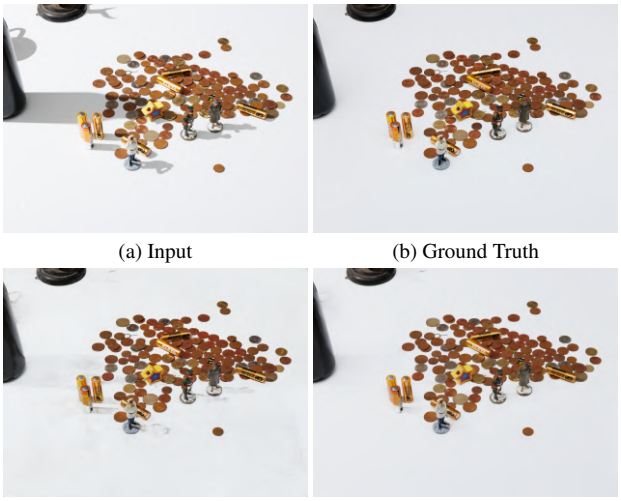
TSRFormer: Transformer Based Two-stage Refinement for Single Image Shadow Removal

Hua-En Chang[2*], Chia-Hsuan Hsieh[4*], Hao-Hsiang Yang[2], I-Hsiang Chen[2],
Yi-Chung Chen[3], Yuan-Chun Chiang[2], Zhi-Kai Huang[2], Wei-Ting Chen[1], Sy-Yen Kuo[2]
[1] Graduate Institute of Electronics Engineering, National Taiwan University, Taiwan
[2] Department of Electrical Engineering, National Taiwan University, Taiwan
[3] Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[4] ServiceNow, USA

Abstract

Single-image shadow removal aims to remove undesired shadow information from captured images. With the development of deep convolutional neural networks, several methods have been proposed to achieve promising performance in shadow removal. However, they still struggle with limited performance due to the non-homogeneous intensity distribution of the shadow. To address this issue, we propose a two-stage shadow removal architecture based on the transformer called TSRFormer. The proposed architecture is divided into shadow removal and content refinement networks. These two stages adopt different transformer architectures and remove the shadow based on different information to achieve effective shadow removal. Experiments performed on challenging benchmark show that the proposed model achieves the 2^{nd} highest SSIM in the NTIRE 2023 Image Shadow Removal Challenge. The source code will be public after the acceptance of this paper.



(a) Input (b) Ground Truth (c) SpAFormer [1] (d) TSRFormer

Figure 1. **Shadow removal results by the state-of-the-art and the proposed methods.** The proposed method TSRFormer can remove more undesired shadows and reconstruct more pixel information.

1. Introduction

When capturing natural images, a shadow is an inevitable phenomenon generated under the condition of the light being partially or completely blocked. Though shadows can provide fruitful information about the captured scenes, they may significantly degrade the image quality of human perception [2] [3] and then further deteriorate the performance of subsequent vision applications such as semantic segmentation, object tracking, and detection [2] [4] [3] [5] [6].

Several shadow removal algorithms have been proposed in past decades and achieved decent performance. These methods can be categorized to prior-based methods [7] [8] [9] and deep learning-based methods [10] [11] [12] [13] [14]. The prior-based strategy focuses on adopting an image formation model and exploiting the prior information of the shadow image to formulate the shadow removal problem. However, designing a comprehensive image formation model is challenging, which is usually restrictive to specific scenes and not general to real-world scenarios. Thus, the performance of this strategy is usually limited.

On the other hand, similar to other low-level vision tasks like single image dehazing and image enhancement [15,16], deep learning-based methods adopt the convolutional neural network and large-scale datasets [17] [18] [19] to learn

*Equally-contributed first authors.

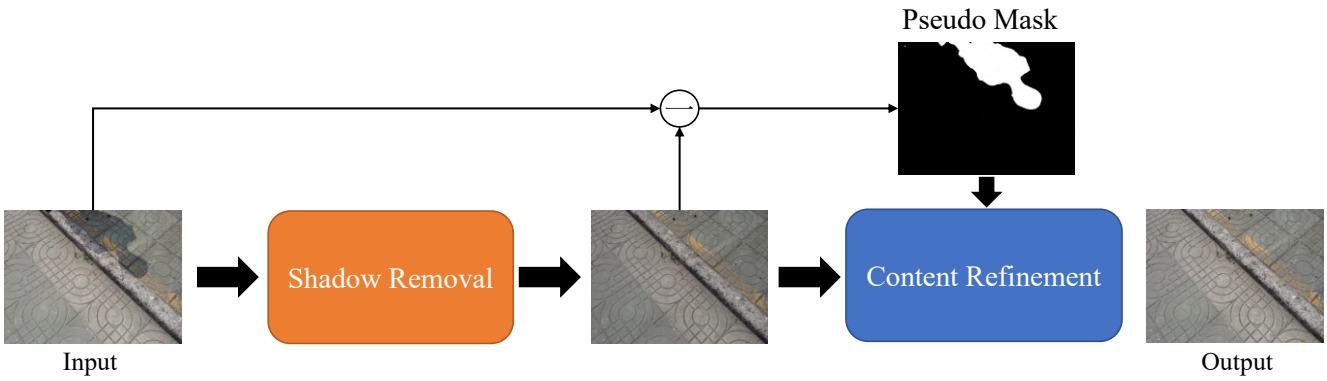


Figure 2. **Overview of the proposed TSRFormer.** The proposed TSRFormer adopts a two-stage recovery pipeline. The first stage is to remove the shadow globally, while the second stage is to recover the residual shadow and refine the content locally.

