Pyramid Ensemble Structure for High Resolution Image Shadow Removal

Shuhao Cui[1], Junshi Huang[1], Shuman Tian[1], Mingyuan Fan[1], Jiaqi Zhang[1],
Li Zhu[1], Xiaoming Wei[1], Xiaolin Wei[1]
Meituan Group
{cuishuhao,huangjunshi}@meituan.com

Abstract

Existing methods for shadow removal in high-resolution images may not be effective due to challenges such as the time-consuming nature of training and the loss of visual data during image cropping or resizing, highlighting the necessity for the development of more efficient methods. In this paper, we propose a novel Pyramid Ensemble Structure (PES) for High Resolution Image Shadow Removal. Our approach takes advantage of multiple scales by constructing pyramid inputs that allow for the capturing of a wide range of shadow sizes and shapes. We then train the network in pyramid stages to enhance global information processing. Furthermore, an ensemble of different shadow removal models is employed, and the maximum value is chosen to indicate the least amount of remaining shadow in the output. Experiments on both validation and testing data sets confirm the effectiveness of our method. In the Image Shadow Removal Challenge competition, our method obtained 22.36 PSNR score (1st place) and 0.70 SSIM score (2nd place) on the test sets.

1. Introduction

Computer vision faces the challenging problem of high resolution shadow removal, which involves eliminating shadows from high-resolution images. Despite their ubiquity in everyday environments, shadows pose a significant challenge to image processing due to their inherently complex nature: they are caused by varying illumination conditions and can have a significant impact on the visibility and quality of the image. In addition, shadows create a shift in pixel intensity, hampering accurate interpretation of the image and rendering it difficult to carry out subsequent analyses. Addressing this issue requires advanced algorithmic techniques and the ability to robustly identify and remove shadows irrespective of context, lighting conditions, and image characteristics.

Recent years have seen promising advances in shadow removal through the use of deep learning-based techniques. Specifically, these techniques employ convolutional neural networks (CNNs) [6, 20] to learn intricate features from data, facilitating the production of high-quality, shadow-free images. In addition to CNN-based approaches, researchers have explored the application of transformer networks [11], which leverage transformer building units to capture global contextual information and produce enhanced results in shadow removal. These novel techniques demonstrate the great potential of deep learning in accurately and effectively addressing the complex problems associated with normal shadow removal in computer vision.

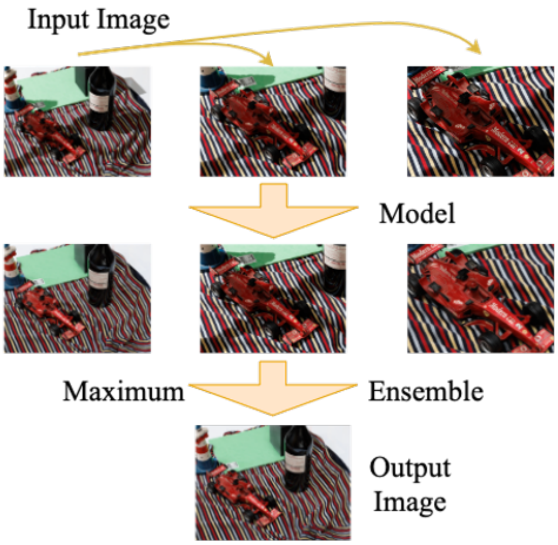

Figure 1. Illustration of Pyramid Ensemble Structure. To achieve shadow-free outputs, multiple scale images are trained across various stages. These outputs are then ensembled by leveraging the maximum results obtained from a range of networks.

However, when it comes to high-resolution shadow removal [23, 24], simply applying existing shadow removal methods may not yield the desired results. Two key issues may arise. Firstly, directly training the methods with input image shapes may prove to be both time-consuming and less effective. This is especially true in high-resolution scenarios where the images tend to be much larger in size. Image cropping or resizing may become a necessity to tackle this, but such measures can result in a loss of important visual data. It is therefore essential to develop new methods that can handle high-resolution images with greater efficiency. The second challenge involves the issue of global information on the image. Directly cropping images as part of the shadow removal process may lead to a loss of important context information. As a result, networks may not be able to effectively construct a well-performed contextual understanding of the image. This may lead to sub-optimal outcomes, making it important to develop methods that take a more comprehensive approach to global information retention.

