Shadow Removal with Paired and Unpaired Learning

Florin-Alexandru Vasluianu
ETH Zürich
fvasluianu@student.ethz.ch

Andrés Romero
ETH Zürich
roandres@ethz.ch

Luc Van Gool
ETH Zürich
vangool@vision.ee.ethz.ch

Radu Timofte
ETH Zürich
radu.timofte@vision.ee.ethz.ch

Abstract

Shadow removal is an important computer vision task aiming at the detection and successful removal of the shadow produced by an occluded light source and a photo-realistic restoration of the image contents. Decades of research produced a multitude of hand-crafted restoration techniques and, more recently, learned solutions from shadowed and shadow-free training image pairs. In this work, we propose a single image shadow removal solution via self-supervised learning by using a conditioned mask. We rely on self-supervision and jointly learn deep models to remove and add shadows to images. We derive two variants for learning from paired images and unpaired images, respectively. Our validation on the recently introduced ISTD and USR datasets demonstrate large quantitative and qualitative improvements over the state-of-the-art for both paired and unpaired learning settings.

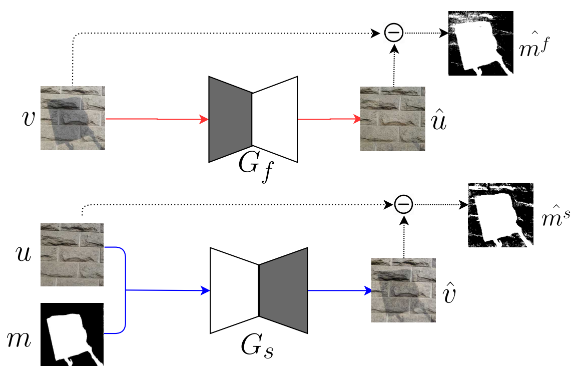
1. Introduction

In an image, a shadow [35] is the direct effect of the occlusion of a light source. By inducing a steep variation in an image region, the shadow impacts the performance of other vision tasks such as image segmentation [7, 1], semantic segmentation [32, 10], object recognition [36, 2, 15] or tracking [21, 24, 4]. In contrast to the unshadowed pixels, the shadow alters the observation of the scene contents by a combination of degradations in illumination, color, detail, and noise levels. The shadow removal task is, essentially, an image restoration task aiming at recovering the underlying content. Many methods [31, 28] have been proposed for detecting and removing shadows from images.

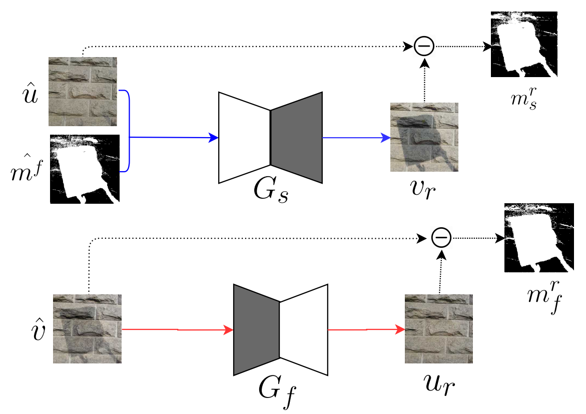
The introduction of large datasets of shadowed and shadow-free image pairs such as SRD [30], ISTD [38] or USR [17] allowed the formulation of the shadow removal process as a regression problem. One of the major challenges is to learn a physically plausible transformation, regardless of the semantic or illumination inconsistencies that may be encountered in the data. Thanks to the advent of Generative Adversarial Networks (GANs) [13] and its flexibility in learning complex distributions, recent efforts [17, 40] have modeled the shadow removal task as an image-to-image translation problem [19]. However, it has been found [29] that the learned transformations are highly prone to artifacts produced in the downsampling/upsampling phases of the translation encoder/decoder model, and moreover, the tendency of the deshadowed image regions to be blurry [17, 41]. In order to circumvent these problems, recent solutions [38, 17, 25] have proposed carefully designed robust loss functions, producing photo-realistic deshadowed results with low pixel-wise restoration errors. However, generally the results are still affected by strong artifacts.

As the shadow removal is a perceptual transformation, the usage of a perceptual score based on learned features, on different levels of complexity [42], enables the exploitation of some invariants over the shadow removal or addition transformation. Increasing the amount of information used in training is expected to induce additional degrees of control such that the learning procedure can be faster, and the results produced will be better, both in terms of fidelity metrics and perceptual scores. As collecting paired images is cumbersome and costly, learning from unpaired shadow and shadow-free images, much cheaper to acquire, is a necessity.

