PAPP: Plug-And-Play Prompts for Shadow Removal

Anonymous ICCV submission

Paper ID 8537

Abstract

Advancements in deep learning have significantly improved shadow removal techniques, yet existing models often struggle to fully leverage semantic information, which is crucial for accurate shadow removal and achieving object-wise consistent results. In this work, we propose a novel approach that extends conventional shadow removal networks with minimal modifications by integrating semantic maps from off-the-shelf segmentation networks as prompts. Utilizing prompt-based learning strategies, we introduce a Plug-and-Play Prompts module (PAPP), which can be seamlessly integrated into various existing shadow removal models to significantly enhance shadow removal performance through the use of semantic-aware prompts in the latent space. Our contributions include the use of semantic information as a prompt in shadow removal, the development of innovative prompt-based learning strategies for generating and exploring semantic-aware prompts in the multi-scale latent space, and substantial advancements in shadow removal efficacy, pushing the state-of-the-art forward.

1. Introduction

Shadows can obscure crucial visual cues, thereby reducing the accuracy of various computer vision tasks, such as object detection, segmentation, and recognition. Consequently, numerous studies have focused on developing methods to remove shadows from images and accurately restore the natural appearance of shadowed areas, ensuring they blend seamlessly with surrounding non-shadowed regions. The advent of deep learning has revolutionized shadow removal approaches, resulting in the development of several sophisticated neural network architectures. Specifically, the introduction of convolutional neural network (CNN)-based approaches [3, 19, 24], generative adversarial network (GAN)-based approaches [5, 13, 35], transformer-based approaches [7, 38], and, more recently, diffusion-based models [8, 9, 16] have led to significant advances. By harnessing the power of deep learning to understand and reconstruct complex image features, these networks have played a critical role in advancing the capabilities of shadow removal systems.

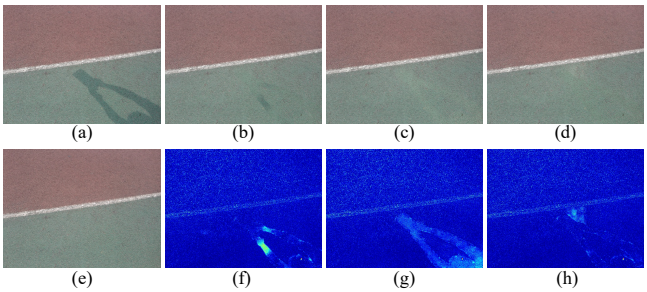


Figure 1. Visual comparison of input prompts evaluated on AISTD. (a) Input shadow image. (b) Result by UFormer. (c) Result by UFormer with additional input mask. (d) Result by UFormer with additional input semantic map and mask. (e) Ground truth image. (f)-(h) are the corresponding error maps for (b)-(d), respectively, with reference to (e).

Research has focused not only on advancing neural network architectures but also on utilizing prior knowledge and other additional inputs. Traditionally, the binary mask has been the most commonly used additional information in shadow removal [5, 7–9, 13, 19, 32]. Furthermore, Le *et al*. [19] introduce physics-based prior based in their novel shadow illumination model, and Retinex theory [18] has also recently been applied as a method to obtain more accurate color maps [45] by decomposing images into reflectance and illumination components [8, 9].

Semantic information are also explored in various way. Semantic information plays a crucial role in shadow removal as it provide insights into the relationship between shadow and non-shadow regions. However, directly applying externally obtained semantic information to the model is challenging. To seamlessly inject semantic information into the baseline model and generate semantically aware feature representations of the input, we use prompt-based learning [2], a technique widely used in NLP [2, 25] to optimize models with contextual information for desired outcomes. The effectiveness of prompt learning has also been demonstrated in several visual tasks that require the integration

of additional information. For example, the degradation-specific prompt for all-in-one image restoration [23], polarization prompt for depth enhancement, and depth estimation prompt for dehazing [33] have been successfully applied. In this paper, we propose utilizing semantic cues as prompts for shadow removal. Proper semantic information helps to better understand shadow images and aids in more accurate shadow removal. Prompt learning allows these semantic details to be transformed into more useful latent prompts, even compensating for suboptimal semantic information.

Also, we propose using semantic maps as additional prompts to restore shadowed regions with semantic information that exhibits shadow-invariant properties. First, to explore how semantic maps aid in shadow removal, we perform preliminary experiments by simply concatenating a binary mask and/or a semantic map obtained from Mask2Former [4] with a given shadow image as input prompts. Figure 1 shows that incorporating these prompts can significantly improve the performance of the shadow removal model. In particular, the binary mask, widely used in the shadow removal task, helps identify the shadow area and significantly enhances their restoration performance, as shown in Figure 1 (c). However, there are still color differences, as shown in Figure 1 (g), due to the lack of contextual information indicating similar areas in the non-shadow region. Alternatively, as shown in Figure in 1 (d), the use of a semantic map can alleviate this issue. The semantic map aids in detecting non-shadowed areas with content similar to the shadowed regions, facilitating the retrieval of appropriate colors. This reduces color disparity by harmonizing the shadowed region with the non-shadowed area based on contextual information, as demonstrated in Figure 1 (h). These observations highlight the advantage of using semantic maps in shadow removal.

The goal of this paper is to further improve existing shadow removal models by leveraging prompt learning while using semantic maps as well as conventional binary masks, as prompts. To effectively exploit the prompts, we propose a Plug-And-Play Prompts (PAPP) for shadow removal consisting of the Prompt Projection Block (PPB), the Prompt Fusion Block (PFB) and the Prompt Interactive Cross-Attention Block (PICAB), which integrates semantic prompts in the latent space to enrich the feature representation of baseline networks.

In particular, our PAPP can be integrated into many conventional shadow removal networks without significant modification. By exploiting semantic cues in latent space across multiple scales, our approach can significantly improve the performance of existing networks. We perform extensive experiments and ablation studies on the AISTD [19], SRD [24], and WSRD+ [31] datasets, which demonstrate the superiority and effectiveness of our proposed method and show state-of-the-art shadow removal results. Our contributions can be summarized as follows:

- We propose a novel method that leverages semantic maps and binary masks in an integrated manner through prompt learning to effectively exploit semantic information for shadow removal.
- We introduce a novel PAPP consisting of effective PPB, PFB and PICAB, which incorporates multiple prompt features and improves baseline feature representation.
- By performing various experiments on benchmark datasets and different types of baseline networks, including the state-of-the-art baseline, we demonstrate the plug-and-play extension of baseline networks.

2. Related works

2.1. Prompt-based Learning

Prompt-based methods have emerged as a powerful paradigm in the field of natural language processing (NLP) [2, 25], and have been developed to provide more specific contextual information to improve the performance of large-scale language models [27, 28].

Useful prompt learning techniques, proven in the field of NLP, are currently gaining attractions in various computer vision tasks [1, 14] or vision language task [43]. For instance, MAE-VQGAN [1] attempts to provide explicit additional input in the form of example prompts to adapt to various downstream tasks such as inpainting, segmentation, and so on, by training its model with these prompts. Especially in the field of image restoration, this approach is mainly used to solve multi-task image restoration. PromptIR [23] successfully exploits degradation-specific prompting, and PromptRestorer [34] uses an external pre-trained network as a prompt to capture perceptually informative features. This strategy allows models to tailor their tasks to the input image, effectively fuse the additional information with the image restoration tasks, and achieve the desired restoration results. In addition, prompt-based approaches allow for seamless integration with existing image restoration frameworks, facilitating the adoption of state-of-the-art approaches and methodologies.

In this work, we aim to elevate the performance of existing shadow removal architectures by using semantic prompts, while identifying relevant information to serve as a prompt. We introduce PAPP, a plug-and-play module that effectively fuses different semantic prompts and can be easily integrated into conventional and state-of-the-art shadow removal networks to boost their performance.

2.2. Single Image Shadow Removal

Removing shadows from images has become more manageable with the advent of deep learning techniques and large datasets [19, 24, 31, 35]. These various deep learning based approaches [3, 7, 8, 13, 19, 24, 40] for shadow

removal attempt to utilize various additional information, including binary mask, semantic information, illumination [9, 19, 21, 40], reflectance [9], and contextual information [3].

ShadowDiffusion [8] and ShadowFormer [7] incorporate binary masks in a semantically meaningful manner within their models. SP+M-Net [19] implement a linear illumination transformation to predict shadow parameters and shadow matte. RIS-GAN [40] utilize an inverse illumination map and negative residual image to estimate the shadow-free image. BCDiff [9] predict shadow-free images in an unsupervised manner, employing a decomposition network to separate the image into reflectance and illumination maps. CANet [3] transfer contextual information from non-shadow regions to shadow regions using a patch matching module. However, these methods rely solely on the decomposition of illumination maps or fail to ensure the overall harmony of the image.

Hu et al.[11] utilize SAM [17] to extract material segmentation information, seamlessly integrating it into test-time adaptation. Additionally, DeshadowNet [24] attempts to leverage semantic information from the VGG16 [29] network, which is pre-trained for image classification. However, since neither VGG16 nor SAM is specifically designed to handle shadow invariance, the extracted semantic information may hinder effective information sharing between shadow and non-shadow regions, particularly for objects with similar characteristics. Furthermore, utilizing SAM [17] to extract semantic information while preserving shadow invariance requires additional training. Moreover, feature maps from VGG16 networks may lack pixel-level granularity, limiting their ability to provide detailed semantic information. In contrast, our method of integrating semantic maps from an off-the shelf model into the shadow removal task offers insights into the relationship between shadow and non-shadow regions based on the positions of objects. Moreover, this enables us to provide consistent semantic information regardless of the presence of shadows. By utilizing a pre-trained network for pixel-wise segmentation, we can offer more accurate semantic information to effectively remove shadows.

3. Proposed Methods

We propose a novel method to exploit semantic information through prompt learning. To achieve this, we first obtain a semantic segmentation dataset suitable for training a shadow removal network. Specifically, we utilize an off-the-shelf image semantic segmentation model [4] trained on the ADE20k dataset [42], which contains a wide range of scenes, and apply it to conventional shadow removal datasets such as AISTD, SRD, and WSRD+ datasets to generate semantic segmentation maps. The semantic segmentation map and the given binary mask are then used as a guide

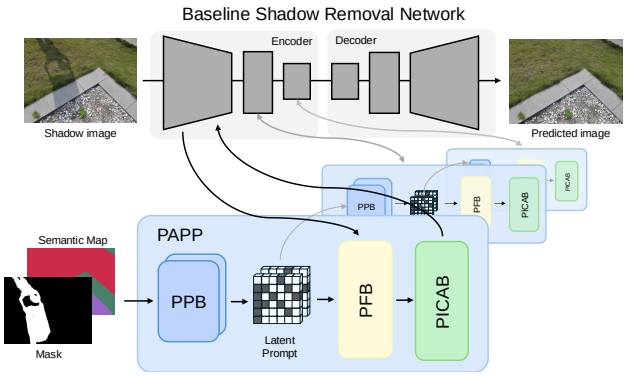


Figure 2. Illustration of the overall process of our method.

to refine shadow regions and produce results that seamlessly integrate with surrounding non-shadow areas. To integrate these prompts within conventional shadow removal networks using a plug-and-play scheme, we propose the Plug-And-Play Prompts (PAPP) module, which includes the Prompt Projection Block (PPB), the Prompt Fusion Block (PFB), and the Prompt Interactive Cross-Attention Block (PICAB). In this section, we present the detailed network architectures for each component.

3.1. Overall Pipeline

In Figure 2, we provide an overview of our framework, which integrates baseline shadow removal architectures with our proposed prompt module. In particular, we present PAPP as an ad hoc, plug-and-play prompt module that is model-agnostic and broadly applicable to various conventional shadow removal networks without requiring significant modifications of the baseline networks.

In general, conventional shadow removal models consist of a U-shaped architecture with encoder and decoder parts. Our PAPP integrates semantic prompt information with baseline features from the encoder to produce semantic-aware features without modifying the baseline decoder. In PAPP, we first extract latent prompts separately through PPB from the semantic segmentation map and the input binary mask, which are constructed using simple but effective convolutional layers. Then, the extracted semantic prompts are fused with baseline feature maps through PFB and PICAB in the latent space across multiple scales, producing a richer and more semantic-specific feature representation.

