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

Chenghua Li[1,2], Bo Yang[2], Zhiqi Wu[2], Gao Chen[2], Yihan Yu[3], Shengxiao Zhou[2]

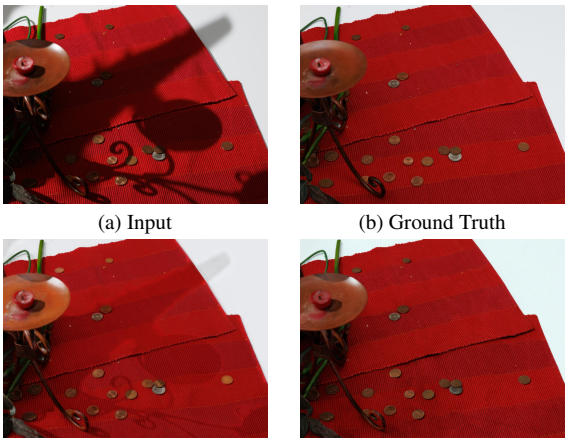
[1]Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]Nanjing Artificial Intelligence Research of IA (AiRiA)
[3]High School Affiliated to Nanjing Normal University Jiangning Campus (NSFJ-CN)

Abstract

Shadow removal is one of essential tasks among image restoration tasks which aims to eliminate the visual semantic information hidden or obscured by the shadow in the image to the largest extent. Variations in lighting and the diverse complexity of shadow depth and color resulting from random background factors are common in the shadow removal task. To address these challenges, this paper proposes a novel interactive shadow removal architecture based on the diffusion model, semantic segmentation and multimodal large language model. Our method utilizes a powerful diffusion model to generate shadow-free images with fewer artifacts and super-resolution models to enhance image details. A universal semantic segmentation model is also involved to reduce percpetual dissonance caused by slicing inference. Furthermore, we integrate the capabilities of multimodal large language models to realize prior rule-based optimization. Leveraging the exceptional generative capability of diffusion model and elaborate cooperation among all the modules, our method achieves outstanding perceptual performance on WSRD dataset. We conduct comprehensive experiments to demonstrate the effectiveness of our approach and share insights gained during the participation in the NTIRE 2024 Image Shadow Removal Challenge.

1. Introduction

In the field of image enhancement, shadow removal is an important and challenging task, which has consistently attracted significant attention due to its impact on the aesthetic quality of images and its potential to interfere with visual tasks such as object recognition, tracking, and image segmentation [4, 8, 15, 22, 42]. The challenges are mainly reflected in two aspects: lighting changes and complex backgrounds. Firstly, the position and intensity of the light source, along with the reflection from the surrounding environment affect the formation and appearance of shadows. As a result, shadows of the same object may show different depths and shapes in different scenes or at different shooting times, which increases the complexity of shadow removal. Secondly, the background color, texture, and lighting conditions influence the formation of shadows, causing the color of the shadow to differ from the actual color of the object, and even exhibit color shifts. In such cases, a single color model or processing method is usually unable to effectively remove shadows. For example, conventional methods based on convolutional neural networks or transformer architecture often lead to varying degrees of artifacts, as shown in Figure1(c).



(a) Input (b) Ground Truth (c) ShadowFormer (d) Ours

Figure 1. Results of ShadowFormer [7] and the our method. The proposed method achieves a much cleaner shadow-free image.

Early shadow removal methods mainly relied on image processing technology and the physical understanding of shadow formation models [6, 35, 40]. Their goal is to restore the original shadow-free image by transferring the local statistics of shadow-free image segments to shadow-

affected areas. With the advancement of deep learning technology and the emergence of large-scale datasets, shadow methods based on deep learning [2, 7, 12] have become the focus of research. However, these methods usually require accurate shadow masks as input, and obtaining shadow region masks is an equally difficult task.

A potential solution is to employ diffusion models to address the above difficulties. Diffusion model has powerful image generation capabilities [20], and has been successfully applied in many fields such as images and videos. Diffusion-model-based methods for shadow removal have demonstrated excellent performance in terms of perceptual quality. Despite the technological advancements in these methods, most existing shadow removal datasets still focus on simplified scenes and primarily involve natural hard shadows formed on flat surfaces under controlled lighting conditions, which limits the model's ability to generalize to more complex situations. In order to solve the problem of generalization, text instructions based image enhancement methods [3, 18] have shown higher flexibility and attractiveness. Methods that combine text instructions with diffusion models [5, 14, 38] further show powerful potential, which generally need to be combined with the Image Controller [41], and then use the control vector to guide image restoration based on the diffusion model.

