IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 October 2023, accepted 17 November 2023, date of publication 21 November 2023,
date of current version 14 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335307

RESEARCH ARTICLE

Trainable Self-Guided Filter for Multi-Focus Image Fusion

LEVENT KARACAN, (Associate Member, IEEE)
Department of Computer Engineering, Iskenderun Technical University, Iskenderun, 31200 Hatay, Turkiye
e-mail: levent.karacan@iste.edu.tr

ABSTRACT Cameras are limited in their ability to capture all-in-focus images due to their limited depth of field. This results in blurriness for objects too far in front of or behind the focused point. To overcome this limitation, multi-focus image fusion (MFIF) approaches have been proposed. Although recent MFIF methods have shown promising results for this task, they still need to be improved in terms of artifacts and color degradation. Motivated by these observations, in this paper, we propose a new Generative Adversarial Network (GAN)–based MFIF model to improve fusion quality by predicting more accurate focus maps thanks to a trainable guided filter we incorporated. The proposed model comprises an encoder-decoder network, and a trainable self-guided filtering (TSGF) module that is specifically designed to enhance spatial consistency in the predicted focus map and to eliminate the requirements of post-processing in existing GAN-based methods. The encoder-decoder network first predicts raw focus maps, which are then passed to the TSGF to produce the final focus maps. To train the proposed model effectively, we define three objectives: L1 loss, GAN loss, and Focal Frequency Loss (FFL) in the frequency domain. L1 loss is defined on ground-truth and predicted focus maps, whereas GAN loss and FFL are defined on ground-truth all-in-focus images and fused images. Experimental results show that the proposed approach outperforms the existing GAN-based methods and achieves highly competitive performance with state-of-the-art methods in terms of standard quantitative image fusion metrics and visual quality on three MFIF benchmark datasets.

INDEX TERMS Multi-focus image fusion, guided filter, generative adversarial networks.

I. INTRODUCTION

Cameras have a limited depth of field (DOF), which means that they cannot take all-in-focus images of scenes containing objects at different depths. More precisely, regions outside the DOF of the cameras appear blurred. To overcome this limitation, modern cameras use a multi-focus image fusion (MFIF) method to fuse multiple images, each focused on a different region, into a single all-in-focus image. MFIF is a valuable task for other computer vision problems such as segmentation and tracking. It has also been utilized in a wide range of fields like microscopy, medical imaging, and remote sensing [52]. However, it is not always straightforward to fuse multi-focus images because it can introduce artifacts such as the halo effect or blurring due to defocus spread effect [30] at object boundaries, which can degrade the quality of the resulting image. Therefore, it is important to use appropriate fusion algorithms and post-processing techniques to minimize these artifacts and obtain high-quality fused images. In this regard, it is expected that a multi-focus image fusion method should fuse multiple images without losing information and not produce artifacts that would distort spatial consistency. Many different approaches have been proposed in the literature for MFIF, which can be broadly classified into spatial domain and transform domain methods.

Spatial domain methods are a type of image processing technique that operates on the pixels of an image directly, either on pixels, in blocks, or regions. Pixel-wise methods, such as the guided filtering-based method (GFF) [21] and dense SIFT-based method (DSIFT) [26], try to identify focused regions in the image and use this information to generate a decision map for the fusion process. However, using these methods, it can be difficult to accurately identify the boundary between focused and defocused regions. Block-wise methods [3] also have their limitations, as they divide the image into blocks and rely on focus measures to

The associate editor coordinating the review of this manuscript and approving it for publication was Miaohui Wang.

identify focused areas, which can lead to inaccuracies in the decision map. Region-wise methods [19], [54], on the other hand, segment the image into different regions and use this information to generate the fused result. However, these methods can also be prone to errors, as the segmentation process can be difficult to perform accurately.

Instead of performing the fusion directly in the image domain, transform domain-based image fusion methods involve converting the input images into another domain, such as the non-subsampled contourlet transform (NSCT) [51], sparse representation (SR) [47], or wavelet transform [16], [38], to perform the fusion. After the fusion process, they inverted the fused transformation back to the image domain. One disadvantage of transform domain-based methods is that they can potentially cause information loss during the transformation process. Additionally, both conventional spatial-domain and transform-domain approaches rely on handcrafted features, rather than learning features from the data. This can limit their performance and flexibility.

Over the last decade, convolutional neural network (CNN)-based methods have also dominated the multi-focus image fusion literature as in other computer vision problems. Liu et al. [24] were the first to propose a method for multi-focus image fusion using CNNs. Their approach involved using a Siamese architecture to predict decision maps like in spatial domain methods for a given set of multi-focus images. Since then, numerous CNN-based methods for MFIF have been proposed, which either predict fused images directly or via decision maps. While direct MFIF methods [17], [42], [50], [55], [56] may not require additional steps, they can suffer from issues such as color degradation and contrast changes. On the other hand, decision map-based methods [8], [18], [24], [28], [29], [30], [40] have generally proven to be more successful, but they need to predict spatially consistent and precise focus maps to avoid boundary artifacts. Additionally, when the background objects are in focus, the boundaries of the foreground objects become indistinct and spread out over a wider area with uncertain widths. This phenomenon is called the Defocus Spread Effect (DSE) and can negatively impact the results of focus map prediction accordingly image fusion if it is not properly addressed. For this reason, some of the recent works [30], [40], [44] particularly tackle the DSE problem to alleviate boundary artifacts.

In recent years, several methods [8], [40], [50] based on Generative Adversarial Networks (GANs) have been proposed to predict focus maps or fusion weights using adversarial learning. FuseGAN [8] was the first method to use GANs to synthesize focus maps from given multi-focus images drawing inspiration from image-to-image translation [14]. MFIF-GAN, proposed by Wang et al. [40], improves the quality of focus maps through architectural modifications, new GAN loss, and new dataset [30] simulating DSE. Despite the significant success of both FuseGAN and MFIF-GAN, they require advanced post-processing steps to refine the focus maps predicted by the network. These steps may include removing small objects, applying morphological operations, and performing spatial filtering. MFF-GAN [50] introduces an unsupervised learning approach for directly fusing input multi-focus images through predicted fusion weights. However, this process can degrade the colors of the input images during fusion.

Focus map generation-based methods have the advantage of preserving more of the original information from the input multi-focus images compared to directly fusing the images. However, these methods are more prone to producing artifacts due to the difficulty in accurately determining the focus/defocus boundary because of DSE. This is because these methods often tend to edges of objects as focus/defocus boundaries but these edges may not always accurately represent the true focus/defocus boundary. This can be explained by that CNNs often rely on edges and objects to make decisions, but they may struggle to determine focus/defocus boundaries by capturing the intricate details related to the transition zone near these boundaries [30]. To address this issue, various post-processing techniques have been developed to improve the accuracy and spatial consistency of the generated focus maps. These techniques can be effective at reducing specific boundary errors and spatial inconsistencies in the focus maps, but they are manually designed with several parameter settings and are not adaptive. In some cases, it may be necessary to use a combination of different post-processing techniques to fully address the issues present in the focus map.

In this paper, we adopt a focus-map-based approach for multi-focus image fusion (MFIF), instead of a direct fusion method like MFF-GAN [50] due to their tendency to degrade color quality. Our primary research objective is to enhance spatial consistency in predicting focus maps for multi-focus image fusion, aiming to reduce blurry artifacts. We address a challenge faced by recent GAN-based methodologies such as FuseGAN [8] and MFIF-GAN [40], where achieving spatially consistent focus maps proves difficult, necessitating additional post-processing steps. To overcome this challenge, our approach introduces a novel method that seamlessly integrates adaptive post-processing into the training process, streamlining focus map generation and ultimately improving the effectiveness of multi-focus image fusion. Specifically, we propose the incorporation of a trainable self-guided filter block [32] at the end of the encoder-decoder network, which leverages the output raw focus map as guidance to refine itself adaptively. Furthermore, unlike FuseGAN [8] and MFIF-GAN [40], which use GAN-loss on real and predicted focus maps, we advise utilizing GAN-loss on real full-focus images and fused images. This decision is prompted by the fact that Isola et al. [14] showed that utilizing a GAN loss function is more efficient for image generation than label prediction. Finally, we propose the use of the additional frequency-domain Focal Frequency Loss (FFL) [15] for multi-focus image fusion (MFIF). The FFL was originally proposed as a way to improve the level of detail in image generation methods, and we show it can be effectively applied to MFIF as well. We conduct thorough experiments with ablation studies to demonstrate the impact of each contribution.

