Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal

Jifeng Wang[1*], Xiang Li[2*], Jian Yang[1,2†]
DeepInsight@PCALab, Nanjing University of Science and Technology
jfwang.cs@gmail.com, {xiang.li.implus, csjyang}@njust.edu.cn

Abstract

Understanding shadows from a single image consists of two types of task in previous studies, containing shadow detection and shadow removal. In this paper, we present a multi-task perspective, which is not embraced by any existing work, to jointly learn both detection and removal in an end-to-end fashion that aims at enjoying the mutually improved benefits from each other. Our framework is based on a novel STacked Conditional Generative Adversarial Network (ST-CGAN), which is composed of two stacked CGANs, each with a generator and a discriminator. Specifically, a shadow image is fed into the first generator which produces a shadow detection mask. That shadow image, concatenated with its predicted mask, goes through the second generator in order to recover its shadow-free image consequently. In addition, the two corresponding discriminators are very likely to model higher level relationships and global scene characteristics for the detected shadow region and reconstruction via removing shadows, respectively. More importantly, for multi-task learning, our design of stacked paradigm provides a novel view which is notably different from the commonly used one as the multi-branch version. To fully evaluate the performance of our proposed framework, we construct the first large-scale benchmark with 1870 image triplets (shadow image, shadow mask image, and shadow-free image) under 135 scenes. Extensive experimental results consistently show the advantages of ST-CGAN over several representative state-of-the-art methods on two large-scale publicly available datasets and our newly released one.

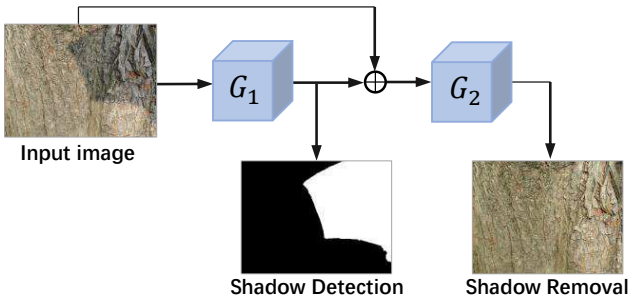


Figure 1. We propose an end-to-end stacked joint learning architecture for two tasks: shadow detection and shadow removal.

1. Introduction

Both shadow detection and shadow removal reveal their respective advantages for scene understanding. The accurate recognition of shadow area (i.e., shadow detection) provides adequate clues about the light sources [25], illumination conditions [38, 39, 40], object shapes [37] and geometry information [19, 20]. Meanwhile, removing the presence of shadows (i.e., shadow removal) in images is of great interest for the downstream computer vision tasks, such as efficient object detection and tracking [3, 32]. Till this end, existing researches basically follow one of the following pipelines for understanding shadows:

Detection only. In the history of shadow detection, a series of data-driven statistical learning approaches [15, 26, 50, 59, 22, 49] have been proposed. Their main objective is to find the shadow regions, in a form of an image mask that separates shadow and non-shadow areas.

Removal only. A list of approaches [7, 5, 58, 10, 47, 1, 54, 29, 43] simply skips the potential information gained from the discovery of shadow regions and directly produces the illumination attenuation effects on the whole image, which is also denoted as a shadow matte [43], to recover the image with shadows removed naturally.

Two stages for removal. Many of the shadow removal methods [11, 12, 23, 8, 51] generally include two *seperated* steps: shadow localization and shadow-free reconstruction by exploiting the intermediate results in the awareness of

*Authors contributed equally.
†Corresponding author.
[1]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, P.R. China.
[2]Jiangsu Key Laboratory of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, P.R. China.

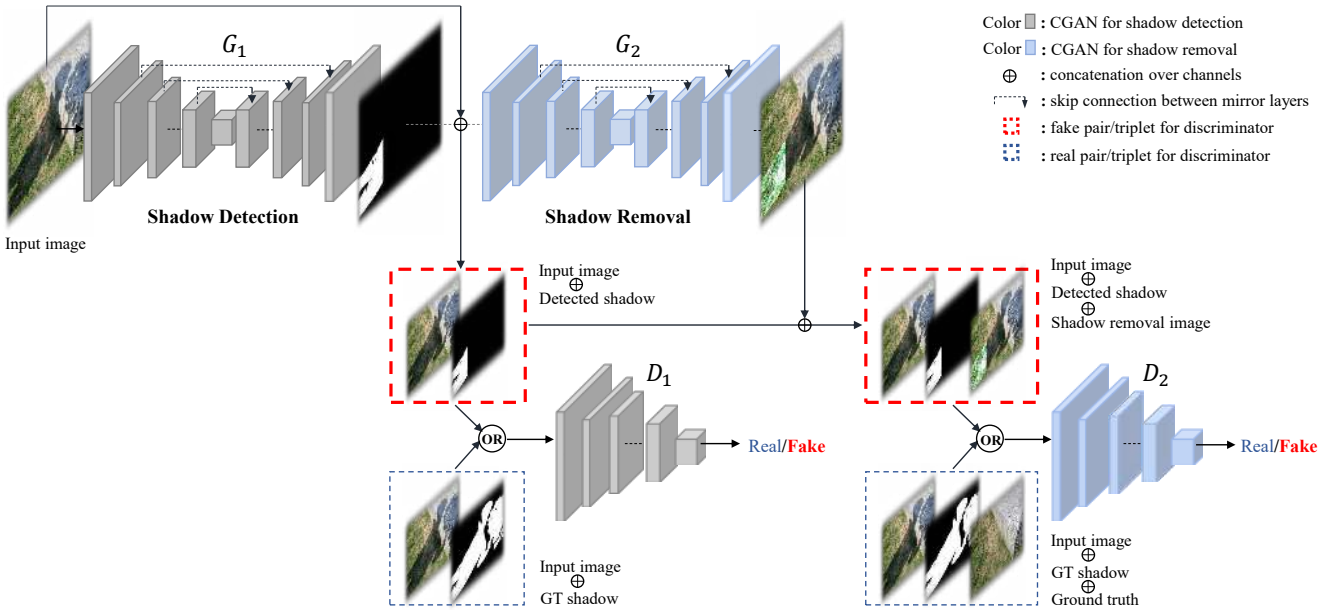


Figure 2. The architecture of the proposed ST-CGAN. It consists of two stacked CGANs: one for shadow detection and another for shadow removal, which are marked in different colors. The intermediate outputs are concatenated together as the subsequent components' input.

shadow regions.

