Task-Customized Mixture of Adapters for General Image Fusion

Supplementary Material

Abstract

In supplementary material, we provide more details, discussions and experiments for the paper "Task-Customized Mixture of Adapters for General Image Fusion", encompassing the following:

(I) A comprehensive overview of the task-customized loss functions along with associated comparative experiments, as detailed in Sec. 7.

(II) More details about the network. This includes the detailed network structure and data flow of TC-MoA, the implementation method of shifted windows, the ablation studies about network design and the analysis of parameters, elaborated in Sec. 8.

(III) Additional analysis and discussion concerning the properties exhibited by different tasks in the experiments and an exploration into the phenomenon of task-specific routing, explored in Sec. 9.

(IV) More quantitative results on TNO dataset, as shown in Sec. 10. And more qualitative results of fused images of multiple tasks, as presented in Sec. 11.

7. Task-Customized Loss Function

Our network structure accommodates the unique requirements of varied tasks, with tailored unsupervised loss functions for each fusion task. In order to generate high-quality fused images, we impose constraints on the structural ($\mathcal{L}_{ssim}$), intensity ($\mathcal{L}_{Pixel}$), and gradient information ($\mathcal{L}_{Grad}$) of the fused images for different fusion tasks.

Specifically, for the VIF task, our primary goal is to retain the most distinct high-frequency and low-frequency information from the source images. With this regard, we introduce $\mathcal{L}_{MaxPixel}$ and $\mathcal{L}_{MaxGrad}$ loss functions in this task. By employing $\mathcal{L}_{MaxPixel}$ [35] loss function, the fused images have more comprehensive shapes of objects in the dark areas and better color saturation. To maintain gradient information, we ensure the gradient's sign (direction) remains unchanged in all related loss functions to avoid any unintended gradient confusion.

$$\mathcal{L}_V = \mathcal{L}_{aux} + \mathcal{L}_{ssim} + \mathcal{L}_{MaxPixel} + \mathcal{L}_{MaxGrad}, \quad (9)$$

$$\begin{aligned}\mathcal{L}_{ssim} = \lambda_1(1 - SSIM(I_{Fusion}, X)) \\ +\lambda_2(1 - SSIM(I_{Fusion}, Y)),\end{aligned} \quad (10)$$

$$\mathcal{L}_{MaxPixel} = \frac{1}{HW} \left\| I_{Fusion} - max(X, Y) \right\|_1, \quad (11)$$

Task	Task-Customized Loss	Q_{abf}	$\mathcal{VIF}$	SSIM	Q_p	Q_c	Q_{cb}
VIF		0.390	0.553	0.400	0.310	0.553	0.413
	✓	**0.601**	**0.726**	**0.455**	**0.412**	**0.637**	**0.494**
MEF		0.521	0.601	0.939	0.555	0.541	0.396
	✓	**0.645**	**0.661**	**0.964**	**0.598**	**0.578**	**0.431**
MFF		0.473	0.761	0.674	0.482	0.696	0.638
	✓	**0.657**	**0.898**	**0.679**	**0.681**	**0.775**	**0.718**

Table 7. Quantitative results of the task-customized loss functions. The SSIM metric in MEF task is replaced by MEF-SSIM metric.

$$\mathcal{L}_{MaxGrad} = \frac{1}{HW} \left\| \nabla I_{Fusion} - absmax(\nabla X, \nabla Y) \right\|_1, \quad (12)$$

where $\mathcal{L}_{aux}$ is the auxiliary loss to avoid unbalanced learning of adapters. $\mathcal{L}_{ssim}$ represents the loss function based on the structural similarity (SSIM) metric, where λ_1 and λ_2 are set to 0.5. The $mean(\cdot)$, $max(\cdot)$, and $absmax(\cdot)$ represent functions that compute the element-wise average, take the maximum selection, and get the maximum absolute value, respectively. The Sobel operator is denoted as ∇. For gradient-related loss functions like $\mathcal{L}_{MaxGrad}$, we retain the sign of the gradient values.

For the MEF task, the fused images should maintain average luminance levels while retaining all gradient information. This strategy drives us to design $\mathcal{L}_{AvgPixel}$ and $L_{MaxGrad}$ loss functions for the MEF task. Additionally, we adopt $\mathcal{L}_{mefssim}$ which is specifically designed for the MEF task, instead of traditional $\mathcal{L}_{ssim}$.

$$\mathcal{L}_E = \mathcal{L}_{aux} + \mathcal{L}_{mefssim} + \mathcal{L}_{AvgPixel} + \mathcal{L}_{MaxGrad} \quad (13)$$

$$\mathcal{L}_{AvgPixel} = \frac{1}{HW} \left\| I_{Fusion} - mean(X, Y) \right\|_1 \quad (14)$$

For the MFF task, each patch of the fused images tends to depend on one specific source image with the maximum gradient. This prevents the objects' edges in defocused areas from being preserved, thereby affecting the quality of the fused images. In practicality, we select only one source patch to calculate the loss function for each patch in fused images. Therefore, the $\mathcal{L}_{MaskPixel}$ and $\mathcal{L}_{MaskGrad}$ loss functions for MFF are designed as,

$$\mathcal{L}_F = \mathcal{L}_{aux} + \mathcal{L}_{ssim} + \mathcal{L}_{MaskPixel} + \mathcal{L}_{MaskGrad} \quad (15)$$

$$\mathcal{L}_{MaskPixel} = \sum_{i=1}^{2} M_i \circ \left\| I_{Fusion} - I_i \right\|_1 \quad (16)$$

$$\mathcal{L}_{MaskGrad} = \sum_{i=1}^{2} M_i \circ \left\| \nabla I_{Fusion} - \nabla I_i \right\|_1 \quad (17)$$