In this paper, we propose a shadow removal method based on diffusion, segmentation and super-resolution models to deal with the above concerns. Our approach leverages a robust diffusion model [14] to generate shadow-free images with minimal artifacts by employing a super-resolution model [1] to enhance image details and utilizing a versatile semantic segmentation model [9] to seamlessly integrate shadows during slice inference. Additionally, we harness the capabilities of a large multi-modal language model [11] to optimize our shadow removal model based on predefined rules. The overall architecture of our method is shown in Figure 2.

The high-resolution image shadow removal challenge in complex situations [27, 28] represents the pinnacle of the shadow removal task. We have conducted sufficient experiments on the WSRD dataset [25] to verify the effectiveness of our method. Experimental results prove that our method can generate shadow removal results with fewer artifacts and no traces of slice splicing thereby offering better visual perception. In summary, the contributions of this article are as follows:

- We proposed an image shadow removal method based on the diffusion model, and further optimized the results by combining super-resolution, segmentation and multi-modal large language models. With the joint effect of all the components, the best perceptual effects are presented.
- We conducted sufficient experiments on the high-resolution image shadow removal dataset WSRD in complex situations, verified the effectiveness of the proposed method and each module, and inspired further research.

2. Related works

2.1. Image Restoration

Image enhancement or restoration refers to improving the visual quality of images or increasing the information content of images, making them more suitable for specific applications or easier for human observation. Traditional methods typically involve operations such as enhancing image contrast, reducing noise, adding details, and adjusting brightness and color balance. Conversely, deep learning methods utilize high-quality (HQ) and low-quality (LQ) images to optimize end-to-end enhancement models. There are many independent tasks in this field, such as image super-resolution [33], denoising [24], haze removal [23], rain removal [37], deblurring [39], shadow removal [30], and dark light enhancement [34], compression enhancement [36], dynamic range enhancement [32]. In recent years, all-in-one image enhancement is an emerging and booming research field [3, 10, 16]. These methods use a single deep blind restoration model to handle degradation of different types and degrees.

2.2. Shadow removal

Shadow removal methods typically involve two key steps: shadow detection and shadow removal, based on the detected shadow mask [7, 19]. Prior-based methods often rely on simple graphical techniques, initially identifying shadow areas and subsequently illuminating these dark regions. In contrast, deep learning-based methods effectively utilize masks or global information to identify shadow regions and remove shadows accordingly.

For example, DeshadowNet mainly focuses on using the CNN to detect and remove the shadow, which is the very first neural network model to solve to deshadow task in an end-to-end manner, it as well proposed the SRD dataset [19]. ShadowFormer has demonstrated state-of-the-art performance on the public shadow removal dataset (SRD [19], ISTD [31]).

With the proposal of DiT [17] and LDM [21], the diffusion model occupies a prominent position in image generation. This kind of method effectively enhances the ability of image processing models, especially those for image generation. For example, general-purpose image restoration IR-SDE [13] achieves highly competitive performance in quantitative comparisons on image deraining, deblurring, and denoising, setting a new state-of-the-art on two deraining datasets.

