WSRD: A Novel Benchmark for High Resolution Image Shadow Removal

Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte
Computer Vision Laboratory, CAIDAS & IFI
University of Würzburg, Germany
{florin-alexandru.vasluianu, tim.seizinger,radu.timofte}@uni-wuerzburg.de

Abstract

Shadow removal is an important computer vision task, whose aim is to successfully detect the shadow affected area appearing through light occlussion, followed by a photo-realistic restoration of the affected image contents, textures, and details. After decades of research, a multitude of hand-crafted restoration techniques were proposed, following different observations on shadow formation models, with scenes altered in particular conditions. However, the increased popularity of deep learning based solutions enabled a significant step forward for the shadow removal solutions, both in terms of reconstruction fidelity and perceptual properties. However, the publicly available datasets remain focused around a particularly low complexity setup, with a low variety of light occluders and affected backgrounds, and with limited representation for more complex light interactions and complex shadow patterns. In this work, we propose WSRD, a novel benchmark for high resolution image shadow removal, characterized by a large variety in terms or represented objects, backgrounds and light occluders. We study more complex interactions, combining self shadows with externally casted shadows, to further extend the study of the phenomenon, its factors and effects. To prove WSRD as a relevant benchmark, we propose DNSR, a novel shadow removal method, comparing the results on WSRD with the performance level observed on other well-established benchmarks like ISTD and ISTD+. We validate our approach comparing with existing state-of-the-art (SOTA) methods, improving both in reconstruction fidelity and perceptual properties, setting a new SOTA for the field.

1. Introduction

Shadows are defined as direct effects of light occlusion [39]. During image acquisition, the sensor reading corresponding to the shadow affected image is directly depending on the amount of light interacting with the objects present in the acquired scene. The geometry of a light occluder will be, thus, projected into the image plane, as a less illuminated region, whose shape and properties depend on a large variety of factors. Along the light intensity, its position in the 3D world, the geometry, it's properties (*e.g.* color, intensity), and the material related properties, defined both for the occluder object and the surface the shadow is casted on, all the aforementioned factors add variety in the shadow patterns set that can be retrieved from the acquired images.

Shadows are usually defined by steep variation in an image region, causing decrease in pixel intensity, without any connection to the variation observed between various homogeneous color regions corresponding to the segments characteristic to each of the objects present in the acquired scene. Thus, the shadow effect brings an additional source of variance in the image color space, thus impacting the other vision tasks such as object recognition [2, 20, 40] or tracking [9, 26, 27], image segmentation [1, 11] or semantic segmentation [15, 36].

Moreover, in contrast to the image pixels from shadow free areas, the shadow phenomenon can be coupled with a series of degradations altering the 3D scene observations, with various ramifications impacting image illumination, color, detail, and noise levels. Thus, shadow removal can be labeled as a particular case of Image Restoration, where the solutions achieving a significant performance level would restore the colors, textures and a majority of the details lost during acquisition.

As the localization information of the shadow affected images is important for the shadow removal task [29,30,47], shadow removal is usually connected with the detection subtask, with publicly available [23, 46] large databases tailored around the shadow phenomenon.

Recently, large-scale databases consisting of shadow-affected and shadow-free image pairs, such as SRD [33], ISTD [47] or USR [22] allowed the formulation of the shadow removal process as a regression problem in the broader supervised learning framework. However, the aforementioned datasets are sharing roughly the same setup, with simple scenes affected only by a shadow casting object. Even though this setup was a good starting point for the already proposed shadow removal methods, a significant number of interactions still remain to be studied.

Therefore, we introduce WSRD, a novel benchmark for High Resolution image shadow removal, based on a large variety of interactions and represented contents. Furthermore, we propose DNSR, a novel shadow removal algorithm, that sets a reference for the introduced benchmark, and improves over the state-of-the art on well established benchmarks, such as ISTD [47] and ISTD+ [29, 47].

2. Related Work

Despite extensive study during decades of research, shadow removal remains a challenging problem. Eventhough the introduction of large datasets [22, 33, 47] enabled a step forward in terms of reconstruction performance, there is still a need for extensive study and solutions able to generalize on a wide variety of shadow formation model parameters.

Early works focused their efforts around determining the underlying physical properties of the shadow formation model. Different image processing techniques [12, 13] aimed at determining a combination of shadow-affected and shadow-free layers, resembling the input image. The importance of the localization information (*e.g.* the positions of the pixels affected by shadows) is acknowledged by authors in [37, 45, 50], where shadow detection was a sub-task of the shadow removal algorithm, the restoration being based on a color transfer procedure, from the shadow-free areas to the shadow affected segments.

A wide variety of factors can be identified as sources of variation in the shadow formation model (*e.g.* materials, shapes, sizes, geometries, illumination, etc). Thus, the large number of involved factors increases the complexity of the shadow formation model, making the shadow removal task a challenging study. Consequently, models built around combinations of the aforementioned parameters are characterized by a lack of generalization ability, when asked to restore images more complex than the particular conditions those solutions were tailored around [19]. Applying shadow detection prior to shadow removal proved to be a winner strategy for early works [16, 19], where hand-crafted features such as pixel intensity, texture, or gradients were used to successfully detect the shadow-affected areas.

