







Unveiling the Influence of Image Super-Resolution on Aerial Scene Classification

Mohamed Ramzy Ibrahim^{1,2}(✉) , Robert Benavente² , Daniel Ponsa² ,
and Felipe Lumbreras² 

¹ Computer Engineering Department, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt

m.ramzy@aast.edu

² Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Barcelona, Spain
{[mramzy](mailto:mramzy@cvc.uab.es), [robert](mailto:robert@cvc.uab.es), [daniel](mailto:daniel@cvc.uab.es), [felipe](mailto:felipe@cvc.uab.es)}@cvc.uab.es

Abstract. Deep learning has made significant advances in recent years, and as a result, it is now in a stage where it can achieve outstanding results in tasks requiring visual understanding of scenes. However, its performance tends to decline when dealing with low-quality images. The advent of super-resolution (SR) techniques has started to have an impact on the field of remote sensing by enabling the restoration of fine details and enhancing image quality, which could help to increase performance in other vision tasks. However, in previous works, contradictory results for scene visual understanding were achieved when SR techniques were applied. In this paper, we present an experimental study on the impact of SR on enhancing aerial scene classification. Through the analysis of different state-of-the-art SR algorithms, including traditional methods and deep learning-based approaches, we unveil the transformative potential of SR in overcoming the limitations of low-resolution (LR) aerial imagery. By enhancing spatial resolution, more fine details are captured, opening the door for an improvement in scene understanding. We also discuss the effect of different image scales on the quality of SR and its effect on aerial scene classification. Our experimental work demonstrates the significant impact of SR on enhancing aerial scene classification compared to LR images, opening new avenues for improved remote sensing applications.

Keywords: Super-resolution · Scene classification · Deep learning · Aerial images · Remote sensing

1 Introduction

Super-resolution (SR) techniques aim to generate a detailed and sharp high-resolution (HR) image from low-resolution (LR) images. The goal of SR is to improve the quality of images by enhancing the information they contain. This

is especially useful in some applications where obtaining HR images is difficult due to the environment (i.e., satellite imaging) or excessive costs (i.e., hardware for HR image acquisition is expensive) [15].

SR, as many other tasks in computer vision, has benefited from the rapid development of deep learning (DL), leading to an impressive improvement in performance in terms of image quality [2]. It seems plausible that improving the quality of input images could improve other subsequent high-level vision tasks and enhance their results. Indeed, some previous works have explored this hypothesis in low-visibility scenarios for object detection [24], and object classification [20]. The conclusions of these studies show that while the pre-processing approaches tested can enhance performance in some scenarios, in others the results can even degrade. Some other works have focused on the impact of SR on specific tasks, obtaining some contradictory results. SR improves object detection results [16], has a very limited favorable impact on object classification [26], but has no positive effect on human pose estimation [6]. Therefore, we can conclude that the impact of SR on other high-level vision tasks is not as clear as could be expected and is highly dependent on the specific task at hand. Among the different image classification problems, scene classification on aerial images could really benefit from an improvement in resolution. This task is particularly challenging due to two main factors. First, the way images are acquired, usually from a camera on an aircraft or satellite, implies that images often have very low resolution. This fact implies that important regions in the image have low detail, and valuable information from the complex spatial distribution of the scene is lost. Second, images are sometimes acquired in tough environmental conditions (rain, clouds, etc.). These factors cause some images from different classes to be very similar (i.e., inter-class similarity), while some images from the same class are quite different (i.e., intra-class diversity). Examples of these problems are shown in Fig. 1. Hence, improving the image resolution of such images could allow the detection of fine-grained details that might facilitate accurate and reliable scene classification.

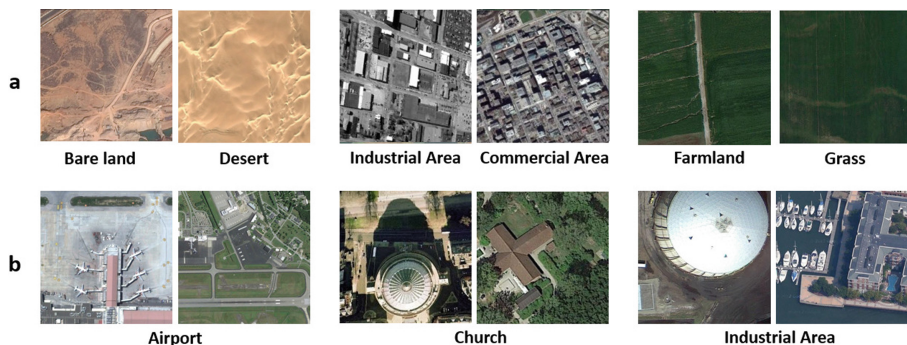


Fig. 1. Examples of (a) inter-class similarity, and (b) intra-class diversity. Images from AID [23], NWPU-RESISC45 [3], and RSSCN7 [27] datasets (left to right in each row).

