Towards Ghost-Free Shadow Removal via Dual Hierarchical Aggregation Network and Shadow Matting GAN

Xiaodong Cun,[1] **Chi-Man Pun,**[1*] **Cheng Shi**[1,2]

[1]Department of Computer and Information Science, University of Macau, Macau, China
[2]School of Computer Science, Xi'an University of Technology, Xi'an, China
{yb87432, cmpun}@umac.mo, chengc_s@163.com

Abstract

Shadow removal is an essential task for scene understanding. Many studies consider only matching the image contents, which often causes two types of ghosts: color in-consistencies in shadow regions or artifacts on shadow boundaries (as shown in Figure. 1). In this paper, we tackle these issues in two ways. First, to carefully learn the border artifacts-free image, we propose a novel network structure named the dual hierarchically aggregation network (DHAN). It contains a series of growth dilated convolutions as the backbone without any down-samplings, and we hierarchically aggregate multi-context features for attention and prediction, respectively. Second, we argue that training on a limited dataset restricts the textural understanding of the network, which leads to the shadow region color in-consistencies. Currently, the largest dataset contains 2k+ shadow/shadow-free image pairs. However, it has only 0.1k+ unique scenes since many samples share exactly the same background with different shadow positions. Thus, we design a shadow matting generative adversarial network (SMGAN) to synthesize realistic shadow mattings from a given shadow mask and shadow-free image. With the help of novel masks or scenes, we enhance the current datasets using synthesized shadow images. Experiments show that our DHAN can erase the shadows and produce high-quality ghost-free images. After training on the synthesized and real datasets, our network outperforms other state-of-the-art methods by a large margin. The code is available: http://github.com/vinthony/ghost-free-shadow-removal/

Introduction

Shadows are a common phenomenon in nature. Cast shadows, which often appear in the opposite direction of a light source, form when the light from an illuminant is blocked by an object. Removing these shadows is essential because shadows may degrade the performance of many computer vision tasks, such as object detection and tracking (Mikic et al. 2000; Khan et al. 2015). Early approaches try to automatically detect shadow regions (Guo, Dai, and Hoiem 2013; Khan et al. 2015) or to do so with user input (Gong and Cosker ; Yang, Tan, and Ahuja 2012) first, after which they
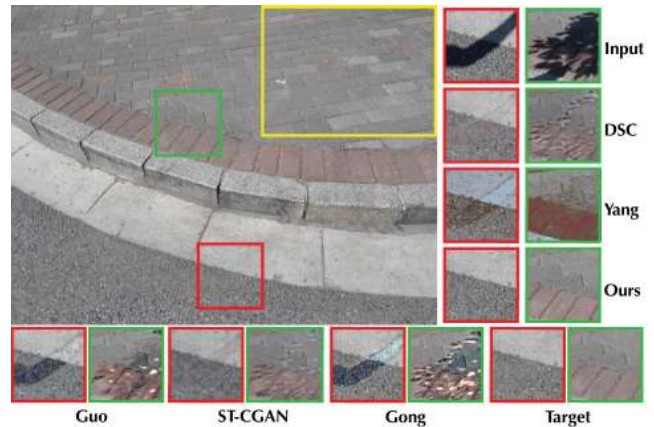
Figure 1: We plot a result of our model with the input shown in the yellow square. From the two zoomed regions, it can be seen that our method successfully removes the shadow and reduces the ghost.

transform the shadows to match the background. For deep learning-based approaches, Qu et al. (2017) train a neural network considering multi-context features (global semantics and low-level details). Then, Wang, Li, and Yang (2018) conduct joint shadow removal/detection using a stacked generative adversarial network. Hu et al. (2018) retrain their shadow detection network to directly train shadow removal network. More recently, Hu et al. (2019) train a network using an unpaired dataset without paired supervisions.

However, current deep learning-based techniques have two major drawbacks. First, we argue that previous network structures are not carefully designed for shadow removal. While the goal of shadow removal is restoring the color from the shadow image, the shadow border also plays a vital role in the visual quality. For example, DeShadowNet (Qu et al. 2017) learns only the shadow matting from multi-context features other than the full image using the pre-trained VGG19, and the artifacts will still show on the border. DSC (Hu et al. 2018) learns using a direction-aware attention model, which well captures the directional details well and also preserves more shadow boundaries. Thus, a

carefully designed model is criterial. In addition, the community lacks a high-quality image pairs (shadow/shadow-free) for shadow removal and joint learning on shadow removal and detection (shadow/shadow-free/mask). Since the largest available dataset contains only 0.1k+ unique scenes, and the performance of unsupervised learning is limited. We argue that such limited scenes will hugely influence the color consistency between the shadow and shadow-free regions because training on small dataset can not understand the scenes well.

In this work, we design a novel network structure to erase shadows and produce high-quality shadow-free images according to the following observations: *the shadow and shadow-free images share the same semantic information and shadow removal specifically and only needs to learn the shadow*. Since the semantic knowledge is the same and the network needs to focus on low-level features, different from others, an image processing network (Chen, Xu, and Koltun 2017) serves as a useful starting point for our method. In detail, this network uses several dilated convolutions (Yu and Koltun 2015) linearly with a growing dilated rate to capture multi-contexts and preserve useful low-level features. Based on this backbone, we hierarchically aggregate dilated convolutions to achieve better feature reuse (Yu et al. 2017). Then, to specifically learn the shadow regions, we design the hierarchical aggregation attention model with the help of multi-contexts and the attention loss from the shadow mask. Non-linear feature aggregation helps our attention to gain the knowledge from multiple previous layers, and dilated convolutions preserve the details well.

