Supplemental Material for "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

Figure 1: Qualitative comparison of various methods in three representative dynamic scenarios. We use the green bounding box to mark the corresponding modality as dominant.

1. Introduction

In this supplementary material, we first discuss the results of dynamic fusion in three representative dynamic scenarios, then explain the justification for dynamic fusion from a gradient perspective, and finally, we show that the proposed dynamic fusion model MoE-Fusion has a significant contribution to downstream tasks through a qualitative evaluation on an object detection task. In addition, we perform a qualitative analysis of the ablation studies to strongly confirm the effectiveness of the proposed components. Moreover, we show the evaluation results of different methods on the LLVIP and M^3FD datasets for the object detection task. To show the performance advantages of the proposed method in local regions, we also perform a quantitative comparison of local regions (foreground and background) of the fused images on three datasets. The above supplementary contents fully reveal that the proposed MoE-Fusion enables sample-adaptive fusion and achieves the most significant advantages on downstream tasks by relying on the powerful dynamic learning capability. At the end of the supplemental material, we also provide details of the two encoders in the proposed MoE-Fusion, details of the fusion loss for optimizing the proposed fusion network, and details of the color space conversion. Our Code will be released for reliably reproducing.

2. Discussion on Dynamic Fusion

2.1. Qualitative Comparison in Complex Scenarios

In complex scenes, different modalities have different characteristics: the texture of an object should not be disturbed by thermal infrared information when it is clearly visible in the visible image; the contrast of the object should not be suppressed by the unfavorable information of the visible image (smoke, darkness, *etc*.) when the visible image is low quality. Therefore, we qualitatively compare different methods on three representative scenes (visible modality-dominated, infrared modality-dominated, and visible-infrared modalities co-dominated).

We mark the focused regions using red rectangular boxes. As shown in Fig. 1, in the visible modality-dominated scenario, our model dynamically learns effective textural details in infrared and visible images, avoiding the redundant infrared contrasts affecting the visible textures. However, the competing methods TarDAL, PIAFusion, IFCNN, SwinFuse, YDTR, AUIF, and DIDFuse failed to dynamically learn the effective information from two different modalities. In daytime scenes, the competing method TarDAL suffered from the over-contrast on the object, which seriously affected the local textures. As a comparison, our fusion results can be adaptively learned with sufficient texture detail and reliable contrast for people.

In the infrared modality-dominated scenario, our model dynamically learns the significant contrast of the object, avoiding the suppression of the thermal information in the infrared image by the smoke in the visible image. Unfortunately, competing methods DenseFuse, AUIF, DIDFuse, RFN-Nest, PIAFusion, IFCNN, SwinFuse, and YDTR suffer from varying degrees of suppression of object contrast from smoke due to indiscriminately and directly fusing information from multiple modalities. Although the competing method TarDAL preserves contrast due to enhanced learning of objects, black shadows, and noise appear in background areas such as the sky and tree branches, which interfere with the overall texture detail. Our approach

