Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The re-identification generative adversarial network for image super-resolution

duwei Hua
Wuhan Institute of Shipbuilding Technology
hanling Li
lhl1231009@126.com

Wuhan Institute of Shipbuilding Technology

Article

Keywords: Image super-resolution, U-net, Realness feedback, Discriminator

Posted Date: July 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9799543/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Highlights

The re-identification generative adversarial network for image super-resolution

Duwei Hua,Hanling Li

- A second-order attention discriminator based on covariance is proposed to extract deep features, enhancing the discriminator's ability to identify authenticity and improving the quality of reconstructed images.
- We propose a loss function designed to identify the local relative authenticity of the input image, thereby enabling pixel-level feedback to the generator and enhancing the quality of the restored images.
- Numerous experiments demonstrate that SOAGAN is cable of achieving superior results.

The re-identification generative adversarial network for image super-resolution

Duwei Hua[a], Hanling Li[a,**]

[a]College of Electrical and Electronic Engineering, Wuhan Institue of Shipbuilding Technology, WuHan, 430050, China

ARTICLE INFO

Keywords:
Image super-resolution
U-net
Realness feedback
Discriminator

ABSTRACT

Although single image super-resolution (SISR) has made significant progress in the field of image restoration, the images reconstructed by it frequently diverge from human visual perception habits. To tackle this challenge, we propose a second-order attention generative adversarial network (SOAGAN) that incorporates re-identification mechanisms for image super-resolution. Specifically, we employ a U-Net architecture integrated with covariance attention as the discriminator to facilitate the restoration of image details. Notably, the U-Net based discriminator is capable of providing not only global image-level relative realness feedback but also pixel-level relative realness feedback to the generator within our framework, thereby facilitating the updating of model parameters and significantly improving the quality of the reconstructed images. Abundant experiments demonstrate that the subjective visual quality of images generated by our method exhibits a marked improvement compared to those produced by representative approaches.

1. Introduction

Single image super-resolution (SISR) aims to recover the corresponding high-resolution (HR) image from the given low-resolution (LR) image(Freeman et al., 2000), which is highly significant for security city building, medical image analysis, remote sensing image processing, et al.(Rasti et al., 2016; Huang et al., 2017b; Yıldırım and Güngör, 2012). In essence, SISR is an ill-posed problem because there exist many mappings given the LR image to obtain HR image, which brings a great challenge for image super-resolution. The main manners for SISR include the interpolation-based method, the model-based method, and the leaning-based method(Chen et al., 2022; Zhang and Wu, 2006; Dong et al., 2011; Ledig et al., 2017; Dong et al., 2014; Zhang et al., 2018a). The interpolation-based method that usually reconstructs the image with loss of high-frequency information and poor subjective visual quality, such as the bicubic interpolation method(Keys, 2003). Simultaneously, the model-based approaches combine various prior information for the SR images, such as (Marquina and Osher, 2008; Sun et al., 2008; Dai et al., 2009). However, the image it generated depends heavily on the selection of prior information and is inefficient with the increase of scale. Recently, with the development of deep leaning, the learning-based techniques learn the mapping functions between HR and LR images, which employs enormous corresponding image pairs mainly depending on neural networks, due to its powerful modeling ability, more and more researchers have proposed enormous relevant methods to promote the quality of recovered SR images(Tai et al., 2017; Zhao et al., 2018; Shocher et al., 2018; Zhang et al., 2020; Ning et al., 2020; Huang et al., 2017a; Yang et al., 2021; Xu et al., 2021).

In 2014, Dong et al.(Dong et al., 2014) first employed the deep convolutional neural network (CNN) to restore SR images, which significantly improved the quality of reconstructed images compared to previous methods. However, the process of model parameter optimization affects image details. To solve the problem, He et al.(He et al., 2016) designed ResNet with skip connections to mitigate model collapse, improving model accuracy. Subsequently, numerous researchers have begun to introduce ResNet to cope with the image SR. For example, Kim et al.(Kim et al., 2016a) designed a 20 layers model that greatly improved the SR image quality. Considering the efficiency of reconstruction process, ESPCN(Shi et al., 2016) designed sub-pixel convolution in the tail of the SR network to increase the image resolution, highly reducing computation costs. Then, DRCN(Kim et al., 2016c) first employed recurrent neural network and residual learning to deepen the network, which got promising results. Similarly, to realize multiple-scale image SR with a neural network, LapSRN(Lai et al., 2017) proposed progressively upsampling strategy and achieved remarkable performance. Furthermore, considering the self-similarity of images, ZSSR(Shocher et al., 2018) designed a zero-shot

*Corresponding author.
lhl12310090126.com (H. Li)

image SR network, which only needs one image to optimize the model parameters and recover image details. However, the reconstruction of ZSSR often need several optimizations, influencing its efficiency. To deal with the problem, MZSR(Soh et al., 2020) introduced meta-learning, which made the test stage only need one update to restore SR images. To further promote SR image quality, SRDenseNet(Tong et al., 2017) utilized the dense connections(Huang et al., 2017a) to improve feature representation abilities of networks. Through detailed research and analysis, RCAN designed the RCAB(Zhang et al., 2018a) module to make full use of intermediate deep features by integrating channel attention mechanism, obtaining greatly performance, simultaneously averting deeper or wider model designing. Furthermore, SAN(Dai et al., 2019) employed covariance-based pooling to replace global average pooling and designed second-order attention covariance-based module for image super-resolution, greatly improving the quality of reconstructed images. Similarly, to further improve the feature representation abilities, researchers (Ronneberger et al., 2015) proposed the U-Net network to deal with image segmentation problems. Considering the powerful capabilities of above mentioned methods, we propose to combine the U-Net network with covariance-based second-order attention to obtain sufficiently features, recovering enough SR image details.

SRGAN(Ledig et al., 2017) suggested that although CNN-based image super-resolution methods had excellent metric indicators, subjective visual quality was poor and the reconstructed image was usually overly smooth and unnatural. Considering the powerful ability of the Generative adversarial network (GAN)(Goodfellow et al., 2014) to generate images, SRGAN introduced the generative adversarial strategy for image reconstruction, while the restored images were more natural with high visual quality. Then, ESRGAN(Wang et al., 2018b) improved the structure of the SRGAN network and introduced relative adversarial loss with reference to RaGAN(Jolicoeur-Martineau, 2018) to further improve the visual quality of the reconstructed image. Although the application of GAN improves the visual quality of SR image to a certain extent, the GAN discriminator only identifies the realness of the global image to update network parameters, which makes it impossible to effectively identify pixel-wise local details of images in the reconstruction process, and the discriminator does not consider the correlation of middle features in the network. Therefore, the ability of current GAN discriminator network to identify the authenticity of generated image to update the model parameters of generator network is still insufficient. In order to solve the problem and reconstruct high quality image with better visual results, we propose a U-Net network with covariance-based second-order channel attention as discriminator to re-identification and greatly improve its ability of identify authenticity. Additionally, to further improve the quality of reconstructed images, we also employ the relative adversarial loss to optimize our network parameters. The overall judgment of the relative authenticity of reconstructed images is not only carried out in the U-Net coding stage but also the pixel-wise relative authenticity of the reconstructed image is identified in the U-Net decoding stage, which meaning the re-identification about image restorations. Numerous results of experiments demonstrate the proposed method could effectively improve the quality of reconstructed images.