We present an innovative solution, the Pyramid Ensemble Structure (PES), to tackle the challenging task of High-Resolution Image Shadow Removal. Figure 1 illustrates our approach, which employs pyramid inputs to capture diverse shadow sizes and shapes at multiple scales. Additionally, we train the network in pyramid stages to improve global information processing. To achieve high-precision shadow removal, we use an ensemble of different shadow removal models and select the maximum output value representing the best possible shadow removal outcome. To further enhance the network and achieve superior shadow removal performance, we employ a model soup technique. Our method has been evaluated on both validation and testing datasets, and the experimental results confirm its effectiveness. PES consistently outperforms other methods in terms of shadow removal accuracy while maintaining high-resolution image quality.

We introduce our novel solution, the Pyramid Ensemble Structure (PES), which excellently addresses the challenging High-Resolution Image Shadow Removal task. As depicted in Figure 1, our approach adopts pyramid inputs to capture various shadow sizes and shapes across multiple scales. Moreover, we train the network in pyramid stages to facilitate global information processing. For precise shadow removal, we leverage an ensemble of diverse shadow removal models and select the output value that yields the optimal shadow removal outcome. To further elevate the network's performance and achieve superior shadow removal results, we apply a model soup technique. Through comprehensive evaluations on validation and testing datasets, our approach demonstrates remarkable effectiveness. In fact, PES consistently outperforms other methods concerning shadow removal accuracy while preserving high-resolution image quality.

The contribution of the paper can be summarized as:

1. Firstly, we propose a novel Pyramid Ensemble Structure (PES) that allows for the effective removal of shadows from high-resolution images by capturing a wide range of shadow sizes and shapes.
2. Secondly, by utilizing a model soup technique and an ensemble of different shadow removal models, we achieve superior shadow removal performance, further enhancing the accuracy and effectiveness of our approach.
3. Thirdly, our method is confirmed to be effective through experiments on both validation and testing data sets, showing promise for future development in the field of image processing.

2. Related Work

Image restoration has been a popular topic in computer vision for many years. One of the classic methods for image restoration is based on the maximum a posteriori (MAP) estimation [10, 21, 27]. The MAP-based methods generally assume a prior on the image structure and use the observed image and the prior to estimate the restored image. Deep neural networks have shown great success in many computer vision tasks [5, 7, 8, 15], including image restoration. Many approaches have been proposed to use deep neural networks for image denoising, deblurring, and super-resolution [1, 22]. One of the critical factors in achieving successful image restoration is the development of a powerful network. To this end, several researchers have proposed and designed high-performing networks, as documented in recent studies [4, 18, 19]. These networks offer effective methods that aim to improve the quality of input images before the restoration process commences. Their effectiveness in generating better input images paves the way for improved image restoration results.

Among restoration tasks, recent works on shadow removal have utilized high-quality ground truth as guidance, as noted in [12–14, 26]. However, some approaches still reconstruct the shadow-free image under a physical illumination model while simultaneously predicting an accurate external shadow matte. For instance, Le [17] enhanced shadow regions by employing a physical linear transformation model for image decomposition. Fu [9], on the other hand, proposed an over-exposure fusion approach for shadow removal. Their approach uses a learnable pixel-wise weighting map to blend a series of over-enhanced shadow images with the original shadow image in an intelligent way.

There have been a number of approaches proposed in the literature for leveraging context information to improve shadow removal performance. One such method, called DeshadowNet [20], employs a multi-level feature combination strategy to increase the network receptive field and incorporates contextual semantic and appearance information. This approach can generate high-quality predictions of shadow mattes with fine local details. Another approach, recently proposed by Cun et al. [6], uses dilated convolutions as a backbone to capture context features. Building on this work, Chen et al. [3] introduced CANet, which incorporates an external patch matting module that selects the top K similar patches to explore potential contextual relationships. In contrast to these methods, ShadowFormer [11] uses transformer building units to capture global contextual information in an end-to-end manner.

3. Method

In Shadow Removal, we are given an input image I_i, possibly with shadows,together with an ground truth image I_g, with shadows removed. The task of shadow removal is to remove the shadows present in an input image I_i, producing an output image I_o that is as similar to I_i as possible, while also ensuring that the shadows are no longer visible.

3.1. Pyramid Inputs

To create a pyramid input image, we start with the input image I_i, which has dimensions $H \times W$. Since H and W may be quite large, training models directly on such images can be time-consuming. To overcome this, we scale the input image to multiple resolutions with ratio σ_1, σ_2, σ_3, resulting pyramid images $P(I)$ as follows:

$$P(I_i, \sigma_1, \sigma_2, \sigma_3) = \{I_{i\frac{H}{\sigma_1}\times\frac{W}{\sigma_1}}, I_{i\frac{H}{\sigma_2}\times\frac{W}{\sigma_2}}, I_{i\frac{W}{\sigma_3}\times\frac{W}{\sigma_3}}\}.$$

By doing so, we can efficiently capture both local and global contents of the image across different resolutions. This, in turn, ensures that our model is robust and effective at detecting features at different levels of resolution. Compared to normal random resize techniques, using a pyramid input image provides even greater robustness and adaptability to varying image resolutions.

