S3R-Net: A Single-Stage Approach to Self-Supervised Shadow Removal

Nikolina Kubiak
University of Surrey
n.kubiak@surrey.ac.uk

Armin Mustafa
University of Surrey
armin.mustafa@surrey.ac.uk

Graeme Phillipson
BBC R&D
graeme.phillipson@bbc.co.uk

Stephen Jolly
BBC R&D
stephen.jolly@bbc.co.uk

Simon Hadfield
University of Surrey
s.hadfield@surrey.ac.uk

Abstract

In this paper we present S3R-Net, the Self-Supervised Shadow Removal Network. The two-branch WGAN model achieves self-supervision[1] *relying on the unify-and-adapt phenomenon - it unifies the style of the output data and infers its characteristics from a database of unaligned shadow-free reference images. This approach stands in contrast to the large body of supervised frameworks. S3R-Net also differentiates itself from the few existing self-supervised models operating in a cycle-consistent manner, as it is a non-cyclic, unidirectional solution. The proposed framework achieves comparable numerical scores to recent self-supervised shadow removal models while exhibiting superior qualitative performance and keeping the computational cost low. The pre-trained models and the code can be found in our github repo.*

1. Introduction

Shadows are physical phenomena that arise when an obstacle appears on the trajectory between a light source and a surface. They can provide necessary guidance for three-dimensional understanding of objects [2] and scenes [7, 57] in monocular settings [20]. However, in other scenarios, they can negatively affect our perception of the world around us. Therefore, a number of shadow removal frameworks have been developed for aesthetic purposes, ranging from de-shadowing casual capture photos [58] to cropping out unwanted objects alongside their shadows [35, 55]. Shadows can also have an adverse affect on the functioning of automated systems. In particular, they can obscure key visual clues, or introduce ambiguities by resembling dark objects [53]. Thus, shadow removal models can be useful as a pre-processing step for other computer vision tasks, *e.g.* for document shadow removal [11, 25] or for improving the accuracy of autonomous vehicle systems [46].

Unfortunately, most of the existing shadow removal frameworks operate in a supervised manner. The models require paired shadowed and shadow-free images, and may also required matching shadow masks. Due to the changing nature of the sun and sky conditions, such ground truth data is difficult to capture consistently for outdoor scenes [38, 47]. Indoor scenarios, while easier, have not gathered significant interest. Finally, synthetic datasets, *e.g.* [41], offer perfect colour and pixel-wise alignment yet they lack the realism and detail of data captured in the wild.

It is obviously possible to limit the training requirements of the shadow removal models and, in fact, a small number of such methods have been published. In the past few years, a few self-supervised methods [21, 23, 33, 43] have exploited cycle-consistency as their main supervisory signal. However, cycleGAN-based models require a secondary proxy task (here: shadow generation). This can negatively affect the model's robustness as it relies on both tasks functioning correctly and in balance. Additionally, this approach increases the number of generators and discriminators used by the system and, thus, the model's complexity. To the best of our knowledge, only a single unsupervised shadow removal framework has been proposed [19] yet its high domain-specificity limits its breadth of applications.

Motivated by the above, we present a novel solution with significantly reduced supervision requirements. The proposed Self-Supervised Shadow Removal Network (*S3R-Net*), contrasted with existing architectures in Fig. 1, is trained without the need for paired shadowed and shadow-free images. The model does not require any ground truth shadow masks nor does it rely on the accuracy of any explicit shadow detection modules. Instead, the desired appearance is learnt via adversarial learning from a collection of shadow-free images. This reference database does not

[1] In this work, we take "self-supervised" to mean a system that does not require ground truth shadow-free pairs for its training inputs, but only a set of shadow-free data from unrelated scenes. This matches the supervision requirements of existing self-supervised techniques, *e.g.* cycleGANs.

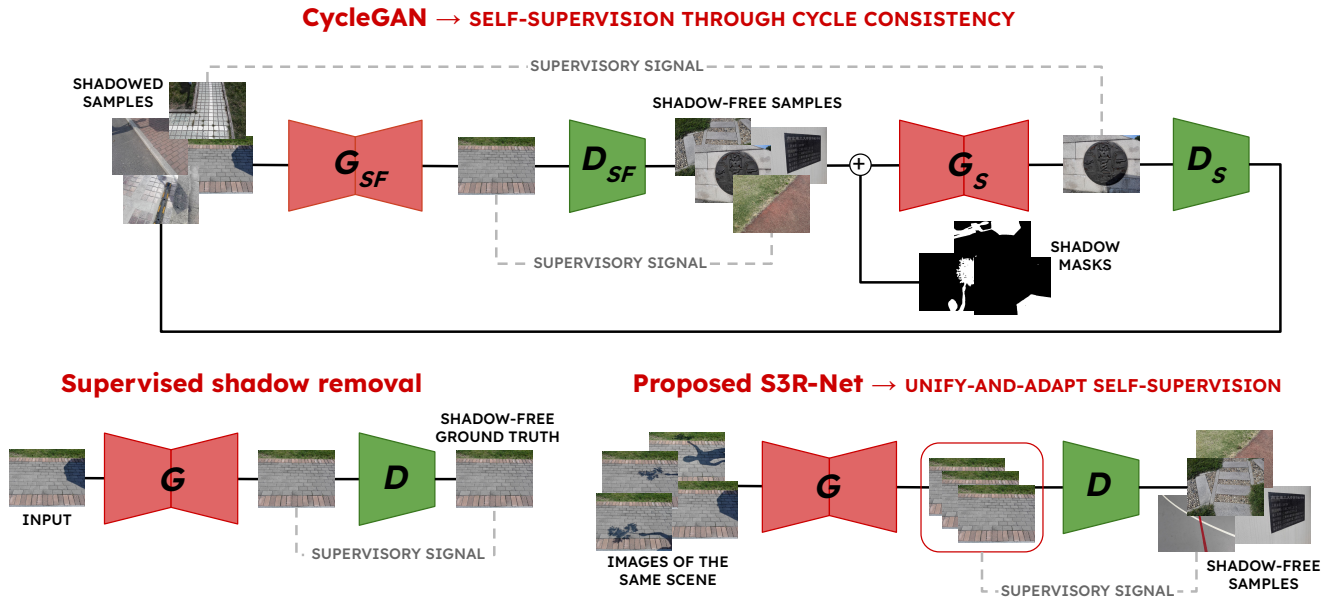

Figure 1. We present the architectures of a standard supervised shadow removal model (bottom left), a cycle-consistent self-supervised model (top) and the proposed S3R-Net (bottom right), exploiting a unify-and-adapt approach to self-supervision. The figure shows model inputs, key modules and the sources of supervisory signal. G and D denote the generator and the discriminator of a GAN, and SF/S subscripts are used to indicate networks generating/discriminating shadow-free/shadowed data.

