LRA&LDRA: Rethinking Residual Predictions for Efficient Shadow Detection and Removal

Mehmet Kerim Yucel[1*] Valia Dimaridou[2*] Bruno Manganelli[1] Mete Ozay[1] Anastasios Drosou[2]
Albert Saà-Garriga[1]

[1]*Samsung Research UK*, [2]*CERTH ITI, Greece*

Abstract

The majority of the state-of-the-art shadow removal models (SRMs) reconstruct whole input images, where their capacity is needlessly spent on reconstructing non-shadow regions. SRMs that predict residuals remedy this up to a degree, but fall short of providing an accurate and flexible solution. In this paper, we rethink residual predictions and propose Learnable Residual Attention (LRA) and Learnable Dense Reconstruction Attention (LDRA) modules, which operate over the input and the output of SRMs. These modules guide an SRM to concentrate on shadow region reconstruction, and limit reconstruction of non-shadow regions. The modules improve shadow removal (up to 20%) and detection accuracy across various backbones, and even improve the accuracy of other removal methods (up to 10%). In addition, the modules have minimal overhead (+<1MB memory) and are implemented in a few lines of code. Furthermore, to combat the challenge of training SRMs with small datasets, we present a synthetic dataset generation pipeline. Using our pipeline, we create a dataset called PITSA, which has 10 times more unique shadow-free images than the largest benchmark dataset. Pre-training models on the PITSA significantly improves shadow removal (+2 MAE on shadow regions) and detection accuracy of multiple methods. Our results show that LRA&LDRA, when plugged into a lightweight architecture pre-trained on the PITSA, outperform state-of-the-art shadow removal (+0.7 all-region MAE) and detection (+0.1 BER) methods on the benchmark ISTD and SRD datasets, despite running faster (+5%) and consuming less memory (×150).

1. Introduction

Shadows are formed due to the interaction between occluder objects and light sources. Shadow intensity and location provide useful clues, such as lighting [36, 52, 66], geometry [32, 50, 59] and camera information [31], but can also harm various vision tasks, either due to poor visibility or shadow-induced *phantom* objects [4, 5, 9, 10, 41, 45, 46, 48, 75, 78]. Therefore, shadow detection and removal remain difficult yet important problems to solve.

Following earlier methods [3, 17, 47, 70, 73], deep learning approaches emerged for shadow detection [82, 81, 27, 79, 8, 49], removal [39, 11, 54, 16, 7, 76], or both [24, 12, 65]. We believe that there are two primary issues in the field; (i) the existing methods fail to focus on shadow regions, and (ii) the available datasets are quite small. The former leads to inefficient use of model capacity, whereas the latter harms generalization ability.

An intuitive fact is that non-shadow regions of shadow and shadow free images *should* be the same after removal. Therefore, reconstructing the whole input image during shadow removal (top diagram of Fig. 1) wastes model capacity on reconstructing the non-shadow regions. We also explore the models predicting residuals [44, 39, 54, 24], namely the *difference* between shadow and shadow-free image (middle diagram of Fig. 1). These methods are encouraged to concentrate on shadow regions only, but tend to produce sub-optimal results. In this paper, we rethink residuals in a stacked CNN paradigm [65] for jointly solving detection and removal; we propose Learnable Residual Attention (LRA) and Learnable Dense Reconstruction Attention (LDRA) modules, which operate over input and output of a shadow removal model (SRM) (bottom diagram of Fig. 1). LRA and LDRA i) guide the SRM to concentrate on shadow region reconstruction and ii) assist the final blending/color-correction. We leverage SRM's ability of concentrating on shadow regions, and use this as an additional supervision for the shadow detection model, which improves its accuracy. LRA&LDRA have negligible overhead, improves existing SRMs and works across various backbones.

We also propose a dataset generation pipeline (see Fig. 3) to address dataset size limitations. In addition to small real-life benchmarks [65, 54, 25], larger synthetic alternatives [29] are present but they have a limited number of unique shadow-free images, which limits their impact. We address this with an idea that scales gracefully; we collect images

*Equal contribution.

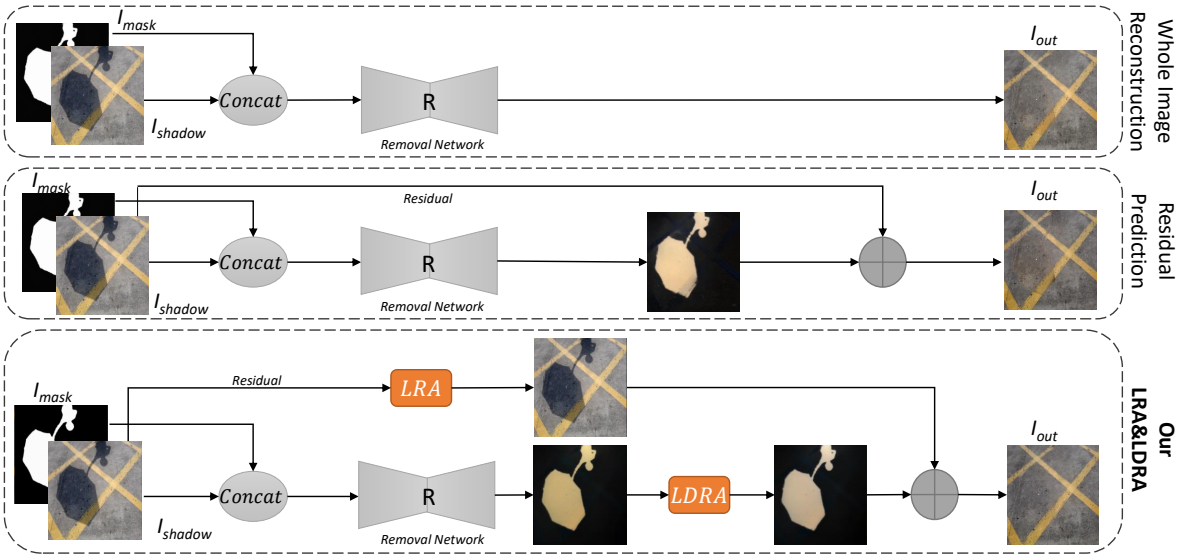

Figure 1. Different shadow removal approaches, where input image I_{shadow} and shadow mask I_{mask} are concatenated, and fed to a shadow removal model (SRM) R to produce the shadow free output I_{out}. Top diagram shows the vanilla approach where a given image is reconstructed (eqn. (2)). Middle diagram shows residual predictions (eqn. (3)), where the SRM is loosely encouraged to predict the *difference* between shadow and shadow-free images. Bottom diagram shows our LRA&LDRA modules (eqn. (4)) that guide the SRM to focus on reconstructing the shadow regions and perform blending/color-correction. For qualitative comparison, please see Fig. 5.