To expedite the training process, we leverage a preprocessing technique to reshape the image in preparation for training. However, given the limited images available, it is crucial to augment the dataset by introducing more diverse images. One effective way to achieve this is by employing a pyramid cropping method where images with varying shapes are cropped into uniform sizes. This effectively bolsters the diversity of the input data. In addition, we further improve the dataset by cropping the images into the same shape of $\frac{H}{\sigma_4} \times \frac{W}{\sigma_4}$, the cropped images I_c based on $C(I_i, \sigma_1, \sigma_2, \sigma_3, \sigma_4)$ can be denoted as:

$$I_c = C(I_i, \sigma_1, \sigma_2, \sigma_3, \sigma_4) = crop(P(I_i, \sigma_1, \sigma_2, \sigma_3), \frac{H}{\sigma_4} \times \frac{W}{\sigma_4})$$

where σ_4 is the crop ratio and *crop* denotes random-crop function.

3.2. Pyramid Stages

The task at hand involves generating an output image I_o from an input image I_i, using a neural network N. The network N is specifically designed for image-to-image translation tasks and has been trained on a large dataset of pairs of images. This allows the network to learn a mapping between different domains, in this case, from the domain of the input image I_i to the domain of the output image I_o, which means $I_o = N(C(I_i, \sigma_1, \sigma_2, \sigma_3, \sigma_4))$.

The network N typically consists of an encoder-decoder architecture, where the encoder extracts high-level features from the input image and the decoder generates the output image from these features. In addition, skip connections are employed in the network to ensure that the low-level details of the input image are preserved in the output image. During training, the network N learns to minimize a loss function that quantifies the difference between the generated output image I_o and the ground truth image I_g. The loss functions involve L1 loss L_{L1}, Mean Squared Error (MSE) loss L_{MSE}, Peak Signal-to-Noise Ratio (PSNR) L_{PSNR} and Structural Similarity Index Measures (SSIM) L_{SSIM}.

The L1 loss is calculated as the absolute difference between the predicted output and the ground truth, as follows:

$$L_{L1}(I_o, I_g) = \frac{1}{n}\sum_{i=1}^{n} |I_o - I_g|.$$

The MSE loss measures the average squared difference between the predicted output and the ground truth, as follows:

$$L_{MSE}(I_o, I_g) = \frac{1}{n}\sum_{i=1}^{n} (I_o - I_g)^2$$

The PSNR is a measure of the quality of the predicted output compared to the ground truth. It is defined as the ratio between the maximum possible power of a signal and the power of the noise that affects the fidelity of its representation, as follows:

$$L_{PSNR}(I_o, I_g) = 10\log_{10}\left(\frac{MAX_I^2}{L_{MSE(I_o, I_g)}}\right)$$

where MAX_I is the maximum possible pixel value of the image. The SSIM measures the similarity between two images by computing a combination of brightness, contrast, and structural similarity, as follows:

$$L_{SSIM}(I_o, I_g) = \frac{(2\mu_{I_o}\mu_{I_g} + c_1)(2\sigma_{I_o I_g} + c_2)}{(\mu_{I_o}^2 + \mu_{I_g}^2 + c_1)(\sigma_{I_o}^2 + \sigma_{I_g}^2 + c_2)}$$

where μ and σ are the mean and standard deviation of the image intensities, $\sigma_{I_o I_g}$ is the covariance between the two images, and c_1 and c_2 are constants to avoid dividing by zero.