It is worth noting that the two targets: shadow mask in detection and shadow-free image in shadow removal, share a fundamental characteristic essentially. As shown in Figure 1, the shadow mask is a two-binary map that segments the original image into two types of region: shadow and non-shadow. Meanwhile, shadow removal mainly focuses on the shadowed area and needs to discover the semantic relationship between the two regions, which indicates strong correlations and possible mutual benefits between two tasks.

Besides, most of the previous methods, including shadow detection [15, 26, 50, 59, 22, 49] and removal [8, 54, 1] are heavily based on local region classifications or low-level feature representations, failing to reason about the global scene semantic structure and illumination conditions. Consequently, a most recent study [36] in shadow detection introduced a Conditional Generative Adversarial Network (CGAN) [33] which is proved to be effective for the global consistency. For shadow removal, Qu et al. [43] also proposed a multi-context architecture with an end-to-end manner, which maintained a global view of feature extraction.

Since no existing approaches have explored the joint learning aspect of these two tasks, in this work, we propose a STacked Conditional Generative Adversarial Network (ST-CGAN) framework and aim to tackle shadow detection and shadow removal problems simultaneously in an end-to-end fashion. Through the stacked adversarial components, the potential mutual promotions between the two tasks can be fully used, and the global perceptions are well preserved. Further, our design of stacked modules is not only to achieve a multi-task purpose, but also inspired from the connectivity pattern of DenseNet [14] and MixNet [53], where outputs of all preceding tasks are used as inputs for all subsequent tasks. The densely connected path has also been proved effective in many low-level vision fields[56, 46]. Specifically, we construct ST-CGAN by stacking two generators along with two discriminators. In Figure 2, each generator takes every prior target of tasks (including the input) and stacks them as its input. Similarly, the discriminator attempts to distinguish the concatenation of all the previous tasks' targets from the real corresponding ground-truth pairs or triplets.

Importantly, the design of the proposed stacked components offers a novel perspective for multi-task learning in the literature. Different from the commonly used multi-branch paradigm (e.g., Mask R-CNN [13], in which each individual task is assigned with a branch), we stack all the tasks that can not only focus on one task once a time in different stages, but also share mutual improvements through forward/backward information flows. Instead, the multi-branch version aims to learn a shared embedding across tasks by simply aggregating the supervisions from each individual task.

To validate the effectiveness of the proposed framework, we further construct a new large-scale Dataset with Image Shadow Triplets (ISTD) consisting of shadow, shadow mask and shadow-free image to match the demand of multi-task learning. It contains 1870 image triplets under 135 distinct scenarios, in which 1330 is assigned for training whilst 540 is for testing.

Extensive experiments on two large-scale publicly available benchmarks and our newly released dataset show that ST-CGAN performs favorably on both detection and removal aspects, comparing to several state-of-the-art methods. Further, we empirically demonstrate the advantages of our stacked joint formula over the widely used multi-branch version for shadow understanding. To conclude, the main contributions of this work are listed as follows:

- It is the first end-to-end framework which jointly learns shadow detection and shadow removal with superior performances on various datasets and on both the two tasks.
- A novel STacked Conditional Generative Adversarial Network (ST-CGAN) with a unique stacked joint learning paradigm is proposed to exploit the advantages of multi-task training for shadow understanding.
- The first large-scale shadow dataset which contains *image triplets* of shadow, shadow mask and shadow-free image is publicly released.

2. Related Work

Shadow Detection. To improve the robustness of shadow detection on consumer photographs and web quality images, a series of data-driven approaches [15, 26, 59] have been taken and been proved to be effective. Recently, Khan et al. [22] first introduced deep Convolutional Neural Networks (CNNs) [45] to automatically learn features for shadow regions/boundaries that significantly outperforms the previous state-of-the-art. A multikernel model for shadow region classification was proposed by Vicente et al. [49] and it is efficiently optimized based on least-squares SVM leave-one-out estimates. More recent work of Vicente et al. [50] used a stacked CNN with separated steps, including first generating the image level shadow-prior and training a patch-based CNN which produces shadow masks for local patches. Nguyen et al. [36] presented the first application of adversarial training for shadow detection and developed a novel conditional GAN architecture with a tunable sensitivity parameter.

Shadow Removal. Early works are motivated by physical models of illumination and color. For instance, Finlayson et al. [5, 7] provide the illumination invariant solutions that work well only on high quality images. Many existing approaches for shadow removal include two steps in general. For the removal part of these two-stage solutions, the shadow is erased either in the gradient domain [6, 35, 2] or the image intensity domain [1, 11, 12, 8, 23]. On the contrary, a few works [47, 55, 42] recover the shadow-free image by intrinsic image decomposition and preclude the need of shadow prediction in an end-to-end manner. However, these methods suffer from altering the colors of the non-shadow regions. Qu et al. [43] further propose a multi-context architecture which consists of three levels (global localization, appearance modeling and semantic modeling) of embedding networks, to explore shadow removal in an end-to-end and fully automatic framework.

