Letter

CDP-GAN: Near-Infrared and Visible Image Fusion Via Color Distribution Preserved GAN

Jun Chen, Kangle Wu, Yang Yu, and Linbo Luo

Dear Editor,

This letter is concerned with dealing with the great discrepancy between near-infrared (NIR) and visible (VS) image fusion via color distribution preserved generative adversarial network (CDP-GAN). Different from the global discriminator in prior GAN, conflict of preserving NIR details and VS color is resolved by introducing an attention guidance mechanism into the discriminator. Moreover, perceptual loss with adaptive weights increases quality of high-frequency features and helps to eliminate noise appeared in VS image. Finally, experiments are given to validate the proposed method.

VS images appear poor details or visual effects in non-ideal lighting conditions such as low light, haze or noisy conditions due to the dependence on object reflection and scene illumination [1]. To solve this problem, the straightforward method is to increase the sensitive area of sensors. However, this requires higher hardware requirements and the improvement of image shadow effect is limited to a very low degree. Another solution is to obtain high signal instantaneous ratio NIR gray image by adding NIR complementary light, and then fuse it with color VS image. The NIR and VS image fusion aims to generate a clean image with considerable detail information, which has important application values in low illumination fields such as night monitoring.

Different fusion schemes have been developed to exploit combining complementary information in VS and NIR images, such as multi-scale decomposition-based, regularization-based and mathematical statistics-based fusion methods. However, most existing methods use manual methods to design transmission model. To improve the fusion effect, the design process is becoming more and more complex. Consequently, it is harder to avoid the implementation and computational efficiency. On the contrary, image fusion methods based on deep learning mostly adopt end-to-end model, which can directly generate fused images using inputs without complicated activity level measurements and fusion rules [2]. Moreover, the deep network is trained with a large number of source images and thus more informative features with specific characteristics could be extracted.

In general, deep learning-based methods have been proved to be effective for image fusion whereas deep learning-based NIR and VS image fusion is seldom studied as far as we have acknowledged. To a certain extent, infrared (IR) and VS image fusion is similar to task of NIR and VS image fusion, for they both motivate to transfer the useful information in IR/NIR image to VS image. And yet, the former aims to transfer the local salient region in IR image, so little attention is paid to preserving color appearance of VS image. On the contrary, NIR and VS image fusion methods are designed to transfer the global details [3]. Therefore, color appearance of the VS color image will inevitably be disturbed while IR and VS image fusion methods are directly applied to fuse NIR and VS images. To verify the above analysis, we present a representative experiment in Fig. 1. The third image is obtained by a recent deep learning-based unified image fusion method (termed as IFCNN). It fails to retain the color appearance compared to the VS image and serious color distortion occurs in its fusion result. Most existing NIR and VS image fusion methods fail to balance these two tasks thus similar failure as IFCNN occurs in their result. Fortunately, due to the capability to fit multiple distribution characteristics [4], the generative adversarial network (GAN) could achieve these two goals and preserve the distribution of detail and color information simultaneously.

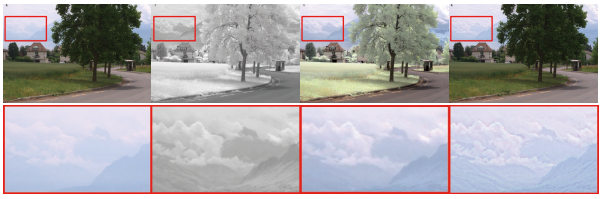

Fig. 1. Schematic illustration of image fusion. From left to right: VS image, NIR image, fusion result of a recent deep learning-based unified image fusion method [5], and fusion result of our proposed CDP-GAN.

On the basis of above analysis, we propose a GAN-based NIR and VS image fusion architecture with color distribution preservation, termed as CDP-GAN. We formulate NIR and VS image fusion as an adversarial game between generator and discriminator. The generator aims to generate a fused image that contains considerate details, whereas the discriminator attempts to constrain the fused image to have similar pixel intensity as noise-free VS image so that final fused image would not suffer from color distortion. To prevent details from being blurred by the global discriminator, an easily realized attention mechanism is combined with discriminator, which enforces discriminator to pay more attention to VS region while ignoring NIR texture region. Therefore, CDP-GAN could produce a noise-free result that not only contain rich textures, but also retain the high-fidelity color information compared with VS image.

