CLIP-SR: Collaborative Linguistic and Image Processing for Super-Resolution

Bingwen Hu, Heng Liu, Zhedong Zheng, and Ping Liu, *Senior Member, IEEE*

***Abstract*—Convolutional Neural Networks (CNNs) have significantly advanced Image Super-Resolution (SR), yet most CNN-based methods rely solely on pixel-based transformations, often leading to artifacts and blurring, particularly under severe downsampling rates (*e.g.*, 8× or 16×). The recently developed text-guided SR approaches leverage textual descriptions to enhance their detail restoration capabilities but frequently struggle with effectively performing alignment, resulting in semantic inconsistencies. To address these challenges, we propose a multi-modal semantic enhancement framework that integrates textual semantics with visual features, effectively mitigating semantic mismatches and detail losses in highly degraded low-resolution (LR) images. Our method enables realistic, high-quality SR to be performed at large upscaling factors, with a maximum scaling ratio of 16×. The framework integrates both text and image inputs using the prompt predictor, the Text-Image Fusion Block (TIFBlock), and the Iterative Refinement Module, leveraging Contrastive Language-Image Pretraining (CLIP) features to guide a progressive enhancement process with fine-grained alignment. This synergy produces high-resolution outputs with sharp textures and strong semantic coherence, even at substantial scaling factors. Extensive comparative experiments and ablation studies validate the effectiveness of our approach. Furthermore, by leveraging textual semantics, our method offers a degree of super-resolution editability, allowing for controlled enhancements while preserving semantic consistency.**

***Index Terms*—Image Super-Resolution, CLIP, Multi-modal Fusion, Language Guidance**

I. Introduction

The advent of Convolutional Neural Networks (CNNs) has significantly advanced the field of image super-resolution (SR) [1]–[6]. Early CNN-based SR methods, which relied solely on low-resolution (LR) images to reconstruct high-resolution (HR) counterparts, often struggled to increase the reconstruction quality of their outputs. To overcome these limitations, subsequent research [7]–[15] introduced prior information to guide the SR process, aiming to compensate for the missing details in LR images. While prior-based approaches have demonstrated improvements, they tend to be restricted to specific types of images, such as those with well-defined structures or attributes (*e.g.*, facial images). Moreover, methods such as SFTGAN [16], which leverage semantic segmentation maps to assist the SR reconstruction procedure, often introduce additional computational costs and are highly dependent on the accuracy of the segmentation process.

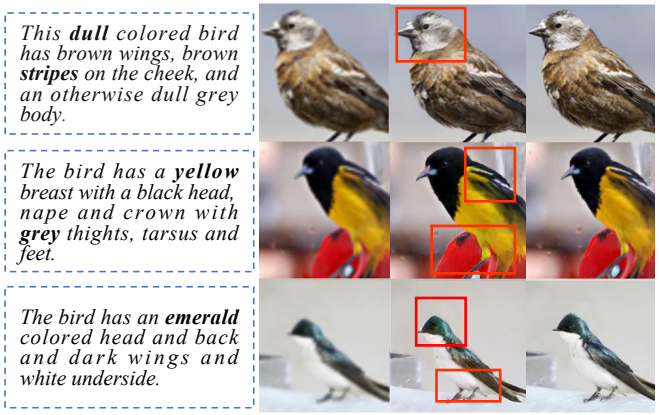


Fig. 1. Visualization of the results recovered by our method from low-resolution (LR) inputs. We highlight the semantic coherence part by aligning the textual guidance with the high-resolution (HR) ground truth.

The use of text descriptions as a form of semantic guidance has emerged as a more flexible and comprehensive alternative for addressing these limitations. Text offers richer and more detailed semantic information, which can guide the super-resolution process across a broader range of images. TGSR [17] was the first method to explore this approach; it uses text to enhance its ability to generate SR image details. However, challenges remain with regard to this method, particularly in terms of achieving effective text-image feature matching and semantic alignment, leading to inconsistencies between the input LR images and the generated SR results. In this paper, we propose a novel approach that ensures semantic consistency while achieving large-scale super-resolution. Our method leverages text descriptions to guide the SR process, ensuring that the reconstructed HR images are both semantically coherent and visually realistic. As shown in Figure 1, our approach addresses the limitations of the previously presented methods, providing a robust solution for conducting high-fidelity SR.

To address the challenges posed by the limitations of prior-based methods and ineffective text-image feature matching techniques, particularly when handling large-scale resolution

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61971004, the Anhui Province Higher Education Institution Collaborative Innovation Project under Grant No. GXXT-2022-044, and in part by the Natural Science Research Project of Anhui Educational Committee under Grant No. 2024AH050161. *(Corresponding author: hengliu@ahut.edu.cn).*

B. Hu is with the School of Computer Science and Technology, Anhui University of Technology, China (e-mail: hu_bingwen@ahut.edu.cn).

H. Liu is with the School of Computer Science and Technology, Anhui University of Technology, China, and the Institute of Artificial Intelligence, Heifei Comprehensive National Science Center, China (e-mail: hengliu@ahut.edu.cn).

Z. Zheng is with FST and ICI, University of Macau, China (e-mail: zhedongzheng@um.edu.mo).

P. Liu is with the Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA (e-mail: pino.pingliu@gmail.com).

Manuscript received April 19, 2005; revised August 26, 2015.

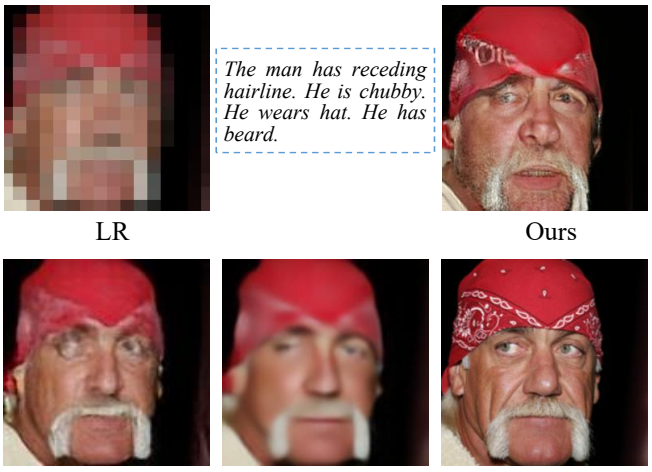


Fig. 2. An example comparison between the 16× image SR results of our method and two SOTA SR methods: DICGAN [18] and CRAFT [19]. Here, DICGAN and CRAFT are retrained on the same dataset as that used by our approach; LR is the input low-resolution image, and GT is the high-resolution ground truth (enlarged appropriately for visualization purposes).

degradation and high semantic ambiguity in real-world scenarios, we introduce a novel approach: Multi-modal Collaborative Semantic Enhancement for Super-Resolution (SR). Rather than treating the relevant text as mere prior guidance, we leverage the text information in conjunction with the LR image as two inputs modalities for SR tasks. Combining these modalities enhances their local semantics and enables high-performance, large-scale SR. Specifically, we introduce a prompt predictor designed to extract essential semantic elements from the input text. Inspired by VPT [20] and GALIP [21], the prompt predictor incorporates a fully connected layer and a self-attention mechanism, serving as a text-driven attention module. Unlike directly inputting raw text vectors into the pre-trained CLIP-ViT, the refined text vectors generated by the prompt predictor enable CLIP-ViT to better align the semantic features between the text and image, thereby improving the quality of the produced cross-modal representation.

To further optimize text-image interactions, we introduce TIFBlock, a novel alignment and fusion module that is specifically designed to enhance the cross-modal integration process. Leveraging pre-trained models such as CLIP [22] for the initial feature extraction step, TIFBlock effectively synthesizes and refines representations, resulting in a significant improvement in its text-to-image matching performance. Building upon TIFBlock, we develop an iterative refinement module, which is a structure dedicated to iterative detail recovery and semantic enhancement. This module progressively refines local details, addresses blurred regions, and maintains semantic consistency across different iterations. A core component of the iterative refinement module is the inclusion of a customized residual connection that is tailored to our framework, which facilitates smooth feature propagation while preserving semantic integrity. The customized residual connection is seamlessly integrated within this module to further optimize the pixel transition and feature propagation tasks, ensuring robust multi-modal fusion. Together, these components align with our design objectives of delivering seamless and effective collaboration between modalities.