In addition, we get our inspiration from the nature in that a cast shadow comes from the opposite direction of the light source. When an object moves, a single shadow-free image can produce countless shadow images. Thus, if we train the network to synthesize shadow images from the corresponding masks and shadow-free images, the network will also produce the novel shadows according to the novel mask. It is much easier to collect the required data for shadow synthesis than for triples of paired shadow/shadow-free/mask. Thus, we design a novel shadow synthesis network using GAN (Goodfellow et al. 2014) to enlarge the current datasets. Instead of learning the shadow image directly, our network learns the shadow matting first and produces the high-quality shadow from the original shadow-free image and matting. Although these images are synthesized using neural network, they still have a similar distributions to the real natural scene compared with those from computer games (Sidorov 2019). In our shadow synthesis, we assume the scenes contain only cast shadows with the objects outside the scene. Experiments shows that our setting also works well in other situations.

We summarize our major contributions as follows.

- We design a joint learning framework, named Dual Hierarchical Aggregation Network (DHAN), for shadow removal with the help of shadow detection. The proposed network hierarchically aggregates the dilated multi-contexts features and attentions, respectively.

- We design and train a Shadow Matting GAN on current paired shadow datasets. Then, the novel paired samples are obtained from the well-trained network with novel masks or scenes. Finally, we train shadow-related tasks using augmented datasets.

- Experiments on real and synthetic data indicate that our algorithm outperforms other state-of-the-art methods, especially on the visual quality.

Related Work

Shadow Removal Early works often erase shadows via user interaction or hand-crafted features (Guo, Dai, and Hoiem 2013; Khan et al. 2015). Inspired by the convolution neural network-based method, Qu et al. (2017) propose a multi-stream neural network for removing the shadows according to global-and-local contexts, and Hu et al. (2018) use a fully convolutional network structure but with direction-aware attention. For learning shadow removal and detection jointly, Wang, Li, and Yang (2018) proposed a stacked GAN where the learned shadow mask serves as the attention for the removal. Ding et al. (2019) extend this method by strong network and learning recurrently.

Shadow Synthesis For it is hard to collect larger shadow removal datasets, Gryka, Terry, and Brostow (2015) generate the shadow images from graphics tools and Sidorov (2019) generate the shadow dataset by controlling the lights in the video game. However, scenes or shadows that are rendered by computer graphics are significantly different from natural scenes. (Le et al. 2018; Le and Samaras 2019) propose a two-stage network for generating the adversarial shadow images before shadow detection, and they also use this idea to remove shadows by enhancing the shadow strength. Focusing on the limitation of a paired dataset, Hu et al. (2019) train a neural network to remove shadows from unpaired shadow/shadow-free images. It also contains an auxiliary network for shadow synthesis. However, their synthesized shadows are not learned using the paired dataset and their shadow generator is used in cycle-consistency loss only.

Shadow Detection Some traditional methods use physical models of illumination and color for shadow detection (Salvador, Cavallaro, and Ebrahimi 2004; Panagopoulos et al. 2011). Others describe image regions by hand-crafted features, such as color (Lalonde, Efros, and Narasimhan 2010), texture (Vicente, Hoai, and Samaras 2017), then recognize these regions by classifier. However, these approaches are not well suited for various natural scenes. Current state-of-the-art shadow detection methods are based on the power of convolution neural network, such as learning shadow edges (Khan et al. 2015; Shen, Wee Chua, and Leman 2015) and designing attention modules by considering the direction (Hu et al. 2018), recurrent attention (Zhu et al. 2018) and distraction (Zheng et al. 2019).

Proposed Methods

Dual Hierarchical Aggregation Network

Given a shadow image I_{shadow}, our method learns the shadow-free image I_{free} using the proposed network with
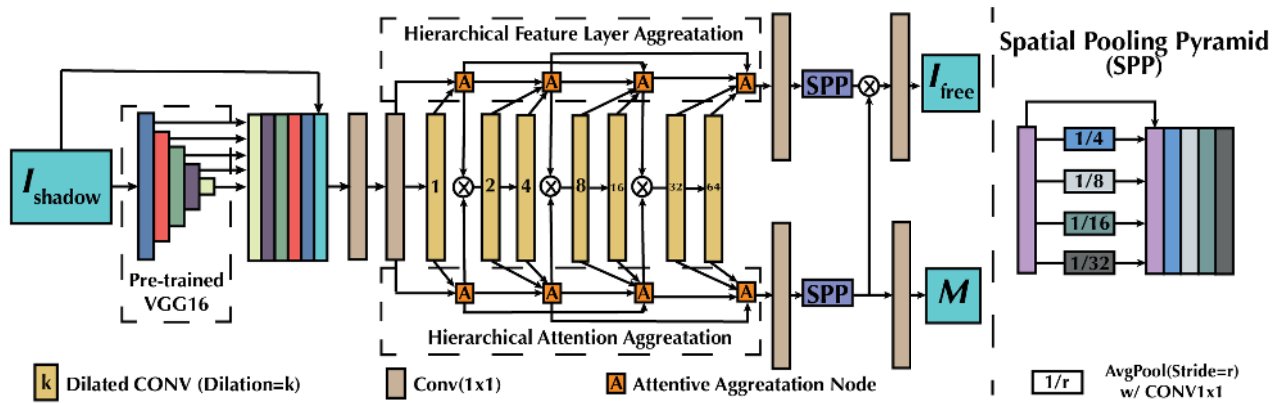
Figure 2: The network structure of the proposed Dual Hierarchical Aggregation Network.