2. Related Work

Recently, deep learning has remarkably stimulated the development of image super-resolution (SR). In the paper, we focus on the deep learning-based SR methods.

Deep Networks for SISR. In 2014, Dong et al.(Dong et al., 2014) first employed CNN to restore SR images in end-to-end manners, and the quality of reconstructed images was greatly improved comparing to traditional methods. Then, enormous CNN-based models have been widely explored. For example, to improve the model efficiency, FSRCNN(Dong et al., 2016) reconstructed the network structures, proposing to up-sample in the tail of networks. Thereby, the middle networks deals with low-resolution images, which significantly decline the model computation costs. Similarly, ESPCN employed sub-pixel convolution at the tail of network to improve the model efficiency. Furthermore, with the network difficult training problem to be solved by residual learning(He et al., 2016), the model layers are further increased. For example, VDSR(Kim et al., 2016b) designed a twenty-layer network by residual structure and gradient clipping manners, considering the multiscale reconstruction problem at the same time which utilized a single model to realize multiscale image super-resolution. Then, to make full use of middle deep features and improve the network performances, Huang et al.(Huang et al., 2017a) proposed a dense connection strategy through exploration of the neural network structures. Therefore, MemNet(Tai et al., 2017) combined residual and dense connection in a network to achieve lasting memory of features, and proposed gate units to filter numerous features, which greatly enhancing the SR images quality. Similarly, RDN(Zhang et al., 2018b) further explored deep network structures and constructed RDN network for image super-resolution. To improve the model efficiency and multi-scale image reconstruction performance, LapSRN(Lai et al., 2017) proposed progressively upsampling strategies. Then,

to sufficiently use deep features, RCAN(Zhang et al., 2018a) proposed residual channel attention for image super-resolution and got great results. Furthermore, SAN(Dai et al., 2019) employed covariance pooling to replace global average pooling with channel attention block, and proposed the second-order channel attention module to enhance the representation ability of feature correlation and significantly improved SR image accuracy. CFSRCNN(Tian et al., 2020) expanded five layers CNN at the tail of the network to further smooth the reconstructed features and achieve coarse-to-fine grained image super-resolution. Additionally, Cutblur(Ahn et al., 2024) proposed a data augmentation strategy for image super-resolution, which introduced data augmentation applied in high-level computer vision tasks into low-level vision tasks to improve the reconstructed image quality. LCSCNet(Yang et al., 2020) explored the linear compression and short connection and employed adaptive feature fusion to reconstruct images. Abundant networks mainly focus on the design of wider or deeper networks for image super-resolution, and achieve remarkable results on objective indicators, such as Peak-Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measurement (SSIM)(Wang et al., 2004). However, the visual quality of reconstructed image is not satisfactory as expected, and there are problems such as unnatural, fuzzy, etc.

The Gan-Based SISR Networks. To improve the subjective qualities of SR images, numerous researchers employed alternate iteration to update the generator and discriminator parameters, while the reconstructed images was more in line with human visual habits. For example, SRGAN(Ledig et al., 2017) employed generate adversarial strategy to image super-resolution. Due to the alternate iteration of generator and discriminator, the reconstructed SR image was difficult to distinguish authenticity. Simultaneously, the perceptual loss based on VGG network(Ledig et al., 2017) was designed to constrain SR images in feature space. Subsequently, abundant researchers have explored various manners to enhance the SR images visual quality. For example, SFTGAN(Wang et al., 2018a) proposed the spatial feature conversion layer, which effectively integrated the semantic category condition information into the reconstructed image. Furthermore, comparing to SRGAN, ESRGAN(Wang et al., 2018b) introduced perceptual loss, relative adversarial loss, and removed batch normalization operation to further improve the SR images quality. Although the visual quality of reconstructed images by perception-driven SR network had been greatly improved, the SR image often suffered geometric changes. Consequently, SPSR(Ma et al., 2020) integrated the gradient branch with the backbone network to further refine the details of reconstructed image. Recently, certain researchers have endeavored to devise sophisticated degradation models for constructing corresponding datasets that simulate real-world degradation factors. These datasets were subsequently employed to train the models, resulting in a significant enhancement of image quality in real-world reconstruction scenarios. For example, Real-ESRGAN(Wang et al., 2021) proposed a high-order degradation model; Zhang et al.(Zhang et al., 2021a) proposed BSRGAN network through redesigning the classical degradation process to imitate the real world degradation model. However, the networks are insufficient to represent complex SR image details, which may cause some details loss.

Attention mechanism. Attention mechanism means adaptive focus areas in human visual system (Itti et al., 1998). Recently, in the field of image classification, object detection and other research fields(Hu et al., 2018; Vaswani et al., 2017; Cao et al., 2020; Liu et al., 2016; Woo et al., 2018), numerous researchers introduce attention mechanisms to solve certain problems. For example, SENet(Hu et al., 2018) designed channel attention to fully tap into deep feature correlations, enhancing the model representation abilities, and significantly improving the image classification accuracy. Similarly, the authors(Vaswani et al., 2017) proposed non-local operation to explore feature information for video classification. Similarly, in the field of image SR, some researchers employed attention mechanism to improve the model performances. For example, RCAN(Zhang et al., 2018a) designed residual channel attention; SERAN(Zhang et al., 2021b) designed channel attention for medical magnetic resonance (MR) image reconstruction, which effectively improved the medical image qualities; SAN(Dai et al., 2019) designed the covariance-based second-order channel attention module to effectively improve the accuracy of image reconstruction. However, the above methods mainly improve the objective metric indicator, such as PSNR and SSIM, while the improvement in visual quality is almost imperceptible. To deal with the problem, we combine the covariance-based second-order channel attention with GAN to improve the texture of reconstructed image and restore heterogeneous image details that are more in line with human subjective visual quality.

