SHARDS: Efficient SHAdow Removal using Dual Stage Network for High-Resolution Images

Mrinmoy Sen[1], Sai Pradyumna Chermala[1], Nazrinbanu Nurmohammad Nagori[1], Venkat Peddigari[1], Praful Mathur[1], B H Pawan Prasad[1], Moonhwan Jeong[2]
[1]Samsung R&D Institute India - Bangalore, [2]Samsung Electronics
{mrinmoy.sen,nazrin.n,p.mathur,pawan.prasad,mh0614.jeong}@samsung.com
{chs.pradyumna,venkatrpeddigari}@gmail.com

Abstract

Shadow Removal is an important and widely researched topic in computer vision. Recent advances in deep learning have resulted in addressing this problem by using convolutional neural networks (CNNs) similar to other vision tasks. But these existing works are limited to low-resolution images. Furthermore, the existing methods rely on heavy network architectures which cannot be deployed on resource-constrained platforms like smartphones. In this paper, we propose SHARDS, a shadow removal method for high-resolution images. The proposed method solves shadow removal for high-resolution images in two stages using two lightweight networks: a Low-resolution Shadow Removal Network (LSRNet) followed by a Detail Refinement Network (DRNet). LSRNet operates at low-resolution and computes a low-resolution, shadow-free output. It achieves state-of-the-art results on standard datasets with 65x lesser network parameters than existing methods. This is followed by DRNet, which is tasked to refine the low-resolution output to a high-resolution output using the high-resolution input shadow image as guidance. We construct high-resolution shadow removal datasets and through our experiments, prove the effectiveness of our proposed method on them. It is then demonstrated that this method can be deployed on modern day smartphones and is the first of its kind solution that can efficiently (2.4secs) perform shadow removal for high-resolution images (12MP) in these devices. Like many existing approaches, our shadow removal network relies on a shadow region mask as input to the network. To complement the lightweight shadow removal network, we also propose a lightweight shadow detector in this paper.

1. Introduction

Shadow Removal from images is a complex problem in computer vision. Prevalence of shadows in an image can adversely affect other computer vision tasks like object recognition. In addition, shadow removal has its application in image editing softwares like Adobe Photoshop where professional photographers relight the shadow areas for shadow removal. This is a cumbersome process and requires domain expertise.

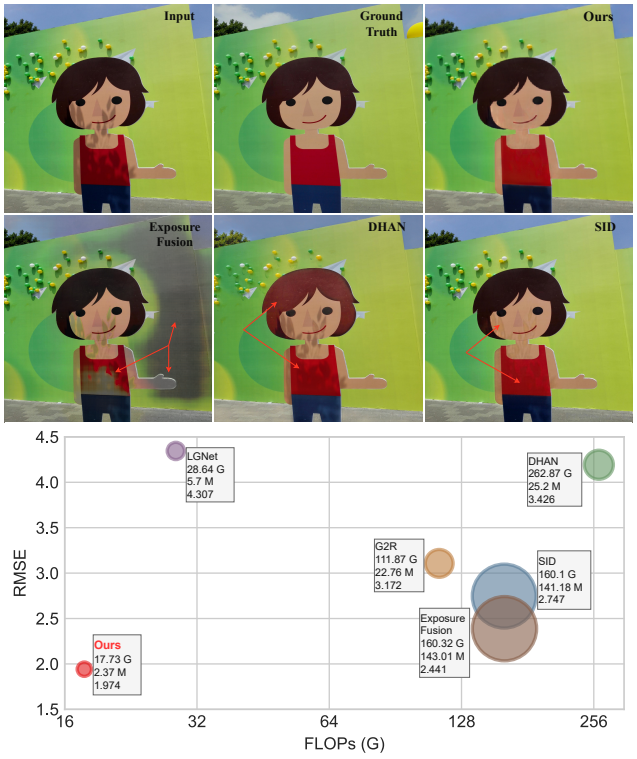

Figure 1. **a.** Existing shadow removal methods produce artefacts or leave shadow traces for high-resolution images. **b.** Model comparison in terms of performance, model parameters and computation with existing methods.

Recent advances in deep learning based approaches [18,

21] have achieved remarkable results in shadow removal. However, a limitation of all the existing approaches is that they work only on low-resolution (<1MP) images. This can be attributed to the fact that CNNs, based on which these networks are built, usually have a limited receptive field. Techniques to increase the receptive field of the network involve adding deeper layers thus increasing the network parameters and effectively making them harder to train. Furthermore, in the context of high-resolution images, it may lead to computational and hardware constraints. In addition, many of these approaches [4, 12] rely on heavy network backbones that cannot be adopted in resource-constrained platforms like smartphones. Thus shadow removal for high-resolution images remains an extremely challenging task and absence of high-resolution datasets adds to the problem.

