A+D Net: Training a Shadow Detector with Adversarial Shadow Attenuation

Hieu Le[1(✉)], Tomas F. Yago Vicente[1,2], Vu Nguyen[1], Minh Hoai[1], and Dimitris Samaras[1]

[1] Stony Brook University, Stony Brook, NY 11794, USA
{hle,tyagovicente,vhnguyen,minhhoai,samaras}@cs.stonybrook.edu
[2] Amazon/A9, Palo Alto, USA

Abstract. We propose a novel GAN-based framework for detecting shadows in images, in which a shadow detection network (D-Net) is trained together with a shadow attenuation network (A-Net) that generates adversarial training examples. The A-Net modifies the original training images constrained by a simplified physical shadow model and is focused on fooling the D-Net's shadow predictions. Hence, it is effectively augmenting the training data for D-Net with hard-to-predict cases. The D-Net is trained to predict shadows in both original images and generated images from the A-Net. Our experimental results show that the additional training data from A-Net significantly improves the shadow detection accuracy of D-Net. Our method outperforms the state-of-the-art methods on the most challenging shadow detection benchmark (SBU) and also obtains state-of-the-art results on a cross-dataset task, testing on UCF. Furthermore, the proposed method achieves accurate real-time shadow detection at 45 frames per second.

Keywords: Shadow detection · GAN · Data augmentation

1 Introduction

Shadows occur frequently in natural scenes, and can hamper many tasks such image segmentation, object tracking, and semantic labeling. Shadows are formed in complex physical interactions between light sources, geometry and materials of the objects in the scene. Information about the physical environment such as sparse 3D scene reconstructions [33], rough geometry estimates [22], and multiple images of the same scene under different illumination conditions [25] can aid shadow detection. Unfortunately, inferring the physical structure of a general scene from a single image is still a difficult problem.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01216-8_41) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11206, pp. 680–696, 2018.
https://doi.org/10.1007/978-3-030-01216-8_41

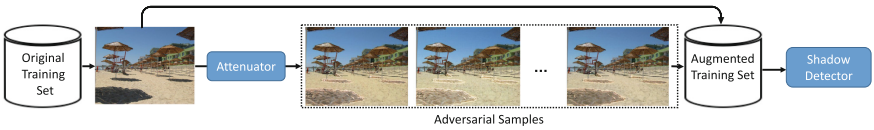


Fig. 1. Adversarial shadow attenuation. The attenuator takes an original shadow image and generates different adversarial shadow samples to train the shadow detector.

The difficulty of shadow detection is exacerbated when dealing with consumer-grade photographs and web images [15]. Such images often come from non-linear camera sensors, and present many compression and noise artifacts. In this case, it is better to train and use appearance-based classifiers [7,13,27,36] rather than relying on physical models of illumination [4,5]. Shadow classifiers, however, require annotated training data, and the performance of a classifier often correlates with the amount of training data. Unfortunately, annotated shadow data is expensive to collect and label. Only recently available training data has increased from a few hundred images [7,36] to a few thousands [30] thus enabling training more powerful shadow classifiers based on deep convolutional neural networks [20,30]. Nevertheless, even a few thousand images is a tiny amount compared to datasets that have driven progress in deep learning [2,16]. It is therefore safe to assume that the performance of deep learning shadow classifiers has not saturated yet, and it can be improved with more training data. Unfortunately, collecting and annotating shadow data is a laborious process. Even a lazy annotation approach [28] takes significant effort; the annotation step itself takes 20 s per image, not including data collection and cleansing efforts.

In this paper, instead of collecting additional data, we propose a method to increase the utility of available shadow data to the fullest extent. The main idea is to generate a set of augmented training images from a single shadow image by weakening the shadow area in the original training image. We refer to this process as shadow attenuation and we train a deep neural network to do so, called A-Net. This network modifies original shadow images so as to weaken the shadow effect, as illustrated in Fig. 1. The generated images serve as additional challenging training samples for a shadow detector D-Net. We present a novel framework, where the shadow attenuator and the shadow detector are trained jointly in an adversarial manner. The output of the attenuation model A-Net provides adversarial training samples with harder-to-detect shadow areas to improve the overall reliability of the detector D-Net.

Recent research also suggests that deep networks are highly sensitive to adversarial perturbations [19,26,34]. By jointly training A-Net and D-Net, we directly enhance the resistance of the detector D-Net to adversarial conditions and improve the generalization of the detector, following the recent trend [3,31,35].

Essentially, what is being proposed here is a data augmentation method for shadow detection. It is different from other data augmentation methods, and

it does not suffer from two inherent problems of general data augmentation approaches, which are: (1) the augmented data might be very different from the real data, having no impact on the generalization ability of the trained classifier on real data, and (2) it is difficult to ensure that the augmented data samples have the same labels as the original data, and this leads to training label noise. A popular approach to address these problems is to constrain the augmented data samples to be close to the original data, e.g., setting an upper bound for the L_2 distance between the original sample and the generated sample. However, it is difficult to set the right bound; a big value would create label noise while a small value would produce augmented samples that are too similar to the original data, yielding no benefit. In this paper, we address these two problems in a principled way, specific to shadow detection. Our idea is to use a physics model of shadows and illumination to guide the data generation process and to estimate the probability of having label noise.

