applied sciences

MDPI

Article

Patch-Wise Infrared and Visible Image Fusion Using Spatial Adaptive Weights

Syeda Minahil †, Jun-Hyung Kim † and Youngbae Hwang *

Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju 28644, Korea; SyedaMinahil@outlook.com (S.M.); mohl@naver.com (J.-H.K.)

* Correspondence: ybhwang@cbnu.ac.kr

† These authors contributed equally to this work.

Abstract: In infrared (IR) and visible image fusion, the significant information is extracted from each source image and integrated into a single image with comprehensive data. We observe that the salient regions in the infrared image contain targets of interests. Therefore, we enforce spatial adaptive weights derived from the infrared images. In this paper, a Generative Adversarial Network (GAN)-based fusion method is proposed for infrared and visible image fusion. Based on the end-to-end network structure with dual discriminators, a patch-wise discrimination is applied to reduce blurry artifact from the previous image-level approaches. A new loss function is also proposed to use constructed weight maps which direct the adversarial training of GAN in a manner such that the informative regions of the infrared images are preserved. Experiments are performed on the two datasets and ablation studies are also conducted. The qualitative and quantitative analysis shows that we achieve competitive results compared to the existing fusion methods.

Keywords: infrared and visible image fusion; Generative Adversarial Network; patchGAN; dual-discriminator; spatial adaptive weights

Citation: Minahil, S.; Kim, J.-H.; Hwang, Y. Patch-Wise Infrared and Visible Image Fusion Using Spatial Adaptive Weights. *Appl. Sci.* **2021**, *11*, 9255. https://doi.org/10.3390/app11199255

Academic Editor: Pavel Lyakhov

Received: 14 September 2021
Accepted: 30 September 2021
Published: 5 October 2021

1. Introduction

In practical applications, a fused image is essential to contain high-quality details for attaining a comprehensive representation of the real scene [1]. Nowadays, various image fusion methods have been proposed and are usually divided into several categories that include sparse representation-based methods [2,3], gradient-based methods [4], wavelet transformation-based methods [5,6], neural network-based methods [7] and deep learning based methods [8,9]. Deep learning based infrared and visible image fusion methods exploit the features of images of different modalities and integrate them into one single image with the composite information. The infrared images reflect the thermal radiation of objects [10] and visible images contain the textural information. Infrared and visible image fusion methods [1,9,11–13] extract the characteristics of both the images and achieve images that improve visual understanding and are beneficial in various fields like computer vision in object detection, recognition and military surveillance [10].

The recent fusion algorithms achieve promising results. Guided filter [14] is widely used for the purpose of image fusion which involves two-scale decomposition of the image. Then, the base layer (containing large scale variations in intensity) and detail layer (capturing small scale details) are fused together using a guided filtering based weighted average method. In Deepfuse [15], there is a Siamese based encoder with two CNN layers that extracts the features of the source images. These maps are fused by the addition strategy. The decoding network with three CNN layers reconstructs the fused image. Although it achieves better results, the network is too simple and cannot extract the salient features properly. In Densefuse [9], Ma et al. proposed a method based on dense block and an auto-encoder module. Denseblock has skip connections which help to preserve more features. The drawback of this approach is that the fusion is not considered in the training

process and only the auto-encoder is trained for the reconstruction of the images. The CNN based methods usually rely on the ground truth and in case of fusion of infrared and visible images there are no predefined standards [16]. Generative Adversarial Network (GAN)-based fusion methods are also very popular recently. These methods use traditional GAN or its variants for fusing the images, but the trained image is trained to be more similar to only one of the source images which catalyze the loss of information existing in the other image. In DDcGAN [12], GAN is applied with dual discriminators where each discriminator is tasked to make the fused image more similar to the infrared and visible image, respectively, without the ground truth.

In this paper, we apply a patch-wise discriminator inspired from the patchGAN [17] in the base structure using dual discriminators [12]. Vanilla GAN validates the authenticity of the entire generated image while the patchGAN verifies the authenticity in units of patches of $N \times N$ size of an image. By doing so, an image is only considered as real if all the patches attain a high probability of being real. This also reduces the computation speed, the number of parameters and is unaffected by the size of the images. In our approach, we also assign adaptive weights to IR image based on the observation that in the IR image only the high activation regions contains the salient information while the remaining background region contains very less or no information. We utilize these weight maps in the loss function.

To summarize, this paper makes the following contributions;

1. We propose a new framework for infrared and visible image fusion in which the patchGAN is applied to the dual discriminator network structure.
2. We introduce a new loss function based on constructing adaptive weight maps based on the IR image to preserve only the important information from both the infrared and visible images.
3. Our method produces competitive results as compared with the existing fusion methods quantitatively as well as qualitatively.

The rest of our paper is structured as follows: related works are briefly reviewed in Section 2. In Section 3, we present our proposed method in detail. In Section 4, we illustrate our experimental results and ablation studies. Finally, the conclusions are drawn in Section 5.

