Fusion from Decomposition: A Self-Supervised Decomposition Approach for Image Fusion (Supplementary Material)

Anonymous ECCV submission

Paper ID 4260

1 Implementation Details

Network Architecture. The whole network follows the U-net architecture as shown in Fig.A 1. The basic module is the ResNeSt block [4]. Specifically, we first employ a simple 3×3 convolutional layer followed by a ReLU layer to transform the source images to obtain the shallow feature representations. And then, we feed these features to the following encoder network E. There are two decoders in the network, denoted as D^u, D^c. The decoders are similar to the encoder, except that the pooling layers are replaced with the bilinear upsample layers. As for the additional encoder E_c, it only has one ResNeSt block. The block numbers and output channel numbers are presented in the figure for better understanding.

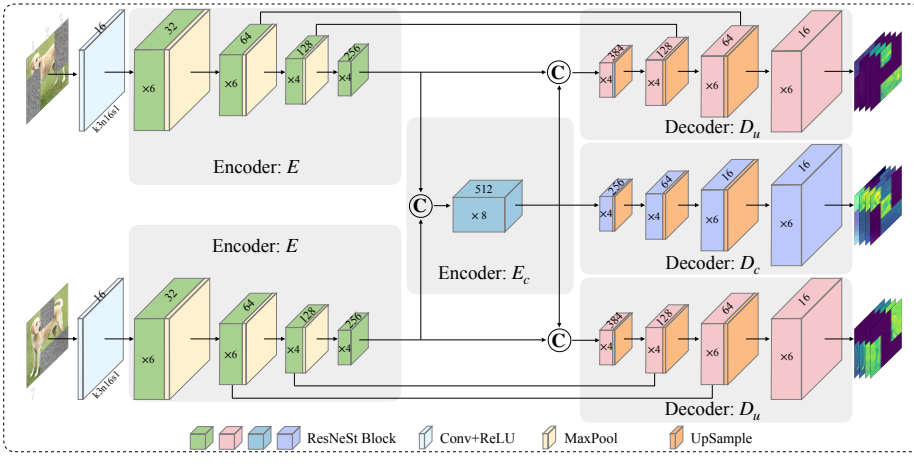

Fig.A 1: The details of network architecture.

Experimental Settings. As briefly introduced in the manuscript, the mask M has the same size as the training image and consists of two types of patches with different sizes, *i.e.*10×10 and 20×20 pixels, respectively. In Tab.A 1, we present more details about the experimental settings. In the training stage, the generated masks perform great diversity due to the randomness introduced in the

preprocessing step. Specifically, the mask ratio of each source image is randomly drawn from a uniform distribution in $\mathcal{U}(0.55, 1.0)$, so that the *common ratio* (i.e. the proportion of the area of common regions to the area of the whole image) varies from 0.01 to 1.

In addition to parameter settings, some tricks are used to improve the training stability. We found that it is hard for the network to converge for directly training the model with four loss items. Therefore, at the early 3,000 iterations, we ask that only the losses of projected images from common and unique components are calculated while the reconstruction losses of P_r are excluded from the loss computation.

Config	Value
Common Ratio	(0.01, 1.0)
Mask Ratio	(0.55, 1.0)
Mask Resolution	$f10\times10,\ 20\times20g$
RandomCrop	(3/4, 4/3)
RandomResize	(0.08, 1.0)

Tab.A 1: Training settings of the data preprocessing.

2 Ablation Studies

To explore the importance of the self-supervised learning with CUD pretext task, we perform comprehensive ablation studies. In general, the ablation studies can be classified into two categories: mask-related and loss-related.

Mask-related Ablation Studies. In the manuscript, the image augumatation is defined as follows:

$$\boldsymbol{x}^i = M_i(\boldsymbol{x}) + \bar{M}_i(n), \tag{1}$$

where n denotes the Gaussian noise, $\boldsymbol{x}$ is original scene, and M_i is the mask. Obviously, the noise patch $\bar{M}_i(n)$ is independent to the counterpart (region) of the other image. To verify the effectiveness of noise patches, we replace the noise patches with zero patches in the ablation study. Recently, the masked image modeling [2] demonstrates that the patch size of the mask is related to the performance of the model to some extent. To verify this opinion, we also conduct an another simple ablation study about the patch size of mask M.

