Letter

NormFuse: Infrared and Visible Image Fusion With Pixel-Adaptive Normalization

Quan Kong, Huabing Zhou, and Yuntao Wu

Dear Editor,

This letter presents a normalization mechanism to effectively fuse infrared and visible images in an encoder-decoder network. Source images are decomposed into source-invariant structure and source-specific detail features. Then, the information of detail features is sufficiently incorporated into the structure features using this normalization mechanism in the decoder, which generates high-contrast fused images with highlighted targets and abundant texture information. Qualitative and quantitative experiments on two challenging datasets demonstrate the superiority of our method over current state-of-the-art methods.

Infrared and visible image fusion (IVIF) is a representative example of image fusion and has wide applications in computer vision, target recognition, and military tasks. Traditional IVIF methods, such as multiscale transform-based methods [1], [2], usually design feature extraction and fusion strategies in a manual way and thus cannot fully capture and understand all source information. Deep learning (DL) based methods [3]–[8] have been trying to overcome these shortcomings of traditional methods. The core of DL methods is to design deep neural networks and introduce loss functions to train the networks, guiding them to extract deep features and then fuse them automatically. For example, in disentangled representation-based DL methods: DIDFuse [5] and DRF [6], source images are decomposed into source-invariant structure features and source-specific detail features via an encoder. Then, the decomposed features are concatenated and fed into a decoder to generate fused images.

However, these DL methods simply adopt concatenation fusion strategy at the reconstruction step. Not leveraging the relationship between different types of extracted features, the networks have to learn to synthesize them only under the guidance of the loss functions, often resulting in unsatisfied effectiveness, such as decreased saliency of highlighted targets and bad resolution of detail textures. To address these issues, we propose a novel fusion network under a normalization mechanism to fuse disentangled source-invariant structure and source-specific detail features, which is termed NormFuse in this letter. The main contributions of this letter are summarized as follows: 1) To the best of our knowledge, this is the first time to treat IVIF as a normalization process in which the structure features are modulated by modulation parameters converted from detail features. 2) Different from existing normalization mechanisms, we propose a pixel-adaptive normalization module, thus each activation of the structure feature maps can be modulated by different modulation values adaptively. 3) The proposed method can help efficiently fuse the disentangled structure and detail information, generating superior fused images over existing comparable state-of-the-art methods.

Problem analysis: A paired infrared and visible images must share the same source-invariant structure part of the scene and meanwhile each of them possesses a source-specific detail part of the scene. The process of disentangling the structure from the detail is invertible and the source image can be fully reconstructed with disentangled structure and detail parts. Conditional normalization, such as adaptive instance normalization (AdaIN) [9] in style transfer and spatially-adaptive normalization (SPADE) [10] in semantic image synthesis, is an effective kind of mechanism to synthesize disentangled representations. In conditional normalization, activations of input features are firstly normalized to zero mean and unit deviation and then denormalized by modulating the feature activations via affine transformation. Because the modulation parameters are converted from external condition information, the condition information is reasonably incorporated into the input features. Normalization mechanism exhibits superior performance than concatenation strategy in both tasks [9], [10].

Inspired by the above analysis, we treat IVIF as a normalization process in which the information of detail features is incorporated into structure features as external condition information. However, AdaIN only uses global statistics as the modulation parameters and thus cannot preserve complete conditional information, which is just the requirement in IVIF tasks. While in SPADE, the modulation parameters mainly depend on semantic classes, not varying spatially in a pixel-wise manner. To completely incorporate the disentangled detail information into structure features and accomplish fully spatial adaptiveness, in this letter we propose pixel-adaptive normalization (PEAN) mechanism, as illustrated in Fig. 1.

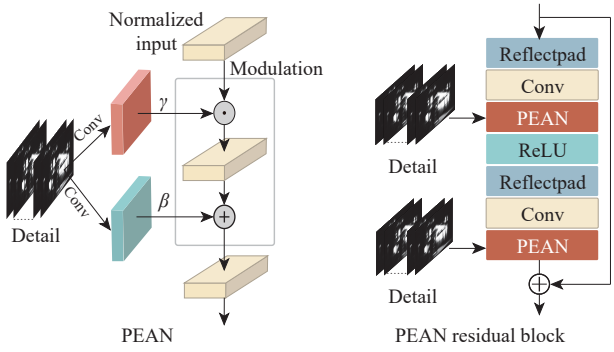

Fig. 1. Illustration diagrams of (left) pixel-adaptive normalization (PEAN) layer and (right) the structure of PEAN residual block (PEANResBlk).