Our contributions can be summarized as follows:

- We propose a novel GAN-based MFIF model that introduces an adaptive post-processing block as a trainable guided filter to improve spatial consistency.
- To leverage the power of GANs more effectively, we redefine the GAN loss on the fused images rather than on focus maps unlike in the previous works.
- We propose to use Focal Frequency Loss (FFL) for MFIF to evaluate fused images also in the frequency domain.
- Experiments on three benchmark datasets demonstrate that the proposed method outperforms existing GAN-based MFIF methods both quantitatively and qualitatively, achieving state-of-the-art results in MFIF.
- Proposed approach speeds up the fusion process compared to existing GAN-based MFIF methods thanks to the proposed end-to-end fusion model.

In the following sections, we first review related GAN-based Image Fusion methods and the use of guided filters for Image Fusion in Section II. Then, we give the details of the proposed method in Section III. In Section IV, we describe the experimental setup, provide qualitative and quantitative results, and present an ablation study. Finally, we conclude the paper in Section V.

II. RELATED WORKS

A. GAN-BASED IMAGE FUSION

Generative Adversarial Networks (GANs) have been widely used to tackle various computer vision problems. Some GAN-based methods have been developed specifically for image fusion. In GAN-based image fusion methods, the generator is trained to produce a fused image from multiple input images, while the discriminator is trained to distinguish between the fused image and the reference images. More precisely, the generator tries to maximize the information fusion between the input images, while the discriminators try to identify the source of the fused image. This creates an adversarial process that helps the generator to produce a fused image that contains relevant information from both input images while being difficult for the discriminators to classify. DDcGAN [31] proposed a GAN-based approach for fusing visible and infrared images. The generator is trained to produce a fused image from the visible and infrared input images, while the discriminators are trained to distinguish the fused image from each of the input images. Similarly, MEF-GAN [43] and GANFuse [49] define GAN objectives for multi-exposure image fusion. Different GAN-based approaches [12], [48] have been proposed for multi-modal medical image fusion, which aims to combine multiple images from different modalities (e.g., MRI, CT, PET) of the same patient. These approaches often introduce specific objectives or constraints to the GAN training process, to ensure that the fused image preserves the relevant information and properties of the input images.

FuseGAN [8] was the first model that was introduced for the task of multi-focus image fusion. The model was inspired by the pix2pix [14] model, which is a well-known conditional GAN (cGAN) approach for image-to-image translation. The generator consists of an image encoder that is shared among the input multi-focus images and a decoder that generates a focus map using the concatenated features of the encoded input images. The generator is trained using the Least Squares GAN (LSGAN) [33] objective, in conjunction with a discriminator network which takes the ground truth or generated focus maps together with the input images and determines if it is real or fake (generated). FuseGAN also uses additional post-processing steps [37] to refine the generated focus map to further improve fusion results. MFIF-GAN [40] improves the FuseGAN with some modifications on architecture, objective functions, and post-processing. Briefly, it includes Squeeze-Excitation(SE) [11] layer into the Residual Network (ResNet) blocks and trains model with WGAN-GP [7] objective instead of LSGAN. It also proposes to use new post-processing steps based on small object removal to speed up focus map refinement. Both FuseGAN and MFIF-GAN need manually designed post-processing steps where various parameters are also set manually. Specifically, FuseGAN uses the convolutional conditional random fields (ConvCRF)-based post-processing method [13], which is effective with ground truth all-in-focus images. However, obtaining such ground truth is often impractical in real-world applications, limiting FuseGAN's performance adaptability. Moreover, ConvCRF introduces computational overhead, hindering real-time or time-sensitive applications. MFIF-GAN [40] employs an object removal method to enhance the spatial consistency of the predicted focus map. However, a drawback lies in the need for manual determination of object size parameters, which adds a level of complexity and subjectivity to the process. This manual intervention may not be ideal, especially when aiming for an automated and adaptable solution.

Recently, MFF-GAN [50] proposes an alternative unsupervised approach for direct image fusion through predicted fusion weights as screen maps. However, it changes the original colors of input images in the fused image.

B. GUIDED FILTER FOR IMAGE FUSION

Guided filter (GF) was first proposed by He et al. [9] as a fast edge-preserving filtering method with a local linear model as opposed to its slow non-linear counterparts such as the Bilateral filter [5]. It has been used for many image processing tasks such as joint upsampling and image matting [9] to provide spatial consistency or detail preservation. GFF [21] is the first image fusion method that utilizes a guided filter to refine weight maps that are constructed using a Laplacian filter on base and detail layers of input images. Guided filtering is a widely utilized technique in MFIF [23],

[36], [53] for refining the focus map or weight map using the input images as guidance. However, this guidance can introduce bias at the boundaries of the focus map toward object edges. Additionally, the presence of texture and edge information within the focus map region may lead to spatial inconsistencies in the focus map prediction.

Recently, a trainable version of a guided filter [41] was proposed to jointly train with a dense prediction network so that an image-based adaptive guided filter can be learned. A multi-exposure fusion method MEFNet [32] utilized such a guided filter layer to jointly upsample low-resolution weight map predictions to a higher resolution, using the input images as guidance, resulting in a faster fusion process.

While existing GAN-based methods in MFIF have shown promising results for focus map prediction, they often struggle with issues related to spatial consistency and boundary artifacts. Although they suggest the use of hand-crafted additional refinement methods, there is still a need for an end-to-end approach to deal with these issues adaptively. To fill this gap, in this study, we propose a new trainable post-processing layer in the form of a trainable guided filter (TGF) [41] for focus map prediction networks. Specifically, we use the output focus map of the generator as a self-guider. The proposed trainable self-guided filter (TSGF) offers two main benefits. First, it improves the spatial consistency of the focus map and the accuracy of the focus/defocus boundary, thereby increasing the quality of the MFIF. Second, it eliminates the need for sophisticated post-processing steps such as those used in FuseGAN [8] and MFIF-GAN [40]. Note that the proposed TSGF can also be used in other focus map prediction-based methods.

III. METHOD

Our objective is to learn a mapping function from input multi-focus images x_1 and x_2 to focus maps $\hat{F}$, $G : \{x_1, x_2\} \rightarrow \hat{F}$ so that we can obtain an all-in-focus image using the following equation:

$$\hat{y} = x_1\hat{F} + x_2(1 - \hat{F}), \tag{1}$$

where it is assumed that x_1 has a focused foreground and a blurred background while x_2 has a blurred foreground and a focused background. Moreover, x_2 may exhibit Defocus Spread Effect (DSE) as mentioned before. The mapping function G is the generator network that is trained by employing a discriminator network D for the GAN objective as shown in Figure 1. We also assume that training data has ground-truth all-in-focus images y and corresponding focus map F for each input pair (x_1, x_2). It should be noted that the proposed method is designed to fuse two input images, but it can also sequentially fuse multiple input images, producing a series of two images each time.