from various sources, extract shadow free patches by automatically filtering shadow regions and synthesize shadows on these patches. With this pipeline we create PITSA, formed of 172K triplets created from 20K unique shadow-free images. Our results show that pre-training models using PITSA significantly improves shadow removal and detection accuracy of various models. We believe that the PITSA is the next step for shadow detection/removal due to its scale and variety. Our contributions are as follows:

- We introduce LRA&LDRA that help guide an SRM to focus on shadow region reconstruction and perform blending/color-correction. LRA&LDRA bring up to 20% improvement over no-LRA&LDRA SRM baselines (including existing methods) with only $<$3ms runtime and $<$1MB memory overhead.
- We propose a new dataset generation pipeline and introduce PITSA [1], which is the largest shadow detection and removal dataset in the literature. Pre-training models using the PITSA introduces significant improvements in shadow removal (+2 MAE points on shadow regions) and detection (+0.1 BER) on the ISTD.
- We combine the PITSA and LRA&LDRA, and present a lightweight design for joint shadow detection and removal, which outperforms state-of-the-art in removal (+0.7 all-region MAE) and detection (+0.1 BER) on ISTD and SRD datasets, despite being significantly smaller ($\times$150 less memory) and faster (+5%).

[1]Our dataset will be made publicly available at https://terabox.com/s/1YQh2fc3SZ3prQZ1hJhejjQ

2. Related Work

Shadow detection. Early detection methods use physical models [51, 55, 62, 15], user input [77, 2], and hand-crafted features [19, 13, 28, 37, 80]. An end-to-end solution is [64], where SBU dataset is proposed. Nguyen *et al.* [49] extend the cGAN paradigm for detection. Le *et al.* [40] use a shadow attenuation network to augment data. Hu *et al.* [27] propose to learn global image context features in a direction-aware manner. Wang *et al.* [65] and Ding *et al.* [12] jointly address detection and removal. Recent studies leverage multi-task learning [8] and intensity-variant/invariant features [82].

Shadow removal. Early methods use illumination/color [3, 48, 14, 42, 68, 47, 58], user input [18, 17, 73] and hand-crafted features [15, 73, 71]. Supervised methods using paired data have attracted attention [57, 33]. Qu *et al.* [54] use localization, semantics and appearance features with a matting loss. Hu *et al.* extend their work to removal [81, 24]. Cun *et al.* [11] use a shadow matting network to create new data and then use hierarchical feature aggregation for removal. Fu *et al.* [16] cast shadow removal as an autoexposure fusion problem, whereas [39] propose a physics-based formulation using three networks. Chen *et al.* [7] present a two-stage method where they transfer non-shadow features to shadow features. Methods training on unpaired data have emerged as well [25, 38, 43, 30, 63].

Residual predictions in shadow removal. Many removal methods take in an image and reconstruct the whole image during removal. Since the difference (i.e. the residual) between shadow and shadow-free image is ideally only on

shadow regions (in practice, all real-life datasets have errors on non-shadow regions [39]), predicting that difference is a solid alternative. A naive summation approach is used in [24, 44, 39]. Ding *et al.* [12] predict residuals with a shadow attention detector, and then feed these to a removal encoder to iteratively remove shadows. Zhang *et al.* [76] use a multi-generator GAN to predict negative residuals, inverse illumination and coarse removal image. In our work, we build on existing residual prediction strategies.

Datasets. Earlier datasets, such as UCF [80] and UIUC [20] for removal, are quite small. ISTD [65] has 1.8K triplets (shadow image, mask and shadow-free image) for detection and removal. For detection, SBU [64] has 4.7K (shadow image and mask) pairs. SRD [54] provides 3K (shadow and shadow-free image) pairs for removal. USR [25] provides $\sim$ 4K unpaired images. The largest detection dataset has 10.5K pairs [26]. Synthetic datasets bypass the labeling and acquisition requirements, but suffer from limitations in shadow-free image variety (1.8K unique shadow-free images, with 10K shadow mattes [29]).

3. Rethinking Residual Predictions

We first introduce the shadow detection and removal tasks, motivate the need for better models to perform these tasks, and present LRA and LDRA modules which are integrated to baseline models to boost their accuracy.

3.1. Preliminaries

Problem formulation. In shadow detection, we employ a function $D(\cdot;\theta_D)$ parameterized by θ_D (implemented by a deep neural network (DNN) model D) to detect the location of shadow(s) observed on an image I_{shadow} by

$$I_{mask} = D(I_{shadow};\theta_D) \tag{1}$$

where I_{mask} is a binary mask representing the shadow location. In shadow removal (Fig. 1), a function $R(\cdot,\cdot;\theta_R)$ parameterized by θ_R (which is implemented using a DNN model R and depicted on the top diagram in Fig. 1) is used to obtain a shadow-free output image I_{out} by

$$I_{out} = R(I_{shadow}, I_{mask};\theta_R). \tag{2}$$

We argue that shadow removal is *localized* image-to-image translation, where *only* a part of I_{shadow}, localized by I_{mask}, will be translated to the target *shadow-free* domain. Most methods use I_{mask} as an additional input to the shadow removal model (SRM) (eqn. (2)), where it is concatenated to I_{shadow} as the fourth channel. This aims to condition the image translation on the masked area, but still leads to whole image reconstruction *in practice*. A way to see if SRMs focus only on shadow areas is to check their non-shadow region error; it should be zero (or equal to dataset error) if SRM focuses only on shadow regions. Table 1 shows the results for various methods on the ISTD dataset. Methods performing whole image reconstruction using eqn. (2) are off from the dataset error; this means these methods i) try to reconstruct the non-shadow region, and ii) they cannot perform this reconstruction accurately.

Residual predictions. A revised shadow removal task can be cast as only predicting *the difference* relative to the input image by (please see the middle diagram in the Fig. 1)

$$I_{out} = R(I_{shadow}, I_{mask};\theta_R) + I_{shadow}. \tag{3}$$

The methods using residual predictions eqn. (3) have lower error in Table 1, but still suffer from sub-optimal results.

