Content-Aware Local GAN for Photo-Realistic Super-Resolution

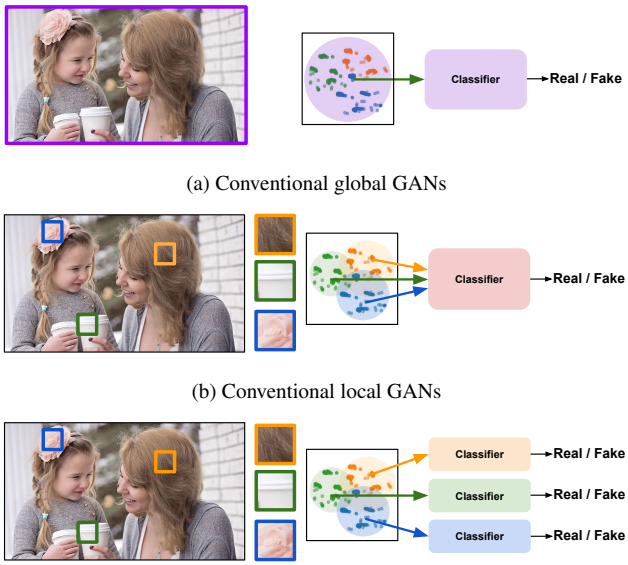
JoonKyu Park[1] Sanghyun Son[1] Kyoung Mu Lee[1,2]
[1]Dept. of ECE&ASRI, [2]IPAI, Seoul National University, Korea
{jkpark0825, thstkdgus35, kyoungmu}@snu.ac.kr

Abstract

Recently, GAN has successfully contributed to making single-image super-resolution (SISR) methods produce more realistic images. However, natural images have complex distribution in the real world, and a single classifier in the discriminator may not have enough capacity to classify real and fake samples, making the preceding SR network generate unpleasing noise and artifacts. To solve the problem, we propose a novel content-aware local GAN framework, CAL-GAN, which processes a large and complicated distribution of real-world images by dividing them into smaller subsets based on similar contents. Our mixture of classifiers (MoC) design allocates different super-resolved patches to corresponding expert classifiers. Additionally, we introduce novel routing and orthogonality loss terms so that different classifiers can handle various contents and learn separable features. By feeding similar distributions into the corresponding specialized classifiers, CAL-GAN enhances the representation power of existing super-resolution models, achieving state-of-the-art perceptual performance on standard benchmarks and real-world images without modifying the generator-side architecture. The codes are available at https://github.com/jkpark0825/CAL_GAN.

1. Introduction

As one of the long-standing challenges in computer vision, single image super-resolution (SISR, or SR in short) aims to reconstruct a super-resolved image $\mathbf{I}_{\text{SR}}$ from its low-resolution (LR) counterpart $\mathbf{I}_{\text{LR}}$. While the inverse problem has multiple possible solutions for a given input due to the ill-posedness, the practical goal of SISR is to generate a perceptually favorable result from the LR input. To this end, earlier approaches [8, 24, 30] have constructed various network architectures specialized for SR and optimized pixel-wise L_2 or L_1 loss terms. Later, perceptual objective functions [22] and adversarial training frameworks [13, 26, 50] are introduced to generate more realistic textures and improve the realism of SR results.



(a) Conventional global GANs

(b) Conventional local GANs

(c) Our CAL-GAN with mixture of classifiers (MoC)

Figure 1: **Overall framework of our CAL-GAN.** Unlike the approaches mentioned in (a) and (b), our CAL-GAN in (c) groups local contents with similar properties and feeds them to their corresponding specialized discriminators.

Previous perceptual SR methods jointly optimize the SR network and discriminator through adversarial training, while the discriminator is responsible for distinguishing real and generated, *i.e.*, fake, SR samples. However, as shown in Figure 1a, the majority of methods [26, 50, 32, 57, 29, 48] adopt a global discriminator that is not specialized to consider the importance of local distribution in SR tasks. To improve local details, recent approaches [48, 51, 6] have incorporated local discriminators, which aim to differentiate between each local texture in HR and SR. However, such discriminators are regarded as content-blind, indicating that they cannot learn the various properties of the local content. Specifically, they naïvely try to learn the whole distributions from each local content using a single discriminator, as shown in Figure 1b. Thus, they struggle to learn the complex distribution of natural images, resulting in sub-optimal solutions. As adversarial training benefits from

finding equilibrium between the generator and discriminator, such a sub-optimality may negatively affect the generator side. Also, in previous approaches, real and fake samples are grouped without considering the diverse properties of different images. For example, separating real and fake tree images would be a nice training example for perceptual SR models, whereas classifying fake tree and real sky may not provide informative gradients. Likewise, as shown in Figure 1b, existing methods enforce the distribution of flowers, cups, and hair images to be learned with a single classifier. Therefore, existing perceptual SR methods [50, 45, 38] usually generate unpleasant artifacts in super-resolved images.

One possible option to alleviate such limitations is to introduce semantic labels [49, 39, 52] to the discriminator side, assuming that patches in the same class have common characteristics. Nevertheless, even though samples from the same category may share specific properties, their low-level statistics can vary. For example, wheels, windows, and bonnets are all labeled as a 'car,' but their contents differ significantly. Another practical issue is that determining the plausible number of semantic classes is infeasible. Natural images contain numerous types of objects, and sometimes it is not trivial to define an exact label, *e.g.*, background, for all possible patches for learning.

To overcome the aforementioned challenges, we propose Content-Aware Local GAN, CAL-GAN, a novel adversarial training framework for handling complicated natural image distributions by their contents. Instead of using a single classifier to learn the complex manifold of natural images, we divide it into smaller subsets according to image contents. As shown in Figure 1c, our mixture of classifiers (MoC) effectively handles much smaller manifolds than previous approaches where features in each group share common properties. Furthermore, we propose an efficient routing module and learning strategy to implicitly determine an appropriate subset of each feature rather than assigning them manually. Combined with orthogonal regularization [21, 54, 55] and LDA-based [36, 10] techniques, we can significantly improve the separability of the learned features for accurate distribution learning.

