CoSeR: Bridging Image and Language for Cognitive Super-Resolution

Haoze Sun[1] Wenbo Li[2*] Jianzhuang Liu[2] Haoyu Chen[3]
Renjing Pei[2] Xueyi Zou[2] Youliang Yan[2] Yujiu Yang[1*]
[1]Tsinghua University [2]Huawei Noah's Ark Lab [3]HKUST (GZ)
shz22@mails.tsinghua.edu.cn liwenbo50@huawei.com yang.yujiu@sz.tsinghua.edu.cn

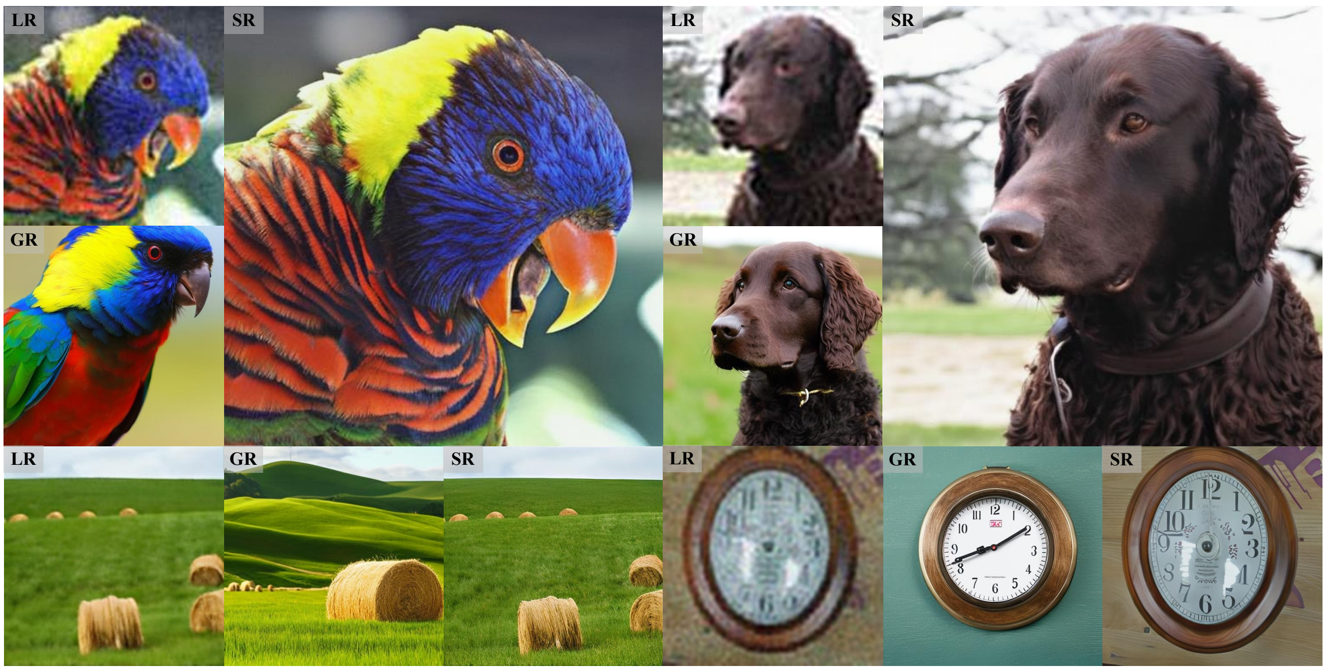

Figure 1. Visualization displaying 4× super-resolution results generated by our Cognitive Super-Resolution (CoSeR) model. CoSeR adeptly extracts cognitive information from a low-resolution (LR) image and utilizes it to generate a high-quality reference image. This reference image, aligning closely with the LR image in terms of semantics and textures, significantly benefits the super-resolution process. For conciseness, we denote the input, generated reference, and restoration result as LR, GR, and SR, respectively. Best viewed zoomed in.

Abstract

Existing super-resolution (SR) models primarily focus on restoring local texture details, often neglecting the global semantic information within the scene. This oversight can lead to the omission of crucial semantic details or the introduction of inaccurate textures during the recovery process. In our work, we introduce the Cognitive Super-Resolution (CoSeR) framework, empowering SR models with the capacity to comprehend low-resolution images. We achieve this by marrying image appearance and language understanding to generate a cognitive embedding, which not only activates prior information from large text-to-image diffusion models but also facilitates the generation of high-quality reference images to optimize the SR process. To further improve image fidelity, we propose a novel condition injection scheme called "All-in-Attention", consolidating all conditional information into a single module. Consequently, our method successfully restores semantically correct and photorealistic details, demonstrating state-of-the-art performance across multiple benchmarks. Project page: https://coser-main.github.io/

*Corresponding author

1. Introduction

Real-world image super-resolution (SR) is a fundamental task in the realm of image processing, aimed at enhancing low-resolution (LR) images to yield the high-resolution (HR) counterparts [36]. Its versatile applicability spans critical domains, including mobile phone photography [5], autonomous driving [31], and robotics [55], while also influencing various computer vision tasks, notably object detection [16], segmentation [54] and recognition [8, 15].

Despite significant advancements in this field in recent years, the processing of complex real-world scenarios continues to pose enduring challenges. Utilizing image priors is a common strategy for tackling real-world SR problems. These priors may be either introduced explicitly in the form of reference images [3, 22, 37, 63, 65, 66, 77, 79, 80], or implicitly leveraged through pre-trained generative models [4, 13, 14, 23, 35, 41, 44, 53, 58, 60, 68, 69, 72]. Especially, the recently emerged text-to-image diffusion models [45, 47–49] exhibit a remarkable capability to generate high-quality images based on user-provided prompts. These models not only possess strong image priors but also allow precise responses to human instructions in the form of language. This opens up the potential to bridge low-level image processing and high-level abstract cognition.

