

Prompt-Aware Controllable Shadow Removal

Kerui Chen, Zhiliang Wu, Wenjin Hou, Kun Li, Hehe Fan* and Yi Yang

ReLER, CCAI, Zhejiang University, China

Abstract

Shadow removal aims to restore the image content in shadowed regions. While deep learning-based methods have shown promising results, they still face key challenges: 1) uncontrolled removal of all shadows, or 2) controllable removal but heavily relies on precise shadow region masks. To address these issues, we introduce a novel paradigm: prompt-aware controllable shadow removal. Unlike existing approaches, our paradigm allows for targeted shadow removal from specific subjects based on user prompts (e.g., dots, lines, or subject masks). This approach eliminates the need for shadow annotations and offers flexible, user-controlled shadow removal. Specifically, we propose an end-to-end learnable model, the *Prompt-Aware Controllable Shadow Removal Network* (PACSRNet). PACSRNet consists of two key modules: a prompt-aware module that generates shadow masks for the specified subject based on the user prompt, and a shadow removal module that uses the shadow prior from the first module to restore the content in the shadowed regions. Additionally, we enhance the shadow removal module by incorporating feature information from the prompt-aware module through a linear operation, providing prompt-guided support for shadow removal. Recognizing that existing shadow removal datasets lack diverse user prompts, we contribute a new dataset specifically designed for prompt-based controllable shadow removal. Extensive experimental results demonstrate the effectiveness and superiority of PACSRNet.

1 Introduction

Shadow removal is a fundamental visual restoration task, which aims to restore the information of the darkness region caused by light occlusion in an image [Liu *et al.*, 2024c; Vasluianu *et al.*, 2024]. Realistic shadow removal can benefit various computer vision tasks [Li *et al.*, 2024a; Li *et al.*, 2025a; Li *et al.*, 2025b; Wang *et al.*, 2025], such as

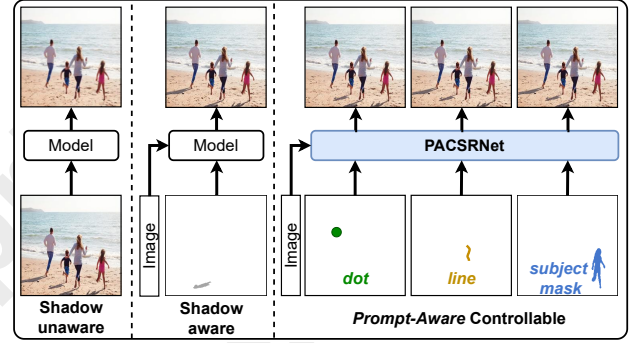


Figure 1: Comparison with existing shadow removal methods. (a) **Shadow unaware** removes all the shadow regions with only raw images as input. (b) **Shadow aware** removes the shadow regions corresponding to the given shadow mask. (c) Our **prompt-aware removal**. Different from (a) and (b), the proposed method allows the removal of any subject’s shadow with various prompts (*i.e.*, dot, line, and subject mask).

face recognition [Du *et al.*, 2022; Kim *et al.*, 2022], content completion [Wu *et al.*, 2023b; Zhang *et al.*, 2023; Wu *et al.*, 2024], and so on. Therefore, shadow removal methods have been extensively studied in computer vision. With the rapid development of deep learning and increasing interest in image editing [Alaluf *et al.*, 2022] and augmented virtual reality [Zhang *et al.*, 2024], shadow removal is getting increased attention in recent years.

Recently, several deep learning-based shadow removal methods have been proposed. These methods generally follow two paradigms: global shadow removal without shadow mask [Hu *et al.*, 2019; Wang *et al.*, 2024c; Li *et al.*, 2024b], and shadow removal requiring a precise shadow mask as prior information [Guo *et al.*, 2023a; Guo *et al.*, 2023b; Xiao *et al.*, 2024]. The former is referred to as *shadow unaware*, while the latter is called *shadow aware*. Although these methods have shown promising results, they still suffer significant challenges in practical applications. On the one hand, shadow unaware methods aim to globally remove shadows from images in a coarse-grained manner, as shown in Fig. 1 (a). They lack the ability to perform fine-grained removal based on user-specific needs, making them uncontrolled. On the other hand, shadow aware methods allow for control over the removed shadows, but they rely on manu-

*Corresponding author.

ally annotated masks of shadow region, as shown in Fig. 1 (b). However, manually annotating precise shadow regions is non-trivial due to the blur boundary between shadow and non-shadow regions, and it is time-consuming and labor-intensive.

In this paper, we propose a new paradigm, named **prompt-aware controllable shadow removal**, to address the aforementioned challenges. Unlike existing shadow unaware methods [Hu *et al.*, 2019; Wang *et al.*, 2024c; Li *et al.*, 2024b] and shadow aware methods [Guo *et al.*, 2023a; Guo *et al.*, 2023b; Xiao *et al.*, 2024], our proposed paradigm enables the removal of shadows from specified subjects based on user prompts, such as dot, line, and subject mask in Fig. 1(c). In this way, we not only achieve controllable shadow removal, but also avoid the non-trivial and tedious tasks of shadow region annotation. These advantages ensure the flexibility and convenience of our paradigm in practical applications.

Specifically, we propose an end-to-end prompt-aware controllable shadow removal network called PACSRNet. It consists of a prompt-aware module and a shadow removal module. The former aims to generate the shadow masks for a specified subject based on the user prompt, while the latter leverages the shadow prior obtained from the former to restore the contents in the shadow regions. We further leverage the feature from the prompt-aware module via a linear operation to provide prompt-aware guidance for the shadow removal module. Among prompt-aware module and a shadow removal module, we design spatial-frequency interaction and dense-sparse local attention blocks as their basic units, respectively. The spatial-frequency interaction block aims to facilitate information interaction between spatial features and frequency features, thereby effectively enhancing the shadow perception and contextual understanding capabilities of the prompt-aware module. The dense-sparse local attention block is used to suppress negative influence of irrelevant pixels, maximizing the utilization of relevant pixels to improve shadow removal performance.

Furthermore, existing shadow removal datasets [Qu *et al.*, 2017; Wang *et al.*, 2018; Le and Samaras, 2019] typically contain only shadow images, shadow-free images, and shadow masks, with a lack of diverse user prompts. As a result, they are not suitable for the proposed prompt-aware controllable shadow removal task. In this paper, we introduce the first prompt-based controllable shadow removal dataset, which provides multiple user prompts for each sample, such as dot, line, and subject mask, effectively simulating real-world shadow removal scenarios. Extensive experimental results demonstrate the effectiveness and superiority of the proposed PACSRNet.