We show qualitative results of mask-related ablation studies in the multi-focus image fusion task in Fig.A 2. The fused result of Fig.A 2c is generated by the model trained with zero patches. As can be seen, the model performs worse than the DeFusion shown in Fig.A 2e. Considering that the performance of multi-focus image fusion mainly depends on the quality of unique information, we can infer that using the noise patches in image transformation may be more conducive to learning unique semantic information. The results of the quantitative evaluation are consistent with the qualitative results. Compared with the performance of DeFuion, the ablation study using the zero patches shows lower

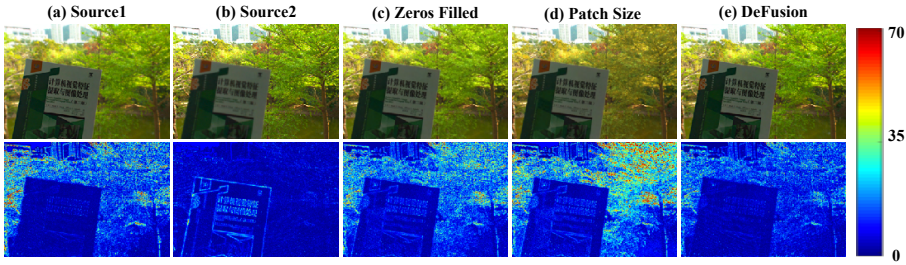

Fig.A 2: Qualitative comparisons of mask-related ablation studies. The (c) is generated by the model whose transformed source images replace the noise patches with zero patches. The resolution of mask patch size is the combination of 16×16 and 32×32 pixels in (d), while the default resolution of DeFusion is the {10×10, 20×20} pixels.

metrics (PSNR:-2.24dB, SSIM:-0.017). The fused results of the model trained with bigger patch sizes are shown in Fig.A 2d. It also produces the worse fused result in this example. The model that uses data with the bigger noise patch to train may generate inappropriately semantic information about the scene for the leaf in Fig.A 2d is incorrectly colored into yellow rather than green. Furthermore, the quantitative evaluation show similar results. Compared with the performance of DeFuion, the ablation study using bigger noise patches shows lower metrics (PSNR:-3.41dB, SSIM:-0.031).

Loss-related Ablation Studies. To verify the effectiveness of decomposition components, we conduct some ablation studies about the decomposed common and unique components. Specifically, we design four groups of ablation studies to analyze the disparity in fusion results of the same example when we remove parts of reconstruction losses. In addition, during training the DeFusion, the ground truth of the projected image $\hat{\boldsymbol{x}}_c = P_c(f_c)$ is the $\boldsymbol{x}_c = M_1(\boldsymbol{x}) \cap M_2(\boldsymbol{x})$, where the projected images are asked not only to keep the useful information in the common regions of source images, but also to predict the zero values in the semantically irrelevant regions (*i.e.*, unique regions). In a similar vein, the $\hat{\boldsymbol{x}}_c = P_c(f_c)$ is also required to predict the zero values in the common regions, except for keeping the unique semantic information of source images. The reason why the network needs to predict the zero values at the semantically irrelevant regions is because we hope that the extracted common information can be more accurate. In order to verify the effectiveness of the setting, we remove the constraint that predicting zero values at the semantically irrelevant regions, and only focus on keeping the information in the masked regions.