the help of the attention loss (shadow mask M) in a single forward process. As previously discussed, our backbone structure is based on the context aggregation network (CAN) (Chen, Xu, and Koltun 2017; Zhang, Ng, and Chen 2018) for image processing. Specifically, we extract the hyper-column features from the pre-trained VGG19 (CONV1_2, CONV2_2, CONV3_2, CONV4_2 and CONV5_2) and the original RGB image serves as an augmented network input. Then, we encode these multi-context features using several dilated convolutions with a growing dilate rate. Each convolutional block follows an Instance Normalization and LeakyReLU layer. Previous studies show that this network structure works well in image processing tasks, such as Nonlocal de-hazing and L_0 smoothing since there is no down-sampling or stride convolution in the context encoding, the low-level details can be well preserved. Although shadow removal is similar to low-level task, it specifically needs to learn the shadow regions. Possible solutions are to learn the particular region using partial convolution (Liu et al. 2018) or gated convolution (Yu et al. 2019). However, these attention mechanisms are designed for image in-painting. Thus, they need to fill the region using global semantic contexts and work in a Unet-like structure.

Thus, to specifically learn the shadows and preserve low-level details, we design **dual hierarchical aggregations** for spatial attentions and mixed layer features, respectively. In detail, as shown in Figure. 2, we build the attention module by aggregating the features from multiple previous layers in a hierarchical layer aggregation style (Yu et al. 2017). Our
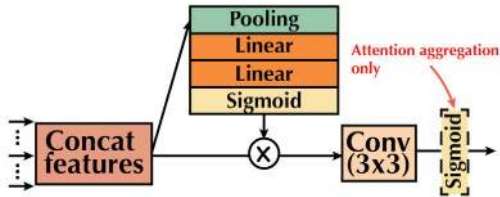

Figure 3: Attentive Aggregation nodes for feature aggregation and attention aggregation.

n layer dual hierarchical aggregation network $DHAN_n = T_n \times AT_n$ merges the features in a tree-like structure for richer context attention. In it, T_n and AT_n can be defined as:

$$T_n = N(R_{n-1}^n(x), ..., R_1^n(x), L_1^n(x), L_2^n(x))$$
$$AT_n = AN(AR_{n-1}^n(x), ..., AR_1^n(x), L_1^n(x), L_2^n(x)) \quad (1)$$

where N and AN are the aggregation node and attention aggregation node, respectively. $N(\cdot)$ means the feature concatenation. R, AR and feature layer L are defined as:

$$L_1^n(x) = DB_{k=2^{n-1}}(R_1^n(x) \times AR_1^n(x))$$
$$L_2^n(x) = DB_{k=2^n}(L_1^n(x))$$

$$R_m^n(x) = \begin{cases} T_m & \text{if m = n-1} \\ T_m(R_{m+1}^n(x)) & \text{otherwise,} \end{cases} \quad (2)$$

$$AR_m^n(x) = \begin{cases} AT_m & \text{if m = n-1} \\ AT_m(AR_{m+1}^n(x)) & \text{otherwise,} \end{cases}$$

where $DB_{k=x}$ represents the x-dilated convolution. Thus, each aggregated feature is spatially reweighed by the multi-context attention. In addition, we also use the ground truth shadow mask as the supervision for the attention.

In each aggregation node (as shown in Figure. 3), we use a squeeze-and-excitation block (Hu, Shen, and Sun 2018) to re-weight the importance of each feature channel. Then, a 3×3 convolution is added to squeeze the features and match the original channels. Notice that, the $Sigmoid(\cdot)$ layer is added only at the end of the attention aggregation nodes (AR). Finally, a spatial pooling pyramid (as shown in Figure. 2) (He et al. 2015) is added at the end of the last aggregation block for multi-context features remixing.