CGAN and Stacked GAN. CGANs have achieved impressive results in various image-to-image translation problems, such as image superresolution [27], image inpainting [41], style transfer [28] and domain adaptation/transfer [18, 60, 30]. The key of CGANs is the introduction of the *adversarial loss* with an informative conditioning variable, that forces the generated images to be with high quality and indistinguishable from real images. Besides, recent researches have proposed some variants of GAN, which mainly explores the stacked scheme of its usage. Zhang et al. [57] first put forward the StackGAN to progressively produce photo-realistic image synthesis with considerably high resolution. Huang et al. [16] design a top-down stack of GANs, each learned to generate lower-level representations conditioned on higher-level representations for the purpose of generating more qualified images. Therefore, our proposed stacked form is distinct from all the above relevant versions in essence.

Multi-task Learning. The learning hypothesis is biased to prefer a shared embedding learnt across multiple tasks. The widely adopted architecture of multi-task formulation is a shared component with multi-branch outputs, each for an individual task. For example, in Mask R-CNN [13] and MultiNet [48], 3 parallel branches for object classification, bounding-box regression and semantic segmentation respectively are utilized. Misra et al. [34] propose "cross-stitch" unit to learn shared representations from multiple supervisory tasks. In Multi-task Network Cascades[4], all tasks share convolutional features, whereas later task also depends the output of a preceding one.

3. A new Dataset with Image Shadow Triplets – ISTD

Existing publicly available datasets are all limited in the view of multi-task settings. Among them, SBU [52] and UCF [59] are prepared for shadow detection only, whilst SRD [43], UIUC [12] and LRSS [10] are constructed for the purpose of shadow removal accordingly.

Dataset	Amount	Content of Images	Type
SRD [43]	3088	shadow/shadow-free	pair
UIUC [12]	76	shadow/shadow-free	pair
LRSS [10]	37	shadow/shadow-free	pair
SBU [52]	4727	shadow/shadow mask	pair
UCF [59]	245	shadow/shadow mask	pair
ISTD (ours)	1870	shadow/shadow mask/shadow-free	triplet

Table 1. Comparisons with other popular shadow related datasets. Ours is unique in the content and type, whilst being in the same order of magnitude to the most large-scale datasets in amount.

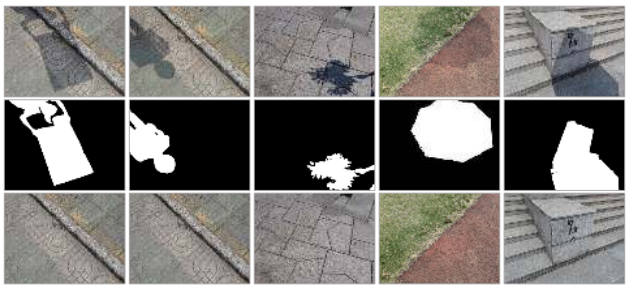


Figure 3. An illustration of several shadow, shadow mask and shadow-free image triplets in ISTD.

To facilitate the evaluation of shadow understanding methods, we have constructed a large-scale Dataset with Image Shadow Triplets called *ISTD*[2]. It contains 1870 triplets of shadow, shadow mask and shadow-free image under 135 different scenarios. To the best of our knowledge, ISTD is the first large-scale benchmark for simultaneous evaluations of shadow detection and shadow removal. Detailed comparisons with previous popular datasets are listed in Table 1.

In addition, our proposed dataset also contains a variety of properties in the following aspects:

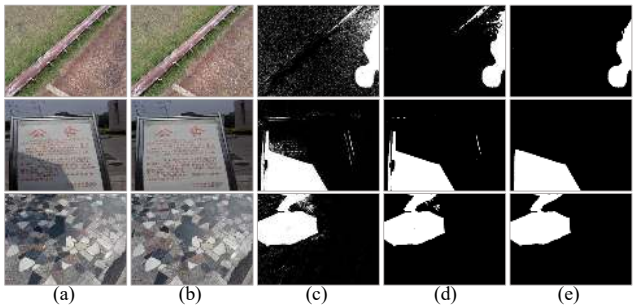
- **Illumination:** Minimized illumination difference between a shadow image and the shadow-free one is obtained. When constructing the dataset, we pose a camera with a fixed exposure parameter to capture the shadow image, where the shadow is cast by an object. Then the occluder is removed in order to get the corresponding shadow-free image. More evidences are given in the 1st and 3rd row of Figure 3.
- **Shapes:** Various shapes of shadows are built by different objects, such as umbrellas, boards, persons, twigs and so on. See the 2nd row of Figure 3.
- **Scenes:** 135 different types of ground materials, e.g., 3th-5th column in Figure 3, are utilized to cover as many complex backgrounds and different reflectances as possible.

Note that even these shadow and shadow-free image pairs are taken within a very short period of time by a fixed camera, illumination noises are unavoidable due to the slight changes of ambient light. As illustrated in Figure 4 (c), 4 (d) and 4 (e), obtaining the final shadow mask for each image generally consists of three steps.

4. Proposed Method

We propose STacked Conditional Generative Adversarial Networks (ST-CGANs), a novel stacked architecture that

[2]ISTD dataset is publicly available at https://github.com/DeepInsight-PCALab/ST-CGAN



(a) (b) (c) (d) (e)

Figure 4. The pipeline for annotating shadow masks of ISTD. (a) – shadow image by camera; (b) – shadow-free image by camera; (c) – thresholding the difference between (a) and (b); (d) – morphological filtering; (e) – manually adjusting label mask for each erroneous pixel.

enables the joint learning for shadow detection and shadow removal, as shown in Figure 2. In this section, we first describe the formulations with loss functions, training procedure, and then present the network details of ST-CGAN, followed by a subsequent discussion.