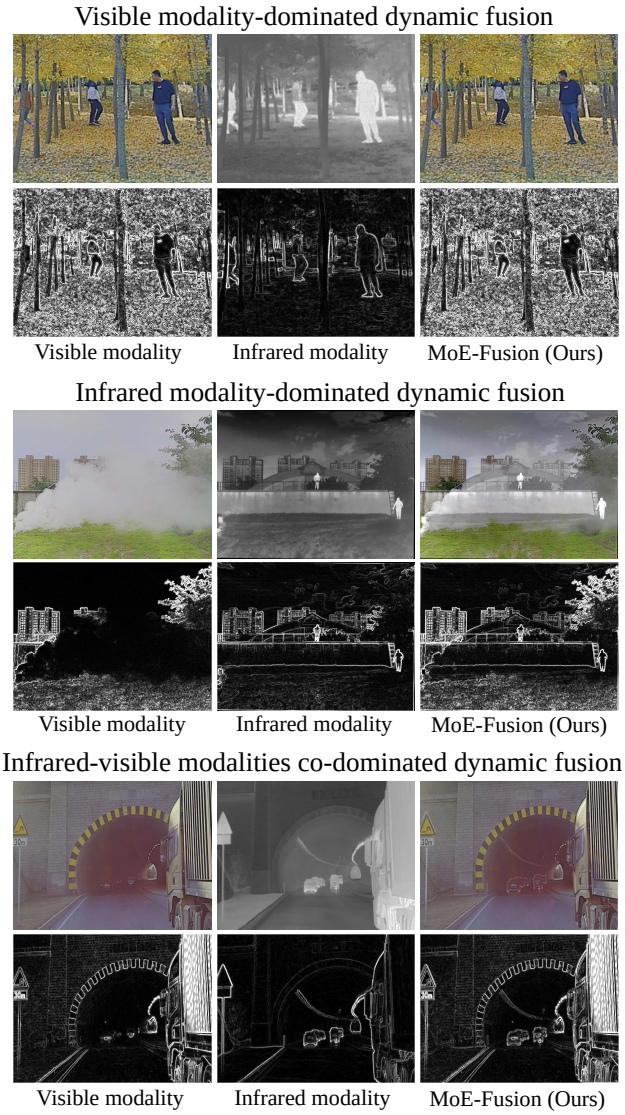

Figure 2: Analysis of dynamic fusion from a gradient perspective. We use the *Sobel* operator to generate the gradient maps of the visible modality, the infrared modality, and the fused results, respectively.

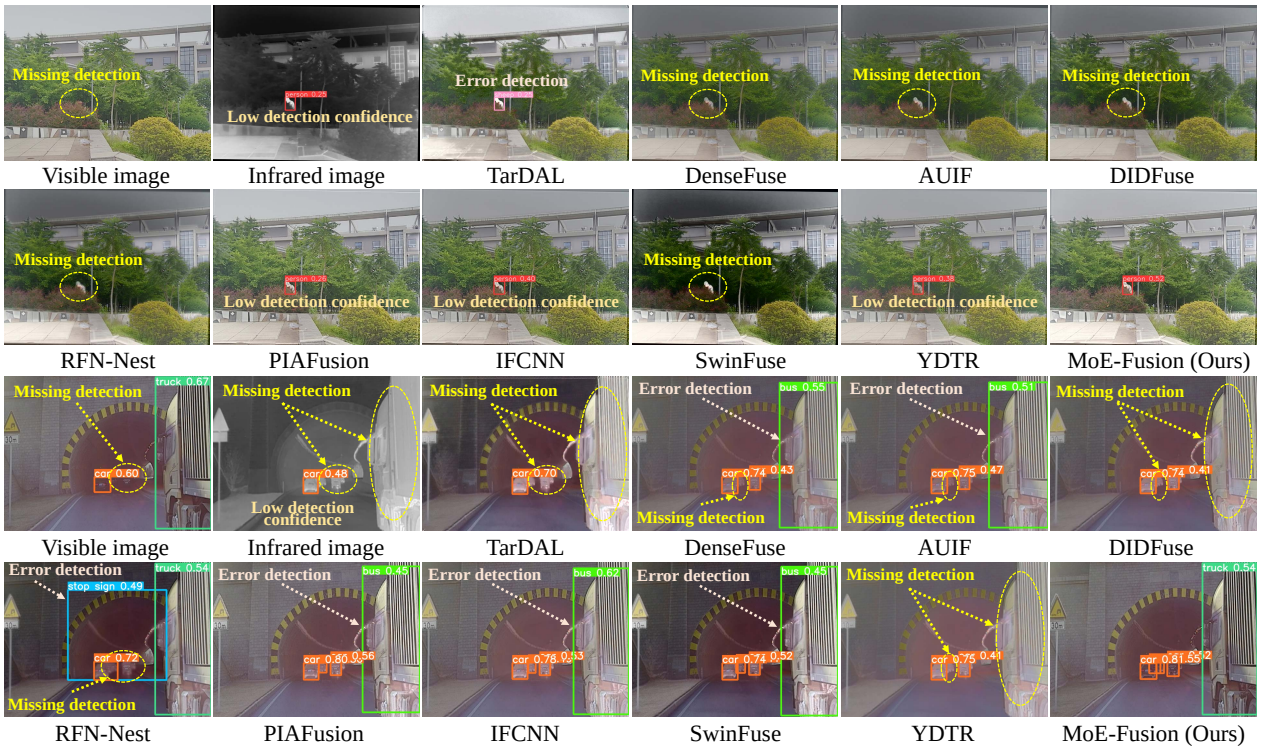

Figure 3: Qualitative comparison of various methods on detection performance.

achieves a sample-adaptive dynamic fusion from local to global, effectively preserving the high-value information of different modalities.

In the visible-infrared modalities co-dominated scenario, our model dynamically preserves the texture details of traffic signs and trucks while effectively learning the contrast information of vehicles in the tunnel. The competing methods TarDAL, YDTR, DenseFuse, AUIF, PIAFusion, and DIDFuse incorrectly fuse thermal information of traffic signs and trucks from infrared images into the final results, leading to serious over-exposure problems. The remaining methods have similar issues, resulting in fusion results with varying degrees of distortion on the trucks and traffic signs. The experimental results demonstrate our effectiveness in dynamically fusing multi-modal knowledge. This is particularly valuable for potential downstream applications, such as image fusion-based object detection.

