 50] proposed the deep learning-based image decomposition methods, which decompose images into background and detail images by high- and low-frequency information, respectively, and then fuse them by designed fusion rules. Recently, some GAN-based methods [26, 25, 18] and Transformer-based methods [40, 36] have also attracted wide attention. These works, despite the different approaches adopted, all focus on learning on the representation of the overall multi-modal features. However, they ignore the dynamic changes in reality, which diminishes the visible texture in good lighting conditions and the infrared contrast in low lighting conditions. We propose a dynamic image fusion framework that enables sample-adaptive fusion from local to global. This approach prompts the fused images to balance the texture details and contrast with specialized experts dynamically.

2.2. Mixture-of-Experts

Mixture-of-Experts (MoE) [10, 30, 27] can dynamically adjust its structure according to different inputs. Shazeer *et al*. [32] constructed a sparsely-gated MoE layer that uses a gate network to select multiple experts and assign weights to each selected expert. The final result of the MoE is a weighted sum of the outputs of the different experts. This work also serves as the basis for subsequent research. Recently, some researchers [13, 27] have focused on exploring learning mechanisms in MoE, trying to solve the problems of unbalanced expert load and the number of activated experts faced by MoE during training. Other researchers [6, 52, 3, 39] focuses on the combination of MoE and Transformer. They expect to use MoE to build sparse models. Zhu *et al*. [52] introduced Conditional MoE into the generalist model and proposed different routing strategies to mitigate the interference between tasks and modalities. Existing MoE-related methods focus on modeling

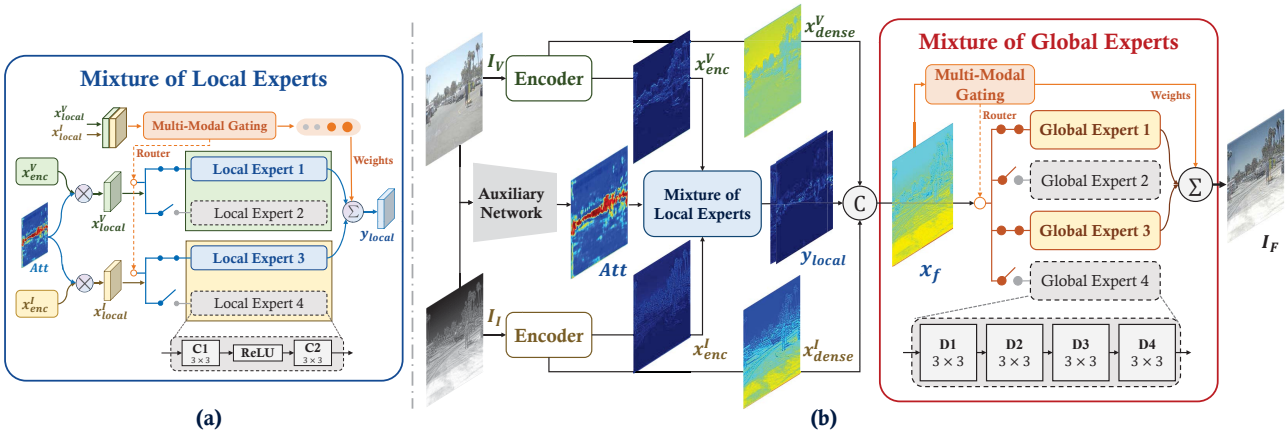

Figure 2. The architecture of MoE-Fusion. MoE-Fusion consists of a Mixture of Local Experts (MoLE), a Mixture of Global Experts (MoGE) guided by the multi-modal gate, a set of infrared and visible feature encoders, and the auxiliary network.

generic knowledge by using the dynamicity and sparsity of MoE, resulting in each expert does not know what they should be expert in. In contrast, we extend the idea of MoE to image fusion tasks for the first time, constructing a multi-modal gated mixture of local-to-global experts, assigning specific tasks to each expert, and enabling sample-adaptive specialized learning, which yields superior performance.

3. Method

3.1. Overall Architecture

In this paper, we propose a dynamic image fusion framework with a multi-modal gated mixture of local-to-global experts, termed MoE-Fusion. In Fig. 2, MoE-Fusion contains two encoders, a Mixture of Local Experts, a Mixture of Global Experts, and the auxiliary network.

In Fig. 2 (b), we send a pair of infrared image $I_{\mathcal{I}} \in \mathbb{R}^{H\times W\times 1}$ and visible image $I_{\mathcal{V}} \in \mathbb{R}^{H\times W\times 3}$ into the infrared and visible encoders ($Enc_{\mathcal{I}}$ and $Enc_{\mathcal{V}}$) to extract the features, respectively. The structure of encoders follows [16]. The output of the encoder has two parts: the feature maps of the last layer ($x^{\mathcal{I}}_{enc}$ and $x^{\mathcal{V}}_{enc}$) and the dense feature maps ($x^{\mathcal{I}}_{dense}$ and $x^{\mathcal{V}}_{dense}$). More details of the structure are provided in the supplementary material. We send the $x^{\mathcal{I}}_{enc}$ and $x^{\mathcal{V}}_{enc}$ to the MoLE along with the attention map which is learning from the auxiliary network. In this paper, we use Faster-RCNN [29] as the auxiliary network. In MoLE, we send visible and infrared features to specific local experts separately to achieve dynamic integration of local features under the guidance of multi-modal gating. We concatenate the outputs of MoLE with the dense feature maps as the input to the MoGE. Each expert in MoGE has the ability to decode global features, and a multi-modal gate network is used to dynamic select which experts are activated to decode multi-modal fusion features. The final fused image $I_{\mathcal{F}} \in \mathbb{R}^{H\times W\times 1}$ is generated by a weighted sum mechanism of the different global experts.