By integrating textual descriptions with LR images, the proposed method enhances SR by leveraging both linguistic semantics and visual features. The traditional SR methods rely solely on visual information and struggle to reconstruct fine details in severely degraded images. In contrast, our collaborative framework uses textual guidance to refine structures and textures, producing SR images that are both visually realistic and semantically aligned with the input text. As shown in Figure 2, our method achieves high-fidelity reconstruction effects for a 16× downsampled facial image, demonstrating competitive performance with state-of-the-art SR techniques. Furthermore, it offers strong interpretability and ensures semantic consistency with the given text descriptions.

The primary contributions of this work are as follows:

- We introduce a new multi-modal semantic coherence approach for large-scale image super-resolution, generating semantically consistent and realistic high-resolution images from severely degraded low-resolution inputs.
- We design a novel Text-Image Fusion Block (TIFBlock) and integrate it with a pre-trained cross-modality model to form an iterative collaborative fusion structure, enabling our framework to progressively restore image details while enhancing local semantics.
- We investigate the impact of diverse textual semantics on image super-resolution. Comprehensive comparative experiments and ablation studies validate the effectiveness of our SR approach, which also maintains semantic coherence.

II. Related Work

A. Prior-Based Image Super-Resolution

Single-image super-resolution (SISR) has become a dynamic area within end-to-end deep learning [23]. The development of diverse models and mechanisms has significantly improved SR methods, particularly in terms of their pixel reconstruction and detail approximation capabilities. Early SR approaches [4], [24]–[28] usually assume that LR image pixels are obtained through bicubic downsampling performed on their HR counterparts. These methods employ various deep mapping networks to directly reconstruct SR image pixels from LR inputs. While these approaches can produce promising results on synthetic data with small-scale degradation, their effectiveness deteriorates significantly in real-world, large-scale degradation scenarios because of the full or partial loss of LR semantics.

To attain improved performance in real-world SR scenarios, numerous prior-based approaches, which deploy explicit or implicit priors to enrich the detail generation process, have been proposed. A representative explicit method is reference-based SR [29]–[32], which leverages one or more high-resolution reference images that share similar textures to those of the input low-resolution image to guide the process of generating an HR output. However, matching the features of

the reference with a low-resolution input could be challenging, and these explicit priors may not be available.

The recent methods, including FSRNet [7], DeepSEE [11], and SFTGAN [8], have shifted toward leveraging implicit priors, yielding improved results by integrating prior information directly into the SR process. For example, FSRNet [7] leverages geometric priors to improve the SR effects produced for facial images, whereas Zhang *et al.* [33] harnessed multi-view consistency. DeepSEE [11] utilizes semantic maps to explore extreme image SR. SFTGAN [8] introduces image segmentation masks as prior features for facial image SR. Although they are effective, these implicit priors are often tailored to specific situations, such as restricted categories [34], [35] or facial images [7], [8], [36], [37], limiting their applicability to more complex, real-world SR tasks. Recent progress in single-image super-resolution has leveraged visual language models and text-guided techniques to achieve increased restoration quality. Methods such as TGSR [17], CoSeR [38], XPSR [39], and TGESR [40] incorporate text semantics as prior conditions, providing additional contextual guidance for the SR reconstruction procedure.

B. Multi-modal Fusion Guided Image Generation

Multi-modal fusion has become an increasingly prevalent approach in various visual tasks, such as image generation, style transfer, and image editing. For example, keypoints are commonly utilized in motion generation [41] and automatic makeup applications [42]. In text-based image synthesis scenarios, GAN-INT-CLS [43] employs text descriptions to generate images using conditional Generative Adversarial Networks (cGANs). To enhance the quality of image, Stack-GAN [44], AttnGAN [45], and DM-GAN [46] leverage multiple generators and discriminators. DF-GAN [47] simplifies the text-to-image synthesis process with a more streamlined and effective approach. LAFITE [48] introduces a contrastive loss based on the CLIP model [22], offering more accurate guidance for generating precise images. In artistic style transfer cases, CLIPstyler [49] enables domain-independent texture transfer from text descriptions to source images, whereas CLVA [50] employs a patchwise style discriminator to extract visual semantics from style instructions, thereby achieving detailed and localized artistic style transfer. SISGAN [51] pioneered the use of an encoder-decoder architecture for conducting text-based semantic editing on images. ManiGAN [52] introduces a two-stage architecture with an attentional cropping module (ACM) and a deformable cropping module (DCM) to facilitate independent network training for text-based image editing. The lightweight GAN [43] further improves the efficiency of the process by applying a word-level discriminator. ManiTrans [53] employs a pre-trained autoregressive transformer, utilizing the CLIP model [22] for addressing semantic losses. More recently, Zeng *et al.* [54] developed a multiround image-editing framework using language guidance.

The emergence of large language models has further spurred advancements in the text-to-image generation field. DALL-E [55] uses VQ-VAE [56] to decompose images into discrete tokens, framing image synthesis as a translation task. LDM [57] applies diffusion models to latent image vectors, enabling an efficient training process with high-quality results. GLIDE [58], which is a diffusion-based text-to-image generation model, uses guided diffusion to enhance the text-conditioned synthesis procedure. GALIP [21] incorporates the CLIP model within adversarial learning for text-to-image synthesis purposes. ControlNet [59], which is introduced by Zhang *et al.*, builds upon the pre-trained Stable Diffusion [57], incorporating a detailed scheme control to guide the image generation process.

Recent advancements in pre-trained diffusion models [57], [58], [60] have significantly improved their image-generation capabilities. While studies [61]–[65] have underscored the generative potential of these models, applying them to SR remains challenging. The high fidelity required for SR demands both speed and efficiency—qualities that diffusion models generally lack due to their multi-step denoising process, which results in slower generation times and complicates latent space manipulation operations.

Compared with the use of pre-trained diffusion models, a GAN-based model is employed in this work for several key reasons. GANs facilitate high-resolution image generation in a single pass, which significantly improves upon the efficiency of diffusion models with an iterative nature. Furthermore, they provide a smooth latent space that enables intuitive control over the generated features, making them particularly well-suited for SR tasks. Additionally, GANs require less training data and computational resources, improving their accessibility for researchers. By leveraging GANs, we aim to achieve high-quality image generation while ensuring the practical applicability of super-resolution.

III. Method

In this section, we present an overview of our proposed CLIP-SR method, followed by detailed descriptions of each component contained within our multi-modal cooperative image super-resolution (SR) network. Finally, we introduce the total loss function used in our approach.

A. Overview

The traditional small-factor SR methods generate HR images from LR images by using deep SR networks. However, large-factor downsampling operations often lead to significant blurring in LR images, making it challenging for SR networks to reconstruct semantically consistent and precise details solely from pixel-space information. To address these challenges, we introduce textual semantics as a complementary input, enabling our network to leverage information derived from both the pixel and textual spaces for generating more accurate details. For clarity, we denote the input low-resolution image as L_{LR}, the complementary text description as T, and the corresponding high-resolution ground truth as I_{GT}. The objective of CLIP-SR, denoted $\mathcal{H}$, is to fuse L_{LR} and T to generate a semantically consistent and visually realistic super-resolution image, which is denoted as I_{SR}.

