information

MDPI

Article

Attentive Generative Adversarial Network with Dual Encoder-Decoder for Shadow Removal

He Wang [1,2], Hua Zou [2,*] and Dengyi Zhang [2]

[1] Information Science Academy of China Electronics Technology Group Corporation, Beijing 100086, China
[2] School of Computer Sciences, Wuhan University, Wuhan 430010, China
[*] Correspondence: zouhua@whu.edu.cn

Abstract: Shadow removal is a fundamental task that aims at restoring dark areas in an image where the light source is blocked by an opaque object, to improve the visibility of shadowed areas. Existing shadow removal methods have developed for decades and yielded many promising results, but most of them are poor at maintaining consistency between shadowed regions and shadow-free regions, resulting in obvious artifacts in restored areas. In this paper, we propose a two-stage (i.e., shadow detection and shadow removal) method based on the Generative Adversarial Network (GAN) to remove shadows. In the shadow detection stage, a Recurrent Neural Network (RNN) is trained to obtain the attention map of shadowed areas. Then the attention map is injected into both generator and discriminator to guide the shadow removal stage. The generator is a dual encoder-decoder that processes the shadowed regions and shadow-free regions separately to reduce inconsistency. The whole network is trained with a spatial variant reconstruction loss along with the GAN loss to make the recovered images more natural. In addition, a novel feature-level perceptual loss is proposed to ensure enhanced images more similar to ground truths. Quantitative metrics like PSNR and SSIM on the ISTD dataset demonstrate that our method outperforms other compared methods. In the meantime, the qualitative comparison shows our approach can effectively avoid artifacts in the restored shadowed areas while keeping structural consistency between shadowed regions and shadow-free regions.

Keywords: attention mechanism; dual encoder-decoder; shadow removal

check for updates

Citation: Wang, H.; Zou, H.; Zhang, D. Attentive Generative Adversarial Network with Dual Encoder-Decoder for Shadow Removal. *Information* **2022**, *13*, 377. https://doi.org/10.3390/info13080377

Academic Editor: Chenglizhao Chen

Received: 9 May 2022
Accepted: 31 July 2022
Published: 5 August 2022

1. Introduction

The presence of shadows in images is one of the main challenges for various computer vision tasks, such as object detection and tracking [1,2]. Many shadow removal methods [3–6] have been proposed to restore shadowed regions to shadow-free regions. With the rapid development of deep learning, Convolutional Neural Networks(CNNs) and Recurrent Neural Networks(RNNs) have been widely used in the detection and removal of shadows [7–9] and achieved remarkable performance. However, most of these methods use one shared network to process both shadowed areas and shadow-free areas, making some restored shadow areas inconsistent with the surroundings in terms of color, brightness, and textures. Since the information contained in different regions is unequal, it is irrational to use a shared network.

In this paper, we propose a novel two-stage framework based on the Generative Adversarial Network (GAN) to enhance shadowed images. Firstly, we introduce an effective attention mechanism in the generator to utilize the contextual information of shadowed regions. Specifically, an Attentive-Recurrent Network (ARN) consisting of residual blocks [10], Long Short-Term Memory (LSTM) [11], and convolutional layers, is proposed to generate the attention map of shadowed areas. Then, the original shadow image concatenated with the attention map is fed into two encoder-decoder networks to remove shadows. The attention map is also used to compute a spatial variant loss [12] so that the generator pays more attention to shadows. In addition, to make the generated image more similar to the ground truth, we apply a perceptual loss at the feature level to constrain the output and the ground truth. Finally, the generated shadow-free image will be delivered into a discriminator composed of seven convolutional blocks and a linear layer for adversarial training. Some examples processed by our approach can be found in Figure 1.

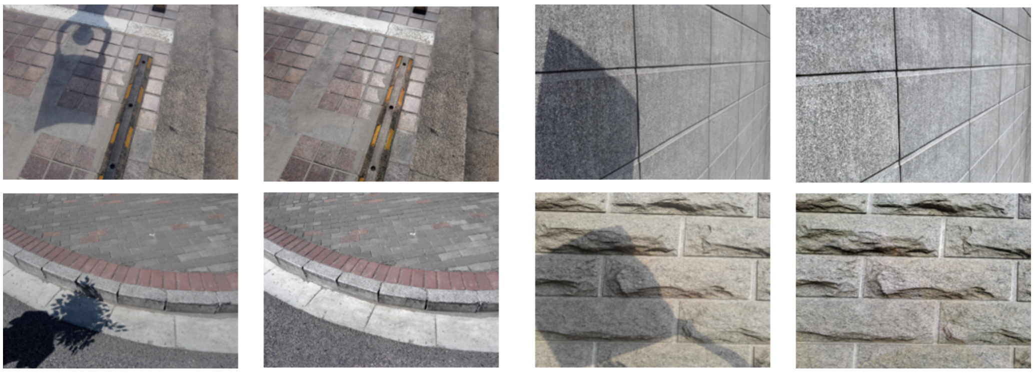

Figure 1. Some shadowed images and the corresponding results produced by our approach. The first column and the third column are the shadow images, while the second column and the fourth column are the results generated by our method. Our predicted images are natural, in which the corresponding lit region of each shadow region is more similar to its surroundings.

