

# Rethinking Super-Resolution as Text-Guided Details Generation

Chenxi Ma

School of Computer Science,  
Shanghai Key Laboratory of  
Intelligent Information Processing,  
Shanghai Collaborative Innovation  
Center of Intelligent Visual  
Computing, Fudan University  
cxma17@fudan.edu.cn

Bo Yan\*

School of Computer Science,  
Shanghai Key Laboratory of  
Intelligent Information Processing,  
Shanghai Collaborative Innovation  
Center of Intelligent Visual  
Computing, Fudan University  
byan@fudan.edu.cn

Qing Lin

School of Computer Science,  
Shanghai Key Laboratory of  
Intelligent Information Processing,  
Shanghai Collaborative Innovation  
Center of Intelligent Visual  
Computing, Fudan University  
18210240028@fudan.edu.cn

Weimin Tan

School of Computer Science,  
Shanghai Key Laboratory of  
Intelligent Information Processing,  
Shanghai Collaborative Innovation  
Center of Intelligent Visual  
Computing, Fudan University  
wmtan14@fudan.edu.cn

Siming Chen

School of Data Science, Fudan  
University  
simingchen3@gmail.com

## ABSTRACT

Deep neural networks have greatly promoted the performance of single image super-resolution (SISR). Conventional methods still resort to restoring the single high-resolution (HR) solution only based on the input of image modality. However, the image-level information is insufficient to predict adequate details and photo-realistic visual quality facing large upscaling factors ( $\times 8$ ,  $\times 16$ ). In this paper, we propose a new perspective that regards the SISR as a semantic image detail enhancement problem to generate semantically reasonable HR image that are faithful to the ground truth. To enhance the semantic accuracy and the visual quality of the reconstructed image, we explore the multi-modal fusion learning in SISR by proposing a **Text-Guided Super-Resolution (TGSR)** framework, which can effectively utilize the information from the text and image modalities. Different from existing methods, the proposed TGSR could generate HR image details that match the text descriptions through a coarse-to-fine process. Extensive experiments and ablation studies demonstrate the effect of the TGSR, which exploits the text reference to recover realistic images.

## CCS CONCEPTS

• **Computing methodologies**  $\rightarrow$  *Reconstruction*.

## KEYWORDS

single image super-resolution, text-guided super-resolution, multi-modal fusion learning

## ACM Reference Format:

Chenxi Ma, Bo Yan, Qing Lin, Weimin Tan, and Siming Chen. 2022. Rethinking Super-Resolution as Text-Guided Details Generation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisbon, Portugal. ACM, New York, NY, USA, 9 pages. <https://doi.org/3503161.3547951>

## 1 INTRODUCTION

Since the LR image is too small to contain enough information for large-factor SR. To further enhance the SR performance, some external priors are introduced to provide more guidance for SR models. FSRNet [5], DeepSEE [2], and Christian et al. [28] exploit the face structure prior (face parsing map, face landmark heatmap) to restore face images. The audio-aided face SR [10] method utilizes the audio prior to guide the face super-resolution problem by extracting facial attributes (age, gender, ethnicity) from the voice of a speaker. These priors are only helpful for face images and can not generalize to natural images. SFTGAN [32] utilizes semantic segmentation maps to super-resolve LR images. However, these image-level semantic priors contain limited information and require additional calculation of existing semantic extraction methods, the accuracy of which greatly affects the super-resolution performance.

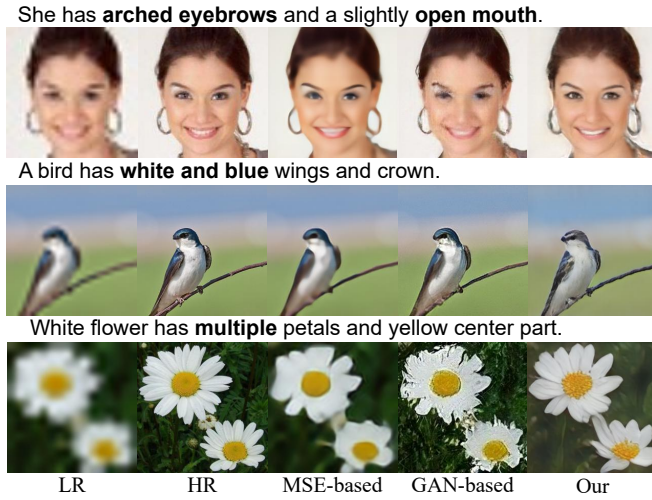
In comparison to the image-level prior, text description of an image contains more abundant semantic information, and describes global image style and local features of main objects, such as, color, shape, species, age, emotion, etc. The text description is easily available, and it can intuitively and flexibly express concepts of an image,

\*Corresponding Author. This work is supported by NSFC (Grant No.: U2001209, 61902076) and Natural Science Foundation of Shanghai (21ZR1406600, 21ZR1403300).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '22, October 10–14, 2022, Lisbon, Portugal.

© 2022 Association for Computing Machinery.  
ACM ISBN 978-1-4503-9203-7/22/10...\$15.00  
<https://doi.org/3503161.3547951>



**Figure 1: The large factor SR results of MSE-based SISR (DRN [11]) and face SR methods (SuperFAN [3]), GAN-based SISR (SPSR [22]) and face SR (DICGAN [21]) methods, and our models. Our model can make better use of text descriptions to restore clear and rational image details.**

it is helpful for us to picture the rough image contents in our mind. By utilizing text information in SR, we can increase the flexibility and controllability of SR, so that we can better meet expectations of people. For old photo recovery, the color or other details are contaminated, additional text information is useful to control the image contents and ensure the SR result conform to the common sense. In addition, for the surveillance video, investigators often draw a profile of characters according to descriptions of witnesses and low-quality surveillance video frames to search for suspects. This requires the SR method to generate face attributes that conform to descriptions, in addition to clear and reasonable textures. However, it is difficult for previous SR methods to restore HR images with specific attributes.