Consider the process by which human experts restore low-quality images [43, 56]: They start by establishing a comprehensive understanding of the image, encompassing scene identification and primary subject recognition. Subsequently, their focus shifts to a meticulous examination and restoration of finer image details. In contrast, conventional image super-resolution techniques [27, 32, 35, 53, 57, 76], adhere to a bottom-up approach, primarily concentrating on local content and direct pixel-level processing. Consequently, these methodologies exhibit inherent limitations in grasping the holistic image context, often failing to restore severely degraded yet semantically vital details. Moreover, given the ill-posed nature of LR images, there is a possibility for introducing semantically erroneous textures [56].

To surmount these challenges, there arises a compelling rationale for imbuing the SR model with "cognitive" capabilities. In this pursuit, we introduce a pioneering SR methodology known as Cognitive Super-Resolution (CoSeR). Our approach aligns with the top-down cognitive process employed by humans in image perception. It commences with the generation of cognitive embeddings, a representation that encapsulates the overarching comprehension of the LR image, containing both scene semantics and image appearance. This cognitive embedding allows us to precisely leverage the implicit prior knowledge embedded in pre-trained text-to-image generation models, resulting in an enhanced capacity to restore image details in a manner akin to human expertise. Previous work [56] uses segmentation maps to offer semantics. However, acquiring ideal segmentation maps for real-world LR images remains difficult, even with advanced models like [26]. Moreover, semantic segmentation is constrained by predefined categories, limiting its applicability to open-world scenes.

Apart from implicitly leveraging diffusion priors, we also advocate for the explicit utilization of image priors. We introduce a novel approach where we employ cognitive embeddings derived from LR inputs to generate reference images through diffusion models. These reference images are subsequently utilized to guide the restoration process. As shown in Figure 1, our cognitive embedding contains language understanding while preserving the color and texture information of the image, thus producing high-quality reference images that are not only semantically aligned but also similar in appearance. This explicit approach brings substantial improvements in capturing high-definition textures compared to relying solely on implicit diffusion knowledge.

We have established both implicit and explicit cognitive priors for LR inputs. Then incorporating these priors effectively into our model is pivotal. Unlike the typical conditional generation methods [20, 42, 74, 78], super-resolution demands a heightened level of fidelity between outputs and low-quality inputs. In order to concurrently ensure texture realism and fidelity, we introduce an "All-in-Attention" design, which integrates multiple information sources via an attention mechanism, including cognitive embeddings, reference images, and LR inputs. This approach allows our model to flexibly use different conditional components, yielding improved results. Our experiments show that our model excels in preserving fidelity compared to previous methods while generating more intricate textures.

The contributions of this paper can be summarized as:

- We introduce CoSeR, a novel framework for high-detail image super-resolution. CoSeR autonomously extracts cognitive embeddings from LR images, harnessing implicit diffusion priors to enhance the LR input.
- We incorporate diffusion priors explicitly by creating semantically coherent reference images, which act as guidance to improve the quality of the restored image.
- To enhance image fidelity, we introduce a novel "All-in-Attention" architecture to integrate conditional information into the SR model. Our method achieves state-of-the-art performance across multiple benchmarks.

2. Related work

2.1. Real-World Image Super-Resolution

Real-world image SR has primarily revolved around two avenues: data utilization and image prior incorporation.

The first category involves the creation of diverse and realistic pairwise data by adapting the physical collection means [2, 6, 61] or improving the generation pipeline [1, 59, 70, 73]. Also, several works [40, 62, 71] combine both paired and unpaired data with weak supervision to enhance performance in real-world scenarios.

The second line focuses on the use of image priors.

While the "learning-from-scratch" approaches [50, 59, 73] demand substantial data and computational resources, using pre-trained generative models with rich texture priors has become a practical and economical practice. Several studies [4, 14, 30, 41, 44, 58, 68] have leveraged pre-trained Generative Adversarial Networks (GANs) to improve the super-resolution process. Nonetheless, these methods occasionally suffer from the generation of unrealistic textures, owing to the inherent limitations of GANs [34, 64]. Consequently, there is a growing interest in utilizing more advanced pre-trained generative models, such as the denoising diffusion models [19, 52], in recent research.

2.2. Diffusion-Based Super-Resolution

Recent approaches [23, 60, 72] utilize implicit knowledge from pre-trained diffusion models [10], yet they typically focus on non-blind degradation [60] or specific domains like facial images [72]. In an alternative strategy proposed by Fei *et al*. [13], the simultaneous estimation of the degradation model is applied to address blind degradation. However, this method relies on test-time optimization and primarily explores SR under linear degradation, thereby exhibiting limitations in handling real-world complexities.

Other approaches [35, 53, 69] leverage recent advancements in large-scale text-to-image diffusion models [45, 47–49]. These models, trained on extensive datasets of high-definition images, provide enhanced capabilities for processing diverse content. StableSR [53] stands as a pioneering work, which harnesses prior information from diffusion model. DiffBIR [35] combines a traditional pixel regression-based image recovery model with the text-to-image diffusion model, mitigating the adverse effects of LR degradation on the generation process. Despite notable advancements in visual quality, these methods have yet to fully harness the potential of text-to-image generation models, mainly due to the limited image content comprehension.

2.3. Reference-Based Super-Resolution

The reference image serves as an explicit prior, ideally containing content relevant to the LR image to facilitate the generation of high-definition details. Recent advancements in reference-based SR can be categorized into two branches [37]. One branch prioritizes spatial alignment, employing techniques like CrossNet [79] and SSEN [3]. However, these methods often encounter challenges in establishing long-distance correspondences. The other branch, represented by SRNTT [77], TTSR [66], MASA-SR [37], and CFE-PatchMatch [63], utilizes patch-matching mechanisms to facilitate the establishment of long-range connections between the reference map and the LR image. Yet, manually specifying reference images in real scenarios is labor-intensive, motivating the development of an automated and high-quality reference generation approach. DWTrans [28] employs a Stable Diffusion model [48], fine-tuned on the SR datasets, to function as a self-referencing image generator. However, it does not fully leverage the pre-trained prior knowledge to generate reference images of sufficiently high definition.