2. Related Works

Generative adversarial networks (GANs) [18] are one of the generative models. GANs have achieved impressive success in generating images from existing images or random noise. In GANs, we have a generator which is purposed to generate real-like fake sampled images with adversarial loss which steers the output image to be indistinguishable from the real images and to be able to fool the discriminator. The discriminator behaves as a classifier and provides the probability of the data being real or fake. The training process of a generator and a discriminator forms an adversarial process and is continued until the discriminator is unable to distinguish the generated samples. The training of GAN is a critical task because of its unstable training. To alleviate this, GANs with a conditional settings were proposed. In conditional GANs (cGAN) [19], the generator and discriminator are conditioned on some auxiliary information. This additional information is generally labeled data. This provides the guidelines to the generator in figuring out what kind of data needs to be generated. Generally, GANs use a cross entropy loss function which may lead to the vanishing gradient problem during training. To overcome this issue, Least Squares Generative Adversarial Networks (LSGANs) [20], which use least square loss for the discriminator, and WGANs [21], which use the Wasserstein loss were proposed.

The first endeavour of utilizing GAN for the task of infrared and visible image fusion was proposed in FusionGAN [22] in which the fused image is compelled to contain more texture details by introducing the discriminator to distinguish the fused image from the visible image . In GAN-based image fusion methods, the discriminator makes the generated image more similar to the visible image and this problem is alleviated in the DDcGAN [12] by the introduction of two discriminators. The generator makes the fused image and

one discriminator distinguishes the fused image from the visible image while the other discriminator distinguishes the fused image from the infrared image. The generator consists of two deconvolution layers and an auto-encoder network. The visible and a low resolution infrared images are passed through the deconvolution layers to get the same resolution. The output of the deconvolution layers are concatenated together and fed into the encoder for the feature extraction and fusion process. The fused feature maps are given to the decoder for the reconstruction of fused image which has the same resolution as the visible image. The encoder has a DenseNet consisting of short connections which improve the feature extraction. Both the discriminators have the same architecture and play an adversarial role against the generator. In DDcGAN, the discriminators are not only supposed to contemplate their adversarial role but also maintain the balance between both the discriminators. The loss of generator is the summation of adversarial loss and the weighted content loss. This adversarial loss comes from the discriminators.

In U2Fusion [23], an end-to-end semi-supervised fusion method is proposed which can fuse multi-focus, multi-exposure and multi-modal images. This method automatically approximates the significance of the source images and suggests and adaptive information preservation degree. This adaptive information preservation degree is utilized to conserve the similarity between the fusion result and source images. NestFuse [24] inspired by the DenseFuse, preserves more multi-scale features from the visible image while enhancing the salient features of infrared images. Their model is based on nest connections and Spatial/Channel attention models. The spatial attention models signifies the importance of each spatial position whereas the channel attention models uses deep features.

In PatchGAN [17], the conditional GANs are used for image-to-image translation. The architecture of the generator and the discriminators differ from the previous works utilizing GANs for the same task. The generator used has a U-Net based architecture and the discriminator is a markovian discriminator [17,25]. PatchGAN does not signify the whole image as fake or real instead of evaluating the local patches from the images. The patch size can be adjusted by changing the size of the kernel in convolution layers or the number of layers in the discriminator. In [17], an input image of 256×256 size is concatenated with the generated 256×256 image and provided to the generator G. The patch size used in this paper is 70×70. The generator provides a feature map of $30 \times 30 \times 1$ which means that each pixel of this map corresponds to the 70×70 patch of the input image. All the values of the $30 \times 30 \times 1$ feature maps are averaged to estimate the probability of the patch being real or fake. This makes patchGAN more attentive to the local features of the images. PatchGAN is now being widely used in many applications. In PGGAN [26], the patchGAN is combined with the globalGAN for the task of image inpainting. The discriminator of PGGAN, first uses the shared layers between the patchGAN and the global GAN to learn the basic low-level features which is later split to generate two separate adversarial losses to preserve both the local and the global features in images. In [27], the author proposes an image text deblurring method using two GAN networks which are used to convert the blurred images to deblurred images and the deblurred images to blurred images which helps in putting the constraints on the generated samples. The discriminator used in this model is patchGAN discriminator. PatchGAN is also used in multilevel feature fusion for underwater image color correction [28]. In this model, the multi-scale features are extracted and then global features are fused together with low-level features at each scale.

3. Proposed Method

In this section, we first elaborate our proposed end-to-end deep learning based fusion network, then we discuss the formation of weight map. At last, we introduce the design of our new loss function.