A. GENERATOR NETWORK WITH TSGF

As illustrated in Figure 1, the generator G is an encoder-decoder network with several ResNet blocks in between like in FuseGAN [8] and MFIF-GAN [40]. We also include the TSGF which is adapted from the study of Wu et al [41] for MFIF that is placed after the decoder to enhance spatial consistency in predicted focus maps and eliminate the need for post-processing. The shared encoder takes two input multi-focus images, x_1 and x_2, and processes them through three Convolution-BatchNorm-ReLU and nine ResNet blocks to extract the bottleneck features. The decoder uses the concatenation of these features to generate the intermediate output focus map $\bar{F}$ through three Deconvolution-BatchNorm-ReLU blocks. In the end, TSGF is applied to further refine the intermediate output $\bar{F}$ and obtain the final focus map $\hat{F}$. The proposed TSGF is designed to improve the accuracy of the focus map by using internal guidance from the network output $\bar{F}$ itself. It is trainable so it

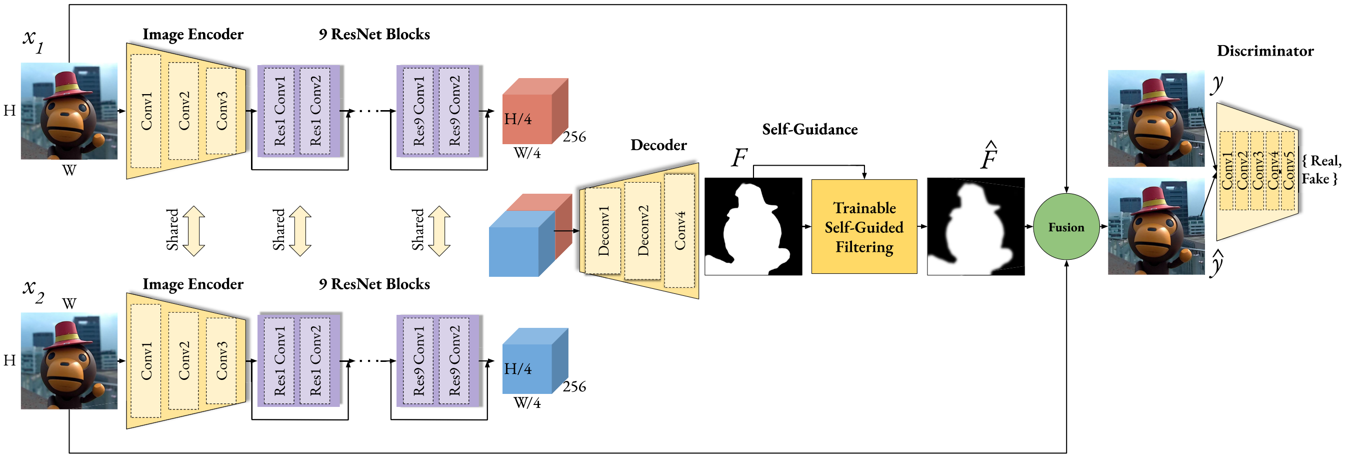

FIGURE 1. Overall architecture of the proposed MFIF model. The image encoder with ResNet blocks encodes input multi-focus images to the features, and the decoder constructs a focus map from those features. The generator network consists of 2D convolution (Conv) and transposed convolution (Deconv) layers. Each ResNet block has two 2D convolution layers, identified by "Res" and a block number. The detailed parameters of each layer are given in Table 1. In addition, we propose to employ a trainable self-guided filtering (TSGF) layer after the output of the decoder. This extension aims to give the model the ability to predict spatially consistent focus maps (Please see Fig. 2 for more details). In the training phase, a discriminator network is utilized for GAN loss on output fused images and target ground-truth full-focus images.

TABLE 1. The generator and discriminator networks' detailed parameter configuration. The generator network comprises 2D convolution (Conv) and transposed convolution (Deconv) layers. Residual network blocks are also denoted by the abbreviation "Res" followed by a block number, and each layer is labeled with a number identifier. For example, "Res1 Conv1" denotes the first convolution layer in the first residual network block. The discriminator network consists of five 2D convolution layers with spectral normalization, and leaky ReLU activation with slope 0.1 in some layers as shown. Please refer to Figure 1.

for the corresponding architectural illustration.

Image Encoder							
Layer	Input Ch.	Output Ch.	Kernel Size	Stride	Padding	Normalization	Activation
Conv1	3	64	7	1	3	BatchNorm(eps $= 1e-5$, momentum $= 0.1$)	ReLU
Conv2	64	128	3	2	1	BatchNorm(eps $= 1e-5$, momentum $= 0.1$)	ReLU
Conv3	128	256	3	2	1	BatchNorm(eps $= 1e-5$, momentum $= 0.1$)	ReLU
Res1-9 Conv1	256	256	3	1	1	BatchNorm(eps $= 1e-5$, momentum $= 0.1$)	ReLU
Res1-9 Conv2	256	256	3	1	1	BatchNorm(eps $= 1e-5$, momentum $= 0.1$)	-
Decoder							
Deconv1	512	256	3	2	1	BatchNorm(eps $= 1e-5$, momentum $= 0.1$)	ReLU
Deconv2	256	128	3	2	1	BatchNorm(eps $= 1e-5$, momentum $= 0.1$)	ReLU
Conv4	128	1	7	1	3	-	Sigmoid
Discriminator							
Conv1	3	64	4	2	2	-	LeakyReLU(slope=0.1)
Conv2	64	128	4	2	2	SpectralNorm	LeakyReLU(slope=0.1)
Conv3	128	256	4	2	2	SpectralNorm	LeakyReLU(slope=0.1)
Conv4	256	512	4	2	2	SpectralNorm	LeakyReLU(slope=0.1)
Conv5	512	1	4	1	2	-	-

can be adjusted during the training process to optimize the performance of the generator. It is also worth noting that all the components (encoder, ResNet blocks, decoder, and TSGF) of the generator network work in a cohesive, end-to-end manner to produce the final output focus map $\hat{F}$. Detailed parameters of each layer of the generator network G including input-output feature dimensions, convolution parameters, normalization, and activation functions are given in Table 1.

B. TRAINABLE SELF-GUIDED FILTER (TSGF) FOR FOCUS MAP REFINEMENT

The guided filter [9] apply a local linear model using a guidance I_i to obtain output value q_i of the ith pixel as:

$$q_i = a_k I_i + b_k, \quad \forall_i \in w_k, \tag{2}$$

where w_k is a kth local square window with radius r and i is a pixel location in window w_k. The linear coefficients a_k and b_k are calculated for each window w_k by minimizing the following reconstruction error:

$$\min_{a_k, b_k} \sum_{i \in w_k} (p_i - (a_k I_i + b_k))^2 + \lambda a_k^2 \tag{3}$$

where p_i is input image and λ is a regularization parameter to control smoothness. Once a_k and b_k is computed for each window using linear regression solver [9], filtering output is obtained as follow:

$$q_i = \bar{a}_i I_i + \bar{b}_i, \tag{4}$$

where $\bar{a}_i$ and $\bar{b}_i$ is the mean coeefficents in w_i surrounding each pixel i.

As the original guided filter is not trainable, Wu et al. [41] defined a differentiable version as a trainable layer so that it can be trained with any prediction network in an end-to-end manner. Instead of using the guided filter as a post-processing step, as has been done in previous works [23], [36], [53], we incorporate a trainable version of the guided filter as a layer within the network that predicts the focus map. This allows the network to learn to produce high-quality focus maps by directly optimizing the output during training. In our proposed MFIF model, the intermediate focus map $\bar{F}$ serves as both guidance and input to the guided filter layer to learn filtering coefficients a_k and b_k. The guidance $\bar{F}$ is also transformed to the task-specific guidance map g_F using a fully convolutional network block [41] as shown in Figure 2. Thus, we call it a trainable self-guided filter (TSGF) layer.