To address above issues, we first reveal the potential of text descriptions in SISR, that the text can provide important reference for SR approaches to restore reasonable results. Here, we propose a **Text-Guided Super-Resolution (TGSR)** method, which uses the multimodal fusion learning to integrate text semantic information into large-factor SR. Since text and image are data of totally different modalities, how to effectively fuse the two modalities and understand semantic information from text descriptions for better SR is a challenging task. The proposed TGSR adopts a coarse-to-fine framework to restore image details. The text information is embedded in the coarse SR stage to enhance image features and restore a rough SR image. In the fine SR stage, the TGSR further refines the visual quality of the final result. We also improve semantic accuracy of the SR result based on the text description during training. Therefore, the text description can guide the SR model to generate more accurate details and meanwhile manipulate the color, shape, texture, background, and other image characteristics.

As shown in Figure 1, if the LR image is too small (large-factor SR) and lack some visual information, traditional SR methods (DRN [11], SuperFAN [3], SPSR [22], DICGAN [21]) produce blur

and fake artifacts due to lacking the understanding of the contextual knowledge. In comparison, the proposed TGSR can effectively understand text descriptions and restore photo-realistic and reasonable results. In addition, the TGSR can generate diverse valid HR solutions for a single LR input by manipulating text descriptions, which enhances the flexibility and the practicability of SR.

The main contributions are as follows:

- This paper rethinks the SR as a semantic detail enhancement problem, that restores semantically reasonable HR details.
- This paper proposes the first text-guided image super-resolution (TGSR) approach based on the multimodal fusion learning, and explores the effectiveness of text descriptions to large-factor SR.
- The proposed TGSR adopts a coarse-to-fine process to restore an HR image, and introduces the text features through the text attention module to modulate image features. The text-guided losses constrain the network to pay more attention on image regions focused by text and to generate semantically reasonable textures.

## 2 RELATED WORK

### 2.1 Single Image Super-Resolution

**Deep Learning Based Super-Resolution:** Deep learning based methods have dominated the development of SISR for their strong representation and fitting capabilities. SRCNN [7], composed of three convolution layers, first introduces the convolutional neural network (CNN) into single image super-resolution (SISR) and leads to a dramatic leap. Based on this work, following SISR methods achieve continuous breakthroughs by proposing different network structures (VDSR [13], EDSR [17], DRN [11], Liu et al. [19]). By exploiting the generative adversarial network (GAN) [8] in SISR, SRGAN [14], ESRGAN [33], SPSR [22] generates visually pleasing results with more high-frequency details.

To further improve the SR performance, some recent methods (FSRNet [5], DeepSEE [2], SFTGAN [32]) introduce external priors (facial landmark heatmaps, parsing maps, semantic maps) to provide guidance for face image SR. To, SFTGAN [32] adopts the semantic segmentation result of a LR image to utilize the semantic prior for SR of natural images.

**Explorative Super-Resolution:** Most traditional SISR methods focus on outputting a unique HR image to approximate the ground truth and ignore the abundance of plausible HR explanations to the input LR image. Due to the ill-posed nature of the SR task, several recent works [1, 23] are proposed to break this limitation and explore infinitely many plausible reconstructions for a given LR image. Explorable Super-Resolution [1] generalizes the traditional SISR task toward an image restoration task, that can output different possible HR images with the observed LR image and can also support editing the output image through user interaction. In specific, the explorable SR method comprises a graphical user interface with a SR network.

PULSE [23] uses the GAN prior to generates a HR face image by optimizing the latent vector of a pre-trained GAN. The DeepSEE [2] also achieves explorative face super-resolution by controlling the semantic maps and the style vector to change the shape and appearance of specific semantic regions. The audio-aided face SR method [10] introduces the audio prior to guide the face SR by

exploiting an encoder-decoder to fuse features from the voice and image of a speaker. In this way, it is able to reconstruct different HR face images that have consistent characteristics with different input audios. These explorative SR works reveal more possibilities for image super-resolution.

## 2.2 Text Guided Image Reconstruction

Though the text prior has never been explored in SR, it is utilized in several computer vision fields, such as text-to-image synthesis, text-guided image colorization, these ideas inspire us to explore the SR task with text information, which is first done in the SR field.

**Text-to-Image Synthesis:** Generating images from text descriptions, a classical multimodal task, has received a lot of attention in academia. [26] first used a GAN conditioned on text features to synthesis images. AttnGAN [30] proposes an attentional generative network and a deep attentional multimodal similarity model to synthesize fine-grained details at different sub-regions of the image by paying attention to the relevant words in the image description. StackGAN [36] generates images in two stages. The first stage focuses on the rough background, color, and contour components and the second stage focuses on image details. In addition to improving the generation quality, MA-GAN [34] guarantees the generation similarity of related sentences describing the same image by proposing a Single-sentence Generation and Multi-sentence Discrimination (SGMD) module. XMC-GAN [35] uses an attentional self-modulation generator and a contrastive discriminator to enforce text-image correspondence. ManiGAN [15] proposes a text-image affine combination module to select and correlate image regions relevant to given text and a detail correction module to rectify mismatched attributes and completes missing contents.

**Text-guided Image Colorization:** Image colorization aims to generate color of a gray-scale image. To utilize text descriptions to edit the image color, LBIE [4] proposes a recurrent attentive model to fuse image and language features. Text2Colors [6] adopts a text-to-palette generation network and a palette-based colorization network to capture semantic information and produce relevant color palettes. Tag2Pix [12] utilizes the text tag related to the color, and proposes a line art colorization method.

## 3 TEXT-GUIDED IMAGE SUPER-RESOLUTION (TGSR)

This section first analyses the effect of text on SISR, proposes the multimodal fusion network for text-guided SISR, and introduces the training constraints.

