Feature Super-Resolution: Make Machine See More Clearly

Weimin Tan, Bo Yan*, Bahetiyaer Bare
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{wmtan14, byan, bahtiyarbari16}@fudan.edu.cn

Abstract

Identifying small size images or small objects is a notoriously challenging problem, as discriminative representations are difficult to learn from the limited information contained in them with poor-quality appearance and unclear object structure. Existing research works usually increase the resolution of low-resolution image in the pixel space in order to provide better visual quality for human viewing. However, the improved performance of such methods is usually limited or even trivial in the case of very small image size (we will show it in this paper explicitly).

*In this paper, different from image super-resolution (ISR), we propose a novel super-resolution technique called **feature super-resolution (FSR)**, which aims at enhancing the discriminatory power of small size image in order to provide high recognition precision for machine. To achieve this goal, we propose a new Feature Super-Resolution Generative Adversarial Network (FSR-GAN) model that transforms the raw poor features of small size images to highly discriminative ones by performing super-resolution in the feature space. Our FSR-GAN consists of two subnetworks: a feature generator network **G** and a feature discriminator network **D**. By training the **G** and the **D** networks in an alternative manner, we encourage the **G** network to discover the latent distribution correlations between small size and large size images and then use **G** to improve the representations of small images. Extensive experiment results on Oxford5K, Paris, Holidays, and Flick100k datasets demonstrate that the proposed FSR approach can effectively enhance the discriminatory ability of features. Even when the resolution of query images is reduced greatly, e.g., 1/64 original size, the query feature enhanced by our FSR approach achieves surprisingly high retrieval performance at different image resolutions and increases the retrieval precision by 25% compared to the raw query feature.*

*This work was supported by NSFC (Grant No.: 61522202; 61772137).

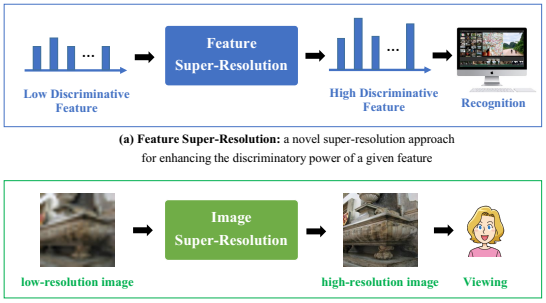
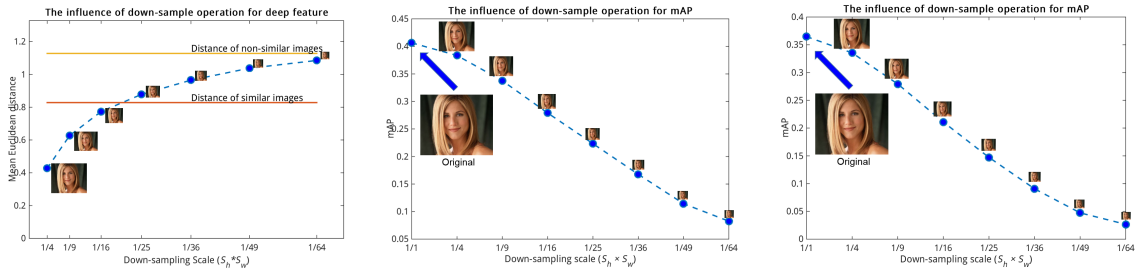


Figure 1. **Feature super-resolution (FSR)** versus **image super-resolution (ISR)**. (a) We propose a novel super-resolution technique called feature super-resolution (FSR), which aims at enhancing the discriminatory power of a given representation (extracted from low-resolution images or small objects) in order to providing high recognition precision for machine. (b) Image super-resolution as a popular technique aims at increasing the resolution of a given image in order to providing better visual quality for human viewing.

1. Introduction

The powerful deep learning framework makes numerous great classification models presented, such as VGG16 [18], GoogLeNet [19], ResNet [7], SeNet [8], *etc.* These models achieve an amazing recognition accuracy on Imagenet dataset [2], even better than human beings do. Actually, they indeed work well on large size images with good-quality appearance and rich object structure. However, they usually fail to identify very small size images since discriminative representations are difficult to learn from their low-resolution and noise representation. Small size images or objects are very common in many real-world scenarios such as small pedestrians in surveillance video, small faces in the crowd, traffic signs, small objects, *etc.* Small size image recognition is much more challenging than normal image recognition and there are rare good solutions so far.

