Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Shadclips:When Parameter-Efficient Fine-Tuning with Multimodal Meets Shadow Removal

Xiaofeng Zhang
Shanghai Jiao Tong University
Zishan Xu
China University of Mining and Technology
Hao Tang
Carnegie Mellon University
Chaochen Gu
jacygu@sjtu.edu.cn
Shanghai Jiao Tong University
Shanying Zhu
Shanghai Jiao Tong University
Xinping Guan
Shanghai Jiao Tong University

Research Article

Keywords: CLIP, Shadow Removal, Segment Anything Model, SONet 1

Posted Date: April 4th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4194150/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Shadclips:When Parameter-Efficient Fine-Tuning with Multimodal Meets Shadow Removal

Xiaofeng Zhang[1], Zishan Xu[2], Hao Tang[3], Chaochen Gu[1*], Shanying Zhu[1], Xinping Guan[1]

[1]Shanghai Jiao Tong University, Shanghai, China.
[2]China University of Mining and Technology, Jiangsu, China.
[3]Carnegie Mellon University, Pittsburgh, USA.

*Corresponding author(s). E-mail(s): jacygu@sjtu.edu.cn;
Contributing authors: framebreak@stju.edu.cn;
TS21170029A31@cumt.edu.cn; bjdxtanghao@gmail.com;
shyzhu@sjtu.edu.cn; xpguan@sjtu.edu.cn;

Abstract

Segment Anything Model (SAM), an advanced universal image segmentation model trained on an expansive visual dataset, has set a new benchmark in image segmentation and computer vision. However, it faced challenges when it came to distinguishing between shadows and their backgrounds. To address this, we proposed ShadClips, which consists of SAM-optimizer and SONet. It has dramatically enhanced SAM's ability to segment shadow images, differentiating between the background and both soft and hard shadows adeptly. Due to its dependence on pixel point inputs, the SAM-Optimizer interface could do better. This method presents challenges, especially when dealing with long, extended shadows. To make the user experience more intuitive and effective, we incorporated the capabilities of CLIPs. Therefore, simple text descriptions like "A photo of a shadow" can be used to guide the SAM-Optimizer, allowing it to select the most relevant shadow mask from SAM's comprehensive category list. Meanwhile, we introduce SONet to shadow removal. A large number of experiments on ISTD/SRD prove that the proposed method is effective and satisfactory. The source code of the ShadClips can be accessed from https://github.com/zhangbaijin/SAM-helps-Shadow.

Keywords: CLIP, Shadow Removal, Segment Anything Model, SONet

1 Introduction

The existence of shadows in natural images can provide information about the environment and illumination conditions, which not only appeals for interpretation of the situation in the image but also makes image processing more complex. Shadows change the color and texture of objects and often vary in shape and size. In addition, hard and soft shadows behave differently, creating additional difficulties in segmentation. Therefore, shadow segmentation is an essential problem in image processing.

There are many works on shadow removal and image restoration, such as semi-supervised models with guidance [1, 4, 5, 6, 7, 8, 9, 10, 11, 2, 3], GAN's methods [12, 13, 14, 15, 16, 17, 18, 19], and some unsupervised methods [20, 21, 22, 23, 24, 25, 26].

In recent developments of foundational models, the Segment Anything Model (SAM) emerges as a prominent universal image segmentation model trained on extensive datasets. Demonstrated to excel in segmenting diverse scenes, SAM signifies a significant advancement in image segmentation and the broader field of computer vision. Consequently, we integrated SAM into the task of image shadow segmentation, addressing both hard and soft shadow segmentation challenges. However, our experiments revealed that SAM's performance in shadow image segmentation could be better. Notably, it frequently needs to be clarified for the background with shadow regions, leading to subpar segmentation results. Hence, we designed a SAM-Optimizer to enhance the segmentation of shadow data. SAM-Optimizer comprises convolutional layers, batch normalization, and MLP layers, capable of specific binary segmentation of shadow categories. It can precisely segment the background and shadow areas when dealing with soft or hard shadows. Additionally, we found that even under good segmentation circumstances, the interactivity of the SAM-Optimizer was poor because SAM only accepts pixel point inputs. For example, inputs like "200, 200" must directly correspond to the position of the shadow area to segment the target area. However, their location cannot be pointed out accurately for some elongated shadows.

To address the challenges above, CLIPs were introduced into the shadow relevance index, ShadCLips, enabling text input. By giving the SAM-Optimizer a command like "A photo of shadow", it can locate the mask most related to "shadow" from SAM's whole category. This integration helps overcome limitations and increases the model's accessibility and usability. Last but not least, in terms of shadow removal, we designed the second-order shadow removal network SONet to achieve better results for shadow removal considering the conditions of light estimation.

Moreover, the difficulty with shadow removal is that the nature of shadows is inherently complex. Not only do shadows change the brightness of an image, but they also change image properties such as color and texture. The shape and intensity of shadows vary depending on the complexity of the light source and the scene. Our approach can effectively solve these challenges through deep unfolding network techniques. Additionally, the second-order deep unfolding network named SONet can better handle the complex task of shadow removal by deconstructing layers, bringing higher accuracy and efficiency.

By integrating the synergistic methodologies of the SAM-optimizer and SONet, we have refined SAM's capabilities in shadow segmentation, underscoring the nuances of shadow attributes inherent to images. The refined approach has not only bolstered its

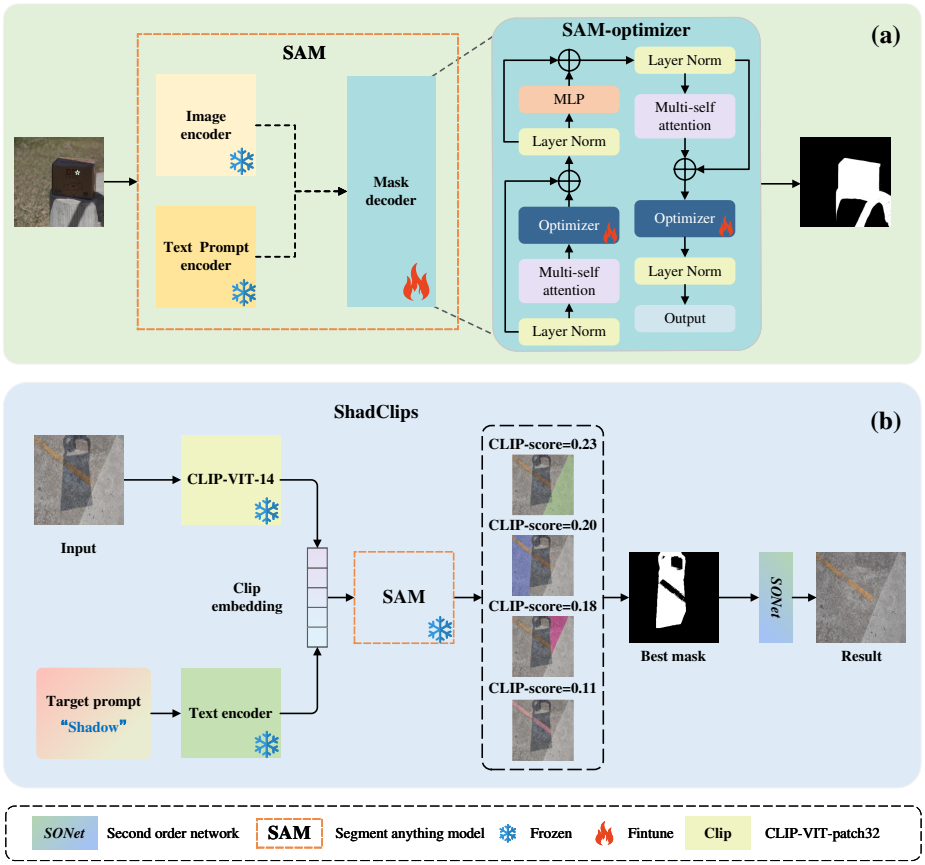

Fig. 1: The structure of ShadClips, which is divided into SAM-optimizer, ShadClips, and SONet, (a) represents finetune, (b) represents inference.

interactivity but also its applicability. Through rigorous experimental validation, we can affirmatively conclude the efficacy of our approach across the dataset, notably on ISTD and SRD benchmarks. In summary, our paper has three main contributions:

1. This paper proposed ShadClips to enhance the interactive selection. The correlation function is introduced to calculate the relationship between masks and shadows, which is helpful for accurate segmentation.
2. This paper proposed a SAM-optimizer that can segment meticulously for the shadow removal task and parameter-efficient fine-tuning.
3. The experimental results on ISTD and SRD datasets proved that the proposed method achieves better performance and demonstrated the effectiveness of the network.