Problem formulation: Given a pair of pre-registration source images, purposes of NIR and VS image fusion are to preserve the useful detail information of both source images and recover the color information of the VS image. In this work, we formulate NIR and VS image fusion problem as an adversarial process between the former two purposes. The schematic framework is shown in Fig. 2. At first, we concat the luminance channel of VS image $I_v^{(l)}$ and NIR image and then input them into the generator. After feature extraction and reconstruction in the generator, the initial fused 1-channel image $I_f^{(l)}$ is obtained as output. Having been restrained by the adaptive perceptual loss, the initial $I_f^{(l)}$ preserves considerable edges and textures which are visible in source images. However, owing to embedding of the NIR image, $I_f^{(l)}$ tends to appear divergent pixel distribution with luminance channel of VS image $I_v^{(l)}$, which will lead to color distortion in the final color result. An initial solution is inputting $I_f^{(l)}$ and $I_v^{(l)}$ into the discriminator and establishing an adversarial relationship between the generator and discriminator by an adversarial constraint loss function. However, low-quality VS

Corresponding author: Jun Chen.

Citation: J. Chen, K. L. Wu, Y. Yu, and L. B. Luo, "CDP-GAN: Near-infrared and visible image fusion via color distribution preserved GAN," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 9, pp. 1698–1701, Sept. 2022.

J. Chen and K. L. Wu are with the School of Automation, Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, China University of Geosciences, Wuhan 430074, China (e-mail: chenjun71983@163.com; wukangle@cug.edu.cn).

Y. Yu is with the Shanghai Institute of Technical Physics, Key Laboratory of Infrared System Detecting and Imaging Technology, Chinese Academy of Sciences, Shanghai 200083, China (e-mail: yuyang@mail.sitp.ac.cn).

L. B. Luo is with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China (e-mail: luolb@hotmail.com).

Digital Object Identifier 10.1109/JAS.2022.105818

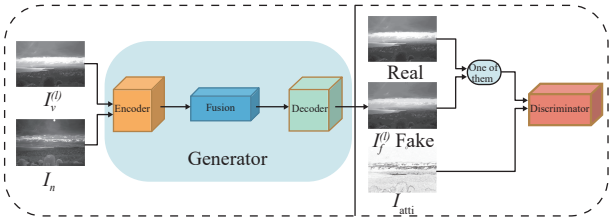

Fig. 2. Framework of the proposed CDP-GAN for NIR and VS image fusion.

images would introduce much noise into $I_f^{(l)}$. Therefore, the mean filtering is performed as initial denoising to obtain the noise-free VS image $I_d^{(l)}$. While training, the discriminator works to distinguish $I_f^{(l)}$ from $I_d^{(l)}$ as much as possible and the generator acts to generate the most realistic $I_f^{(l)}$ to fool discriminator. Pixel intensity of $I_f^{(l)}$ will be gradually more and more similar to $I_d^{(l)}$ in the adversarial process. Consequently, the final color image reconstructed by $I_f^{(l)}$ and chrominance channel of I_v will appear natural and similar color information as the color VS image. Noting that the adopted color space is YCbCr in which color components Cb and Cr represent blue and red chromaticity components, respectively [6]. Moreover, in order to prevent details of generated results from being blurred, we introduce an attention-guided architecture into the discriminator.

Network architecture: As shown in Fig. 3, the generator of CDP-GAN contains two encoder networks, a feature fusion module and a decoder network. To reduce computational complexity while training, these two encoder networks have the same architecture that consists of 4 convolution layers that adopt 3×3 filter consistently, with weights shared between them. In order to reduce gradient loss, compensate for feature loss and reuse previously calculated features, the encoder adopts dense connection and establishes short direct connection between each layer and all layers in feedforward mode [7]. The batch normalization is adopted on the heels of each layer to quicken the training and avoid gradient explosion [8]. The stride of each convolution layer is empirically set to 1. To avoid dropping out detail information in the source images, we remove downsampling operation in convolution. Moreover, all convolution layers except the last one use leaky ReLU activation function which could deal with dying neurons in traditional ReLU and speed up the convergence.

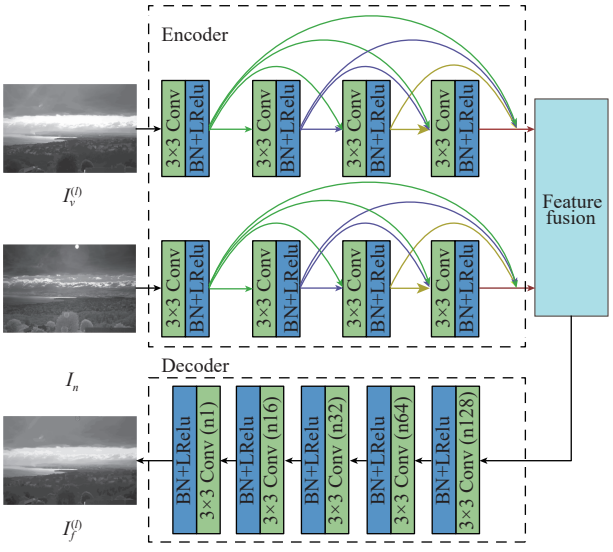

Fig. 3. The overall architecture of generator.