3. Methods

In thie study, we design a SOAGAN network tailored for single image super-resolution with the overall framework of SOAGAN depicted in Fig. 1. Specifically, the generator G takes a low-resolution image as input to generate a super-resolution image, which along with the target image, is subsequently fed into the discriminator. The encoding

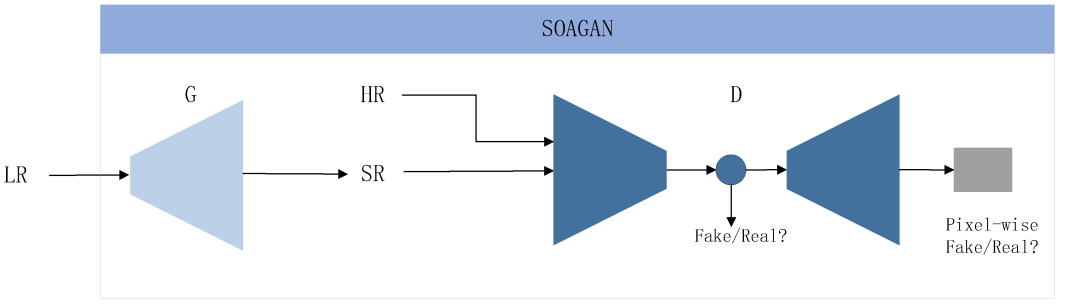


Figure 1: The overall architectural framework of SOAGAN is depicted herein, where G represents the generator, D signifies the discriminator component.

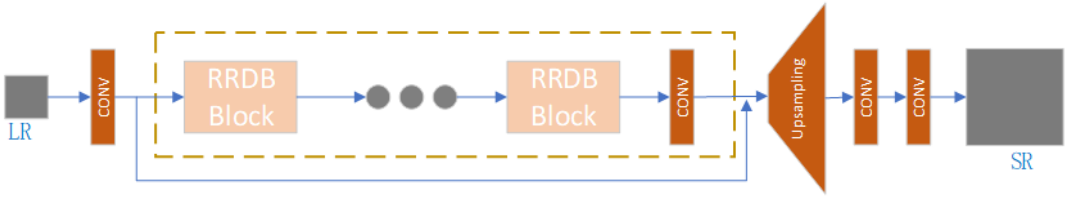


Figure 2: The overall architectural framework of Generator.

component of discriminator D assesses the global relative authenticity of the image, while the decoding component re-identifies the relative authenticity at the pixel level. The relative adversarial loss is employed to update the model parameters and improve the quality of the reconstructed image. The total loss function comprises L_1 loss, perceptual loss and relative adversarial loss. This approach enables the reconstruction of substantial high-frequency information, thereby improving the subjective visual quality of the Super-resolved images.

3.1. Network architectures

In SOAGAN, the generator employs the same generator network architecture as ESRGAN, which consists of multiple cascaded modules of Residual-in-Residual Dense Block (RRDB), as depicted in Fig. 2. Here, the upsampling module is implemented using a sub-pixel convolution network. Given an LR image, we initially extract shallow features through a 3×3 convolution, and expand the number of channels to facilitate the capture of deep features.

$$\mathrm{F}_0 = \phi(\mathrm{I}, \delta), \tag{1}$$

Subsequently, these deep features are extracted via a cascade network composed of RRDB modules. Fusing these deep features withe the shallow features enhances the model training process, while maintaining unchanged dimensions during the feature extraction phase.

$$\begin{aligned} \mathrm{F_d} &= \psi(\mathrm{F}_0), \\ \mathrm{F_d} &= \eta(\Psi_{10}(\cdots(\Psi_1(\mathrm{F}_0)))), \end{aligned} \tag{2}$$

The extracted deep features are then fed into the reconstruction module for upsampling, and an SR image is obtained after a two-layer convolution.

$$\begin{aligned} \bar{\mathrm{I}} &= \varphi(\mathrm{I}, \mathrm{F}_0, \mathrm{F_d}), \\ &= \varphi_o(\varphi_{u1}(\mathrm{F_d} + \mathrm{F}_0)) + \varphi_{u2}(\mathrm{I_{LR}}), \end{aligned} \tag{3}$$

To improve the subjective visual quality of the reconstructed image, the discriminator employs a U-Net network with skip connections. Furthermore, a covariance-based second-order attention module is integrated into the skip connection process to learn the correlation of intermediate features and improve feature representation. In the attention fusion module, the features from the decoder network are first upsampled and then processed with a 1×1 convolution to align the channel dimension with that of the encoder part. Finally, the features are concatenated along the channel dimension and output through convolution, reducing the channel dimension to half of the preceding layer. The encoding part of the U-Net discriminator outputs the relative authenticity of the global image, while the decoding part re-performs pixel-level relative authenticity discrimination. The re-identification of pixel-level relative authenticity is provided as feedback to the generator to restore more realistic image details. The covariance-based second-order

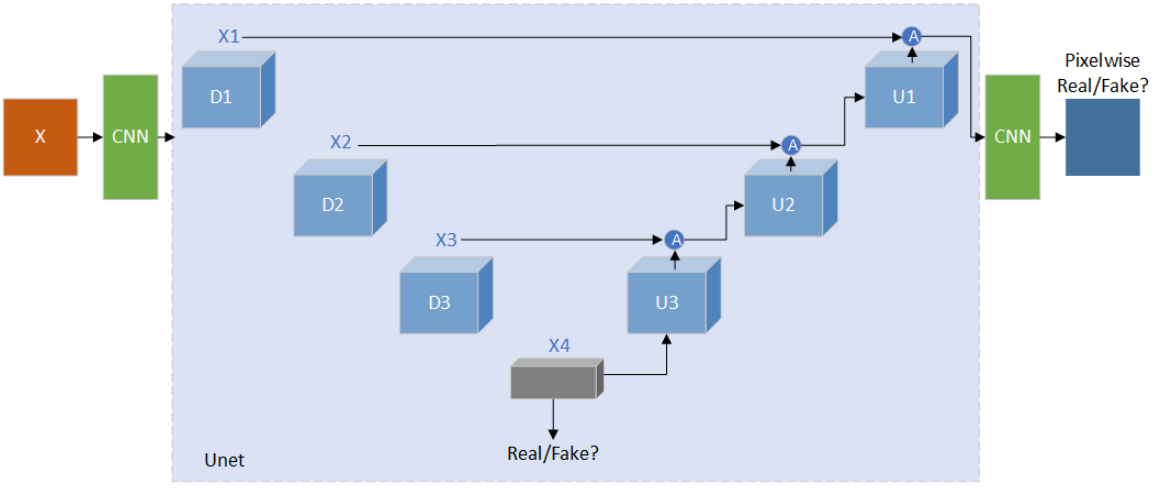


Figure 3: The overview structure of our proposed discriminator.