To sum up, our contributions are summarized as follows:

- We propose a new paradigm for shadow removal: prompt-based controllable shadow removal. It enables the removal of shadows from specified subjects based on user prompts, effectively alleviating the need for shadow masks. To the best of our knowledge, this is the first work to explore prompt-based shadow removal.
- We develop a prompt-aware controllable shadow removal network. This network consists of a prompt-aware module and a shadow removal module, with

spatial-frequency interaction and dense-sparse local attention blocks as their basic units.

- We customize the first prompt-based controllable shadow removal dataset. This dataset consists of 11,900 shadow and shadow-free samples, along with various prompts (*e.g.*, dot, line, and subject mask). We will release it to facilitate subsequent research.

2 Related Work

2.1 Shadow Removal

Classic shadow removal methods [Finlayson *et al.*, 2005; Shor and Lischinski, 2008; Finlayson *et al.*, 2009; Yang *et al.*, 2012] typically utilized hand-crafted prior knowledge, such as illumination, region-based characteristics, and density, to remove shadow regions. However, such prior knowledge often lacks high-level semantic features, which results in the above methods usually failing in complex scenarios. Recently, learning-based approaches [Wang *et al.*, 2018; Zhu *et al.*, 2022; Li *et al.*, 2023; Guo *et al.*, 2023b; Guo *et al.*, 2023a; Wang *et al.*, 2024c; Mei *et al.*, 2024] have significantly advanced removal effects using large-scale datasets. On the one hand, [Wang *et al.*, 2018; Wang *et al.*, 2024c] realize end-to-end shadow unaware removal of all shadow regions. However, these approaches are uncontrolled and cannot target specified shadow regions for shadow removal. On the other hand, some studies [Li *et al.*, 2023; Guo *et al.*, 2023a; Guo *et al.*, 2023b; Luo *et al.*, 2025] present a shadow aware removal network removing shadows by global shadow masks as prior input. Meanwhile, [Mei *et al.*, 2024] achieves instance-level removal by using a single object’s shadow mask. Despite achieving impressive results, these shadow aware networks still heavily rely on accurate shadow masks to indicate removal regions, which can be cumbersome and unfriendly in practical applications. To achieve controllable shadow removal and enhance the user experience, we propose a new prompt-aware shadow removal network that can accept diverse prompts (*i.e.*, dot, line and subject mask), offering a more flexible and intuitive interface.

2.2 Shadow Detection

Shadow detection [Huang *et al.*, 2011; Vicente *et al.*, 2017; Khan *et al.*, 2014; Chen *et al.*, 2020] aims to identify and segment shadow regions within a given image. Early studies [Huang *et al.*, 2011; Vicente *et al.*, 2015; Shen *et al.*, 2015; Vicente *et al.*, 2017] focus mainly on utilizing various hand-crafted heuristic cues, such as color priors, image texture, or chromaticity, to identify shadow regions. However, when dealing with complex shadow scenes, hand-crafted features often struggle to provide adequate information and are limited in their ability to describe shadows, leading to significant degradation. Recently, deep learning-based shadow detection methods [Khan *et al.*, 2014; Le *et al.*, 2018; Chen *et al.*, 2020; Liao *et al.*, 2021; Yang *et al.*, 2023] have achieved impressive results and have become mainstream. For detecting all the shadow regions in the given image more accurately, [Khan *et al.*, 2014] first uses CNNs to learn semantic features for shadow detection, [Chen *et al.*, 2020] presents a semi-supervised shadow detection algorithm by

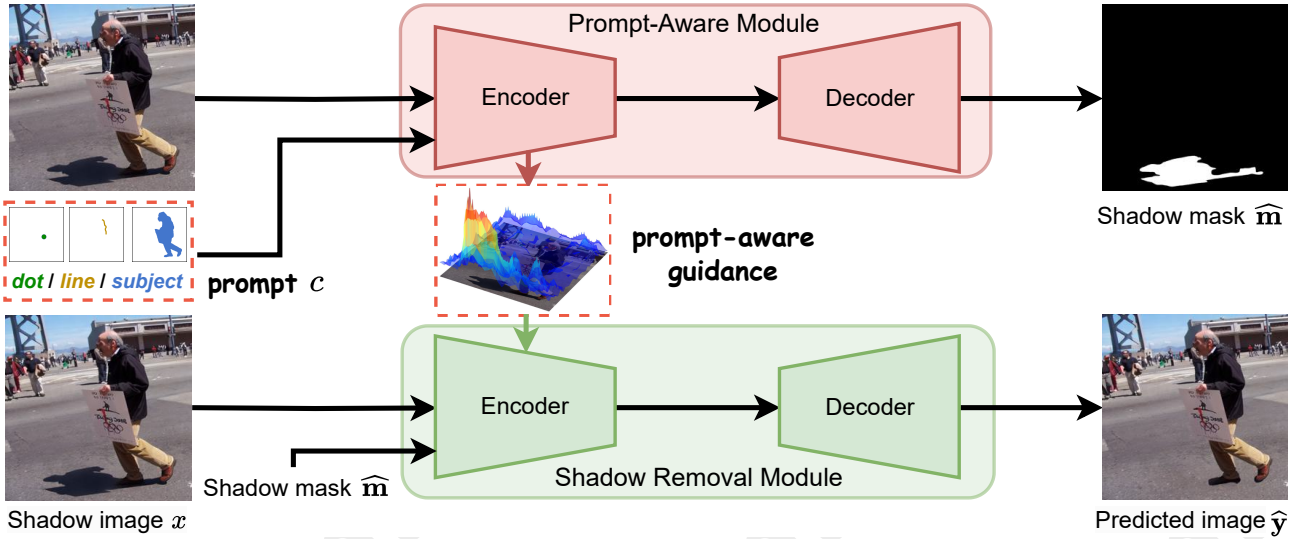


Figure 2: **The overview of the PACSRNet.** PACSRNet is composed of a prompt-aware module and a shadow removal module. Prompt-aware module takes a shadow image and a prompt c as inputs generating corresponding shadow mask $\hat{m}$ and prompt-aware guidance feature. The generated shadow mask will be served as explicit guidance and fed into the shadow removal module along with the shadow image for shadow removal. The prompt-aware guidance feature is applied to encoder of shadow removal module guiding the shadow removal implicitly.