Loss Functions By considering the semantic measures and low-level details in multiple contexts, we train our network using the multi layer perception loss (Zhang, Ng, and Chen 2018; Johnson, Alahi, and Fei-Fei 2016) L_Φ in the pixel and feature domain as follows:

$$L_\Phi = \sum_{k=0}^{5} \lambda_l ||\Phi_k(I'_{free})) - \Phi_k(I_{free})||_1 \quad (3)$$

where Φ is the VGG16 that is pre-trained on ImageNet. We estimate the differences from the CONVk_2(k = 1...5) and
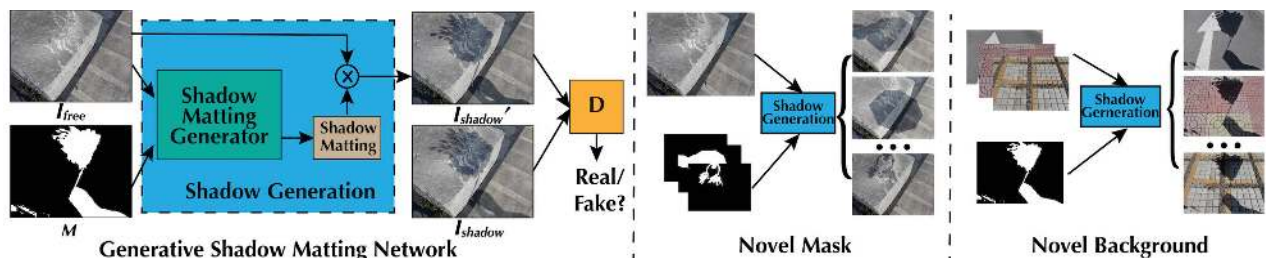
Figure 4: The network structure of Shadow Matting GAN.

the original images ($k = 0$ in Equation. 3) between the natural shadow-free image I_{free} and the generated I'_{free} using the L_1 loss with weight λ_l. For the attention loss, we use the binary cross entropy (BCE) loss between the predicted mask M' and the ground truth mask M as follows:

$$L_m = \frac{1}{N} \sum -[Mlog(M') + (1 - M)log(1 - M')] \quad (4)$$

Where N is the total number of pixels. In addition, the convolution neural network might also preserve the shadow border in the images, which will also degrades the shadow removal quality. Thus, we use the adversarial loss as (Isola et al. 2017) in our approach. By considering our DHAN as the generator, the discriminator D is constructed with several convolution blocks. We simply discriminate the patches in the real shadow-free image I_{free} and the predicted I'_{free} given the shadow image I_{shadow}. The discriminator is optimized to identify the generated image that pushes the generator towards realistic shadow-free images. The discriminator loss is as follows:

$$L_{cGAN} = \sum [log(D(I_{shadow}, G(I_{shadow})) \\ -log(D(I_{shadow}, I_{free}))] \quad (5)$$

Overall ,the final objective of our method G^* is as follows: $\arg\min_G \arg\max_D L_{cGAN} + \lambda L_\Phi + \alpha L_m$. We set α equals to 100 and β equals to 20 empirically.

SM-GAN: Shadow Matting GAN

As discussed in the introduction, one of the key issues in current shadow removal methods is the limitation of the paired dataset. However, with respect to the various types of shadow masks and scenes, it is hard to collect a large dataset with paired shadow/shadow-free images. The state-of-the-art SRD dataset (Qu et al. 2017) and ISTD (Wang, Li, and Yang 2018) contains only 100+ different scenes and 10+ unique shadow types. In addition, the environmental light changes so fast that it is impossible to take pairs of shadow/shadow-free images under the same light conditions, even in a flash. The uncertainty of the light transforming between shadow/shadow-free images influences the accuracy of the network. For these reasons, we argue that the application range of current learning-based shadow removal algorithms is limited and training using limited scenes will cause color inconsistencies between shadow and shadow-free regions. However, it is far easier to collect the natural shadow-free images and random masks than to collect

the corresponding shadow/shadow-free/shadow-mask sets. Thus, we synthesize the shadow image using the random shadow mask and the shadow-free image via the GAN to improve the accuracy of shadow networks and extend their capabilities to larger natural domains.

As shown in Figure.4, using the knowledge of the shadow image I_{shadow}, the shadow mask M and the corresponding shadow-free image I_{free} in the current dataset, we train shadow synthesis as a paired task (Isola et al. 2017) using the GAN. Specifically, we train a generator to synthesize the shadow image I'_{shadow} from the corresponding shadow-free image I_{free} and mask M. Then, the discriminator is optimized to identify whether the synthesized shadow image I'_{shadow} is real or not. Since there is a corresponding ground truth image I_{shadow} in the dataset, apart from adversarial loss from the discriminator, we restrict the similarity of the I'_{shadow} and I_{shadow} in a supervised manner. Instead of generating the shadow image directly, we learn the shadow matting $I_{matting}$ following the former observation (Qu et al. 2017): $I_{shadow} = I_{matting} \times I_{free}$. Thus, our shadow synthesis network can be written as follows:

$$I'_{shadow} = G(I_{free}, M) \times I_{free} \quad (6)$$

where $G(\cdot)$ is the generator of our shadow matting GAN. The supervised losses in the shadow matting generator are similar to our shadow removal: We use the multi-scale perceptual loss to measure the prediction quality and adversarial loss to guarantee the realism of the output. Then, we update the parameters in the discriminator network D to recognize whether the generated shadow image is real or not. The generator of our SM-GAN has a similar structure with the CycleGAN (Zhu et al. 2017). For simplicity, the discriminator is constructed with 5 convolution layers with ReLU non-linear activation and Batch Normalization layer, which is similar to our shadow removal network.