In our fusion network, we apply patchGAN [17] in the dual discriminator structure framework [12], as shown in Figure 1. Our model has one generator G and two discriminators D_i and D_v. Given the infrared image i and visible image v, the task of the generator

is to generate a fused image which should be able to fool the discriminators. D_v aims to distinguish the generated image from the visible image, while D_i is trained to discriminate between the original infrared image and the fused image. In general, the output of the discriminators is a scalar value that approximates the probability of the input from the source data rather than generated data G.

Inspired by the patchGAN [17], the generator as well as the discriminator is of the form convolution-BatchNorm-ReLu [12,29,30]. The generator's architecture is based on U-Net [31] with the auto-encoder system with skip connections. The discriminators used for our model are Markovian discriminators [17,25], "PatchGAN", that only penalize structure at the scale of patches. This discriminator tries to signify if each $N \times N$ patch in an image is real or fake. In vanilla GAN, given an input image, the output is a single probability of the image being real or fake, but here we get an $N \times N$ array of output X where each X_{ij} represents the probability of each patch of an image being real or fake.

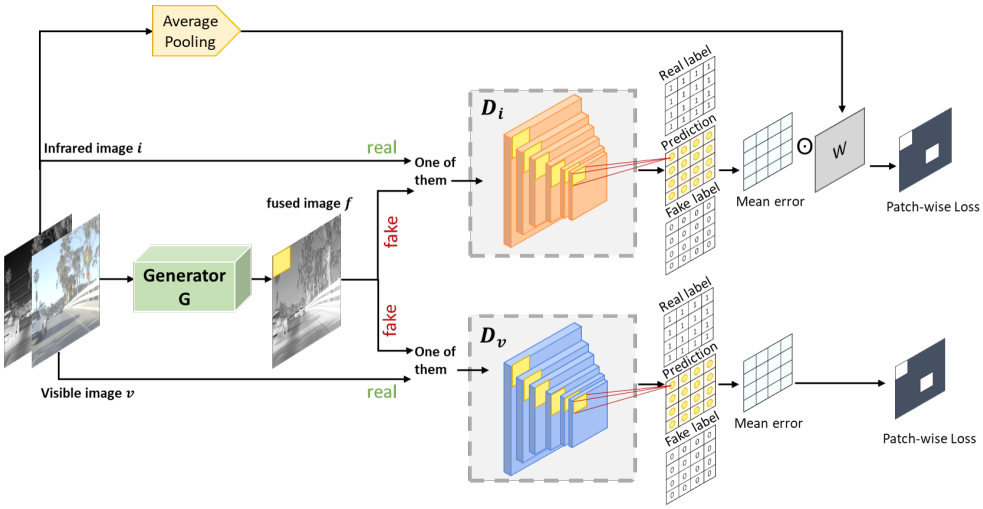

Figure 1. The overall architecture of our proposed method.

We have observed that in infrared image, only the high activation region contains significant information. The complexity in IR-visible image fusion research lies in correctly extracting the information on thermal targets from the IR image, and trying to keep obvious background information of the visible image in the fusion image [32]. We intend to give more weight to the informative regions of IR. For this purpose, we generate an adaptive weight map W based on the IR image and utilize this weight map in the adversarial loss during training in a fashion which could preserve the informative region of the IR image. For the construction of the weight map W, we take the IR image and apply average pooling. We can choose average pooling as well as max-pooling strategy to create the weight maps equal to the size of the output of the discriminator. The reason to choose average pooling is that the output of the average pooling can be considered as a smooth version of the input image. In this way, more weights can be given to high activation regions of the input infrared image. In contrast, max pooling can cause an abrupt change in weights of adjacent pixels. We have also compared these two strategies in ablation studies.The size of the weight map should match the size of the output of the discriminator.

Loss Function

Loss function plays a principle role in the learning of any model. In usual GAN, the discriminator is trained on the real and the generated images but the generator is trained indirectly via discriminator. In infrared and visible image fusion methods, the dual purposed generator is not only tasked to fool the discriminator but it should also keep the correspondence between the generated image and the source images. This is supervised by the loss function of G. Each discriminator is first trained with the input patches (i.e.,

the visible image patches in D_v and the infrared image patches in D_i) and then the fused patches. The total loss of each discriminator is the sum of both the losses which aims to discriminate between the fused image by G and the source images.

For the discriminators, the patches from the source images are real data while the patches from the generated/fused images are fake data. The discriminator D_i is first trained on the loss between the patch from input infrared image and the real label vector with each value 1. Then, this D_i is trained on the loss between the patch from fused image and the fake label vector with each value 0. Similarly, the discriminator D_v learns the loss between the patch from input visible image and the real label vector. After this, this D_v is trained on the loss between the patch from fused image and the fake label vector. We are using the mean square error loss for the discriminators.