In this work, we propose a single image shadow removal solution via self-supervised learning and derive two variants for learning either from paired or from unpaired training images. Our method exploits several observations made on the shadow formation process and employs the cyclic consistency and the GAN paradigms as inspired by the CycleGAN [43], a seminal work for learning image-to-image translations between two image domains from unpaired im-



(a) Forward step



(b) Reconstruction step

Figure 1: The forward (top) and the reconstruction step (bottom). As a convention, red lines were used for the manipulation involving shadow affected input, blue lines for the shadow free input, and the black lines for the mask computation operations. *u* and *v* could or could not be paired data. Our training framework uses *v*'s mask (*m*) to insert it in G_s. In a paired setting $\hat{v}$ should resemble v, while not the case for an unpaired one.

ages. An overview of our method is depicted in Figure 1. Given a dataset with shadow images and respective masks, either on paired or unpaired settings, the core of our method is to exploit the given mask information in a self-supervised fashion by using, in the unpaired setting, randomly sampled shadow masks into the training framework (replacing m by $\hat{m}^*$ during the forward step in Figure 1.a), and reconstructing the original input, imposing the cycle-consistency (Figure 1.b). Sampling from intermediary results implies a need to control their quality, using all the information available, and imposing phenomenon characteristic properties during training. A critical observation is that we do not need to impose strong pixel-wise fidelity losses in our solution, but rather capture contents and general texture and colors, which are inherently perceptual.

2. Related Work

Shadow removal is not a recent problem in the computer vision field, and despite recent efforts in deep learning and generative modeling it is still a challenging problem.

Early methods tackled this problem by using the underlying physical properties of the shadow formation. They were based on image decomposition as a combination of shadow and shadow-free layers [9, 8], or on an early shadow detection followed by a color transfer from the shadow-free region to the shadow affected region in the local neighborhood [39, 33, 37].

The variety of shadow generation systems (*e.g.* shapes, size, scale, illumination, *etc.*) implies an increased complexity in shadow model parameters computation, and consequently, models parameterized with these properties are known for not being able to handle shadow removal in complex situations [23]. A step forward in this direction is a two-stage model for shadow detection and removal, respectively, thus increasing the generalization performance. In order to successfully detect the shadow, earlier works [14, 12] proposed to include hand-crafted features such as image intensity, texture, or gradients.

The evolution of the Convolutional Neural Networks (CNNs) enabled the propagation of these learnable features along with the layers of the model, and [23, 22] proposed solutions using CNNs for shadow detection and a Bayesian model for shadow removal. Moreover, [12] pioneered using an unsupervised end-to-end auto-encoder model to learn a cross-domain mapping between shadow and shadow-free images. However, the need for manual labeling in the preprocessing step in order to produce an estimation over the shadow mask limits this method, both in terms of the complexity of the addressed light-occluder systems and the bias injection.

Qu *et al.* [30] proposed a model based on three networks extracting relevant features from multiple views and aggregating them to recover the shadow-free image. The importance of the localization information is acknowledged by Hu *et al.* [16], where the shadows were detected and removed using the idea of a Spatial Recurrent Neural Network [3] by exploring the direction-aware context.

Subsequently, since the introduction of GANs, the dominant strategy is to learn an image-to-image mapping function using an encoder/decoder architecture. The de-facto methodologies for image-to-image translation for paired and unpaired data are pix2pix [19] and CycleGAN [43], respectively. In the former, it is assumed a single transformation between shadow and deshadowed regions, while in the latter, in order to deal with the unsupervised nature of the data, there are two different models for shadow removal and addition, and a cycle-constraint loss ensures the flow from one domain to another.

Following this trend, [38] proposed a model based on

Conditional GANs [27] using paired data, where they deployed 2 stacked conditional GANs aiming detection and then using the partial results for shadow removal. Recently, Le *et al*. [25, 26] proposed a model based on two neural networks able to learn the shadow model parameters and the shadow matte. The main limitation of the model is the usage of a simple linear relation as the light model. However, by using the same occluder there can be multiple light sources to produce non-homogeneous shadow areas that can not be described by a linear model, and therefore, the performance of the model is expected to drop. Nonetheless, if the assumption made about the uniqueness of the light source holds, the method is able to produce realistic results.

Despite the recent effort in shadow removal literature, all prior methods rely on the assumption of paired datasets for shadow and shadow-free images. Going in an unsupervised direction, [17] developed MaskShadowGAN using a vanilla CycleGAN [43] approach, where the shadow masks are computed as a binarization of the image difference, by thresholding it using Otsu's algorithm. This estimation over the shadow mask is used in order to compute the reconstructed images, then enforcing a cycle consistency over the produced outputs. Later, in [6], authors proposed the usage of a learnt attention map to provide the information over the properties of the shadow affected area, used to recover a shadow-free image in a residual learning setting.

In contrast, we formulate a component in the training objective that is going to offer guidance over the quality of the synthetically-generated shadow masks used as input in the step. The control over the properties of the shadow mask will be proven to be crucial in order to learn a realistic mapping, and so, using such a component will increase the degree of control in the training procedure.

3. Proposed Method

3.1. Problem Formulation

Considering the shadow image domain X and the shadow-free image domain set Y, we are mainly interested to learn the mapping function $G_f : X \rightarrow Y$. Existing techniques [19] rely on a critical dataset assumption of having access to paired images, *i.e*. the same scene with/without shadows. As we will show in Section 4, this assumption does not always hold, and having an unsupervised approach leads, surprisingly, to better performance. To this end, we assume a subset of unpaired images $T = \{(u, v) | u \in Y, v \in X\}$.