### 3.1 Problem Formulation

Given a LR image as well as its text description, the goal of the text-guided super-resolution is reconstructing a HR image with rational textures. Different from the traditional SR, the key to TGSR is reasonably utilizing the text features and giving the SR network more guidance. To make better use of the text guidance, we need to solve two problems: what kind of information text can provide and how such information works.

The text information can be the global description or the object feature of an image. First, text can indicate salient objects in an image and guide the SR network to find these objects and focus on

important image regions. Therefore, we introduce a Text Attention Module to exploit the response of keywords in text to corresponding image regions. Second, text describes visual features (color, shape, etc.) and semantic features (species, gender, age, etc.) of an image. Therefore, we adopt the text-image consistency loss to enforce the semantic consistency between SR results and text descriptions. Thus, the text is able to tell the SR network the significant image feature and is helpful for the network to restore more real textures and details by utilizing the visual relationship between text and image. which ensures the abundance of details

### 3.2 Overview

In Figure 2, the TGSR adopts a coarse-to-fine process through a dual-branch network structure to restore more precise image details. The text features are extracted and embedded in the coarse SR stage to obtain a rough SR result consistent with the text description. The fine SR stage takes the output of the first stage as input to refine the final SR result.

Suppose  $(I, text)$  represents an image-text pair, the text feature  $t$  is extracted from the text description  $text$  of the LR image  $I^{LR}$ . Then, the TGSR network, composed of a text-aware global branch and a refine branch, utilizes the external text guidance and the internal image information from  $t$  and  $I^{LR}$  to restore visually compromising HR images. The text-aware global branch uses the text guidance to generate a rough SR result  $I_G^{HR}$  by incorporating the text feature  $t$  with image features. Based on the  $I_G^{HR}$ , the refine branch focuses on refining photo-realistic details and increases the fidelity of the final HR result  $I^{HR}$ . The TGSR is optimized under constraints at image and semantic levels.

### 3.3 Multi-Modal Fusion SR Network

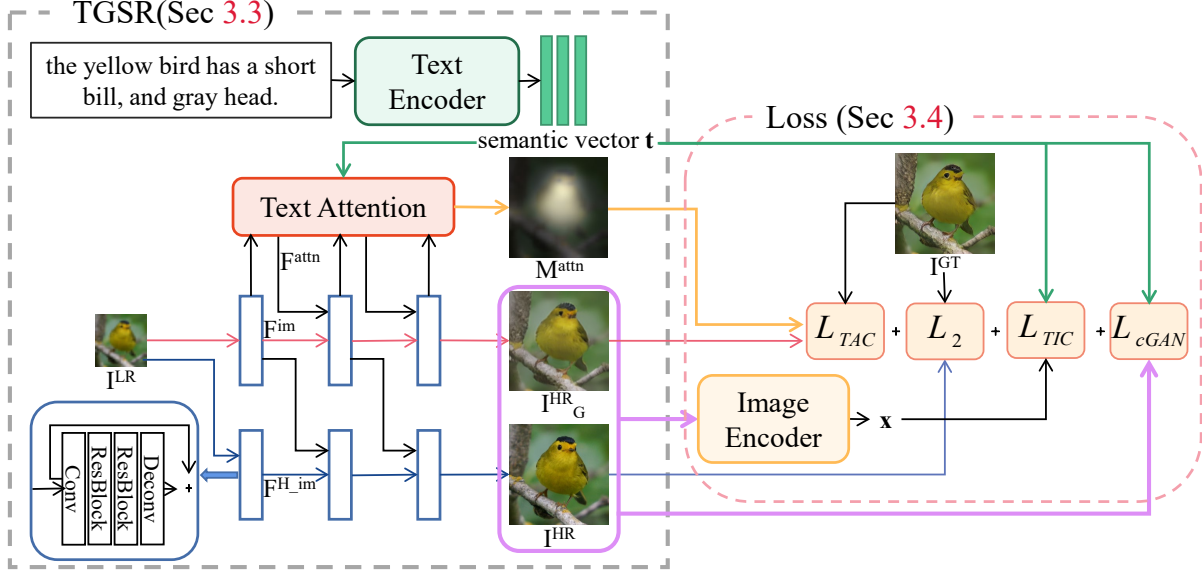
**Text Encoder:** The text encoder chooses the bi-directional Long Short-Term Memory network (LSTM) [27] to extract the text feature, inspired by the AttnGAN [30]. In specific, the text description is transferred to the text feature extractor to extract a preliminary text feature  $t \in R^{D*T}$ , where T is the number of words and D is the dimension of each word vector in the text feature.

$$t = TextEncoder(text). \quad (1)$$

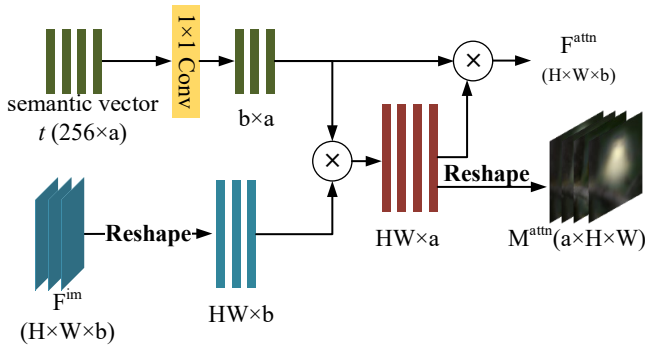
**SR Network:** As shown in Figure 2, the dual-branch TGSR network contains a global branch and a refine branch to generate a coarse SR result and enrich image details, respectively. Both the global branch and the refine branch adopt a pyramid structure to reconstruct  $\times 2$ ,  $\times 4$ ,  $\times 8$  upscaled image features.

In the global branch, we extract and upscale feature  $(F_1^{im})$  from the input LR image with a convolution layer, a residual block, and a deconvolution layer, which are wrapped together by the blue block in Figure 2. The residual block is composed of two convolution layers and a residual connection.