The main framework of the proposed discriminator is the same as that in the original GAN which acts as a classifier to distinguish whether input image comes from $I_v^{(l)}$ or $I_d^{(l)}$ data distribution. The discriminator is mainly composed of 4 convolution layers and a full connection layer. The stride of all layers is set to 2 thus the size of extracted feature maps gradually halves. Being influenced by the discriminator, generated result $I_f^{(l)}$ will be gradually closer to $I_v^{(l)}$. However, we might as well summarize the role of the generator as embedding useful detail information in $I_v^{(l)}$ and I_n into the generated result $I_f^{(l)}$. The global discriminator treats every pixel in $I_f^{(l)}$ fairly and thus textures and edges that come from NIR image will be blurred.

To handle problems of detail loss caused by the global discriminator, we introduce an easy-to-use attention mechanism for the discriminator. As shown in Fig. 4, along with input images, a corresponding attention map is also fed into the discriminator and then multiplied with all feature maps before every convolution layer in channel dimension. Noting that the downsampling operation is adopted before the latter three convolution layers to fit each feature map. The attention map is calculated ahead as follows:

$$I_{\text{atti}} = 1 - \nabla I_n. \tag{1}$$

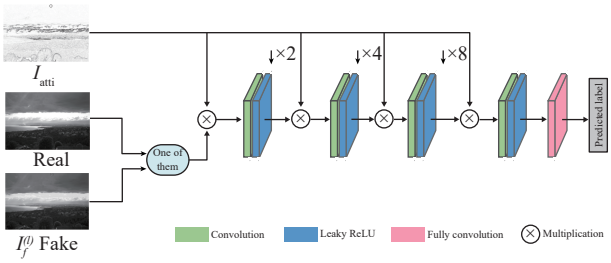

Fig. 4. The overall architecture of discriminator.

The attention map I_{atti} is reflection of gradient map of I_n. Noting that values in ∇I_n are scaled to range $[0,1]$. Larger gradient values mean more detailed information in this local area in I_n and little attention of discriminator should be paid here for preventing details coming from I_n. Guided by the attention map, convolution layers pay more attention to the feature extraction of non-NIR feature areas and thus NIR detail loss is eliminated.

Loss function: The loss function of generator in this letter consists of two parts: content loss L_{con} and adversarial loss $L_{adv}(G)$.

$$L_g = L_{con} + \lambda L_{adv}(G). \tag{2}$$

where hyperparameter λ is introduced to trade-off influence of generator and discriminator. For context loss, pixel-wise loss is widely used and calculates the loss between pixels of the generated and target images. It makes the generated result have a high signal-to-noise ratio (PSNR) but lack of high-frequency information, which results in over-smooth texture. The generated image should be close to input images in terms of both low-level pixel values and high-level abstract features. Moreover, the NIR and VS image have great discrepancy that cannot be handled by pixel-wise loss.

To preserve both high-frequency and low-frequency information of source images, we introduce the perceptual loss as our loss function. Thanks to the large and diverse datasets, other visual tasks such as target recognition and segmentation CNN models have more powerful feature extraction ability. Thus, we adopt the VGG-16 network which has been well trained with ImageNet dataset to extract features. At first, we duplicate the input into 3 channels and then feed them into the VGG16. Feature maps are separated before the pooling layer and their number of channels gradually doubles whereas size halves. $\varphi_j(I)$ denotes feature maps of input I before the j pooling layer. To prevent introduction of noise and loss of details, we choose $\varphi_3(I)$ which has the closest size to the input in the content loss L_{con}. In generator, fused result $I_f^{(l)}$ is reconstructed from $I_v^{(l)}$ and I_n by constraining distance of feature map φ_3 between inputs and output. Thus, the content loss L_{con} is formulated as follows:

$$L_{con}=\frac{1}{N}\times\sum_{i=1}^{N}[w^i_{I_v^{(l)}}\times\|\varphi^i_3(f)-\varphi^i_3(I_v^{(l)})\| + w^i_{I_n}\times\|\varphi^i_3(f)-\varphi^i_3(I_n)\|] \tag{3}$$