the mapping function from shadow images to shadow-free images. Though these methods have achieved promising performance in shadow removal, they still suffer from undesired results. Specifically, existing shadow removal methods do not address the non-homogeneous distribution of intensity of shadow well. The regions with strong intensity of shadow cannot be removed clearly, which limits the performance of shadow removal. As shown in Figure 1 the reconstructed results of existing methods are not pleasant due to the non-homogeneous distribution of intensity of the shadow. To this end, we proposed a novel two-stage refinement single image shadow removal method which consists of shadow removal and content refinement stages. The first stage aims to remove most shadow regions globally and the second stage focuses on removing the residual shadow based on the results of the first stage and compensating for the missing information locally. In each stage, we adopt a vision transformer architecture with different information as inputs to conduct shadow removal sequentially. As shown in Figure 1, the network can learn more robust shadow removal with this two-stage training paradigm. Extensive experimental results show that the proposed TSRFormer performs robustly and favorably against the state-of-the-art schemes for single-image shadow removal. We make the following contributions to this work:

- We present the TSRFormer for single-image shadow removal. Our method uses a two-stage reconstruction pipeline and two transformer architectures that can learn robust shadow removal.
- We test our proposed method on the dataset provided by NTIRE 2023 shadow removal challenge and achieve the 2^{nd} highest SSIM in this challenge. Moreover, several experiments demonstrate the effectiveness of the proposed TSRFormer.

2. Related Works

2.1. Image Shadow Removal

Existing image shadow removal can be mainly divided into two classes:

Prior-based Methods [20] [21] [22] [23] [9] [7] [24] make assumptions based on statistical analysis and human observation such as region information [22, 25], illumination [8, 9, 26, 27], and image gradients [7, 20] to design shadow removal formulation. For example, Guo *et al.* [22] find the variation between shadow and shadow-free regions, and they adopt this property to distinguish and reconstruct shadow-free results. Finlayson *et al.* [20] proposed to leverage illuminant invariance and gradient information to remove shadow edge and then adopt image inpainting technique to recover results. To further improve this work, they proposed introducing the entropy for pixel values to suppress the shadow in captured images. Gryka *et al.* [7] adopt a mapping function for image patches to perform shadow removal. However, this kind of strategy may suffer from limited performance when the assumptions of the scenes are failed, which causes poor generalization ability in comprehensive scenarios.

Deep Learning-based Methods [19] [18] [13] [28] [29] [30] [10] [30] leverage the large-scale datasets to train the conventional neural networks, and it can be split to supervised and unsupervised strategies.

For the supervised strategy, Le *et al.* [29] proposed a physical illumination model and an image decomposition formulation to remove the shadow. Zhu *et al.* [31] proves the shadow removal and generation process can benefit the whole training procedure. They proposed a unified network to train these two processes simultaneously. Chen *et al.* [12] developed a context-aware network to integrate the information of shadow-free regions and shadow regions in the latent feature spaces. Wan *et al.* [32] found the style incon-

