HirFormer: Dynamic High Resolution Transformer for Large-Scale Image Shadow Removal

Xin Lu, Yurui Zhu, Xi Wang, Dong Li, Jie Xiao,
Yunpeng Zhang, Xueyang Fu*, Zheng-Jun Zha
University of Science and Technology of China
luxion@mail.ustc.edu.cn, xyfu@ustc.edu.cn

Abstract

*Existing image restoration models have limited performance in high-resolution image shadow removal tasks, particularly in handling complex background information and unevenly distributed shadows. To address this challenge, we propose a novel two-stage approach called HirFormer for high-resolution image shadow removal. The first stage, Dynamic High Resolution Transformer, reconstructs the high-resolution background information and removes a significant portion of the shadows based on the Transformer architecture. The second stage, Large-scale Image Refinement, incorporates the NAFNet model to further eliminate residual shadows and address block artifacts introduced by the first stage. Experimental results on official datasets validate the superiority of our method compared to existing approaches, and our approach emerged as the winner in the fidelity track of the NTIRE 2024 Shadow Removal Challenge during the final testing competition (**1st place**).*

1. Introduction

Shadows are commonly observed in various natural scenes when a light source is partially or completely obstructed by objects. While shadows in images can provide rich natural information in specific contexts, they inevitably degrade the perception quality of background information. However, correspondingly, shadows in images also introduce a series of challenges for subsequent high-level vision tasks, *e.g.*, object tracking [33] and detection [31], semantic segmentation [53], and face recognition [50]. Therefore, shadow removal, as one of the fundamental tasks in computer vision, has been extensively studied. In recent years, the removal of shadows in high-resolution images has emerged as a challenging aspect in the field of image restoration research.

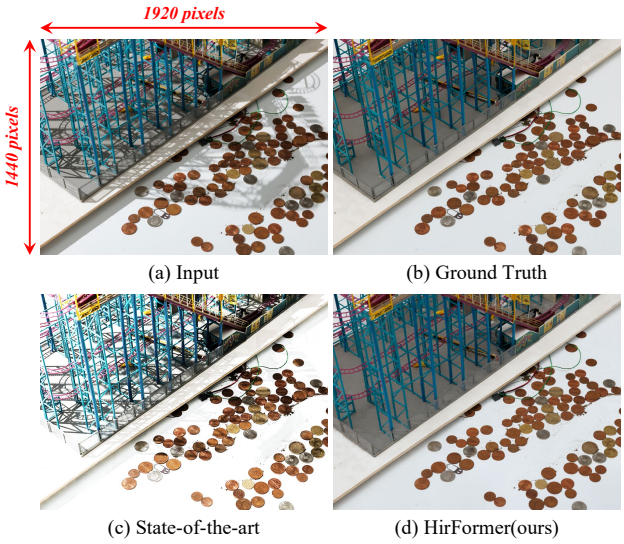

Figure 1. Visual Results of Shadow removal by the SOTA ShadowFormer [17] and the proposed methods. When dealing with complex shadow scenes at high resolution, our method Reshaormer can remove more undesired shadows and reconstruct more pixel information, while also mitigating the disruption to the original background information.

The current approaches for removing shadows from images can be broadly categorized into two types: traditional methods based on physical models and solutions based on deep learning. Traditional shadow removal methods rely on prior knowledge of the physical properties of the image, *e.g.*, image gradients [16], illumination [43], and regions [18, 38], *etc*. However, due to the limitations imposed by these physical characteristics, traditional methods face challenges in accurately modeling and extending to complex shadow removal scenarios in the real world [23].

Moreover, the success of deep learning approaches in diverse computer vision tasks [6, 11–13, 27, 54, 55] has led to their gradual dominance. This trend extends to low-level

* : Corresponding author.

visual tasks, *e.g*., single-image rain removal, image deblurring, and image reflection removal, *etc*. Le *et al*. [25, 26] employed shadow illumination modeling to infer the mapping relationship between the shadow image ($\mathbf{I}_s$) and the clean image ($\mathbf{I}_{sf}$). On the other hand, Hu *et al*. [19] introduced a direction-aware spatial attention module and a growing dilated convolution approach to facilitate shadow removal, effectively utilizing multiple contextual features. Liu *et al*. [31] enhanced the training of their network by utilizing a large dataset of synthetic shadow images, resulting in improved shadow removal performance. Additionally, Zhu *et al*. [54] proposed a novel bidirectional mapping network (BMNet) that incorporates auxiliary supervision considering shadow generation, leading to more effective restoration of the underlying background content during the shadow removal process. While these image restoration methods have achieved good performance in shadow removal tasks, there are still some challenging cases. For instance, when shadows have non-uniform and concentrated distributions, the models may not thoroughly restore the image, and they can even disrupt the contour information in the background region [54]. Moreover, the current methods have not been sufficiently optimized for high-resolution large-scale images, thereby constraining the restoration effectiveness for high-resolution shadow images.