3. Methodology

Our Cognitive Super-Resolution (CoSeR) model employs a dual-stage process for restoring LR images. Initially, we develop a cognitive encoder to conduct a thorough analysis of the image content, conveying the cognitive embedding to the diffusion model. This triggers the activation of image priors within the pre-trained Stable Diffusion model [48], facilitating the restoration of intricate details. Additionally, our approach utilizes cognitive understanding to generate high-fidelity reference images that closely align with the input semantics. These reference images serve as auxiliary information, contributing to the enhancement of super-resolution results. Ultimately, our model simultaneously applies three conditional controls to the pre-trained Stable Diffusion model: the LR image, cognitive embedding, and reference image, as detailed in Figure 2.

3.1. Cognitive Encoder

To distill cognitive information from LR images, our model commences with LR preprocessing aimed at mitigating the impact of degradation. We use a lightweight SRResnet [27] for $4\times$ super-resolution, without additional supervision. We then utilize a pre-trained CLIP [46] image encoder to extract features from the enhanced image. It is crucial to underscore that, although CLIP adeptly aligns image and language content, a significant disparity persists between the image embedding and the language embedding. These two components focus on different points, where image features inherently capture spatially variant details, while language features encapsulate comprehensive information. Thus, a single language token may correspond to multiple subjects dispersed across diverse regions of an image.

To overcome the challenge of aligning image and language representations, prior methods [38, 39] have often focused on aligning the class token of the image embedding and the class token of the corresponding language embedding, neglecting other tokens. However, relying solely on this single class token has been observed to introduce cognitive bias. As shown by the generated reference images from language tokens in the first row of Figure 3 (left part), cognitive bias diminishes gradually as the token number (before and including the class token) increases. To simultaneously address information misalignment and inaccurate cognition, we introduce a cognitive adapter that is tailored to extract multi-token cognitive embedding from image features, shown in Figure 3 (right part). Drawing inspiration from the Q-Former structure [29], originally de-

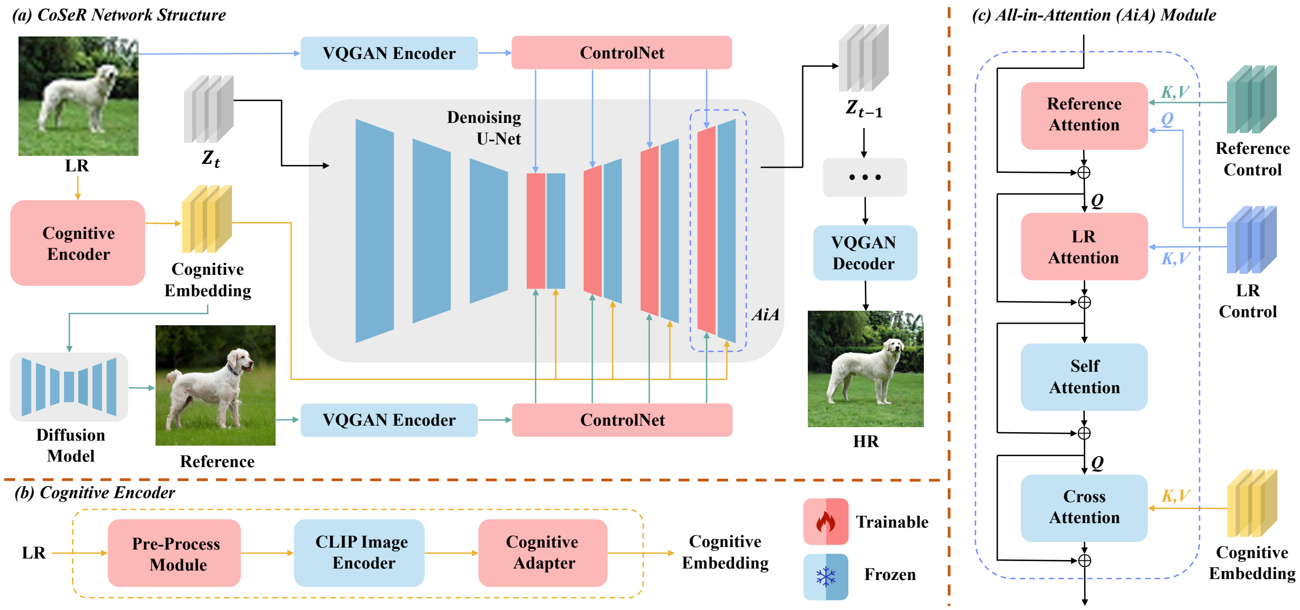

Figure 2. Framework of the proposed cognitive super-resolution (CoSeR) network. Given a low-resolution (LR) image, we employ a cognitive encoder to extract cognitive embedding containing semantic and textural information, which is then used to generate a high-quality reference image. The LR input, cognitive embedding, and reference image are integrated into the denoising U-Net using the all-in-attention (AiA) module, represented by blue, gold, and cyan lines, respectively. The structures of the cognitive encoder and AiA module are detailed in (b) and (c). Trainable modules are highlighted in red, while frozen modules are indicated in blue.

vised for vision-language representation learning, our approach employs learnable queries to interact with spatially-arranged image information, thereby reshaping information organization and facilitating feature compression. Our approach also incorporates a novel form of supervision, enhancing the adapter's capacity not only to reorganize image features but also to function as a modality transformer.