As shown in Fig.A 3, we employ the multi-exposure image fusion task to verify the capability of each reconstruction loss item. The first row in Fig.A 3 is generated by the model trained without the common construction loss. We can see that the whole visualized common feature map is activated. As for the second row, since we remove the unique reconstruction loss item during training,

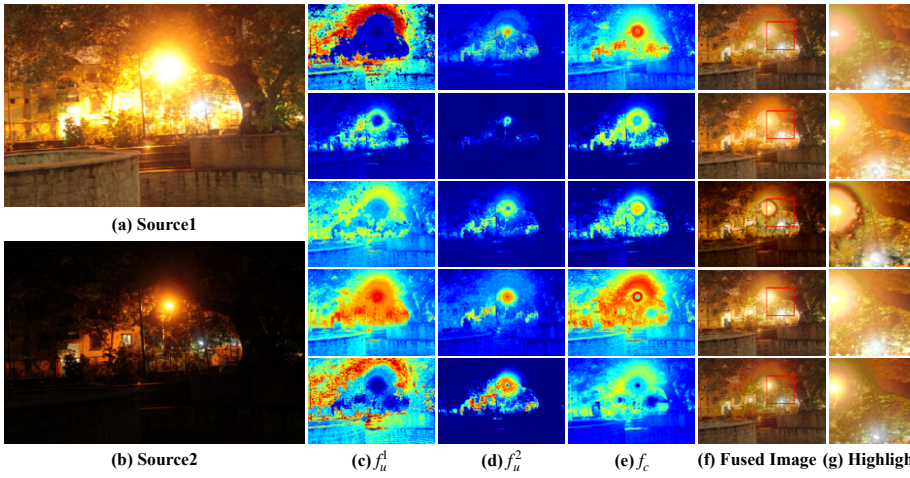

Fig.A 3: Qualitative comparisons of loss-related ablation studies. (First row: training without common reconstruction loss; second row: training without unique reconstruction losses; third row: training with only original scene reconstruction loss; fourth row: training with only predicting the masked regions; fifth row: DeFusion in the manuscript.)

the two visualized unique feature maps look similar. It denotes that both the over-exposed and under-exposed images all focus on the same region in the extracted unique features, which may be redundant for fusion. In the third row, we only keep one loss to reconstruct the original scene during training and remove the other reconstruction losses. In this case, the quality of feature representation obviously decreases, *e.g.*the visualized unique feature map f_u^1 exhibits some artifacts that even present in the final fusion results. In addition, the above-mentioned three ablation studies all show inconsistent edge. The worst edge artifacts are generated by the model train without any common or unique reconstruction loss. As for the fourth row, the corresponding model is trained to focus on keeping the information in the masked region. Compared to the DeFusion models, the visualized feature maps in this case cannot distinguish the common and unique region, so that the corresponding fused result preserves insufficient details of under-exposed image while remaining too much brightness information of over-exposed image.

3 Fused Results

In this section, we present more qualitative fused results on both the MEFB benchmark [6] and SICE dataset [1] for the multi-exposure task, as shown in Fig.A 4. The Fig.A 5 presents more qualitative fused results on the Real-MFF dataset [5] for the multi-focus image fusion task. The Fig.A 6 presents more qualitative fused results on the RoadScene dataset [3] for the visible-infrared image fusion task.