Note that we aim to attenuate the shadow areas, not to remove them. Shadow removal is an important problem, but training a good shadow removal network would require many training pairs of corresponding shadow/shadow-free images, which are not available. Furthermore, completely removed shadows would correspond to having label noise, and this might hurt the performance of the detector.

Experimental results show that our shadow detector outperforms the state-of-the-art methods in the challenging shadow detection benchmark SBU [30] as well as on the cross-dataset task (training on SBU and testing on the UCF dataset [36]). Furthermore, our method is more efficient than many existing ones because it does not require a post-processing step such as patch averaging or conditional random field (CRF) smoothing. Our method detects shadows at 45 frames per second for 256×256 input images.

2 Related Work

Single image shadow detection is a well studied problem. Earlier work focused on physical modeling of illumination [4,5]. These methods render illumination invariant representations of the images where shadow detection is trivial. These methods, however, only work well for high quality images taken with narrow-band sensors [15]. Another early attempt to incorporate physics based constraints with rough geometry was the approach of Panagopoulos *et al.* [21] where the illumination environment is modeled as a mixture of von Mises-Fisher distributions [1] and the shadow pixels are segmented via a graphical model. Recently, data-driven approaches based on learning classifiers [8,11,13,27] from small annotated datasets [7,36] have shown more success. For instance, Vicente *et al.* [27,29] optimized a multi-kernel Least-Squares SVM based on leave-one-out estimates. This approach yielded accurate results on the UCF [36] and UIUC [7] datasets, but its underlying training procedure and optimization method cannot handle a large amount of training data.

To handle and benefit from a large amount of training data, recent shadow detection methods have been developed based on the stochastic gradient descent

training of deep neural networks. Vicente *et al.* [30] proposed a stacked-CNN architecture, combining an image-level Fully Convolution Neural Network (FCN) with a patch-CNN. This approach achieved good detection results, but it is cumbersome as the Fully Connected Network (FCN) has to be trained before its predictions are used to train the patch-CNN. Similarly, testing was computationally expensive as it requires the FCN prediction followed by predictions of densely sampled patches covering the testing image. Recently, Nguyen *et al.* [20] presented scGAN, a method based on Generative Adversarial Networks (GANs) [6]. They proposed a parametric conditional GAN [17] framework, where the generator was trained to generate the shadow mask, conditioned on an input RGB patch and a sensitivity parameter. To obtain the final shadow mask for an input image, the generator must be run on multiple image patches at multiple scales and the outputs are averaged. Their method achieved good results on the SBU dataset, but the detection procedure was computationally expensive at test time. Our proposed method also uses adversarial training for shadow detection, but it is fundamentally different from scGAN. scGAN uses the generator to generate a binary *shadow mask* conditioned on the input image, while our method uses the generator to generate augmented training images in RGB space. Furthermore, while scGAN uses the discriminator as a regulator to encourage global consistency, the discriminator in our approach plays a more prominent role for shadow pixel classification. In contrast to scGAN, our method does not require post processing or output averaging, leading to real-time shadow detection. Another method that uses GAN for shadow detection is Stacked Conditional GAN [32]. This method, however, requires the availability of shadow-free images. Another recent approach [10] proposes to use contextual information for a better shadow detection. Contextual information is incorporated by having several spatial-directional recurrent neural networks. While this method yields excellent results on shadow detection benchmarks, it also requires running a CRF as a post-processing step.

We propose a method to improve shadow detection with augmented training examples, in sync with recent trends on data augmentation. For example, Zhang *et al.* [35] proposed a simple augmentation method by enriching the dataset with the linear combinations of pairs of examples and their labels to improve the generalization of the network and its resistance toward adversarial examples. Another approach that used adversarial examples for training a network was proposed by Shrivastava *et al.* [24]. They adversarially trained a Refiner network that inputs synthetic examples and outputs more realistic images. The refined examples can be used as additional training data. In a similar way, our proposed Attenuator (A-Net) takes original training images and generates realistic images with attenuated shadows that act as additional training examples for our shadow detector. The generation of adversarial examples is an integral part of the joint training process with the detector (D-Net), in contrast to [24] where the generated data is a preprocessing step to enrich the training set. The effects of the shadow Attenuator can also be seen as related to adversarial perturbations [18]: A-Net modifies the input images so as to fool the predictions of the

shadow detector D-Net. Adversarial examples also can be used to improve the generalization of the network for domain adaptation [31] in which a conditional GAN is used to perform feature augmentation.

3 Adversarial Training and Attenuation

3.1 Framework Overview

We present a novel framework for shadow detection based on adversarial training and shadow attenuation. Our proposed model contains two jointly trained deep networks. Figure 2 illustrates the flow diagram of our framework. The shadow attenuation network, called Attenuator or A-Net, takes as input a shadow image and its corresponding shadow mask. Based on these inputs, the Attenuator generates a version of the input image where the shadows have been attenuated. Attenuation can be thought of as partial shadow removal. The image generated by the Attenuator is fed into a shadow detection network, called Detector or D-Net, which predicts the shadow areas. On each training iteration, D-Net also takes the original input image, and learns to predict the corresponding annotated ground-truth shadow mask.