Our CAL-GAN is applicable to existing SR frameworks without modifications. Thus, we demonstrate that CAL-GAN improves the performance of state-of-the-art SR methods [30, 50, 28] to reconstruct more photo-realistic results on diverse benchmarks. Extensive ablation studies also validate that the design principle of CAL-GAN helps SR models to generate perceptually fine details and textures. Our several-fold contributions can be organized as follows:

- We present a powerful adversarial training framework that utilizes multiple expertized classifiers. A proper patch-wise routing policy can be learned during training with the proposed balancing loss.
- We adopt orthogonality constrain and distribution-based inter/intra-class loss terms to minimize correlation between different discriminators to implement an expert system.
- Our CAL-GAN achieves superior SR performance on various benchmarks and real-world images, achieving high perceptual metrics and visual quality.

2. Related Works

Deep learning-based SISR. Starting from SRCNN [8], early methods [24, 30] have implemented very deep architectures with residual structures to increase the number of learnable parameters. After then, numerous novel building blocks and architectures have been proposed [62, 61, 11, 43] to enhance the performance of the SISR. Specifically, RCAN [61] adopts residual channel attention network, and RDN [62] leverages global feature fusion to fully utilize hierarchical features. Stepping forward, SCN [11] proposes scale-wise convolution to aggregate multi-scale features locally and gradually. Motivated by recent implicit neural representation techniques [42], SR models handle irregular and asymmetric scaling factors [5] as well as arbitrary transformations [43]. Furthermore, motivated by recent success in attention mechanisms, [28, 56] adopts Transformer to enhance super-resolution performance. On the other hand, SRDD [33] explicitly learns a deep dictionary of high-resolution to make the network robust for out-of-domain test images. Nevertheless, the aforementioned methods do not consider the perceptual quality of the super-resolved images, but focus on achieving high PSNR and SSIM only.

Photo-Realistic Super-Resolution. Instead of minimizing pixel-wise reconstruction error, recent studies try to model the perceptual convenience by adversarial training [13]. Since SRGAN [26] first employ GAN framework with VGG-based perceptual loss function, [50, 45, 38, 53, 4, 32, 60, 29] suggested adversarial learning strategy by modifying GAN framework. ESRGAN [50] introduced a relativistic discriminator, while SRResCycGAN++ [45] utilized cycle consistency [64] to match the domain consistency between low-resolution and high-resolution images. DGP [38] adopted a progressive approach to the layers of the generator and discriminators, and DASR [53] utilized wavelet transform to feed high-frequency components into the GAN network. Stepping forward, SPSR [32] proposed gradient loss to guide the model with structural information, and LDL [29] explicitly discriminated visual artifacts from realistic details to regularize the GAN training framework. In addition, SFT-GAN [49] and SROBB [39] used explicit semantic priors based on segmentation.

On the other hand, a few methods [35, 58] utilize different approaches for photo-realistic super-resolution. Focusing on that outputs of the SR network should be down-

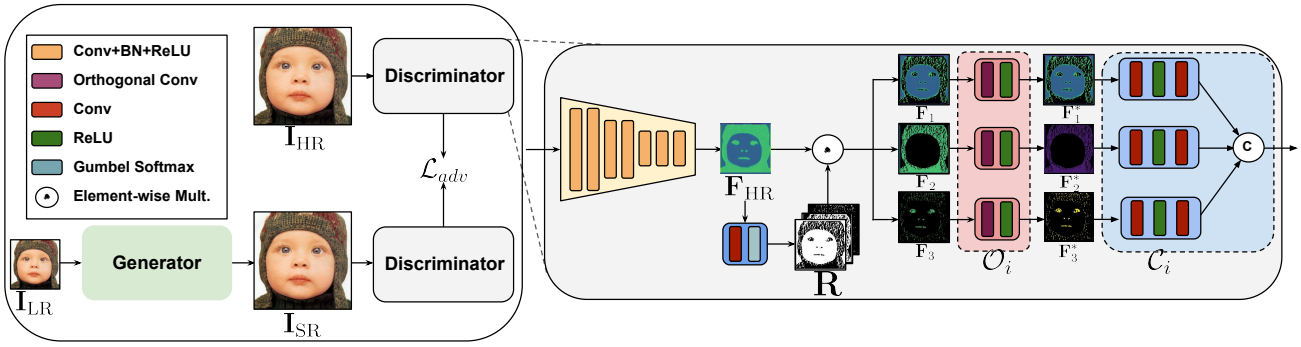


Figure 2: **The overall architecture of Content-Aware Local GAN (CAL-GAN).** Different from the previous GAN, router network $\mathcal{R}$ allocates features to N-many classes. For convenience, we visualize when $N = 3$. Then, the corresponding orthogonal convolution $\mathcal{O}_i$ and mixture of classifiers $\mathcal{C}_i$ distinguish real and fake to form the adversarial loss $\mathcal{L}_{adv}$. Note that routing mask $\mathbf{R}$ is only obtained from $\mathbf{I}_{\text{HR}}$. © denotes concatenation.

scaled to the correct LR inputs, PULSE [35] maps recovered images to the realistic images and downscale domain using GAN. And BSRGAN [58] proposes augmentation techniques by randomly shuffling degradation sequences, imitating the real ISP process. Despite these advancements, visual artifacts from adversarial training still remain problematic in SR tasks.

Mixture of experts. Rather than relying on a single model to process complicated large-scale data, distributing the workload to multiple workers can be a more effective strategy. From this motivation, Jacob *et al*. [19] first adopt a gating strategy to divide information between different models. Since each model observes only a part of training data, it is referred to as an *expert*, and the entire system becomes a mixture of experts (MoE). Recent attempts [41, 63, 12] have demonstrated the superiority of MoE in deep learning, while there are two main shortcomings when combining multiple experts: limited computational resources and stability. Without advanced techniques such as expert capacity [12] or parallelization [14], the processing time of conventional systems is proportional to the number of experts. Also, traditional routing strategy [41] can results in unstable training of the MoE system when appropriate regularization is not used. In this paper, we also employ the concept of MoE in the GAN framework during training. By using the novel loss functions, features can be regularized and effectively routed without instability.