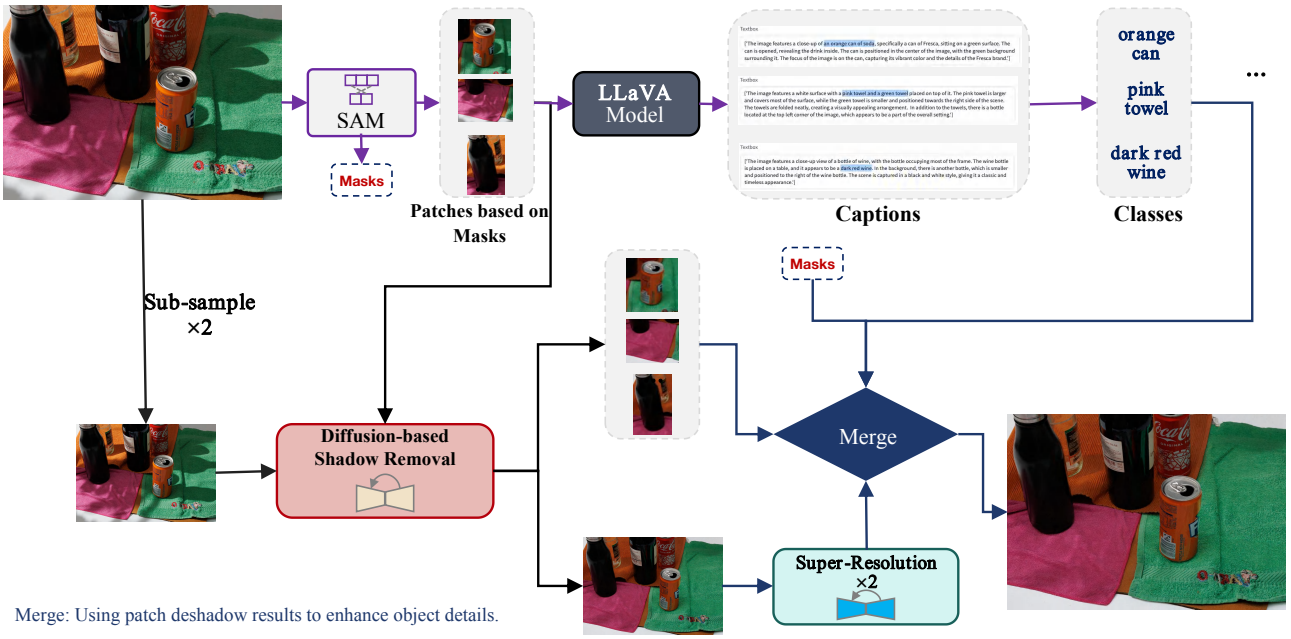


Figure 2. Framework of our method. Generally, we have separate pipelines for our image processing. The primary operation begins by sub-sampling the input image to resize it in a controlled manner, which is then fed into the shadow removal model. For the output, we employ a super-resolution model to restore the image to its original size, enhancing it with additional features and details. Another pipeline is the patch processing. And for the unique foreground object like keyboard, we utilize SAM and LLaVA to filter them out. These patches and objects are then integrated with the output from the previous step.

2.3. NTIRE Challenges

The NTIRE Image Shadow Removal Challenge builds a high resolution image shadow removal dataset, named WSRD [26]. In this dataset, the complexity of the foreground increases and the placement of foreground objects is more complex and independent. Lighting conditions also vary widely in different images. In addition, the interaction between objects and lights is enhanced. Moreover, objects in images often have diverse forms of details, which poses great difficulties for model learning.

The report [27, 28] highlights numerous innovative methodologies in the shadow removal task. In the latest edition, LUMOS introduced a two-stage method employing a ViT-based model for initial deshadowing followed by a NAFNet-based network for refining the output; the Shadow R team developed a ConvNext-based U-Net architecture for initial shadow processing and incorporated a transformer-based enhancement and refinement module to further improve the images; the LVGroup HFUT employed a NAFNet backbone with a creative merging strategy, integrating outputs from different checkpoints by the well-designed loss function.

It's worth noting that our method also participated in this competition and ranked 9th in the perceptual track[1]. While it may not have achieved top rankings among all the entries, we have conducted more study and analysis in this paper and provide valuable insights for this track. For example, we introduce the SAM masks [9] 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.

[1] NTIRE24 Image Shadow Removal Challenge-Track 2 (perceptual)

3. Methods

As shown in Figure 2, our method consists of four modules, namely the diffusion-based shadow removal network [14], SAM-based foreground optimization [9], Super-resolution based detail enhancement [1]), and the multimodal-large-language-model (LLaVA [11]) based specific regions selection. In this section, we will introduce the pipeline in detail.

For any input image I_{in}, the model's processing process is mainly divided into two branches: shadow removal and foreground optimization.

3.1. Shadow removal branch

The input image $I_{in} \in \mathbb{R}^{H\times W\times 3}$ is sub-sampled using bilinear interpolation method with factor 2 with the results denoted as $I_{ds} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$. The main motivation of this operation is that we believe the full image inference is nec-

essary for the shadow removal task. However, the GPU memory is always limited during inference. For example, the resolution of images in WSRD [26] is 1920×1440, which is too high for GPUs weaker than NVIDIA A100.