4.1. STacked Conditional Generative Adversarial Networks

Generative Adversarial Networks (GANs) [9] consists of two players: a generator G and a discriminator D. These two players are competing in a zero-sum game, in which the generator G aims to produce a realistic image given an input $\mathbf{z}$, that is sampled from a certain noise distribution. The discriminator D is forced to classify if a given image is generated by G or it is indeed a real one from the dataset. Hence, the adversarial competition progressively facilitates each other, whilst making it hard for D to differentiate G's generation from the real data. Conditional Generative Adversarial Networks (CGANs) [33] extends GANs by introducing an additional observed information, named conditioning variable, to both the generator G and discriminator D.

Our ST-CGAN consists of two Conditional GANs in which the second one is stacked upon the first. For the first CGAN of ST-CGAN in Figure 2, both the generator G_1 and discriminator D_1 are conditioned on the input RGB shadow image $\mathbf{x}$. G_1 is trained to output the corresponding shadow mask $G_1(\mathbf{z}, \mathbf{x})$, where $\mathbf{z}$ is the random sampled noise vector. We denote the ground truth of shadow mask for $\mathbf{x}$ as $\mathbf{y}$, to which $G_1(\mathbf{z}, \mathbf{x})$ is supposed to be close. As a result, G_1 needs to model the distribution $p_{data}(\mathbf{x}, \mathbf{y})$ of the dataset. The objective function for the first CGAN is:

$$\mathcal{L}_{CGAN_1}(G_1, D_1) = \mathbf{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\log D_1(\mathbf{x}, \mathbf{y})] + \mathbf{E}_{\mathbf{x}\sim p_{data}(\mathbf{x}),\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D_1(\mathbf{x}, G_1(\mathbf{z}, \mathbf{x})))]. \quad (1)$$

We further eliminate the random variable $\mathbf{z}$ to have a deterministic generator G_1 and thus the Equation (1) is sim-

Network	Layer	Cv_0	Cv_1	Cv_2	Cv_3	Cv_4 (×3)	Cv_5	CvT_6	CvT_7 (×3)	CvT_8	CvT_9	CvT_{10}	CvT_{11}
G_1/G_2	#C_in	3/4	64	128	256	512	512	512	1024	1024	512	256	128
	#C_out	64	128	256	512	512	512	512	512	256	128	64	1/3
	before	–	LReLU	LReLU	LReLU	LReLU	LReLU	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
	after	–	BN	BN	BN	BN	–	BN	BN	BN	BN	BN	Tanh
	link	$\rightarrow \mathrm{CvT}_{11}$	$\rightarrow \mathrm{CvT}_{10}$	$\rightarrow \mathrm{CvT}_9$	$\rightarrow \mathrm{CvT}_8$	$\rightarrow \mathrm{CvT}_7$	–	–	$\mathrm{Cv}_4 \rightarrow$	$\mathrm{Cv}_3 \rightarrow$	$\mathrm{Cv}_2 \rightarrow$	$\mathrm{Cv}_1 \rightarrow$	$\mathrm{Cv}_0 \rightarrow$

Table 2. The architecture for generator G_1/G_2 of ST-CGAN. Cv_i means a classic convolutional layer whilst CvT_i stands for a transposed convolutional layer that upsamples a feature map. Cv_4 (×3) indicates that the block of Cv_4 is replicated for additional two times, three in total. "#C_in" and "#C_out" denote for the amount of input channels and output channels respectively. "before" shows the immediate layer before a block and "after" gives the subsequent one directly. "link" explains the specific connections that lie in U-Net architectures [44] in which $\rightarrow$ decides the direction of connectivity, i.e., $\mathrm{Cv}_0 \rightarrow \mathrm{CvT}_{11}$ bridges the output of Cv_0 concatenated to the input of CvT_{11}. LReLU is short for Leaky ReLU activation [31] and BN is a abbreviation of Batch Normalization [17].

Network	Layer	Cv_0	Cv_1	Cv_2	Cv_3	Cv_4
D_1/D_2	#C_in	4/7	64	128	256	512
	#C_out	64	128	256	512	1
	before	–	LReLU	LReLU	LReLU	LReLU
	after	–	BN	BN	BN	Sigmoid

Table 3. The architectures for discriminator D_1/D_2 of ST-CGAN. Annotations are kept the same with Table 2.

plified to:

$$\mathcal{L}_{CGAN_1}(G_1, D_1) = \mathbf{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\log D_1(\mathbf{x},\mathbf{y})] + \mathbf{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1 - D_1(\mathbf{x}, G_1(\mathbf{x})))]. \quad (2)$$

Besides the adversarial loss, the classical data loss is adopted that encourages a straight and accurate regression of the target:

$$\mathcal{L}_{data_1}(G_1) = \mathbf{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}||\mathbf{y} - G_1(\mathbf{x})||. \quad (3)$$