3. Method

Our goal is to reconstruct a photo-realistic SR result $\mathbf{I}_{\text{SR}} \in \mathbb{R}^{H\times W\times 3}$ from a given LR image $\mathbf{I}_{\text{LR}} \in \mathbb{R}^{h\times w\times 3}$. Here, (h, w) and $(H, W) = (sh, sw)$ represent (height, width) of LR and SR images under a fixed scaling factor of s, respectively. To this end, we adopt adversarial training framework [13, 26] and perceptual loss functions [50, 29] to reconstruct photo-realistic outputs. Our training objective for the SR model $\mathcal{L}_{\text{SR}}$ is defined as follows:

$$\mathcal{L}_{\text{SR}} = \lambda_{\text{r}}\mathcal{L}_{\text{recon}} + \lambda_{\text{p}}\mathcal{L}_{\text{per}} + \lambda_{\text{g}}\mathcal{L}_{\text{gen}}, \tag{1}$$

where $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{per}}$, $\mathcal{L}_{\text{gen}}$ refer to pixel-wise L_1 reconstruction loss [30], perceptual loss [1], and adversarial loss [23] from the following discriminator, respectively. Hyperparameters λ_{r}, λ_{p}, and λ_{g} determine weights of each loss term. In the adversarial training framework, SR and discriminator networks are optimized alternately [26, 50].

3.1. Content-aware local discriminator

While previous perceptual SR methods have achieved significant successes, one concern exists regarding their discriminator architectures. Typically, those discriminators have a single classifier [50, 26] that has to classify the realism of complex natural images, without considering the complexity of natural image distributions. Such a formulation may negatively impact the model's ability to classify real and fake samples and further decrease the performance of the preceding SR network. Therefore, we propose Content-Aware Local GAN, CAL-GAN, to alleviate the burden of the classifier and improve the overall SR framework. Specifically, we divide complex distributions into several subsets based on image contents and employ multiple expert classifiers. As illustrated in Figure 2, each expert is specialized to the corresponding subsets.

3.2. Dynamic routing in discriminator

To formulate our Content-Aware Local discriminator $\mathcal{D}$, we first apply a sequence of Conv-BN-ReLU to extract features $\mathbf{F}_{\text{HR}}, \mathbf{F}_{\text{SR}} \in \mathbb{R}^{H/4\times W/4\times c}$ from $\mathbf{I}_{\text{HR}}$ and $\mathbf{I}_{\text{SR}}$, respectively. Here, $c = 512$ denotes the number of channels. Then, we propose to adopt multiple classifiers in discriminator $\mathcal{D}$, which are specialized in different content distributions as in Figure 2. To allocate different features to N-many classifiers, we define a router network $\mathcal{R}$. The module

is responsible for deciding which classifier processes a spatial feature $\mathbf{F}$. For a given feature $\mathbf{F}$, Our router is designed to produce a point-wise one-hot vector $\mathbf{R} \in \mathbb{R}^{H/4 \times W/4 \times N}$ and N binary masks $\mathbf{R}_i \in \mathbb{R}^{H/4 \times W/4}$ by following:

$$\begin{aligned} \mathbf{R} &= \texttt{gumbel\_softmax}\left(\text{Conv}_{1\times1}\left(\mathbf{F}\right)\right), \\ \mathbf{R}_i\left[\frac{H}{4}, \frac{W}{4}\right] &= \begin{cases} 1, & \text{if } \mathbf{R}\left[\frac{H}{4}, \frac{W}{4}, i\right] = 1. \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{2}$$

where Gumbel-Softmax [20] is applied across the channel dimension to enable differentiation of the one-hot vector $\mathbf{R}$. The 1×1 convolution is optimized during training.

Using the predicted routing mask $\mathbf{R}$, we split the spatial feature $\mathbf{F}_{\text{HR}}$ to N disjoint domain-specific features $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_N\}$ as follows:

$$\mathbf{F}_i = \mathbf{F}_{\text{HR}} \odot \mathbf{R}_i, \tag{3}$$

where $\odot$ denotes element-wise multiplication. We note that $\mathbf{R}_i$ is broadcasted across channel dimension c when (3) is calculated.

Balancing the routing. Although feature $\mathbf{F}$ is spatially divided into N subsets $\mathbf{F}_i$, dynamic routing without any constrain can cause load-imbalance problem [27, 40, 63] which assigns feature disproportionately. As load-imbalance leads specific subsets of features to contain a plethora or scarce information, we introduce balancing loss $\mathcal{L}_{\text{b}}$ to alleviate the issue [40, 27, 12].

$$\mathcal{L}_{\text{b}} = \frac{N}{M} \sum_{i,j} \max_k \mathbf{R}'\left[i, j, k\right], \tag{4}$$

where $\mathbf{R}'$ is obtained by replacing `gumbel_softmax` in (2) with a standard Softmax function, and $M = \frac{HW}{16}$ represents the number of spatial pixels. The balancing loss decreases when the maximum value of Softmax distribution $\mathbf{R}'$ becomes smaller, and probability values are evenly distributed so that optimizing the loss enforces uniform routing across N subsets. However, as uniform routing produces lower maximum values when bigger N is used, we scale the loss to compensate for situations where the maximum value is too small.

3.3. Meditating class property

So far, we have uniformly disentangled N-many subsets $\mathbf{F}_i$ from $\mathbf{F}_{\text{HR}}$ using balancing loss $\mathcal{L}_b$. However, $\mathcal{L}_b$ does not guarantee that features assigned to different classifiers have unique characteristics. Therefore, we adopt the concept of orthogonality [21] and LDA [36] to meditate feature properties in within-classes and between-classes.

