2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Efficient Shadow Detection and Removal using Synthetic Data with Domain Adaptation

Rui Guo *
Toyota Motor North America
R&D InfoTech Labs
Mountain View, CA, USA
rui.guo@toyota.com

Babajide Ayinde *
Toyota InfoTechnology Center
Mountain View, CA, USA
bayinde@us.toyota-itc.com

Hao Sun
Toyota InfoTechnology Center
Mountain View, CA, USA
hsun@us.toyota-itc.com

***Abstract*—In recent years, learning based shadow detection and removal approaches have shown prospects and, in most cases, yielded state-of-the-art results. The performance of these approaches, however, relies heavily on the construction of training database of shadow images, shadow-free versions, and shadow maps as the ground truth. This conventional data gathering is time-consuming, expensive, or even practically intractable to realize especially for outdoor scenes with complicated shadow patterns, thus limiting the size of the data available for training. In this paper, we leverage on large high quality synthetic image database and domain adaptation to mitigate the bottlenecks resulting from insufficient training samples and domain bias. Specifically, our approach utilizes adversarial training to predict near-pixel-perfect shadow map from synthetic shadow image for downstreaming shadow removal steps. At inference time, we capitalize on domain adaptation via image style transfer to map the style of real-world scene to that of synthetic scene for the purpose of detecting and subsequently removing shadow. Comprehensive experiments indicate that our approach outperforms state-of-the-art methods on selected benchmark datasets.**

***Index Terms*—shadow detection, shadow removal, domain adaptation**

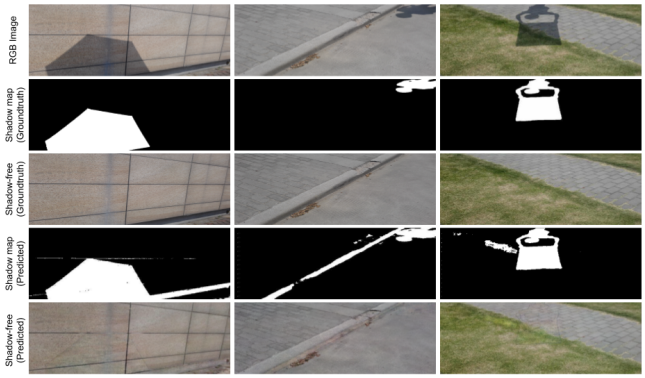

Fig. 1. First row consists of three shadow image samples from recently released Image Shadow Triplets Dataset (ISTD) [48] and the second and third rows are the corresponding groundtruth shadow maps and shadow-free, respectively. The fourth and fifth rows consist the shadow maps and shadow-free images predicted by our model.

I. Introduction

Shadow removal has been a challenging and long-standing computer vision problem due to its complexity and importance in many other tasks such as semantic segmentation, object detection and tracking, and depth estimation [4], [13], [14], [33], [35], [42]. Fundamentally, a shadow removal system takes a shadow image as the input and outputs a shadow-free version of the same image with consistent perceptual quality. Many prior arts have modeled the shadow removing process as pixel-wise scaling such that shadow image X can be obtained by scaling all pixels by illumination attenuation factors (or shadow matte) ξ to produce a shadow-free image X_{sf} in accordance to the relation $X = \xi \odot X_{sf}$, where $\odot$ is element-wise multiplication. This abstraction converts a complicated shadow removal task to a more convenient problem of unmixing illumination attenuation factors [1], [12], [13], [16], [18], [23], [29].

A peculiar drawback common to some of these methods is that, they cannot automatically estimate shadow matte without combining other prior knowledge such as shadow map [16], [23], [48] or user input [1], [12], [13]. The main challenge with this approach is that, shadow detection is a difficult standalone task and if not robustly addressed, it can lead to the deterioration of the entire shadow removal pipeline [16], [25], [51]. In addition, the size of the triplet dataset (shadow, shadow-free, and shadow mask), which is required in the supervised learning, is usually small, thereby limiting the architectural size of deep neural network that can sufficiently learn the mapping [23], [40]. Similar to [48], our pipeline incorporates shadow detection stage and addresses the problem of data shortage by training on large corpus of synthetic data.

Other limitation shared by most of the existing methods is that, their models are trained on dataset with simple outdoor scenarios, thereby undermining the generalizability and applicability of their solution. The main reason is that the dataset to train models to combat complex scenario is extremely difficult to collect, thus, existing models trained on images with relatively simply pattern and texture do not generalize to scenes with clutter background and shadow patterns. For instance, the shadow patterns encountered in autonomous driving are very complicated due to complex background and varying illumination attenuation origining from trees and buildings of varying heights [31]. Besides, creating a triplet dataset for simple shadow patterns still involves inaccurate and

* Authors have equally contributions.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413254

time-consuming thresholding and morphological filtering [48].

One underemphasized fact is that the same technology that facilitates the large-scale training of deep neural network models can also help synthesize data for efficient training of these models [2], [38], [39]. Similar image rendering engines used for gaming to capture photorealistic images can also be used to synthesize shadow images and their corresponding shadow-free versions for training a removal model. Thus, paving way for the training of shadow removal models for applications/scenarios where shadow patterns are complicated and dataset is impractical to realize.