The increased popularity of the Convolutional Neural Networks (CNNs) and the introduction of large databases for shadow detection and removal [22, 33, 46, 47] enabled a new category of solutions, characterized by a step forward in terms of performance. Qu *et al.* [33] proposed system of three sub-networks solving different sub-tasks of the shadow removal procedure. The G-net extracts a set of high-level features, the A-net models the appearance of the scene, and the S-net computes a shadow matte characterizing the shadow pattern affecting the scene.

Recently, Le *et al.* [29] proposed a system based on two neural networks able to learn an approximation of the shadow model, by estimating a corresponding shadow matte. However, the system is limited by using a simple linear geometry encoding the illumination system. So, even in the condition on a single light occluder object, the existence of multiple light sources will lead to non-homogeneous opacity shadow areas that can not be learnt using a linear model. Moreover the nature of the shadow phenomenon itself is not linear, with the geometry of the scene with respect to the light position and intensity directly linked to non-homogeneous opacity.

On one side, other works focused on end-to-end learnt shadow removal [14, 18, 19, 21, 35, 47, 51]. Hu *et al.* [21], proposes a type of solution coding the localization information in a learning procedure, where the shadow detection and removal have as backbone a Spatial Recurrent Neural Network [3], exploring the concept of a direction-aware context. These larger complexity models can indeed, when provided high quality training data, learn complex shadow models, producing high quality restored images. One particular case is [14], where the manipulation of an exposure parameter in the shadow-affected and shadow free areas enable the learning of a high reconstruction fidelity mapping.

Moreover, the introduction of the attention [44] mechanism in the computer vision tasks enabled a new set of solutions [4, 18, 53] based usually on variations of Channel Attention [49] or different types of feature fusion functions. The availability for larger receptive fields, coupled with better normalized gradients continues to provide a performance boost for this category of solutions. Moreover, reducing the number of parameters enabled a design shift towards more complex operations, requiring a larger numbers of FLOPs. But, given the development of the currently available hardware, this category of models achieve real time performance both for consumer level GPUs and smartphone deployed hardware [4, 6, 7, 35, 41].

On the other side, the introduction, followed then by a deeper understanding, of the Generative Adversarial Networks (GANs) [17] enabled a new category of solutions solving the shadow removal as a image-to-image translation procedure. Given the proposed pix2pix [24] and CycleGAN [54] there are two approaches that can be distinguished. While the former assumes there exists a single transformation between shadow-affected and shadow-free images, the latter defining assumptions are less restrictive, implying the existence of a forward transformation and it's inverse. This increases the complexity of the learnt model, with different generators learning the direct and the inverse transformation from the shadow-affected to shadow-free domain, guided by a cycle-constraint to ensure the approximation of a realistic mapping.

Representative for the category of solutions is the work in [47], where a Conditional GANs [32] based model was trained in the fully supervised setup, with two stacked conditional GANs aiming at shadow detection and removal. Another work is represented by [8], where a GAN based model learns a large variety of shadow-matte images. Then, DHAN [8], a model where downsamplings are replaced by dilated convolutions, coupled with multi-context feature aggregation for attention, encodes the complex shadow formation learned model in a solution characterized by a significant performance level in terms of perceptual properties.

One advantage of the CycleGAN [54] approach is the prospect of learning a domain-to-domain transformation without the need of paired training data. Obviously, paired data acquisition is a tedious process, and the possible considered setups are rather simple in nature. This possibility of learning a transformation in an unsupervised setup was successfully explored by [22, 42], proposing high performance solutions characterized by improved perceptual properties.

However, the common simple setup used by the aforementioned image databases still represents the main limitation of the mentioned shadow removal methods, these models being unable to generalize outside the condition of an external object casting a shadow in a scene. The newly introduced WSRD will provide extensive proof that more complex light-object interactions will produce scenes that are more difficult to deshadow.

3. Method

3.1. WSRD Benchmark Particularities

As we already mentioned, the main drawback of the already publicly available shadow removal datasets comes from the simplicity of the setup used for data acquisition. As it can easily be observed for datasets such as SRD [33], ISTD [47], USR [22] there is a limited light-object interaction, with limited representativity of the multitude of factors impacting the shadow formation model. The setup consists of a surface characterized by an increased degree of consistency in terms of colors and textures, on which a shadow is casted using an object whose geometry is not captured, the occluder object being kept outside the acquired scene. Even if the simplicity of the aforementioned setup allows for the acquisition of an increased number of image pairs, the representativity of the possible scenarios remains low, thus affecting the generalization ability of the solutions tailored around these databases.

Adopting a more difficult setup is problematic, since interactions such as self casted shadows would imply the usage of a light system to successfully eliminate them, and the introduction of new lights will undoubtedly lead to color inconsistencies between the shadow affected input and the shadow free ground truth. Therefore, a color correction strategy has to be taken into consideration, directly depending on the used setup.