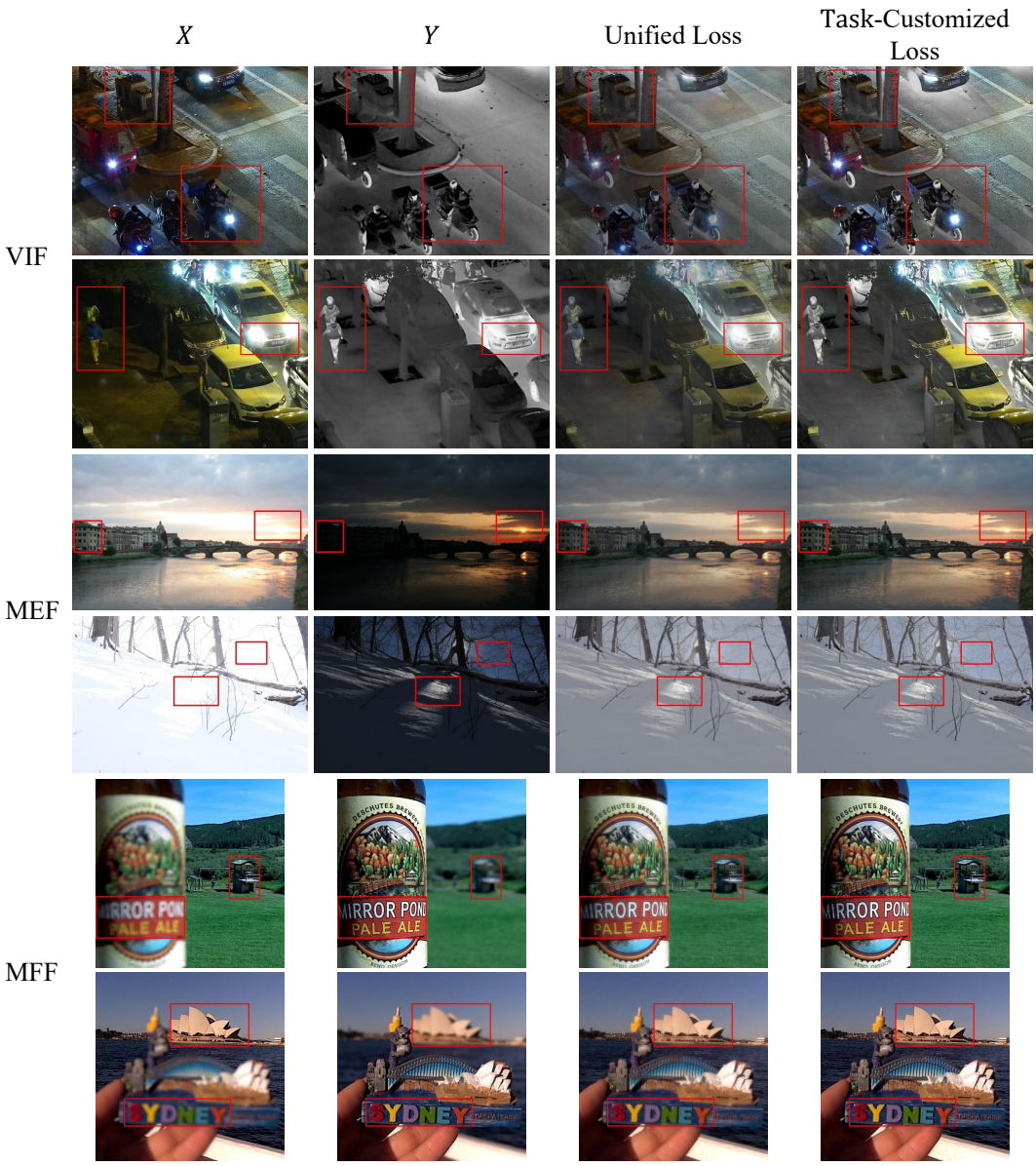

Figure 9. Qualitative comparisons on task-customized loss functions.

where M represents the mask map $M \in \mathbb{R}^{pH \times pW \times 1}$, and i denotes the index of image in source images tuple (X, Y). If the patch's maximum gradient in current image exceeds the other image, we allot the mask map value of the location of this patch as 1. Conversely, it is set to 0.

Qualitative and Quantitative Comparisons. We compare the model trained with task-customized loss functions and that trained with unified loss function. The unified loss function follows the PMGI combined with $\mathcal{L}_{ssim}$. The quantitative results are shown in Table 7, and the qualitative results are reported in Fig. 9. The quantitative results show that the fused images obtained by our task-customized loss functions are rich in high frequency and structural details, conforming to human perception across all tasks. The qualitative results show that our fused images display superior contrast, color saturation, and textural detail when compared to images produced by unified loss function. It is worth noting that previous MFF methods often require post-processing on the fused images or features to obtain clear images, but our model can directly reconstruct the fused images with task-customized loss functions.

8. Details of Network.

Detail Architecture of TC-MoA. We illustrate the detailed network structure, data flow of the prompt generation and prompt-driven fusion stages of TC-MoA in Fig. 10.