To build our model, we construct pyramid stages based on the loss functions available. In stage 1, we focus more

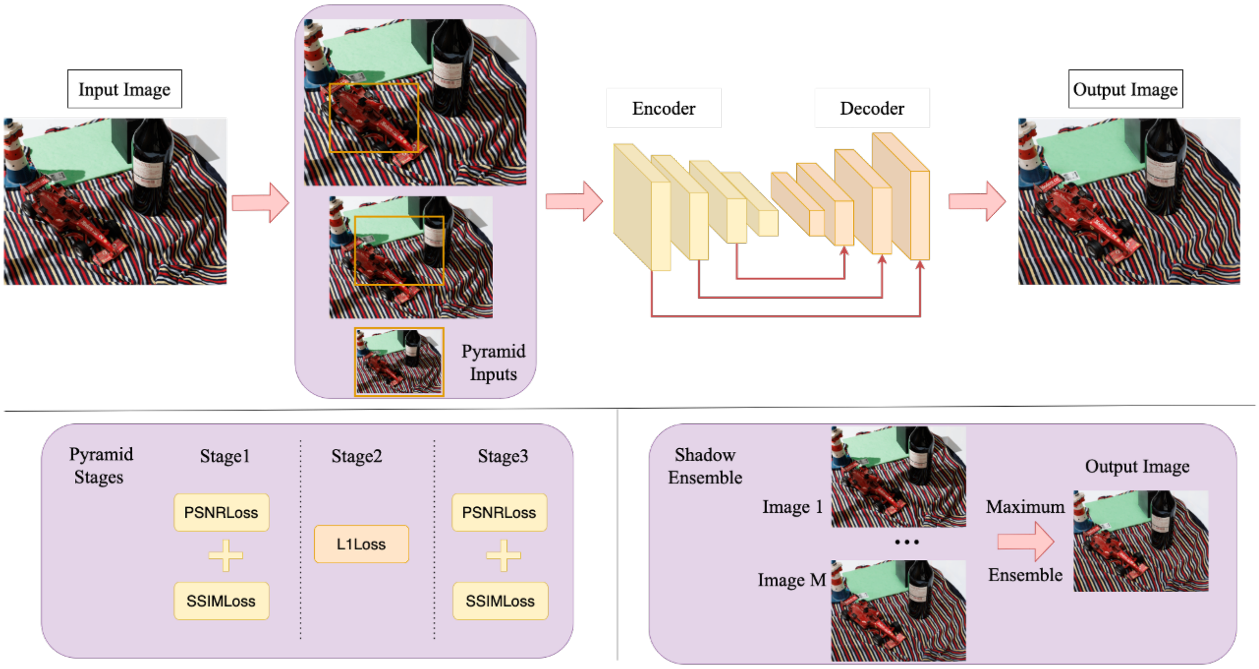

Figure 2. Overall framework of Pyramid Ensemble Structure (PES) . First, we implement Pyramid Inputs, which entails resizing and cropping the input images into various sizes and shapes. Once adjusted, the input images are then forwarded to the network for processing, which is trained based on diversity loss functions in Pyramid Stages. Finally, the output images are ensembled by selecting the maximum result from the various options available.

on local regions by using larger values of the ratios σ_1, σ_2, σ_3, and σ_4. To optimize the model, we adopt PSNR loss and SSIM loss functions during stage 1. The loss functions for stage 1 can be derived using the following formulas:

$$\begin{aligned} L_{Stage1} &= L_{PSNR}(N(C(I_i, \sigma_1, \sigma_2, \sigma_3, \sigma_4)), I_g) \\ &\quad + L_{SSIM}(N(C(I_i, \sigma_1, \sigma_2, \sigma_3, \sigma_4)), I_g). \end{aligned}$$

Moving on to stage 2, we refine our model by adopting the L1 loss function, which further improves its accuracy. To achieve this, we reduce the ratios by half as compared to stage 1. The loss functions for stage 2 can be computed using the following formulas:

$$L_{Stage2} = L_{L1}(N(C(I_i, \frac{\sigma_1}{2}, \frac{\sigma_2}{2}, \frac{\sigma_3}{2}, \frac{\sigma_4}{2})), I_g).$$

In the final stage, we continue to refine the model by using both PSNR loss and SSIM loss functions, which have proven to be effective in enhancing image quality. Additionally, we further reduce the ratios by half compared to stage 2 to fine-tune the model for better performance. The loss functions for stage 3 can be derived using the following formulas:

$$\begin{aligned} L_{Stage3} &= L_{PSNR}(N(C(I_i, \frac{\sigma_1}{4}, \frac{\sigma_2}{4}, \frac{\sigma_3}{4}, \frac{\sigma_4}{4})), I_g) \\ &\quad + L_{SSIM}(N(C(I_i, \frac{\sigma_1}{4}, \frac{\sigma_2}{4}, \frac{\sigma_3}{4}, \frac{\sigma_4}{4})), I_g). \end{aligned}$$

By incorporating the pyramid stages described above, the network is able to capture both local and global information, allowing for a more comprehensive analysis and optimization of the input data. The adoption of L1, PSNR, and SSIM loss functions during the training process ensures a balance between local and global similarities, including aspects such as brightness, contrast, and structural similarities. By achieving this balance, we can ensure a better outcome across multiple training iterations, resulting in improved performance overall.