We use convolution layers to compute the modulation parameters γ and β, which have the same size as the structure S and detail D feature maps. Therefore, each activation of structure features can be modulated pixel-wise. Let h^i denote the activations of input (structure) features with the height H^i and width W^i of the i-th layer of a deep convolutional network for a batch of N samples. Let C^i be the number of channels in the layer. The normalized activation values are directly rescaled and biased with γ and β in a channel-wise manner respectively ($n \in N$, $c \in C^i$, $y \in H^i$, $x \in W^i$)

$$\gamma^i_{c,y,x}(D)\frac{h^i_{n,c,y,x}-\mu^i_c}{\sigma^i_c}+\beta^i_{c,y,x}(D) \tag{1}$$

where $h^i_{n,c,y,x}$ is the activation before batch normalization (BN) and μ^i_c and σ^i_c are the mean and deviation of the activations in channel c: $\mu^i_c=\frac{1}{NH^iW^i}\sum_{n,y,x}h^i_{n,c,y,x}$ and $\sigma^i_c=\sqrt{\frac{1}{NH^iW^i}\sum_{n,y,x}(h^i_{n,c,y,x}-\mu^i_c)^2}$.

Network architecture: Fig. 2 illustrates our IVIF network architecture based on PEAN. There are three convolution layers and two residual blocks (ResBlk) in the encoder. The decoder consists of two PEANResBlk and one convolution layer. The structure of PEANResBlk is shown in the right of Fig. 1. The outputs of the two ResBlks are concatenated with the outputs of the two PEANResBlks along channels, respectively. The size of all feature maps is kept the same

Corresponding author: Yuntao Wu.

Citation: Q. Kong, H. B. Zhou, and Y. T. Wu, "NormFuse: Infrared and visible image fusion with pixel-adaptive normalization," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 12, pp. 2190–2192, Dec. 2022.

Q. Kong is with the School of Art and Design and School of Computer Science and Engineering Artificial Intelligence, Wuhan Institute of Technology, Wuhan 430205, China (e-mail: witkongquan@wit.edu.cn).

H. B. Zhou and Y. T. Wu are with the School of Computer Science and Engineering Artificial Intelligence and Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430205, China (e-mail: zhouhuabing@gmail.com; ytwu@wit.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.106112

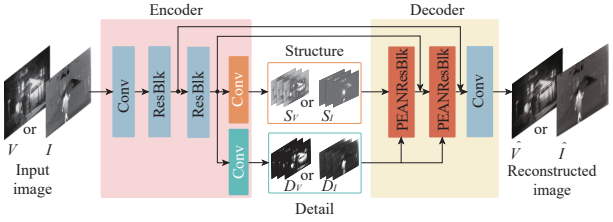

Fig. 2. The encoder-decoder network of NormFuse.

as the input images.

In the training phase, an infrared I or a visible V image is fed to the encoder and decomposed into structure features and modal detail features. Then, the structure feature maps are directly fed to the decoder, whereas the detail feature maps are converted into parameters to modulate the structure feature maps through PEANResBlks. Consequently, the decoder generates the reconstructed image. While in the testing phase, we firstly sum the structure (detail) feature maps from different sources channel-wise: $S_F = S_I + S_V$ ($D_F = D_I + D_V$). Then, a fused image is generated through the decoder with the input of S_F and D_F.

Our loss function consists of decomposition loss and reconstruction loss: $L_{\text{total}} = L_{de} + L_{rec}$. To minimize the difference between the infrared and visible structure feature maps and maximize the difference between their detail feature maps, the decomposition loss is

$$L_{de} = \Phi\Big(\sum_c \|S_I - S_V\|_2^2\Big) - \alpha_1 \Phi\Big(\sum_c \|D_I - D_V\|_2^2\Big). \tag{2}$$

Φ is tanh function. The decoder aims to reconstruct the infrared and visible images $\hat{I}$, $\hat{V}$ as close to I, V as possible in pixel intensity, structural similarity index (SSIM). The abundant gradient information from visible images is expected to retain. Thus the reconstruction loss is

$$L_{rec} = \alpha_2 f(I, \hat{I}) + \alpha_3 f(V, \hat{V}) + \alpha_4 \|\nabla V - \nabla \hat{V}\|_1. \tag{3}$$

Here ∇ denotes the gradient operator, and $f(X, \hat{X}) = \|X - \hat{X}\|_2^2 + \lambda L_{\text{SSIM}}(X, \hat{X})$, where $X = I$ or V and $L_{\text{SSIM}}(X, \hat{X}) = [1 - \text{SSIM}(X, \hat{X})]/2$.