After training, the generator in SMGAN can be easily used to sample novel domain examples, such as the novel background and novel shadow mask (as shown in the novel mask/background of Figure.4). Here, we use the shadow-free images in USR (Hu et al. 2019) and shadow masks from ISTD training set to synthesize the shadow images for data augmentation. For each shadow-free image in the USR dataset, we randomly choose 3 images from the training mask of the ISTD dataset to generate shadow. The experiments show that this kind of setting is sufficient for the shadow removal and detection tasks.
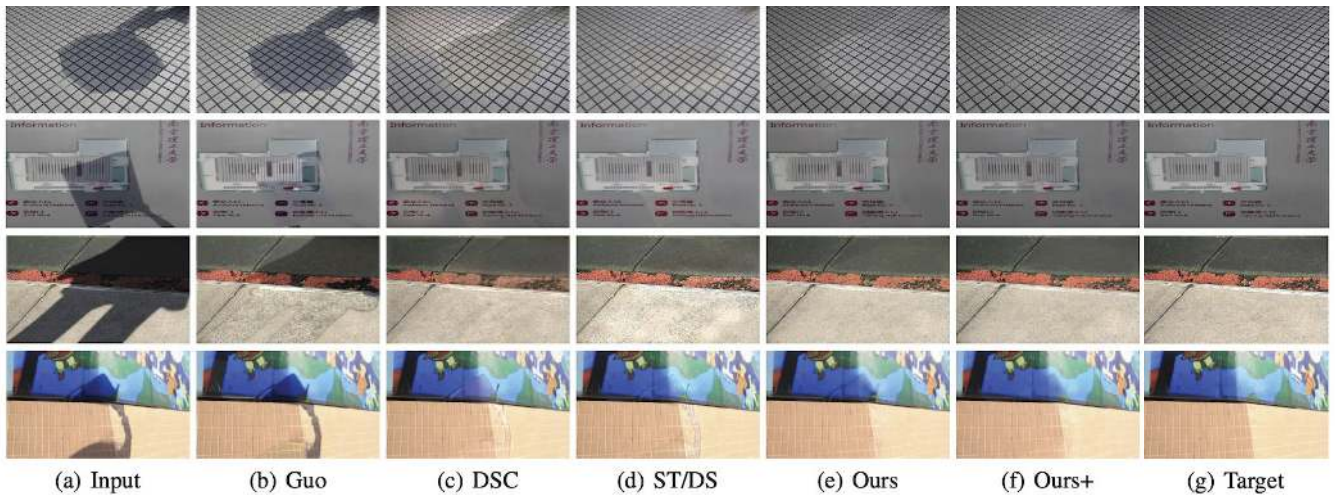
Figure 5: Shadow Removal Comparison. The top and bottom two samples are from ISTD and SRD dataset, respectively. In (d), the top and bottom two results are from ST-CGAN (ST) and DeShadowNet (DS), respectively. Best view with zoom-in.

Table 1: Comparison of Shadow removal results. The subscripts represent the years of the compared methods.

Method	S	NS	RMSE	SSIM-S	PSNR-S	S	NS	RMSE	SSIM-S	PSNR-S
	ISTD dataset					SRD dataset				
Original	32.67	6.83	10.97	98.19	32.87	46.87	5.60	14.28	88.54	19.35
Yang$_{12'}$	19.82	14.83	15.63	93.25	28.01	23.43	22.26	22.57	-	-
Guo$_{13'}$	18.95	7.46	9.30	96.04	27.49	29.89	6.47	12.60	91.46	23.69
Gong$_{14'}$	14.98	7.29	8.53	96.97	29.79	19.58	4.92	8.73	-	-
DeShadowNet$_{17'}$	-	-	-	-	-	10.81	4.85	6.10	94.68	31.97
ST-CGAN$_{18'}$	10.33	6.93	7.47	94.93	32.23	-	-	-	-	-
DSC$_{18'}$	9.48	6.14	6.67	96.66	33.45	10.89	4.99	6.23	93.75	31.69
Ours (DHAN)	8.14	6.04	6.37	98.29	34.50	8.94	4.80	5.67	95.29	33.36
Ours (DHAN)+DA	**7.52**	**5.43**	**5.76**	**98.36**	**34.98**	**8.39**	**4.67**	**5.46**	**95.31**	**33.72**

Experiments

In this section, first, we conduct the shadow removal experiments using standard benchmarks, the ISTD (Wang, Li, and Yang 2018) and the SRD (Qu et al. 2017). The SRD contains 2k+ and 408 images for training and testing, respectively. In ISTD, there is 1.3k and 540 images for training and testing, respectively. On both datasets[1], we jointly learn the shadow removal and attention (shadow mask) using the proposed DHAN. In addition, we retrain these tasks on the dataset containing both real and synthesized shadow images. Then, we evaluate the effectiveness of our synthesized shadows for shadow detection on SBU (Vicente et al. 2016) dataset, which contains 4k+ and 0.6k+ images for training and testing, respectively. Finally, we conduct the ablution study on the network structure, the quality of shadow synthesis and the network capabilities.