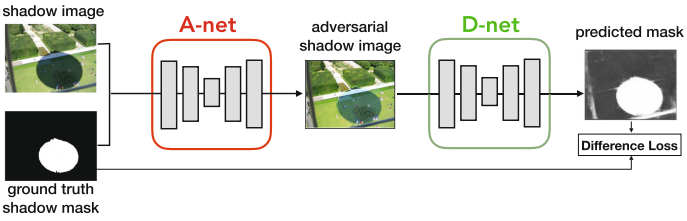


Fig. 2. Adversarial training of a shadow detector. A-Net takes a shadow image and its corresponding shadow mask as input, and generates an adversarial example by attenuating the shadow regions in the input image. The attenuated shadows are less discernible and therefore harder to detect. D-Net takes this image as input and aims to recover the original shadow mask.

A-Net is trained to attenuate shadow regions so as to fool the shadow detector. In particular, for pixels inside the provided shadow mask, A-Net manipulates the values of the pixels to disguise them as non-shadow pixels so that they cannot be recognized by D-Net. We further constrain the attenuation transformation using a loss that incorporates physics-inspired shadow domain knowledge. This enhances the quality of the generated pixels, improving the generalizability of the detector. At the same time, A-Net learns not to change the values or the pixels outside the shadow mask. We enforce this with a loss that penalizes the difference between the generated image and the input image on the area outside of the shadow mask (non-shadow pixels). The adversarial training process with all the aforementioned constraints and the back propagation error from the shadow detection network guides A-Net to perform shadow attenuation.

The detector network, D-Net, takes the adversarial examples generated by A-Net and predicts shadow masks. Shadow areas in the images generated by A-Net are generally harder to detect than in the input images, since A-Net is trained to attenuate the shadows to fool D-Net. As a result, D-Net is trained with challenging examples in addition to the original training examples. As D-Net improves its ability to detect shadows, A-Net also improves its ability to attenuate shadows to confound D-Net with tougher adversarial examples. This process strengthens the shadow detection ability of D-Net.

3.2 Physics-Based Shadow and Illumination Model

We use a physics-based illumination model to guide the data generation process and avoid label noise. We use the simplified illumination model used by Guo *et al.* [7,8] where, each pixel is lit by a combination of direct and environment lights: $I_i = (k_i L_d + L_e)\, R_i$, where I is an image and I_i denotes the color of the i^{th} pixel of the image. R_i is the surface reflectance corresponding to the i^{th} pixel. L_d and L_e are 3×1 vectors representing the colors and intensities of the direct light and the environment light (which models area sources and inter reflections), respectively. $k_i \in [0, 1]$ is the shadowing factor that indicates how much of the direct light reaches the pixel i. k_i remains close to 0 for the umbra region of the shadow, while it gets increasingly close to 1 in the penumbra region. For pixels inside shadow-free areas $k_i = 1$. We can relate the original shadow region and its corresponding shadow-free version by the ratio:

$$\frac{I_i^{\text{shadow-free}}}{I_i^{\text{shadow}}} = \frac{L_d + L_e}{k_i L_d + L_e}.$$

By taking the ratio between the shadow-free and in-shadow values, we have eliminated the unknown reflectance factor. We assume that the direct light is constant over the scene depicted by the image, and the effects of the environment light are similar for all pixels. We incorporate this model into the training process of both A-Net and D-Net:

- **A-Net:** We design the *physics loss* to enforce the illumination ratios for pixels inside an attenuated shadow area to have a small variance.
- **D-Net:** We directly estimate the illumination ratio between the areas inside and outside the shadow mask to measure shadow strength in the attenuated images to avoid training label noise.

3.3 A-Net: Shadow Attenuator Network

The shadow attenuator network A-Net is trained to re-illuminate only the shadow areas so that they cannot be detected by the detector network D-Net. To obtain useful and realistic attenuated shadows, A-Net aims to fool D-Net while respecting a physical illumination model. Figure 3 shows the training process of A-Net, which attenuates shadow areas under the following constraints and

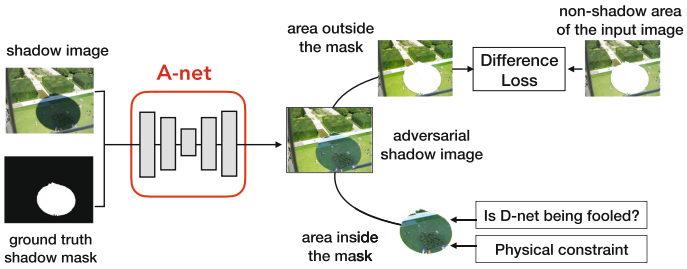


Fig. 3. A-Net. The area outside the shadow mask is constrained by the difference loss with respect to the input image. The area inside the shadow mask is constrained by the feedback from D-Net and the physics based constraint.

objectives: (1) Values of non-shadow pixels are preserved. (2) Shadow pixels are re-illuminated such that D-Net cannot recognize them as shadow pixels. (3) The resulting pixel transformation obeys physics-inspired illumination constraints.