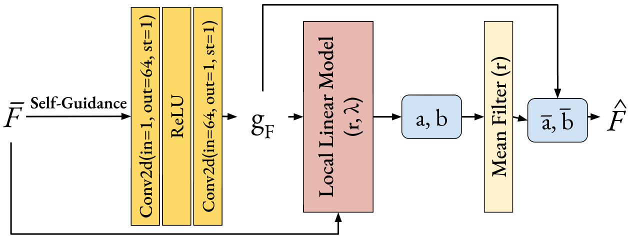
$\bar{F}$
Self-Guidance
Conv2d(in=1, out=64, st=1)
ReLU
Conv2d(in=64, out=1, st=1)
g_F
Local Linear Model (r, λ)
a, b
Mean Filter (r)
$\bar{a}, \bar{b}$
$\hat{F}$

FIGURE 2. Trainable self-guided filtering. The output of the decoder network is employed as self-guidance to improve itself in terms of spatial consistency. It is first transformed into the task-specific guidance map g_F using trainable two convolutional layers. Then the local linear model is applied to obtain coefficients a and b to further compute the output focus map.

In that case, final focus map $\hat{F}$ can obtained as follows:

$$\hat{F}_i = \bar{a}_i g_F + \bar{b}_i. \tag{5}$$

As illustrated in Figure 2, our proposed MFIF model introduces a new self-guidance approach, which uses the intermediate output of the model itself as guidance, rather than relying on the input images as guidance as in previous works [32], [41] using trainable guided filter. Our approach is particularly beneficial for focus map prediction in MFIF. Because, as mentioned in Section II-B, using input multi-focus images as guidance can introduce bias at the boundaries of the focus map toward object edges. Additionally, the presence of texture and edge information within the focus map region may lead to spatial inconsistency in the focus map prediction. In contrast, our self-guidance approach provides all network components to be trained in coordination to adaptively learn to refine the final focus map.

In summary, the Trainable Self-Guided Filtering (TSGF) module plays a pivotal role in significantly improving focus map prediction accuracy. During training, TSGF collaborates with the entire network, ensuring the predicted focus map maintains spatial consistency. Notably, TSGF effectively addresses discrepancies within the predicted focus map in textured regions or around small objects by filling holes and reducing noise. This adaptability enhances focus map refinement, resulting in a more precise alignment with genuine focus regions in the input images. Moreover, TSGF demonstrates adaptability by intelligently extending focus and defocus boundaries, considering the natural tendency of focus regions to extend beyond object edges due to the defocus spread effect (DSE). Consequently, it provides a more accurate representation of focus regions, ultimately enhancing the overall precision and accuracy of the final focus map.

C. LOSS FUNCTION

Our generator model G is designed to predict a focus map $\hat{F}$ for a given multi-focus input image pair (x_1, x_2) that is as close as possible to the ground truth focus map F. To train our model, we use a combination of three losses: reconstruction loss, adversarial loss, and focal frequency loss (FFL) [15].

We define the reconstruction loss as $L1$ distance between ground truth(F) and predicted ($\hat{F} = G(x_1, x_2)$) focus maps:

$$\mathcal{L}_{rec}(G) = \mathbb{E}_{(x_1,x_2,F)}\|F - G(x_1, x_2)\|_1 \tag{6}$$

For adversarial loss, we employ a multi-scale 70 × 70 PatchGAN discriminator D network [14], [39]. We employ three discriminators, each operating at a different scale of the image. To create an image pyramid of three scales, the real and synthesized high-resolution images are downsampled by a factor of 2 and 4. The discriminators are trained to distinguish between real all-in-focus and fused images at these respective scales. All discriminators share the same network structure which is given in Table 1. We use leaky ReLUs with a slope of 0.1. Additionally, to stabilize and enhance the training process, we incorporate spectral normalization [34] in the discriminator network (D). This technique helps control the Lipschitz constant of the discriminator, promoting smoother convergence during adversarial training [34].

Our approach to adversarial loss differs from the FuseGAN [8] and MFIF-GAN [40] methods as illustrated in Figure 3. Specifically, both FuseGAN and MFIF-GAN adopt a conditional GAN approach using input multi-focus images x_1 and x_2 as conditions to the real (F) or fake ($\hat{F}$) focus maps. Instead, we use the original all-in-focus image y and fused image $\hat{y}$ (See Eq. 1) as real and fake inputs to the Discriminator. Thus, we incorporate all fusion processes into the adversarial learning rather than just focus map prediction. Moreover, it is shown that using GAN for photographic image generation is more effective than using it for label prediction [14].

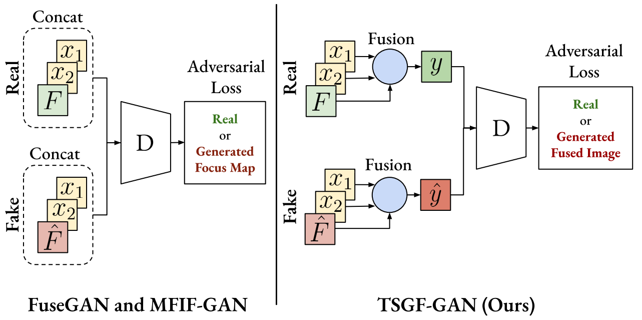

FIGURE 3. Comparison of adversarial loss uitilization in FuseGAN [8], MFIF-GAN [40], and the proposed TSGF-GAN. Our approach illustrates a distinct strategy where the fused results are provided to the discriminator during training, differing from FuseGAN [8] and MFIF-GAN [40], where the discriminator receives the concatenation of focus maps and source images.

For our GAN objective, we utilize the Least Squares Generative Adversarial Networks (LSGAN) [33] as follows:

$$\mathcal{L}_{GAN}(D) = \mathbb{E}_{(x_1,x_2,F)}(D(y) - 1)^2 + \mathbb{E}_{(x_1,x_2)}D(\hat{y})^2 \tag{7}$$

$$\mathcal{L}_{GAN}(G) = \mathbb{E}_{(x_1,x_2)}(D(\hat{y}) - 1)^2 \tag{8}$$

$$\mathcal{L}_{GAN} = \mathcal{L}_{GAN}(D) + \mathcal{L}_{GAN}(G) \tag{9}$$

Recently, Jiang et al. [15] have proposed a frequency domain loss called focal frequency loss (FFL) to address the issue of missing frequencies in image synthesis methods. We adapt FFL to the MFIF to recover missing frequencies in the fused image as follows:

$$FFL = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} w(u, v)|\mathcal{F}_y(u, v) - \mathcal{F}_{\hat{y}}(u, v)|^2 \tag{10}$$

where $\mathcal{F}_y(u, v)$ and $\mathcal{F}_{\hat{y}}(u, v)$ denote $M \times N$ dimensional 2D Discrete Fourier Transforms of ground truth all-in-focus image y and fused image $\hat{y}$ respectively. The FFL is calculated as the average of the squared differences between the 2D discrete Fourier transforms weighted by the spectrum weight matrix $w(u, v)$. The spectrum weight matrix down-weights easy frequencies, and its values are determined dynamically based on the current loss of each frequency during training using a non-uniform distribution:

$$w(u, v) = |\mathcal{F}_y(u, v) - \mathcal{F}_{\hat{y}}(u, v)|^{\alpha} \tag{11}$$

where α is defined as the scaling factor. We found $\alpha = 2$ in the experiments for MFIF.

Our final objective is the combination of these three losses:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{GAN} + FFL \tag{12}$$

In our training process, each loss function contributes distinctly to enhancing our multi-focus image fusion model. L1 loss ensures accurate focus map reconstruction. Meanwhile, the GAN loss significantly improves focus map prediction quality through adversarial training, encouraging the more realistic fusion results through focus maps. FFL loss in the frequency domain assists the GAN loss, particularly in managing high-frequency components. This is particularly beneficial in regions susceptible to blurring, such as those near the focus/defocus boundary. The combined effect of these functions results in a superior multi-focus image fusion model.

IV. EXPERIMENTS

In this section, we first give information about the datasets and implementation details used during the training and testing phases. We then provide both qualitative and quantitative results, comparing them to the performance of baseline MFIF models. Finally, we present the results of the ablation study, which aims to evaluate the individual contributions and impacts of each element of the proposed method. In the following sections, we call our method TSGF-GAN.