Implementation Detail We train all our models using TensorFlow framework and optimize the network on Adam optimizer (Kingma and Ba 2014) with a fixed learning rates

[1]Since there is no ground truth shadow mask in the SRD, we calculate the shadow matting using the shadow/shadow-free pairs and generate the binary mask via thresh-holding.

of $2*10^{-4}$ and 10^{-4} in the generator and the discriminator, respectively. In training, we randomly resize the original image from 256p to 480p, thereby retaining the original aspect radio rather than cropping the image. We set the batch size to 1 and train the network in over 100 epochs for shadow synthesis and 150 epochs for shadow removal (60 epochs for the augmented datasets). For testing, we evaluate all the images in their original size. It takes 0.2s for our network to generate a shadow-free image with a resolution of 640x480 on a single NVIDIA TITAN V GPU. On shadow removal, we follow the previous works and evaluate the RMSE on the LAB color space in the shadow (S) and non-shadow (NS) regions. In addition, we also report the PSNR and SSIM for the shadow region (SSIM-S and PSNR-S, respectively) to evaluate whether it can produce high-quality shadow removal results. To evaluate on shadow detection, we use the Balance Error Rate (BER) for the shadow (S) and non-shadow region (NS) as in previous works. The BER is defined as follows: $BER = 1 - \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$, where TP, TN, FP and FN are the total numbers of the true positives, true negatives, false positives and false negatives, respectively.
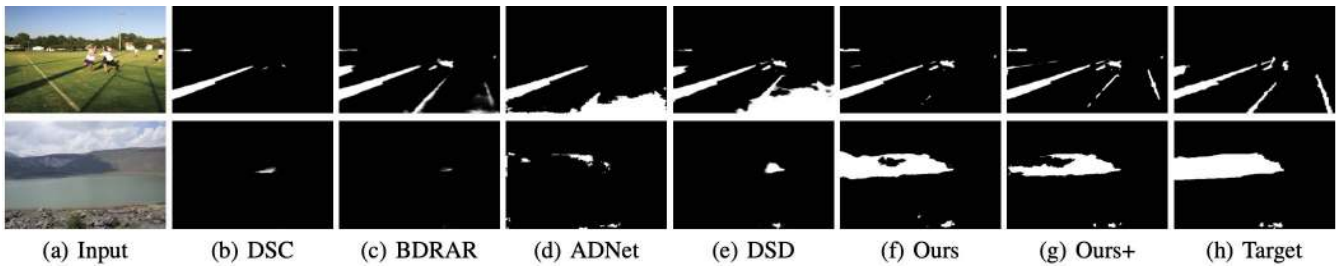
Figure 6: Comparison the shadow detection results with state-of-the-art methods on the SBU dataset.

Comparison with the State-of-the-art Methods

Shadow Removal   We compare ours results with those of the state-of-the-art shadow removal methods, including ST-CGAN (Wang, Li, and Yang 2018) that jointly conducts shadow removal and detection, DSC (Hu et al. 2018) that removes shadows using a direction-aware attention mechanism, and DeShadowNet (Qu et al. 2017) conducts shadow removal using multi-context features and the shadow matting loss. Apart from deep learning-based methods, we also compare our algorithm with some open-sourced traditional methods (Guo, Dai, and Hoiem 2013; Gong and Cosker ; Yang, Tan, and Ahuja 2012). Note that, in order to conduct a fair comparison, all the results and figures are provided by the authors or taken from the original papers. Since the training codes of ST-CGAN and DeShadowNet are not publicly available, we evaluate their performances only on their own dataset. As shown in Figure. 5, our method provides much better visual quality than the others on both two standard datasets. In these samples, previous works produce many ghosts because of the remaining shadow boundaries or color inconsistencies. For example, DeShadowNet learns the shadow matting only where the restored color might not match the background. In addition, the DSC and De-ShadowNet will cause sharp shadow boundaries since their network structures need to down-sample and up-sample in the multi-contexts. Meanwhile, due to the limited dataset, they all lack better scene understandings. However, the proposed method (Ours) has minimal artifacts in the shadow boundaries and the synthesized dataset used as augmentation (Ours+DA) provides much better color consistency. For the numerical evaluation in Table. 1, all the methods can successfully learn the shadow-free region with the help of network; however, it is clear that our DHAN outperforms the others, especially on the shadow regions, because of the effectiveness of our attention model and attention loss. Additionally, training on the augmented synthesized shadow dataset helps a lot in shadow and non-shadow regions since it successfully enlarges the knowledge of current dataset. Note that, on SRD dataset, even though ISTD dataset contains larger color differences between the training pairs in the shadow-free region and our network try to fit these mapping in shadow synthesis, the synthesized shadows also help the scenes understanding on SRD dataset.

Shadow Synthesis for Detection   Our synthesized dataset also benefits to shadow detection. In detail, we modify our

Table 2: Comparison on shadow detection.

Method	backbone	S↓	NS↓	BER↓
scGAN[17']	-	8.39	9.69	9.04
ST-CGAN[18']	-	3.75	12.53	8.14
DSC[18']	VGG19	9.76	**1.42**	5.59
ADNet[18']	Adversarial	4.45	6.30	5.37
Ours	VGG19	2.80	6.32	4.56
Ours+DA	VGG19	**2.71**	5.87	4.29
BDRAR[18']	**ResNeXt101**	3.40	3.89	3.64
DSDNet[19']	**ResNeXt101**	3.33	3.58	**3.45**

network structure to match the detection by removing the output of attention loss. Thus, we only use the mask as output for back-propagation. We compare our results with state-of-the-art learning-based shadow detection methods, such as attention-based methods DSC (Hu et al. 2018), BDRAR (Zhu et al. 2018), DSD (Zheng et al. 2019), joint learning method ST-CGAN (Wang, Li, and Yang 2018) and adversarial shadow-based methods ADNet (Le et al. 2018). As shown in Table. 2, although our network is not particularly designed for shadow detection, it still gets better results than those other methods under the same backbone. Our network also obtains comparable results with the state-of-the-art methods using the powerful pre-trained ResNext101 model and multi-contexts supervisions. The shadows in the SBU dataset are much more challenging and complex than simple cast shadows. However, our synthesized shadow still helps, as in Figure. 6 and Table. 2.