Specifically, we introduce a text-image fusion block (TIFBlock) within a multi-modal iterative refinement model, which integrates CLIP [22] and a TIFBlock to effectively perform

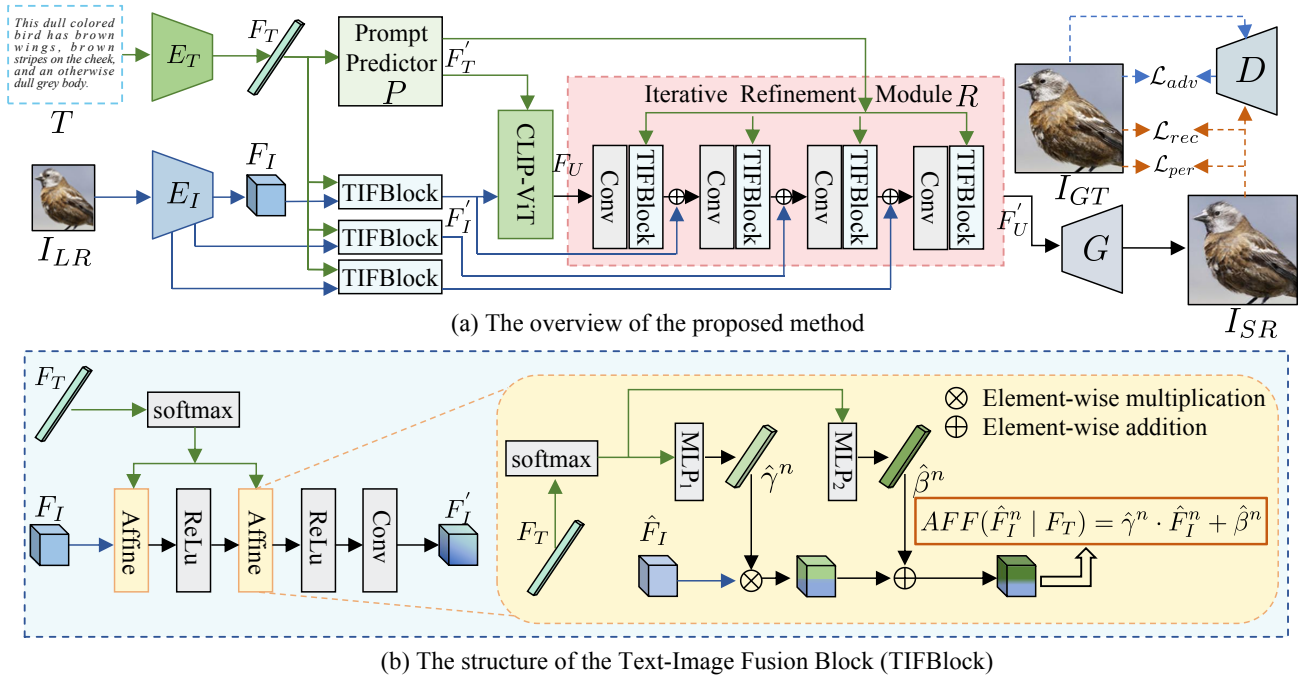


Fig. 3. The architecture of our proposed multi-modal cooperative semantic enhancement model for large-factor image super-resolution (see subfigure (a)). Given an input low-resolution image I_{LR} and text guidance T, features F_I and F_T are first extracted using an image encoder E_I and a text encoder E_T, respectively. The text feature vector F_T is further refined through a prompt predictor module P and then processed by the CLIP-ViT model to enhance textual guidance. The refined text and image features are subsequently integrated using a Text-Image Fusion Block (TIFBlock), which aligns and combines the two modalities (see subfigure (b)). Within the TIFBlock, an affine transformation is applied in its text fusion module. This transformation employs two consecutive MLPs that generate channel-wise scaling parameters ($\hat{\gamma}^n = MLP_1(Softmax(F_T))$) and shifting parameters ($\hat{\beta}^n = MLP_2(Softmax(F_T))$). These parameters adaptively modulate the channel-wise features of the visual representation $\hat{F}_I^n$. Finally, the fused multi-modal features undergo iterative refinement and semantic enhancement through a continuous Conv-TIFBlock structure, which is referred to as an iterative refinement module R. This iterative process ensures that progressively improved super-resolution outputs with enhanced details and semantic coherence are obtained.

large-factor SR. To efficiently combine information derived from different modalities, *i.e.*, text and images, we design a robust fusion strategy that preserves essential textual details while avoiding the information losses observed in simpler approaches [44], [51], [66] that directly merge text vectors with image features. Our TIFBlock employs an affine transformation alignment strategy to increase the accuracy of text-to-image fusion and retain critical semantic details. Given the inherent differences between text and image features, precise alignment is crucial for achieving semantic coherence. To further reduce cross-modal inconsistencies, a prompt predictor is employed to process the text vectors prior to conducting alignment. Additionally, the CLIP model [22] is integrated within our framework as a supplementary alignment tool, ensuring a contextually precise and semantically coherent text-image fusion process for SR. To ensure coherence with the LR content contained in the generated SR image, we design two additional mechanisms that build on our fusion strategy. Specifically, we incorporate residual connections to preserve the essential LR details, particularly in cases where semantic conflicts may arise. Additionally, text semantics are integrated at each layer of the multi-modal iterative refinement module, progressively guiding the SR process with fine-grained adjustments. These refined semantic fusion strategies ensure that the generated SR image remains both structurally and semantically consistent with the LR input. Figure 3 provides an overview of the overall network architecture and the details of the TIFBlock.

B. Network Architecture

In this section, we present the key components of our proposed multi-modal large-factor image super-resolution model. The model primarily comprises five components: text and image encoders, a prompt predictor, a text-image fusion block (TIFBlock), an iterative refinement module, and a CLIP-based discriminator.

In essence, the text and image encoders extract text vectors and image features, respectively, providing foundational representations for the following steps. The TIFBlock aligns and fuses these features, enabling the cohesive integration of textual and visual information. CLIP-ViT and the prompt predictor effectively enhance the textual guidance provided throughout the generation process. The iterative refinement module progressively restores image details and enhances local semantics through multiple iterations, ensuring alignment between different modalities. Finally, the CLIP-based discriminator comprehensively evaluates the fidelity, semantic quality, and coherence of the generated image. By leveraging the synergistic interaction among these five components, our method generates semantically consistent and realistically reconstructed high-resolution images, even from severely degraded low-resolution inputs (*e.g.*, with 8× or 16× downsampling).

1) Text and Image Encoders: We utilize two distinct encoders to process the input modalities. The text encoder, which is denoted as E_T, follows the architecture of CLIP [22] and encodes textual inputs T into feature vectors F_T, where

$F_T = E_T(T)$, to effectively capture semantic information. For the input LR image I_{LR}, the image encoderE_I employs a series of convolutional layers to progressively transform the input into an 8×8 feature map F_I, where $F_I = E_I(I_{LR})$. These encoders allow our model to generate compatible feature representations for both text and image inputs, preparing them for the subsequent fusion step within the network.

2) Prompt Predictor: Before leveraging the pre-trained CLIP-ViT model to align image features with corresponding text vectors, we introduce a prompt predictor inspired by VPT [20] and GALIP [21]. The prompt predictor, which is denoted as P, comprises a fully connected (FC) layer and a self-attention layer; the predictor functions as a text-driven attention mechanism. It predicts text-conditioned prompts, $F_T^{'} = P(F_T)$, which are appended to the visual patch embeddings in CLIP-ViT. This design enables the generated images to more effectively capture the semantic content of the input text while maintaining alignment with the visual information encoded by the CLIP-ViT model.

The prompt predictor leverages the output of the text encoder to selectively focus on salient textual elements, which are then fused with the visual features. This integration process enables the generator to more accurately interpret and translate the given text into detailed, coherent visual representations, enhancing the degree of alignment between the text descriptions and the generated images in terms of both content and quality.

3) Text-Image Fusion Block (TIFBlock): To further enhance the influence of text information on images, we introduce a Text-Image Fusion Block (TIFBlock) that integrates textual semantics as a complementary feature source. As shown in Figure 3 (b), the TIFBlock incorporates an affine transformation within its text fusion module. Following the design principles of DF-GAN [47], we introduce a ReLU layer after each affine layer to increase the diversity of text-fused images by introducing nonlinear relationships. Additionally, to improve the ability of the model to comprehend text descriptions, we apply a Softmax function to re-weight the text features before passing them to the affine layer. This re-weighting strategy allows for a smoother and more reliable integration of the text and image domains.