The MoE-Fusion is optimized mainly by calculating the pixel loss and gradient loss between the fused image $I_{\mathcal{F}}$ and two source images ($I_{\mathcal{I}}$ and $I_{\mathcal{V}}$). In addition, we also introduce the load loss to motivate each expert to receive roughly equal numbers of training images. The auxiliary detection networks are optimized independently by the detection loss.

3.2. Mixture of Local Experts

In infrared-visible image fusion tasks, specialized learning of multi-modal local information by a sample adaptive manner helps to overcome the challenge of multi-modal fusion failure in complex scenes. To realize this vision, we need to address two questions: (1) How to find local regions in multi-modal images; (2) How to learn the local features dynamically due to the differences in various samples.

As shown in Fig. 2 (a), we propose a Mixture of Local Experts (MoLE) to dynamically learning the multi-modal local features. We use the auxiliary detection networks with a spatial attention module to learn the attention maps. Then we can find the local regions in multi-modal images according to the guidance of the learned attention maps. Specifically, we introduce attention modules in two auxiliary detection networks for extracting visible attention maps and infrared attention maps, respectively. The modal-specific attention map $Att_{\mathcal{V}}/Att_{\mathcal{I}}$ is generated by the attention module between the feature extractor and the detection head in the detection networks, which consists of a Conv(1×1)-BN-ReLU layer and a Conv(1×1)-BN-Sigmoid layer. The $Att_{\mathcal{V}}$ and $Att_{\mathcal{I}}$ are concatenated and fed into 2 convolutional layers, the maximum of which output is the Att. In MoLE, we multiply the $x^{\mathcal{V}}_{enc}$ and $x^{\mathcal{I}}_{enc}$ with Att to obtain $x^{\mathcal{V}}_{local}$ and $x^{\mathcal{I}}_{local}$, respectively. Then we concatenate the $x^{\mathcal{V}}_{local}$ and $x^{\mathcal{I}}_{local}$ to get the x_{local}, which is the input of the multi-modal gating network. The MoLE is consist of

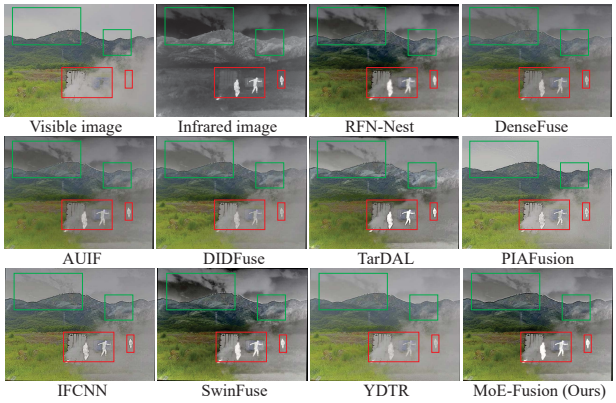

Figure 3. Qualitative comparisons of various methods on representative images selected from the M^{3}FD dataset.

a multi-modal gating network G_{local} and a set of N expert networks $\{E_1^{local}, ..., E_N^{local}\}$. The structure of each expert network is 2 convolution layers and 1 ReLU layers.

In MoLE, we flatten the input $x_{local} \in \mathbb{R}^{H\times W\times C}$ to $s_{local} \in \mathbb{R}^D$. The gate network G_{local} takes the vector s_{local} as input and produces the probability of it with respective to N experts. The formalization of the gate network is as follows,

$$G_{local}(s_{local}) = softmax(topK(s_{local} \cdot W_{local})), \quad (1)$$

where $W_{local} \in \mathbb{R}^{D\times N}$ is a learnable weight matrix and the top K outputs are normalized via $softmax$ function.

To achieve specialized learning of different modalities, we input visible local features $x_{local}^{\mathcal{V}}$ into one set of expert networks $\{E_1^{local}, ..., E_{N/2}^{local}\}$ and infrared local features $x_{local}^{\mathcal{I}}$ into another set of non-overlapping expert networks $\{E_{(N/2)+1}^{local}, ..., E_N^{local}\}$. Each expert network produces its own output $E_i^{local}(x_{local}^j)$. The final output y_{local} of the MoLE is calculated as follows,

$$y_{local} = \sum_{i=1}^{N} G_{local}(s_{local})_i E_i^{local}(x_{local}^j), \quad (2)$$

where j represents $\mathcal{I}$ or $\mathcal{V}$. Then we concatenate the y_{local}, $x_{dense}^{\mathcal{I}}$, and $x_{dense}^{\mathcal{V}}$ to obtain the global multi-modal fusion feature x_f.