Orthogonal strategy. Feeding unrelated features into the independent classifiers can improve classification performance in our CAL-GAN. Therefore, to prevent different subsets $\mathbf{F}_j$ and $\mathbf{F}_k$ $(j \neq k)$ from sharing some properties in common, we introduce the 1×1 orthogonal convolution $\mathcal{O}_i$ [47, 55, 21] to maximize independency between different splits of $\mathbf{F}_i$ as follows:

$$\mathbf{F}_i^* = \mathcal{O}_i(\mathbf{F}_i), \tag{5}$$

where $\mathbf{F}_i^* \in \mathbb{R}^{H/4 \times W/4 \times c}$ is a linear projection of $\mathbf{F}_i$. Since we expect $\mathbf{F}_i^*$ and $\mathbf{F}_j^*$ to be independent for $i \neq j$, we propose an orthogonal loss $\mathcal{L}_{\text{o}}$ and optimize $\mathcal{O} \in \mathbb{R}^{N \times c \times c \times 1 \times 1}$, which is a stack of weights $\mathcal{O}_i \in \mathbb{R}^{c \times c \times 1 \times 1}$, as follows:

$$\mathcal{L}_{\text{o}} = \left\| \psi(\mathcal{O})\psi(\mathcal{O})^T - I_N \right\|_F^2, \tag{6}$$

where ψ is a view operator that reshapes $\mathcal{O} \in \mathbb{R}^{N \times c \times c \times 1 \times 1}$ to $\mathbb{R}^{N \times c^2}$, and I_N is an $N \times N$ identity matrix and $\|\cdot\|_F^2$ represents the Frobenius norm. By forcing $\psi(\mathcal{O})\psi(\mathcal{O})^T$ to be an identity matrix, each $\mathcal{O}_i$ can deliver unique information.

LDA strategy. We further define distribution-base loss $\mathcal{L}_{\text{dist}}$ to gather similar information in the same index, while splitting different information to disjoint classes. The proposed distribution loss $\mathcal{L}_{\text{dist}}$ is calculated as a sum of between-class loss $\mathcal{L}_{\text{btw}}$ and within-class loss $\mathcal{L}_{\text{wi}}$ as follows:

$$\begin{aligned} \mathcal{L}_{\text{btw}} &= \sum_{i<j}^{N} \left(\frac{\bar{\mathbf{F}}_i^* \cdot \bar{\mathbf{F}}_j^*}{\left\|\bar{\mathbf{F}}_i^*\right\| \left\|\bar{\mathbf{F}}_j^*\right\|} \right)^2, \\ \mathcal{L}_{\text{wi}} &= \sum_{i=1}^{N} \text{Var}\left(\mathbf{F}_i^*\right), \\ \mathcal{L}_{\text{dist}} &= \mathcal{L}_{\text{btw}} + \mathcal{L}_{\text{wi}}, \end{aligned} \tag{7}$$

where the mean vector $\bar{\mathbf{F}}_i^* \in \mathbb{R}^c$ and variance are calculated over spatial dimensions. Minimizing $\mathcal{L}_{\text{btw}}$ decreases the correlation between different classes, while optimizing $\mathcal{L}_{\text{wi}}$ reduces the variance of the points assigned to the same class. Therefore, the proposed distribution loss $\mathcal{L}_{\text{dist}}$ effectively adjusts distances between features as intended.

3.4. Training domain-aware discriminator

Mixture of classifiers. Rather than using a single classifier, the proposed content-aware discriminator architecture leverages N parallel classifiers $\mathcal{C}_i$ as shown in Figure 2. Therefore, the discriminator output $\mathbf{D}$ is defined as a summation of the N spatially *disjoint* outputs, since different discriminators process different pixels in the original feature $\mathbf{F}$. We calculate $\mathbf{D}$ as follows:

$$\mathbf{D} = \sum_{i=1}^{N} \mathcal{C}_i\left(\mathbf{F}_i^*\right). \tag{8}$$

By utilizing the spatially-disjoint property, the output $\mathbf{D}$ can be calculated more efficiently. Still, the overhead from our implementation is minor since the discriminator is used only for training, not inference.

Baseline		EDSR		SwinIR		RRDB				
Metric	Dataset	+GAN	**+CAL-GAN**	+GAN	**+CAL-GAN**	ESRGAN	USRGAN	SPSR	LDL	**+CAL-GAN**
$\text{LPIPS}_{\downarrow}$	Set5	0.155	**0.088**	0.068	**0.058**	0.076	0.080	0.065	0.067	**0.061**
	BSD100	0.297	**0.136**	0.163	**0.147**	0.162	0.174	0.161	0.153	**0.151**
	Urban100	0.205	**0.154**	0.107	**0.098**	0.123	0.133	0.118	0.110	**0.108**
	General100	0.158	**0.104**	0.076	**0.074**	0.088	0.094	0.087	0.079	**0.077**
	DIV2K (val)	0.220	**0.074**	0.093	**0.087**	0.115	0.133	0.101	0.101	**0.091**
$\text{FID}_{\downarrow}$	Set5	42.364	**27.196**	**25.570**	26.205	27.215	37.006	30.904	25.288	**24.490**
	BSD100	76.872	**54.379**	43.373	**40.170**	50.691	48.347	47.304	**40.806**	43.217
	Urban100	29.149	**22.930**	19.219	**17.447**	20.345	21.555	18.672	17.758	**17.550**
	General100	50.187	**37.035**	26.545	**26.504**	29.843	32.959	30.159	**27.506**	29.417
	DIV2K (val)	19.464	**15.863**	13.270	**12.097**	13.557	14.031	13.754	12.145	**11.772**
$\text{BRISQUE}_{\downarrow}$	Set5	29.383	**16.711**	16.944	**14.066**	16.164	16.667	**9.457**	17.704	14.125
	BSD100	26.502	**11.917**	13.632	**11.406**	11.993	12.372	**9.475**	13.202	9.948
	Urban100	21.771	**12.629**	18.701	**17.608**	22.392	15.684	**15.584**	17.467	15.624
	General100	17.957	**17.905**	16.778	**15.221**	16.764	17.625	**12.836**	18.084	15.091
	DIV2K (val)	27.589	**11.555**	11.852	**11.406**	14.140	14.516	**11.220**	12.680	13.576
$\text{DISTS}_{\downarrow}$	Set5	0.138	**0.113**	0.093	**0.085**	0.095	0.105	0.092	0.092	**0.091**
	BSD100	0.184	**0.151**	0.137	**0.128**	0.124	0.137	0.126	0.127	**0.118**
	Urban100	0.160	**0.127**	0.089	**0.083**	0.088	0.098	0.085	**0.079**	0.082
	General100	0.130	**0.105**	**0.080**	0.081	0.085	0.093	0.088	**0.083**	**0.083**
	DIV2K (val)	0.112	**0.078**	0.051	**0.048**	0.059	0.065	0.055	0.053	**0.049**