We represent the CLIP image embedding extracted from LR as $\boldsymbol{I} \in \mathbb{R}^{B\times T_i\times C_i}$, where B, T_i, C_i denote batch size, token number, and channel number, respectively. Additionally, $\boldsymbol{L} \in \mathbb{R}^{B\times T_l\times C_l}$ denotes the CLIP language embedding extracted from the ground-truth caption (extracted from HR images using BLIP2 [29]). In our cognitive adapter, we employ T_e learnable queries ($T_e \leq T_l$) such that the resulting cognitive embedding is denoted as $\boldsymbol{E} \in \mathbb{R}^{B\times T_e\times C_l}$. We propose to use T_e tokens preceding the class token $\boldsymbol{L}[t_{cls}]$ (inclusive) for supervision, as these tokens retain all previous information [46]. If there are insufficient supervision tokens, we use the class token for end-filling. Therefore, our supervision $\boldsymbol{L'}$ can be regarded as a more comprehensive representation than the class token. The loss function for training the cognitive encoder is expressed as:

$$\mathcal{L}_{CE} = \|\boldsymbol{E} - \boldsymbol{L'}\|_2^2, \tag{1}$$

where

$$\boldsymbol{L'} = \begin{cases} \text{Padding}\left(\boldsymbol{L}\left[: t_{cls}\right], \boldsymbol{L}\left[t_{cls}\right]\right), & \text{if } t_{cls} < T_e; \\ \boldsymbol{L}\left[(t_{cls} - T_e) : t_{cls}\right], & \text{if } t_{cls} \geqslant T_e. \end{cases} \tag{2}$$

A more extensive explanation of the supervision strategy can be found in the supplementary materials.

Discussion. We opt for feature embedding in the cognition process over direct LR image captioning for several reasons. Firstly, although guided by language embedding, our cognitive embedding retains fine-grained image features, proving favorable in generating reference images with high semantic similarity. In the second row of Figure 3 (left part), we show the BLIP2 captions generated from LR images. They fail to identify the precise taxon, color, and texture of the animals, leading to suboptimal generations compared to our cognitive adapter. Secondly, employing a pre-trained image caption model requires a large number of parameters, potentially reaching 7B [29]. In contrast, our cognitive adapter is notably lighter, with only 3% parameters, resulting in favorable efficiency. Thirdly, pre-trained image caption models may yield inaccurate captions for LR images due to disparities in the input distribution. In contrast, our cognitive adapter is more robust for LR images, shown in the third row of Figure 3 (left part).

3.2. Reference Image Generation and Encoding

We propagate the cognitive embedding to the pre-trained Stable Diffusion model for generating reference images without incurring additional parameters. The resulting reference images empower our SR model to explicitly leverage image priors. As depicted in Figure 1, our cognitive embedding excels in producing well-aligned reference images.

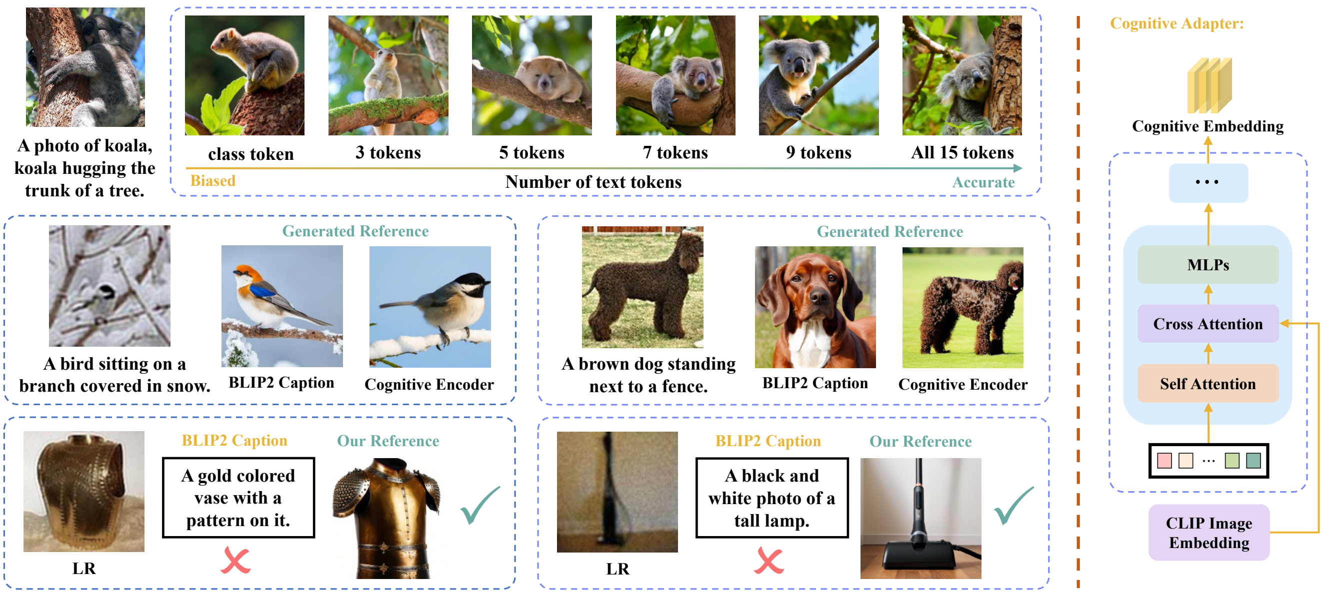

Figure 3. (**Left**) Generated reference images by BLIP2 captions and cognitive encoder. The first row shows the need to augment the token number. The last two rows show the drawbacks of directly employing captions for cognition. (**Right**) Structure of our cognitive adapter.

We employ a pre-trained VQGAN [12] for encoding images into latent codes, as opposed to a trainable CNN like [74], given the robust encoding capabilities exhibited by VQGAN. Subsequently, we follow the ControlNet [74] approach by utilizing the U-Net encoder to generate multi-scale control features. We represent the LR control and reference image control as $\{\boldsymbol{X}_i\}_{i=1}^4$ and $\{\boldsymbol{R}_i\}_{i=1}^4$, respectively. Notably, we observe that using a single control encoder for both LR and reference images is sufficient for achieving satisfactory results, enhancing the parameter efficiency of our model. The generated controls are then input into the All-in-Attention module, as elaborated in the following section. In fact, when automatically generating reference images, we only need to use Stable Diffusion to generate latent codes, and subsequently input them into the control encoder, thereby circumventing the process of decoding and encoding in Figure 2.