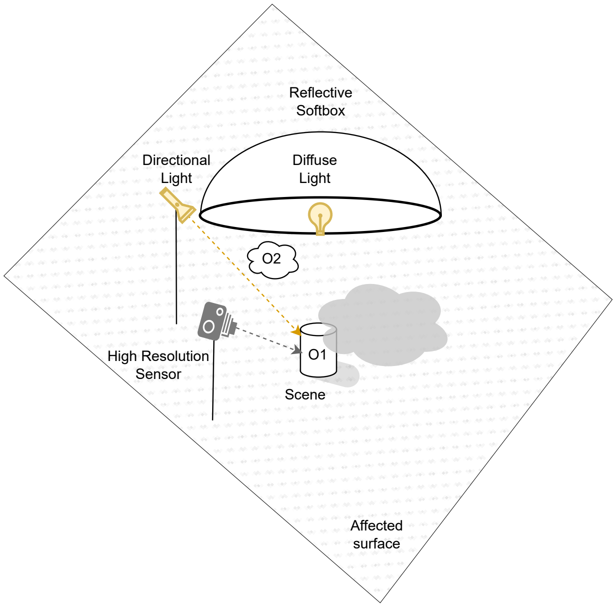

Figure 1. A visual representation of the setup used for data acquisition. The two occludes O1 and O2 cast different shadows into the scenes. Based on the nature of the light sources and the occluder objects, the corresponding shadow pattern is characterized by different properties.

Moreover, given the fact that data acquisition is done in outdoors conditions, with an obvious delay between the captured image pair components, caused by the necessity to change the scene, additional lighting inconsistencies will be present in the captured data. For example, such inconsistencies were observed for the ISTD dataset [47], with Le *et al.* proposing a correction method [29] aiming to further decrease the error observed between the image pair components in the shadow free areas. Additional semantic differences found between the input and the corresponding ground truth images [42] come as another reason to invest more effort into the study of alternative setups.

Unlike previous methods, the proposed image database was fully built around a set of controllable conditions, using controllable artificial light sources. Figure 1 provides a visual representation of the used data acquisition scenario, where the objects in the scene, coupled with the outside-the-scene light occluders co-participate in the shadow formation process. By extending the study to capture the interactions between different shadow types, we aim at a better generalization of the model when deployed on real data.

The environment is built around a set of two lights. One fixed flash acting as spotlight is at an elevated position over the scene, and a 45° angle from the base surface plane of the setup. This will be the light considered for shadow casting, being triggered only during the acquisition shadow affected images. The other light is a diffuse flash, pointed towards a reflective softbox, aiming at an optimal distribution of the light in the captured scene, countering all the possible self-shadows created by the high complexity surfaces appearing in the captured scene.

To capture the shadow affected input frame, both the spotlight and the diffuse light were triggered. The spotlight causes objects in the scene to create a particular shadow pattern, according to the light orientation and their geometry, while the diffuse flash helps for the acquisition of detail-rich images, helpful later, during the image restoration task. To capture the shadow-free ground truth images, we only activate the diffuse flash. The softbox will uniformly distribute the light across the scene, avoiding the appearance of shiny over-exposed areas. As the light setup changes between the images forming a pair, we are matching the exposure of the input and the ground truth images by post-processing the RAW data in Adobe Lightroom.

As shown in Figure 1, the images are captured using a high resolution sensor, the Canon EOS R6 II, set up in a slightly elevated position with respect to the surface plane. The first light, serving as directional spotlight, is fixed at a 45° vertical angle towards the scene and a 90° horizontal angle with respect to the camera position. During capturing of the input and ground truth frames, the camera ISP parameters were fixed in manual mode, to avoid exposure or white balance changes between corresponding frames. The light sensitivity was set to ISO100, minimizing noise based color fringes, and the aperture of the used 70mm lens system was fixed to F11, thus maximizing the depth of field.

The WSRD [1] is based on a large variety of captured surfaces (see Figure 2), with a multitude of colors and textures. Moreover, the captured objects are characterized by different geometries, with different thickness, height, or depth. The dataset is characterized by various types of materials, with opaque, translucent and transparent materials, characterized by different conductivity. The dataset contains 1200 high resolution image pairs (1920×1440px), with 1000 pairs used for training, 100 in the validation split and another 100 representing the benchmark. Even though the acquisition setup was kept consistent for the aforementioned splits, the data splitting was made with respect to the represented contents, with 20% of the testing/validation samples representing objects unseen during training and 50% of the testing/validation acquired scenes are characterized by base plane surfaces that are unseen during training.

A particular difficulty posed by the described setup is the appearance of soft self-shadows. These are caused by the increased complexity of surfaces that, in some cases, cast diffuse shadows on different surfaces. This can be easily circumvented by the adoption a translucent scene based that is illuminated from below. Additional softboxes placed on the sides of the scene volume can additionally improve the uniform light distribution. However, the explored setup is characterized by high flexibility when it comes to the properties of the lights, captured materials, and light occluder combinations, enabling to further extend the study of the shadow formation model.

3.2. DNSR: Distill-Net Shadow Removal

Figure 3 provides a schematic representation of the proposed architecture (DNSR). The model is built in a classical UNet [34] fashion, with skip connections between the encoder blocks and their decoder counterparts. Given the high resolution of the input data, there exists an advantage in performing a progressive downsampling of the feature space used for shadow removal.