exploring unlabeled data through a multitask mean teacher framework, [Liao *et al.*, 2021] introduces the confidence map prediction network to combine the prediction results by multiple methods, and the latest method [Yang *et al.*, 2023] uses iterative label tuning to refine noisy labels helping the network better recognize non-shadow regions and alleviate overfitting. Although these methods can efficiently detect global shadow regions, they lack attention to controllability. Therefore, [Wang *et al.*, 2020; Wang *et al.*, 2021; Wang *et al.*, 2022] shift to instance-level shadow detection, focusing on a finer granularity by identifying shadow regions corresponding to a specific object. However, these methods require an accurate object mask for shadow detection. In this paper, we attempt to simultaneously perform shadow removal and detection within diverse and flexible prompts.

3 Method

3.1 Problem Formulation and Overview

Problem Formulation. Given a shadow image $x \in \mathbb{R}^{h \times w \times 3}$ with width w and height h . The goal of shadow removal is to generate shadow-free image $\hat{y} \in \mathbb{R}^{h \times w \times 3}$, which should be spatially consistent with the ground truth $y \in \mathbb{R}^{h \times w \times 3}$. Existing methods generally follow two technical pipelines, *shadow unaware removal* and *shadow aware removal*, to solve this task. Despite significant progress, they still suffer challenges in practical applications. For example, shadow unaware removal methods [Hu *et al.*, 2019; Wang *et al.*, 2024c; Li *et al.*, 2024b] lack the ability to perform controllable fine-grained shadow removal, and shadow aware removal methods [Guo *et al.*, 2023a; Guo *et al.*, 2023b; Xiao *et al.*, 2024] rely on precise shadow region mask $m \in \mathbb{R}^{h \times w \times 1}$ as a prerequisite. These challenges significantly limit the practical application space of shadow removal methods.

In this paper, we propose a new paradigm, *prompt-aware*

controllable shadow removal. It aims to enable removal of shadow from the specified subject based on user prompt c . Unlike existing methods, our paradigm can directly learn the mapping $\mathcal{G}$ from x to $\hat{y}$ guided by user prompt c , without requiring any annotations of the shadow region m . The proposed prompt-aware controllable shadow removal paradigm can be formulated as follows:

$$\hat{y} = \mathcal{G}(x; c), \quad (1)$$

where c can be diverse prompt clues, such as dot, line, subject mask, and so on. This paradigm not only enables controllable fine-grained shadow removal by user prompts, but also effectively alleviates the need for shadow masks during inference. **Network Design.** In this paper, we design an end-to-end prompt-aware controllable shadow removal network, named PACSRNet. As illustrated in Fig. 2, the proposed PACSRNet comprises two key components: a prompt-aware module and a shadow removal module. First, the prompt-aware module generates a shadow mask and prompt-aware guidance for a specified subject based on user prompts, indicating the content restoration regions for the shadow removal module. Then, the shadow removal module aggregates valid contextual information from shadow-free regions using the shadow prior obtained from the prompt-aware module to restore the content of shadow regions. In our network, the prompt-aware module and shadow removal module are closely correlated and mutually constrained. The former assists the latter to locate the shadow regions of a specific subject based on user prompt, while the latter regularizes the former by the reconstruction loss, enforcing it focuses on the shadow region of the specified subject.

3.2 Prompt-Aware Module

The prompt-aware module is designed to generate a shadow mask and provide prompt-aware guidance of a specified subject based on user prompt, indicating the restoration regions

for the shadow removal module. In the prompt-aware module, it is crucial to locate the shadow boundary of a specific subject based on user prompt. In fact, there is a significant frequency difference between the shadow and non-shadow regions in an image. On the one hand, shadow regions tend to exhibit lower frequencies compared to non-shadow ones. On the other hand, the transition from the shadow region to the non-shadow region is abrupt, creating a distinct gradient change at the shadow boundary. Consequently, integrating frequency information into spatial features helps the shadow perception module predict the shadow mask.

For this purpose, we design a spatial-frequency interaction (SFI) block by introducing discrete wavelet transform (DWT) [Mallat, 1989]. It decomposes the features into frequency components using DWT and interacts with the spatial domain features to enhance the ability to perceive shadow boundaries. Specifically, for the feature f corresponding to the image x , we first feed it into two different branches to extract the frequency feature e and the spatial feature z , respectively.

$$e = \mathcal{D}(f), \quad z = \mathcal{C}_1(f), \quad (2)$$

where $\mathcal{D}(\cdot)$ and $\mathcal{C}_1(\cdot)$ denote a DWT layer and a 3×3 convolution layer. In this way, the feature e can capture the frequency variation details in the image, while feature z learn the image semantic information in the spatial domain.

After obtained frequency feature e and spatial feature z , we interact the information between them to enhance the representation ability of shadow regions. Such a strategy can better utilize the complementary information between frequency feature e and the spatial feature z to facilitate shadow perception of the prompt-aware module. The whole feature interaction process can be formulated as:

$$g = e \oplus \mathcal{C}_2(e), \quad h = f \oplus \mathcal{C}_3(z), \quad (3)$$

where g and h are output of frequency and spatial branches, respectively. $\mathcal{C}_2(\cdot)$ and $\mathcal{C}_3(\cdot)$ denote two different 1×1 convolutional layers. $\oplus$ represents the addition operation.

Finally, the output g of the frequency branch is converted into the spatial domain by an inverse DWT layer $\mathcal{I}(\cdot)$, and aggregated with the features from the spatial branch to obtain the final interaction feature t .

$$t = \mathcal{A}(\mathcal{I}(g), h), \quad (4)$$

where $\mathcal{A}(\cdot, \cdot)$ is aggregation convolutional layer. The final predicted shadow mask $\hat{m}$ can be obtained by decoding t with the decoder. Additionally, we further leverage the feature from prompt-aware module via a linear operation to provide prompt-aware guidance for shadow removal module.