Current research works such as image super-resolution (ISR) focus on increasing the resolution of a given image. Its most common application is to provide better visual quality after resizing a digital image for human viewing,



(a) (b) (c)

Figure 2. We explore the impact of different low-resolutions on deep representations, which is evaluated on Oxford5K [14] dataset. The S_h and S_w denote the height and width down-sampling ratios, respectively. We set S_h equals S_w. (a) decreasing image size seriously impacts the deep representation. When the down-sample scale is smaller than 1/20, the feature distance between low-resolution images and high-resolution ones is greater than the distance of similar images. (b) image retrieval results of searching Oxford5K dataset. (c) large-scale image retrieval results of searching Oxford5K plus Flick100k. (b) and (c) show that with the decrease of the resolution of query images, the retrieval precision is decreased rapidly.

as shown in Fig. 1(b). In recent years, numerous image super-resolution approaches have been proposed to restore high frequency information in order to generate high quality images with rich details, and have achieved great success [21, 3, 10, 11, 12, 5, 20]. This process is referred to pixel-space field enhancement in the literature. Those ISR based approaches are able to recover the object details in small size image and can improve the identification accuracy in some extent. In the experiment, we will compare ISR based approaches with our FSR algorithm and demonstrate their limitations. Intuitively, as shown in Fig. 1, if we perform super-resolution for machine recognition instead of human viewing, the paradigm should change accordingly. Therefore, for machine recognition, we propose feature super-resolution for improving the discriminative ability of features, as illustrated in Fig. 1(a).

In order to practically explore the necessity and feasibility of feature super-resolution, we conduct several experiments to understand the impact of low-resolution image on deep representations. For these experiments, we evaluate the effect of down-scaling operation on the deep representations extracted by using the popular VGG16 model [18] on Oxford5K dataset [14]. The Oxford5K dataset contains 5,063 images and 55 query images. For convenience, the deep representations mentioned below refers to the neural activations in the 36^{th} layer of VGGnet model [18] if there is no special notification.

Firstly, we summarize the experimental results of mean Euclidean distance between the deep features extracted from high-resolution images and their corresponding low-resolution ones. The low-resolution images are obtained by performing uniform down-sampling operation with different scaling ratios. Figure. 2(a) shows the change of deep feature with the decrease of image resolution. From Fig. 2(a) we observe that with the decrease of image resolution, the difference between the deep features extracted from low-resolution images and high-resolution ones is growing wider. This result implies that decreasing image resolution is capable of impacting deep representation, though the powerful VGG16 model [18] is carefully trained on the large-scale Imagenet dataset [2] and adds some useful training tricks.

Furthermore, we conduct image retrieval experiment to understand how low-resolution images affect matching/retrieval accuracy when using deep features. We summarize the experimental results of the mean average precision (mAP) as a function of the down-scaling ratio in Fig. 2(b) and Fig. 2(c). The results demonstrate that with the decrease of the resolution of the image, the retrieval precision is decreased rapidly. This phenomenon is reasonable because the low-resolution image has lost many detail information, which results in failing to extract discriminative features even using the powerful very deep convolutional neural network.

From Fig. 2(a) to Fig. 2(c) we have known that the low-resolution images not only impact the extracted deep features, but also seriously decrease the matching/retrieval accuracy. To find a solution to the problem, we conduct the third experiment to further understand the relationship between the deep features extracted from different resolution images. Specifically, we calculate the variance of Euclidean distance for the same down-scaling ratio. Table 1 summarizes the experiment result, which reports that the variances of Euclidean distance across different down-scaling ratios are very small. This result means that the extracted deep features are changed regularly with down-scaling ratios, $i.e.$, the change of deep features mainly depends on the amount of information lost instead of specific image content. This key observation provides an important basis for our FSR approach to solve the problem mentioned above.

Based on the key observation mentioned above, we propose a novel Feature Super-Resolution Generative Adver-

Table 1. Variances of Euclidean distance across different down-scaling ratios. We conduct this experiment to further understand the relationship between the deep features extracted from different resolution images. The result reports that the extracted deep features are changed regularly with down-scaling ratios.

Down-scaling ratio	1/4	1/9	1/16	1/25	1/36	1/49	1/64	Average
Variance of Euclidean distance	0.0070	0.0103	0.0107	0.0101	0.0091	0.0078	0.0068	0.0088

sarial Network (FSR-GAN) model to enhance the discriminatory ability of the representation of small size images, achieving similar attributes as images with clear appearance and thus more discriminative for better identification. Our FSR-GAN consists of two subnetworks: a feature generator network **G** and a feature discriminator network **D**. The **G** is a simple convolution neural network which maps the raw poor representations of small size images to highly discriminative ones by discovering the latent distribution correlations between small size and large size images, achieving "super-resolution" on the feature space. The **D** estimates the probability that a representation comes from the real data or the fake data generated by **G**. It actually provides guidance for updating **G**. Note that different from traditional Generative Adversarial Network (GAN), our proposed FSR-GAN includes a new focal loss tailored for scale-invariant feature enhancement.