3.3. Mixture of Global Experts

Traditional image fusion algorithms use the same network structure and parameters to learn the fusion features of different samples. In contrast, we propose the MoGE to dynamically integrate multi-modal global features, which can adaptively adjust its own structure and parameters when dealing with different samples, thus showing superior advantages in terms of model expressiveness and self-adaptability. The main components of the MoGE include a

Table 1. Quantitative comparison of our MoE-Fusion with 9 state-of-the-art methods. Bold red indicates the best, Bold blue indicates the second best, and Bold cyan indicates the third best.

	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
M^{3}FD Dataset [18]								
DenseFuse [16]	6.4134	0.0364	8.5987	2.9524	0.6572	3.0700	1.5069	0.3838
RFN-Nest [17]	6.9208	0.0345	9.2984	2.9301	0.7806	3.1698	1.5410	0.3772
IFCNN [48]	6.6555	0.0599	9.2456	2.9954	0.7522	5.0932	**1.5448**	**0.5755**
PIAFusion [35]	6.8167	**0.0707**	**10.1228**	**3.8337**	**0.8447**	**5.6560**	1.3065	**0.5540**
DIDFuse [49]	6.6116	0.0420	9.3409	2.9955	0.7382	3.5668	**1.5875**	0.4342
AUIF [50]	6.5233	0.0399	8.8759	2.9793	0.6796	3.3224	1.5314	0.4124
SwinFuse [40]	**6.9819**	**0.0696**	9.6400	3.2004	**0.9114**	**5.6234**	1.5395	0.5166
YDTR [36]	6.5397	0.0496	9.2631	3.2128	0.7276	3.8951	1.5076	0.4812
TarDAL [18]	**7.1347**	0.0528	**9.6820**	**3.2853**	0.8347	4.1998	1.5334	0.3858
MoE-Fusion	**7.0018**	**0.0715**	**10.1406**	**4.1949**	**1.0034**	**5.6742**	**1.5433**	**0.6661**
FLIR Dataset [44]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
DenseFuse [16]	6.9479	0.0304	10.5928	2.9254	0.6413	3.0524	1.3132	0.2947
RFN-Nest [17]	**7.4277**	0.0267	**10.8747**	2.9669	0.7260	2.7664	**1.6731**	0.2603
IFCNN [48]	7.1367	**0.0638**	10.6668	2.9263	0.7567	**6.1081**	1.3521	**0.4894**
PIAFusion [35]	6.9968	0.0551	10.6563	**3.0927**	**0.8178**	5.1849	1.1615	**0.4388**
DIDFuse [49]	7.2754	**0.0670**	**11.6318**	2.5022	0.6649	**5.8815**	1.4913	0.3469
AUIF [50]	7.2853	0.0463	10.2108	2.8893	0.7099	4.4471	1.5823	0.3240
SwinFuse [40]	7.4163	0.0582	10.6077	2.9404	**0.8104**	5.7587	**1.6739**	0.3751
YDTR [36]	6.8948	0.0364	10.6571	**3.0820**	0.6749	3.2993	1.3379	0.3333
TarDAL [18]	**7.4866**	0.0588	10.6948	3.0228	0.7665	5.1955	1.3182	0.3896
MoE-Fusion	**7.4925**	**0.0603**	**10.7007**	**3.1209**	**0.8212**	**5.7604**	**1.6818**	**0.4991**
LLVIP Dataset [11]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
DenseFuse [16]	6.8314	0.0426	9.3800	2.6764	0.6894	3.2640	1.2109	0.3093
RFN-Nest [17]	7.1408	0.0300	9.7184	2.5042	0.7294	2.7853	**1.4612**	0.2287
IFCNN [48]	7.2139	**0.0688**	**9.7633**	2.9479	**0.7797**	**5.4136**	1.4269	**0.5845**
PIAFusion [35]	**7.3954**	**0.0787**	9.7320	**3.3690**	**0.8860**	**6.0846**	**1.5300**	**0.5789**
DIDFuse [49]	6.0372	0.0550	7.8074	2.5137	0.5054	3.4474	1.2574	0.2436
AUIF [50]	6.1947	0.0636	7.8418	2.3966	0.5533	3.8588	1.2840	0.2764
SwinFuse [40]	5.9973	0.0608	7.6525	2.1846	0.5962	3.7344	1.2629	0.2620
YDTR [36]	6.6922	0.0474	8.8701	2.9152	0.6322	3.2043	1.0881	0.2907
TarDAL [18]	**7.3504**	0.0647	**9.7676**	**3.4655**	0.7769	4.6094	1.3607	0.4431
MoE-Fusion	**7.3523**	**0.0862**	**9.8664**	**3.1061**	**0.9202**	**6.1316**	**1.7841**	**0.5932**

global multi-modal gating network G_{global} and a set of N expert networks $\{E_1^{global}, ..., E_N^{global}\}$. In MoGE, we flatten x_f to get s_f and feed it into G_{global}. The corresponding gating weights of the N expert networks are calculated as follows,

$$G_{global}(s_f) = softmax(topK(s_f \cdot W_{global})), \quad (3)$$

where W_{global} is a learnable weight matrix and the top K outputs are normalized via $softmax$ distribution.