3.2. LRA&LDRA Modules

In order to solve the aforementioned problems, we rethink/extend residual predictions of eqn. (3) via LRA and LDRA modules by (see the bottom diagram Fig. 1)

$$\begin{aligned} I_{out} = LDRA(R(I_{shadow}, I_{mask};\theta_R);\theta_{LRA}) \\ + LRA(I_{shadow};\theta_{LRA}) \end{aligned} \tag{4}$$

where $LDRA(\cdot;\theta_{LDRA})$ and $LRA(\cdot;\theta_{LRA})$ functions operate in image space. Note that eqn. 3 is a special-case of eqn. 4, where identity function is employed for both LRA&LDRA; this is what we refer as the *rethinking/extending* of residual predictions. Specifically, $LRA\&LDRA$ should i) guide $R(\cdot,\cdot;\theta_R)$ to focus on shadow regions and ii) perform blending/color-correction.

When implementing, we aim LRA and LDRA to i) be efficient, ii) have a strong spatial component to better guide R to perform localized translation, and iii) have a strong channel-wise component to transform output and input of R for blending. We implement LRA&LDRA using [23]

$$\begin{aligned} LRA(i,j;\theta_{LRA}) &\triangleq X(i,j) \odot g^h(i) \odot g^w(j), \\ LDRA(i,j;\theta_{LDRA}) &\triangleq X(i,j) \odot g^h(i) \odot g^w(j), \end{aligned} \tag{5}$$

where X denotes a matrix of an input image, i and j denote the spatial locations of a pixel, $\odot$ denotes element-wise multiplication, $\theta_{LDRA} \neq \theta_{LRA}$, and g^h and g^w are attention vectors with horizontal and vertical components. We omit the channel dimension for brevity. We implement LRA&LDRA with eqn. (5) and train them jointly with R.

Why [23]? We use [23] as it meets our criteria; i) minimal overhead, ii) strong spatial component due to global-pooling-free and direction-aware design and iii) ability to capture cross-channel information. Note that our formulation can use any implementation (learnable or hand-crafted) for LRA and $LDRA$, not just [23] (Sec. 5.3 for ablation).

How does LRA&LDRA guide R? Note that vanilla residual predictions guide R to focus on shadow regions already (see Fig. 1), and since LRA&LDRA extend them (LRA specifically, as it extends the residual), they inherently perform this guidance as well. We later show LRA&LDRA

Methods	*Original*	STC [65]	DSN [54]	MSG[25]	DCS[30]	LGS [43]	AE[16]	G2R[44]	SP-I-M[39]
Non-shadow	2.6	7.7	6.0	4.0	3.5	3.4	3.8	2.9	3.1

Table 1. MAE of the state-of-the-art methods on non-shadow regions using the Adjusted-ISTD. *Original* denotes the intrinsic error of A-ISTD. Residual predictions are used in [44, 39] and whole image reconstruction is performed in the others.

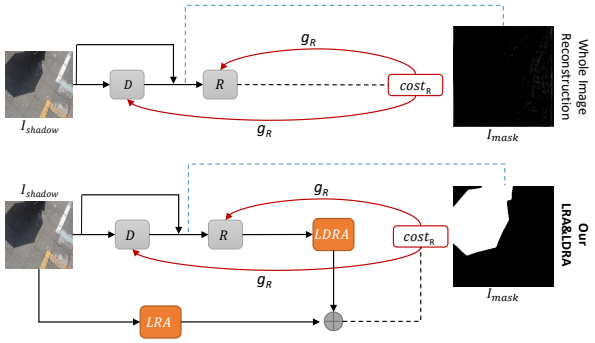

Figure 2. The positive effect of LRA&LDRA on the shadow detection model D, when D is *not detached* from R. Top diagram shows the whole image reconstruction where the gradients g_R of $cost_R$ with respect to R do not help D to improve its accuracy, because R does not predict localized outputs. Bottom diagram shows LRA&LDRA that force R to produce localized outputs, where the gradients g_R help D. I_{mask} shows outputs of the model D trained in a weakly-supervised manner by only using the gradients g_R (see supplementary material for details).

does a better job in guiding R to focus on shadow regions, and what these modules do individually (see Figure 4).

Blending and color correction. LRA&LDRA can be thought as a refinement network, where it operates over both the input and the output of R. This is in contrast with existing removal methods with secondary refinement networks [39, 44], which operate only on the output of R. We hypothesize (and later show) LRA&LDRA to be a much efficient alternative to such approaches.

3.3. LRA&LDRA Improves Shadow Detection

Preliminaries. Shadow removal without an explicit localization prior (i.e. mask) is possible, but we consider shadow detection as a necessary step. Therefore, similar to [65], we jointly train the models R and D. Assume that we have a cost function $cost$ (identified by the ℓ_1 loss) defined by

$$cost(y, y'; \theta) \triangleq \|y - y'\|_1 \tag{6}$$

where y denotes the shadow-free image I_{out} predicted by a model with parameters θ and y' denotes the ground-truth shadow-free image I_{free} for the shadow removal cost $cost_R$, and y denotes the predicted mask I_{mask} and y' denotes the ground-truth mask $I_{mask-gt}$ for the shadow detection cost $cost_D$. During training of the model R, gradients g_R of $cost_R$ with respect to parameters of R are calculated and used to update the parameters of R. Similarly, during the training of D, gradients g_D of $cost_D$ with respect to D are calculated and used to update the parameters of D. Since D and R are stacked, where the output of D is an input to R, a common practice is to *detach* the first model (in this case, D) when updating the parameters of the second model (in this case, R). In other words, the gradients g_R are not backpropagated to D; each model is updated separately with their own cost functions. Alternatively, *not detaching* means that the gradients g_R will be backpropagated to D and used to update the parameters of D; essentially, $cost_D$ will have an additional term in eqn. (6).

Benefits of LRA&LDRA in shadow detection. In the baseline removal model R implementing eqn. (2), the *detach* operation can be useful for D; D outputs I_{mask} which is a binary mask with shadow localization information, and R outputs the entire image I_{out} (see the top diagram of Fig. 1). These two images do not have much in common, especially in terms of the information they have. In contrast, when R implements eqn. (4) with LRA&LDRA, we conjecture that *not detaching* might be useful for D; in this case, R outputs a region that has the same localization information as the output I_{mask} of D (see bottom diagram of Fig. 1). We confirm the usefulness of *not detaching* D from R with LRA and LDRA in Section 5 by experimenting with and without detaching. See Fig. 2 for a detailed visualization. Also see supplementary material Section 3.8 for weakly-supervised detection experiments supporting the benefits of LRA&LDRA.