In this paper, we propose the FSR concept using the framework of GAN to form a local loss function for representation enhancement. The main contributions of this work are:

- Based on the key observation, we propose a novel concept of feature super-resolution that is different from the image super resolution. This technique is expected to make a breakthrough in the challenging task of identifying small size images or objects.
- We are the first to successfully introduce the FSR concept into the GAN framework, called FSR- GAN, achieving large improvement comparing with existing approaches.
- We introduce a new focal loss for generative network, making it put more effort into hard examples with large downscales and preventing it from being affected by easy examples with small downscales, thus the optimal solution can be obtained.
- Several successful applications explicitly show that our FSR-GAN is far superior to the comparison approaches.

2. Related Work

2.1. Image Super-Resolution

Image super-resolution (ISR) approaches aim to estimate a high-resolution image from low-resolution images. Recently, convolutional neural network (CNN) based ISR methods have shown excellent performance. In Wang $et\ al.$ [21] the authors propose to combine the merits of deep CNN and sparse coding for ISR, because they observe that domain expertise represented by the sparse coding model is still valuable and can be effectively implemented with a LISTA network [21]. Dong $et\ al.$ [3] propose to train an end-to-end deep fully convolutional network with three layers for ISR. This work enables the network to learn the upscaling filters directly, which is helpful in increasing the performance in terms of speed and accuracy. Following this strategy, numerous works have proposed more deep and complex networks for improving the performance of ISR [10, 11, 12, 5, 20].

2.2. Generative Adversarial Network

Goodfellow $et\ al.$ [6] propose an interesting framework named generative adversarial network (GAN) for generating plausible-looking images. The GAN framework consists of two models: a generative model G and a discriminative model D. The model G captures the data distribution. The model D estimates the probability that a sample came from the training data rather than G. There two models will be trained simultaneously for estimating generative model. Unfortunately, the preliminary GAN is not stable in training [6]. To improve it, Arjovsky $et\ al.$ [1] propose Wasserstein GAN (WGAN) by modifying loss function and network design. Our approach benefits from WGAN. GAN has been used in variety of applications such as image super-resolution [12], unsupervised representation learning [16], image super-resolution [12], text to image synthesis [17], dialogue generation [13], machine translation [23], $etc.$

3. Our Proposed FSR-GAN

3.1. Overview

In FSR, the aim is to estimate a highly discriminative feature F^{SR} from a low-resolution input image I^{LR}. Correspondingly, we use F^{HR} to denote the high-resolution image. I^{LR} is obtained by performing down-sampling operation with different downscaling factors. Figure. 3 shows the architecture of our proposed FSR-GAN. It consists three blocks, $i.e.$, general feature extraction model, feature generative network, and feature discriminative network. The first block is to extract good representations for input images of I^{LR} and I^{HR}. Note that this block can be a traditional feature extraction model or a powerful deep neural network. In this work, we employ the trained VGG16 model [18] as

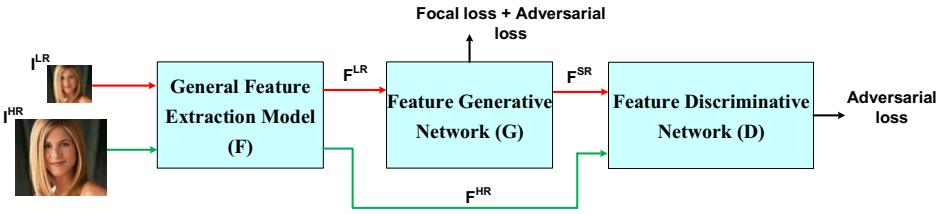


Figure 3. Overview of our FSR-GAN for implementing the proposed FSR approach. It consists three blocks, $i.e.$, general feature extraction model, feature generative network, and feature discriminative network. The first block is to extract good representations for input images of I^{LR} and I^{HR}. After extracting representations, the feature generative network G transforms the raw poor features F^{LR} of input low-resolution images to highly discriminative ones, called super representations F^{SR}. Finally, the feature discriminative network serves as a supervisor to distinguish the currently generated super representations F^{SR} for the small size images and the original representations F^{HR} from the large size images.

the feature extraction model to represent input images. We use F^{LR} and F^{HR} to denote the extracted representations of I^{LR} and I^{HR}, respectively. It can formally be written as:

$$F^{LR} = F(I^{LR}) \tag{1}$$
$$F^{HR} = F(I^{HR}) \tag{2}$$

where F denotes the general feature extraction model. Obviously, F^{LR} and F^{HR} are different in the discrimination ability.