Therefore, in this paper, we introduce a novel two-stage approach called HirFormer (Dynamic High Resolution Transformer for Large-Scale Image Shadow Removal) to overcome this limitation and enhance the restoration quality specifically for high-resolution shadow images. In the first stage, we employ vision transformer [7] blocks to construct a U-shaped encoder-decoder framework. The input high-resolution image is divided into 16 smaller blocks, arranged in a 4×4 grid. Patch embeddings are applied to each patch, which are subsequently processed by the network for restoration. The restored patches are then sequentially stitched together to reconstruct the high-resolution image. This initial stage effectively removes a substantial portion of the shadows. In the second stage, we utilize multiple NAFNet blocks [3] without global residual connections, to refine the high-resolution image. The primary goal of this stage is to eliminate the remaining shadows while minimizing boundary artifacts that may arise due to the block effect of the Transformer. By adopting this approach, we achieve dynamic high-resolution shadow removal. As shown in Figure 1, our method exhibits superior robustness in handling high-resolution shadow removal tasks. Extensive experiments further validate that HirFormer outperforms the state-of-the-art single-image shadow removal approaches in terms of fidelity metrics for the final restoration results. In summary, our contributions are as follows:

- We introduce a novel approach for dynamic high-resolution image processing, enabling vision transformer to effectively restore large-scale images by fully harnessing the benefits of global self-attention in the context of image shadow removal tasks.
- We adopt a two-stage shadow removal strategy to integrate the advantages of vision transformer and NAFNet in low-level visual tasks. This fusion approach enhances the robustness of high-resolution image shadow removal, while effectively mitigating boundary artifacts resulting from the block effect of the Transformer.
- Our method is confirmed to be effective through experiments on both validation and testing data sets provided in the NTIRE 2024 Shadow Removal Challenge and our approach emerged as **the winner in the fidelity track** during the final testing competition.

2. Related work

2.1. Image Shadow Removal

Existing methods for shadow removal in images can be broadly categorized into two types, traditional physics-based techniques and deep learning-based approaches:

Traditional techniques. Early shadow removal methods [8, 9, 14, 18, 48] mostly relied on prior knowledge of the physical characteristics of images, *e.g*., lighting conditions, gradients, regions, and user interactions. Guo *et al*. [18] restored shadow-free images by establishing illumination conditions between individual regions. Finlayson *et al*. [8, 9] utilized the characteristic of gradient consistency to remove shadows. Gong *et al*. [14] improved the robustness of shadow removal algorithms by incorporating two user interaction inputs.

Deep learning-based approaches. In recent years, an escalating number of researchers have employed large-scale datasets to train deep neural networks [6, 19, 26, 29, 30, 32, 40] for the purpose of shadow removal. These approaches usually leverage supervised and unsupervised training strategies.

In supervised methods, Chen *et al*. [4] proposed a context-aware network that integrates information from both shadow and shadow-free regions in the feature space. Le *et al*. [26] employed an illumination model and image decomposition techniques to accomplish the task of shadow removal. Fu *et al*. [10] addressed shadow removal by manipulating the weight map during image exposure fusion, utilizing exposure fusion techniques. Wan *et al*. [39] tackled the challenge of inconsistent static styles between the shadow and shadow-free regions before and after restoration by introducing a style-guided shadow removal network, which aligns the restored shadow regions with the background style. Zhu *et al*. [54] introduced the concept of mutual assistance between the shadow generation and shadow removal processes and designed a parameter-unified net-