The structure of each expert network consists of 4 convolution layers. Each expert takes the global multi-modal fusion feature x_f as input to produce its own output $E_i^{global}(x_f)$. The final output $I_{\mathcal{F}}$ of MoGE is the linearly weighted combination of each expert's output with the corresponding gating weights. The formalization is as follows,

$$I_{\mathcal{F}} = \sum_{i=1}^{N} G_{global}(s_f)_i E_i^{global}(x_f). \quad (4)$$

3.4. Loss Function

In MoE-Fusion, we use the fusion loss $\mathcal{L}_{fusion}$ to guide the optimization of the image fusion network, and the auxiliary detection networks are optimized by their respective detection loss [29] ($\mathcal{L}_{det}^{\mathcal{V}}$ or $\mathcal{L}_{det}^{\mathcal{I}}$). We end-to-end train the entire framework through these three loss functions. Specif-

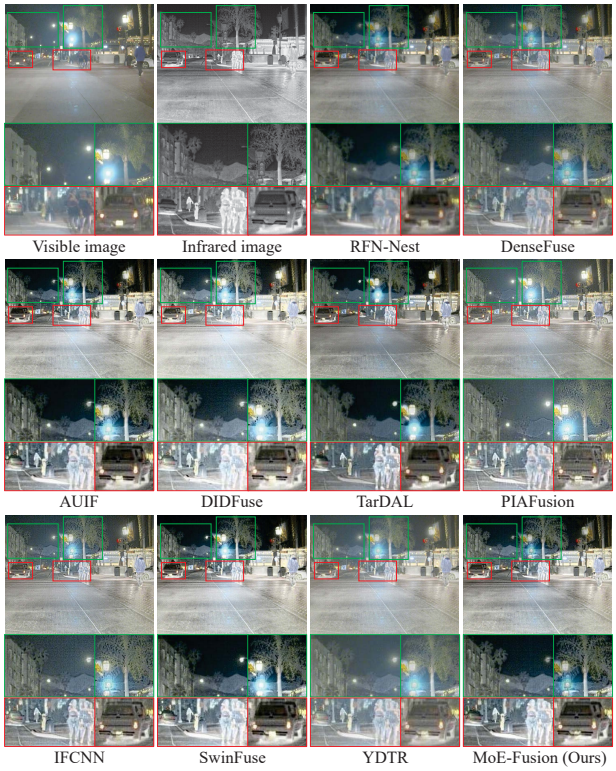

Figure 4. Qualitative comparisons of various methods on representative images selected from the FLIR dataset.

ically, the formula of fusion loss is as follows,

$$\mathcal{L}_{fusion} = \mathcal{L}_{pixel} + \alpha\mathcal{L}_{grad} + \mathcal{L}_{load}, \tag{5}$$

where the pixel loss $\mathcal{L}_{pixel}$ constrains the fused image to preserve more significant pixel intensities originating from the target images, while the gradient loss $\mathcal{L}_{grad}$ forces the fused image to contain more texture details from different modalities. $\mathcal{L}_{load}$ represents load loss, which encourages experts to receive roughly equal numbers of training examples [32]. More details about pixel loss, gradient loss and load loss are provided in the supplementary material. α is used to strike a balance between the different loss functions.

4. Experiments

4.1. Experimental Setting

Datasets and Partition Protocol. We conducted experiments on three publicly available datasets: (M^{3}FD [18], LLVIP [11] and FLIR [37]).

M^{3}FD: It contains $4,200$ infrared-visible image pairs captured by on-board cameras. We used $3,900$ pairs of images for training and the remaining 300 pairs for evaluation.

FLIR: We used the "aligned" version [44] of FLIR in this work. It contains $5,142$ infrared-visible image pairs captured by on-board cameras. We used $4,129$ image pairs for training and $1,013$ image pairs for evaluation.

LLVIP: The LLVIP dataset contains $15,488$ aligned infrared-visible image pairs, which is captured by the surveillance cameras in different street scenes. We trained the model with $12,025$ image pairs and performed evaluation on $3,463$ image pairs.

Competing methods. We compared the 9 state-of-the-art methods on three publicly available datasets (M^{3}FD [18], LLVIP [11] and FLIR [37]). In these comparison methods, DenseFuse [16] and RFN-Nest [17] are the autoencoder-based methods, PIAFusion [35] and IFCNN [48] are the CNN-based methods, TarDAL [18] is the GAN-based methods. DIDFuse [49] and AUIF [50] are the deep learning-based image decomposition methods. SwinFuse [40] and YDTR [36] are the Transformer-based methods.

Implementation Details. We performed experiments on a computing platform with two NVIDIA GeForce RTX 3090 GPUs. We used Adam Optimization to update the overall network parameters with the learning rate of 1.0×10^{-4}. The auxiliary network Faster R-CNN [29] is also trained along with the image fusion pipeline. The training epoch is set to 24 and the batch size is 4. The tuning parameter α is set to 10. For MoLE and MoGE, we set the number of experts is 4, and sparsely activate the top 2 experts.