3.2. Overall scheme

The overall scheme of our method is presented in Figure 1. Our system is based on the CycleGAN approach [43], where, for each domain, a generator learns the transformation to the other domain, and a discriminator is used to guide the learning procedure. The shadow addition generator receives two inputs: a shadow free image and a binary mask providing the information about the position, size, and shape of the hallucinated area. The shadow removal network only receives the shadow affected image, learning to localize the affected area and then, to restore the contents. Formally, $\hat{u} = G_f(v)$ and $\hat{v} = G_s(u, m)$, for removal and addition respectively (Figure 1a), m being either the provided mask, if available, or a randomly sampled mask from a collection of synthetically generated masks, computed using the previous partial results in the forward shadow removal step. To close the cycle-consistency loop, we use self-supervision to reconstruct the original inputs (Figure 1b), using the mask computed in the forward shadow removal to attempt a successful reconstruction. We extensively explain this process in Section 3.3.7.

Besides our self-supervised training framework, an important ingredient of our method relies on the carefully designed loss functions we proceed to explain in the following section. The loss objective is carefully developed in order to adapt the learning procedure to the unpaired setting, where the information regarding the properties of the shadow affected area is not available. Thus, the shadow detection has to be learnt by the shadow removal generator, using only the partial information available on the shadow addition track of the cycle (*e.g*. synthetical shadow mask used to produced the reconstructed result, and how successful the reconstruction was, using the provided synthetical mask). Here, we rely on deep feature based loss terms, used either to exploit the content invariance along the suffered transformations, or to enhance the visual properties of the results.

3.3. Objectives and losses

For simplicity and sake of clarity, it is important to mention that we define our losses regardless of the training settings (paired or unpaired) and the transformation mapping (inserting or removing shadow), and instead we use placeholders. We will make a clear distinction at the end of this section.

3.3.1 Pixel-wise losses

Our main motivation to build an unsupervised shadow removal comes from an observation we illustrate in Figure 2. On the one hand, there are pixel-wise inconsistencies (*e.g*., different lighting conditions outside the shadow or content misalignment) for paired images in the ISTD dataset [38], so building a model under this assumption compromises the performance. On the other hand, using loss functions based solely on a pixel-wise level (L1, L2, etc.) is also not a suitable learning indicator, as it can lead to producing quite blurry outputs while minimizing the function.

3.3.2 Perceptual losses

We aim at removing shadows while preserving the non-shadowed areas as unaltered as possible. Therefore, inspired by recent literature in photo-enhancement [18], style transfer [11] and perceptual super resolution [20], we form a perceptual ensemble loss for color, style, and content, respectively. The parameters used were empirically chosen in relation to the amplitude of each loss on a subset of the train data: $\alpha_1 = 1$, $\alpha_2 = 0.1$ and $\alpha_3 = 10000$.

$$L_{perceptual} = \alpha_1 \cdot L_{color} + \alpha_2 \cdot L_{content} + \alpha_3 \cdot L_{style}, \quad (1)$$

3.3.3 Color loss

The introduction of a color loss can be explained, firstly, by the need to capture and preserve color information in the image. Under ideal settings, this could be done by imposing a pixel-level loss (*e.g.*, L1, L2, MSE). However, we consider that the color is a lower frequency component than the textural information of the image (our eyes are less sensitive to color than to intensity changes) and the pixel-level observations are generally noisy (*e.g.*, pixel-wise inconsistencies in ISTD image pairs). To this end, inspired by [5], we perform a Gaussian filter over the real and fake image, and compute the Mean Squared Error (MSE),

$$L_{color} = MSE(I^1_{smoothed}, I^2_{smoothed}) \quad (2)$$

3.3.4 Content loss

Building on the assumption that an image with shadows and that one without shadows should have similar content in terms of semantic relevant regions, the $L_{content}$ is defined as

$$L_{content} = \frac{1}{N_l}\sum_{i=1}^{N_l} MSE(C^i_{I^1}, C^i_{I^2}), \quad (3)$$

where C^i is the feature vector representation extracted in the i-th target layer of the ImageNet pretrained VGG-16 network [34], for each input image I^n.

3.3.5 Style loss

L_{style}, is defined as

$$L_{style} = \frac{1}{N_l}\sum_{i=1}^{N_l} MSE(H^i_{I^1}, H^i_{I^2}) \quad (4)$$

$$H^l_{I\,i,j} = \sum_{k=1}^{D} C^l_{I\,i,k} C^l_{I\,k,j} \quad (5)$$

where the Gram matrix H^l_I of the feature vector extracted by every i-th layer of the VGG-16 net. The Gram matrix H^i_I defines a style for the feature set extracted by the i-th layer of the VGG-16 net, using as input the I image. By minimizing the mean square error difference between the styles computed for feature sets at different levels of complexity, the results produced will be characterized by better perceptual properties.

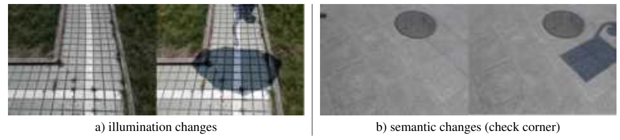


Figure 2: Examples of ISTD paired images that are not perfectly aligned and consistent.

3.3.6 Adversarial Losses

The formulation of the problem using adversarial learning implies the introduction of two new components, D_f and D_s. The main idea behind the learning procedure is that, for each domain, the discriminator will distinguish between the synthetic and the real results, forcing the counterpart generator to produce a better output in terms of semantic content and image properties.