2.2. Analysis from the Gradient Perspective

In three representative dynamic scenarios, we use the *Sobel* operator to extract the gradient information for the visible modality, the infrared modality, and our fusion results, respectively. As shown in Fig. 2, we can find that in scenes dominated by visible modality, the visible modality has the richest texture details compared to the infrared modality, so our method tends to mainly learn information from the visible modality. In scenes dominated by infrared modality, the infrared modality can also provide rich high-frequency information because it is not affected by smoke occlusion, so our method takes the infrared modality as the main learning reference to make up for the shortcomings of the visible modality. In scenes where infrared-visible modalities co-dominate, trucks and traffic signs have rich high-frequency gradient information in the visible modality, while the gradient information of vehicles in the tunnel needs to be reflected in the infrared modality. Our fusion results dynamically learn the complementary information of gradients in the two modalities, preserving the complete gradient information for trucks, traffic signs, vehicles, *etc*. According to the analysis of gradient perspective, the amount of effective information contained in different modalities is promising to be the basis of model dynamic learning, and in future work, we will take this as a clue to explore the trustworthy dynamic fusion of multi-modal images.

2.3. Qualitative Evaluation on the Object Detection

Following [5], we utilize YOLOv5 as the detection model. As shown in Fig. 3, our model dynamically learns the knowledge of multi-modal images in a sample-adaptive manner and achieves the best detection performance in the two example scenes. Competing methods directly combined the texture details and object contrast of different

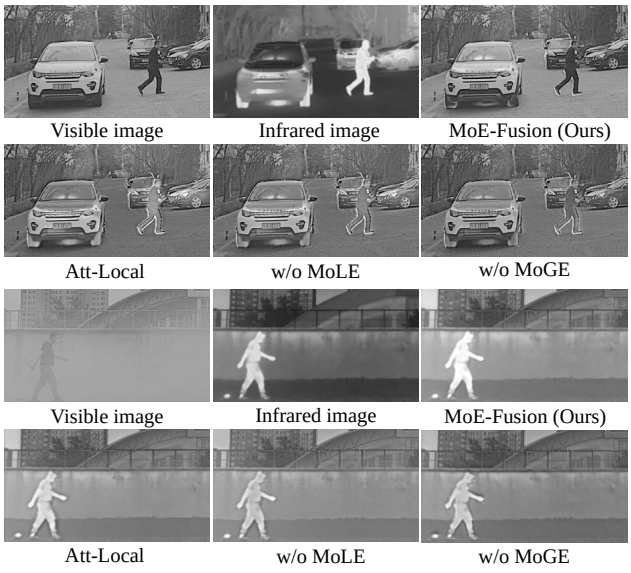

Figure 4: Qualitative analysis of ablation studies for the proposed MoE-Fusion on M^{3}FD dataset.

modalities, ignoring the dynamic changes in reality. For example, in the first scenario, the competing method TarDAL over-learned the thermal information of infrared, making the detection model incorrectly detect a person as a sheep. The rest of the competing methods do not make full use of the valid information of the different modalities, leading to the problem that the detection model cannot detect the object or has low detection confidence (below 0.5). In contrast, our method has the highest detection confidence, outperforming the unimodal detection results. In the second detection scenario, most competing methods also suffer from infrared modalities that do not adequately preserve the texture details of the trucks and cars, leading to missed and error detection problems. Our method dynamically learns key information of different modalities from local to global by the sample-adaptive manner to generate fused images that are more suitable for downstream tasks.

2.4. Qualitative Analysis of Ablation Studies

As shown in Fig. 4, we qualitatively analyze the results of the ablation studies on the M^{3}FD dataset. We use grayscale images to show the two scenarios of visible modality dominance and infrared modality dominance. In visible modality-dominated scenes, the fusion results of MoE-Fusion can effectively preserve the rich texture details of the visible modality and avoid the effects of overexposure of the infrared modality on pedestrians. Removing MoLE from MoE-Fusion clearly shows that the lack of dynamic learning of multi-modal local information causes the over-exposure problem of pedestrians. Replacing MoLE with Att-Local also leads to bad results, which strongly

Table 1: Object detection evaluation on the M^{3}FD dataset. Red indicates the best. "Motor" is short for Motorcycle.