3.3 Shadow Removal Module

The shadow removal module aims to restore the content in the predicted shadow regions $\hat{m}$ by aggregating effective contextual information. This is consistent with the goal of shadow aware removal settings. Recently, benefiting from the advantages of long-range feature capture, transformer-based methods have achieved superior performance. To enable the local attention mechanism [Wang *et al.*, 2024b; Wang *et al.*, 2024a] of the transformer to capture global representations, the latest method [Xiao *et al.*, 2024] introduces

a random shuffle strategy to enable global interactions. Although significant improvements have been made, this strategy introduces new challenges. Specifically, the random shuffle strategy gives each pixel an equal probability of appearing in the same local token. However, this means that more irrelevant content will be introduced in the local tokens. In this scenario, aggregating the features using all attention relations based on query-key pairs will introduce redundant or irrelevant content into shadow regions, resulting in blurry or compromised results [Zhou *et al.*, 2024; Wu *et al.*, 2025]. Therefore, we design a dense-sparse local attention (DSL) block to alleviate the above challenges.

For the local token l_i obtained after using the random shuffle strategy, we first map it into query (q_i), key (k_i), and value (v_i) by three different linear layers. Then, the standard dense-attention score d_i between q_i and k_i can be calculated by a softmax layer:

$$d_i = \text{Softmax} \left(\frac{q_i \cdot (k_i)^T}{\sqrt{d}} \right), \quad (5)$$

where $\text{Softmax}(\cdot)$ denotes the softmax layer. Since the shuffle strategy introduces irrelevant content, Eq.(5) will inevitably reduce the attention scores of relevant content, resulting in sub-optimal shadow removal results.

To mitigate the negative impact of irrelevant content, we develop a sparse attention by a screening operation. The sparse attention score s_i can be calculated as follows:

$$s_i = \text{Softmax} \left(\text{mask} \left(\frac{q_i \cdot (k_i)^T}{\sqrt{d}} \right) \right), \quad (6)$$

where mask denotes a screening operation. It sets negative similarity values to negative infinity and retains the positive similarity values.

Note that simply relying on s_i will result in over sparsity, which in turn causes the encoded features insufficient to restore the shadow region. Only using d_i will inadvertently introduce irrelevant content into shadow regions, leading to distorted and blurry results. Therefore, we introduce two learnable weights (ω_1 and ω_2) to balance d_i and s_i . The aggregated feature $\hat{l}_i$ can be calculated as follows:

$$\hat{l}_i = (\omega_1 \odot d_i + \omega_2 \odot s_i) \cdot v_i, \quad (7)$$

where $\odot$ denotes the multiply operation. Finally, to preserve the original semantic information of the image, all aggregated features will be restored to the original order by the inverse shuffle operation.

Additionally, to obtain more guidance from the prompt-aware module, we connect the multi-scale features extracted by the prompt-aware module to the encoder of the shadow removal module through a linear layer. In this way, the prompt-aware module can implicitly guide the feature learning of the shadow removal module, enabling controllable shadow removal based on user prompts.

3.4 Loss Functions

We train our PACSRNet by minimizing the following loss:

$$\mathcal{L} = \lambda \mathcal{L}_{re} + \mathcal{L}_{pr}, \quad (8)$$

where $\mathcal{L}_{re}$ and $\mathcal{L}_{pr}$ denote shadow removal loss and shadow prediction loss, respectively. λ is a trade-off parameter. In real implementation, we empirically set $\lambda = 3$.

Methods	Shadow Regions			Non-Shadow Regions			All Regions		
	PSNR $\uparrow$	SSIM $\uparrow$	RMSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	RMSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	RMSE $\downarrow$
PCSRD dataset									
BMNet [Zhu <i>et al.</i> , 2022]	44.459	0.9955	5.9428	48.832	0.9961	0.6401	41.627	0.9881	0.7743
Inpaint4Shadow [Li <i>et al.</i> , 2023]	45.443	0.9958	8.6692	45.643	0.9955	0.8817	41.969	0.9884	0.9723
OmniSR [Xu <i>et al.</i> , 2025]	44.632	0.9956	6.1321	49.013	0.9961	0.6377	41.743	0.9881	0.8538
RASM [Liu <i>et al.</i> , 2024a]	45.127	0.9957	5.6430	49.327	0.9962	0.5703	42.103	0.9887	0.6213
ShadowFormer [Guo <i>et al.</i> , 2023a]	45.318	0.9957	<u>5.1687</u>	49.774	0.9962	<u>0.5018</u>	42.302	<u>0.9889</u>	<u>0.6183</u>
ShadowDiffusion [Guo <i>et al.</i> , 2023b]	44.593	0.9896	12.507	46.094	0.9889	1.0726	40.823	0.9869	1.2135
HomoFormer [Xiao <i>et al.</i> , 2024]	45.256	0.9957	5.3159	49.593	0.9962	0.5042	42.219	0.9888	0.6251
PACSRNet[†] (Ours)	45.559	0.9959	4.9987	49.784	0.9964	0.4927	42.494	0.9892	0.6038
PACSRNet (Ours) w/ Dot	43.382	0.9952	6.1311	48.622	0.9961	0.4864	40.956	0.9878	0.6341
PACSRNet (Ours) w/ Line	43.479	0.9953	6.0847	48.867	0.9961	0.4861	41.056	0.9879	0.6333
PACSRNet (Ours) w/ Subject Mask	44.354	0.9957	5.5541	49.061	0.9961	0.5022	41.592	0.9884	0.6263
ISTD+ dataset									
BMNet [Zhu <i>et al.</i> , 2022]	37.87	0.991	5.62	37.51	0.985	2.45	33.98	0.972	2.97
Inpaint4Shadow [Li <i>et al.</i> , 2023]	38.10	0.990	6.09	37.66	0.981	2.82	34.16	0.967	3.35
ShadowFormer [Guo <i>et al.</i> , 2023a]	39.48	0.992	5.23	38.82	0.983	2.30	35.46	0.971	2.78
ShadowDiffusion [Guo <i>et al.</i> , 2023b]	<u>39.69</u>	<u>0.992</u>	4.97	38.89	0.987	2.28	<u>35.67</u>	0.975	2.72
HomoFormer [Xiao <i>et al.</i> , 2024]	39.49	0.993	4.73	38.75	0.984	2.23	35.35	0.975	<u>2.64</u>
PACSRNet[†] (Ours)	40.32	0.993	<u>4.89</u>	39.18	<u>0.985</u>	<u>2.27</u>	36.02	<u>0.972</u>	2.63

Table 1: Comparison with the state-of-the-art methods on the PCSRD and ISTD+ [Le and Samaras, 2019] datasets. PACSRNet[†] denotes the model only uses the shadow removal module. The best and second best results are **boldfaced** and underlined.

