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

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*

1. Outline

This supplementary material presents additional content for our paper with the title "LRA&LDRA: Rethinking Residual Predictions for Efficient Shadow Detection and Removal". In Section 2, we provide additional details on our PITSA dataset. In Section 3, we provide additional results and analyses of experiments on our proposed LRA&LDRA modules which are presented in the main text.

2. Details on the PITSA Dataset

We first provide additional details on the pipeline used to generate the PITSA dataset, and then provide detailed statistics about the PITSA dataset.

2.1. Further Details on Generating Shadow Augmentation

In Section 4 of our manuscript, we provide details on the shadow superimposition stage of our dataset creation pipeline, and mention the parameters we alter to approximate the shadow areas; warmth, hue, saturation and darkness. Below, we further expand the operations we perform (i.e. to acquire I_{dark} from I_{free}, as mentioned in the manuscript). We use the following operations to alter the warmth

$$\begin{aligned} I^R_{cool} &= \alpha \times DC(I^R_{free}) + (1-\alpha) \times I^R_{free} \\ I^G_{cool} &= I^G_{free} \\ I^B_{cool} &= \alpha \times IC(I^B_{free}) + (1-\alpha) \times I^B_{free} \end{aligned} \tag{1}$$

where $\times$ denotes element-wise multiplication, I_{free} is a shadow-free image/patch extracted at stage 1 (shadow-free patch extraction via shadow detection model) of our pipeline, superscripts R, G, B denote individual color channels of the relevant image, α is the weighting factor that controls the intensity of the cooling effect (randomly sampled from a uniform distribution [0, 0.5)), and DC& IC are non-linear functions that decrease and increase the intensity values, as shown in Fig. 1, respectively. For the latter (hue, saturation and darkness), we first normalize I_{cool} by

$$\begin{aligned} I^R_{cool\prime} &= I^R_{cool}/255 \\ I^G_{cool\prime} &= I^G_{cool}/255 \\ I^B_{cool\prime} &= I^B_{cool}/255 \end{aligned} \tag{2}$$

and then calculate

$$\begin{aligned} C_{\max} &= \max(I^R_{cool\prime}, I^G_{cool\prime}, I^B_{cool\prime}) \\ C_{\min} &= \min(I^R_{cool\prime}, I^G_{cool\prime}, I^B_{cool\prime}) \\ \Delta &= C_{\max} - C_{\min} \end{aligned} \tag{3}$$

We then use $C_{\max}, C_{\min}$ and Δ to convert RGB images to HSV. We obtain H channel by

$$H_{dark} = \beta \times \begin{cases} 60^\circ \times \left(\frac{I^G_{cool\prime} - I^B_{cool\prime}}{\Delta} \bmod 6 \right) & , \text{if } C_{\max} = I^R_{cool\prime} \\ 60^\circ \times \left(\frac{I^B_{cool\prime} - I^R_{cool\prime}}{\Delta} + 2 \right) & , \text{if } C_{\max} = I^G_{cool\prime} \\ 60^\circ \times \left(\frac{I^R_{cool\prime} - I^G_{cool\prime}}{\Delta} + 4 \right) & , \text{if } C_{\max} = I^B_{cool\prime} \end{cases} \tag{4}$$

and S and V channels are obtained by

$$\begin{aligned} S_{dark} &= \gamma \times \begin{cases} 0 & , \text{if } C_{\max} = 0 \\ \frac{\Delta}{C_{\max}} & , \text{if } C_{\max} \neq 0 \end{cases} \\ V_{dark} &= \delta \times C_{\max} \end{aligned} \tag{5}$$

where β, γ, δ are the three parameters controlling the distance in color shift and the intensity of saturation and darkness changes. For the PITSA dataset, they were randomly sampled for each triplet from the empirically-determined ranges [0.8, 1.2], [0, 1] and [0.35, 0.75], respectively. Finally, we map H_{dark}, S_{dark} and V_{dark} back into the RGB space, and we randomly apply a further darkening step as detailed in [6]. The resulting output is named I_{dark}, which

*Equal contribution.

Figure 1. On the left hand side, DC (i.e. red channel mapping) and IC (i.e. blue channel mapping) functions are plotted, which are used to decrease/increase intensity values in warmth altering phase of our dataset generation pipeline (see eqn. (1)). On the right hand side, example failure images are shown where the pre-trained shadow detector fails to find all shadows, and as a result, some of our shadow-free (I_{free}) images/patches actually have shadows on them.

is used for final shadow blending operation that creates I_{shadow} (please see eqn. 9 in the manuscript).

As mentioned in the main manuscript, the primary bottleneck of the pipeline is the accuracy of the shadow detection model M used for patch extraction. We show several failure cases in Fig. 1. Inevitably, there are some nighttime and low-light images in our dataset; such images are going to add noise to our data. Furthermore, the failure of the pretrained shadow detection model leads to shadow-contaminated shadow-free images. Manual vetting, or repeated patch filtering via M is a desirable way forward, at the expense of additional cost. However, our results (shown in the main manuscript) show that despite this noise, the PITSA dataset significantly improves the results of our method and existing methods. Finally, any improvement in M is likely to directly translate to the quality of the produced dataset, which shows the value our dataset generation pipeline brings.

2.2. Statistics of the PITSA

Image variety. Example (shadow) images from our PITSA dataset are shown in Fig. 2; PITSA has significantly more variety in scenes compared to existing datasets such as ISTD and SRD. ISTD and SRD, since they are real-life datasets where pairs or triplets are generated under controlled environments, primarily contain ground-directed images taken with low-pitch camera angles (i.e. pavement, grass, etc.) with relatively fixed distance to the subject. Our PITSA dataset, on the other hand, exploits images taken with any camera angle or position, and significantly scales the scene variety ($\sim$ 10 times).