3.3. Shadow Ensemble

In the context of shadow removal tasks, we have made a crucial observation that bright regions rarely ever contain shadows. These areas are more likely to be non-shadow regions, which means we can assume that they are free from shadows. By utilizing this insight, we can directly obtain the maximum prediction among all the shadow models, which allows us to select the areas with the least shadow coverage. Based on this, we propose a simple yet highly effective approach to shadow removal that involves selecting the maximum prediction from all of the shadow models.

To formalize this, we consider a setting where we have a total of M shadow models based on input of cropped image I_c, denoted by $N_1(I_c), N_2(I_c), \ldots, N_M(I_c)$. Each shadow model predicts the regions of the input that contain shadows. We then combine the predictions of all the shadow models to generate a final prediction, denoted by $N(I_c)$, that accurately identifies the shadow regions in I_c. for an image I_c, Then $N(I_c)$ can be calculated as follows:

$$N(I_c) = \max_{j=1}^{M} \{N_j(I_c)\} .$$

Directly obtaining the maximum prediction among the shadow models is equivalent to selecting the regions with the least shadow coverage. Overall, the above approach enables us to effectively combine the predictions of multiple shadow models N_j to generate a final prediction that is both accurate and robust in identifying the regions of an input that contain shadows.

Accordingly, we formulate the framework of Pyramid Ensemble Structure (PES) in Figure 2, which is composed of Pyramid Inputs, Pyramid Stages and Shadow Ensemble. The three main components of the framework work together seamlessly to ensure optimal performance. This multi-component approach has proven to be highly effective in producing accurate, high-quality shadow-free images. The effectiveness of the PES framework can be attributed to the seamless cooperation between its three components, resulting in an optimal and comprehensive solution for shadow removal.

3.4. Model Soup Finetune

When considering shadow removal tasks, let N_j represent the parameters of the jth model, with a total of M models being considered. Building upon the methodology proposed by Wortsman [25], we calculate the average parameter values in the following way:

$$N_{soup} = \frac{1}{M} \sum_{j=1}^{M} N_i.$$

To further enhance the performance of the network, we employ a technique whereby we use the newly generated network parameters as initialization and retrain the network. This process involves leveraging the strengths of multiple models and combining them into a single, more powerful entity, resulting in a network that performs better than any individual model. By utilizing the average parameter values of multiple models, we create a superior starting point for the retraining process. This leads to faster and more efficient convergence, as the network builds upon the collective knowledge of the multiple models. The retrained network is thus able to identify and remove shadows more accurately and efficiently, resulting in a significant improvement in overall performance.

4. Experiments

4.1. Experiment Settings

We chose to utilize NAFNet as our basic network architecture [4] due to its implementation of only nonlinear activation functions. We specifically trained our model using two variations of NAFNet: NAFNet32 and NAFNet64. During our experiments, we found that NAFNet32 provided the best results and thus used it for the majority of our results. We chose Lion [2] optimizer to train the model for its superior performance over Adam [16].

We use the official shadow removal competition dataset, where the image shape is (1440, 1920). As shown in Table 3.1, the input images are cropped into different sizes under different stages. Stage 1 involves three different image sizes for the input, which are reshaped to a crop size of (240, 320). The loss function used is a combination of PSNR Loss and SSIM Loss, which are both measures of the quality of the reconstructed image. The batch size used is 144, and the model is trained for 50,000 iterations. In stage 2, the same three input image sizes are used, but they are cropped to a larger size of (480, 640). The loss function used in this stage is L1 Loss, which is a different measure of image quality than PSNR or SSIM. The batch size is smaller, at 64, and the model is trained for 20,000 iterations. Note that in stage 2 of the training process, we also reshape the images to (480, 640), and crop to (480, 640), to ensure the global information in networks. In Stage 3: This final stage uses only two input image sizes, which are both larger than the largest input size used in the previous stages. These images are cropped to the same size as the larger input size in Stage 2, which is (960, 1280). The loss function used here is again a combination of PSNR Loss and SSIM Loss, and the batch size used is further reduced to 16. The model is trained for 10,000 iterations.

Overall, the three stages involve varying input sizes, crop sizes, loss functions, batch sizes, and numbers of iterations, which are systematically adjusted to optimize the performance of the image reconstruction model. The use of multiple stages with different settings allows the model to learn effectively from the data and achieve good results on a range of input sizes.