Between the encoder and decoder blocks, our features suffer a distillation phase, based on the input shadow map and variations of the input shadow affected image. Two learnt parameters, α and ω, are used to manipulate the exposure in the shadow/shadow-free areas. The idea of exposure manipulation was explored by Le *et al.* in [14], achieving state-of-the-art results. This variation of the input shadow image will be processed by the Distiller module, and the output feature map will be added to the Channel Attention [49] weighted feature map, after it passes through a Dynamic Convolution model and Layer Normalization [4]. The dynamic convolution operation, that appears in the Distill blocks and the Decoder block is defined as an auto-weighted average of a fixed number of 3×3 convolutions, with the set of weighting parameters being learnt by the model. The idea is similar to the already explored Malleable Convolution [25], or Dynamic Convolution [5], but with far lower constraints over the set of blending parameters. The Fused Pooling operation form the Distiller Module represents two parallel Average/Max Pooling operations, that are blended by a 1×1 convolution.

In the decoder block, the updated propagated feature map is upscaled by a 2×2 transposed convolution operation, then the output of this operation is fused to the skip connection information through the Stereo Channel Attention Module [4]. An additional block is then used to refine the feature map, before it gets propagated to the next level of the architecture.

Overall, the model is characterized by a number of 47.22 Million learned parameters, with 67 GMACs needed to compute the shadow free estimation, at the original resolution of the ISTD [47] dataset, of 480×640 px.

3.3. Experimental Setup

We based our work on the Pytorch framework, using for training and evaluation four NVIDIA RTX3090Ti GPUs with 24 GB VRAM.

[1] https://github.com/fvasluianu97/WSRD-DNSR

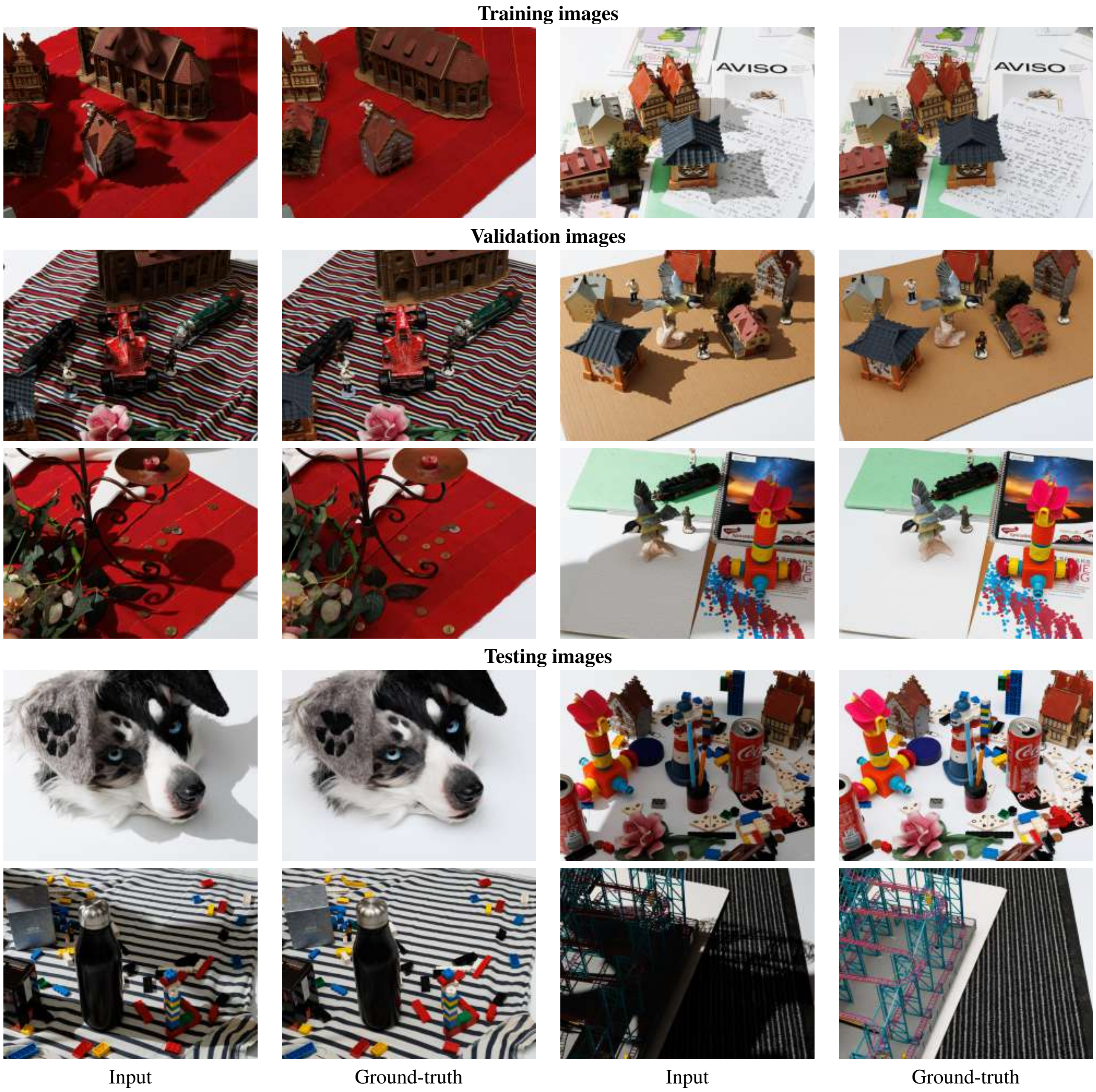

Figure 2. Samples from the introduced WSRD benchmark. Best zoom-in on screen in the electronic version.

