NTIRE 2023 Image Shadow Removal Challenge Report

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
Shuchen Xia

Abstract

This work reviews the results of the NTIRE 2023 Challenge on Image Shadow Removal. The described set of solutions were proposed for a novel dataset, which captures a wide range of object-light interactions. It consists of 1200 roughly pixel aligned pairs of real shadow free and shadow affected images, captured in a controlled environment. The data was captured in a white-box setup, using professional equipment for lights and data acquisition sensors. The challenge had a number of 144 participants registered, out of which 19 teams were compared in the final ranking. The proposed solutions extend the work on shadow removal, improving over the performance level describing state-of-the-art methods.

Florin-Alexandru Vasluianu, Tim Seizinger and Radu Timofte are the NTIRE 2023 challenge organizers. The other authors participated in the challenge.
Appendix.A contains the authors' team names and affiliations.
https://cvlai.net/ntire/2023

1. Introduction

The Shadow Removal task is a complex problem in computer vision. The main challenge of this task is to propose an algothm able to cope with the wide range of factors involved in the shadow creation model. Shadows are direct effects of light occlusion [56]. Depending on the shape of the occluder, its material, the position and intensity of the occluded directional/diffuse light, shadows have different shapes and intensities. The other source of variation comes with the properties of the surface the shadow is being casted, with surfaces characterized by complex material properties and textures, and a large variety in terms of colors. All the aforementioned factors are imposing additional conditions to an already difficult problem.

Light occlusion induces a steep variation in an image region in terms in the luminance component of the acquired images, producing a change in terms of represented colors without semantic support. Thus, shadow impacts the performance of other vision tasks such as image segmentation [1, 22], semantic segmentation [23, 54], object recognition and tracking [5, 21, 27, 30, 33, 57].

The shadow removal task is, essentially, an image restoration task aiming at recovering the information lost by light occlusion, using the information available in the shadow affected image. Lately, large image databases, consisting of shadow/shadow-free image pairs (SRD [50], ISTD [64]) or unpaired data (USR [28]) enabled the formulation of the shadow removal problem in the deep learning framework. Since both paired and unpaired data is publicly available and well described in literature, the shadow removal solutions adopted both the fully supervised approach [35, 50, 53, 64, 78], and the unsupervised (weakly-supervised) learning framework [28, 59]. Lately, the introduction of the attention mechanism [62] in the image restoration solutions [8, 70] further improved the shadow removal performance, with more of the contemporary works [26, 78] adopting different variations of feature fusing strategies.

One particular challenge of the shadow removal task is learning a physically explainable transformation, achieving a significant performance level in terms of perceptual metrics. The semantic or illumination inconsistencies observed in the already introduced datasets increase the complexity of this task, impacting the performance of the solutions trained in the fully-supervised setup. The introduction of the Generative Adversarial Networks (GANs) framework [24] with their ability to learn complex distributions, enabled a new class of solutions, with recent efforts [28, 59, 75] solving the shadow removal task in the broader image-to-image translation framework [29].

In [46], the authors observed that the usage of the classical encoder-decoder UNet structure [51] has a tendency to produce artifacts-affected results. Moreover, there is a tendency of the de-shadowed image regions to appear blurry [28, 76]. One of the solutions proposed to the aforementioned limitations was to design robust loss functions, with recent solutions [26, 28, 35, 36, 53, 59, 64, 78] able to produce photo-realistic deshadowed results with high pixelwise fidelity.

Similar to previous editions similar challenges [2, 3], the NTIRE 2023 Image Shadow Removal Challenge represents a step forward in establishing a more complex single image shadow removal benchmark. It is based on a novel dataset capturing a wider range of light object interactions, with a large variety of light occluders and shadow affected surfaces. The *WSRD* [1] dataset [60] consists of 1200 shadow affected high resolution images and their corresponding ground truth images of the same scene. We perform an objective evaluation by comparing the restored output of the methods with the ground truth images of the dataset, with a high focus on the perceptual performance of the proposed solutions.

This challenge is one of the NTIRE 2023 Workshop [2] series of challenges on: night photography rendering [55], HR depth from images of specular and transparent surfaces [74], image denoising [39], video colorization [31], shadow removal [61], quality assessment of video enhancement [40], stereo super-resolution [66], light field image super-resolution [68], image super-resolution ($\times 4$) [79], 360° omnidirectional image and video super-resolution [6], lens-to-lens bokeh effect transformation [15], real-time 4K super-resolution [18], HR nonhomogenous dehazing [4], efficient super-resolution [38].

2. Image Shadow Removal Challenge

The objectives of the NTIRE 2023 challenge on Image Shadow Removal are:

- to gauge and push the state-of-the-art in image shadow removal;
- to compare and promote the state-of-the-art solutions;
- to promote the *WSRD* [60] dataset as a novel challenging benchmark for high resolution image shadow removal.

2.1. Image Shadow Removal Dataset

The main limitation of other datasets proposed for the shadow removal task like ISTD [64], USR [28] and SRD [63] is the limited number of interactions they capture. The data is mainly a collection of image pairs capturing different scenes in outdoors conditions, consisting of a surface on witch a shadow is casted by an occluder object, which itself does not appear in the image. The main reason behind adopting this setup is its simplicity, since a cluttered scene or a scene composed of high-complexity textures would need the addition of more lights to counter self-shadows, and additional post-processing would be needed for color alignment. Moreover, since most of the data is acquired in outdoors conditions, additional lighting inconsistencies were observed in the ISTD dataset [64], with an additional correction method being proposed in [35], further decreasing the fidelity error in the regions that are not affected by shadow. Additional semantic differences found between the input and the corresponding ground truth images [59] come as another reason to further investigate the shadow formation models in a more difficult setup.

Opposed to these previous methods, our capturing setup is built around controllable artificial light sources. As can be observed in Figure 1, we positioned the camera, a Canon EOS R6 II, in a 45 degree vertical angle towards the scene. The first flash serving as directional light source was mounted in a 45 degree vertical angle towards the scene and

[1] https://github.com/fvasluianu97/WSRD-DNSR

[2] https://cvlai.net/ntire/2023/

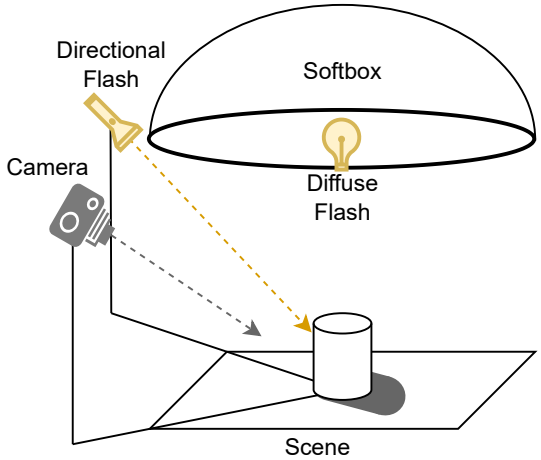

Figure 1. A visual representation of the setup used for data acquisition.

a 90 degree horizontal angle towards the camera. A second flash mounted in a softbox directly above the scene is providing diffuse illumination without casting shadows. Both flashes were always used on their maximum brightness setting.

For capturing the input and shadow free ground truth frames the camera was used in manual mode, meaning that all the camera settings like exposure or white balance were fixed and did not change in between frames. Its light sensitivity was set to ISO 100 to minimize noise, and the aperture of the 70mm lens was set to F11 to maximize the depth of field.

To capture a input frame with shadows, both the directional and the diffuse flash was used. The directional flash causes objects in the scene to cast a shadow, while the diffuse flash ensures that the shadow areas still contain enough detail for later recovery. To capture the shadow-free ground truth frames, only the diffuse flash was activated. As only one flash is active in this case, we matched the exposure of the input and the ground truth by processing the RAW data in Adobe Lightroom.

A problem that still remains in our setup is that objects with complex surfaces still possess soft self-shadows and, in some cases, cast diffuse shadows on the surfaces below them. An improved setup could remedy this by using a translucent scene surface that could be lit from below, or using additional soft boxes on the sides. Given the fixed camera pose and the lights positioning, additional sources of variation can be found to extend the current image database.

2.2. Evaluation

Being a predominantly perceptual task, we evaluate the NTIRE2023 Image shadow Removal submissions given the following criteria:

1. The Mean Opinion Score (MOS) for the submitted predictions;
2. The LPIPS [77] distance between the predictions and the ground-truth images. We used the ImageNet pretrained AlexNet [34] for the LPIPS feature extraction;
3. The Structured Similarity Index (SSIM) [69] score;
4. The reconstruction fidelity in terms of PSNR values;
5. An efficiency metric stating the fidelity gain per million of learned parameters characterizing the proposed solution, with respect to the PSNR value of 15.03 dB chrcterizing the test split pairs.

The user study consisted of amateur and professional photographers, whose feedback was quantified as an integer value starting from 0 (worst) to 10 (best). The evaluated dataset consisted of the predictions submitted by participants for the samples $s \in \{1, 19, 36, 47, 55, 64, 82, 89\}$ of the test split, where the index of the first testing sample is 0. The Mean Opinion Score (MOS) reported represents the average of the study participants feedback for the aforementioned set of samples.

2.3. Challenge Phases

1. **Development phase:** In this phase, the training split, consisting of 1000 image pairs, was available on the challenge webpage. The participants tailored their solutions based on the properties of the data, deploying additional preprocessing operations.
2. **Validation phase:** The validation split, consisting of 100 shadow affected images was published on the challenge website. The participants used the Codalab validation server [48] to submit their predictions, without having access to the ground truth images.
3. **Testing phase:** The participants had access to the 100 images testing split of the dataset. Using the test server [48], they provided the predictions of their model for evaluation. The submission with the best results in terms of PSNR and SSIM was considered for evaluation in the final ranking. Finally, an user study involving professional and amateur photographers was conducted, and the Mean Opinion Score (MOS) was used to show the best performers in terms of perceptual properties.