In addition to training on different stages, we also generated five high-performing networks by varying hyperparameters and architectures, as summarized in Table 4. The table has four columns: ”Num,” ”Network,” ”Train Stage,” and ”Iteration”. The ”Num” column simply enumerates the models in the ensemble from 1 to 5. The ”Network” column specifies the NAFNet architecture used for each model, with varying depths and widths (NAFNet32 or NAFNet64). The ”Train Stage” column indicates the training stage from Stage 1, Stage 3, or a combination of Stage 1 plus additional training to improve structural similarity (large SSIM). The

Stage Num	**Reshape size**	**Crop size**	**Loss**	**Batch size**	**Iteration**
Stage1	(480, 640), (960, 1280),(1440, 1920)	(240, 320)	PSNR Loss + SSIM Loss	144	50000
Stage2	(480, 640), (960, 1280),(1440, 1920)	(480, 640)	L1 Loss	64	20000
Stage3	(960, 1280),(1440, 1920)	(960, 1280)	PSNR Loss + SSIM Loss	16	10000

Table 1. Different settings in the Pyramid Stages.

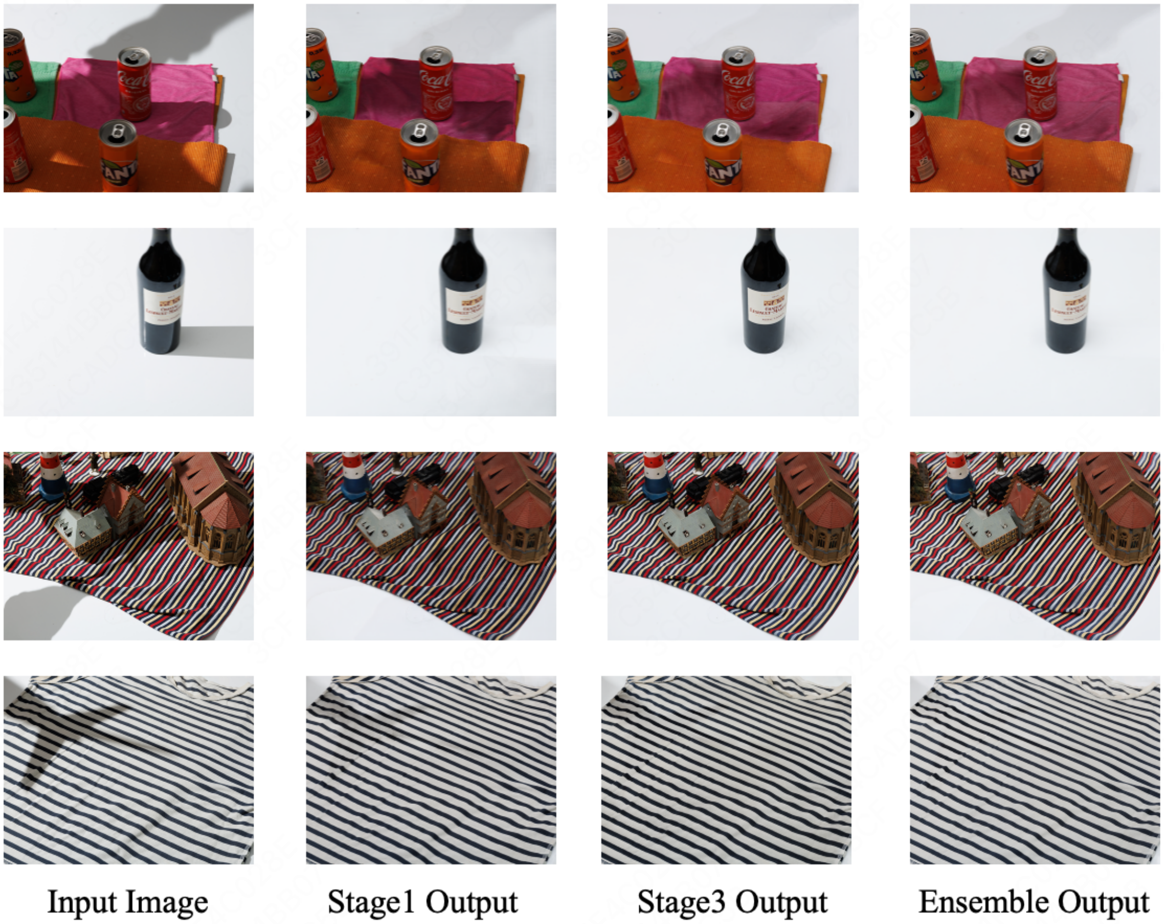

Figure 3. Example of images under different stages of processing. From the original input image to the final ensemble output, we can observe that more shadows have been effectively removed, resulting in an overall improvement in image clarity.

