

# Supplementary Materials

## Enhancing Images with Coupled Low-Resolution and Ultra-Dark Degradations: A Tri-level Learning Framework

Anonymous Authors

### 1 SUPPLEMENTARY MATERIAL OVERVIEW

In the following supplementary material, we first provide the detailed network architectures of our proposed approach (Sec. 2). Then we describe the details of benchmarking datasets and ablation studies (Sec. 3). Finally, we provide additional experiments and ablation analysis (Sec. 4).

### 2 DETAILED NETWORK ARCHITECTURE

#### 2.1 Architecture: Illumination Regulator

As shown in Fig. 1, the network architecture of our illumination regulator  $N_{ir}(\omega_i)$  is based on a shallow U-net structure, comprising three encoding layers with max pooling for downsampling, and three decoding layers with bilinear interpolation for upsampling. This strategic design integrates three distinct encoding layers employing max pooling to efficiently reduce the spatial dimensions for downsampling, capturing the essential low-level features while progressively diminishing computational complexity. Conversely, the architecture boasts three symmetric decoding layers, utilizing bilinear interpolation for upsampling, effectively restoring the resolution of the processed features. These layers are adept at reconstructing the intricate spatial hierarchies of illumination attributes. We extract illumination features (i.e.,  $f_i^{[o]}$ ) from the final three decoding layers to prepare for modulation of the reflectance map  $V$  in the subsequent stage. In the last layer, we introduce an activation function with Sigmoid to obtain the modulation factors (i.e.,  $\alpha$  and  $\beta$ ). This is then used in a weighted operation, followed by a point-wise division with the input image. Finally, a clamping operation is performed to obtain the reflectance map  $V$ . This deliberate choice facilitates a nuanced adjustment of the incoming features, allowing for a refined weighting operation that precedes a methodical point-wise division with the original input image. This division operation is key to disentangling the reflectance components from the raw captured data. Our illumination regulator is thus a testament to a harmonious fusion of structural ingenuity and functional precision, crafted to empower the subsequent stages of our computational framework with refined and contextually informed illumination features for superior image enhancement results.

#### 2.2 Architecture: Feature Refinement Network

As depicted in Fig. 2, the network architecture for our feature refinement  $N_{fs}(\omega_s)$  is built upon a shallow U-net framework. It consists of three encoder layers incorporating transformer-based Encoder Blocks [6, 12] and max pooling for downsampling, complemented by three decoder layers that utilize IHM modules and bilinear interpolation for upsampling. In the last three layers of the decoder, the IHM modules receive illumination prior features from the previous stage and multi-layer semantic hint features from frozen

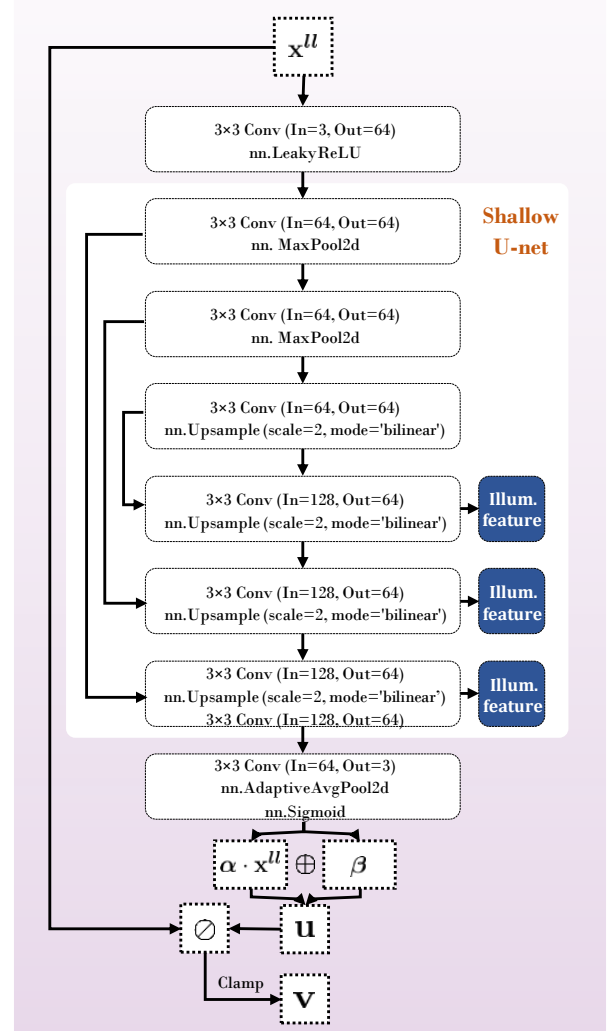


Figure 1: Network architecture of our illumination regulator.