3. Challenge Results

The challenge registered 144 participants, with number of 19 teams providing their code and results, thus being ranked in the final phase. The submission prepared by each team consists of codes, testing split results and a factsheet

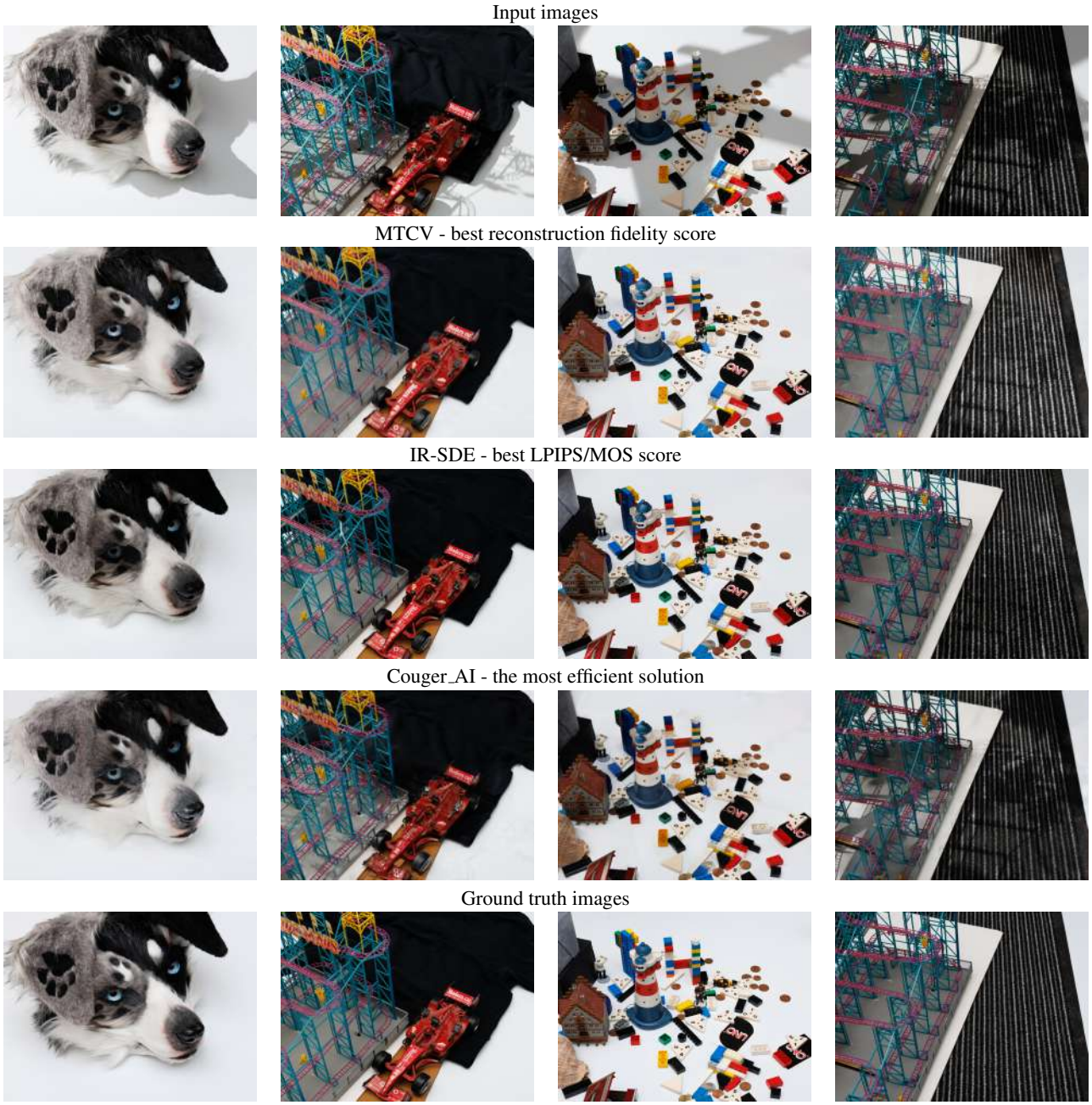

Figure 2. Visual results provided for best performing solution on each of the metrics deployed. Best zoom-in on screen in the electronic version.

containing a description of the proposed solution and the set of results for visual comparison. Section 4 provides details about each of the solutions ranked in the final phase.

Table 1 provides the quantitative evaluation of the submitted results, along with the ranking with respect to each of the metrics used for evaluation. Results corresponding to the solutions ranked first for one of the metrics are provided in the Figure 2. Given a recent trend in the community to focus on more efficient solutions [16, 17, 26, 58], we explicitly report the reconstruction fidelity performance gain per milion learnt parameters of the proposed solutions. Moreover, we list the reported inference time per full-resolution sample (1920×1440 px), on the device each of the teams used to develop their model.

Checking Table 1, the metric with the highest correlation to the Mean Opinion Score (MOS) is the LPIPS,

while PSNR and SSIM can be used to differentiate solutions achieving similar performance levels. However, the top solution achieved consistent results along the set of compared metrics.

4. Challenge Methods

4.1. MTCV

The MTCV team has proposed a new model called Pyramid Ensemble Structure (PES) [19] that is based on the NAFNet architecture introduced in [8]. PES comprises three main components: Pyramid Inputs, Pyramid Stages, and Shadow Ensemble. The team's innovative approach, as illustrated in Figure 3, has the potential to enhance existing models and improve overall performance.

To enable their model to capture diverse shadow sizes and shapes, the MTCV team leverages multiple scales through a technique called Pyramid Inputs. This involves constructing pyramid inputs to enable the network to capture information at different scales. By using multiscale image training, the team is able to better preserve global information from the reconstructed frame. To ensure efficient training, the pyramid input images are further cropped into the same shape.

To further enhance the ability of their model to process global information, the MTCV team utilizes a training approach known as Pyramid Stages. This involves training the network using PSNR and SSIM as the main performance metrics. The team then applies a fine-tuning step using L1 loss at a larger training resolution. Finally, the training procedure is continued using SSIM loss and Mean Squared Error for a resolution comparable to the input resolution of the training data. By using this multi-stage training process, the team is able to achieve superior performance in their model.

Amid the realm of shadow removal task, they have made a crucial observation: bright regions that rarely contain shadows are more likely to be non-shadow regions. This discernment enables them to assume that such regions are free from shadows. Leveraging this novel insight, they propose a simple yet highly effective approach to shadow removal - selecting the maximum prediction from all the shadow models. Their innovative technique builds on this observation, demonstrating superior results in terms of accuracy and performance.

4.2. IR-SDE

They propose Refusion [44] that uses the IR-SDE [43] as the base diffusion framework (see Figure 4), which can naturally transform the high-quality image to its degraded counterpart, without caring how complicated the degradation is (even for real-world degradation). As shown in Figure 4, IR-SDE is a mean-reverting SDE in which the forward process is defined as:

$$dx = \theta_t (\mu - x)dt + \sigma_t dw, \tag{1}$$

where θ_t and σ_t are time-dependent positive parameters that characterize the speed of the mean-reversion and the stochastic volatility, respectively. Since it is an Ito SDE, we could derive a reverse-time SDE:

$$dx = \left[\theta_t (\mu - x) - \sigma_t^2 \nabla_x \log p_t(x)\right]dt + \sigma_t d\hat{w}. \tag{2}$$

At test time, the only unknown part is the score $\nabla_x \log p_t(x)$ of the marginal distribution at time t. As other diffusion-based models, we employ a CNN network to estimate the score to backward from the low-quality image to the high-quality image.

Unlike other L_1 loss normally trained networks which usually produce smooth/blurry results, the proposed Reffusion aims to achieve a highly competitive perceptual performance as well as the distortion scores (PSNR). To counter training pairs position shifts and their non aligned luminance, they train the Reffusion model to implicitly learn to align with the ground truth, which is more flexible and efficient.

In addition, they further improve the results by updating the score-network from U-Net to NAFNet [8], which is more efficient and also has a good performance compared with recent Transformers. To adaptively insert the scalar time into the network, they construct a simple multi-layer perceptron to learn two pairs of scale-shift parameters and apply them to the features with affine transforms. Such a network leads to better learning of score function conditioned on current state x_t, original low-quality image x_T, and timestamp t.

4.3. SRDM

As shown in Figure 5, the proposed model, SRDM, is a three-stage model to recover shadow-free images from shadow affected ones. The first stage will downsample the input shadow image and its shadow mask by factor $f = 3$ using bicubic interpolation. The shadow mask is calculated by a pretrained BDRAR [80] shadow detector. Then the second stage adopts SR3 [52], a diffusion model for super-resolution, to generate a shadow-free image conditioned on the shadow-affected image and its shadow mask. At last, they apply a pretrained ESRGAN [67] model as the third stage to restore the original resolution of the shadow-free image.

4.4. SYU-HnVLab

Team SYU-HnVLab started with a method called CAIR [72], which excels at image restoration, as a baseline. This method is good at removing image filters by color attention. Given the properties of the shadow affected areas,a