”Iteration” column specifies the number of iterations used for each model. By aggregating the maximum prediction of the ensemble, we successfully eliminated shadows from the images.

4.2. Experiment Results

The final results are shown in Table 2. Our evaluation results demonstrate that our approach achieves the best performance in terms of PSNR and ranks second in SSIM among all the methods tested. This indicates the effectiveness of our approach in improving the quality of image reconstructions. However, our performance on SSIM is not as high as our performance on PSNR. This may be attributed to the smaller weights assigned to the loss function of SSIM compared to other loss functions used in our approach. Although SSIM is an important measure of the perceptual quality of reconstructed images, its contribution to the overall loss function might have been comparatively smaller. We would like to emphasize that our approach still achieved competitive results on SSIM despite the smaller weight of the loss function, hence demonstrating the robustness and effectiveness of our approach overall. Further analyses could be carried out in future work to better balance the weights of loss functions and achieve even better

Name	PSNR	SSIM
Goring	11.83	0.37
zwl	13.38	0.45
fvasluianu	15.03	0.47
CD_luo	17.36	0.53
Concentration7	17.57	0.55
jiangchengzhi	17.74	0.5
nbyqh	17.78	0.55
zjuShen	17.83	0.53
Yuki-11	18.08	0.53
amaguri	18.08	0.53
some	18.58	0.59
Concentrate-Silence	18.73	0.56
try22	18.87	0.56
duchongyang	19.14	0.6
Jaszheng	19.14	0.6
nann	19.18	0.59
Concentration	19.23	0.6
to42	19.55	0.61
ir-sde	19.60	0.58
priyakansal	19.67	0.63
shrutiphutke	19.71	0.63
jane_j	19.82	0.64
user118	20.40	0.62
SabariNathan	20.56	0.63
chong40	20.68	0.62
BowenZhao	20.73	0.62
thea	20.75	0.65
leeyeoreum01	21.02	0.66
WangtaekOh	21.08	0.66
tiger	21.13	0.65
Krocy	21.24	0.66
Una	21.25	0.67
HuanZheng	21.43	0.68
Rebecca	21.58	0.68
leaves	21.68	0.69
codalab123	21.69	0.69
daylight	21.70	0.69
mrchang87	21.79	**0.70**
xyz123	22.20	0.69
PES (ours)	**22.36**	**0.70**

Table 2. PSNR and SSIM of Testing Results. Compared with other teams, we obtain better results on PSNR and SSIM.

results on both PSNR and SSIM.

Ablation Study	**PSNR**	**SSIM**
NaFNet	21.69	0.699
+Pyrimid Inputs+Lion	23.19	0.735
+Pyrimid Stages	23.46	0.749
+Model soup finetune	23.59	0.750
Ensemble	23.64	0.760

Table 3. Ablation study results in terms of PSNR and SSIM on the validation.

Num	Network	Train Stage	Iteration
1	NAFNet32	Stage1	40000
2	NAFNet32	Stage3	10000
3	NAFNet32	Stage1+large SSIM	60000
4	NAFNet32	model soup	10000
5	NAFNet64	Stage1	120000

Table 4. The ensembled networks.

The ablation study process can be analyzed based on the changes in PSNR and SSIM values as we move from one stage to another in Table 3. The initial model, NAFNet, has a PSNR of 21.69 and an SSIM of 0.699. The addition of Pyrimid Inputs and Lion results in a significant improvement in PSNR to 23.19 and SSIM to 0.735. Moving on to Pyrimid Stages, we see a further improvement in PSNR to 23.46 and SSIM to 0.749. The addition of Model soup finetune produces a slightly higher PSNR of 23.59 and an SSIM of 0.750. Finally, the ensemble of all the models results in the highest values of PSNR and SSIM at 23.64 and 0.760, respectively. Therefore, the ablation study process indicates that each stage of the model development improves the PSNR and SSIM values, with the ensemble of all the models providing the best results. This information can be useful for optimizing the model development process to achieve better results for similar tasks in the future.