To address these limitations, we propose a novel shadow removal method that can remove shadows even from high-resolution images. Specifically, this is achieved using a dual stage approach with two lightweight networks LSRNet and DRNet. To elaborate, given a high-resolution shadow image, the first network (LSRNet) is trained to remove the shadow from the image at low-resolution. Solving the shadow removal problem at low-resolution ensures that the network has a large receptive field to efficiently aggregate global context from shadow-free regions to relight shadow regions in a consistent manner. To restore the details in the low-resolution output, the second network (DRNet) is trained to retrieve the details using the high-resolution shadow image as guidance. In this work, we demonstrate that the proposed LSRNet achieves state-of-the-art performance on the challenging ISTD [21] shadow removal dataset with significantly lesser number of network parameters than existing methods proving that earlier techniques have not been designed keeping network efficiency into consideration. In addition, we also create new high-resolution shadow datasets to prove the effectiveness of our proposed architecture on high-resolution images. In Figure 1, we show a sample result of our proposed network compared against existing state-of-the-art shadow removal methods on a high-resolution shadow image. While our method is able to remove the shadow completely other methods struggle on high-resolution images due to their limited receptive field. Additionally, in Figure 1, we compare the performance of the proposed method against existing methods [4, 12, 6, 15, 16] in terms of model parameters, computation and shadow removal quality (RMSE metrics). Our network outperforms the existing techniques while being significantly more parameter and computation efficient.

Our proposed shadow removal architecture also uses a shadow mask as an input to the networks. To this effect, we propose a fast and efficient shadow detection network. We demonstrate that the proposed shadow detector network performs comparably against existing networks even though it has less network parameters. Finally, due to the lightweight nature of the shadow detection and removal networks, we show that the proposed architecture can be deployed on latest smartphones and achieve extremely fast shadow removal with processing time of approximately 2.4secs for 12MP (4032x3024) images.

To summarize, the major contributions of this work are: 1. We propose SHARDS, a novel shadow removal method for high-resolution images using two lightweight networks and prove the effectiveness of the method on high-resolution images.
2. We also demonstrate that on existing low-resolution benchmark ISTD dataset, our shadow removal network achieves state-of-the-art performance with a magnitude of order less parameters and computations than the existing networks.
3. We also propose a fast and efficient shadow detection network that achieves comparable performance with state-of-the art methods with lesser network parameters.
4. Additionally, we demonstrate that the proposed method can be deployed on modern smartphone devices and achieves extremely fast shadow removal (2.4 secs) for high-resolution (12MP) images.

2. Related Works

2.1. Shadow Removal

Early traditional works on shadow removal were aimed to model the physical properties of shadow, based on image decomposition of shadow and shadow-free layers [5], or a color transfer from the non-shadow region to the shadow region [19]. With the availability of larger datasets, a number of deep-learning based shadow removal techniques have been proposed. In DeShadowNet [18] by Qu *et al*., the network is trained to remove shadows in an end-to-end manner by predicting the shadow matte. Wang et.al [21] use two stacked cGANs to jointly train a shadow detector and shadow removal model. DHAN [4] uses attention and hierarchical aggregation of features to address the boundary artifacts termed 'ghosting' observed in earlier techniques. In Mask-ShadowGAN [7], the CycleGAN [25] framework is used to train the model in an unsupervised manner. In [12], Le *et al*. propose SID, which uses two networks to predict the shadow parameters and shadow-matte respectively. The predictions are combined using a linear illumination model to get the final shadow-free output. Although SID claims to adapt to high-resolution images using the matte interpolation technique, the simple linear model proposed produces visible artefacts around very detailed shadow boundaries which is further pronounced when scaled to high-resolution images as shown in Figure 1. In [6]. Fu *et al*. proposes shadow removal as a multi-exposure image fusion problem

(AEF). The simple illumination based models proposed in SID and AEF [12, 6] ignore or constrain the spatially-variant properties of shadows thus limiting their generalization capability. While dual stage networks [2, 14] exists for other tasks like image segmentation and matting, they cannot be directly adopted for the shadow removal task.

2.2. Shadow Detection

Similar to shadow removal, early works on shadow detection also relied on physical models of illumination. Recent deep learning based methods have outperformed these early techniques and achieved remarkable results. Vicente *et al*. [20] use a stacked CNN approach with noisily annotated data. Nyugen *et al*. [17] incorporate a sensitivity parameter in the scGAN network for shadow detection. Le *et al*. [13] use GANs to generate adversarial training samples to improve the shadow detection performance. Zheng *et al*. [24] propose incorporating distraction semantics to the network to predict false positives and false negatives and the distraction features are fused to each layer for better prediction. Hu *et al*. in DSCNet [9] and FSDNet [8] uses RNN to incorporate direction-aware features in four directions for better context aggregation. FSDNet [8] is among the first networks designed specifically to be efficient for mobile deployment.