Details on Fused Results. With each passing TC-MoA module, the fusion features are added to each net-

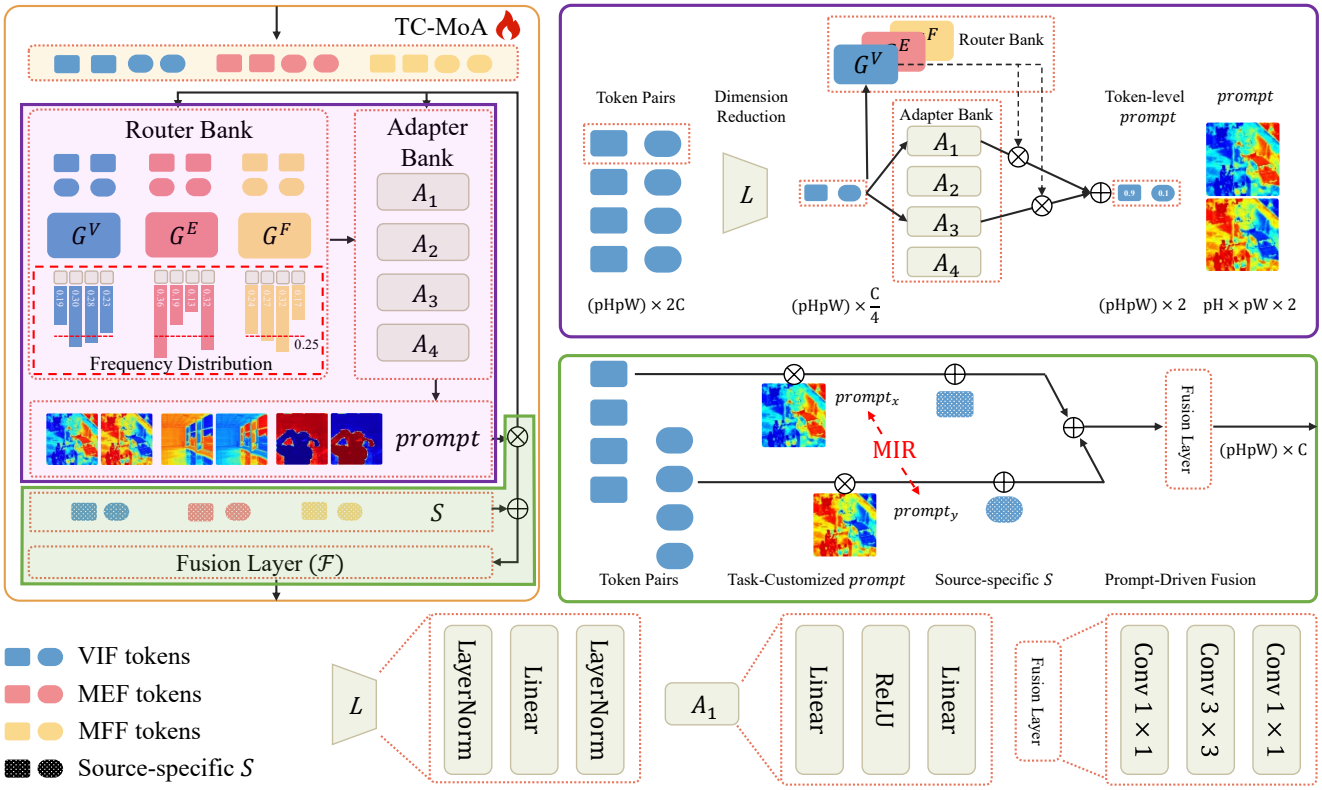

Figure 10. Detail architecture for TC-MoA. The purple box represents the process of generating task-customized prompt, while the green box illustrates the prompt-driven fusion process. The MIR stands for mutual information regularization constraint. The routing results of all samples for a specific task by a task-specific router constitute the frequency distribution of the mixture of adapters.

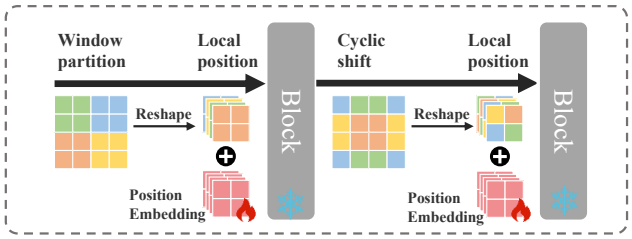

Figure 11. Workflow of the *shifted windows*.

work branch according to hyperparameter settings. We consider that the features in each branch increasingly resemble those of the fused image, and ultimately the outputs of the two branches to be approximately identical, thus arbitrarily choosing one branch to obtain the fused image.

Shifted Windows. We integrate windows shifting into the frozen ViT blocks by incorporating learnable relative position embeddings, as shown in Fig. 11. To reduce the computational cost of transformer, most previous approaches only support fixed-size inputs. These methods employ additional pre-processing and post-processing steps to crop and stitch the image together, resulting in the checkerboard artifacts. To address this issue, we partition the features into multiple windows of 14×14 patches (to maintain consistency with the token length during pre-training). Then, we introduce learnable local position embeddings to enable the model's perception of token positions within the windows, ensuring the model is spatially aware of the local windows. Subsequently, we apply cyclic shifts to other blocks, acquiring a global receptive field. To this end, *shifted windows* not only allows for efficiently handling of different input sizes, but also captures the global receptive field of the image to effectively avoid checkerboard artifacts.