Experiments: The hyperparameters are set as: $\alpha_1 = 0.5$, $\alpha_2 = 2$, $\alpha_3 = 3$, $\alpha_4 = 20$ and $\lambda = 5$. In training phase, the network is optimized by Adam over 120 epochs with a batch size of 24. The learning rate is set to 10^{-3} and decreased by 10 times every 40 epochs.

We conduct experiments on TNO [11] and RoadScene [12] datasets. The training set contains 180 RoadScene image pairs. The test set contains 38 TNO and 37 RoadScene image pairs. Before training, all images are transformed into grayscale and center-cropped with 128×128 pixels. We compare our method with HMSD [1], HMSD_GF [2], FusionGAN [4] and DIDFuse [5] qualitatively and quantitatively according to entropy (EN), standard deviation (SD), spatial frequency (SF), visual information fidelity (VIF), mutual information (MI) and mean gradient (MG) metrics.

To examine the decomposition effect, we visualize one channel of the structure and detail feature maps in Fig. 3. S_V and S_I are overall similar, reflecting the fundamental structure information of the same scene. Obtain exactly similar structures from distinct source images is very difficult due to their vastly distinct image representations. In contrast, D_V and D_I are remarkably different and mainly reflect their complementary modal information. For the top row results, D_V mainly contains information of the path and the ground uncovered by grass, while D_I mainly contains information of the truck and bright spots. The obvious benefit of such complementarity is that their simple addition makes all source-specific valuable information retained and highlighted in subsequent fusion process.

To understand how the PEAN module works, we visualize one channel of $BN(h)$, γ, β, and the multiplication of γ with $BN(h)$ in the second PEANResBlk in the reconstruction process of visible images. As shown in Fig. 4, although γ and β are converted from the same D_V, they are very different. For example, in the top row, the roof, road and lamp are prominent in γ while dark in β. Meanwhile, γ and β modulate $BN(h)$ in different ways. By rescaling $BN(h)$ with γ, the information of the roof, salient in both γ and $BN(h)$, is enhanced,

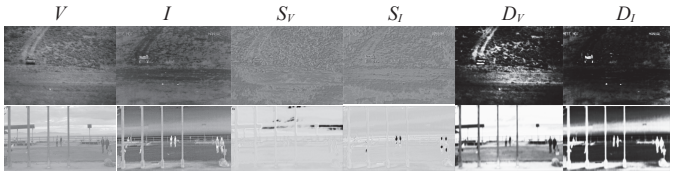

Fig. 3. Visualization of one channel of the disentangled structure and the detail feature maps. Top: TNO test dataset and bottom: RoadScene test dataset.

while the information of the road and lamp, only prominent in γ, is suppressed. β contains the information of the pedestrian, the umbrella, and the trees, which are missing in $\gamma BN(h)$. The information contained in β is a needful complement to $\gamma BN(h)$.

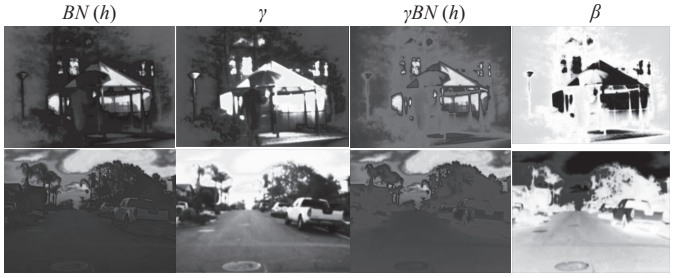

Fig. 4. Visualization of one channel of $BN(h)$, γ, $\gamma BN(h)$, and β in the reconstruction of visible images.

Fusion results: Figs. 5 and 6 show the six metrics for TNO and RoadScene test datasets, respectively. Overall, NormFuse achieves the best results on SD, VIF, MI, and EN. Particularly, it performs much better than all other methods on SD and VIF. For MG and SF, NormFuse obtains comparable results on TNO datasets. These results demonstrate that our method provides the highest contrast, the least distortion, and the most information. We attribute this superiority of NormFuse to the effective fusion of the structure and all the modal information under normalization mechanism.