SAM [5, 14] of the general semantic model. The architecture’s synergy is encapsulated in the blue and pink box annotations within the figure, symbolizing the dual modulation strategy that enables the adaptive recalibration and finessing of reflectance characteristics. This process is pivotal for the elucidation of intricate feature details in the reconstructed image. The final layer introduces a dynamic up-sampling module [3] that yields the super-resolved output. In the nuanced construction of the FFN (Feed-Forward Network) within the encoder transformer and the IHM, we opt for

a novel dual-pathway design. Each path undergoes a linear transformation, with one being subjected to the GELU (Gaussian Error Linear Unit) non-linearity [4], fostering a fusion of feature maps with diversified receptive fields. Depth-wise convolution ingrains the positional relevance of adjacent pixels within the information-rich feature space, thereby accentuating the enrichment of features with pertinent contextual data. This strategic formulation is instrumental in the network's capability to discern and harness local image structures pivotal for the restoration task at hand. This refined network architecture, therefore, is not merely a series of layers but a thoughtful orchestration of feature flow and refinement, tailored to learn and enhance the salient structural components of images for high-fidelity restoration.

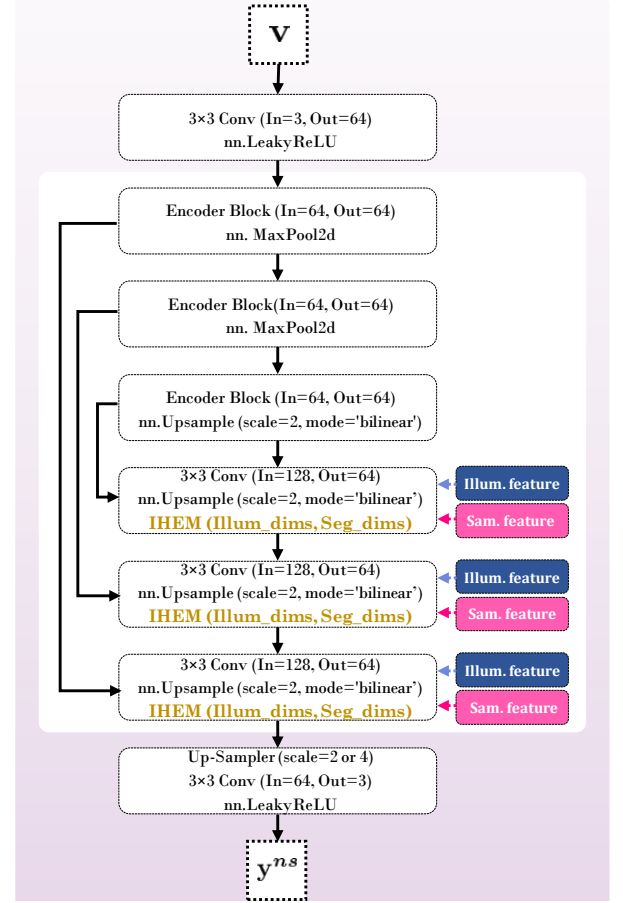
### 3 BENCHMARKING DETAILS

**Data Preparation.** We evaluated the benchmark performance of all compared methods across four widely recognized datasets: 1) RELISUR<sup>1</sup>, 2) DarkFace [11]<sup>2</sup>, 3) Dark-Zurich [10], and 4) Cityscapes [2]. We commenced our experimentation using the authentic RELISUR dataset as our primary training set, with an input dimension of 624x624 pixels. This dataset encompasses 1045 image pairs spanning three different resolution scales ( $\times 1$ ,  $\times 2$ , and  $\times 4$ ) and five distinct low-light levels (ranging from -2.5EV to -5.0EV), consisting of low-resolution, low-illumination images alongside their high-resolution, normal-light counterparts. To broadly assess the real-world performance of our proposed method under various authentic low-light open scenes and different exposure conditions, the remaining three datasets were employed as extensive validation sets. Specifically, the DarkFace dataset, comprising 1000 images with dimensions of 1024x720, was utilized to further gauge the generalization capability of our method in real-world nighttime scenarios. Additionally, we selected a subset of 151 nighttime autonomous driving scenes from the Dark-Zurich open dataset, with dimensions downsampled to 480x270, to serve as a test set. Lastly, data from the Cityscapes dataset pertaining to the Strasbourg location during nighttime were chosen and subjected to a darkening process, adopting a day-to-night domain transfer simulation scheme [7], with dimensions set to 1024x512. These diverse datasets facilitated a comprehensive evaluation of our approach across a variety of challenging low-light conditions.