Methods	People	Car	Bus	Motor	Lamp	Truck	mAP
Visible	17.9	70.7	78.1	34.8	33.2	49.9	47.4
Infrared	59.2	64.1	54.5	19.3	18.3	42.7	43.0
DenseFuse [3]	43.4	73.0	76.7	38.7	21.0	45.6	49.7
RFN-Nest [4]	31.8	68.4	74.5	31.1	19.8	41.4	44.5
IFCNN [11]	49.2	74.6	77.4	40.0	21.2	48.3	51.8
PIAFusion [7]	49.6	73.5	73.2	41.0	29.5	51.0	52.9
DIDFuse [12]	44.4	72.7	76.2	34.3	23.3	44.7	49.3
AUIF [13]	46.8	74.4	78.7	40.0	26.1	47.9	52.3
SwinFuse [9]	52.3	75.5	79.3	42.2	29.9	47.1	54.4
YDTR [8]	44.5	74.0	76.2	33.6	22.0	52.3	50.4
TarDAL [5]	59.1	66.7	67.2	30.8	12.5	42.3	46.4
MoE-Fusion	**59.7**	**77.1**	**80.2**	**42.7**	**34.1**	**56.1**	**58.3**

demonstrates the importance of MoLE to dynamically learn local information in multi-modal images. In addition, removing MoGE in MoE-Fusion also causes the fusion results to be disturbed by the infrared information of the object due to the lack of global dynamic learning, which also proves the effectiveness of MoGE. In infrared modality-dominated scenes, the fusion results of MoE-Fusion also preserve the most significant contrast information, and removing either MoLE or MoGE leads to a loss of contrast information, making the model vulnerable to suppression by smoke in visible images. Replacing MoLE in MoE-Fusion with Att-Local also weakens the contrast of pedestrians but is better than w/o MoLE, which also proves the effectiveness of the proposed MoLE. With the combination of MoLE and MoGE, MoE-Fusion can dynamically balance texture details and contrast from local to global with specialized experts, which achieves superior fusion performance.

3. Detection Evaluation

According to Table 1, the proposed MoE-Fusion achieves the most superior performance on the M^{3}FD dataset, 3.9% ahead of the next best method SwinFuse on mAP, and more than 10% ahead of the results for the infrared/visible modalities. Our method also achieves the best performance in each category. In particular, our performance on Lamp (34.1%) is substantially ahead of all competing methods and infrared modalities. These illustrate that MoE-Fusion fully exploits the power of multi-modal fusion by dynamically learning valid information from multi-modal images. According to Table 2, our MoE-Fusion outperforms all the competing methods and achieves the highest mAP. We achieved a 3.7% advantage over the second-best method PIAFusion. Considering that LLVIP is a night scene dataset and thus the detection performance of visible modality is poor, by dynamically fusing infrared and

Table 2: Object detection evaluation on the LLVIP dataset. Red indicates the best. LLVIP only contains the category "person", so only mAP is shown.

Methods	Visible	Infrared	DenseFuse [3]	RFN-Nest [4]	IFCNN [11]	PIAFusion [7]
mAP	51.8	85.6	66.2	67.5	85.9	87.3
Methods	DIDFuse [12]	AUIF [13]	SwinFuse [9]	YDTR [8]	TarDAL [5]	**MoE-Fusion**
mAP	69.9	64.1	53.7	77.4	85.0	**91.0**

visible images, the upper limit of infrared modality can be broken and achieved beyond the detection performance of single modality. The results of these detection evaluations demonstrate the proposed dynamic image fusion method is more beneficial for downstream tasks.

4. Local Quantitative Comparisons

To show the advantages of the proposed method in local regions of fused images, we also perform the local quantitative comparisons of the fused images on three datasets. We use the foreground and background areas to represent the local regions separately. Firstly, we use the annotated bounding boxes in the dataset to generate foreground and background masks. Then, we use these masks to extract the foreground and background images of the fusion results separately. In the foreground image, the region except the foreground objects is masked, and their pixel values are set to 0. In the background image, all the foreground objects are masked, and their pixel values are set to 0. On three datasets, we quantitatively evaluate the foreground and background images of the different methods separately to compare the fusion performance of the different methods in local regions intuitively. The results are reported in Table 3 and Table 4, respectively.

4.1. Local Comparison on the M^{3}FD Dataset

According to Table 3, our method achieves the best results on 7 metrics. In particular, it shows overwhelming advantages on VIF, MI, and Q_{abf}, which indicates that the fusion results of the proposed method in the local foreground region are more favorable to the visual perception effect of human eyes and also contain more valuable information. The highest EN, SF, SD, and AG also show that our method effectively preserves the texture details and contrast in the local foreground region of the multi-modal image. The experimental results reveal the effectiveness of the proposed MoE-Fusion to dynamically learn the local information of multi-modal images in a sample-adaptive manner.