have to be paired or aligned with the input sequences, and can represent any scene, including scenes which do not appear in the shadowed training data. The driving force of our GAN framework is the *unify-and-adapt* approach. The de-shadowed outputs are created by a unidirectional, two-branch network that attempts to map multiple differently shadowed versions of a scene to a uniform shadow-free output (the *unify* step). The style of this unified output domain is then adapted to that of the reference style via a discriminator which distinguishes between the generated and the real shadow-free samples (the *adapt* step). This approach helps us ensure good colour-consistency and overall quality of the reconstructions while exploiting both cross-scene and self-consistency information.

To sum up, the contributions of this paper are as follows:

1. We present a new *unify-and-adapt* self-supervised shadow removal model that achieves competitive scores on the ISTD and AISTD datasets without relying on cycle-consistency or domain-specific priors;
2. We demonstrate that S3R-Net achieves superior qualitative performance when contrasted with the best performing and most recent self-supervised shadow removal frameworks;
3. We prove the efficiency of the proposed system via a model parameter count and train-time GFLOPS comparisons between existing self-supervised shadow removal models (see: Fig. 5).

2. Literature review

Shadow removal is a not a new computer vision problem. Historically, the literature in the field looked at colour and illumination statistics to create physics-based solutions, *e.g.* [6, 8, 40]. Other works relied on user input to guide the shadow detection and removal steps [12, 14]. However, in recent years, we have observed the emergence of large-scale shadow removal datasets, coupled with the rising popularity of deep learning. These changes have led to the creation of a number of learning-based solutions that have produced state-of-the-art results in shadow removal and its sister task of shadow detection [44, 48, 49, 51, 59]. In this literature review, we will focus on the learnt de-shadowing frameworks and, in particular, their approach to reducing supervision requirements.

Learnt shadow removal. A number of works draw inspiration from physical illumination models that find the mapping between shadowed and shadow-free pixels [10, 28]. Such a function is used to over-expose the shadowed data so that its dark areas match the shadow-free regions. Then, the original and over-exposed images are blended to achieve a de-shadowed result. SP-Net [28] learns the shadow parameters and uses them to combine the natural and over-exposed shadowed data using a matte. Fu *et al*. [10] formulate the task as an auto-exposure fusion problem and smartly weigh a number of over-exposed shadowed regions to de-shadow the input.

Another group of papers exploits semantic scene understanding contained in pre-trained backbones. DeShadowNet [38] uses features from shallower and deeper layers of the VGG classifier [42] to decode appearance and semantic scene information that can be combined to guide shadow removal. Cun *et al*. [4] fuse the features with the input and use hierarchical feature aggregation to combine spatial attention with the information from earlier layers. Hu *et al*. [22] use CNN features to learn the direction-aware spatial context used to guide shadow removal. CANet [3] matches contextual patches between shadow-free and shadowed regions, and transfers the features at different scales from the former to the latter. In DeS3 [24] vision transformer features are used alongside attention and colour constancy constraints. PRNet [52] links the features with RNNs.

Generative Adversarial Networks (GANs) have also been a common choice for shadow removal. Wang *et al*. [47] stacked two GANs to detect the shadow and then use its mask as conditioning information for the removal step. RIS-GAN [56] consists of three parallel GANs for shadow removal as well as residual and illumination estimation. ARGAN [5] uses attention to recursively detect and remove shadows, making it robust to shadows of varying strength and complexity. SHARDS [39] deshadows low-resolution images and then uses them as guidance for full-resolution shadow removal. More recently, solutions using state-of-the-art techniques such as transformers [16, 54] or diffusion models [17, 32, 36] have also been proposed.

Some authors have also explored new ways of thinking about shadow removal. Li *et al*. [30] demonstrated that the task of inpainting is compatible with shadow removal, and linking the two decreases the prominence of shadow remnants. The authors also propose a system [31] relying on attentional fusion of 2 task branches – one for shadow-free region information relay and one for deshadowing. Wan *et al*. [45], on the other hand, pose shadow removal as a intra-frame style transfer problem.

Unlike the above, some systems do not require ground truth shadow-free images. Le and Samaras [29] build on SP-Net [28] and require only paired shadow masks to train their weakly-supervised model. These are used to crop out partly-shadowed and shadow-free patches from an image and limit the dependence on paired shadow-free data. In contrast, Liu *et al*. [34] create a train set by masking out shadowed and shadow-free areas. Their G2R-ShadowNet generates shadows, guided by the real shadow area masks, and then learns to de-shadow them using the shadow-free input regions. Zhong *et al*. [60] expand on this solution by improving the realism of the generated shadows. Guo *et al*. [18] use masks to train an image decomposition module being part of their diffusion-based shadow removal system.

The shadow removal literature is rich yet all of the above methods rely on ground truth in the form of shadow-free images and/or shadow masks. To improve generality, it is important for models to be trained on large-scale, real-world datasets. Unfortunately, capturing ground truth shadow-free data is error-prone and time-consuming. This is apparent in the existing shadow datasets [38, 47] which are known to contain slight scene framing and colour inconsistencies between different images corresponding to the same scene [22, 28, 43]. Obtaining the masks for real-life data is also laborious as it requires per-pixel annotations.

Un- and self-supervised frameworks Motivated by the discussed challenges, this paper aims to create a shadow removal network that does not require aligned ground truth data. Instead of relying solely on pixel-wise error computation between the input and the reference, the system should guide the shadow removal using adversarial losses and exploit other information present in the data.