These constraints and objectives can be incorporated in the training of A-Net by defining a proper loss function. Let I denote an input image, and $M(I)$ be the shadow mask of I. Let $A(I)$ denote the output of A-Net for the input pair of I and $M(I)$ (here we write $A(I)$ as the short form for $A(I, M(I))$). Let $D(I)$ denote the output of D-Net for an input image I, i.e. the predicted shadow mask. Ideally, the output should be 1 for shadow pixels and 0 otherwise. The objective of A-Net's training is to minimize a weighted combination of three losses:

$$\mathcal{L}_A(I) = \lambda_{nsd}\mathcal{L}_{nsd}(I) + \lambda_{sd}\mathcal{L}_{sd}(I) + \lambda_{ph}\mathcal{L}_{ph}(I), \tag{1}$$

where $\mathcal{L}_{nsd}$ is the loss that penalizes the modification of values for pixels outside the shadow mask $M(I)$ for the input image I: $\mathcal{L}_{nsd}(I) = \text{mean}_{i \notin M(I)} \|A(I)_i - I_i\|_1$. $\mathcal{L}_{sd}$ is the adversarial loss. It penalizes the correct recognition of D-Net for shadow pixels on the generated image, restricted to the area inside the training shadow mask $M(I)$: $\mathcal{L}_{sd}(I) = \text{mean}_{i \in M(I)}[D(A(I))_i]$. $\mathcal{L}_{ph}$ is a physics-inspired loss to ensure that the shadow area in the generated image is re-illuminated in a physically feasible way. Based on the illumination model described in Sect. 3.2, we want the ratio $\frac{A(I)_i}{I_i}$ to be similar for all pixels i inside a re-illuminated shadow area. We model this by adding a loss term for the variance of the log ratios

$$\mathcal{L}_{ph}(I) = \sum_{c \in \{R,G,B\}} \underset{i \in M(I)}{\text{Variance}} \left[\log(A(I)_i^c) - \log(I_i^c)\right].$$

where $(\cdot)^c$ denotes the pixel value in the color channel c of the RGB color image.

Figure 4 shows some examples of attenuated shadows that were generated by A-Net during the adversarial training process. The two original input images contain easy to detect shadows with strengths 3.46 and 2.63. The heuristic to measure these shadow strength values are described in Sect. 3.4. The outputs of A-Net given these input images and shadow masks are shown in columns (c, d, e),

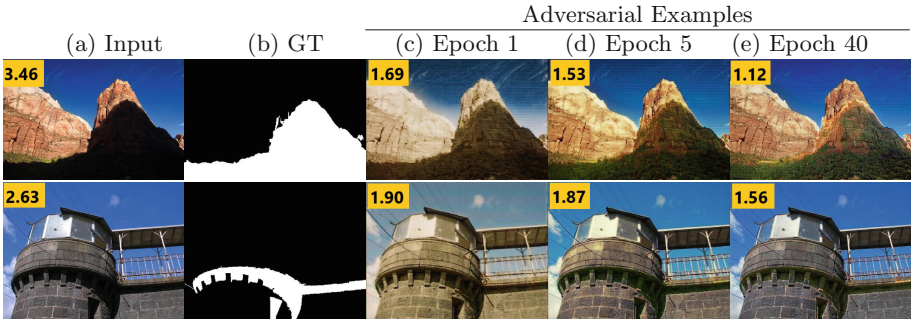


Fig. 4. Examples of attenuated shadows. (a) Input image. (b) Ground truth shadow mask. (c, d, e): adversarial examples with attenuated shadows generated by A-Net from epoch 1, 5, and 40 respectively. The corresponding *shadow strength* are shown as black text on the top-left corner of each image.

obtained at epochs 1, 5, and 40 during training. The shadows in the generated images become harder to detect as training progresses. Numerically, the shadow strength of the attenuated shadows decreases over time. Moreover, A-Net also learns to not change the non-shadow areas.

3.4 D-Net: Shadow Detector Network

The D-Net is central to our framework. It learns to detect shadows from adversarial examples generated by the A-Net as well as original training examples. On each training iteration, both the original input and the adversarially attenuated image are used to train D-Net. The learning objective for D-Net is to minimize the following loss function:

$$\mathcal{L}_D(I) = \lambda_{real} \left\| D(I) - M(I) \right\|_1 + \lambda_{adv}(A(I)) \left\| D(A(I)) - M(I) \right\|_1, \qquad (2)$$

where λ_{real} and $\lambda_{adv}(A(I))$ control how much D-Net should learn from the real sample I and the adversarial example $A(I)$ respectively. $\lambda_{adv}(A(I))$ depends on how much the shadow in I has been attenuated. If $A(I)$ is the completely shadow-free version of I, $\lambda_{adv}(A(I))$ should ideally be zero. Otherwise, this loss function corresponds to having label noise as it requires the output of the shadow detector D-Net for the input $A(I)$ to be the same as the shadow mask $M(I)$, while $A(I)$ is a shadow-free image.

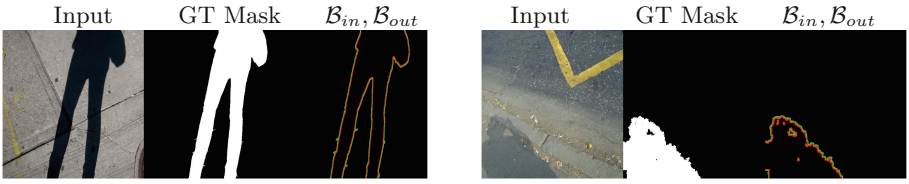
To determine if $A(I)$ is a shadow-free image, we derive a heuristic based on the illumination model described in Sect. 3.2. We first define two areas alongside the shadow boundary, denoted as $\mathcal{B}_{in}$ and $\mathcal{B}_{out}$, illustrated in Fig. 5. $\mathcal{B}_{out}$ (green) is the area right outside the boundary, computed by subtracting the shadow mask from its dilated version. The inside area $\mathcal{B}_{in}$ (red) is computed similarly with the eroded shadow mask. We define the *shadow strength* $k_{strength}$ as the ratio of average pixel intensities of the two boundary areas: $k_{strength}(A(I)) = \frac{\text{mean}_{i \in \mathcal{B}_{out}}[A(I)_i]}{\text{mean}_{i \in \mathcal{B}_{in}}[A(I)_i]}$. Figure 5 shows two examples of images with

two different shadow strengths; an image with a darker shadow (relative to the non-shadow area) has a higher value of $k_{strength}$ and vice versa.