3. Proposed Method

3.1. Shadow Removal

Like many recent works [4, 7] in Shadow Removal, the proposed network is trained using the GAN framework in which two networks, the Generator and the Discriminator are jointly trained in an adversarial setup. We treat shadow removal as an image-to-image translation problem where the shadow images and the shadow free images constitute the two domains. Specifically, in the proposed architecture, shadow removal on high-resolution images is carried out using two lightweight networks that we denote by LSRNet and DRNet. Given a shadow image and the corresponding shadow mask, LSRNet is trained to output the shadow-free image at low-resolution similar to existing methods. At low-resolution the network has a large receptive field to capture non-local contextual cues from shadow-free regions which is important to relight the shadow regions. DRNet is trained independently of LSRNet to restore the details into a low-resolution shadow-free image. We propose that since the high-frequency details are already present in the high-resolution input shadow image, it can act as a guidance to the DRNet. Specifically, the network takes an upsampled low-resolution shadow-free image along with the high-resolution shadow image and shadow mask as inputs, and is trained to reproduce the ground-truth high-resolution shadow-free image. In addition, since the primary task of shadow removal is solved at a low-resolution by LSRNet, it allows DRNet to have an extremely lightweight architecture. The proposed dual stage architecture is depicted in Figure 2.

LSRNet is based on an encoder-bottleneck-decoder architecture as shown in Figure 3. Employing four strided convolutions in the encoder, it downsamples the feature space to $1/16^{th}$ of the input resolution thereby capturing rich and deep features. This design replaces the heavy backbones (like VGG, ResNext) in some of the SOTAs (DHAN [4], SID [12]). We argue that effective shadow removal requires extensive non-local information to maintain consistency between the shadow region and shadow-free region in the output. The lack of it could lead to conspicuous artefacts. To overcome this, we propose to use self-attention in the residual bottleneck module. The decoder uses skip connections and upsampling convolutions to scale the output back to the original resolution. In addition, we use Convolution Block Attention (CBAM) [23] layers in the decoder to let the network dynamically weigh the relevant channels and spatial locations. To adapt to different shadow intensities, we append three additional inputs by doing gamma correction on the input image with different gammas $\gamma = 0.7, 0.5, 0.35$, converting them to LAB color space and taking their (L) channels. Finally, we add a multi-level perceptual loss to improve the output quality. The discriminator is a slightly modified multi-scale patch implementation [11] with each scale housing a self-attention layer. DRNet uses a similar network architecture to that of LSRNet with the exception of self-attention and CBAM layers. To train both the networks, triplets of shadow image, shadow mask and the corresponding ground truth shadow free image are needed. Let I_s^{hr}, I_m^{hr} , I_{sf}^{hr} and I_s, I_m and I_{sf} be such a high-resolution and corresponding downsampled low-resolution triplet in the dataset. LSRNet is trained to transform I_s to a shadow free image I'_{sf} expressed as,

$$I'_{sf} = LSR_{Net}(I_s, I_{s^\gamma}, I_m) \tag{1}$$

We perform gamma correction on the input shadow image to get I_{s^γ} as described above and provide it as an additional input to LSRNet. The Generator LSR_{Net} and Discriminator D_{LR} are trained to jointly optimize the following objective function,

$$\begin{aligned} L_{GAN}(LSR_{Net}, D_{LR}) = E_{I_{sf}}[\log(D_{LR}(I_{sf}))] + \\ E_{I_s, I_{s^\gamma}, I_m}[\log(1 - D_{LR}(LSR_{Net}(I_s, I_{s^\gamma}, I_m)))] \end{aligned} \tag{2}$$

Along with the adversarial loss, an additional multi-layer perceptual loss L_{percep} is used for the Generator and is represented as,

$$L_{percep} = \sum_{k=0}^{5} \lambda_k ||\phi_k(I'_{sf}) - \phi_k(I_{sf})||_1 \tag{3}$$