Then, the down-sampled image I_{ds} is processed by the diffusion-based shadow removal network (section 3.1.1), which needs to be trained on the experimental dataset, then we get the reconstructed shadow free image, denoted as $I_{recon} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 3}$. Next, the resolution of the reconstructed image is increased by the super-resolution model (section 3.1.2) to be consistent with the input image, where the output is recorded as $I_{SR} \in \mathbb{R}^{H \times W \times 3}$.

3.1.1 Diffusion-based Shadow Removal Network

The diffusion-based shadow removal network shown in Figure 2 is built from DA-CLIP [14] and IR-SDE [13]. Inspired by the multitasking image restoration model, we combine DA-CLIP and ICB module [3] to make sure shadow removal model not only have the creativity from diffusion but also be strongly guided in processing different image restoration tasks.

DA-CLIP utilizes a pre-trained vision-language model and a text guided diffusion model to enhances the details of the generated images. The pre-trained CLIP is used to encode low quality images into text descriptions of their corresponding degradation types. However, we remove the CLIP module from DA-CLIP and just use its diffusion module as our diffusion-based shadow removal network.

3.1.2 Super-Resolution Up-scaling

The reverse process of $\times 2$ down-sample operation is super-resolution. The simplest solution is bi-linear interpolation up-sampling. Further, the current state-of-the-art super-resolution model HAT [1] could enhance more details of the low resolution images. The original spatial dimensions of the images are restored by applying a $\times 2$ super-resolution to the results I_{recon} output by the shadow removal network.

3.2. Foreground enhancement branch

This branch is mainly based on the general segmentation network SAM [9]. For special foreground objects like keyboards, we observe significant deficiencies compared to the ground truth. The keyboard has multiple black keys and white letters. However, diffusion-model-based shadow removal methods tend to reconstruct tiny details in the image, which causes a large loss in consistency with the ground truth.

Thus, we introduce the powerful segmentation model SAM to generate the masks of objects. The masks would be used for two optimization operations: patch foreground enhancement and input foreground enhancement. The former is related to the patch splitting during inference (3.2.1) and the latter is based on the built prior rules (3.2.2).

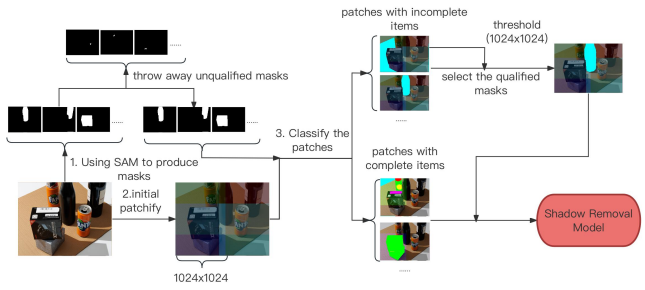


Figure 3. A general pipeline for the proposed patch processing. Step 1, the SAM initially extracts the masks from the input image, and we filter out unqualified masks which are meaningless. Step 2, the input image is cropped into four 1024×1024 intertwined slices. Step 3, the filtered masks are utilized for the classification of the patches. Step 4, the patches are separated into 2 classes, one for the patches with complete foreground items inside and the other for the items exceed the boundaries to some extent. For the second class of patches, we calculate the size of it and check whether its size is within 1024×1024 and enough for a new patch the contain. If it is qualified in patch size, we will create a new patch of the same size for it and include it to the original patches. Finally, these patches will be sent to the shadow removal model.

3.2.1 Patch foreground enhancement

In this branch, the input image $I_{in} \in \mathbb{R}^{H \times W \times 3}$ is first fed into the SAM [9] to generate foreground masks $M = \{m_i \in \mathbb{R}^{H \times W}, i = 0, 1, 2, \cdots\}$. Then, we split the input image into overlapping patches $I_{patches} = \{I^k_{patch} \in \mathbb{R}^{1024 \times 1024 \times 3}, k = 0, 1, 2, \cdots\}$ with size 1024×1024. For images in WSRD[26] dataset, our splitting rule results in four image patches including the upper-left, the upper-right, the lower-left, and the lower-right patches. The size of the patch is determined by the GPU memory size.

Next, we enrich the number of patches by the segmentation masks to ensure that all salient objects in the input image I_{in} are completely contained in at least one patch. This needs the following two steps. One is cleaning the output masks M, retaining only sufficiently significant and appropriately sized blocks to control model inference complexity and GPU memory usage. The other is adding a patch according to the selected masks by judging whether a mask lies in a patch completely or not. If not, we generate a new patch with size 1024×1024 centered on this mask.