4 Experiment

4.1 Experiment Setups

Dataset Generation. Existing shadow removal datasets [Qu *et al.*, 2017; Wang *et al.*, 2018; Le and Samaras, 2019] are not suitable for prompt-aware controllable shadow removal task. On the one hand, these datasets typically include only shadow images, shadow-free images, and shadow masks, lacking diverse user prompts such as a dot, line, and subject mask. On the other hand, the image scenes in these datasets typically contain only one shadow region, failing to simulate complex real-world scenes.

In this paper, we customize the prompt-based controllable shadow removal dataset, named PCSRD. Specifically, we first sample shadow-free images, shadow images, and shadow masks with multiple subject scenes from DESOBAv2 [Liu *et al.*, 2024b] dataset, and collect the corresponding subject mask as one of the user prompts. Subsequently, to obtain diverse user prompts, we employ the subject mask to automatically generate the corresponding dot and line prompts by a dynamic programming strategy. In this way, customized PCSRD dataset can effectively simulate complex real-world scenes with multiple subjects and diverse prompts.

The final dataset PACSRD¹ consists of 11,900 complex scenes samples with resolution of 256×256 , which are randomly divided into 10,000 for training, 1,000 for validation, and 900 for testing. To ensure the validity of the results, we carefully prevented any data leakage between the training and testing sets. To the best of our knowledge, this is the first dataset for prompt-aware controllable shadow removal task.

Testing Dataset. In addition to our customized PCSRD dataset, we also introduce the ISTD+ [Le and Samaras, 2019] dataset to validate the effectiveness of our shadow removal module. ISTD+ [Le and Samaras, 2019] dataset includes

1,330 training and 540 testing triplets. Each triplet consists of a shadow image, shadow-free image, and shadow region mask. The ISTD+ dataset is obtained by reducing the illumination inconsistency between the shadow and shadow-free image of ISTD [Wang *et al.*, 2018] dataset.

Baselines. To the best of our knowledge, there is no existing work focusing on prompt-based controllable shadow removal task. Therefore, we use seven state-of-the-art shadow removal methods as our baselines to evaluate the performance of our PACSRNet, including BMNet [Zhu *et al.*, 2022], ShadowFormer [Guo *et al.*, 2023a], ShadowDiffusion [Guo *et al.*, 2023b], Inpaint4Shadow [Li *et al.*, 2023], RASM [Liu *et al.*, 2024b], HomoFormer [Xiao *et al.*, 2024] and OmniSR [Xu *et al.*, 2025]. To ensure the comparability of experimental results, these baseline methods used for comparison are fine-tuned on our PCSRD dataset through their released codes.

Evaluation Metrics. Following previous works [Wang *et al.*, 2018; Qu *et al.*, 2017; Le and Samaras, 2019], we employ the peak signal-to-noise ratio (PSNR) [Wu *et al.*, 2023a], structural similarity (SSIM) [Zhang *et al.*, 2023], and root mean square error (RMSE) [Wu *et al.*, 2021] as quantitative evaluation metrics. Specifically, PSNR and SSIM are two popular metrics to evaluate image fidelity of shadow removal results. The higher value ($\uparrow$) of PSNR and SSIM metrics indicates better performance. RMSE measures the mean error between the pixels of the results and the ground truth. The lower value ($\downarrow$) of RMSE indicates better shadow removal. Furthermore, to perform detailed analysis of the removal effects, we calculate all the above metrics for shadow regions, non-shadow regions, and all regions, comparing the ground truth with the generated results for each region separately.

4.2 Experimental Results and Analysis

Quantitative Results. The quantitative results are reported in Tab. 1. As shown in the table, our method achieves performance comparable to ShadowDiffusion [Guo *et al.*, 2023b]

¹Dataset: https://drive.google.com/drive/folders/1hAQJ4pfpC-1m77vpGihdNZJlebrxVTyf?usp=drive_link

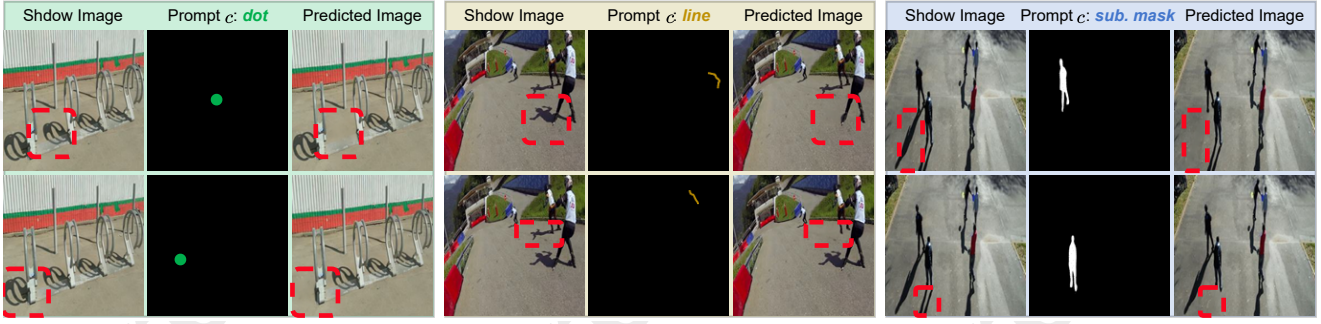


Figure 3: Examples of shadow removal results based on dot, line and subject mask prompts. In the same image, we use the prompt to specify the different subjects and perform corresponding controllable shadow removal.

Method	PSNR $\uparrow$	SSIM $\uparrow$	RMSE $\downarrow$
PACSRNet w/o SFI	42.265	0.9888	0.6330
PACSRNet w/o LDSA	42.384	0.9890	0.6075
PACSRNet w/o Guidance	41.215	0.9881	0.6262
PACSRNet (Full)	42.494	0.9892	0.6038

Table 2: Ablation results for the key blocks of PACSRNet.

in terms of PSNR across the entire image under three different user prompts. Specifically, the PSNR of our method is 40.95 dB, 41.06 dB, and 41.59 dB under point, line, and subject mask prompts, respectively. These results outperform the PSNR of ShadowDiffusion (40.82 dB) [Guo *et al.*, 2023b], which requires precise shadow region masks as priors. This demonstrates the effectiveness of the proposed prompt-aware controllable shadow removal framework.