2 Related work

2.1 Segment Anything Model and Variants

Segment Anything Model (SAM) [27] is a basic model that can handle various types of image segmentation tasks without retraining and has been evaluated to exhibit impressive zero-sample performance. Following the introduction of SAM, FastSAM [28], mobileSAM[29], and Semantic-SAM [30] have also been proposed; FastSAM speed has reached the real-time range; mobile-SAM is a lightweight SAM, focusing on the image encoder this part, that is, the image encoder and segmentation mask decoder decoupling optimization. Grounded-SAM [31] combines Grounding DINO and SAM while connecting ChatGPT and Stable Diffusion to achieve actual human-computer interaction.
Grounded-SAM [31] MedSAM [32] and [33] investigate the application of SAM to medical images. [32] curates a large-scale medical image dataset and develops a simple fine-tuning method to adapt SAM to general medical image segmentation. [33] uses a low-rank-based (LoRA) fine-tuning strategy to fine-tune the SAM image encoder with a cue encoder and a mask decoder for fine-tuning on a labeled medical image segmentation dataset. [34] evaluates the potential of SAM for medical image segmentation tasks, and the results show that the model has a powerful zero-sample learning capability and generalizes well to abdominal CT data. [35] develops an efficient pipeline for generating a large-scale Remote Sensing (RS) image segmentation dataset called SAMRS, using the SAM and existing RS object detection datasets. [36] proposes SAM-Adapter to significantly improve the performance of SAM by incorporating domain-specific information or visual cues into the segmentation network. IA (Inpainting Anything) [37] is a powerful and user-friendly pipeline that solves issues related to repair, such as object removal, content filling, and background replacement. It combines the advantages of segmenting any model (SAM), state-of-the-art image painters, and AI-generated content (AIGC) models.

2.2 CLIP Model

The full name of CLIP [38] is Contrastive Language-Image Pre-training, i.e., a pre-training method or model based on contrastive text-image pairs.CLIP is a multimodal model based on contrastive learning. Unlike some contrastive learning methods in CV, such as moco and similar, the training data of CLIP is text-image pairs: an image and its corresponding text description; here, it is hoped that the model can learn the matching relationship of text-image pairs through contrastive learning. Data are text-image pairs: an image and its corresponding text description; here, the model is hoped to learn the matching relationship of text-image pairs through contrast learning. CLIP consists of two models: Text Encoder and Image Encoder.

2.3 Deep Unfolding CNN Image Shadow Removal

Shadow removal is one of the essential tasks in the field of computer vision. Many works have adopted deep CNN to remove shadows and made significant breakthroughs. Some studies suggest that public large-scale datasets (such as ISTD) train

networks in an end-to-end manner. At the same time, research using deep unfolding methods has also achieved many results.
Deep unfolding methods typically combine model-driven methods (such as clear interpretability) with the advantages of CNN-based methods (such as strong learning ability). They typically include the following steps: (1) unfolding specific iterative optimization algorithms; (2) Parametric expansion model; (3) Parameters can be learned through CNN updates. Currently, many other methods for computer vision tasks, such as Cu[17], synthesize realistic shad images through adversarial learning to improve network performance. Fu [39] Transformed the shadow removal problem into a multi-exposure image fusion problem.

3 Methodology

ShadClips is shown in Figure 1 and Algorithm 1, it is inspired by [36, 37]. It is divided into two parts: SAM-optimizer with CLIP model and the second part is the second-order variational network (SO-Net).
(1) SAM contains a wealth of knowledge that can be transferred to any segmentation task.
(2) The fine-grained segmentation results provided by SAM can be used as the first stage of pre-processing for the shadow removal task, with the expectation that subsequent modules can be optimized to obtain better results.
(3) The prior knowledge related to the shadow removal task can be introduced through a careful design of the prompt to improve the final result and the interpretability of the method.

The SO-Net model integrates the design constraint term of shadow mask to achieve the exact entity mapping between pixels in the unshaded region and adapts the fine-grained mapping between pixels in the shadowed area and unshaded pixels within one stage. Second, based on the new illumination model, we reformulate the de-shadowing task as a variational optimization model consisting of favorable data fidelity term with shading related to the previous term.

3.1 SAM-Optimizer and Parameter-Efficient Fine-Tuning

We introduce an approach to image shadow segmentation using SAM-Optimizer. SAM-optimizer is designed for Parameter-Efficient Fine-Tuning. We chose the SAM-optimizer as the core part of our shadow segmentation, mainly based on its excellent performance in complex and challenging shadow detection tasks. When dealing with shadow detection problems, conventional segmentation models such as SAM often perform poorly in solving the fine segmentation of shadow regions and accurately determining shadow boundaries. SAM-optimizer significantly improves the accuracy of shadow detection by introducing task-specific visual cues and domain knowledge. We deployed two Adapters for each ViT block in the SAM encoder. For the standard ViT block, we have placed the first Adapter after the multicast attention and before the residual connection (as shown in Figure 2, and we have set the second Adapter after the multicast attention in the residual path in the MLP layer. In the SAM decoder,

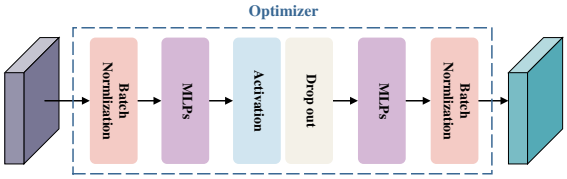

Fig. 2: The overview of SAM-optimizer.

we deploy 1 Adapter for each ViT block. The first Adapter is deployed, with residual cues embedded in the cue to the image embedded in the image after multi-head cross attention.

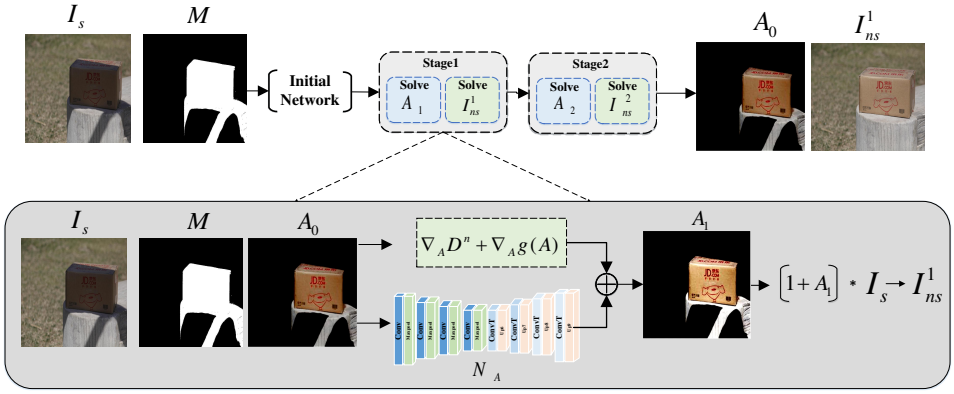

Fig. 3: The structure of SONet.