We quantitatively compare the image variety of ISTD, SRD and PITSA datasets; we first feed all three datasets to a ResNet model pre-trained on the ImageNet dataset [3], and collect the (top-5) predictions of the model for each image in the training split of each dataset. Afterwards, we simply calculate the histograms of these predictions and visualize them in heatmaps. Middle row of Fig. 3 shows these (40×25) heatmaps for each dataset, where each pixel represents a class, color black indicates no predictions for the class and brighter colors indicate more frequent predictions of the class. It is visible that ISTD is quite limited in variety whereas SRD is a bit better. The last heatmap shows that our PITSA dataset has significantly more variety in scenes.

We used the training splits for each dataset when calculating the heatmaps (of Fig.3), and PITSA is significantly larger than the others in terms of image quantity. For a fairer comparison, we take the ten most frequent predictions made by the pre-trained ResNet model for each dataset, obtain the number of samples of each prediction, and calculate their ratio to the number of images in the training split. The results are shown in Fig. 4 for all three datasets. The three most frequent predictions of ISTD make up around 30% of the entire training set, and the first two are *doormat* and *manhole* classes, both referring to a certain viewpoint (i.e. looking at the ground). SRD is slightly better, where the three most frequent predictions account for around 14% of the entire training set. Please note that *doormat* and *manhole* classes are in top three for SRD like ISTD, which is another proof of the limited variety. PITSA, on the other hand, shows that its variety does not solely come from the large number of images it has; the three most frequent predictions are *monastery*, *palace* and *lakeside* classes, showing the variety in scenery. Furthermore, the three most frequent predictions make up a mere 5% of the entire training set. This shows that PITSA does have a better variety, but also it has a more balanced representation of classes.

Statistics of shadow regions on images. In addition to its scene diversity, PITSA dataset also shines with its shadow distribution, both in terms of size and location. The top row of Fig.3 shows the histogram of shadow mask sizes as a percentage of the original image sizes, for ISTD, SRD

Figure 2. Example images from the PITSA dataset (shadow image I_{shadow} examples). Note the diversity of images, as well as the shadow mask shapes.

and PITSA datasets. PITSA shows a more uniform distribution in the shadow sizes, whereas SRD and ISTD have distinct peaks around 10% and 20%, respectively. Furthermore, PITSA has a better distribution of shadows with large mask sizes (over 50%).

In addition to the shadow size, another important information is the location of the shadows observed in images. The bottom row of Fig. 3 shows the location distribution of shadows in images, where brighter areas correspond to areas with greater likelihood of having shadows. ISTD and SRD have limited variety in shadow locations, where SRD is slightly more diverse. PITSA, on the other hand, has a significantly larger variety in shadow locations, showing that PITSA images tend to have shadows in any location of an image.

Comparison with other datasets. We compare PITSA with the (previous) largest removal synthetic dataset in terms of pretraining performance; we pretrain our best model either on PITSA or on [6], and then finetune on ISTD. Table 1 shows PITSA provides a performance boost across all metrics, especially visible in detection BER and shadow-region MAE.

Pre-train on:	[6]	PITSA
S	6.2	**5.6**
NS	**2.4**	**2.4**
All	3.0	**2.9**
BER	1.70	**1.47**

Table 1. Comparison with [6] in terms of pretraining performance. Models are trained on either [6] or PITSA, and then finetuned on ISTD; results shown are from A-ISTD test set. PITSA provides a bigger boost to performance than [6].

2.3. The hypothesis behind PITSA

From another perspective, we can think of PITSA as a dataset for a task more abstract than shadow detection&removal, which is the detection and removal of altered colors of image regions. This task definition differs from shadow detection and removal; i) altered color is not necessarily a shadow, but can be of any color, ii) alteration-free image (i.e. input image) may have shadows (i.e. shadow-free images *should* not have shadows) and iii) altered color regions are not limited with the realism of shadows (i.e. we