Then, the text attention module (TAM) [30], shown in Figure 3, is adopted to combine the text feature and the image feature. The TAM calculates a word attention map  $M_i^{attn}$  and a text-embedding image feature  $F_i^{attn}$  based on the text feature  $t$  and  $F_i^{im}$  at  $i$ -th stage. With the explainable text attention maps  $M^{attn}$  from the TAM (see Figure 4), the text feature can accurately focus on different image regions and is helpful for the TGSR to well understand the key



**Figure 2: Architecture of the text-guided super-resolution (TGSR) framework, composed of a global branch and a refine branch. The pre-trained text encoder first extracts a semantic vector  $t$  from the text description. The global branch generates a rough HR image  $I_G^{HR}$  by incorporating  $t$  and image features with a text attention module. The refine branch restores more high-frequency details of the final HR image  $I^{HR}$ . The total loss of the TGSR is composed of the L2 loss, the text-attention reconstruction loss, the text-image consistency loss and the conditional GAN loss.**



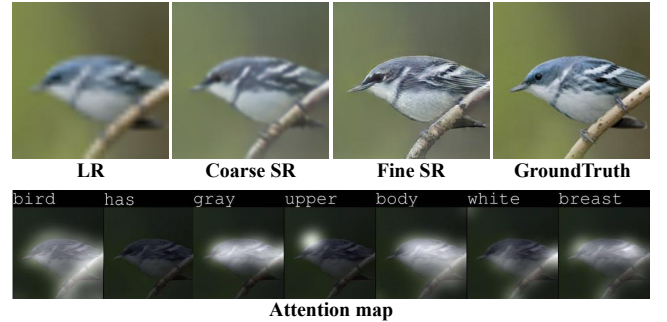
**Figure 3: Structure of the text attention module (TAM). The TAM generates a text-embedding image feature  $F^{attn}$  and an attention map  $M^{attn} \in R^{a \times H \times W}$  corresponding to  $a$  words in text.**

semantic information.

$$M_i^{attn}, F_i^{attn} = TAM(t, F_i^{im}). \quad (2)$$

Next, two image features  $F_i^{attn}$  and  $F_i^{im}$  are concatenated and delivered to a convolution layer, a residual block, and a deconvolution layer to obtain the image feature  $F_{i+1}^{im}$  at next stage. By progressively incorporating text features and image features, we can fully utilize the text guidance. Last, one convolution layer with kernel size  $3 \times 3 \times 3$  is operated on  $F_i^{im}$  to obtain a coarse HR images ( $I_G^{HR}$ ).

This bird has gray upper body and white breast.



**Figure 4: The outputs of the global branch and the refine branch and the word attention map corresponding to each word in the input text description.**

The refine branch takes the LR image  $I^{LR}$  and the image feature  $F_i^{im}$  at different scales of the global branch as input to focus on refining authentic image details. Similar to the global branch, the refine branch extracts the feature of  $I^{LR}$  and progressively upscales it to obtain  $F_i^{H-im}$ .

Then, the feature  $F_i^{H-im}$  at each stage are cascaded with  $F_i^{im}$ . The final HR image ( $I^{HR}$ ) is generated through a convolution layer with kernel size  $3 \times 3$  and 3 filters. As shown in Figure 4, the fine SR result of the refine branch is able to enhance the image details and restore more faithful textures based on the coarse SR result of the global branch.

The convolution layers, used to output images, have kernel size 3 and output channel number 3. Other convolution layers are followed by a Rectified Linear Unit (ReLU) activation layer and have kernel of size 3×3 and 64 filters. Each deconvolution layer has kernel of size 6×6 with stride 2 and 64 filters.

### 3.4 Training Constraints

The goal of our TGSR is simultaneously restoring accurate image details and rational image contents, that are consistent with the semantic information of the text prior. Since it is difficult to optimize the SR network at pixel level and semantic level directly, we separate the final optimization objective into two stages. To enable the network in the coarse SR stage to focus on rough image structure and high-level semantic accuracy, the low-frequency component  $\hat{I}^{GT}$  of the ground-truth HR image  $I^{GT}$  serves as the training label of the global branch, which has complete semantic information but few high-frequency textures. In this way, the refine branch, which is supervised by  $I^{GT}$ , has less training difficulty and can largely improve the pixel accuracy. The  $\hat{I}^{GT}$  is obtained by applying the low-pass filter on  $I^{GT}$ .

The TGSR is trained in an end-to-end way. Suppose  $(I_n, text_n)_{n=1:M}$  denote a batch of image-text pairs. The global branch and the refine branch are jointly optimized with a global constraint and a fine constraint  $(\mathcal{L}_{global}, \mathcal{L}_{fine})$ .

$$\mathcal{L} = \mathcal{L}_{global} + \mathcal{L}_{fine}. \quad (3)$$

As shown in Figure 2, the  $\mathcal{L}_{global}$ , calculated on the  $I_G^{HR}$  of the global branch, contains a L2 reconstruction loss, a text-image consistency loss  $\mathcal{L}_{TIC}$  (DAMSM loss [30]), and a text-adaptive conditional GAN loss  $\mathcal{L}_{cGAN}$ . The  $\mathcal{L}_{fine}$ , calculated on the output  $I^{HR}$  of the refine branch, is composed of a text-attention reconstruction loss  $\mathcal{L}_{TAR}$  in addition to  $\mathcal{L}_{cGAN}$  and  $\mathcal{L}_{TIC}$ .

$$\begin{aligned} \mathcal{L}_{global} &= \lambda_{L2} \|I_G^{HR}, \hat{I}_G^{GT}\|_2 + \lambda_{cGAN} \mathcal{L}_{cGAN}(I_G^{HR}, t) \\ &\quad + \lambda_{TIC} \mathcal{L}_{TIC}(I_G^{HR}, t), \\ \mathcal{L}_{fine} &= \lambda_{TAR} \mathcal{L}_{TAR}(I^{HR}, I^{GT}) + \lambda_{cGAN} \mathcal{L}_{cGAN}(I^{HR}, t) \\ &\quad + \lambda_{TIC} \mathcal{L}_{TIC}(I^{HR}, t), \end{aligned} \quad (4)$$

where  $\lambda_{L2}$ ,  $\lambda_{cGAN}$ ,  $\lambda_{TIC}$ , and  $\lambda_{TAR}$  represent the weights used to balance contributions of different losses.

