The 29th Meeting on Image Recognition and Understanding

CCDiT: One-Step Diffusion Transformer for Color-Consistent Image Super-Resolution

Yitian MA[1,a)] Longhao SUN[1,b)] Peilin XIONG[1,c)] Keiji YANAI[1,d)]

Abstract

Real-world image super-resolution (SR) aims to recover high-resolution images from low-resolution inputs degraded by unknown blur, noise, compression, and sensor processes. Diffusion-based SR methods can synthesize realistic textures, but multi-step sampling is expensive. Recent one-step variants greatly accelerate inference, yet we observe that they often introduce color drift while hallucinating high-frequency details, a problem that has received little explicit attention in one-step SR. We propose CCDiT, a one-step Diffusion Transformer framework that adapts Qwen-Image-Edit to real-world SR with LoRA fine-tuning and treats color consistency as an explicit restoration objective. CCDiT further introduces Alpha Mask fusion in the YCbCr space, preserving the luminance channel of the Qwen SR output and adaptively mixing chrominance from the Qwen SR output and the bicubic reference image. Experiments on RealSR, DRealSR, and DIV2K-Val show that CCDiT improves several reconstruction and perceptual metrics with one-step inference, while the color-fusion study shows that the bicubic reference image is useful as a chrominance anchor when its colors are reliable.

1. Introduction

Image super-resolution restores high-resolution images from low-resolution inputs and is used when image detail must be recovered from limited visual evidence [1], [14], [19]. Real-world SR is challenging because degradations are usually unknown and may include blur, noise, compression artifacts, sensor degradation, and missing texture. Diffusion models provide strong restoration priors for this setting [7], [15], [18], [19], [22], [23], but standard diffusion restoration requires many denoising steps. Recent one-step methods such as OSEDiff, TSD-SR, and ODTSR reduce this cost by restoring an image with a single denoising step [4], [5], [27].

However, we find that this efficiency gain comes with an overlooked failure mode: color drift. Although one-step diffusion SR substantially accelerates inference, its restored images often exhibit chrominance shifts when hallucinating high-frequency textures. Such shifts are not merely cosmetic: color can affect scene understanding, sign and text reading, product inspection, food analysis, and other visually sensitive tasks [2], [21]. To our knowledge, existing one-step SR studies mainly optimize speed, texture realism, and perceptual quality, but have not explicitly formulated color consistency as the central problem. A practical one-step SR method should therefore preserve diffusion acceleration while limiting avoidable color shifts.

We propose CCDiT, a color-aware one-step Diffusion Transformer framework for real-world SR. CCDiT first adapts Qwen-Image-Edit to one-step SR through LoRA fine-tuning. It uses a bicubic reference image as the restoration start point and predicts the Qwen SR output with one denoising step. CCDiT then freezes the Qwen SR backbone and trains Alpha Mask Net, a lightweight module that performs adaptive YCbCr chrominance fusion between the Qwen SR output and the bicubic reference image.

The contributions are threefold. First, we identify color drift as an underexplored failure mode of one-step diffusion SR and formulate color consistency as an explicit objective. Second, we adapt Qwen-Image-Edit into a one-step Diffusion Transformer SR backbone and introduce Alpha Mask fusion, which preserves the luminance of the Qwen SR output and adaptively fuses CbCr chrominance from the Qwen SR output and the bicubic reference image. Third, we evaluate CCDiT on RealSR, DRealSR, and DIV2K-Val and analyze when Alpha Mask fusion improves color consistency and when unreliable reference colors limit its effect.

2. Proposed Method

2.1 Overview

CCDiT separates detail restoration from color anchoring in two stages, as summarized in Fig. 1. Given a high-resolution image I_{HR}, we generate a low-resolution input I_{LR} through an online degradation process $\mathcal{D}(\cdot)$, and obtain a reference image I_{ref} with bicubic interpolation:

$$I_{LR} = \mathcal{D}(I_{HR}), \qquad I_{ref} = \mathrm{Bicubic}(I_{LR}). \tag{1}$$

In the first stage, Qwen-Image-Edit is adapted to map the bicubic reference image to the Qwen SR output I_q with one denoising step. In the second stage, this adapted SR backbone is frozen. Alpha Mask Net then predicts a spatial fu-

[1] The University of Electro-Communications
[a)] ma-y@mm.inf.uec.ac.jp
[b)] sun-l@mm.inf.uec.ac.jp
[c)] xiong-p@mm.inf.uec.ac.jp
[d)] yanai@cs.uec.ac.jp