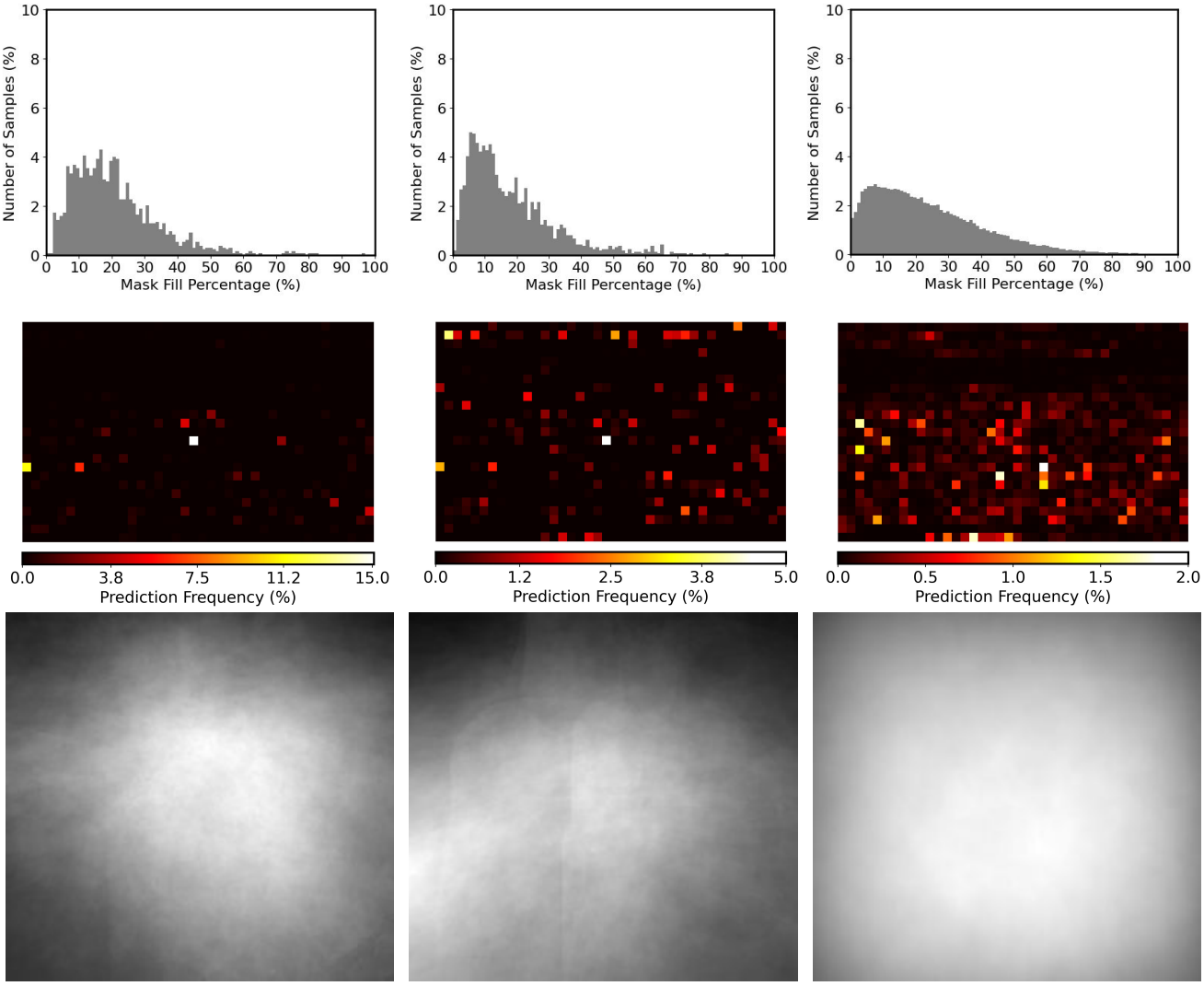

Figure 3. Statistics for the ISTD (first column), SRD (second column) and PITSA datasets (third column). First row shows the distribution of mask fill percentage (i.e. percentage of regions occupied by shadow on images) against number of samples (i.e., percentage of the overall number of samples). Second row shows the (histogram) heatmaps for the predictions made by a ResNet model pre-trained on ImageNet [3], when training samples of each dataset are fed to the network. Each class is represented by a pixel (40×25 heatmap for all 1K classes), where black color indicates no predictions for that class and brighter colors indicate more frequent predictions. The last row shows the shadow location distribution per pixel, where brighter values mean a greater likelihood of shadows being present on that pixel. Note that in each row, the PITSA dataset shows favourable characteristics; a balanced shadow size (row 1), a greater scene diversity (row 2) and a greater shadow location diversity (row 3).

can alter the color of a clear sky, whereas we can not cast shadows on a clear sky). The latter two lets us relax two important requirements; i) shadow/alteration-free image *purity* (i.e. absolutely no shadows in shadow-free image) and shadow/alteration mask realism. With these relaxations, we can use i) more shadow/alteration-free images and ii) a more diverse set of shadow/alteration masks. Thanks to this, PITSA can be scaled even further, and provides a boost to nearly all shadow removal/detection methods when used in pretraining. The potential downside of PITSA is that we do not recommend its use as a benchmark dataset for evaluation, as above points make it suboptimal for the more concrete/constrained task of shadow detection/removal.

3. Details on the LRA&LDRA

In this section we present additional details on experiments provided in the main manuscript.

3.1. Why [4] for LRA&LDRA?

As mentioned in Section 3.2 of the main text, we choose [4] to implement both LRA&LDRA modules due to its

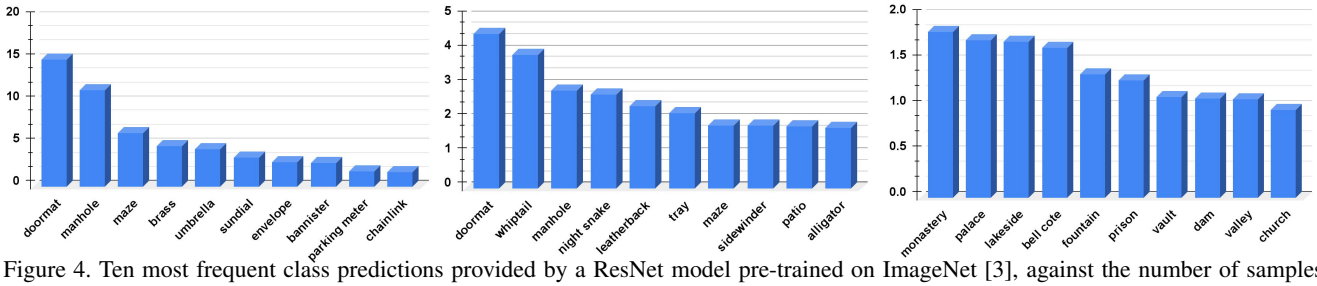

Figure 4. Ten most frequent class predictions provided by a ResNet model pre-trained on ImageNet [3], against the number of samples (per class) as a ratio of the overall number of training samples, for ISTD (left), SRD (middle) and PITSA (right) datasets. Note that the predicted labels show greater diversity for the PITSA dataset. Furthermore, the ratios of class predictions are significantly lower for PITSA, showing that the PITSA dataset is not biased towards a specific scene category (i.e. ground images for ISTD and SRD). Best viewed when zoomed in.