3.2. Plug-And-Play Prompt Module (PAPP)

To seamlessly integrate prompt features with those of baseline shadow removal networks, we propose the PAPP module. Our PAPP is positioned in the middle of the encoder in the baseline architecture, as shown in Figure 2, allowing the multi-scale semantic prompt to persist and interact with the baseline feature maps. The detailed key components of our PAPP including the Prompt Projection Block (PPB), the Prompt Fusion Block (PFB), and the Prompt Interactive

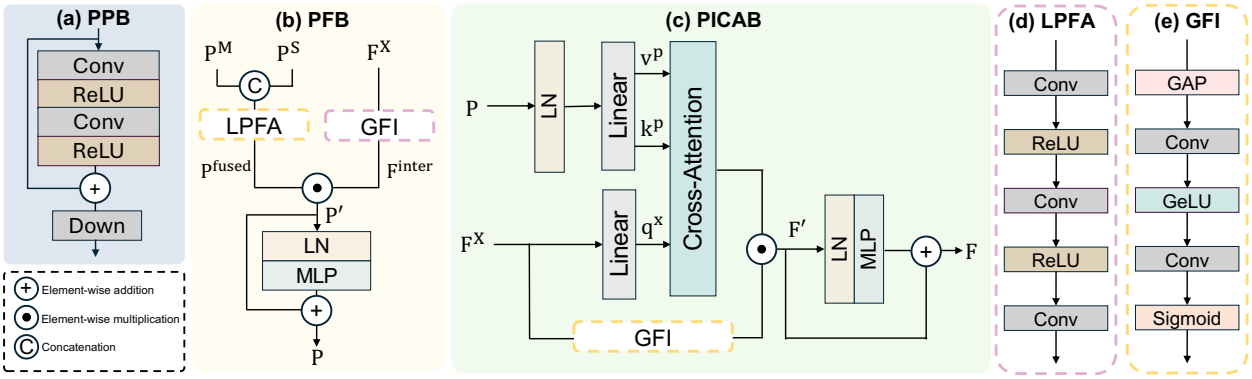


Figure 3. Detailed PAPP architecture. (a) Prompt Projection Block (PPB). (b) Prompt Fusion Block (PFB). (c) Prompt Interactive Cross-Attention Block (PICAB). (d) Local Prompt Feature Aggregation (LPFA). (e) Global Feature Interaction (GFI).

Cross-Attention Block (PICAB), are shown in Figure 3.

Prompt Projection Block (PPB). To generate a semantic-specific prompts in the latent space, we input the binary mask (commonly used as input in shadow removal task) along with an additional semantic segmentation map as prompts.

Our PAPP module then leverages these prompts in the latent space through a multi-scale approach. Specifically, in the Prompt Projection Block (PPB) at the first scale (i.e., level 1), the binary mask P_0^M and semantic segmentation map P_0^S are projected onto the latent space, generating the latent prompts P_1^M and P_1^S, respectively. For this projection, we employ two consecutive convolutional layers followed by ReLU activations, including skip connections, and subsequent $2\times$ down-sampling operation, as illustrated in Figure 3(a).

At the next scale (i.e., level 2), the latent prompts generated in the previous scale are further processed by the PPB. Consequently, the l-level prompt features can be expressed as follows:

$$\begin{cases} P_l^S = PPB(P_{l-1}^S; \theta^S), \\ P_l^M = PPB(P_{l-1}^M; \theta^M), \quad l = 1, 2, 3. \end{cases} \tag{1}$$

Note that the semantic prompt and the mask prompt are projected separately at each scale onto P_l^S and P_l^M, where θ^S denotes the projection parameters for the semantic prompt and θ^M denotes the projection parameters for the mask prompt. Through these projections, semantic information and the presence of shadows are embedded in the latent space, and the latent prompts are then passed as input to the next stage, the Prompt Fusion Block (PFB).

Prompt Fusion Block (PFB). Our PFB is designed to integrate the latent prompts P_l^M and P_l^S from the PPB to generate a more enriched semantic information feature P_l. The PFB consists of two key components: local prompt fusion aggregation (LPFA), which locally fuses the prompt features using convolutions, and global feature interaction (GFI), which is inspired by the attention mechanism [10] to globally attend to the fused prompt alongside the baseline feature.

Once the latent prompts P_l^M and P_l^S are passed from PPB, we concatenate them before passing them through the LPFA. As shown in Figure 3(d), the LPFA consists of three convolutional layers and two ReLU activation functions, including channel reduction, to effectively aggregate the local information of the concatenated two latent prompts. This process combines the extracted semantic and mask prompts, resulting in the fused prompt feature map P_l^{fused} at scale level l as follows:

$$P_l^{fused} = LPFA([P_l^M, P_l^S]). \tag{2}$$

Then, to further improve P_l^{fused} by integrating global context information from the baseline shadow removal networks, we construct the GFI to extract global information from the encoder of the baseline models. Specifically, given the baseline features at scale level l, denoted F_l^X, the GFI first projects these features into the interactable feature F_l^{inter}, as shown in Figure 3 (e). Specifically, the GFI includes global average pooling, two convolutional layers, GeLU activation, and a final sigmoid function that processes the baseline feature F_l^X to produce F_l^{inter} as follows:

$$F_l^{inter} = GFI(F_l^X). \tag{3}$$

Finally, we merge the prompts P_l^{fused} with the global context information from the baseline F_l^{inter}, resulting in enriched prompts. To achieve this, we first combine P_l^{fused} and F_l^{inter} as follows:

$$P_l' = P_l^{fused} \odot F_l^{inter}, \tag{4}$$

where $\odot$ denotes the element-wise multiplication. Then, to capture long-range correlation, inspired by [7], we apply layer normalization followed by a multi-layer perceptron (MLP), incorporating a skip connection, to generate the fused feature map P_l from P_l', as shown in Figure 3(b).

To summarize, in our PFB, we produce the enhanced semantic-aware prompt P_l as follows:

$$P_l = PFB(F_l^X, [P_l^M, P_l^S]). \tag{5}$$

Prompt Interactive Cross-Attention Block (PICAB). We present the Prompt Interactive Cross-Attention Block (PICAB) to enhance the baseline features F_l^X by interacting with the semantic-aware prompt feature P_l generated by the PFB.

First, PICAB locally captures the correlation between the baseline feature F_l^X and the prompt feature P_l using a window-based attention mechanism [22]. Specifically, in PICAB, the window-based multi-head self-attention (W-MSA) [37] is modified and employs a cross-attention mechanism (i.e., W-MCA) to achieve spatial correlation between different features, as shown in Figure 3(c). Notably, for the cross-attention, the baseline feature F_l^X is projected into a query vector q_l^x, while the prompt feature P_l is projected into key and value vectors k_l^p and v_l^p, respectively, using learnable parameters $W_{q_l}^F$, $W_{k_l}^P$, and $W_{v_l}^P$, and this process yields:

$$\begin{cases} q_l^x = p_\gamma(F_l^X) \times W_{q_l}^F, \\ k_l^p = p_\gamma(P_l) \times W_{k_l}^P, \\ v_l^p = p_\gamma(P_l) \times W_{v_l}^P, \end{cases} \tag{6}$$

where p_γ represents the γ-size window partition function [37], which transforms the $H \times W \times C$ input feature into $\gamma^2 \times \frac{H}{\gamma} \times \frac{W}{\gamma} \times C$ feature map. Then, the cross-attention mechanism is formulated as follows:

$$\text{W-MCA}(q_l^x, k_l^p, v_l^p) = \sigma\Big(\frac{q_l^x k_l^p}{\sqrt{d}} + B\Big) v_l^p, \tag{7}$$

where d is scaling parameter, B denotes the relative positional encoding [22, 37]. Similar to the approach used in PFB, we apply the GFI mechanism in PICAB to capture global information from the baseline features, addressing the limitations of W-MCA, which, being divided into windows, captures only local information. The GFI compensates for this by enabling the model to access global context. The outputs of W-MCA and GFI are then integrated by the element-wise multiplication, resulting in an extended baseline feature map F_l'. Then, a layer normalization and an MLP are applied to F_l', incorporating a skip connection, to finally produce the semantic-aware enhanced feature F_l. Note that, the enhanced feature F_l from PICAB is then fed into the encoder of the baseline as shown in Figure 2.

4. Experiments

We apply the proposed PAPP to conventional baseline models, including ShadowFormer [7] and HomoFormer [38], where HomoFormer is the current state-of-the-art method for shadow removal. In addition, we also use the diffusion-based SR3 [26] and UFormer [37] as baselines due to their strong performance in various image restoration tasks.

4.1. Implementation Details

We use the official ShadowFormer, HomoFormer, and SR3 network architectures as our baselines. While for UFormer, we built the baseline by reducing the depth of layers from 4 to 3, which are suitable for shadow removal. In addition, for fair comparisons, we also concatenate the binary mask with the input shadow image for the UFormer and SR3 baselines. To measure the performance of the baselines, we use the official parameters when available, otherwise we train the parameters. In particular, we use the official parameters of ShadowFormer and HomoFormer trained on the AISTD dataset. For all experiments, we use Mask2Former-base [4] model to collect semantic segmentation.

We conduct experiments on three benchmark datasets. First, we use the commonly used Adjusted ISTD (AISTD) dataset [19], which contains 1330 training and 540 test triplets, each consisting of shadow images, masks, and shadow-free images. Next, we employ the SRD dataset [24], which consists of 2,680 training pairs and 408 testing pairs of shadow and shadow-free images. Since the SRD dataset does not provide binary masks, we use the predicted masks generated by DHAN [5] during both the training and testing phases, for comparison with other methods [5, 6, 8]. Lastly, to demonstrate the effectiveness of our method on challenging datasets, we conduct experiments on the WSRD+ dataset [31], which contains a wide variety of captured surfaces with diverse colors and textures. The WSRD+ dataset consists of 1200 high-resolution image pairs including 1000 training pairs, 100 validation pairs, and 100 test pairs. We evaluate on the validation pairs because the test ground-truth is unavailable. Due to the lack of binary masks in WSRD+, we construct a PAPP module that uses only the segmentation map as prompt, verifying that PAPP can be applied in various shadow removal settings. Notably, when training the baseline models combined with our PAPP module, we use the same loss function as the original baseline model, without any modifications.

For the evaluation metrics, we utilize the root-mean-square error (RMSE) in the LAB color space to compare ground truth images with the predicted images. Additionally, for a comprehensive comparison, we measure the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) in the RGB color space between the ground truth images and the predicted results. Moreover, we measure the perceptual quality of the output images using LPIPS [41] for WSRD+. More details regarding the experimental settings can be found in the supplementary material.

4.2. Quantitative Results

First, we quantitatively compare various shadow removal models, and the comparison results on the SRD, AISTD, and WSRD+ test sets are summarized in Table 1, Table 2, and Table 3, respectively. In particular, 'baseline name +

Method	Shadow			Non-Shadow			All		
	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Input Image	18.96	0.871	36.69	31.47	0.975	4.83	18.19	0.830	14.05
DHAN [5]	33.67	0.978	8.94	34.79	0.979	4.80	30.51	0.949	5.67
Fu *et al*. [6]	32.26	0.966	9.55	31.87	0.945	5.74	28.40	0.893	6.50
BMNet [45]	35.05	0.981	6.61	36.02	0.982	3.61	31.69	0.956	4.46
ShadowDiffusion [8]	38.72	0.987	4.98	37.78	0.985	3.44	34.73	0.970	3.63
Liu *et al*. [21]	36.51	0.983	5.49	37.71	0.986	3.00	33.48	0.967	3.66
SR3 [26]	39.29	0.991	4.99	37.67	0.989	2.93	34.56	0.974	3.35
SR3 + PAPP	**40.04**	**0.991**	**4.88**	**37.73**	**0.989**	**3.00**	**34.93**	**0.975**	**3.39**
UFormer [37]	36.86	0.986	5.44	37.29	0.990	2.92	33.30	0.971	3.79
UFormer + PAPP	**36.90**	**0.986**	**5.40**	**37.39**	**0.990**	**2.89**	**33.36**	**0.971**	**3.74**
ShadowFormer [7]	36.45	0.986	5.59	37.35	0.990	2.87	33.08	0.970	3.77
ShadowFormer + PAPP	**36.64**	**0.986**	**5.47**	**37.49**	**0.990**	**2.84**	**33.22**	**0.970**	**3.72**
HomoFormer [38]	39.12	0.991	4.30	39.73	0.994	2.44	35.57	0.981	3.07
HomoFormer + PAPP	**39.21**	**0.991**	**4.24**	**39.79**	**0.994**	**2.43**	**35.65**	**0.981**	**3.05**

Table 1. Quatiative results of shadow removal on SRD datasets.