Ablation Studies about Network Design. We explore the effect of three network architectures in our framework. The qualitative results are shown in Fig. 12. i) Predicting images in isolated windows neglects the inter-window information exchange, thereby yielding inconsistencies across the entire image. Our method enables the interaction of information across the entire image by shifting windows, thereby maintaining the overall coherence of the image. ii) The transformers are not able to effectively transmit information to adjacent patches solely based on long-range dependencies, resulting in the occurrence of the fusion image checkerboard effect. To address this, we introduce local receptive fields via convolutional layers to enhance the interaction of local features and alleviate the checkerboard effect between adjacent patches. iii) Using only gradient constraints with absolute values can lead the network to blindly learn the gradients of the image, resulting in the generation of fusion images with incorrect gradient directions, which do not

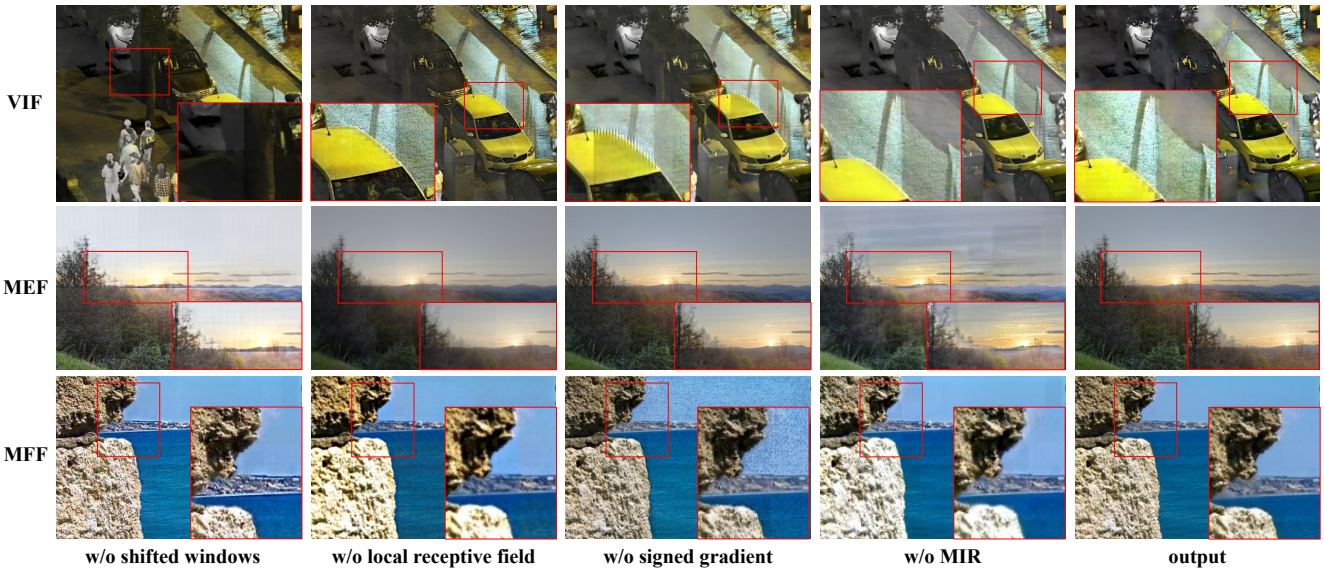

Figure 12. Ablation studies about network design.

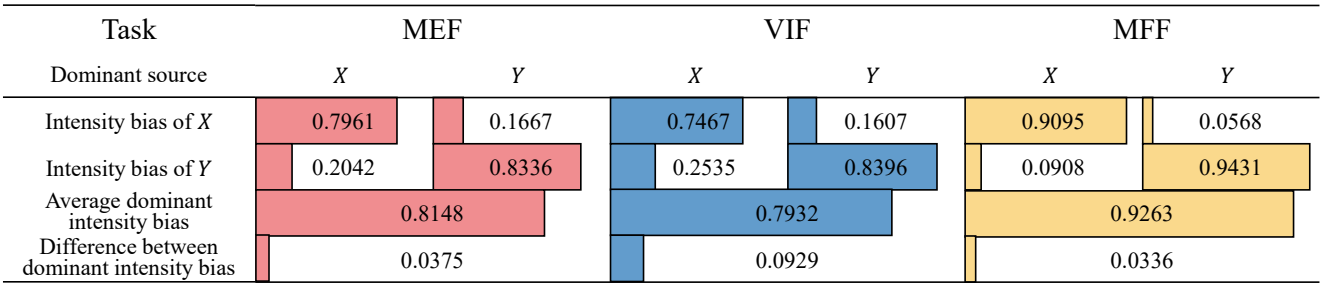

Figure 13. Statistical analysis of intensity bias in different tasks. The "Average dominant intensity bias" refers to the average value of the dominant intensity bias for the X and Y sources. The "Difference between dominant intensity bias" denotes to the absolute value of the difference between the dominant intensity bias of the X and Y sources.

All Parameters			
348.7 M			
Frozen	Trainable		
	9.58 M (2.82%)		
	Position Embedding	TC-MoA	
		9.39 M (2.77%)	
329.54 M (97.18%)	0.19 M (0.06%)	Prompt Generation	Prompt-Driven Fusion
		2.15 M (0.63%)	7.24 M (2.13%)

Table 8. Detailed statistics of the parameters. The data is sourced from the TC-MoA model based on ViT-large network structure with $\tau = 4$ and $N = 4$.

align with human visual perception. The loss function based on the signed value of the gradient avoids confusion in the direction of the gradient.

More Analysis of Parameters. Our network is an efficient parameter fine-tuning method inserted into the frozen ViT framework with pre-trained parameters. As depicted in Table 8, we use mere 2.77 % trainable parameters to bridge the gap between pre-trained model and image fusion tasks. The introduced shifted windows only demand 0.06 % parameters, rendering them practically negligible. In TC-MoA, although we have multiple routing networks and adapters, a mere 0.63% parameters suffice to generate prompt efficiently. In fact, most of the parameters are contributed by the convolutional layers in the fusion layer. In addition, our model offers the potential to further reduce the number of parameters by compressing the number of channels in the convolutional layers.