Finally, the patches $I_{patches}$ of the input image are fed into the diffusion-based shadow removal network, and output shadow-free image patches $I^{clean}_{patches}$ with size $1024 \times$

1024.

3.2.2 Input foreground enhancement

By analyzing the model output I_{recon}, we found that the shadow removal model cannot handle some specific objects, like keyboards, towels, and black cloths. To solve this problem, we decided to keep the original regions of these objects and merge them with the output I_{recon}.

Thus, we filter out the masks containing specific objects with the powerful visual question-answer model LLaVA [11]. For example, for input images I_{in} or image patches I_{patch}, we set the question *"Is there a keyboard in the picture?"*. According to the answer of LLaVA, we could find the images or patches containing specific objects. The following rules are found to be useful for our model. However, these prior rules may be varied when the shadow removal model is updated.

- Keep the keyboard region unchanged.
- Keep the towels region unchanged.
- Keep the black cloths region unchanged.

The masks obtained by the prior rules is denoted as M_{priors} and will be used for the final merging step.

3.2.3 Merging

Now, we have I_{SR}, $I^{clean}_{patches}$, and M_{priors}. The merging operation, shown in Figure 2, takes them all as inputs, and output an image I_{output}, where $I_{output} = Merge(I_{SR}; I^{clean}_{patches}, M_{priors})$.

The merging operation is straightforward, that is, replacing part of the reconstructed image with the object area of patches or input according to the corresponding mask.

4. Experiments

4.1. Settings

The experiments are mainly conducted on the validation set of the NTIRE 2024 Image Shadow Removal Challenge [27, 29]. The dataset is updated for improved pixel alignment and some new contents based on the WSRD dataset [26].

To validate the effectiveness of our method, we compare its performance with the current state-of-the-art methods, including SpA-Former [43], ShadowFormer [7]. We train ShadowFormer on the challenge dataset using its official training code[2]. Because of the necessity of the masks of shadow for the method of ShadowFormer, we leverage HSV thresholding and morphological operations to provide masks under elaborated parameters for WSRD dataset, which is not equipped with masks originally. For SpA-Former, we use the random cropping strategy for training.

[2] https://github.com/GuoLanqing/ShadowFormer

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.

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**

Experiments are conducted on a high-performance workstation featuring an NVIDIA A100 GPU, enabling swift training and inference. Our shadow removal model is implemented using the PyTorch deep learning framework, capitalizing on its versatility and computational efficiency. During training, we utilized a batch size of 8 and employed the AdamW optimizer with a learning rate of 2×10^{-4}. Additionally, we incorporated cosine annealing learning rate decay to enhance training stability and convergence. The input patch size for training was set to 570×570, with random horizontal and vertical flips applied as augmentations to diversify the training datasets and enhance robustness.

We use four metrics to evaluate various aspects of the quality of the generated images, such as PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), LPIPS (Learned Perceptual Image Patch Similarity), and FID (Fréchet Inception Distance), which is the same as to DA-CLIP [14].

4.2. Results

The results are presented in Table 1. Our method outperforms the compared methods in both PSNR and FID metrics. As PSNR provides a quantitative measure of reconstruction quality, our method has higher fidelity to the ground-truth images. Lower FID values indicate that the generated images by our method are closer to the ground-truth images in terms of both visual appearance and statistical properties.

However, ShadowFormer achieves the highest SSIM, which indicates that ShadowFormer holds the best structural similarity between the ground-truth images by comparing their luminance, contrast, and structure. While LPIPS captures more nuanced differences between images, SpA-Former shows much more similarity of deep features between the output images and the ground-truth images. Our method fails these two metrics, suggesting that our method cannot balance the luminance, contrast, structure, deep feature embedding factors between the shadow removal and the image construction. This could be considered a limitation of diffusion-based image generation methods, and be addressed by further controlling strategies.

On the other hand, these metrics alone cannot compre-

Figure 4. Comparison with ShadowFormer, SpA-Former and GT. Our method can remove the shadow accurately without significantly affecting the information in the shadow regions and we retain more details and textures based on patch processing and utilization of segmentation.