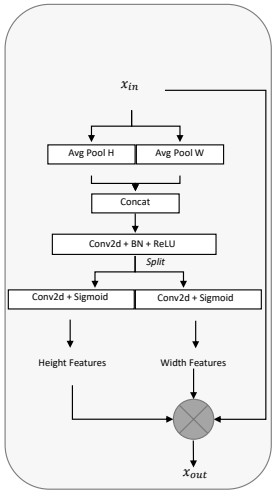

Figure 5. Diagram of [4] used to implement LRA&LDRA.

minimal overhead and strong spatial/cross-channel components.We now explain our choice in more detail.

Spatial components. [4] avoids global-pooling by factorizing it into two 1D feature encoding operations, and doing so preserves spatial/positional information better than alternatives. This is important in our use case as we want LRA&LDRA to guide R to focus on shadow regions, and also filter out non-shadow region prediction should R fail to do so (see such failure occurs for vanilla residual predictions in Figure 4, main text).

Cross-channel relations. [4], building on previous attention methods like [12], models inter-channel information. This is especially important for blending and color-correction; recall from main text that LRA prepares/transforms the input for blending and LDRA performs color-correction. Since LRA&LDRA operate on images, the ability of choosing the most important color channel for such transforms is important. This is realized thanks to the strong inter-channel information modeling of [4].

3.2. Detailed ablations on LRA&LDRA.

Having justified the addition of LRA&LDRA in Table 2 main text, we now present the full set of ablation experiments we conducted when designing LRA&LDRA. The results are shown in Table 2.

We note that table 2 is an extended version of Table 2 in main text. This version also shows that [4] is a better option than [12] (spatial attention) for implementing LRA&LDRA, which further justifies our choice. Furthermore, we also show that alternatives to the additive residual formulation, such as multiplication and convolution layer, do not work as well (visible in shadow region MAE) as the additive residual formulation we adopt in LRA&LDRA.

3.3. Mask Generation for the SRD

Since we present a joint solution for detection and removal, we also leverage mask information during evaluation. The mask information is available for ISTD but not for SRD dataset. To this end, we generate masks for SRD dataset to be able to evaluate the shadow removal model R, which requires a mask input. This lets us evaluate the performance of the removal model in more detail; shadow and non-shadow accuracy can be evaluated separately.

During mask generation, in addition to adaptively thresholding the difference between shadow/shadow-free images [2], we manually inspect the masks and either i) automatically correct the ones with minor errors or ii) manually annotate from scratch the ones with incorrect masks. For the former, we perform mask filling via flood fill and noise reduction via median filtering, while we use tools for pixel-level annotation for the latter.

3.4. Additional Details on the Training Phase

Visualization of our pipeline. Our overall pipeline is shown in Fig. 6. Both networks are trained with ℓ_1 loss using their respective ground-truth labels, but the detection network also leverages the gradients g_R of the removal network with respect to the cost of the removal network $cost_R$.
Training duration. We note that the training takes around a day for full 2000 epochs for ISTD and slightly more for

	Ablation on different (LRA, LDRA)									Ablation on eqn (4	
	$\mathbb{B}$	(1,1)	(1-I_m,I_m)	([4],1)	(1,[4])	([12],1)	(1,[12])	([12],[12])	([4],[4])	([4],[4])†	([4],[4])‡
S ↓	7.94	8.69	**7.32**	7.73	8.45	8.27	7.91	8.25	7.54	9.80	8.51
NS ↓	3.20	2.66	2.97	2.71	**2.55**	2.73	2.68	2.65	**2.55**	2.77	2.67
All ↓	3.86	3.56	3.54	3.45	3.40	3.55	3.50	3.47	**3.29**	3.80	3.52
BER ↓	2.84	1.91	1.81	1.69	1.85	1.77	2.30	1.78	**1.56**	1.96	1.73

Table 2. Accuracy (MAE and BER) obtained for D and R, equipped with various (LRA, $LDRA$). I_m denotes the I_{mask}. Results with † and ‡ use multiplication and convolution instead of the summation operator in eqn. (4). The baseline $\mathbb{B}$ formulates R with eqn.(2). *S*, *NS* and *All* stand for shadow, non-shadow and all regions, respectively. This table is an extended version of Table 2 in main text.