After extracting representations, the feature generative network G transforms the raw poor features F^{LR} of input low-resolution images to highly discriminative ones, called super representations F^{SR}. It is defined as:

$$F^{SR} = G(F^{LR}) \tag{3}$$

Then, the feature discriminative network serves as a supervisor to distinguish the currently generated super representations F^{SR} for the small size images and the original representations F^{HR} from the large size images. Our ultimate goal is to train a generative network **G** that learns to transfer the poor representations of low-resolution images to super representations similar to those of high-resolution images. To achieve it, we propose a focal loss for training **G** network by considering the distribution of down sampling scales and the imbalance of examples, which is significantly different from the generative network of preliminary GAN [6, 1]. The focal loss function is described in more detail in Section 3.2.

3.2. Focal Loss Function

Goodfellow $et\ al.$ [6] propose a great and interesting idea to generate examples by unsupervised learning the input data distribution. By training the **G** and **D** networks in an adversarial way, the network **G** can successfully learn the distribution of input data.

The loss functions of the generative network and the discriminative network can formally be written as:

$$L(G) = E_{x\sim P_g}[1 - logD(x)] \tag{4}$$
$$L(D) = -E_{x\sim P_r}[logD(x)] - E_{x\sim P_g}[1 - logD(x)] \tag{5}$$

The problem of preliminary GAN [6] is that the better the classifier, the more serious the generator gradient disappears. This problem easily results in unsatisfied training result. In order to improve the stability of GAN training, Arjovsky $et\ al.$ [1] propose WGAN by modifying loss function and network design. The loss functions of the generative network and the discriminative network are defined as follows:

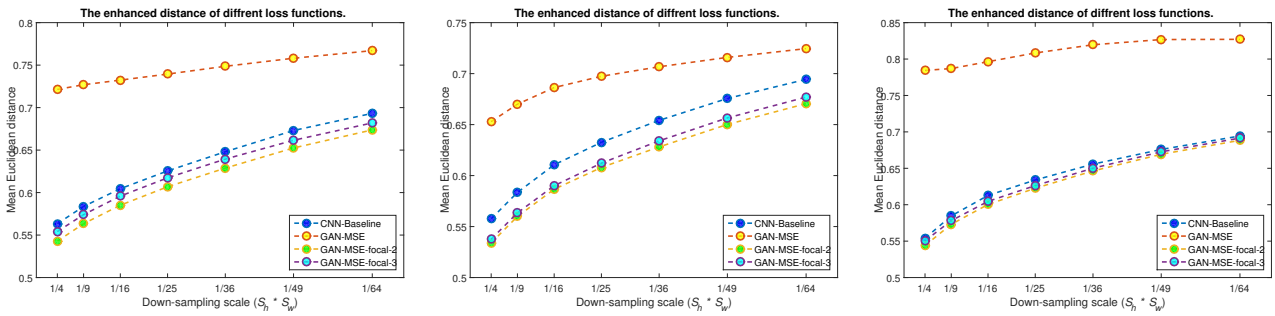
$$L(G) = -E_{x\sim P_g}[D(x)] \tag{6}$$

$$L(D) = E_{x\sim P_g}[D(x)] - E_{x\sim P_r}[D(x)] \tag{7}$$

These two functions can guide the training process of **G** and **D** networks. The smaller the loss function, the smaller the Wasserstein distance between the real distribution and the generative distribution, $i.e.$, the better the GAN training.

Naively, we can use WGAN to implement our proposed idea of representation enhancement. However, in the experiment, we find that directly using WGAN can not enhance the imputed poor representation. The possible reason is that the constraints on the generative network are too loose. Actually, some works demonstrate that by adding a stronger constraint to the generative network, it can help guide **G** to converge better. Therefore, we add a mean squared error (MSE) term to the Eq. (6). The Eq. (6) can be rewritten as:

$$L(G) = -E_{x\sim P_g}[D(x)] + \frac{1}{m}\sum_{i=1}^{m}(||F_i^{SR} - F_i^{HR}||_2) \tag{8}$$



(a) Holidays (b) Paris (c) Oxford5K

Figure 4. Demonstration of the distance between the enhanced representations and the original ones from large size images using different loss functions on Oxford5K [14], Paris [9], and INRIA Holidays [9] datasets. The S_h and S_w denote the height and width down-sampling ratios, respectively. We set S_h equals S_w. The results consistently show that by incorporating the focal loss in GAN can improve the performance of feature super-resolution. When the parameter r of focal loss in Eq. (9) equals to 2, we obtain the best results.

where m represents the number of examples.

At this point, Eq. (8) does not take into account the imbalance of examples with different down sampling scales. Inspired by the work [14] using focal cross entropy loss for dense object detection, we propose a new focal loss for representation enhancement. By incorporating the focal loss in Eq. (8), which can be rewritten as:

$$L(G) = -E_{x\sim P_g}[D(x)] + \frac{1}{m}\sum_{i=1}^{m}(||F_i^{SR} - F_i^{HR}||_2)^r \tag{9}$$

where r denotes the weight of focal loss. The larger the value of r, the greater the weight of hard examples.