Qualitative Evaluation. To visually validate the effectiveness of the proposed PACSRNet, we present the shadow removal results of PACSRNet under three different prompts in Fig. 3. Specifically, the left of Fig. 3 shows an example of shadow removal based on a dot prompt, the middle displays the shadow removal case based on a line prompt, and the right demonstrates the shadow removal case based on a subject mask prompt. From Fig. 3, we can clearly observe that the proposed PACSRNet is capable of effectively removing the shadows of the user-specified subject under all three types of prompts. These results highlight the flexibility and robustness of PACSRNet in effectively handling a wide range of user-specified prompts, demonstrating its ability to perform high-quality shadow removal across different scenarios. In addition, we can see in Fig. 3 that PACSRNet can remove shadows of different subjects using the dot, line, and subject mask prompts. For instance, in the example of the point prompt, PACSRNet effectively removes the shadows from different subjects by processing prompt placed at various locations across the image. This flexibility further emphasizes the robustness of PACSRNet in controllable shadow removal.

4.3 Ablation Study

Effectiveness of Spatial-Frequency Interaction (SFI). We conduct ablation study to verify the effectiveness of SFI block. Specifically, we compare the full model with SFI block and the model without SFI block in Tab. 2. From the results, we can observe that SFI block significantly improve the performance of the model. Its PSNR value increases by 0.2326

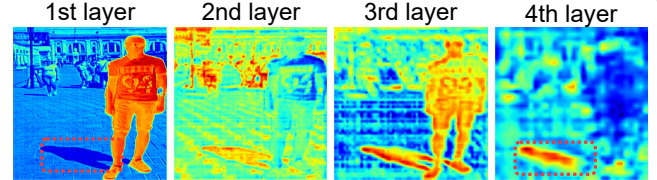


Figure 4: Visualization of prompt-aware guidance feature maps. It implicitly guides the shadow removal module to focus on shadow regions marked in red dashed box.

Method	IoU $\uparrow$	BCE $\downarrow$
PACSRNet - dot	0.816	0.018
PACSRNet - line	0.825	0.016
PACSRNet - subject mask	0.872	0.013

Table 3: Evaluation of shadow masks predicted using different types of prompts in PACSRNet.

dB. These results illustrate that frequency features are beneficial for the prediction and restoration of shadow regions.

Effectiveness of Dense-Sparse Local Attention (DSLAs). In Tab. 2, we conduct an ablation study to validate the effectiveness of DSLAs. From the table, we can observe that the standard dense attention model without sparse branch achieves the worst removal performance due to the introduction of irrelevant content. Its PSNR is improved by 0.11 dB compared to the dense-sparse local attention model with sparse attention branch. These results demonstrate that using sparse attention branch to mitigate the negative impact of irrelevant content on the attention computation is effective.

Effectiveness of Prompt-Aware Guidance Strategy. As mentioned in the section 3.3, we connect the multi-scale features from the prompt-aware module to the encoder of the shadow removal module, implicitly guiding its feature learning. To verify the effectiveness of the prompt-aware guidance strategy, we compare the full model with those without the prompt-aware guidance strategy. As shown in Tab. 2, the full model using the prompt-aware guidance strategy achieves better shadow removal performance. Furthermore, we visualize the multi-scale feature maps in the prompt-aware module used to guide the feature learning of the shadow removal module in Fig. 4. As shown in Fig. 4, we can see that the features at different scales guide different aspects of the shadow removal module. For example, low-level features focus on the subject, while high-level features emphasize the shadow

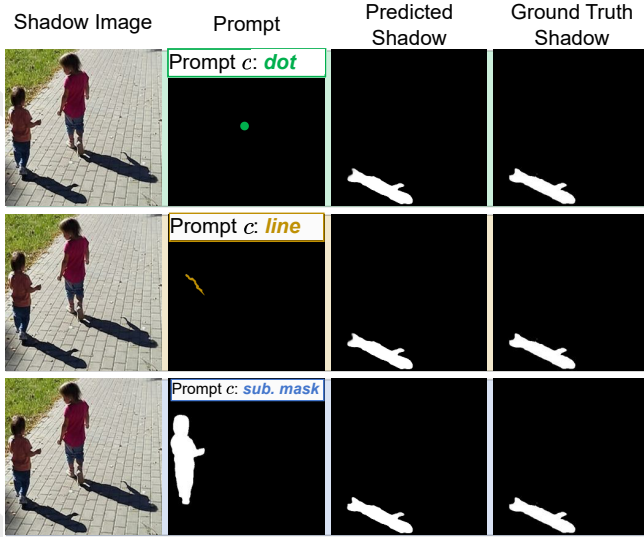


Figure 5: Visual comparison of the predicted shadow region masks for one subject under three different prompts. Our PACSRNet is robust to different types of user prompts.

regions related to the subject. These results verify the necessity and effectiveness of the prompt-aware guidance strategy. **Robustness to User Prompts.** In our design, the proposed PACSRNet supports various user prompts, such as dots, lines, and subject masks. Here, we verify its robustness to user prompts. As shown in Tab. 3, we compare the shadow region masks predicted by PACSRNet under three different prompts. From the table, we can observe that the three different prompts yield similar intersection over union (IoU) and binary cross entropy (BCE) values. In addition, we also present the shadow region prediction results of PACSRNet under three different user prompts. As shown in Fig. 5, PACSRNet obtains shadow region masks consistent with ground truth under three different user prompts. These results further demonstrate that PACSRNet is robust to user prompts.

Effectiveness of Prompt-Aware Module. In Fig. 6, we visualize the shadow region prediction results of the prompt-aware module under three different prompts. From the figure, we can see that the prompt-aware module can predict the shadow mask of the user-specified subject among multiple subjects under three different prompts. For example, in the first case, the prompt-aware module predicts the shadow regions of the schoolbag based on the given dot prompt. These results demonstrate the effectiveness and superiority of the designed prompt-aware module in shadow prediction.