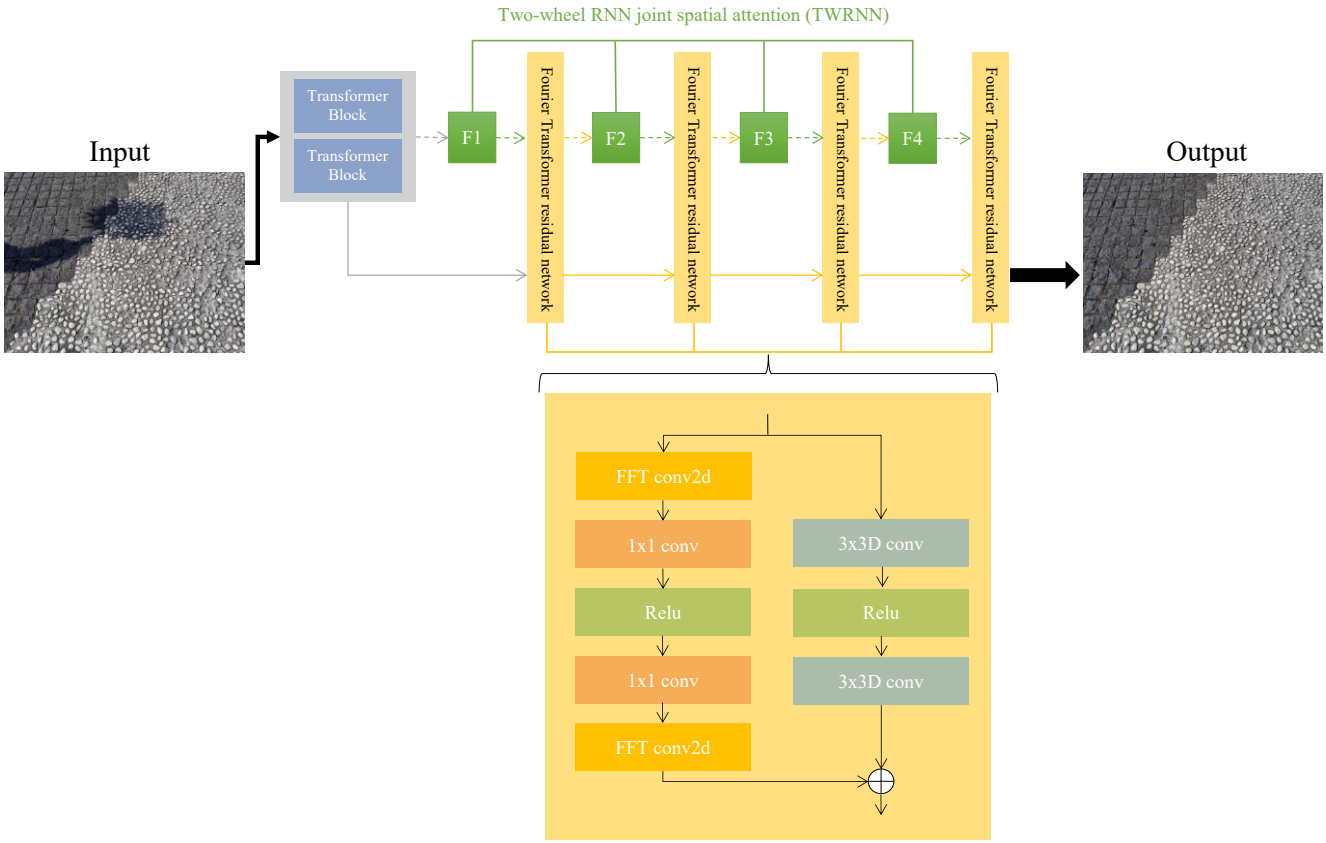


Figure 3. **Architecture of the shadow removal network.**

sistency problem between shadow and shadow-free regions after the shadow removal. To solve this problem, they proposed a style-guided shadow removal network to exploit the shadow-free regions to adjust the recovered shadow regions for consistent style. Fu *et al.* [33] solved the shadow removal problem by reformulating it to an exposure fusion problem. By generating the weight maps for images with different exposures, shadow-free results can be obtained by the fusion technique. Moreover, some methods [1] [34] adopted the transformer [35] [36] as the backbone to capture global information for shadow removal.

For unsupervised strategy, several shadow removal methods [37] [10] [38] [39] have emerged. They are generally based on a generative adversarial network (GAN) to train the network with unpaired shadow and clean images. Jin *et al.* [37] exploited an unsupervised domain classifier to guide the shadow removal network and proposed physics-based shadow-free chromaticity to constrain the network.

Although these methods achieve remarkable performance for shadow removal in most scenes, they still suffer from limited performance since they neglect to consider the non-homogeneous intensity distribution of shadow. Thus, developing a solution to cope with this problem is of great importance.

2.2. Transformer-Based Image Restoration

Initially, the transformer [40] containing the self-attention module has shown impressive performance in natural language processing (NLP) tasks. Besides NLP tasks, self-attention modules containing spatial and channel attention are leveraged in many computer vision applications [16, 41–43]. Recently, the vision transformer has achieved significant success in the computer vision community. This architecture decomposes an image into a series of patches with sequences and learns mutual relationships. Based on this technique, several vision tasks such as object detection [44] [45] [46], segmentation [47, 48], and image recognition [35, 36, 49]. Recently, several image restoration techniques have adopted this architecture as the backbone and achieve state-of-the-art performance, including super-resolution [50, 51], denoising [52, 53], deraining [53], dehazing [54, 55], and desnowing [56, 57].