4. Large Scale Shadow Detection and Removal

In this section, we motivate the need for a new dataset and present our pipeline.

4.1. Motivation.

Real-life removal and detection datasets are small (10K [26] and 4K samples [25]), and even large synthetic datasets [29] have a limited amount of unique shadow-free images (1.8K). One can assume any image to be shadow-free to scale a dataset, but these images are likely to have shadows already and might lead to suboptimal models.

4.2. Our Proposed PITSA Dataset

We propose a new pipeline to find a working compromise; ability to leverage *any* image for dataset generation while keeping the noise (i.e. existing shadows) at minimum. We aim to create a dataset of triplets (I_{shadow}, I_{free}, $I_{mask-gt}$), corresponding to shadowed, shadow-free and shadow mask images. This is done via a two-stage process; shadow-free patch extraction and shadow superimposition.

Shadow-free patch extraction. First, we collect a database of images $\mathbb{D}$ from various sources. For each image $I_{src} \in \mathbb{D}$, we run a pre-trained shadow detection model M [8] (not to be confused with D) to obtain a shadow mask. This shadow mask is refined via a CRF [35] model (CRF) to

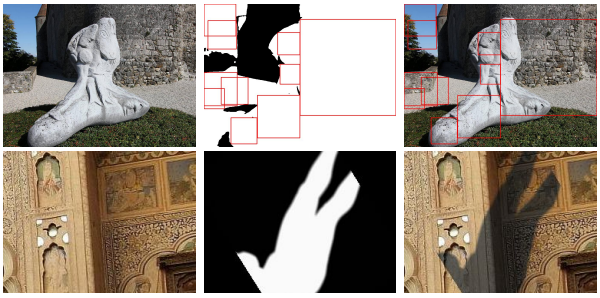

Figure 3. Our dataset creation pipeline. Top images show an input (left), predicted shadow mask (eqn. (7)), shadow-free patch candidates (red boxes) and resulting shadow-free image patches (right). Bottom images show an extracted shadow-free patch (left), sampled mask (middle) and synthesized shadow image (right).

	Ablation on different (LRA, LDRA)					
	$\mathbb{B}$	($\mathbb{1}$,$\mathbb{1}$)	(1-I_m,I_m)	([23],$\mathbb{1}$)	($\mathbb{1}$,[23])	([23],[23])
S ↓	7.94	8.69	**7.32**	7.73	8.45	7.54
NS ↓	3.20	2.66	2.97	2.71	**2.55**	**2.55**
All ↓	3.86	3.56	3.54	3.45	3.40	**3.29**
BER ↓	2.84	1.91	1.81	1.69	1.85	**1.56**

Table 2. Accuracy (MAE and BER) obtained for D and R, equipped with various (LRA, $LDRA$). I_m denotes the I_{mask}. The baseline $\mathbb{B}$ formulates R with eqn. (2). *S*, *NS* and *All* stand for shadow, non-shadow and all regions, respectively.

improve its precision, and it is thresholded by a function $thresh$ to remove low-confidence areas by

$$m_{src} = thresh(CRF(M(I_{src}), I_{src})). \quad (7)$$

We identify all the regions in m_{src} without shadows; we look for the largest square bounding boxes in the mask that do not intersect shadow pixels, and we filter out highly overlapping candidates according to intersection-over-union (IOU) metric and require a minimum patch size. In our pipeline, we set a minimum IOU threshold of 0.3, a minimum size of 128 pixels, and we extract up to 10 patches per input image. The coordinates of the resulting boxes are used to sample the input image, giving us N I_{free}, where N is the number of patches/images that meet the above criteria. The process is visualized in Fig. 3.

Shadow superimposition. We then proceed to generate a modified version of I_{free} images by applying a sequence of operations aimed at approximating the shadow region. Unlike [29], we do not attempt to only constrain the light model to ambient light. Instead, we additionally allow the following parameters to be altered: warmth, hue, saturation and lightness. The warmth is altered by modifying the red and blue channels of the image. For the latter, the image is first converted into HSV format, each channel is modified independently, and then it is converted back to its BGR format. We call the resulting output I_{dark}.

While the produced results could appear unrealistic under certain combinations, we are able to approximate different light colors and improve the model resilience to slight color variations in the shadow regions. Finally, we use a shadow mask database $\mathbb{D}^{mask}$ to obtain a shadow mask by

$$I_{mask-gt} = F(\sigma(\mathbb{D}^{mask})) \quad (8)$$

where F is a function that applies random transformations (flips and rotations) to the mask randomly sampled by $\sigma(\cdot)$. We use $I_{mask-gt}$ to determine the shadow image via the following blending operation

$$I_{shadow} = I_{dark} \odot I_{mask-gt} + I_{free} \odot (1 - I_{mask-gt}) \quad (9)$$

where $\odot$ denotes element-wise multiplication. The mask database is initially comprised of masks of [29], but during the generation, it is expanded by the masks m_{src} produced by M. At the end, we obtain over 20000 masks.

Dataset	Number of samples	Detection	Removal	Paired
ISTD [65]	1870	✓	✓	✓
SRD [54]	3088	×	✓	✓
USR [25]	4215	×	✓	×
SBU [64]	4087	✓	×	✓
CUHK-Shadow [26]	10500	✓	×	✓
PITSA (ours)	**172539**	✓	✓	✓

Table 3. Our PITSA dataset is the largest detection and removal dataset by a significant margin. It is also significantly diverse, both in terms of scenery and shadow mask shapes/locations. Number of samples refer to number of image pairs or triplets.

Discussion. Using HR-WSI [69] and MIT-Adobe-5K [6], we create PITSA (Patch Isolation Triplets with Shadow Augmentations); it consists of 172539 triplets created using 20416 unique images and over 20000 masks (see Fig. 3 for an example). Our pipeline is similar to [38, 29], but has key differences; i) [38] use small overlapping patches with fixed size, whereas we analyse entire masks and extract largest patches to include more context, ii) unlike [29], we generate new shadow-free patches and iii) do not limit ourselves to *realistic* shadow masks, we also use mask outputs of the filtering detection model, and further increase mask diversity. A limitation of our pipeline is the error of the shadow detector M. This can be improved by repeating the detection process or updating M with a better model. We show in Section 5 that the volume of data overcomes the potential noise. Table 3 shows that the PITSA dataset is the largest shadow detection and removal dataset by a large margin. See supplementary material for details on PITSA.