Figure. 4 shows the enhanced results using different loss functions on three datasets. "CNN-Baseline" denotes that only the feature generative network is used, $i.e.$, the feature discriminative network is not used. Besides, it employs the standard MSE as loss function. "GAN-MSE-focal-2" means that the parameter r in Eq. (9) equals to 2. To demonstrate the performance of different loss functions, we calculate the distance between the enhanced representations produced by **G** network and the original ones from large size images. From Fig. 4 we observe, the GAN plus MSE loss function in Eq. (8) obtains the worst performance, even worse than "CNN-Baseline". However, by incorporating the focal loss in GAN, the performance is significantly improved, even better than "CNN-Baseline". When the parameter r of focal loss in Eq. (9) equals to 2, we obtain the best results. Finally, in the experiment, we employ Eq.(7) and Eq.(9) as loss functions for our discriminative network and generative network, respectively.

3.3. Implementation Details

Architecture of feature generative network: The feature generative network aims to generate super representations for small size images to improve identification accuracy. To achieve this goal, we design the generator as a deep CNN learning network. As shown in Table 2, our feature generative network is a normal convolutional neural network, which consists of 6 convolutional layers, 1 dropout layer, and 1 fully connected layer. The first layer has 8×8 kernel, 4 channels, and 1 stride. Note that we employ a large kernel in the first convolution layer in order to fully exploit the latent information in the input representations. All layers employ Leaky ReLU activation function [22].

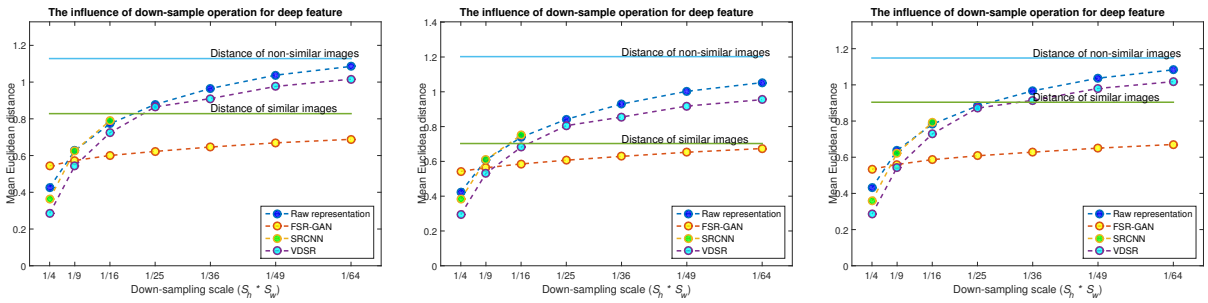
Table 2. Architecture of feature generative network

type	kernel size	stride	channel	output size
convolution	8×8	1	4	$64 \times 64 \times 4$
convolution	5×5	2	8	$32 \times 32 \times 8$
convolution	5×5	1	16	$32 \times 32 \times 16$
convolution	5×5	2	32	$16 \times 16 \times 32$
convolution	5×5	1	64	$16 \times 16 \times 64$
convolution	5×5	2	128	$8 \times 8 \times 128$
dropout(70%)				$1 \times 64 \times 128$
linear				1×4096

Table 3. Architecture of feature discriminative network

type	kernel size	stride	channel	output size
convolution	5×5	2	8	$32 \times 32 \times 8$
convolution	5×5	2	16	$16 \times 16 \times 16$
convolution	3×3	2	32	$8 \times 8 \times 32$
convolution	3×3	1	64	$8 \times 8 \times 64$
linear				1

Architecture of feature discriminative network: The feature discriminative network is to differentiate between the generated super representation for small size image and the original one from the real large size image. To achieve this purpose, we design a simple deep CNN network. As shown in Table 3, our feature generative network is a convolutional neural network, which consists of 4 convolutional layers and 1 fully connected layer. All these layers employ Leaky ReLU activation function [22] except for the final layer. Following Arjovsky $et\ al.$ [1], the final layer is a ful-



(a) Oxford5K (b) Holidays (c) Paris

Figure 5. Comparison of Euclidean distance on three datasets: Oxford5K [14], Paris [9], and INRIA Holidays [9]. Here, the Euclidean distance measures the difference between the enhanced representations produced by different approaches and the highly discriminative ones from high-resolution images. This satisfying results prove that the idea of representation enhancement is feasible, which is a good news for small size images or objects identification researchers.

ly connected layer with no activation function and outputs a value for predicting real or fake.