We have also included Figure 3 to present some of our partial results. The figure provides a comprehensive view of the entire process, from the input images to the final ensemble output images. As can be observed from the figure, after the completion of the training process in Stage 1, a significant improvement had been achieved as the shadows on the background were almost entirely removed, while the shadows on the cloth appeared to be relatively weaker. Subsequently, after undergoing the training process in Stage 3, notable progress had been made as the shadows on the background were completely eliminated and the shadows on the cloth appeared to be less pronounced than in Stage 1. Finally, following the model ensemble, the shadows on the cloth seemed to have been almost fully removed, thereby attesting to the effectiveness of our approach.

4.3. Discussions

To enhance the robustness of our network, we first attempted to initialize it by training on a well-known and widely used public dataset for shadow removal [20]. However, despite our efforts, this approach did not yield the desired results, and we were compelled to explore alternative options. As such, we evaluated two highly effective algorithms, RLFN and BSRN [18], but unfortunately, even these algorithms did not produce satisfactory results. Recognizing the need for further improvement, we also experimented with the incorporation of shadow detection into our approach. However, we ran into the obstacle of an inaccurate definition of shadow, which prevented us from resolving the issue.

Upon reflection, we acknowledge that we have some regrets concerning our approach to the competition. Looking back, we believe that our results could have been improved if we had employed a technique involving the forwarding of our networks on localized regions and subsequently ensemble them with the full images. However, due to the constraints of time, we were unable to explore this approach in detail. In retrospect, we recognize that by implementing this technique, we could have potentially improved the overall performance of our network. Nevertheless, we remain proud of the effort we put forth and the results we were able to achieve given the constraints of the competition.

5. Conclusion

In this paper, we introduced a new method, the Pyramid Ensemble Structure (PES), to address the challenging problem of High Resolution Image Shadow Removal. Our approach exploits multiple scales by constructing pyramid inputs, which helps capture a wide range of shadow sizes and shapes. The employment of pyramid stages during neural network training improves global information processing, while the use of an ensemble of different shadow removal models and model soup technique further refines the network to ensure high-precision shadow removal. Our experimental results demonstrate the superior performance of PES, achieving state-of-the-art results in shadow removal accuracy while preserving high-resolution image quality. We believe that our approach can contribute to a wide range of applications where shadow removal from high-resolution images is crucial.

References

[1] Harshit Burger, Christian Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *Computer Vision–ECCV 2012*, pages 13–27. Springer, 2012. 2

[2] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023. 5

[3] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 4743–4752, 2021. 3

[4] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2022. 2, 5

[5] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[6] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proc. AAAI Conf. on Artificial Intelligence*, pages 10680–10687, 2020. 1, 3

[7] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9716–9725, June 2021. 2

[8] Zhengcong Fei, Shuman Tian, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Meta-ensemble parameter learning. *arXiv preprint arXiv:2210.01973*, 2022. 2

[9] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 10571–10580, 2021. 2

[10] Dan Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 349–356. IEEE, 2009. 2

[11] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 1, 3

[12] Lanqing Guo, Siyu Huang, Haosen Liu, and Bihan Wen. Fino: Flow-based joint image and noise model. *arXiv preprint arXiv:2111.06031*, 2021. 2

[13] Lanqing Guo, Renjie Wan, Guan-Ming Su, Alex C Kot, and Bihan Wen. Multi-scale feature guided low-light image enhancement. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 554–558. IEEE, 2021. 2

[14] Lanqing Guo, Renjie Wan, Wenhan Yang, Alex Kot, and Bihan Wen. Enhancing low-light images in real world via cross-image disentanglement. *arXiv preprint arXiv:2201.03145*, 2022. 2

[15] 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. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 8578–8587, 2019. 2

[18] Yawei Li, Kai Zhang, Radu Timofte, Luc Van Gool, Fangyuan Kong, Mingxi Li, Songwei Liu, Zongcai Du, Ding Liu, Chenhui Zhou, et al. Ntire 2022 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1062–1102, 2022. 2, 7

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2

[20] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 4067–4075, 2017. 1, 2, 7

[21] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992. 2

[22] Christian J Schuler, Ulrich Lemmin, and Stefan Harmeling. A benchmark for image restoration with a spatially variant blur. *IEEE Transactions on Image Processing*, 29:6508–6521, 2020. 2

[23] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[24] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. Ntire 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[25] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022. 5

[26] Rongkai Zhang, Lanqing Guo, Siyu Huang, and Bihan Wen. Rellie: Deep reinforcement learning for customized low-light image enhancement. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2429–2437, 2021. 2

[27] Xiao-Ping Zhang, Yu-Mei Yang, and En-Min Feng. Multichannel singular spectrum analysis for signal separation and image restoration. *IEEE Transactions on Image Processing*, 17(10):1911–1926, 2008. 2