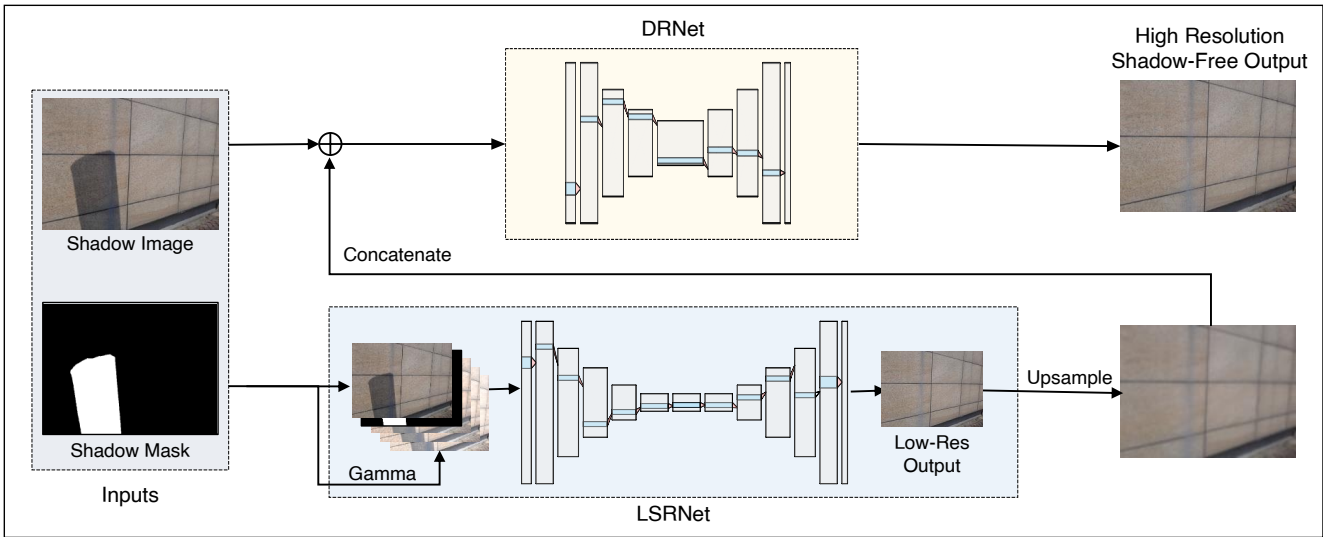

Figure 2. Proposed Shadow Removal using Dual Stage (SHARDS) Network architecture for high-resolution images.

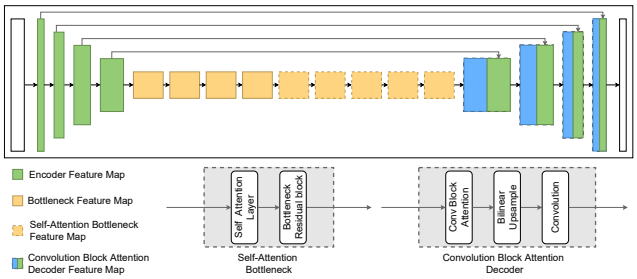

Figure 3. Architecture of Proposed LSRNet for shadow removal at low-resolution

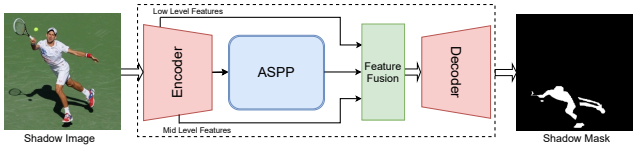

Figure 4. Proposed Shadow Detector Network Architecture.

where ϕ is the pre-trained VGG-19 feature map and the loss is calculated for layers $Convk\_2$ for $(k = 1, 2, 3, 4, 5)$. For $(k = 0)$, it is the pixel-wise difference between I'_{sf} and I_{sf}, the generated and the ground-truth shadow free image weighted by λ_k. To improve the shadow removal quality, we also force the output from bottleneck block to be shadow-free by using the above perceptual loss formulation. We use a 3-channel convolutional layer followed by TanH activation to first map the output from bottleneck block to RGB space. An appropriately downsampled version of the original shadow-free image is then used as a ground-truth. This results in a multi-level perceptual loss setting with weights of λ_{high} and $1 - \lambda_{high}$ for the final and bottleneck outputs respectively. Thus, the overall loss function for the network is expressed as,

$$L_{total} = L_{GAN} + \theta L_{percep\_multilevel} \tag{4}$$

We use $\theta = 1$ in our experiments. A similar setup is used to train DRNet as well. Specifically, DRNet is trained to transform I_{sf} to a high-resolution shadow-free image $I^{hr'}_{sf}$ using I^{hr}_{s} and I^{hr}_{m} as guidance, which is expressed as,

$$I^{hr'}_{sf} = DR_{Net}(I_{sf}, I^{hr}_{s}, I^{hr}_{m}) \tag{5}$$

3.2. Shadow Detection

Like many previous works [12, 6] the proposed shadow removal method takes a shadow mask as input. Most shadow detection networks [26, 24, 9, 3] rely on heavy network backbones and cannot be deployed in an embedded device like smartphone due to their high computational complexity. We propose a lightweight shadow detector network that complements the proposed shadow removal network. Our network consists of an encoder, ASPP and decoder module similar to the DeepLabV3 [1] architecture. We adopt the efficient MobileNetV2 backbone as the encoder, followed by the ASPP module to extract features at multiple scales and to incorporate global context. Finally, the decoder consists of three convolution layers followed by upsampling at each stage to predict the shadow mask. At only 3.2 million parameters, the network has significantly less parameters than existing methods. Figure 4 shows the architecture of the proposed network. Recently in [8] Hu *et al*. proposed FSDNet, a lightweight shadow detection network with only 4.4M network parameters. We outperform

FSDNet across all datasets and show that FSDNet with its direction-aware spatial context module (DSC) having RNN formulation does not offer any additional context than the ASPP module. To train the network we use the weighted cross entropy loss (WCE). Let p_i be the probability of the shadow for the i-th pixel as predicted by the network where y_i is the ground-truth ($y_i = 1$, if it is a shadow pixel and $y_i = 0$ otherwise). The WCE loss for an image is then calculated by,

$$L_{wce} = \sum_i (-\frac{N_n}{N_n + N_p} y_i \log(p_i) - \frac{N_p}{N_n + N_p}(1 - y_i)\log(1 - p_i)) \quad (6)$$

where, N_p and N_n denote the summation of false positives and false negatives in the image.