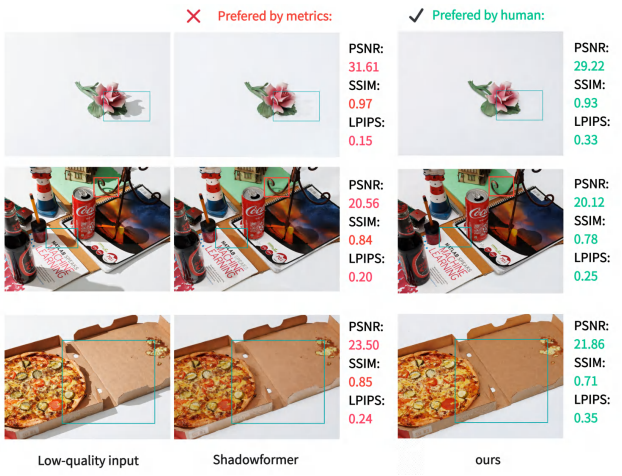


Figure 5. These illustrations highlight the discrepancy between metric and perceptual. Our method produces better pictures despite maintaining lower metrics.

hensively evaluate a model's performance. Therefore, we conducted a qualitative comparison of the results on the validation set, as shown in Figure 4. From the visual inspection, it can be observed that our method maintains better shadow removal performance than the compared methods. In contrast, the compared methods still leave traces of shadows in most result images.

It should be emphasized that metrics such as PSNR, SSIM, and LPIPS may not fully reflect the effect of shadow removal. Figure 5 shows several examples of shadow removal effects. Our method lags behind in these metrics but has better shadow removal effects, which also demonstrates the importance of qualitative evaluation in performance evaluation of shadow removal methods. Through this visual comparison, we gain valuable insights into the true effectiveness of our approach, which may not be adequately captured by numerical metrics alone.

Overall, despite that the quantitative metrics show some discrepancies, the qualitative assessment indicates the superiority of our method in terms of visual perception. This underscores the importance of considering both quantitative metrics and qualitative evaluations when assessing model performance.

4.3. Ablation study

The proposed method for shadow removal consists four modules, as shown in Table 2, namely the diffusion-based shadow removal network (*Diffusion* [14]), SAM-based for inference based on original image patches (*SAM* [9]), Super-resolution based detail enhancement (*bilinear-SR* or *HAT-SR* [1]), and multimodal large language models based specific regions selection (*LLaVA* [11]). This section brings the ablation studies of these modules.

Figure 6. Visual effects of ablation study. (a) *Diffusion+path-splitting* demonstrates clear stitching traces. (b) and (c) shows much better shadow removal results than (a).

Firstly, we utilize patch splitting inference due to the limited GPU memory, named *Diffusion + patch-splitting* in the first line of the results in Table 2. This method achieves the best results for all metrics. However, most of the result images show traces of cleaned slices and splicing, which greatly affects the visual effect, as shown in Figure 6.

We adopt down-sampling strategy to address this problem, which down-scales the input image by $\times 2$, shadow remove, and up-scale the results by the HAT-SR model. Starting from this baseline model *Diffusion+HAT-SR*, we gradually add *SAM* and then *LLaVA* (see section 3). The PSNR metric is improved by 0.439 dB and 0.0034 dB respectively, indicating that the Diffusion-based shadow removal models prefers to generate higher fidelity results using the patch-based inference method.

Next, we focus on explaining the slight improvement based on *LLaVA*. We carefully compared the shadow removal results of *the initial model* and found that the initial model did not perform well in removing shadows on black objects, such as keyboards, black clothes, etc. We used the multimodal large language model LLaVA to identify these images and segmented the black object areas based on SAM, directly pasting them onto the result image without any processing. We select two images with black keyboards[3] from the validation set of the NTIRE2024 shadow removal challenge. The above operation improves the results by 0.1178 dB and 0.2226 dB respectively, showing nice improvement.

Finally, we replaced the HAT model with a simpler bilinear up-sampling method, resulting in an average performance improvement of 0.0492 dB. Although this improvement is relatively small, through comparing with model's output results, we discovered valuable insights, where HAT

[3] 0056.png, 0057.png

Table 2. Ablation study of proposed pipeline. The comparison indicates that SAM is an useful technique, while bilinear upsampling is much better than HAT-SR model.