We perform our training and evaluation runs on the ISTD [47], ISTD+ [29, 47] and our proposed WSRD datasets. Given the particularities of the aforementioned datasets, the hyperparameters of the model training will be adapted for optimal convergence.

The model was trained using a combination of L_2 and a perceptual loss [42] as the minimized objective. The definition of the training loss for the prediction $\hat{y}$ and the ground-truth y is provided in the Equation 1.

$$L(\hat{y}, y) = L_2(\hat{y}, y) + \alpha_{perc.} L_{perc.}(\hat{y}, y) \tag{1}$$

The perceptual loss (Equation 2) is a combination of a color loss, a VGG19 [38] based content loss, and a style loss [10]. Each of the aforementioned loss terms are blended in the final objective through a set of weights determined by the magnitude of each loss function when the model reaches

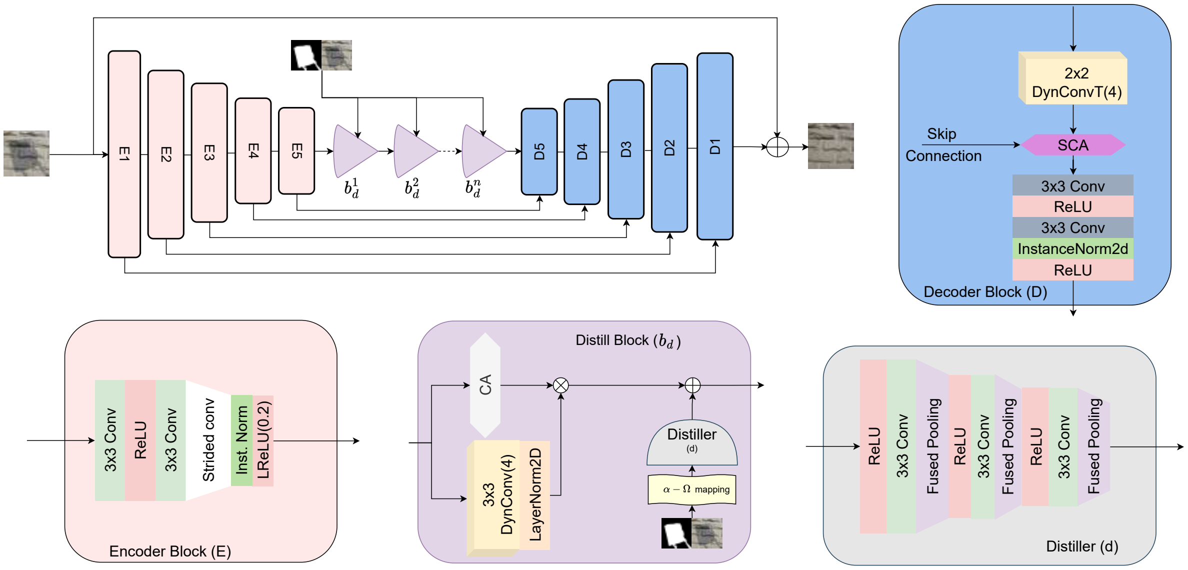

Figure 3. A visual representation of the proposed DNSR architecture (*top left*), along with the decoder block (D) (*top right*). On the bottom row, we provide schematics representing the encored block (E) (*left*), the architecture for the Distill Block (d_b), (*center*) and a visual representation of the Distiller (d) module (*right*).

training convergence.

$$L_{perc.}(\hat{y}, y) = \alpha_c L_c(\hat{y}, y) + \alpha_{vgg} L_{vgg}(\hat{y}, y) + \alpha_s L_s(\hat{y}, y) \tag{2}$$

The definition of the color loss L_c is provided in the Equation 3, where the smoothing operation is performed using a gaussian filter.

$$L_c(\hat{y}, y) = L_2(\hat{y}_{smoothed}, y_{smoothed}) \tag{3}$$

In Equation 4, we define the VGG loss L_{vgg} as the average L_2 loss observed for the feature maps F^i extracted from the VGG19 [38], after each batch normalization operation. The index i is covering the set of aforementioned layers.

$$L_{vgg}(\hat{y}, y) = \frac{1}{N_l} \sum_{i=1}^{N_l} L_2(F^i_{\hat{y}}, F^i_y), \tag{4}$$

Using the 2D nature of the extracted feature maps, in Equation 5, for the fixed i, j index pair, we define the Gramm matrix $G^l_{i,j}$ characteristic to the layer l and it's corresponding dimension D. Using an elementwise product operation, we can use every Gramm matrix as a set of similarity measures, characterizing the set of VGG features extracted for the prediction $\hat{y}$ and for the ground-truth y.

$$G^l_{i,j}(x) = \sum_{k=1}^{D} {F^l}_{i,k}(x) {F^l}_{k,j}(x) \tag{5}$$

Then, the style loss L_s can be defined, as shown in the Equation 6.

$$L_s(\hat{y}, y) = \frac{1}{N_l} \sum_{i=1}^{N_l} L_2(G^i(\hat{y}), G^i(y)) \tag{6}$$

In Table 1, we specify the values for the loss terms blending weights for each of the datasets used for training. Given the significant differences between the aforementioned datasets, there are differences in the other hyperparameters characterizing the performed training runs.