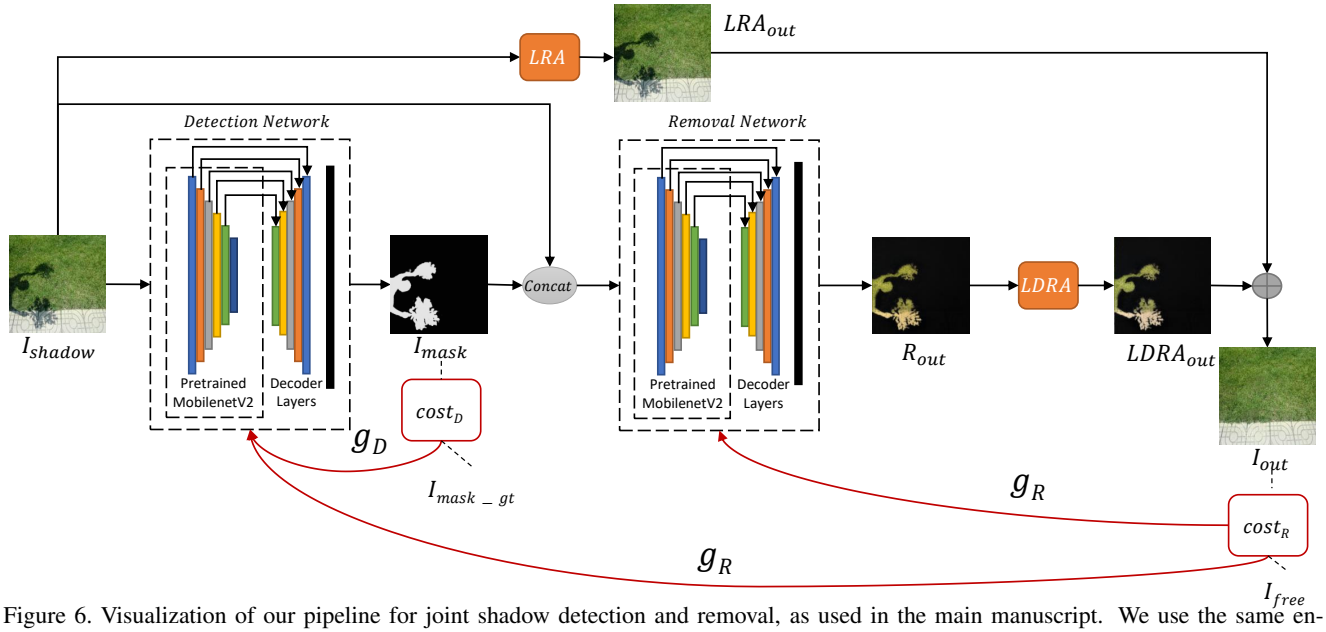

Figure 6. Visualization of our pipeline for joint shadow detection and removal, as used in the main manuscript. We use the same encoder/decoder architecture [13] for detection and removal networks. The removal network is trained with ℓ_1 loss ($cost_R$), whereas detection is trained with ℓ_1 loss ($cost_D$) and the gradients g_R of the removal network with respect to $cost_R$. Best viewed when zoomed in.

SRD datasets, although, due to early stopping, we generally stop the training before 2000 epochs. Training a model for an epoch on the PITSA dataset with a batch size of 16 on a single GPU, takes around an hour. Note that the timings are based on an RTX 3090 GPU.

3.5. Additional Analyses

In Tables 6 and 7 of the main manuscript, we show that we manage to outperform competing removal methods in various metrics. However, many of them [7, 8] use pretrained detection networks to produce their masks (during evaluation phase), whereas we train (and evaluate) our own detection network jointly, as stated in our manuscript. We note that our detection network reaches state-of-the-art accuracy on ISTD, therefore, it is better than the detection methods used by other removal methods. This leads us to the following question: *are our improvements in shadow removal related to the improvements in shadow detection?* We provide answers to this question in two different ways.

First, we look back at the main manuscript. In Table 4, the last three columns show our ablation on existing methods, and their change when we plug LRA&LDRA into these methods. For all three methods we have tested, we see consistent improvements in shadow removal accuracy in all regions. For [10], the improvement in removal is more pronounced compared to [7, 8], since [10] includes a detection network as well, and it improves considerably due to our LRA&LDRA. However, two removal-only methods [7, 8] still improve with the introduction of LRA&LDRA, especially in shadow regions.

Method	Baseline	LRA&LDRA
S	7.3	**6.8**
NS	2.7	**2.3**
All	3.3	**3.0**

Table 3. Accuracy of shadow removal models on the ISTD dataset. Baseline denotes the baseline method (which is described in eqn. (2) in the main manuscript. Both methods are trained without a detection network; only removal networks are trained with ground-truth masks as the input mask during training and evaluation.

Second, we train a baseline without LRA&LDRA (formulated as eqn. (2) in the main manuscript) without a detection network, where ground-truth masks are used as input to the removal network. We compare this with our LRA&LDRA without a detection network, again where ground-truth masks are used as input to the removal network. Table 3 shows that the improvements brought by

Figure 7. Visualizations of what LRA and LDRA attend on an input image. From left to right, input image, LDRA output and LRA output are shown. Brighter areas show the areas attended by our LRA and LRDA modules. Note that LDRA primarily performs spatial attention, whereas LRA does channel-wise transformations. The bottom figures show a failure case, where LRA does unnecessary spatial transformation and results into an inaccurate result.

LRA&LDRA are not only due to the improvements in the detection network accuracy, since both networks use ground-truth masks in Table 3.

3.6. Visualization of LRA&LDRA

Fig. 7 visualizes the locations overlaid on input images to which LRA&LDRA pay more attention. We first recall from Section 5.3 of main text several key points; i) LRA&LDRA jointly guide the removal network to produce localized (i.e. only on shadow regions) outputs, ii) LDRA takes in the output of the removal network, and refines/color-corrects it for the final blending, and iii) LRA takes in the input image and performs primarily channel-wise transformations to prepare the input for the final blending with the output of the LDRA.