Effectiveness of Shadow Removal Module. To further verify the effectiveness of the shadow removal module, we compare it with five baselines on two datasets (PCSRD and ISTD+ [Le and Samaras, 2019]). As shown in Tab. 1, our shadow removal module achieves superior performance on both PCSR and ISTD+ [Le and Samaras, 2019] datasets. In particular, for the shadow regions, our shadow removal module outperforms the five baselines by a large margin on both the PCSR and ISTD+ [Le and Samaras, 2019] datasets. The specific earnings of it are 0.116 dB and 0.63 dB in shadow regions on both PCSR and ISTD+ [Le and Samaras, 2019]

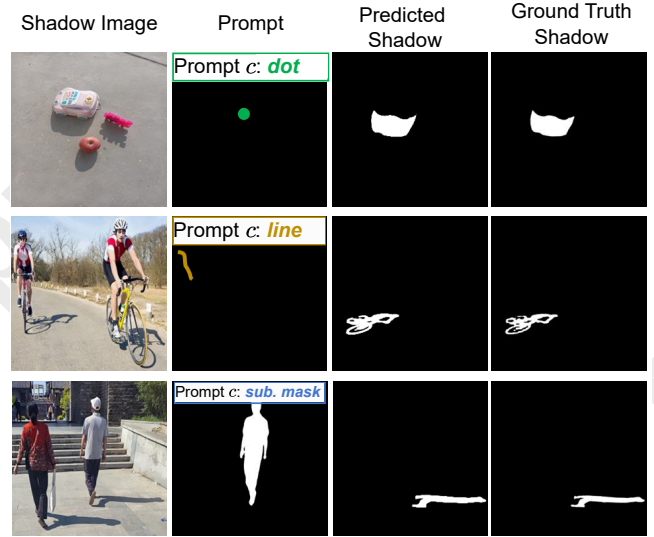


Figure 6: Visual examples of our PACSRNet predicting shadow regions under three different prompts.

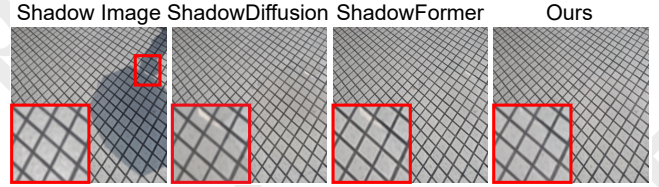


Figure 7: Example of shadow removal results on the ISTD+ [Le and Samaras, 2019] dataset. The input shadow image, the estimated results of ShadowDiffusion [Guo *et al.*, 2023b], ShadowFormer [Guo *et al.*, 2023a], and ours, respectively. The slice of the shadow image corresponds to the ground truth shadow-free image.

datasets, respectively. Furthermore, Fig. 7 presents an example comparing our shadow removal module with two competitive baselines: ShadowFormer [Guo *et al.*, 2023a] and ShadowDiffusion [Guo *et al.*, 2023b]. As can be observed, the shadow regions restored by our shadow removal module are more plausible. These results demonstrate the superiority of our shadow removal module in shadow removal.

5 Conclusion

This paper develops a prompt-aware controllable shadow removal network, which consists of two key components: a prompt-aware module and a shadow removal module. The former aims to generate the shadow masks for a specified subject based on user prompts (*e.g.*, dot, line, and subject mask), while the latter aggregates relevant content by a dense-sparse local attention block to restore the shadow regions predicted by the former. Such a design not only achieves controllable shadow removal, but also avoids the non-trivial and tedious annotations of shadow region. Furthermore, we introduce the first dataset suitable for the prompt-aware controllable shadow removal task, which can effectively facilitate subsequent research. Extensive experimental results demonstrate the superiority and flexibility of our method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62472381), the Natural Science Foundation of Zhejiang Province (LDT23F02023F02) and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

- [Alaluf *et al.*, 2022] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proc. CVPR*, pages 18511–18521, 2022.
- [Chen *et al.*, 2020] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *Proc. CVPR*, pages 5611–5620, 2020.
- [Du *et al.*, 2022] Hang Du, Hailin Shi, Dan Zeng, Xiao-Ping Zhang, and Tao Mei. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM CSUR*, 54(10s):1–42, 2022.
- [Finlayson *et al.*, 2005] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE TPAMI*, 28(1):59–68, 2005.
- [Finlayson *et al.*, 2009] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *IJCV*, 85(1):35–57, 2009.
- [Guo *et al.*, 2023a] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023.
- [Guo *et al.*, 2023b] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proc. CVPR*, pages 14049–14058, 2023.
- [Hu *et al.*, 2019] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE TPAMI*, 42(11):2795–2808, 2019.
- [Huang *et al.*, 2011] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *Proc. ICCV*, pages 898–905. IEEE, 2011.
- [Khan *et al.*, 2014] Salman Hameed Khan, Mohammed Benamoun, Ferdous Sohel, and Roberto Togneri. Automatic feature learning for robust shadow detection. In *Proc. CVPR*, pages 1939–1946. IEEE, 2014.
- [Kim *et al.*, 2022] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proc. CVPR*, pages 18750–18759, 2022.
- [Le and Samaras, 2019] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proc. ICCV*, pages 8578–8587, 2019.
- [Le *et al.*, 2018] Hieu Le, Tomas F Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. Addnet: Training a shadow detector with adversarial shadow attenuation. In *Proc. ECCV*, pages 662–678, 2018.
- [Li *et al.*, 2023] Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor Tsang, and Song Wang. Leveraging inpainting for single-image shadow removal. In *Proc. ICCV*, pages 13055–13064, 2023.
- [Li *et al.*, 2024a] Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. Mmad: Multi-label micro-action detection in videos. *arXiv preprint arXiv:2407.05311*, 2024.
- [Li *et al.*, 2024b] Xinjie Li, Yang Zhao, Dong Wang, Yuan Chen, Li Cao, and Xiaoping Liu. Controlling the latent diffusion model for generative image shadow removal via residual generation. *arXiv preprint arXiv:2412.02322*, 2024.
- [Li *et al.*, 2025a] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang. Prototypical calibrating ambiguous samples for micro-action recognition. In *Proc. AAAI*, pages 4815–4823, 2025.
- [Li *et al.*, 2025b] Kun Li, Xinge Peng, Dan Guo, Xun Yang, and Meng Wang. Repetitive action counting with hybrid temporal relation modeling. *IEEE TMM*, 2025.
- [Liao *et al.*, 2021] Jingwei Liao, Yanli Liu, Guanyu Xing, Housheng Wei, Jueyu Chen, and Songhua Xu. Shadow detection via predicting the confidence maps of shadow detection methods. In *Proc. ACM MM*, pages 704–712, 2021.
- [Liu *et al.*, 2024a] Hengxing Liu, Mingjia Li, and Xiaojie Guo. Regional attention for shadow removal. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5949–5957, 2024.
- [Liu *et al.*, 2024b] Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. Shadow generation for composite image using diffusion model. In *Proc. CVPR*, pages 8121–8130, 2024.
- [Liu *et al.*, 2024c] Yuhao Liu, Zhanhan Ke, Ke Xu, Fang Liu, Zhenwei Wang, and Rynson WH Lau. Recasting regional lighting for shadow removal. In *Proc. AAAI*, volume 38, pages 3810–3818, 2024.
- [Luo *et al.*, 2025] Jinting Luo, Ru Li, Chengzhi Jiang, Xiaoming Zhang, Mingyan Han, Ting Jiang, Haoqiang Fan, and Shuaicheng Liu. Diff-shadow: Global-guided diffusion model for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5856–5864, 2025.
- [Mallat, 1989] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE TPAMI*, 11(7):674–693, 1989.
- [Mei *et al.*, 2024] Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, and Vishal M Patel. Latent feature-guided diffusion models for shadow removal. In *Proc. WACV*, pages 4313–4322, 2024.