4. Experimental Results

4.1. Shadow Removal

In this section, we thoroughly examine our method and ablate on the individual components used. We then compare our work with relevant state-of-the-arts from recent literature and demonstrate the effectiveness of our lightweight shadow removal network. All our experiments are trained only on the baseline dataset and do not use any additional synthetic data. Finally, we demonstrate the superiority of our dual stage architecture for high-resolution shadow removal.

4.1.1 Network Implementation

Both LSRNet and DRNet can be divided into the following three conceptual blocks – encoder, bottleneck and decoder. The encoder consists of five convolutional layers – first with a kernel of 7x7 and the rest four with 3x3 kernels. The number of channels used in the five convolutional layers are [32, 64, 128, 128, 256]. The bottleneck block uses a series of 9 bottleneck residual blocks. The latter five of these blocks use self-attention. The decoder mirrors the encoder and maps the output back to 3 channel RGB space at the input resolution. The decoder has concatenation skip connections with CBAM attention layers. We use Batch Normalization [10] and ReLU activation in both the networks with TanH as the last activation layer.

4.1.2 Datasets

ISTD: The Image Shadow Triplets Dataset or ISTD [21] is a large-scale dataset containing 1870 triplets of shadow, shadow mask and shadow-free images at a resolution of 640x480, split into 1330 training and 540 testing triplets. We resize the images to 400x400 before training but retain the original resolution for testing. To reduce the illumination and color discrepancies arising in the paired-image capture, we adopt pre-processing step from [12]. Due to a lack of publicly available high-resolution datasets, we created ISTD-HQ to test our high-resolution inference framework. We use the super-resolution network from [22] and produce upscaled images at 2560x1920 resolution from the original ISTD images. Although this might not match a dataset captured in high-resolution, we believe it will be useful in establishing the applicability of the proposed method.

Figure 5. Sample shadow images from Shadow Food-HQ (SFHQ) dataset.

Shadow Food-HQ (SFHQ): We constructed a new Shadow Triplet dataset comprising of high-resolution food images captured at 12MP (4032x3024) resolution. It consists of 14520 shadow triplets. Shadow mask ground-truths are manually annotated to include externally cast shadows only. The images consist of diverse scenes captured with varying lighting conditions and perspectives. The dataset is divided into 14K training and 520 testing triplets. Similar to ISTD dataset, we use 400x400 images for training and testing. Sample shadow images from the dataset are shown in Figure 5.

4.1.3 Training Details

We train both our networks for 220K iterations with a batch size of 4 using Adam optimizer with a beta 1 of 0.5 and beta 2 of 0.999. We start with a learning rate of 0.0002 and reduce it by half every 80K iterations. We update our generator and discriminator alternatively. During training, we randomly flip the images with 50% probability and add color jitter to augment the data. For LSRNet, the training takes a little over two and a half days on an Nvidia Tesla P40 GPU and three days for DRNet.

4.1.4 Evaluation Metrics

For evaluating Shadow Removal performance, we use Root Mean Squared Error (RMSE) metric. We compute RMSE on the whole image and on shadow region individually and report our numbers in the Lab color space.

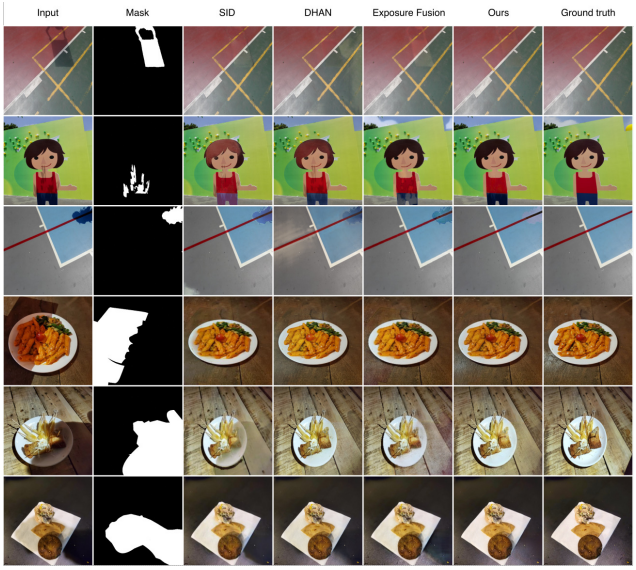

Figure 6. Shadow Removal result comparison of proposed network with SOTA methods SID [12], DHAN [4] and AEF [6] on ISTD and SFHQ. (**Zoom-in for better visualization).**