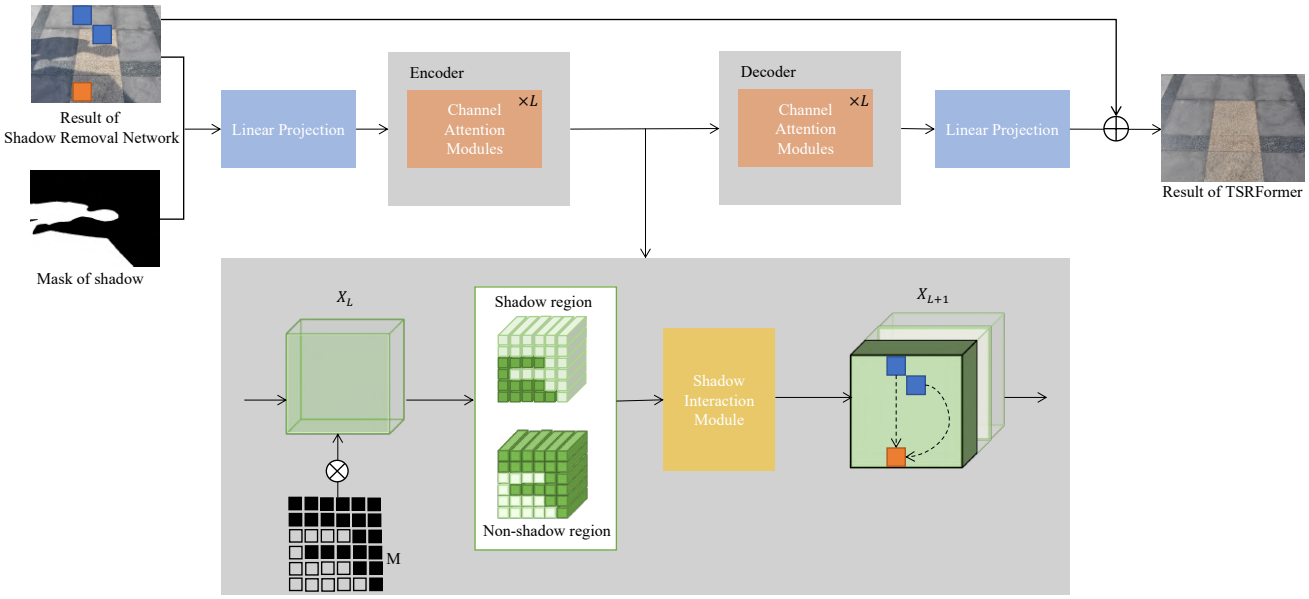


Figure 4. **Architecture of the content refinement stage.**

3. Proposed Methods

3.1. Two-stage Shadow Removal Pipeline

Existing shadow removal cannot generate a shadow-free image with decent quality due to the non-homogeneous intensity of the shadow. To address this issue, inspired by other image restoration tasks [15, 58], we proposed a novel two-stage shadow removal pipeline. As shown in Figure 2, our shadow removal can be divided into two parts: (i) the shadow removal network (SRN) and (ii) the content refinement network (CRN).

- **Shadow Removal Network** aims to remove the shadow from the input shadow image globally. This network removes most parts of the shadow and generates a rough shadow removal result at this stage.
- **Content Refinement Network** focuses on removing the residual shadow and compensating for the missing pixel information locally at this stage. We compute the shadow region based on the difference between the input shadow image and the recovered result obtained by the shadow removal network. Based on the shadow region, we can conduct content refinement and residual shadow removal locally.

3.2. Shadow Removal Network

The structure of the shadow removal network is shown in Figure 3. We adopt the architecture of SpA-Former [1] as the backbone of our shadow removal network. It contains a deep Transformer and CNN network.

First, the input image is passed to the Transformer block to capture local and global information. Then, a 3×3 convolution block is adopted to extract the feature maps. These feature maps are passed to two wheel RNN joint spatial attention network (TWRNN) and the bottleneck network. The TWRNN aims to help the network focus on specific elements by generating attention maps from the inputs.

Transformer block enables the network to capture global and local information. This module complements the TWRNN network since the result of the transformer does not depend on the previous step, which enables the network can process in a parallel fashion. By this design, more accurate feature learning can be achieved. The encoder of the transformer reduces the dimension and increases the number of channels. The low-level features of the encoder and high-level features of the decoder can be aggregated in this block. This operation can benefit the network to keep both structural and textural details in the results.

Two-Wheel RNN joint spatial Attention is a two-round four-way RNN architecture that can project the features in four directions, which can extract spatial context information to focus on desired shadow features. First, three vanilla residual blocks are adopted to extract features for the attention residual block to remove the shadow. Then, the feature map without shadow is fed into two residual blocks to generate the shadow-free image.

Joint Fourier Transform Residuals Module [59] is an improved module by the ResBlock, and it can compute the difference between shadow and clear images. The vanilla Resblock tends to have limited performance in extracting low-frequency information. It is not beneficial to reconstruct the shadow-free image since the reconstruction process includes low and high-frequency information recovery.

As shown in Figure 2, high-frequency and low-frequency information can be captured in the reconstruction process using the Fourier transform residual modules. Also, to improve the ability of the network to capture long-range information, the residual information is integrated with the attention map to improve the residual learning. Then, the result of the shadow removal network is generated.

3.3. Content Refinement Network

In this stage, the content refinement network aims to remove residual shadow and recover the missing pixel information locally. To this end, we adopt the Shadowformer [34] as the backbone at this stage. We adopt the image generated by the shadow removal network and the mask of shadow as inputs in this stage. The mask of shadow can be computed by:

$$M(x,y) = \theta(|I_{Input}(x,y) - I_{Shadow}(x,y)|) \quad (1)$$

where M, I_{Input}, and I_{Shadow} denote the shadow mask, the input shadow image, and the result generated by the shadow removal network. $\theta(\cdot)$ denotes the threshold operation. Then, based on these inputs, the content refinement network generates the final result of the shadow removal.