- [Qu *et al.*, 2017] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proc. CVPR*, pages 4067–4075, 2017.
- [Shen *et al.*, 2015] Li Shen, Teck Wee Chua, and Karianto Leman. Shadow optimization from structured deep edge detection. In *Proc. CVPR*, pages 2067–2074, 2015.
- [Shor and Lischinski, 2008] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum*, volume 27, pages 577–586. Wiley Online Library, 2008.
- [Vasluianu *et al.*, 2024] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, Wei Dong, Han Zhou, Yuqiong Tian, Jun Chen, et al. Ntire 2024 image shadow removal challenge report. In *Proc. CVPR*, pages 6547–6570, 2024.
- [Vicente *et al.*, 2015] Tomás F Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection. In *Proc. ICCV*, pages 3388–3396, 2015.
- [Vicente *et al.*, 2017] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE TPAMI*, 40(3):682–695, 2017.
- [Wang *et al.*, 2018] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proc. CVPR*, pages 1788–1797, 2018.
- [Wang *et al.*, 2020] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *Proc. CVPR*, pages 1880–1889, 2020.
- [Wang *et al.*, 2021] Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, and Pheng-Ann Heng. Single-stage instance shadow detection with bidirectional relation learning. In *Proc. CVPR*, pages 1–11, 2021.
- [Wang *et al.*, 2022] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *IEEE TPAMI*, 45(3):3259–3273, 2022.
- [Wang *et al.*, 2024a] Fei Wang, Dan Guo, Kun Li, Zhun Zhong, and Meng Wang. Frequency decoupling for motion magnification via multi-level isomorphic architecture. In *Proc. CVPR*, pages 18984–18994, 2024.
- [Wang *et al.*, 2024b] Jianan Wang, Zhiliang Wu, Hanyu Xuan, and Yan Yan. Text-video completion networks with motion compensation and attention aggregation. In *Proc. ICASSP*, pages 2990–2994, 2024.
- [Wang *et al.*, 2024c] Xinrui Wang, Lanqing Guo, Xiyu Wang, Siyu Huang, and Bihan Wen. Softshadow: Leveraging penumbra-aware soft masks for shadow removal. *arXiv preprint arXiv:2409.07041*, 2024.
- [Wang *et al.*, 2025] Fei Wang, Kun Li, Yiqi Nie, Zhangling Duan, Peng Zou, Zhiliang Wu, Yuwei Wang, and Yanyan Wei. Exploiting ensemble learning for cross-view isolated sign language recognition. *arXiv preprint arXiv:2502.02196*, 2025.
- [Wu *et al.*, 2021] Zhiliang Wu, Kang Zhang, Hanyu Xuan, Jian Yang, and Yan Yan. Dapc-net: Deformable alignment and pyramid context completion networks for video inpainting. *SPL*, 28:1145–1149, 2021.
- [Wu *et al.*, 2023a] Zhiliang Wu, Changchang Sun, Hanyu Xuan, and Yan Yan. Deep stereo video inpainting. In *Proc. CVPR*, pages 5693–5702, 2023.
- [Wu *et al.*, 2023b] Zhiliang Wu, Hanyu Xuan, Changchang Sun, Weili Guan, Kang Zhang, and Yan Yan. Semi-supervised video inpainting with cycle consistency constraints. In *Proc. CVPR*, pages 22586–22595, 2023.
- [Wu *et al.*, 2024] Zhiliang Wu, Changchang Sun, Hanyu Xuan, Gaowen Liu, and Yan Yan. Waveformer: Wavelet transformer for noise-robust video inpainting. In *Proc. AAAI*, volume 38, pages 6180–6188, 2024.
- [Wu *et al.*, 2025] Zhiliang Wu, Kerui Chen, Kun Li, Hehe Fan, and Yi Yang. BVINet: Unlocking blind video inpainting with zero annotations. *arXiv preprint arXiv:2502.01181*, 2025.
- [Xiao *et al.*, 2024] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. Homoformer: Homogenized transformer for image shadow removal. In *Proc. CVPR*, pages 25617–25626, 2024.
- [Xu *et al.*, 2025] Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. Omnir: Shadow removal under direct and indirect lighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8887–8895, 2025.
- [Yang *et al.*, 2012] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE TIP*, 21(10):4361–4368, 2012.
- [Yang *et al.*, 2023] Han Yang, Tianyu Wang, Xiaowei Hu, and Chi-Wing Fu. Silt: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In *Proc. ICCV*, pages 12687–12698, 2023.
- [Zhang *et al.*, 2023] Yanni Zhang, Zhiliang Wu, and Yan Yan. Pfta-net: Progressive feature alignment and temporal attention fusion networks for video inpainting. In *Proc. ICIP*, pages 191–195, 2023.
- [Zhang *et al.*, 2024] Zixuan Zhang, Xinge Guo, and Chengkuo Lee. Advances in olfactory augmented virtual reality towards future metaverse applications. *NC*, 15(1):6465, 2024.
- [Zhou *et al.*, 2024] Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proc. CVPR*, pages 2952–2963, 2024.
- [Zhu *et al.*, 2022] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proc. CVPR*, pages 5627–5636, 2022.