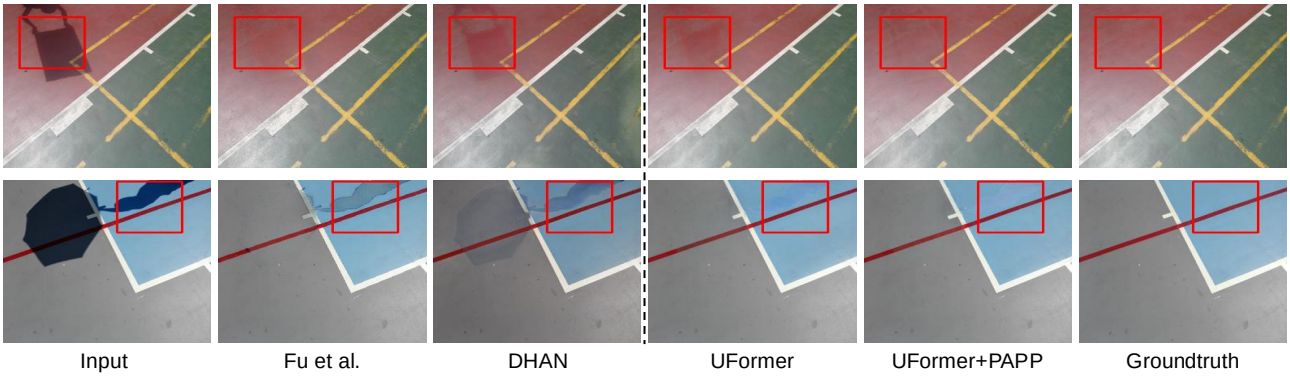


Figure 4. Visual comparison on AISTD datasets. We compare our UFormer+PAPP with Fu *et al*. [6] and DHAN [5].

Method	Shadow PSNR↑ / RMSE↓	Non-Shadow PSNR↑ / RMSE↓	All PSNR↑ / RMSE↓
Input Image	20.83 / 40.2	37.46 / 2.6	20.46 / 8.5
DeshadowNet [24]	- / 15.9	- / 6.0	- / 7.6
DHAN [5]	32.92 / 11.2	27.15 / 7.1	25.66 / 7.8
Fu *et al*. [6]	36.04 / 6.6	31.16 / 3.8	29.45 / 4.2
BMNet [45]	- / 5.6	- / 2.5	- / 3.0
ShadowDiffusion [8]	39.82 / 4.9	38.90 / 2.3	35.72 / 2.7
Liu *et al*. [21]	38.04 / 5.7	39.15 / 2.3	34.96 / 2.9
SR3 [26]	37.72 / 5.98	37.52 / 2.57	33.99 / 3.11
SR3 + PAPP	**37.86 / 5.96**	**38.16 / 2.51**	**34.37 / 3.05**
UFormer [37]	39.37 / 5.32	38.62 / 2.24	35.29 / 2.74
UFormer + PAPP	**39.73 / 4.98**	**38.76 / 2.23**	**35.54 / 2.68**
ShadowFormer [7]	39.21 / 5.32	38.83 / 2.24	35.30 / 2.74
ShadowFormer + PAPP	**39.49 / 5.15**	**39.03 / 2.21**	**35.58 / 2.69**
HomoFormer [38]	39.51 / 4.92	38.66 / 2.27	35.32 / 2.68
HomoFormer + PAPP	**40.45 / 4.69**	**39.19 / 2.18**	**36.06 / 2.57**

Table 2. Quantitative comparison on AISTD datasets.

PAPP′ indicates baseline models equipped with the proposed PAPP.

In the tables, the results of our method are highlighted for clarity. For comparisons with other methods on the SRD and AISTD datasets, we adhere to the experimental settings with a resolution of 256 × 256, as adopted in previous works [5, 7, 46].

Method	Resolution	PSNR↑	SSIM↑	RMSE↓	LPIPS↓
SR3 [26]	256×256	21.13	0.916	12.21	0.059
SR3 + PAPP		**22.51**	**0.927**	**10.64**	**0.050**
UFormer [37]	Full	22.54	0.885	9.82	0.068
UFormer + PAPP		**22.80**	**0.888**	**9.57**	**0.065**
HomoFormer [38]	Full	23.57	0.900	8.71	0.053
HomoFormer + PAPP		**23.75**	**0.902**	**8.59**	**0.052**

Table 3. Quantitative comparison on WSRD+ dataset. Due to lack of resources when training the SR3 model, we use down-scaled images for training and evaluation.

As can be seen from the data presented in each table, our method consistently improves the performance of the baseline models in terms of PSNR, SSIM, and RMSE metrics on all benchmark datasets over different baselines, and LPIPS on the WSRD+ dataset, demonstrating perceptual quality. Specifically, on the AISTD dataset, we observe a significant improvement of 0.74 dB and 0.94 dB in PSNR for the All and Shadow regions, respectively, while achieving state-of-

the-art results when using the HomoFormer baseline. These results demonstrate that our PAPP can be easily integrated into conventional baseline networks, including state-of-the-art models, while consistently improving performance. Notably, our method also showed enhanced performance on the WSRD+ dataset, where the complexity of scenes and the presence of various objects make shadow region restoration particularly challenging. This suggests that the semantic map provides direct cue for restoring shadows in complex scenes and indicates that PAPP can be applied to various datasets even without a binary mask. Please see our supplementary material for more results from various baselines and datasets.

4.3. Qualitative Results

To demonstrate the advantages of our method over others, we provide visual comparisons with other shadow removal approaches. Figure 4 shows visual results on the AISTD dataset. Compared to other models, our results show significantly better reconstruction of shadow regions, with reduced color differences and minimal boundary effects. In particular, compared to the baseline results, the colors of the non-shadow and shadow regions are highly consistent in our method. This suggests that our method effectively predicts the shadow region by leveraging semantic information from the non-shadow regions as shown by UFormer+PAPP. In contrast, other methods that do not simultaneously utilize both semantic and mask information tend to produce results where the shadow and non-shadow regions are not harmoniously integrated as shown other methods in Figure 4.

The effectiveness of PAPP on challenging datasets, characterized by complex shadow interactions and increasingly intricate surfaces, is clearly demonstrated in Figure 6, where it outperforms various baseline models.

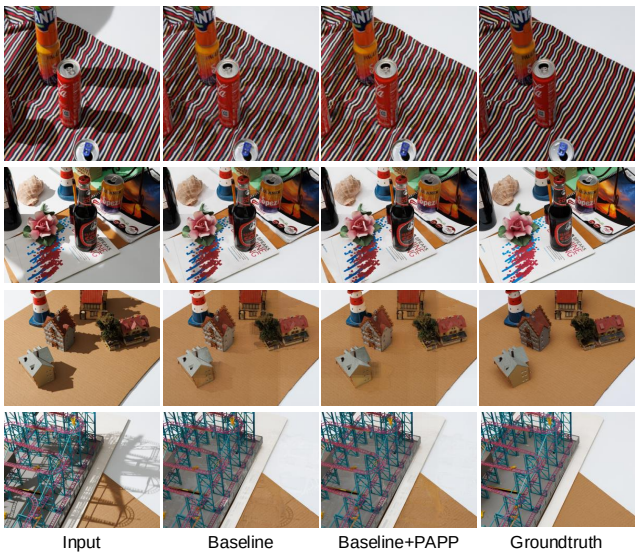


Figure 6. Visual comparison of shadow removal results on WSRD+. The top row shows results based on SR3, the second row presents results based on UFormer, and the bottom two rows display results based on HomoFormer.

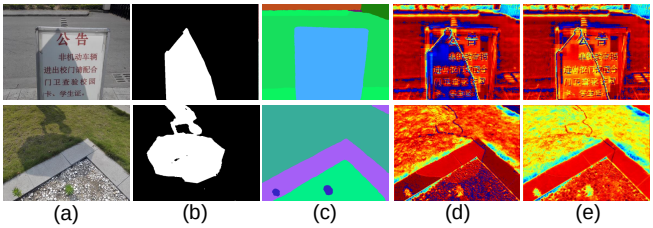


Figure 7. Visual comparison of feature maps with and without PAPP using ShadowFormer. (a) Input shadow image. (b) Input mask. (c) Input semantic map. (d) Feature map F^X without PAPP. (e) Feature map F with PAPP.

4.4. Ablations

For the ablation study, we fix the baseline model as UFormer and conduct experiments on the AISTD dataset, measuring PSNR and RMSE in the shadow region for comparison.

Impact of Prompt Injection. We first discuss how the prompt is injected into the baseline model and how it improves the quality of shadow removal. Note that settings (a) and (b) in Table 6 correspond to conventional models. Comparing (c) and (d) to (a) in Table 6, using any prompt with PAPP can improve the performance. Furthermore, by comparing (b) and (d), we observe that (d) shows an improvement of 0.29 dB in PSNR and 0.22 in RMSE over the baseline (b). This result demonstrates that using mask information as a latent prompt through PAPP enables more accurate shadow removal than using it as a simple concatenated input. By exploiting the semantic prompt and mask prompt in PAPP as in (e), the performance improves by 0.32 dB PSNR and 0.40 RMSE over the baseline (b), demonstrating that our PAPP improves the shadow removal performance. Finally, our full model in (f), which also takes mask information as network input like the baseline model (*i.e.* (b)), achieves the highest PSNR and the second-best RMSE performance.

Impact of PFB and PICAB. Further ablation study on the components of PAPP is performed as shown in Table 4 and Table 5. When the LPFA in PFB is removed, the performance drops by 0.07 dB, demonstrating that our local prompt fusion strategy contributes to improved performance. In addition, we see that when the GFI in PFB is removed, performance decreases by 0.13 dB, highlighting the critical role that the globally interacted feature plays in effective prompt fusion. As shown in Table 7, we present more detailed ablation studies on PFB and PICAB, exploring performance variations by changing their channel sizes and number of heads in attention mechanism. Moreover, to

Experiment	PSNR↑ / RMSE↓
PFB w/o LPFA	39.66 / 5.01
PFB w/o GFI	39.60 / 5.23
PFB (Ours)	**39.73 / 4.98**

Table 4. Ablation study for PFB.

W-MSA	W-MCA	GFI	PSNR↑ / RMSE↓
✓	-	-	39.45 / 5.20
-	✓	-	39.55 / 5.09
-	✓	✓	**39.73 / 4.98**

Table 5. Ablation study for PICAB.

Figure 5. Our failure case in UCF [44].

	Concat.	*PAPP*	PSNR↑ / RMSE↓
(a)	-	-	36.83 / 6.93
(b)	*M*	-	39.37 / 5.32
(c)	-	*S*	37.66 / 6.21
(d)	-	*M*	39.66 / 5.10
(e)	-	*M+S*	39.69 / **4.92**
(f)	*M*	*M+S*	**39.73** / 4.98

Table 6. Ablation study on injection and selection of prompts. *S* and *M* indicate semantic map and binary mask.

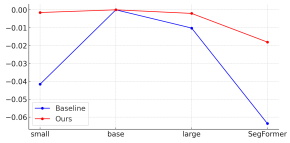


Figure 8. The graph compares the variation in PSNR (y-axis) across different semantic segmentation maps (x-axis).