In our shadow segmentation phase, we feed the input image into the SAM-optimizer for segmentation. The part of the SAM-optimizer uses our visual cues and domain knowledge designed explicitly for shadow detection to capture better and localize shadows. After SAM-optimizer processing, the shadow part of the image will be segmented accurately. After segmentation, we feed the shadowed portion of the image to our designed shadow removal algorithm for processing.

Prompt Training: In the training part, we only train the mask-decoder and Prompt encoder part; in addition, we embed the CLIP model into the Prompt encoder to replace the point encoder of the original SAM model compared to the original SAM; we are more accurate for the detection of shadows because the shadows are irregular and cannot accurately locate the shadow pixels in the image, so we choose to embed CLIP as a vector to guide the SAM-optimizer for accurate segmentation.

In the testing stage, we establish the relationship between CLIP and image with the following expression,

$$\begin{aligned} \mathrm{I}' &= \mathrm{M} \odot \mathrm{I}, \\ \text{probs} &= g \circ f \circ p\left(T, I'\right), \end{aligned} \tag{1}$$

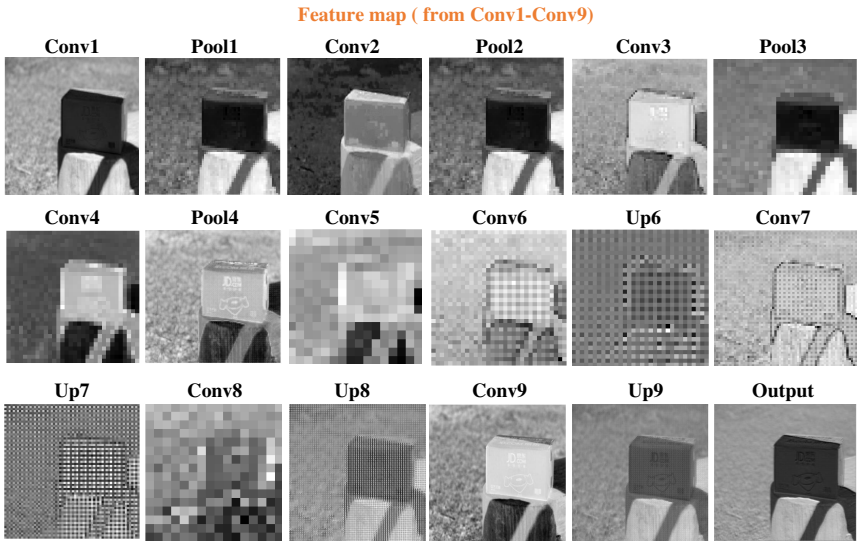

Fig. 4: The visualization of SONet's feature map.

Algorithm 1 The processing via the ShadClips.

Require: Training set
Ensure: Shadow removal result.

1: Fine-tuning the SAM-optimizer framework with the ISTD/SRD training set and predicting the results of shadow segmentation tasks.
2: **for** I_s (shadow class) in I_{class} (SAM class) **do**
3: Update the Binary loss via $Loss_{BCE}$.
4: Adjust the optimizer parameters.
5: **end for**
6: Train the SO-Net.
7: Add the ShadClips and SO-Net to one stage.
8: I_s input SAM-optimizer, the full category I_{class} is output.
9: Text prompt "shadow" is input into the CLIP model, and text matching is done for each category in I_{class}.
10: The most relevant mask M_{best} is reserved and then sent into SONet.

where $\circ$ represents a combination of functions.

$$\text{decode}\left(\underset{i}{\text{argmax}}\left(\text{probs}_i\right)\right) = T, \tag{2}$$

let M be the mask, I be the original image, T be the target text, f be the CLIP model, p be the processor, g be the softmax function, $argmax$ be the operation of indexing to get the maximum value, and decode be the operation of decoding from the category index to the text.

3.2 Second-order Variational Image Shadow Removal Network (SO-Net)

Second-order light enhancement is proposed by [40]:

$$\mathrm{LE}_t(\mathrm{x}) = \mathrm{LE}_{t-1}(\mathrm{x}) + \mathbf{A}_t(\mathrm{x})\mathrm{LE}_{t-1}(\mathrm{x})\left(1 - \mathrm{LE}_{t-1}(\mathrm{x})\right), \tag{3}$$

where x is the pixel coordinate, $\mathrm{LE}(I(\mathrm{x})$; To enhance the output; $\mathbf{A} \in [-1, 1]$ is a trainable curve parameter (modifies the size of the curve and controls exposure): each pixel is normalized.

3.2.1 Shadow Removal Lighting Model

SONet is shown in Figure 3, we propose a shadow removal illumination model to effectively remove shadows from images while maintaining their spatial detail features. In this model, we first use a segmentation model based on the Segment Anything Model (SAM) to accurately locate shadow areas. Then, we expand it using spatial variation characteristics and existing shadow masking information to improve the accuracy of the target mask.

Generally, we formulate this problem as a maximum aposteriori estimation problem, which is written as

$$\begin{aligned} \mathbf{I}_{ns} &= \arg\max_{\mathbf{I}_{ns}} \log P\left(\mathbf{I}_{ns} \mid \mathbf{I}_s\right) \\ &= \arg\max_{\mathbf{I}_{ns}} \log P\left(\mathbf{I}_s \mid \mathbf{I}_{ns}\right) + \log P\left(\mathbf{I}_{ns}\right), \end{aligned} \tag{4}$$

where $\log P\left(\mathbf{I}_s \mid \mathbf{I}_{ns}\right)$ represents the data likelihood and $\log P\left(I_{ns}\right)$ represents characterizes the intrinsic prior knowledge of I_{ns}. We have designed a basic convolutional module called Dynamic Mapping Residual Block for each network scale. And use two CNN networks to calculate the initial and gradient results, respectively. Formalize this problem as a maximum posterior estimation problem. We transform the shadow removal problem into solving the minimum variational model and use it to further transform it into an optimization problem without constrained optimization:

$$\begin{cases} \arg\min_{\mathbf{I}_{ns}} D\left(\mathbf{I}_s, \mathbf{I}_{ns}, \mathbf{A}\right) + \alpha\phi\left(\mathbf{I}_{ns}, \mathbf{A}\right), \\ \text{s.t. } g(\mathbf{A}) = 0, \end{cases} \tag{5}$$

where α represents the weight parameter, $\phi()$ represents the function and $\arg\min D()$ represents the favorable data fidelity term:

$$D\left(\mathbf{I}_s, \mathbf{I}_{ns}, \mathbf{A}\right) = \frac{1}{2}\left\|\mathbf{I}_s - \frac{\mathbf{I}_{ns}}{\mathbf{1} + \mathbf{A}}\right\|_2. \tag{6}$$

We further convert it into a non-constrained optimization problem as:

$$\arg\min_{\mathbf{I}_{ns}} D\left(\mathbf{I}_s, \mathbf{I}_{ns}, \mathbf{A}\right) + \beta g(\mathbf{A}) + \alpha\phi(\boldsymbol{A}), \tag{7}$$

where β represents the weight parameter. We need to alternately update A and I_{ns} as follows:

$$\begin{cases} A^2 = A^1 - \eta_A \left(\nabla_A D^1 + \beta \nabla_A g\left(A^1\right) + \alpha \nabla_A \phi\left(A^1\right)\right) \\ I_{ns}^2 = \left(1 + A^2\right) \star I_s, \end{cases} \tag{8}$$

among them, ∇ represents the gradient operator; n represents the current number of iterations, with a maximum of K.