Further in the second CGAN of Figure 2, by applying the similar formulations above, we have:

$$\mathcal{L}_{data_2}(G_2|G_1) = \mathbf{E}_{\mathbf{x},\mathbf{r}\sim p_{data}(\mathbf{x},\mathbf{r})}||\mathbf{r} - G_2(\mathbf{x}, G_1(\mathbf{x}))||, \quad (4)$$

$$\mathcal{L}_{CGAN_2}(G_2, D_2|G_1) = \mathbf{E}_{\mathbf{x},\mathbf{y},\mathbf{r}\sim p_{data}(\mathbf{x},\mathbf{y},\mathbf{r})}[\log D_2(\mathbf{x},\mathbf{y},\mathbf{r})] + \mathbf{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1 - D_2(\mathbf{x}, G_1(\mathbf{x}), G_2(\mathbf{x}, G_1(\mathbf{x}))))], \quad (5)$$

where $\mathbf{r}$ denotes for $\mathbf{x}$'s corresponding shadow-free image and G_2 takes a combination of $\mathbf{x}$ and $G_1(\mathbf{x})$ as inputs whereas D_2 differentiates the concatenation of outputs from G_1 and G_2, conditioned on $\mathbf{x}$, from the real pairs. Till this end, we can finally conclude the entire objective for the joint learning task which results in solving a mini-max problem where the optimization aims to find a saddle point:

$$\min_{G_1,G_2}\max_{D_1,D_2} \mathcal{L}_{data_1}(G_1) + \lambda_1\mathcal{L}_{data_2}(G_2|G_1) + \lambda_2\mathcal{L}_{CGAN_1}(G_1, D_1) + \lambda_3\mathcal{L}_{CGAN_2}(G_2, D_2|G_1). \quad (6)$$

It is regarded as a two-player zero-sum game. The first player is a team consisting of two generators (G_1, G_2). The second player is a team containing two discriminators (D_1, D_2). In order to defeat the second player, the members of the first team are encouraged to produce outputs that are close to their corresponding ground-truths.

4.2. Network Architecture and Training Details

Generator. The generator is inspired by the U-Net architecture [44], which is originally designed for biomedical image segmentation. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. The detailed structure of G_1/G_2, similar to [18], is listed in the Table 2.

Discriminator. For D_1, it receives a pair of images as inputs, composed of an original RGB scene image and a shadow mask image that generates 4-channel feature-maps as inputs. The dimensionality of channels increases to 7 for D_2 as it accepts an additional shadow-free image. Table 3 gives more details of these two discriminators.

Training/Implementation settings. Our code is based on pytorch [21]. We train ST-CGAN with the Adam solver [24] and an alternating gradient update scheme is applied. Specifically, we first adopt a gradient ascent step to update D_1, D_2 with G_1, G_2 fixed. We then apply a gradient descent step to update G_1, G_2 with D_1, D_2 fixed. We initialize all the weights of ST-CGAN by sampling from a zero-mean normal distribution with standard deviation 0.2. During training, augmentations are adopted by cropping (image size 286 $\rightarrow$ 256) and flipping (horizontally) operations. A practical setting for λ, where $\lambda_1 = 5, \lambda_2 = 0.1, \lambda_3 = 0.1$, is used. The Binary Cross Entropy (BCE) loss is assigned for the objective of image mask regression and L1 loss is utilized for the shadow-free image reconstruction respectively.

4.3. Discussion

The stacked term. The commonly used form of multi-task learning is the multi-branch version. It aims to learn a shared representation, which is further utilized for each task in parallel. But our stacked design differs quite a lot from it. We conduct the multi-task learning in such a way that each task can focus on its individual feature embeddings, instead of a shared embedding across tasks, whilst they still enhance each other through the stacked connections, in a

Using ISTD Train	Detection Aspects	StackedCNN [52]	cGAN [36]	scGAN [36]	ours
SBU [52] (%)	Shadow	11.29	24.07	9.1	9.02
	Non-shadow	20.49	13.13	17.41	13.66
	BER	15.94	18.6	13.26	11.34
UCF [59] (%)	Shadow	10.56	23.23	9.09	8.77
	Non-shadow	27.58	15.61	23.74	23.59
	BER	18.67	19.42	16.41	16.18
ISTD (%)	Shadow	7.96	10.81	3.22	2.14
	Non-shadow	9.23	8.48	6.18	5.55
	BER	8.6	9.64	4.7	3.85

Table 4. Detection with quantitative results using BER, smaller is better. For our proposed architecture, we use image triplets of ISTD training set. These models are tested on three datasets. The best and second best results are marked in red and blue colors, respectively.

Using SBU Train	Detection Aspects	StackedCNN [52]	cGAN [36]	scGAN [36]	ours
SBU [52] (%)	Shadow	9.6	20.5	7.8	3.75
	Non-shadow	12.5	6.9	10.4	12.53
	BER	11.0	13.6	9.1	8.14
UCF [59] (%)	Shadow	9.0	27.06	7.7	4.94
	Non-shadow	17.1	10.93	15.3	17.52
	BER	13.0	18.99	11.5	11.23
ISTD (%)	Shadow	11.33	19.93	9.5	4.8
	Non-shadow	9.57	4.92	8.46	9.9
	BER	10.45	12.42	8.98	7.35

Table 5. Detection with quantitative results using BER, smaller is better. For our proposed architecture, we use image pairs of SBU training set together with their roughly generated shadow-free images by Guo et al. [12] to form image triplets for training. The best and second best results are marked in red and blue colors, respectively.

form of a forward/backward information flow. The following experiments also confirm the effectiveness of our architecture on the two tasks, compared with the multi-branch one, which can be found in Table 8.

The adversarial term. Moreover, Conditional GANs (CGANs) are able to effectively enforce higher order consistencies, to learn a joint distribution of image pairs or triplets. This confers an additional advantage to our method, as we implement our basic component to be CGAN and perform a stacked input into the adversarial networks, when compared with nearly most of previous approaches.

5. Experiments

To comprehensively evaluate the performance of our proposed method, we perform extensive experiments on a variety of datasets and evaluate ST-CGAN in both detection and removal measures, respectively.

5.1. Datasets

We mainly utilize two large-scale publicly available datasets including SBU [52] and UCF [59], along with our newly collected dataset ISTD.

SBU [52] has 4727 pairs of shadow and shadow mask image. Among them, 4089 pairs are for training and the rest is for testing.