**The text-adaptive conditional GAN loss:**  $\mathcal{L}_{cGAN}$  utilizes the conditional GAN, where the conditional discriminator  $cD(\cdot)$  judges the realism of image  $I(I^{HR}, I_G^{HR})$  conditioned on the text feature  $t$ .

$$\mathcal{L}_{cGAN}(I, t) = \mathcal{E}[\log(cD(I|t))]. \quad (5)$$

**The text-image consistency loss:**  $\mathcal{L}_{TIC}$  constrains semantic consistency between an image and its text description. AttnGAN [30] proposes the deep attentional multimodal similarity module to map image and text into a common space and to calculate similarity between them. We introduce this module into our TGSR network to calculate the text-image matching score (TIM) and the  $\mathcal{L}_{TIC}$ .

As described before, the text encoder maps a text description into the semantic vector  $t$ . Then, we map the super-resolved image  $I(I^{HR}, I_G^{HR})$  into the image feature  $x \in 768 \times 17 \times 17$  with an image feature extractor (Inception-v3 [29]). The  $x$  is reshaped to  $R^{768 \times 289}$ , and  $x_i \in R^{768}$  denotes the visual feature vector for  $i_{th}$  image region.

dataset	CUB		Oxford-102		CelebA		COCO	
	Train	Test	Train	Test	Train	Test	Train	Test
Images	8855	2933	7,034	1,155	162,770	39,829	80,000	40,000
Text per image	10	10	10	10	10	10	5	5

Table 1: The datasets statistics.

To evaluate the semantic accuracy of the SR results, the TIM measures the matching degree between an image and its text description by the following calculations:

First, we generate and normalize the similarity matrix,  $s \in R^{T \times 289}$ , where  $s_{i,j}$  is similarity between  $i_{th}$  word and  $j_{th}$  image sub-region).

$$\begin{aligned} s &= t^T x, \\ s'_{i,j} &= \exp(s_{i,j}) / \sum_{k=0:T-1} \exp(s_{k,j}). \end{aligned} \quad (6)$$

Then, we obtain the region-context vector  $c$ , where  $c_i$  denotes the relation between  $i_{th}$  word and image subregions.

$$\begin{aligned} a_j &= \exp(s'_{i,j}) / \sum_{k=0:288} \exp(s'_{i,k}), \\ c_i &= \sum_{j=0:288} a_j x_j. \end{aligned} \quad (7)$$

Last, the relevance  $R(c_i, t_i)$  between  $i_{th}$  word and image is generated to obtain the TIM  $(I_n, text_n)$ .

$$\begin{aligned} R(c_i, t_i) &= (c_i^T t_i) / (\|c_i\| \cdot \|t_i\|), \\ TIM(I_n, text_n) &= \log(\sum_{i=1:T-1} \exp(\cdot R(c_i, t_i))). \end{aligned} \quad (8)$$

The  $\mathcal{L}_{TIC}$  is the negative log posterior probability between the image and text.

$$\begin{aligned} P(text_n | I_n) &= \frac{\exp(R(c_{n_i}, t_{n_i}))}{\sum_{j=1:M} \exp(R(c_{n_j}, t_{n_j}))}; \\ \mathcal{L}_{TIC} &= -(\sum_{n=1:M} \log P(text_n | I_n) \\ &\quad + \sum_{n=1:M} \log P(I_n | text_n)). \end{aligned} \quad (9)$$

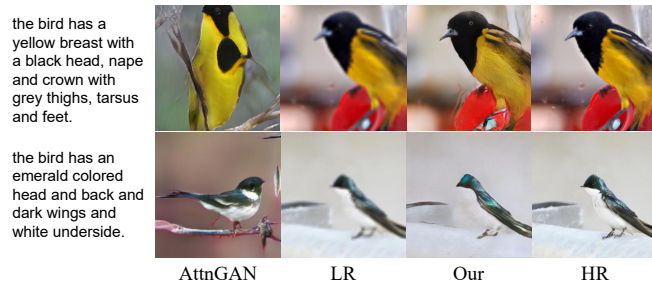
**The text-attention reconstruction loss:**  $\mathcal{L}_{TAR}$  can be considered as a weighted L2 loss, which calculates the pixel-level similarity between the ground truth  $I^{GT}$  and the super-resolved image  $I^{HR}$ . The  $\mathcal{L}_{TAR}$  assigns pixels different weights based on five attention maps with the largest activation values in  $M^{attn}$ , that corresponding to attention maps of five keywords, to make the network focus on the accuracy of pixels in visually important regions.

$$\mathcal{L}_{TAR} = \frac{1}{5} \sum_{i=1:5} M_i^{attn} \|I^{HR}, I^{GT}\|_2. \quad (10)$$

The text feature encoder and the image feature encoder are pre-trained following the methods [30] and keep fixed when training the TGSR model.

## 4 EXPERIMENTS AND ANALYSIS

In this section, we first introduce the datasets and implementation details. Then, we conduct the ablation study to evaluate the contributions of different designs and compare the proposed TGSR with state-of-the-art SR methods. At last, the diversity of SR results is demonstrated by manipulating the text descriptions.