PFB	dimension	12	24	36 (Ours)	48
	PSNR↑ / RMSE↓	39.53/5.17	39.78/5.02	**39.73/4.98**	39.64/4.97
PICAB	# of Heads	1	2	4 (Ours)	8
	PSNR↑ / RMSE↓	39.64/5.25	39.61/5.05	**39.73/4.98**	39.63/5.17

Table 7. Ablation study results for PFB with varying dimensions and PICAB with different numbers of heads.

Figure 9. Challenging case illustrating the impact of inaccuracies in semantic maps on the WSRD+ [30] dataset. We measure PSNR in the bottom-left region of images.

validate the effects of PICAB, we compare window-based multi-head cross-attention (W-MCA) and self-attention (W-MSA). We see that using W-MCA over W-MSA improves the PSNR by 0.1 dB and the RMSE by 0.11, demonstrating that cross-attention can effectively capture semantic prompt information and improve performance. Furthermore, fusing GFI with W-MCA increases PSNR by 0.18 dB and reduces RMSE by 0.11, indicating that injecting the global contextual information from the baseline feature into locally shared features improves overall performance.

Analysis We visualize the feature map before and after applying PAPP in Figure 7 to validate its effectiveness. By comparing the output of the PAPP module with the input, we observe that the feature map in the shadow region becomes more similar to that in the surrounding non-shadow region, which aligning with the semantic information. This shows that our PAPP effectively improves the baseline feature representation by integrating semantic information and mask information. With better feature representation, the baseline model can remove shadows more effectively, producing a smoother image, especially in the boundary regions.

Additionally, our PAPP effectively utilizes semantic prompts while remaining robust to lower-quality inputs. Figure 8 shows variation in PSNR across different semantic segmentation maps Mask2Former [4] variants (small, base, large) and SegFormer [39] measured in terms of PSNR deviation across all regions. As shown in Figure 8, the performance of models with PAPP shows only a slight variation, as PAPP refines the poor semantic information. In contrast, the baseline model's performance varies more significantly when using the same low-quality mask. This demonstrates that our approach is less dependent on the quality of the semantic information provided.

We also conduct experiments with poor-quality semantic information. As shown in Figure 9, the segmentation maps inherently contain some errors, yet our methods have been trained with such maps, making them notably robust to inaccuracies. Despite these imperfections, our approach effectively utilize the provided information, ensuring reliable performance in challenging scenarios.

5. Conclusion

In this work, we presented key insights into using Prompt learning methods with semantic information and binary masks as latent prompts, and demonstrated the superiority of the PAPP module that processes semantic prompts in shadow removal. We introduced the use of semantic maps in this field, validating their positive impact, which has been effectively implemented in our PAPP. Additionally, PAPP is a plug-and-play module, that can be easily integrated into existing or future shadow removal works, with demonstrated performance gains on various benchmark datasets. Moreover, we are the first to fuse multiple prompts simultaneously in shadow removal, allowing for further performance improvements by incorporating additional explicit prompts beyond masks or semantic maps.

Limitations. Our method, which utilizes semantic segmentation maps and binary masks, encounters challenges when the shadow region is significantly larger than the non-shadow region. In experiments on the real-world UCF dataset [44], our method struggles to effectively leverage semantic information in such cases, as illustrated in Figure 5. There may be additional useful prompts, such as illumination maps, reflectance maps, or other types of semantic information, that have yet to be explored and could help networks remove shadows more effectively. We identify this as a potential area for future research.

References

[1] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. 2022. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. 2020. 1, 2

[3] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *ICCV*, 2021. 1, 2, 3

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 3, 5, 8

[5] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *AAAI*, 2020. 1, 5, 6, 2, 3

[6] 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 *CVPR*, 2021. 5, 6, 1, 2, 3

[7] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. In *AAAI*, 2023. 1, 2, 3, 4, 5, 6, 12, 13, 20, 21

[8] 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 *CVPR*, 2023. 1, 2, 3, 5, 6

[9] Lanqing Guo, Chong Wang, Wenhan Yang, Yufei Wang, and Bihan Wen. Boundary-aware divide and conquer: A diffusion-based solution for unsupervised shadow removal. In *ICCV*, 2023. 1, 3

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 4

[11] Shilin Hu, Hieu Le, ShahRukh Athar, Sagnik Das, and Dimitris Samaras. Shadow removal refinement via material-consistent shadow edges. *arXiv preprint arXiv:2409.06848*, 2024. 3

[12] 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. 2, 3

[13] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, 2019. 1, 2

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2

[15] Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *ICCV*, 2021. 3

[16] Yeying Jin, Wei Ye, Wenhan Yang, Yuan Yuan, and Robby T Tan. Des3: Adaptive attention-driven self and soft shadow removal using vit similarity. In *AAAI*, 2024. 1

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[18] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971. 1

[19] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *ICCV*, 2019. 1, 2, 3, 5, 4, 6, 7, 8, 9, 10

[20] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *ECCV*, 2020. 1

[21] Yuhao Liu, Zhanghan Ke, Ke Xu, Fang Liu, Zhenwei Wang, and Rynson WH Lau. Recasting regional lighting for shadow removal. In *AAAI*, 2024. 3, 6

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[23] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. 2024. 2

[24] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, 2017. 1, 2, 3, 5, 6, 18, 19, 20, 21, 22, 23, 24, 25

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 2

[26] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 45(4): 4713–4726, 2022. 5, 6, 1, 7, 8, 14, 15, 22, 23

[27] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020. 2

[28] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 2

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[30] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In *CVPR*, 2023. 8

[31] 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 *CVPR*, 2024. 2, 5

[32] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *ECCV*, 2022. 1, 2, 3, 9, 10, 16, 17, 24, 25

[33] Cong Wang, Jinshan Pan, Wanyu Lin, Jiangxin Dong, Wei Wang, and Xiao-Ming Wu. Selfpromer: Self-prompt dehazing transformers with depth-consistency. In *AAAI*, 2024. 2

[34] Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, and Junyang Chen. Promptre-

storer: A prompting image restoration method with degradation perception. 2024. 2

[35] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 2018. 1, 2, 3, 11, 12, 13, 14, 15, 16, 17

[36] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 2018. 3

[37] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 5, 6, 1, 11, 19

[38] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. Homoformer: Homogenized transformer for image shadow removal. In *CVPR*, 2024. 1, 5, 6, 4, 18

[39] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 8

[40] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *AAAI*, 2020. 2, 3

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[42] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 3

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2

[44] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, 2010. 8, 1

[45] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *CVPR*, 2022. 1, 6, 3

[46] Yurui Zhu, Zeyu Xiao, Yanchi Fang, Xueyang Fu, Zhiwei Xiong, and Zheng-Jun Zha. Efficient model-driven network for shadow removal. In *AAAI*, 2022. 6, 3

PAPP: Plug-And-Play Prompts for Shadow Removal

Supplementary Material

6. Implementation Details

6.1. Evaluation

For comparison purposes, we evaluate all metrics on MATLAB, following the conventions of previous works[8, 20, 32], with image resolution of 256x256 on ISTD [35], AISTD [19], and SRD [24] datasets. To demonstrate that PAPP can be applied to various models, we applied it to SG-Shadow [32], SR3 [26], UFormer [37], ShadowFormer [7], and HomoFormer [38]. The models with PAPP applied are referred to as baseline+PAPP.

6.2. Training details

We implement all proposed networks using PyTorch on NVIDIA GeForce RTX 3090 GPU cards. We primarily adopted the corresponding baseline training strategies to train PAPP. The detailed training strategies are described in Table 11. Exceptionally, we train HomoFormer for 1200 epochs on the SRD dataset to ensure faithful reproduction for both the baseline and the applied PAPP.

	Params (M)↓	PSNR↑/RMSE↓
ShadowFormer (Big) [7]	14.5	39.37/5.32
ShadowFormer + PAPP	**12.6**	**39.49/5.15**

Table 8. Ablation study for number of parameters.

7. Additional Results

7.1. Visual consistency

To compare visual consistency, we compute the mean variance of the error map (*i.e.* the difference between the ground-truth clean image and the network result) on the AISTD dataset. As shown in in Table 9, applying PAPP results in much lower error variance, indicating that PAPP produces more visually consistent predictions by using semantic information, thereby improving shadow removal performance.

	UFormer	UFormer + PAPP
Variance↓	39.019	**36.484**

Table 9. Visual consistency comparison on AISTD.

7.2. Number of parameters and FLOPs

We conduct experiments with a similar number of parameters as the baseline. For comparison, we increase depths of ShadowFormer [7] to 3 for all layer. Both models are trained on the AISTD dataset. As shown in table 8, our approach achieves higher scores despite using fewer parameters. This demonstrates that our method enhances baseline models not merely through an increase in parameters.

To validate the efficiency of our PAPP module, we provide a comparison of FLOPs and the number of parameters between baseline and PAPP in Table 10. The FLOPs are calculated using an image size of 256×256. For a fair comparison with previous methods, we adjusted the window size of UFormer and ShadowFormer from 10 to 8.

	FLOPs (G)	Params (M)
SP+M-Net [19]	39.8	141.2
DHAN [5]	262.9	21.8
Auto-exposure [6]	104.8	142.2
SG-Shadow	39.7	6.2
SG-Shadow + PAPP	58.0	8.1
UFormer	31.4	5.2
UFormer + PAPP	47.6	7.9
ShadowFormer	64.6	11.4
ShadowFormer + PAPP	77.3	12.6
HomoFormer	35.6	17.8
HomoFormer + PAPP	53.6	22.8

Table 10. Comparison of FLOPs and the number of parameters after applying PAPP.

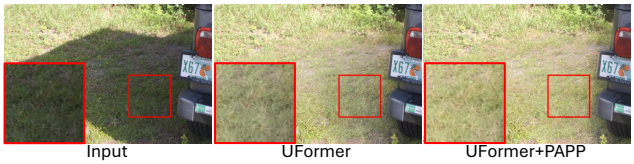


Figure 10. Visual comparison between baseline and ours in Real world datasets [44].

7.3. Visual comparison in Real world datasets.

To ensure comprehensive evaluation, we provide visual examples from the UCF [44] , demonstrating our method's robustness and effectiveness in real-world, diverse, and complex scenarios, as shown in Figure 10.

models	SG-Shadow	SR3	UFormer	ShadowFormer	HomoFormer
base	CNN	Diffusion	TransFormer		
batch size	1	16	7	7	8
epoch	200	3000	500	500	600
inference image size	full	256×256	full	full	full
train patch size	400×400	160×160	320×320	320×320	384×384
optimizer	Adam	Adam	AdamW	AdamW	Adam
initial learning rate	2×10^{-4}	3×10^{-5}	2×10^{-4}	2×10^{-4}	2×10^{-4}
gpus	RTX 3090	RTX 3090	RTX 3090×2	RTX 3090	RTX 3090×2
sampling steps	-	25	-	-	-

Table 11. Training details about whole experiments, more details can found each baseline papers.

7.4. More results in ISTD dataset and SG-shadow

To further demonstrate the effectiveness of PAPP, we provide quantitative results of additional dataset and baseline in Table 12. The ISTD dataset [35] consists of 1330 training and 540 test triplets, each consisting of shadow images, masks, and shadow-free images. Due to the redundancy with AISTD [19] dataset, the results for the ISTD dataset are presented in the supplementary material, where the general superiority of PAPP is also demonstrated.

Also, we conduct additional experiments with CNN-based model SG-shadow [32], which is described in Table 13. Given that PAPP works well with the CNN-based method SG-Shadow, it is evident that PAPP can be effectively combined with a variety of methods.

7.5. More Visual Results

To provide further validation of our methods, we present additional visual comparisons with other shadow removal networks, DSC [12], Auto-exposure [6], and DHAN [5]. We also include visual results for our baselines and applied PAPPs on the ISTD [35], AISTD [35] and SRD [24] datasets. Each visual result showcases the optimal visual effects achieved by our methods.

Additiaonal visual results on AISTD. Figure 11 presents the visual results of HomoFormer+PAPP on the AISTD dataset, while Figure 12, 13 show the results of ShadowFormer+PAPP. Figure 14, 15 display the results of SR3+PAPP, and Figure 16, 17 illustrate the results of SG-Shadow+PAPP. As seen in these figures, our PAPP method enhances the restored images, consistently reducing boundary artifacts and achieving more harmonious predictions.