Parameter Settings: The loss functions of Eq.(7) and Eq.(9) are optimized using Adam algorithm with an initial learning rate of 0.0008. The parameter r of focal loss in Eq. (9) is set to 2. Typically, our FSR-GAN takes 6 epochs for training, and is capable of producing high-discriminative features. We implement our model using the tensorflow framework. We train the network using an Nvidia GPU Quadro M4000 on subsets of Oxford5K, Holidays, and Paris datasets.

4. Experimental Results

In this section, we evaluate the enhancing ability of our FSR-GAN when trained on the subsets of Oxford5K [14], Paris [9], and Holidays [9] datasets. The new ISR methods including SRCNN [4] and VDSR [10] are compared with our approach. We choose these two ISR approaches to compare because they report excellent performance in many datasets. We use S_h and S_w to denote the height and width down-sampling ratios, respectively. We set S_h equals S_w, $e.g.$, $S_h = S_w = 1/2$. Note that we only have results of this method in 1/4, 1/9, and 1/16 down-sampling ratios. Therefore, In the following experiments, we show the results of SRCNN method at these three ratios.

4.1. Experimental Setup

For our experiments we evaluate the effect of using FSR on the following datasets: Oxford5K [14], Paris [9], INRIA Holidays [9], and Flickr 100k [14]. These datasets cover a wide variety of scene types, which is helpful to comprehensively evaluate the performance of the proposed algorithm.

Oxford5K dataset [14]: consists of 5,062 high-resolution (1024 $\times$ 768) images and 55 query images (11 landmarks). This dataset is collected by searching Flickr.

Holidays dataset [9]: is a set of images which mainly contains holidays photos. This dataset contains 1,491 images, 500 queries, and 991 corresponding relevant images. It includes a very large variety of scene types.

Paris dataset [9]: consists of 6,412 images collected from Flickr by searching for particular Paris landmarks. Similar to Oxford5K, it contains 55 query images. Each query corresponds to a landmark in Paris.

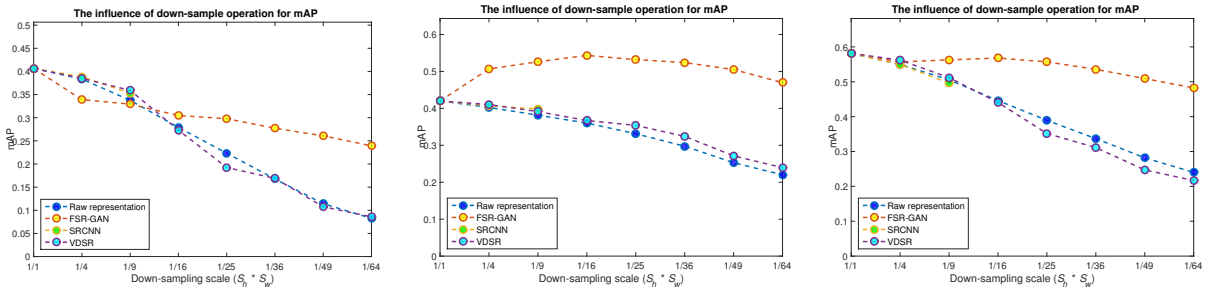
Flickr 100k dataset [14]: consists of 100,071 images collected from Flickr by searching for popular Flickr tags [14]. This dataset is used as a distractor dataset.

In our experiments, the officially provided train/test split is used for experiments. For Oxford5K dataset [14], 4,500 images are randomly selected for training and 562 images for evaluating. For Holidays dataset, we use the provided 500 queries to form the test set, and the rest as training set. For Paris dataset [9], we randomly sample 612 images to form the test query set, and use the rest as training set.

4.2. The Effectiveness of Super Representations by FSR-GAN

We calculate the Euclidean distance between the enhanced representations generated by mentioned approaches and the ones from large size images, for example, the Euclidean distance between the enhanced representations by our approach and the ones from large size images are denoted as FSR-GAN in Fig. 5. The Euclidean distances of similar images and non-similar images in original dataset are also demonstrated in Fig. 5.

We find that the proposed FSR-GAN approach can significantly reduce the gap between the representation of low-resolution images and high-resolution ones. This implies that the enhanced representations are more close to the high discriminative features extracted from high-resolution images. The discriminative ability of representation is largely enhanced by our FSR-GAN approach. Further, we observe that even the image resolution has been down-scaled to 1/64 original size, we still achieve about 0.62 distance, which is much smaller than the distance of similar images. The ISR



(a) Oxford5K (b) Holidays (c) Paris

Figure 6. Comparison of image retrieval performance on three datasets: Oxford5K [14], Paris [9], and INRIA Holidays [9].