Figure 4: Covariance-based second order channel attention block (SOCA).

channel attention module employs the second-order channel attention (SOCA) module from the SAN(Dai et al., 2019) method. Firstly, feature covariance is calculated to investigate feature distribution, thereby enhancing the network's discriminative capabilities. To thoroughly explore the correlation between features, global covariance pooling is utilized instead of the global average pooling operation employed in the RCAN channel attention network. As shown in Fig. 4, the input feature in the SOCA module first obtains a $C \times 1 \times 1$ feature description through global covariance pooling (HGCP). Subsequently, the channel dimension of the feature is reduced and the computational complexity of the model is decreased through a convolution network with parameter W_D. The channel dimension is then restored by using a convolution network with parameter W_U. Finally, the weight value is output by the sigmoid function f and multiplied with the initial input feature via dot product to obtain the enhanced representation. An ablation study demonstrates the effectiveness of this approach.

3.2. Relative discriminator

In order to enhance the subjective visual quality of the generated images, as depicted in Fig. 3, SOAGAN not only utilizes the U-Net network as a discriminator, but also introduces a global and pixel-level relative adversarial loss based on RaGAN to further augment the discriminator's capability in identifying authenticity. Unlike the the standard GAN network, which predicts the authenticity probability of an image, the relative discriminator is capable of predicting the relative authenticity of a true image x_r in comparison to a false image x_f generated by the generator, and vice versa. The U-Net encoder and decoder are employed to estimate the relative authenticity probabilities at both the coding and decoding networks, and is formulated as follows:

$$L_D = L^{Ra}_{D_{enc}} + L_{D^{Ra}_{dec}}, \tag{4}$$

The total relative adversarial loss incurred by the coding network is expressed as follows:

$$\begin{aligned} D_{ra}(x_r, x_f) &= \delta(c(x_r) - E_{x_f}[c(x_f)])1, \\ D_{ra}(x_f, x_r) &= \delta(c(x_f) - E_{x_r}[c(x_r)])0, \\ L^{Ra}_{D_{enc}} &= -E_{x_r}[log(D_{ra}(x_r, x_f))] - E_{x_f}[log(1 - D_{ra}(x_f, x_r))] \end{aligned} \tag{5}$$

E_{x_f} denotes a batch of data averaging, $x_f = D(x_i)$, x_i represents the input low resolution image, δ represents sigmoid function, and $c(x)$ represents the output of the discriminate network. The corresponding decoding network relative

adversarial loss $L^{Ra}_{D_{dec}}$, using the mean of all pixel prediction results:

$$L^{Ra}_{D_{dec}} = -E_{x_r}[\sum_{i,j} log(D_{ra}(x_r, x_f))] - E_{x_f}[\sum_{i,j} log(1 - D_{ra}(x_f, x_r))], \tag{6}$$

$[D_{ra}(x_r, x_f))]_{i,j}$ and $[D_{ra}(x_f, x_r))]_{i,j}$ represent the prediction of pixels (i, j). The output of each pixel comes from the prediction results after the fusion of high-order and low-order features, which is realized by the skip connection and the second-order channel attention mechanism of covariance-based. The corresponding generator relative adversarial loss is as follows:

$$\begin{aligned} L_G &= L^{Ra}_{G_{enc}} + L^{Ra}_{G_{dec}}, \\ L^{Ra}_{G_{enc}} &= -E_{x_f}[log(D_{ra}(x_f, x_r))] - E_{x_r}[log(1 - D_{ra}(x_r, x_f))], \\ L^{Ra}_{G_{dec}} &= -E_{x_f}[\sum_{i,j} log[D_{ra}(x_f, x_r)]_{i,j}] - E_{x_r}[\sum_{i,j} log[1 - D_{ra}(x_r, x_f)]_{i,j}], \end{aligned} \tag{7}$$

The relative adversarial loss of the generator and the target loss of the discriminator present a symmetric form to focus on the global and local details of the reconstructed image, thus the generated image looks more realistic and natural.

3.3. Loss function

To generate an image that better aligns with subjective human perception, the loss function incorporates a combination of the L_1 content loss, perceptual loss L_{percep} , and relative adversarial loss L_G combination. The $L1$content loss is defined as:

$$\begin{aligned} L_1 &= E_{x_i} \left\| G(x_i) - y \right\|_1, \\ L_{percep} &= E_{x_i} \left\| \phi(G(x_i)) - \phi(y) \right\|_2^2, \end{aligned} \tag{8}$$

where $\phi(\cdot)$ is the VGG network. According to the method of ESRGAN network, the feature before activation layer is employed to calculate the perceptual loss. x_i is the input low-resolution image, and y is the target image. In summary, the total objective function L_{Gt} for the generator is as follow:

$$L_{Gt} = L_{percep} + \lambda L_G + \eta L_1, \tag{9}$$

4. Experiments

In this section, the details of model training are described first, and then the comparison results of visual and objective indicators between this algorithm and some represent super-resolution reconstruction methods with the basic test set are given. What's more, ablation experiments were carried out to verify the effectiveness of the fusion covariance-based second-order attention (SOCA) module in the discriminator.

4.1. Datasets and Metrics

Following previous works of (Li et al., 2022; Zhou et al., 2023), we merge the public training images from Flickr2K (Lim et al., 2017a) and DIV2K (Agustsson and Timofte, 2017) to train our network for various scales. Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2010), BSD100 (Martin et al., 2001), Urban100 (Huang et al., 2015) and Manga109(Matsui et al., 2017) are employed to validate our model performances. Besides, the average peak signal-to-noise ratio (PSNR) and structural similarity (Wang et al., 2004) (SSIM) on the Y (i.e., luminance) channel of YCbCr space are adopted to evaluate the effectiveness of our model. We apply random rotations of 90°, 180°, 270° and horizontal flipping for data augmentation in training. Following SRGAN and ESRGAN, all experiments were performed on a scaling factor of 4. In this paper, we adopt the DIV2K dataset with 800 high resolution images as training dataset and construct pairs of training images by bicubic interpolation algorithm. In training process, random 90-degree rotation, horizontal and vertical flip are used to enhance the data, and the size of HR image is set to 128×128, which reduces the cost of model calculation. We test the SR performance with the metric indicator PSNR and SSIM on five standard datasets: Set5, Set14, Urban100, BSD100, Manga109. Following previous work of LCSCNet(Yang et al., 2020), CASGCN(Yang and Qi, 2021), RCAN(Zhang et al., 2018a), only Y channel on the YCbCr color space is utilized to calculate the value of PSNR/SSIM. In addition, Blau et al.(Blau and Michaeli, 2017) proved that the

Table 1
Quantitative comparison of SOAGAN with different methods.