3.3. All-in-Attention Module

In image super-resolution, preserving fidelity to LR inputs is important. Our experiments in Section 4.4 demonstrate that the introduction of LR control $\{\boldsymbol{X}_i\}_{i=1}^4$ through attention mechanisms leads to enhanced fidelity. Consequently, we advocate for the comprehensive integration of all conditional information into our model, achieved through the design of an All-in-Attention (AiA) module. Beyond accommodating LR inputs, this design facilitates reference patch-matching for the establishment of long-range connections [66]. The cognitive embedding is seamlessly incorporated via the cross-attention mechanism of Stable Diffusion.

Illustrated in Figure 2 (c), the AiA module enhances the original attention module in Stable Diffusion by introducing trainable reference attention and LR attention, while maintaining the frozen state of the self-attention and cross-attention components. This structural augmentation is applied across all attention modules within the middle and decoder of the denoising U-Net. We denote the query, key, and value features in the attention mechanism as $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. Regarding LR attention, $\boldsymbol{Q}$ is derived from the denoising U-Net feature $\boldsymbol{Z}$, while $\boldsymbol{K}$ and $\boldsymbol{V}$ originate from the LR control $\boldsymbol{X}_i$. In reference attention, we opt to use the LR control as $\boldsymbol{Q}$ for better fidelity, with $\boldsymbol{K}$ and $\boldsymbol{V}$ coming from the reference control $\boldsymbol{R}_i$. In the original cross-attention, we use cognitive embedding $\boldsymbol{E}$ as inputs for $\boldsymbol{K}$ and $\boldsymbol{V}$. Notably, to counteract the potential blurring effect of the conventional attention mechanism in reference-based SR [66], we introduce "one-hot attention" to enhance the LR image with the most relevant reference feature, and additional details are available in the supplementary materials.

4. Experiments

4.1. Implementation Details

Our CoSeR is built based on Stable Diffusion 2.1-base[1]. The model is trained with a batch size of 192 over 20000 steps on 8 V100 GPUs. We use Adam [25] optimizer with a learning rate of 5×10^{-5}. Following StableSR [53], we train our model on 512×512 resolution and apply DDPM sampling [19] with 200 timesteps for inference.

The training process involves two stages: Firstly, we train the cognitive encoder using the defined loss func-

[1]https://huggingface.co/stabilityai/stable-diffusion-2-1-base

Datasets	Metrics	RealSR	Real-ESRGAN+	BSRGAN	DASR	FeMaSR	LDM	StableSR	CoSeR (Ours)
ImageNet Test2000	FID↓	86.36	32.68	41.11	39.15	31.25	34.54	<u>22.53</u>	**19.41**
	DISTS↓	0.2649	0.1739	0.1946	0.1931	0.1597	0.1664	<u>0.1527</u>	**0.1482**
	LPIPS↓	0.4519	0.2943	0.3381	0.3346	0.3027	0.3289	<u>0.2871</u>	**0.2863**
	CLIP-Score↑	0.6242	0.8132	0.7719	0.7838	0.8253	0.8119	<u>0.8622</u>	**0.8755**
	MANIQA↑	0.0796	0.1370	0.1115	0.0914	<u>0.1936</u>	0.1830	0.1556	**0.2133**
	MUSIQ↑	50.18	57.52	52.33	48.98	**67.20**	64.15	60.20	<u>65.51</u>
RealSR [2]	FID↓	157.85	87.00	111.03	107.38	91.45	92.43	<u>84.06</u>	**80.82**
	DISTS↓	0.2529	0.2028	0.2545	0.2171	0.2131	0.2055	<u>0.1867</u>	**0.1826**
	LPIPS↓	0.3672	0.2803	0.3224	0.3056	0.2683	0.2924	<u>0.2536</u>	**0.2438**
	CLIP-Score↑	0.7458	0.8345	0.8074	0.8332	0.8108	0.8330	<u>0.8517</u>	**0.8545**
	MANIQA↑	0.1474	0.1776	0.1696	0.1803	0.2033	0.1986	<u>0.2144</u>	**0.2522**
	MUSIQ↑	60.40	61.90	60.82	60.90	66.47	<u>67.27</u>	67.08	**70.29**
DRealSR [61]	FID↓	148.58	<u>74.72</u>	107.76	96.42	86.81	87.16	75.83	**71.22**
	DISTS↓	0.2673	0.2216	0.2238	0.2345	0.2231	0.2179	<u>0.2048</u>	**0.1977**
	LPIPS↓	0.4212	0.3239	0.3972	0.3534	0.2981	0.3258	<u>0.2920</u>	**0.2702**
	CLIP-Score↑	0.7360	0.8504	0.8157	0.8510	0.8332	0.8459	<u>0.8681</u>	**0.8766**
	MANIQA↑	0.1090	0.1742	0.1491	0.1739	0.1998	0.1890	<u>0.2241</u>	**0.2575**
	MUSIQ↑	54.28	62.80	57.72	62.14	66.57	67.03	<u>68.27</u>	**70.18**

Table 1. Quantitative comparisons on both ImageNet Test2000 and real-world benchmarks (RealSR and DRealSR). The best results are highlighted in **bold** and the second best results are in <u>underlined</u>.

tion in Eq. 1. The cognitive encoder employs 50 learnable queries, a choice substantiated in the supplementary materials. Then, we freeze the cognitive encoder and train the SR model. Following [74], we initialize ControlNet with Stable Diffusion weights. To maximize the utilization of the pre-trained model, the reference attention and LR attention modules are initialized using self-attention weights. In the inference phase, cognitive information is enhanced via classifier-free guidance [18], utilizing a scaling factor of 3. To optimize the trade-off between realism and fidelity, we adopt the pre-trained feature wrapping module in [53], which is integrated with the VQGAN decoder.

4.2. Experimental Settings

Training and testing datasets. We aim to develop an image super-resolution model empowered with cognitive capabilities adaptable to diverse real-world scenarios. To this end, we utilize the extensive ImageNet dataset [9] for training, renowned for its wide array of scenarios and objects. We acquire over 900K HR images with 512×512 resolution and employ Real-ESRGAN [59] degradation to generate corresponding LR images. We employ BLIP2 [29] to generate three descriptive captions for each HR image, filtering out captions with CLIP scores [46] below 0.28.