5. Experiments

5.1. Datasets and Evaluation Metrics

Datasets. We examine LRA&LDRA on the benchmark datasets ISTD [65] and SRD [54]. ISTD consists of 1870 image triplets (1330 train, 540 test) and is used for training & evaluation of models for detection and removal. We use the color-corrected version of ISTD test set (A-ISTD) [39]. SRD is formed of 3088 image pairs (2680 train, 408 test), and it is used for training & evaluation of models.

Evaluation metrics. To evaluate removal accuracy, we use MAE in the LAB space for shadow, non-shadow and all

	Ablation on different backbones					Ablation on existing methods		
	RNXt[72]-50	RNXt[72]-101	MNas[60]	EffNet[61]	Ghost[21]	STC[65]	G2R [44]	SP-I-M [39]
S ↓	**7.26** / 7.39	**7.11** / 7.27	8.09 / **7.86**	7.82 / **7.80**	**7.91** / 8.11	8.08 / **7.78**	10.3 / **9.90**	5.84 / **5.55**
NS ↓	3.33 / **2.87**	3.13 / **2.90**	3.45 / **2.89**	3.88 / **2.63**	3.70 / **3.11**	3.87 / **3.47**	3.87 / **3.81**	2.59 / **2.57**
All ↓	3.91 / **3.58**	3.73 / **3.58**	4.11 / **3.62**	4.46 / **3.39**	4.37 / **3.77**	4.53 / **4.16**	4.84 / **4.70**	3.11 / **3.01**
BER ↓	1.99 / **1.94**	**2.09** / 2.20	2.26 / **2.20**	2.06 / **1.61**	**2.60** / 2.77	3.88 / **3.65**	–	–

Table 4. Accuracy obtained with different backbones (columns 1 to 5) and existing methods (columns 6 to 8), with (right) and without (left) using LRA&LDRA. First five columns use backbones pre-trained on the ImageNet for implementing R and D. We retrain all existing methods [65, 44, 39] with official code, when available. For [44], we evaluate images with 480×640 resolution, following the official codebase. The model proposed in [39] was trained using GT masks.

regions; we note that although many methods claim to report RMSE, they actually report MAE [1]. For detection, we use the balanced error rate (BER). Images are resized to 256×256 for evaluation. Unlike [82, 27, 40, 8, 79], we do not post-process the masks predicted by D.

5.2. Architecture and Implementation Details

Network architecture. LRA&LDRA can be plugged into any model, but as our primary solution, we use the architecture of [74], which is an efficient dense prediction network based on MobileNetv2 [56] and FBNet [67]. We use this architecture, as our preliminary experiments show it has a good efficiency/accuracy trade-off. We use the same architecture for both R and D, where the differences between the two are the number of input channels (3 and 4 for D and R, respectively), and R having LRA&LDRA.

Implementation details. We initialize the encoders of R and D with ImageNet-pretrained weights, and the rest with [22]. Models are jointly trained using PyTorch [53] for 2K epochs with batch size 16, where learning rates for both are set to 2e-4. Images are resized to 286×286, randomly cropped to 256×256 and augmented (random horizontal flipping). Adam [34] optimizer is used with ℓ_1 loss for training both models. We use early stopping with a validation split from the training set (20% hold-out ratio). For the ISTD and SRD, we train the models separately. When indicated, the models are pre-trained for 350 epochs on the PITSA using the same hyperparameters.

5.3. Ablation Studies

We conduct detailed ablation studies to show the effect of several components of our method on accuracy. We experiment on the ISTD dataset using the lightweight architecture described in Section 5.2 (unless stated otherwise).

Component Analyses of LRA&LDRA. Fig. 4 shows the outputs obtained at each stage of our pipeline. The addition of LRA does not seem to change much (spatially) visually, but it improves the shadow region accuracy (see Table 2); it prepares the input for blending with R_{out}. Note the ineffective blending without LRA in rows 2 and 3 in Fig. 4; artefacts are visible in the final outputs. Furthermore, as mentioned earlier in the paper, the addition of LRA&LDRA guides R to produce a localized results. Note that R_{out} is sharpest with LRA&LDRA, showing that it does the guidance better than vanilla residual predictions. Another effect of LDRA, as seen in $LDRA_{out}$, is color correction; the shape is the same with R_{out} but LDRA *refines* the colors, making them suitable for the final blending. Finally, note that I_{mask} is sharpest in LRA&LDRA, verifying the usefulness of the gradients provided by R in the training of D.

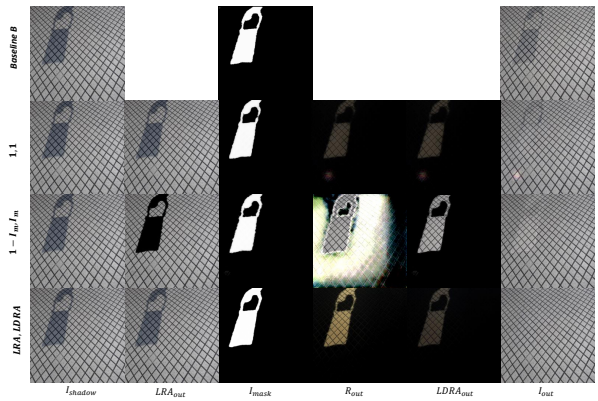

Figure 4. Rows show the output obtained using baseline w/o residuals, vanilla residuals (i.e. identity for LRA&LDRA), (1- I_{mask}. I_{mask}) and the final LRA&LDRA. Columns show stage-wise outputs of our method; input, output of LRA, predicted mask, output of the SRM R, output of the LDRA and the final result. Our LRA&LDRA produce sharper masks, removal outputs and artefact-free results.

Choosing LRA&LDRA. We test alternatives for implementing $LDRA$ and LRA, such as the identity function $\mathbb{1}$, I_{mask} predicted by D and coordinate attention [23]. Table 2 shows that, regardless of the type of LDRA&LRA, both D (+1 BER) and R (+0.6 MAE) are improved. LRA and LDRA modules (with [23]) introduce improvements individually, and even more so when they are used together, which justifies their presence. Note that the best shadow region accuracy comes from choosing I_{mask}, whereas the best overall accuracy is obtained from LRA&LDRA that use coordinate attention [23]. This shows the flexibility of our approach, where different functions can be chosen for different goals. See supplementary material Section 3.2 for an extended version of this ablation.