Our purpose is to train the network in such a way that it gives more weight to the high activation region of infrared images which contains the significant information. For this, we define a new loss function for each patch by infusing the weights in the loss. We propose two methods of doing this; one is to use this new weighted loss only in the discriminator D_i as $L^*_{D_i}$, and the other method is to use it in the discriminator D_i as $L^*_{D_i}$ and generator as $L^{adv*}_{D_i}$ simultaneously. Generally, the loss of the discriminator is defined as follows:

$$L_D = \min_D V(D) = \frac{1}{2}\mathbb{E}[(D(x) - a)^2] + \frac{1}{2}\mathbb{E}[(D(G(x)) - b)^2] \tag{1}$$

where $D(x)$ and $D(G(x))$ is the output of discriminator given the image x and the generated image $G(x)$, respectively . 'a' is the target label vector which is 1 in the case of source images and 'b' is the target label vector which is 0 in the case of a fused image.

Using the new loss in D_i, the losses of both the discriminators become

$$L^*_{D_i} = \min_D V(D_i) = \frac{1}{2}\mathbb{E}[W * (D_i(i) - a)^2] + \frac{1}{2}\mathbb{E}[W * (D_i(f) - b)^2] \tag{2}$$

$$L_{D_v} = \min_D V(D_v) = \frac{1}{2}\mathbb{E}[(D_v(v) - a)^2] + \frac{1}{2}\mathbb{E}[(D_v(f) - b)^2] \tag{3}$$

The loss function of the generator consists of the content loss and the adversarial loss:

$$L_G = L_{con} + L_{adv} \tag{4}$$

L_{ssim} and L_{mse} are the structural similarity loss and the mean square error loss, respectively, between the input images and the generated/fused image. They are used as the content loss.

$$L_{con} = L_{ssim} + L_{mse} \tag{5}$$

$L^{adv}_{D_i}$ and $L^{adv}_{D_v}$ are the adversarial losses provided by the discriminators D_i and D_v. Here, mean square error loss is used as an adversarial loss. So the total loss of G becomes

$$L_{Total} = L_{ssim} + L_{mse} + \gamma L^{adv}_{D_i} + \lambda L^{adv}_{D_v} \tag{6}$$

In the case of using the new loss in the generator, we replace the mean square error loss $L^{adv}_{D_i}$ with the new loss $L^{adv*}_{D_i}$. Here, the output of the discriminator is compared with real label only.

$$L^{adv*}_{D_i} = \min_G V(G) = \frac{1}{2}\mathbb{E}[W * (D_i - a)^2] \tag{7}$$

$$L^{adv}_{D_v} = \min_G V(G) = \frac{1}{2}\mathbb{E}[(D_v - a)^2] \tag{8}$$

4. Experiments

In this section, we describe our experimental results. We have conducted extensive evaluation and comparison study against state-of-the-art algorithms including U2Fusion [23],

NestFuse [24] and DDcGAN [12]. We first conduct our experiments on the images taken from the RoadScene [23] and our private dataset. In the original patchGAN paper [17], they used an input image of size 256 × 256 and a patch size of 70 × 70. However, for our experiments, for the input image of 512 × 256, we have taken a patch size of 65 × 65 and a learning rate of 0.0001. We also change the values of γ, but for the qualitative analysis both λ and γ are fixed at 0.5. For the quantitative comparison, we select 20 images with different conditions, including indoor and outdoor, and day and night. For the verification of our results, we choose six quality metrics; Correlation coefficient (CC) that measures the degree of linear correlation between the source images and the fused image, sum of correlation of differences (SCD) [33], FMI_{dct} [34] calculates the mutual information for discrete cosine feature, modified Structural Similarity ($SSIM_a$) and multi scale SSIM (MSSSIM) for no reference image which models the loss and distortion between two images according to their similarities in light, contrast and structure information and Peak signal-to-noise ratio (PSNR).

4.1. Qualitative Analysis

The fused images attained by the three state-of-the-art algorithms and our proposed method are shown in Figures 2 and 3. We analyze the relative performance on three images from RoadScene and three from our private dataset using the new loss in discriminator $L^*_{D_i}$ as well as generator $L^{adv*}_{D_i}$.

We can see that images created by our method preserve the thermal targets from the infrared images as in the third example of Figure 2. It can also preserve more textural information from visible images such as sky (red box) in the first example of Figure 3. If we look at the overall images created by these methods, we would observe that the DDcGAN creates blurry images. NestFuse gives better results in first example of Figure 3, but some salient features are not clear such as in the red box in second image of Figure 3. Generally our proposed method tries to preserve as much information from both the infrared as well as the visible images and tries to maintain the overall good quality of images, simultaneously visible in the third and the fourth images.

4.2. Quantitative Analysis

For the quantitative comparison, we take the average value of 20 fused images for each metric. In this comparison, we analyze the new weighted loss in discriminator D_{ir} and generator G, and also in discriminator D_{ir} only. In Tables 1 and 2, we have taken fixed values of $\lambda = 0.5$ and $\gamma = 0.1$. Table 1 displays the values of different metrics for RoadScene dataset while Table 2 displays the values for our private dataset. The best values are indicated in red, the second best values are indicated in blue, and the third highest values in cyan.