UCF [59] has 245 shadow and shadow mask pairs in total, which are all used for testing in the following experiments.

ISTD is our new released dataset consisting of 1870 triplets, which is suitable for multi-task training. It is randomly divided into 1330 for training and 540 for testing.

5.2. Compared Methods and Metrics

For detection part, we compare ST-CGAN with the state-of-the-art StackedCNN [52], cGAN [36] and scGAN [36]. To evaluate the shadow detection performance quantatively, we follow the commonly used terms [36] to compare the provided ground-truth masks and the predicted ones with the main evaluation metric, which is called Balance Error Rate (BER):

$$\mathrm{BER} = 1 - \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \qquad (7)$$

along with separated per pixel error rates per class (shadow and non-shadow).

For removal part, we use the publicly available source codes [12, 55, 8] as our baselines. In order to perform a quantitative comparison, we follow [12, 43] and use the root mean square error (RMSE) in LAB color space between the ground truth shadow-free image and the recovered image as measurement, and then evaluate the results on the whole

Dataset	Removal aspects	Original	Guo et al. [12]	Yang et al. [55]	Gong et al. [8]	ours
ISTD	Shadow	32.67	18.95	19.82	14.98	10.33
	Non-shadow	6.83	7.46	14.83	7.29	6.93
	All	10.97	9.3	15.63	8.53	7.47

Table 6. Removal with quantitative results using RMSE, smaller is better. The original difference between the shadow and shadow-free images is reported in the third column. We perform multi-task training on ISTD and compare it with three state-of-the-art methods. The best and second best results are marked in red and blue colors, respectively.

Task Type	Aspects	Ours	Ours (-D_1)	Ours (-D_2)	Ours (-G_1 -D_1)	Ours (-G_2 -D_2)
Removal	Shadow	10.33	10.36	10.38	12.12	–
	Non-shadow	6.93	6.96	7.03	7.45	–
	All	7.47	7.51	7.56	8.19	–
Detection (%)	Shadow	2.14	2.62	2.49	–	3.4
	Non-shadow	5.55	6.18	6.03	–	5.1
	BER	3.85	4.4	4.26	–	4.25

Table 7. Component analysis of ST-CGAN on ISTD by using RMSE for removal and BER for detection, smaller is better. The metrics related to shadow and non-shadow part are also provided. The best and second best results are marked in red and blue colors, respectively.

image as well as shadow and non-shadow regions separately.

5.3. Detection Evaluation

For detection, we utilize the cross-dataset shadow detection schedule, similar in [36], to evaluate our method. We first train our proposed ST-CGAN on the ISTD training set. The evaluations are thus conducted on three datasets with three state-of-the-art approaches in Table 4. As can be seen, ST-CGAN outperforms StackedCNN and cGAN by a large margin. In terms of BER, we obtain a significant 14.4% error reduction on SBU and 18.1% on ISTD respectively, compared to scGAN.

Next, we switch the training set to SBU's training data. Considering our framework requires image triplets that SBU cannot offer, we make an additional pre-processing step. In order to get the corresponding shadow-free image, we use the shadow removal code [12] to generate them as coarse labels. We also test these trained models on the three datasets. Despite the inaccurate shadow-free ground-truths, our proposed framework still significantly improves the overall performances. Specifically, on the SBU test set, ST-CGAN achieves an obvious improvement with 10.5% error reduction of BER over the previous best record from scGAN.

In Figure 5, we demonstrate the comparisons of the detection results qualitatively. As shown in Figure 5 (a) and 5 (b), ST-CGAN is not easily fooled by the lower brightness area of the scene, comparing to cGAN and scGAN. Our method is also precise in detecting shadows cast on bright areas such as the line mark in Figure 5 (c) and 5 (d). The proposed ST-CGAN is able to detect more fine-grained shadow details (e.g., shadow of leaves) than other methods, as shown in Figure 5 (e) and 5 (f).

5.4. Removal Evaluation

For removal, we compare our proposed ST-CGAN with the three state-of-the-art methods on ISTD dataset, as shown in Table 6. The RMSE values are reported. We evaluate the performance of different methods on the shadow regions, non-shadow regions, and the whole image. The proposed ST-CGAN achieves the best performance among all the compared methods by a large margin. Notably, the error of non-shadow region is very close to the original one, which indicates its strong ability to distinguish the non-shadow part of an image. The advantage of removal also partially comes from the joint learning scheme, where the well-trained detection block provides more clear clues of shadow and shadow-free areas.

We also demonstrate the comparisons of the removal results. As shown in Figure 5, although Yang [55] can recover shadow-free image, it alters the colors of both shadow and nonshadow regions. Guo [11] and Gong [8] fail to detect shadow accurately, thus both of their predictions are incomplete especially in shadow regions. Moreover, due to the difficulty of determining the environmental illuminations and global consistency, all the compared baseline models produce unsatisfactory results on the semantic regions.