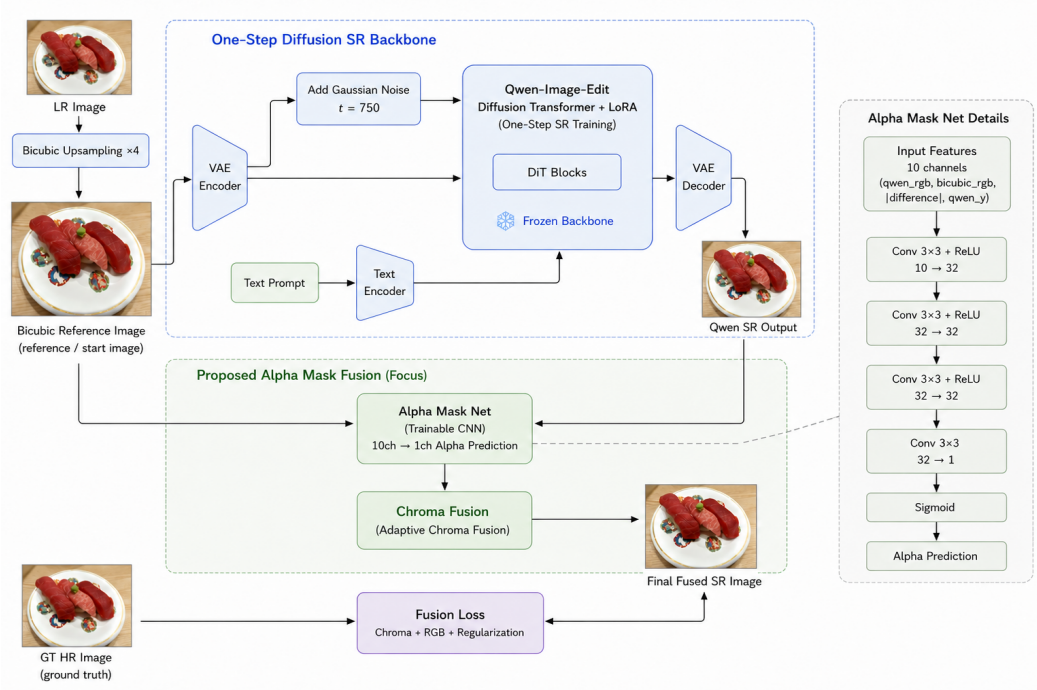


Fig. 1 Overview of CCDiT. Qwen-Image-Edit is first adapted for one-step SR from a bicubic reference image. The adapted Qwen SR backbone is then frozen, and Alpha Mask Net learns to fuse chrominance channels from the Qwen SR output and the bicubic reference image while preserving the restored luminance channel.

sion map that mixes the CbCr channels of I_q and I_{ref} while retaining the luminance channel of I_q. The final image is denoted as I_f. This single overview figure therefore covers both the trained SR backbone and the later frozen-backbone color-fusion stage.

2.2 One-Step Diffusion Transformer SR

The first-stage backbone restores details without iterative sampling. Instead of starting from pure Gaussian noise, CCDiT starts from the bicubic reference image. This keeps the low-frequency layout and approximate color distribution close to the input while asking the Diffusion Transformer to recover missing high-frequency content. Once this one-step SR mapping is learned, the backbone is kept fixed for the color-fusion stage, so subsequent training changes only the lightweight Alpha Mask Net.

We map the target image and the reference image into the latent space:

$$z_{gt} = E(I_{HR}), \qquad z_{ref} = E(I_{ref}), \tag{2}$$

where $E(\cdot)$ denotes the VAE encoder. Gaussian noise is added to the reference latent at a fixed one-step noise level:

$$z_t = \mathrm{AddNoise}(z_{ref}, \epsilon, t), \qquad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

Following ODTSR [5], we use $t = 750$, selected by `one_step_t_index`=250 in a 1000-step scheduler. The Qwen-Image-Edit Diffusion Transformer takes z_t, z_{ref}, the text condition c, and timestep t as inputs and predicts the noise term:

$$\epsilon_\theta = \epsilon_\theta(z_t, z_{ref}, c, t). \tag{4}$$

The predicted latent is obtained by a one-step update $z_q = z_t - \sigma_t \epsilon_\theta$ and decoded as $I_q = D(z_q)$. To reduce training cost and preserve the pretrained generative prior, we apply LoRA to selected attention projection and MLP-related layers [9].