In Table 4, our method achieves the best results on six metrics, the second best on SCD and the third best on EN. Our method dynamically learns the effective information of multi-modal images from local to global and thus can preserve the best high-frequency texture information on the local background region, as demonstrated by the results on SF, SD, and AG. The significant advantages on VIF, MI, and Q_{abf} also show that the proposed method can retain more valuable local background information in multi-modal images while producing fused images that are more suitable for human visual perception. Such superior performance is attributed to the proposed dynamic learning framework from local to global, which achieves state-of-the-art fusion performance through the sample adaptive approach. The excellent performance in the local region demonstrates that MoE-Fusion can balance global and local learning, which will benefit the performance of downstream tasks.

4.2. Local Comparison on the FLIR Dataset

In Table 3, our method outperforms all the compared methods on 6 metrics and achieved the third best results on the remaining 2 metrics, respectively. Specifically, the highest EN, MI, and SD indicate that the proposed method preserves the richest information and the highest contrast on the foreground image. The best results on SCD and Q_{abf} show that our method can prompt the foreground of the fused image to learn the most valuable complementary and edge information from the multi-modal images. In addition, the highest VIF also demonstrates the advantages of the foreground images generated by our method on visual effects. These results effectively demonstrate that the proposed MoE-Fusion has significant advantages in the local foreground quality of fused images due to the specialized learning of multi-modal local information.

As shown in Table 4, our method achieves superiority on 5 metrics, where the highest EN, MI, SF, and AG represent that our method preserves the richest texture details and pixel intensity information on the local background region of fused images. As well as the highest VIF indicates that our fusion results are also best suited for human vision on the local background region. The results on Q_{abf} and SCD also illustrate that our method is competitive in preserving multi-modal image complementary and edge information. These results also demonstrate that our fusion results are significantly superior to other methods on local background quality due to dynamic specialized learning of multi-modal local information.

Table 3: Local Quantitative comparison of our MoE-Fusion with 9 state-of-the-art methods for **the foreground regions of fused images**. 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^3FD Dataset [5]								
DenseFuse [3]	0.7273	0.0274	3.5816	0.7874	1.1452	0.5994	0.4655	0.5420
RFN-Nest [4]	0.7469	0.0265	3.5744	0.7865	1.1289	0.5973	0.5111	0.5090
IFCNN [11]	0.7544	0.0349	3.5772	0.7995	1.1732	0.8649	0.8726	**0.6492**
PIAFusion [7]	**0.7650**	**0.0373**	3.5740	**0.8208**	**1.1845**	**0.9019**	**1.0557**	**0.6296**
DIDFuse [12]	0.7381	0.0292	3.5809	0.7923	1.1566	0.6575	0.5813	0.5651
AUIF [13]	0.7324	0.0286	**3.5821**	0.7869	1.1436	0.6415	0.5566	0.5700
SwinFuse [9]	0.7561	**0.0361**	3.4939	0.8157	1.0825	**0.8891**	0.5435	0.5860
YDTR [8]	0.7350	0.0332	**3.5823**	0.8055	**1.1845**	0.7159	**1.0427**	0.5840
TarDAL [5]	**0.7608**	0.0339	3.5768	**0.8195**	1.1492	0.7683	**1.0278**	0.5277
MoE-Fusion	**0.7688**	**0.0377**	**3.5884**	**0.8844**	**1.2664**	**0.9166**	0.8671	**0.6964**
FLIR Dataset [10]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
DenseFuse [3]	1.0133	0.0379	4.3260	0.9867	1.1369	0.8175	0.3184	0.5382
RFN-Nest [4]	1.0472	0.0333	4.3269	0.9998	1.0739	0.7711	0.6335	0.5086
IFCNN [11]	1.0570	**0.0455**	**4.3319**	1.0143	1.1680	1.1543	**0.6894**	**0.6434**
PIAFusion [7]	**1.0615**	0.0430	4.3311	**1.0197**	**1.1706**	**1.1900**	0.6867	**0.6214**
DIDFuse [12]	1.0439	**0.0505**	4.3315	0.9947	1.1637	1.1286	0.6428	0.5602
AUIF [13]	1.0579	0.0301	4.1833	1.0060	0.7889	0.8856	0.4076	0.3687
SwinFuse [9]	**1.0617**	0.0436	4.3250	1.0141	1.1453	**1.1897**	**0.6930**	0.5973
YDTR [8]	1.0301	0.0407	**4.3321**	1.0065	**1.1698**	0.9667	0.6484	0.5746
TarDAL [5]	1.0605	0.0381	4.3305	**1.0181**	1.0580	0.9666	0.5872	0.5428
MoE-Fusion	**1.0621**	**0.0444**	**4.3322**	**1.0203**	**1.1710**	**1.1548**	**0.6980**	**0.6460**
LLVIP Dataset [2]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
DenseFuse [3]	0.5527	0.0172	2.9714	0.6360	1.0694	0.3460	0.5545	0.4465
RFN-Nest [4]	0.5613	0.0175	2.9028	0.6275	1.0402	0.3286	0.7709	0.3657
IFCNN [11]	0.5711	0.0252	3.0012	0.6471	1.1080	0.5475	0.8209	**0.6708**
PIAFusion [7]	**0.5850**	**0.0281**	2.9529	**0.6737**	**1.1472**	**0.5818**	**1.1314**	**0.6791**
DIDFuse [12]	0.5315	0.0205	2.5029	0.6175	0.8252	0.4030	0.1454	0.3497
AUIF [13]	0.5364	0.0199	2.6871	0.6210	0.9058	0.3824	0.1922	0.3306
SwinFuse [9]	0.5072	0.0196	2.4646	0.5753	0.8285	0.3697	0.0557	0.2830
YDTR [8]	0.5544	0.0213	**2.9899**	0.6470	1.0960	0.4096	0.8406	0.5286
TarDAL [5]	**0.5818**	**0.0283**	**2.9914**	**0.6713**	**1.1403**	**0.5820**	**1.1316**	0.5957
MoE-Fusion	**0.5725**	**0.0285**	**3.0074**	**0.6522**	**1.1475**	**0.5824**	**1.1317**	**0.6821**