The process of the TIFBlock starts by feeding the LR image I_{LR} into the image encoder network E_I, extracting an image feature vector F_I. Moreover, the text is encoded via the pre-trained CLIP encoder E_T, producing a text vector F_T. The text features are then re-weighted via the Softmax function before being passed through the affine transformation layer. Within this layer, the re-weighted text vector is processed through two consecutive Multi-Layer Perceptrons (MLPs), which generate a channel-wise scaling parameter $\hat{\gamma} = MLP_1(Softmax(F_T))$ and the channel-wise shifting parameter $\hat{\beta} = MLP_2(Softmax(F_T))$. The affine transformation then adaptively adjusts the channel-wise features of the visual feature $\hat{F}_I^n$. The affine transformation is defined as follows:

$$AFF(\hat{F}_I^n \mid F_T) = \hat{\gamma}^n \cdot \hat{F}_I^n + \hat{\beta}^n, \tag{1}$$

where AFF denotes the affine transformation, $\hat{F}_I^n$ represents the n-th channel of the visual feature map $\hat{F}_I$, F_T represents the text vector, and γ^n and β^n are learnable scaling and shifting parameters, respectively. This mechanism enables the model to dynamically adjust the feature response to the textual context, leading to more accurate and meaningful alignment results.

The TIFBlock performs the initial alignment and integration steps on the text and image features by fusing these modalities through affine transformations, ensuring semantic consistency and accurate feature combinations. These fused multi-modal features are then passed to the Iterative Refinement Module, which progressively enhances the quality of the image by refining local details and reinforcing semantic coherence through multiple iterations. The iterative process builds on the fused features provided by the TIFBlock, enabling the model to generate outputs with higher resolution and realistic textures. Together, the TIFBlock establishes the foundational alignment of the two modalities, whereas the Iterative Refinement Module further optimizes and restores the image details in a step-by-step manner.

4) Iterative Refinement Module: To ensure that the generated image aligns closely with the given text, we iteratively refine the image features derived from CLIP-ViT by using a residual structure to fuse text-image features in a process that is guided by the text vector. Initially, the prompt predictor leverages the output of the text encoder to bridge the semantic gap between the text and image modalities. The low-resolution image features F_I are subsequently combined with the text vector F_T within the TIFBlock to further align the image and text features. CLIP-ViT is then employed to reconcile any inconsistencies between the image and text, ensuring that the final image features match the knowledge existing in both modalities. Finally, the outputs acquired from the prompt predictor, TIFBlock, and CLIP-ViT model are iteratively merged via the residual structure to generate a high-resolution image that is semantically consistent with the provided text.

Throughout the entire pipeline, we utilize text information at three key stages. First, we employ a simple convolutional network to extract features from the low-resolution image, which are integrated with the text information using the TIFBlock. This integration scheme ensures that the combined features encapsulate both detailed visual cues and semantic information, enabling precise guidance for the information flow within the CLIP-ViT network. Next, a text attention mechanism processes the textual features to address the inherent differences between the text and image modalities, facilitating an effective cross-modal alignment process. Additionally, the textual information serves as the input of a prompt predictor that feeds into the CLIP-ViT model, further enhancing the fusion results obtained for visual and semantic features. Finally, after obtaining preliminary image features from CLIP-ViT, the iterative refinement module progressively restores detailed image information by iteratively fusing it with textual semantics and enlarging the image through an additional upsampling module G. The upsampling module G consists of multiple blocks, each of which contains a 3×3 convolutional layer (with a kernel size of 3, a stride of 1, and a padding of 1) followed by a PixelShuffle layer (with an upscaling factor of 2). The number of blocks is determined by

the super-resolution scale factor.

5) CLIP-Based Discriminator: We utilize the CLIP-based discriminator proposed in GALIP [21], which extracts more informative visual features from complex images, enabling the discriminator to more effectively identify unrealistic image regions. This, in turn, prompts the generator to produce more realistic images. The structure of the discriminator provides a deep understanding of complex scenes by integrating additional visual information into the CLIP framework, making it particularly well-suited for its role as a discriminator. Specifically, the CLIP-based discriminator is designed to incorporate the language-image pre-training process of CLIP [22], with enhancements tailored to improving its effectiveness at evaluating the quality of generated images.

During training, the discriminator aims to distinguish between generated and real images. The superior performance of the CLIP model in terms of aligning text and images derived from different modalities allows the CLIP-based discriminator to gain a comprehensive and nuanced understanding of the image content, contributing to the generation of higher-quality and semantically consistent outputs by our proposed method.

C. Optimization Objectives

Reconstruction Loss. To ensure consistency in the content of the reconstructed images, we employ the pixel-wise $\mathcal{L}_1$-norm, which is defined as follows:

$$\mathcal{L}_{rec} = \mathbb{E}[\|\mathcal{H}(I_{LR}, T) - I_{GT}\|_1], \tag{2}$$

where $\mathcal{H}(I_{LR}, T)$ denotes the output generated by the full super-resolution network $\mathcal{H}$ proposed in this work, F_T represents the text description, and I_{GT} represents the high-resolution ground truth that corresponds to the input low-resolution image I_{LR}.

Perceptual Loss. Additionally, we use the perceptual loss [67] to encourage visual consistency between the generated super-resolution results and the real high-resolution images. The perceptual loss is defined as follows:

$$\mathcal{L}_{per} = \mathbb{E}\left[\sum_{i=0}^{5} \sigma_i \|\phi_i(\mathcal{H}(I_{LR}, T)) - \phi_i(I_{GT})\|_1\right], \tag{3}$$

where $\phi_i(.)$ denotes the feature map derived from the i-th layer of the pre-trained perception network ϕ. We employ the pre-trained VGG-19 network [68] as our ϕ and select five activation layers for computing the perceptual loss. The hyper-parameters σ_i modulate the contribution of the i-th layer to the total loss term in Equation 3.

Text-Constrained Adversarial Loss. To constrain the semantic information contained in the input text, we utilize the text-constrained adversarial loss [21]. Here, I_{LR} represents a given low-resolution image, and F_T is the text vector extracted from the corresponding text input. Both the low-resolution image I_{LR} and the text vector F_T are fed into the super-resolution network $\mathcal{H}$, resulting in an output $\mathcal{H}(I_{LR}, T)$. Let C and $\mathcal{V}$ represent the frozen CLIP-ViT model and the image feature extractor model contained in the CLIP-based discriminator, respectively. $Sim(.,.)$ denotes the cosine similarity between the generated HR image $\mathcal{H}(I_{LR}, F_T)$ and the text vector F_T.

The text-constrained adversarial loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv} = &- \mathbb{E}_{(\mathcal{H}(I_{LR},T)\sim\mathbb{P}_g)}\left[D(C(\mathcal{H}(I_{LR}, T), F_T))\right] \\ &- \alpha\mathbb{E}_{(\mathcal{H}(I_{LR},T)\sim\mathbb{P}_g)}\left[Sim(\mathcal{V}((\mathcal{H}(I_{LR}, T)), F_T)\right],\end{aligned} \tag{4}$$

where α is a hyper-parameter that controls the weight of the text-image similarity, and $\mathbb{P}_g$ denotes the synthetic data distribution.

Total Loss. Considering all of the above loss functions, the total objective function is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}, \tag{5}$$

where the hyper-parameter λ_{adv} controls the weight of the adversarial loss $\mathcal{L}_{adv}$.

IV. Experiment

A. Implementation Details

Dataset. We evaluate our method on the COCO [69], Caltech-UCSD Birds 200 (CUB) [70], and CelebA [71] datasets, each of which contains images paired with textual descriptions. For training, all images are cropped to the resolution of 256×256, with low-resolution images generated by performing bicubic downsampling on high-resolution counterparts. The utilized CLIP-ViT backbone is the ViT-B/32 model.

Setting. We train the proposed method on an NVIDIA RTX A5000 by using the Adam optimizer with parameters of $\beta_1 = 0.0$ and $\beta_2 = 0.9$ over 220 epochs. The hyper-parameter λ_{adv} is set to 0.01. Moreover, following the setup in GALIP [21], we set α to 4. Since the official code for TGSR [17] is unavailable, we use TGSR$^{\#}$ to represent results reproduced on the basis of the visual examples and quantitative metrics provided in the paper that propose TGSR for comparison with other methods.