Equation (8) has two problems that must be solved during the above iteration process. The first problem is to calculate the gradient of the objective function. The corresponding gradients of quadratic terms $D(\cdot)$ and $g(\cdot)$ are easy to calculate. Then, the corresponding $\phi()$ can be achieved through deep CNN A. $\nabla_A \phi\left(A^1\right)$ can be directly learned from the training dataset. Another issue is how to obtain the initial values of the iterative algorithm. Due to the ability of CNN to produce reasonable results, we use CNN technology to estimate the initial values A_0 and I_{ns}.

3.2.2 Variational Model of Iterative Algorithm CNN Framework

We have designed a deep convolutional neural network based on iterative algorithms, combined with the U-net architecture and utilizing advanced deep learning technologies such as ViT and AIGC Models. At each iteration stage, we first update A and then update the unshadowed image I_{ns}. Specifically, this process can be expressed as the following equations:

$$\begin{aligned} \mathbf{A}^0, \mathbf{I}_{ns}^0 &= N_{\text{init}}\ \left(\mathbf{I}_s, \mathbf{M}\right), \\ N_A^{n+1} &= \nabla_{A\varphi}\left(\mathbf{A}^n\right), \end{aligned} \tag{9}$$

where I_s represents a shadow image, and the input image contains shadows. M represents a mask, identifying a binary image of shaded areas in the input image. A means ambient light estimation, which is the light intensity in the input image's shaded areas. I_{ns} represents a shadow-removed image and outputs the resulting image after removing shadows.

N_{an+1} and $Ninit$: represent different iteration stages, $Ninit$ is the initialization stage, N_{an+1} represents the value updated in the nth stage. ϕ is a loss function that optimizes ambient light estimation A. Gradient ϕ represents the loss function of optimized ambient light estimation ϕ. The gradient is used to update the value of A.

So, the entire iterative process mainly optimizes the shadow removal effect in the input image by updating the ambient light estimation A and the shadow removal image I_{ns}.

3.2.3 Feature map visual results of SONet

In Figure 4, we present a comprehensive feature visualization analysis of the SONet architecture to elucidate the internal feature transformations and transmission

Method	ISTD			SRD		
	RMSE-all ↓	SSIM-all ↑	PSNR-all ↑	RMSE-all ↓	SSIM-all ↑	PSNR-all ↑
Yang [41]	15.63	14.83	19.82	-	-	-
Guo [42]	9.3	0.919	23.07	-	-	-
Gong [43]	8.53	0.908	24.07	-	-	-
DeShadowNet [12]	7.83	0.943	29.54	6.15	0.864	27.53
STC-GAN [13]	7.47	0.929	27.43	5.33	0.943	30.12
SID [14]	7.96	0.948	25.01	5.65	0.932	30.13
DSC [15]	6.67	0.845	26.62	17.43	0.648	20.65
RIS-GAN [16]	6.62	0.859	26.42	5.54	0.922	29.33
Mask-ShadowGAN [8]	7.63	0.837	27.53	7.32	0.853	27.35
DHAN [17]	6.28	0.921	27.88	4.80	0.928	28.93
LG-shadow [9]	6.67	0.906	25.83	6.93	0.912	25.90
Auto-Exp [44]	5.88	0.845	27.19	6.75	0.856	26.72
G2R [20]	7.84	0.932	24.72	7.56	0.925	21.03
Bijetive [26]	**4.78**	0.928	29.02	4.15	0.944	30.35
SpA-Former [45]	6.86	0.931	27.73	6.86	0.931	27.73
ShadowFormer [46]	4.79	0.934	**30.16**	**4.09**	0.951	**32.38**
ShadClips (Ours)	5.09	**0.959**	29.85	4.79	**0.952**	31.72

Table 1: Performance comparison of results on ISTD and SRD dataset.

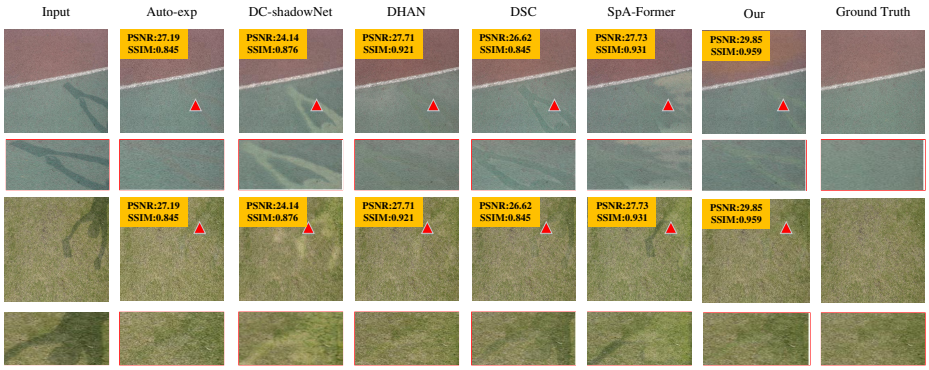

Fig. 5: The compare results of ShadClips with other SOTA methods on the ISTD dataset.

dynamics throughout the shadow removal process. SONet is constructed with 10 convolutional layers, 4 pooling layers, and 4 upsampling layers, processing inputs that consist of a shadow-bearing image and its associated mask.

Initially, the convolutional layers serve to extract foundational image characteristics, such as edges and textures, which are essential for the subsequent steps of shadow detection and removal. This hierarchical processing enables the network to incrementally refine sophisticated and abstract features, vital for pinpointing shadowed regions.

As the architecture advances, the pooling layers play a pivotal role by performing downsampling operations. These operations not only streamline the feature maps

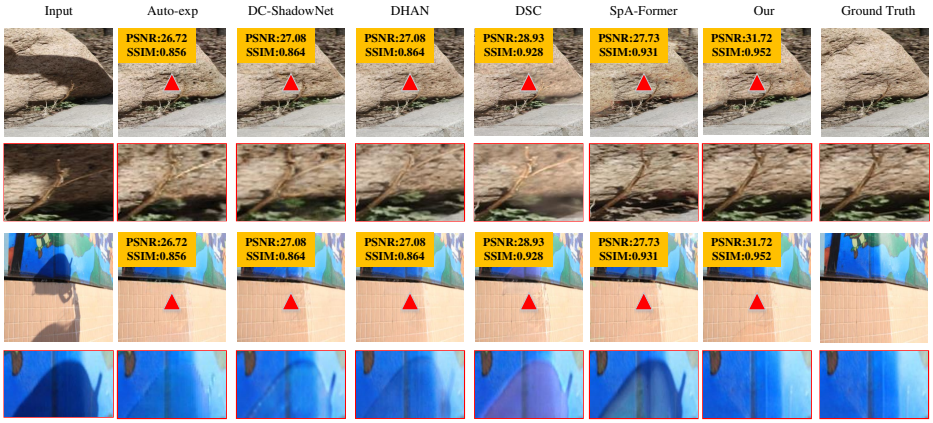

Fig. 6: The compare results of ShadClips with other SOTA methods on the SRD dataset.