A. DATASETS

To generate a training dataset, we utilize the alpha-matte model [30] as it considers the DSE. We use the PASCAL VOC 2012 dataset [6] which includes pairs of images (I_i, S_i), where I_i and S_i represent the RGB images and corresponding binary segmentation maps in which pixels with a value of 1 denote the foreground and pixels with a value of 0 represent the background. In the alpha-matte model [30], it is assumed that the binary segmentation maps S_i are ground-truth focus maps F_i. According to this, the multi-focus input images x_{1i} and x_{2i} are obtained by:

$$x_{1i} = F_i I_i + (1 - F_i)\Big(G(\sigma) * \big((1 - F_i)I_i\big)\Big) \tag{13}$$

$$x_{2i} = G(\sigma) * (F_i I_i) + \Big(1 - \big(G(\sigma) * F_i\big)\Big)\big((1 - F_i)I_i\big) \tag{14}$$

where $G(\sigma)$ denotes a Gaussian filter applied to mimic DefocusSpread Effect(DSE), and $*$ denotes the convolution operation. The parameter σ controls the level of blurriness or the intensity of the DSE. It is set to have a value of $\sigma = 1$.

To evaluate the performance of the trained model, we employ three datasets: the widely used Lytro dataset [35], MFI-WHU [50], and the recently introduced MFFW dataset [45] which is particularly affected by DSE. The Lytro dataset, which is widely recognized as a benchmark in the MFIF domain, consists of 20 pairs of multi-focus images captured using a light field camera. The recently constructed MFI-WHU dataset comprises 120 image pairs, with 30 pairs used for testing and 90 pairs used for training. It involves Gaussian blur and a decision map based on the public COCO dataset [22]. We evaluate 30 images allocated for testing in this study. To overcome the DSE limitation of the Lytro dataset, Xu et al. [45] presented the MFFW dataset, which includes 13 real multi-focus image pairs with intense DSE. At present, MFFW is the only multi-focus image dataset available with such distinct Depth of Field (DoF) variation characteristics. We conducted a comprehensive evaluation of the image fusion performance by utilizing the varied attributes of these datasets, which provided valuable insights into the effectiveness and generalization capability of our proposed approach.

B. IMPLEMENTATION DETAILS

We train the Generator G and Discriminator D using the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and without weight decay. We use a constant learning rate of 2×10^{-5} and a batch size of 2 for 8 epochs. Each batch consists of two pairs of multi-focus input images (x_1, x_2) along with their corresponding ground truth focus maps F and original all-in-focus image y. Generator G is responsible for generating the focus map $\hat{F}$, while Discriminator D evaluates the realism of the fused images $\hat{y}$ compared to the real full focus images y.

During the test time, we no longer use the Discriminator D. The output focus map $\hat{F}$ of the Generator model is utilized to fuse input images x_1 and x_2 using Equation 1 after thresholding with the value of 0.5 is applied.

C. EVALUATION METRICS

Since there is no ground truth for testing image fusion, it is challenging to evaluate the quality of fused images. Various image fusion metrics have been proposed to measure fusion quality from different perspectives. Those metrics are broadly divided into four main categories by Liu et al. [27]: information theory-based metrics, image feature-based metrics, structural similarity-based metrics, and human perception-based metrics. In this study, we select five representative metrics from these categories by following the FuseGAN [8] to evaluate the proposed model. These include the information theory-based *Normalized Mutual Information* Q_{NMI} [10], image feature-based *Gradient-based Fusion Metric* Q_G [46] and *Spatial Frequency* Q_{SF} [57], structural similarity-based *Yang's metric* Q_Y [20], and human-perception-based *Chen-blum metric* Q_{CB} [2].

We can briefly explain the representative metrics as follows:

- Normalized mutual information Q_{NMI}: Q_{NMI} is a measure based on information theory that aims to improve the stability of traditional mutual information (MI).

TABLE 2. **Quantitative comparisons with state-of-the-art MFIF methods. The average scores on Lytro and MFFW datasets according to the five MFIF metrics are presented. The best, second-best, and third-best results are shown in blue, red, and green colors respectively. ↑ denotes higher is better.**

Datasets	Lytro Dataset					MFFW Dataset					MFI-WHU Dataset				
Methods/Metrics	$\mathrm{Q_{NMI}}$ ↑	$\mathrm{Q_{G}}$ ↑	$\mathrm{Q_{SF}}$ ↑	$\mathrm{Q_{Y}}$ ↑	$\mathrm{Q_{CB}}$ ↑	$\mathrm{Q_{NMI}}$ ↑	$\mathrm{Q_{G}}$ ↑	$\mathrm{Q_{SF}}$ ↑	$\mathrm{Q_{Y}}$ ↑	$\mathrm{Q_{CB}}$ ↑	$\mathrm{Q_{NMI}}$ ↑	$\mathrm{Q_{G}}$ ↑	$\mathrm{Q_{SF}}$ ↑	$\mathrm{Q_{Y}}$ ↑	$\mathrm{Q_{CB}}$ ↑
GFF [21]	1.0501	0.6948	-0.0308	0.9723	0.7881	0.6989	0.3462	-0.0596	0.7931	0.6815	1.1030	0.6719	-0.0308	0.9716	0.8084
MWGF [58]	1.0666	0.7076	-0.0387	0.9758	0.7856	0.7084	0.4107	-0.0652	0.8050	0.6894	1.1299	0.6790	-0.0422	0.9828	0.8112
QuadTree [1]	1.0983	0.6993	-0.0255	0.9763	0.7984	0.6969	0.3061	-0.0528	0.7994	0.6903	1.1714	0.6843	-0.0248	0.9838	0.8193
DSIFT [26]	1.0986	0.7012	-0.0244	0.9762	0.7989	0.6826	0.3330	-0.0475	0.7830	0.6805	1.1706	0.6856	-0.0256	0.9835	0.8192
CSR [25]	0.9185	0.6656	-0.0337	0.9190	0.7028	0.6958	0.2894	-0.0622	0.7211	0.6114	1.1045	0.6689	-0.0292	0.9734	0.8015
CNN [4]	1.0846	0.7076	-0.0342	0.9758	0.7961	0.6998	0.3450	-0.0667	0.8007	0.6798	1.1561	0.6790	-0.0317	0.9834	0.8184
FuseGAN [8]	1.1108	0.7063	-0.0225	0.9770	0.7978	0.7052	0.3288	-0.0487	0.7750	0.6860	1.1606	0.6797	-0.0247	0.9829	0.8186
GFDF [36]	1.0925	0.6985	-0.0289	0.9766	0.7990	0.6946	0.3557	-0.0552	0.7971	0.6901	1.1604	0.6795	-0.0278	0.9840	0.8200
MMFNet [30]	0.9083	0.6415	-0.0085	0.9444	0.7500	–	–	–	–	–	–	–	–	–	–
SESF [29]	1.0991	0.7117	-0.0242	0.9766	0.7956	0.6822	0.3346	-0.0168	0.7931	0.6824	1.1606	0.6776	-0.0274	0.9811	0.8132
IFCNN [55]	0.8824	0.6966	-0.0271	0.9445	0.7229	0.6973	0.3079	-0.0634	0.7735	0.6270	0.8849	0.5937	-0.0283	0.9573	0.7343
DRPL [18]	1.1061	0.7184	-0.0226	0.9753	0.7961	0.6801	0.3216	-0.0292	0.7741	0.6724	1.0311	0.6618	-0.0231	0.9757	0.8049
U2Fusion	0.7389	0.5326	-0.2946	0.8327	0.6482	0.6587	0.3232	-0.3538	0.6510	0.5730	0.6608	0.5021	-0.0667	0.7858	0.5164
MFF-GAN [50]	0.7431	0.5922	0.0518	0.8517	0.6294	0.6560	0.3078	-0.0853	0.6886	0.5623	0.7409	0.5352	0.0877	0.8883	0.6331
MFIF-GAN [40]	1.1111	0.7179	-0.0232	0.9770	0.7976	0.7049	0.3264	-0.0545	0.7964	0.6875	1.1673	0.6799	-0.0228	0.9836	0.8191
GACN [28]	1.1041	0.7106	-0.0288	0.9756	0.7954	0.6749	0.3292	-0.0500	0.7741	0.6755	1.1677	0.6798	-0.0309	0.9837	0.8171
TSGF-GAN(Ours)	1.1115	0.7194	-0.0237	0.9772	0.7990	0.7081	0.3301	-0.0439	0.8056	0.6903	1.1706	0.6799	-0.0244	0.9839	0.8201