A common approach to self-supervised learning relies on cycleGANs [61] and cycle-consistency training. The images used in this bidirectional process still come from two domains – shadowed and shadow-free – yet they no longer have to be paired, which lowers the data requirements. Mask-ShadowGAN [21] is the first self-supervised shadow removal framework and it operates based on the generic cycleGAN losses. Liu *et al*. build on this solution in LG-ShadowNet [33] and introduce two key changes: they first learn shadow removal in the L channel of the Lab colour space and then warmstart the all-channel network with those weights, and propose a vector-based colour loss. Vasluianu *et al*. [43] focus on the colour and pixel-wise inaccuracies in the existing datasets. To tackle them, they blur the inputs and outputs for colour-consistency enforcement, and rely purely on perceptual losses to control the content and style. DC-ShadowNet [23] expands the cycleGAN idea to soft and hard shadow understanding, and uses shadow domain classifiers alongside physics-based constraints.

Cycle-consistency is not the only way of reducing the supervision requirements. He *et al*. [19] propose an unsupervised portrait-specific solution that uses GAN inversion and leverages the generative facial priors embedded in the pre-trained StyleGAN2 [26]. While such priors are readily available for the popular portrait domain, the method cannot be easily interpolated to other more arbitrary problems.

In contrast, this paper follows a multi-branch *unify-and-adapt* approach to self-supervision. This avoids the dependency on a paired inverse task, and instead exploits the commonality of de-shadowed outputs across different shadowed training inputs. This removes the need to train an inverse task, improving compactness, efficiency and robustness of the system.

3. Methodology

In the following paragraphs we discuss the implementation details of the proposed method. First, we explain our ap-

proach to self-supervision. This is followed by a discussion of the losses driving our model. Finally, we provide more details on the implementation of the GAN used as part of our model.

3.1. The unify-and-adapt approach to self-supervision and other training losses

To limit the need for ground truth, we propose a self-supervised solution. The few existing domain-independent models that do not require paired data to train, all operate in a cycle-consistent (*i.e*. bidirectional) manner. In contrast, we build S3R-Net based on an emerging *unify-and-adapt* approach to self-supervision from the field of relighting, which uses a unidirectional, single-stage 2-branch network [27] (Fig. 2). In such a GAN-based architecture (described in more detail in Section 3.2) the input (shadowed) images $\mathbf{I}$ corresponding to the same scene are paired and fed into the generator $\mathcal{G}$ in parallel. The generator produces a shadow-correction residual that is added to the input before going through the final activation layer. The final reconstruction can therefore be described as $\hat{\mathbf{I}} = \mathbf{I} + \mathcal{G}(\mathbf{I})$. Using residuals is intended to limit the region of change within the image and keep the shadow-free areas intact. Instead of relying on a pixel-wise aligned ground truth, the system exploits the knowledge that the correct de-shadowed solution must be consistent across all differently shadowed versions of the input scene. Once the generator has learnt to enforce uniformity across the different variations of the input, the discriminator $\mathcal{D}$ helps to *adapt* this uniform output towards the correct output style through its adversarial losses (Eq. 10-11). Here, this target style is inferred from a collection of shadow-free samples $\mathbf{I}^*$ showing arbitrary scenes.

In addition to the control achieved through $\mathcal{D}$, we exploit a number of generator losses. As discussed in the Introduction, the available datasets have colour- and pixel-wise misalignments between the inputs and shadow-free equivalents of the same scene. With this in mind, we want to control our training using a mixture of pixel-wise and feature-based losses. The former provide strong guidance signals yet are prone to innate dataset errors. The latter are not as strong yet are more resilient to pixel-wise discrepancies.

The first pair of proposed losses controls the *unify* aspect of the *unify-and-adapt* approach, *i.e*. ensures that both outputs of our 2-branch network look the same, regardless of the initial shadows present. We enforce this using an L1 loss $\mathcal{L}_{os}$ defined as

$$\mathcal{L}_{os} = \|(\mathbf{I}_A + \mathcal{G}(\mathbf{I}_A)) - (\mathbf{I}_B + \mathcal{G}(\mathbf{I}_B))\|_1. \quad (1)$$

In the above equation, we use the A and B subscripts to refer to the images associated with each of the two network branches. It is also important to note that the output similarity loss is not applied directly to the shadow-correction residuals (*i.e*. $\mathcal{G}(\mathbf{I})$), but rather to the re-composited de-shadowed images.

The uniformity of the outputs is additionally controlled using the perceptual loss $\mathcal{L}_{perc}$. This compares the features extracted from both outputs using a pretrained VGG-19 backbone vgg. The loss can be formalised as

$$\mathcal{L}_{perc} = \sum^{i} \left\| vgg_i\left(\hat{\mathbf{I}}_A\right) - vgg_i\left(\hat{\mathbf{I}}_B\right) \right\|_1 \quad (2)$$

and *i* represents different feature scales within the network.

While $\mathcal{L}_{os}$ and $\mathcal{L}_{perc}$ aim to equate the outputs, any colour or framing differences between the paired images will affect the output quality. We can counteract this by preserving the information present in the shadow-free regions. To this end, we calculate a shadow mask $\mathbf{M}$ for each branch's input-output pair by applying Otsu tresholding [37] to a greyscale version of the images, *i.e*. $\mathbf{M} = Otsu(\mathbf{I} - \hat{\mathbf{I}})$. In the mask, 1s denote shadowed regions and 0s - shadow-free. To focus on the shadow-free areas, we invert the mask and obtain $\hat{\mathbf{M}}$. We then use $\hat{\mathbf{M}}$ to mask out the shadowed region in the input and the output, and compare the visible shadow-free areas. The resulting shadow-free region loss $\mathcal{L}_{sfr}$ is given as

$$\begin{aligned}\mathcal{L}_{sfr} = &\left\| \left(\hat{\mathbf{M}}_A \odot \hat{\mathbf{I}}_A\right) - \left(\hat{\mathbf{M}}_A \odot \mathbf{I}_A\right) \right\|_2 + \\ &\left\| \left(\hat{\mathbf{M}}_B \odot \hat{\mathbf{I}}_B\right) - \left(\hat{\mathbf{M}}_B \odot \mathbf{I}_B\right) \right\|_2,\end{aligned} \quad (3)$$

where $\odot$ symbolises the Hadamard product.