B. Quantitative Evaluation

To quantitatively assess the quality of the SR images generated by different methods, we utilize two primary evaluation metrics: the Natural Image Quality Evaluator (NIQE) [72] and the Perceptual Index (PI) [73]. The NIQE evaluates the overall quality of SR images, with lower scores indicating more natural and realistic results. The PI, on the other hand, measures the perceptual quality of the images, where lower PI values correspond to better visual quality. We specifically choose the NIQE and PI for our experiments (except for Table II) instead of traditional metrics such as PSNR and SSIM, which focus more on image distortion but overlook objective quality and perceptual experience. In the context of SR, the NIQE and PI are more aligned with assessing the realism and naturalness of images, making them better suited for this task.

Table I presents our experimental results obtained on the CUB and COCO datasets. For the smaller CUB dataset, we compare the NIQE and PI scores with those of several state-of-the-art super-resolution methods, including EDSR [1], ESRGAN [2], SPSR [12], and TGSR$^{\#}$ [17]. Our method achieves the second-best NIQE score, closely following that of ESRGAN, while outperforming both Bicubic interpolation and EDSR in terms of the PI. On the larger COCO dataset, our

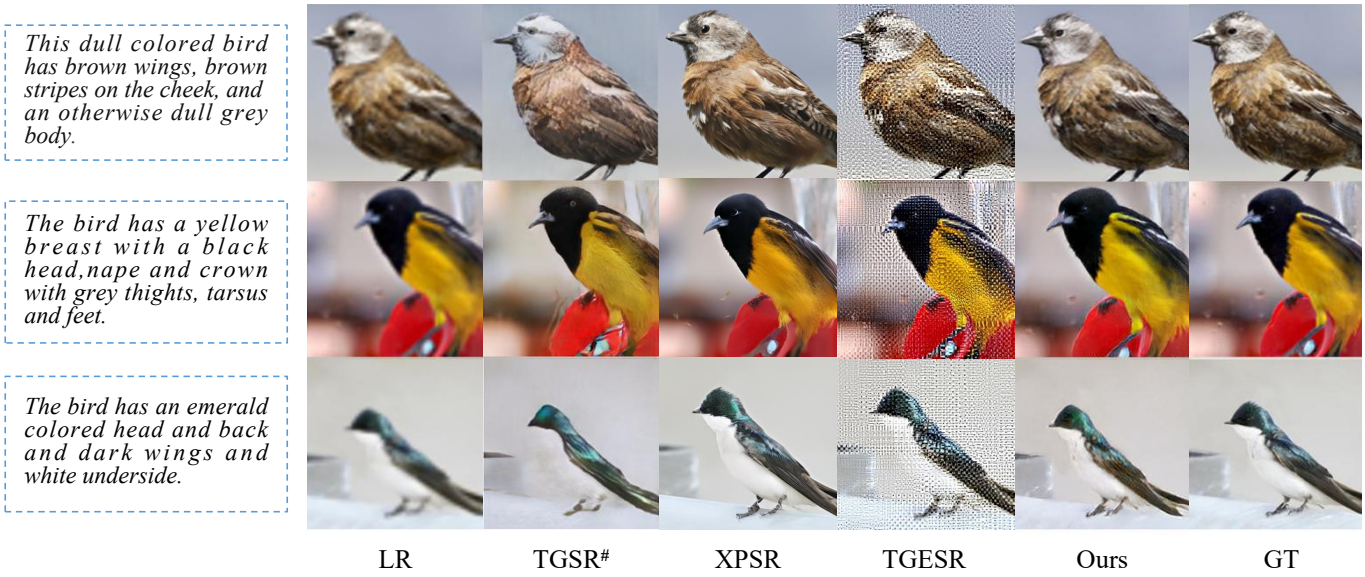


Fig. 4. Visual comparison among the $4\times$ SR results obtained with three SOTA SR methods, *i.e.,* TGSR [17], XPSR [39], and TGESR [40], and our method on the CUB dataset. Notably, # indicates the results reported in the corresponding original paper.

TABLE I
QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND THE COMPARATIVE METHODS ON THE CUB AND COCO DATASETS. THE SYMBOL ↓ DENOTES THAT LOWER VALUES OF THE ASSOCIATED ARE BETTER.

Dataset	Metrics	Bicubic	EDSR [1]	ESRGAN [2]	SPSR [12]	TGSR# [17]	Ours
CUB	NIQE ↓	12.374	10.684	**5.465**	5.885	6.623	**5.825**
	PI ↓	9.747	8.168	**2.644**	3.345	**2.560**	4.167
COCO	NIQE ↓	11.110	9.683	6.816	**6.378**	6.484	**4.706**
	PI ↓	9.373	8.515	7.135	6.060	**4.922**	**3.610**

TABLE II
QUANTITATIVE COMPARISONS RESULTS OBTAINED ON THE CELEBA DATASET.

Metrics	Bicubic	SuperFAN [9]	DICGAN [18]	TGSR# [17]	XPSR [39]	TGESR [40]	Ours
PSNR ↑	25.81	28.91	**33.61**	23.48	26.76	4.24	**28.974**
SSIM ↑	**0.844**	0.815	**0.895**	0.766	0.778	0.447	0.808
NIQE ↓	14.514	6.459	**5.755**	8.846	6.511	10.335	**5.172**
PI ↓	9.676	5.345	5.599	7.165	**5.235**	6.113	**4.476**

approach significantly outperforms all the comparison methods in both the NIQE and the PI, demonstrating superior generalizability. The observed performance degradations exhibited by the other approaches on COCO further underscore the robustness and versatility of our method.

Table II provides quantitative comparisons among the PSNR, SSIM, NIQE, and PI metrics produced on the CelebA dataset. Our method is evaluated against several baseline approaches, including Bicubic interpolation, SuperFAN [9], DICGAN [18], TGSR# [17], XPSR [39], and TGESR [40]. The results demonstrate that the proposed method achieves competitive performance across all the metrics. Specifically, compared with Bicubic interpolation, SuperFAN, and DICGAN, which rely solely on single-modality input, our approach incorporates supplementary textual information to achieve cross-modal semantic alignment, resulting in superior super-resolution performance. Moreover, in comparison with TGSR#, XPSR [39], and TGESR [40], which also utilize text guidance, our multi-modal collaborative semantic enhancement mechanism produces high-resolution images that are both semantically consistent and visually realistic. In summary, our method consistently delivers competitive results across three datasets, underscoring its effectiveness in image super-resolution tasks.

C. Qualitative Evaluation

To further validate the effectiveness of the proposed method, we conduct additional qualitative experiments. As illustrated in Figure 4, the experimental results demonstrate that our method achieves satisfactory visual outcomes even with this modification. These findings further confirm that the proposed multi-modal collaborative framework can consistently generate high-quality SR images with clear details and strong semantic coherence.

Concurrently, we conduct a $4\times$ SR experiment, upscaling low-resolution images from 64×64 to 256×256. As shown in Figure 5, our method, along with SuperFAN [9], DICGAN [18], XPSR [39], and TGESR [40], achieves commendable visual quality. However, SuperFAN, DICGAN, and TGESR exhibit noticeable artifacts, whereas our approach