As the discriminators are characteristic to the shadow domain X and the shadow free domain Y, the adversarial losses are defined, for the synthetic results produced in the forward step ($\hat{I}_f$,$\hat{I}_s$), as stated in Equation 6 and 7. The image pair (I_s, I_f) is the ground truth shadow-shadow free pair used as input, and $(\hat{I_s}^*, \hat{I_f}^*)$ is a pair of randomly sampled synthetic results.

$$\begin{aligned} L^s_{GAN}(I_s, I_f) = \frac{1}{2}(&MSE(J, D_s(\hat{I}_s, I_s)) \\ &+ MSE(O, D_s(\hat{I_f}^*, I_s))),\ \forall \hat{I_f}^* \notin X \end{aligned} \quad (6)$$

$$\begin{aligned} L^f_{GAN}(I_s, I_f) = \frac{1}{2}(&MSE(J, D_f(\hat{I}_f, I_f)) \\ &+ MSE(O, D_f(\hat{I_s}^*, I_f))),\ \forall \hat{I_s}^* \notin Y \end{aligned} \quad (7)$$

The standard output for the discriminators for the positive and negative examples used in training was defined as J and O, as a consequence of the usage of the *patchGAN* concept. So, J and O are defined as the all one matrix, and all zero matrix, respectively, with a size equal to the size of the output size of the discriminator.

3.3.7 Self-Supervised Shadow Loss

The last and core ingredient of our system is the self-supervised shadow loss (S3 loss). In order to generate an image with shadows, we couple the shadow insertion generator to receive both a binary mask in addition to the RGB image. Our rationale comes from realizing that shadow addition can be at any random shape, scale, size, and position, then tackling this problem in an unconditional way is ill-posed. Additionally, by guiding the shadow insertion using a conditional mask, we can use a randomly inserted shadow mask into a deshadowed image in order to perform a cycle consistency loss to recover the mask.

Considering u the shadow free image and v the shadow image, then the generated images have the form of $\hat{u} = G_f(v)$ and $\hat{v} = G_s(u, m)$, where m is the shadow mask between the images u and v. Similarly, we compute the reconstructed cycle-consistency images as $u_r = G_f(\hat{v})$ and $v_r = G_s(\hat{u}, \hat{m^f})$, where $\hat{m^f} = Bin(\hat{u} - v)$ is the synthetic shadow mask computed, as the median value binarization of the difference between the synthetic shadow free image and the true shadow image. As the generator uses this result in the reconstruction step, the quality of this result is crucial for a qualitative reconstruction. Even if a good mask will be the result of a realistic transformation, the usage of this proxy loss will enable faster learning, and also, the learning of more realistic mappings.

In the unpaired setting, since the u and v images don't represent the same scene, the mask used for the forward shadow addition, $\hat{m}^*$, is sampled from a fixed size memory buffer. This contains masks produced in the forward step of the shadow removal process. So, $\hat{m}^* = Bin(\hat{u}^* - v^*)$, where v^* is the shadow affected input image used to compute the synthetic shadow free image $\hat{u}^*$. The quality of the used shadow masks can be directly linked to a realistic looking output. As a result, the mask loss L_{mask} was introduced, exploiting the invariance property of the shadow mask during the transformations performed along the cycles. L_{mask} was defined as the L1 difference between either the provided mask in the paired setup, or the randomly sampled mask in the unpaired, and the masks computed with the synthetically generated images, after the forward pass, conditioning on the input images. Additional terms were added to enforce the consistency of the shadow mask also for the reconstruction procedure, conditioning on the outputs of the forward step.

3.3.8 Total Loss

As the images that are to be fed to the model in the unpaired setting do not represent the same scene, the loss function has to be carefully chosen such that the equilibrium point can be reached, and so, the learning procedure will converge to the required solution. As we defined the GAN loss in Equation 6 and 7, we are using replay buffers such that the discriminators are less likely to rely on the simple difference between frames representing similar contents. Thus, the degree of control is weaker, as we are sampling from intermediary results to feed the negative samples $\hat{I}_f^*$ and $\hat{I}_s^*$ in the binary classification problem. So, additional available information can be exploited by using different complexity features extracted by a pretrained neural network, to guide the learning to a natural transformation.

Moreover, another crucial goal is controlling, as much as possible, the quality of the intermediary results, by the way of minimizing the distance between the topology information localizing the hallucinated area in the forward pass, to the one observed in the reconstruction pass of the training step. So, we are exploiting the observation that, under convergence conditions, both shadow removal and shadow addition procedures are inverse to each other, as we stated shadow removal as a bijectivity in the problem formulation.

$$\begin{aligned} L_{gen}(u,v) = {} & \gamma_1 \cdot (L^f_{GAN}(\hat{u}, u) + L^s_{GAN}(\hat{v}, v) \\ & + L^f_{GAN}(u_r, u) + L^s_{GAN}(v_r, v)) \\ & + \gamma_2 \cdot (L_{content}(u, \hat{v}) + L_{content}(v, \hat{u})) \\ & + \gamma_3 \cdot (L_{pix}(u, u_r) + L_{pix}(v, v_r)) \\ & + \gamma_4 \cdot (L_{perceptual}(u, u_r) + L_{perceptual}(v, v_r)) \\ & + \gamma_5 \cdot (L_{mask}(\hat{m^f}, m^f_r) + L_{mask}(\hat{m^s}, m^s_r) \\ & + \beta_2 L_{mask}(\hat{m}^*, \hat{m^f})) \end{aligned} \tag{8}$$