From Table 1, we can see that as compared to the state-of the-art methods, our method achieves the top-2 results in CC and $SSIM_a$. The second and third best results in MS-SSIM, SCD and PSNR. The second best result has a narrow margin of only 0.018 from the best result in PSNR, a difference of only 0.0252 from the best result in MS-SSIM and a difference of 0.0386 from the best result in SCD.

In Table 2, as compared to the other state-of the-art methods, our method acquires the top-2 ranks in MS-SSIM, FMI_{dct} which calculates the mutual information for discrete cosine features and $SSIM_a$. The second best and the third best results in CC, SCD and PSNR. CC achieves the second best result with a difference of 0.0017 from the best result. SCD has a difference of only 0.0002 and PSNR has a margin of 0.16 from the best results. The highest values in FMI_{dct} and $SSIM_a$ indicate that our method attains more features and structural information. Highest value in PSNR indicates that the fused image is more similar to the source images and is of higher quality with less distortion. In general, our method gives better performance than DDcGAN by simply replacing the GAN with PatchGAN and adding weights. These results prove the effectiveness of our method.

Figure 2. Qualitative comparison of the proposed method with other state-of-the-art methods on image pairs taken from the private dataset. The first row contains visible images. Second row contains infrared images. The last row contains the fused results of our proposed method.

We also compare the different values of γ in Tables 3 and 4. Here, the best values are indicated in red and the second best values are indicated in blue. From Tables 3 and 4, we can witness that most of the highest results are achieved with $\lambda = 0.5$ and $\gamma = 0.1$.