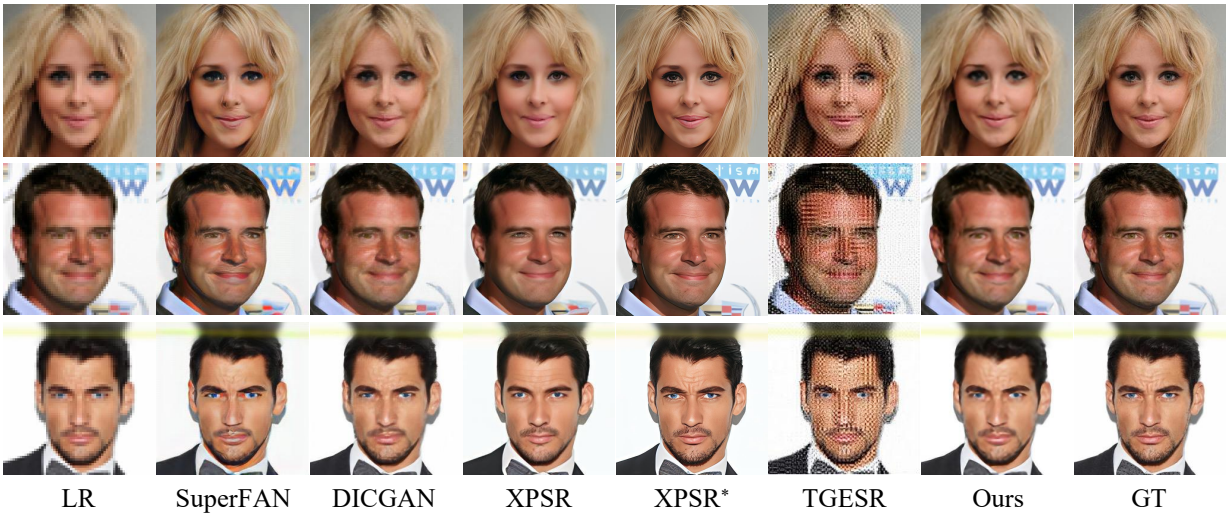


Fig. 5. Visual comparison among the $4\times$ SR results obtained with four SOTA SR methods, *i.e.,* SuperFAN [9], DICGAN [18], XPSR [39], TGESR [40], and our method on the CelebA dataset. $*$ denotes that $4\times$ SR is applied on the basis of the settings of XPSR, where an input image with a 128×128 resolution is upscaled to 512×512.

Fig. 6. Visual comparison among the $16\times$ SR results obtained with CRAFT [19], XPSR [39], TGESR [40], and our method on the CelebA dataset. $*$ indicates that $16\times$ SR is applied on the basis of the settings of XPSR, where an input image with a 32×32 resolution is upscaled to 512×512.

produces smoother and more visually appealing results, highlighting its advantage with regard to leveraging text guidance for performing cross-modal semantic alignment. Figure 6 further presents the qualitative results of a $16\times$ SR task conducted on the CelebA dataset. CRAFT [19] generates overly smoothed images, failing to recover fine details. While TGESR produces visually plausible results, it struggles to preserve the semantic integrity of the source images. XPSR, though effective at $4\times$ SR, undergoes severe distortions at $16\times$ SR, even under its original experimental settings, demonstrating a substantial performance drop in large-scale SR cases with heavily degraded images. In contrast, our method successfully super-resolves images to 256×256, achieving two key objectives: (1) restoring the essential semantic information and (2) maintaining high consistency with the original low-resolution input.

To compare the complexity and efficiency of our proposed SR model, we evaluate its number of parameters and inference time against those of several state-of-the-art models, including SuperFAN [9], XPSR [39], and TGESR [40]. The comparisons are conducted under the same conditions on an NVIDIA RTX 3090 GPU. As shown in Table III, our method significantly outperforms the diffusion-based XPSR [39] and TGESR [40] models in terms of model size and inference efficiency. However, owing to the incorporation of CLIP-ViT and the iterative refinement module, our model has a larger parameter count than SuperFAN [9] does and has a longer inference time. Nevertheless, given the superior SR performance of our approach, this trade-off is acceptable.

TABLE III
COMPLEXITY AND RUNTIME EFFICIENCY COMPARISONS AMONG DIFFERENT METHODS. THE RUNTIME REPRESENTS THE TIME CONSUMED FOR INFERRING EACH IMAGE.

Metrics	SuperFAN [9]	XPSR [39]	TGESR [40]	Ours
Parameters	1.3 M	1.9 B	5.5 B	632.2 M
Runtime	4.1 ms	6.3 s	39.9 s	24.1 ms

D. Ablation Studies and Further Discussion

To evaluate the effectiveness of each component included in our proposed method, we conducted ablation studies on the CUB dataset. We consider four variants: (1) a baseline U-Net for single image super-resolution, where $\mathcal{L}_{total} =$

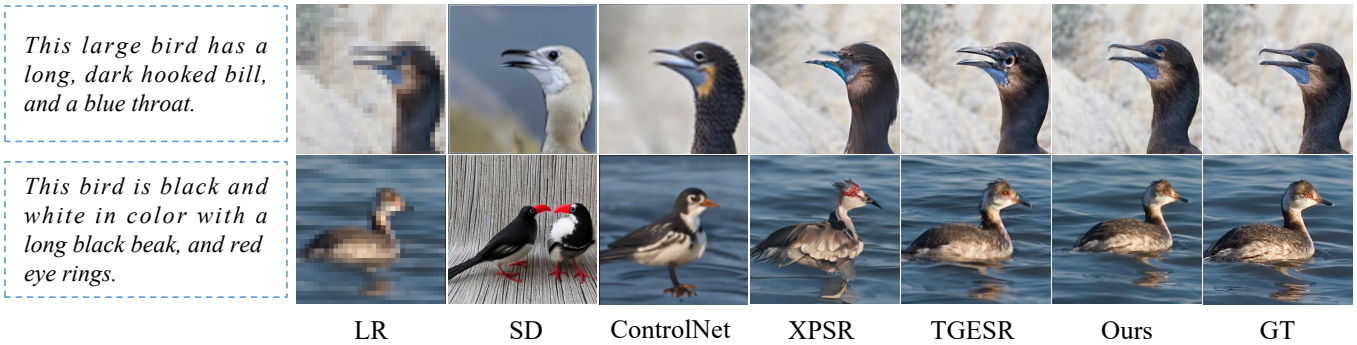


Fig. 7. Visual comparison among the 8× SR results obtained with four diffusion model-based methods, *i.e.,* Stable Diffusion (SD) [57], ControlNet [59], XPSR [39], TGESR [40], and our method on the CelebA dataset.

TABLE IV
COMPARISON AMONG THE QUANTITATIVE RESULTS OBTAINED WITH DIFFERENT COMPONENTS OF OUR METHOD ON THE CUB DATASET.

Variants	U-Net	Text	CLIP-ViT	D	NIQE ↓	PI ↓	SSIM ↑	PSNR ↑
1	✓	✗	✗	✗	13.057	10.384	0.391	16.150
2	✓	✓	✗	✗	6.244	5.855	0.834	26.987
3	✓	✓	✓	✗	6.578	6.020	0.835	27.891
4	✓	✓	✗	✓	6.178	6.004	0.835	27.210
Ours	✓	✓	✓	✓	**5.825**	**4.167**	**0.845**	**28.495**

$\mathcal{L}_{rec} + \mathcal{L}_{per}$; (2) variant 1 with additional text supervision that incorporates our proposed multi-modal fusion architecture (including the TIFBlock and iterative refinement module), where $\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{per}$; (3) variant 2 with a pre-trained CLIP-ViT model, where $\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{per}$; and (4) variant 2 with a CLIP-based discriminator D, where $\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}$.

As shown in Table IV, when the full model removes the text description module, CLIP-ViT, or the CLIP-based discriminator, the corresponding variants exhibit higher NIQE scores, indicating noticeable declines in performance. These experimental results demonstrate the effectiveness of incorporating textual information for enhancing the performance of the model. Additionally, they validate that the proposed Text-Image Fusion Block (TIFBlock) and Iterative Refinement Module effectively align textual and visual features, providing crucial semantic guidance for generating semantically consistent and realistic high-resolution images.

E. Analysis of the Capability of LR-to-SR

To evaluate the text-guided SR performance of our method, we conduct a comparison with Stable Diffusion (SD) [57], ControlNet [59], XPSR [39], and TGESR [40] on the 8× SR task. These methods take low-resolution (32 × 32) images along with textual descriptions as inputs. As shown in Figure 7, these models often introduce unwanted modifications, distorting the original visual information. In contrast, our method consistently produces sharper, more detailed SR images while effectively preserving both semantic coherence and fine-grained textures.

Our method outperforms SD and ControlNet in SR tasks for two main reasons. First, SD and ControlNet rely on iterative denoising, which struggles with performing 8× upscaling on severely degraded low-resolution images. While ControlNet introduces LR images as conditions, it lacks a mechanism for conducting semantic enhancement at extreme scaling factors. In contrast, our TIFBlocks iteratively refine both local textures and global semantics, producing more realistic and coherent SR results. Second, SD and ControlNet process image and text inputs separately, which can lead to misalignment between the textual descriptions and the generated images. Our approach employs a prompt predictor and iterative refinement module that leverage CLIP-based multi-modal alignment, ensuring semantic consistency and generating text-guided high-resolution outputs.