Table 1: $\times$**4 SR performance comparison about different perceptual metrics on various datasets. Best** and second-best numbers are denoted with **bold** and underline, respectively. Models with +GAN are trained using naïve adversarial loss [13], while +CAL-GAN denotes networks optimized with the proposed distribution-aware discriminator architecture.

Objective function for discriminator. To train our discriminator, we combine a standard cross-entropy loss $\mathcal{L}_{\text{dis}}$ and all of the novel objective functions listed above to construct the total objective $\mathcal{L}_{\text{D}}$ by following:

$$\begin{aligned} \mathcal{L}_{\text{dis}} &= -\log \mathbf{D}_{\text{real}} - \log\left(1 - \mathbf{D}_{\text{fake}}\right), \\ \mathcal{L}_{\text{D}} &= \lambda_{\text{b}}\mathcal{L}_{\text{b}} + \lambda_{\text{o}}\mathcal{L}_{\text{o}} + \lambda_{\text{d}}\mathcal{L}_{\text{dist}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}}, \end{aligned} \tag{9}$$

where $\lambda_{\text{b}}, \lambda_{\text{o}}, \lambda_{\text{d}}, \lambda_{\text{dis}}$ are hyperparameters of the corresponding loss functions. $\mathbf{D}_{\text{real}}$ and $\mathbf{D}_{\text{fake}}$ are discriminator outputs given HR and SR images as input, respectively. We note that the proposed term $\mathcal{L}_{\text{D}}$ can be generally applicable to diverse discriminator architectures.

Metric	Dataset	ESRGAN	USRGAN	SPSR	LDL	**+CAL-GAN**
$\text{PSNR}^{\uparrow}$	Set5	30.438	30.910	30.397	30.985	**31.177**
	BSD100	25.288	**25.974**	25.495	25.929	25.925
	Urban100	24.365	24.891	24.804	**25.498**	25.290
	General100	29.425	30.001	28.561	**30.232**	30.182
	DIV2K (val)	28.175	28.787	28.182	**28.951**	28.863
$\text{SSIM}^{\uparrow}$	Set5	0.852	**0.866**	0.844	0.862	0.863
	BSD100	0.650	0.676	0.658	**0.679**	0.676
	Urban100	0.734	0.750	0.742	**0.767**	0.763
	General100	0.810	0.824	0.809	**0.827**	0.825
	DIV2K (val)	0.776	0.794	0.772	**0.795**	0.790
User Study$^{\uparrow}$ (%)		10.7	10.8	20.0	19.7	**39.2**

Table 2: **Distortion metrics comparison with photo-realistic SR methods.**

4. Experiments

Datasets. To train SR models, we adopt widely-used DIV2K [44], which contains 800 high-quality images. The network takes 64×64 random crops as inputs with a batch size of 48. For evaluation, we use standard Set5 [2], BSD100 [34], and Urban100 [17] benchmarks. We also compare various methods on General100 [9] and DIV2K [44] validation datasets for extensive study.

Metrics. Our primary goal is to reconstruct photo-realistic SR results. Therefore, we adopt four widely-used perceptual metrics to measure image quality rather than conventional PSNR/SSIM. For full-reference evaluation using ground-truth images, we use LPIPS [59] and DISTS [7]. We also introduce non-reference and distribution metrics, BRISQUE [37] and FID [16], respectively, to show the effectiveness of the proposed CAL-GAN framework. We note that models with the best perceptual quality may not show high PSNR/SSIM under the limited network capacity [3].

Experimental configurations. We first train the PSNR-based SR model without using the proposed discriminator. Then, the final SR model is fine-tuned with the discriminator under (9). Adam [25] optimizer is used to update our SR and discriminator. For the generator, $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{per}}$, and $\mathcal{L}_{\text{gen}}$ are set to 0.01, 1, and 0.005, respectively. And for the discriminator, λ_{b}, λ_{o}, λ_{d}, and λ_{dis} are set to 0.05, 10, 10, and 1, respectively. In order to determine the most appropriate weight for our proposed loss functions, we have performed a greedy search. Our model is trained for 300,000 iterations, where the learning rate starts from 10^{-4} and is halved after 150,000 updates. It takes about 32 hours to train our model using four RTX 2080 Ti GPUs.