Evaluation Metrics. We evaluated the performance of the proposed method based on qualitative and quantitative results. The qualitative evaluation is mainly based on the visual effect of the fused image. A good fused image needs to have complementary information of multi-modal images. The quantitative evaluation mainly uses quality evaluation metrics to measure the performance of image fusion. We selected 8 popular metrics, including the entropy (EN) [31], spatial frequency (SF) [4], standard deviation (SD), mutual information (MI) [28], visual information fidelity (VIF) [7], average gradient (AG) [2], the sum of the correlations of differences (SCD) [1], and gradient-based similarity measurement (Q_{abf}) [43]. We also evaluate the performance of the different methods on the typical downstream task, infrared-visible object detection.

4.2. Evaluation on the M^{3}FD dataset

Quantitative Comparisons. Table 1 presents the results of the quantitative evaluation on the M^{3}FD dataset, where our method achieves the best in 7 metrics and the second and third best performance in the remaining metrics, respectively. In particular, it shows overwhelming advantages on VIF, MI,and Q_{abf}, which indicates that our fusion results contain more valuable information and are more beneficial to the visual perception effect of human eyes. The highest SF, SD and AG also indicate that our fusion results preserve sufficient texture detail and contrast. Such superior performance is attributed to the proposed dynamic learning frame-

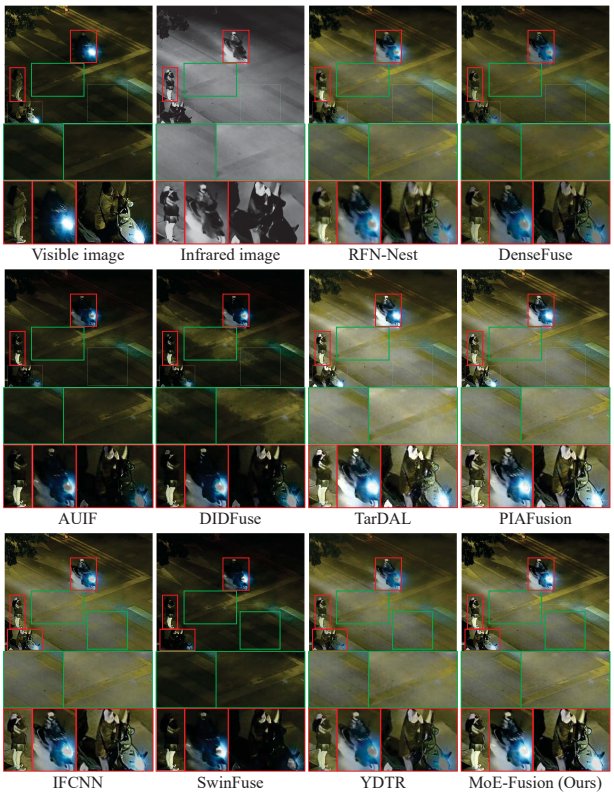

Figure 5. Qualitative comparisons of various methods on representative images selected from the LLVIP dataset.

work from local to global, which achieves state-of-the-art fusion performance through the sample adaptive approach.

Qualitative Comparisons. To better show the superiority of our model, we assigned the color information of the 3-channel visible image to the single-channel fused image through the color space conversion between RGB and YCbCr. We mark the background region with the green rectangular box and the foreground region with the red rectangular box. As shown in Fig. 3, our fusion results have the best results in both local and global regions. Compared to PIAFusion, YDTR, AUIF, IFCNN and DIDFuse, our fusion results show better contrast, and our fusion results show better texture details compared to TarDAL and SwinFuse. In daytime scenes, our model adaptively learned sufficient texture detail and significant contrast, such as containers and people. Especially for people, our method significantly avoids the effect of smoke and preserves the contrast information of infrared. In local regions, we successfully preserve the rich texture information of containers, outperforming other methods. In global regions such as mountains, grass and sky, our fusion results also effectively retain texture information, indicating that our method has better visual effects due to dynamic learning of local and global information of different modalities.

4.3. Evaluation on the FLIR dataset

Quantitative Comparisons. Table 1 reports the performance of the different methods on the FLIR dataset for 8 metrics. Our method achieves the best results in 5 metrics. Among them, the highest EN and MI indicate that our method can preserve abundant information of the multi-modal images well. The best performance of our method on SCD and Q_{abf} also indicates that our fusion results can better learn the multi-modal complementary information and edge information, which makes our fusion results have better foreground-background contrast and richer texture details. Moreover, the highest VIF also demonstrates that our method can generate fused images with better visual effects that are beneficial to human observation. In addition, the third best results on SF, SD, and AG also indicate that our method is highly competitive. The quantitative results on the FLIR dataset also validate the superiority of our method in dynamically fusing multi-modal complementary information from the local to the global.

Qualitative Comparisons. We mark the background region with the green rectangular box and the foreground region with the red rectangular box. We also show their zoomed-in effects for easier comparison. As shown in Fig. 4, our fusion results have the best results in both local and global regions. In night scenes, our fusion results adaptively learn sufficient texture detail and contrast, for example on buildings, trees, mountains and traffic lights. Especially for the traffic lights, our method effectively avoids the effects of glare and best preserves the entire outline of the traffic lights. On the local regions, our fusion results preserve the best contrast information of the pedestrians and the rich detail information of the vehicles. These comparisons illustrate that our method has better visual effects due to the effective dynamic learning of local information from different modalities. The superiority of the proposed MoE-Fusion also reveals that specialized knowledge of multi-modal local and global in fusion networks can effectively improve fusion performance.