Rank	Team	Username	PSNR↑	SSIM↑	LPIPS↓	MOS↑	Params.(M)	Gain/Mil. params↑	Runtime(s)	Device	Extra data
1	MTCV	cuishuhao	$22.36_{(1)}$	$0.7_{(1)}$	$0.182_{(2)}$	$8.31_{(3)}$	46	$0.159_{(12)}$	0.002	A100	No
2	IR-SDE	ir-sde	$19.6_{(14)}$	$0.58_{(15)}$	$0.149_{(1)}$	$8.94_{(1)}$	74	$0.062_{(14)}$	34	A100	No
3	SRDM	xyz123	$22.2_{(2)}$	$0.69_{(3)}$	$0.269_{(9)}$	$7.44_{(5)}$	151.7	$0.047_{(16)}$	136	A100	No
4	SYU-HnVLab	Una	$21.25_{(8)}$	$0.67_{(8)}$	$0.217_{(6)}$	$6.84_{(6)}$	13.9	$0.447_{(7)}$	0.004	A100	No
5	MegSRD	CD_luo	$17.36_{(18)}$	$0.53_{(17)}$	$0.198_{(4)}$	$8.81_{(2)}$	104.6	$0.022_{(17)}$	18	RTX2080Ti	No
6	UM-JTG	daylight	$21.7_{(4)}$	$0.69_{(4)}$	$0.283_{(10)}$	$6.81_{(7)}$	27	$0.247_{(10)}$	60	RTX3090	No
7	LVGroup_HFUT	HuanZheng	$21.43_{(7)}$	$0.68_{(7)}$	$0.231_{(7)}$	$6.79_{(8)}$	38.88	$0.165_{(11)}$	0.25	2×RTX3090Ti	No
8	IIM_TTI	Yuki-11	$18.08_{(16)}$	$0.53_{(16)}$	$0.196_{(3)}$	$7.44_{(4)}$	55	$0.055_{(15)}$	1.01	A100	ImageNet
9	NTU607-shadow	mrchang87	$21.79_{(3)}$	$0.7_{(2)}$	$0.236_{(8)}$	$6.15_{(11)}$	11.85	$0.57_{(6)}$	1.81	RTX3090	ISTD
10	MM911	codalab123	$21.69_{(5)}$	$0.69_{(5)}$	$0.293_{(11)}$	$5.51_{(12)}$	25.8	$0.258_{(9)}$	0.005	A100	No
11	leaves	leaves	$21.68_{(6)}$	$0.69_{(6)}$	$0.309_{(13)}$	$6.51_{(9)}$	16	$0.416_{(8)}$	0.005	2×RTX2080Ti	No
12	MegSRF	jiangchengzhi	$17.74_{(17)}$	$0.5_{(18)}$	$0.203_{(5)}$	$6.45_{(10)}$	34.2	$0.079_{(13)}$	1.3	RTX2080Ti	No
13	Couger_AI	SabariNathan	$20.56_{(12)}$	$0.63_{(11)}$	$0.306_{(12)}$	$4.75_{(15)}$	0.861	$0.006_{(1)}$	0.001	K80/T4	No
14	Noir	Krocy	$21.24_{(9)}$	$0.66_{(9)}$	$0.389_{(14)}$	$4.73_{(16)}$	6.73	$0.923_{(5)}$	0.22	RTX3090	No
15	CVPR_IITRPR	shrutiphutke	$19.71_{(13)}$	$0.63_{(12)}$	$0.414_{(16)}$	$5.0_{(14)}$	0.97	$4.825_{(2)}$	0.25	RTX2080	No
16	NUSSZ-ShadowRemove	tiger	$21.13_{(10)}$	$0.65_{(10)}$	$0.411_{(15)}$	$4.13_{(18)}$	5	$1.22_{(4)}$	1.6	RTX3090	No
17	Concentration	Concentration	$19.23_{(15)}$	$0.6_{(14)}$	$0.418_{(17)}$	$5.48_{(13)}$	1.17	$3.59_{(3)}$	0.003	GTX1080Ti	No
18	ZJUME251	BowenZhao	$20.73_{(11)}$	$0.62_{(13)}$	$0.48_{(18)}$	$4.31_{(17)}$	190	$0.03_{(18)}$	0.037	RTX2080Ti	No
19	Imsensor	Goring	$11.83_{(19)}$	$0.37_{(19)}$	$0.984_{(19)}$	$0_{(19)}$	9.18	$-0.349_{(19)}$	0.13	TitanX	No

Table 1. Quantitative results of the challenge final submission on the *WSRD* test split. Using naming convention $n_{(m)}$, where n is the value of the metric evaluated and (m) is the rank in the list of submissions sorted by the evaluated metric value.

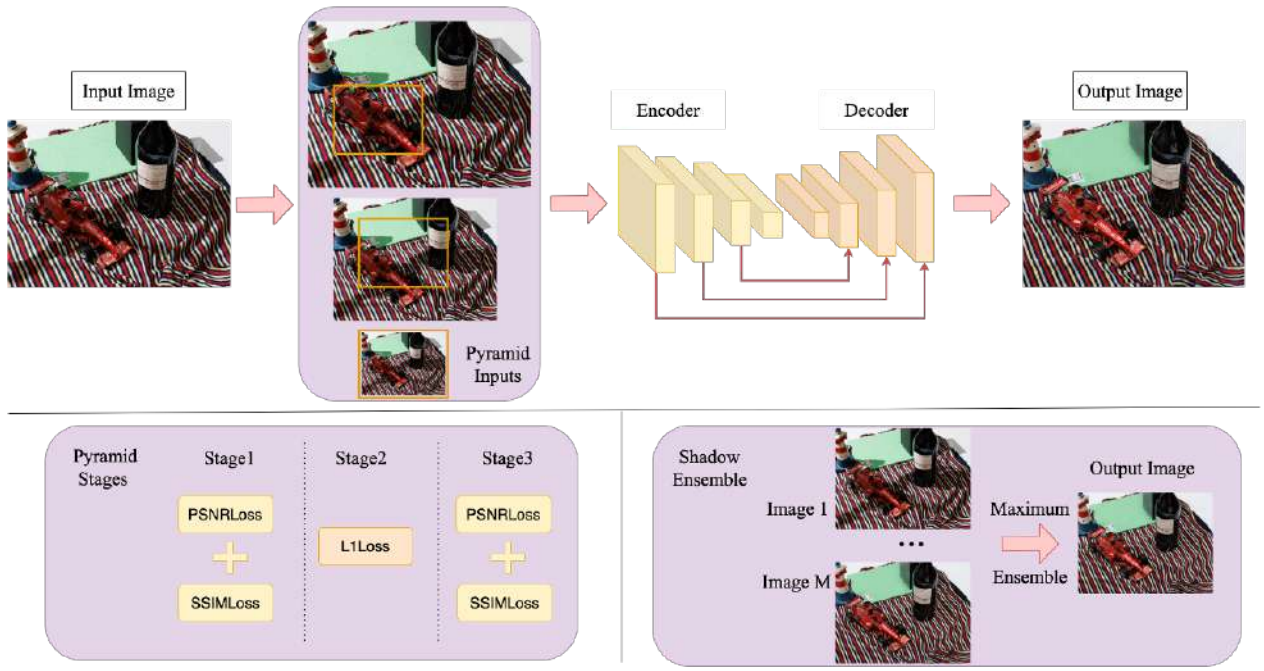

Figure 3. Overall framework of Pyramid Ensemble Structure (PES) proposed by the MTCV Team. First, they implement Pyramid Inputs, which entails resizing and cropping the input images into various sizes and shapes. Once adjusted, the input images are then forwarded to the network for processing, which is trained based on diversity loss functions in Pyramid Stages. Finally, the output images are ensembled by selecting the maximum result from the various options available.

model that can learn information about color would be suitable for removing shadows.

In general, shadows are areas of an image where less light is present, resulting in a lower level of luminance. And since the depth and location of shadows depend on the intensity and position of light, the CIELab color space, which can represent light information, was preffered over the RGB color space. In particular, luminance attention is proposed to focus on the light information. Team SYU-HnVLab starts with a color attention mechanism as baseline, which takes RGB input and outputs RGB color attentive features. Then, the different features (luminances feature and color features) are fused using the color space fusion inspired by [20]. They ultimately proposed two network architecture: luminance attention network for shadow removal (LASR) and color-luminance attention network (CLAN) (see Figure 6). LASR only uses luminance attention, and CLAN takes advantage of both chroma-component and lu-

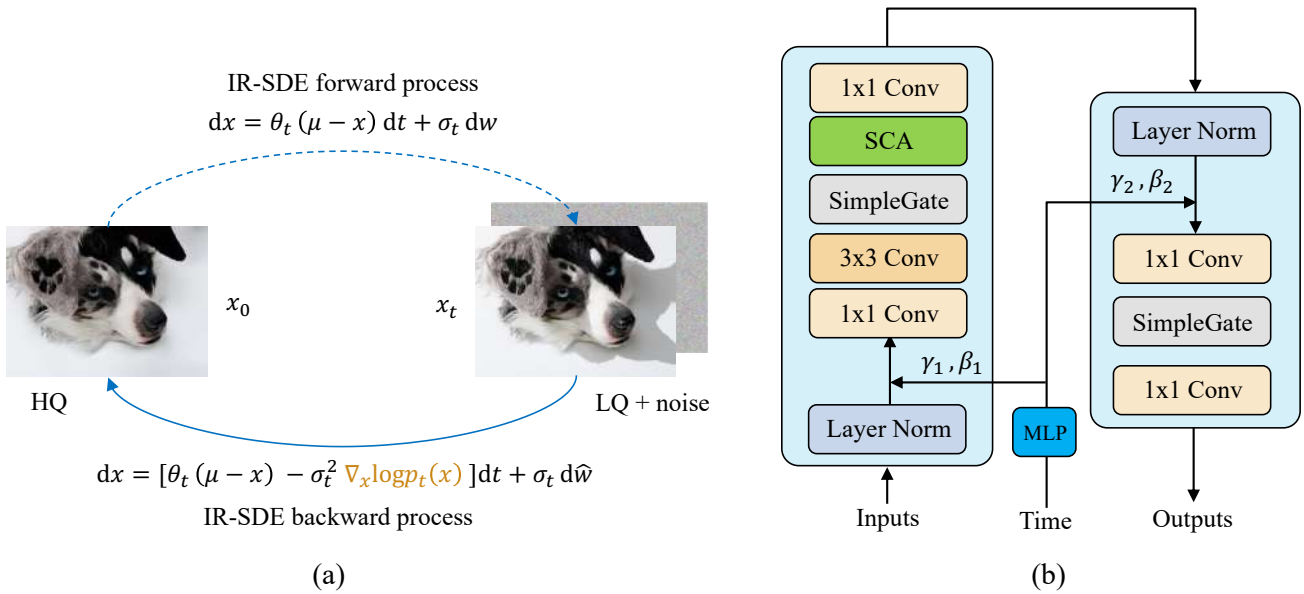

Figure 4. (a) Image restoration based on the IR-SDE [43], which uses a mean-reverting stochastic differential equation (SDE) to recover images. (b) The modified NAFBlock. Here "SCA" is the simple channel attention, and "SimpleGate" is an element-wise operation that splits feature channels into two parts and then multiplies them as output.