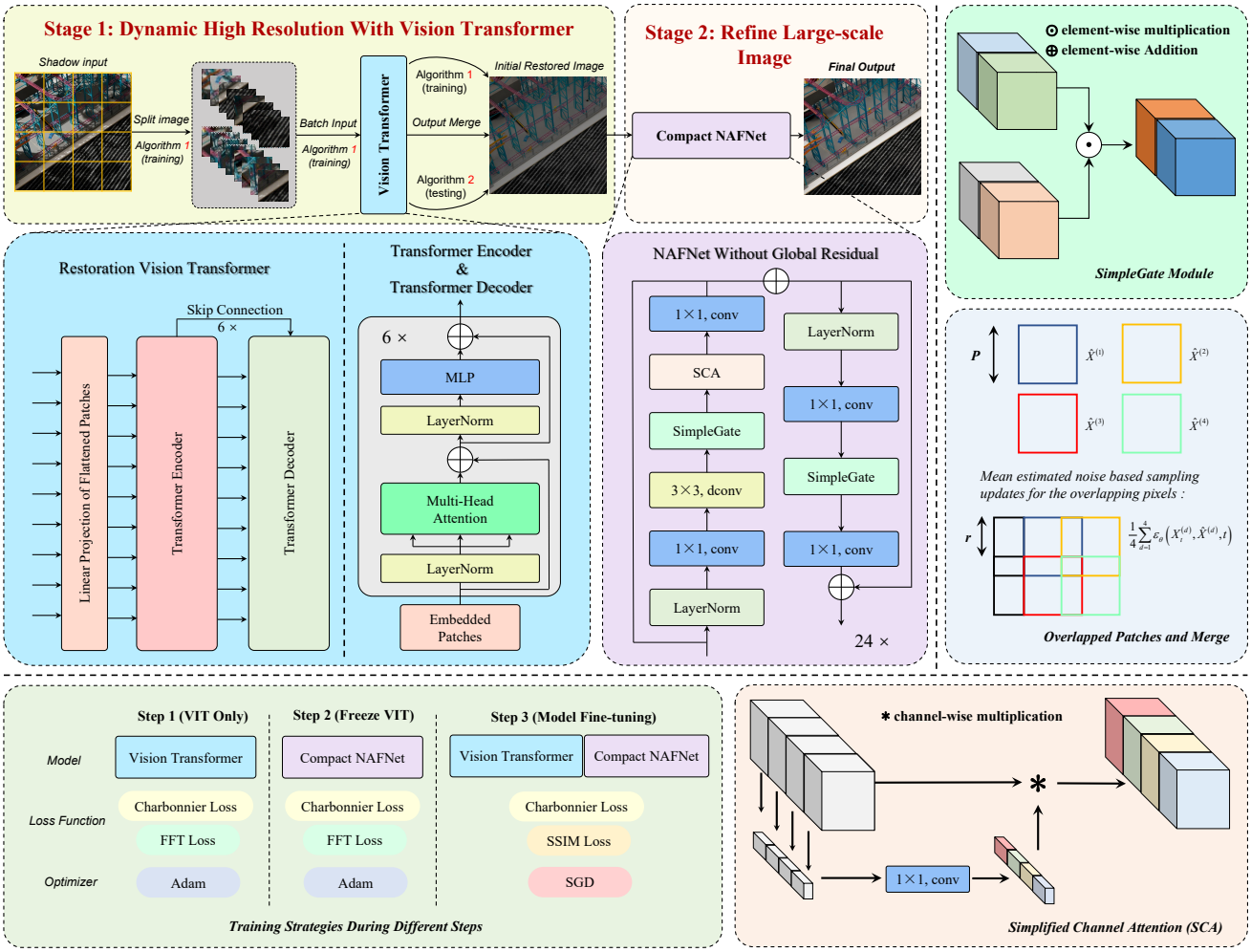

Figure 2. Architecture of HirFormer for high-resolution image shadow removal. The shadow images undergo the initial restoration process using the Dynamic High Resolution Vision Transformer, yielding preliminary restoration results. Subsequently, the Initial Restored Image is further refined for large-scale images by passing it through a compact NAFNet network, ultimately producing the final clean image.

work for synchronous training. In the latest state-of-the-art (SOTA) approach [17], transformer [7, 34] are employed as fundamental modules for the encoder and decoder, enhancing the ability to capture contextual information.

In unsupervised methods, several notable works [20, 21, 30, 31] primarily utilize generative models, *e.g.*, GAN or Diffusion, training the models using unpaired shadow images. Jin *et al*. [21] introduced a method that incorporates shadow-free chromaticity as a constraint to guide the network, leveraging an unsupervised domain classifier for shadow removal.

Although the aforementioned methods have demonstrated promising results in their specific domains for shadow removal, they often face challenges in achieving high performance in complex real-world environments. This is primarily attributed to the substantial variations in shadow distribution and the complexity of non-shadow regions in the background. Consequently, there is a pressing need to develop shadow removal models that can effectively address complex scenes and high-resolution images.

2.2. Image Restoration Using Transformer

The Transformer model [37], which utilizes self-attention mechanisms, was initially applied in the field of natural language processing (NLP) and demonstrated remarkable performance in modeling long sequences. Furthermore, various attention modules have found extensive applications in computer vision tasks [44–46, 56, 57]. The emergence of Vision Transformer (VIT) [7] has made it possible to establish a unified framework bridging NLP and computer vision. Subsequently, the Transformer model has gradually been adopted in various visual tasks, including

image recognition [7, 34], object detection [1], and segmentation [41, 52]. In low-level visual tasks, the Transformer model has also achieved state-of-the-art performance in image restoration domains, *e.g.*, image deblurring [3, 45, 46], deraining [44, 57], dehazing [22], desnowing [57], and super-resolution [28, 47].

In this paper, we designed a Dynamic High Resolution Algorithm 1 to fully leverage the contextual understanding capabilities of Transformer, achieving high-performance restoration of complex high-resolution shadow images.

3. Method

The input for the high-resolution image shadow removal task is a shadow image $\mathbf{I}_s$. The optimization objective is to obtain a shadow-free image $\mathbf{I}_{sf}$ through model processing, minimizing the discrepancy with the ground truth image $\mathbf{I}_{gt}$, while ensuring both effective shadow removal and accurate background reconstruction.

3.1. Two-stage Image Restoration Pipeline

Due to the high resolution of shadow images, uneven shadow distribution, and complex background information, existing shadow removal networks struggle to effectively reconstruct pixel information. Therefore, we propose a novel two-stage shadow removal pipeline called HirFormer, which consists of two components: the Dynamic High Resolution Transformer and the Large-scale Image Refinement Network. The overall framework of HirFormer is illustrated in Figure 2.

- The Dynamic High Resolution Transformer utilizes the processing capabilities of a transformer model to effectively handle long sequences and reconstruct pixel-level details. As a result, it produces the preliminary restoration result $\mathbf{I}_v$, which represents a clean image after the initial removal of shadows.
- The Large-scale Image Refinement Network utilizes the excellent restoration capabilities of NAFNet to further remove shadows, while also helping to eliminate artifacts caused by the block effect of VIT, resulting in the final refined image $\mathbf{I}_{sf}$.

3.2. Dynamic High Resolution Transformer

During this stage, we employ vision transformer blocks to construct a U-shaped encoder-decoder framework. The input $\mathbf{I}_s$ $\left(\in \mathbb{R}^{N\times C\times H\times W}\right)$ is divided into 16 smaller blocks, arranged in a 4×4 grid. Patch embeddings are applied to each patch $\mathbf{I}_i$ $\left(\in \mathbb{R}^{16N\times C\times \frac{H}{4}\times \frac{W}{4}}\right)$, which are subsequently processed in parallel by the network for restoration. The restored patches are then sequentially stitched together to reconstruct the high-resolution image $\mathbf{I}_v$ $\left(\in \mathbb{R}^{N\times C\times H\times W}\right)$, they will then be sent to the Large-scale image Refinement Network for further refinement. The detailed steps of this algorithm during the training and testing processes are illustrated in Algorithm 1 and Algorithm 2.