Methods	Set5	Set14	BSD100	Urban100	Manga109
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Ours	30.25/0.8444	26.37/0.7843	25.44/0.6545	24.60/0.7376	28.41/0.8550
ESRGAN	30.44/0.8523	26.28/0.7852	25.33/0.6533	24.35/0.7340	28.41/0.8595
SRGAN	29.92/0.8478	26.52/0.7861	25.57/0.6566	24.40/0.7301	28.11/0.8632
SRResNet	32.18/0.8955	28.58/0.7861	27.59/0.7381	26.11/0.7871	30.43/0.9086
EDSR	32.48/0.8994	28.80/0.7861	27.73/0.7438	26.64/0.8038	31.00/0.9156
FASRGAN	30.04/0.8412	26.20/0.7861	25.29/0.6466	24.44/0.7325	28.10/0.8486
SRCNN	30.20/0.8603	27.38/0.7861	26.78/0.7090	24.28/0.7140	27.07/0.8451
BSRGAN	27.67/0.8081	25.89/0.7861	25.60/0.6538	23.38/0.6916	25.96/0.8249
Bicubic	28.40/0.8113	26.08/0.7861	25.95/0.6693	23.13/0.6578	24.87/0.7867

Table 2
Qualitative comparison of SOAGAN with different methods.

Methods	Set5	Set14	BSD100	Urban100	Manga109
	PI	PI	PI	PI	PI
Ours	3.4639	2.6597	2.2566	3.4933	3.3237
ESRGAN	3.6543	2.8652	2.4188	3.7112	3.4623
SRGAN	3.4071	2.8748	2.2917	3.4199	3.3137
SRResNet	5.9872	5.3449	5.2984	5.1546	4.8784
EDSR	6.0300	5.3141	5.1608	4.9588	4.7320
FASRGAN	3.2318	2.7174	2.2585	3.4651	3.2754
SRCNN	6.6513	6.1827	6.0389	6.1287	5.7505
BSRGAN	4.4798	3.6713	3.1784	4.1169	4.0060
Bicubic	7.3184	7.0027	6.9066	6.9027	6.6632

visual perception effect of the reconstructed image is not always improved with the increase of PSNR, so we introduce Perceptual Index (PI) to evaluate the SR results that is more in line with human subjective visual quality and it combines the Ma(Mittal et al., 2013) and NIQE(Ma et al., 2017): PI=1/2 * ((10-Ma)+NIQE). The lower value of PI means the better visual quality of reconstructed image and closer to the ground truth. The whole training process is divided into two stages. Firstly, the PSNR-oriented model is trained by L_1 loss, and the learning rate is set to 2×10^{-4}. The learning rate is halved by every 2×10^5 iteration, and the total number of iterations is 1×10^6. Then, the trained model parameters are used to initialize the generator to train our model. The hyperparameters $\lambda = 5\times10^{-3}, \eta = 1\times10^{-2}$ of the loss L_{Gt} , the learning rate of generator is set to 2×10^{-4}, of discriminator is set to 2×10^{-5} and halved at [50k, 100k, 200k, 300k] iteration, the total iteration in the experiment is 4×10^5. In this paper, Adam optimizer is used to optimize the model parameters, and the corresponding parameters $\beta_=0.9, \beta_2 = 0.99$. The generator and discriminator network are iteratively updated, and the batch size is set to 8. Since we mainly focus on improving the ability of discriminator to identify the authenticity of SR image texture, the generator still uses the Residual in Residual Dense Block (RRDB) module in the ESRGAN network as the backbone network. Sub-pixel convolution is used for up-sampling at the end of the network, and then the SR image is outputted through a layer of convolution. We use Pytorch to implement our model with a NVIDIA V100 GPU.

4.2. Comparison with Other Methods

We juxtapose our results with several representative models that focus on PSNR and visual quality, such as Bicubic, SRCNN(Dong et al., 2014), SRResnet(Ledig et al., 2017), EDSR(Lim et al., 2017b), SRGAN(Ledig et al., 2017), ESRGAN(Wang et al., 2018b), FASRGAN(Yan et al., 2022), BSRGAN(Zhang et al., 2021a). As demonstrated in Table 1, our proposed method exhibits a certain degree of improvement in terms of PSNR/SSIM metrics when compared to SRGAN, ESRGAN, and BSRGAN, indicating its capability to reconstruct more accurate images.