4. Experimental Results

4.1. Evaluation measures

For the quantitative evaluation of our method on the ISTD [47] related datasets, we stick to the already established method of reporting the Lab image representation RMSE error between the predicted outputs and their corresponding ground truths. Given the fact that ISTD provides also a set of shadow maps as the localization information for the shadow affected areas, we report the Lab space RMSE for the shadow affected/shadow free areas.

As the WSDR dataset was used as challenge data for the NTIRE2023 Image Shadow Removal challenge [43], we adopt their reported metrics to compare against the top performing solutions. So, we report the recovery fidelity in terms of PSNR, the Structured Similarity Index (SSIM) [48] and, since the shadow removal is a highly perceptual task, we report the LPIPS distance [52]. Considering that we used a VGG19 [38] based loss in our training objective, the reported LPIPS uses AlexNet [28] as feature extractor.

	Training objective				Training hyperparam.					Evaluation hyperparam.		
Training Dataset	$\alpha_{perc.}$	α_c	α_{vgg}	α_s	Train res.	rCrop	rRotate	rHflip	rVflip	Eval. res.	Orig res.	Rescale
WSRD	0.25	1.0	0.1	10^7	800×800	$p = 1$	$p = 0.5$	$p = 0.5$	$p = 0.5$	960×1280	1440×1920	✓
ISTD [47], ISTD+ [29,47]	0.10	1.0	0.1	10^7	448×448	$p = 1$	$p = 0.5$	$p = 0.5$	$p = 0.5$	480×640	480×640	✗

Table 1. Hyperparameters characteristic to the trainig/evaluation runs, depending on the dataset.

4.2. Quantitative results

In Table 2, we provide a comparison between the proposed DNSR and other top performing solutions from the NTIRE2023 Image Shadow Removal challenge [43]. Also, we provide information about the well-established methods that were used as backbone for the proposed solutions. As you can see, DNSR is able to produce high fidelity results with consistent perceptual properties.

Method	Team	Backbone	PSNR↑	SSIM↑	LPIPS↓
PES	MTCV	NAFNet [4]	22.36	0.70	0.182
IR-SDE	IR-SDE	IR-SDE [31]	19.60	0.58	0.149
SRDM	SRDM	-	22.20	0.69	0.269
MFDSNSR	MM911	SHARDS [35]	21.69	0.69	0.293
ShadowFormer+	IIM_TTI	ShadowFormer [18]	18.08	0.53	0.196
DNSR (*ours*)	-	-	22.92	0.65	0.285

Table 2. Quantitative results of the challenge final submission on the *WSRD* test split. We compare against top performing NTIRE 2023 Image Shadow Removal Challenge [43] solutions.

To compare the proposed method against other shadow removal proposed methods, we also report results on the ISTD dataset [47] and its corrected variant [29, 47]. Here, we compare against well-established solutions like STCGAN [47], DHAN [8], PULSr [42], and AEF [14]. Also, we report results for some contemporary works, such as ShadowFormer [18] and SHARDS [35]. We report the Lab space RMSE, with the reported values computed over the set of publicly available results, at their original resolution, or with our own implementations of the described methods, based on the publicly available software and descriptions.

Method name	Eval. res.	Lab space RMSE		
		Shadow region	Shadow free region	Total
unprocessed	640×480	15.07	3.86	6.80
STCGAN [47]	256×256	4.83	3.44	4.05
DHAN [8]	640×480	4.65	3.13	3.43
PULSr gen. [42]	512×512	4.48	3.03	3.33
AEF [14]	256×256	3.75	2.79	3.1
ShadowFormer [18]	640×480	3.25	2.38	2.43
DNSR (*ours*)	640×480	4.39	2.47	2.84

Table 3. Quantitative results of DNSR (*ours*), compared to state-of-the-art solutions on the ISTD [47] dataset.

As it can be observed in the Table 3 and Table 4, DNSR is achieves top performance on the ISTD benchmarks [29,47], improving over the results achieved by well-established methods, and with a comparable level of performance compared to other contemporary works.

Method name	Eval. res.	Lab space RMSE		
		Shadow region	Shadow free region	Total
unprocessed	640×480	17.53	1.82	7.15
SP-M Net [29]	512×512	4.79	4.27	4.37
DHAN [8]	640×480	4.04	2.97	3.19
PULSr gen. [42]	512×512	4.12	2.39	2.82
AEF [14]	256×256	3.23	2.05	2.31
SHARDS [35]	640×480	3.16	1.61	1.98
ShadowFormer [18]	640×480	2.93	1.66	1.93
DNSR (*ours*)	640×480	3.92	1.80	2.24

Table 4. Quantitative results of DNSR (*ours*), compared the aforementioned top performing solutions on the ISTD+ [29, 47] test split.