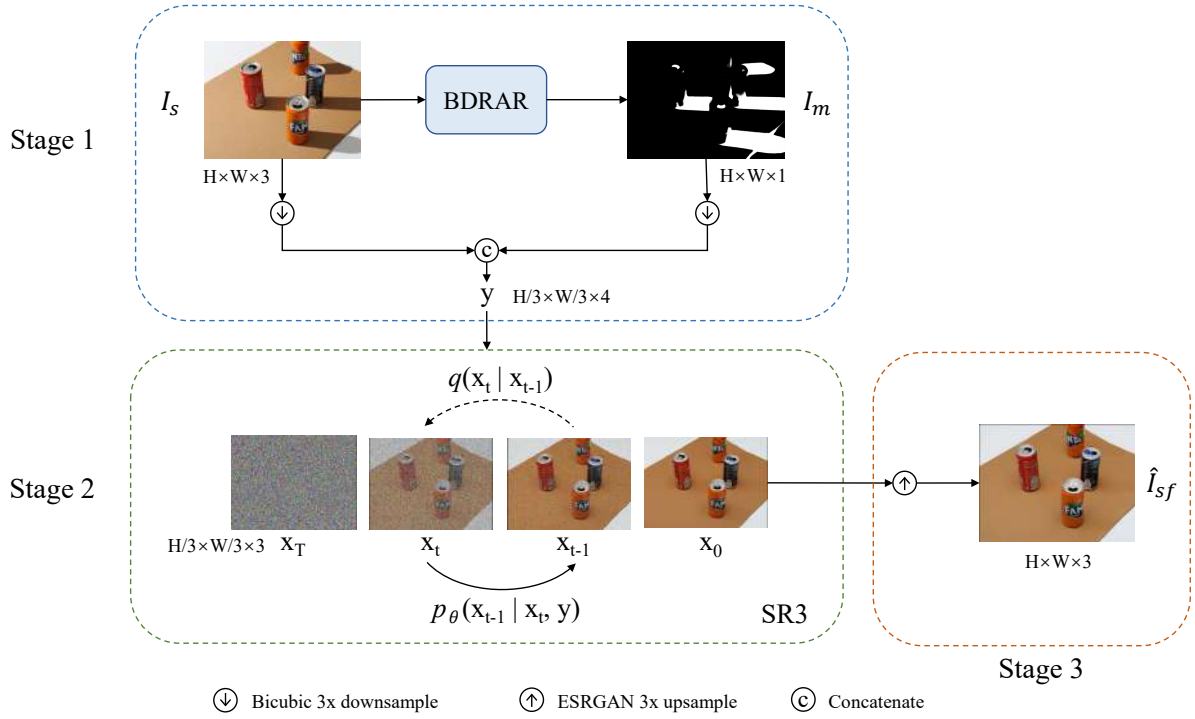

Figure 5. The architecture of the solution proposed by the SRDM team. The first stage consists of shadow detection, followed by shadow removal at a lower scale. In the last step, the shadow free prediction is upscaled back to original resolution.

minance attention, effectively fusing those two information sources. Both models were trained with an additional loss term, to account for the data misalignment.

4.5. MegsRD

Their approach (see Figure 7) extends on previous work in image restoration [47], where image restoration is achieved by smoothing noise estimation of overlapping patches during inference.

In the training process, they first apply a pre-trained alignment network to warp the input such that the input shadow affected image and the gound-truth are aligned. Then, the original resolution image is sampled from the aligned data distribution and sliced into patches of size 128×128, as a condition to learn the conditional diffusion model. In the testing phase, they start by padding the original resolution image to fit the sampling procedure, and then decompose the image into multiple overlapping patches of size 512×512.

The set of overlapping patches is then used as input to the trained noise estimation network, that has to merge the output noise estimation to act on the sampling process of diffusion, such that the overlapping blending can effectively smooth merging artifacts between patches.

4.6. UM-JTG

Team UM-JTG (see Figure 8) started by extracting both the shallow features of the shadow images and the corresponding edge maps using strided convolutions ($s = 4$). After feature concatenation and a simple convolutional layer, the feature set is processed by six Swin Transformer blocks, in a strategy similar to the previous work in [41]. Based on the observation that large window size can activate more input information and the Hybrid Attention Block (HAB) can increase the receptive field of the transformer model [12], they added the Residual Hybrid Attention Group (RHAG) to their model. This consists of additional CNN blocks working in parallel with the window-based Transformer blocks, aiming to further enhance the representation ability of the original Swin blocks. Finally, after the skip connection is added of the network's output, *PixelShuffle* is adopted to reconstruct the high-resolution shadow-free images.

The training objective is a combination of L1 loss, color loss and spatial loss, aiming to suppress artifacts of the edges shadow areas and the overall exposure deviation in the process of optimization.

4.7. LVGroup_HFUT

A visual representation of the LVGroup_HFUT proposed solution is detailed in Figure 9. The backbone architecture is based on NAFNet [8], but with a different configuration of the resolution-altering and dense-encoding operations. Specifically, they use five up/down convolution layers followed by a sequence of NAFBlocks as building blocks of their model.

Due to the pixel offset between shadow-affected images and the corresponding ground-truth, the heavy blur will occurred during training. To tackle misalignment, they deployed pixel offset correction as the first stage of their model, to alleviate this particular effect.

The training procedure is further optimized for convergence, with an early-stopping mechanism avoiding the overfitting behaviour of the proposed model.

4.8. IIM_TTI

Team IIM_TTI proposed a processing pipeline (see Figure 10), in which every step is tailored around one challenge of the proposed task . They start with a homography estimation step, to account for the roughly aligned data used for training. Then, they perform a semi-automatic shadow mask estimation, using their method (MASMA), which detects the shadow affected regions in the HSV image space. In the next step, they use MTMT [13] for shadow detection, using the pretrained ImageNet ResNeXt as backbone.

In the last step, MTMT and Shadowformer [26] are trained together for shadow detection and shadow removal. The methosd achieves significant performance in the perceptual domain, with notable results for the LPIPS and MOS metrics.

A combination of the SSIM loss and a Structure Preservation Loss is used as the training objective of the proposed model.

4.9. NTU607-shadow

Their solution and details are published in [7]. Shortly, team NTU607-shadow proposes a two-stage recovery strategy [9–11] which contains a pseudo-free generator and an image shadow remover. In the first stage, they train a GAN-based free image generator, which can roughly remove shadows. This stage is based on SpA-Former [78]. They use the first model to remove the shadow and obtain the shadow mask by calculating the difference between the original images and shadow-free images simultaneously. The second stage considers shadow masks and original images to obtain refined shadow-free images. To this end, they deploy ShadowFormer [26] for this step of their proposed solution. The ShadowFormer leverages original images and shadow masks as inputs and generates shadow-free images.

To train their version of SpA-Former [78], Adam optimizer [32] is adopted and the batch size is set to $b = 3$. The network is trained for 200 epochs with the momentum β_1 = 0.5, β_2 = 0.999. The learning rate is initialed as 4×10^{-4}. They use the L1, L2 and the softplus function as loss terms blended in the optimized objective.

In the second step, for ShadowFormer [26], the image

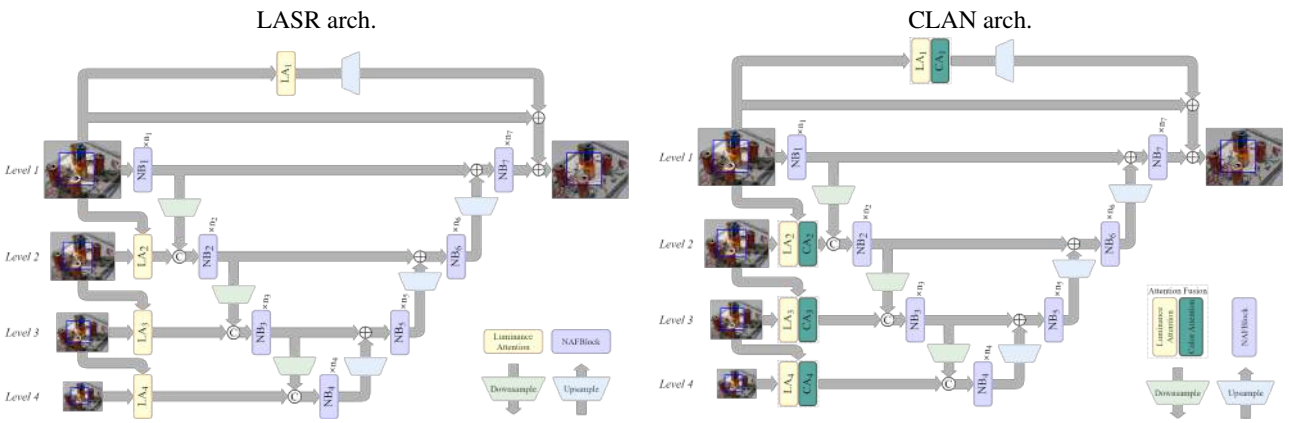

Figure 6. The schematic representation of the architecture of the SYU-HnVLab proposed solution.

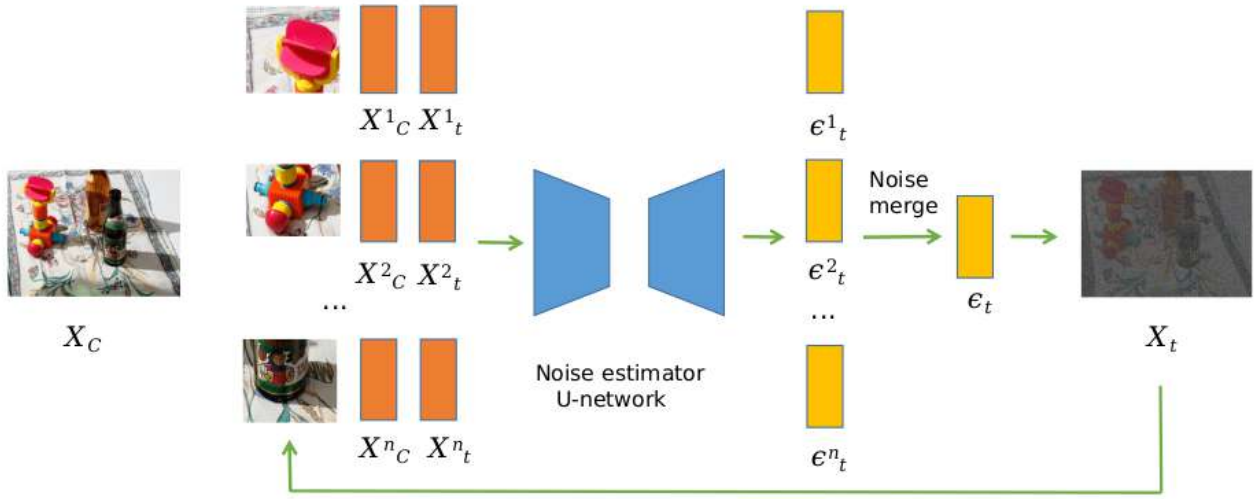

Figure 7. A visual representation of the solution proposed by MegSRD team.

is randomly cropped as 320 $\times$ 320, without data augmentation. The AdamW optimizer [42] is utilized with a batch size of $b = 8$ to train the network. The network is trained for 500 epochs with the momentum β_1 = 0.5, β_2 = 0.999 and a weight decay of 0.02. The learning rate is initialed as 2×10^{-4}, and the training objective is based on the Charbonnier loss function.