Diffusion	patch-splitting	HAT-SR	SAM	LLaVA	bilinear-SR	PSNR	SSIM	LPIPS	FID
✓	✓					**24.8280**	**0.7820**	**0.2123**	**45.8000**
✓		✓				23.4157	0.7211	0.3575	60.7929
✓		✓	✓			23.8547	0.7368	0.3089	52.3047
✓		✓	✓	✓		23.8581	0.7370	0.3085	52.4310
✓			✓		✓	23.9039	0.7770	0.3105	49.5052
✓			✓	✓	✓	23.9074	0.7772	0.3101	49.5479

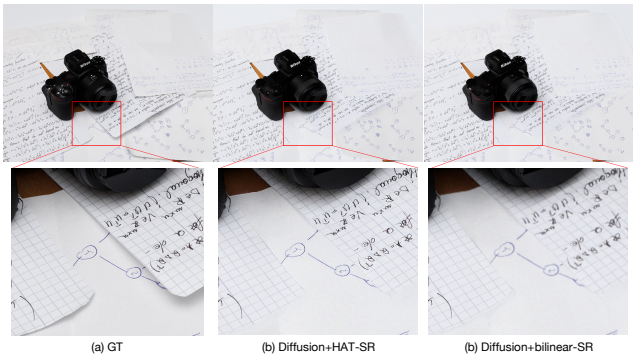


Figure 7. These illustrations highlight the difference between HAT-SR and bilinear-SR. Comparing (b) and (a), we can infer that the HAT-SR model outperforms the interpolation-based bilinear-SR method (c) in super-resolving text image content.

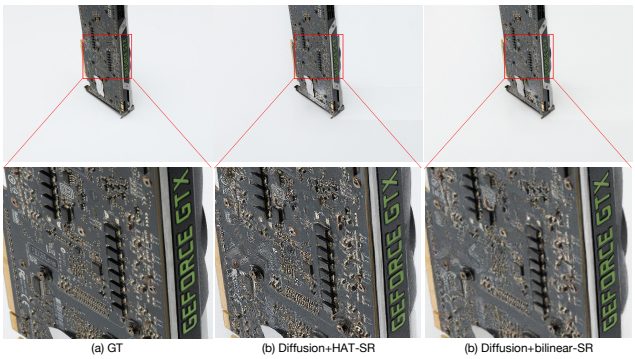


Figure 8. These illustrations highlight the difference between HAT-SR and bilinear-SR. HAT-SR model amplifies errors in the reconstructed images from the diffusion model, leading to greater deviations from the ground truth (GT).

model performs well on text image content, but after replacing it with bilinear, the performance of this image decreases (see Figure 7). However, for content with extremely complex details, since there may be errors after shadow removal by the Diffusion model, HAT model amplifies these errors, leading to a decrease in overall performance (see Figure 8).

In summary, we found that SAM contributes to improving the performance of the Diffusion-based shadow removal network, while bilinear upsampling operation relative to the massive HAT super-resolution model helps improve average performance. The HAT super-resolution model amplifies the errors reconstructed by the Diffusion model, resulting in performance degradation. Moreover, combining multimodal large language models to carefully process the results of the above models according to specific rules will further improve the effects of specific images. The deeper mechanism will be further studied in future work.

In addition, this article focuses on exploring how to improve the visual effect of shadow removal. The parameters of several modules introduced are relatively large and time-consuming, as shown in Table 3. The efficiency optimization is also one of our future research works.

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

Table 3. Number of Parameters and inference times of each module in our framework

5. Conclusion

In this study, we proposed a novel shadow removal framework based on diffusion, super-resolution, segmentation, and LLaVA. Through extensive experiments, we have demonstrated the effectiveness of our approach in addressing the challenging task of shadow removal in images. Our method contributes to this research domain by presenting an effective solution to the challenging task of shadow removal. We believe that our proposed method has the potential to facilitate advancements in various applications.

Acknowledgements. This work was supported by the Key Projects of Industrial Foresight and Key Core Technology in Jiangsu Province Science and Technology Plan (Grant No. SBE2023010012).