4.4. Evaluation on the LLVIP dataset

Quantitative Comparisons. The quantitative results of the different methods on the LLVIP dataset are reported in Table 1. Our method outperforms all the compared methods on 6 metrics and achieved the second and third best results on the remaining 2 metrics, respectively. Specifically, the highest SF and AG we achieved indicate that the proposed method preserves richer texture details in the multi-modal images. As well as the highest SD also indicates that our fusion results can contain the highest contrast information between the foreground and the background. SCD and Q_{abf} denote the complementary information and edge information transferred from multi-modal images to fused image, respectively, and our highest results on these two metrics

Table 2. Ablation studies on three infrared-visible datasets.

	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
M^{3}FD Dataset [18]								
w/o MoLE	6.7856	0.0692	9.1636	2.7214	0.8100	5.5509	1.5153	0.6535
w/o MoGE	6.8351	0.0695	9.2491	2.8138	0.8261	5.6521	1.5346	0.6375
Att-Local	6.8656	0.0693	9.1894	2.7320	0.8271	5.5809	1.5190	0.6390
MoE-Fusion	**7.0018**	**0.0715**	**10.1406**	**4.1949**	**1.0034**	**5.6742**	**1.5433**	**0.6661**
FLIR Dataset [44]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
w/o MoLE	6.9021	0.0587	10.2859	2.3666	0.5371	5.6961	1.0976	0.3187
w/o MoGE	7.3172	0.0548	10.6654	2.9208	0.7747	5.2525	1.6629	0.4796
Att-Local	7.3070	0.0602	10.6403	2.6673	0.7448	5.7265	1.4788	0.4838
MoE-Fusion	**7.4925**	**0.0603**	**10.7007**	**3.1209**	**0.8212**	**5.7604**	**1.6818**	**0.4991**
LLVIP Dataset [11]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
w/o MoLE	6.9077	0.0661	9.8090	2.5039	0.5812	4.9358	1.1739	0.4732
w/o MoGE	7.2740	0.0847	9.6034	2.7067	0.7621	5.5960	1.6260	0.5110
Att-Local	7.2528	0.0858	9.5483	2.7129	0.8599	6.1099	1.6393	0.5735
MoE-Fusion	**7.3523**	**0.0862**	**9.8664**	**3.1061**	**0.9202**	**6.1316**	**1.7841**	**0.5932**

indicate that our method can learn more valuable information from multi-modal images. Moreover, the highest VIF also means that our method can generate the most appealing fused images that are more suitable for human vision. These quantitative results demonstrate that the proposed MoE-Fusion can effectively learn multi-modal knowledge and generate informative and appealing fusion results.

Qualitative Comparisons. We mark the background region with the green rectangular box and the foreground region with the red rectangular box. We also show their zoomed-in effects for easier comparison. As shown in Fig. 5, we can find that the proposed method best preserves the texture details of the local and global in the multi-modal image compared with the state-of-the-art methods, while highlighting the contrast information of the local dynamically. Specifically, for the background region, our fusion results show the sharpest effect on the edge texture of the zebra crossing. For the foreground region, our fusion results preserve the most significant contrast and the richest texture detail in pedestrians and cyclists. Qualitative comparisons show that our MoE-Fusion can balance the texture details and contrasts with the specialized experts dynamically.

4.5. Ablation Study

We conducted ablation studies on the M^{3}FD, LLVIP, and FLIR datasets and reported the results in Table 2.

MoLE. To verify the effectiveness of MoLE, we removed MoLE from MoE-Fusion and then extracted multi-modal features by two encoders only, and they were sent to MoGE after concatenation. As shown in Table 2, all the metrics show a significant decrease after removing MoLE, indicating that MoLE is very effective in learning multi-modal texture details and contrast information. Among others, the decrease on SCD also shows it is difficult to learn complementary multi-modal local information sufficiently without MoLE, which strongly supports our motivation to design MoLE. The local dynamic experts of MoLE adaptively boost dominant local (foreground and background) information induced from the auxiliary detector of different modalities for dynamic fusion, specialized to local regions.

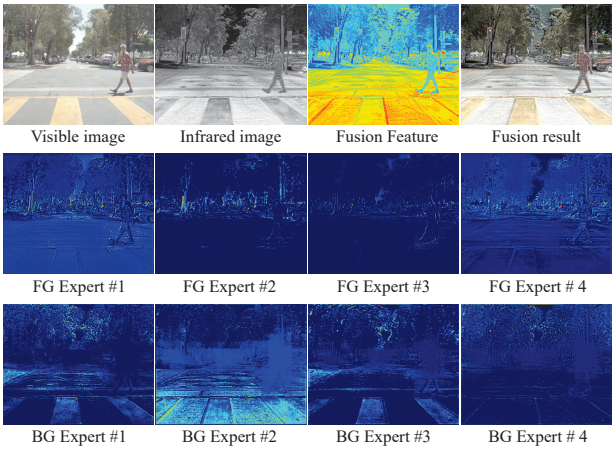

Figure 6. Visualization of the features learned by each expert in MoLE. FG denotes foreground, BG denotes background.