As established in many recent studies, the idea of using synthetic data to train models for variety of machine vision problem (from depth estimation to multi-object tracking) where the groundtruth is difficult or even impossible to obtain has empirically been shown to be widely effective and plausible [2], [8], [26], [36], [39]. Although this idea is promising, however, other problem arises due to the reason that a model trained on data from one domain does not generalize to data from another domain. This is in part due to the distinctive, descriptive, and inherent attributes of the two domains.

Motivated by [2], we train a shadow detection model on synthetic data to obtain a pixel-perfect detection map and use style transfer for domain adaptation during inference on real-world data. The problem addressed here is four-fold: (i) *synthetic shadow detection* - we propose an adversarial-learning-based supervised model using light-weight architecture with skip-connections that takes a shadowed image as the input and directly outputs a pixel-perfect shadow map, (ii) *domain adaptation* - we perform domain adaptation via style transfer to transfer the style of real-world shadow image to synthetic style in order to minimize the problem of domain bias in the shadow detection stage, (iii) *shadow removal* - we pre-train a shadow removal model using adversarial learning on our synthetic data that has been transformed to real-world style and then fine-tune on a newly released Image Shadow Triplets Dataset (ISTD) by [48], and lastly (iv) we show the effectiveness of leveraging large number of synthetic data (which can easily be obtained) for training pixel-perfect shadow detection and ultimately for removing shadow.

II. Related Work

Existing methods in context of shadow detection/removal and domain adaptation/style transfer are separately discussed in the following subsections.

A. Shadow Detection/Removal

The motivation of earlier approaches stems from physical models of illumination and color. Examples of such methods develop algorithms that are invariant to illumination [5], [7]. However, the methods are only effective for images with high quality. Many existing shadow removal methods follow a two-step paradigm that first localizes shadow and then performs shadow removal. In order to localize shadow regions, these approaches require either shadow detection algorithms [16], [23] or human annotations [1], [12], [13], [50]. Subsequently in the shadow removal stage, the shadows in umbra and penumbra regions are then respectively removed using two handcrafted-feature-based algorithms.

Also, some methods focus on removing shadow in gradient [3], [6], [32] or image intensity domain [1], [12], [15], [17], [23]. The performance of these algorithms is highly reliant on the shadow detection phase and the task of detecting shadow in itself is challenging hugely because of limited training data. Shadow detection and removal based on deep convolutional neural network were first introduced in [23], [40]. To avoid overfitting, however, these algorithms are restricted to using small-sized network architecture due to data scarcity. They also require global post-processing to ensure the consistency in their predictions.

A more recent approach develops a fully convolutional neural network model pre-trained end-to-end on synthetic data to output shadow matte and in order to avoid domain bias issue, the model is then fine-tuned on small amount of real-world data [35]. One of the first work that introduced adversarial training for shadow detection utilized a conditional Generative Adversarial Network (CGAN) architecture with a tunable sensitivity [34]. The most closely related work to ours is the recent end-to-end framework that jointly learns shadow detection and removal based on stacked CGAN [48], from newly created triplet dataset, but this method still suffers from the problem of data scarcity and low quality groundtruth shadow map and shadow-free image version as observed in Figure 1.

Our method, on the other hand, uses two standalone modules to detect and remove shadow. In addition to using adversarial training to learn to predict the shadow map from a lot of synthetic shadow images for subsequent shadow removal, our pipeline also utilizes image style transfer to minimize the inter-domain bias.

B. Domain Adaptation via Style Transfer

In general, model trained on one dataset does not generalize well to other datasets due to the intrinsic bias between training dataset and the other data. Consequently, the performance of model trained on synthetic data may not be good when tested on data obtained from real-world. Typically, the most prevalent solution to the problem of domain variation is to fine-tune the trained model on the new data [35]. However, this method requires large amount of new data, which can be very time-consuming, expensive, or even practically intractable to come by resulting in the use of synthetic data in the first place [2].

Other methods enabling models trained on one dataset (source) to equally perform well another dataset (target) minimize the distance between the source and target feature distributions using maximum discrepancy approach [30], [43], [44] or adversarial training [2], [9], [10], [30], [45]. Since established in [28], style transfer can be conceptualized as a process of aligning distribution from the content image to the style image. Transferring the style of one image (from the source domain) to another image (from the target domain) is essentially a minimization of the distance between the source

and target distributions [2], [28].

Other important and popular approaches pre-train a model with large amount of data from the target style to avoid directly altering the image [22], [27], [46], [52]. We exploit this idea to adapt our real-world shadow images to our shadow removal pipeline trained on synthetic images.

III. Proposed Method

The proposed shadow removal is a three-stage scheme, powered by three models and processed in a cascade. The first stage involves training the shadow detection model on synthetic data generated from simulated environment with Unity game engine [11]. Since the detection model is likely to be biased towards the style of synthetic data, in the second stage we train another model to transfer style from real-world to the synthetic. Finally in the third stage, the output of the shadow detection is one of the essential inputs to the shadow removal model trained on both synthetic data, that has been transformed to real-world style and real-world data. In the next subsections, specifics of the shadow detection, removal and domain adaptation are discussed.