where $N=256$ denotes the number of channels. $w^i_{I_v^{(l)}}$ and $w^i_{I_n}$ present the weight parameters of the i-th feature map for $I_v^{(l)}$ and I_n. Obviously, it is difficult to set desired values of $w^i_{I_v^{(l)}}$ and $w^i_{I_n}$ that are suitable for all the feature maps. Thus, we design an adaptive strategy for $w^i_{I_v^{(l)}}$ and $w^i_{I_n}$ which aims to assign larger values for the feature map with considerable useful information. For two feature maps of $I_v^{(l)}$ and I_n in the i-th channel, we attempt to measure the information contained in them and their gradients are adopted for measurement. Compared with entities in general information theory, image gradient is a small receptive field measure based on local spatial structure. While used in deep learning framework, gradients are more efficient in both computation and storage. Therefore, they are more suitable for CNN information measurement. Comparing amounts of information, the trade-off weights are defined by weighted averaging as follows:

$$w^i_{I_v^{(l)}}=\frac{\left\|\nabla\varphi^i_3(I_v^{(l)})\right\|}{\left\|\nabla\varphi^i_3(I_v^{(l)})\right\|+\left\|\nabla\varphi^i_3(I_n)\right\|}$$
$$w^i_{I_n}=1-w^i_{I_v^{(l)}}. \tag{4}$$

In CDP-GAN, we introduce the attention-guided discriminator which acts to keep similar pixel distribution with $I_d^{(l)}$ thus color distortion in the generated result could be eliminated. The adversarial loss $L_{adv}(G)$ between the generator and discriminator is designed to distinguish $I_f^{(l)}$ from $I_d^{(l)}$ and is formulated as follows:

$$L_{adv}(G)=-\mathbb{E}_{I_f^{(l)}\sim p_{I_f^{(l)}}}\left[D\left(I_f^{(l)}\right)\right] \tag{5}$$

$p_{I_f^{(l)}}$ denotes the generated image dataset and $D\left(I_f^{(l)}\right)$ represents possibility that input $I_f^{(l)}$ comes from the generated dataset.

By optimizing loss of generator, the generator could produce a 1-channel noise-free result in which considerable edges and textures are preserved. Whereas, there is great discrepancy in pixel distribution between the result and luminance channel of VS image. Compared with the original color VS image, the color brightness of the fused color image will be extremely different once inversely transformed back to RGB space. Thus, adopting a discriminator to constrain the generated result to appear similar pixel distribution with luminance channel of VS image could help solve problem of color distortion caused by global pixel diversity. To establish the adversarial game between generator and discriminator, we formulate a loss function of discriminator based on WGAN as follows:

$$L_D=-\mathbb{E}_{x\sim p_{I_d^{(l)}}}[D(x)]+\mathbb{E}_{I_f^{(l)}\sim p_{I_f^{(l)}}}\left[D\left(I_f^{(l)}\right)\right] + \varphi\mathbb{E}_{\tilde{x}}[\|\nabla_{\tilde{x}}D(\tilde{x})\|_2-1]. \tag{6}$$

The first two terms in (6) adopt Wasserstein distance estimation to solve the problem of gradient vanishing and stabilize the training process. The last term denotes the gradient penalty factor which is designed to satisfy the Lipschitz continuity condition, φ is a hyperparameter to control the trade-off between these terms.

Experiment and analysis: In this letter, we construct a large-scale training set by uniformly sampling training sequence of RGB-NIR dataset [9]. The training set is composed of 27 264 pairs of image patches with size 64×64. Two test image pairs are chosen from RGB-NIR dataset for comparing with different methods qualitatively and another 20 test image pairs for quantitative evaluation. We conduct the comparison experiments with seven state-of-the-art fusion methods including VSM [10], CVN [11], WLP [9], CNI [12], GF [13], U2fusion [6] and SDNet [14]. The parameters are set as follows: training epoch is set to 20, and number of batch images is set to 12. For best fusion performance, λ and φ are set to 5 and 1, respectively.