Architecture. As shown in Figure 4, first, a linear projection block is adopted to extract the low-level feature from the input. These features are fed into the transformer with several channel attention modules to obtain the multi-scale global features. The encoder consists of several down-sampling blocks and channel attention blocks, while the decoder contains several up-sampling blocks and channel attention blocks. The channel attention block integrates the spatial information and long-range correlation by a multi-layer perceptron [60]. The channel attention modules repeat L times to obtain the hierarchical features. Then, we adopt Shadow-Interaction Module [34] to capture the global contextual correlation using previous features. The generated features are fed into the decoder and passed several channel attention modules with the features skip-connected from the encoder. Last, the final result of the shadow removal is obtained by a linear projection operation.

3.4. Loss Functions

To train the proposed TSRFormer, we apply different loss functions in different stages. For the shadow removal network, we adopt L_1 loss to calculate the difference between the recovered image and the shadow-free image. In addition, we adopt the L_2 loss function for mask and attention, and the softplus function for GAN loss and discriminator. We adopt the Charbonnier loss [61] that can be regarded as the robust L_1 loss function for the content refinement stage. The Charbonnier loss is expressed as

$$L_{Cha}(I, I_{Content}) = \frac{1}{T}\sum_{i}^{T}\sqrt{(I_i - I_{Content})^2 + \epsilon^2} \quad (2)$$

where I and $I_{Content}$ represent the ground truth and de-shadowed images generated from the content refinement network, respectively, and e is seen as a tiny constant (e.g., 10^{-6}) for stable and robust convergence. L_{Cha} can restore global structure [61] and can be more robust to handle outliers.

4. Experiments

4.1. Dataset

The 2023 NTIRE image shadow removal dataset is a novel dataset called WSRD [62] with a large diversity of contents. This dataset consists of shadow-affected images and shadow-free images. There are 1000 images for training, 100 for validation, and 100 for testing. Additionally, shadow-free images are only provided in the training set. The size of all images is 1440 × 1920 × 3. Furthermore, we also adopt the extra ITDS [18] dataset. ITDS contains 1330 training and 540 testing triplets (e.g., shadow images, masks, and shadow-free images). We initially used ITDS to train our model before we adopted the 2023 NTIRE image shadow removal dataset to implement two-stage training.

4.2. Experimental Setting

There are two stage training phases. First, for SpA-Former, the image is randomly cropped as 360 × 480, and we do not use other data augmentation tricks. The Adam optimizer [63] is adopted and the batch size is set to 3 per card. We train the network for 200 epochs with the momentum β_1 = 0.5, β_2 = 0.999. The learning rates of the generator and the discriminator are initialed as 4×10^{-4} and 3.2×10^{-3}. Second, for ShadowFormer, the image is randomly cropped as 320 × 320, and we do not use other data augmentation tricks. The AdamW optimizer [64] is utilized with a batch size of 8 to train the network. We train the network for 500 epochs with the momentum β_1 = 0.5, β_2 = 0.999 and the weight delay = 0.02. The learning rate is initialed as 2×10^{-4}. We use the Charboinner loss function to optimize the network. We perform our experiments on four Nvidia RTX 3090 graphic cards based on the PyTorch platform. The model takes two days to train.

4.3. Ablation Study

To prove the effectiveness of the proposed TSRFormer, we conduct ablation studies. We adopt the training and testing set of NTIRE 2023 shadow removal dataset for training and evaluation. We use the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) as metrics for quantitative evaluation.

Figure 5. **The visual comparison of the proposed method and other existing methods in shadow removal on the WSRD dataset.**

The ablation experiments consist of three experimental settings:

- The TSRFormer is only with shadow removal network.
- The TSRFormer is only with content refinement network.
- The TSRFormer with the proposed two-stage recovery pipeline.

The quantitative and qualitative results are reported in

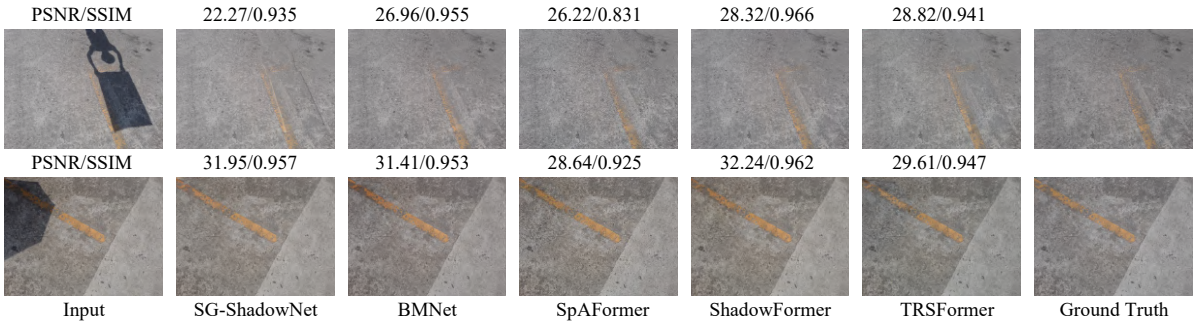


Figure 6. **The visual comparison of the proposed method and other existing methods in shadow removal in ITDS [18] dataset.**