The images in Fig. 7 show that LDRA attends on shadow regions, but not in a tight manner; the regions around the

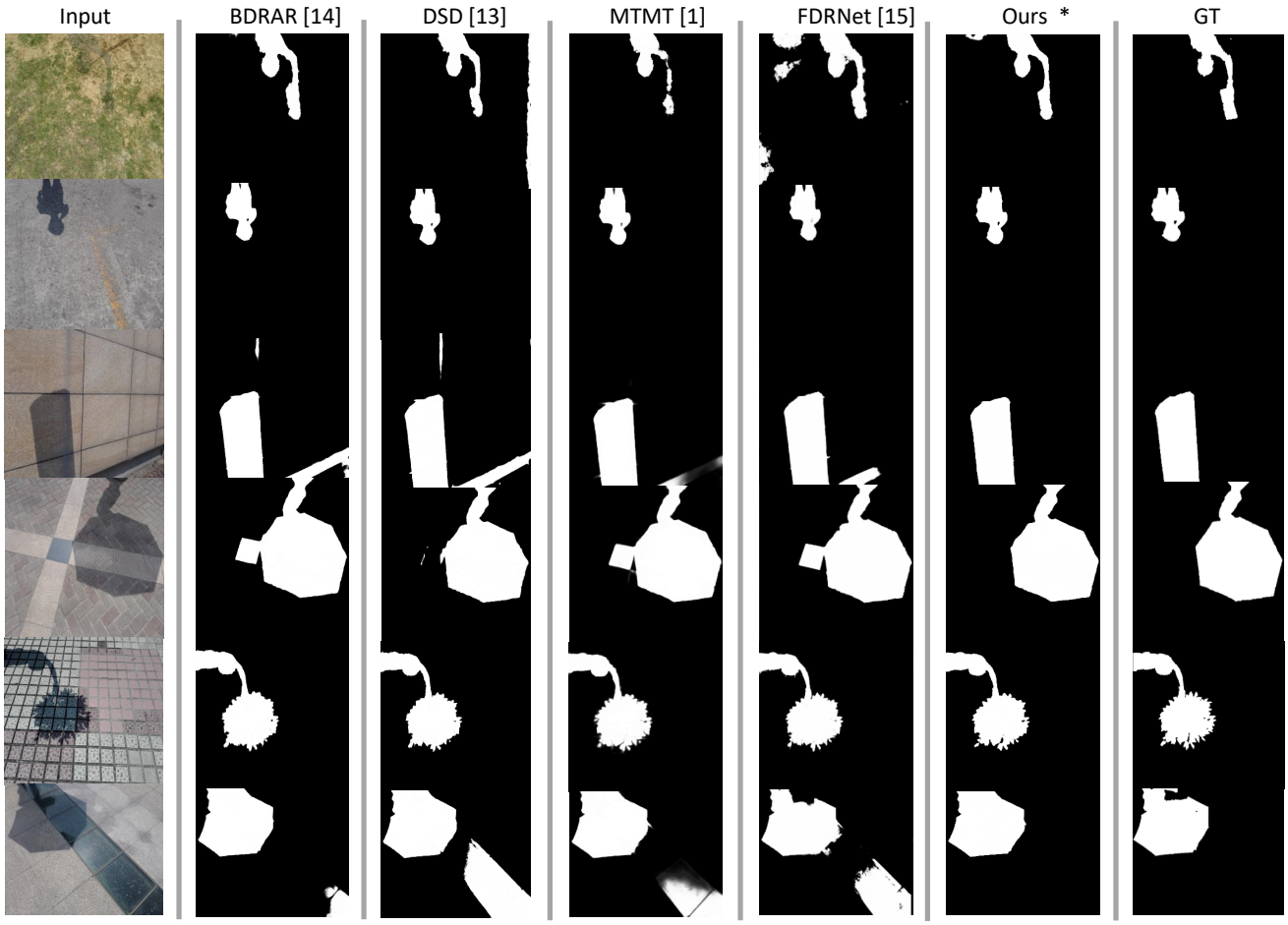

Figure 8. Qualitative comparison of LRA&LDRA and other shadow detection methods [16, 15, 14, 1]. Our model is pre-trained on the PITSA dataset. Note the several cases where our masks are not distracted by other dark (not shadow) regions, resulting into fewer false positives (rows 1, 3, 4 and 6). Please also note the sharp details (i.e. bright areas in between the shadow mask in row 5) successfully recovered by our method in rows 2 and 5.

shadow region are also attended. We believe that this is caused by the fact that our LRA&LDRA implementations are lightweight; for a *tighter* attention operation, a higher capacity model can be a better alternative. However, the current implementation is helpful for the final *blending*; this slightly larger attention region essentially acts like a dilation so the boundary of shadow region can be better combined/blended with the non-shadow regions. The operation of LRA is less intuitive compared to the operation of LDRA, since LRA primarily performs a channel-wise attention, rather than a spatial one. This channel-wise attention helps the blending operation with the output of the LDRA. However, please note that in the first two images (rows 1 and 2), some non-shadow areas are highlighted as well, especially along the borders of the shadow areas. This verifies the last point we made above; LRA helps prepare the input for blending, primarily via channel-wise transformations but also via some spatial transformations. The last image (row 3) shows a failure case, where the LDRA does a sufficient job, but LRA performs unnecessary spatial transformations, providing a sub-optimal result.

3.7. Qualitative Results for Shadow Detection

We provide additional qualitative results using our detection network (pre-trained on the PITSA dataset) for the ISTD dataset, and compare it with several state-of-the-art methods in Fig. 8. Our detection network outperforms other methods qualitatively as well; note that for various images, our method successfully avoids false positives (rows 1, 3, 4 and 6). Furthermore, our method is capable of recovering sharp details in complex shadow formations, such as the small bright spot in between the neck and the arm of the person in second row and the bright spots between the leaves in the fifth row. This shows that LRA&LDRA, in addition to its good accuracy in shadow removal, is also a strong performer in shadow detection despite a simple and compact network design.