A. Shadow Detection and Removal

Both shadow detection and removal procedures are formulated as cascaded image-to-image translations that seeking to map shadow image X to shadow map K and using both X and K to predict a shadow-free image X_{sf}. However, this is challenging especially for image with complicated and distributed shadow patterns and also because we are not only interested in estimating X_{sf} from both K and X at pixel-level, but also, it is desirable that the X_{sf} to be realistic and sharp just like the groundtruth. One plausible way to achieve this high level specification is to allow model to automatically learn appropriate objective function for the task. Based on a comparative review, the most cutting-edge prediction-based methods for image-to-image translation uses adversarial training because of its tendency to reproduce sharp and photo realistic output compared to its counterparts that are based solely on l_1 or l_2 normed reconstruction loss minimization.

Generative Adversarial Networks (GANs) are trained to automatically learn a loss function by letting the discriminator justify its input is real or fake while simultaneously training a generative model to minimize the reconstruction loss. For the shadow detection, the discriminator D_{sd} aims to maximize $\mathbf{E}_{K\sim p_r(X,K)}[logD_{sd}(K)]$ when K is sampled from real distribution and given a fake image sample $G_{sd}(z)$, $z \sim p_z(z)$, it is trained to output probability, $D_{sd}(G_{sd}(z))$, close to zero by maximizing $\mathbf{E}_{z\sim p_z(z)}[log(1-D_{sd}(G_{sd}(z)))]$, where p_z is a noise distribution. The generator network G_{sd}, however, is trained to maximize the chances of D_{sd} producing a high probability for a fake image sample $G(z)$ by thus minimizing $\mathbf{E}_{z\sim p_z(z)}[log(1-D_{sd}(G_{sd}(z)))]$. The adversarial cost is obtained by combining the objectives of both D_{sd} and G_{sd} in a min-max game and from game theory point of view, GANs achieve Nash equilibrium by modeling a distribution $p_z(z)$ that is close to the empirical distribution $p_r(X,K)$.

Our approach uses conditional variant of GAN that utilizes additional observed information as input to both generator and discriminator. Specifically, generator G_{sd} and discriminator D_{sd} are conditioned on the RGB shadow image X. Generator G_{sd} is trained to synthesize shadow map $G_{sd}(z,X)$ or $\tilde{K}$ that is close as much as possible to K. At equilibrium, the expectation is that $G_{sd}(z,X)$ will converge to the Likelihood $(p_r(X,K))$. In order to ensure that the generator is deterministic, the randomly sampled noise vector z is eliminated . The adversarial cost for shadow detection module, J_{adv}, then becomes:

$$\begin{aligned} J_{adv} = \min_{G_{sd}} \max_{D_{sd}} \quad & \mathbf{E}_{X,K\sim p_r(X,K)}[log(D_{sd}(X,K))] \\ & + \mathbf{E}_{X\sim p_r(X)}[log(1-D_{sd}(X,G_{sd}(X)))] \end{aligned} \tag{1}$$

In addition to the adversarial loss, we also utilize the conventional L_1 norm reconstruction loss in (2) to encourage G_{sd} to synthesize images that are closed to the ground truth, both in terms of context and structure.

$$J_{recon} = ||K - G_{sd}(X)||_1 \tag{2}$$

The overall learning cost for training our shadow detection system then becomes:

$$J = \lambda_1 J_{adv} + \lambda_2 J_{recon} \tag{3}$$

where λ_1 and λ_2 are experimentally determined hyper-parameters that control the trade-off between the two objectives.

For the shadow removal module, the input to the generator G_{sr} is the output of G_{sd}, conditioned on RGB image X. The input of the discriminator D_{sr} is the output of both generators G_{sd} and G_{sr}, conditioned on X. Similar to the learning cost for shadow detection, the adversarial cost for shadow detection, L_{adv}, is thus given as:

$$\begin{aligned} L_{adv} = \min_{G_{sr}} \max_{D_{sr}} \quad & \mathbf{E}_{X,K,X_{sf}\sim p_r(X,K,X_{sf})}[log(D_{sr}(X,K,X_{sf}))] \\ & + \mathbf{E}_{X\sim p_r(X)}[log(1-D_{sr}(X,G_{sd}(X),G_{sr}(X,G_{sd}(X))))] \end{aligned} \tag{4}$$

The overall learning cost for training our shadow removal system is:

$$L = \lambda_3 L_{adv} + \lambda_4 L_{recon} \tag{5}$$

where λ_3 and λ_4 are hyper-parameters and,

$$L_{recon} = ||X_{sf} - G_{sr}(X, G_{sd}(X))||_1 \tag{6}$$

As opposed to [48] that concurrently trains both shadow detection and removal, our shadow detection and removal are standalone modules and are trained and used in tandem.