Our network is trained and tested on the widely used the Image Shadow Triplets dataset(ISTD) [9] which contains shadow images, masks, and shadow-free images as shown in Figure 2. To the best of our knowledge, this study is the first work to recover the shadow image with high consistency between shadowed regions and shadow-free regions. We highlight our contributions as follows:

- We design a two-stage network, in which the result of shadow detection is regarded as the attention map to guide the shadow removal. In particular, the attention map is used to compute the spatial variance loss [12] so that the network can focus more on shadows.
- Novel dual encoder-decoder modules are proposed to process shadowed regions and shadow-free regions separately in order to reduce the inconsistency. The input of encoder-decoder modules is the concatenation of the attention map and shadowed image.
- A feature-level perceptual loss is applied to ensure the similarity between the generated image and the ground truth.

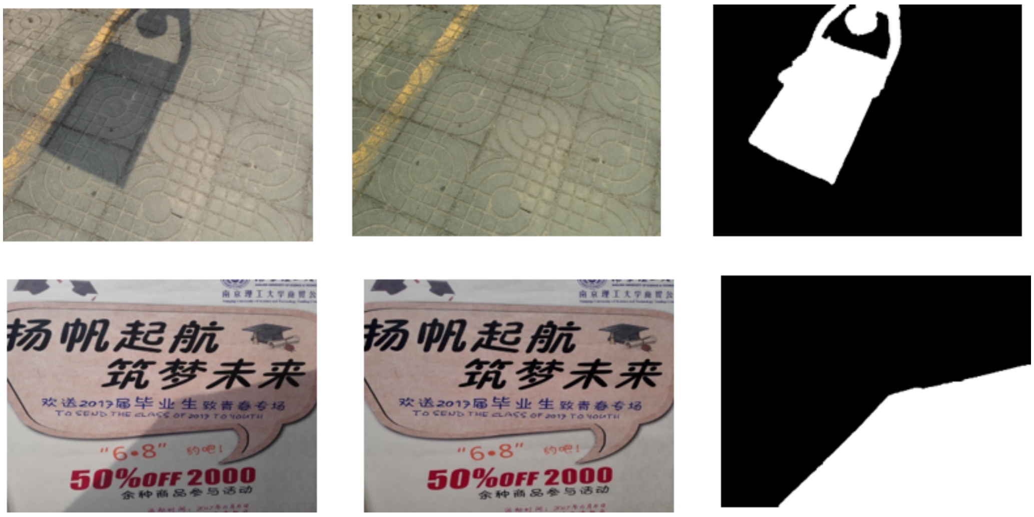

Figure 2. Some samples from the Image Shadow Triplets dataset (ISTD) [9]. The first column is the shadowed image, the second column represents the ground truth, and the last column is the mask of shadows. The purpose of our work is to restore a shadowed image into a shadow-free image and keep it similar to the ground truth as much as possible.

The rest of this paper is organized as follows. We review some related works in Section 2. Section 3 describes the details of our approach for shadow removal. Then, we present the experimental results to evaluate the superiority of the proposed method in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

2.1. Shadow Detection

The traditional shadow removal methods are based on physical modeling of illumination and color [13–15]. They learn the shadow properties through hand-crafted features such as color [16–18], texture [17–19], and edges [16,19,20] on annotated shadow images. These methods can be classified into two categories: decision tree [16,19] and SVM [17,18,20]. Due to the difficulty of designing and choosing hand-crafted features, they can only process simple situations.

Because of the powerful learning ability of the convolution network, more and more methods [21–25] are based on it. These methods outperform previous approaches greatly. Khan et al. [26] introduced deep CNNs to automatically learn features for shadow regions/boundaries. Vicente et al. [27] trained stacked-CNN using a large dataset with noisy annotations. Hu et al. [7] proposed a novel network for shadow detection by harvesting direction-aware spatial context, in which the direction-aware attention mechanism is introduced in a spatial recurrent neural network (RNN). The study [28] further extended [7] on more datasets to evaluate the performance. These methods only consider shadow regions, thus, they omit the consistency of the whole image. Unlike the previous methods, we introduce an attention mechanism in shadow detection, which can make use of the contextual information and focus on the shadow edge simultaneously.

2.2. Shadow Removal

Early works are motivated by physical models of illumination and color. For instance, Finlayson et al. [13,14] provided illumination-invariant solutions that work well only on high-quality images. More recently, there are some methods [29–32] to remove shadows using deep networks. Khan et al. [5,26] adopted CNN to detect shadows followed by a Bayesian model to remove shadows. Qu et al. [8] developed three sub-networks to extract features from multiple views and remove shadows. Wang et al. [9] used two Conditional Generative Adversarial networks (CGAN) to detect shadows and remove shadows, respectively. Hu et al. [7] explored the direction-aware spatial context to detect and remove shadows. However, these methods have problems in some situations. Because the shadowed regions and shadow-free regions are trained together, some image properties between shadowed areas and shadow-free areas like color, texture, and illumination are inconsistent. In contrast, our method processes the two regions separately to make the image more natural.

Our proposed method is also based on generative adversarial networks. However, compared with the existing mainstream deep learning methods [7,8], our proposed method is more concerned with solving the inconsistency in the shadowed and shadow-free regions. A dual encoder-decoder structure is used in the generation phase to reduce the inconsistency by processing the shadowed and shadow-free regions separately.

3. Proposed Method

Figure 3 shows the overall architecture of the network. As we can see, the network contains two essential parts: the generator and the discriminator. The generator aims to remove shadows to enhance the input image. Specifically, the generative part can be divided into two stages: shadow detection and shadow removal. We utilize the detection result as the attention map to guide shadow removal. Finally, the enhanced image and the attention map are fed into the discriminator for adversarial training. We detail the whole network as described below.