Algorithm 1 Dynamic High Resolution During Training

Input: the shadow images $\mathbf{I}_s : tensor(N, C, H, W)$
Output: the clean images $\mathbf{I}_{sf} : tensor(N, C, H, W)$

$o, N, C, H, W = 4, \mathbf{I}_s.shape$
for $image$ in $\mathbf{I}_s\ (N)$ **do**
  $\mathbf{I}\,(i) : tensor\left(o^2, C, \frac{H}{o}, \frac{W}{o}\right) \leftarrow split\,(\mathbf{I}_s\,(i))$
  if $(h, w)\,\%\,\left(\frac{H}{4}, \frac{W}{4}\right) == 0$ **then**
    $\mathbf{Position}\,(\mathbf{a}, \mathbf{b}) \leftarrow (\mathbf{h}, \mathbf{w})$
/* Store the segmented image along the channel */
$\mathbf{I}_l : tensor\left(o^2 \cdot N, C, \frac{H}{o}, \frac{W}{o}\right) \leftarrow reshape\left(\sum_{i=1}^{N} \mathbf{I}\,(i)\right)$
/* $\mathbf{Model_1}$ is the Vision Transformer in stage 1 */
$\mathbf{I}_i : tensor\left(o^2 \cdot N, C, \frac{H}{o}, \frac{W}{o}\right) \leftarrow \mathbf{Model}_1\,(\mathbf{I}_l)$
$\mathbf{I}_v : tensor(N, C, H, W) \leftarrow merge\,(\mathbf{I}_i, \mathbf{Position})$
$Loss\,(\mathbf{I}_v, \mathbf{I}_{gt})\,.backward()$
/* $\mathbf{Model_2}$ is the Compact NAFNet in stage 2 */
$\mathbf{I}_{sf} : tensor(N, C, H, W) \leftarrow \mathbf{Model}_2\,(\mathbf{I}_v)$
$Loss\,(\mathbf{I}_{sf}, \mathbf{I}_{gt})\,.backward()$
return $\mathbf{I}_{sf}$

3.3. Large-scale Image Refinement Network

During the refinement stage of shadow removal, the model requires fine-tuning on high-resolution images to effectively eliminate residual shadows and address artifacts caused by the block effect of VIT. As illustrated in Figure 2, we adopted a reduced-scale variant of the NAFNet network, composed of 24 NAFNet modules. Notably, due to the absence of shadow residual learning in the refinement process, the global residual connections were omitted to ensure a lightweight network design.

3.4. Loss Functions

The training process of HirFormer is divided into three steps, as illustrated in Figure 2. Each step utilizes a different loss function to optimize the restoration performance, with a particular focus on preserving image fidelity. To this end, we employ the Charbonnier loss [2], which is mathematically defined as follows:

$$\mathcal{L}_{content} = \frac{1}{n}\sum_{n=1}^{n}\sqrt{\|\mathbf{I}_{gt} - \mathbf{I}_c\|^2 + \epsilon^2}, \tag{1}$$

where $\mathbf{I}_{gt}$ and $\mathbf{I}_c$ represent the ground truth and shadow-free images generated from different networks, respectively. In addition, ϵ is seen as a tiny constant (*e.g.*, 10^{-5}) for stable and robust convergence, and n represents the total number of input images in a single iteration.

In addition to the pixel-level content loss, we employ auxiliary losses based on frequency domain information to complement our network. To enhance the restoration of frequency domain information, we further utilize the FFT loss, which is mathematically defined as follows:

$$\mathcal{L}_{frequency} = \frac{1}{n}\sum_{n=1}^{n} \left\| \mathcal{F}\left(\mathbf{I}_{gt}\right) - \mathcal{F}\left(\mathbf{I}_{c}\right) \right\|_1. \tag{2}$$

The structural similarity index (SSIM) is a metric that quantifies the similarity between two images by considering a combination of luminance, contrast, and structural similarities. In order to improve the fidelity of the restored images, we additionally utilize the SSIM loss:

$$\mathcal{L}_{ssim} = \frac{\left(2\mu_{\mathbf{I}_{gt}}\mu_{\mathbf{I}_{c}} + c_1\right)\left(2\sigma_{\mathbf{I}_{gt}\mathbf{I}_{c}} + c_2\right)}{\left(\mu^2_{\mathbf{I}_{gt}} + \mu^2_{\mathbf{I}_{c}} + c_1\right)\left(\sigma^2_{\mathbf{I}_{gt}} + \sigma^2_{\mathbf{I}_{c}} + c_2\right)}, \tag{3}$$

where μ and σ denote the mean and standard deviation of image intensities, respectively. The term $\sigma_{\mathbf{I}_{gt}\mathbf{I}_{c}}$ represents the covariance between the two images. Furthermore, the constants c_1 and c_2 are included to prevent division by zero.

During the initial training step, we exclusively focus on training the encoder-decoder architecture of VIT. This step aims to primarily eliminate shadows from the image and reconstruct background pixels. The total loss utilized during this stage is defined as:

$$\mathcal{L}_{step1} = \mathcal{L}_{content} + \lambda\mathcal{L}_{frequency}, \tag{4}$$

where λ denotes the balanced weight and we empirically set λ to 0.1 as default.

During the second training step, we keep the parameters of VIT frozen and focus solely on training the Refine Large-scale Image network (*i.e.*, compact NAFNet). The total loss utilized during this stage remains consistent with the first stage:

$$\mathcal{L}_{step2} = \mathcal{L}_{step1} = \mathcal{L}_{content} + \zeta\mathcal{L}_{frequency}, \tag{5}$$

where ζ is assigned a constant value of 0.02.

During the last training step, we perform simultaneous fine-tuning of both stages of the model (e.g., VIT and compact NAFNet) to achieve the ultimate refinement effect. The total loss function utilized at this stage is defined as:

$$\mathcal{L}_{step3} = \mathcal{L}_{content} + \tau\mathcal{L}_{ssim}, \tag{6}$$

where τ represents a weight coefficient, which is set to 0.05.