Even though comparing information from the output and its corresponding input is robust against pixel-wise discrepancies, there is potential for error stemming from imperfect mask calculation. Therefore, we also add a feature-based counterpart to $\mathcal{L}_{sfr}$ – a feature loss $\mathcal{L}_{feat}$, originally proposed in [23]. In their paper, Jin *et al*. conduct a study on features extracted from shadowed and shadow-free images of the same scene, and discovered that the features extracted at a particular network layer (Conv22 in the pretrained VGG-16 backbone – vgg_{22}) are the most shadow-invariant. Thus, we can extract features from the model input and the produced output, and ensure their feature-level consistency regardless of shadows. The aforementioned loss can be described as

$$\begin{aligned}\mathcal{L}_{feat} = &\left\| vgg_{22}\left(\hat{\mathbf{I}}_A\right) - vgg_{22}\left(\mathbf{I}_A\right) \right\|_1 + \\ &\left\| vgg_{22}\left(\hat{\mathbf{I}}_B\right) - vgg_{22}\left(\mathbf{I}_B\right) \right\|_1.\end{aligned} \quad (4)$$

While $\mathcal{L}_{sfr}$ is focused on the actual pixel values, $\mathcal{L}_{feat}$ is more concerned with general scene structure preservation, and the losses complement each other.

Finally, we want to prevent the shadow removal model from uniformly brightening the entire image. Therefore, we add a constraint that makes $\mathcal{G}$ de-shadow only the shadowed regions and leave the shadow-free areas intact. The

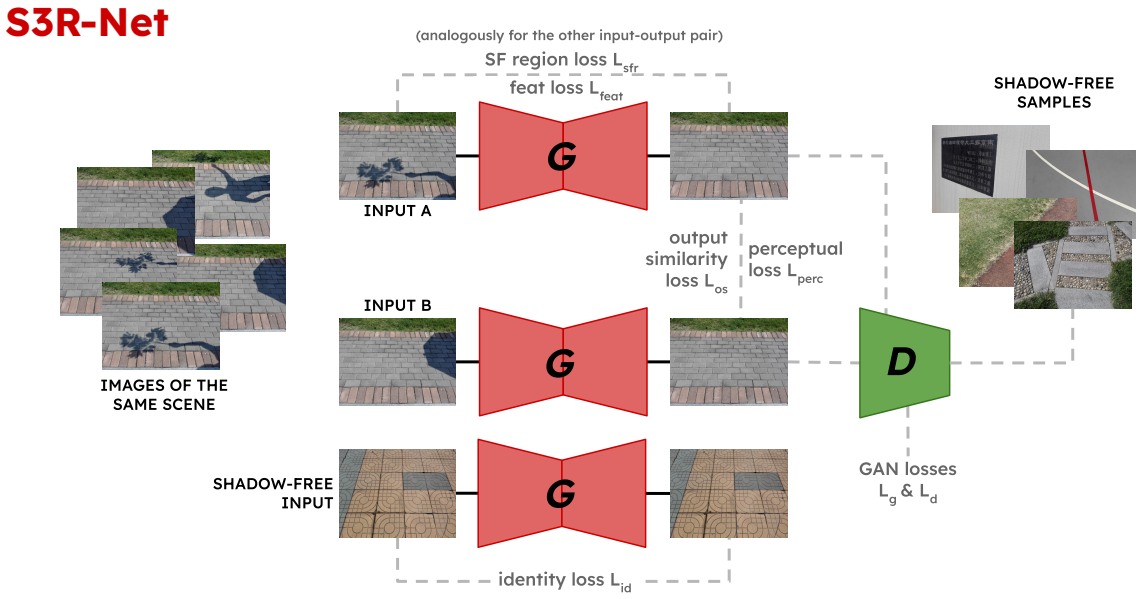

Figure 2. Our S3R-Net system and its losses. The generators (G) shown above are the same exact model with the same weights.

shadow-free region loss goes some way towards enforcing this. However, in the absence of ground truth shadow masks, $\mathcal{L}_{sfr}$ must rely on the network output and the calculated shadow masks $\hat{\mathbf{M}}$, which limits its robustness. We mitigate this by feeding $\mathcal{G}$ a shadow-free image $\mathbf{I}_{sf}$ and expecting it to produce a virtually empty residual, leading to a no-adjustment reconstruction $\hat{\mathbf{I}}_{sf}$. We enforce this using the identity loss

$$\mathcal{L}_{id} = \left\|\mathbf{I}_{sf} - \hat{\mathbf{I}}_{sf}\right\|_1. \tag{5}$$

We would like to emphasise that $\mathbf{I}_{sf}$ never represents the same scene as the $\{\mathbf{I}_A, \mathbf{I}_B\}$ input pair.

Since the losses described above serve different purposes and address different problems, we add scaling λ to each sub-loss. The total generator loss thus becomes

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_g + \lambda_{os}\mathcal{L}_{os} + \lambda_{perc}\mathcal{L}_{perc} + \\ \lambda_{sfr}\mathcal{L}_{sfr} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{id}\mathcal{L}_{id}.\end{aligned} \tag{6}$$

The 2-branch approach described in this section is necessary for loss calculation during training. However, at test time, the losses are no longer used and, thus, branch duplication is not required. Consequently, this means that during inference, we only need to feed a single image into a single $\mathcal{G}$ branch and there is no need to pair up the inputs.

3.2. The adversarial shadow removal model

As outlined above, the core of the proposed S3R-Net, shown in Fig. 2, is a GAN. Our system builds on the classic pix2pixHD model [50]. The framework uses a fully-convolutional encoder-decoder network as its generator and has a fully-convolutional multi-scale discriminator.