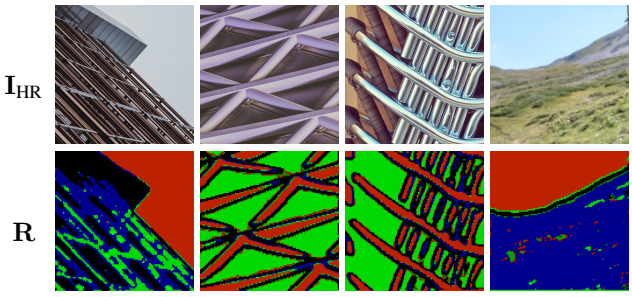
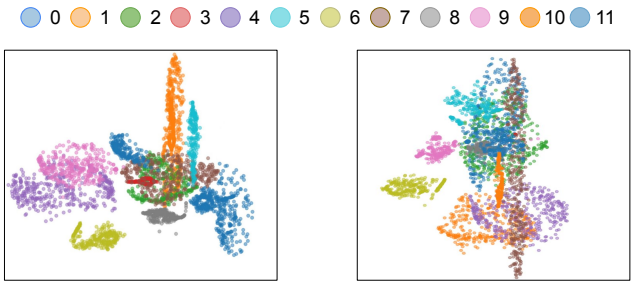


Figure 3: **Visualization of the routing mask R.** Our routing module assigns one of N classes to each point in HR and SR images. We show distinctive colors to indicate different one-hot vectors.



(a) Image A (b) Image B

Figure 4: **Distribution of differently routed features.** We visualize the t-SNE [46] plots on two individual images. Each data point refers to a spatial feature from the discriminator, *i.e.*, $\mathbf{F}_i^*(x, y)$. Different colors indicate different classes learned from our class-aware local discriminator.

4.1. CAL-GAN with state-of-the-art methods

We first show that the proposed CAL-GAN framework is generally applicable to existing state-of-the-art SR models. For extensive comparison, we use two CNN-based methods, EDSR [30] and RRDB [50], and one Transformer-based SwinIR [28]. We note that the term 'RRDB' is used to refer to a class of models based on Residual-in-Residual Dense Blocks [50], rather than a specific network architecture. When integrated with EDSR [30] and SwinIR [28], CAL-GAN consistently improves the perceptual quality of super-resolved images compared to the plain GAN counterparts as shown in Table 1. The proposed approach is effective regardless of the baseline model and evaluation datasets, showing that CAL-GAN can be generally applied to both traditional CNNs and emerging attention-based models.

The 'RRDB' section in Table 1 compares performance among different GAN-based solutions which use similar backbone. Compared to ESRGAN [50] and USRGAN [57], which adopt plain adversarial training framework, the proposed method shows significant performance gain on all datasets and metrics. While SPSR [32] has good non-reference BRISQUE scores, it is due to the model having an additional image gradient estimation branch to which BRISQUE is sensitive. Recent LDL [29] is designed to discriminate visual artifacts from adversarial training from real high-frequency details and achieves comparable performance to CAL-GAN. Still, due to the ambiguity in artifacts and details, our method provides much sharper and more accurate reconstructions, as shown in Figure 6. Furthermore, we also show the qualitative comparisons on RealSR [58] dataset compared to previous methods [48, 58, 29].

We further provide traditional distortion-based metrics, *i.e.*, PSNR and SSIM, and user preference study in Table 2. For the user study, we randomly selected 12 images from the evaluation datasets, and 10 participants are instructed to choose the most visually pleasing SR result from five different models. The results show that approximately 40% of the participants preferred the images generated by our CAL-GAN, indicating superior visual quality compared to other methods. While our CAL-GAN may not achieve the best PSNR and SSIM scores on various benchmarks, the user preference study confirms that these metrics are not always strongly correlated with human perception of image quality.

4.2. Visualizing content-aware features

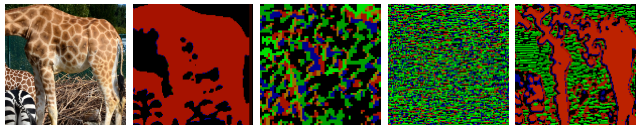
We visually analyze the learned discriminator features to demonstrate how our routing module works. As shown in Figure 3, every pixel in the feature map is assigned to a certain class based on its content. Specifically, patches in the plain area, *e.g.*, sky, are labeled with red, while regions with detailed textures are marked with blue and black. The visualization validates that the proposed routing strategy can effectively discriminate image semantics without requiring any ground-truth class labels. Also, the assigned labels are well-balanced due to our balancing loss. Furthermore, we adopt the t-SNE [46] plot to show that features with different indices form distinct clusters as shown in Figure 4. The above analysis justifies that our routing module can assign proper labels to each local feature by image context and group them into the same category.

4.3. Effect of the proposed loss combination

Table 3 provides a quantitative analysis of our loss functions: balancing loss $\mathcal{L}_{\text{b}}$, orthogonal loss $\mathcal{L}_{\text{o}}$, and distribution loss $\mathcal{L}_{\text{dist}}$. Our CAL-GAN achieves the best perceptual scores when using all three terms together. Visualization in Figure 5 further provides insight regarding the role of each loss function. When the balancing loss $\mathcal{L}_{\text{b}}$ is not used, a load-imbalance issue occurs, and only a few classifiers are used during training. Specifically, our routing module can separate a forward object and background to some extent, but the class diversity decreases significantly, as shown in Figure 5b. Without the orthogonal loss $\mathcal{L}_{\text{o}}$, we can observe patch-like artifacts as shown in Figure 5c. While the discriminator can group neighboring contents that have similar properties, the lack of orthogonality cannot guarantee that

$\mathcal{L}_{\text{b}}$	$\mathcal{L}_{\text{o}}$	$\mathcal{L}_{\text{dist}}$	LPIPS$_{\downarrow}$	DISTS$_{\downarrow}$	FID$_{\downarrow}$
✗	✓	✓	0.100	0.053	13.660
✓	✗	✓	0.109	0.057	13.807
✓	✓	✗	0.106	0.052	14.046
✓	✓	✓	**0.091**	**0.049**	**11.772**

Table 3: **Comparison of models on DIV2K validation set with subtracting proposed loss functions.** The cross marker (✗) indicates that the corresponding loss is not used for training.