Model	Precision ↑	Recall ↑	F1 score ↑	IoU ↑	PSNR ↑	SSIM ↑	RMSE ↓
SAM	0.7923	0.7206	0.7168	0.6320	0.907	27.53	6.48
Grounded-SAM	0.8932	0.9047	0.8638	0.9122	0.901	28.42	6.01
Our	**0.9622**	**0.9871**	**0.9731**	**0.9521**	**0.959**	**29.85**	**5.09**

Table 2: Abaltion study of SAM-optimizer

to reduce computational demands but also expand the receptive field of the features, facilitating the capture of shadow attributes across broader areas. Additionally, pooling layers enhance the model's resilience to minor spatial variations, thereby bolstering its generalization capacity.

The integration of upsampling layers aims to reconstruct the original image dimensions, empowering the network to generate high-quality, shadow-free images through in-depth feature analysis. This phase involves transforming low-resolution feature maps into high-resolution outputs, thereby reinstating image fidelity and mitigating shadow artifacts.

Throughout its architecture, SONet leverages an input mask to furnish spatial context, significantly aiding in the precise localization of shadows. This mask is instrumental during feature transmission, ensuring the effective elimination of shadows while preserving the authenticity and clarity of the unshadowed regions. By leveraging the characteristics of convolutional layers, pooling layers, and upsampling layers, SONet efficiently identifies and removes shadows from input images while maintaining the naturalness and detail of other areas.

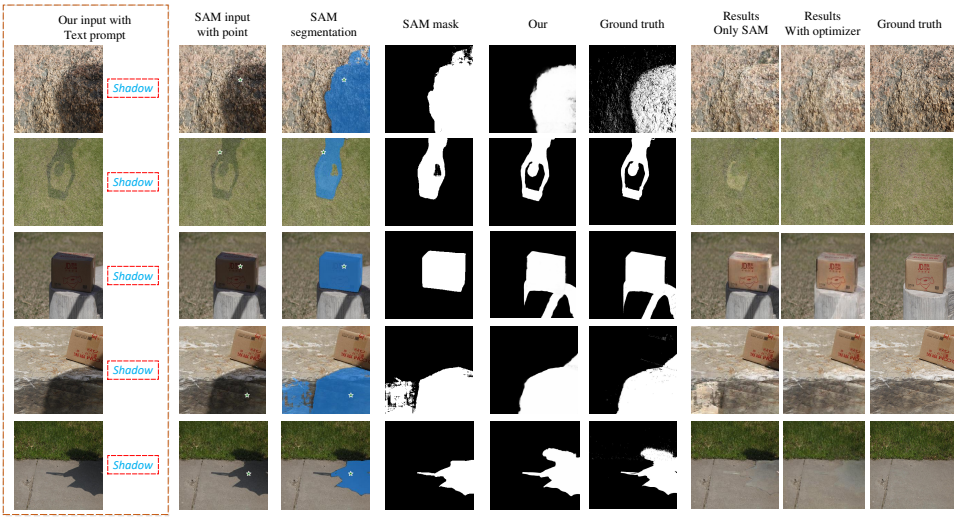

Fig. 7: The ablation study of ShadClips about segmentation results, the comparison of segmentation results between SAM and SAM-optimizer is shown from left to right (third and fourth columns), and then we put the segmentation results into SO-Net for shadow removal in order (sixth and seventh columns).

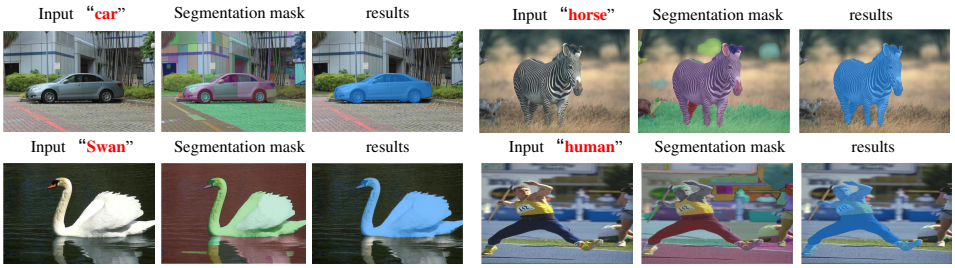

Fig. 8: The ablation study of ShadClips about segmentation results(The leftmost is the input image, the second column is the fully segmented image, and the third column is the segmented image after adding the CLIP prompt).

4 Experiment

4.1 Dataset and Training Details

This paper uses the ISTD [13] and dataset SRD. In this paper, we only need to fine-tune the SAM-Optimizer; we use the ViT-H version of the SAM model. Balanced Binary Categorical Cross Entropy (BCE) loss was used. The initial learning rate was set to 2e-4. Cosine decay was applied to the learning rate. The training was performed for 100 epochs. Experiments were implemented using PyTorch on one NVIDIA Tesla A100 GPU.

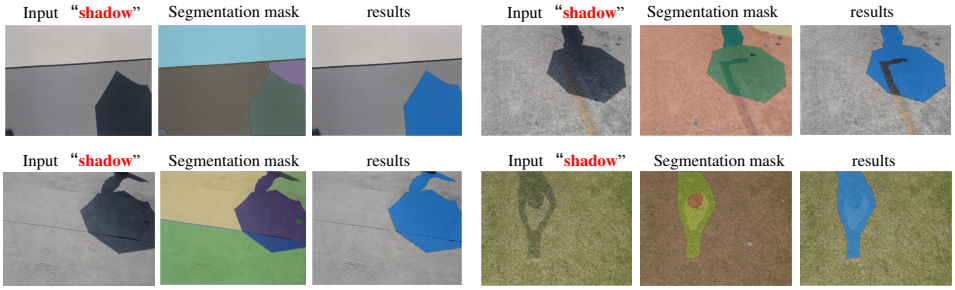

Fig. 9: The ablation study of ShadClips about segmentation results(The leftmost is the input image, the second column is the fully segmented image, and the third column is the segmented image after adding the CLIP prompt).

4.2 Experiment Design

In terms of experimental design, we divided it into three parts to verify the superiority of shad-clips. The first part is the detailed segmentation effect of SAM-optimizer+CLIP, which can be referred to as an ablation study, and the second part compares the interactivity we chose with the state-of-the-art method, Grounded-SAM+gpt2. Under the same friendly interactivity of text input, our segmentation effect is better than Grounded-SAM, and the third part compares the shadow removal effect.

4.3 Comparison with State-of-the-Art Methods

Our method is compared with existing methods including Yang [41], Guo [42], Gong [43], DeShadowNet[12], STC-GAN [13], DSC [15], Mask-ShadowGAN [8], RIS-GAN [16], DHAN [17], SID [14], LG-shadow [9], G2R [20], SG-ShadNet [47], Auto-exp [44], Bijective [26], SpA-Former [45]. We adopt RMSE, SSIM, and PSNR in the LAB color space as evaluation metrics.

The proposed ShadClips demonstrates satisfactory performance compared to other methods, as depicted in Figure 5. Additionally, we present the shadow removal effectiveness of ShadClips on the ISTD and SRD datasets, detailed in Table 1. This table shows that ShadClips achieves favorable RMSE values in shadow areas, non-shadow regions, and across entire images on the ISTD dataset. Traditional methods (Guo et al. and Gong et al.) can not distinguish this region as a shadow, thereby generating the predictions that are similar as input. Most data-driven methods can remove the shadow of the central part to some extent, while they all tend to generate the ghost shadow over the original shadow boundary. In comparison, our Shadclips can yield reasonable exposure estimation and smooth removal results on the shadow boundary. This suggests that the shadow-removed images produced by ShadClips closely resemble their corresponding ground-truth shadow-free counterparts. Models like DSC [15] and SpA-Former [45] sometimes misinterpret dark non-shadowed areas, leading to inaccuracies. Their inability to fully utilize shadow mask information, despite including it in their network input, is a notable limitation. DC-ShadowNet [12], and Auto-exp

Method	DHAN	Auto-exp	G2R	Our(SONet)	Our(SAM+SONet)
Inference time	0.3450s	0.0376s	0.0176s	0.0197s	1.525S
Parameters	21.8M	43.2M	102.5M	10.1M	645.9M

Table 3: Comparision of inference time and Parameters

are unable to remove the high degree of fusion of soft shadows and background information, and artifacts still exist, while the results of this paper and DHAN [17] are relatively satisfactory. Unlike the supervised approaches that rely on ternary datasets (input/mask/target), our method presented in this paper is unsupervised and only requires a binary dataset (input/mask). This makes our approach more practical for real-world scenarios, where it's challenging to simultaneously obtain both shadowed and unshadowed images. By leveraging the segmentation capabilities of SAM, our method not only simplifies the process but also offers a forward-thinking solution for shadow removal that aligns better with future applications.