B. Adaptation of Domain through Style Transfer

Since our shadow detection module was trained on data obtained from a synthetic environment, topnotch performance is not guaranteed on real-world images. This is mainly due to the high disparity between real-world images and synthetic images. The goal of domain adaptation is to minimize this

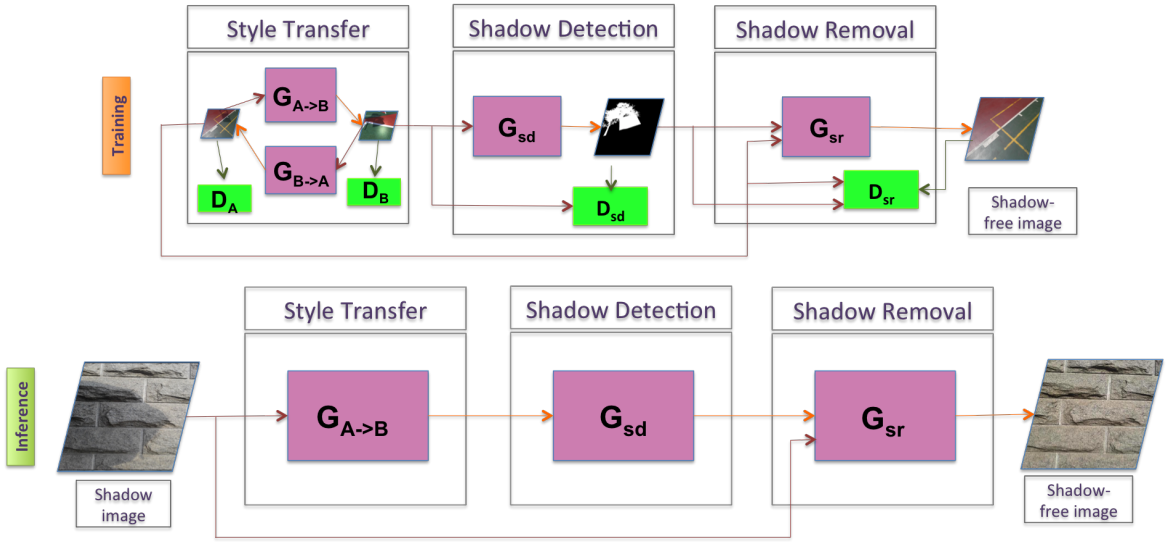

Fig. 2. Proposed shadow removal pipeline. Real-world image (Domain A) is first translated into the style of synthetic database (Domain B) using a CycleGAN framework [21]. Shadow region (shadow map) is detected from the styled image using shadow detection model trained only on synthetic data. Finally, both shadow image and map are used to predict a shadow-free image version.

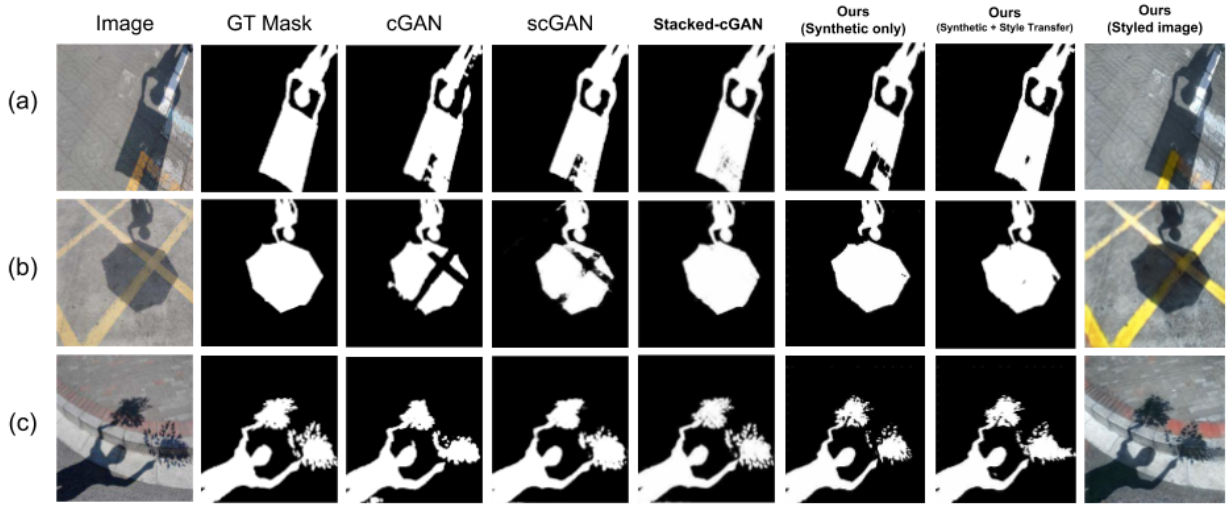

Fig. 3. Qualitative comparison of shadow detection on select ISTD test set

disparity. Specifically, we want to learn a function $\mathcal{F} : A \rightarrow B$ that maps images in real-world (domain A) to synthetic world (domain B) such that the distribution of $\mathcal{F}(A)$ is very similar to that of B. This invariably means that an image in domain A needs to be translated into domain B before estimating the shadow map with a model trained on samples from B.