The first-stage objective combines a latent reconstruction loss and a perceptual loss:

$$\mathcal{L}_{stage1} = \lambda_z \|z_q - z_{gt}\|_2^2 + \lambda_p \mathrm{LPIPS}_{alex}(D(z_q), D(z_{gt})), \tag{5}$$

where $\lambda_z = \lambda_p = 1.0$.

2.3 Adaptive Chrominance Fusion

The Qwen SR output can contain sharp texture, but its chrominance may differ from the input scene. We therefore design Alpha Mask Net as an original lightweight fusion module, motivated by alpha compositing and image matting [11], [16], spatial attention/gating [26], and the luminance–chrominance separation in YCbCr color space [17]. Unlike natural image matting, our alpha map is not a foreground opacity; it is a spatial reliability gate that decides whether each region should trust the restored chrominance or the input-aligned reference chrominance. In this design, the Qwen SR output supplies structure and detail, while the bicubic reference provides a low-frequency color anchor. We perform fusion in the YCbCr color space, preserving the luminance channel of the Qwen SR output:

$$(Y_q, Cb_q, Cr_q) = \mathrm{YCbCr}(I_q), \quad (Y_r, Cb_r, Cr_r) = \mathrm{YCbCr}(I_{ref}), \tag{6}$$

$$Y_f = Y_q. \tag{7}$$

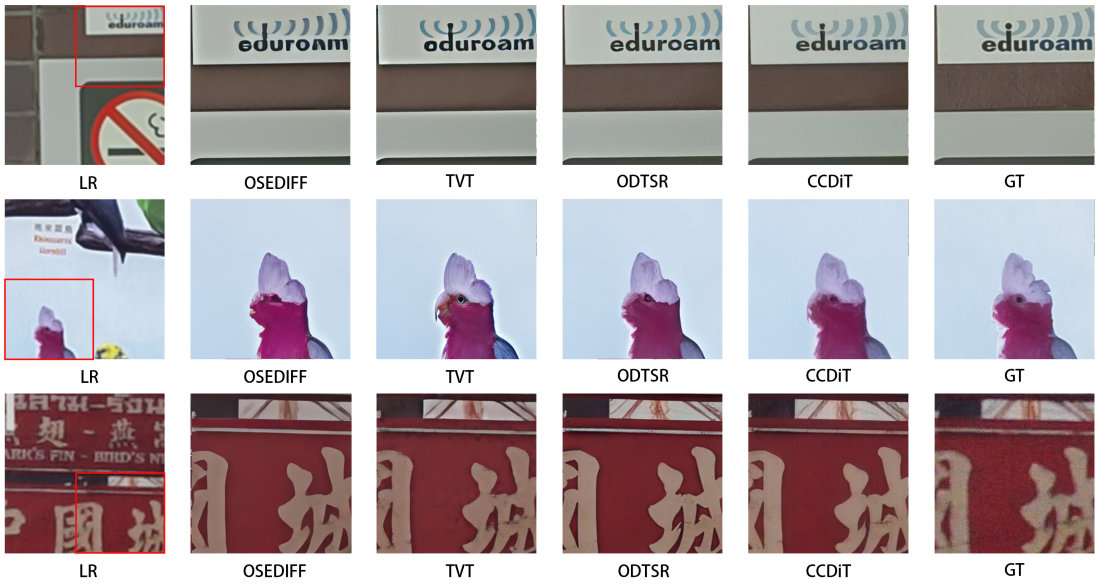


Fig. 2. Qualitative comparison with existing one-step SR methods on RealSR and DRealSR.

Table 1. Quantitative comparison on RealSR, DRealSR, and DIV2K-Val. CCDiT uses one denoising step. Red indicates the best result, and blue indicates the second-best result.