To further assess the effectiveness of text-guided SR methods, we conduct extensive experiments on multiple datasets, evaluating the tested methods, including XPSR and TGESR, across 4×, 8×, and 16× SR tasks. As shown in Figure 5, XPSR achieves satisfactory results at 4× SR for facial images but results in significant artifacts and structural distortions when it is applied to 8× and 16× SR tasks (see Figure 6 and Figure 7, respectively), indicating its limitations in terms of handling extreme upscaling. Conversely, TGESR performs well at higher scaling factors (8× and 16× SR) but struggles to maintain fine-grained details at 4× SR, suggesting an inconsistency in its ability across different upscaling levels. In comparison, our proposed CLIP-SR approach consistently generates high-quality super-resolved images across all scales, preserving both their semantic integrity and visual fidelity. As summarized in Table III, CLIP-SR not only outperforms XPSR and TGESR in terms of reconstruction quality but also has superior computational efficiency and reduced model complexity. These results highlight the robustness and scalability of our approach, making it well-suited for high-fidelity SR applications implemented under varying degradation conditions.

F. Analysis of the Editability of LR-to-SR Transformation

To evaluate the editability of our model in low-to-high-resolution transformations, we manipulate a subset of the CUB test images, as shown in Figure 8. Figure 8(b) presents color modifications in the nape, crown, and abdomen regions. Owing to low-resolution constraints, the network prioritizes pixel-level accuracy over high-level semantics, leading to slight blurring in the black region around the bird's head. Nevertheless, our method successfully adjusts the wing color in the abdomen area. When prompted with "yellow" (Figure 8(c)), the model effectively alters the wing hue, exhibiting variations across the outputs. This diversity underscores its ability to perform semantically consistent, controllable edits.

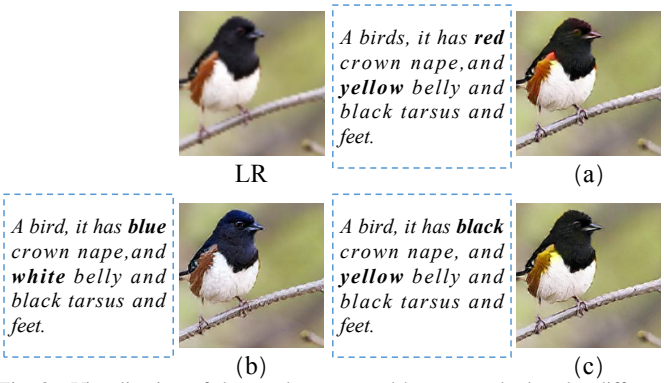


Fig. 8. Visualization of the results generated by our method under different text prompts. Our method demonstrates the ability to generate diverse and semantically consistent results.

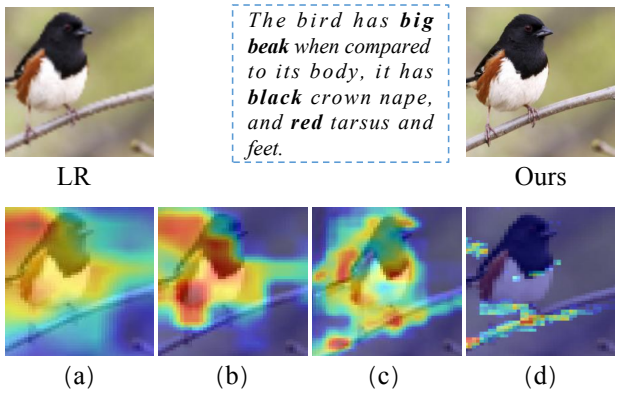


Fig. 9. Visualization of the heatmaps produced for low-resolution images during the super-resolution process. Subfigures (a), (b), (c), and (d) correspond to the results of performing text fusion at the first, second, third, and fourth layers, respectively, within the iterative refinement module

G. Analysis of the Effectiveness of Multi-modal Fusion and the Number of Iteration Layers

To evaluate the effectiveness of our multi-modal fusion module and the impact of different numbers of iterative fusion layers, we analyze the heatmap outputs produced across different layers within the multi-modal fusion module. As shown in Figure 9, each text input is paired with a corresponding low-resolution image. Figure 9 (a) shows the output derived from the initial text fusion layer, where the network begins by generating an image that is loosely aligned with the bird. In the subsequent layers, the attention of the model is progressively refined: in Figure 9 (b), the focus shifts to the bird's neck and body, whereas the further iterations shown in Figures 9 (c) and (d) progressively enhance finer details, including the bird's feet and tarsus. These findings empirically confirm the effectiveness of our iterative refinement module, demonstrating that four iterations are sufficient for achieving high-quality, semantically consistent text-to-image super-resolution results.

H. Limitations

Despite the superiority of the proposed method in the text-to-image super-resolution task, certain limitations warrant consideration in future research. The CLIP-ViT-B/32 model effectively leverages textual information to achieve enhanced image quality, particularly in the realm of semantic-guided super-resolution. It effectively bridges the gap between textual and visual data, enabling precise control over the process

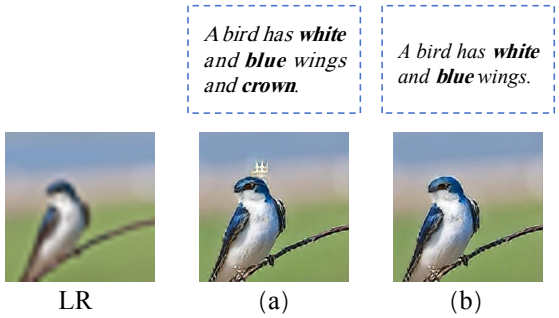


Fig. 10. Visualization of the results generated by our method under different text prompts. Given a low-resolution input image (a), our method produces super-resolution images guided by two distinct text descriptions: (b) and (c). While our method yields impressive results, certain details in the generated images may exhibit deviations due to the inherent ambiguity of natural language semantics.

of generating high-resolution images. However, despite these strengths, the model can occasionally misinterpret ambiguous descriptions. For example, as illustrated in Figure 10 (a), when instructed to generate an image featuring a "crown" on a bird, the model may incorrectly interpret the "crown" as a royal crown rather than the bird's crest. This misinterpretation underscores the necessity of including precise language in prompts. As demonstrated in Figure 10 (b), removing the term "crown" and providing a more specific context often yields the desired image. Future research could focus on enhancing the ability of the model to disambiguate homonyms and develop a deeper understanding of context-specific semantics.

V. Conclusion

We introduce a multi-modal semantic consistency framework for large-scale image super-resolution (SR) that leverages text-image fusion to enhance both the visual fidelity and semantic coherence of images. Our approach integrates a pre-trained cross-modal model within an iterative refinement process, enabling progressive detail recovery and text-guided enhancements. Extensive experiments demonstrate that our model achieves superior performance to that of the existing methods, particularly in terms of preserving fine-grained textures and maintaining semantic alignment. Despite these advancements, our method struggles with ambiguous textual inputs, which can lead to inconsistencies in the SR results. Future work will focus on refining text preprocessing techniques to attain improved instruction clarity, further enhancing the controllability and reliability of text-guided SR.