LRA&LDRA as plug-and-play. Table 4 shows the comparisons of different methods and backbones with and without using LRA&LDRA. The first five columns show that across different backbones, LRA&LDRA improve over-

	Ablation on D		Ablation on Pre-training on PITSA		
	($\mathbb{B}$, $\mathbb{B}$†)	(LL, LL†)	LL	SP-I-M [39]	STC[65]
S ↓	7.77 / **7.43**	**7.53** / 7.82	(7.54, **5.67**)	(5.84, **5.02**)	(8.08, **6.49**)
NS ↓	**3.14** / 3.28	**2.49** / 2.76	(2.55, **2.40**)	(2.59, **2.48**)	(3.87, **2.77**)
All ↓	**3.87** / 3.94	**3.28** / 3.56	(3.29, **2.91**)	(3.11, **2.85**)	(4.53, **3.30**)
BER ↓	2.55 / **2.04**	**1.82** / 1.93	(1.56, **1.47**)	–	(3.88, **2.01**)

Table 5. Accuracy of D and R across various experiments. The first two columns show results where the model D is detached (marked by †) from the model R, where R implements the baseline (B) (eqn. (2)) or our LRA&LDRA (LL). Columns 3 to 5 show results where different methods are pretrained on our PITSA dataset (right) or not (left).

all and non-shadow MAE consistently, whereas they improve shadow MAE and BER in most cases. The last three columns show that existing methods are consistently improved with LRA&LDRA (+0.4 MAE).

Improved shadow detection with LRA&LDRA. Tables 2 and 4 show that LRA&LDRA improve accuracy of D. We also experiment with and without LRA&LDRA where the model D is *detached*, or *not detached*, from R. The first two columns of Table 5 show that *detaching* R from D improves the accuracy of D for whole image reconstruction (eqn. (2)), but is detrimental for D and R with LRA&LDRA. This supports our claim in Section 3.3 that with LRA&LDRA, backpropagating the gradients g_R to D improves D.

Pre-training models using the PITSA. Columns 3 to 5 of Table 5 show that pre-training models on PITSA significantly improves (+1.2 MAE *all*) accuracy of all methods. In non-shadow regions, improvements are slight as both [39] and us are already close to dataset error. However, shadow regions are improved significantly. In detection, accuracy is improved slightly with LRA&LDRA, and that of [65] is improved greatly (+1.8 BER). We credit this to LRA&LDRA having improved detection already, so there is a *smaller* room for improvement compared to [65]. Furthermore, PITSA outperforms the (previously) largest synthetic dataset (see supplementary material Table 1).

5.4. Comparison with State-of-the-art

Shadow removal. We compare LRA&LDRA with handcrafted methods [73, 17], ST-CGAN [65], DHAN [11], DeShadowNet [54], G2R-ShadowNet [44], SP-I-M [39], DC-ShadowNet [30] and several other methods.

Table 6 shows removal accuracy for the SRD test set. Our methods outperform others considering accuracy on the shadow regions and overall accuracy. Furthermore, pre-training on the PITSA dataset shows significant accuracy improvements. Table 7 shows shadow removal accuracy for the A-ISTD. Our methods (LL) produce the best non-shadow (2.5 MAE) and overall (3.2 MAE) accuracy, which shows that LRA&LDRA enable models to focus on shadow removal rather than reconstructing non-shadow regions. In shadow regions, our methods perform quite competitively, despite others using significantly larger networks for removal. We note that LRA&LDRA actually improve these methods as well (last two columns of Table 4).

	Input	ISR[17]	DSC[24]	DSN[54]	DCS[30]	$\mathbb{B}$	LL	LL †
S ↓	37.4	25.4	8.8	3.5	7.7	7.6	**7.5**	**6.5**
NS ↓	3.9	6.9	**3.2**	8.8	3.4	4.0	3.5	**3.4**
All ↓	13.7	12.3	4.8	5.1	4.6	4.8	**4.4**	**4.0**

Table 6. Removal performance on SRD. $\mathbb{B}$ denotes the baseline method (described in eqn. (2)). † indicates pre-training models on the PITSA. LL indicates LRA&LDRA.

LRA&LDRA, once pre-trained on our PITSA dataset, significantly outperform (+0.7 all-regions MAE) others.

Shadow detection. We compare our model D against MTMT [8], DSD [79], stacked-CNN [64], scGAN [49], ST-CGAN [65], DSC [24], BD-RAR [81] and FDRNet [82]. Table 8 shows that FDRNet [82] outperforms us by only 0.01 BER despite performing post-processing [35]. Once pre-trained on PITSA, our method outperforms FDRNet (+0.1 BER). These results verify the usefulness of LRA&LDRA in improving detection accuracy. See supplementary material for qualitative results.

5.5. Qualitative Results and Discussions

Qualitative results. Fig. 5 shows that LRA&LDRA barely affect the non-shadow regions. This is especially visible in the first, fourth and fifth rows. Even high complexity SP-I-M [39] attempt to recover the non-shadow regions and fail, when non-shadow regions are complex. Our models are competitive on shadow regions; they produce minimal ghosting and consistent colors. Finally, all rows show the significant effect of PITSA pre-training; LRA&LDRA pretrained on the PITSA significantly outperform all others, both in shadow and non-shadow regions. See supplementary material for more results using in-the-wild images.

Analysis of performance. Table 9 shows that LRA&LDRA have minimal overhead; 0.1 MFLOPs complexity, 0.7MB memory and 2.7ms runtime (0.5ms for [39]). We also show that LRA&LDRA plugged into our architecture (Table 9 second column) is smaller, faster and consumes less memory compared to others, despite outperforming them. Compared to [39], our method is faster (5%), smaller ($\times$150 less memory) and more accurate despite performing detection and removal jointly.

Why not just..copy-paste (CP) the non-shadow region? A natural question against LRA&LDRA is; instead of these, why don't we just copy-paste the non-shadow region from the input image to the output image? In ideal cases, CP might provide 0 MAE on non-shadow regions. In practice, however, our solution is better due to several reasons; i) CP relies on the availability of perfect shadow masks, which may not be feasible ii) LRA&LDRA performs blending/color correction, iii) LRA&LDRA provides robustness to minor mask errors and iv) also improves shadow detection performance. It is also plausible to use CP and then do blending, but LRA&LDRA is an end-to-end, learnable alternative that can scale with data/capacity. See supplementary material for further details and discussions.