**Implementation Details.** We adhere to the tri-level learning strategy as outlined in Alg. 1 for our network training, with the total number of iterations set to 150,000. We utilize the Adam optimizer with beta values configured at [0.9, 0.999]. The initial learning rates for the upper and lower layers of the two stages are set to  $\gamma_u = 1e-4$ ,  $\gamma_l = 2e-4$ ,  $\sigma_u = 1e-4$  and  $\sigma_l = 1e-4$ , respectively. A cosine annealing restart strategy is implemented for cyclic learning rate scheduling. The dataset  $\{\mathcal{D}\}$  is partitioned into  $\{\mathcal{D}_{ul}\} \cup \{\mathcal{D}_{ll}\}$  and at a distribution ratio of 1 : 5. Experiments are conducted using PyTorch version 2.0.1, which supports CUDA 11.7, on a single NVIDIA RTX A6000 GPU with 48GB of RAM. A curriculum progressive training process is introduced, where the size of input image patches and the batch size are progressively adjusted in accordance with the training progress. In the initial stages, smaller image

<sup>1</sup><https://vap.aau.dk/rellisur/>

<sup>2</sup><https://flyywh.github.io/CVPRW2019LowLight/>



**Figure 2: Network architecture of our feature refinement network.**

patches and larger batch sizes are used, while in later stages, the sizes of the image patches increase and batch sizes decrease. Three data augmentation operations are incorporated: random cropping, random rotation, and random flipping.

## 4 ADDITIONAL EXPERIMENTS

### 4.1 Results on DarkFace Dataset

The comparative visualization presented in Fig. 3 demonstrates the performance of various methods on the challenging DarkFace dataset, including PAN [15], SwinIR [6], MIRNet [13], Restormer [12], SRFormer [16], and HAT [1]. The input images, characterized by their low-light conditions, pose significant difficulty for feature discernment, color fidelity, and texture preservation. The PAN method appears to over-enhance the scene, leading to unnatural color saturation and loss of detail in brighter areas. MIRNet's output shows improved visibility; however, it introduces a red tint across the image, potentially compromising natural color representation. SwinIR delivers better color balance but still struggles with artifact preservation around light sources. Restormer offers a notable improvement in terms of color accuracy and detail enhancement, yet some