9. Additional Analysis and Discussion

The Properties of Various Fusion Tasks. The properties of a fusion task, as well as the relationship between multiple tasks, can be depicted through the task prompt. For each token pair, if $prompt_x > prompt_y$, then we refer to X as the dominant source and Y as the auxiliary source. The dominant intensity bias of X is defined as the average value of $prompt_x$ for all token pairs where X dominates. Conversely, the auxiliary intensity bias of X is the mean value of $prompt_x$ for all token pairs where X is the auxil-

Table 9. Quantitative results of the VIF task on TNO dataset.

Method	$\mathcal{VIF}$	Q_c	EN	SD	$Q_{cv}\downarrow$	MS-SSIM	FMI	Q_w
DeFusion [20] [ECCV'22]	0.513	0.569	6.579	8.862	500.767	0.840	0.900	0.563
DDFM [55] [ICCV'23]	0.276	0.390	6.853	9.219	976.884	0.685	0.878	0.294
MoE-Fusion [3] [ICCV'23]	**0.757**	0.541	**7.008**	9.158	743.774	**0.901**	0.907	0.770
TC-MoA *Base*	0.748	**0.631**	7.003	**9.335**	**393.571**	0.895	**0.911**	**0.770**
TC-MoA *Large*	**0.793**	**0.659**	**7.026**	**9.392**	**414.068**	**0.908**	**0.910**	**0.772**

Table 10. Ablation experiments on the value of K.

Hyperparameter	VIF			MEF			MFF		
	Q_{abf}	Q_p	SSIM	Q_{abf}	Q_p	SSIM	Q_{abf}	Q_p	SSIM
$K=1$	0.608	0.731	0.456	0.655	0.897	**0.679**	**0.652**	**0.674**	0.413
$K=2$	0.608	**0.739**	**0.458**	0.656	0.896	0.678	0.649	0.665	0.411
$K=3$	**0.610**	0.734	**0.457**	**0.656**	**0.898**	0.679	0.652	0.669	**0.413**
$K=4$	**0.609**	**0.736**	0.457	**0.658**	**0.897**	**0.680**	**0.653**	**0.669**	**0.413**

iary source. As shown in Fig. 13, for an instance, the dominant and auxiliary intensity biases of X for the VIF task are 0.7467 and 0.1607, respectively.

Fig. 13 reveals two fusion patterns: 1) The mean dominant intensity bias for MFF is higher than those of MEF and VIF. This indicates that fused images from MFF tasks typically draw information heavily from one source image per token pair, rendering the fusion for MFF extremely unbalanced at the token-level. In contrast, MEF and VIF display a more balanced fusion. From a token-level fusion perspective, MEF and VIF tasks share more similarity. 2) The difference between dominant intensity bias of VIF is significantly larger than that of the other fusion tasks. For the VIF task, when the infrared images (Y source) dominate, the dominant intensity bias is found to greatly surpass that of the visible images (X source). Once the network identifies the dominance of infrared images, it assigns higher weights to its features. This results in VIF being more unbalanced at the source-level, while MEF and MFF exhibit more balance. We believe that part of the reason for this phenomenon is that the LLVIP dataset for the VIF task, composed primarily of nocturnal scenes, where infrared images might provide more information than visible images.

To this end, the average dominant intensity bias reflects the fusion pattern of the tasks at the token-level, while the difference between dominant intensity bias illustrates the pattern at the source-level. The empirical evidence of our method shows variations and connections among these fusion tasks, thus demonstrating our effectiveness in handling multiple fusion tasks with a unified model.

Details of Task-Specific Routing. Different task-specific routers indeed customized various mixtures of adapters, as shown in Fig. 14. The following patterns can be observed: 1) For the VIF task, if source X is the dominant source on one token pair, it is inclined to utilize the yellow adapter for processing. Conversely, the green adapter is primarily used when source Y predominates. 2) For the MEF task, the blue adapter is used when source Y is the dominant source. Other colored adapters are utilized under conditions where source X is dominant. 3) For the MFF task, the yellow adapter tends to deal with cases where X dominates, while red and green tend to address situations where Y is dominant. Notably, high-frequency areas are more frequently handled by the red adapter, while low-frequency areas are inclined to handle by the green adapter.

Obviously, by task-specific routing, the network tends to select different mixtures of adapters to accommodate varying tasks. Therefore, these adapters have task tendencies and different divisions of labor (such as high and low frequency works). This is an interesting finding that can be further explored for more controllable fusions.

Analysis on Top K. We performed ablation experiments on the value of K under four adapters, as shown in Table 10. The following two findings can be observed: i) the more the number of experts chosen for routing, the better the overall performance. Intuitively, the more adapters are routed to, the greater the dynamism of the network, and consequently the better the performance, which is consistent with empirical results. ii) The choice of K has a minor impact on performance, indicating that the network is not sensitive to this parameter. In summary, we have struck a balance between performance and inference cost by setting Top $K=2$.

10. More Quantitative Comparisons

TNO Dataset. Table 9 shows additional results on TNO dataset. Our method achieves superior overall performance compared with the most recent methods.