We use the shadow strength of the attenuated image to decide if D-Net should learn from the attenuated shadow image. Heuristically, the shadow might be completely removed if the shadow strength $k_{strength}$ is too close to 1, i.e., the two areas on the two sides of the shadow boundary have the same average intensities. Based on this heuristic, we set the weight for the adversarial example $A(I)$ as follows:

$$\lambda_{adv}(A(I)) = \begin{cases} \lambda^0_{adv} & \text{if } k_{strength}(A(I)) > 1 + \epsilon \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where λ^0_{adv} is a tunable baseline factor for adversarial examples and ϵ is a small threshold which we empirically set to 0.05.



(a) Strong shadow ($k_{strength} = 4.16$) (b) Weak shadow ($k_{strength} = 1.15$)

Fig. 5. Estimating the shadow strength. From the ground-truth shadow mask, we define two area $\mathcal{B}_{in}$ (red) and $\mathcal{B}_{out}$ (green) obtained by dilation and erosion of the shadow mask. The shadow strength $k_{strengh}$ is computed as the ratio between the average intensity of pixels in $\mathcal{B}_{out}$ over the average intensity of pixels in $\mathcal{B}_{in}$. (a) an image with very a strong dark shadow, $k_{strength} = 4.16$. (b) light shadow $k_{strength} = 1.15$. (Color figure online)

3.5 Network Architectures

Both A-Net and D-Net were developed based on the U-Net architecture [23]. Following [12], we created networks with seven skip-connection modules, each of which contains a sequence of Convolutional, BatchNorm, and Leaky-ReLu [9] layers. The A-Net input is a four channel image, which is the concatenation of the RGB image and the corresponding shadow mask. The A-Net output is a three channel RGB image. The input to D-Net is an RGB image, and the output is a single channel shadow mask.

4 Experiments and Results

We experiment on several public shadow datasets. One of them is the SBU Shadow dataset [30]. This dataset consists of pairs of RGB images and corresponding annotated shadow binary masks. The SBU dataset contains 4089

training images, and 638 testing images, and is currently the largest and most challenging shadow benchmark. We also perform cross-dataset experiments on the UCF testing set [36], which contains 110 images with corresponding shadow masks. We quantitatively evaluate shadow detection performance by comparing the testing ground-truth shadow masks with the prediction masks produced by D-Net. As is common practice in the shadow detection literature, we will use the Balanced Error Rate (BER) as the principal evaluation metric. The BER is defined as: $\text{BER} = 1 - \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$, where TP, TN, FP, FN are the total numbers of true positive, true negative, false positive, and false negative pixels respectively. Since natural images tend to overwhelmingly more non-shadow pixels, the BER is less biased than mean pixel accuracy. We also provide separate mean pixel error rates for the shadow and non-shadow classes.

Training and Implementation Details. We use stochastic gradient descent with the Adam solver [14] to train our model. We use mini batch SGD with batch size of 64. On each training iteration, we perform three forward passes consecutively: forward the input shadow image I to A-Net to get the adversarial example $A(I)$, then separately forward the adversarial image and shadow input image to D-Net. We alternate one parameter update step on D-Net with one update step on A-Net, as suggested by [6]. Before training and testing, we transform the images into log-space. We experimentally set our training parameters as: $(\lambda_{nsd}, \lambda_{sd}, \lambda_{ph}, \lambda_{real}, \lambda^0_{adv}) := (30, 1, 100, 0.8, 0.2)$. We implemented our framework on PyTorch. More details can be found at: www3.cs.stonybrook.edu/~cvl/projects/adnet/index.html.

4.1 Shadow Detection Evaluation

We evaluate the shadow detection performance of the proposed D-Net on the SBU and UCF datasets. To detect shadows in an image, we first resize the image to 256×256. We input this image to D-net to produce a shadow mask of size 256×256, which will be compared with the ground-truth shadow mask for evaluation (in the original size).

In Table 1, we compare the performance of our method with the state-of-the-art methods Stacked-CNN [30], scGAN [20], ST-CGAN [32], and DSC [10]. We also consider a variant of D-Net, trained without the attenuated shadow images from A-Net. All methods are trained on the SBU training set. Performance is reported in terms of BER, as well as shadow and non-shadow error rates. Note that DSC [10] only reported BER numbers on the SBU dataset and its cross-domain results were obtained on testing data that is different from the commonly used UCF test dataset (as proposed by [36]).

On the SBU test set, our detector (D-Net) outperforms the previous state-of-the-art methods. Compared to the Stacked-CNN we obtain a 51% error reduction. Compared to scGAN and ST-CGAN, D-Net brings a 41% error reduction and a 33% error reduction respectively. D-Net outperforms DSC by 0.2% BER, even though it is significantly simpler. D-Net is fully convolutional, without the need of for running recurrent neural networks and CRF post processing.