Method	Dataset	Shadow	Non-Shadow	Whole
SID	ISTD	4.788	3.175	3.482
DHAN	ISTD	4.649	3.137	3.426
AEF	ISTD	3.465	2.11	2.388
Ours	ISTD	**3.164**	**1.606**	**1.955**
SID	SFHQ	7.497	3.558	4.755
DHAN	SFHQ	6.826	2.207	3.767
AEF	SFHQ	7.572	3.182	4.582
Ours	SFHQ	**6.017**	**1.928**	**3.32**

Table 1. Quantitative result comparison of our proposed network against SID [12], DHAN [4] and AEF [6] in terms of Lab color space RMSE on ISTD and SFHQ dataset.

Configuration	Params	ISTD	SFHQ
Base Network	1.97M	3.32	6.55
+ Gamma Augmentation	1.973M	3.31	6.52
+ CBAM	2.0M	3.18	6.42
+ Self-Attention	2.371M	3.14	6.19
+ Discriminator Attention	2.371M	3.16	6.02

Table 2. LSRNet: Ablation Study – Number of network parameters, Shadow RMSE in Lab color space in ISTD and SFHQ datasets.

4.1.5 Qualitative Comparison

We first provide qualitative image comparison on low-resolution images between our proposed LSRNet and existing state-of-the-art shadow removal methods SID [12], DHAN [4] and AEF [6] in Figure 6. Results are shown on standard ISTD dataset and resized low-resolution images from SFHQ dataset. As depicted in the figure our method produces artefact-free results across different scenarios.

In Figure 7, we provide qualitative image comparison between our proposed method and existing state-of-the-art shadow removal methods on progressively increasing image resolutions (256x256, 512x512, 1024x1024, 2048x2048). As shown in the figure the shadow removal quality deteriorates gradually as we increase the image resolution proving that existing methods do not adapt well to high-resolution images. In contrast, our proposed dual-stage approach using LSRNet and DRNet can reliably scale to high-resolution images.

In the inference phase for ISTD dataset we obtain the shadow masks using the proposed shadow detector for our method. For SID, DHAN and AEF, low resolution ISTD results and metrics are obtained using the official models and results released by the authors. High-resolution ones are inferred using the official weights shared. For SFHQ dataset the models are trained using the default parameter choices of the respective code-bases. Additionally, to ensure fairness, for testing on SFHQ dataset, ground-truth shadow masks is used for all the methods including ours.

4.1.6 Quantitative Comparison

In Table 1, we provide quantitative evaluation between our proposed method and existing shadow removal methods. The proposed LSRNet outperforms existing methods on both ISTD and SFHQ datasets. In addition, at 2.4 million parameters and 18 GFLOPs our proposed method is significantly lightweight and computationally efficient than existing methods.

4.1.7 Ablation Study

In this ablation study, we analyze the effects of our network components and benchmark them on the ISTD [21] and SFHQ datasets. For each design, we provide the network parameter count and the RMSE values for shadow regions in Lab color space. Our baseline network without any attention blocks and without using any gamma augmented images has an RMSE of 3.32 and 6.55 respectively on these datasets. Adding three gamma augmented images to the input improves the numbers marginally while also not affecting the complexity much. Adding Convolution Block Attention [23] in the decoder brings down the shadow RMSE by 4% and 1.5% respectively. With self-attention in the generator and discriminator networks, on the SFHQ dataset the model sees an improvement of 6% in the shadow region whereas on the ISTD dataset it improves by 1%. The results are summarized in Table 2 with qualitative results shown in Figure 8.

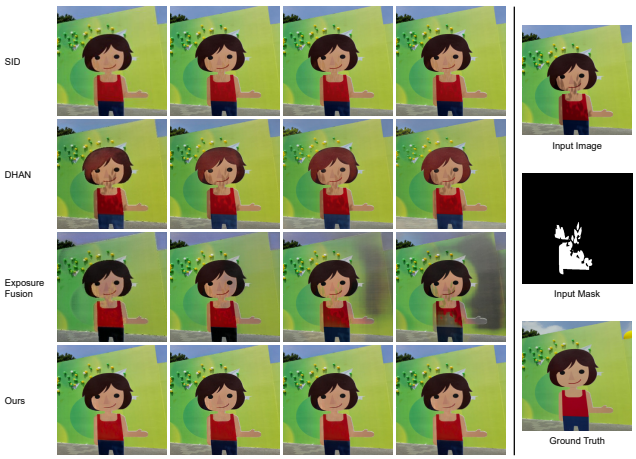

Figure 7. Visual comparison showing shadow removal results at progressively increasing image resolutions (left to right)(256, 512, 1024, 2048). While shadow removal quality degrades as we increase the resolution for other methods it remains consistent for our proposed method. (**Zoom-in for better visualization).**