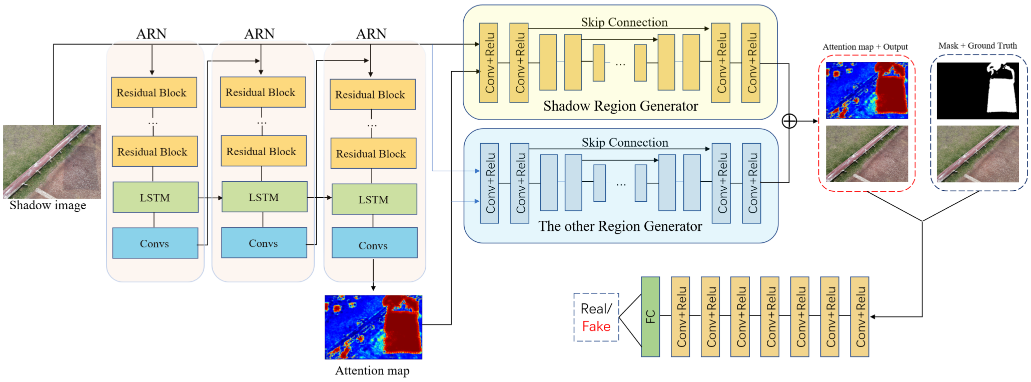

Figure 3. The architecture of our proposed network. The generator consists of an Attentive Recurrent Network (ARN) and two contextual autoencoders. The discriminator contains seven convolution layers and a fully connected layer.

3.1. *Generative Netwoks*

As we can see in Figure 3, an Attentive-Recurrent Network (ARN) is used as the detection network. Meanwhile, the removal network is a dual autoencoder following the encoder-decoder architecture. The detection result is regarded as the attention map and concatenated with the input image as prior. After that, the concatenated results are fed into the autoencoder so that the autoencoder will focus on the shadowed regions to produce a better restored image.

3.1.1. Attentive-Recurrent Network

As shown in Figure 4, each block (of each time step) in ARN comprises five layers of ResNet [10], a convolutional LSTM [11] unit, and convolutional layers for generating the 2D attention maps. The shortcuts in residual blocks can avoid vanishing gradients and overfitting since they acquire more features than the normal convolution units. In order to improve the detection performance, we use LSTMs to focus on spatial information and contextual relations of shadows.

The value of generated attention map is between 0 and 1. The higher the value is, the greater attention it suggests. Unlike the existing detection methods, attentive networks can concentrate on spatial features and contextual relations. We empirically assume that the network delivers higher performance if we share the attention weights rather than using different weights. In order to make the weight of each block can be shared, we make the output of LSTM in each block as input to the LSTM of the next block. In each part of the ARN, the i^{th} attention map will be fed to the $(i+1)^{th}$ part with the shadow image. More details can be seen in Figure 4.

We set the initialized value of the first attention map to 0.5 when training. For each block in ARN, we concatenate the attention map with the shadow image and then feed them into the next block. Particularly, the later attention maps have larger values indicating the increase in confidence. The loss function is defined as the sum of mean squared error (MSE) between the output attention map and the binary mask at all time steps shown in Formula (1).

$$\mathcal{L}_{ATT}(\{A\}, M) = \sum_{t=1}^{N} \theta^{N-t} L_{MSE}(A_t, M) \tag{1}$$

where A_t means the attention map of step t. When t is 1, the input of the ARN block is the shadow image stacked with an initial attention map whose value is 0.5. We set the total time step N to 3 and θ to 0.8. In fact, a higher N will produce a better attention map, but it will need more memory.

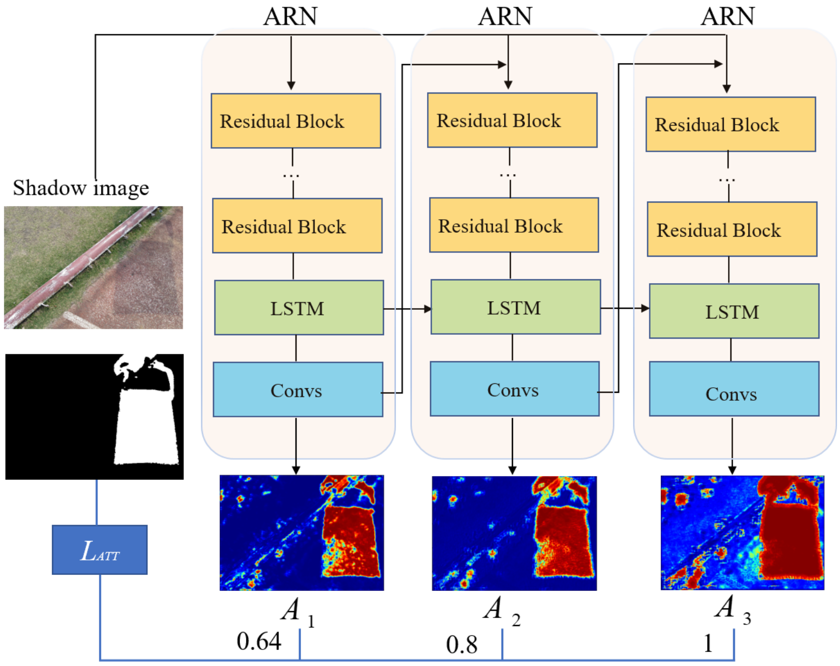

Figure 4. The architecture of our attentive-recurrent network. The input is a shadowed image, and the output of each ARN is fed to the next ARN. *A*1, *A*2, and *A*3 are the attention maps of each ARN. The color means the attention weight of different regions.