(a) $\mathbf{I}_{\text{SR}}$ (b) w/o $\mathcal{L}_{\text{b}}$ (c) w/o $\mathcal{L}_{\text{o}}$ (d) w/o $\mathcal{L}_{\text{dist}}$ (e) **Ours**

Figure 5: **Visualization of the routing mask $\mathbf{R}$ under different loss combinations.** Same as Figure 3, different colors in the mask map indicate different one-hot vectors $\mathbf{R}$.

different categories can represent independent information. Finally, when we do not include the distribution loss $\mathcal{L}_{\text{dist}}$ to our training objective, features in different subsets cannot be distinguished as they are mixed in the high-dimensional manifold. Figure 5d illustrates that the estimated mask $\mathbf{R}$ looks like noise rather than representing low-level context. Our CAL-GAN is trained using all three terms together and therefore predicts well-balanced class labels which deliver low-level color and texture information of the given image as shown in Figure 5e.

4.4. Analyzing our discriminator design

The discriminator plays a significant role in our CAL-GAN. Therefore, we perform an extensive ablation about the relationship between its design and SR performance.

The number of classifiers. First, Table 4 compares discriminators with the different numbers of classifiers N. We note that $N = 1$ corresponds to a standard GAN configuration, which performs the worst in terms of perceptual metrics, *i.e.*, LPIPS, DISTS, and FID. Using $N = 8$ or 12 classifiers yield the best results, while the minimum FID is achieved when $N = 12$. Since each classifier will receive fewer training samples and be too specific with much larger N, increasing the number of classifiers to over $N = 16$ does not bring a performance gain compared to $N = 12$. Therefore, we use $N = 12$ classifiers in our framework.

Patch size. According to Isola *et al*. [18], it is beneficial for the discriminator to assign a single class to $p \times p$ pixels in the input image. In our design, the patch size p is determined by the size of discriminator feature $\mathbf{F}_*$. While we use $p = 4$ throughout the main manuscript, other design choices are also available. Table 5 shows how the patch size p affects to the SR performance. Using $p = 4$ brings the best perceptual scores, and densely labeling all the pixels ($p = 1$) has the worst performance due to the classification noise. Although larger patch sizes have advantages in training-time computational costs, the performance decreases with $p = 8$ or $p = 16$. The performance degradation happens because assigning a single label to a broader content with mixed frequency components is more challenging. Thus, we select $p = 4$ in our final design.

# of classifiers N	1	4	8	**12**	16
LPIPS	0.110	0.106	0.091	**0.091**	0.098
DISTS	0.056	0.052	0.051	**0.049**	0.050
FID	13.953	13.369	12.965	**11.772**	12.520

Table 4: **Effects of number of classifiers on DIV2K (val).**

patch size	1	**4**	8	16
LPIPS	0.104	**0.091**	0.098	0.099
DISTS	0.060	**0.049**	0.055	0.055
FID	14.116	**11.772**	13.630	13.614

Table 5: **Effects of patch size on DIV2K (val).**

Dataset	Set5	Set14	BSD100	Urban100	General100	DIV2K (val)
Segmentation-based	0.106	0.166	0.233	0.167	0.116	0.137
Content-based (Ours)	**0.061**	**0.131**	**0.151**	**0.108**	**0.077**	**0.091**

Table 6: **LPIPS comparison with segmentation-based routing.**

4.5. Effect of the learned routing strategy

The proposed content-aware local discriminator learns to classify patches of different distributions without handcrafted class labels. As visualized in Figure 3 and Figure 5, the learned assignments group similar contents and colors together. To validate the effectiveness of the proposed approach, we compare the learned labels with predictions from a pre-trained segmentation model. Since DIV2K dataset does not provide segmentation labels, we apply pre-trained Mask-RCNN [15] to get pseudo labels for the training samples. For fair comparison, we introduce COCO [31] dataset and manually reorganize the existing 80 classes into 12 high-level classes. For example, we group {horse, zebra} and {handbag, suitcase} together, respectively. Table 6 shows that our CAL-GAN achieves perceptually better results than segmentation-based routing.

While segmentation-based routing can consider high-level semantics, we have discovered some possible limitations. First, the pre-trained segmentation model does not perform well on high-quality DIV2K images and provides incorrect routing due to domain mismatch. We note that it is not suitable to use COCO images to train the SR model instead, as they contain lots of noise. Also, semantically the same pixels may not have the same low-level property. Specifically, a door and tire belong to the same class, 'car,' but their contents and textures vary. On the other hand, our routing strategy can group pixels with similar low-level characteristics, *e.g.*, contents and colors, without requiring ground-truth labels for training.