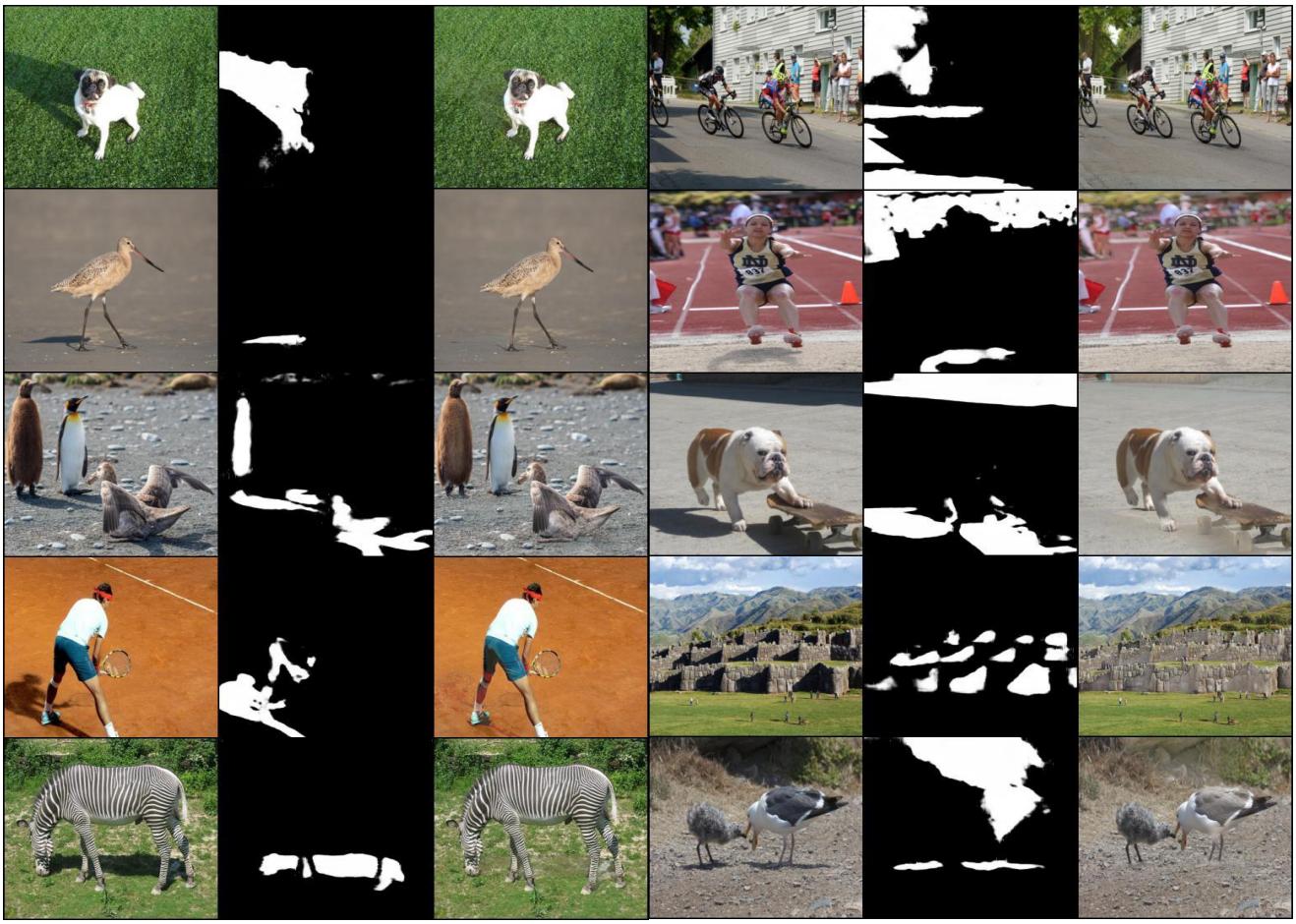

Figure 9. Zero-shot evaluation for detection and removal. (Input-detection-removal) on SOBA [11] and SBU [9] not used in training.

3.8. Weakly Supervised Shadow Detection

We explained in Section 3.3 of the main text that LRA&LDRA improves shadow detection results, when gradients G_R of the removal model R is backpropagated to the detection model D. The results in Table 5 main text rows 1&2 main text already verified this claim.

Here, we take a step towards weakly supervised shadow detection and ask this question; *what would happen if we trained D without ground-truth masks, and only with the gradients G_R of removal model R?* In other words, we do not *detach* D from R, and only use removal supervision (i.e. weak supervision).

Table 4 shows that whole image reconstruction $\mathbb{B}$ provides no useful signal to D; detection accuracy is random (50 BER). With LRA&LDRA using I_{mask} or LRA&LDRA provide useful information to model D, and detection accuracy is significantly improved (17 BER) compared to $\mathbb{B}$ (see images in Figure 2 main text). Note that these results are inferior to the D trained fully-supervised, but they act as further proof that LRA&LDRA help improve D.

	Ablation on WSL		
	$\mathbb{B}$ WSL	(1-I_m,I_m) WSL	LL WSL
BER ↓	50.63	18.88	**17.26**

Table 4. Weakly-supervised detection results, where no ground-truth mask supervision is provided to D, and it is trained indirectly with removal supervision. $\mathbb{B}$ is the baseline methods implementing eqn. (2) in main text, and I_m is the mask predicted by D. *LL* indicates LRA&LDRA.