4.3.1 Comparison study with SOTA methods of Inference time and Parameters

We compare several current leading de-shadowing methods, including DHAN [17], Auto-exp [44], and G2R [20]. Regarding inference time, these methods necessitate 0.345s, 0.0376s, and 0.0176s, respectively. In contrast, our SONet requires merely 0.0197 seconds, showcasing a processing speed comparable to, or even surpassing, that of G2R. When integrated with the SAM-Optimizer for precise shadow segmentation, the total processing time extends to 1.525 seconds. This increase is attributed to the intricate shadow recognition and segmentation tasks.

SONet exhibits considerable superiority over competing methodologies in terms of model complexity. Specifically, DHAN, Auto-exp, and G2R possess 21.8M, 43.2M, and 102.5M parameters, respectively, whereas SONet is designed with only 10.1M parameters. This stark contrast underscores SONet's heightened efficiency and refined network architecture. Upon incorporating the SAM-Optimizer, the aggregate parameter count escalates to 645.9M. Despite this surge, the synergistic fusion of SONet with the SAM-Optimizer achieves unparalleled precision in shadow removal, heralding innovative advancements in the realm of image processing.

4.4 Ablation Study

4.4.1 Ablation Study about SAM-optimizer

In our generalization experiments, as illustrated in Figure 7 and Table 2, shadows were categorized into soft (shallow) and hard (deep) shadows, with the latter bearing resemblance to inpainting tasks. SAM's segmentation performance on deep shadows was found lacking, primarily due to the indistinguishability between shadows and backgrounds. This is attributed to the deep fusion of deep shadows with background information, as evident in the first row of the results. Moreover, SAM struggled with distinguishing contrasting background colors, as illustrated in the third row where the

differentiation between lawn and street colors resulted in uniform shadow classification across distinct areas.

Furthermore, we employed four evaluation metrics: Precision, Recall, F1 score, and Intersection over Union (IoU), detailed in Table 2. The original SAM model scored 0.7923, 0.7206, 0.7168, and 0.6320, respectively. In stark contrast, our enhanced SAM-Optimizer dramatically improved these figures to 0.9622, 0.9871, 0.9731, and 0.9521, respectively, showcasing a significant advancement over the original model.

Our ablation studies extended to shadow removal efficacy. Initially, SAM required precise pixel coordinates, e.g., "200, 200", to segment shadows effectively. Given the inherent irregularity of shadows—some being exceedingly slender—accurately pinpointing their locations posed a considerable challenge. To address this, we incorporated a text-based approach using the CLIP model, facilitating shadow segmentation without necessitating exact pixel coordinates. By simply specifying "shadow" as the target, this method enhances the model's flexibility and utility. Additionally, the synergy between SAM-Optimizer and CLIP enabled precise segmentation across various categories by inputting descriptive text, such as "human, shadow", showcasing the results in Figure 8 and Figure 9.

Visualizing similarity scores in Figure 10 further underscored this efficacy, with the shadow class achieving the highest score of 0.30, particularly noted along the diagonal where the score was paramount.

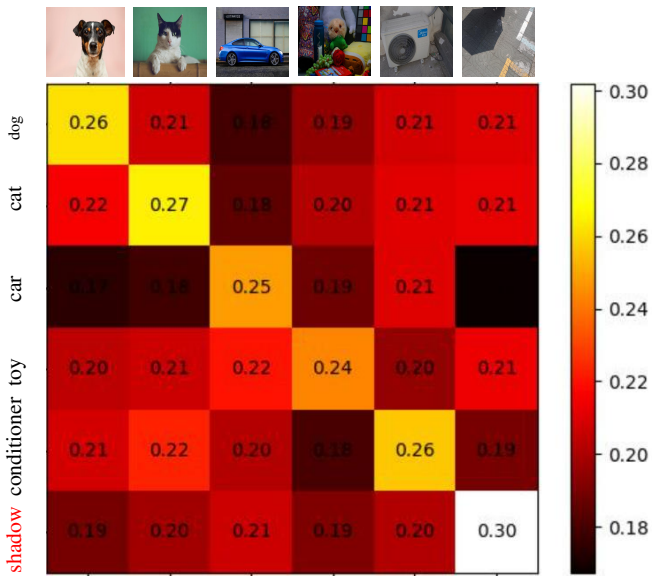

Fig. 10: The CLIP-score of shadow image

In our study, we delve deeply into the effect of the SAM-Optimizer on shadow detection, dedicating substantial effort towards the refinement of visualization techniques. The initial SAM model utilized a point-pixel input strategy, a novel yet

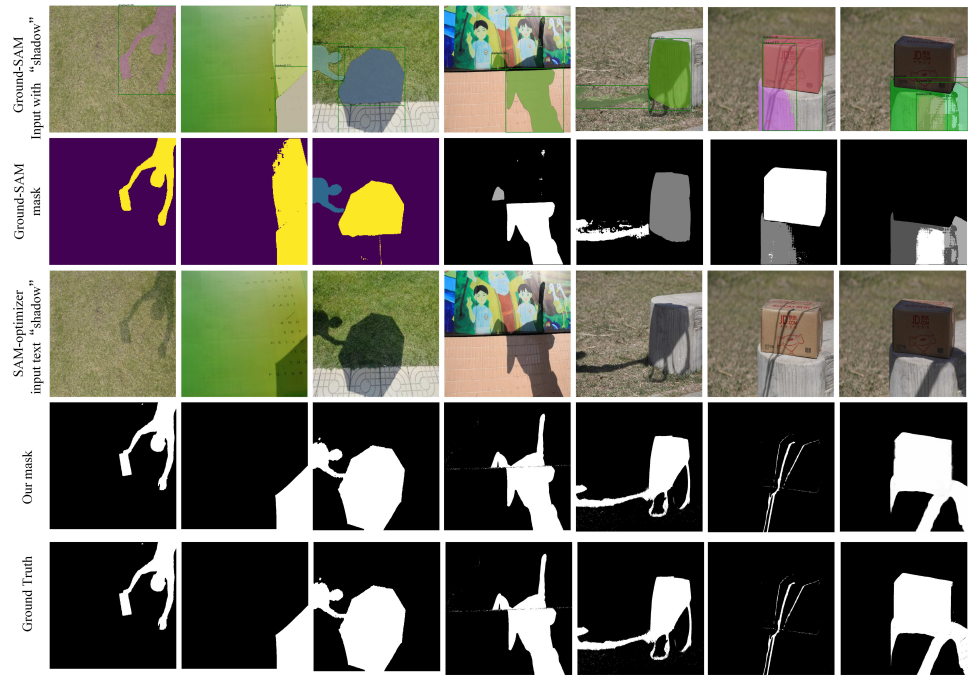

Fig. 11: The compare results of SAM-optimizer+CLIP prompt with Grounded-SAM (DINO + ChatGPT) by text prompt "shadow".

limited approach, especially in the precise localization crucial for effective shadow segmentation. The intricacies and subtle distinctions often remained elusive or poorly defined. To address these limitations, we introduced an optimizer, significantly enhancing SAM's operational capabilities. Post-modification, the segmentation outcomes not only improved but also attained a level of accuracy that far exceeded our initial expectations. This stark improvement highlights the critical role of the optimizer, showcasing its pivotal influence on refining the segmentation process.