References

[1] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023. 2, 3, 4, 7

[2] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4743–4752, 2021. 2

[3] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*, 2024. 2, 4

[4] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence*, 25(10):1337–1342, 2003. 1

[5] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[6] Maciej Gryka, Michael Terry, and Gabriel J Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics (TOG)*, 34(5):1–15, 2015. 1

[7] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: global context helps shadow removal. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2023. 1, 2, 5

[8] Cláudio Rosito Jung. Efficient background subtraction and shadow removal for monochromatic video sequences. *IEEE Transactions on Multimedia*, 11(3):571–577, 2009. 1

[9] 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. 2, 3, 4, 7

[10] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17452–17462, 2022. 2

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3, 5, 7

[12] 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. 2

[13] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjolund, and Thomas B. Schon. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23045–23066. PMLR, 2023. 2, 4

[14] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjolund, and Thomas B. Schon. Controlling vision-language models for universal image restoration. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4, 5, 7

[15] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 26(8):1079–1087, 2004. 1

[16] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5815–5824. IEEE, 2023. 2

[17] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2

[18] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. PromptIR: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[19] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4067–4075, 2017. 2

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[22] Andres Sanin, Conrad Sanderson, and Brian C Lovell. Improved shadow removal for robust person tracking in surveillance scenarios. In *2010 20th International Conference on Pattern Recognition*, pages 141–144. IEEE, 2010. 1

[23] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 2

[24] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020. 2

[25] 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 (CVPR) Workshops*, pages 1826–1835, 2023. 2

[26] 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*, pages 1825–1834, 2023. 3, 4, 5

[27] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, Shuhao Cui, Junshi Huang, Shuman Tian, Mingyuan Fan, Jiaqi Zhang, Li Zhu, Xiaoming Wei, Xiaolin Wei, Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön, Xiaoyi Dong, Xi Sheryl Zhang, Chenghua Li, Cong Leng, Woon-Ha Yeo, Wang-Taek Oh, Yeo-Reum Lee, Han-Cheol Ryu, Jinting Luo, Chengzhi Jiang, Mingyan Han, Qi Wu, Wenjie Lin, Lei Yu, Xinpeng Li, Ting Jiang, Haoqiang Fan, Shuaicheng Liu, Shuning Xu, Binbin Song, Xiangyu Chen, Shile Zhang, Jiantao Zhou, Zhao Zhang, Suiyi Zhao, Huan Zheng, Yangcheng Gao, Yanyan Wei, Bo Wang, Jiahuan Ren, Yan Luo, Yuki Kondo, Riku Miyata, Fuma Yasue, Taito Naruki, Norimichi Ukita, Hua-En Chang, Hao-Hsiang Yang, Yi-Chung Chen, Yuan-Chun Chiang, Zhi-Kai Huang, Wei-Ting Chen, I-Hsiang Chen, Chia-Hsuan Hsieh, Sy-Yen Kuo, Li Xianwei, Huiyuan Fu, Chunlin Liu, Huadong Ma, Binglan Fu, Huiming He, Mengjia Wang, Wenxuan She, Yu Liu, Sabari Nathan, Priya Kansal, Zhongjian Zhang, Huabin Yang, Yan Wang, Yanru Zhang, Shruti S. Phutke, Ashutosh Kulkarni, MD Raqib Khan, Subrahmanyam Murala, Santosh Kumar Vipparthi, Heng Ye, Zixi Liu, Xingyi Yang, Songhua Liu, Yinwei Wu, Yongcheng Jing, Qianhao Yu, Naishan Zheng, Jie Huang, Yuhang Long, Mingde Yao, Feng Zhao, Bowen Zhao, Nan Ye, Ning Shen, Yanpeng Cao, Tong Xiong, Weiran Xia, Dingwen Li, and Shuchen Xia. Ntire 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1788–1807, 2023. 2, 3, 5

[28] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2, 3

[29] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 5

[30] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *European Conference on Computer Vision*, pages 361–378. Springer, 2022. 2

[31] 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. 2

[32] Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8874–8895, 2021. 2

[33] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020. 2

[34] Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, and Heng Tao Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1662–1671, 2023. 2

[35] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma. Fast shadow removal using adaptive multi-scale illumination transfer. In *Computer Graphics Forum*, pages 207–218. Wiley Online Library, 2013. 1

[36] Jianquan Yang, Guopu Zhu, Yao Luo, Sam Kwong, Xinpeng Zhang, and Yicong Zhou. Forensic analysis of jpeg-domain enhanced images via coefficient likelihood modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1006–1019, 2021. 2

[37] Wenhan Yang, Robby T Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. Single image deraining: From model-based to data-driven and beyond. *IEEE Transactions on pattern analysis and machine intelligence*, 43(11):4059–4077, 2020. 2

[38] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[39] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022. 2

[40] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015. 1

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2

[42] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):611–624, 2018. 1

[43] 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. 5