For the cross-dataset experiments, the detectors are trained on the SBU training set, but they are evaluated on the test set of the UCF dataset [36]. These datasets are disjoint; while SBU covers a wide range of scenes, UCF focuses on images where dark shadows as well dark albedo objects are present. Again, we compare our method with the previous state-of-the-art methods: Stacked-CNN [30], scGAN [20], and ST-CGAN [32]. In terms of BER, our proposed D-Net yields significant error reductions of 18% and 16% with respect to scGAN and ST-CGAN, respectively. The performance gap between D-Net trained with and without attenuated shadow images is very significant, highlighting the benefits of having attenuated shadow examples for training.

Table 1. Evaluation of shadow detection methods on the SBU Shadow dataset [30] and for cross-dataset detection on UCF [36]. All methods are trained on the SBU training data. Both Balanced Error Rate (BER) and per class error rates are shown. DSC [10] only reported BER numbers, and used a different UCF test dataset, so cross-domain performance cannot be compared. Best performances is printed in bold.

Method	Evaluated on SBU testset [30]			Evaluated on UCF testset [36]		
	BER	Shadow	Non shad.	BER	Shadow	Non shad.
stacked-CNN [30]	11.0	9.6	12.5	13.0	9.0	17.1
scGAN [20]	9.1	7.8	10.4	11.5	7.7	15.3
ST-CGAN [32]	8.1	**3.7**	12.5	11.2	**5.0**	17.5
DSC [10]	5.6	–	–	–	–	–
D-Net (w/o A-Net)	8.8	8.1	9.3	11.8	8.9	14.7
D-Net (with A-Net)	**5.4**	5.3	**5.5**	**9.4**	7.0	**11.8**

4.2 Qualitative Results

In Fig. 6(i) and (ii), we show shadow detection results on the SBU dataset. The columns show input images, ground truth shadow masks, and D-Net outputs, respectively. In Fig. 6(i), we see how the D-Net correctly predicts shadows on different types of scenes such as desert, mountain, snow, and under different weather conditions from sunny to cloudy and overcast. In Fig. 6(ii), notice how the D-Net accurately predicts shadows in close-ups as well as long-range shots, and in aerial images. Figure 7 shows qualitative comparisons with the shadow detection results of scGAN [20]. In general, D-Net produces more accurate shadows with sharper boundaries.

4.3 Failure Cases

Some failure cases of our method are shown in Fig. 8. Many are due to dark albedo material regions being incorrectly classified as shadows. We also investigate the locations of wrongly classified pixels to understand the causes of failure.

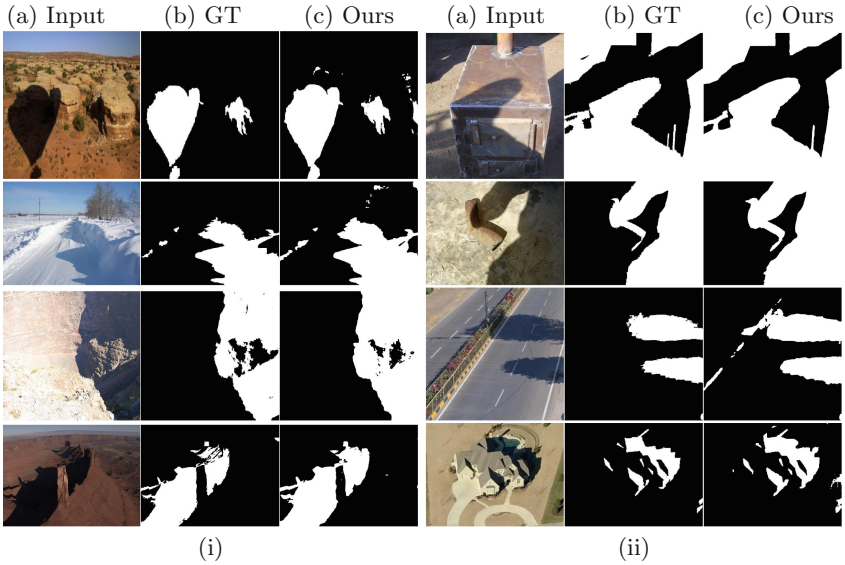


Fig. 6. Shadow detection results. Our proposed method accurately detects shadows on: (i) different scenes, and illumination conditions; (ii) close-ups and long-range shots, as well as aerial images.