approaches such as SRCNN [4] and VDSR [10] try to increase the discriminative ability of representation from low-resolution images by applying enhancement in pixel space. These approaches only obtain good performance at relatively large size image, $e.g.$, 1/4 scale, but it fails at other small size image, $e.g.$, 1/16, 1/25, 1/36, $etc.$

5. Applications

The surprising results on three widely used datasets [14, 9] in Section 4 have shown the effectiveness of FSR-GAN. Here, we show that the enhanced representations learned by our FSR-GAN approach also perform well in image retrieval applications. The focus in this section is to show the retrieval performance of extending FSR-GAN approach to retrieval application. Specifically, we employ three applications to evaluate FSR-GAN performance: (i) Content Based Image Retrieval, (ii) Large-Scale Image Retrieval, (iii) Low Bit-Rate Mobile Visual Search. The goal is to measure the quality of enhanced features when we use them as query to search images in a database. Actually, these applications can further verify the potentiality and robustness of our FSR-GAN.

5.1. Experimental Setup

In the experiment, we use three benchmark datasets including Oxford5K [14], Paris [9], and INRIA Holidays [9] mentioned in Section 4.1. We compare it with the raw features from low-resolution images. For evaluation criterion, we use the mean Average Precision (mAP) metric as a function of down-scaling ratio, and the mAP as a function of query bits for evaluating low bit-rate retrieval performance. The mAP score is a common used measure, which summarizes rankings from multiple queries by averaging mean-precisions.

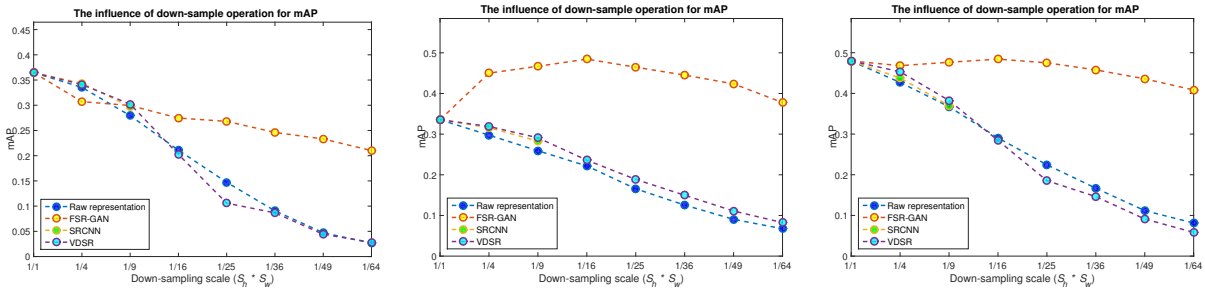
5.2. Content based Image Retrieval

In this section, we use content-based image retrieval to evaluate the retrieval performance of our FSR-GAN approach, looking at the retrieval precision as well as the down-scaling ratio. Specifically, the resolution of query images in Oxford5K [14], Holidays, and Paris [15] are first reduced to different low-resolutions by using uniform down-sample method. Then, these low-resolution queries are upsampled to 224 $\times$ 224 using bicubic interpolation method in order to satisfy the input size of VGG16. Finally, we extract deep features (36^{th} layer in VGG16) from these upsampled images.

Figure. 6 demonstrates mAP as a function of the down-sample ratio for Oxford5K [14], Holidays [9], and Paris [9] datasets, respectively. From Fig. 6, we observe that our FSR-GAN outperforms "Raw representation" and ISR methods by large margins in most cases. The "Raw representation" denotes the bicubic interpolation approach. Surprisingly, even when the query images have been down-sampled to 1/64 original size, our FSR approach still achieves considerable retrieval accuracy and significantly outperforms Raw representation. Interestingly, we find that although the resolution of query images is drastically changed, the retrieval performance of FSR-GAN is relatively stable. For Holidays dataset in Fig. 6, the strange phenomenon is that the retrieval accuracy increases slowly with the decrease of resolution. This phenomenon is caused by the characteristic of Holidays dataset. In Holidays dataset, the number of images associated with each query is small (about 4 images), which easily results in the fluctuation of retrieval accuracy. Overall, the results well imply that our FSR-GAN is capable of enhancing the discriminatory power of features extracted from low-resolution images.