Ablation Studies

Network Structure   We evaluate the contributions of our network structure on the SRD dataset. As shown in Table. 3 and Figure. 7, our baseline is the *context aggregation network* (CAN) (Zhang, Ng, and Chen 2018) that is designed for image processing and reflection removal. It has a poor performance on the shadow removal task both on shadow and non-shadow regions. However, our feature layer aggregation (+ LA) technique greatly improves the quality of the shadow regions. This is because the hierarchical mixed features will force the network to better understand multiple contexts. Then, when training using the dual aggregation network with attention aggregation (+DLA), our method gains far better results with the supervision of the shadow mask. In addition, we find that the local gated con-

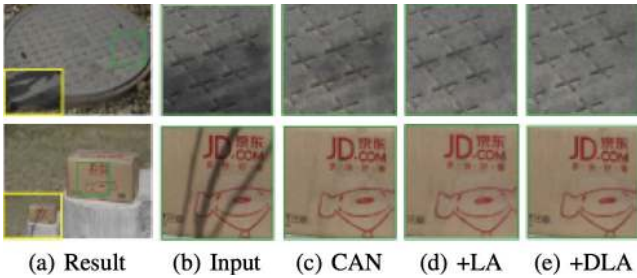(a) Result   (b) Input   (c) CAN   (d) +LA   (e) +DLA

Figure 7: Ablation study on our network structure. We plot the results of our full methods with the input in yellow region. The baseline network CAN can remove the shadow but still have color in-consistency, and our feature LA(+LA) and dual LA(+DLA) can further remove the artifacts.

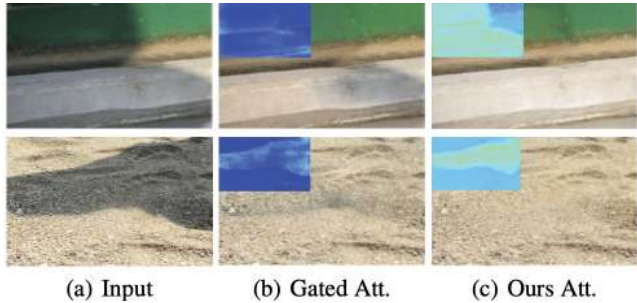(a) Input   (b) Gated Att.   (c) Ours Att.

Figure 8: Comparison with the attention from gated convolution. We show the results and corresponding attention map. It is obviously that gated attention learns the feature locally while our attention predict the shadow region successfully.

volution (Yu et al. 2019) is quantitatively less effective on our task. Here, we plot an example in Figure. 8 and compare our attention with the results on local gated convolution. We visualize the mean responses on the last attention features after $Sigmoid$. The figure shows that when the shadow region is large enough, the local gated convolution only gains the knowledge from its sibling convolution, causing the failure to capture the whole image context.

Comparison on Shadow Synthesis Since there are few deep learning-based methods for shadow synthesis particularly, we compare the auxiliary network for unpaired shadow removal in the Mask-ShadowGAN (Hu et al. 2019) and the ShadowGAN (Zhang, Liang, and Wang 2019), which synthesizes shadows using the Pix2Pix (Isola et al. 2017). Note that, the shadow generator in Mask-ShadowGAN is just an auxiliary helper for unpaired shadow removal, other than synthesizing the paired dataset for supervised learning as

Table 3: Ablation Study on SRD dataset.

Methods	S↓	NS↓	RMSE↓
CAN	10.05	5.11	6.14
+ LA	9.31	5.10	5.98
+ DLA	**8.93**	**4.80**	**5.67**

Table 4: Comparison on the quality of shadow synthesis.

Method	LPIPS ↓	SSIM ↑	PSNR ↑
Masked-ShadowGAN	25.30	80.18	23.18
Pix2Pix	11.51	88.82	26.01
Ours	**5.38**	**91.67**	**26.31**

(a) Input   (b) ISTD   (c) ISTD+S

Figure 9: Training using our synthesized dataset (ISTD+S) helps the network works better on novel domains.

data augmentation. The ShadowGAN trains using only the low-resolution images from the computer rendered dataset. For comparison, we train our shadow synthesis network and others on the ISTD training set and test by synthesizing the shadow images using the ISTD test set. We evaluate the results using the SSIM and PSNR. In addition, since perceptual quality is also important for image generation, we evaluate the generated shadow using the image perceptual quality metrics LPIPS (Zhang et al. 2018). As shown in Table. 4, our method outperforms other methods in all the metrics.