Additional visual results on ISTD. Figure 18 shows the visual comparison of UFormer+PAPP. Figure 19, 20 represent the visual results of ShadowFormer+PAPP. Figure 21, 22 present the visual results of SR3+PAPP. Figure 23, 24 display the visual comparison of SG-Shadow+PAPP. Across these various baselines, we observe that the predictions with PAPP are closer to the ground truth.

Additional visual results on SRD. Figure 25 presents the visual comparison of HomoFormer+PAPP on SRD dataset. Figure 26 shows the visual comparison of UFormer+PAPP, while Figure 27, 28 display the visual comparison of ShadowFormer+PAPP. The visual comparison of SR3+PAPP and SG-Shadow+PAPP are shown in Figure 29, 30 and Figure 31, 32, respectively. These figures consistently demonstrate that applying PAPP results in fewer artifacts and significantly more effective shadow restoration.

	Shadow Region(S)			Non-Shadow Region (NS)			All image (ALL)		
	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Input Image	22.40	0.936	32.10	27.32	0.976	7.09	20.56	0.893	10.88
ST-CGAN [36]	33.74	0.981	9.99	29.51	0.958	6.05	27.44	0.929	6.65
DSC [12]	34.64	0.984	8.72	31.26	0.969	5.04	29.00	0.944	5.59
DHAN [5]	35.53	0.988	7.49	31.05	0.971	5.30	29.11	0.954	5.66
Fu *et al*. [6]	34.71	0.975	7.91	28.61	0.880	5.51	27.19	0.945	5.88
DC-ShadowNet [15]	31.69	0.976	11.43	28.99	0.958	5.81	26.38	0.922	6.57
Zhu et al. [46]	36.95	0.987	8.29	31.54	0.978	4.55	29.85	0.960	5.09
BMNet [45]	35.61	0.988	7.60	32.80	0.976	4.59	30.28	0.959	5.02
SR3 (2022)	35.97	0.989	7.67	31.74	0.974	5.27	29.75	0.958	5.64
SR3 + PAPP	**37.05**	**0.990**	**6.65**	**32.77**	**0.978**	**4.47**	**30.74**	**0.964**	**4.82**
UFormer (2022)	38.35	0.990	6.06	32.70	0.980	4.06	31.19	0.967	4.39
UFormer + PAPP	**38.48**	**0.990**	**5.985**	**34.57**	**0.981**	**3.73**	**32.48**	**0.968**	**4.08**
ShadowFormer (2023)	38.23	0.990	6.18	34.62	0.981	3.68	32.42	0.967	4.09
ShadowFormer + Prompt	**38.29**	**0.991**	**6.03**	**34.72**	**0.981**	**3.62**	**32.50**	**0.968**	**4.01**

Table 12. Quantitative results of shadow removal on ISTD [35] dataset.

Datasets	Shadow Region(S)			Non-Shadow Region (NS)			All image (ALL)		
	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
ISTD	37.57	0.990	6.17	32.79	0.980	4.17	30.88	0.965	4.50
ISTD (+ PAPP)	**37.89**	**0.990**	**5.80**	**32.92**	**0.979**	**4.22**	**31.10**	**0.966**	**4.48**
AISTD	37.60	0.990	5.82	37.42	0.985	2.44	33.90	0.971	3.00
AISTD (+ PAPP)	**38.58**	**0.991**	**5.33**	**37.63**	**0.985**	**2.40**	**34.50**	**0.973**	**2.88**
SRD	35.50	0.983	5.98	37.05	0.988	3.16	32.58	0.964	3.95
SRD (+ PAPP)	**36.42**	**0.985**	**5.37**	**37.33**	**0.989**	**3.10**	**33.22**	**0.967**	**3.74**

Table 13. Quantitative results of shadow removal with SG-Shadow [32]. (+PAPP) indicates the method where PAPP is applied on the respective dataset.

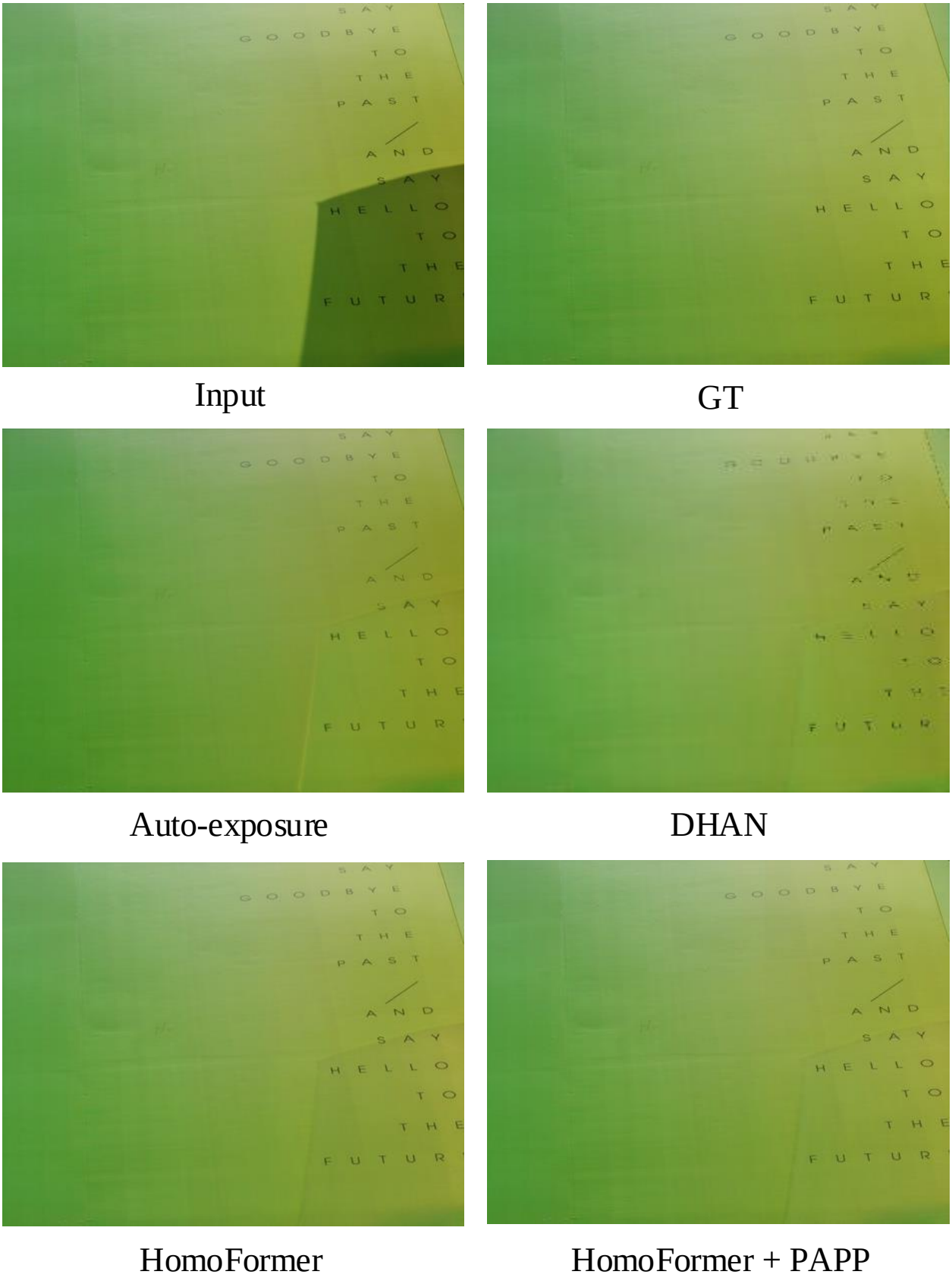


Figure 11. Visual comparison of shadow removal results on AISTD [19] datasets with HomoFormer [38] baseline. Best viewed on high-resolution display.

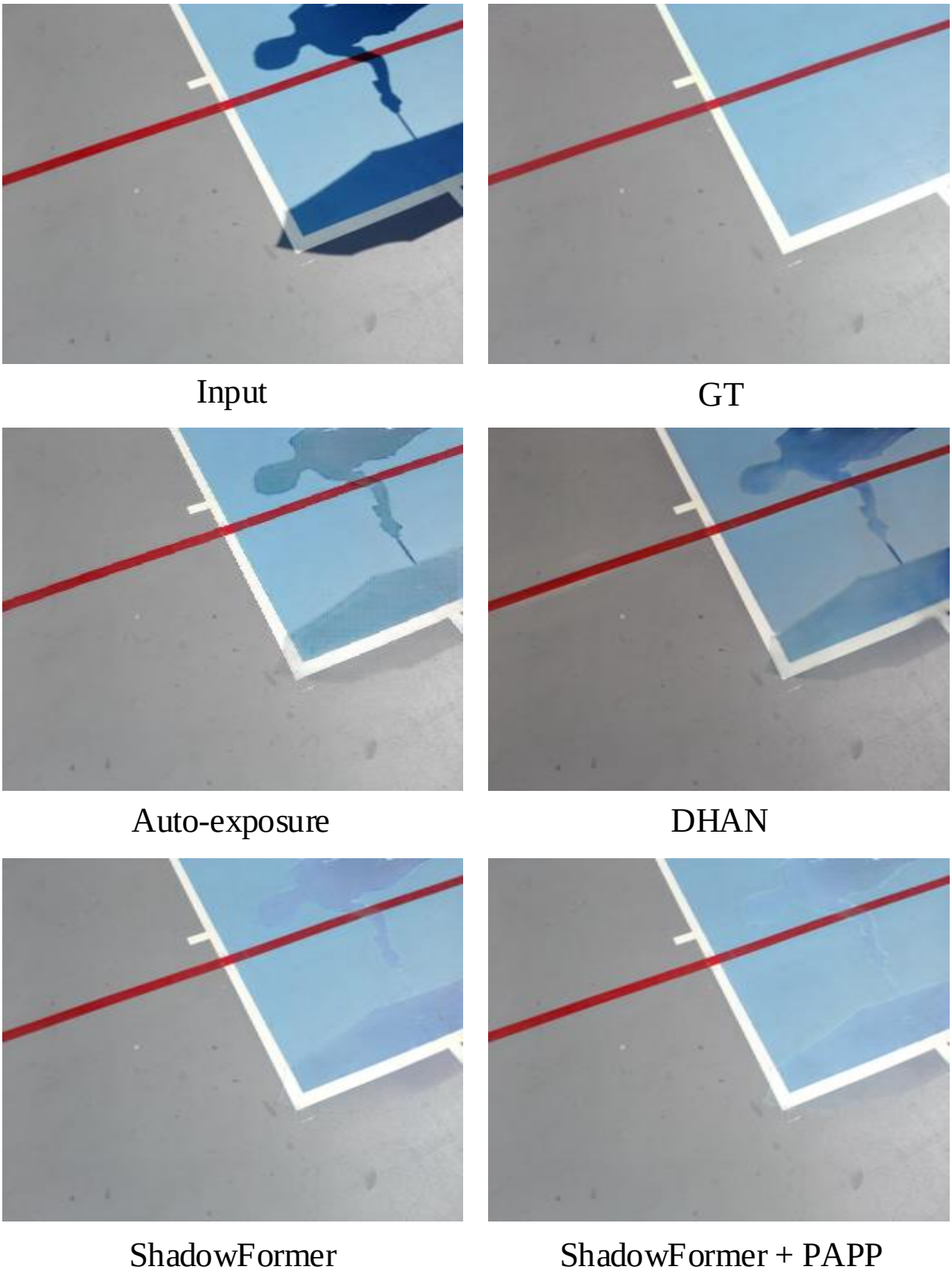
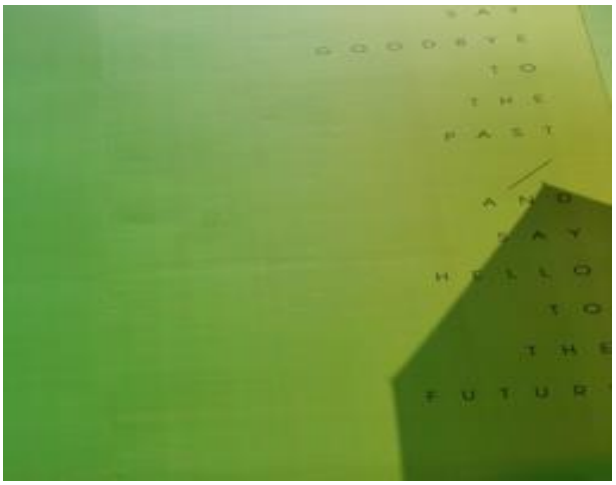
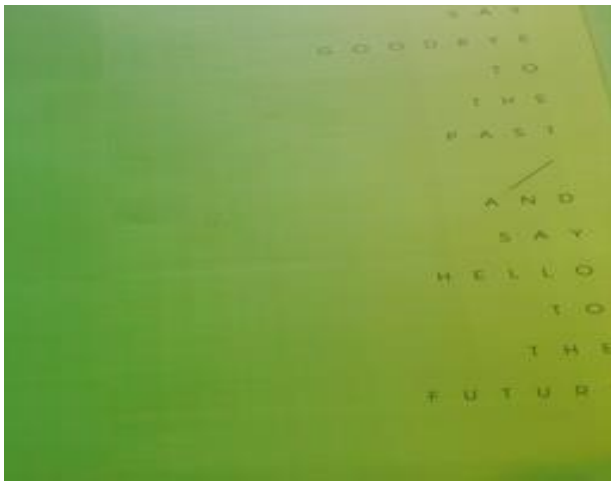