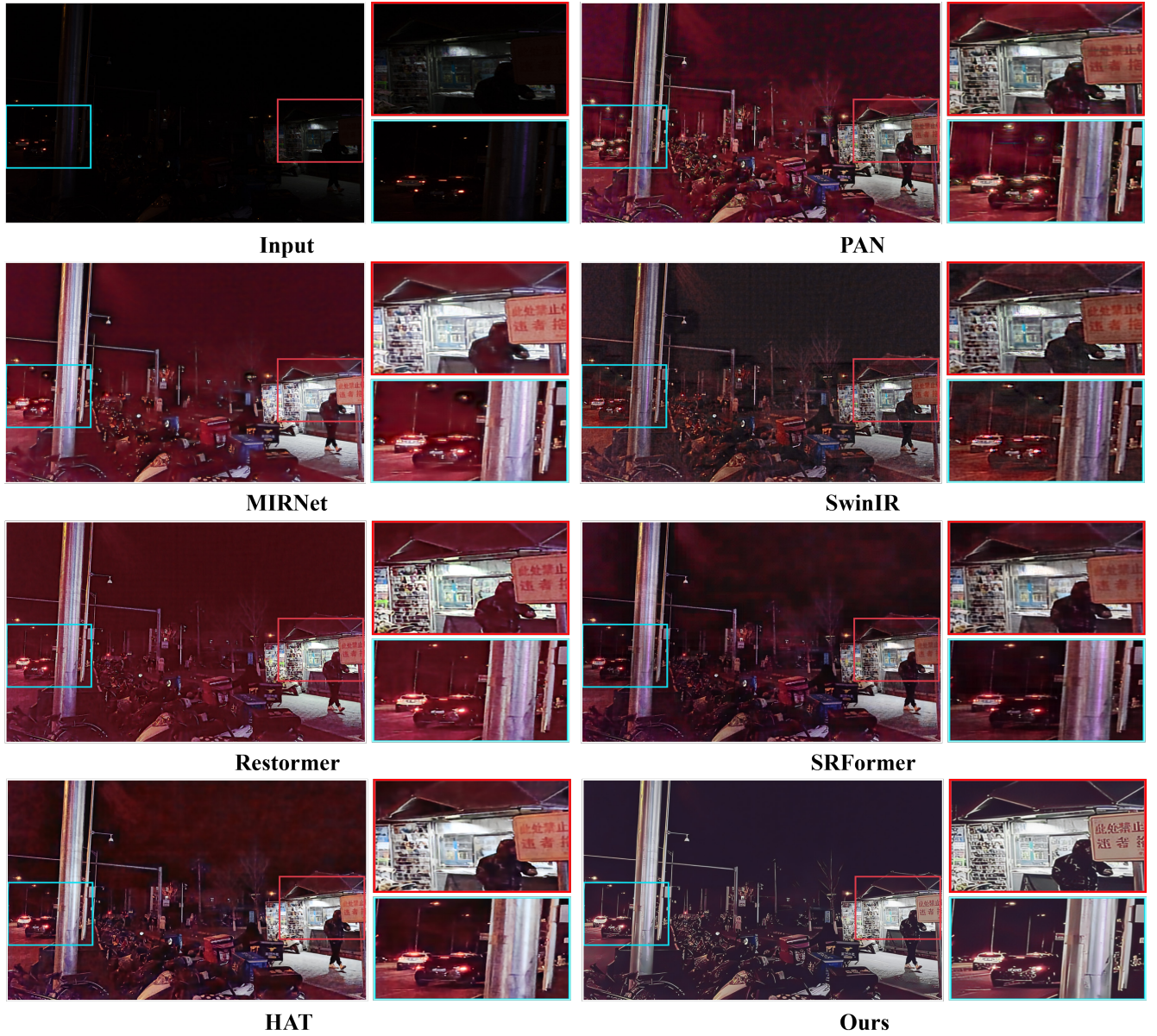


Figure 3: Qualitative segmentation results on the DarkFace [11] dataset.

regions remain underexposed, affecting the visibility of finer details. SRFormer, while adept at illumination enhancement, seems to smooth out some of the textures, leading to a loss of detail in certain areas. The HAT method provides a balanced enhancement, although some parts of the image appear slightly washed out, indicating a potential compromise in dynamic range and contrast. Our proposed solution, in contrast, stands out for its ability to maintain a natural appearance while significantly improving overall visibility. It demonstrates superior texture clarity and color rendering, preserving both the integrity of the scene and the details necessary for accurate perception and analysis. Overall, our approach yields a comprehensive improvement in image quality, addressing the

intrinsic complexities of low-light image processing with greater efficacy than other methods showcased.

## 4.2 Results on Darken Cityscapes Dataset

Fig. 4 showcases a side-by-side comparison of enhancement methods applied to the Darken Cityscapes Dataset, which is characterized by its challenging low-light conditions. The 'Raw' images serve as our baseline, depicting urban scenes with underexposed areas that obscure critical details. When artificially darkened, these 'Darken' images further accentuate the challenge, with essential features becoming nearly indiscernible. The 'Restormer' method makes notable improvements in brightness and contrast, yet it

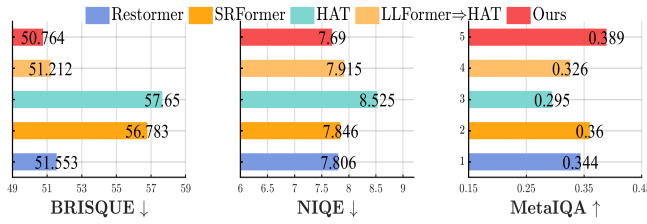




**Figure 4: Qualitative segmentation results on the darken Cityscapes [2] dataset.**

tends to introduce a slight yellow bias, which can be observed in the over-saturation of the yellow car and the building façades. The 'SRFormer' algorithm, while it enhances overall luminance, appears to struggle with noise, particularly in the darker regions, leading to a grainy texture. 'HAT' provides a more balanced enhancement, correcting the exposure with fewer color shifts. However, it falls short in recovering finer details, as evidenced by the somewhat blurred appearance of architectural elements and street surfaces. The 'LLFormer  $\Rightarrow$  HAT' combination displays a marked improvement in color accuracy and detail preservation. Nonetheless, it exhibits some

oversmoothing, particularly in areas requiring subtle texture differentiation. Our method, in contrast, achieves a harmonious balance, significantly enhancing visibility without compromising natural color tones or introducing undue artifacts. It adeptly restores the vibrancy of the scene, including the clarity of line markings on the road and the intricate brickwork of the buildings, which are now distinct and sharp. The enhancement effectively illuminates the scene, bringing forth details that were previously lost in the shadows while maintaining the integrity of the original colors and textures. Overall, our approach stands out for its ability to deliver a clear and authentic representation of the nighttime cityscape,



**Figure 5: Quantitative results on three metrics for darken Cityscapes [2] dataset.**

offering an improved visual experience that is closest to what one would expect in naturally lit conditions.