	Input	SRB[73]	ISR[17]	STC[65]	DHN[11]	DSN[54]	PRS[20]	DSC[24]	G2R[44]	SP-I-M[39]	AE[16]	BMN [83]	$\mathbb{B}$	LL	LL †
S ↓	40.2	24.7	13.3	13.4	11.4	15.9	22.0	7.6	7.3	**6.0**	6.5	6.1	7.9	7.5	**5.6**
NS ↓	2.6	14.4	2.6	7.7	7.2	6.0	3.1	3.2	2.9	3.1	3.8	2.9	3.2	**2.5**	**2.4**
All ↓	8.5	16.0	4.3	8.7	7.9	7.6	6.1	3.9	3.6	3.6	4.2	3.5	3.8	**3.2**	**2.9**

Table 7. Removal performance on A-ISTD. $\mathbb{B}$ is the baseline method. † indicates pre-training on the PITSA. LL indicates LRA&LDRA.

	SCN[64]	SCG[49]	STC[65]	BDR[81]	DSC[24]	DSD[79]	MTM[8]	FDR[82]	LL	LL †
BER ↓	8.60	4.70	3.85	2.69	3.42	2.17	1.72	**1.55**	1.56	**1.47**

Table 8. Detection performance on A-ISTD. † indicates pre-training on the PITSA. LL indicates LRA&LDRA.

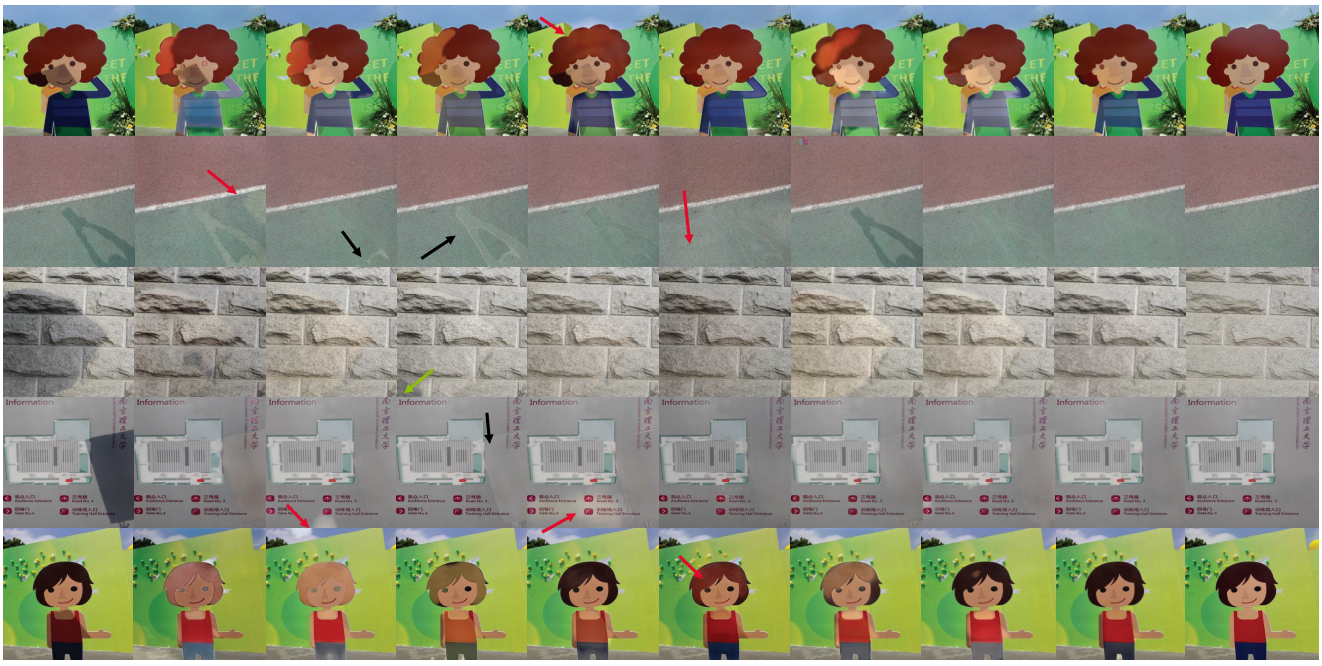

Figure 5. Qualitative comparison of LRA&LDRA and others. * indicates models pre-trained on the PITSA. We highlight some examples where alternatives unnecessarily alter non-shadow areas (→), miss shadow areas (→) or produce artefacts in shadow areas (→). Note that LRA&LDRA barely touches non-shadow areas and has competitive shadow area performance. Best viewed when zoomed in.

	$\mathbb{B}$	LL	STC[65]	STC[65] †	SP-I-M[39]	SP-I-M[39] †	G2R[44]	G2R[44] †
Runtime (ms)	37	39.7	280	281.4	40.4	40.9	116	118.9
Memory (GB)	0.061	0.062	2.74	2.74	10.5	10.05	0.141	0.142
FLOPs (G)	0.6835	0.6836	721.2651	721.2652	25.8018	25.8019	94.5756	94.5757
ISTD MAE (All)	3.86	3.29	4.53	4.16	3.60	3.49	4.84	4.70

Table 9. The overhead of LRA&LDRA (LL). $\mathbb{B}$ is the baseline method (eqn. (2)). Methods with † are trained with LRA&LDRA. FLOPs are for removal networks only. Memory and runtime values are for D and R for the first four columns, and for R only for the last four. Measurements are performed with an RTX 3090 using PyTorch. Note that the overhead brought by LL (and †) is negligible.

6. Conclusions

We address shadow detection and removal tasks; we rethink residual predictions with LRA&LDRA modules that operate over the input and output of a shadow removal model. These modules guide the model to concentrate on shadow regions, and perform color-correction and blending. Our experiments show that LRA&LDRA achieve state-of-the-art accuracy on detection & removal with a significantly smaller and faster network. LRA&LDRA work across various backbones and even improve existing methods. Finally, we propose a new dataset generation pipeline and the PITSA dataset for detection & removal, which is 10 times more diverse than the largest dataset. Our results show that pre-training models on the PITSA further improves LRA&LDRA and other methods significantly.