5.5. Component Analysis of ST-CGAN

To illustrate the effects of different components of ST-CGAN, we make a series of ablation experiments by progressively removing different parts of it. According to both the removal and the detection performances in Table 7, we find that each individual component is necessary and indispensable for the final excellent predictions. Moreover, the last two columns of Table 7 also demonstrate that without the stacked joint learning, a single module consisting of one generator and one discriminator performs worse consistent-

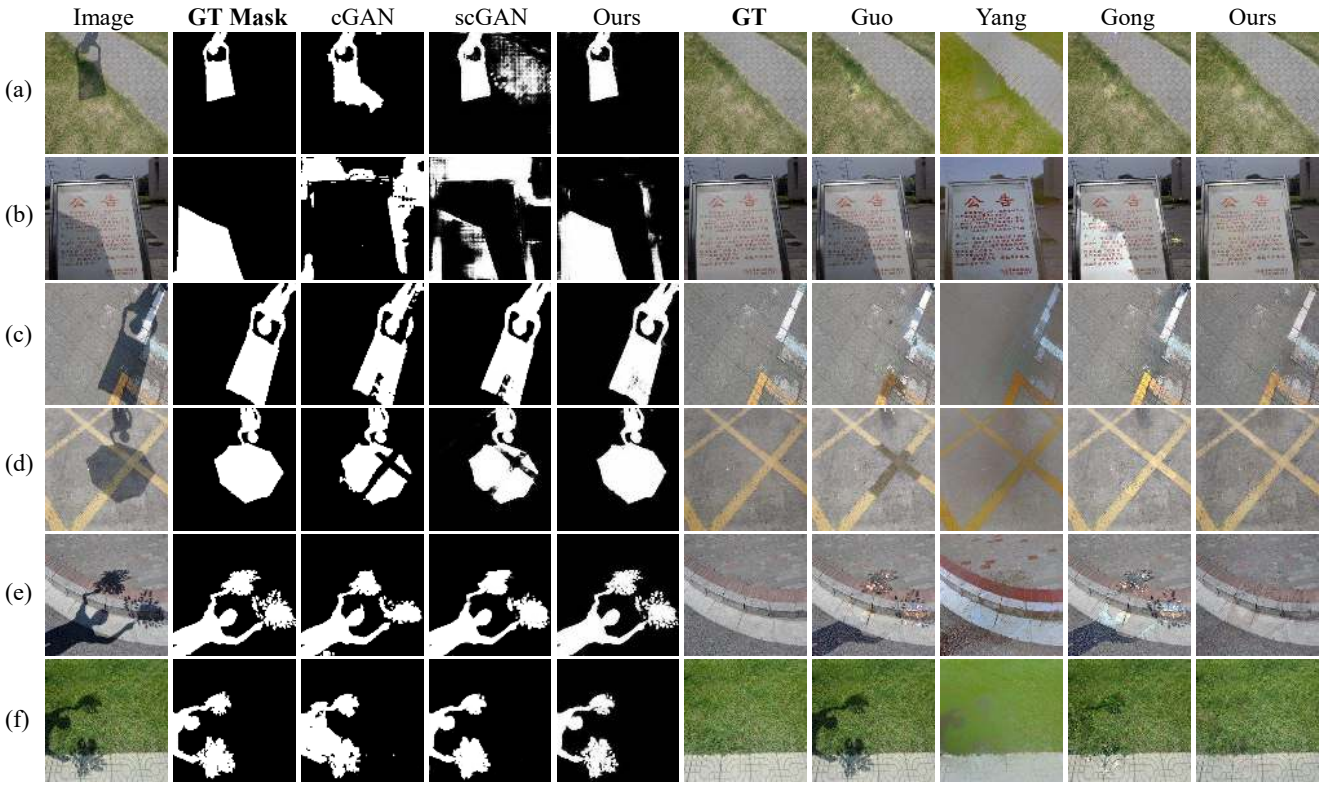


Figure 5. Comparison of shadow detection and removal results of different methods on ISTD dataset. Note that our proposed ST-CGAN simultaneously produces the detection and removal results, whilst others are either for shadow detection or for shadow removal.

Task Type	Aspects	Multi-branch	Ours
Removal	Shadow	11.54	**10.33**
	Non-shadow	7.13	**6.93**
	All	7.84	**7.47**
Detection (%)	Shadow	2.34	**2.14**
	Non-shadow	7.2	**5.55**
	BER	4.77	**3.85**

Table 8. Comparisons between stacked learning (ours) and multi-branch learning with removal and detection results on ISTD dataset.

ly. It further implies the effectiveness of our multi-task architecture on both shadow detection and shadow removal.

5.6. Stacked Joint *vs.* Multi-branch Learning

We further modify our body architecture into a multi-branch version, where each branch is designed for one task respectively. Therefore, the framework aims to learn a shared embedding which is supervised by two tasks. For a clear explanation, the illustration of comparisons between ours and the multi-branch one is also given. With all other training settings fixed, we fairly compare our proposed ST-CGAN with the multi-branch version quantitatively on the measurements of both detection and removal on ISTD dataset. Table 8 reports that our stacked joint learning paradigm consistently outperforms the multi-branch version in every single aspect of the metrics.

6. Conclusion

In this paper, we have proposed STacked Conditional Generative Adversarial Network (ST-CGAN) to jointly learn shadow detection and shadow removal. Our framework has at least four unique advantages as follows: 1) it is the first end-to-end approach that tackles shadow detection and shadow removal simultaneously; 2) we design a novel stacked mode, which densely connects all the tasks in the purpose of multi-task learning, that proves its effectiveness and suggests the future extension on other types of multiple tasks; 3) the stacked adversarial components are able to preserve the global scene characteristics hierarchically, thus it leads to a fine-grained and natural recovery of shadow-free images; 4) ST-CGAN consistently improves the overall performances on both the detection and removal of shadows. Moreover, as an additional contribution, we publicly release the first large-scale dataset which contains shadow, shadow mask and shadow-free image triplets.

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Science Fund of China under Grant Nos. U1713208 and 61472187, the 973 Program No.2014CB349303, and Program for Changjiang Scholars.