Table 4: Local Quantitative comparison of our MoE-Fusion with 9 state-of-the-art methods for **the background regions of fused images**. 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^3FD Dataset [5]								
DenseFuse [3]	6.2466	0.0413	9.3869	3.4312	0.8207	2.9234	1.2845	0.4216
RFN-Nest [4]	6.7317	0.0396	9.7022	3.4181	0.9304	3.0006	1.4032	0.4115
IFCNN [11]	6.4615	0.0611	10.1760	3.4543	0.9117	4.7119	1.3914	**0.5948**
PIAFusion [7]	6.6101	**0.0697**	**10.6753**	**4.2777**	**0.9967**	**5.1920**	1.1991	**0.5724**
DIDFuse [12]	6.4335	0.0463	9.9337	3.4657	0.8963	3.3748	1.3576	0.4667
AUIF [13]	6.3530	0.0439	9.5516	3.4630	0.8432	3.1355	1.3180	0.4426
SwinFuse [9]	**7.0682**	**0.0674**	9.7274	3.6591	**1.0340**	**5.1878**	**1.4323**	0.5267
YDTR [8]	6.3607	0.0541	**10.2420**	3.6616	0.8902	3.7014	**1.4174**	0.5120
TarDAL [5]	**6.9323**	0.0551	10.1222	**3.7133**	0.9856	3.9125	1.3664	0.4201
MoE-Fusion	**6.7860**	**0.0710**	**10.6973**	**4.5515**	**1.1308**	**5.2203**	**1.4284**	**0.6721**
FLIR Dataset [10]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
DenseFuse [3]	6.7076	0.0455	10.9301	3.5629	0.7998	2.9991	1.0409	0.3475
RFN-Nest [4]	**7.1380**	0.0395	10.9758	3.5958	0.8616	2.6326	1.1256	0.3046
IFCNN [11]	6.8659	0.0674	10.8812	3.5080	0.9094	**5.3579**	1.1797	**0.5624**
PIAFusion [7]	6.6917	0.0648	**11.8801**	3.5812	**0.9594**	4.8385	1.0870	**0.4676**
DIDFuse [12]	6.9978	**0.0680**	**11.8999**	3.1025	0.8305	**5.3311**	**1.4191**	0.3856
AUIF [13]	7.0057	0.0468	10.1748	3.5235	0.7845	3.9412	0.6594	0.3244
SwinFuse [9]	7.1181	**0.0679**	10.9948	3.5513	**0.9579**	5.2656	**1.5214**	0.4090
YDTR [8]	6.6353	0.0534	11.0152	**3.6865**	0.8334	3.2230	1.2674	0.3743
TarDAL [5]	**7.2068**	0.0640	10.9833	**3.7278**	0.9042	4.8738	0.8469	0.4323
MoE-Fusion	**7.2187**	**0.0681**	**11.0548**	**3.7368**	**0.9667**	**5.3910**	**1.4059**	**0.5246**
LLVIP Dataset [2]								
	EN	SF	SD	MI	VIF	AG	SCD	Q_{abf}
DenseFuse [3]	6.7613	0.0434	9.5516	3.0287	0.7521	3.1273	1.1330	0.3181
RFN-Nest [4]	7.0563	0.0330	9.8548	2.8719	0.7955	2.6958	**1.4070**	0.2460
IFCNN [11]	7.1119	**0.0680**	**9.9317**	3.2737	**0.8395**	**5.1106**	1.3752	**0.5958**
PIAFusion [7]	**7.3004**	**0.0784**	9.8792	**3.6756**	**0.9380**	**5.7825**	**1.5195**	**0.5772**
DIDFuse [12]	5.9448	0.0527	7.6424	2.8137	0.5465	3.1695	1.0767	0.2340
AUIF [13]	6.1165	0.0617	7.7777	2.7252	0.6080	3.6030	1.1095	0.2725
SwinFuse [9]	5.8519	0.0589	7.5822	2.4033	0.6421	3.4821	1.0832	0.2630
YDTR [8]	6.6232	0.0484	9.1427	3.2533	0.6902	3.0566	1.0642	0.2918
TarDAL [5]	**7.2692**	0.0679	**9.8993**	**3.7244**	0.8328	4.4663	1.3791	0.4398
MoE-Fusion	**7.2048**	**0.0853**	**9.9427**	**3.3231**	**0.9678**	**5.8198**	**1.7371**	**0.5899**