Network	Params	FLOPs(G)			
		256	512	1024	2048
LSRNet	2.37M	17.73	73.0	326.78	1859.76
DRNet	0.49M	5.34	21.28	85.53	342.12

Table 3. Computation efficiency comparison between LSRNet and DRNet in terms of GFLOPs at different input resolutions of (256x256), (512x512), (1024x1024) and (2048x2048)

In Figure 9, we show the effect of using only LSRNet for shadow removal at different resolutions. As described earlier having a large receptive field in the network is important to obtain shadow-remnant free outputs. At low-resolution the receptive field of the network is the largest and as the resolution increases the output quality degrades progressively as shown in the figure. Moreover, in the proposed 2-stage approach as DRNet is only used for detail refinement of the shadow-free output, the network at only 0.49M parameters is computationally more efficient than LSRNet. As shown in Table 3, more than 5x computational gain is achieved with DRNet over LSRNet for high-resolution images (2048x2048). This also helps in deploying the solution in resource constrained environments like embedded devices.

4.2. Shadow Detection

4.2.1 Training Details

The proposed shadow detection network is trained for 35k iterations with a learning rate of 0.005, batch size of 16 and SGD optimizer with a momentum of 0.9.

Method	Parameters	FLOPs (G)	SBU	ISTD
A+D Net	54.41M	22.31	5.37	3.23
DSC	79.03M	212.87	5.59	3.42
DSDNet	58.16M	106.43	3.45	2.17
BDRAR	42.46M	117.56	3.64	2.69
MTMT-Net	44.12M	142.32	**3.15**	**1.77**
FSDNet	4.4M	8.74	8.8	3.67
Ours	**3.2M**	**8.54**	5.59	2.23

Table 4. Quantitative comparison of our Shadow Detection method against state-of-the-art methods in terms of BER.

4.2.2 Evaluation Metrics

We evaluate our network using one of the standard and largest publicly available shadow datasets, namely SBU [20] in addition to ISTD. Quantitative evaluation for the shadow detection performance is done by calculating the BER (Balanced Error rate) between the ground truth mask and predicted shadow mask.

$$BER = (1 - \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})) * 100 \quad (7)$$

Where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives respectively. Lower BER values indicate better shadow detection result.

4.2.3 Results

We compare our proposed network with several shadow detection techniques: BDRAR [26], DSDNet [24], DSC [9], MTMT-Net [3] and FSDNet [8] and across different datasets. We obtain BER of all networks except FSDNet by directly taking the results from the authors' and for FSDNet by training the network ourselves. As depicted in Table 4, although some of the recent techniques such as DSDNet and MTMT-Net are better in BER metrics when compared to our proposed network, it comes with a cost in efficiency due to the number of network parameters used. FSDNet and our proposed network are the only lightweight networks with 4.4M and 3.2M training parameters respectively. Our method outperforms FSDNet in accuracy across different datasets with 36% and 39% improvement in BER on SBU and ISTD dataset respectively. Sample results comparing our network qualitatively with some of the existing state-of-the-art techniques are shown in Figure 10.

4.3. On-device implementation

We deploy the proposed architecture consisting of shadow detection and shadow removal networks on a latest Android Smartphone device (Samsung Galaxy S22) powered by the Qualcomm Snapdragon 8 Gen 1 chipset and having 12GB of RAM. The networks are converted to

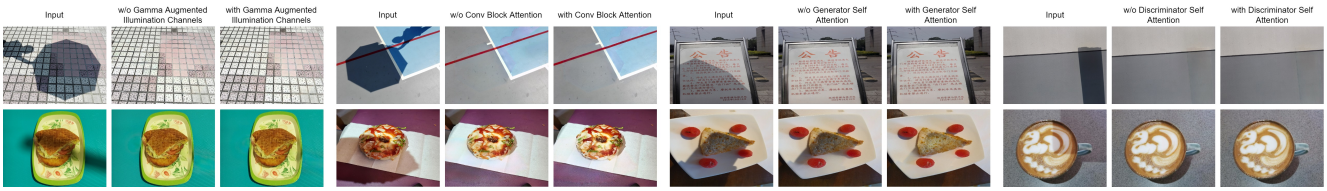

Figure 8. Visual results showing the effect of the different network components in the proposed shadow removal network.

