Shadow Removal based on Diffusion, Segmentation and Super-resolution Models

We would like to thank all reviewers for your constructive comments. The point-to-point responses to concerns raised by each reviewer are as follow. To make clear answers, we apology for adjusting the sequence of some responses. We have included a list of changes in section F.

A. To Reviewer #1

a) The proposed solution was not very competitive among solutions submitted to NTIRE 2024 Image Shadow Removal Challenge.

Thanks for your suggestions. We have update detailed descriptions of our method's ranking and contribution to this track in section 2.3. Specially, we have conducted more study and analysis in this paper and provide valuable insights for this track. For example, we introduce the SAM masks [1] to eliminate edge artifacts caused by stitching during slice inference, resulting in a performance increase of 0.4 dB. Also, diffusion models may not perform well when handling shadows on black objects or deep shadows.

b) Only one method ShadowFormer is compared in the section of Experiment.

Thanks for pointing it out. We have updated Table 1 and Figure 4 in the revised paper and add SpA-Former [3] for comparison, which also achieves best SSIM metric in the listed methods. We also put this table here, as shown in Table 1.

Table 1. Results on the validation dataset of the NTIRE24 Image Shadow Removal Challenge. It is important to note that the results with * presented here are trained by ourselves using their official training code. Ours (patch) indicates the *Diffusion + patch splitting* in Table 2 in our camera ready paper.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
SpA-Former*	22.09	0.7436	**0.1471**	78.92
ShadowFormer*	23.82	**0.8156**	0.2190	60.62
Ours	23.91	0.7772	0.3101	49.55
Ours (patch)	**24.8280**	0.7820	0.2123	**45.80**

c) Cannot find any quantitative and qualitative comparisons for these two datasets for ISTD and SRD dataset. Thanks for your suggestions. We have done experiments on ISTD and SRD dataset for ShadowFormer model to ensure that its official training code can reproduce the results in its paper, and the answer is yes. Then, we trained this model on the challenge dataset, and compared the results with our method, as shown in Table 1 and Figure 4 in our revised paper. We have removed the ambiguous descriptions about these two datasets.

d) For Fig.5, the proposed method tends to generate blurred results with less details, thus leading to lower metrics.

Thanks for your instructive suggestions. In our revised paper, we have modified the expression mistakes. For the images in Figure 5, our method lags behind in metrics but has better shadow removal effects, which proves by the less shadow artifacts of our method.

e) Author should also check their citations carefully to avoid repeated references.

Thanks for point it out. In our revised paper, we have carefully examined the citations and modified the repeated references.

B. To Reviewer #2

a) The effect is not outstanding enough to contribute to the community: Insufficient PSNR and MOS, not good enough visual effect, rank 9th on Challenge. The effect of the proposed method is average, the contribution is not significant, and the ranking is relatively low.

Thanks for your suggestions. In our revised paper, section 2.3 gives detailed descriptions of the contribution of our method. Especially for shadow removal task, our framework generates much better shadow free images than other methods, which is shown in Figure 4 in the revised paper. More importantly, we discuss the usage of the diffusion models in the shadow removal task and give valuable study for the combination with SAM [1], LLaVA [2], which may be useful to future works in this task.

b) The paper mentions the method with the parameter quantity of 5 Millions, the parameter size is so small when using diffusion and LLaVA (Large Language Vision Assistant) models.

Thanks for your detailed suggestions. We have updated the number of parameters in our revised paper. Moreover, we give the parameters and inference time of the key modules in our model (also listed below in Table 2). This helps to balance the increase in inference cost with the performance gain obtained.

Model	Parameters (Million)	Inference (seconds)
Diffusion	52.21	44.97
HAT-SR	40.70	10.55
SAM	641.09	75.60

Table 2. Parameters of each module in our framework

C. To Reviewer #3

a) For the typos and writing errors, I hope that the authors will see this comment and improve on their writing before the camera ready deadline.

Thanks for pointing it out and the detailed suggestions. We have polished our paper and fixed some typos in our camera ready paper. We have rewritten all paragraphs, carefully revised the sentence structures, and corrected any unclear or erroneous expressions.

b) A comparison in terms of computational complexity would be needed, in order to understand the trade-off between the identified advantage per added complexity.

Thanks for your suggestions. We give the parameters and inference time of the key modules of our framework in Table 3 in the camera ready paper (also listed below in Table 2). This will help us to understand the trade-off between the identified advantage per added complexity.

c) The evaluations seem to be performed on the data provided for the Development Phase on the challenge. However, this is not clearly stated in the paper. Thanks for your detailed suggestions. In the first paragraph of section 4.1 in our revised paper, we have added the dataset description and clearly state that our experiments are performed on the validation set of the challenge.

D. To Reviewer #4

a) there are some errors like Figure 6 being referenced as showing the effects of LLaVa, but actually showing the differences when using HAT according to its description.

Thanks for pointing this out and detailed suggestions. In our revised paper, we have modified this error, where Figure 7 and Figure 8 are both showing the effects of SR methods.

b) The visual qualitative performance appears to be good, while in quantitative terms the results are somewhat inconclusive compared to ShadowFormer.

Thanks for your suggestions. In our revised paper, we make this clear that our method is much better than ShadowFormer in quantitative terms, as shown in Table 1. Our diffusion shadow removal model achieves 1 dB higher than ShadowFormer in PSNR metric. However, ShadowFormer wins among all the methods compared in the SSIM metric, which shows stronger consistency with the ground-truth images in luminance, contrast, and structure.

c) Overall the method seems to have some merit, but it seems to suffer a bit from the overall design complexity of multiple separate networks interacting.

Thanks for your suggestions. Our aim of design for each module is to achieve better shadow removal effects. Regarding complexity, we have added detailed comparisons (Table 3 in our camera ready paper, also list them here in Table 2 here) to clearly illustrate parameters quantity and inference time. We hope to address related optimizations in future work.

E. To Reviewer #5

a) Predefined rules are not clear in the paper. How do they impact the capacity of the method to different shadow types?

Thanks for your suggestions. In our revised paper, we have presented a more detailed description for the predefined rules in section 3.2.2, which are mainly used for foreground enhancement.

F. A list of changes

We list the main changes between the first round paper and the final version as follow:

- We rewrote the introduction section with more references supplemented.
- We added the SpA-Former method and corresponding comparisons and listed the contributions of this paper to make it more clear, see Figure 4 and Table 1 in our camera ready paper.
- We added a brief discussion on the complexity of the proposed method including the number of parameters and inference times of the modules of our model, see Table 3 in our camera ready paper.
- We rewrote section 3 to make the whole pipeline more clear.
- We conducted extensive ablation studies on the modules of the proposed pipeline using both quantitative analysis and qualitative analysis. For a better understanding, please refer to Figre 6 and Table 2 in our revised paper.
- We fixed the typos/wrong caption and polished the writing.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[3] Xiaofeng Zhang, Yudi Zhao, Chaochen Gu, Changsheng Lu, and Shanying Zhu. Spa-former:an effective and lightweight transformer for image shadow removal. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. 1