Algorithm 2 Dynamic High Resolution During Testing

Input: the shadow images $\mathbf{I}_s : tensor(N, C, H, W)$
Output: the clean images $\mathbf{I}_{sf} : tensor(N, C, H, W)$

$N, C, H, W = \mathbf{I}_s.shape$
/* PS and OS are the size of patch and overlap */
$PS, OS = Crop\_Patch\_Size, Overlap\_Size$
$m = (H//\left(PS - OS\right) + 1)$
$n = (W//\left(PS - OS\right) + 1)$
/* M is the number of cropped images after overlap. */
$M = (H//\left(PS - OS\right) + 1)\left(W//\left(PS - OS\right) + 1\right)$
for $image$ in $\mathbf{I}_s\left(N\right)$ **do**
  if $(h, w)\,\%\left(PS - OS, PS - OS\right) == 0$ **then**
    $\mathbf{I}\left(i\right) : tensor(M, C, PS, PS) \leftarrow split\left(\mathbf{I}_s\left(i\right)\right)$
    $\mathbf{Position}\left(\mathbf{a}, \mathbf{b}\right) \leftarrow \left(\mathbf{h}, \mathbf{w}\right)$
$\mathbf{I}_l : tensor\left(M \cdot N, C, \frac{H}{m}, \frac{W}{n}\right) \leftarrow reshape\left(\sum_{i=1}^{N} \mathbf{I}\left(i\right)\right)$
$\mathbf{I}_i : tensor\left(M \cdot N, C, \frac{H}{m}, \frac{W}{n}\right) \leftarrow \mathbf{Model}_1\left(\mathbf{I}_l\right)$
$\mathbf{I}_v : tensor(N, C, H, W) \leftarrow merge\left(\mathbf{I}_i, \mathbf{Position}\right)$
$\mathbf{I}_{sf} : tensor(N, C, H, W) \leftarrow \mathbf{Model}_2\left(\mathbf{I}_v\right)$
return $\mathbf{I}_{sf}$

4. Experiments

4.1. Dataset

The NTIRE image shadow removal dataset provided by the organizing committee for the year 2024 is partitioned into three sections: the training set (ntire24-train), the validation set (ntire24-valid), and the test set (ntire24-test). This dataset comprises paired images, consisting of shadowed images and their corresponding ground truth images without shadows. All images have dimensions of $3 \times 1440 \times 1920$. The training set consists of 1000 pairs of input and ground truth images, while the validation set comprises 100 pairs of input and ground truth images. The test set includes only 75 input images, and the final evaluation will be based on the scores obtained from submissions to the competition server. Therefore, in this study, we will utilize the training set for training and validation purposes, while the validation set will be used to assess the model's performance. Furthermore, the organizing committee has also released the NTIRE image shadow removal dataset for the year 2023 [35], which has a similar composition. We will incorporate this dataset for training concurrently.

In addition to the official datasets mentioned above, we incorporated synthetic shadow datasets to improve the generalization performance of the shadow removal model. To accomplish this, we utilized the ntire24-train and ntire23-train datasets to train a shadow generator using a GAN network [15]. Subsequently, we generated 1000 paired synthetic shadow datasets on ntire24-train-GT and ntire23-train-GT, respectively. These synthetic shadow datasets were randomly mixed with the officially released shadow datasets for training, thereby diversifying the training data.

4.2. Implementation Details

We implement our proposed HirFormer image restoration network via the PyTorch 1.8 platform. Adam [24] optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is

Table 1. Officially quantitative evaluations for the submissions corresponding to the *Track 1 (fidelity)* of the NTIRE 2024 Image Shadow Removal Challenge on the *ntire24-test* dataset [36]. Our method achieved outstanding performance, winning the competition(**1st place**).

Rank	Team	PSNR↑	SSIM↑	LPIPS↓	Params (M)	Runtime	Device
1$_{/18}$	**LUMOS(ours)**	24.78$_{(2)}$	0.832$_{(2)}$	0.110$_{(4)}$	23	6.00 s	RTX3060
2$_{/18}$	Shadow_R	24.58$_{(3)}$	0.832$_{(1)}$	0.098$_{(2)}$	376	2.55 s	RTX2080Ti
3$_{/18}$	ShadowTech_Innovators	24.81$_{(1)}$	0.832$_{(3)}$	0.111$_{(5)}$	26	3.15 s	A40
4$_{/18}$	LVGroup_HFUT	24.35$_{(4)}$	0.823$_{(6)}$	0.082$_{(1)}$	17	3.46 s	RTX4090
5$_{/18}$	USTC_ShadowTitan	24.04$_{(5)}$	0.827$_{(4)}$	0.104$_{(3)}$	83	6.00 hours	A40
6$_{/18}$	GGBond	23.87$_{(6)}$	0.824$_{(5)}$	0.127$_{(6)}$	8.895	4.50 s	A6000

adopted to optimize HirFormer. Additionally, we introduce the progressive training strategy [58] and the specific training phase of HirFormer could be divided into three steps:

Step 1: Dynamic High Resolution Transformer. We use progressive training strategy at first. We start training with patch size 256×256 and batch size 24 for 40K iterations. The patch size and batch size pairs are updated to [(512, 12),(1024 , 5),(1408, 3)] at iterations [36K, 24K, 24K]. The initial learning rate is 4×10^{-4} and remains unchanged when patch size is 1408. Later the learning rate changes with Cosine Annealing scheme to 8×10^{-5}. Due to the high computational complexity of vision transformers, it is generally challenging to use large patch sizes. Therefore, to enhance the model's performance on large-scale images, we employed a dynamic high-resolution approach. This approach involves dividing the input image into smaller patches at a certain ratio. Among the three options of 2×2, 4×4, and 8×8 divisions, we selected the 4×4 division scheme that exhibited better performance during the validation phase. After processing all the smaller patches, a merging method is applied to gradually reconstruct the high-resolution image, thus enabling error backpropagation for training the restoration model. The first step performs on the NVIDIA 4090 device. We obtain the best model at this step as the initialization of the second step.