To comprehensively assess our model's performance across diverse scenarios, we curate a non-overlapped ImageNet test set consisting of 2000 LR-HR pairs via Real-ESRGAN. We choose two images from each category, ensuring the test set's diversity and balance. In addition to our constructed test set, we conduct evaluations on established real-world benchmarks such as RealSR [2] and DRealSR [61]. LR images are acquired at the same resolution used during training, specifically 128×128. For RealSR and DRealSR, we initially resize LR images, adjusting the shorter sides to 128, followed by center cropping.

Compared methods. We compare CoSeR with some state-of-the-art real-world SR methods, including RealSR [21], Real-ESRGAN+ [59], BSRGAN [73], DASR [33], FeMaSR [7], latent diffusion models (LDM) [48], StableSR [53]. To ensure fair comparisons, we retrain all these models using our ImageNet training set except RealSR and BSRGAN, which share the network structure with Real-ESRGAN+ but employ different degradation pipelines.

Evaluation metrics. To better align with human perception, we employ six perceptual metrics: FID [17], DISTS [11], LPIPS [75], CLIP-Score [46], MANIQA [67] and MUSIQ [24]. FID, DISTS, and LPIPS measure perceptual distance, while CLIP-Score estimates semantic accuracy by evaluating scores between HR images and SR results. Given our focus on real-world scenarios where ground-truth HR data might be unavailable, we include non-reference image quality assessments, MANIQA and MUSIQ. Notably, pixel-level image quality assessments like PSNR and SSIM are presented in the supplementary materials solely for reference. Prior research [11, 51, 67, 75] has indicated their weak correlation with human perception regarding image quality in real-world contexts.

4.3. Comparison with State of the Arts

Quantitative comparison. We perform an extensive quantitative comparison on both the ImageNet Test2000 dataset and real-world benchmarks (RealSR and DRealSR), as pre-

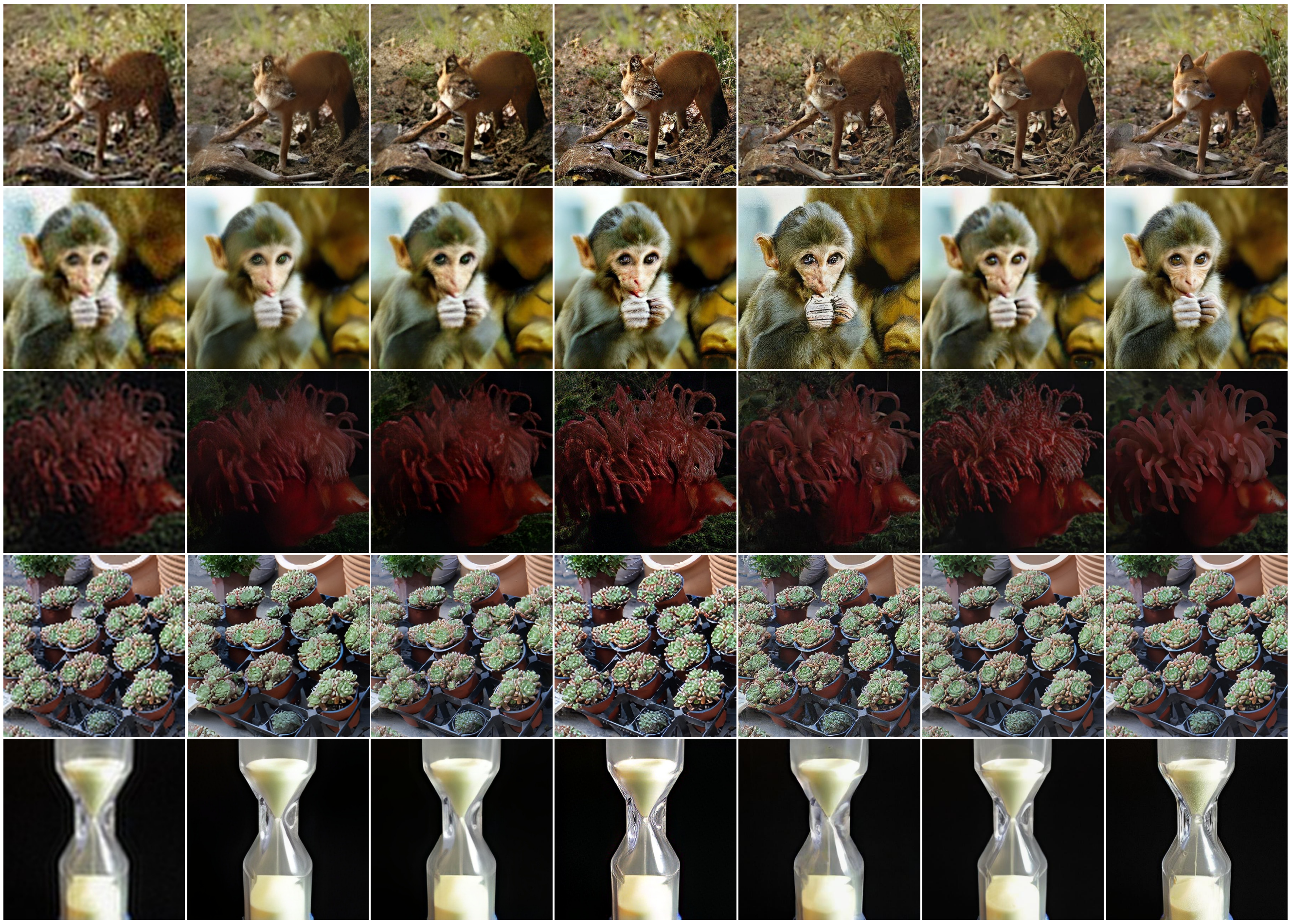

Figure 4. Qualitative comparisons on both synthesized and real-world test datasets. Our CoSeR obtains the best visual performance.

sented in Table 1. As mentioned previously, we retrain the comparison models using the ImageNet training dataset to ensure fair comparisons. Additionally, the results of officially released models are provided in the supplementary materials. Our method consistently demonstrates superior performance across nearly all datasets and metrics, highlighting its robustness and superiority. Notably, our FID scores surpass the second-best performance by $\mathbf{13.8}\%$, $\mathbf{3.8}\%$, and $\mathbf{4.7}\%$ on the ImageNet Test2000, RealSR, and DRealSR, respectively. While FeMaSR exhibits better performance in MUSIQ on the ImageNet Test2000, as depicted in Figure 4, it introduces numerous unrealistic artifacts that might not be reflected by the non-reference metric MUSIQ.