Dataset	Method	Step(s)	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑	FID↓	PSNR↑	SSIM↑
RealSR	OSEDiff	1	0.2921	0.2128	69.09	0.6326	123.49	25.15	0.7341
	PiSA-SR	1	0.2672	0.2044	70.15	0.6560	124.09	25.50	0.7417
	TSD-SR	1	0.2743	0.2104	71.19	0.6347	114.45	24.81	0.7172
	TVT	1	0.2597	0.2061	69.89	0.6232	109.90	25.81	0.7596
	ODTSR	1	0.2398	0.1894	68.29	0.6622	101.49	25.07	0.7361
	CCDiT	1	0.1949	0.1733	64.76	0.6286	82.13	26.31	0.7553
DRealSR	OSEDiff	1	0.2968	0.2165	64.65	0.5895	135.29	27.92	0.7835
	PiSA-SR	1	0.2960	0.2169	66.11	0.6146	130.48	28.32	0.7804
	TSD-SR	1	0.2966	0.2136	66.61	0.5850	135.32	27.77	0.7559
	TVT	1	0.2899	0.2205	65.56	0.5776	134.28	28.27	0.7899
	ODTSR	1	0.2592	0.1926	62.86	0.6227	119.86	28.14	0.7736
	CCDiT	1	0.2347	0.1962	60.79	0.5959	109.89	28.67	0.7766
DIV2K-Val	OSEDiff	1	0.2942	0.1975	67.96	0.6147	26.34	23.72	0.6109
	PiSA-SR	1	0.2823	0.1934	69.68	0.6401	25.09	23.87	0.6058
	TSD-SR	1	0.2673	0.1821	71.69	0.6226	23.29	23.02	0.5808
	TVT	1	0.2773	0.1860	68.67	0.6060	24.78	24.23	0.6292
	ODTSR	1	0.2761	0.1700	67.94	0.6438	21.47	23.92	0.6108
	CCDiT	1	0.2648	0.1957	63.78	0.6121	27.14	24.56	0.6178

Table 2. Color fusion ablation and cross-model generalization. The upper part reports results on CCDiT, and the lower part reports generalization results on OSEDiff outputs. Red indicates the best result within each group.

Backbone	Dataset	Method	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑	FID↓	PSNR↑	SSIM↑	Mean $\Delta E2000$ ↓
Color Fusion Ablation on CCDiT										
CCDiT	RealSR	CCDiT w/o Fusion	0.1926	0.1777	64.96	0.6323	79.80	24.79	0.7233	4.6296
		Fixed CbCr Fusion	0.2076	0.1758	63.86	0.6236	92.99	24.64	0.7138	4.7790
		CCDiT	0.1949	0.1733	64.76	0.6286	82.13	24.79	0.7195	4.5566
	DRealSR	CCDiT w/o Fusion	0.2271	0.1965	61.51	0.6059	105.28	26.69	0.7360	4.4442
		Fixed CbCr Fusion	0.2469	0.1996	59.65	0.5818	123.85	26.44	0.7267	4.6888
		CCDiT	0.2347	0.1962	60.79	0.5959	109.89	26.65	0.7318	4.4473
Cross-Model Generalization on OSEDiff										
OSEDiff	RealSR	OSEDiff	0.2921	0.2128	69.09	0.6335	123.50	23.59	0.6833	5.1556
		OSEDiff + Fixed CbCr	0.2954	0.2106	68.62	0.6182	140.46	23.54	0.6821	5.0414
		OSEDiff + Alpha Mask	0.2888	0.2104	69.25	0.6274	129.16	23.63	0.6851	4.9010
	DRealSR	OSEDiff	0.2968	0.2165	64.65	0.5899	135.29	25.85	0.7245	4.8380
		OSEDiff + Fixed CbCr	0.3047	0.2174	63.49	0.5613	148.70	25.78	0.7235	4.8097
		OSEDiff + Alpha Mask	0.2977	0.2169	64.38	0.5794	138.98	25.89	0.7260	4.6749

Fixed CbCr replacement is a natural but overly rigid baseline: it assumes that the bicubic reference is always color-correct and that the Qwen SR output is always color-unreliable. In real degradations, this assumption can fail because compression artifacts, sensor noise, or demosaicing errors may already change the low-resolution color. Alpha

Mask Net is designed to make this decision locally instead of globally. The cited alpha blending, matting, and attention works provide the conceptual basis for soft blending and spatial gating, while the 10-channel mask predictor and the chrominance-specific training objective are designed for CCDiT. It predicts a spatially adaptive fusion weight map:

$$\alpha = M_\phi\left(\mathrm{Concat}(I_q, I_{ref}, |I_q - I_{ref}|, Y_q)\right), \qquad \alpha \in [0,1]^{H\times W}. \tag{8}$$