Step 2: Large-scale Image Refinement Network. Due to the presence of residual shadows in the overall image after restoration using the vision transformer, along with the edge artifacts caused by the block effect, the visual perception is significantly affected. To address this, we introduced a scaled-down version of the NAFNet network for refinement, added after the restoration network. This two-stage model, called HirFormer, was formed. In the second training step, the parameters of the vision transformer blocks were no longer updated, focusing solely on training the added NAFNet network for further refinement. We start training with patch size 1408 and batch size 1. The initial learning rate is 8×10^{-5} and changes with Cosine Annealing scheme to 1×10^{-7}, including 40K iterations in total. Exponential Moving Average (EMA) is applied for the dynamic adjustment of model parameters. The second step performs on the NVIDIA 4090 device.

Step 3: Fine-tuning HirFormer. In order to further improve the ultimate performance of the entire model, we performed synchronous fine-tuning on both stages of the network. This enabled us to optimize the parameters of the entire model concurrently. We start training with patch size 1408 and batch size 1. The initial learning rate is 5×10^{-5} and changes with Cosine Annealing scheme to 4×10^{-8}, including 40K iterations in total. The last step performs on the NVIDIA A40 device.

The training strategy for the three steps is also illustrated in Figure 2. In the testing phase, we adopt the model after fine-tuning to achieve the best performance. 12G GPU memory is enough to infer our model, and we use one NVIDIA 3060 GPU with 12G memory for testing.

4.3. Comparations

Based on prior method [58], we employ three reference-based metrics to verify the effectiveness of our method: Peak Signal-to-Noise Ratio (PSNR), the structural similarity (SSIM) [42], and Learned Perceptual Image Patch Similarity (LPIPS) [49]. For the PSNR and SSIM metrics, higher is better. For the LPIPS metric, lower means better.

Results of Challenge. Table 1 presents a comparative analysis of our proposed method with the **top six** competing teams in the fidelity track of the NTIRE 2024 image shadow removal challenge. The metrics PSNR, SSIM, and LPIPS represent the average values computed on the entire ntire24-test dataset after the final testing submission. It can be observed from Table 1 that our method exhibits outstanding performance across all metrics among the participating teams, securing the **top position** overall in the track. Additionally, our model boasts a small parameter size of 23M.

Comparison with Previous SOTA Methods To compare the performance of our proposed HirFormer with several state-of-the-art image restoration methods that have demonstrated superior performance, we selected three models published in different journals: ShadowFormer [17], SwinIR [28], and ShuffleFormer [46]. In order to ensure fair evaluation, we used the same training data and methodology, and assessed their performance on the same test set (ntire24-valid). Their visual results are shown in Figure 3, while the metrics are presented in Table 2. It proves