3.1.2. Contextual Encoder

The contextual encoder is used to convert the shadow image into a clean image. The existing methods use the same network to process shadowed and shadow-free areas. However, the initial information on the two areas is different. Therefore, it is unreasonable to treat the shadow regions and other regions in the same way. To solve this problem, we apply two separate networks with the same structure, which contains 16 Conv-Relu blocks, to generate the result. Additionally, we also use three skip connections to prevent overfitting and retain more details. The input of the autoencoder is the concatenation of the last attention map and the shadow image. It is worth noting that the restored areas (shadowed regions in the original image) in the output of encoder 1 and non-shadowed areas (shadow-free regions in the original image) in the output of encoder 2 are fused to obtain the final enhanced image. The network architecture can be seen in Figure 5.

We use a reconstruction loss to make the whole image similar to the ground truth as below:

$$\mathcal{L}_{RES}(\bar{I}, I_{gt}) = \mathcal{L}_{MSE}(\bar{I}, I_{gt}) \tag{2}$$

where $\mathcal{L}_{MSE}(\bar{I}, I_{gt})$ means the mean square error [33] between the generative result and the ground truth.

In order to make our networks concentrate on the shadow region and the surrounding structures, we further use the attention map M_{ATT} to compute the spatial variant loss [12] as Formula (3). The attention maps from the attentive-recurrent network are not only detection of shadows, but also spatial locations and their relative order by considering confidence on both known and unknown pixels.

$$\mathcal{L}_{SVL}(M_{ATT}; \bar{I}, I_{gt}) = M_{ATT} \bigodot L_{MSE}(\bar{I}, I_{gt}) \tag{3}$$

We also introduce a perceptual loss at the feature level as Formula (4) to measure the global discrepancy between the autoencoder's output and the ground truth. The features are extracted by a well-trained VGG-16 [34].

$$\mathcal{L}_{VGG}(\bar{I}, I_{gt}) = \mathcal{L}_{MSE}(VGG(\bar{I}), VGG(I_{gt})) \tag{4}$$

The loss function of generative network is:

$$\mathcal{L}_G = 10^{-2}\mathcal{L}_{GAN}(\bar{I}) + \mathcal{L}_{ATT}(A, M) + \mathcal{L}_{SVL}(M_{ATT}; \bar{I}, I_{gt}) + \mathcal{L}_{VGG}(\bar{I}, I_{gt}) \tag{5}$$

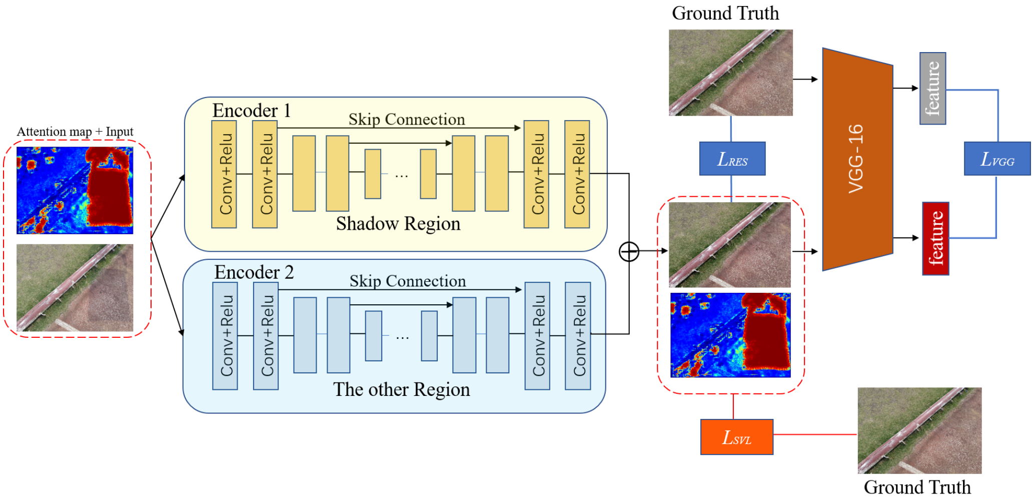

Figure 5. The architecture of the autoencoder. It contains two encoder-decoders with three skip connections. The final output will be used to compute $\mathcal{L}_{MSE}$, $\mathcal{L}_{RES}$, $\mathcal{L}_{SVL}$.

3.2. Discriminative Network

The purpose of the discriminative network is to distinguish whether the image is fake or real. We set the generated image concatenated with the attention map as fake (0), and the ground truth concatenated with the mask as real (1). Our discriminative network contains seven 3 × 3 convolution layers and a fully connected layer with a sigmoid activation function. The result of the discriminator is 0 or 1, which means fake or real, respectively. Figure 6 shows the structure of the discriminator.