We train the aforementioned GAN as a Wasserstein GAN (WGAN) [1]. This approach brings in a few noteworthy changes. WGANs abandon the 0-1 (fake-real) labels of vanilla GANs [13]. Instead, their underlying loss metric, the Wasserstein distance $\mathbb{W}$ (*a.k.a.* Earth-mover's distance),

$$\mathbb{W}(\mathbf{I}, \mathbf{I}^*) = \mathbb{E}[\mathcal{D}(\mathcal{G}(\mathbf{I}))] - \mathbb{E}[\mathcal{D}(\mathbf{I}^*)] \tag{7}$$

can be understood as the minimum cost of aligning all of one distribution's samples with the other's.

To ensure the desired behaviour, we enforce 1-Lipschitz continuity on the discriminator function using a gradient penalty $\mathcal{E}_{gp}$ [15]. To calculate $\mathcal{E}_{gp}$, the discriminator is fed an image $\bar{\mathbf{I}}$ created by mixing real and generated examples with a weight sampled from a uniform distribution,

$$\bar{\mathbf{I}} = \epsilon\mathbf{I} + (1-\epsilon)\mathbf{I}^* \quad \text{where} \quad \epsilon \sim \mathcal{U}(0,1). \tag{8}$$

Then, the gradients w.r.t. the sample $\nabla_{\bar{\mathbf{I}}}$ are constrained to be close to 1, *i.e.*

$$\mathcal{E}_{gp} = \mathbb{E}[(\left\|\nabla_{\bar{\mathbf{I}}}\mathcal{D}(\bar{\mathbf{I}})\right\|_2 - 1)^2]. \tag{9}$$

With all of this in mind, the adversarial $\mathcal{G}$ and $\mathcal{D}$ losses - $\mathcal{L}_g$ and $\mathcal{L}_d$ - can be described as

$$\mathcal{L}_g = -\mathbb{E}[\mathcal{D}(\mathcal{G}(\mathbf{I}))] \quad \textit{and} \tag{10}$$

$$\mathcal{L}_d = \mathbb{E}[\mathcal{D}(\mathcal{G}(\mathbf{I}))] - \mathbb{E}[\mathcal{D}(\mathbf{I}^*)] + \lambda_{gp}\mathcal{E}_{gp}. \tag{11}$$

Unlike in a traditional GAN, WGAN's $\mathcal{G}$ and $\mathcal{D}$ are not trained for the same number of iterations. Multiple $\mathcal{D}$ iterations are performed for each $\mathcal{G}$ pass; we follow the official WGAN advice and set the $\mathcal{D}$:$\mathcal{G}$ iterations ratio to 5:1. Finally, while the pix2pixHD baseline uses InstanceNorm in

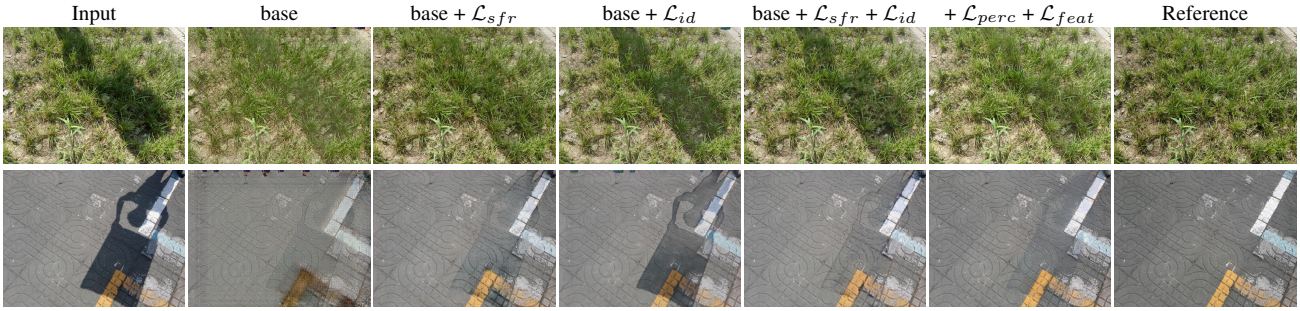

Figure 3. Ablation study: visual impact of S3R-Net's losses.

both the generator and discriminator, our S3R-Net instead uses learnable affine parameters with the normalisation layers in $\mathcal{D}$.

4. Experiments

The proposed S3R-Net was written in PyTorch. The loss scaling values mentioned in the Methodology were set as follows: λ_{gp} = 10, λ_{os} = 1, λ_{perc} = 2, λ_{sfr} = 5, λ_{feat} = 2 and λ_{id} = 1. The model was trained for 30 epochs and the best checkpoint from this range was chosen. Adam was used as the optimiser, with betas set to (0.0, 0.9). During training, we used the StepLR scheduler with an initial learning rate of 5×10^{-4}, a step of 10 and a gamma of 0.1.

For the ablation study (Sec. 4.1) and the first part of the experiments (Sec. 4.2), all models were trained and evaluated on full-size images (640×480) from the ISTD [47] dataset. The training set contains 1331 samples, yet due to our two-branch approach, we form approx. 10.3k training pairs. The test set consists of 540 images which are fed into our generator individually, as the technique does not require multiple inputs at test time. We also train and test our model on the adjusted ISTD (AISTD) dataset [28] (Sec. 4.3) with the same parameters.

In the following sections the performance of S3R-Net and other models is evaluated qualitatively and quantitatively. To follow the standard practise in the shadow removal domain, the numerical evaluation is presented in terms of RMSE calculated in the Lab colour space. However, we note that this is a mislabelling by previous authors, and that the evaluation script used by all cited authors actually calculates the error in terms of MAE (mean absolute error). This is a known, previously identified error in the literature, *e.g.* [9, 21]. To follow the standards and facilitate future comparisons, we also report MAE values but label them as RMSE in the relevant tables.

4.1. Ablation study

In this section we review the model's losses and demonstrate the compactness of the proposed S3R-Net.

4.1.1 Loss ablation study

We first evaluate the influence of each model loss on the overall system performance. During the ablation tests we do not remove the GAN losses - $\mathcal{L}_g$ and $\mathcal{L}_d$ - as well as the output similarity $\mathcal{L}_{os}$ as they are absolutely crucial to our model; we denote this case as 'base'. This is then expanded by gradually adding the other model losses. The differences stemming from each model change are shown in Fig. 3 and Table 1.