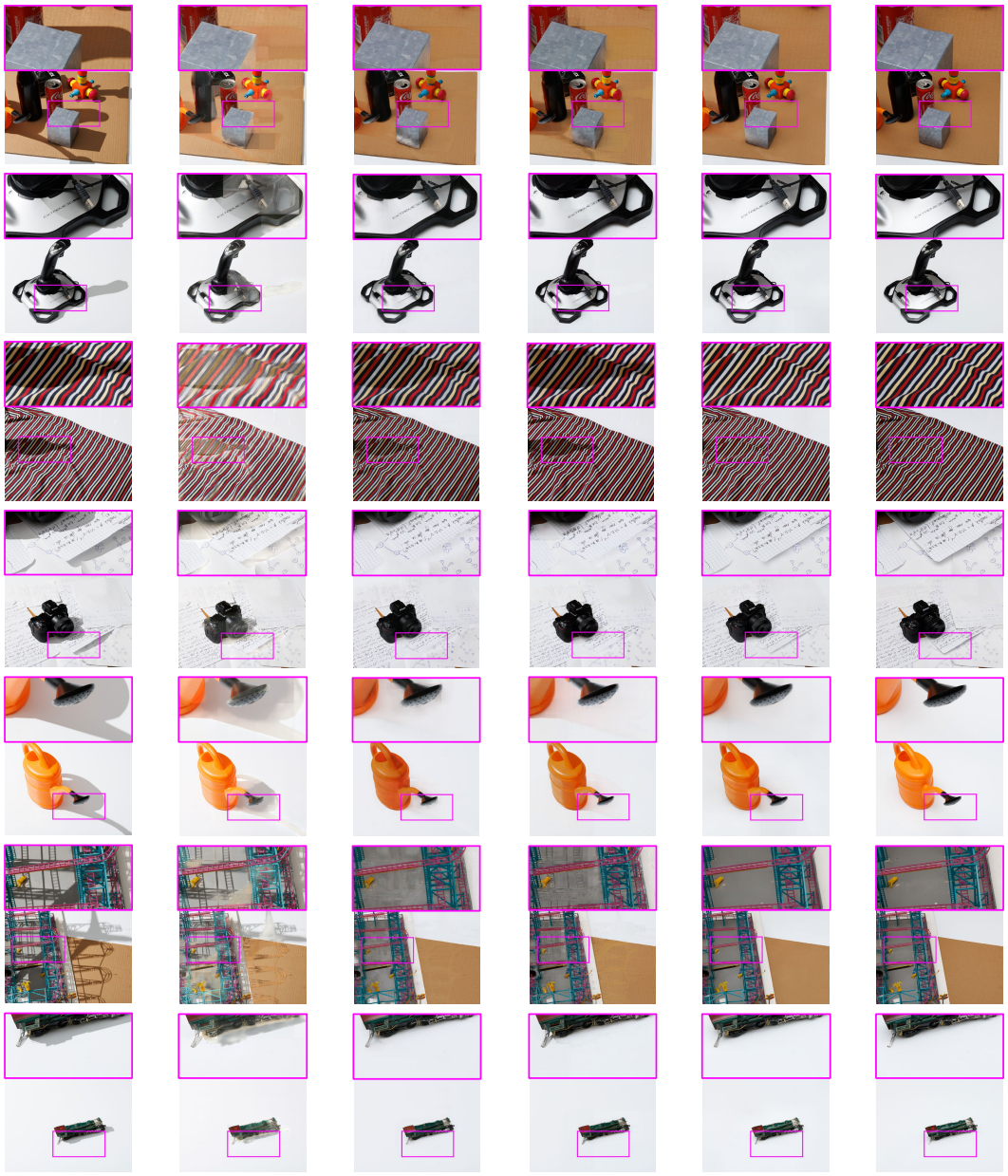

Figure 3. Visual comparison results of shadow removal on the ntire24-valid dataset. (a) to (c) are the estimated results from previous SOTA methods: ShadowFormer [17], SwinIR [28] and ShuffleFormer [46], respectively.

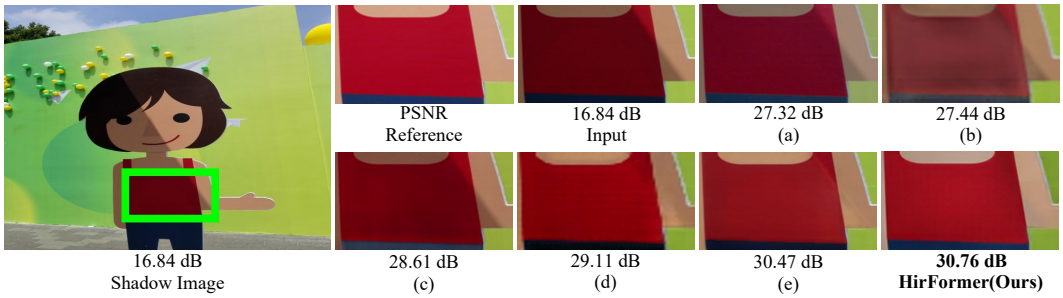

Figure 4. Visual comparison results on ISTD [40] dataset. (a) to (e) are the results from previous SOTA methods: DSC [19], SpA-Former [51], DHAN [5], DC-ShadowNet [21] and ShadowFormer [17], respectively. We did not train on ISTD [40] dataset, but only generalized validation, demonstrating that our method has excellent generalization ability on low resolution shadow datasets.

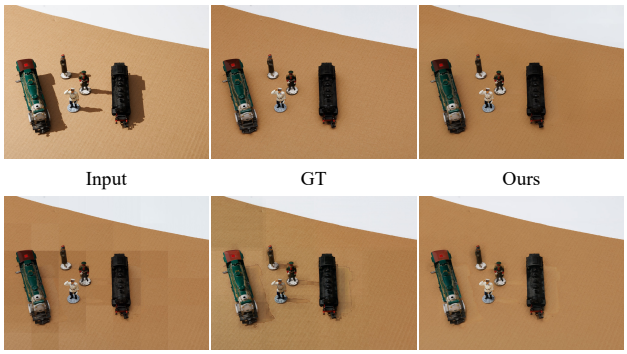

Figure 5. Visual results of ablation study for different settings in Stage 1 and Stage 2.

that HirFormer outperforms existing methods in terms of objective metrics and achieves superior shadow removal results, exhibiting a closer approximation to the ground truth in the restoration of complex background information. In addition, Figure 4 illustrates that the untrained HirFormer can still easily generalize to the low-resolution dataset ISTD [40], surpassing the performance of the previous SOTA methods.

4.4. Ablation Study

Analysis of the effects of Refine Large-scale Image(Stage 2). Our proposed method is a two-stage model, with Stage 2 serving for further refinement of the images. We experimentally examined the impact of Stage 2 on the overall performance of the model. As observed in Table 3, removing Stage 2 significantly reduces the PSNR and RMSE metrics. Additionally, in Figure 5, the visual results demonstrate an increased presence of residual artifacts and block effects when Stage 2 is excluded.

Table 2. Quantitative evaluation of HirFormer in comparison to existing methods on the ntire24-valid dataset.

Method	PSNR↑	SSIM↑
ShadowFormer [17]	22.907	0.819
SwinIR [28]	23.257	0.814
ShuffleFormer [46]	24.724	0.821
Ours	**26.319**	**0.845**

Table 3. Ablation for different settings in Stage 1 and Stage 2.

Models	PSNR↑	SSIM↑
HirFormer & w./o Stage 2	23.024	0.807
HirFormer & 2 × 2 grid in Stage 1	24.691	0.832
HirFormer & 8 × 8 grid in Stage 2	25.337	0.811
HirFormer	**26.319**	**0.845**

Figure 6. Visual results of HirFormer limitations.

Analysis of the effects of grid splitting type used in Dynamic High Resolution Algorithm 1(Stage 1). The experiments validated the impact of different grid splitting types in the Dynamic High Resolution Algorithm 1 on the performance of HirFormer. As shown in Table 3, using both 2×2 and 8×8 grid splitting methods resulted in a decrease in objective metrics. Furthermore, Figure 5 indicates that changing the grid splitting type exacerbates the residual shadows.

4.5. Limitation

Our proposed HirFormer effectively removes shadows from high-resolution images and reconstructs pixel details. However, it still has its limitations. As shown in Figure 6, in two complex scenes, although our approach successfully removes a significant portion of the shadows in the images, it also introduces color darkening and slight distortion in the background information. This may be attributed to the model leveraging shadow diffusion in the background when dealing with images where there is minor color difference between the shadows and the background.