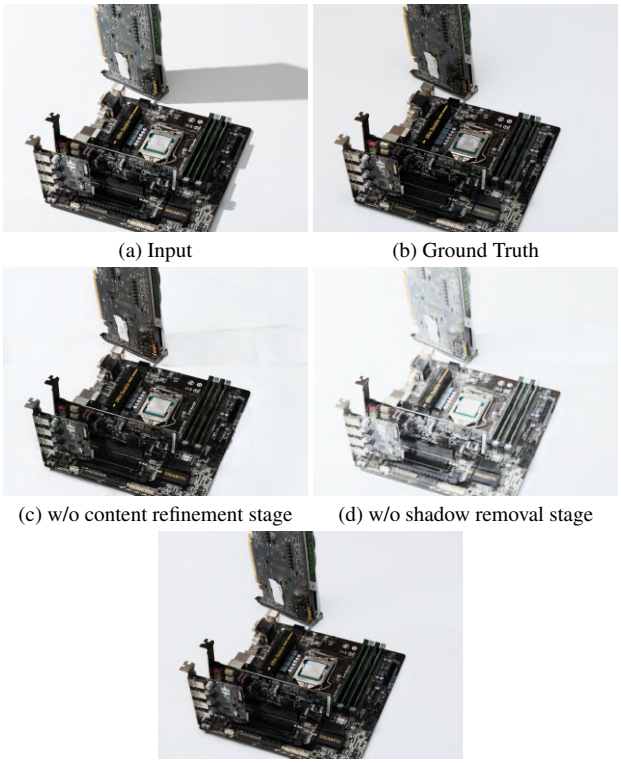


Figure 7. **Ablation study of two-stage shadow removal strategy.** We compare the result of each proposed stage qualitatively.

Table 1 and Figure 7. The PSNR and SSIM scores of the model without using shadow removal have limited performance in both PSNR and SSIM since the content refinement network cannot effectively generate correct shadow removal results without appropriate shadow removal results. Also, the model's performance without content refinement is worse than the proposed two-stage pipeline since the residual shadow and the missing information are not recovered by this local reconstruction stage.

4.4. Results of Challenge

We list the results of the proposed TSRFormer compared with other competing entries in the image shadow removal challenge of NTIRE 2023 workshop [65] in Table 2. The PSNR and SSIM values are averaged across the entire test set of each submission to evaluate the performance of all submissions. As shown in Table 2, our results obtained competitive performance in terms of PSNR and SSIM.

4.5. Comparison with Existing Methods

We adopt several existing shadow removal methods to compare the performance of the proposed method, including BMNet [31], SG-ShadowNet [32], SpAFormer [1]. For fair evaluation, we retrain their models based on their official implementation on the websites with the same training dataset adopted in this work and evaluate the performance on the same test dataset. The results are reported in Table 3. The proposed TSRFormer achieves the best performance compared to other baselines. Also, we show the visual comparison in Figure 5 and Figure 6.

4.6. Limitations and Discussion

The proposed method consists of two recovery stages. Although this method achieves competitive performance in this competition, it may fail under certain conditions. For example, if the output of the shadow removal network is not pleasant, the result of content refinement cannot generate a decent result. We show an example in Figure 8. We think this limitation can be addressed by improving the robustness of the shadow removal to ensure the quality of the recovered result in the first stage.

For future works, we think the shadow mask generation of the content refinement can be improved since we compute the shadow mask by using the difference between the recovered result by the shadow removal network and the input image directly. It is not robust since the quality of the mask depends on the reconstructed image of the shadow removal network. We think the shadow mask generation can be replaced with an independent network and optimized

Table 1. **Ablations.** The comparison of using different shadow removal stages in NTIRE 2023 shadow removal dataset.

Module	Metrics		Computational Complexity		
	PSNR	SSIM	FLOPs	Parameters	Inference Time
Only with Shadow Removal Stage (Stage 1)	23.116	0.779	88.044G	0.530M	0.252 sec/image
Only with Content Refinement Stage (Stage 2)	14.821	0.661	100.960 G	11.352M	1.810 sec/image
Two-stage Pipeline (Ours)	24.650	0.822	189.004	11.882M	2.062 sec/image

Table 2. **The results of the challenge of four methods over NTIRE 2023 Image Shadow Removal validation and testing dataset.** Our proposed method can achieve the competitive result in terms of SSIM.

User Name	Validation		Testing	
	PSNR	SSIM	PSNR	SSIM
HuanZheng	24.04 (1)	0.78 (1)	21.43 (7)	0.68 (7)
xyz123	23.94 (2)	0.76 (2)	22.20 (2)	0.69 (3)
cuishuhao	23.60 (3)	0.76 (4)	22.36 (1)	0.70 (1)
mrchang87	22.93 (6)	0.76 (3)	21.79 (3)	0.70 (2)

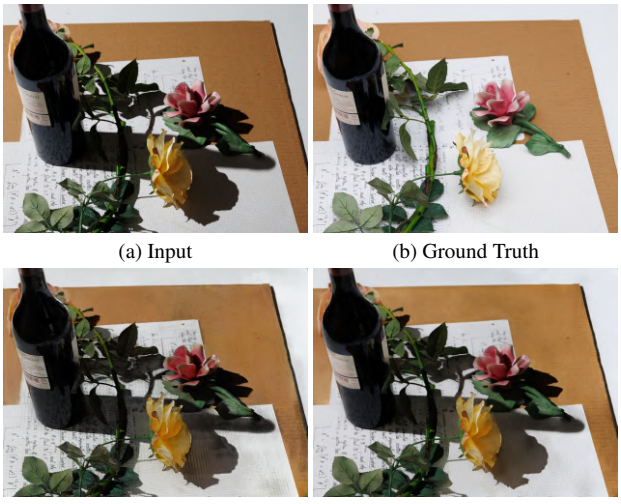
Table 3. **Quantitative Evaluation with existing shadow removal methods on NTIRE 2023 shadow removal dataset.** The proposed TSRFormer can achieve better performance compared to other baselines.