In the base case scenario, the error is spread between the shadow (S) and shadow-free (N) regions. The resulting images have some artefacts and their colours are slightly muted. The introduction of $\mathcal{L}_{sfr}$ leads to a significant drop in the shadow-free region error, which perfectly demonstrates the loss's purpose; the general quality also improves. $\mathcal{L}_{id}$ halves the shadow-free region error and leads to a slight increase in RMSE(S). The goal of identity loss is to prevent changes from being made to the shadow-free region. Since the loss is calculated on fully shadow-free samples, it does not introduce any awareness of shadows and, thus, might decrease performance in this area. Combining both of the newly introduced losses further improves the the RMSE(N) and results in an RMSE(S) score somewhere between the others'. Finally, we add the feature-based losses $\mathcal{L}_{perc}$ and $\mathcal{L}_{feat}$. The losses do not have a significant impact on the numerical results yet they improve the visual quality of the outputs and, in particular, decrease the appearance of shadow edges due to their robustness to misalignment (zoom in on bottom row). This may contribute to the slight drop in RMSE(S) in the final version of the model.

Table 1. Loss ablation study: impact of gradual loss addition on the performance of the proposed S3R-Net. A/A = "as above".

Method	RMSE(A)	RMSE(S)	RMSE(N)
base	13.17	15.84	12.82
base + $\mathcal{L}_{sfr}$	7.80	<u>12.59</u>	7.12
base + $\mathcal{L}_{id}$	7.66	16.91	<u>6.15</u>
base + $\mathcal{L}_{sfr}$ + $\mathcal{L}_{id}$	**7.09**	14.99	**5.94**
A/A + $\mathcal{L}_{perc}$ + $\mathcal{L}_{feat}$	<u>7.12</u>	**12.16**	6.38

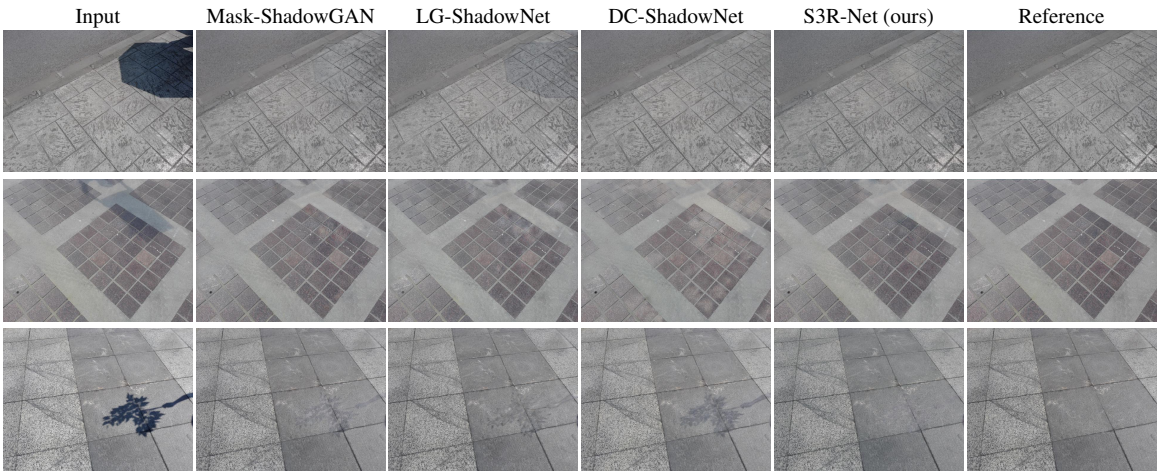

Figure 4. Visual results on the ISTD dataset.

4.1.2 Model compactness study

Recently, the growing accessibility of powerful GPUs has accelerated the development of top-accuracy models. However, this improvement is usually coupled with an increase in model size and computational requirements. Therefore, in this section, we wish to show that the proposed S3R-Net achieves good performance without excessively inflating the network.

The results of our study are visualised in Fig. 5. The graph plots the models' performance in terms of RMSE(A) vs a total number of train-time GFLOPS. The marker used to represent each solution has a radius corresponding to its total number of parameters. The presented values represent the GFLOPS/parameters of the generator and discriminator networks forming a given de-shadowing system.

In this comparison we consider 3 self-supervised, cycle-consistency based models: Mask-ShadowGAN [21], LG-ShadowNet [33] and DC-ShadowNet [23]. We take the numerical RMSE results for the first two models from their papers and run the outputs of the pre-trained DC-ShadowNet through the official evaluation script (the authors only report results on cropped data). Our S3R-Net is the least computationally expensive self-supervised shadow removal model reviewed here. The network also comes second in terms of RMSE(A), just after LG-ShadowNet. In terms of model parameters, the SqueezeNet-based LG-ShadowNet is closely followed by our S3R-Net and Mask-ShadowGAN. The most recent DC-ShadowNet is the largest and most computationally-heavy model in our evaluation.

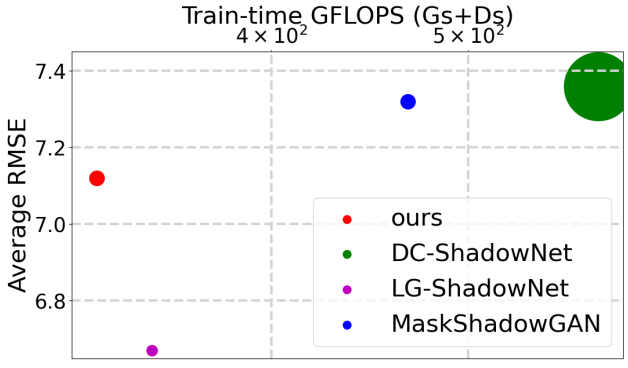

Figure 5. Model error vs train-time GFLOPS comparison between each model's generator(s)+discriminator(s) trained on full-size ISTD images. Circle radius is proportional to the total number of model parameters.

4.2. Results on ISTD

Next, we undertake a more in-depth comparison of our S3R-Net model with the relevant self-supervised state-of-the-art methods introduced in Sec. 4.1.2. The images used for the comparisons were provided by the authors (LG-ShadowNet) or generated using the published pre-trained models (the other 2 methods).