So, we choose the total loss for the unpaired case as a linear combination of the losses described, where γ and β parameters control the contribution of each loss. Note that each component is easily extracted from Figure 1. As the shadow position and shape are invariant under the transformations implied, by adding this term in the training objective, the transformations will be naturally plausible. As the datasets are not characterized by a high variation in terms of shadow regions shapes and positions, the model will benefit from adding a loss term such that the mask produced in the reconstruction procedure is similar to the sampled mask $\hat{m}^*$

3.3.9 Total Loss for Paired Data

As our method can be extended for paired datasets, in Equation 9 we show the modifications to the loss functions in this scenario.

$$\begin{aligned} L_{gen}(u,v) = {} & \gamma_1 \cdot (L^f_{GAN}(\hat{u}, u) + L^s_{GAN}(\hat{v}, v) \\ & + L^f_{GAN}(u_r, u) + L^s_{GAN}(v_r, v)) \\ & + \gamma_2 \cdot (L_{content}(u, \hat{v}) + L_{content}(v, \hat{u})) \\ & + \gamma_3 \cdot (L_{pix}(u, u_r) + L_{pix}(v, v_r) + \beta_1 L_{pix}(u, \hat{u})) \\ & + \gamma_4 \cdot (L_{perceptual}(u, u_r) + L_{perceptual}(v, v_r)) \\ & + \gamma_5 \cdot (L_{mask}(\hat{m^f}, m^f_r) + L_{mask}(\hat{m^s}, m^s_r) \\ & + \beta_2 L_{mask}(m, \hat{m^f})) \end{aligned} \tag{9}$$

When training using paired data, a constraint can be used to speed-up the convergence process, by adding the term $L_{pix}(u, \hat{u})$ as the L1 pixel-wise loss between the input shadow free image and the shadow free image generated in the forward step of the cycle. The β_2 parameter is needed to force the model to create a suitable transformation in the forward step of the cycle, as the reconstruction process will be using this intermediate representation of the shadow free image. As the mask shape and position should be the same, the shadow masks would not differ along the cycle, so L_{mask}, the L1 distance between the two shadow masks, was introduced. The weights used in the linear combination of GAN, L1, perceptual, or mask losses were determined with respect to the magnitude of each term, and their speed of decrease. Exact values were provided in Table 1, for both the paired and the unpaired settings.

Table 1: Parameters of the total loss function (Equation 8 and 9) defined for our training framework.

Setting	γ_1	γ_2	γ_3	γ_4	γ_5	β_1	β_2
Unpaired training	250	10	100	30	60	0	100
Paired training	250	20	60	50	60	10	100

3.4. Implementation details

Generator. The generator consists of eight pairs of downsampling/upsampling blocks, with skip-connections from the downsampling block to the upsampling module, on the same dimensionality level. As operations, the direct/transposed convolution with kernel size 4 and zero padding 1 was used. The result is passed through a LeakyReLU activation ($\alpha = 0.2$) for the downsampling blocks, and ReLU for the upsampling part. The final output is passed through a $\tanh$ activation. Dropout with 50% probability was used in order to improve the behaviour of the generators during training.

The discriminator. consists of four convolutional blocks, each having the convolution operator with $k = 4$, instance normalization and LeakyReLU as activation ($\alpha = 0.2$). The final output size of the discriminator will be the size of the patch described as the "perceptive field" of the model. The depth of the initial input tensor can be explained by the fact that the discriminator will receive as input a pair of images, each of them with three channels, as they are RGB images.

Initialization. As the initialization, the weights in both the discriminators and the generators were drawn from a Gaussian distribution with 0 mean and 0.2 variance.

4. Experimental Results

4.1. Setup

Datasets. We validate our system over *ISTD* [38], and *USR* [17] datasets. On the one hand, the ISTD dataset contains paired data for shadow and shadow-free images. Given the illumination inconsistency problem in this dataset, Le *et al*. [25] proposed a compensation method, creating thus the *ISTD+* dataset. On the other hand, the USR dataset is a collection of unpaired shadow and shadow free images used for unsupervised tasks. For unpaired training and testing on the ISTD dataset, the shadow and shadow free images were randomly sampled. The random mask inserted in each iteration comes from a buffer bank of synthetic masks.

Experimental Framework. We train our system during 100 epochs, learning rate 0.005 with λ-decay scheduling after the first 40 epochs. We use Adam optimizer with $\beta = (0.9, 0.999)$.

For both the paired and unpaired settings, the masks were computed as a binarization of the difference between the shadow free image and the shadow image, by a thresholding procedure using the median value of the difference.

Evaluation measures. For the quantitative evaluation of our method, we use the Root Mean Square Error (RMSE) and Peak Signal to Noise Ratio (PSNR) between the output deshadowed image and the reference/ground truth image. We compute these pixel-wise fidelity measures in both RGB and Lab color space, respectively. It is well established that RMSE and PSNR do not correlate well with perceptual quality, so complementary to the fidelity measures, we also employ LPIPS [42] score in order to assess the photorealism of the produced deshadowed images with respect to the ground truth.