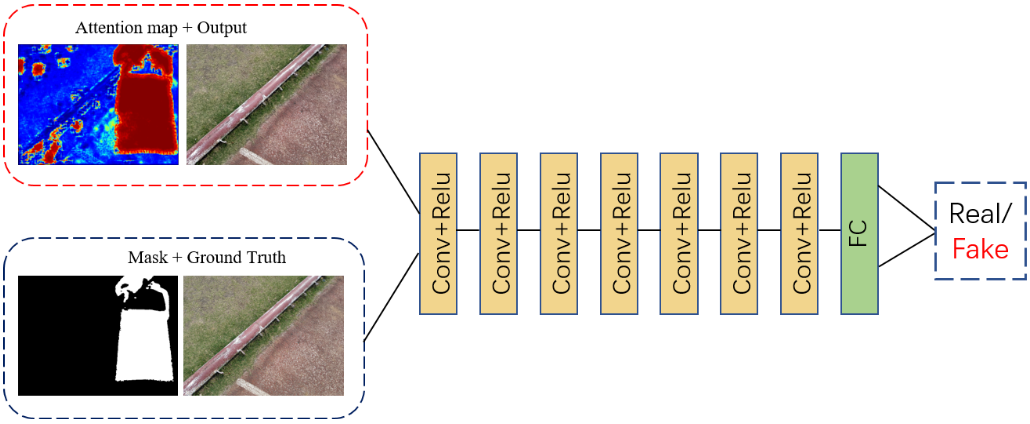

Figure 6. The architecture of the discriminator. The input of the network is the generation concatenated with attention map or the ground truth concatenated with the mask. The output of the discriminator is 0/1, which means fake or real.

The loss function in the discriminative network is defined as:

$$\mathcal{L}_D = \underset{G}{min}\underset{D}{max} E_{I_{gt} \sim P_{clean}}[\log D(I_{gt})] + E_{I \sim P_{shadow}}[\log(1 - D(G(I)))] \tag{6}$$

where G represents the generative network, and D represents the discriminative network. I_{gt} is the ground truth, and I means the input shadowed image.

3.3. Implement Details

The VGG-16 [34] which is used to extract features to calculate the perceptual loss is initialized with the model pre-trained on ImageNet-1k [35]. Meanwhile, other layers are randomly initialized to accelerate the training process and reduce over-fitting. As for the generative network, we use an SGD with a momentum of 0.95 to optimize it. The initial learning rate is set to 0.001 with a decay rate of 0.9 after 1000 iterations. And the discriminative network is optimized by an Adam, in which the initial learning rate is 0.02 with a decay rate of 0.9 after 3000 iterations. We train our networks with a batch size of 1 for 200 k iterations in total. Our method is implemented by TensorFlow on an NVIDIA GTX 1070 (8.00 GB).

4. Experiments

4.1. Dataset

There are many public datasets related to shadows, but some of them are prepared for shadow detection only (SBU [27], UCF [19]), and some of them only have shadowed and shadow-free pairs (SRD [8], UIUC [4], LRSS [3]). We take the widely used ISTD [9] dataset in shadow removal to train and test our model. ISTD [9] contains triplets of shadowed images, shadow-free images, and shadow masks. There are 1330 training triplets and 540 testing triplets in total, covering various shadow shapes for 135 different scenes. The input image size in our network is 240 × 320, and we augment the training dataset by flipping or cropping to avoid overfitting.

4.2. Evaluation Metrics

We follow the recent work to evaluate shadow removal results by Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [36]. We first calculate the MSE between the ground truth I_{gt} and the generated image $\bar{I}$ $(m \times n)$ as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I_{gt}(i,j) - \bar{I}(i,j)]^2 \tag{7}$$

Based on the MSE, $PSNR$ is defined as:

$$PSNR = 10 \cdot lg(\frac{MAX_{I_{gt}}^2}{MSE}) \tag{8}$$

where $MAX_{I_{gt}}^2$ is the max value of each pixel in the ground truth. A higher $PSNR$ means better image quality.

$SSIM$ [36] is an index to measure the similarity of the ground truth I_{gt} and the generated image $\bar{I}$ as below:

$$SSIM(I_{gt}, \bar{I}) = \frac{(2\mu_{I_{gt}}\mu_{\bar{I}} + C_1)(2\sigma_{I_{gt},\bar{I}} + C_2)}{(\mu_{I_{gt}}^2 + \mu_{\bar{I}}^2 + C_1)(\sigma_{I_{gt}}^2 + \sigma_{\bar{I}}^2 + C_2)} \tag{9}$$

where I_{gt} and $\bar{I}$ are the ground truth and the restored image, respectively, $\mu_{I_{gt}}$ is the average of I_{gt}, and $\mu_{\bar{I}}$ is the average of $\bar{I}$. Finally, $\sigma_{I_{gt}}^2$ is the variance of I_{gt}, $\sigma_{\bar{I}}^2$ is the variance of $\bar{I}$, $\sigma_{I_{gt},\bar{I}}$ is the covariance of I_{gt} and $\bar{I}$. We set $c_1 = (0.01 \times (2^8 - 1))^2, c_2 = (0.03 \times (2^8 - 1))^2$.

The value of $SSIM$ lies in the range [0, 1]. Please note that higher $SSIM$ is what we expected.

4.3. Results Comparison

In this section, we will compare our methods with other state-of-the-art approaches in quantity and quality. We also provide ablation studies to verify the effectiveness of each component in our framework.