The results of this state-of-the-art experiment are presented in Table 2 and Fig. 4. Our model is outperformed only by LG-ShadowNet, which surpasses all other models numerically. However, when it comes to visual results, both it and Mask-ShadowGAN struggle with shadow edges. These are less prominent in outputs generated with DC-ShadowNet and even less so in our S3R-Net. We attribute this to the use of perceptual losses, taking care of the overall scene structure alongside the removal.

Additionally, we note that the competitors' models have a tendency to unnecessarily lighten darker image regions. This is particularly clear in the middle row of Fig. 4, where a number of floor tiles have their centres lightened (DC-ShadowNet) or small cloud-like areas appear (LG-ShadowNet). The issue does not seem to be present in S3R-Net outputs.

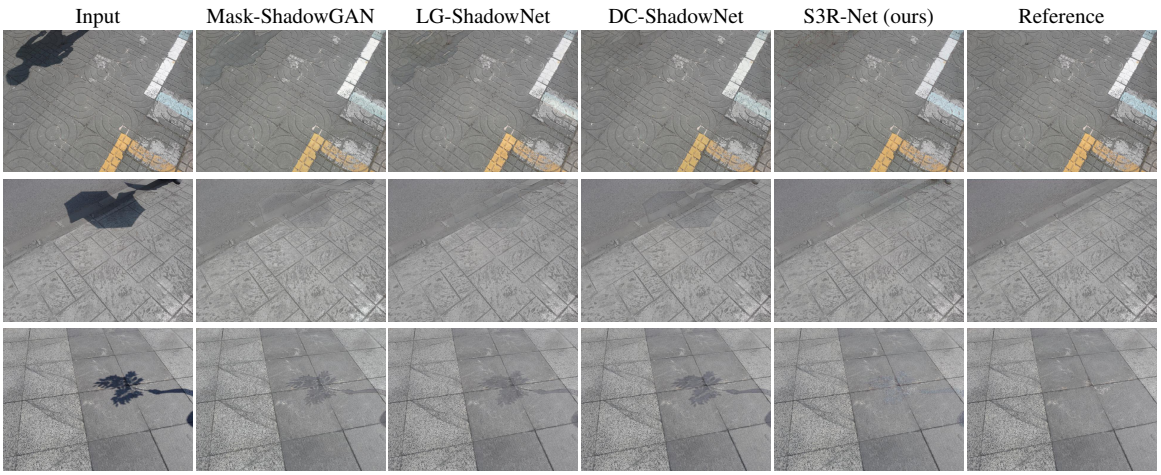

Figure 6. Visual results on the AISTD dataset.

Table 2. Results on the ISTD dataset (full-size images).

Method	RMSE(A)	RMSE(S)	RMSE(N)
Mask-ShadowGAN [21]	7.32	12.65	6.57
LG-ShadowNet [33]	**6.67**	11.63	**5.91**
DC-ShadowNet [23]	7.36	**11.21**	6.64
S3R-Net (ours)	7.12	12.16	6.38

4.3. Results on AISTD

We retrain our model and test it on the AISTD [28] dataset. Once again, we compare S3R-Net with SOTA on full-size images, *i.e.* 640×480 pixels. Mask-ShadowGAN was not trained on AISTD, so we retrain the model using the code from the official github repo. The numerical and visual results for LG-Shadow come directly from the authors' github and paper. Finally, DC-ShadowNet reports performance only on cropped data, so again we use their pre-trained model and the original evaluation script to get the numerical results for full-size inputs.

The results of this comparison are presented in Table 3 and Fig. 6. Just like before, LG-ShadowNet is the top performer in the AISTD comparison, with the remaining 3 models achieving similar performance. In Fig. 6 we can see that the models, including the top-performing LG-ShadowNet, suffer from similar faults as on the ISTD dataset: In the top 2 rows, the outputs generated by our competitors have clear shadow edges while our reconstructions have significantly smoother shadow boundaries. The appearance of the shadow fill is also most visibly reduced in the S3R-Net samples. Finally, our model can successfully detect and remove shadow in cases where other models fail (bottom row). Despite coming third quantitatively, we believe we have shown that S3R-Net can generate results that look the most pleasant to the human visual system.

Table 3. Results on the AISTD dataset (full-size images).

Method	RMSE(A)	RMSE(S)	RMSE(N)
Mask-ShadowGAN [21]	5.84	12.28	4.82
LG-ShadowNet [33]	**5.02**	**10.64**	**4.02**
DC-ShadowNet [23]	5.64	12.63	4.33
S3R-Net (ours)	5.71	12.86	4.43

5. Conclusions & future work

In this paper we presented S3R-Net – a shadow removal model with a novel *unify-and-adapt* approach to self-supervision. The proposed network achieves similar quantitative performance to the state-of-the-art self-supervised frameworks at a low computational cost. Additionally, and most importantly, we demonstrate superior qualitative performance: S3R-Net generates de-shadowed images with virtually imperceptible shadows, both in terms of edges as well as the inside fill, while maintaining the scene colours.

In future work, it would be interesting to further research the problem of colour adjustments in the shadow region. Adding the feature-based losses, we managed to address a common issue of leftover shadow edges. However, matching the colour of the de-shadowed region to the rest of the scene is still an ongoing problem – both for us and the other models. Moreover, the proposed S3R-Net does not need ground truth for training, yet the method still relies on paired input data. Therefore, we could investigate ways of further reducing our supervision requirements, *e.g.* by creatively using data augmentations. Finally, the (A)ISTD dataset exhibits some misalignment between shadowed and shadow-free samples. Our S3R-Net deals with this well due a mix of pixel-wise as well as perceptual losses. However, during model development we have not experimented with any higher levels of misalignment, so it is unclear how the network would perform given worse quality data.