Compared methods. We directly compare our proposed solution to two other methods capable to learn from unpaired data: *CycleGAN* [43] and *Mask Shadow GAN* [17]. Moreover, in order to compare with prior systems in paired learning, we report qualitative and quantitative results for the following methods:*DSC* [16], *ST-CGAN* [38] and *DeShadowNet* [30].

4.2. Ablation Study

In Table 2 we report our results on ISTD for different settings. Each configuration was trained for 100 epochs, on ISTD dataset. When switching from a learning procedure based on both fidelity and perceptual losses to an only-fidelity loss based objective ($\gamma_4 = 0$) the results improve in fidelity and lack in perceptual terms. The removal of the mask loss ($\gamma_5 = 0$) produces similar results in terms of both perceptual and fidelity measures, but the standard deviation

Table 2: The impact of various loss function parameters on the performance.

Method	LPIPS↓ avg	LPIPS↓ stddev	RMSE↓ RGB	RMSE↓ Lab	PSNR↑ RGB	PSNR↑ Lab
Paired setting	**0.031**	0.025	15.05	4.18	26.88	37.91
Paired setting ($\gamma_2 = 0$)	0.032	0.029	14.75	4.15	27.1	38.05
Paired setting ($\gamma_4 = 0$)	0.033	0.022	14.07	4.01	27.39	38.31
Paired setting ($\gamma_5 = 0$)	0.031	0.027	**13.80**	**3.90**	**27.70**	**38.58**
Paired setting ($\beta_2 = 0$)	0.035	**0.019**	14.70	4.11	27.26	38.2
Paired setting[1] ($\beta_1 = 0$)	0.021	0.046	5.96	2.61	34.27	41.67

over the LPIPS score is higher due to a more pronounced difficulty of the model to deal with more complex textures.

The forward pixel-wise loss ($\beta_1 = 0$) is very important in order to produce realistic results in the forward step of the cycle, even though the latent representation learnt $(\hat{u}, \hat{v})$, for both shadow and shadow free domains, produce the best results in terms of reconstruction error (either fidelity loss or perceptual score). As a better mask is a consequence of a better reconstruction, the dropping of the forward mask ($\beta_2 = 0$), produce results characterized by similar pixel-wise properties, but lacking in perceptual terms.

Table 3: Ablative results in terms of pixel-wise loss (RMSE and PSNR, both on Lab space) and perceptual quality loss (LPIPS) for different settings of the loss function.

Setting description	RMSE↓	PSNR↑	LPIPS↓
default set of parameters	5.73	33.40	**0.045**
default and $\gamma_2 = 0$	5.34	34.35	0.082
default and $\gamma_3 = 10$	5.98	32.87	0.104
default and $\gamma_4 = 0$	**3.71**	**37.31**	0.056
default and $\gamma_5 = 0$	4.42	36.56	0.074

In Table 3, we investigated the behaviour of our model under different configurations of the unpaired setting. A trade-off between improving in terms of fidelity score vs. the perceptual properties can be observed, concluding that both the mask loss and the perceptual loss yield better results in terms of perceptual score.

4.3. Quantitative results

To quantitatively evaluate the performance of our shadow removal solution, we adhere to the ISTD and USR benchmarks [17, 38] and report the results in Table 4. For all the reported results, we used our models trained on the training partition of the ISTD dataset, for 100 epochs for both the paired and the unpaired settings. For the unpaired setting, the shadow and shadow free training images were sampled without replacement.

[1]This configuration was not considered, due to learning a non-realistic mapping.

Table 4: Comparison with state-of-the-art methods on ISTD and USR datasets.

Method	ISTD test images LPIPS↓ avg	ISTD LPIPS↓ stddev	ISTD RMSE↓ RGB	ISTD RMSE↓ Lab	ISTD PSNR↑ RGB	ISTD PSNR↑ Lab	USR test images LPIPS↓ avg	USR LPIPS↓ stddev	USR RMSE↓ RGB	USR RMSE↓ Lab	USR PSNR↑ RGB	USR PSNR↑ Lab
Unpaired data training												
MaskShadowGAN[17]	0.25	0.09	28.34	7.32	19.78	31.65	0.31	0.11	27.53	7.06	19.97	31.76
CycleGAN [43]	0.118	0.07	25.4	6.95	20.59	31.83	0.147	0.07	30.04	9.66	19.04	29.06
ours (unpaired)	**0.041**	**0.033**	**7.58**	**5.12**	**31.18**	**34.45**	**0.009**	**0.004**	**5.70**	**2.21**	**33.26**	**41.06**
Paired data training												
DeShadowNet [30]	0.080	0.055	31.96	7.98	19.30	31.27	-	-	-	-	-	-
DSC [16]	0.202	0.087	23.36	6.03	21.85	33.63	-	-	-	-	-	-
ST-CGAN [38]	0.067	0.043	22.11	5.93	22.66	34.05	-	-	-	-	-	-
ours (paired)	**0.031**	**0.025**	**15.05**	**4.18**	**26.88**	**37.90**	-	-	-	-	-	-

Table 5: Lab color space results for both shadow and shadow-free pixels on ISTD[38] and ISTD+[25] datasets.