Method	Metrics	
	PSNR	SSIM
BMNet [31]	20.362	0.713
SG-ShadowNet [32]	20.023	0.675
SpAFormer [1]	23.116	0.779
Ours	24.650	0.822

with the whole architecture to improve the effectiveness of this method. Moreover, based on this improved model, we will adopt more datasets to train and conduct more comprehensive evaluations on existing datasets such as ITDS [18].

5. Conclusion

In this paper, we propose an effective single-image shadow removal solution called TSRFormer. This method contains two stages, including the shadow removal network and the content refinement network. The former part aims to remove the shadow globally, and the latter part focuses on suppressing the residual shadow and refining the content information. These two networks are both based on transformer architectures. Experimental results prove the effectiveness of the proposed method compared to existing shadow removal approaches. Moreover, in the NTIRE 2023 Image Shadow Removal Challenge, the proposed TSRFormer achieves competitive performance in terms of SSIM.



(a) Input (b) Ground Truth

(c) Result of SRN (d) Result of TSRFormer

Figure 8. **The failure case of shadow removal for the proposed TSRFormer.** The unpleasant shadow removal result will be generated when the shadow removal stage fails.

6. Acknowledgement

We thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources. This work was supported by the National Science and Technology Council, Taiwan, under Grant MOST 108-2221-E-002-072-MY3, Grant MOST 108-2638-E-002-002-MY2, and Grant NSTC 111-2221-E-002-136-MY3.

References

[1] X. F. Zhang, C. C. Gu, and S. Y. Zhu, "Spa-former: Transformer image shadow detection and removal via spatial attention," *arXiv preprint arXiv:2206.10910*, 2022. 1, 3, 4, 7, 8

[2] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1337–1342, 2003. 1

[3] S. Nadimi and B. Bhanu, "Physical models for moving shadow and object detection in video," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 8, pp. 1079–1087, 2004. 1

[4] C. R. Jung, "Efficient background subtraction and shadow removal for monochromatic video sequences," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 571–577, 2009. 1

[5] A. Sanin, C. Sanderson, and B. C. Lovell, "Improved shadow removal for robust person tracking in surveillance scenarios," in *ICPR*, 2010, pp. 141–144. 1

[6] W. Zhang, X. Zhao, J.-M. Morvan, and L. Chen, "Improving shadow suppression for illumination robust face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 611–624, 2018. 1

[7] M. Gryka, M. Terry, and G. J. Brostow, "Learning to remove soft shadows," *ACM Trans. Graph.*, vol. 34, no. 5, p. 153, 2015. 1, 2

[8] L. Zhang, Q. Zhang, and C. Xiao, "Shadow remover: Image shadow removal based on illumination recovering optimization," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4623–4636, 2015. 1, 2

[9] C. Xiao, R. She, D. Xiao, and K.-L. Ma, "Fast shadow removal using adaptive multi-scale illumination transfer," in *Comput. Graph. Forum*, 2013. 1, 2

[10] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang, "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE Trans. Image Process.*, vol. 30, pp. 1853–1865, 2021. 1, 2, 3

[11] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020. 1

[12] Z. Chen, C. Long, L. Zhang, and C. Xiao, "Canet: A context-aware network for shadow removal," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4743–4752. 1, 2

[13] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng, "Direction-aware spatial context features for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1, 2

[14] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, and B. Wen, "Shadowdiffusion: When degradation prior meets diffusion model for shadow removal," *arXiv preprint arXiv:2212.04711*, 2022. 1

[15] W.-T. Chen, Z.-K. Huang, C.-C. Tsai, H.-H. Yang, J.-J. Ding, and S.-Y. Kuo, "Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 653–17 662. 1, 4

[16] H.-H. Yang, K.-C. Huang, and W.-T. Chen, "Laffnet: A lightweight adaptive feature fusion network for underwater image enhancement," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 685–692. 1, 3

[17] T. F. Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras, "Large-scale training of shadow detectors with noisily-annotated shadow examples," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 816–832. 1

[18] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2, 5, 7, 8

[19] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau, "Deshadownet: A multi-context embedding deep network for shadow removal," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 2

[20] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 59–68, 2005. 2

[21] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *International Journal of Computer Vision*, vol. 85, no. 1, pp. 35–57, 2009. 2