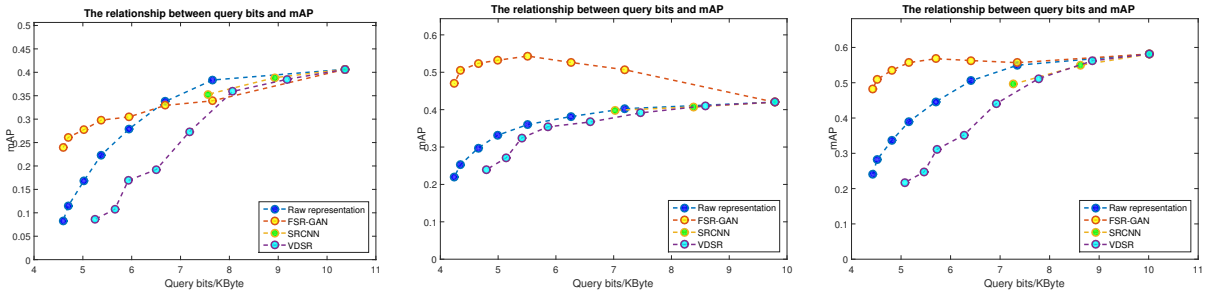
5.3. Large-Scale Image Retrieval

In order to evaluate the robustness of FSR-GAN approach, we conduct large-scale image retrieval experiment by mixing Flick100k dataset [14] as distractor with Oxford5K, Holidays, and Paris datasets. We summarize the experimental results on these three datasets in Fig. 7. From this figure we observe that our approach significantly outperforms Raw representation and ISR methods. At the similar down-sample ratio, our approach is capable of providing higher retrieval precision than raw feature. Moreover, our FSR-GAN still shows stable retrieval performance even



(a) Oxford5K (b) Holidays (c) Paris

Figure 7. Comparison of large-scale image retrieval performance on three datasets: Oxford5K [14], Paris [9], and INRIA Holidays [9] plus Flick100k dataset as distractors.



(a) Oxford5K (b) Holidays (c) Paris

Figure 8. Demonstration of low bit-rate image retrieval performance on Oxford5K [14], Paris [9], and INRIA Holidays [9] datasets. Our FSR-GAN approach shows surprising potential performance at low query bit.

when the resolution of original queries changes dramatically. Interestingly, we still observe that the retrieval accuracy increases slowly with the decrease of resolution at Holidays dataset. This phenomenon is caused by the characteristic of Holidays dataset as mentioned in Section 5.2. The experimental result well demonstrates the significant improvement achieved by our approach. We believe this is a significant progress in low bit-rate large-scale image retrieval.

5.4. Low Bit-Rate Mobile Visual Search

Camera equipped mobile devices, such as mobile phones are becoming ubiquitous platforms for deployment of visual search and augmented reality applications. With relatively slow wireless links, the response time of the retrieval system critically depends on how much information must be transferred. Therefore, reducing the upstream query data is an essential requirement for typical client-server visual search architectures. We can extend the FSR-GAN approach to the application of low bit-rate image retrieval. The user-end only requires down-sampling the query image to a small size. The cloud-end firstly exploits our FSR-GAN approach to recover the discriminatory ability for the uploaded small size image. Then, it uses the enhanced representation to search database and returns retrieval results to the user-end. Thus, the user-end only performs a surprisingly simple operation of down-sampling to reduce the bits of query, and can enjoy low latency and high precision mobile-visual-search (*i.e.*, has better user experience).

To demonstrate this goal, we have done more experiments to verify it. Fig. 8 shows the mAP as a function of query bits on three retrieval datasets. Note that in order to eliminate the influence of image coding, all resized query images are saved in PNG format. From Fig. 8 we observe that our FSR-GAN approach demonstrates surprising potential performance at low query bit.

6. Conclusion

We have presented a novel feature super-resolution technique for improving the discriminatory power of representations extracted from low-resolution images. By analyzing the impact of down-scaling operation on the deep features, we have two key conclusions. One is that low-resolution images not only impact the extracted deep features, but also seriously decrease the retrieval accuracy. The other is that deep features extracted from low-resolution images are changed regularly with down-scaling ratios, which inspires us to develop a feature super-resolution model to learn the mapping relationship between low-discriminative features and high-discriminative features. Extensive experiment results suggest that our proposed FSR-GAN approach is not only an effective solution for enhancing features, but also shows its great potential in many applications.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 38(2):295–307, 2016.

[4] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'2016)*, 38(2):295–307, Feb 2016.

[5] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision (ECCV)*, pages 391–407. Springer, 2016.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, pages 2672–2680, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[8] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[9] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV'2008*, pages 304–317. Springer, 2008.

[10] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[11] J. Kim, J. Kwon Lee, and K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016.

[12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[13] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR'2007*, pages 1–8, 2007.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR'2008*, 2008.

[16] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[17] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1–9, 2015.

[20] Y. Wang, L. Wang, H. Wang, and P. Li. End-to-end image super-resolution via deep and shallow convolutional networks. *arXiv preprint arXiv:1607.07680*, 2016.

[21] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 370–378, 2015.

[22] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[23] Z. Yang, W. Chen, F. Wang, and B. Xu. Improving neural machine translation with conditional sequence generative adversarial nets. *arXiv preprint arXiv:1703.04887*, 2017.