Method	Setup Train	Setup Test	All RMSE	All PSNR	Shadow RMSE	Shadow PSNR	Shadow free RMSE	Shadow free PSNR
ARGAN[6]	ISTD	ISTD	5.89	N/A	6.65	N/A	5.41	N/A
ours (unpaired)	ISTD	ISTD	5.12	34.45	6.98	32.65	4.94	34.71
ours (paired)	ISTD	ISTD	4.18	37.90	4.63	36.87	4.07	38.22
ours (paired)	ISTD+	ISTD+	**3.04**	**41.10**	**4.15**	**38.16**	**2.77**	**42.05**
[25] (paired)	ISTD+	ISTD+	3.8	N/A	7.4	N/A	3.1	N/A

The USR dataset provides a collection of shadow free images and two splits of shadow images, for training and validation, which are not representing the same scene as the shadow-free images. The same sampling procedure was deployed for the USR dataset.

As shown in Table 4 our models largely improve the state-of-the-art in both fidelity (RMSE, PSNR) and perceptual measures (LPIPS) on both benchmarks.

4.4. Qualitative results

Figure 3 shows visual results obtained on randomly picked ISTD test images. We note that the results achieved by our solutions are the closest to the reference shadow free images, while the other methods generally produce strong artifacts. Our results clearly improve over the unpaired state-of-the-art Mask Shadow GAN, producing more appealing and artifact-free images. On the paired counterpart, our method completely removes the shadow while the related methods produce visible traces.

4.5. Discussion

For both paired and unpaired settings, our system produces the best perceptual metrics (lower LPIPS) and the best pixel-wise error metrics (PSNR and RMSE) with respect to state-of-the-art methods by large margins.

Figure 4 shows results and the square L2 norm of the residuals in the image space for our models. We observe that the paired version of the model has problems in recovering the unshadowed region on the neighborhoods characterized by sharp variations in terms of color and illumination variations. This could be due to the ISTD dataset used for training the model. ISTD has a limited shadow formation diversity in its pairs. Therefore, the model provides poorer results on images representing much more complex scenes.

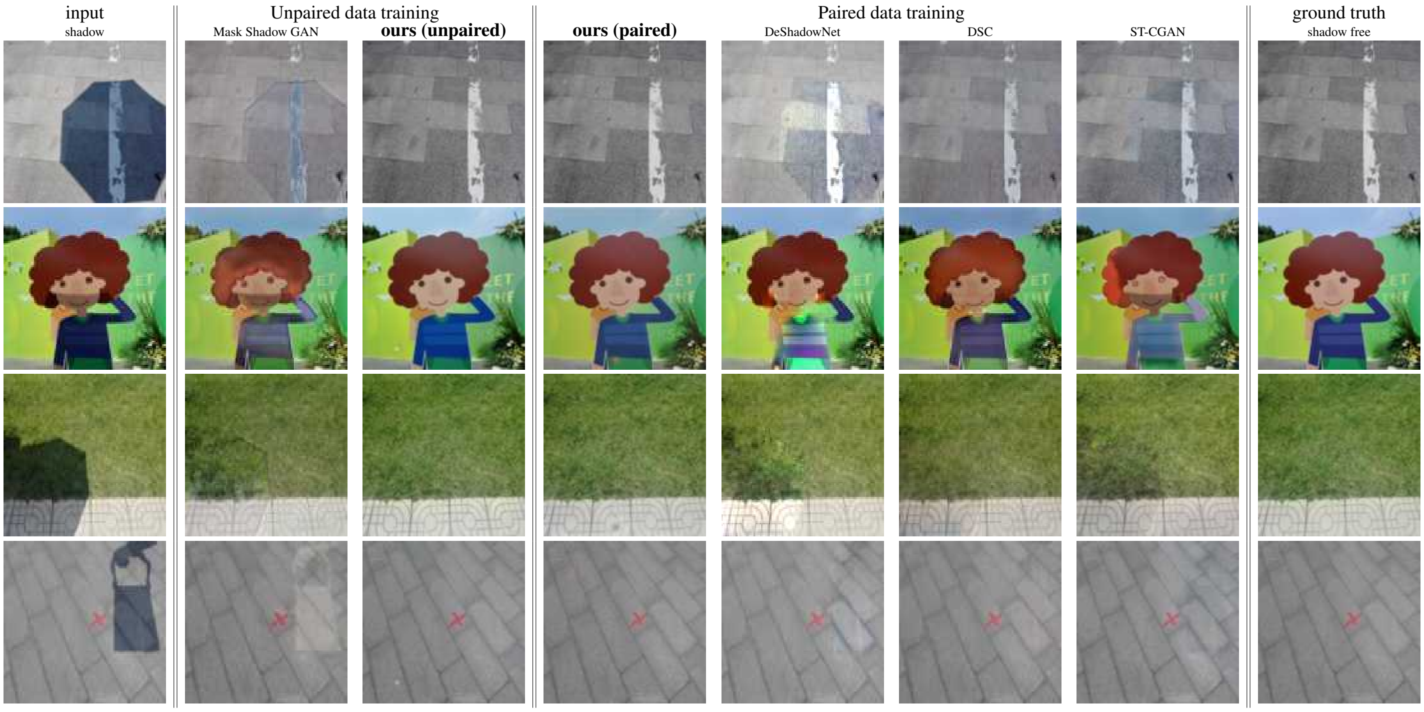


Figure 3: Visual results for the proposed solution and comparison with state-of-the-art learned methods. Best zoom in on screen for better details.