11. More Qualitative Comparisons

More qualitative results of various fusion tasks are presented in Figs. 15 to 17. In the case of the VIF task, our fusion results maintain the detailed information of the visible images to the greatest extent (such as license plate

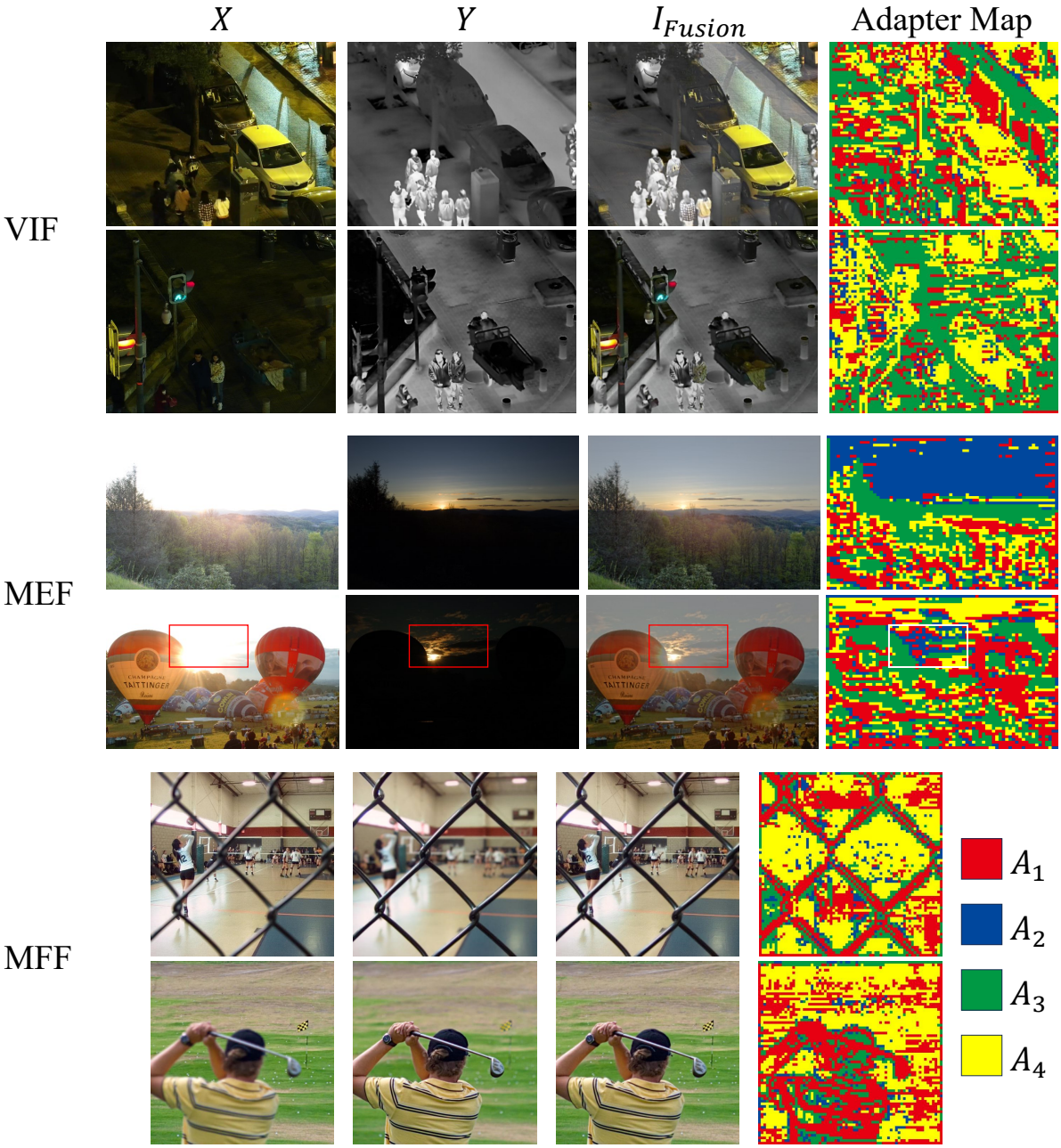

Figure 14. The visualization of adapter selection for the last TC-MoA in the ViT encoder. We visualize the index of the adapters with the highest routing weight as the adapter map. Different colors in the adapter map represent the selection of different adapters. It is worth noting that this is the adapters selection situation of the last TC-MoA in the encoder, which can only reflect the general trend and may contain unclear noise.

numbers), while clearly highlighting the information from the infrared images. The overall images possess excellent brightness and contrast. For the MEF task, our fusion results adequately preserve the detailed structural and color information from both sources, especially avoiding blurry halos in overexposed areas. In terms of the MFF task, our model avoids distortions in structural details and color, particularly in the areas with text. In terms of the overall visual perception, our method is comparable to the IFCNN on MFF, a supervised training method. In addition, our method provides a substantial degree of fusion controllability, making it practicable to manipulate fusion results based on particular requirements.