We further exhibit five representative fused images generated by

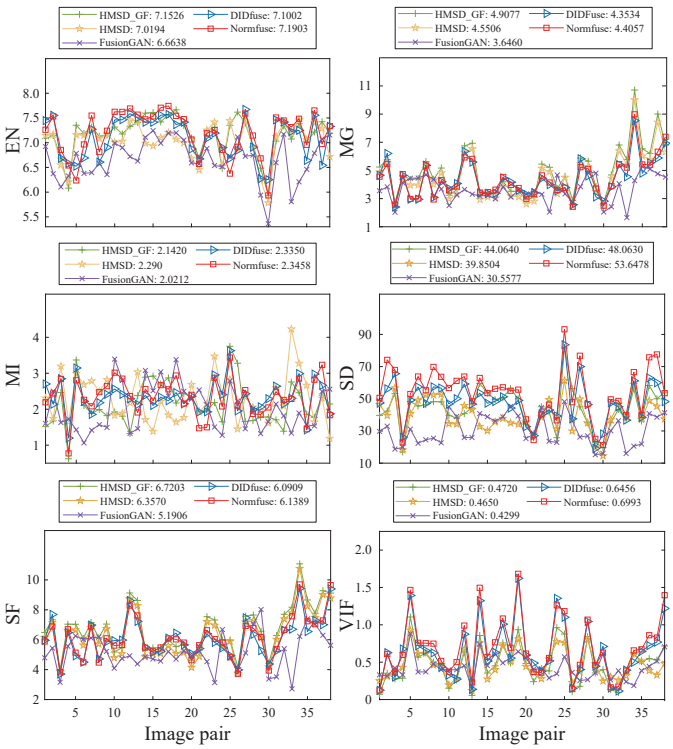

Fig. 5. Quantitative comparison of different methods on six metrics for TNO test dataset.

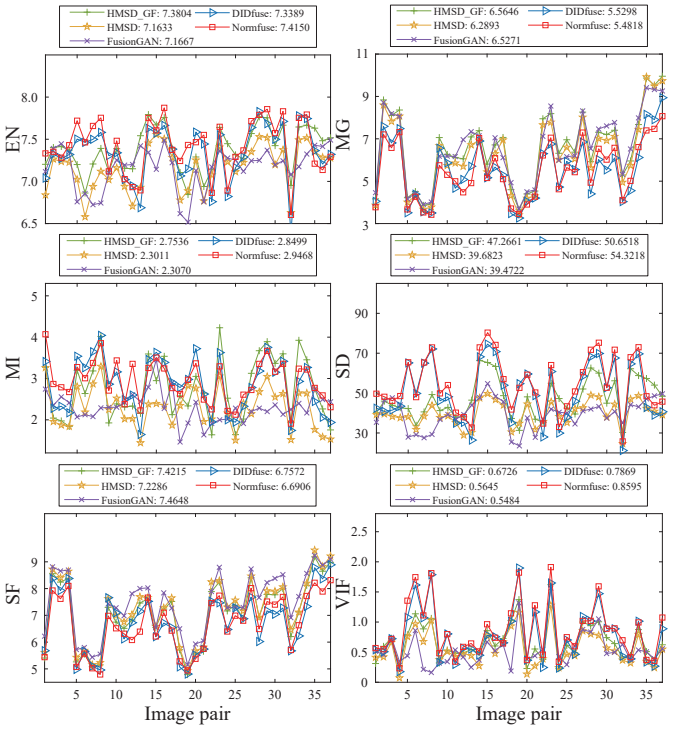

Fig. 6. Quantitative comparison of different methods on six metrics for RoadScene test dataset.

different methods in Fig. 7. Compared to other methods, NormFuse is superior in three aspects. First, NormFuse generates more highlighted targets, such as the bright light spot, the head of the car and the street lamp in the first, third and fifth columns respectively. Second, NormFuse achieves higher contrast. As shown in the left three columns, the contrast of the wall, the bunker and the body of the soldier in NormFuse are stronger than those in other methods, which are more likely to attract human attention and provide better visual effect. Third, our results have better information fidelity. In the fourth column, the sky contains a gradual transition from the bright to the dark regions in NormFuse, reflecting more real information from both sources. While the transition is very sharp in DIDFuse or even does not occur in other methods.