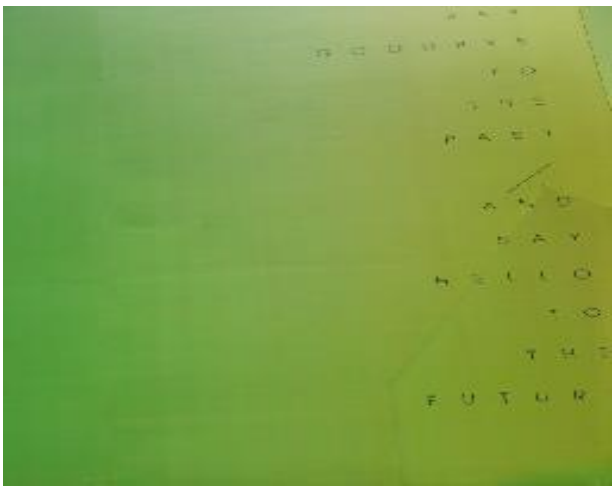
Figure 12. Visual comparison of shadow removal results on AISTD [19] datasets with ShadowFormer [7] baseline. Best viewed on high-resolution display.



Input

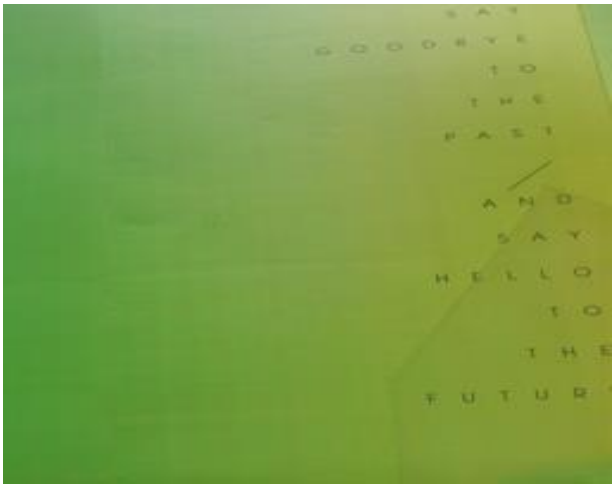


GT

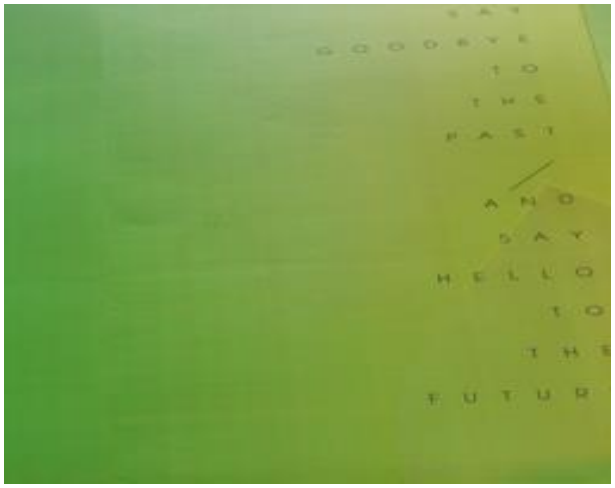


Auto-exposure

DHAN



ShadowFormer



ShadowFormer + PAPP

Figure 13. Visual comparison of shadow removal results on AISTD [19] datasets with ShadowFormer [7] baseline. Best viewed on high-resolution display.

Input

GT

Auto-exposure

DHAN

SR3

SR3 + PAPP

Figure 14. Visual comparison of shadow removal results on AISTD [19] datasets with SR3 [26] baseline. Best viewed on high-resolution display.

Input

GT

Auto-exposure

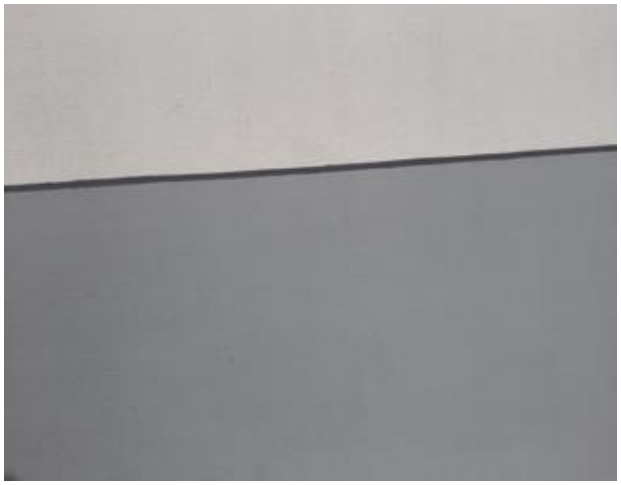
DHAN

SR3

SR3 + PAPP

Figure 15. Visual comparison of shadow removal results on AISTD [19] datasets with SR3 [26] baseline. Best viewed on high-resolution display.

Input



GT

Auto-exposure

DHAN

SGshadow

SGshadow + PAPP

Figure 16. Visual comparison of shadow removal results on AISTD [19] datasets with SGshadow [32] baseline. Best viewed on high-resolution display.

Input

GT

Auto-exposure

DHAN

SGshadow

SGshadow + PAPP

Figure 17. Visual comparison of shadow removal results on AISTD [19] datasets with SGshadow [32] baseline. Best viewed on high-resolution display.

Figure 18. Visual comparison of shadow removal results on ISTD [35] datasets with UFormer [37] baseline. Best viewed on high-resolution display.

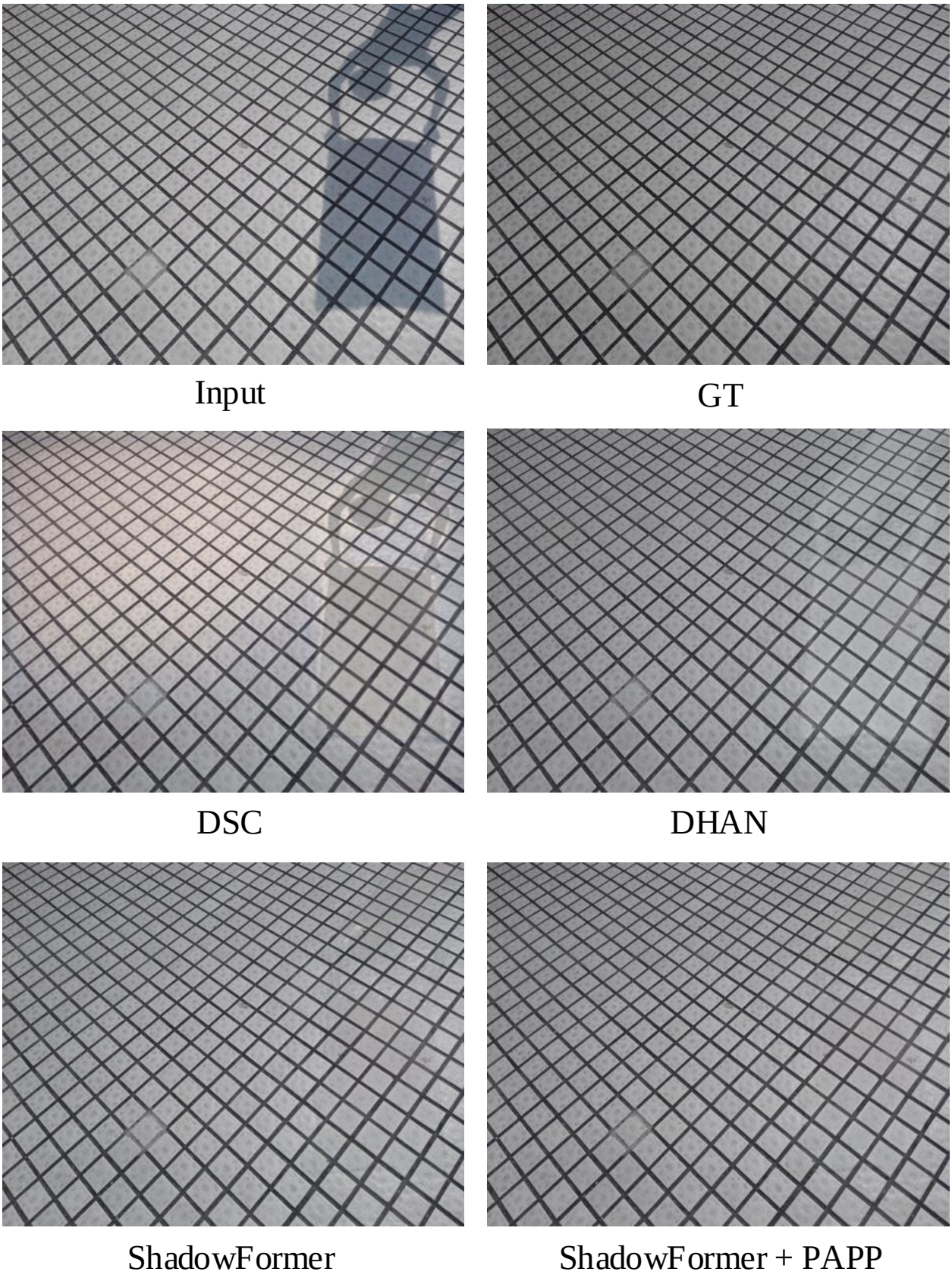


Figure 19. Visual comparison of shadow removal results on ISTD [35] datasets with ShadowFormer [7] baseline. Best viewed on high-resolution display.