[22] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2956–2967, 2012. 2

[23] Q. Yang, K.-H. Tan, and N. Ahuja, "Shadow removal using bilateral filtering," *IEEE Transactions on Image processing*, vol. 21, no. 10, pp. 4361–4368, 2012. 2

[24] L.-Q. Ma, J. Wang, E. Shechtman, K. Sunkavalli, and S.-M. Hu, "Appearance harmonization for single image shadow removal," in *Computer Graphics Forum*, vol. 35, no. 7. Wiley Online Library, 2016, pp. 189–197. 2

[25] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 682–695, 2017. 2

[26] Y. Shor and D. Lischinski, "The shadow meets the mask: Pyramid-based shadow removal," *Comput. Graph. Forum*, vol. 27, pp. 577–586, 04 2008. 2

[27] Q. Yang, K.-H. Tan, and N. Ahuja, "Shadow removal using bilateral filtering," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4361–4368, 2012. 2

[28] B. Ding, C. Long, L. Zhang, and C. Xiao, "Argan: Attentive recurrent generative adversarial network for shadow detection and removal," in *Int. Conf. Comput. Vis.*, 2019. 2

[29] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Int. Conf. Comput. Vis.*, 2019, pp. 8578–8587. 2

[30] Y.-H. Lin, W.-C. Chen, and Y.-Y. Chuang, "Bedsr-net: A deep shadow removal network from a single document image," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 12 905–12 914. 2

[31] Y. Zhu, J. Huang, X. Fu, F. Zhao, Q. Sun, and Z.-J. Zha, "Bijective mapping network for shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5627–5636. 2, 7, 8

[32] J. Wan, H. Yin, Z. Wu, X. Wu, Y. Liu, and S. Wang, "Style-guided shadow removal," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 7, 8

[33] L. Fu, C. Zhou, Q. Guo, F. Juefei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang, "Auto-exposure fusion for single-image shadow removal," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 571–10 580. 3

[34] L. Guo, S. Huang, D. Liu, H. Cheng, and B. Wen, "Shadowformer: Global context helps image shadow removal," *arXiv preprint arXiv:2302.01650*, 2023. 3, 5

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021. 3

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021. 3

[37] Y. Jin, A. Sharma, and R. T. Tan, "Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5027–5036. 3

[38] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-shadowgan: Learning to remove shadows from unpaired data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2472–2481. 3

[39] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From shadow generation to shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4927–4936. 3

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017. 3

[41] H.-H. Yang, C. Chen, I-Hsiang, C.-H. Hsieh, Hua-En, Y.-C. Chiang, Y.-C. Chen, W.-T. Huang, Zhi-Kai Chen, and S.-Y. Kuo, "Semantic guidance learning for high-resolution non-homogeneous dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 3

[42] H.-H. Yang, C.-H. H. Yang, and Y.-C. F. Wang, "Wavelet channel attention module with a fusion network for single image deraining," in *IEEE International Conference on Image Processing (ICIP)*, 2020. 3

[43] K.-C. Huang, H.-H. Yang, and W.-T. Chen, "Multi-scale aggregation with self-attention network for modeling electrical motor dynamics," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 3

[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020. 3

[45] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," *arXiv:2010.04159*, 2020. 3

[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv:2103.14030*, 2021. 3

[47] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021. 3

[48] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021. 3

[49] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv:2101.11986*, 2021. 3

[50] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *CVPR*, 2020. 3

[51] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *ICCV Workshops*, 2021. 3

[52] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *CVPR*, 2021. 3

[53] Z. Wang, X. Cun, J. Bao, and J. Liu, "Uformer: A general u-shaped transformer for image restoration," *arXiv:2106.03106*, 2021. 3

[54] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2353–2363. 3

[55] Y. Zhou, Z. Chen, R. Li, B. Sheng, L. Zhu, and P. Li, "Eha-transformer: Efficient and haze-adaptive transformer for single image dehazing," in *The 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, 2022, pp. 1–8. 3

[56] T. Zhang, N. Jiang, J. Lin, J. Lin, and T. Zhao, "Desnowformer: an effective transformer-based image desnowing network," in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2022, pp. 1–5. 3

[57] S. Chen, T. Ye, Y. Liu, E. Chen, J. Shi, and J. Zhou, "Snowformer: Scale-aware transformer via context interaction for single image desnowing," *arXiv preprint arXiv:2208.09703*, 2022. 3

[58] W.-T. Chen, H.-L. Lou, H.-Y. Fang, I.-H. Chen, Y.-W. Chen, J.-J. Ding, and S.-Y. Kuo, "Desmokenet: A two-stage smoke removal pipeline based on self-attentive feature consensus and multi-level contrastive regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3346–3359, 2021. 4

[59] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual fourier transformation for single image deblurring," *arXiv preprint arXiv:2111.11745*, 2021. 4

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020," *arXiv preprint arXiv:2010.11929*, 2010. 5

[61] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[62] F.-A. Vasluianu, T. Seizinger, and R. Timofte, "Wsrd: A novel benchmark for high resolution image shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014. 5

[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. 5

[65] F.-A. Vasluianu, T. Seizinger, R. Timofte *et al.*, "Ntire 2023 image shadow removal challenge report," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7