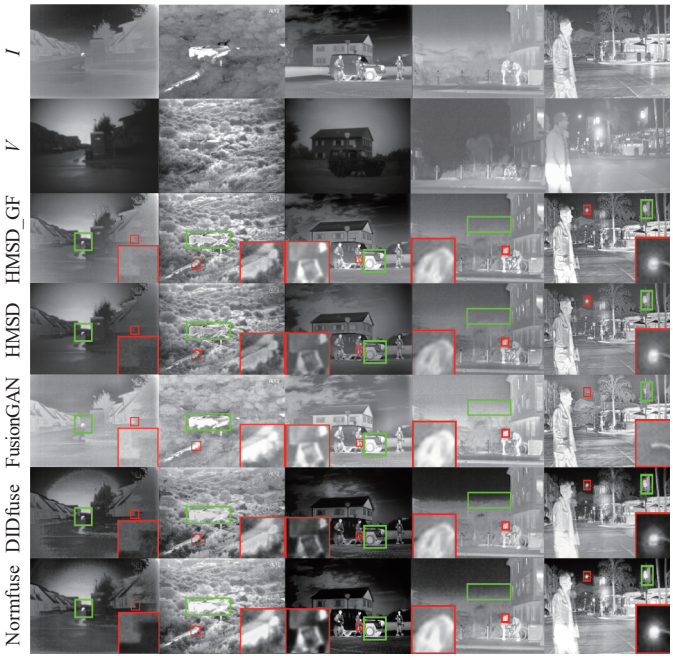

Fig. 7. Qualitative results for different methods. The left three columns: TNO test dataset and the right two columns: RoadScene test dataset.

To directly verify the effectiveness of PEAN module, we conduct ablation study where the PEAN modules are replaced by BN and the detail and structure features are concatenated as input to the decoder. We refer to this method as ResCat. As shown in Table 1, NormFuse exceeds ResCat on EN, SD, VIF and MI, and obtains very close results on SF and MG, verifying the effectiveness of PEAN mechanism over concatenation strategy.

Table 1. Quantitative Comparison Between ResCat and NormFuse. Better Values are Shown in Bold

	TNO dataset		RoadScene dataset	
Metrics	ResCat	NormFuse	ResCat	NormFuse
EN	7.1512	**7.1903**	7.4020	**7.4150**
SD	47.5677	**53.6478**	51.1525	**54.3218**
SF	**6.2642**	6.1389	**6.7738**	6.6906
VIF	0.6376	**0.6993**	0.8248	**0.8595**
MI	2.2054	**2.3458**	2.7438	**2.9468**
MG	**4.4193**	4.4057	**5.5790**	5.4818

Conclusion: In this letter, we leverage the relationship between the disentangled structure and detail features transformed from infrared and visible source images and employ normalization mechanism to fuse the disentangled features. To guarantee effective information fusion, we propose pixel-adaptive normalization as the fusion mechanism. An encoder-decoder network based on this mechanism generates high-quality fusion images, providing a new benchmark for IVIF tasks.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (62171327, 61771353), the first batch of application basic technology and science research foundation in Hubei Nuclear Power Operation Engineering Technology Research Center (B210610), and the Hubei Three Gorges Laboratory Open Fund (SC215001).

References

[1] Z. Zhou, W. Bo, L. Sun, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Inform. Fusion*, vol. 30, pp. 15–26, 2016.

[2] Z. Zhou, M. Dong, X. Xie, and Z. Gao, "Fusion of infrared and visible images for night-vision context enhancement," *Appl. Opt.* , vol. 55, pp. 6480–6490, 2016.

[3] H. Li and X. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, 2018.

[4] J. Ma, W. Yu, Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inform. Fusion*, vol. 48, pp. 11–26, 2019.

[5] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, and P. Li, "DIDFuse: Deep image decomposition for infrared and visible image fusion," in *Proc. IJCAI*, 2020, pp. 970–976.

[6] H. Xu, X. Wang, and J. Ma., " DRF: Distengled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[7] L. Tang, J. Yuan, and J. Ma., "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inform. Fusion*, vol. 82, pp. 28–42, 2022.

[8] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[9] X. Huang and S. Belongie, " Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV*, 2017, pp. 1510–1519.

[10] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. CVPR*, 2019, pp. 2332–2341.

[11] A. Toet and M. Hogervorst, "Progress in color night vision," *Opt. Eng.*, vol. 51, no. 1, pp. 1–20, 2012.

[12] H. Xu, J. Ma, Z. Le, and X. Guo, " FusionDN: A unified densely connected network for image fusion," in *Proc. AAAI*, vol. 34, no. 7, pp. 12484–12491, 2020. DOI: 10.1609/aaai.v34i07.6936.