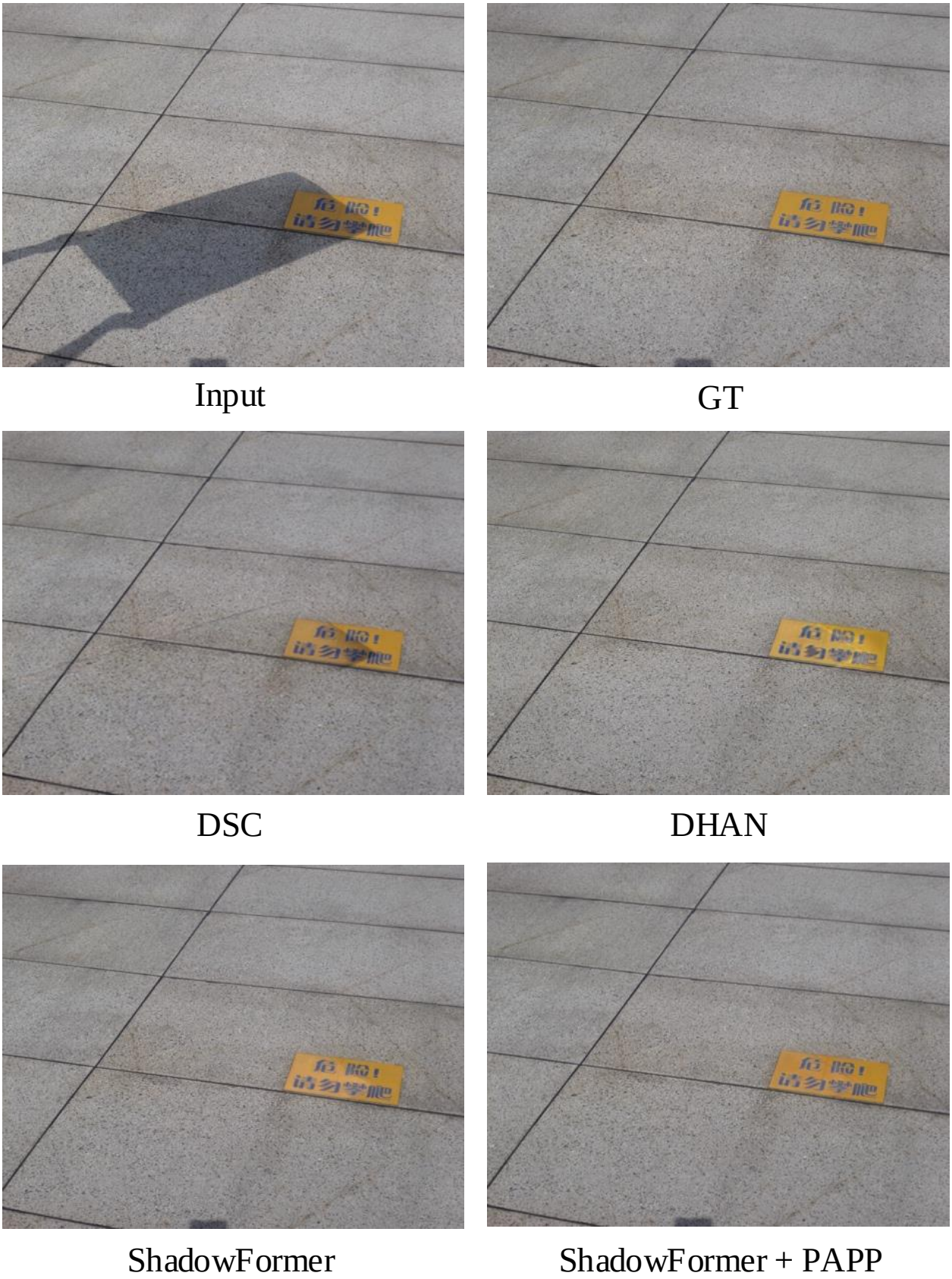


Figure 20. Visual comparison of shadow removal results on ISTD [35] datasets with ShadowFormer [7] baseline. Best viewed on high-resolution display.

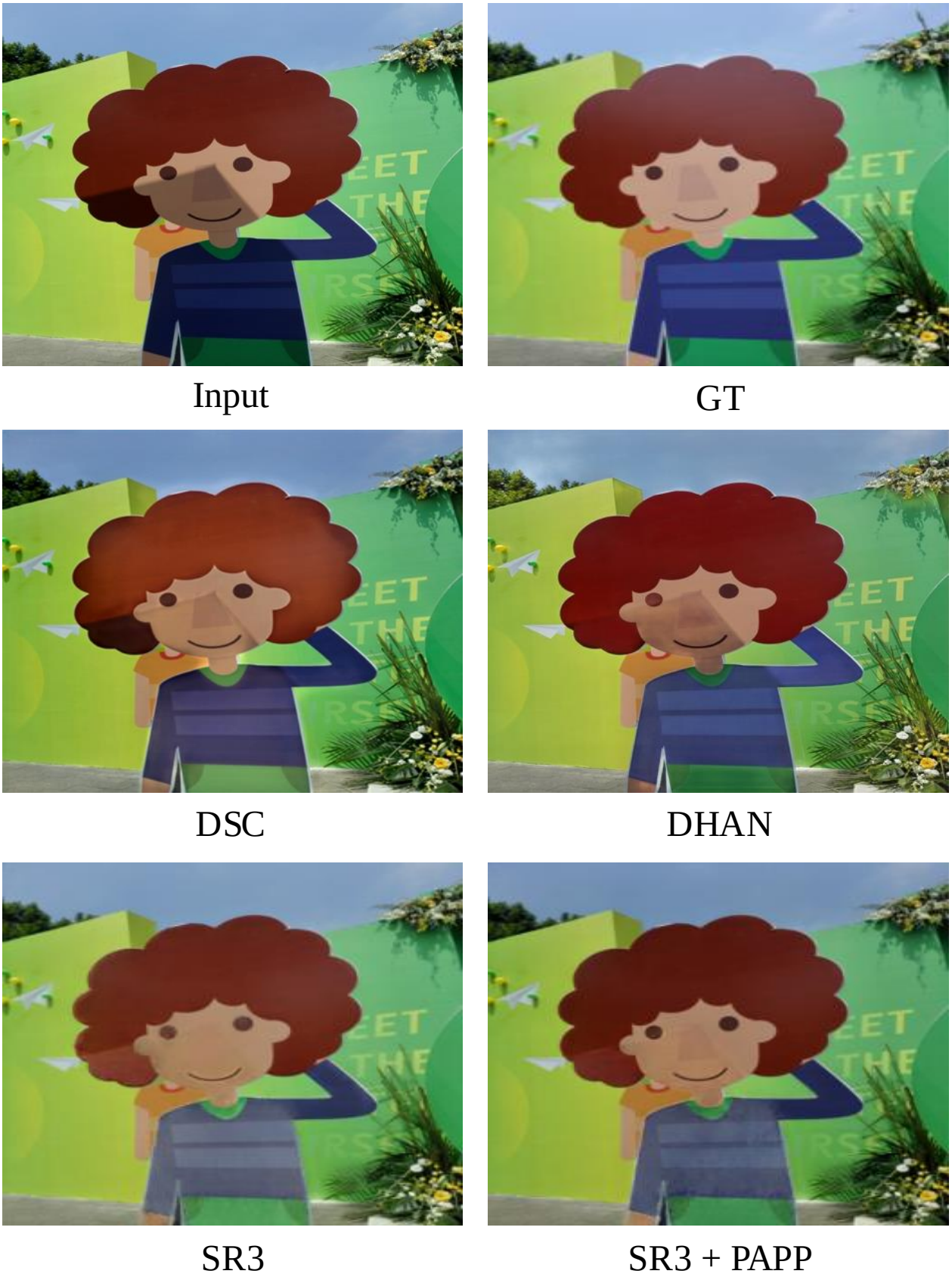


Figure 21. Visual comparison of shadow removal results on ISTD [35] datasets with SR3 [26] baseline. Best viewed on high-resolution display.

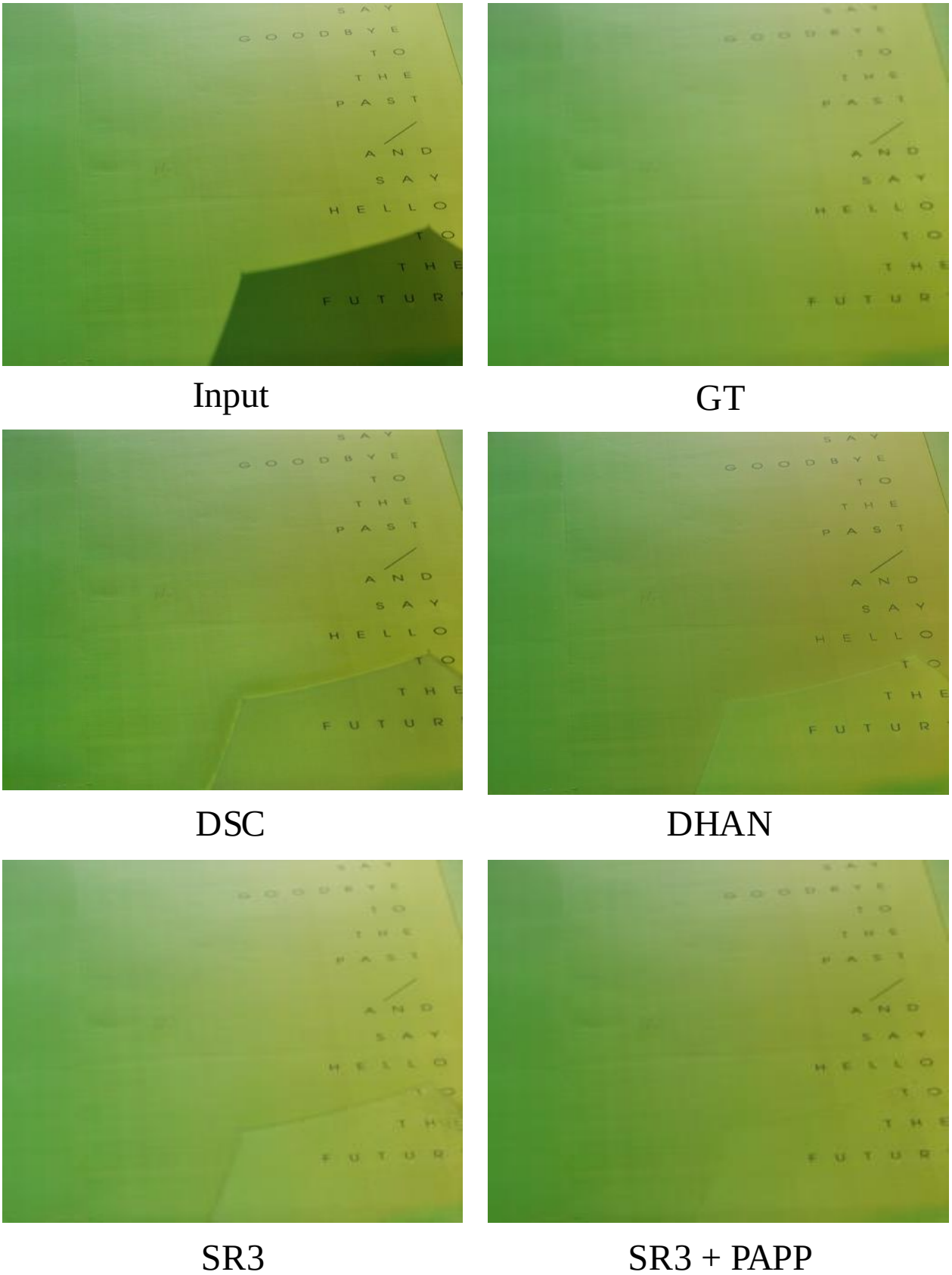


Figure 22. Visual comparison of shadow removal results on ISTD [35] datasets with SR3 [26] baseline. Best viewed on high-resolution display.