**Figure 5: A visual comparison to the text-to-image synthesis model AttnGAN [30].**

	baseline	+ TAM	+ $\mathcal{L}_{TIC}$	+ coarse-to-fine	+ $\mathcal{L}_{TAR}$
NIQE↓	10.801	4.657	4.650	3.424	<b>3.179</b>
PI↓	9.048	3.513	3.840	3.239	<b>2.663</b>
TIM↑	1.453	2.643	<b>3.641</b>	3.501	3.553

**Table 2: Experimental results of ablation study. The average NIQE/PI/TIM on CUB for  $\times 8$  SR. TIM shows the semantic accuracy of images. ↓ denotes the lower is the better.**

#### 4.1 Datasets and Protocols

We train and test our models on Caltech-UCSD Birds 200 (CUB) [31], Oxford-102 [25], CelebA [20] and COCO [18] datasets, where all images are annotated with several natural language captions. The details of datasets are listed in Table 1. All images are resized and cropped into patches of size 256. To train SR models, LR images are obtained by downscaling HR images with bicubic interpolation.

All models are trained on the machine with 2.20 GHz Intel (R) Xeon (R) CPU, and GTX1080Ti GPU (128G RAM). The initial learning rate is set to  $1e-4$ . We adopt Adam optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 1e-8$ . The loss weights  $\lambda_{L2}$ ,  $\lambda_{cGAN}$ ,  $\lambda_{TIC}$ , and  $\lambda_{TAR}$  are set as 1, 0.1, 0.5, and 1 respectively.

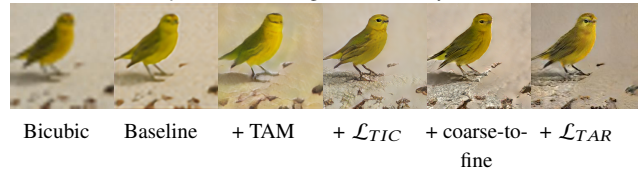
The PSNR and SSIM [37] measure the distortion degree of images and ignore the subjective quality. As stated in [9], NIQE [24] and Perceptual Index (PI) (the lower, the better) cannot fairly reflect the subjective performance, since they cannot distinguish the GAN generated noises and real details and prefer images with obvious unrealistic artifacts produced by GAN-based methods. Since our TGSR aims to output better visual results with accurate semantic features, we choose the R-precision (TIM), which is used to measure the consistency between images and text descriptions, by following [16, 30] to evaluate the recovery degree of the semantic information in the super-resolved image according to the text.

#### 4.2 Ablation Study

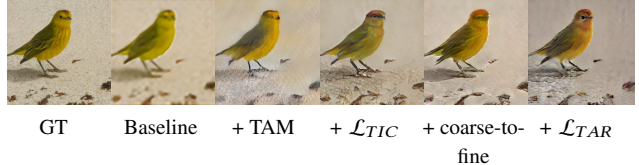
To verify the effect of text to SR, we first evaluate models with different inputs, including text, LR image, and the combination of text and LR image. Then, we construct different models to illustrate contributions of different designs.

**Text only.** Figure 5 compares our model with the text-to-image synthesis model, AttnGAN [30]. The AttnGAN uses the same text descriptions as ours to generate images, as we can observe that the

A small bird with a **yellow** head, a light brown and yellow throat.



A small bird with a **red** head, a light brown and yellow throat.



**Figure 6: A visual comparison of ablation study. The upper and lower lines denote two SR results of different models on one LR image. Different results are obtained by inputting two texts with few changes to the same model.**

output image of the AttnGAN has accurate semantic features but lacks pixel-wise accuracy.

**LR to HR.** To verify the contribution of text to SR, we first remove the text input from our model and construct a baseline, which is similar to traditional SISR models and only requires LR images as input, by cascading several convolution layers, residual blocks and deconvolution layers. The baseline model is trained with the L2 loss only. As shown in Table 2, the NIQE/PI/TIM scores of outputs of the baseline model are lower than others. In addition, as shown in Figure 6, the baseline can not use the external text information and only outputs a single SR result for a LR image input.

**Text and LR to HR.** Then, we introduce the text input into the baseline SR model with the proposed TAM modules. Other ablation studies are also conducted by constructing models with different configurations to analyze the effectiveness of different designs of the TGSR, including the coarse-to-fine structure, the  $\mathcal{L}_{TIC}$  and  $\mathcal{L}_{TAR}$ , by progressively added them on the baseline.

Figure 6 denotes different SR results of one LR image with different text inputs. The text description has no effect on the baseline. To activate the text guidance, ‘+ TAM’ adds the text attention module (TAM) after all deconvolution layers of the baseline. In Table 2, the model ‘+ TAM’ outputs images with better perceptual quality (NIQE/PI) and semantic accuracy (TIM). From Figure 6, we see that image details of ‘+ TAM’ can be slightly changed by altering keywords in text descriptions. As expected, employing the TAM enables model to generate more textures related to text descriptions and to enhance the TIM score. Though SR results can correctly correspond to text descriptions, some real details are missing.

After employing the text-image consistency loss ( $\mathcal{L}_{TIC}$ ) on the ‘+ TAM’ model, we obtain a model ‘+  $\mathcal{L}_{TIC}$ ’. In Table 2, the ‘+  $\mathcal{L}_{TIC}$ ’ has similar NIQE/PI as the ‘+ TAM’, but higher TIM, which illustrates that the loss  $\mathcal{L}_{TIC}$  can improve the semantic accuracy of images relative to text. From the Figure 6, the ‘+  $\mathcal{L}_{TIC}$ ’ restores more realistic bird textures.

To refine image details, the model ‘+ coarse-to-fine’ introduces the coarse-to-fine structure by adding the refine branch based on the ‘+  $\mathcal{L}_{TIC}$ ’ and regarding the ‘+ TAM’ as the global branch. Lower



Figure 7: Visual comparison of different SR methods (EDSR [17], SPSR [22], and ours) on scale 4, 8, 16.