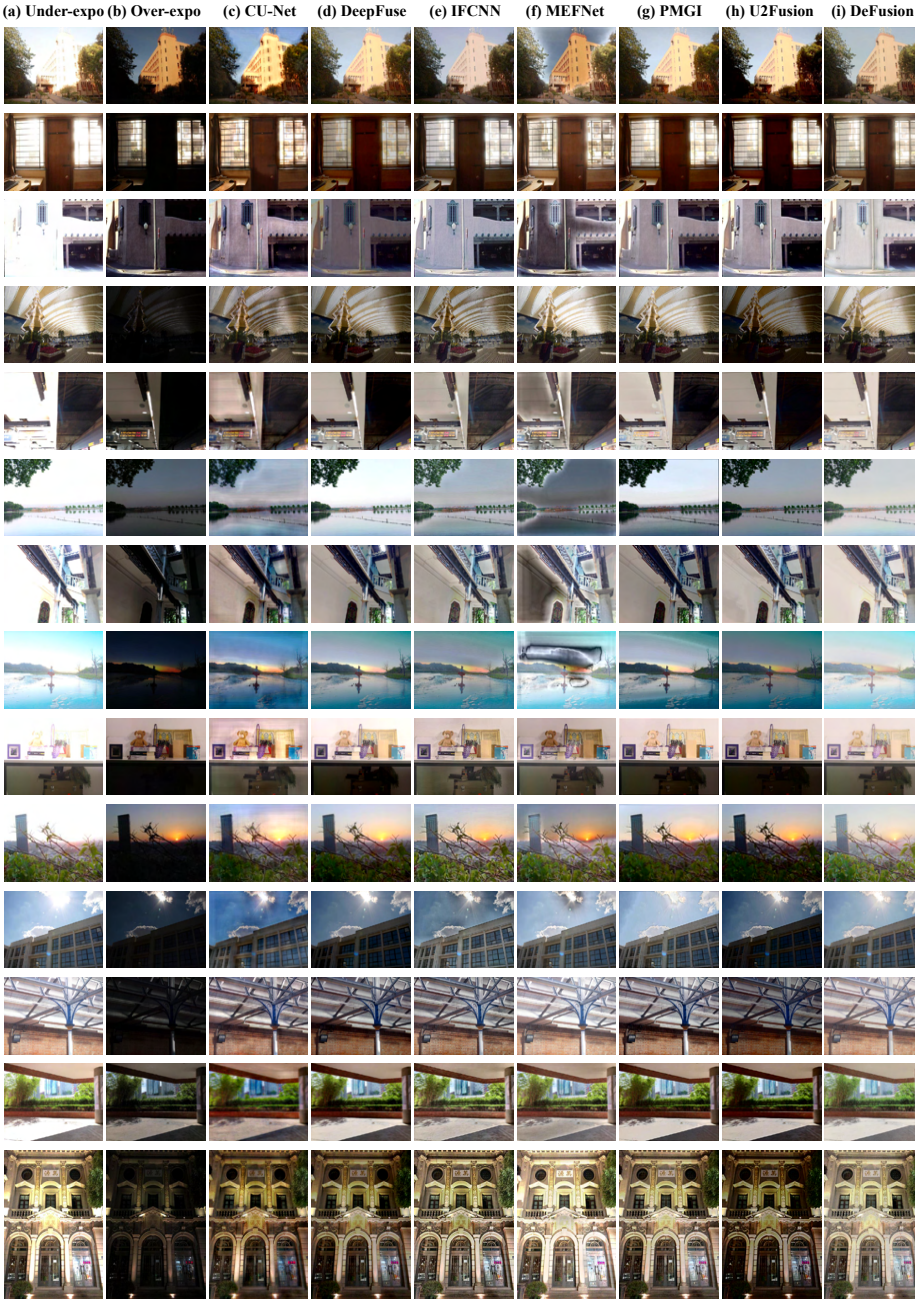

Fig.A 4: Fused results on the multi-exposure image fusion task.

Fig.A 5: Fused results on the multi-focus image fusion task. We provide the enhanced residual maps for each result of comparison and input images to highlight the difference with GT.

Fig.A 6: Fused results on the visible infrared image fusion task.

References

1. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing **27**(4), 2049–2062 (2018)
2. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. arXiv preprint arXiv:2111.09886 (2021)
3. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2Fusion: A unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
4. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Muller, J., Manmatha, R., Li, M., Smola, A.: ResNeSt: Split-attention networks. arXiv preprint arXiv:2004.08955 (2020)
5. Zhang, J., Liao, Q., Liu, S., Ma, H., Yang, W., Xue, J.H.: Real-MFF: A large realistic multi-focus image dataset with ground truth. Pattern Recognition Letters **138**, 370–377 (2020)
6. Zhang, X.: Benchmarking and comparing multi-exposure image fusion algorithms. Information Fusion (2021)