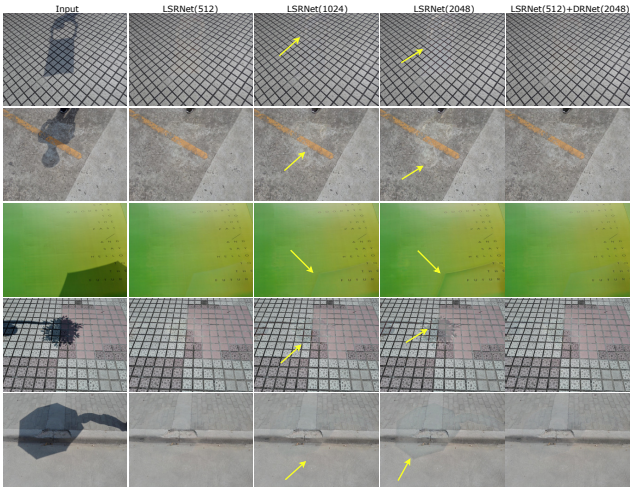

Figure 9. Qualitative result comparison to demonstrate the relationship between resolution and shadow removal quality. Using only LSRNet, results progressively degrade with increase in resolution. With the proposed dual-stage approach (LSRNet + DRNet) the output quality is retained at high-resolution. (**Zoom-in for better visualization).**

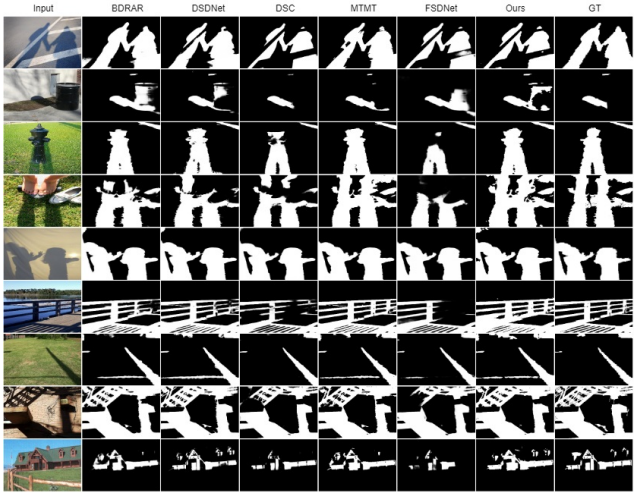

Figure 10. Qualitative results of our proposed Shadow Detection method, compared with other shadow detection methods.

Network	Execution Time (ms)	Count	Total (ms)
Detector	70	1	70
LSRNet	290	1	290
DRNet	170	12	2040
Total	-	-	**2400**

Table 5. On-device performance on latest Smartphone device for 12MP (4032x3024) resolution image using the proposed shadow detection and removal networks.

Tensorflow-Lite format and GPU delegation is used during inference. We report the execution numbers with the proposed shadow removal architecture on a 12MP (4032x3024) image. In the on-device implementation, for DRNet, we use the image tiling approach and perform multiple inferences (12 tiles for 12MP image) to reduce memory and computation footprint. As shown in Table 5, the end-to-end execution time is around 2.4 secs on a modern smartphone device proving the efficiency of the proposed technique.

5. Conclusion

In this paper, we have proposed a novel shadow removal architecture SHARDS, that can perform shadow removal on high-resolution images. The proposed architecture, even though has lesser network parameters than existing techniques, outperforms them on existing low-resolution datasets and can further adapt to high-resolution images as well. In addition, we propose an efficient and lightweight shadow detection network that compares favorably against existing techniques. It is also shown that the proposed method can be efficiently deployed in a modern smartphone device to remove shadows from high-resolution images proving the real-world applicability of the proposed solution. Like other similar methods, the requirement of a paired dataset limits the generalization of the approach. However, the ideas proposed are equally applicable to unpaired methods as well. While the existing datasets are very simplistic, real world shadows often involve self-cast shadows removing which would not be ideal. The detection and removal methods should be robust against these which would be explored in a future work.

References

[1] Liang-Chieh Chen, G. Papandreou, Florian Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[2] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[3] 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 *CVPR*, 2020.

[4] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10680–10687, Apr. 2020.

[5] Graham D. Finlayson, Steven D. Hordley, and Mark S. Drew. Removing shadows from images. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *Computer Vision — ECCV 2002*, pages 823–836, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[6] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10566–10575, 2021.

[7] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[8] 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. *arXiv preprint arXiv:1911.06998*, 2019.

[9] X. Hu, L. Zhu, C. Fu, J. Qin, and P. Heng. Direction-aware spatial context features for shadow detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018.

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[11] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[12] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8577–8586, 2019.

[13] 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 European Conference on Computer Vision*, 2018.

[14] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8758–8767, 2021.

[15] 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.

[16] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4925–4934, 2021.

[17] V. Nguyen, T. F. Y. Vicente, M. Zhao, M. Hoai, and D. Samaras. Shadow detection with conditional generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4520–4528, 2017.

[18] 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, 2017.

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

[20] 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.

[21] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 2018.

[22] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing.

[23] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 3–19, Cham, 2018. Springer International Publishing.

[24] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. Distraction-aware shadow detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[25] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[26] 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 Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 122–137, Cham, 2018. Springer International Publishing.