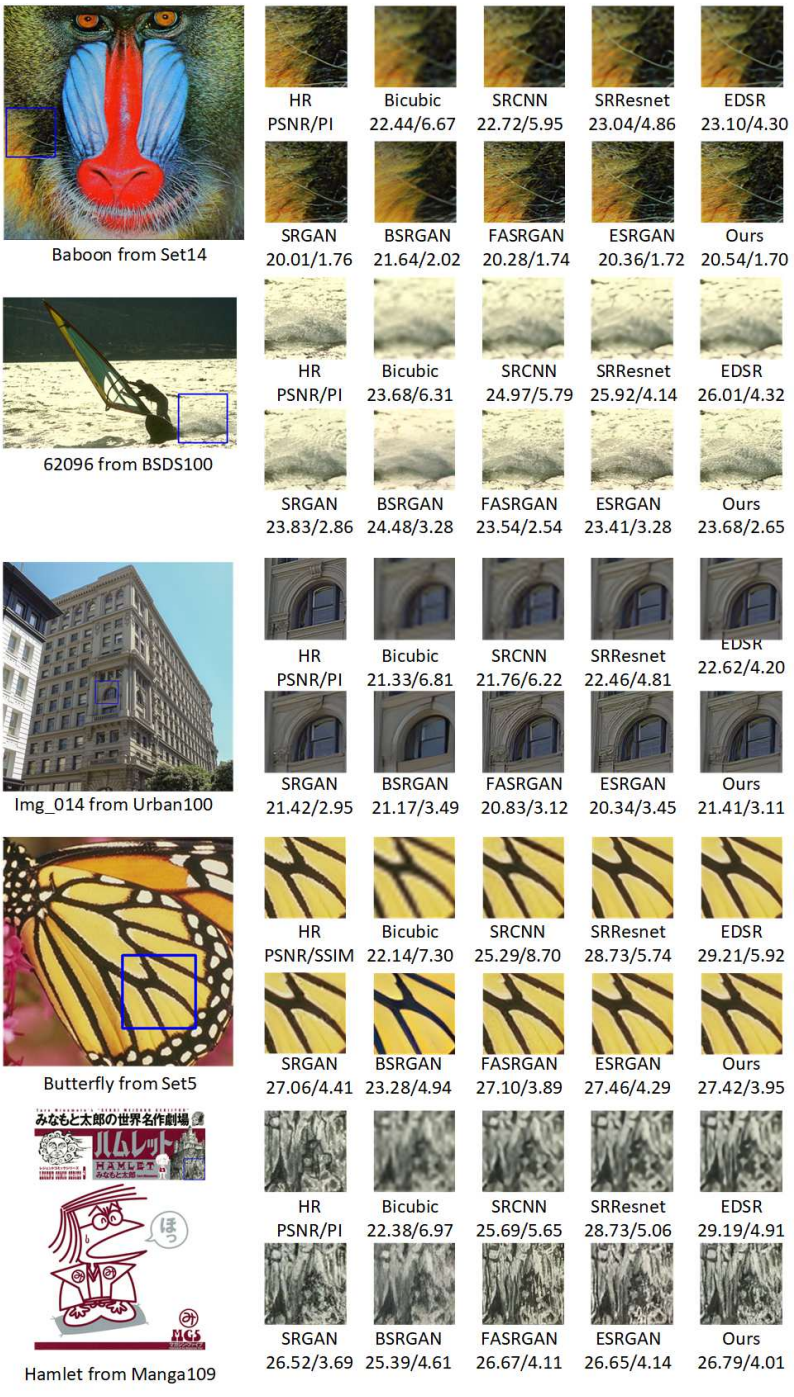


Figure 5: Comparative visual comparison of diverse super-resolution (SR) methodologies.

However, when juxtaposed with PSNR-based algorithms such as SRCNN, EDSR, and SRResnet, our method still demonstrates relatively high PSNR values. In comparison to the ESRGAN network, our algorithm generally achieves superior PSNR/SSIM indices, despite both networks employing the same generator. This further underscores the significance of introducing a UNet discriminator with global and local relative adversarial loss. Specifically, the U-Net network is utilized to construct a discriminator capable of performing global and local pixel-wise relative authenticity

discrimination, providing feedback to update the generator network parameters and thereby enhancing the quality of the reconstructed image. As shown in Table 2, our method outperforms the ESRGAN model in terms of PI index, achieving remarkable results in most cases, and indicating a higher visual quality of the reconstructed image that closely aligns with human subjective perception. We also conduct a comparative analysis of the subjective visual quality of the reconstructed images with state-of-the-art (SOTA) algorithms. As illustrated in Fig. 5, the first column displays the ground truth image from the test dataset, with HR denoting the corresponding target image, while the remaining columns present the reconstructed results of each method. Fig. 5 reveals that in comparison to previous methods, our algorithm could recover more texture and details, resulting in reconstructed images that are more realistic and natural without excessive smoothing and sharpening. For example, in the case of the Baboon image in Fig. 5, our algorithm could clearly reconstructs Baboon's beard, without introducing additional noise and artifacts, nor does it generate extraneous beard textures. In contrast, ESRGAN and FASRGAN reconstruct an additional beard that deviates from the real image. Furthermore, the images generated by EDSR and SRResnet methods appear overly smooth and unnatural, with blurred subjective visual quality. The Img014 and Hamlet images in Fig. 5 also demonstrate that compared to SRResnet and EDSR methods, our approach could reconstruct more accurate and realistic architectural structures. Additionally, ESRGAN introduces multiple noise points above the left side of Img014, damaging the reconstructed window edge structure. However, our method could effectively eliminates this drawback and outperforms the ESRGAN network in terms of quantitative indicators. In summary, although the reconstructed images generated by SRResnet, EDSR and SRCNN exhibit higher PSNR values than ours, our method demonstrateds superior performance in terms of the PI indicator, closely aligning with human visual quality, and more accurately replicating the texture of the real target image.

4.3. Discussions

To validate the efficacy of the fusion covariance-based second-order attention mechanism, ablation experiments are meticulously designed. These encompassed the discriminator model SOCA + UNet incorporating second-order attention, the discriminator model UNet devoid of any attention mechanism, and the discriminator model CAL + UNet employing global average pooling with a first-order channel attention framework. As depicted Fig. 4, global average pooling is utilized as a substitute for covariance pooling, constituting the channel attention layer (CAL) module. As illustrated in Fig. 6, the first column represents the ground truth image, while the right side displays the restored image with the network trained by corresponding identifier, wit HR designated as the target image for the respective region. It is evident from Fig. 6 that the covariance-based second-order attention mechanism could reconstruct the target image with rich and precise texture details. For example, in the case of Img042, there exists a significant discrepancy in building orientation reconstruction between the model without attention and the one with utilizing global average pooling for first-order attention. Conversely, our proposed method achieves largely accurate image reconstruction. Furthermore, the image labeled 119082 in Fig. 6 reveals that both UNet and CAL + UNet introduce extraneous white texture details, whereas our method yields improved and comparatively more accurate results. The experimental results substantiate the effectiveness of integrating covariance-based second-order channel attention into the discriminator.

5. Conclusion

In this paper, we propose a generative adversarial network (SOAGAN) that incorporated covariance-based second-order attention mechanism to address the challenge of single image super-resolution(SISR). Specifically, we adopt the UNet architecture as the discriminator, enhanced with re-identification capabilities, to facilitate the comprehensive reconstruction of image details. To enable pixel-level feedback for refining the generator's network parameters, we employ both global and local relative adversarial losses. Extensive experimental results substantiate that our proposed method significantly enhances the visual quality of reconstructed images, yielding results that are more natural, texturally nuanced and realistic. Nevertheless, the exploration of a more effective attention model remains a promising avenue for not only further improving the visual quality but also enhancing the accuracy of reconstructed images, warranting continued research in this direction.

CRediT authorship contribution statement

Duwei Hua: Conceptualization, Methodology, Software, Formal analysis, Writing — original draft, Writing — review & editing, Data curation. **Hanling Li:** Supervision, Validation.