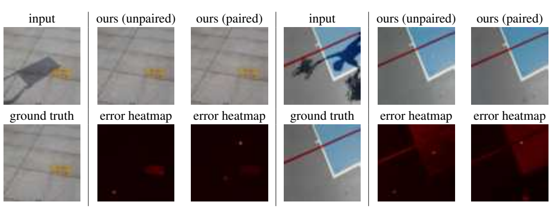


Figure 4: Visual results for the proposed solutions trained with unpaired and paired, and corresponding error heatmaps. Best zoom in on screen for better details.

Furthermore, as we show in Figure 2, the semantic differences and the differences in illumination are expected to induce a certain degree of uncertainty when using largely weighted L1 loss terms between ground truth images and the synthetically generated images in the same domain. Therefore, as it can be seen in Figure 4, the error is not concentrated in the shadow affected area, but, in steep variations in terms of texture, and some peaks in error can be observed in that area. When training in an unpaired manner, by simply dropping this loss term we can overcome this issue, improving our results on the ISTD dataset, compared to the paired setting.

The unpaired version also benefits from both sampling processes deployed, *i.e.*, for the shadow mask (using a mask buffer) and the negative examples for discriminator training. Since the sample sets are dynamically generated from synthetic data, the variation of the provided examples is expected to be higher. Therefore, the generalization ability of the model increases (as it can be observed in Table 4) producing better results in terms of both pixel-wise loss and perceptual metrics. This behaviour can be explained by the model benefiting from the variety of random localization/shape combinations characterising the shadowed region.

Although the degree of control is weak under the unpaired setting, the exploitation of both deep features, and the proxy loss defined for the transformed region provides sufficient information for the learnt mapping to be realistic.

5. Conclusions

In this work we proposed a novel single image shadow removal solution. We rely on self-supervision and jointly learn shadow removal from and shadow addition to images. Even if the degree of control is significantly weaker in the unpaired setting, using learnt deep-features, we are able to control the learning procedure to converge to a realistic mapping. As our experimental results show on ISTD and USR datasets, we set a new state-of-the-art in single image shadow removal, by largely outperforming prior works in both fidelity (RMSE, PSNR) and perceptual quality (LPIPS) for both paired and unpaired settings.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 1

[2] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014. 1

[3] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016. 2

[4] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014. 1

[5] Etienne de Stoutz, Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, and Luc Van Gool. Fast perceptual image enhancement. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 4

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10213–10222, 2019. 3, 7

[7] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 1

[8] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 2

[9] Graham D Finlayson, Steven D Hordley, and Mark S Drew. Removing shadows from images. In *European conference on computer vision*, pages 823–836. Springer, 2002. 2

[10] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018. 1

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 4

[12] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *Proceedings of the British Machine Vision Conference*, 2014. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[14] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2956–2967, 11 2013. 2

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[16] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. to appear. 2, 6, 7

[17] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-ShadowGAN: Learning to remove shadows from unpaired data. In *ICCV*, 2019. 1, 3, 6, 7

[18] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3277–3285, 2017. 4

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2, 3

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 4

[21] Pakorn KaewTraKulPong and Richard Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*, pages 135–144. Springer, 2002. 1

[22] Salman Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1–1, 01 2015. 2

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

[24] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–23, 2015. 1

[25] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 3, 6, 7

[26] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *The IEEE European Conference on Computer Vision (ECCV)*, August 2020. 3

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, abs/1411.1784, 2014. 3

[28] Vu Nguyen, Yago Vicente, F Tomas, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *Proceedings of the*

IEEE International Conference on Computer Vision, pages 4510–4518, 2017. 1

[29] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016. 1

[30] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2308–2316, July 2017. 1, 2, 6, 7

[31] Andres Sanin, Conrad Sanderson, and Brian C Lovell. Shadow detection: A survey and comparative evaluation of recent methods. *Pattern recognition*, 45(4):1684–1695, 2012. 1

[32] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2016. 1

[33] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. *Computer Graphics Forum*, 27(2):577–586, Apr. 2008. 2

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[35] Marc Stamminger and George Drettakis. Perspective shadow maps. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 557–562, 2002. 1

[36] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 1

[37] T. F. Y. Vicente, M. Hoai, and D. Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):682–695, March 2018. 2

[38] Jifeng Wang, Xiang Li, and Jian 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. 1, 2, 3, 6, 7

[39] Tai-Pang Wu, Chi-Keung Tang, Michael S. Brown, and Heung-Yeung Shum. Natural shadow matting. *ACM Trans. Graph.*, 26(2):8–es, June 2007. 2

[40] Ling Zhang, Chengjiang Long, Xiao-Long Zhang, and Chunxia Xiao. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *AAAI*, 2020. 1

[41] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 1

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1, 6

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1, 2, 3, 6, 7