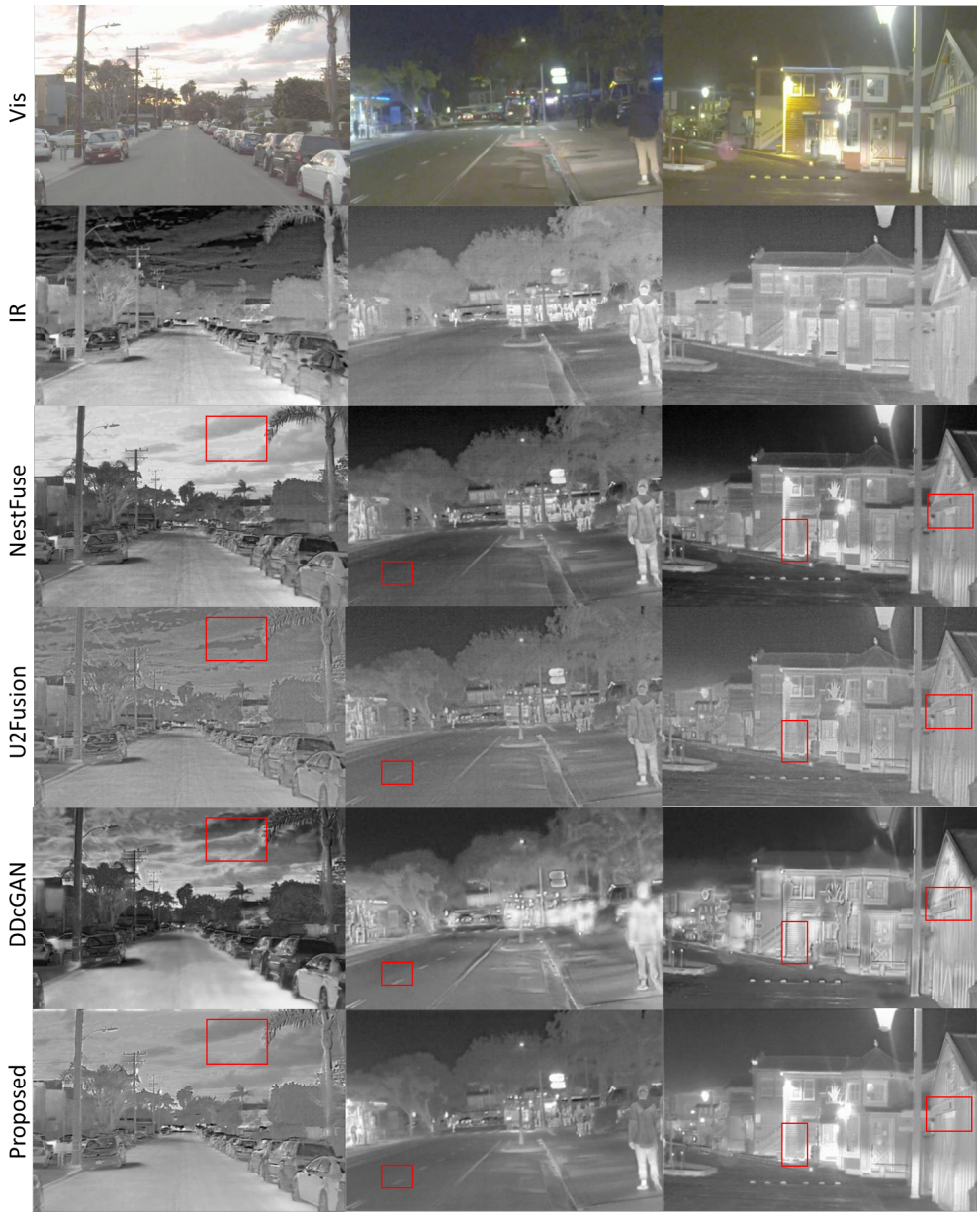

Figure 3. Qualitative comparison of the proposed method with other state-of-the-art methods on image pairs taken from the RoadScene dataset. The first row contains visible images. Second row contains infrared images. The last row contains the fused results of our proposed method.

Table 1. The average values of quality metrics for 20 fused images of our RoadScene dataset with $\lambda = 0.5$ and $\gamma = 0.1$ used in the loss function. The best values are indicated in red, the second best values are indicated in blue and the third highest values in cyan.

Methods		CC	*MSSSIM*	SCD [33]	FMI_{dct} [34]	$SSIM_a$	PSNR
U2Fusion [23]		1.2199	0.8907	1.3236	0.3057	0.7204	16.0903
NestFuse [24]		1.1889	0.8376	1.6574	0.2969	0.6598	13.8161
DDcGAN [12]		1.1752	0.7067	1.5041	0.3589	0.5965	13.8019
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.2675	0.8655	1.6188	0.2691	0.7381	16.0222
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.2587	0.8625	1.5609	0.2736	0.7396	16.0723

Table 2. The average values of quality metrics for 20 fused images of our private dataset with $\lambda = 0.5$ and $\gamma = 0.1$ used in the loss function. The best values are indicated in red, the second best values are indicated in blue and the third highest values in cyan.

Methods		CC	$MSSSIM$	SCD [33]	FMI_{dct} [34]	$SSIM_a$	PSNR
U2Fusion [23]		1.3785	0.8954	0.9211	0.1953	0.7426	19.2835
NestFuse [24]		1.4732	0.899	1.4616	0.2369	0.7322	18.7873
DDcGAN [12]		1.2685	0.7785	1.1741	0.1879	0.5409	11.5284
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.4715	0.9068	1.4614	0.2684	0.7594	19.0826
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.4701	0.905	1.3959	0.2714	0.7541	19.1149

Table 3. The average values of quality metrics for 20 fused images of our RoadScene dataset. Different values of λ and γ are used in the loss function. The best values are indicated in red and the second best values are indicated in blue.

Methods		CC	$MSSSIM$	SCD [33]	FMI_{dct} [34]	$SSIM_a$	PSNR
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.2675	0.8655	1.6188	0.2691	0.7381	16.0222
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.5$	1.2549	0.8785	1.6598	0.2501	0.7319	15.8065
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 5$	1.2379	0.8601	1.6986	0.2346	0.7139	15.2591
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.2587	0.8625	1.5609	0.2736	0.7396	16.0723
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.5$	1.2567	0.8764	1.6566	0.2569	0.7326	15.8283
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 5$	1.2165	0.8547	1.6916	0.24	0.6998	14.7879

Table 4. The average values of quality metrics for 20 fused images of our private dataset. Different values of λ and γ are used in the loss function. The best values are indicated in red and the second best values are indicated in blue.

Methods		CC	$MSSSIM$	SCD [33]	FMI_{dct} [34]	$SSIM_a$	PSNR
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.4715	0.9068	1.4614	0.2684	0.7594	19.0826
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.5$	1.4787	0.8933	1.3735	0.2702	0.7615	19.17
ours ($L^*_{D_i}$, $L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 5$	1.3997	0.8427	0.8844	0.2508	0.7515	19.2911
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.1$	1.4701	0.905	1.3959	0.2714	0.7541	19.1149
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 0.5$	1.4702	0.8841	1.3156	0.2527	0.7542	19.1338
ours ($L^{adv*}_{D_i}$)	$\lambda = 0.5$, $\gamma = 5$	1.4239	0.856	0.956	0.2198	0.7543	19.2189

4.3. Ablation Studies

In order to illustrate the effect of the gamma on each metric for both type of losses, we perform extensive experiments and the results are summarized in the Figure 4. Gamma indicates different weight for adversarial loss of D_i, see Equation (6). We alter gamma as we intend to see the effect of infrared images. For our experiments, we choose γ = 0.1, 0.2, 0.3, 0.5, 0.8, 2, 3 and 5. We can observe that out of the six metrics, four gives highest results for gamma equal to 0.1 and 0.2. However, for each metric except SCD the values are highest between 0 and 1. After 1, the values start to decrease.