Similar to [2], we adopt the method of using adversarial training with cycle-consistency [21], [52] to adapt these two domains via image style transfer. This approach utilizes two generators ($G_{A\rightarrow B}$ and $G_{B\rightarrow A}$) and two discriminators (D_A and D_B) to model two functions that map samples from domain A to B, and vice versa. D_A is trained to decipher between sample X from domain A and the output of $G_{B\rightarrow A}$, while D_B discriminates between sample Y from domain B and synthesized output from $G_{A\rightarrow B}$. The complete objective function used for training the style transfer model comprises of adversarial losses ($J_{adv}^{A\rightarrow B}$ and $J_{adv}^{B\rightarrow A}$), cycle-consistency loss (J_{cycle}), and identity mapping loss as given in 7:

$$J_{net} = J_{adv}^{A\rightarrow B} + J_{adv}^{B\rightarrow A} + \gamma J_{cycle} + J_{identity} \tag{7}$$

where

$$\begin{aligned} J_{adv}^{A\rightarrow B} = \min_{G_{A\rightarrow B}} \max_{D_B} \; & \mathbf{E}_{Y\sim p_r(Y)}[log(D_B(Y))] \\ & + \mathbf{E}_{X\sim p_r(X)}[log(1 - D_B(G_{A\rightarrow B}(X)))], \end{aligned} \tag{8}$$

$$\begin{aligned} J_{adv}^{B\rightarrow A} = \min_{G_{B\rightarrow A}} \max_{D_A} \; & \mathbf{E}_{X\sim p_r(X)}[log(D_A(X))] \\ & + \mathbf{E}_{Y\sim p_r(Y)}[log(1 - D_A(G_{B\rightarrow A}(Y)))], \end{aligned} \tag{9}$$

$$\begin{aligned} J_{cycle} = \; & ||G_{B\rightarrow A}(G_{A\rightarrow B}(X)) - X||_1 \\ & + ||G_{A\rightarrow B}(G_{B\rightarrow A}(Y)) - Y||_1. \end{aligned} \tag{10}$$

and

$$J_{identity} = ||G_{A\to B}(Y)) - Y||_1 + ||G_{B\to A}(X)) - X||_1. \tag{11}$$

The purpose of the cycle consistency is to penalize the model for producing random images that are contextually incoherent. By the end of a looped training cycle, $J_{cycle} \to 0$, then, $G_{B\to A}(G_{A\to B}(X)) \approx X$ and $G_{A\to B}(G_{B\to A}(Y)) \approx Y$. On the other hand, $J_{identity}$ enforces near identity mapping whenever real samples of target domain are given as input to the generator. It encourages the generator to preserve color composition between input and output as much as possible. Figure 2 depicts the end-to-end shadow removal pipeline including the detection, style transfer and shadow removal.

C. Network Structure and Implementation Details

All experiments were performed on Intel(r) Core(TM) i7-5930K CPU @ 3.50Ghz and a 64GB of RAM running a 64-bit Ubuntu 16.04 edition. The software implementation has been in Pytorch on four Titan X 12GB GPUs. ADAM [24] optimizer was used in all the experiment with momentum $\beta_1 = 0.5$, $\beta_2 = 0.999$, and learning rate of 0.0002.

1) Shadow detection: As earlier mentioned, we train the shadow detection module on synthetic data only. Our synthetic data was captured by a camera that moved around in a virtual environment. The virtual environment has been created in Unity game engine version 2018.2.0f2 64-bit Pro. Shadow was casted on projection planes by shining different light sources on different occluding objects with randomly varying heights, field of view, and light source direction. Shadow images and their corresponding shadow maps were captured every 20 frames. We captured 135,000 480×640 triplets of shadow image, shadow map, and shadow-free image. The dataset was split into 120,000 and 15,000 in training and testing sets, respectively.

The backbone generator design follows the encoder-decoder structure with skip-connections, originally introduced in U-Net [37]. This concept of skipping connection is important because it enables the exchange of geometric information between corresponding encoder and decoder layers. The design of generator G_{sd} follows the approach detailed in [21]. The discriminator D_{sd}, on the other hand, takes as input a 4-channel feature map (obtained by concatenating the shadow map with its corresponding synthetic RGB image). Both generator and discriminator use BatchNorm [20] and slightly different activation function after each convolution; generator use ReLU while in the discriminator, leaky ReLU with slope 0.2 is used. Hyperparameters λ_1 and λ_2 are empirically set to 0.01 and 0.99, respectively.

2) Shadow removal: The architectural setup for the shadow removal part is similar to the aforementioned detection module. Similar to hyperparameters for training shadow detection, $\lambda_3 = \lambda_1$ and $\lambda_4 = \lambda_2$. The major difference is that:

- As opposed to training on synthetic data, we pre-trained the shadow removal module first on our synthetic data that has been transformed to real-world style using $G_{B\to A}$. We then fine-tuned on ISTD [48] - real-world triplet dataset consisting of 1330 triplets of shadow-map, shadow, shadow-free images.
- The input to the generator G_{sr} and discriminator D_{sr} are 4 and 7-channel feature maps, respectively.

3) Style transfer: The architectures of D_A and D_B are similar to the discriminator in shadow detection stage but differ in the number of input channel. Both D_A and D_B take as input a 3-channel feature map corresponding to styled image ($G_{B\to A}(Y)$ or $G_{A\to B}(X)$) or original image (X or Y). Also, both D_A and D_B are updated based on not only the last generated output but on 10 generator outputs [21], [41], [52]. The structures of both generators $G_{A\to B}$ and $G_{B\to A}$ are similar to the architecture in [2], [22] where the input is processed through two layers of convolution followed by six residual layers [19] and lastly through two additional layers of convolution to recover original image size. Hyperparameter γ is experimentally set to 10.