4.3. Local Comparison on the LLVIP Dataset

According to Table 3, it can be seen that we achieved significant advantages on 6 metrics. Specifically, the highest SF, AG, and SD illustrate that our fusion results can have the richest texture detail and the highest contrast information in the local foreground region. The superior performance on SCD and Q_{abf} also indicates that our fusion results on the local foreground region can effectively learn complementary and edge information from multi-modal images. Moreover, the highest VIF also demonstrates that our fusion results are more favourable to human observation on the local foreground region. These results illustrate that the proposed method can effectively preserve valuable information in the local foreground region of the multi-modal images, which benefits from the specialized learning of local information.

As seen in Table 4, we achieved the best on 5 metrics and the second best on 1 metric. The highest SF, AG, SD, and SCD indicate that our method preserves the richest texture details and valuable multi-modal complementary information on the local background region of fused images. The highest VIF also indicates that our fused image is closer to human vision on the local background region. In addition, the second best on Q_{abf} also illustrates that our method is competitive in preserving multi-modal edge information. These results demonstrate that our method can better motivate fused images to preserve local background detail information of multi-modal images through sample-adaptive dynamic learning.

5. Encoder Architecture

The detailed architecture of the two encoders in the proposed MoE-Fusion is shown in Fig. 5. By convention, we grayscale the 3-channel visible image $I_{\mathcal{V}}$ to obtain a single-channel image, and then send it and the infrared image $I_{\mathcal{I}}$ to the two encoders ($Enc_{\mathcal{V}}$ and $Enc_{\mathcal{I}}$) separately for feature extraction. These two encoders have the same structure, but the parameters are not shared. Each encoder contains one convolutional layer $C1$ and three densely connected convolutional layers ($DC1$, $DC2$, $DC3$). Referring to [1], the output of each convolutional layer is fed into each subsequent convolutional layer in the densely connected mechanism, which facilitates the preservation of deep features as much as possible. Finally, we obtain the feature maps of $DC3$ layer ($x^{\mathcal{I}}_{enc}$ and $x^{\mathcal{V}}_{enc}$) and dense feature maps ($x^{\mathcal{I}}_{dense}$ and $x^{\mathcal{V}}_{dense}$) from the two encoders, respectively.

6. Fusion Loss

In MoE-Fusion, the fusion loss $\mathcal{L}_{fusion}$ contains the pixel loss $\mathcal{L}_{pixel}$, gradient loss $\mathcal{L}_{grad}$, and load loss $\mathcal{L}_{load}$. We provide more details on their formalization in this supplementary material.

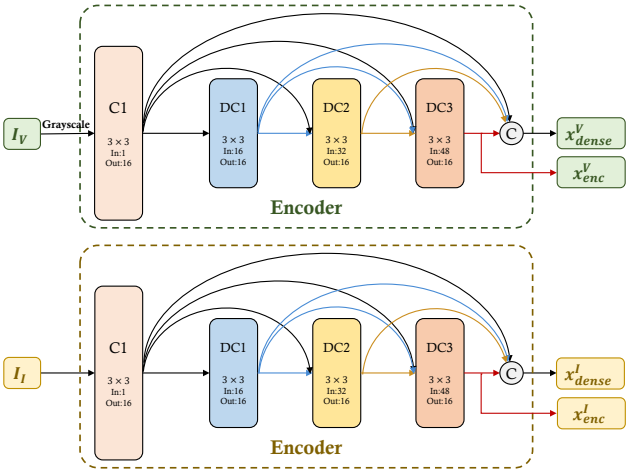

Figure 5: The architecture of the two encoders in the proposed MoE-Fusion.