- Gradient-based fusion metric Q_G: Using image characteristics, Q_G evaluates the transfer of edge information from source pictures to fused images.
- Image fusion metric-based on spatial frequency Q_{SF}: The metric is a relative metric that can be applied to many picture types and evaluates spatial frequency.
- Yang's metric Q_Y: This metric is a metric based on picture structural similarity that assesses the preservation of structural information from source images in the fused image.
- Chen-blum metric Q_{CB}: Q_{CB} is a human perception-inspired metric that evaluates the fused image by comparing contrast features to those in the source images.

More information about these metrics, including their mathematical definitions, can be found in the review work by Liu et al. [27] and the original papers.

D. QUANTITATIVE AND QUALITATIVE COMPARISONS

In this study, we quantitatively compare our method to state-of-the-art techniques in the field based on the five standard metrics [8] on Lytro, MFFW, and MFI-WHU datasets. The scores show the average performance across the images of each dataset. Our method achieves the best or comparable performance among the other methods on three datasets, as shown in Table 2. For the Lytro dataset, our method outperforms the other methods on four out of five metrics, for the MFFW dataset, it obtains the best score on two metrics and the second-best score on one metric, and for the MFI-WHU dataset, it achieves the best score on one metric and the second-best score on two metrics. The outcomes show that our method is highly competitive and possesses the potential to perform above existing state-of-the-art methods. It is also worth noting that proposed method consistently outperforms the GAN-based methods on three MFIF datasets.

We present the visual MFIF results for qualitative comparison using samples from Lytro and MFFW datasets in Figures 4 and 5. To provide better observation of boundary artifacts and color degradation across the results from various approaches, we also provide difference images, which are obtained by subtracting x_2 from the fused image $\hat{y}$ (difference values are also boosted by multiplication by 5 to highlight artifacts) as well as zoomed MFIF results. Moreover, we provide the human-perception-based metric Q_{CB} (Chen-blum) [2] under the fusion results for each method to assess the visual results better, as Q_{CB} offers a meaningful assessment of the fused outcomes from a perceptual perspective. In general, we've found that focus map prediction-based methods (CNN [4], GFDF [36], MMFNet [30], SESF [29], FuseGAN [8], DRPL [18], MFIF-GAN [40], GACN [28]) commonly generate more boundary artifacts than direct fusion methods (GFF [21], CSR [25], DenseFuse [17], IFCNN [55], U2Fusion [42], MFF-GAN [50]). On the other hand, direct fusion methods tend to degrade color information. As for the proposed focus map-based method (TSGF-GAN), it produces fewer artifacts near the focus/defocus boundary than other focus map-based methods and does not degrade color information like direct fusion method MFF-GAN [50]. This can more evidently be

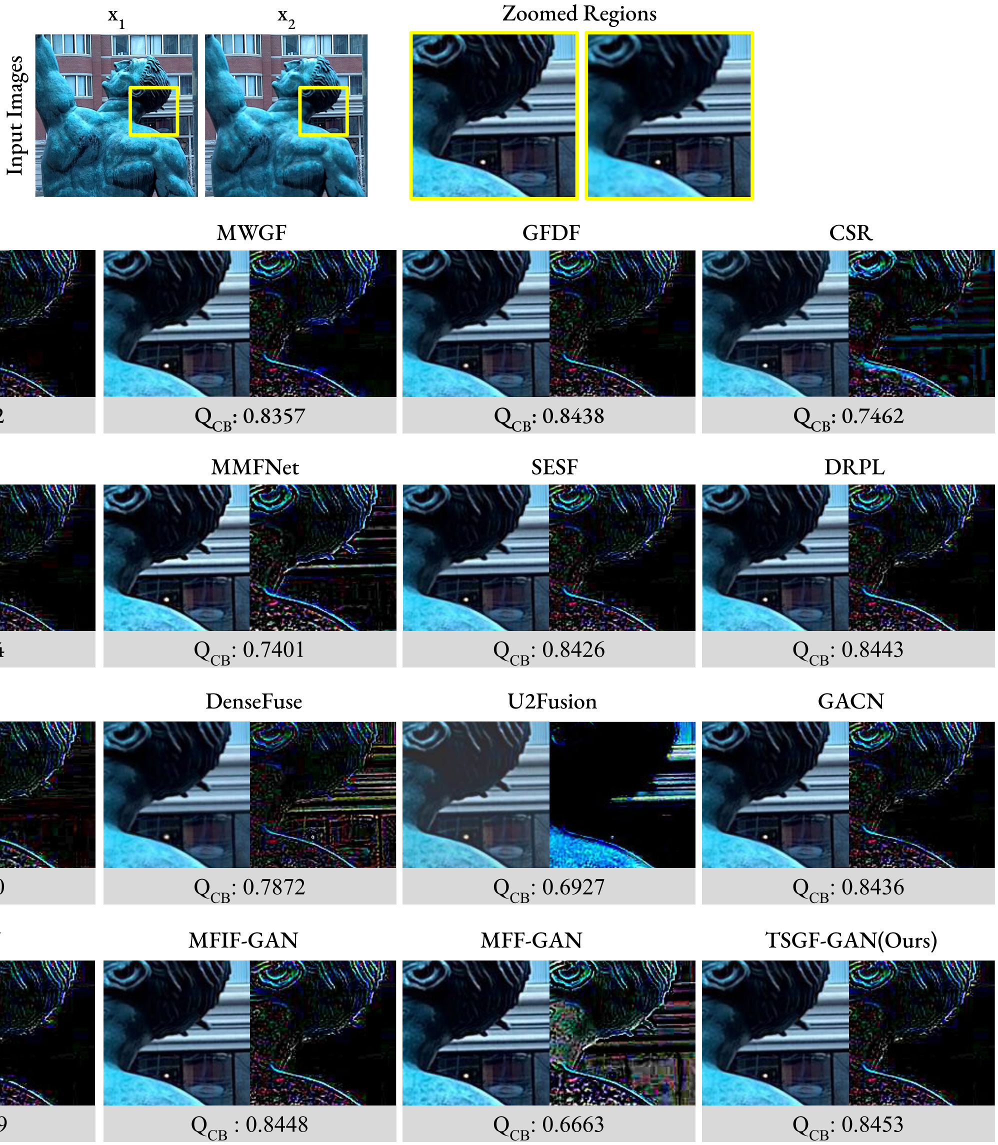

FIGURE 4. **Qualitative comparisons with state-of-the-art MFIF methods on a Lytro dataset sample. The results are given for zoomed the yellow-marked region. The resulting fused images are shown alongside the difference images ($\hat{y} - x_2$) obtained by subtracting the x_2 image from the fused images $\hat{y}$. We also provide the human-perception-based Q_{CB} (Chen-blum) metric under the fusion results of each method to assess the fusion results.**

observed in Figure 5 as the sample from the MFFW dataset is highly affected by DSE.

Overall, the findings of this study highlight the promising nature of our method, demonstrating its ability to achieve state-of-the-art or comparable performance across the evaluated metrics and visual quality on three datasets.