We show the performance results with two techniques of using the new loss function. We also perform experiments with different values of gamma. We witness that the γ between 0 and 1 yields the best results for each quality metrics expect SCD in Figure 4. We also observe the values for both techniques of using the new loss in D_i only, and in both D_i and G and see that the loss used in both D_i and G delivers good performance overall.

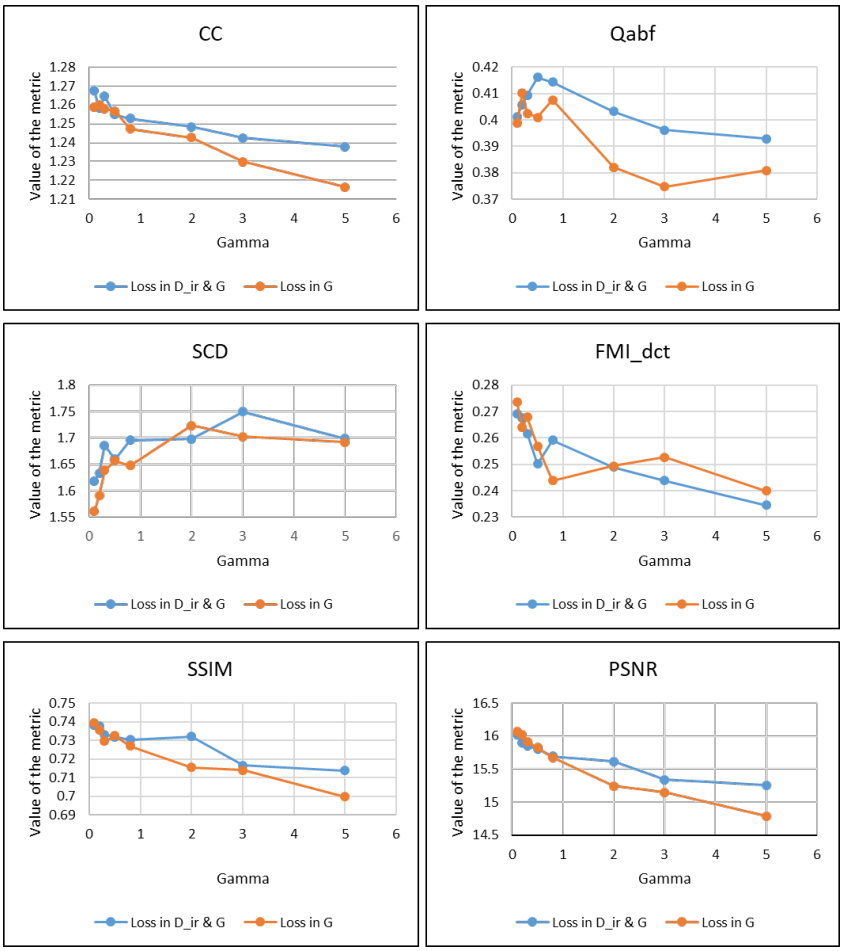

Figure 4. Values of each metric for images from RoadScene Dataset with different γ. Lambda is fixed to 0.5 for this analysis. Type of the loss used is shown in the legends.

There are two ways to construct the weight maps equal to the size of the output of discriminator, average pooling and maximum pooling. In Table 5, we analyze the effect of using both pooling strategies for the construction of the weights.The settings for this analysis include the weighted loss used in both D_i and G with $\lambda = 0.5$ and $\gamma = 0.1$. Based on these results, it is quite evident that average pooling performs better than maximum pooling.

Table 5. The average values of quality metrics for 20 fused images on two datasets with different pooling methods. The best values are indicated in blue.

Dataset	Methods	CC	$MSSSIM$	SCD [33]	FMI_{dct} [34]	$SSIM_a$	PSNR
RoadScene	Avg-Pooling	1.2675	0.8655	1.6188	0.2691	0.7381	16.0222
RoadScene	Max-Pooling	1.2637	0.8617	1.5528	0.2671	0.7375	16.0355
Private	Avg-Pooling	1.4715	0.9068	1.4614	0.2684	0.7594	19.0826
Private	Max-Pooling	1.4467	0.8988	1.3488	0.2594	0.7508	19.224

5. Conclusions

This paper presents a new end-to-end trainable framework where patchGAN is used with dual discriminators. The advantage of using the patchGAN is that it tries to signify if each $N \times N$ patch in an image is real or fake rather than determining the entire image, allowing us to observe features that are otherwise hard to perceive. The fundamental characteristic of the proposed method is the weights derived from the infrared images. These weights are utilized in defining the new loss function which can be used in two ways; (1) Use the new weighted loss in infrared discriminator and as an adversarial loss

in generator, (2) Use the new loss as adversarial loss in generator only. These weight maps can help achieve our objective of accurately extracting the information from the highly informative regions of infrared as well as visible images. The experiments are conducted on RoadScene and our private dataset to evaluate the performance of our proposed method qualitatively as well as quantitatively. The experimental results indicate that the images fused by the proposed method contain more details and are more vivid. The proposed technique is simple yet effective and achieves better results than the state-of-the-art methods.