1) Qualitative analysis: To prove superiority of CDP-GAN subjectively, we provide two sets of representative qualitative results in Figs. 5 and 6. Their results could be classified into two categories. The first category is more like VS image thus natural color appears, such as VSM and CNI. However, textures are not rich enough in their results. More specifically, mountains in Figs. 5(c) and 5(f) are unclear compared with other results. In conclusion, VSM and CNI have little ability to retain details in NIR image. On the contrary, the second category contains much detail information about mountains in Figs. 5(d), 5(e) and 5(g)−5(i), but their color appearance is unnatural and color distortion occurs, such as GF, CVN, WLP, U2Fusion and SDNet. CDP-GAN can not only preserve details under non-ideal scene, but also has hardly any color distortion.

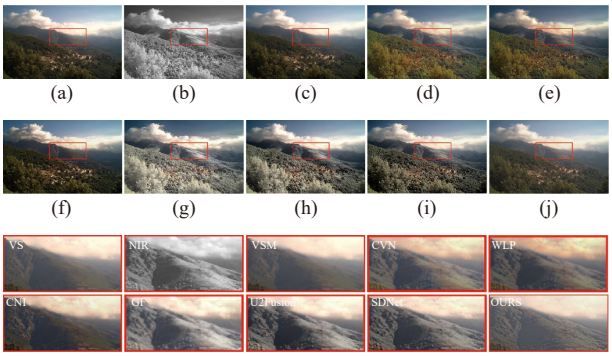

Fig. 5. Qualitative comparison of different fusion algorithms on image Img1. (a) VS image; (b) NIR image; The fused images are obtained by (c) VSM; (d) CVN; (e) WLP; (f) CNI; (g) GF; (h) U2fusion; (i) SDNet; (j) our algorithm.

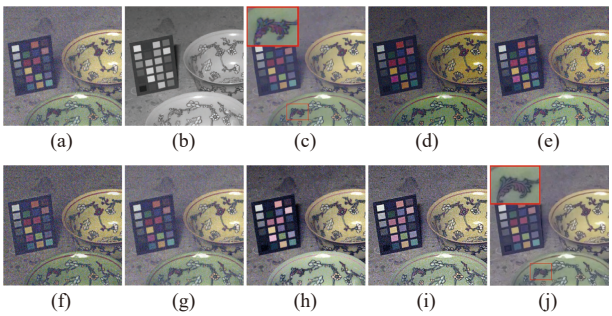

Fig. 6. Qualitative comparison of different fusion algorithms on image Img2. (a) VS image; (b) NIR image; The fused images are obtained by (c) VSM; (d) CVN; (e) WLP; (f) CNI; (g) GF; (h) U2fusion; (i) SDNet; (j) our algorithm.

Limited by device, VS image in Fig. 6 contains much noise. In CDP-GAN, adaptive perceptual loss measures feature rather than pixel-wise similarity. Features obtained by VGG-16 are robust to noise in protecting structure [15] thus generator could produce noise-free result. Moreover, to prevent introducing noise, initial denoised VS image is adopted in discriminator. Initial denoising would remove details in VS color image along with the noise. However, adaptive perceptual loss can transfer details of captured NIR gray image to fused result. Therefore, the final fused image can be both noise-free and detail-preserved. As shown in Fig. 6, only VSM and CDP-GAN could produce noise-free image. However, as highlighted by red rectangles in Fig. 6, the vein is more clear in our result and thus CDP-GAN could not only eliminate noise in VS images but also retain considerable details which are of benefit to human visual system.

2) Quantitative analysis: To demonstrate effectiveness of CDP-

GAN, quantitative experiments are also provided. The objective evaluations of selected metrics EN, SD, MI, $Q^{AB/F}$, SSIM and PSNR [6] are shown in Table 1. Our CDP-GAN could obtain the largest average values on EN, SD, SSIM and PSNR. Slightly worse, CDP-GAN wins the second largest values on MI and $Q^{AB/F}$. For MI, the CNI achieves the largest values because it only acts as a dehazing processing so its result is mostly similar to VS image but abandons the details in NIR image. For $Q^{AB/F}$, the GF algorithm gets the best performance for its motivation to retain the details in source images. However, GF ignores the discrepancy between color and gray image which has been analyzed in the qualitative experiments.