IV. Experimental Results

This section evaluates and reports the performance of the proposed method on two benchmark datasets: SBU [47] and ISTB [48]. SBU database contains 4089 training and 638 testing pairs of shadow image and shadow mask. ISTB database on the other hand has 1330 and 540 pairs for training and testing, respectively. For the shadow detection stage, we benchmark our method with four existing techniques namely: stackedCGAN [48] stackedCNN [34], cGAN [47], and scGAN [47] using Balance Error Rate (BER) in Eq. 12.

$$BER = 1 - \frac{1}{2}\Big(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\Big) \tag{12}$$

For the shadow removal, we also benchmark our pipeline with recently proposed heuristics [12], [16], [48], [49] using root mean square error in LAB color space between groundtruth and the predicted shadow-free image.

A. Evaluation of Shadow detection

We design multiple experiments on ISTD/SBU benchmarks to receive the comprehensive evaluation of the algorithm. The performance of our proposed shadow detection models (with synthetic data and/or style transfer) is compared with the state-of-the-art on the basis of their BER measure in 12 and reported in the section. The first two experiments are conducted with training and testing within the same dataset respectively. It is notable that our model was trained on style transferred data, and in the inference, the real-world style testing images are also transferred into synthetic domain. The quantitative results are recorded in Table III-B. On the BER basis, the proposed algorithm received competitive results.

To better understand the generalization capability so as the power of style transfer, we also conduct the experiments with training solely on synthetic data and at inference, we transferred the real-world style of ISTD test set to synthetic style. All other benchmark models were trained using SBU

TABLE I

QUANTITATIVE RESULTS ON SHADOW DETECTION USING BER (SMALLER IS BETTER). ALL MODELS WERE TRAINED AND TESTED WITHIN THE SAME ISTD/SBU DATASET. MARK S INDICATES THE TRAINING WITH SYNTHETIC DATA, AND ST INDICATES THE TESTING WITH STYLE TRANSFER.

Train/Test Dataset	Detection Aspects	StackedCNN [34]	cGAN [47]	scGAN [47]	stackedCGAN [48]	ours (S+ST)
ISTD/ISTD	Shadow	7.96	10.81	3.22	**2.14**	2.89
	Non-shadow	9.23	8.48	6.18	**5.55**	5.82
	BER	8.6	9.64	4.7	**3.85**	4.35
SBU/SBU	Shadow	9.6	20.5	7.8	3.75	**3.45**
	Non-shadow	12.5	**6.9**	10.4	12.53	11.83
	BER	11.0	13.6	9.1	8.14	**7.64**

TABLE II

QUANTITATIVE RESULTS ON SHADOW DETECTION USING BER. ALL BENCHMARK MODELS WERE TRAINED ON SBU DATASET AND TESTED ON ISTD

Train/Test Dataset	Detection Aspects	StackedCNN [34]	cGAN [47]	scGAN [47]	stackedCGAN [48]	ours (S)	ours (S+ST)
SBU/ISTD	Shadow	11.33	19.93	9.5	**4.8**	9.13	**7.05**
	Non-shadow	9.57	**4.92**	8.46	9.9	**3.17**	7.04
	BER	10.45	12.42	8.98	7.35	**6.15**	**7.05**

TABLE III

QUANTITATIVE RESULTS ON SHADOW DETECTION USING BER. ALL BENCHMARK MODELS WERE TRAINED ON ISTD DATASET AND TESTED ON SBU

Train/Test Dataset	Detection Aspects	StackedCNN [34]	cGAN [47]	scGAN [47]	stackedCGAN [48]	ours (S)
ISTD/SBU	Shadow	11.29	24.07	9.1	9.02	**7.50**
	Non-shadow	20.49	**13.13**	17.41	13.66	13.37
	BER	15.94	18.60	13.26	11.34	**10.57**

TABLE IV

QUANTITATIVE RESULTS ON SHADOW REMOVAL USING RMSE. ALL BENCHMARK MODELS WERE TRAINED AND TESTED ON ISTD

Test Dataset	Detection Aspects	Original	Yang [49]	Guo [16]	Gong [12]	stackedCGAN [48]	ours
ISTD	Shadow	32.67	19.82	18.95	14.98	10.33	**5.30**
	Non-shadow	6.83	14.83	7.46	7.29	6.93	**6.81**
	BER	10.97	15.63	9.30	8.53	7.47	**5.78**

database and tested on ISTD test set to compare their generalization. As can be seen in Table II that our methods (with and without domain adaptation via style transfer) outperform all benchmark methods in terms of BER for the entire shadow map even with only synthetic data. However, the detection rate of [48] is better than ours in the shadow region. Our defensive argument is that, it happens because the groundtruth of ISTD database is inaccurately labeled as shown in Figure 1. Qualitative results in Figure 3 also show that our approach is highly competitive with state-of-the-art.