Author Contributions: Conceptualization, J.-H.K. and Y.H.; methodology, S.M. and J.-H.K.; software, S.M.; validation, S.M. and J.-H.K.; writing—original draft preparation, S.M.; writing—review and editing, Y.H.; visualization, S.M.; supervision, J.-H.K. and Y.H.; project administration, Y.H.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Grand Information Technology Research support program (IITP-2020-0-01462), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1077110).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y.; Dong, L.; Ji, Y.; Xu, W. Infrared and Visible Image Fusion through Details Preservation. *Sensors* **2019**, *19*, 4556. [CrossRef] [PubMed]
2. Liu, Y.; Yang, X.; Zhang, R.; Albertini, M.K.; Celik, T.; Jeon, G. Entropy-Based Image Fusion with Joint Sparse Representation and Rolling Guidance Filter. *Entropy* **2020**, *22*, 118. [CrossRef]
3. Jiang, W.; Yang, X.; Wu, W.; Liu, K.; Ahmad, A.; Sangaiah, A.K.; Jeon, G. Medical images fusion by using weighted least squares filter and sparse representation. *Comput. Electr. Eng.* **2018**, *67*, 252–266. [CrossRef]
4. Shao, Z.; Wu, W.; Guo, S. IHS-GTF: A Fusion Method for Optical and Synthetic Aperture Radar Data. *Remote Sens.* **2020**, *12*, 2796. [CrossRef]
5. Chipman, L.; Orr, T.; Graham, L. Wavelets and image fusion. In Proceedings of the International Conference on Image Processing, Washington, DC, USA, 23–26 October 1995; Volume 3, pp. 248–251.
6. Lewis, J.; O'Callaghan, R.; Nikolov, S.; Bull, D.; Canagarajah, N. Pixel- and region-based image fusion with complex wavelets. *Inf. Fusion* **2007**, *8*, 119–130. [CrossRef]
7. Xiang, T.Z.; Yan, L.; Gao, R. A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking PCNN in NSCT domain. *Infrared Phys. Technol.* **2015**, *69*, 53–61. [CrossRef]
8. Li, H.; Wu, X.J.; Kittler, J. Infrared and visible image fusion using a deep learning framework. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2705–2710.
9. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef] [PubMed]
10. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [CrossRef]
11. Xu, D.; Wang, Y.; Xu, S.; Zhu, K.; Zhang, N.; Zhang, X. Infrared and Visible Image Fusion with a Generative Adversarial Network and a Residual Network. *Appl. Sci.* **2020**, *10*, 554. [CrossRef]
12. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.P. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [CrossRef]
13. Zhao, F.; Zhao, W.; Yao, L.; Liu, Y. Self-supervised feature adaption for infrared and visible image fusion. *Inf. Fusion* **2021**, 76, 189–203. [CrossRef]
14. Li, S.; Kang, X.; Hu, J. Image Fusion With Guided Filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875. [PubMed]
15. Prabhakar, K.; Srikar, V.; Babu, R. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4724–4732.
16. Sun, C.; Zhang, C.; Xiong, N. Infrared and Visible Image Fusion Techniques Based on Deep Learning: A Review. *Electronics* **2020**, *9*, 2162. [CrossRef]
17. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-To-Image Translation With Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
18. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27: 28th Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
19. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.

20. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
21. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sudney, Australia, 7–9 August 2017; pp. 214–223.
22. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
23. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
24. Li, H.; Wu, X.J.; Durrani, T. NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656. [CrossRef]
25. Li, C.; Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 702–716.
26. Demir, U.; Ünal, G.B. Patch-Based Image Inpainting with Generative Adversarial Networks. *arXiv* **2018**, arXiv:1803.07422.
27. Wu, C.; Du, H.; Wu, Q.; Zhang, S. Image Text Deblurring Method Based on Generative Adversarial Network. *Electronics* **2020**, *9*, 220.
28. Liu, X.; Gao, Z.; Chen, B.M. MLFcGAN: Multilevel feature fusion-based conditional GAN for underwater image color correction. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1488–1492. [CrossRef]
29. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2016**, arXiv:1511.06434.
30. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
31. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceeding of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
32. Zuo, Y.; Liu, J.; Bai, G.; Wang, X.; Sun, M. Airborne Infrared and Visible Image Fusion Combined with Region Segmentation. *Sensors* **2017**, *17*, 1127. [CrossRef] [PubMed]
33. Aslantas, V.; Bendes, E. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU-Int. J. Electron. Commun.* **2015**, *69*, 1890–1896. [CrossRef]
34. Haghighat, M.; Razian, M.A. Fast-FMI: Non-reference image fusion metric. In Proceedings of the 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan, 15–17 October 2014; pp. 1–3.