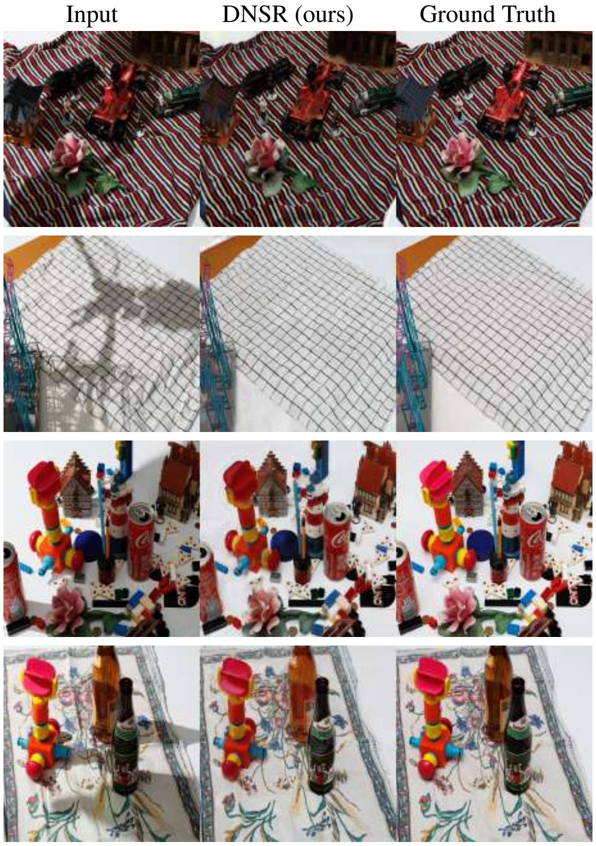

Figure 4. Visual results of DNSR on the test split of WSRD dataset.

4.3. Qualitative results

In Figure 4, we provide samples of the shadow free predictions of the proposed DNSR on the WSRD test split.