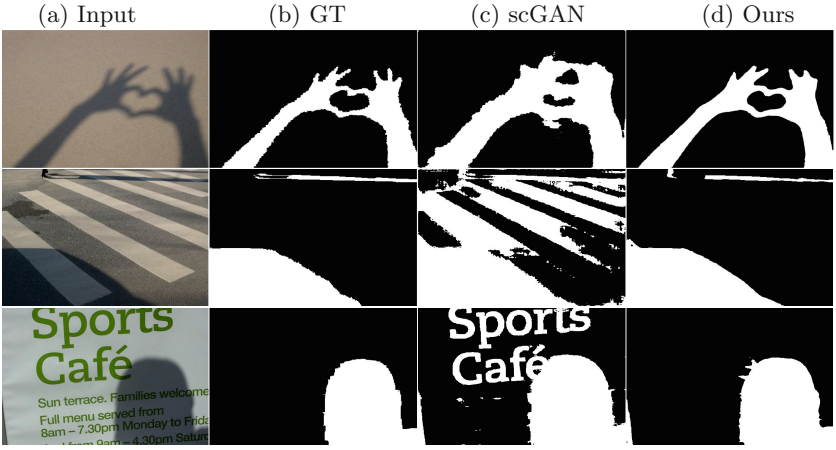


Fig. 7. Comparison of shadow detection on SBU dataset. Qualitative comparison between our method and the state-of-the-art method scGAN [20]. (a) Input image. (b) Ground-truth shadow mask. (c) Predicted shadow mask by scGAN [20]. (d) Predicted shadow mask by our method.

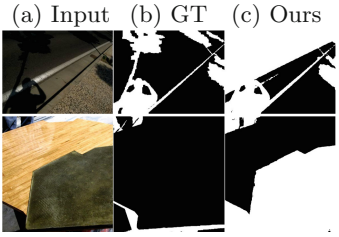


Fig. 8. Failed shadow detection examples. Failure cases of our method due to non-shadow dark albedo regions. (a) Input image. (b) Ground-truth mask. (c) Predicted shadow mask by our method.

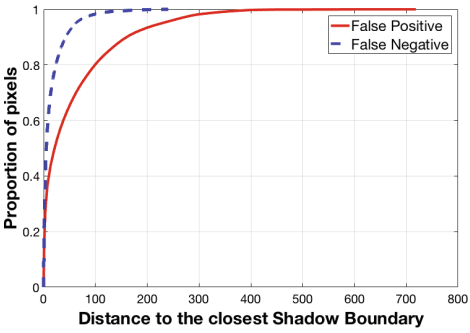


Fig. 9. Cumulative curve of the distance of wrongly predicted pixels to the closest shadow boundary on the SBU testing set.

Figure 9 shows the proportion of wrongly predicted pixels with respect to their distances to the closest ground-truth shadow boundary on the SBU testing set. A large portion of missed shadow pixels is within a small distance to a boundary. Specifically, 65% of false negative cases are within 10 pixels of a shadow boundary. This means the shadow pixels missed by our method are probably either around the shadow boundaries or inside very small shadow regions. Meanwhile a large portion of false positive prediction is far away from a shadow boundary. This is perhaps due to the misclassifications of dark objects as shadows.

4.4 Ablation Study and Parameter Analysis

We conducted experiments to analyze the impact of the physics-based loss (L_{ph}) and the weight function λ_{adv} in our framework. We trained our model with two additional scenarios for comparison: (1) without the physics-based loss and without the weight function λ_{adv}, and (2) with the physics-based loss but without the weight function λ_{adv}. We denote these two configurations as ($-L_{ph}$, $-\lambda_{adv}$) and ($+L_{ph}$, $-\lambda_{adv}$) respectively. Table 2 shows the shadow detection results of the models trained with these modified conditions. We tested the models, trained on SBU, on both the UCF and SBU testing sets. As can be seen from Table 2, dropping the weight function λ_{adv} increased error rates slightly, while dropping the physics-based loss drastically increased error rates. In Fig. 10, we compare adversarial examples generated by the model trained with and without the physics-based loss. Incorporating this loss produces images with more realistic attenuated shadows. Thus, the produced examples aid the training of the shadow detector D-Net. In our experiments, at the 50^{th} training epoch, approximately 6% of all images generated by A-Net, were not used based on λ_{adv}.

Table 2. Ablation study. Comparison of shadow detection results of our framework with and without inclusion of the physics based loss L_{ph}. Detection performance significantly profits from incorporating the physics based loss L_{ph} into the training process: 20% reduction of BER in SBU [30] testing set, and 27% error reduction in UCF [36] (cross-dataset task)

Method	Evaluated on SBU testset			Evaluated on UCF testset		
	BER	Shadow	Non shad.	BER	Shadow	Non shad.
D-Net ($+L_{ph}, +\lambda_{adv}$)	**5.4**	**5.3**	5.5	**9.4**	**7.0**	**11.8**
D-Net ($+L_{ph}, -\lambda_{adv}$)	5.7	6.2	**5.2**	9.9	7.3	12.5
D-Net ($-L_{ph}, -\lambda_{adv}$)	7.1	7.6	6.7	13.6	15.9	11.3

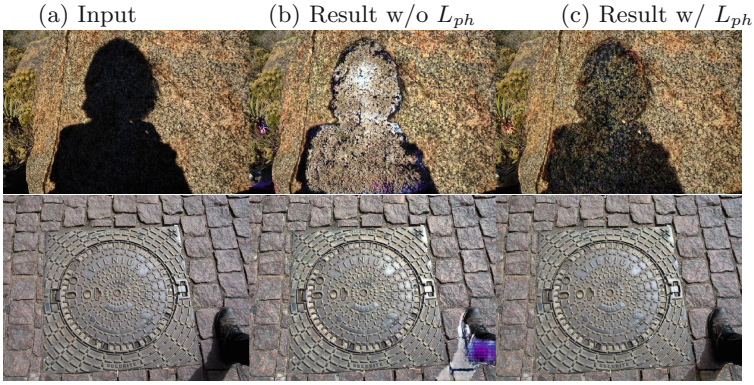


Fig. 10. Examples of adversarial examples generated with and without physics. (a) Input image I. (b) Adversarial example generated by A-Net trained without physics based loss. (c) Adversarial example generated by A-Net trained with physics based loss.