4.10. MM911

Team MM911 builds the proposed solution around the current state-of-the-art on the corrected [35] version of the ISTD [64] dataset. The solution is called SHARDS [53], and proposes shadow removal at lower scales, followed by a refinement step that integrates the upscaling of the shadow free prediction to the original resolution of the input.

Without provided shadow masks, they use the Mask-free-LSRNet for low resolutions shadow removal, performing the task at a lower resolution of the shadow affected inputs (400$\times$400). The other step is performed by the second component, the Mask-free-DRNet, upscaling the predicted images from 400$\times$400 to 1920$\times$1440 and learning and an additional fine-tuning procedure, thus increasing the quality of the predicted full-resolution shadow-free images.

4.11. leaves

Team leaves emphasizes the importance of the shadow mask in the shadow removal operation. Thus, they proposed a solution based on SpA-former [78], that they trained in or-

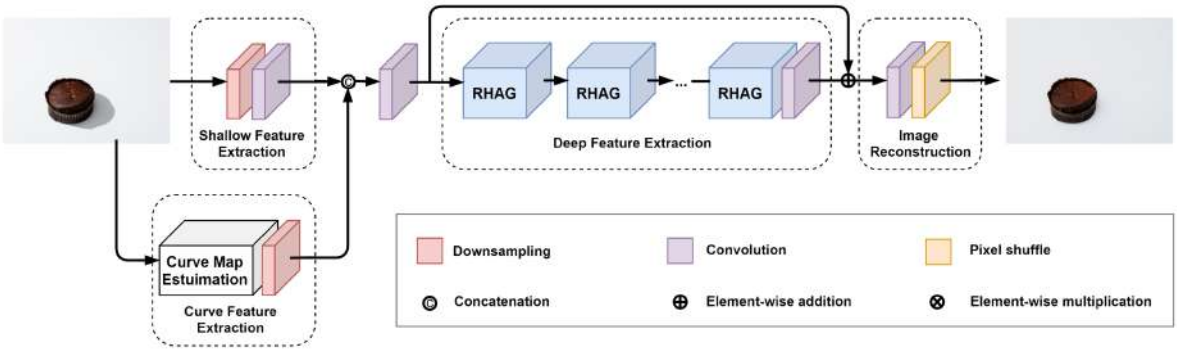

Figure 8. The overall architecture of the solution proposed by Team UM-JTG.

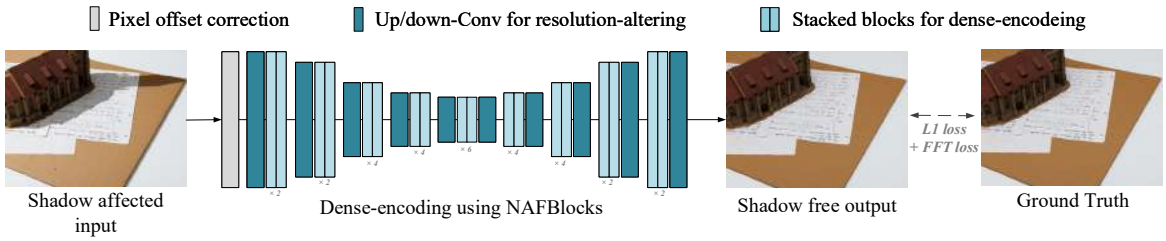

Figure 9. The structure of the LVGroup_HFUT proposed solution.

der to estimate the shadow masks based on a binary thresholding of the difference between the shadow affected image and its corresponding shadow-free prediction. Using the set of shadow masks estimated in the first step of their solution, they deployed ShadowFormer [26], to get their final predictions.

The prediction step uses an ensemble strategy, with multiple snapshots taken during training being aggregated in an averaging strategy, predicting the final shadow-free reconstructed frames.

4.12. MegSRF

The solution proposed by MegSRF is shown in Fig. 11. In the training process, the shadow-effected image I is downsampled ($2\times$) with area interpolation to I_s. Following downsampling, I_s is fed into the NAFNet [8] to generate the returning factor f_s with low resolution. Then, the factor is upsampled ($2\times$), to the original resolution of the image I.

Finally, the input image variants I and I_s are multiplied by f and f_s respectively, getting the shadow-free images $\hat{I}$ and $\hat{I}_s$. The model of the proposed solution is described in the Equation 3 and Equation 4.

$$\begin{aligned} \hat{I} &= I * f \\ f &= \mathcal{U}(f_s) \end{aligned} \tag{3}$$

$$\begin{aligned} \hat{I}_s &= I_s * f_s \\ f_s &= \psi(I_s) \\ I_s &= \mathcal{D}(I) \end{aligned} \tag{4}$$

Here $\mathcal{U}$ and $\mathcal{D}$ are upsampling and downsampling operations. ψ donotes the NAFNet model. The training objective is a combination of $\mathcal{L}_1$ loss on I and the perceptual loss $\mathcal{L}_p$.

Specifically, for a pair of the network output $\hat{I}$ and its ground truth I^{GT} , they compute the loss terms defined in Equation 5.

$$\begin{aligned} \mathcal{L}_1 &= ||\hat{I} - I^{\mathrm{GT}}||_1 \\ \mathcal{L}_p &= \|\phi_j(\hat{I}_s) - \phi_j(I_s^{\mathrm{GT}})\|_1 \end{aligned} \tag{5}$$

Here, ϕ_j is a pre-trained VGG-16 with normalization loss in the j-th layer. The final loss $\mathcal{L}$ is a combination of the two aforementioned loss terms.

In the evaluation process, the unsharp mask (USM) method is implemented on the factor f, as described in the Equation 6.

$$\hat{I} = I * USM(f) \tag{6}$$

4.13. Couger_AI

Team Couger_AI proposed an end-to-end model lightweight model (see Figure 12), aiming to remove the shadow and restore the shadow-free images using a low inference cost. The proposed approach is mainly inspired

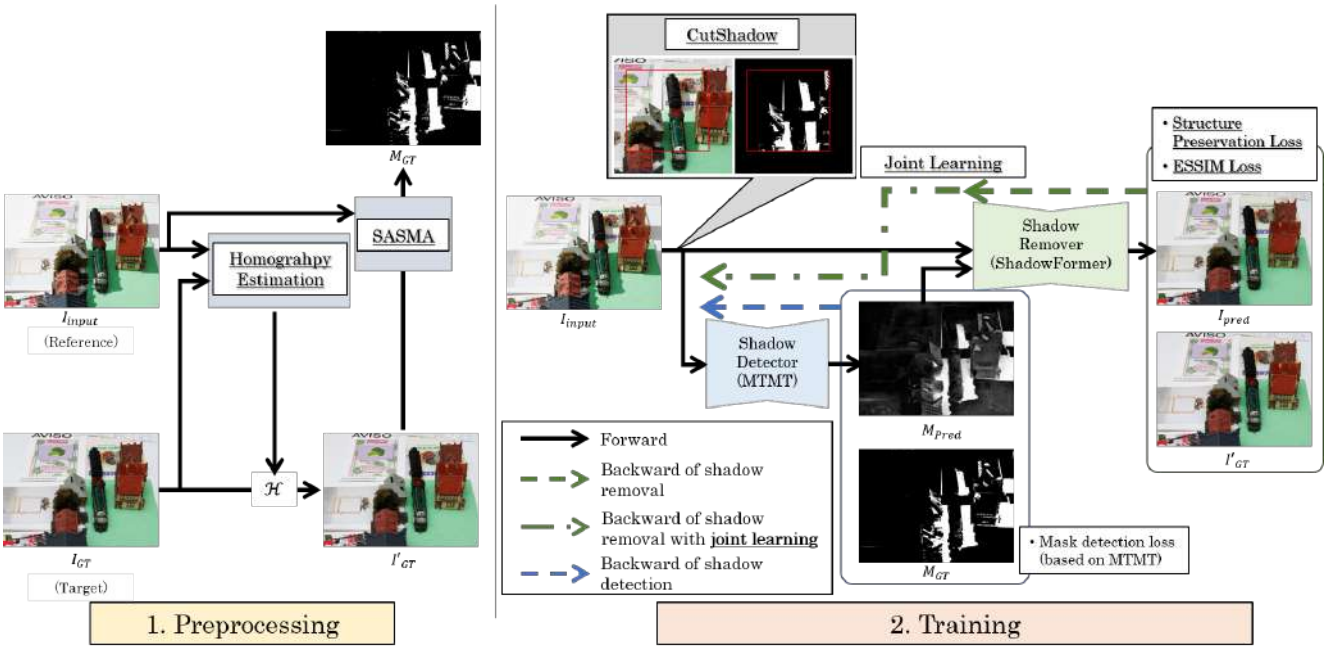

Figure 10. Overall diagram of the IM_TTI team proposed methodology. In preprocessing, they account for image misalignment by a Homography Estimation, followed by the ground-truth shadow mask estimation by SAMSA. After the shadow mask estimator, ShadowFormer [26] is used to predict shadow-free reconstructions and a shadow mask compared to the prediciton of the shadow detector. In the last step, CutShadow is used for augumentation, improving the quality of the reconstructed images.

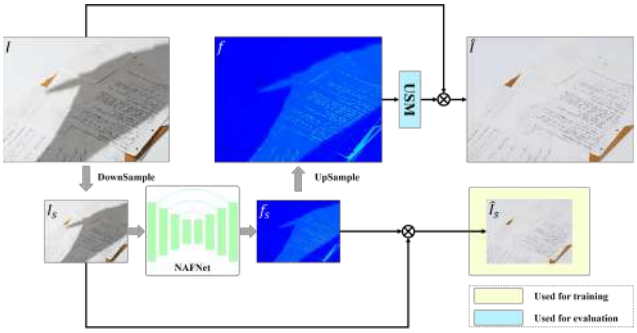

Figure 11. The pipeline of the method proposed by Team MegSRF. The shadow-effected image I is downsampled (I_s) and fed into the NAFNet [8] to generate the retuning factor f_s at low resolution. Then, the factor is upsampled (f) to the same size of I, and an unsharp mask (USM) method is applied to f. Then the factor f is multiplied to I to get a shadow-free predicted image.