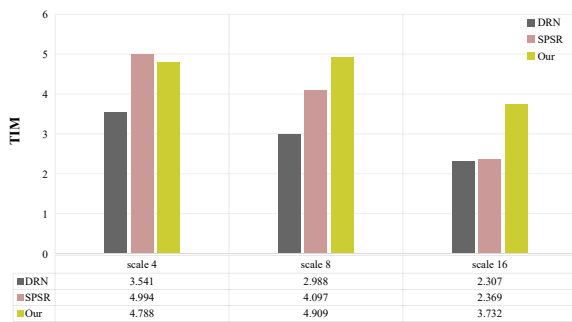


Figure 8: TIM variation of different SR methods on different scales. TIM scores reflect amounts of semantic information contained in SR images. The proposed TGSR keeps a well stability on different scales in terms of semantic accuracy.

metrics	Bicubic	SuperFAN [3]	DICGAN [21]	Ours	GroundTruth
TIM↑	0.049	0.013	0.378	<b>0.885</b>	0.815
PSNR↑	25.81	24.85	27.42	23.48	-
SSIM↑	0.844	0.859	0.862	0.766	-
NIQE↓	14.514	10.347	5.657	8.8464	7.473
PI↓	9.676	8.606	5.002	7.165	5.256
FID↓	105.232	120.666	92.901	93.919	-

Table 3: Quantitative comparison with state-of-the-art face SR methods on the CelebA dataset.

NIQE/PI illustrates the refine branch can effectively promote visual SR results. The final model ‘+  $\mathcal{L}_{TAR}$ ’, which further employs the text-attention reconstruction loss  $\mathcal{L}_{TAR}$ , obtains the best NIQE/PI and restores more authentic textures. As shown in Figure 6, the results of the ‘+  $\mathcal{L}_{TAR}$ ’ have abundant and photo-realistic image details, such as the head and feather of the bird.

### 4.3 Comparison with State-of-the-arts

As we stated before, the network structure is not our main contribution, so we compare with several representative MSE-based and GAN-based SISR and Face SR methods, including EDSR [17], ESRGAN [33], SPSR [22], SuperFAN [3], DICGAN [21], that have public codes and models. Note that most SR methods only provide  $\times 4$  models, we retrain  $\times 8$  SR models of the EDSR and ESRGAN

Dataset	metrics	Bicubic	EDSR	ESRGAN	SPSR	Ours	GroundTruth
CUB	TIM↑	0.920	1.090	2.482	2.045	<b>2.841</b>	3.189
	NIQE↓	12.374	10.684	<b>5.465</b>	5.885	6.623	6.734
	PI↓	9.747	8.168	2.644	3.345	<b>2.560</b>	2.302
COCO	TIM↑	4.967	5.708	6.280	6.650	<b>7.353</b>	7.649
	NIQE↓	11.110	9.683	6.816	6.378	<b>6.4844</b>	3.840
	PI↓	9.373	8.515	7.135	6.060	<b>4.922</b>	3.657

Table 4: Comparison with state-of-the-arts on CUB and COCO datasets.

Dataset	scale	metrics	Bicubic	EDSR	ESRGAN	SPSR	Ours	GroundTruth
Oxford-102	$\times 8$	TIM↑	2.446	3.541	3.980	4.097	<b>4.778</b>	5.112
		NIQE↓	11.089	9.405	4.860	<b>4.465</b>	5.282	4.506
		PI↓	10.333	8.564	4.183	<b>3.221</b>	5.100	4.688
Oxford-102	$\times 16$	TIM↑	1.039	2.307	2.108	2.369	<b>3.732</b>	5.112
		NIQE↓	13.495	11.685	7.541	5.993	<b>4.506</b>	4.159
		PI↓	12.076	9.376	6.012	4.688	<b>3.654</b>	2.476

Table 5: Quantitative comparison with state-of-the-arts on Oxford-102 dataset.

The bird has **brown** crown, **striped** feathers.



A bird has **orange** belly and a **small black** bill.



**Zebras** grazing in a field.

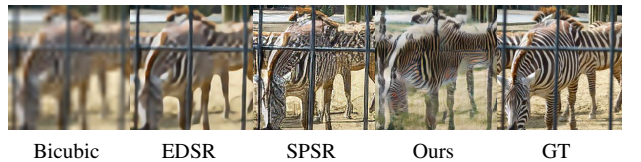


Figure 9: Comparison with SOTA on CUB and COCO.

on our training set for fair comparison by adding an additional  $\times 2$  upsampling layer at the end of their original models.

**Stability to different scales:** The visual performance and the quantitative TIM variation of different SR methods at different scales are demonstrated in Figure 7 and Figure 8. We can observe that the performance drop is obvious for existing SR methods, facing larger scale factors. In Figure 7, the compared SR methods (DRN and SPSR) either generate blur textures or fake artifacts for scale 8 and 16. The proposed TGSR can generate clear image details and accurate semantic information. In Figure 8, the TIM results of DRN and SPSR decrease significantly accompanied by increasing scale factors. For a small scale, the LR image can provide enough information to restore semantically accurate images, therefore these SR models can well deal with factor by only using pixel-wise constraint and

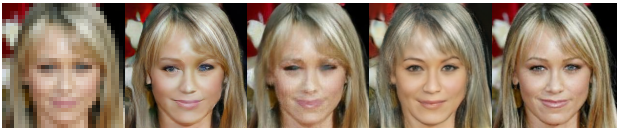
The **young** man is smiling.



The **double chined** man has high cheekbones.



She has straight hair which is **blond**.



LR SuperFAN DCGAN Ours GT

**Figure 10: Visual comparison with state-of-the-art face SR methods on the CelebA. The text inputs are above images.**

generate images, that are more consistent with the text descriptions and have higher TIM scores. For larger scale, these SR methods become less efficiency for LR image fail to provide enough information, which leads to worse visual quality and text-image matching score. In comparison, the proposed TGSR keeps stable TIM score, for it can restore the intelligibility of an image with accurate semantic information even for a large scale. The proposed TGSR has an obvious advantage for large scale factors.

Above experiments demonstrate that traditional single-modal SISR methods are significantly influenced by scale factors as they are trained to decrease the pixel-wise reconstruction error, and only utilize the spatial information from the LR image, and the further improvement of perceptual quality is restricted.