When α is close to 1, the model keeps the Qwen chrominance; when it is close to 0, it uses the reference chrominance. The chrominance channels are fused as

$$Cb_f = \alpha \odot Cb_q + (1-\alpha) \odot Cb_r, \tag{9}$$

$$Cr_f = \alpha \odot Cr_q + (1-\alpha) \odot Cr_r. \tag{10}$$

The final image is obtained by converting (Y_f, Cb_f, Cr_f) back to RGB. Alpha Mask Net has four convolutional layers with channel sizes $10 \to 32 \to 32 \to 32 \to 1$ and a final sigmoid activation. The 10-channel input contains I_q, I_{ref}, their absolute difference, and Y_q, so the network can detect local color disagreement while preserving restored structure. We keep the architecture fully convolutional and shallow because the target is not to synthesize new texture, but to estimate local chrominance reliability. Freezing the Qwen SR backbone further decouples color correction from detail restoration, making Alpha Mask Net a lightweight post-restoration color-consistency module.

Only the parameters ϕ of Alpha Mask Net are optimized in the second stage. Its loss is

$$\begin{aligned}\mathcal{L}_{mask} = {} & \lambda_c\mathcal{L}_c + \lambda_r\mathcal{L}_{rgb} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_m\mathcal{L}_{mean} \\ & + \lambda_v\mathcal{L}_{var} + \lambda_t\mathcal{L}_{teacher} + \lambda_o\mathcal{L}_{oracle}.\end{aligned} \tag{11}$$

Here, $\mathcal{L}_c$ and $\mathcal{L}_{rgb}$ make the fused result close to the ground truth in chrominance and RGB space, while $\mathcal{L}_{tv}$ avoids noisy pixel-level mask changes. The mean and variance constraints prevent degenerate masks that always choose only one source. We further use a luminance-gradient teacher mask as a region-level prior: flat regions tend to use more reference chrominance, while edge and texture regions may trust the Qwen chrominance more. An oracle alpha target computed from ground-truth chrominance is used only during training to guide the best local mixing coefficient. At inference time, CCDiT requires only one denoising step and one lightweight CNN forward pass.

3. Experiments

Settings. We train CCDiT on LSDIR [13] with an OSEDiff/Real-ESRGAN-style 4× high-order degradation pipeline [24], [27], including blur, resizing, noise, and JPEG compression. BLIP captions [12] are appended with a restoration suffix. The first stage fine-tunes Qwen-Image-Edit with LoRA on 6 GPUs using learning rate 5×10^{-5} and rank 16; the second stage trains Alpha Mask Net on 2 GPUs using learning rate 2×10^{-4} and 32 hidden channels. We evaluate on RealSR, DRealSR, and DIV2K-Val, and report PSNR/SSIM [8], [25], LPIPS/DISTS [3], [29], MUSIQ/MANIQA [10], [28], FID [6], and Mean $\Delta E2000$ [20].

Qualitative comparison. Fig. 2 compares CCDiT with one-step baselines on examples containing text-like structures and saturated colors. OSEDiff, TVT, and ODTSR recover local details but can introduce chrominance changes around signs and high-contrast boundaries. CCDiT keeps the restored detail from the Qwen SR backbone while reducing avoidable color drift through Alpha Mask fusion, which is most visible in regions where text or object colors should remain stable.

Main comparison. Table 1 compares CCDiT with recent one-step SR methods. On RealSR, CCDiT reduces LPIPS to 0.1949 and FID to 82.13, and reaches the highest PSNR of 26.31. On DRealSR, it obtains the best LPIPS, FID, and PSNR. On DIV2K-Val, CCDiT obtains the best LPIPS and PSNR, while ODTSR remains better on some perceptual/no-reference metrics. Thus, CCDiT improves reconstruction and distribution quality without dominating every metric.

Color fusion study. Table 2 tests whether color anchoring should be fixed or adaptive. Fixed CbCr replacement worsens several metrics, showing that the bicubic reference is not always a reliable color source. Alpha Mask fusion gives a safer alternative: on RealSR, it reduces DISTS from 0.1777 to 0.1733 and Mean $\Delta E2000$ from 4.6296 to 4.5566. It also improves OSEDiff generalization results, reducing Mean $\Delta E2000$ from 5.1556 to 4.9010 on RealSR and from 4.8380 to 4.6749 on DRealSR. The remaining DRealSR limitation suggests that future color-aware SR should estimate reference reliability jointly with restoration.