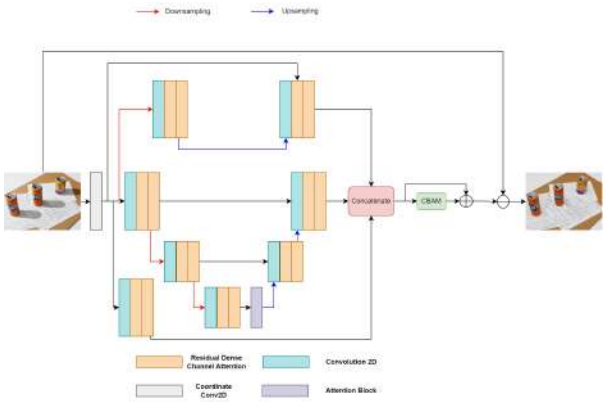

Figure 12. Architecture diagram of the shadow removal solution proposed by Team Couger_AI

by [45]. The network consists of three branches, fusing different complexity features. Similar to [45], a residual dense attention block is used as a backbone in an encoder-decoder structure as the first branch. The residual dense attention block uses a combination of dense connections, feature attention, and residual learning to extract and enhance image features. The output of this branch is concatenated with the two other branches to enhance the feature detailing shadow-free area of the original images. A gradient loss is used to guide the training procedure, improving the quality of the predicted images.

4.14. Noir

They propose an image shadow removal encoder-decoder model based the UNet structure [51]. The model mainly consists of two blocks, the Conv2dGN block, and the Attention block. The Conv2dGN block is composed of a convolutional layer, a group normalization layer, and a

SeLU activation function. The Attention block is similar to the multi-head attention mechanism. Neither of the aforementioned blocks affects the resolution of the feature maps, keeping it consistent between the blocks used to forward propagate the image. The structure of the model and additional information about the resolution manipulation are provided in Figure 13.

The encoder first downsamples the image by bilinear interpolation and then processes it twice using the Conv2dGN block, increasing the number of channels characterizing the feature maps. This step is repeated several times until the size of the feature map becomes (45×60×128). Then, the Attention module is used to process and fuse the global information of the image at the current scale feature map.

The decoder restores the image to its original size. Here, UNet skip connections are used to concatenate the corresponding image features, providing different scale information at the decoder level. After passing through the Conv2dGN block, the image is upscaled by bilinear interpolation. Finally, the image is restored to its original size further refined through two convolutional layers.

Using the Generative Adversarial Network framework, team Noir applied a discriminator to enhance the shadow removal efficiency of their proposed model. The discriminator is similar to the encoder, using downsampling and Conv2dGN blocks to change the outout size to 11×15×3. Then, the output is flatten and processed by a multilayer perceptron structure, predicting the probability of the sample to be real or fake. The training procedure is guided by the WGAN-GP [25] loss function, which uses a gradient penalty to improve the training procedure stability.

The loss function of the proposed generator id defined in Equation 7, and the training objective for the discriminator is defined in Equation 8.

$$L_{proposed} = MSELoss(r, t) - D(r) \tag{7}$$

$$L_d = D(r) - D(t) + \lambda_{gp} L_{gp}^{D}(interp.) \tag{8}$$

Here, r and t represent the output of proposed model and the ground truth image, and $interp.$ is an interpolated image between r and t. λ_p and λ_{gp} are hyper-parameters of the proposed model.

On one hand, the loss function of the proposed model aims to achieve a high fidelity reconstruction in terms of mean Squared error, while being characterized by improved semantic properties (guided by the discriminator component). The discriminator and the proposed generator are trained in an adversarial framework to increase the perceptual properties of the results produced by the proposed solution.

4.15. CVPR_IITRPR

Team CVPR_IITRPR proposed a computationally efficient, lightweight network (see Figure 14) for image shadow removal, characterized by a low number of parameters (0.97M). This solution is based on previous work [49] for image inpainting, which uses the localization information about the affected regions as a mask as one of the inputs pushed through the model. Unlike [49], the only input provided to the model is the shadow-affected iamge.

The proposed architecture consists of the multi-encoder level feature fusion module, auxiliary decoder, and actual reconstruction decoder. The encoder multi-level feature fusion module extracts relevant information from each of the encoder levels. This information is then processed with an auxiliary decoder, followed by a space depth correlation module to assist the actual reconstruction decoder for the shadow removal task. The weights of the network are optimized by calculating the loss ($Loss_{1-4}$) at each level characterizing the decoder.

4.16. NUSSZ-ShadowRemove

The proposed solution is based on SpA-Former [78], consisting of transformer layer and a series of joint Fourier transform residual blocks and two-wheel joint spatial attention.

First of all, from the Transformer network, the feature map is processed by a 3×3 convolution, then fed to the bottleneck structure and a Two-Wheel RNN Joint Spatial Attention (TWRNN).

The Two-Wheel RNN Joint Spatial Attention module is designed to make the network emphasize specific shadow patterns, as it can discover and find the focus map from the input element map. The attention graph is a two-dimensional matrix, in which the value of each element is a continuous value, indicating the activation characteristic to each pixel.

Deep convolution is introduced to emphasize the local context global attention map before computing feature covariance to generate from a layer normalized tensor. The transformer block first generates the predictions for the query (Q), key (K), and value (V), enriching the local context. This is achieved by applying 1×1 convolution to aggregate pixel-by-pixel cross-channel context, followed by 3×3 deep convolution to encode channel space context. Next, the query and key projections are reshaped such that their dot product operation generates an output map with the size of the transposed attention map.

The RNN model is used to project the descent in four main directions. Three standard residual blocks are first used to extract features, used to guide the three subsequent attention residual blocks. The task of these blocks is to eliminate shadows by learning negative residual. Finally,

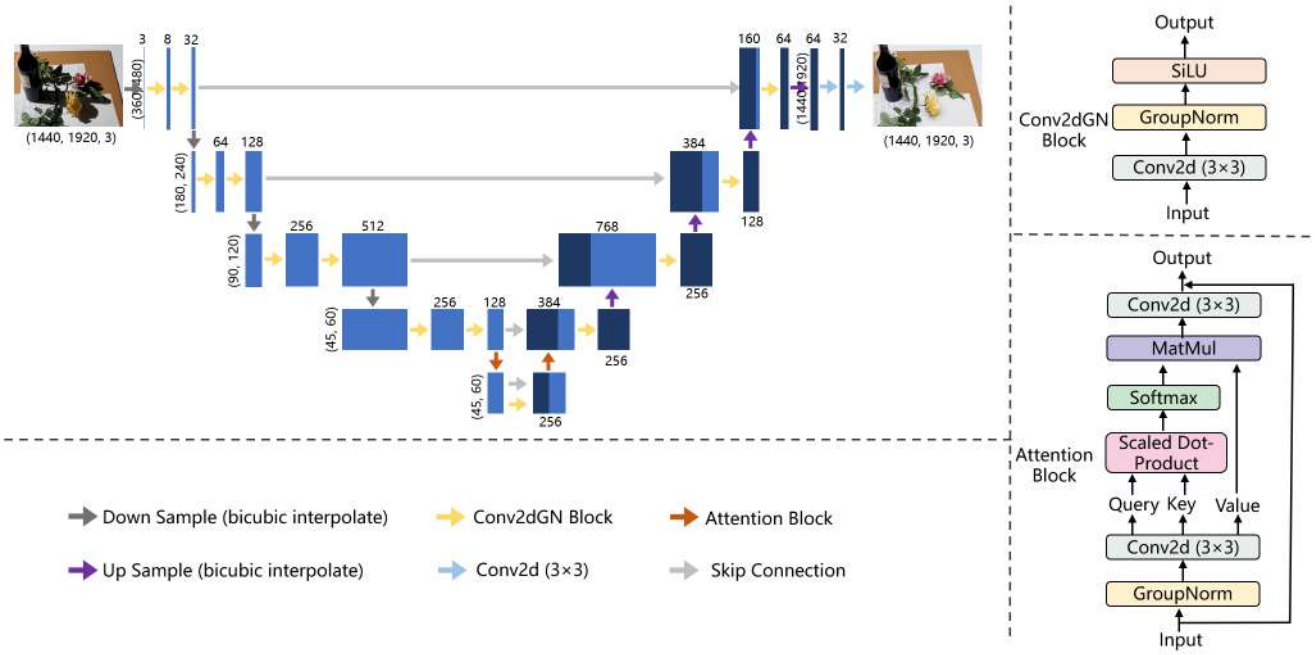

Figure 13. The structure of the team Noir proposed encoder-decoder model.

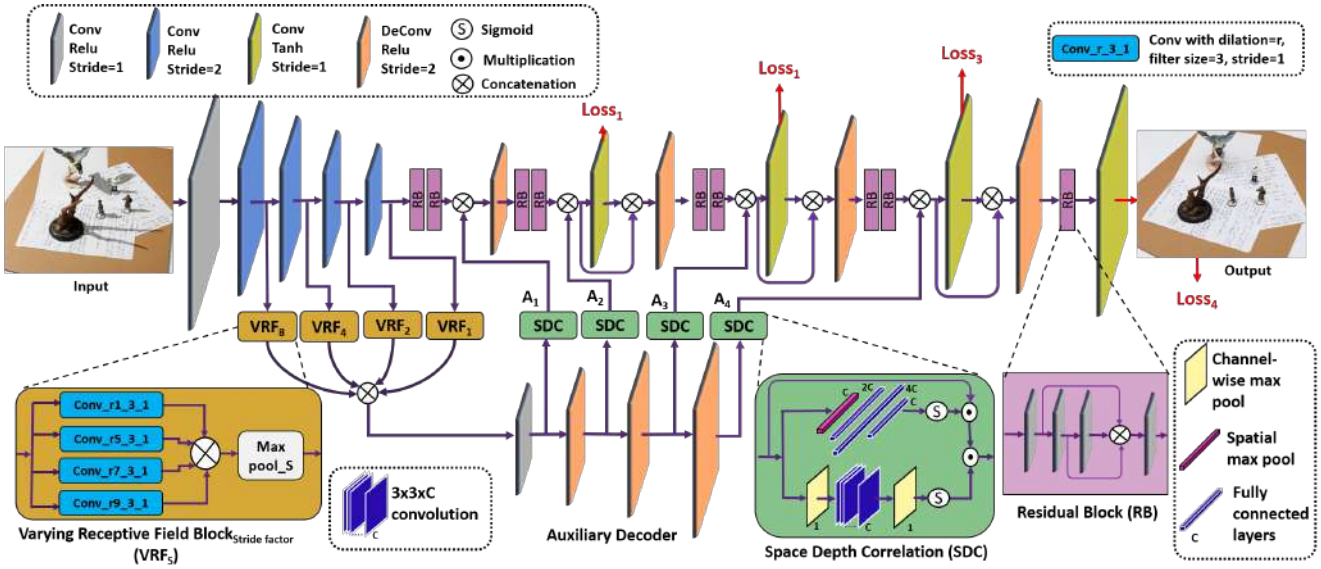

Figure 14. A visual representation of the solution proposed by Team CVPR_IITRPR.

the generated feature map is fed into two standard residual blocks to reconstruct the final shadow-removed image.