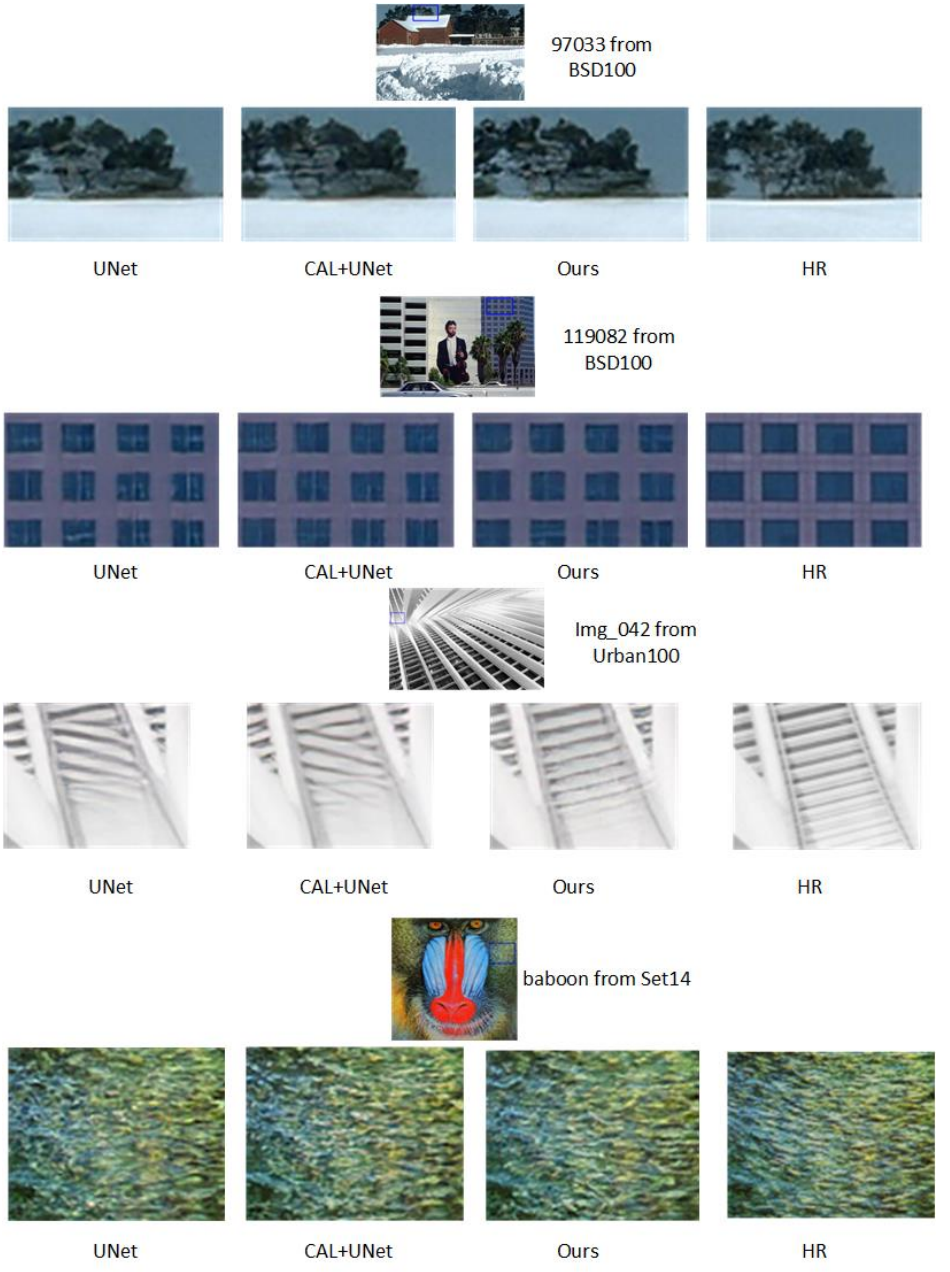


Figure 6: Visual comparison within the ablation study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The research of this paper is supported by Wuhan Institue of Shipbuilding Technology 2026 School-level Scientific Research Project (2026z07).

Data Available

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Author contributings

Duwei Hua: Conceptulization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing-original draft, Visualization; Hanling Li: Formal analysis, Investigation, Supervision, Validation

References

Agustsson, E., Timofte, R., 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. workshops, pp. 126–135.

Ahn, N., Yoo, J., Sohn, K.A., 2024. Data augmentation for low-level vision: Cutblur and mixture-of-augmentation. Int. J. Comput. Vis 132.

Bevilacqua, M., Roumy, A., Guillemot, C.M., Alberi-Morel, M.L., 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding, in: Proc. Brit. Mach. Vis. Conf., pp. 1–10.

Blau, Y., Michaeli, T., 2017. The perception-distortion tradeoff. arXiv abs/1711.06077.

Cao, J., Chen, Q., Guo, J., Shi, R., 2020. Attention-guided context feature pyramid network for object detection. arXiv abs/2005.11475.

Chen, H.G., He, X.H., Qing, L.B., Wu, Y.Y., Ren, C., Sheriff, R., Zhu, C., 2022. Real-world single image super-resolution: A brief review. Information Fusion 79, 124–145.

Dai, S., Han, M., Xu, W., Wu, Y., Gong, Y., Katsaggelos, A.K., 2009. Softcuts: a soft edge smoothness prior for color image super-resolution. IEEE Trans. Image process. 18, 969–981.

Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L., 2019. Second-order attention network for single image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 11065–11074.

Dong, C., Loy, C.C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution, in: Proc. Eur. Conf. Comput. Vis., Springer. pp. 184–199.

Dong, C., Loy, C.C., Tang, X., 2016. Accelerating the super-resolution convolutional neural network, in: Proc. Eur. Conf. Comput. Vis., Springer. pp. 391–407.

Dong, W.S., Zhang, L., Shi, G.M., Wu, X.L., 2011. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. IEEE Trans. Image process. 20, 1838–1857.

Freeman, W., Pasztor, E., Carmichael, O., 2000. Learning low-level vision. Int. J. Comput. vis. 40, 25–47.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Proc. 27th Int. Conf. Neural Inf. Process. Syst. 27.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017a. Densely connected convolutional networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4700–4708.

Huang, J.B., Singh, A., Ahuja, N., 2015. Single image super-resolution from transformed self-exemplars, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5197–5206.

Huang, Y., Shao, L., Frangi, A., 2017b. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. workshops, pp. 6070–6079.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20, 1254–1259.

Jolicoeur-Martineau, A., 2018. The relativistic discriminator: a key element missing from standard gan. arXiv preprint arXiv:1807.00734 .

Keys, R.G., 2003. Cubic convolution interpolation for digital image processing. IEEE Trans. Acoust. Speech Signal Process. 29, 1153–1160.

Kim, J., Lee, J.K., Lee, K.M., 2016a. Accurate image super-resolution using very deep convolutional networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1646–1654.

Kim, J., Lee, J.K., Lee, K.M., 2016b. Accurate image super-resolution using very deep convolutional networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1646–1654.