References

[1] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 2017, pp. 136–144. 1, 6, 7

[2] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision workshops*, 2018, pp. 0–0. 1, 6, 7

[3] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, "Coarse-to-fine cnn for image super-resolution," *IEEE Transactions on Multimedia*, vol. 23, pp. 1489–1502, 2020. 1

[4] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *Advances in Neural Information Processing Systems*, vol. 38, no. 2, pp. 295–307, 2015. 1, 2

[5] Y. Zhang, X. Yu, X. Lu, and P. Liu, "Pro-uigan: Progressive face hallucination from occluded thumbnails," *IEEE Transactions on Image Processing*, vol. 31, pp. 3236–3250, 2022. 1

[6] Y. Zhang, I. W. Tsang, J. Li, P. Liu, X. Lu, and X. Yu, "Face hallucination with finishing touches," *IEEE Transactions on Image Processing*, vol. 30, pp. 1728–1743, 2021. 1

[7] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2492–2501. 1, 3

[8] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8260–8269. 1, 3

[9] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 109–117. 1, 7, 8

[10] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6163–6172. 1

[11] M. C. Buhler, A. Romero, and R. Timofte, "Deepsee: Deep disentangled semantic explorative extreme super-resolution," in *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 3

[12] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7769–7778. 1, 6, 7

[13] W. Li, J. Li, G. Gao, W. Deng, J. Zhou, J. Yang, and G.-J. Qi, "Cross-receptive focused inference network for lightweight image super-resolution," *IEEE Transactions on Multimedia*, vol. 26, pp. 864–877, 2023. 1

[14] Y. Zhao, Q. Teng, H. Chen, S. Zhang, X. He, Y. Li, and R. E. Sheriff, "Activating more information in arbitrary-scale image super-resolution," *IEEE Transactions on Multimedia*, 2024. 1

[15] D. Liu, X. Wang, R. Han, N. Bai, J. Hou, and S. Pang, "Cte-net: Contextual texture enhancement network for image super-resolution," *IEEE Transactions on Multimedia*, 2024. 1

[16] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 606–615. 1

[17] C. Ma, B. Yan, Q. Lin, W. Tan, and S. Chen, "Rethinking super-resolution as text-guided details generation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3461–3469. 1, 3, 6, 7

[18] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5569–5578. 2, 7, 8

[19] A. Li, L. Zhang, Y. Liu, and C. Zhu, "Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 514–12 524. 2, 8

[20] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proceedings of the European Conference on Computer Vision*, 2022. 2, 5

[21] M. Tao, B.-K. Bao, H. Tang, and C. Xu, "Galip: Generative adversarial clips for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 214–14 223. 2, 3, 5, 6

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. 2, 3, 4, 6

[23] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019. 2

[24] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654. 2

[25] Kim, Jiwon and Lee, Jung Kwon and Lee, Kyoung Mu, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645. 2

[26] W. Shi, J. Caballero, F. Huszr, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883. 2

[27] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690. 2

[28] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301. 2

[29] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 88–104. 2

[30] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7982–7991. 2

[31] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5791–5800. 2

[32] Y. Jiang, K. C. Chan, X. Wang, C. C. Loy, and Z. Liu, "Robust reference-based super-resolution via c2-matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2103–2112. 2

[33] X. Zhang, Z. Zheng, D. Gao, B. Zhang, Y. Yang, and T.-S. Chua, "Multi-view consistent generative adversarial networks for compositional 3d-aware image synthesis," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 2219–2242, 2023. 3

[34] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "Glean: Generative latent bank for large-factor image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 245–14 254. 3

[35] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7474–7489, 2021. 3

[36] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9168–9178. 3

[37] T. Yang, P. Ren, X. Xie, and L. Zhang, "Gan prior embedded network for blind face restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 672–681. 3

[38] H. Sun, W. Li, J. Liu, H. Chen, R. Pei, X. Zou, Y. Yan, and Y. Yang, "Coser: Bridging image and language for cognitive super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 868–25 878. 3

[39] Y. Qu, K. Yuan, K. Zhao, Q. Xie, J. Hao, M. Sun, and C. Zhou, "Xpsr: Cross-modal priors for diffusion-based image super-resolution," in *Proceedings of the European Conference on Computer Vision*. Springer, 2024, pp. 285–303. 3, 7, 8, 9

[40] K. V. Gandikota and P. Chandramouli, "Text-guided explorable image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 900–25 911. 3, 7, 8, 9

[41] Y. Suo, Z. Zheng, X. Wang, B. Zhang, and Y. Yang, "Jointly harnessing prior structures and temporal consistency for sign language video generation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 6, pp. 1–18, 2024. 3

[42] Z. Huang, Z. Zheng, C. Yan, H. Xie, Y. Sun, J. Wang, and J. Zhang, "Real-world automatic makeup via identity preservation makeup net," in *International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence, 2021. 3

[43] B. Li, X. Qi, P. Torr, and T. Lukasiewicz, "Lightweight generative adversarial networks for text-guided image manipulation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 020–22 031, 2020. 3

[44] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked

generative adversarial networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 5907–5915. 3, 4

[45] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324. 3

[46] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5802–5810. 3

[47] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 515–16 525. 3, 5

[48] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Towards language-free training for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 907–17 917. 3

[49] G. Kwon and J. C. Ye, "Clipstyler: Image style transfer with a single text condition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 062–18 071. 3

[50] T.-J. Fu, X. E. Wang, and W. Y. Wang, "Language-driven artistic style transfer," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 717–734. 3

[51] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 5706–5714. 3, 4

[52] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, "Manigan: Text-guided image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7880–7889. 3

[53] J. Wang, G. Lu, H. Xu, Z. Li, C. Xu, and Y. Fu, "Manitrans: Entity-level text-guided image manipulation via token-wise semantic alignment and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 707–10 717. 3

[54] L. Zeng, Z. Zheng, Y. Wei, and T.-s. Chua, "Instilling multi-round thinking to text-guided image generation," *arXiv preprint arXiv:2401.08472*, 2024. 3

[55] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022. 3

[56] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017. 3

[57] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695. 3, 9

[58] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021. 3

[59] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847. 3, 9

[60] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022. 3

[61] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. 3

[62] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4296–4304. 3

[63] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3

[64] J. Wang, P. Liu, J. Liu, and W. Xu, "Text-guided eyeglasses manipulation with spatial constraints," *IEEE Transactions on Multimedia*, vol. 26, pp. 4375–4388, 2024. 3

[65] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4492–4501. 3

[66] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1060–1069. 4

[67] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 694–711. 6

[68] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. 6

[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755. 6

[70] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011. 6

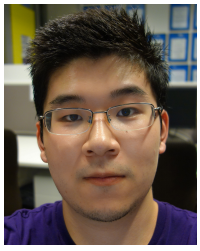
[71] D. Cheng, B. Price, S. Cohen, and M. S. Brown, "Beyond white: Ground truth colors for color constancy correction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015, pp. 298–306. 6

[72] Z. Wang, "Image quality assessment: Form error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 604–606, 2004. 6

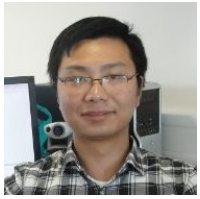
[73] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Proceedings of the European Conference on Computer Vision workshops*, 2018, pp. 334–355. 6

Bingwen Hu is a Lecturer with the Anhui University of Technology. He received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2023. From 2018 to 2020, he was a Joint Ph.D. Student with the Center for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW, Australia. His research interests include computer vision, image processing, and multimedia content analysis.

Heng Liu is a professor at the School of Computer Science and Technology, Anhui University of Technology, China. He received his Ph.D. in Pattern Recognition and Intelligent System from Shanghai Jiao Tong University in 2008. His current research interests include computer vision, biometrics, and deep learning. He has contributed over 100 research papers and has served as a program committee member for AAAI and IJCAI, as well as a reviewer for IEEE TIP, IEEE TGRS, IEEE TMM, ACM MM, ICCV, ECCV, etc.



Zhedong Zheng is an Assistant Professor with the University of Macau. He received the Ph.D. degree from the University of Technology Sydney in 2021 and the B.S. degree from Fudan University in 2016. He was a postdoctoral research fellow at the School of Computing, National University of Singapore. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He served as the senior PC for IJCAI and AAAI, and the area chair for ACM MM'24 and ICASSP'25.



Ping Liu is an Assistant Professor in the Computer Science department, University of Nevada, Reno, USA. He was a senior scientist in the Center for Frontier AI Research, A*STAR, Singapore, from 2020 to 2024. From 2018 to 2020, he was a Research Staff with the Center for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW, Australia. He received the bachelor's degree in electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2005, the master's degree from the Huazhong University of Science and Technology, Wuhan, in 2008, and the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2015. His research interests include computer vision and deep learning.