FTR [71] is a common practice in end-to-end image recovery architectures, consisting of employing a ResBlock which learns the difference between blurred and clear image pairs.

4.17. Concentration

Team Concentration extends on their previous work [73], proposing a novel cleanness-navigated-shadow network (CNSNet), with a shadow-oriented adaptive normalization (SOAN) module and a shadow-aware aggregation with transformer (SAAT) module based on the shadow mask information. Under the guidance of the localization information provided by the shadow mask, the SOAN mod-

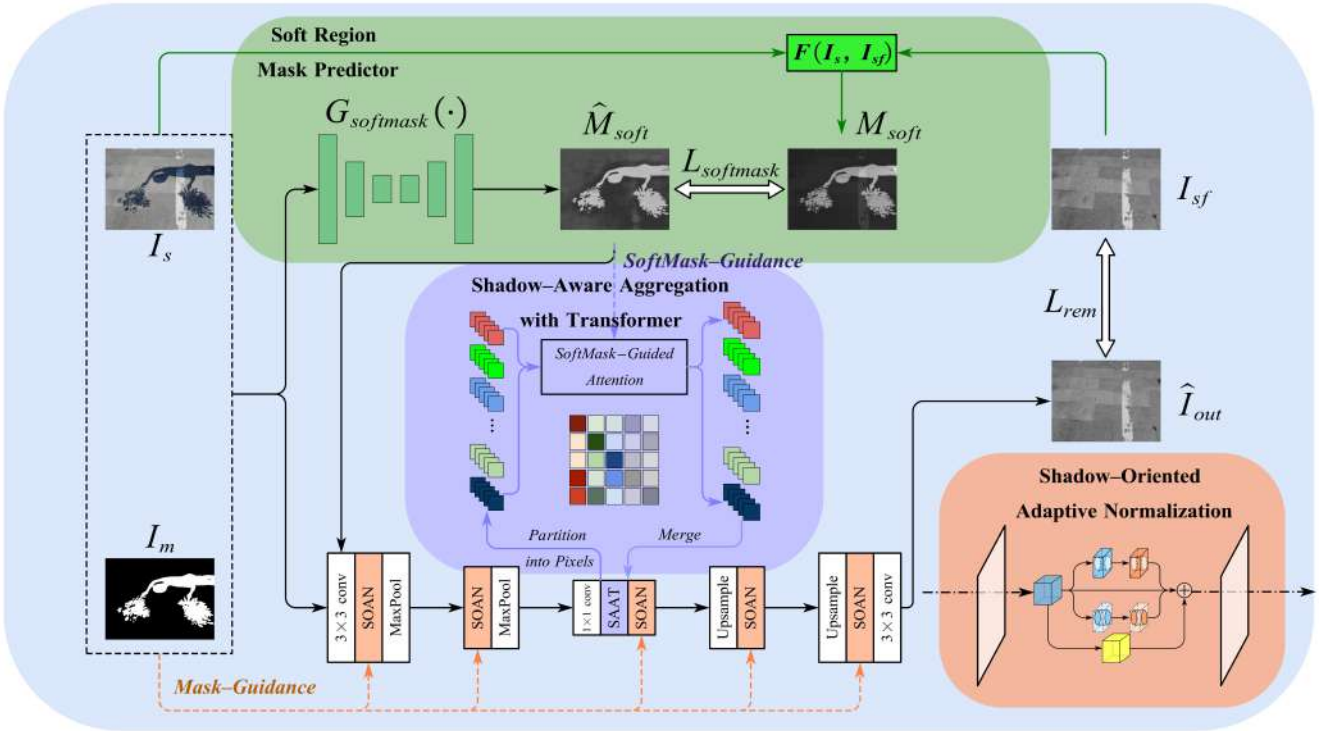

Figure 15. Illustration of the Concentration Team proposed cleanness-navigated-shadow network (CNSNet) for shadow removal. It involves three key elements: soft-region mask predictor (green box), shadow-oriented adaptive normalization (SOAN) module (orange box), and shadow-aware aggregation with transformer (SAAT) module (purple box). First, the predictor takes in a shadow image and its corresponding shadow mask to obtain a soft-region mask. Then, both hard and soft masks are concatenated with the input image, entering the UNet-like network to produce the shadow-free results. Note that the guidance (dotted arrows) of both hard and soft masks is applied in the region-wise SOAN and pixel-wise SAAT modules, respectively.

ule formulates the statistics from the non-shadow region and adaptively applies them to the shadow region for region-wise restoration. The SAAT module utilizes the shadow mask to precisely guide the restoration of each shadowed pixel by considering the highly relevant pixels from the shadow-free regions for global pixel-wise restoration.

4.18. ZJUME251

Unlike the previous ISTD dataset [65], most of the shadows in the challenge dataset are produced by the object itself under the influence of illumination rather than the projection of an external object. Generating shadow mask images using a fixed threshold segmentation method is challenging. Therefore, Team ZJUME251 proposes a shadow mask image generation algorithm with better robustness. By converting the image to YUV color space, extracting its Y-channel image, and calculating the difference between $I_{\text{shadow-free}}$ and I_{shadow}, they produced more accurate shadow masks. Inspired by [81], they retrain the shadow detector network using the masks estimated using the luminance component if the YUV representation.

Following the design approach of [37], they propose the Shadow Image Decomposition and Reconstruction Network (SIDRN) as a solution for the single image shadow removal task. Considering a linear model between the shadow-free pixels and the pixels in shadow affected regions, a shadow parameter estimator is designed to predict the (w, b) pair, that can be used to estimate a relit image in a similar strategy as the one used in [35]. Using the described linear model, the relit image can be computed as shown in Equation 9.

$$I_{\text{relit}} = w \cdot I_{\text{shadow}} + b. \tag{9}$$

According to well-known image decomposition system [14], the shadow-free image can be expressed as in Equation 10.

$$I_{\text{shadow-free}} = I_{\text{shadow}} \cdot (1 - \alpha) + I_{\text{relit}} \cdot \alpha. \tag{10}$$

The relit image, together with the shadow affected image and the shadow mask are input into the shadow matte prediction network to get the shadow-free image. Finally, considering non-linear and variable illumination conditions, an inpainting network is designed to further refine the shadow-free recovered images.

4.19. Imsensor

In terms of the overall architecture, the Imsensor proposed two-stage model achieves coarse-grained to fine-grained optimization of the entire training process through the design of two UNets [51]. Firstly, they use an adaptive method to generate a shadow mask of the original image, that is then used to generate the complementary non shadow mask. This set of masks is used to guide an attention mechanism performing the information exchange between the two UNet structures. Finally, the set of learnt feature maps is pushed through a confrontation network that estimates the final shadow free recovered image.

5. Conclusion

For the first edition, the NTIRE23 Challenge for Image Shadow Removal enjoyed significant attention from the computer vision community. A number of 144 teams participated in the NTIRE 2023 Image Shadow Removal Challenge, of which 19 teams were ranked in the final phase. The described solutions show novelty in proposing new architectures, or tailoring well established models to the particularities of the provided data. Several novel solutions were proposed by our participants, improving over the existing state-of-the art.

The final ranking of the teams focuses on the perceptual quality of their submitted results, being based on the resulting MOS of our user study, the LPIPS distance, and the SSIM score. The reconstruction fidelity is shown by the PSNR values. We aim at encouraging the research community into the direction of more efficient solutions, with a larger deployment potential. The feedback from the challenge participants provided insightful ideas for the following editions of the challenge.

Acknowledgments

This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2023 workshop and challenges sponsors: Sony Interactive Entertainment, Meta Reality Labs, ModelScope, ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab).

A. Teams and Affiliations

NTIRE 2023 Team

Title:
NTIRE 2023 Challenge on Image Shadow Removal
Members:
Florin-Alexandru Vasluianu[1], Tim Seizinger[1], Radu Timofte[1,2]
Affiliations:
[1] Computer Vision Lab, IFI & CAIDAS, University of Würzburg
[2] Computer Vision Lab, ETH Zürich, Switzerland

MTCV Team

Title:
Pyramid Structure for High Resolution Image Shadow Removal.
Members:
Shuhao Cui[1], Junshi Huang, Shuman Tian, Mingyuan Fan, Jiaqi Zhang, Li Zhu, Xiaoming Wei, Xiaolin Wei.
Affiliations:
[1] Meituan Group.

IR-SDE Team

Title:
Refusion: Enabling Large-Size Realistic Image Restoration with Latent-Space Diffusion Models [44].
Members:
Ziwei Luo[1], Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön
Affiliations:
[1] Department of Information Technology, Uppsala University, Sweden

SRDM Team

Title:
SRDM: Shadow Removal Diffusion Model.
Members:
Xiaoyi Dong[1], Xi Sheryl Zhang[1,2], Chenghua Li[1,2], Cong Leng[1,3]
Affiliations:
[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Nanjing Artificial Intelligence Research of IA (AiRiA), Nanjing, China
[3] MAICRO, Nanjing, China

SYU-HnVLab Team

Title:
CLAN: Color and Luminance Attention Network for Shadow Removal / LASR: Luminance Attention Network for Shadow Removal.
Members:
Woon-Ha Yeo[1], Wang-Taek Oh, Yeo-Reum Lee, Han-Cheol Ryu
Affiliations:
[1] Department of Artificial Intelligence Convergence, Sahmyook University, Seoul, Republic of Korea

MegSRD Team

Title:
Shadow Removal with Noise Merge Denoising Diffusion Models.
Members:
Jinting Luo[1], Chengzhi Jiang[1], Mingyan Han[1], Qi Wu[1], Wenjie Lin[1], Lei Yu[1], Xinpeng Li[1], Ting Jiang[1], Haoqiang Fan[1], and Shuaicheng Liu[1,2]
Affiliations:
[1] Megvii Technology
[2] University of Electronic Science and Technology of China

UM-JTG Team

Title:
Curve-Guided Swin Transformer for Image Shadow Removal.
Members:
Shuning Xu[1], Binbin Song, Xiangyu Chen, Shile Zhang, Jiantao Zhou
Affiliations:
[1] University of Macau