4. Conclusion

We presented CCDiT, a one-step Diffusion Transformer framework for real-world image super-resolution that addresses color drift, an underexplored failure mode of fast diffusion SR. CCDiT adapts Qwen-Image-Edit through LoRA fine-tuning and restores images with one denoising step from a bicubic reference image. It further uses Alpha Mask fusion to preserve the luminance of the Qwen SR output while adaptively mixing chrominance in the YCbCr space. Experiments show that CCDiT improves several reconstruction and perceptual metrics with one-step inference. The color-fusion experiments show a conditional result: Alpha Mask fusion is better than fixed CbCr replacement and improves color difference on RealSR, but its benefit is limited when the bicubic reference image already contains unstable colors. Overall, CCDiT suggests that one-step diffusion SR should treat color consistency as an explicit design objective, rather than as a side effect of perceptual restoration.

References

[1] Chen, X. et al.: Activating More Pixels in Image Super-Resolution Transformer, *CVPR* (2023).
[2] Chiu, M.-C. et al.: ColorSense: A Study on Color

Vision in Machine Visual Recognition, arXiv preprint arXiv:2212.08650 (2022).
[3] Ding, K. et al.: Image Quality Assessment: Unifying Structure and Texture Similarity, *IEEE TPAMI* (2020).
[4] Dong, L., Fan, Q., Guo, Y., Wang, Z., Zhang, Q., Chen, J., Luo, Y. and Zou, C.: TSD-SR: One-Step Diffusion with Target Score Distillation for Real-World Image Super-Resolution, *CVPR* (2025).
[5] Fang, Y. et al.: One-Step Diffusion Transformer for Controllable Real-World Image Super-Resolution, arXiv preprint arXiv:2511.17138 (2025).
[6] Heusel, M. et al.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, *NeurIPS* (2017).
[7] Ho, J. et al.: Denoising Diffusion Probabilistic Models, *NeurIPS* (2020).
[8] Hore, A. and Ziou, D.: Image Quality Metrics: PSNR vs. SSIM, *ICPR* (2010).
[9] Hu, E. et al.: LoRA: Low-Rank Adaptation of Large Language Models, *ICLR* (2022).
[10] Ke, J. et al.: MUSIQ: Multi-scale Image Quality Transformer, *ICCV* (2021).
[11] Levin, A., Lischinski, D. and Weiss, Y.: A Closed-Form Solution to Natural Image Matting, *IEEE TPAMI*, Vol. 30, No. 2, pp. 228–242 (2008).
[12] Li, J. et al.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, *ICML* (2022).
[13] Li, J. et al.: LSDIR: A Large Scale Dataset for Real-World Image Super-Resolution, arXiv preprint arXiv:2308.11578 (2023).
[14] Liang, J. et al.: SwinIR: Image Restoration Using Swin Transformer, *ICCV Workshops* (2021).
[15] Lin, X. et al.: DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior, arXiv preprint arXiv:2308.15070 (2023).
[16] Porter, T. and Duff, T.: Compositing Digital Images, *Computer Graphics*, Vol. 18, No. 3, pp. 253–259 (1984).
[17] Poynton, C.: *Digital Video and HDTV Algorithms and Interfaces*, Morgan Kaufmann (2003).
[18] Rombach, R. et al.: High-Resolution Image Synthesis with Latent Diffusion Models, *CVPR* (2022).
[19] Saharia, C. et al.: Image Super-Resolution via Iterative Refinement, *IEEE TPAMI* (2023).
[20] Sharma, G. et al.: The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations, *Color Research and Application* (2005).
[21] Singh, A. et al.: Assessing The Importance Of Colours For CNNs In Object Recognition, arXiv preprint arXiv:2012.06917 (2020).
[22] Song, J. et al.: Denoising Diffusion Implicit Models, *ICLR* (2021).
[23] Wang, J. et al.: Exploiting Diffusion Prior for Real-World Image Super-Resolution, arXiv preprint arXiv:2305.07015 (2023).
[24] Wang, X. et al.: Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data, *ICCV Workshops* (2021).
[25] Wang, Z. et al.: Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE TIP* (2004).
[26] Woo, S., Park, J., Lee, J.-Y. and Kweon, I. S.: CBAM: Convolutional Block Attention Module, *ECCV* (2018).
[27] Wu, R. et al.: One-Step Effective Diffusion Network for Real-World Image Super-Resolution, arXiv preprint arXiv:2406.08177 (2024).
[28] Yang, S. et al.: MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment, *CVPR* (2022).
[29] Zhang, R. et al.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, *CVPR* (2018).