4.3.1. Qualitative and Quantitative Evaluation

First of all, we compare our network with state-of-the-art shadow removal methods including DSC [7], Gong et al. [37], and ST-CGAN [9]. Compared with the traditional method [37], we can improve the PSNR and SSIM by 19.2% and 44.94%, respectively. These results demonstrate that deep learning has a significant impact on shadow removal. For the recent methods which use CNNs [7,9], our approach has a 3.24% increase in SSIM and 2.77% improvement in PSNR. This is mainly attributed to the fact that our method pays more attention to the structural characteristics of shadows, resulting in higher consistency between the restored image and the ground truth. Overall, our method performs better in PSNR and SSIM than existing methods. The detailed results are in Table 1.

Aside from the quantitative evaluation, we visualize the results of different methods to evaluate them qualitatively. As we can see in Figure 7, the result of Gong et al. [37] is the worst, indicating the traditional method cannot deal with complex conditions. DSC [7] and ST-CGAN [9] are better than the traditional method since their results are more natural, but we can still find some differences between the shadowed regions and the shadow-free regions. Besides the loss of image information, DSC [7] and ST-CGAN [9] pay less attention to the image details, which is common in existing methods because they use the same network to process both shadowed regions and shadow-free regions. This problem can be solved perfectly with our method. We use a dual-autoencoder to separately process the shadowed regions and the shadow-free regions, so the consistency of our results is more remarkable.

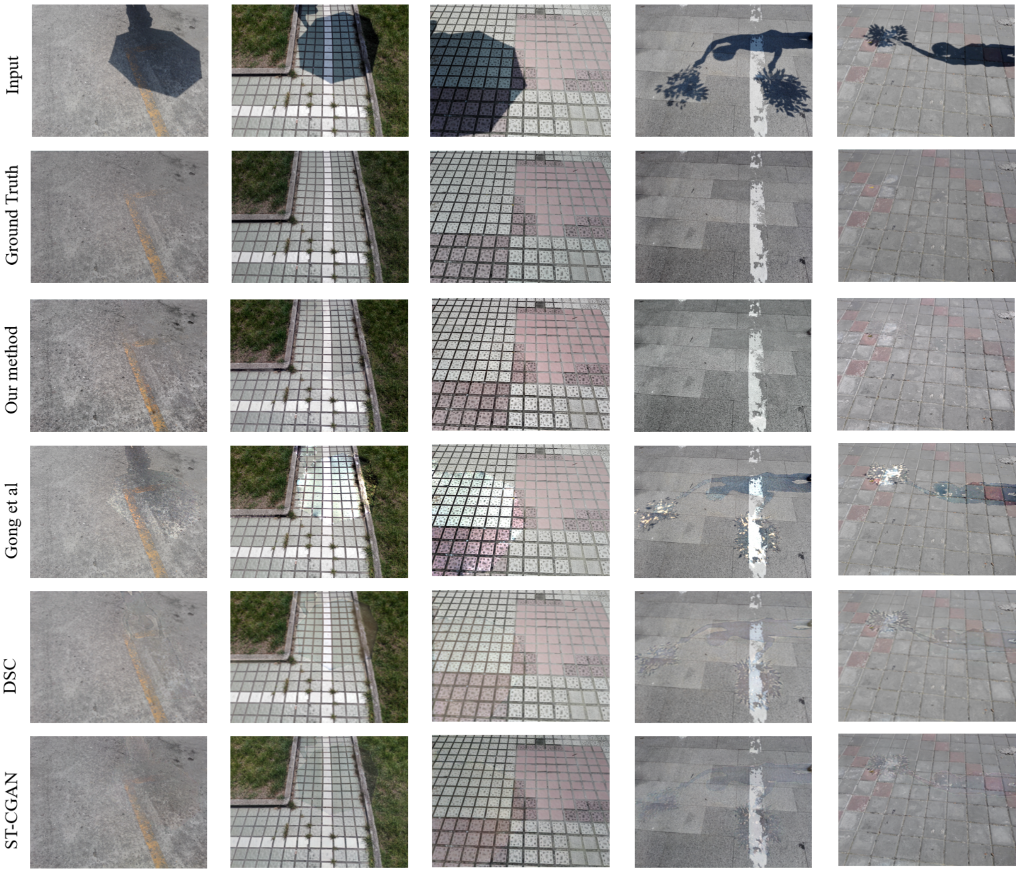

Figure 7. Visualization results of different methods. From top to bottom: shadowed image (input), the ground truth, our method, Gong et al. [37], DSC [7] and ST-CGAN [9]. Nearly all shadows are removed by our method. Furthermore, the color, brightness, and texture are more similar to the ground truth.

Table 1. Quantitative evaluation results between the existing methods and our work.

Methods	PSNR	SSIM
Gong et al. [37]	19.89	0.6103
DSC [7]	23.07	0.8639
ST-CGAN [9]	23.63	0.8568
Our Method	23.71	0.8846

4.3.2. Ablation Study

To prove each part of our network is indispensable, four variants of our method are designed to perform various ablation studies. Particularly, O is the contextual autoencoder along with one encoder-decoder net. O+A means the generator that contains one encoder-decoder net and the attention mechanism. T is the autoencoder containing two encoder-decoder nets. T+A is our complete architecture: attentive autoencoder using the attention mechanism with two encoder-decoder nets. As shown in Table 2, when using the attention mechanism, the PSNR and SSIM are obviously improved. Since the attention map can highlight shadows, it is beneficial for recovering. Furthermore, it is unreasonable to use the same network to process shadowed regions and shadow-free regions. Unlike existing methods, we apply the separate processing of shadowed areas and the shadow-free areas to keep the consistency.

Table 2. Quantitative evaluation results between each part and the whole network.