By integrating the TSGF as an adaptive post-processing block, our model significantly improves spatial consistency. Our model's GAN loss definition on fusion results instead of focus maps achieves better results than previous focus map-based GAN losses (FuseGAN and MFIF-GAN) as GAN loss is more successful in evaluating real images than focus maps [14]. Furthermore, incorporating the Focal Frequency Loss efficiently considers fine details and high-frequency information of the fused images, leading to improved focus map prediction. The end-to-end learning approach also contributes to the model's success, enabling it to adapt and generalize to diverse multi-focus scenarios during training.

E. ABLATION STUDY

In Figure 6, we investigate the impact of different loss functions and the trainable self-guided filter (TSGF) on the baseline model. Initially, we examine the use of GAN loss in two different aspects: GAN loss applied to the focus map (referred to as GAN-F) and GAN loss applied to the fused image (referred to as GAN-Y). FuseGAN [8] and MFIF-GAN [40] define the GAN loss based on the predicted focus map. We criticize this approach, considering that GAN loss is

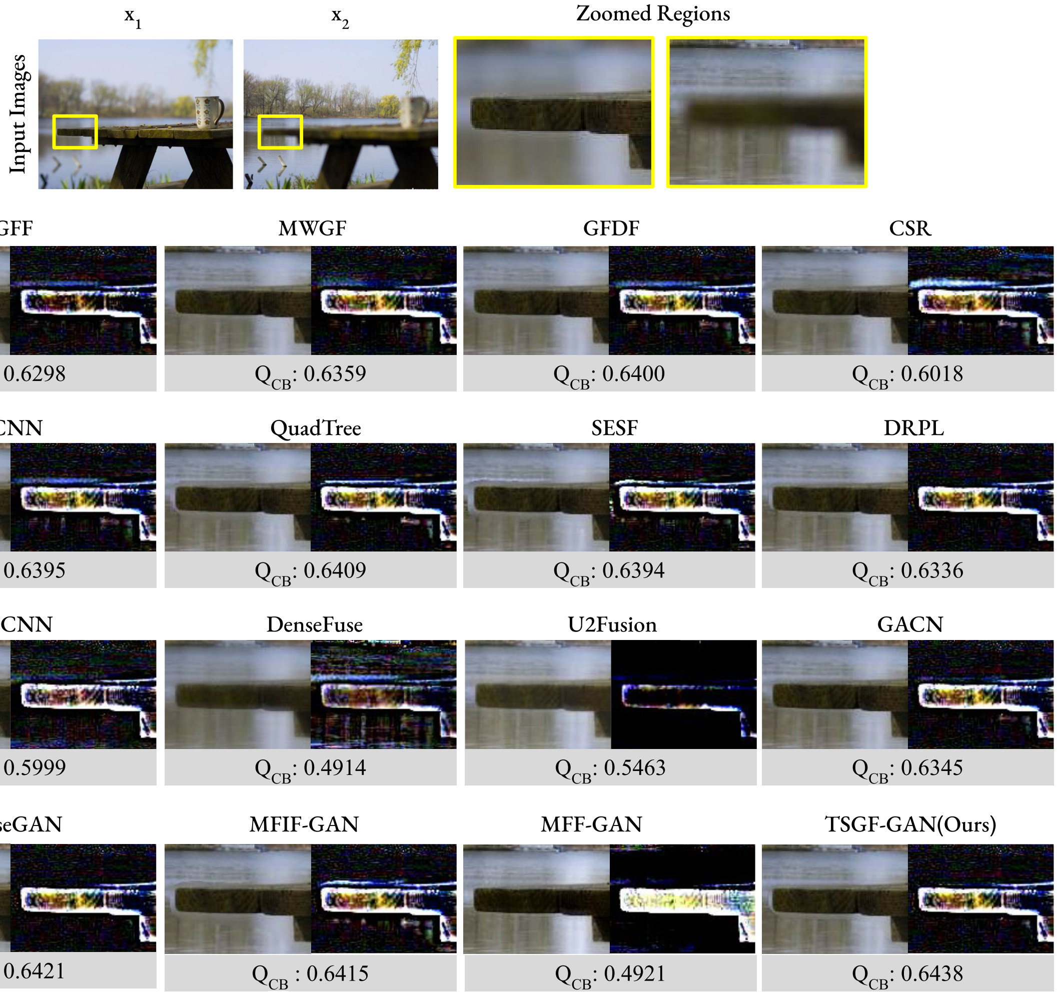

FIGURE 5. Qualitative comparisons with state-of-the-art MFIF methods on a MFFW dataset sample. The results are given for zoomed the yellow-marked region. The resulting fused images are shown alongside the difference images ($\hat{y} - x_2$) obtained by subtracting the x_2 image from the fused images $\hat{y}$. We also provide the human-perception-based Q_{CB} (Chen-blum) metric under the fusion results of each method to assess the fusion results.

known to be particularly effective for image generation tasks, where the emphasis lies on recovering fine details. Because GAN loss may not be as suitable for label prediction tasks, as discussed by Isola et al. [14]. We found the results in this direction and observed that GAN-Y produces better results (an improvement on four out of five metrics), especially on Gradient-based Fusion Metric Q_G.

We meticulously examined the impact of incorporating the trainable self-guided filter (TSGF) into the baseline GAN model during the ablation study. The TSGF is intended to improve spatial consistency and achieve adaptive focus map adjustment, particularly along the focus/defocus boundary. As shown in Figure 6, the integration of TSGF resulted in significant improvements in overall results, except Q_G (Gradient-based Fusion Metric). To preserve the performance on Q_G, we propose using the Focal Frequency Loss (FFL) [15]. The results show that employing the FFL loss further improves the model's overall performance.

As a result, our extensive ablation study demonstrates the significant positive impact of incorporating TSGF into the baseline GAN model. Furthermore, our proposed inclusion of the FFL loss allows us to maintain the gradient metric's performance while improving the overall quality of the model's output.

F. EXECUTION TIME

We compare the execution times of recent baseline GAN-based MFIF methods and the proposed TSGF-GAN in Table 3 in which we report the average required time for samples of the Lytro dataset. The experiments were carried out on a computer equipped with an Intel(R) Xeon(R) Bronze 3104 CPU @ 1.7 and a Quadro RTX 8000 GPU. Our method outperforms baseline FuseGAN [8] and MFIF-GAN [40] in terms of runtime as they use additional post-processing steps. We also include the runtime of recent

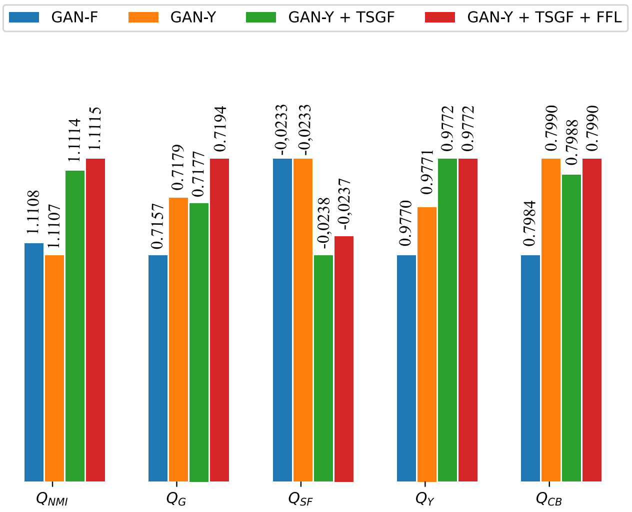

FIGURE 6. Ablation study.

TABLE 3. Required average time to fuse input images on Lytro dataset.

Method	Time(Seconds)
FuseGAN [50]	0.5576
MFF-GAN [50]	0.6548
MFIF-GAN [40]	0.2675
TSGF-GAN(Ours)	0.2309

GAN-based MFF-GAN [50] which is also slower than the proposed TSGF-GAN.