MoGE. We replaced the MoGE with a common decoder with the same structure as that of a single expert in the MoGE. As shown in Table 2, all the metrics appear to be significantly decreased, strongly verifying that the MoGE can help fused images preserve more contrast and texture detail information. Moreover, these results also demonstrate that MoGE can effectively motivate the image fusion network to dynamically adapt to different samples, learning better feature representations, and thus achieving better fusion performance. The global dynamic experts of MoGE further enhance the dominant modality from a global perspective while refining the potential errors that may occurr in local fusion.

Attention-based Local Feature. We want to explore how the fusion performance changes when dynamic learning is not performed for local features and only local feature priors constructed by attention maps are used as local features. We designed an attention-based local feature learning module (Att-Local), and to keep the output channels consistent, we followed the Att-Local module with a 1×1 convolution layers. In Table 2, all metrics cannot exceed the results of MoE-Fusion with the use of Att-Local, but most of them are higher than *w/o MoLE*, which demonstrates on the one hand that our proposed MoLE is indeed effective, and on the other hand that the dynamic integration of the effective information from the respective modalities is beneficial to improve the performance of multi-modal image fusion.

4.6. Analysis and Discussion

Visualization of MoLE. In the MoLE, we can obtain local feature priors according to the attention map Att, which we define as foreground local features (FG), and we also use $1 - Att$ to obtain background local features (BG). We visualize what each expert has learned in MoLE. As shown

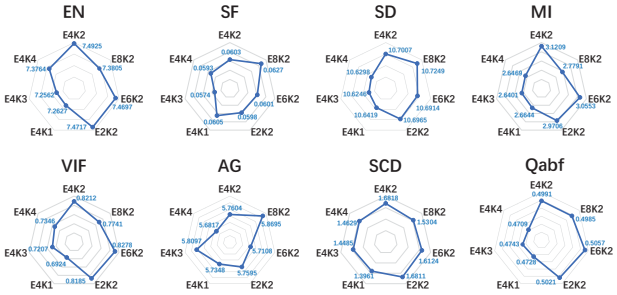

Figure 7. Experiments on the number of experts.

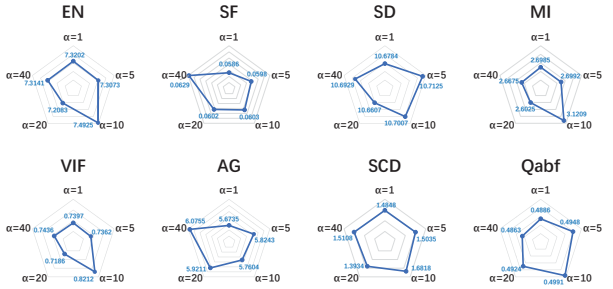

Figure 8. Hyperparameter sensitivity analysis on the FLIR dataset.

Table 3. Object detection evaluation on the FLIR dataset.

Methods	car	person	bicycle	mAP
Visible	55.90	27.20	33.20	38.70
Infrared	68.70	45.60	30.80	48.40
DenseFuse [16]	65.80	44.70	31.40	47.30
RFN-Nest [17]	58.80	36.50	26.20	40.50
IFCNN [48]	67.80	45.70	34.70	49.40
PIAFusion [35]	62.60	40.90	31.50	45.00
DIDFuse [49]	64.20	41.70	32.10	46.00
AUIF [50]	61.70	36.80	29.60	42.70
SwinFuse [40]	65.60	41.30	29.70	45.60
YDTR [36]	65.00	44.20	32.50	47.20
TarDAL [18]	62.50	35.40	30.80	42.90
MoE-Fusion	**72.90**	**55.00**	**39.50**	**55.80**

in Fig. 6, we see that the four foreground local experts can clearly learn foreground information, while the four background local experts can also learn rich background features. These results show that MoLE can successfully let each expert know what it should specialize in.

Detection Evaluation. Good fused images should have better performance in downstream tasks. For different image fusion methods, we perform the evaluation on the object detection task and use the mean average precision (mAP) [5] as the metric. Following [51], we first train the object detection model using infrared images and visible images, and then we input the fused images generated by different image fusion methods into the object detection model for inference and evaluate their detection performance. In this paper, we use Faster R-CNN [29] as the object detection algorithm and set the IoU (Intersection over Union) threshold for evaluation to 0.5. According to Table 3, our MoE-Fusion outperforms all the compared methods and achieves the highest mAP. It is worth noting that our method has an overwhelming advantage on all categories, which demonstrate the proposed dynamic image fusion method is more beneficial for downstream tasks. The detection evaluation on other datasets is provided in the supplementary material.