Methods	PSNR	SSIM
O	20.30	0.8174
O+A	21.01	0.8625
T	20.52	0.8570
T+A (Our Method)	23.71	0.8846

The visualization results are shown in Figure 8. It can be seen that the network detection result is more accurate after applying the attention mechanism. Moreover, the boundary is integrated with the surrounding very well. After adopting the two encoder-decoders, the color, texture, and illumination of the image are more natural. Our method can recover the shadow image to a clear image effectively. The results shown in Figure 8 demonstrate the essence of each part of our network module.

Like ST-CGAN [9] and DSC [7], our proposed method adopts a multi-branch structure. The difference is that our method detects shadows first, and then processes shadowed and shadow-free regions separately. In contrast, ST-CGAN treats shadow detection and shadow removal jointly. Similarly, the DSC method also uses multi-branch structure, the difference is that the method uses multi-branch structure to extract features at different scales, and then the DSC module to extract contextual features. The method can obtain good results in both shadow detections. ST-CGAN and DSC process both shadowed areas and shadow-free areas as a whole, making some restored shadow areas inconsistent with the surroundings. Different from them, our proposed method is more concerned with solving the inconsistency in the shadowed and shadow-free regions. The experimental results show that our proposed method can achieve better metrics and more natural results.

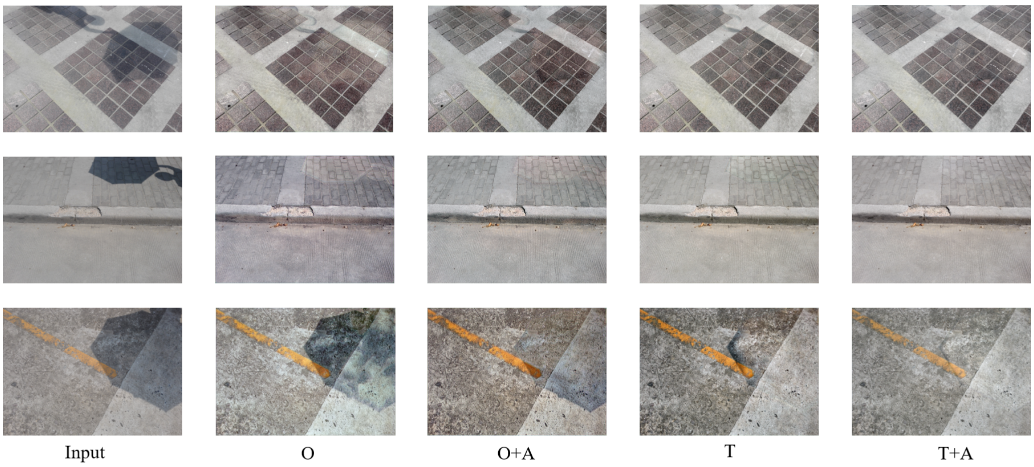

Figure 8. Visualization results of each variant of our method. From left to right: Input, O (one encoder-decoder without attention mechanism), T (two encoder-decoder nets without attention mechanism), O+A (one encoder decoder with attention mechanism) and T+A (the whole network, two encoder-decoder nets with attention mechanism).

4.4. Discussion

4.4.1. Application

Our method can be applied in many ways. For example, the detected shadows in the first stage can play an important role in various applications of visual scene understanding, such as scene geometry depiction, camera location, object relighting, and scene illumination inference. Meanwhile, the proposed shadow-removal model can be used to boost the performance of many computer vision tasks through data augmentation, such as image classification, object detection, and intrinsic image decomposition.

4.4.2. Limitations

On the other hand, there exist some limitations to the proposed method. For example, the two-stage training and dual branches may lead to higher computational costs, and thus cannot be deployed on lightweight and mobile devices.

5. Conclusions

In this work, we present a novel two-stage generative adversarial framework for shadow removal. In the shadow detection stage, we develop an attentive-recurrent network to generate the visual attention map. In the shadow removal stage, two auto-encoders are used to process the shadowed regions and shadow-free regions separately to keep consistency. The attention map is fed into the auto-encoders as a prior to make the generator focus on shadows. We compare our approach with the state-of-art methods to show its superiority both quantitatively and qualitatively. Furthermore, various ablation studies demonstrate that each part of our work is essential. In the future, we expect to collect more shadowed and shadow-free images with more complex scenes as well as improve the ability of the network to remove shadows in complex conditions.

Author Contributions: Conceptualization, H.W. and H.Z.; methodology, H.W. and H.Z.; software, H.Z.; validation, H.Z.; formal analysis, H.W. and H.Z.; data curation, H.W. and H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Bingtuan Science and Technology Program (grants no. 2019BC008 and no. 2022DB005).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were used in this study. The ISTD dataset can be found here: https://drive.google.com/file/d/1I0qw-65KBA6np8vIZzO6oeiOvcDBttAY/view (accessed on 20 December 2021).

Acknowledgments: We would like to thank Nanjing University of Science and Technology for providing the ISTD dataset. Moreover, we would like to thank the authors of these methods we compared, including DSC, ST-GAN, ISR. Our deepest gratitude goes to the reviewers and editors for their careful work and thoughtful suggestions that have helped improve this paper substantially.