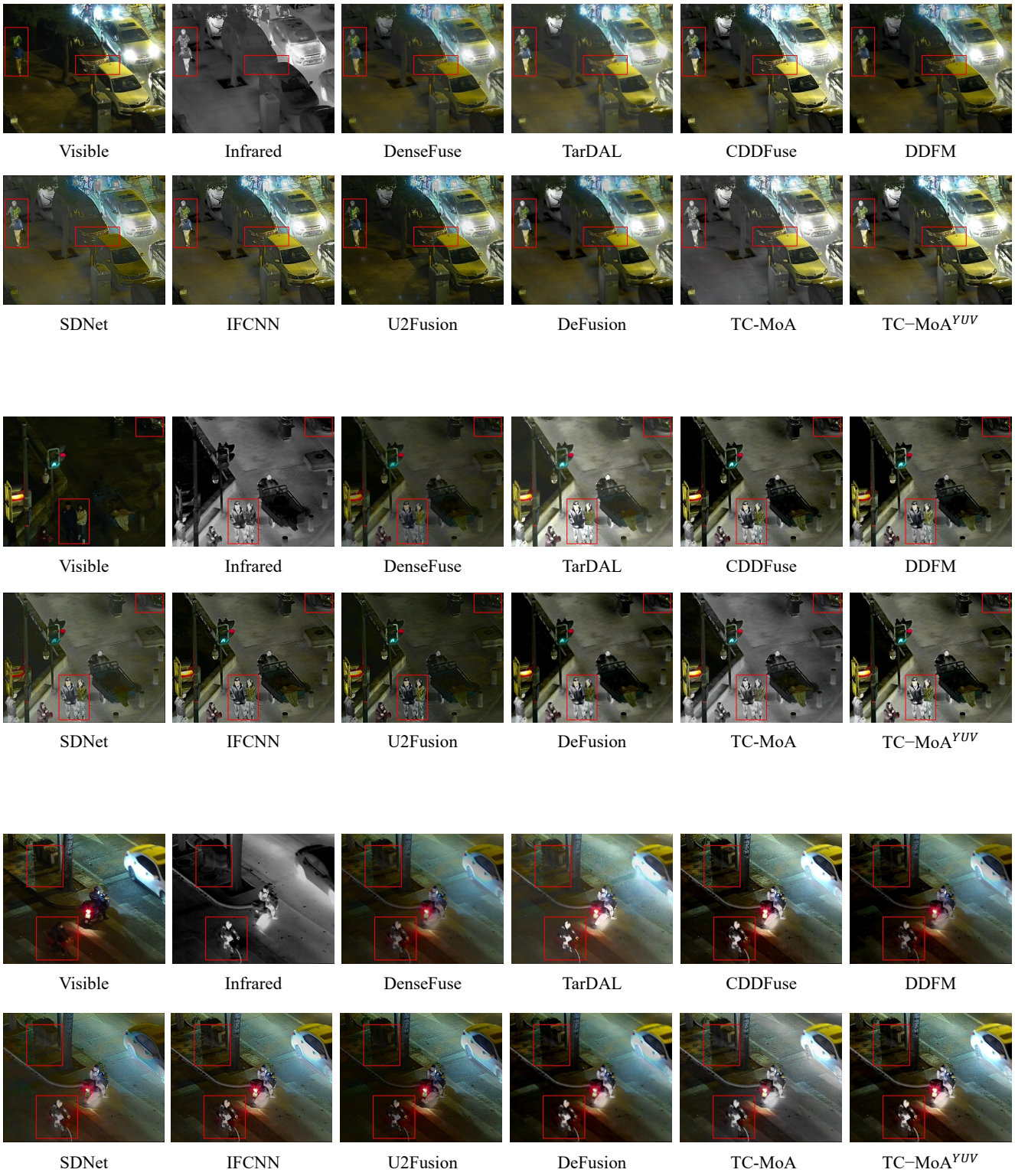

Figure 15. More qualitative comparisons in the VIF task.