Kim, J., Lee, J.K., Lee, K.M., 2016c. Deeply-recursive convolutional network for image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1637–1645.

Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H., 2017. Deep laplacian pyramid networks for fast and accurate super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 624–632.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4681–4690.

Li, Z., Liu, Y., Chen, X., Cai, H., Gu, J., Qiao, Y., Dong, C., 2022. Blueprint separable residual network for efficient image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 833–843.

Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K., 2017a. Enhanced deep residual networks for single image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. workshops, pp. 136–144.

Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K., 2017b. Enhanced deep residual networks for single image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. workshops, pp. 136–144.
Liu, X., Xia, T., Wang, J., Lin, Y., 2016. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. arXiv abs/1603.06765.
Ma, C., Rao, Y., Cheng, Y., Chen, C., Lu, J., Zhou, J., 2020. Structure-preserving super resolution with gradient guidance, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Computer Vision Foundation / IEEE. pp. 7766–7775.
Ma, C., Yang, C.Y., Yang, X., Yang, M.H., 2017. Learning a no-reference quality metric for single-image super-resolution. Comput. Vis. Image Underst. 158, 1–16.
Marquina, A., Osher, S.J., 2008. Image super-resolution by tv-regularization and bregman iteration. J. Sci. Comput. 37, 367–382.
Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proc. Int. Conf. Comput. Vis., IEEE. pp. 416–423.
Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K., 2017. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools Appl. 76, 21811–21838.
Mittal, A., Soundararajan, R., Bovik, A.C., 2013. Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. 20, 209–212.
Ning, Q., Dong, W.S., Shi, G.M., Li, L.D., Li, X., 2020. Accurate and lightweight image super-resolution with model-guided deep unfolding network. IEEE J. Sel. Topics Signal Process. 15, 240–252.
Rasti, P., Uiboupin, T., Escalera, S., Anbarjafari, G., 2016. Convolutional neural network super resolution for face recognition in surveillance monitoring, in: Articulated Motion and Deformable Objects, Springer. pp. 175–184.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., Springer. pp. 234–241.
Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1874–1883.
Shocher, A., Cohen, N., Irani, M., 2018. "zero-shot" super-resolution using deep internal learning, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3118–3126.
Soh, J., Cho, S., Cho, N., 2020. Meta-transfer learning for zero-shot super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3516–3525.
Sun, J., Xu, Z., Shum, H.Y., 2008. Image super-resolution using gradient profile prior, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., IEEE. pp. 1–8.
Tai, Y., Yang, J., Liu, X., Xu, C., 2017. Memnet: A persistent memory network for image restoration, in: Proc. Eur. Conf. Comput. Vis., pp. 4539–4547.
Tian, C., Xu, Y., Zuo, W., Zhang, B., Fei, L., Lin, C.W., 2020. Coarse-to-fine cnn for image super-resolution. IEEE Trans. Multimedia 23, 1489–1502.
Tong, T., Li, G., Liu, X.J., Gao, Q., 2017. Image super-resolution using dense skip connections, in: Proc. Int. Conf. Comput. Vis., pp. 4799–4807.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Proc. Adv. Neural Inf. Process. Syst, pp. 5998–6008.
Wang, X., Xie, L., Dong, C., Shan, Y., 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, IEEE. pp. 1905–1914.
Wang, X., Yu, K., Dong, C., Loy, C., 2018a. Recovering realistic texture in image super-resolution by deep spatial feature transform, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Computer Vision Foundation / IEEE Computer Society. pp. 606–615.
Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C., 2018b. Esrgan: Enhanced super-resolution generative adversarial networks, in: Proc. Eur. Conf. Comput. Vis., pp. 0–0.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image process. 13, 600–612.
Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. CBAM: convolutional block attention module, in: Proc. Eur. Conf. Comput. Vis., Springer. pp. 3–19.
Xu, X.Q., Wang, Z.Y., Shi, H., 2021. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. arXiv preprint arXiv:2103.12716 .
Yan, Y., Liu, C., Chen, C., Sun, X., Jin, L., Peng, X., Zhou, X., 2022. Fine-grained attention and feature-sharing generative adversarial networks for single image super-resolution. IEEE Trans. Multim. 24, 1473–1487.
Yang, W., Zhang, X., Tian, Y., w. Wang, Xue, J., Liao, Q., 2020. Lcscnet: Linear compressing-based skip-connecting network for image super-resolution. IEEE Trans. Image Process. 29, 1450–1464.
Yang, X., Zhang, Y.F., Li, T., Guo, Y.Q., Zhou, D., 2021. Image super-resolution based on the down-sampling iterative module and deep cnn. Circuits Syst. Signal Process. 40, 3437–3455.
Yang, Y., Qi, Y., 2021. Image super-resolution via channel attention and spatial graph convolutional network. Pattern Recognit. 112, 107798.
Yıldırım, D., Güngör, O., 2012. A novel image fusion method using ikonos satellite images. Journal of Geodesy and Geoinformation 1, 75–83.
Zeyde, R., Elad, M., Protter, M., 2010. On single image scale-up using sparse-representations, in: Proc. Int. Conf. Curves Surfaces, Springer. pp. 711–730.
Zhang, K., Gool, L.V., Timofte, R., 2020. Deep unfolding network for image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3217–3226.
Zhang, K., Liang, J., Gool, L.V., Timofte, R., 2021a. Designing a practical degradation model for deep blind image super-resolution, in: Proc. IEEE/CVF Int. Conf. Comput., IEEE. pp. 4771–4780.
Zhang, L., Wu, X.L., 2006. An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE Trans. Image process. 15, 2226–2238.
Zhang, Y., Li, K., Li, K., Fu, Y., 2021b. MR image super-resolution with squeeze and excitation reasoning attention network, in: Proc. Conf. Comput. Vis. Pattern Recognit., pp. 13425–13434.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., 2018a. Image super-resolution using very deep residual channel attention networks, in: Proc. Eur. Conf. Comput. Vis., pp. 286–301.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y., 2018b. Residual dense network for image super-resolution, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2472–2481.

Zhao, M., Wu, G.C., Li, Y.P., Hao, X.Y., Fang, L., Liu, Y.B., 2018. Cross-scale reference-based light field super-resolution. IEEE Trans. Computational Imaging 4, 406–418.

Zhou, L., Cai, H., Gu, J., Li, Z., Liu, Y., Chen, X., Qiao, Y., Dong, C., 2023. Efficient image super-resolution using vast-receptive-field attention, in: Proc. Eur. Conf. Comput. Vis. Workshop, Springer. pp. 256–272.