5. Conclusion

In this paper, we propose HirFormer, a novel approach for efficient shadow removal in high-resolution images. HirFormer consists of two stages: Dynamic High Resolution Transformer and Large-scale Image Refinement. The first stage is based on the VIT model to reconstruct fine pixel details and remove a substantial portion of shadows. The second stage incorporates the NAFNet model to further eliminate residual shadows and address block artifacts introduced by the first stage. Experimental analysis demonstrates the strong competitiveness of HirFormer in shadow removal and high-resolution background reconstruction. It effectively restores large-scale degraded images affected by shadows, outperforming existing methods across various objective metrics. In addition to winning the NTIRE2024 image shadow removal challenge, we believe that HirFormer also holds great potential for other high-resolution image restoration tasks.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 4

[2] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, pages 168–172 vol.2, 1994. 4

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines fornbsp;image restoration. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, page 17–33, Berlin, Heidelberg, 2022. Springer-Verlag. 2, 4

[4] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4723–4732, 2021. 2

[5] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan, 2019. 7

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10212–10221, 2019. 1, 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3, 4

[8] Graham D. Finlayson, Steven D. Hordley, Cheng Lu, and Mark S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:59–68, 2006. 2

[9] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85:35–57, 2009. 2

[10] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10566–10575, 2021. 2

[11] Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2090–2104, 2021. 1

[12] Xueyang Fu, Xi Wang, Aiping Liu, Junwei Han, and Zheng-Jun Zha. Learning dual priors for jpeg compression artifacts removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4066–4075, 2021.

[13] Xueyang Fu, Menglu Wang, Xiangyong Cao, Xinghao Ding, and Zheng-Jun Zha. A model-driven deep unfolding method for jpeg artifacts removal. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6802–6816, 2022. 1

[14] Han Gong and Darren P. Cosker. Interactive removal and ground truth for difficult shadow scenes. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 33 9:1798–811, 2016. 2

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, 2014. 5

[16] Maciej Gryka, Michael Terry, and Gabriel J. Brostow. Learning to remove soft shadows. *ACM Trans. Graph.*, 34(5), 2015. 1

[17] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: global context helps shadow removal. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2023. 1, 3, 6, 7, 8

[18] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 2012. 1, 2

[19] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018. 2, 7

[20] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2472–2481, 2019. 3

[21] Yeying Jin, Aashish Sharma, and Robby T. Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5007–5016, 2021. 3, 7

[22] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2343–2353, 2022. 4

[23] Salman Hameed Khan, Bennamoun, Ferdous Sohel, and Roberto B. Togneri. Automatic shadow detection and removal from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:431–446, 2016. 1

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[25] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, page 264–281, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[26] Hieu M. Le and Dimitris Samaras. Shadow removal via shadow image decomposition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8577–8586, 2019. 2

[27] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. Edge-aware regional message passing controller for image forgery localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8222–8232, 2023. 1

[28] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 4, 6, 7, 8

[29] Yun-Hsuan Lin, Wen-Chin Chen, and Yung-Yu Chuang. Bedsr-net: A deep shadow removal network from a single document image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12902–12911, 2020. 2

[30] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing*, 30:1853–1865, 2021. 2, 3

[31] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4925–4934, 2021. 1, 2, 3

[32] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2308–2316, 2017. 2

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

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2020. 3, 4

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

[36] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 6

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Tomas F. Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:682–695, 2018. 1

[39] J. Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Y. Liu, and Song Wang. Style-guided shadow removal. In *European Conference on Computer Vision*, 2022. 2

[40] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 2, 7, 8

[41] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021. 4

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[43] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma. Fast shadow removal using adaptive multi-scale illumination transfer. *Computer Graphics Forum*, 32, 2013. 1

[44] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zhengjun Zha. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:12978–12995, 2022. 3, 4

[45] Jie Xiao, Xueyang Fu, Feng Wu, and Zhengjun Zha. Stochastic window transformer for image restoration. In *Neural Information Processing Systems*, 2022. 4

[46] Jie Xiao, Xueyang Fu, Man Zhou, HongJiang Liu, and Zhengjun Zha. Random shuffle transformer for image restoration. In *International Conference on Machine Learning*, 2023. 3, 4, 6, 7, 8

[47] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5790–5799, 2020. 4

[48] Qingxiong Yang, K. H. Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image Processing*, 21:4361–4368, 2012. 2

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[50] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41 (3):611–624, 2019. 1

[51] Xiaofeng Zhang, Yudi Zhao, Chaochen Gu, Changsheng Lu, and Shanying Zhu. Spa-former:an effective and lightweight transformer for image shadow removal. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. 7

[52] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886, 2021. 4

[53] Yurui Zhu, Xueyang Fu, Chengzhi Cao, Xi Wang, Qibin Sun, and Zheng-Jun Zha. Single image shadow detection via complementary mechanism. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 6717–6726, New York, NY, USA, 2022. Association for Computing Machinery. 1

[54] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng–Jun Zha. Bijective mapping network for shadow removal. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5617–5626, 2022. 1, 2

[55] Yurui Zhu, Zeyu Xiao, Yanchi Fang, Xueyang Fu, Zhiwei Xiong, and Zhengjun Zha. Efficient model-driven network for shadow removal. In *AAAI Conference on Artificial Intelligence*, 2022. 1

[56] Yurui Zhu, Xueyang Fu, Zheyu Zhang, Aiping Liu, Zhiwei Xiong, and Zhengjun Zha. Hue guidance network for single image reflection removal. *IEEE transactions on neural networks and learning systems*, PP, 2023. 3

[57] Yurui Zhu, Tianyu Wang, Xueyang Fu, X. Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiao hua Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21747–21758, 2023. 3, 4

[58] Yurui Zhu, Xi Wang, Xueyang Fu, and Xiaowei Hu. Enhanced coarse-to-fine network fornbsp;image restoration fromnbsp;under-display cameras. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, page 130–146, Berlin, Heidelberg, 2023. Springer-Verlag. 6