LVGroup_HFUT Team

Title:
Rethinking resolution-altering and dense-encoding in NAFNet for image shadow removal.
Members:
Zhao Zhang[1], Suiyi Zhao, Huan Zheng, Yangcheng Gao, Yanyan Wei, Bo Wang, Jiahuan Ren, Yan Luo
Affiliations:
[1] Hefei University of Technology (HFUT)

IM_TTI Team

Title:
ShadowFormer+ : Rethinking the ShadowFormer learning methods and removing the effect of external camera parameter changes
Members:
Yuki Kondo[1], Riku Miyata, Fuma Yasue, Taito Naruki, Norimichi Ukita
Affiliations:
[1] Toyota Technological Institute, Japan

NTU607-shadow Team

Title:
ShadowFormer+ : Rethinking the ShadowFormer learning methods and removing the effect of external camera parameter changes
Members:
Hua-En Chang[2], Hao-Hsiang Yang[2], Yi-Chung Chen[3], Yuan-Chun Chiang[2], Zhi-Kai Huang[2], Wei-Ting Chen[1], I-Hsiang Chen[2], Chia-Hsuan Hsieh[4], Sy-Yen Kuo[2]
Affiliations:
[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

MM911 Team

Title:
Mask Free Dual-Stage Network for Shadow Removal.
Members:
Li Xianwei[1], Huiyuan Fu, Chunlin Liu, Huadong Ma, Binglan Fu, Huiming He
Affiliations:
[1] Beijing University of Post and Teleconmunication, Beijing, China

leaves Team

Title:
A New Method to Generate Mask of Shadow Image to Help Shadow Removal Task.
Members:
Mengjia Wang[1], Wenxuan She, Yu Liu
Affiliations:
[1] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education

MegSRF Team

Title:
Shadow Removal through Luminance Retuning.
Members:
Chengzhi Jiang[1], Wenjie Lin[1], Jinting Luo[1], Ting Jiang[1], Qi Wu[1], Mingyan Han[1], Xinpeng Li[1], Lei Yu[1], Haoqiang Fan[1] and Shuaicheng Liu[1,2]
Affiliations:
[1] Megvii Technology
[2] University of Electronic Science and Technology of China

Couger_AI Team

Title:
Light-DeShadowNet: A Light-weight CNN for Shadow-Removal Network.
Members:

Sabari Nathan[1], Priya Kansal
Affiliations:
[1] Couger Inc.

Noir Team

Title:
Image Shadow Removal with U-net based model and Discriminator.
Members:
Zhongjian Zhang[1], Huabin Yang, Yan Wang, Yanru Zhang
Affiliations:
[1] University of Electronic Science and Technology of China

CVPR_IITRPR Team

Title:
A lightweight Architecture with Auxiliary Decoder for Image Shadow Removal.
Members:
Shruti S. Phutke[1], Ashutosh Kulkarni[1], MD Raqib Khan[1], Subrahmanyam Murala[1], Santosh Kumar Vipparthi[1]
Affiliations:
[1]Computer Vision and Pattern Recognition Lab, Indian Institute of Technology Ropar, Rupnagar Punjab, INDIA

NUSSZ-ShadowRemove

Title:
Image Shadow Removal with U-net based model and Discriminator.
Members:
Heng Ye[1], Zixi Liu[1], Xingyi Yang[2], Songhua Liu[2], Yinwei Wu[3], Yongcheng Jing[4]
Affiliations:
[1] National University of Singapore (Suzhou) Research Institute
[2] National University of Singapore
[3] National University of Singapore (Chongqing) Research Institute
[4] University of Sydney

Concentration Team

Title:
A Cleanness-Navigated-Shadow Network for Shadow Removal.
Members:
Qianhao Yu[1], Naishan Zheng, Jie Huang, Yuhang Long, Mingde Yao, Feng Zhao
Affiliations:
[1] Brain-Inspired Vision Laboratory, Information Science and Technology Institution, University of Science and Technology of China

ZJUME251 Team

Title:
Shadow Image Decomposition and Reconstruction Network.
Members:
Bowen Zhao[1,2], Nan Ye[1,2], Ning Shen[1,2], Yanpeng Cao[1,2]
Affiliations:
[1] State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, 310027, China
[2] Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, 310027, China

Imsensor Team

Title:
StepShadowMaskGAN
Members:
Tong Xiong[1], Weiran Xia,Dingwen Li,Shuchen Xia
Affiliations:
[1] South China University of Technology

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 1

[2] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2020 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 490–491, 2020. 2

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2021 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 627–646, 2021. 2

[4] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[5] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014. 1

[6] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Ying Shan, Gen

Li, Radu Timofte, et al. NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[7] Hua-En Chang, Chia-Hsuan Hsieh, Hao-Hsiang Yang, I-Hsiang Chen, Yi-Chung Chen, Yuan-Chun Chiang, Wei-Ting Huang, Zhi-Kai Chen, and Sy-Yen Kuo. TSRFormer: Transformer based two-stage refinement for single image shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 8

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 2, 5, 8, 10, 11

[9] Wei-Ting Chen, Kuan-Yu Chen, I-Hsiang Chen, Hao-Yu Fang, Jian-Jiun Ding, and Sy-Yen Kuo. Missing recovery: Single image reflection removal based on auxiliary prior learning. *IEEE Transactions on Image Processing*, 32:643–656, 2022. 8

[10] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen 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*, pages 17653–17662, 2022. 8

[11] Wei-Ting Chen, Hao-Lun Lou, Hao-Yu Fang, I-Hsiang Chen, Yi-Wen Chen, Jian-Jiun Ding, and Sy-Yen 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*, 32(6):3346–3359, 2021. 8

[12] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 8

[13] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *CVPR*, 2020. 8

[14] Yung-Yu Chuang, Dan B Goldman, Brian Curless, David H Salesin, and Richard Szeliski. Shadow matting and compositing. In *ACM SIGGRAPH 2003 Papers*, pages 494–500. 2003. 14

[15] Marcos V Conde, Manuel Kolmet, Tim Seizinger, Thomas E. Bishop, Radu Timofte, et al. Lens-to-lens bokeh effect transformation. NTIRE 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[16] Marcos V Conde, Florin Vasluianu, Sabari Nathan, and Radu Timofte. Real-time under-display cameras image restoration and hdr on mobile devices. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 747–762. Springer, 2023. 4

[17] Marcos V. Conde, Florin Vasluianu, Javier Vazquez-Corral, and Radu Timofte. Perceptual image enhancement for smartphone real-time applications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1848–1858, January 2023. 4

[18] Marcos V Conde, Eduard Zamfir, Radu Timofte, et al. Efficient deep models for real-time 4k image super-resolution. NTIRE 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[19] Shuhao Cui, junshi huang, Shuman Tian, Mingyuan Fan, jiaqi zhang, Li Zhu, Xiaoming Wei, and Xiaolin Wei. Pyramid ensemble structure for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 5

[20] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion, 2020. 6

[21] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014. 1

[22] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 1

[23] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018. 1

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[25] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 12

[26] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 2, 4, 8, 10, 11

[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

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

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[30] Pakorn KaewTraKulPong and Richard Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*, pages 135–144. Springer, 2002. 1

[31] Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. NTIRE 2023 video col-

orization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[33] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–23, 2015. 1

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. 3

[35] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 9, 14

[36] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *The IEEE European Conference on Computer Vision (ECCV)*, August 2020. 2

[37] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9088–9101, 2021. 14

[38] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[39] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[40] Xiaohong Liu, Xiongkuo Min, Wei Sun, Yulun Zhang, Kai Zhang, Radu Timofte, Guangtao Zhai, Yixuan Gao, Yuqin Cao, Tengchuan Kou, Yunlong Dong, Ziheng Jia, et al. NTIRE 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[41] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 9

[43] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023. 5, 7

[44] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 5, 15

[45] Sabari Nathan and Priya Kansal. Skeletonnetv2: A dense channel attention blocks for skeleton extraction. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2142–2149, 2021. 11

[46] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016. 2

[47] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 8

[48] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*, 2022. 3

[49] Shruti S Phutke and Subrahmanyam Murala. Pseudo decoder guided light-weight architecture for image inpainting. *IEEE Transactions on Image Processing*, 31:6577–6590, 2022. 12

[50] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2308–2316, July 2017. 2

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 11, 15

[52] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv:2104.07636*, 2021. 5

[53] Mrinmoy Sen, Sai Pradyumna Chermala, Nazrinbanu Nurmohammad Nagori, Venkat Peddigari, Praful Mathur, B H Pawan Prasad, and Moonhwan Jeong. Shards: Efficient shadow removal using dual stage network for high-resolution images. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1809–1817, 2023. 2, 9

[54] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2016. 1

[55] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan A Ermakov, Nikola Banic, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, et al. NTIRE 2023 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[56] Marc Stamminger and George Drettakis. Perspective shadow maps. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 557–562, 2002. 1

[57] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 1

[58] Florin Vasluianu and Radu Timofte. Efficient video enhancement transformer. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4068–4072, 2022. 4

[59] Florin-Alexandru Vasluianu, Andrés Romero, Luc Van Gool, and Radu Timofte. Shadow removal with paired and unpaired learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 826–835, 2021. 2

[60] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu 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. 2

[61] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. NTIRE 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[62] 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. 2

[63] Jifeng Wang, Xiang Li, Le Hui, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2017. 2

[64] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 2, 9

[65] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 14

[66] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[67] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 5

[68] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[69] 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. 3

[70] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[71] Mao Xintian, Liu Yiming, Liu Fengze, Li Qingli, Shen Wei, and Wang Yan. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023. 13

[72] Woon-Ha Yeo, Wang-Taek Oh, Kyung-Su Kang, Young-Il Kim, and Han-Cheol Ryu. Cair: fast and lightweight multi-scale color attention network for instagram filter removal. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 714–728. Springer, 2023. 5

[73] Qianhao Yu, Naishan Zheng, Jie Huang, and Feng Zhao. Cn-snet: A cleanness-navigated-shadow network for shadow removal. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 221–238. Springer, 2023. 13

[74] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, et al. NTIRE 2023 challenge on hr depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

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

[76] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[77] 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. 3

[78] Xiao Feng Zhang, Chao Chen Gu, and Shan Ying Zhu. Spaformer: Transformer image shadow detection and removal via spatial attention. *arXiv e-prints*, pages arXiv–2206, 2022. 2, 8, 9, 12

[79] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, et al. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[80] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 2018. 5

[81] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018. 14