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 (TPAMI)*, 33(6):1202–1216, 2011. 1, 2, 3

[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(8):1670–1687, 2015. 3

[3] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti. Improving shadow suppression in moving object detection with hsv color information. In *IEEE Intelligent Transportation Systems (ITSC)*, pages 334–339, 2001. 1

[4] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[5] G. D. Finlayson, M. S. Drew, and C. Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision (IJCV)*, 85(1):35–57, 2009. 1, 3

[6] G. D. Finlayson, S. D. Hordley, and M. S. Drew. Removing shadows from images. In *European Conference on Computer Vision (ECCV)*, pages 823–836. Springer, 2002. 3

[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 (TPAMI)*, 28(1):59–68, 2006. 1, 3

[8] H. Gong and D. Cosker. Interactive shadow removal and ground truth for variable scene categories. In *British Machine Vision Conference (BMVC)*. University of Bath, 2014. 1, 2, 3, 6, 7

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 4

[10] M. Gryka, M. Terry, and G. J. Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics (TOG)*, 34(5):153, 2015. 1, 3, 4

[11] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2033–2040, 2011. 1, 3, 7

[12] R. Guo, Q. Dai, and D. Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2956–2967, 2013. 1, 3, 4, 6, 7

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017. 2, 3

[14] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016. 2

[15] X. Huang, G. Hua, J. Tumblin, and L. Williams. What characterizes a shadow boundary under the sun and sky? In *IEEE International Conference on Computer Vision (ICCV)*, pages 898–905, 2011. 1, 2, 3

[16] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie. Stacked generative adversarial networks. *arXiv preprint arXiv:1612.04357*, 2016. 3

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning(ICML)*, pages 448–456, 2015. 5

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 3, 5

[19] I. N. Junejo and H. Foroosh. Estimating geo-temporal location of stationary cameras using shadow trajectories. In *European conference on computer vision (ECCV)*, pages 318–331. Springer, 2008. 1

[20] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):157, 2011. 1

[21] N. Ketkar. Introduction to pytorch. In *Deep Learning with Python*, pages 195–208. Springer, 2017. 5

[22] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Automatic feature learning for robust shadow detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1946, 2014. 1, 2, 3

[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 and Machine Intelligence (TPAMI)*, 38(3):431–446, 2016. 1, 3

[24] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Estimating natural illumination from a single outdoor image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 183–190, 2009. 1

[26] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *European Conference on Computer Vision (ECCV)*, pages 322–335. Springer, 2010. 1, 2, 3

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 3

[28] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 702–716. Springer, 2016. 3

[29] F. Liu and M. Gleicher. Texture-consistent shadow removal. In *European Conference on Computer Vision (ECCV)*, pages 437–450. Springer, 2008. 1

[30] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017. 3

[31] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013. 5

[32] I. Mikic, P. C. Cosman, G. T. Kogut, and M. M. Trivedi. Moving shadow and object detection in traffic scenes. In *In-*

ternational Conference on Pattern Recognition (ICPR), volume 1, pages 321–324. IEEE, 2000. 1

[33] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2, 4

[34] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[35] A. Mohan, J. Tumblin, and P. Choudhury. Editing soft shadows in a digital photograph. *IEEE Computer Graphics and Applications*, 27(2):23–31, 2007. 3

[36] V. Nguyen, T. F. Yago Vicente, M. Zhao, M. Hoai, and D. Samaras. Shadow detection with conditional generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4510–4518, 2017. 2, 3, 6, 7

[37] T. Okabe, I. Sato, and Y. Sato. Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1693–1700, 2009. 1

[38] A. Panagopoulos, D. Samaras, and N. Paragios. Robust shadow and illumination estimation using a mixture model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 651–658, 2009. 1

[39] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios. Illumination estimation and cast shadow detection through a higher-order graphical model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 673–680, 2011. 1

[40] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios. Simultaneous cast shadows, illumination and geometry inference using hypergraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(2):437–449, 2013. 1

[41] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 3

[42] L. Qu, J. Tian, Z. Han, and Y. Tang. Pixel-wise orthogonal decomposition for color illumination invariant and shadow-free image. *Optics express*, 23(3):2220–2239, 2015. 3

[43] 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)*, 2017. 1, 2, 3, 4, 6

[44] 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 (MICCAI)*, pages 234–241. Springer, 2015. 5

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[46] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[47] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(9):1459–1472, 2005. 1, 3

[48] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv preprint arXiv:1612.07695*, 2016. 3

[49] Y. Vicente, F. Tomas, M. Hoai, and D. Samaras. Leave-one-out kernel optimization for shadow detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3388–3396, 2015. 1, 2, 3

[50] Y. Vicente, F. Tomas, M. Hoai, and D. Samaras. Noisy label recovery for shadow detection in unfamiliar domains. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3783–3792, 2016. 1, 2, 3

[51] Y. Vicente, F. Tomas, M. Hoai, and D. Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, PP(99):1–1, 2017. 1

[52] Y. Vicente, F. Tomas, 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 (ECCV)*, pages 816–832. Springer, 2016. 3, 4, 6

[53] W. Wang, X. Li, J. Yang, and T. Lu. Mixed link networks. *arXiv preprint arXiv:1802.01808*, 2018. 2

[54] T.-P. Wu, C.-K. Tang, M. S. Brown, and H.-Y. Shum. Natural shadow matting. *ACM Transactions on Graphics (TOG)*, 26(2):8, 2007. 1, 2

[55] Q. Yang, K.-H. Tan, and N. Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image Processing (TIP)*, 21(10):4361–4368, 2012. 3, 6, 7

[56] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[57] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016. 3

[58] L. Zhang, Q. Zhang, and C. Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing (TIP)*, 24(11):4623–4636, 2015. 1

[59] J. Zhu, K. G. Samuel, S. Z. Masood, and M. F. Tappen. Learning to recognize shadows in monochromatic natural images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 223–230, 2010. 1, 2, 3, 4, 6

[60] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 3