To combat the effect of testing on data with inaccurate annotation, we again compare our detection model with the state-of-the-art models trained on ISTD but tested on a more accurate benchmark test database (SBU) using the BER. Here, we clearly show in Table III-B that, except for the cGAN on non-shadow region, our model trained on synthetic data outperforms benchmark methods even without the need for domain adaptation. This reinforces the general hypothesis that more data is always better and that creating accurate groundtruth database for shadow removal task is arduous.

B. Evaluation of Shadow Removal

The performance of our shadow removal model is also compared with four recent methods on the basis of RMSE using the ISTD test set for three regions (shadow region, non-shadow region, and whole image). As can be observed in Figures 1 and 4 that our approach is both qualitative better and more accurate than existing methods. This performance is attributed in part to accurate localization of shadow region by our shadow detection model that leveraging large quantity of synthetic data to predict near-pixel-perfect. For quantitative evaluation, we also benchmark our model with other methods in Table III-B using RMSE. Our approach again outperforms all existing methods in terms of RMSE between groundtruth and predicted shadow-free images for shadow region and entire image. This implies that our method works better than state-of-the-art on ISTD test set. However, for non-shadow regions, RMSE of ours is much higher than other methods. This is well expected because we detect and remove shadow in regions that still have shadows in the groundtruth.

For the completion of the evaluation, we demonstrate the

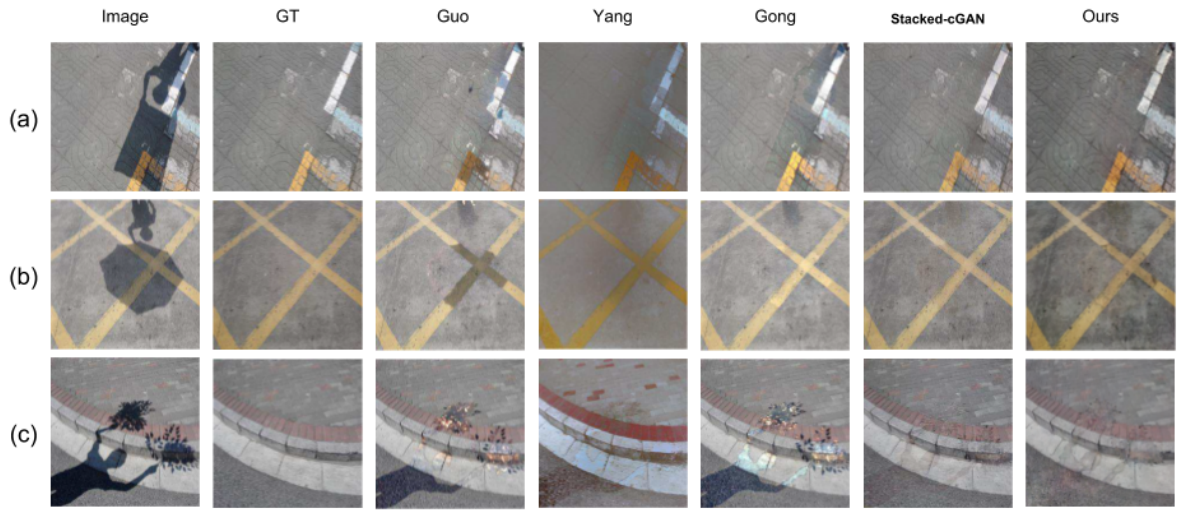

Fig. 4. Qualitative comparison of shadow removal on select ISTD test samples

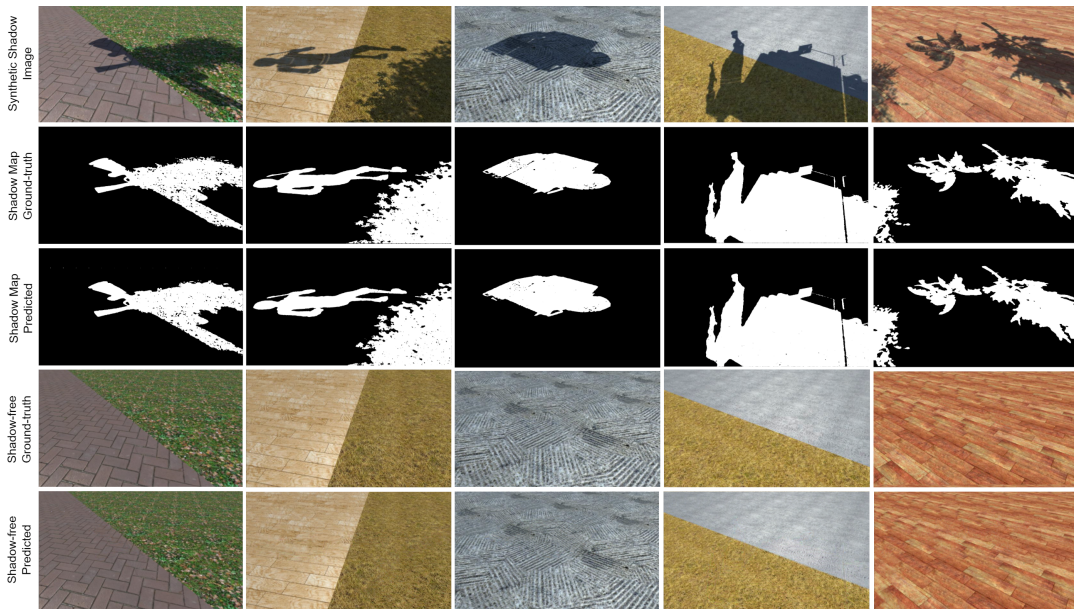

Fig. 5. Qualitative demonstration of the synthetic images and the shadow detection/removal process on them.

synthetic data we created and the shadow detection/removal process on it in Figure 5. The qualitative illustration reveals the advantages of the algorithm.