3.9. Further Discussions

Zero-shot performance. We show several qualitative results on unseen distributions; results are shown in Figure 9. The model used to acquire these images are pretrained on PITSA; the results show strong detection and removal performance, verifying the effectiveness of PITSA pretraining and LRA&LDRA formulation.

Why not just..copy-paste (CP) the non-shadow region? We continue our discussion on LRA&LDRA versus simple copy-pasting. As shown in the main text, LRA&LDRA has key advantages over simply copy-pasting non-shadow regions of input to the output, such as robustness to minor mask errors and ability to perform blending/color-correction . We visualize such cases in Figure 10.

The importance of non-shadow regions. LRA and LDRA

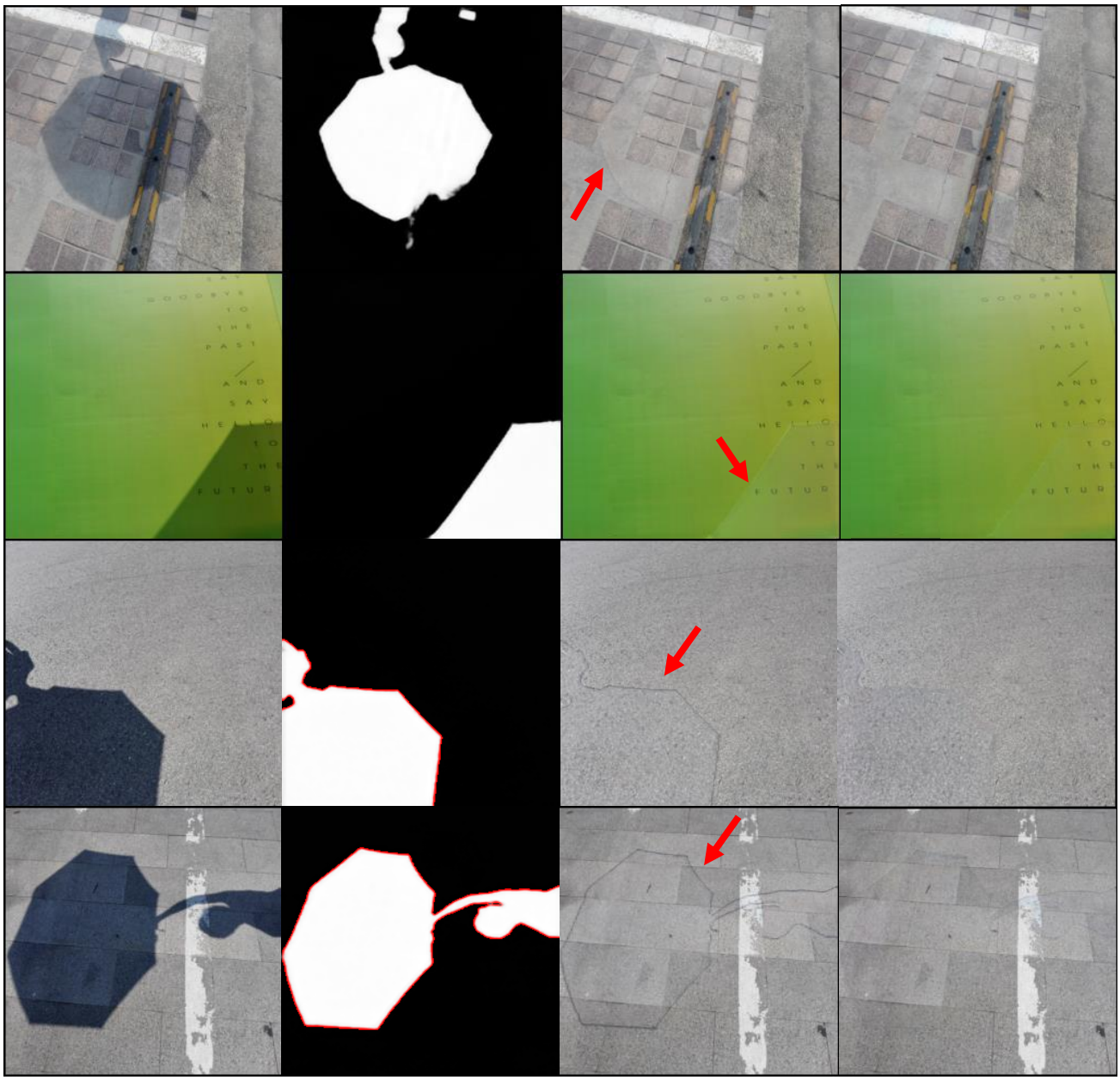

Figure 10. Advantages of LRA&LDRA over simply copy-pasting. (Rows 1-2): *Copy-paste* (3rd col.) with the blending artefacts (red arrows), which are addressed by our method (4th col.). (Rows 3-4): *Copy-paste* where mask errors (shown in red over masks) cause artefacts, which are addressed by our method.

improve overall shadow removal accuracy, but also give a solution to avoid focusing on non-shadow regions. This is important because in removal datasets (and real-life scenarios), shadow regions are generally smaller than non-shadow regions [5]. Therefore, models without LRA&LDRA can prioritize non-shadow regions during training, especially with common regression losses (i.e. ℓ_1), which can hinder shadow region performance. Furthermore, although many methods rightfully focus on shadow regions, non-shadow regions are equally important in practice; unsuccessful non-shadow region reconstructions are equally undesirable.

References

[1] 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.

[2] 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.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[13] 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.

[14] 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.

[15] 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.

[16] 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.