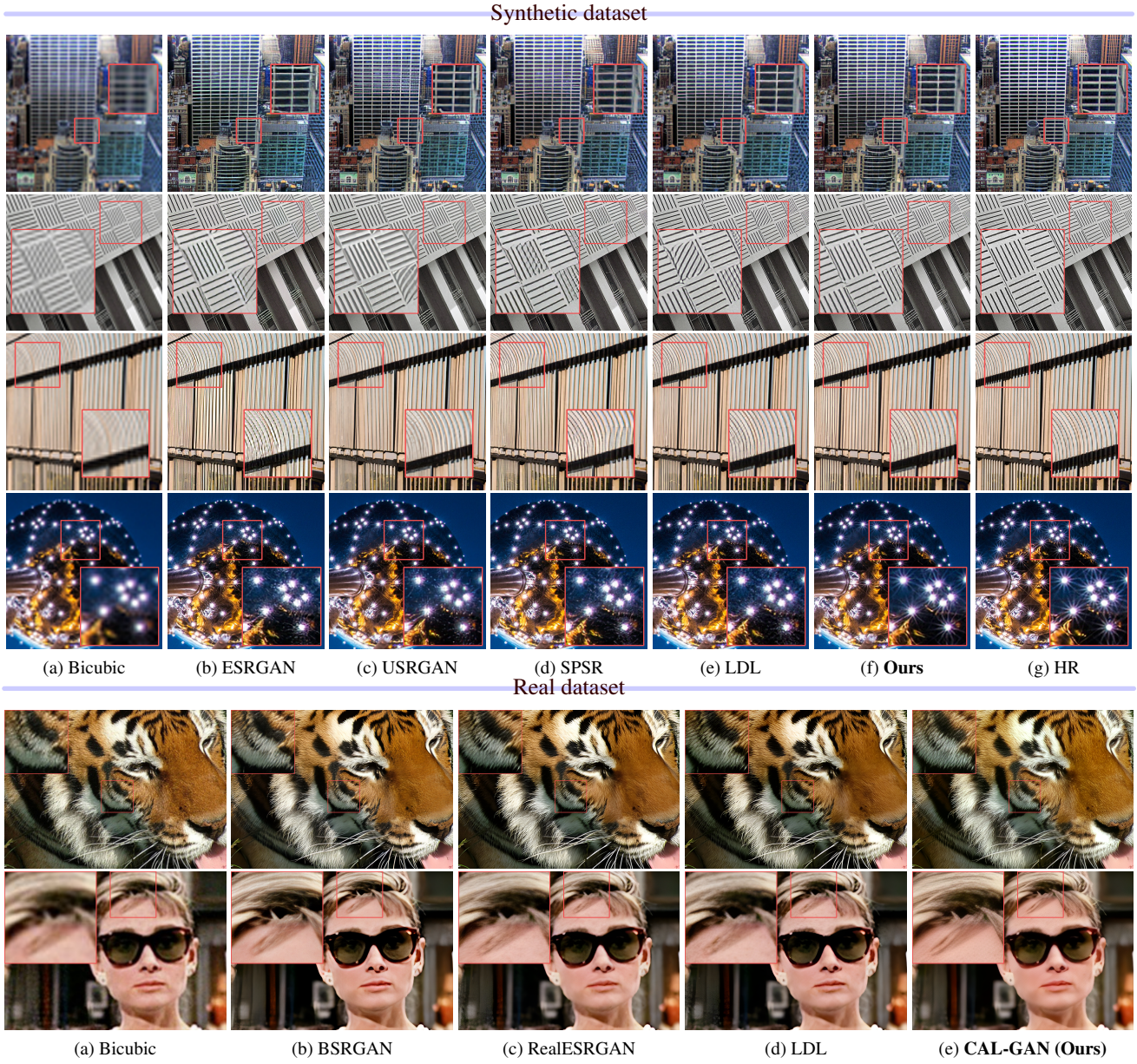


Figure 6: **Qualitative comparison on $\times$4 SR with state-of-the-art methods.**

5. Conclusion

In this work, we propose a novel GAN framework, CAL-GAN, which utilizes a mixture of classifiers (MoC) to address the challenges of handling complex natural image distribution in perceptual SR. While previous GAN-based SR methods learn the distribution of real and fake images unconditionally, our method enables multiple classifiers to learn and distinguish specialized distribution based on its content using innovative loss formulations. Our extensive experiments demonstrate that using multiple classifiers leads to superior handling of complicated natural image distribution in SR and can be extended to other image restoration tasks such as deblurring and denoising.

Acknowledgments This work was supported in part by the IITP grants [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University), No. 2021-0-02068, and No.2023-0-00156], and the NRF grant [No. 2021M3A9E4080782] funded by the Korea government (MSIT).

References

[1] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR*, 2015.

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVA*, 2012.

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018.

[4] Angela Castillo, María Escobar, Juan C Pérez, Andrés Romero, Radu Timofte, Luc Van Gool, and Pablo Arbelaez. Generalized real-world super-resolution through adversarial robustness. In *CVPR*, 2021.

[5] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, 2021.

[6] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[7] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 2020.

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2015.

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, 2016.

[10] Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. Deep linear discriminant analysis. *arXiv preprint arXiv:1511.04707*, 2015.

[11] Yuchen Fan, Jiahui Yu, Ding Liu, and Thomas S Huang. Scale-wise convolution for image restoration. In *AAAI*, 2020.

[12] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 27, 2014.

[14] Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*, 2021.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

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

[17] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[19] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 1991.

[20] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[21] Myeongho Jeon, Daekyung Kim, Woochul Lee, Myungjoo Kang, and Joonseok Lee. A conservative approach for unbiased learning on unknown biases. In *CVPR*, 2022.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[23] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[24] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

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

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

[27] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

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

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

[30] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR*, 2017.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[32] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *CVPR*, 2020.

[33] Shunta Maeda. Image super-resolution with deep dictionary. *arXiv preprint arXiv:2207.09228*, 2022.

[34] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

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

[36] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *SiPS*, 1999.

[37] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *TIP*, 2012.

[38] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *TPAMI*, 2021.

[39] Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Srobb: Targeted perceptual loss for single image super-resolution. In *ICCV*, 2019.

[40] Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, et al. Mesh-tensorflow: Deep learning for supercomputers. In *NeurIPS*, 2018.

[41] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[42] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020.

[43] Sanghyun Son and Kyoung Mu Lee. Srwarp: Generalized image super-resolution under arbitrary transformation. In *CVPR*, 2021.

[44] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR Workshops*, 2017.

[45] Rao Muhammad Umer and Christian Micheloni. Deep cyclic generative adversarial residual convolutional networks for real image super-resolution. In *ECCV*, 2020.

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[47] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *IJCAI*, 2015.

[48] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021.

[49] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018.

[50] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018.

[51] Yin Wang. Single image super-resolution with u-net generative adversarial networks. In *IMCEC*, 2021.

[52] Yifan Wang, Lijun Wang, Hongyu Wang, and Peihua Li. Resolution-aware network for image super-resolution. *TCSVT*, 2018.

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

[54] Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *ICCV*, 2021.

[55] Di Xie, Jiang Xiong, and Shiliang Pu. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In *CVPR*, 2017.

[56] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022.

[57] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, 2020.

[58] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021.

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

[60] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Super resolution generative adversarial networks with learning to rank. *arXiv preprint arXiv:2107.09427*, 2021.

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

[62] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018.

[63] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. *arXiv preprint arXiv:2202.09368*, 2022.

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017.