Table 1. Quantitative Comparison of Different Fusion Algorithms

Methods	EN	SD	MI	$Q_{AB/F}$	SSIM	PSNR
VSM	2.0870	2.6034	2.2859	3.0830	1.3270	3.1421
CVN	2.3222	1.8187	1.3745	3.9422	1.3262	2.9885
WLP	2.3041	1.7950	1.3133	3.7387	1.3276	3.0880
CNI	2.2156	2.4835	**2.3105**	4.2134	1.3268	3.3092
GF	2.3456	1.8128	1.7840	**4.2442**	1.3246	3.0696
U2fusion	2.4849	2.1777	1.3600	3.7701	1.3270	2.8385
SDNet	2.5540	1.9928	1.0516	3.6495	1.3196	3.0042
Ours	**2.5923**	**2.6628**	2.2934	4.2348	**1.3277**	**4.0659**

3) Validation experiment: To verify effect of improvements, ablation experiments are shown in Fig. 7. While pixel-wise loss is adopted in generator, result contains less details than our result. As framed in Fig. 7(c), mountains in the distance are not clear enough, which proves that details have not been transmitted to fused image. On the contrary, we can see from Fig. 7(g) that mountains obtained by proposed loss are more complete and larger detail preservation degree is achieved. While no attention mechanism, although result appears similar color information, detail blurred and halo artifacts which are schematically shown in Fig. 7(d) occur. Moreover, it can be concluded that discriminator with multiplication-attention mechanism could achieve greater details and color retention by comparing Figs. 7(e) and 7(g). For feature fusion strategy, result obtained with addition achieves goal of preserving detail and color information. However, concatenation is more suitable for feature fusion as generator with this strategy could recover clearer textures compared with the addition strategy as shown in highlighted region.

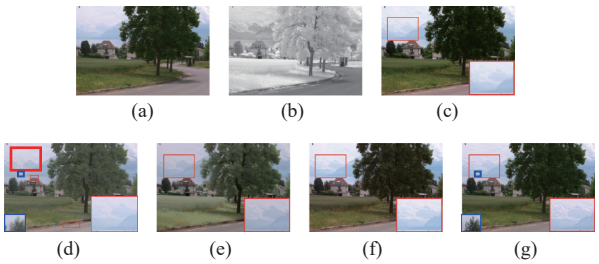

Fig. 7. Ablation experiment. (a) VS image; (b) NIR image; The fused images obtained (c) with pixel-wise loss; (d) without attention mechanism in discriminator; (e) with concatenation-attention mechanism in discriminator; (f) with addition strategy in feature fusion; (g) with multiplication-attention mechanism in discriminator and concatenation strategy in feature fusion.

Conclusion: In this letter, we propose a new framework for NIR and VS image fusion termed as CDP-GAN. It can simultaneously keep color distribution in VS image and detail information in both source images. Specifically, an adaptive perceptual loss is introduced to increase detail preservation degree. Moreover, we unite the global discriminator with proposed attention mechanism, which effectively eliminates color distortion. Experiments verify that our CDP-GAN can not only retain useful information in source images but also act as a denoising method with the aid of NIR image. Compared with other methods on publicly available datasets, our CDP-GAN is superior in both qualitative and quantitative aspects.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (62073304, 41977242, 61973283).

References

[1] W. Wang and X. Yuan, "Recent advances in image dehazing," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 3, pp. 410–436, 2017.

[2] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021.

[3] J. Ma, Y. Wei, P. Liang, L. Chang, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019.

[4] H. Zhang, J. Yuan, X. Tian, and J. Ma, "GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual markovian discriminators," *IEEE Trans. Computational Imaging*, vol. 7, pp. 1134–1147, 2021.

[5] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.

[6] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.

[7] Q. Lian, W. Yan, X. Zhang, and S. Chen, "Single image rain removal using image decomposition and a dense network," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 6, pp. 1428–1437, 2019.

[8] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[9] A. V. Vanmali and V. M. Gadre, "Visible and NIR image fusion using weight-map-guided Laplacian-Gaussian pyramid for improving scene visibility," *Sadhana*, vol. 42, no. 7, pp. 1063–1082, 2017.

[10] Q. Yan, X. Shen, L. Xu, S. Zhuo, X. Zhang, L. Shen, and J. Jia, "Cross-field joint image restoration via scale map," in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 1537−1544.

[11] A. V. Vanmali, S. G. Kelkar, and V. M. Gadre, "A novel approach for image dehazing combining visible-NIR images," in *Proc. National Conf. Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2015, pp. 1–4.

[12] L. Schaul, C. Fredembach, and S. Süsstrunk, "Color image dehazing using the near-infrared," in *Proc. IEEE Int. Conf. Image Processing*, 2009, pp. 1629–1632.

[13] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[14] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Computer Vision*, vol. 129, pp. 2761–2785, 2021.

[15] P. Dai, Z. Li, Y. Zhang, S. Liu, and B. Zeng, "PBR-Net: Imitating physically based rendering using deep neural network," *IEEE Trans. Image Processing*, vol. 29, pp. 5980–5992, 2020.