Fig. 5 illustrates the quantitative performance of enhancement techniques in terms of Natural Image Quality Evaluator (NIQE) [9], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [8] and MetaQA [17]. Our method registers a BRISQUE score improvement of approximately 1% over the second-best, Restormer, suggesting marginally enhanced naturalness in image quality. In the NIQE assessment, our score improves by roughly 3% compared to the HAT method, indicating a closer resemblance to natural image statistics. Notably, our method achieves a MetaQA score that is about 16% higher than LLFormer  $\Rightarrow$  HAT, the next closest competitor, highlighting a significant leap in perceived image quality. These metrics collectively affirm the superior performance of our approach in rendering high-fidelity images from the Darken Cityscapes dataset.

### 4.3 Ablation Studies

In Tab. 1, we dissect the impact of the LII module and the  $\mathcal{L}_{srl}$  loss on our model’s performance. These studies were conducted on the RELISUR dataset for the  $2 \times$  upscaling task, with findings quantified across three metrics: PSNR, SSIM, and LPIPS. The full model, encapsulating all designed features and loss functions, sets a high bar with a PSNR of 22.456, an SSIM score of 0.744, and an LPIPS value of 0.304, illustrating a robust enhancement capacity. When the LII module is excluded (w/o LII), there’s a noticeable dip in performance across all metrics: PSNR drops by approximately 2%, SSIM by around 4%, and LPIPS increases by roughly 28%, which correlates to a decrease in image quality and structural accuracy. Moreover, the variant without the  $\mathcal{L}_{srl}$  loss (w/ LII (w/o  $\mathcal{L}_{srl}$ )) also displays a performance reduction, with a PSNR decrease by nearly 1% and SSIM by about 3%, but a less pronounced increase in LPIPS by approximately 23%. These results underline the significance of the  $\mathcal{L}_{srl}$  loss in fine-tuning the perceptual details and achieving closer alignment with human visual perception. The ablation study clearly demonstrates the integral roles that both the LII module and  $\mathcal{L}_{srl}$  loss play in our model’s superior performance. The LII module’s impact is particularly noteworthy in maintaining structural integrity and perceptual quality, which are critical for real-world applications of super-resolution.

**Table 1: Ablation studies for LII module and  $\mathcal{L}_{srl}$  loss. We report quantitative results on three metrics on the  $2 \times$  task based on RELISUR dataset.**

NO	Model	PSNR↑	SSIM↑	LPIPS↓
1.	Ours (Full model)	22.456	0.744	0.304
2.	w/o LII	21.969↓0.487	0.710↓0.034	0.390↑0.086
3.	w/ LII (w/o $\mathcal{L}_{srl}$ )	22.04↓0.416	0.712↓0.032	0.396↑0.092

### REFERENCES

- [1] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. 2023. Activating More Pixels in Image Super-Resolution Transformer. In *CVPR*. 22367–22377.
- [2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Endzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*.
- [3] Yutong Dai, Hao Lu, and Chunhua Shen. 2021. Learning Affinity-Aware Upsampling for Deep Image Matting. In *CVPR*. 6841–6850.
- [4] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
- [5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).
- [6] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *CVPR*. 1833–1844.
- [7] Rundong Luo, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2023. Similarity min-max: Zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8104–8114.
- [8] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708.
- [9] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* 20, 3 (2012), 209–212.
- [10] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2020. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2020), 3139–3153.
- [11] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Zhang, and et al. 2020. Advancing Image Understanding in Poor Visibility Environments: A Collective Benchmark Study. *IEEE Transactions on Image Processing* 29 (2020), 5737–5752.
- [12] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*. 5728–5739.
- [13] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2020. Learning enriched features for real image restoration and enhancement. In *ECCV*. 492–511.
- [14] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv preprint arXiv:2306.14289* (2023).
- [15] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. 2020. Efficient image super-resolution using pixel attention. In *ECCV*. 56–72.
- [16] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. 2023. SRFormer: Permuted Self-Attention for Single Image Super-Resolution. In *CVPR*.
- [17] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. 2020. MetaQA: Deep Meta-Learning for No-Reference Image Quality Assessment. In *Computer Vision and Pattern Recognition*. 14143–14152.