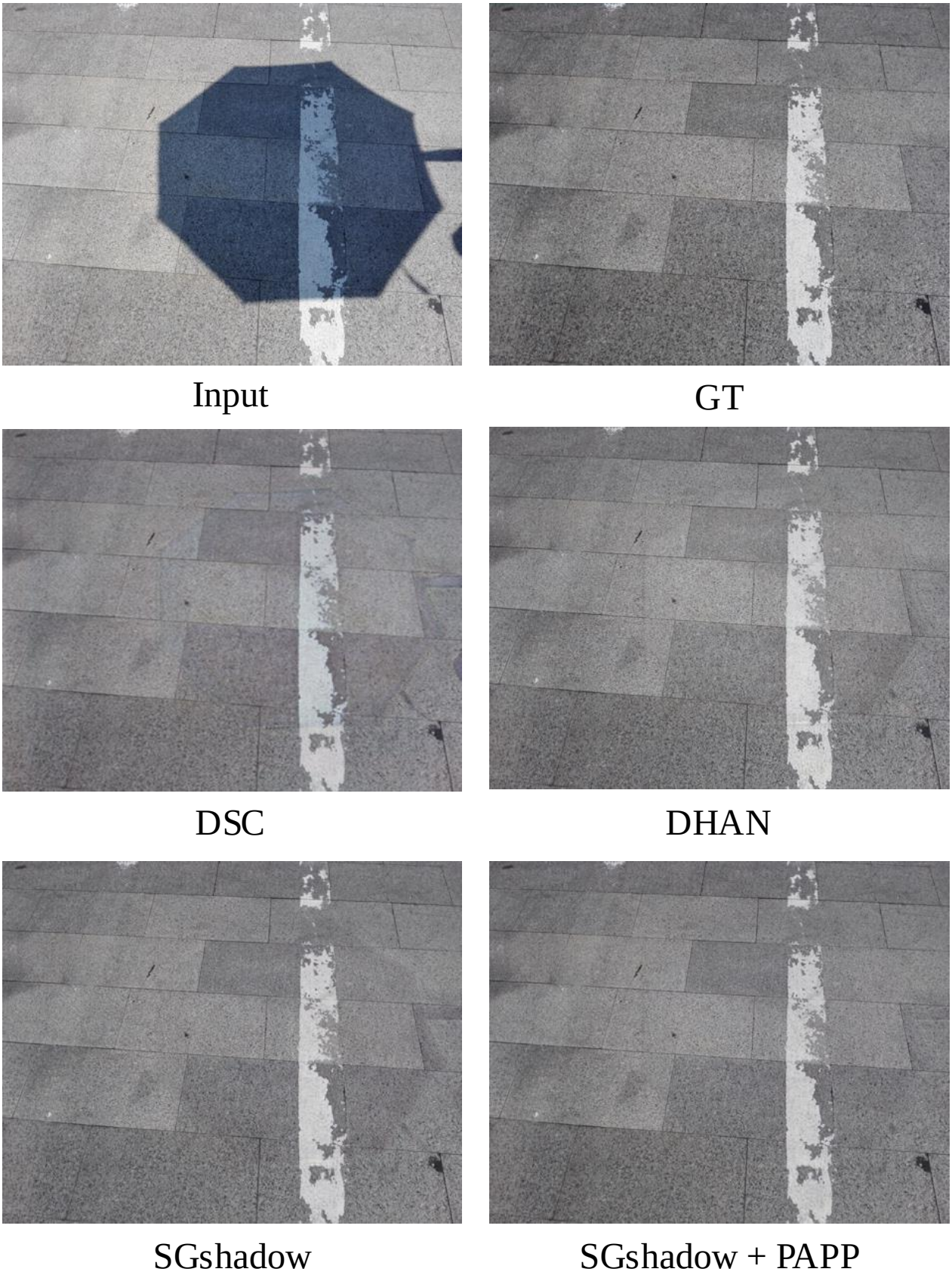


Figure 23. Visual comparison of shadow removal results on ISTD [35] datasets with SGshadow [32] baseline. Best viewed on high-resolution display.

Input

GT

DSC

DHAN

SGshadow

SGshadow + PAPP

Figure 24. Visual comparison of shadow removal results on ISTD [35] datasets with SGshadow [32] baseline. Best viewed on high-resolution display.

Input

GT

DSC

Auto-exposure

HomoFormer

HomoFormer + PAPP

Figure 25. Visual comparison of shadow removal results on SRD [24] datasets with HomoFormer [38] baseline. Best viewed on high-resolution display.

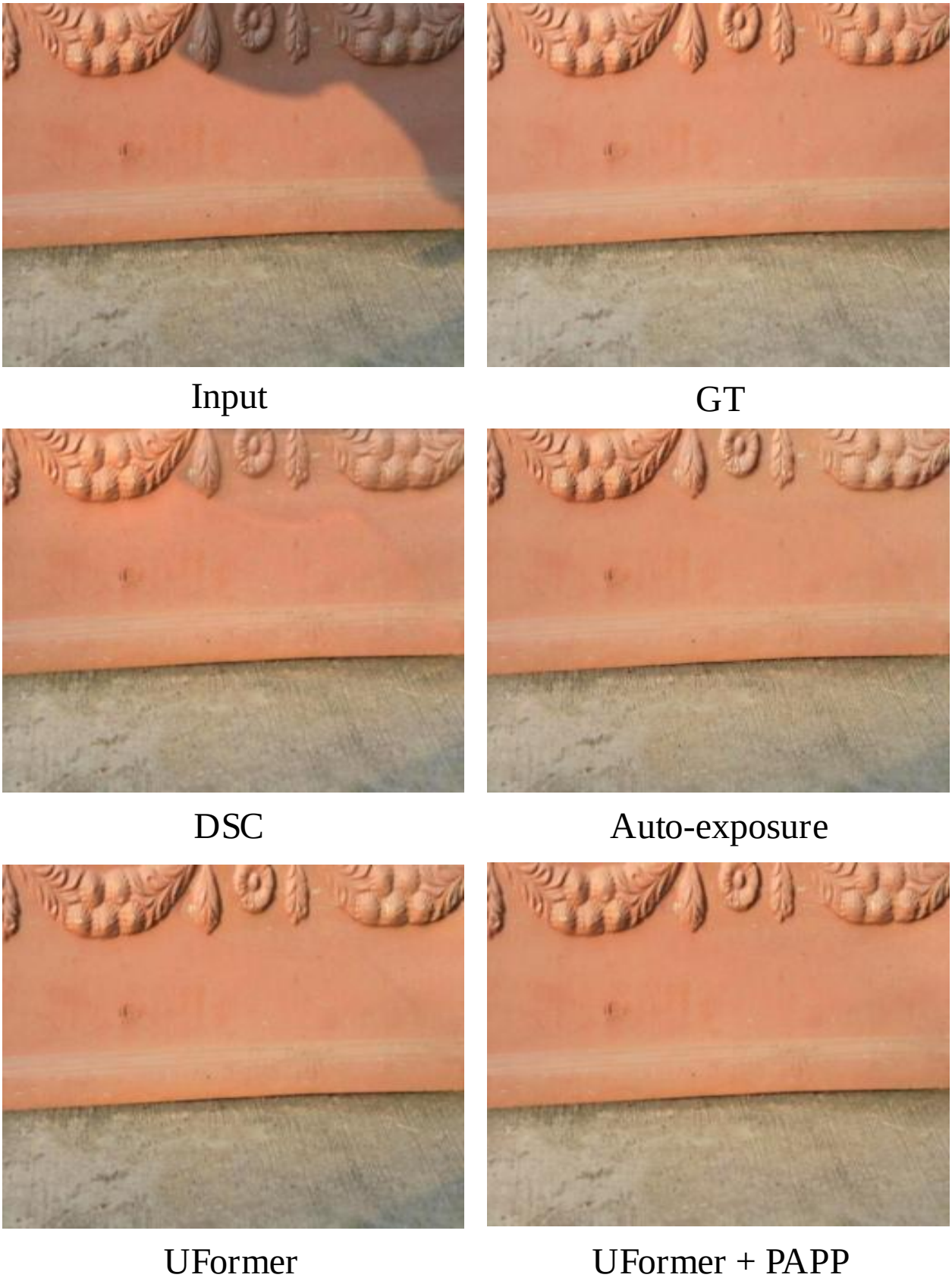


Figure 26. Visual comparison of shadow removal results on SRD [24] datasets with UFormer [37] baseline. Best viewed on high-resolution display.

Input

GT

DSC

Auto-exposure

ShadowFormer

ShadowFormer + PAPP

Figure 27. Visual comparison of shadow removal results on SRD [24] datasets with ShadowFormer [7] baseline. Best viewed on high-resolution display.

Input

GT

DSC

Auto-exposure

ShadowFormer

ShadowFormer + PAPP

Figure 28. Visual comparison of shadow removal results on SRD [24] datasets with ShadowFormer [7] baseline. Best viewed on high-resolution display.

Figure 29. Visual comparison of shadow removal results on SRD [24] datasets with SR3 [26] baseline. Best viewed on high-resolution display.

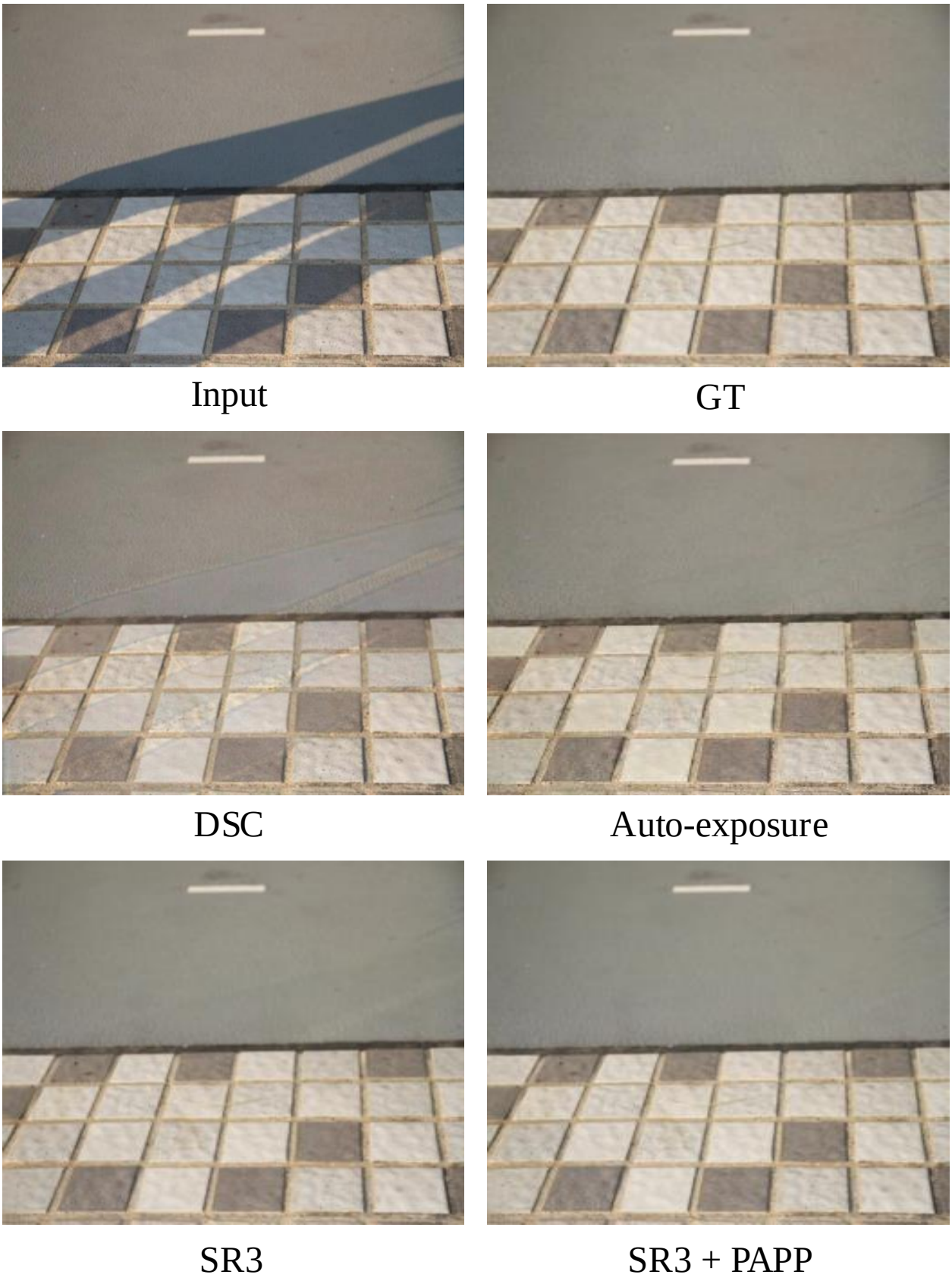


Figure 30. Visual comparison of shadow removal results on SRD [24] datasets with SR3 [26] baseline. Best viewed on high-resolution display.

Input

GT

DSC

Auto-exposure

SGshadow

SGshadow + PAPP

Figure 31. Visual comparison of shadow removal results on SRD [24] datasets with SGshadow [32] baseline. Best viewed on high-resolution display.

Input

GT

DSC

Auto-exposure

SGshadow

SGshadow + PAPP

Figure 32. Visual comparison of shadow removal results on SRD [24] datasets with SGshadow [32] baseline. Best viewed on high-resolution display.