V. Conclusion

This paper addresses the possibility of leveraging large-scale high quality synthetic image database to train shadow detection and removal models since conventional data gathering method is time-consuming, expensive, or even practically intractable to realize in some outdoor scenarios thereby limiting the size of training database. Since models trained on synthetic data may not generalize well on real-world data, we leverage on GAN-based style transfer to align the real-world data distribution with the distribution approximated by the shadow detection and removal models. The performance of the proposed method was competitively advantageous compared with the state-of-the-art approaches in terms of Balance Error Rate and root mean square error for shadow detection and removal, respectively.

References

[1] E. Arbel and H. Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1202–1216, 2011.

[2] A. Atapour-Abarghouei and T. P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 18, page 1, 2018.

[3] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2015.

[4] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence*, 2003.

[5] G. D. Finlayson, M. S. Drew, and C. Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009.

[6] G. D. Finlayson, S. D. Hordley, and M. S. Drew. Removing shadows from images. In *European conference on computer vision*, pages 823–836. Springer, 2002.

[7] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):59–68, 2006.

[8] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[9] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[10] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[11] W. Goldstone. *Unity game development essentials*. Packt Publishing Ltd, 2009.

[12] H. Gong and D. Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC 2014-Proceedings of the British Machine Vision Conference 2014*. University of Bath, 2014.

[13] M. Gryka, M. Terry, and G. J. Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics (TOG)*, 34(5):153, 2015.

[14] R. Guo, B. Ayinde, H. Sun, H. Muralidharan, and K. Oguchi. Monocular depth estimation using synthetic images with shadow removal. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1432–1439. IEEE, 2019.

[15] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2033–2040.

[16] R. Guo, Q. Dai, and D. Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2013.

[17] R. Guo and H. Qi. Partially-sparse restricted boltzmann machine for background modeling and subtraction. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pages 209–214. IEEE, 2013.

[18] R. Guo, W. Wang, and H. Qi. Hyperspectral image unmixing using autoencoder cascade. In *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4. IEEE, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 448–456. JMLR. org, 2015.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[22] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[23] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Automatic shadow detection and removal from a single image. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):431–446, 2016.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *In Proc. Int. Conf. learning Representations*, 2014.

[25] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *European conference on computer vision*, pages 322–335. Springer, 2010.

[26] T. A. Le, A. G. Baydin, R. Zinkov, and F. Wood. Using synthetic data to train neural networks is model-based reasoning. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 3514–3521.

[27] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[28] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[29] F. Liu and M. Gleicher. Texture-consistent shadow removal. In *European Conference on Computer Vision*, pages 437–450. Springer, 2008.

[30] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[31] Y. Meng, Y. Lu, A. Raj, S. Sunarjo, R. Guo, T. Javidi, G. Bansal, and D. Bharadia. Signet: Semantic instance aided unsupervised 3d geometry perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9810–9820, 2019.

[32] A. Mohan, J. Tumblin, and P. Choudhury. Editing soft shadows in a digital photograph. *IEEE Computer Graphics and Applications*, 27(2), 2007.

[33] S. Nadimi and B. Bhanu. Physical models for moving shadow and object detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 26(8):1079–1087, 2004.

[34] V. Nguyen, T. F. Y. Vicente, M. Zhao, M. Hoai, and D. Samaras. Shadow detection with conditional generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4520–4528. IEEE, 2017.

[35] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.

[36] P. S. Rajpura, H. Bojinov, and R. S. Hegde. Object detection using deep cnns trained on synthetic images. *arXiv preprint arXiv:1706.06782*, 2017.

[37] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] A. Ruano Miralles. An open-source development environment for self-driving vehicles. 2017.

[39] S. Shah, D. Dey, C. Lovett, and A. Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*, pages 621–635. Springer, 2018.

[40] L. Shen, T. Wee Chua, and K. Leman. Shadow optimization from structured deep edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2067–2074, 2015.

[41] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, volume 2, page 5, 2017.

[42] J. Stander, R. Mech, and J. Ostermann. Detection of moving cast shadows for object segmentation. *IEEE Transactions on multimedia*, 1(1):65–76, 1999.

[43] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[44] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[45] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[46] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016.

[47] T. F. Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *European Conference on Computer Vision*, pages 816–832. Springer, 2016.

[48] J. Wang, X. Li, and J. Yang. 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*, pages 1788–1797, 2018.

[49] Q. Yang, K.-H. Tan, and N. Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image processing*, 21(10):4361–4368, 2012.

[50] L. Zhang, Q. Zhang, and C. Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015.

[51] J. Zhu, K. G. Samuel, S. Z. Masood, and M. F. Tappen. Learning to recognize shadows in monochromatic natural images. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 223–230. IEEE, 2010.

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.