**Quantitative metrics:** Table 3, 4, 5 report quantitative comparisons with state-of-the-art SISR methods and face SR methods. In these Tables, NIQE/PI of some GAN-based SR methods are lower than that of HR images, which goes against the common sense and demonstrates NIQE/PI cannot fairly evaluate GAN-based image restoration algorithms, as stated in [9]. The PSNR/SSIM are not appropriate to evaluate SR images in terms of the human perception. Our approach makes significant progress on TIM and obtains NIQE/PI closer to the GT.

**Visual results:** Figure 1, 9, and 10 demonstrate the subjective SR results. It is difficult for traditional SR networks to predict accurate HR contents by only referring to the finite pixel information in tiny LR images. In comparison, the proposed TGSR exploits the text guidance to restore better visual results and faithful textures (e.g., the feathers of birds) that are consistent with text descriptions.

#### 4.4 Editable Super-Resolution

We present one interesting application of TGSR, a text-guided editable image reconstruction, where users can manually edit SR results based on the text input. The possibility to tweak a model’s output is



Bicubic arched eyebrows old man GT



Bicubic high cheekbones attractive, pale skin GT



Bicubic Zebras Giraffes GT



Bicubic white head. brown head. GT

**Figure 11: Examples of manipulating details in SR results with different keywords in text descriptions.**

important for making super-resolution more useful and controllable. As shown in Figure 11, the proposed text guided SR model is able to flexibly control image details, such as skin, texture, age, and generates different HR image contents based on different keywords in text descriptions.

**Limitation:** The main contribution of this paper is introducing the text descriptions in SISR. We observe some limitations in our work. First, since text descriptions in existing datasets may not cover enough details, the text becomes invalid sometimes. Besides, text may correspond to multi visual representations, which influences fidelity of the final result.

## 5 CONCLUSION

This paper aims to solve the larger factor super-resolution by regarding the single image super-resolution problem as a text-guided detail generation problem, and proposes a multimodal text-guided super-resolution algorithm, which restores realistic and rational image details by utilizing the guidance of text descriptions in a coarse-to-fine process. The text-guided losses enable the network to focus on objects concentrated by the text descriptions, and keep the consistency between the super-resolved image and the corresponding text description. Experimental results demonstrate that the proposed approach can flexibly incorporate the text prior to facilitate the rich-detail super-resolution, which is practical in reality.



## REFERENCES

- [1] Yuval Bahat and Tomer Michaeli. 2020. Explorable Super Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [2] Marcel Christoph Bühler, Andrés Romero, and Radu Timofte. 2020. DeepSEE: Deep Disentangled Semantic Explorative Extreme Super-Resolution.
- [3] Adrian Bulat and Georgios Tzimiropoulos. 2018. Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [4] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-Based Image Editing with Recurrent Attentive Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [5] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. 2018. FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2492–2501.
- [6] Wonwoong Cho, Hyojin Bahng, David Keetae Park, Seungjoo Yoo, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. 2018. Text2Colors: Guiding Image Colorization through Text-Driven Palette Generation. In *European Conference on Computer Vision (ECCV)*.
- [7] C. Dong, C. C. Loy, K. He, and X. Tang. 2016. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (Feb 2016), 295–307.
- [8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 3 (2014), 2672–2680.
- [9] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. 2020. PIPAL: a Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. In *European Conference on Computer Vision (ECCV)*.
- [10] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. 2020. Learning Meta Face Recognition in Unseen Domains. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [11] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhong Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. 2020. Closed-loop Matters: Dual Regression Networks for Single Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [12] Hyunsu Kim, Ho Young Jho, Eunhyeok Park, and Sungjoo Yoo. 2019. Tag2Pix: Line Art Colorization Using Text Tag With SECat and Changing Loss. In *IEEE International Conference on Computer Vision (ICCV)*.
- [13] J. Kim, J. K. Lee, and K. M. Lee. 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1646–1654.
- [14] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4681–4690.
- [15] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020. Manigan: Text-guided image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7880–7889.
- [16] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019. Object-driven Text-to-Image Synthesis via Adversarial Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 4.
- [18] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*.
- [19] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. 2020. Residual Feature Aggregation Network for Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [20] Z. Liu, P. Luo, X. Wang, and X. Tang. 2015. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*. 3730–3738. <https://doi.org/10.1109/ICCV.2015.425>
- [21] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. 2020. Deep Face Super-Resolution with Iterative Collaboration between Attentive Recovery and Landmark Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [22] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. 2020. Structure-Preserving Super Resolution with Gradient Guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [23] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [24] A. Mittal, R. Soundararajan, and A. C. Bovik. 2013. Making a “Completely Blind” Image Quality Analyzer. *IEEE Signal Processing Letters* 20, 3 (2013), 209–212.
- [25] M-E. Nilsback and A. Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*.
- [26] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text-to-Image Synthesis. In *International Conference on Machine Learning (ICML)*.
- [27] M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681. <https://doi.org/10.1109/78.650093>
- [28] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. 2018. Deep Semantic Face Deblurring. *CoRR* abs/1803.03345 (2018).
- [29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [30] Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaoqi Huang, Xiaodong He, Tao Xu, Pengchuan Zhang. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report CNS-TR-2011-001, California Institute of Technology.
- [32] X. Wang, K. Yu, C. Dong, and C. Change Loy. 2018. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 606–615. <https://doi.org/10.1109/CVPR.2018.00070>
- [33] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *European Conference on Computer Vision (ECCV)*.
- [34] Yanhua Yang, Lei Wang, De Xie, Cheng Deng, and Dacheng Tao. 2021. Multi-Sentence Auxiliary Adversarial Networks for Fine-Grained Text-to-Image Synthesis. *IEEE Transactions on Image Processing* 30 (2021), 2798–2809. <https://doi.org/10.1109/TIP.2021.3055062>
- [35] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [36] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaoqi Huang, and Dimitris Metaxas. 2017. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*.
- [37] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.