References

[1] Official GitHub page of "shadow removal via shadow image decomposition", ICCV 2019. `https://github.com/cvlab-stonybrook/SID`.

[2] Eli Arbel and Hagit Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1202–1216, 2010.

[3] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2(3-26):2, 1978.

[4] K Berker Logoglu, Hazal Lezki, M Kerim Yucel, Ahu Ozturk, Alper Kucukkomurler, Batuhan Karagoz, Erkut Erdem, and Aykut Erdem. Feature-based efficient moving object detection for low-altitude aerial platforms. In *Proceedings*

of the IEEE International Conference on Computer Vision Workshops, pages 2119–2128, 2017.

[5] Yunus Can Bilge, Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, and Pinar Duygulu. Red carpet to fight club: Partially-supervised domain transfer for face recognition in violent videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3358–3369, 2021.

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[7] 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*, pages 4743–4752, 2021.

[8] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5611–5620, 2020.

[9] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence*, 25(10):1337–1342, 2003.

[10] Rita Cucchiara, Costantino Grana, Massimo Piccardi, Andrea Prati, and Stefano Sirotti. Improving shadow suppression in moving object detection with hsv color information. In *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585)*, pages 334–339. IEEE, 2001.

[11] 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*, volume 34, pages 10680–10687, 2020.

[12] 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.

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

[14] 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.

[15] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):59–68, 2005.

[16] 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*, pages 10571–10580, 2021.

[17] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC*, pages 1–11. Citeseer, 2014.

[18] Maciej Gryka, Michael Terry, and Gabriel J Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics (TOG)*, 34(5):1–15, 2015.

[19] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR 2011*, pages 2033–2040. IEEE, 2011.

[20] 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, 2012.

[21] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1580–1589, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[23] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13713–13722, 2021.

[24] 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):2795–2808, 2019.

[25] 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*, pages 2472–2481, 2019.

[26] Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing*, 30:1925–1934, 2021.

[27] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018.

[28] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *2011 international conference on computer vision*, pages 898–905. IEEE, 2011.

[29] Naoto Inoue and Toshihiko Yamasaki. Learning from synthetic shadows for shadow detection and removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4187–4197, 2020.

[30] 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*, pages 5027–5036, 2021.

[31] Imran N Junejo and Hassan Foroosh. Estimating geo-temporal location of stationary cameras using shadow trajectories. In *European conference on computer vision*, pages 318–331. Springer, 2008.

[32] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011.

[33] Salman H 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(3):431–446, 2015.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[35] John D Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[36] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating natural illumination from a single outdoor image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 183–190. IEEE, 2009.

[37] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *European conference on computer vision*, pages 322–335. Springer, 2010.

[38] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *European Conference on Computer Vision*, pages 264–281. Springer, 2020.

[39] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021.

[40] Hieu Le, Tomas F Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+ d net: Training a shadow detector with adversarial shadow attenuation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 662–678, 2018.

[41] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9039–9048, 2018.

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

[43] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing*, 30:1853–1865, 2021.

[44] 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*, pages 4927–4936, 2021.

[45] Chengjiang Long and Gang Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2839–2847, 2015.

[46] Chengjiang Long and Gang Hua. Correlational gaussian processes for cross-domain visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 118–126, 2017.

[47] Ankit Mohan, Jack Tumblin, and Prasun Choudhury. Editing soft shadows in a digital photograph. *IEEE Computer Graphics and Applications*, 27(2):23–31, 2007.

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

[49] Vu Nguyen, Tomas F Yago Vicente, 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.

[50] Takahiro Okabe, Imari Sato, and Yoichi Sato. Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1693–1700. IEEE, 2009.

[51] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. Illumination estimation and cast shadow detection through a higher-order graphical model. In *CVPR 2011*, pages 673–680. IEEE, 2011.

[52] Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. Simultaneous cast shadows, illumination and geometry inference using hypergraphs. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):437–449, 2012.

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[54] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4067–4075, 2017.

[55] Elena Salvador, Andrea Cavallaro, and Touradj Ebrahimi. Cast shadow segmentation using invariant color features. *Computer vision and image understanding*, 95(2):238–259, 2004.

[56] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

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

[58] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum*, volume 27, pages 577–586. Wiley Online Library, 2008.

[59] Kalyan Sunkavalli, Todd Zickler, and Hanspeter Pfister. Visibility subspaces: Uncalibrated photometric stereo with shadows. In *European Conference on Computer Vision*, pages 251–264. Springer, 2010.

[60] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

[61] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[62] Jiandong Tian, Xiaojun Qi, Liangqiong Qu, and Yandong Tang. New spectrum ratio properties and features for shadow detection. *Pattern Recognition*, 51:85–96, 2016.

[63] 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 (CVPR) Workshops*, pages 826–835, June 2021.

[64] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *European Conference on Computer Vision*, pages 816–832. Springer, 2016.

[65] 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.

[66] 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*, pages 1880–1889, 2020.

[67] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019.

[68] Tai-Pang Wu, Chi-Keung Tang, Michael S Brown, and Heung-Yeung Shum. Natural shadow matting. *ACM Transactions on Graphics (TOG)*, 26(2):8–es, 2007.

[69] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 611–620, 2020.

[70] Chunxia Xiao, Donglin Xiao, Ling Zhang, and Lin Chen. Efficient shadow removal using subregion matching illumination transfer. In *Computer Graphics Forum*, volume 32, pages 421–430. Wiley Online Library, 2013.

[71] Yao Xiao, Efstratios Tsougenis, and Chi-Keung Tang. Shadow removal from single rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3011–3018, 2014.

[72] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[73] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image processing*, 21(10):4361–4368, 2012.

[74] Mehmet Kerim Yucel, Valia Dimaridou, Anastasios Drosou, and Albert Saa-Garriga. Real-time monocular depth estimation with sparse supervision on mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2428–2437, 2021.

[75] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Xin Yang, Haiyang Mei, and Baocai Yin. Multi-context and enhanced reconstruction network for single image super resolution. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[76] 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*, volume 34, pages 12829–12836, 2020.

[77] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015.

[78] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):611–624, 2018.

[79] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Distraction-aware shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2019.

[80] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *2010 IEEE Computer Society conference on computer vision and pattern recognition*, pages 223–230. IEEE, 2010.

[81] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018.

[82] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson WH Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4702–4711, 2021.

[83] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5627–5636, 2022.