Extend to Novel Domain Training using our synthesized dataset enlarges the capabilities of the network. Here, we show the potential of our methods in novel scenes. Particularly, we test the pre-trained ISTD model on the SRD and SBU datasets **without** retraining on their specific training datasets. We show an examples from shadow detection datasets SBU in Figure. 9. With the help of the synthesized dataset (ISTD+S in Figure. 9), our network can better understand the scenes and remove the shadows in these regions without ghosts or color inconsistencies. In the SRD dataset, when training using the ISTD+S dataset, the RMSE of our network is 9.82; however, if we train using only the ISTD dataset, the RMSE is 11.9.

Conclusion

In this paper, we aim to create a ghost-free shadow removal method in two aspects. First, we propose a novel network that starts with context aggregation network and hierarchically aggregates the features and attentions. Our network structure can produce high-quality border-free results from multi-context features and attentions. Then, to enlarge the scenes in the current shadow removal dataset and reduce the color inconsistencies, we train a shadow matting GAN to synthesize the shadow images from the corresponding shadow-free images and masks. The proposed shadow generator is used to augment the data of the current dataset. Experiments show both our methods are helpful for generating the ghost-free shadow-free images.

Acknowledgments

References

Chen, Q.; Xu, J.; and Koltun, V. 2017. Fast image processing with fully-convolutional networks. In *ICCV*, 2497–2506.

Ding, B.; Long, C.; Zhang, L.; and Xiao, C. 2019. AR-GAN: Attentive Recurrent Generative Adversarial Network for Shadow Detection and Removal. *ICCV*.

Gong, H., and Cosker, D. Interactive shadow removal and ground truth for variable scene categories.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Gryka, M.; Terry, M.; and Brostow, G. J. 2015. Learning to remove soft shadows. *ACM TOG*.

Guo, R.; Dai, Q.; and Hoiem, D. 2013. Paired Regions for Shadow Detection and Removal. *IEEE TPAMI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI* 37(9):1904–1916.

Hu, X.; Fu, C.-W.; Zhu, L.; Qin, J.; and Heng, P.-A. 2018. Direction-aware Spatial Context Features for Shadow Detection and Removal. *IEEE TPAMI*.

Hu, X.; Jiang, Y.; Fu, C.-W.; and Heng, P.-A. 2019. Mask-shadowgan: Learning to remove shadows from unpaired data. *ICCV*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*, 7132–7141.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 1125–1134.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 694–711. Springer.

Khan, S. H.; Bennamoun, M.; Sohel, F.; and Togneri, R. 2015. Automatic shadow detection and removal from a single image. *IEEE TPAMI* 38(3):431–446.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lalonde, J.-F.; Efros, A. A.; and Narasimhan, S. G. 2010. Detecting ground shadows in outdoor consumer photographs. In *ECCV*, 322–335. Springer.

Le, H., and Samaras, D. 2019. Shadow removal via shadow image decomposition. In *ICCV*.

Le, H.; Vicente, T. F. Y.; 0004, V. N.; Hoai, M.; and Samaras, D. 2018. A+D Net - Training a Shadow Detector with Adversarial Shadow Attenuation. *ECCV*.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; and Catanzaro, B. 2018. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 85–100.

Mikic, I.; Cosman, P. C.; Kogut, G. T.; and Trivedi, M. M. 2000. Moving shadow and object detection in traffic scenes. In *ICPR*, volume 1, 321–324. IEEE.

Panagopoulos, A.; Wang, C.; Samaras, D.; and Paragios, N. 2011. Illumination estimation and cast shadow detection through a higher-order graphical model. In *CVPR*, 673–680.

Qu, L.; Tian, J.; He, S.; Tang, Y.; and Lau, R. W. 2017. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, 4067–4075.

Salvador, E.; Cavallaro, A.; and Ebrahimi, T. 2004. Cast shadow segmentation using invariant color features. *CVIU* 95(2):238–259.

Shen, L.; Wee Chua, T.; and Leman, K. 2015. Shadow optimization from structured deep edge detection. In *CVPR*.

Sidorov, O. 2019. Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In *CVPR Workshop*, 0–0.

Vicente, T. F. Y.; Hou, L.; Yu, C.-P.; Hoai, M.; and Samaras, D. 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, 816–832.

Vicente, T. F. Y.; Hoai, M.; and Samaras, D. 2017. Leave-one-out kernel optimization for shadow detection and removal. *IEEE TPAMI* 40(3):682–695.

Wang, J.; Li, X.; and Yang, J. 2018. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 1788–1797.

Yang, Q.; Tan, K.-H.; and Ahuja, N. 2012. Shadow removal using bilateral filtering. *IEEE TIP* 21(10):4361–4368.

Yu, F., and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2017. Deep Layer Aggregation. *CVPR*.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. *ICCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, S.; Liang, R.; and Wang, M. 2019. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media* 5(1):105–115.

Zhang, X.; Ng, R.; and Chen, Q. 2018. Single Image Reflection Separation with Perceptual Losses. *CVPR*.

Zheng, Q.; Qiao, X.; Cao, Y.; and Lau, R. W. 2019. Distraction-aware shadow detection. In *CVPR*, 5167–5176.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *ICCV*.

Zhu, L.; Deng, Z.; Hu, X.; Fu, C.-W.; Xu, X.; Qin, J.; and Heng, P.-A. 2018. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 121–136.