V. CONCLUSION

In this paper, we present a novel GAN-based Multi-Focus Image Fusion (MFIF) method for focus map prediction. Our approach incorporates a trainable self-guided filter (TSGF) layer to generate the focus map and uses frequency domain loss to improve focus map prediction, outperforming baseline GAN-based models. Unlike previous GAN-based methods, we define the GAN loss on the fused image rather than the predicted focus map, resulting in more effective utilization of the GANs. Testing on two multi-focus image datasets demonstrates that our method achieves superior or comparable results to state-of-the-art approaches in terms of quantitative metrics and image quality. In particular, our approach eliminates the need for post-processing steps, and the adaptive enhancement provided by TSGF improves computational efficiency.

In conclusion, we believe that our proposed method is a step forward in GAN-based multi-focus image fusion methods and it can simplify the process while improving the overall performance. We hope that this work will inspire further research in this field and pave the way for more efficient and effective GAN-based multi-focus image fusion methods using trainable filters.

Regarding the limitations of our proposed method, we have adopted the focus map prediction approach, which involves using a focus map for image fusion. While this approach is advantageous for preserving the properties of input images, it has the drawback of potentially introducing artifacts near the focus/defocus boundary. Conversely, direct fusion methods can mitigate artifacts at the boundary, but they may struggle to preserve input image properties, such as color information. To address this limitation, future research will explore a hybrid approach that combines focus map-based and direct fusion methods. Integrating these techniques may achieve a more precise and seamless multi-focus image fusion (MFIF) process, effectively mitigating artifacts while preserving the essential properties of the input images.

DECLARATION OF CONFLICT OF INTEREST

The authors have no conflict of interests to declare that are relevant to the content of this article.

REFERENCES

[1] X. Bai, Y. Zhang, F. Zhou, and B. Xue, "Quadtree-based multi-focus image fusion using a weighted focus-measure," *Inf. Fusion*, vol. 22, pp. 105–118, Mar. 2015.

[2] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1421–1432, Sep. 2009.

[3] I. De and B. Chanda, "Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure," *Inf. Fusion*, vol. 14, no. 2, pp. 136–146, Apr. 2013.

[4] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," *IEEE Access*, vol. 5, pp. 15750–15761, 2017.

[5] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, Jul. 2002, pp. 257–266.

[6] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5769–5779.

[8] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1982–1996, Aug. 2019.

[9] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[10] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on 'information measure for performance of image fusion,'" *Electron. Lett.*, vol. 44, no. 18, pp. 1066–1067, 2008.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[12] J. Huang, Z. Le, Y. Ma, F. Fan, H. Zhang, and L. Yang, "MGMDcGAN: Medical image fusion using multi-generator multi-discriminator conditional generative adversarial network," *IEEE Access*, vol. 8, pp. 55145–55157, 2020.

[13] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[15] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13899–13909.

[16] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.

[17] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[18] J. Li, X. Guo, G. Lu, B. Zhang, Y. Xu, F. Wu, and D. Zhang, "DRPL: Deep regression pair learning for multi-focus image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4816–4831, 2020.

[19] M. Li, W. Cai, and Z. Tan, "A region-based multi-sensor image fusion scheme using pulse-coupled neural network," *Pattern Recognit. Lett.*, vol. 27, no. 16, pp. 1948–1956, Dec. 2006.

[20] S. Li, R. Hong, and X. Wu, "A novel similarity based quality metric for image fusion," in *Proc. Int. Conf. Audio, Lang. Image Process.*, Jul. 2008, pp. 167–172.

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

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 740–755.

[23] S. Liu, J. Ma, L. Yin, H. Li, S. Cong, X. Ma, and S. Hu, "Multi-focus color image fusion algorithm based on super-resolution reconstruction and focused area detection," *IEEE Access*, vol. 8, pp. 90760–90778, 2020.

[24] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36, pp. 191–207, Jul. 2017.

[25] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[26] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Inf. Fusion*, vol. 23, pp. 139–155, May 2015.

[27] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.

[28] B. Ma, X. Yin, D. Wu, H. Shen, X. Ban, and Y. Wang, "End-to-end learning for simultaneously generating decision map and multi-focus image fusion result," *Neurocomputing*, vol. 470, pp. 204–216, Jan. 2022.

[29] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, and M. Mukeshimana, "SESF-fuse: An unsupervised deep model for multi-focus image fusion," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 5793–5804, Jun. 2021.

[30] H. Ma, Q. Liao, J. Zhang, S. Liu, and J.-H. Xue, "An a-Matte boundary defocus model-based cascaded network for multi-focus image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8668–8679, 2020.

[31] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[32] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, and Z. Wang, "Deep guided learning for fast multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 2808–2819, 2020.

[33] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[34] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

[35] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Inf. Fusion*, vol. 25, pp. 72–84, Sep. 2015.

[36] X. Qiu, M. Li, L. Zhang, and X. Yuan, "Guided filter-based multi-focus image fusion through focus region detection," *Signal Process., Image Commun.*, vol. 72, pp. 35–46, Mar. 2019.

[37] M. T. T. Teichmann and R. Cipolla, "Convolutional CRFs for semantic segmentation," 2018, *arXiv:1805.04777*.

[38] J. Tian and L. Chen, "Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure," *Signal Process.*, vol. 92, no. 9, pp. 2137–2146, Sep. 2012.

[39] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[40] Y. Wang, S. Xu, J. Liu, Z. Zhao, C. Zhang, and J. Zhang, "MFIF-GAN: A new generative adversarial network for multi-focus image fusion," *Signal Process., Image Commun.*, vol. 96, Aug. 2021, Art. no. 116295.

[41] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1838–1847.

[42] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[43] H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 7203–7216, 2020.

[44] S. Xu, L. Ji, Z. Wang, P. Li, K. Sun, C. Zhang, and J. Zhang, "Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1561–1570, 2020.

[45] S. Xu, X. Wei, C. Zhang, J. Liu, and J. Zhang, "MFFW: A new dataset for multi-focus image fusion," 2020, *arXiv:2002.04780*.

[46] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, p. 308, 2000.

[47] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010.

[48] Z. Yang, Y. Chen, Z. Le, F. Fan, and E. Pan, "Multi-source medical image fusion based on Wasserstein generative adversarial networks," *IEEE Access*, vol. 7, pp. 175947–175958, 2019.

[49] Z. Yang, Y. Chen, Z. Le, and Y. Ma, "GANFuse: A novel multi-exposure image fusion method based on generative adversarial networks," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 6133–6145, Jun. 2021.

[50] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Inf. Fusion*, vol. 66, pp. 40–53, Feb. 2021.

[51] Q. Zhang and B.-L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, Jul. 2009.

[52] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4819–4838, Sep. 2022.

[53] Y. Zhang, P. Zhao, Y. Ma, and X. Fan, "Multi-focus image fusion with joint guided image filtering," *Signal Process., Image Commun.*, vol. 92, Mar. 2021, Art. no. 116128.

[54] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," *Inf. Fusion*, vol. 35, pp. 81–101, May 2017.

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

[56] W. Zhao, D. Wang, and H. Lu, "Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1102–1115, Apr. 2019.

[57] Y. Zheng, E. A. Essock, B. C. Hansen, and A. M. Haun, "A new metric based on extended spatial frequency and its application to DWT based fusion algorithms," *Inf. Fusion*, vol. 8, no. 2, pp. 177–192, Apr. 2007.

[58] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," *Inf. Fusion*, vol. 20, pp. 60–72, Nov. 2014.

LEVENT KARACAN (Associate Member, IEEE) received the M.Sc. and Ph.D. degrees in computer engineering from Hacettepe University, Turkey, in 2012 and 2019, respectively. He has been an Assistant Professor with İskenderun Technical University, Turkey, since 2020. His research interests include computer vision and pattern recognition, image filtering, image and video processing, and deep generative models.