Qualitative comparison. We provide visual comparisons in Figure 4. Enriched by a comprehensive understanding of scene information, CoSeR excels in enhancing high-quality texture details. As demonstrated in the first and second rows, our results exhibit significantly clearer and more realistic fur and facial features in the animals. Similarly, in the third and fourth rows, our method adeptly reconstructs realistic textures such as the anemone tentacles and succulent leaves—achievements unmatched by other methods. Particularly, our model's cognitive capabilities enable the recovery of semantic details almost lost in low-resolution inputs. Notably, in the first row, only our model successfully restores the dhole's eyes, while in the fifth row, only our method can reconstruct the sand within the hourglass. These visual cases distinctly showcase our model's capacity to comprehend scenes and produce high-quality images.

User Study. To further substantiate the effectiveness of CoSeR in real-world scenarios, a user study is conducted on 20 real-world LRs collected from the Internet or captured by mobile phones. 23 subjects are asked to select the visually superior result from the four HRs generated by Real-ESRGAN+, FeMaSR, StableSR, and CoSeR. A total of 20×23 votes are collected, with approximately $\mathbf{80}\%$ of participants concurring that our method exhibited the best visual effect. This underscores the superiority and robustness of CoSeR in real-world scenarios. Detailed voting results are available in the supplementary materials.

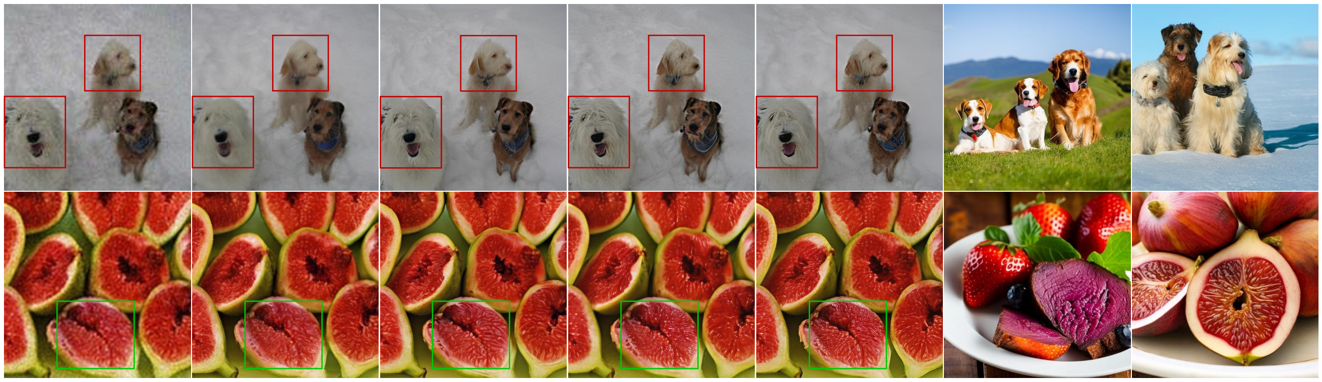

Figure 5. Visual comparisons were conducted to assess the impact of cognitive information, comparing scenarios with no utilization of cognitive information, employing different cognitive encoders, and incorporating the generated reference image.

4.4. Ablation Study

We dissect the individual contributions of different components in our framework. Given the diverse focuses of different components, we employ the most appropriate evaluation metrics to measure their respective utilities. We exclude reference generation from the evaluation of other modules.

Cognitive information. Prior research, such as [38], aligned class tokens between CLIP image and text encoders using MLP. In our method, we denote the cognitive encoder with class token alignment MLP and our cognitive encoder as the "class token encoder" and "multi-token encoder", respectively. Table 2 shows that incorporating cognitive information markedly improves FID and CLIP-Score, indicating better semantic and texture accuracy. Our studies reveal that the class token encoder may introduce semantic and texture biases, as evident in the quality of the generated reference images in Figure 5 (penultimate column). To quantitatively evaluate cognitive bias, we introduce the "Gen-score", calculated as the CLIP-Score between the generated reference image and the ground-truth image. Both the metrics and the visuals highlight our superior cognitive ability. This advantage extends to the final SR results, notably visible in the detailed hair and pulp texture in Figure 5.

Reference guidance. The explicitly introduced reference image significantly contributes to enhancing the texture details in the SR results (Figure 5). To better correlate with human perception, our evaluation primarily focuses on assessing restoration quality using FID and two non-reference image quality assessments. Table 2's fifth column reveals that adding a generated reference image elevates the overall visual quality of the SR results without sacrificing fidelity. Additionally, compared to real-world ImageNet references, our generated images achieve similar or superior results.

All-in-Attention (AiA) module. To mitigate fidelity loss from excessive generation, we introduced the All-in-Attention (AiA) module, which integrates various conditions to improve input consistency. For fidelity assessment, we utilize ground-truth-involved FID, DISTS, and LPIPS metrics. Compared to spatial feature transform (SFT) [56] integrated in StableSR [53], our AiA module achieves a 5.7% lower FID score, along with superior DISTS and LPIPS results. This manifests the effectiveness of our AiA module in enhancing result fidelity.