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

References

1. Simonelli, A.; Bulo, S.R.; Porzi, L.; Antequera, M.L.; Kontschieder, P. Disentangling Monocular 3D Object Detection: From Single to Multi-Class Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1219–1231. [CrossRef] [PubMed]
2. Ong, J.; Vo, B.T.; Vo, B.N.; Kim, D.Y.; Nordholm, S. A Bayesian Filter for Multi-View 3D Multi-Object Tracking With Occlusion Handling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2246–2263. [CrossRef] [PubMed]
3. Gryka, M.; Terry, M.; Brostow, G.J. Learning to remove soft shadows. *ACM Trans. Graph. (TOG)* **2015**, *34*, 1–15. [CrossRef]
4. Guo, R.; Dai, Q.; Hoiem, D. Paired regions for shadow detection and removal. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 2956–2967. [CrossRef] [PubMed]
5. Khan, S.H.; Bennamoun, M.; Sohel, F.; Togneri, R. Automatic shadow detection and removal from a single image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 431–446. [CrossRef] [PubMed]
6. Vicente, T.F.Y.; Hoai, M.; Samaras, D. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 682–695. [CrossRef] [PubMed]
7. Hu, X.; Zhu, L.; Fu, C.-W.; Qin, J.; Heng, P.-A. Direction-aware spatial context features for shadow detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 7454–7462.
8. Inoue, N.; Yamasaki, T. Deshadownet: Learning from Synthetic Shadows for Shadow Detection and Removal. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4187–4197. [CrossRef]
9. Wang, J.; Li, X.; Yang, J. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 1788–1797.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
11. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–10 December 2015; pp. 802–810.
12. Wang, Y.; Tao, X.; Qi, X.; Shen, X.; Jia, J. Image inpainting via generative multi-column convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 331–340.
13. Finlayson, G.D.; Hordley, S.D.; Lu, C.; Drew, M.S. On the removal of shadows from images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 59–68. [CrossRef] [PubMed]
14. Finlayson, G.D.; Drew, M.S.; Lu, C. Entropy minimization for shadow removal. *Int. J. Comput. Vis.* **2009**, *85*, 35–57. [CrossRef]
15. Tian, J.; Qi, X.; Qu, L.; Tang, Y. New spectrum ratio properties and features for shadow detection. *Pattern Recognit.* **2016**, *51*, 85–96. [CrossRef]
16. Lalonde, J.-F.; Efros, A.A.; Narasimhan, S.G. Detecting ground shadows in outdoor consumer photographs. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 322–335.
17. Guo, R.; Dai, Q.; Hoiem, D. Single-image shadow detection and removal using paired regions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 21–23 June 2011; pp. 2033–2040.
18. Vicente, Y.; Tomas, F.; Hoai, M.; Samaras, D. Leave-one-out kernel optimization for shadow detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 3388–3396.
19. Zhu, J.; Samuel, K.G.; Masood, S.Z.; Tappen, M.F. Learning to recognize shadows in monochromatic natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 223–230.
20. Huang, X.; Hua, G.; Tumblin, J.; Williams, L. What characterizes a shadow boundary under the sun and sky? In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 898–905.
21. Le, H.; Vicente, Y.; Tomas, F.; Nguyen, V.; Hoai, M.; Samaras, D. A+d net: Training a shadow detector with adversarial shadow attenuation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 662–678.
22. Zheng, Q.; Qiao, X.; Cao, Y.; Lau, R. Distraction-Aware Shadow Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5162–5171.

23. Zhu, L.; Deng, Z.; Hu, X.; Fu, C.-W.; Xu, X.; Qin, J.; Heng, P.-A. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 121–136.
24. Nguyen, V.; Vicente, T.F.Y.; Zhao, M.; Hoai, M.; Samaras, D. Shadow detection with conditional generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4520–4528.
25. Wang, Y.; Zhao, X.; Li, Y.; Hu, X.; Huang, K.; Cripac, N. Densely cascaded shadow detection network via deeply supervised parallel fusion. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 1007–1013.
26. Khan, S.H.; Bennamoun, M.; Sohel, F.; Togneri, R. Automatic feature learning for robust shadow detection. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1939–1946.
27. Vicente, T.F.Y.; Hou, L.; Yu, C.-P.; Hoai, M.; Samaras, D. Large-scale training of shadow detectors with noisily-annotated shadow examples. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 816–832.
28. Hu, X.; Zhu, L.; Fu, C.-W.; Qin, J.; Heng, P.-A. Direction-Aware Spatial Context Features for Shadow Detection and Removal. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2795–2808. [CrossRef]
29. Sahoo, S.; Nanda, P.K. Adaptive Feature Fusion and Spatio-Temporal Background Modeling in KDE Framework for Object Detection and Shadow Removal. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1103–1118. [CrossRef]
30. Liu, Z.; Yin, H.; Mi, Y.; Pu, M.; Wang, S. Shadow Removal by a Lightness-Guided Network With Training on Unpaired Data. *IEEE Trans. Image Process.* **2021**, *30*, 1853–1865. [CrossRef] [PubMed]
31. Fu, L.; Zhou, C.; Guo, Q.; Xu, F.; Yu, H.; Feng, W.; Liu, Y.; Wang, S. Auto-Exposure Fusion for Single-Image Shadow Removal. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 10571–10580.
32. Chen, Z.; Long, C.; Zhang, L.; Xiao, C. CANet: A Context-Aware Network for Shadow Removal. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4723–4732.
33. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
35. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
36. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
37. Gong, H.; Cosker, D. Interactive shadow removal and ground truth for vari-able scene categories. In Proceedings of the BMVC, Nottingham, UK, 1–5 September 2014; pp. 1–11.