4.4.2 Ablation Study with Grounded-SAM + ChatGPT

In our comparative analysis of shadow segmentation capabilities between SAM-Optimizer+CLIP and Grounded-SAM (ChatGPT+SAM), we utilized visualization techniques as depicted in Figure 11 and Table 2. Grounded-SAM, upon receiving the "shadow" text prompt, employs GroundDINO to identify the shadow area, conducts confidence assessments, and subsequently segments the mask. Notably, Grounded-SAM operates through a two-step procedure of detection and segmentation upon the same prompt. While capable of detecting shadow masks in a majority of scenarios, Grounded-SAM's decision-making core is powered by GroundDino, which fuses a Transformer-based detector, DINO, with truth value pre-training, integrating language processing to facilitate open-set concept generalization.

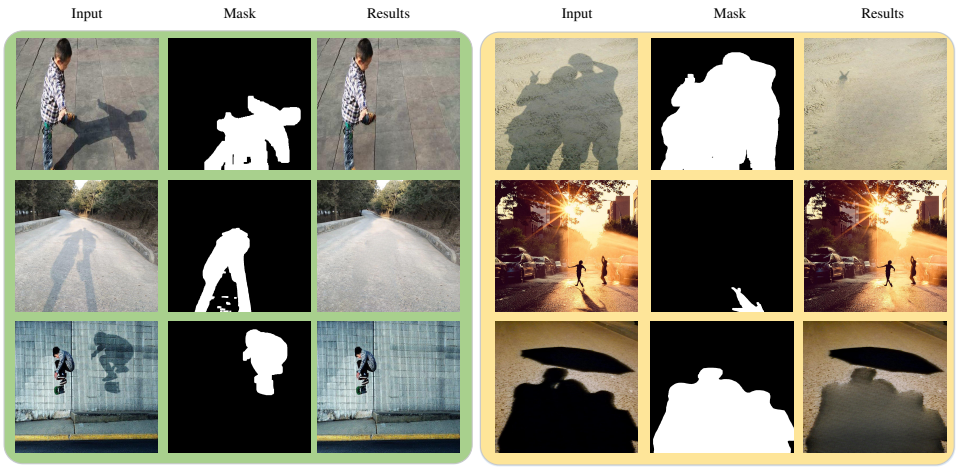

Fig. 12: The results on real image, left images are easy samples, right images are difficult samples.

Conversely, our model simplifies this process by leveraging CLIP for vector correlation analysis, thereby obviating the need for additional language model training. This approach markedly diminishes computational costs.

Moreover, we observed that the detection and segmentation timing of GroundSAM significantly lags behind that of our SAM-Optimizer, rendering it less suitable for real-time applications. Although both GroundSAM+SONet and SAM-Optimizer+CLIP+SONet demonstrate comparable capabilities in mask area detection, GroundSAM's performance on intricate samples remains suboptimal. The precision of segmentation fundamentally influences the shadow removal's success — accurate segmentation leads to effective shadow elimination. However, our experimental findings indicate that GroundSAM+SONet's performance metrics, namely SSIM, PSNR, and RMSE, are 0.901, 28.42, and 6.01, respectively, which are inferior to those achieved by our SAM-Optimizer+SONet.

4.4.3 Visual results on real shadow image

In our study, we present compelling evidence of realistic segmentation and deshadowing, as depicted in Figure 12. On the left, a straightforward example showcases the method's effectiveness in segmenting and eliminating shadows, affirming the technique's viability and precision. This result underscores the ability to preserve the intrinsic aesthetic of images, even when faced with pronounced shadows, particularly in well-illuminated settings. Here, the technique notably enhances image quality and visual appeal.

Conversely, the complexity escalates with samples characterized by challenging conditions, such as the large, dark shadows observed at night or the multifaceted shadow effects encountered at dusk, illustrated on the right. These scenarios, marked by intricate lighting and multiple shadows, pose significant hurdles. Although our

algorithms retain the capability to identify and address some shadows in these conditions, preserving the natural look and intricate details of the image proves to be more daunting.

Looking forward, our research aims to refine and advance our algorithms, enabling them to adeptly navigate the nuances of complex lighting environments and substantial dark shadows. We are confident that ongoing technological evolution and enhancements will empower us to offer increasingly robust and versatile solutions. These advancements will address a broad spectrum of lighting and shadow challenges, pushing the frontiers of image processing technology ever forward.

5 Conclusion

In this paper, we introduce ShadClips, a novel approach to tackling the task of image shadow removal. While the SAM model exhibits exceptional proficiency in image segmentation, it falls short in accurately segmenting shadowed images. To address this limitation, we have devised the SAM-Optimizer, specifically crafted to enhance shadow segmentation across diverse shadow profiles. However, this methodology requires pixel point inputs, posing a challenge in identifying certain elongated shadows. To elevate user experience and interaction, we have integrated CLIPs, facilitating text-based prompts such as "A photo with a shadow". This integration enables the SAM-Optimizer to precisely pinpoint the most relevant shadow mask within SAM's extensive classification. Concurrently, we have designed a second-order shadow removal network, SO-Net. Moving forward, our research will focus on zero-shot tasks for larger models. Through these advancements and new research directions, we aim to further enhance the precision and usability of shadow removal, contributing significantly to the field of image processing.

6 Acknowledgement

This work was supported by the National Natural Science Foundation NO. 62273235, the Joint Fund of the Ministry of Education NO. 8091B022101, Deep Blue Program Fund Project, Second Institute of Oceanography, Ministry of Natural Resources.

7 Declarations

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] Bin Sheng, Ping Li, Xiaoxin Fang, Ping Tan, and Enhua Wu. "Depth-aware motion deblurring using loopy belief propagation". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.4 (2019), pp. 955–969.

[2] Mrinmoy Sen, Sai Pradyumna Chermala, Nazrinbanu Nurmohammad Nagori, Venkat Peddigari, Praful Mathur, BH Prasad, and Moonhwan Jeong. "Shards: Efficient shadow removal using dual stage network for high-resolution images". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 1809–1817.

[3] Mehmet Kerim Yücel, Valia Dimaridou, Bruno Manganelli, Mete Ozay, Anastasios Drosou, and Albert Saà-Garriga. "Lra&ldra: Rethinking residual predictions for efficient shadow detection and removal". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 4925–4935.

[4] Yang Wen, Bin Sheng, Ping Li, Weiyao Lin, and David Dagan Feng. "Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution". In: *IEEE Transactions on Image Processing* 28.2 (2018), pp. 994–1006.

[5] Zhihua Chen, Ting Gao, Bin Sheng, Ping Li, and CL Philip Chen. "Outdoor shadow estimating using multiclass geometric decomposition based on BLS". In: *IEEE transactions on cybernetics* 50.5 (2018), pp. 2152–2165.

[6] Yadang Chen, Ren Jiang, Yuhui Zheng, Bin Sheng, Zhi-Xin Yang, and Enhua Wu. "Dual Branch Multi-Level Semantic Learning for Few-Shot Segmentation". In: *IEEE Transactions on Image Processing* 33 (2024), pp. 1432–1447.