CoSeR	FID↓	CLIP-Score↑	Gen-score↑
w/ multi-token encoder	**20.27**	**0.8674**	**0.5953**
w/ class token encoder	21.35	0.8628	0.4881
w/o cognitive encoder	23.18	0.8484	–
CoSeR	FID↓	MUSIQ↑	MANIQA↑
w/ generated reference	19.80	**64.21**	**0.2107**
w/ ImageNet reference	**19.72**	63.49	0.2056
w/o reference	20.27	61.82	0.1874
CoSeR	FID↓	DISTS↓	LPIPS↓
w/ AiA	**20.27**	**0.1502**	**0.3076**
w/ SFT	21.50	0.1530	0.3101

Table 2. Ablation studies on cognitive information, reference guidance, and All-in-Attention (AiA) module on ImageNet Test2000. We remove controllable feature wrapping [53] for evaluation.

5. Conclusion

This paper introduces a pioneering approach to endow super-resolution (SR) with cognitive abilities. Our model excels in producing high-definition reference images that aid the SR process. We also introduce an All-in-Attention module to enhance fidelity. Our extensive experiments demonstrate the practical effectiveness of this approach.

Acknowledge: This work was partly supported by the National Natural Science Foundation of China (Grant No. 61991451) and the Shenzhen Science and Technology Program (JCYJ20220818101001004).

References

[1] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *ECCV*, pages 185–200, 2018. 2

[2] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pages 3086–3095, 2019. 2, 6

[3] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *ECCV*, pages 325–342. Springer, 2022. 2, 3

[4] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021. 2, 3

[5] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, pages 1652–1660, 2019. 2

[6] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, pages 1652–1660, 2019. 2

[7] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *ACMMM*, pages 1329–1338, 2022. 6

[8] Jin Chen, Jun Chen, Zheng Wang, Chao Liang, and Chia-Wen Lin. Identity-aware face super-resolution for low-resolution face recognition. *SPL*, 27:645–649, 2020. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 3

[11] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 44(5):2567–2581, 2020. 6

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 5

[13] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pages 9935–9946, 2023. 2, 3

[14] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, pages 3012–3021, 2020. 2, 3

[15] Bahadir K Gunturk, Aziz Umit Batur, Yucel Altunbasak, Monson H Hayes, and Russell M Mersereau. Eigenface-domain super-resolution for face recognition. *TIP*, 12(5): 597–606, 2003. 2

[16] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *ICONIP*, pages 387–395. Springer, 2021. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 6

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 3, 5

[20] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[21] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, pages 466–467, 2020. 6

[22] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, pages 2103–2112, 2021. 2

[23] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *NeurIPS*, 35: 23593–23606, 2022. 2, 3

[24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 6

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *ICCV*, 2023. 2

[27] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 2, 3

[28] Guangyuan Li, Wei Xing, Lei Zhao, Zehua Lan, Jiakai Sun, Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, and Zhijie Lin. Self-reference image super-resolution via pre-trained diffusion large model and window adjustable transformer. In *ACMMM*, pages 7981–7992, 2023. 3

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 4, 6

[30] Wenbo Li, Kun Zhou, Lu Qi, Liying Lu, and Jiangbo Lu. Best-buddy gans for highly detailed image super-resolution. In *AAAI*, pages 1412–1420, 2022. 3

[31] Yu-Jhe Li, Shawn Hunt, Jinhyung Park, Matthew O'Toole, and Kris Kitani. Azimuth super-resolution for fmcw radar in autonomous driving. In *CVPR*, pages 17504–17513, 2023. 2

[32] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 2

[33] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, pages 574–591. Springer, 2022. 6

[34] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *CVPR*, pages 5657–5666, 2022. 3

[35] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2, 3

[36] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *TPAMI*, 45(5):5461–5480, 2022. 2

[37] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, pages 6368–6377, 2021. 2, 3

[38] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 2023. 3, 8

[39] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 3

[40] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, pages 291–300, 2020. 2

[41] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2437–2445, 2020. 2, 3

[42] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[43] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2

[44] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *TPAMI*, 44(11):7474–7489, 2021. 2, 3

[45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3, 4, 6

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 3

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3, 6

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 2, 3

[50] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 45(4):4713–4726, 2022. 3

[51] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang. Region-adaptive deformable network for image quality assessment. In *CVPRW*, pages 324–333, 2021. 6

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[53] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 2, 3, 5, 6, 8

[54] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Dual super-resolution learning for semantic segmentation. In *CVPR*, pages 3774–3783, 2020. 2

[55] Ruoxi Wang, Dandan Zhang, Qingbiao Li, Xiao-Yun Zhou, and Benny Lo. Real-time surgical environment enhancement for robot-assisted minimally invasive surgery based on super-resolution. In *ICRA*, pages 3434–3440. IEEE, 2021. 2

[56] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018. 2, 8

[57] 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 *ECCVW*, pages 0–0, 2018. 2

[58] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, pages 9168–9178, 2021. 2, 3

[59] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 2, 3, 6

[60] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 2, 3

[61] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, pages 101–117. Springer, 2020. 2, 6

[62] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*, pages 13385–13394, 2021. 2

[63] Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *AAAI*, pages 2768–2776, 2022. 2, 3

[64] Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. Desra: Detect and delete the artifacts of gan-based real-world super-resolution models. In *ICML*, pages 38204–38226. PMLR, 2023. 3

[65] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 2

[66] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 2, 3, 5

[67] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pages 1191–1200, 2022. 6

[68] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021. 2, 3

[69] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 2, 3

[70] Tao Yang, Peiran Ren, Lei Zhang, et al. Synthesizing realistic image restoration training pairs: A diffusion approach. *arXiv preprint arXiv:2303.06994*, 2023. 2

[71] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, pages 701–710, 2018. 2

[72] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 2, 3

[73] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 2, 3, 6

[74] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 5, 6

[75] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[76] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 2

[77] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, pages 7982–7991, 2019. 2, 3

[78] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023. 2

[79] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *ECCV*, pages 88–104, 2018. 2, 3

[80] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *NeurIPS*, 33:3499–3509, 2020. 2