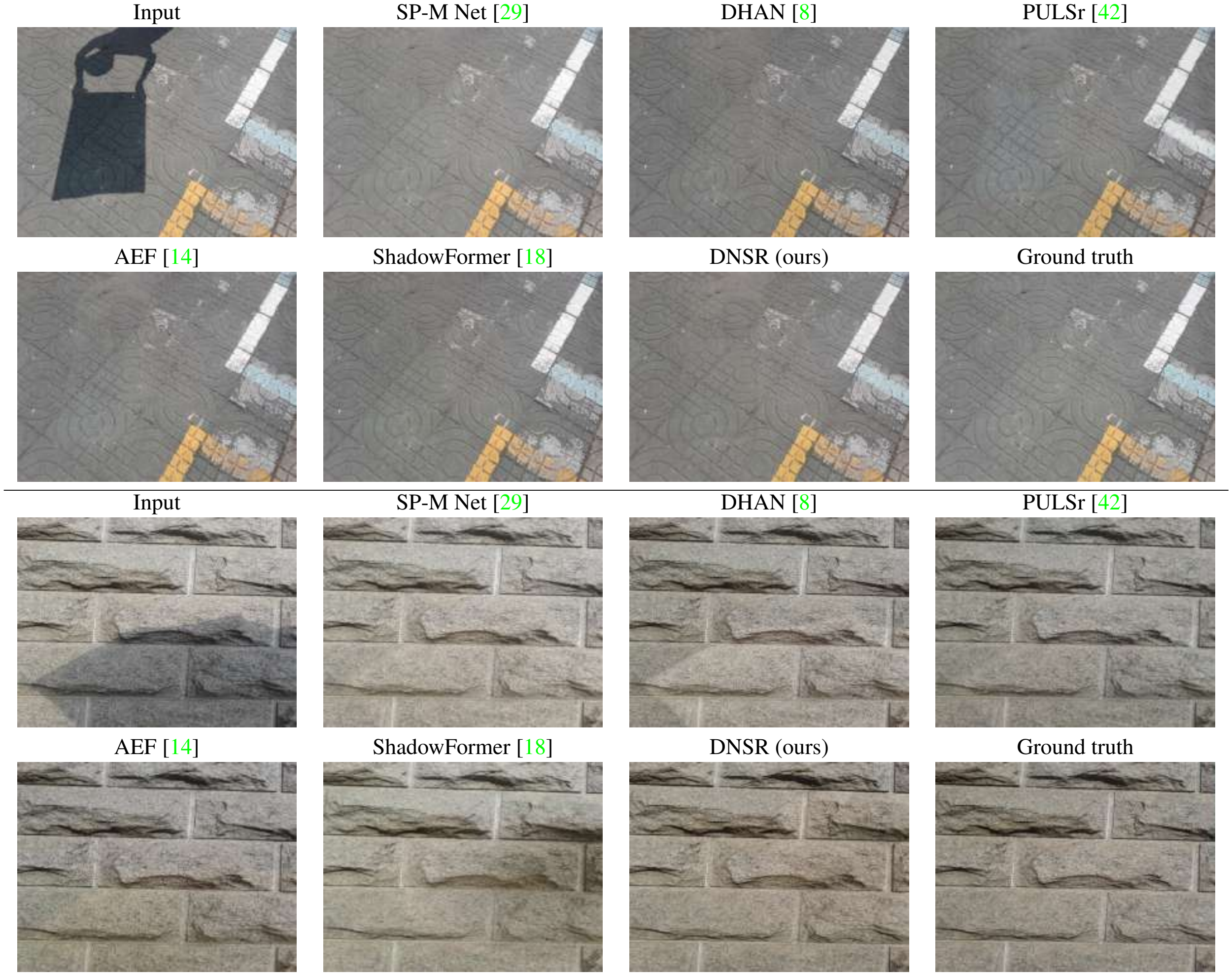

Figure 5. Visual results from the test split of the ISTD [47] dataset. Here, we compare our DNSR against SP-M Net [29], DHAN [8], PULSr [42], AEF [14] and ShadowFormer [18]. All the results were upscaled back to the original resolution of the input data. Note that SP-M Net [29] is trained on the corrected data (using the authors proposed method). Best zoom-in on screen in the electronic version.

DNSR is able to successfully remove the shadows, with a high perceptual quality. It can handle complex conditions with cluttered scenes described by a wide range of interractions and shadow patterns.

To validate the proposed method against other well-described solutions, available in the public literature, we report equivalent predictions (see Figure 5) on samples from the ISTD [47] test split. DNSR is able to produce results characterized by a low level of visible artifacts, with correct colors and textures, and naturally looking reconstructions. The quality of the provided results is supported by the performance level quantified in Table 3 and Table 4.

5. Conclusions

In this work we proposed a novel benchmark for the Image Shadow Removal task, extending the study of shadow formation models by increasing the representation of a wide range of interactions and altered surfaces. The image database is characterized by a large variety of surfaces, characterizing a multitude of colors and textures. To evaluate the differences between the newly introduced benchmark and other well-established datasets, we propose DNSR, a novel solution for Image Shadow Removal, achieving state-of-the art results on the proposed benchmark, and being able to achieve a similar or better level of performance, compared to other contemporary works tailored around the existing datasets.

Acknowledgements

This work was partially supported by the Humboldt Foundation. We thank the sponsors of the NTIRE 2023 workshop and challenges.

References

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

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

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

[4] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 2, 4, 7

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020. 4

[6] Marcos V Conde, Florin Vasluianu, Sabari Nathan, and Radu Timofte. Real-time under-display cameras image restoration and hdr on mobile devices. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 747–762. Springer, 2023. 2

[7] Marcos V. Conde, Florin Vasluianu, Javier Vazquez-Corral, and Radu Timofte. Perceptual image enhancement for smartphone real-time applications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1848–1858, January 2023. 2

[8] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan, 2019. 3, 7, 8

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

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

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

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

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

[14] 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. 2, 4, 7, 8

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

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

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[18] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 2, 7, 8

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

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

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

[22] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-ShadowGAN: Learning to remove shadows from unpaired data. In *ICCV*, 2019. 1, 2, 3

[23] 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. 1

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

[25] Yifan Jiang, Bartlomiej Wronski, Ben Mildenhall, Jonathan T Barron, Zhangyang Wang, and Tianfan Xue. Fast and high quality image denoising via malleable convolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 429–446. Springer, 2022. 4

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

[27] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *Proceed-*

ings of the IEEE international conference on computer vision workshops, pages 1–23, 2015. 1

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 6

[29] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 3, 5, 7, 8

[30] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *The IEEE European Conference on Computer Vision (ECCV)*, August 2020. 1

[31] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023. 7

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

[33] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2308–2316, July 2017. 1, 2, 3

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[35] Mrinmoy Sen, Sai Pradyumna Chermala, Nazrinbanu Nurmohammad Nagori, Venkat Peddigari, Praful Mathur, B H Pawan Prasad, and Moonhwan Jeong. Shards: Efficient shadow removal using dual stage network for high-resolution images. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1809–1817, 2023. 2, 7

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

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

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

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

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

[41] Florin Vasluianu and Radu Timofte. Efficient video enhancement transformer. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4068–4072, 2022. 2

[42] Florin-Alexandru Vasluianu, Andrés 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*, pages 826–835, 2021. 3, 5, 7, 8

[43] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Ntire 2023 image shadow removal challenge report. In *New Trends in Image Restoration (NTIRE 2023) Workshop.*, 2023. 6, 7

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

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

[46] 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 Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 816–832, Cham, 2016. Springer International Publishing. 1, 2

[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 Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[49] 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. 2, 4

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

[51] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image Processing*, 21(10):4361–4368, 2012. 2

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

[53] Xiao Feng Zhang, Chao Chen Gu, and Shan Ying Zhu. Spaformer: Transformer image shadow detection and removal via spatial attention. *arXiv e-prints*, pages arXiv–2206, 2022. 2

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