6.1. Pixel Loss

We perform the different operations on the foreground and background of the multi-modal images to facilitate the fusion network to learn more valuable pixel intensity information. $\mathcal{L}_{pixel}$ is defined as,

$$\mathcal{L}_{pixel} = \mathcal{L}^{fg}_{pixel} + \mathcal{L}^{bg}_{pixel}, \tag{1}$$

where $\mathcal{L}^{fg}_{pixel}$ represents the pixel loss on the foreground regions, and $\mathcal{L}^{bg}_{pixel}$ represents the pixel loss on the background regions. Their formalizations are as follows,

$$\mathcal{L}^{fg}_{pixel} = \frac{1}{HW} \left\| I_m \circ \left(I_{\mathcal{F}} - max\left(I_{\mathcal{V}}, I_{\mathcal{I}} \right) \right) \right\|_1, \tag{2}$$

$$\mathcal{L}^{bg}_{pixel} = \frac{1}{HW} \left\| (1 - I_m) \circ \left(I_{\mathcal{F}} - mean\left(I_{\mathcal{V}}, I_{\mathcal{I}} \right) \right) \right\|_1, \tag{3}$$

where H and W represent the height and width of the image, respectively, I_m is the foreground mask generated according to the ground-truth bounding boxes of the auxiliary detection network. $\|\cdot\|_1$ stands for the l_1-norm, the operator $\circ$ denotes the element-wise multiplication, max$(\cdot)$ denotes the element-wise maximum operation, and mean$(\cdot)$ denotes the element-wise average operation.

6.2. Gradient Loss

We expect the fused image to preserve the richest texture details of the images from both modalities. So the gradient loss $\mathcal{L}_{grad}$ is formulated as,

$$\mathcal{L}_{grad} = \frac{1}{HW} \left\| |\nabla I_{\mathcal{F}}| - max\left(|\nabla I_{\mathcal{V}}|, |\nabla I_{\mathcal{I}}| \right) \right\|_1, \tag{4}$$

where ∇ denotes the Sobel gradient operator, which measures the texture detail information of an image. $|\cdot|$ stands for the absolute operation.

6.3. Load Loss

In the Mixture-of-Experts (MoE), the load loss is mainly used to encourage experts to receive roughly equal numbers of training examples. The proposed MoE-Fusion contains MoLE and MoGE. Therefore, the load loss $\mathcal{L}_{load}$ in our work consists of two parts, namely $\mathcal{L}_{load}^{local}$ and $\mathcal{L}_{load}^{global}$. The $\mathcal{L}_{load}$ can be calculated as:

$$\mathcal{L}_{load} = \mathcal{L}_{load}^{local} + \mathcal{L}_{load}^{global}. \tag{5}$$

Following [6], the $\mathcal{L}_{load}^{local}$ and $\mathcal{L}_{load}^{global}$ are defined as the square of the coefficient of variation of the load vector. We also follow [6] to initialize the weight matrix of the gate network in each MoE to zero, which keeps the expert load of MoE in an approximately equal state during the initial phase.

7. Color Space Conversion of Fused Images

By convention, the output of a fusion network is a single-channel image. We can better present the fusion results by supplementing color information to the fused image through color space conversion. The color information is mainly preserved in the visible images. We first transfer the visible image from the RGB color space to the YCbCr color space and extract the Cb and Cr channels of the visible image, which contain the color information. Then we concatenate the single-channel fused image with the Cb and Cr channels of the visible image to obtain the 3-channel fused image. Finally, we convert the 3-channel fused image back to the RGB color space to obtain the color fusion result.

8. Limitation

In this work, the total number of experts N, as well as the number of sparsely activated experts K in MoLE and MoGE are selected empirically. The potential of our model may not be fully exploited due to the limitation of the data scale and computational resources. Considering the dynamic scenarios in reality, the empirical selection may restrict the fusion performance in broader scenarios. Therefore, how to adaptively select the most suitable numbers of experts N and sparse activated experts K according to the scenario is still a challenging problem. We will study to overcome it in future work.

References

[1] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 8

[2] 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. 6, 7

[3] 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. 4, 5, 6, 7

[4] 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. 4, 5, 6, 7

[5] 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. 3, 4, 5, 6, 7

[6] 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. 9

[7] 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. 4, 5, 6, 7

[8] 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. 4, 5, 6, 7

[9] 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. 4, 5, 6, 7

[10] 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. 6, 7

[11] 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. 4, 5, 6, 7

[12] 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. 4, 5, 6, 7

[13] 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. 4, 5, 6, 7