Number of Experts. In Fig. 7, we performed 4 sets of experiments on the FLIR dataset, E2k2, E4K2, E6K2, and E8K2, to explore the effect of the number of experts on the fusion results. As an example, E4K2 means that the MoE contains 4 experts and sparsely selects top 2 experts for integration. We found that E4K2 was higher than E2K2 in 7 metrics, E6K2 in 6 metrics, and E8K2 in 5 metrics, suggesting that a higher number of experts may not be better. In addition, we also set up 3 sets of experiments, E4K1, E4K3 and E4K4, to verify the effect of sparse selection of experts in MoE on the fusion results. In Fig. 7, we find that E4K2 can outperform E4K1 and E4K3 in 7 metrics and E4K4 in all metrics. Therefore, in this work, we set 4 experts for each MoE and sparsely select 2 experts for integration.

Hyperparameter. As shown in Fig. 8, we choose five values for tuning parameter α (1, 5, 10, 20, and 40) and experiment with them in turn. When α is less than 10, the fusion results fail to exceed the performance on all metrics with α equal to 10. When α is greater than 10, there is an improvement in only 2 metrics (SF and AG) compared to α equal to 10, but the other 6 metrics show a decrease. Therefore, in this work, we set α to 10 to obtain better results.

5. Conclusion

In this paper, we propose a novel dynamic image fusion framework with a multi-modal gated mixture of local-to-global experts (MoE-Fusion), which can produce reliable infrared-visible image fusion results. Our framework focuses on dynamically integrating effective information from different source modalities by performing sample-adaptive infrared-visible fusion from local to global. The MoE-Fusion model dynamically balances the texture details and contrasts with the specialized local experts and global experts. The experimental results on three challenging datasets demonstrate that the proposed MoE-Fusion outperforms state-of-the-art methods in terms of visual effects and quantitative metrics. Moreover, we also validate the superiority of our MoE-Fusion in the object detection task. In future work, we will explore leveraging the uncertainty of different images to guide the fusion, and investigate developing an uncertainty-gated MoE paradigm for dynamic image fusion.

Acknowledgments

This work was supported in part by the National Key R&D Program of China 2022ZD0116500, in part by the National Natural Science Foundation of China under Grant 62222608, 62106171, and 61925602, in part by the Haihe Lab of ITAI under Grant 22HHXCJC00002, and in part by the Tianjin Natural Science Foundation under Grant 21JCYBJC00580.

References

[1] V Aslantas and Emre Bendes. A new image quality metric for image fusion: the sum of the correlations of differences. *Aeu-international Journal of electronics and communications*, 69(12):1890–1896, 2015.

[2] Guangmang Cui, Huajun Feng, Zhi hai Xu, Qi Li, and Yue ting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015.

[3] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Z. Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*, 2022.

[4] Ahmet M. Eskicioglu and Paul S. Fisher. Image quality measures and their performance. *IEEE Trans. Commun.*, 43:2959–2965, 1995.

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[6] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

[7] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14:127–135, 2013.

[8] 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.

[9] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015.

[10] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[11] 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 (ICCV) Workshops*, pages 3496–3504, 2021.

[12] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3783–3791, 2022.

[13] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam M. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *ICLR*, 2021.

[14] Hui Li, Xiaojun Wu, and Tariq S. Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69:9645–9656, 2020.

[15] Hui Li, Xiaojun Wu, and Josef Kittler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29:4733–4746, 2020.

[16] 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.

[17] 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.

[18] 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 (CVPR)*, pages 5802–5811, 2022.

[19] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):105–119, 2021.

[20] Jinyuan Liu, Yuhui Wu, Zhanbo Huang, Risheng Liu, and Xin Fan. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Processing Letters*, 28:1818–1822, 2021.

[21] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Transactions on Image Processing*, 30:1261–1274, 2020.

[22] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1600–1608, 2021.

[23] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016.

[24] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019.

[25] 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.

[26] 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.

[27] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

[28] Guihong Qu, Dali Zhang, and P. Yan. Information measure for performance of image fusion. *Electronics Letters*, 38:313–315, 2002.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[30] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, 2021.

[31] John Roberts, Jan A. N. van Aardt, and Fethi B. Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2, 2008.

[32] Noam M. Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR*, 2017.

[33] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4003–4011, 2022.

[34] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:6700–6713, 2022.

[35] 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*, 83-84:79–92, 2022.

[36] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, pages 1–16, 2022.

[37] FA Team et al. Free flir thermal dataset for algorithm training, 2019.

[38] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, 2022.

[39] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *NeurIPS*, 2022.

[40] Zhishe Wang, Yanlin Chen, Wenyu Shao, Hui Li, and Lei Zhang. Swinfuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Transactions on Instrumentation and Measurement*, pages 1–1, 2022.

[41] 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*, 44:502–518, 2022.

[42] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19647–19656, 2022.

[43] Costas S. Xydeas and Vladimir S. Petrovic. Objective image fusion performance measure. *Electronics Letters*, 36:308–309, 2000.

[44] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280. IEEE, 2020.

[45] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021.

[46] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *AAAI*, 2020.

[47] Qiang Zhang, Yi Liu, Rick S Blum, Jungong Han, and Dacheng Tao. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information Fusion*, 40:57–75, 2018.

[48] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

[49] 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 970–976. ijcai.org, 2020.

[50] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2021.

[51] Huabing Zhou, Wei Wu, Yanduo Zhang, Jiayi Ma, and Haibin Ling. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. *IEEE Transactions on Multimedia*, pages 1–1, 2021.

[52] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *NeurIPS*, 2022.