This work was partially supported by the BBC and the EPSRC's iCASE project "Computational lighting in video" (voucher 19000034).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017. 5

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8), 2014. 1

[3] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[4] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

[5] 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 (ICCV)*, 2019. 3

[6] Mark S Drew, Graham D Finlayson, and Steven D Hordley. Recovery of chromaticity image free from shadows via illumination invariance. In *IEEE Workshop on Color and Photometric Methods in Computer Vision (ICCVW)*, 2003. 2

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1

[8] G.D. Finlayson and M.S. Drew. 4-sensor camera calibration for image representation invariant to shading, shadows, lighting, and specularities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2001. 2

[9] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal - official github repo. https://github.com/tsingqguo/exposure-fusion-shadow-removal. Accessed: 2022-03-03. 6

[10] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[11] Konstantinos Georgiadis, M Kerim Yucel, Evangelos Skartados, Valia Dimaridou, Anastasios Drosou, Albert Saa-Garriga, and Bruno Manganelli. Lp-ioanet: Efficient high resolution document shadow removal. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 1

[12] Han Gong and Darren Cosker. Interactive removal and ground truth for difficult shadow scenes. *JOSA A*, 33(9), 2016. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems (NIPS)*, 27, 2014. 5

[14] Maciej Gryka, Michael Terry, and Gabriel J. Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics*, 34(5), 2015. 2

[15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 5

[16] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 3

[17] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[18] Lanqing Guo, Chong Wang, Wenhan Yang, Yufei Wang, and Bihan Wen. Boundary-aware divide and conquer: A diffusion-based solution for unsupervised shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[19] Yingqing He, Yazhou Xing, Tianjia Zhang, and Qifeng Chen. Unsupervised portrait shadow removal via generative priors. In *ACM International Conference on Multimedia (ACM MM)*, 2021. 1, 3

[20] Ian P Howard. *Perceiving in depth, volume 1: basic mechanisms*. Oxford University Press, 2012. 1

[21] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-ShadowGAN: Learning to Remove Shadows from Unpaired Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 6, 7, 8

[22] 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*, 42(11), 2020. 3

[23] Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 3, 4, 7, 8

[24] Yeying Jin, Wenhan Yang, Wei Ye, Yuan Yuan, and Robby T Tan. Des3: Attention-driven self and soft shadow removal using vit similarity and color convergence. *arXiv preprint arXiv:2211.08089*, 2022. 3

[25] Seungjun Jung, Muhammad Abul Hasan, and Changick Kim. Water-filling: An efficient algorithm for digitized document shadow removal. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018. 1

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[27] Nikolina Kubiak, Armin Mustafa, Graeme Phillipson, Stephen Jolly, and Simon Hadfield. Silt: Self-supervised lighting transfer using implicit image decomposition. In *Pro-*

ceedings of the British Machine Vision Conference (BMVC), 2021. 4

[28] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 6, 8

[29] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[30] Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor Tsang, and Song Wang. Leveraging inpainting for single-image shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[31] Xiaoguang Li, Qing Guo, Pingping Cai, Wei Feng, Ivor Tsang, and Song Wang. Learning restoration is not enough: Transfering identical mapping for single-image shadow removal. *arXiv preprint arXiv:2305.10640*, 2023. 3

[32] Yuhao Liu, Zhanghan Ke, Ke Xu, Fang Liu, Zhenwei Wang, and Rynson Lau. Recasting regional lighting for shadow removal. In *Proceedings of the AAAI Conference on Artificial Inteligence (AAAI)*, 2024. 3

[33] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. Shadow Removal by a Lightness-Guided Network with Training on Unpaired Data. *IEEE Trans. on Image Process.*, 30, 2021. 1, 3, 7, 8

[34] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[35] Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T Freeman, and Michael Rubinstein. Omnimatte: associating objects and their effects in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[36] Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, and Vishal M Patel. Latent feature-guided diffusion models for shadow removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[37] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1), 1979. 4

[38] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau. DeshadowNet: A Multi-context Embedding Deep Network for Shadow Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3

[39] Mrinmoy Sen, Sai Pradyumna Chermala, Nazrinbanu Nurmohammad Nagori, Venkat Peddigari, Praful Mathur, BH Prasad, and Moonhwan Jeong. Shards: Efficient shadow removal using dual stage network for high-resolution images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 3

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

[41] Oleksii Sidorov. Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 1

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 3

[43] Florin-Alexandru Vasluianu, Andres Romero, Luc Van Gool, and Radu Timofte. Shadow removal with paired and unpaired learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 1, 3

[44] Tomas F. Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[45] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[46] Chunxiang Wang, Hanqing Xu, Zhiyu Zhou, Liuyuan Deng, and Ming Yang. Shadow detection and removal for illumination consistency on the road. *IEEE Transactions on Intelligent Vehicles*, 5(4), 2020. 1

[47] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 6

[48] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[49] Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, and Pheng-Ann Heng. Single-stage instance shadow detection with bidirectional relation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[50] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[51] Xiao Wang, Siyuan Yao, Pengwen Dai, Rui Wang, and Xiaochun Cao. Updated paired regions for shadow detection from single image. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. 2

[52] Yonghui Wang, Wengang Zhou, Hao Feng, Li Li, and Houqiang Li. Progressive recurrent network for shadow removal. *Computer Vision and Image Understanding*, 238, 2024. 3

[53] Qi Wu, Wende Zhang, and B.V.K. Vijaya Kumar. Strong shadow removal via patch-based shadow edge detection. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, 2012. 1

[54] Qianhao Yu, Naishan Zheng, Jie Huang, and Feng Zhao. Cnsnet: A cleanness-navigated-shadow network for shadow removal. In *Proceedings of the European Conference on Computer Vision (Workshops) (ECCVW)*, 2022. 3

[55] Edward Zhang, Ricardo Martin-Brualla, Janne Kontkanen, and Brian L Curless. No shadow left behind: Removing objects and their shadows using approximate lighting and geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[56] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

[57] Ruo Zhang, Ping-Sing Tsai, J.E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8), 1999. 1

[58] Xuaner Cecilia Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. Portrait Shadow Manipulation. *ACM Transactions on Graphics*, 2020. 1

[59] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. Distraction-aware shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[60] Yunshan Zhong, Mingbao Lin, Lizhou You, Yuxin Zhang, Luoqi Liu, and Rongrong Ji. Shadow removal by high-quality shadow synthesis. *arXiv preprint arXiv:2212.04108*, 2022. 3

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 3