We conducted experiments to study the effect of the parameters of our framework. We started from the parameter settings reported in Sect. 4. When we chose $\lambda_{sd} = 10$, D-Net achieved 6.5% BER. As λ_{sd} increases, A-Net attenuates the shadow more dramatically but also tends to change the non-shadow part, generating lower quality images in general. In the second experiment, we rescaled the ratio between the real and adversarial images being input to D-Net. When we chose $\lambda^0_{adv} = 0.5$ and $\lambda_{real} = 0.5$, D-Net achieved 7.0% BER.

5 Summary

In this paper, we have presented a novel framework for adversarial training of a shadow detector using shadow attenuation. We have shown experimentally how our model is able to effectively learn from both real shadow training examples as well as adversarial examples. Our trained model outperforms the previous state-of-art shadow detectors in two benchmark datasets, demonstrating the

effectiveness and generalization ability of our model. Furthermore, to the best of our knowledge, this is the first shadow detector that can detect shadows accurately at real-time speed, 45 fps.

Acknowledgements. This work was supported by the Vietnam Education Foundation, a gift from Adobe, NSF grant CNS-1718014, the Partner University Fund, and the SUNY2020 Infrastructure Transportation Security Center. The authors would also like to thank NVIDIA for GPU donation.

References

1. Banerjee, A., Dhillon, I.S., Ghosh, J., Sra, S.: Clustering on the unit hypersphere using von Mises-Fisher distributions. J. Mach. Learn. Res. **6**, 1345–1382 (2005)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2009)
3. Erraqabi, A., Baratin, A., Bengio, Y., Lacoste-Julien, S.: A3T: Adversarially augmented adversarial training. arXiv:1801.04055 (2018)
4. Finlayson, G., Hordley, S., Lu, C., Drew, M.: On the removal of shadows from images. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 59–68 (2006)
5. Finlayson, G., Drew, M., Lu, C.: Entropy minimization for shadow removal. Int. J. Comput. Vis. **85**, 35–57 (2009)
6. Goodfellow, I.J., et al.: Generative adversarial networks. In: Advances in Neural Information Processing Systems (2014)
7. Guo, R., Dai, Q., Hoiem, D.: Single-image shadow detection and removal using paired regions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2011)
8. Guo, R., Dai, Q., Hoiem, D.: Paired regions for shadow detection and removal. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2956–2967 (2012)
9. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the International Conference on Computer Vision (2015)
10. Hu, X., Zhu, L., Fu, C.W., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
11. Huang, X., Hua, G., Tumblin, J., Williams, L.: What characterizes a shadow boundary under the sun and sky? In: Proceedings of the International Conference on Computer Vision (2011)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
13. Khan, H., Bennamoun, M., Sohel, F., Togneri, R.: Automatic feature learning for robust shadow detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (2015)
15. Lalonde, J.-F., Efros, A.A., Narasimhan, S.G.: Detecting ground shadows in outdoor consumer photographs. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6312, pp. 322–335. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15552-9_24

16. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
17. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
18. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
19. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
20. Nguyen, V., Vicente, T.F.Y., Zhao, M., Hoai, M., Samaras, D.: Shadow detection with conditional generative adversarial networks. In: Proceedings of the International Conference on Computer Vision (2017)
21. Panagopoulos, A., Samaras, D., Paragios, N.: Robust shadow and illumination estimation using a mixture model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2009)
22. Panagopoulos, A., Wang, C., Samaras, D., Paragios, N.: Simultaneous cast shadows, illumination and geometry inference using hypergraphs. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 437–449 (2013)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
25. Sunkavalli, K., Matusik, W., Pfister, H., Rusinkiewicz, S.: Factored time-lapse video. In: Proceedings of the ACM SIGGRAPH Conference on Computer Graphics (2007)
26. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: attacks and defenses. In: Proceedings of the International Conference on Learning Representations (2018)
27. Vicente, T.F.Y., Hoai, M., Samaras, D.: Leave-one-out kernel optimization for shadow detection. In: Proceedings of the International Conference on Computer Vision (2015)
28. Vicente, T.F.Y., Hoai, M., Samaras, D.: Noisy label recovery for shadow detection in unfamiliar domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
29. Vicente, T.F.Y., Hoai, M., Samaras, D.: Leave-one-out kernel optimization for shadow detection and removal. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 682–695 (2018)
30. Vicente, T.F.Y., Hou, L., Yu, C.-P., Hoai, M., Samaras, D.: Large-scale training of shadow detectors with noisily-annotated shadow examples. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 816–832. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_49
31. Volpi, R., Morerio, P., Savarese, S., Murino, V.: Adversarial feature augmentation for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)

32. Wang, J., Li, X., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
33. Wehrwein, S., Bala, K., Snavely, N.: Shadow detection and sun direction in photo collections. In: Proceedings of 3DV (2015)
34. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the International Conference on Computer Vision (2017)
35. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: Proceedings of the International Conference on Learning Representations (2018)
36. Zhu, J., Samuel, K., Masood, S., Tappen, M.: Learning to recognize shadows in monochromatic natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2010)