[7] Jixue Tang, Xiang-ning Wang, Xiaolong Yang, Yang Wen, Bo Qian, Tingli Chen, and Bin Sheng. "TSNet: Task-specific network for joint diabetic retinopathy grading and lesion segmentation of ultra-wide optical coherence tomography angiography images". In: *The Visual Computer* (2023), pp. 1–12.

[8] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. "Mask-shadowgan: Learning to remove shadows from unpaired data". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2472–2481.

[9] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. "Shadow removal by a lightness-guided network with training on unpaired data". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 1853–1865.

[10] Hieu Le and Dimitris Samaras. "Physics-based shadow image decomposition for shadow removal". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2021), pp. 9088–9101.

[11] Qianhao Yu, Naishan Zheng, Jie Huang, and Feng Zhao. "Cnsnet: A cleanness-navigated-shadow network for shadow removal". In: *European Conference on Computer Vision*. Springer. 2022, pp. 221–238.

[12] 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*. 2017, pp. 4067–4075.

[13] 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*. 2018, pp. 1788–1797.

[14] Hieu Le and Dimitris Samaras. "Shadow removal via shadow image decomposition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 8578–8587.

[15] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. "Direction-aware spatial context features for shadow detection and removal". In: *IEEE transactions on pattern analysis and machine intelligence* 42.11 (2019), pp. 2795–2808.

[16] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. "Risgan: Explore residual and illumination with generative adversarial networks for shadow removal". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12829–12836.

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

[18] 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*. 2021, pp. 4743–4752.

[19] Subhrajyoti Dasgupta, Arindam Das, Sudip Das, Ciarán Eising, Andrei Bursuc, Ujjwal Bhattacharya, and Senthil Yogamani. "UnShadowNet: Illumination Critic Guided Contrastive Learning For Shadow Removal". In: *IEEE Access* (2023).

[20] Florin-Alexandru Vasluianu, Andres Romero, Luc Van Gool, and Radu Timofte. "Self-supervised shadow removal". In: *arXiv preprint arXiv:2010.11619* (2020).

[21] Hieu Le and Dimitris Samaras. "From shadow segmentation to shadow removal". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer. 2020, pp. 264–281.

[22] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. "From shadow generation to shadow removal". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 4927–4936.

[23] Chao Tan and Xin Feng. "Unsupervised shadow removal using target consistency generative adversarial network". In: *arXiv preprint arXiv:2010.01291* (2020).

[24] Yingqing He, Yazhou Xing, Tianjia Zhang, and Qifeng Chen. "Unsupervised portrait shadow removal via generative priors". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 236–244.

[25] Yurui Zhu, Zeyu Xiao, Yanchi Fang, Xueyang Fu, Zhiwei Xiong, and Zheng-Jun Zha. "Efficient model-driven network for shadow removal". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 3. 2022, pp. 3635–3643.

[26] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. "Bijective mapping network for shadow removal". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5627–5636.

[27] 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". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[28] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. "fastsam". In: *arXiv preprint arXiv:2306.12156* (2023).

[29] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. "Faster segment anything: Towards lightweight sam for mobile applications". In: *arXiv preprint arXiv:2306.14289* (2023).

[30] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. "Semantic-sam: Segment and recognize anything at any granularity". In: *arXiv preprint arXiv:2307.04767* (2023).

[31] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. "Grounded sam: Assembling open-world models for diverse visual tasks". In: *arXiv preprint arXiv:2401.14159* (2024).

[32] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment anything in medical images". In: *Nature Communications* 15.1 (2024), p. 654.

[33] Kaidong Zhang and Dong Liu. "Customized segment anything model for medical image segmentation". In: *arXiv preprint arXiv:2304.13785* (2023).

[34] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. "Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model". In: *arXiv preprint arXiv:2304.05396* (2023).

[35] D Wang, J Zhang, B Du, D Tao, and L Zhang. "Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. arXiv". In: *arXiv preprint arXiv:2305.02034* (2023).

[36] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. "SAM fails to segment anything?–SAM-Adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, and more". In: *arXiv preprint arXiv:2304.09148* (2023).

[37] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. "Inpaint anything: Segment anything meets image inpainting". In: *arXiv preprint arXiv:2304.06790* (2023).

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[39] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. "Auto-exposure fusion for single-image shadow removal". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 10571–10580.

[40] 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*. 2021, pp. 2429–2437.

[41] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. "Shadow removal using bilateral filtering". In: *IEEE Transactions on Image processing* 21.10 (2012), pp. 4361–4368.

[42] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. "Paired regions for shadow detection and removal". In: *IEEE transactions on pattern analysis and machine intelligence* 35.12 (2012), pp. 2956–2967.

[43] Han Gong and DP Cosker. "Interactive shadow removal and ground truth for variable scene categories". In: *BMVC 2014-Proceedings of the British Machine Vision Conference 2014*. 2014.

[44] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. "Auto-exposure fusion for single-image shadow removal". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 10571–10580.

[45] Xiao Feng Zhang, Chao Chen Gu, and Shan Ying Zhu. "Spa-former: Transformer image shadow detection and removal via spatial attention". In: *arXiv preprint arXiv:2206.10910* (2022).

[46] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. "Shadowformer: Global context helps image shadow removal". In: *arXiv preprint arXiv:2302.01650* (2023).

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

[48] Zhihua Chen, Yu Zhou, Ran Li, Ping Li, and Bin Sheng. "SCPA-Net: Self-calibrated pyramid aggregation for image dehazing". In: *Computer Animation and Virtual Worlds* 33.3-4 (2022), e2061.

[49] Yu Zhou, Zhihua Chen, Ping Li, Haitao Song, CL Philip Chen, and Bin Sheng. "FSAD-Net: Feedback spatial attention dehazing network". In: *IEEE transactions on neural networks and learning systems* (2022).

[50] Zhihua Chen, Yu Zhou, Ping Li, Xiaoyu Chi, Lei Ma, and Bin Sheng. "DCNet: Dual-task cycle network for end-to-end image dehazing". In: *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2021, pp. 1–6.

[51] Yeying Jin, Ruoteng Li, Wenhan Yang, and Robby T Tan. "Estimating reflectance layer from a single image: Integrating reflectance guidance and shadow/specular aware learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 1. 2023, pp. 1069–1077.

[52] Xinkui Mei, Xiufen Ye, Xiaofeng Zhang, Yusong Liu, Junting Wang, Jun Hou, and Xuli Wang. "UIR-Net: A Simple and Effective Baseline for Underwater Image Restoration and Enhancement". In: *Remote Sensing* 15.1 (2022), p. 39.

[53] Chao Tan and Xin Feng. "Unsupervised shadow removal using target consistency generative adversarial network". In: *arXiv preprint arXiv:2010.01291* (2020).

[54] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. "From shadow generation to shadow removal". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 4927–4936.

[55] Xiaofeng Zhang, Feng Chen, Cailing Wang, Ming Tao, and Guo-Ping Jiang. "Sienet: Siamese expansion network for image extrapolation". In: *IEEE Signal Processing Letters* 27 (2020), pp. 1590–1594.

[56] Runhan Shen, Xiaofeng Zhang, and Yonggang Xiang. "AFFNet: attention mechanism network based on fusion feature for image cloud removal". In: *International Journal of Pattern Recognition and Artificial Intelligence* 36.08 (2022), p. 2254014.

[57] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Zhihao Liu, and Song Wang. "Crformer: A cross-region transformer for shadow removal". In: *arXiv preprint arXiv:2207.01600* (2022).

[58] Yeying Jin, Aashish Sharma, and Robby T Tan. "Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5027–5036.

[59] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International conference on machine learning*. PMLR. 2022, pp. 12888–12900.

[60] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.