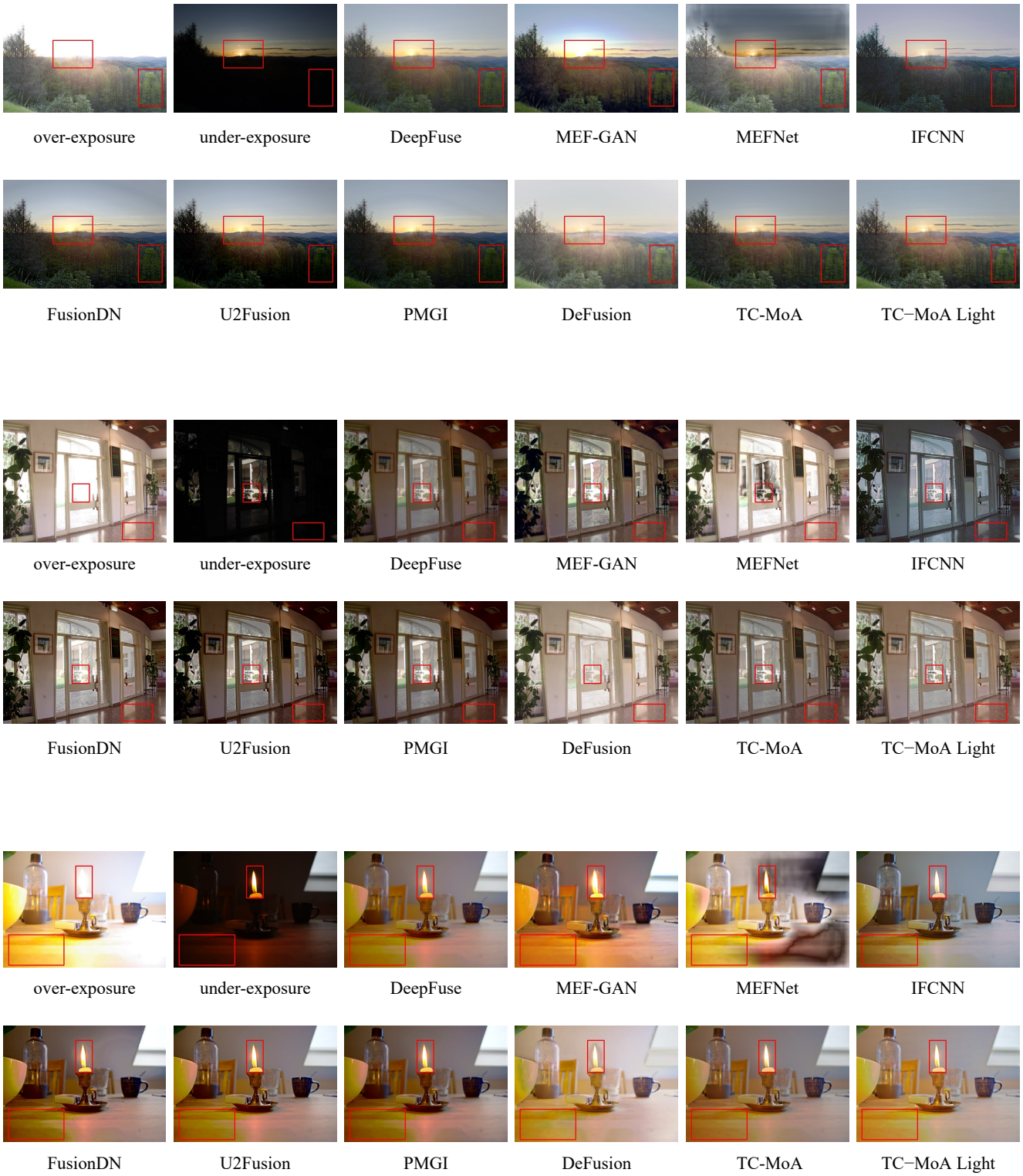

Figure 16. More qualitative comparisons in the MEF task.

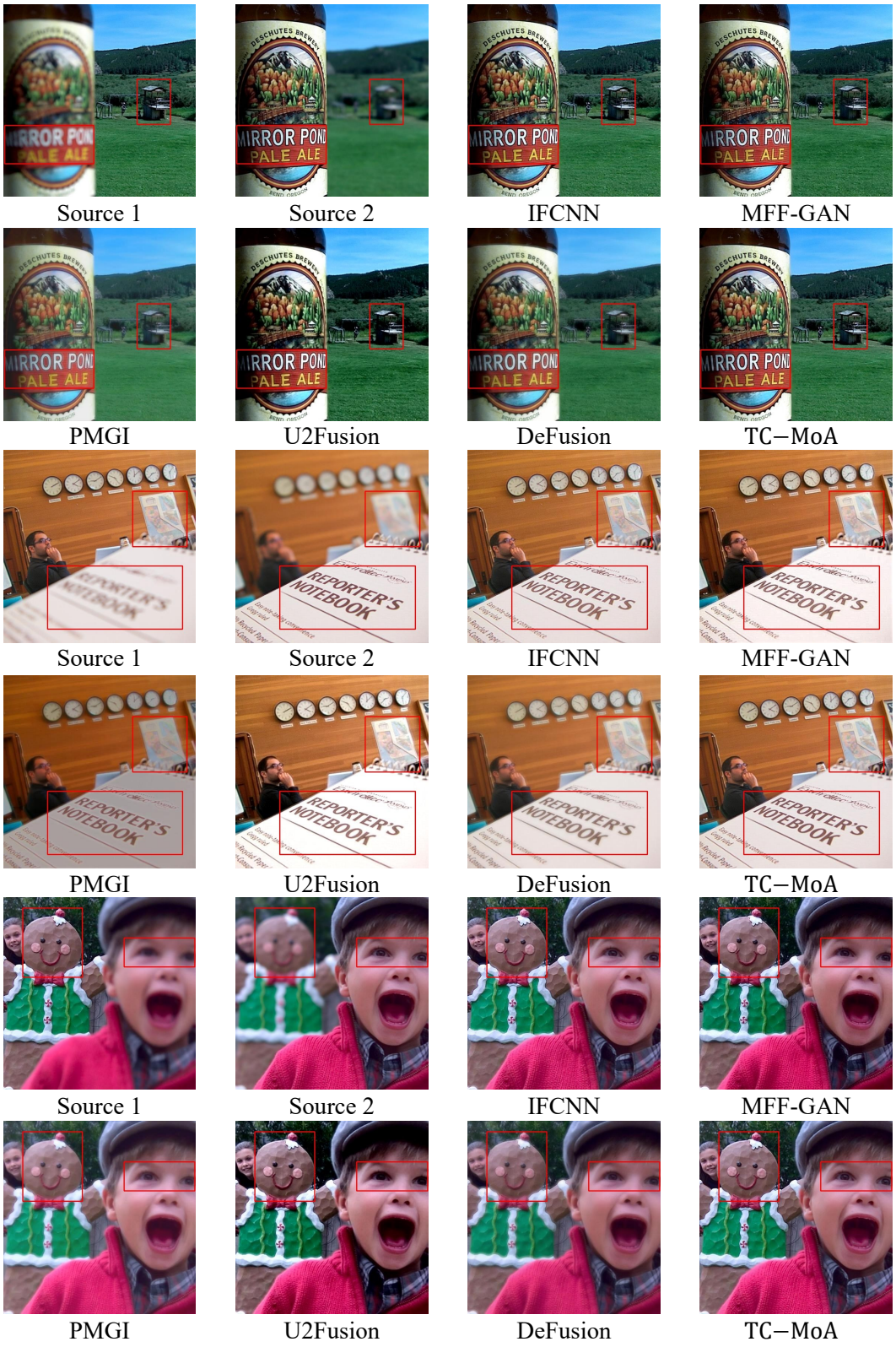

Figure 17. More qualitative comparisons in the MFF task.