In this paper, we present a comparative study to test the hypothesis that increasing image resolution can improve the results of aerial scene classification. The aim of this study is to complement previous works that studied the impact of SR on high-level tasks and to evaluate the impact of SR in the field of aerial scene classification. To achieve this goal, we have conducted experiments on (1) different SR models, from the traditional bicubic algorithm to convolutional neural networks (CNNs) and vision transformers (ViTs), (2) different state-of-the-art classifiers based on DL, (3) different benchmark datasets of aerial scenes with complex classes, and (4) simulated LR images with different scaling factors ($\times 0.5$ and $\times 0.25$).

2 Super-Resolution

A number of SR methods have been proposed in the last few decades. Initially, traditional methods were based on interpolation, optimization, and learning. However, in terms of resolution and restoration quality, these methods have been superseded by DL approaches [4]. In recent years, different models based on CNNs have been applied to SR tasks. Dong *et al.* [5] proposed the first SR model based on CNN (SRCNN). It is a simple model trained to learn an end-to-end mapping between LR and HR image pairs. Later, Tai *et al.* [18] proposed a Deep Recursive Residual Convolution Neural Network (DRRCNN) consisting of 52 convolution layers with residual connections. They also proposed a memory network (MemNet) [19] based on a recursive and a gate units for explicitly mining persistent memory through an adaptive learning process.

Generative adversarial networks (GANs) have also been used in SR tasks. Ledig *et al.* [11] proposed SRGAN, a general GAN for SR, and Sajjadi *et al.* [14] presented EnhanceNet GAN, which uses automatic texture synthesis and perceptual loss to build realistic textures without focusing on ground truth pixels. Wang *et al.* [21] proposed the Enhanced Super-Resolution Generative adversarial network (ESRGAN) which includes a residual-in-residual dense block based on the original residual dense block [25].

Indeed, the development of residual dense blocks has shown a potential in the SR field. Following this idea, the 3DRRDB model [8] was designed to stack large number of features and bypass them between network layers which helps in generating a good HR image.

Finally, after the success of ViTs in different vision tasks, it was also adopted in the SR field [12, 13]. Currently, the leading ViTs architecture in SR is SwinIR [12] that has a superior performance in image SR. In the following sections, we review in detail the methods used in our study.

2.1 Super-Resolution Convolution Neural Network (SRCNN)

SRCNN [5] was the first successful super-resolution CNN, and it pioneered many subsequent approaches. SRCNN structure is simple, consisting only of three convolution layers, each of which (except for the last layer) is followed by a rectified

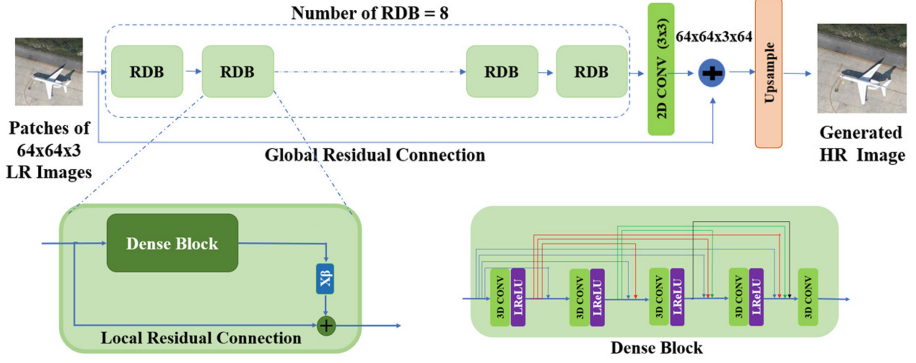


Fig. 2. Modified 3D Residual-in-residual dense block (m3DRRDB) with illustration of dense block.

linear unit (ReLU) non-linearity. The first convolution layer, referred to as feature extraction, is responsible for creating feature maps from the input images. The second convolution layer, known as non-linear mapping, is responsible for converting the feature maps into high-dimensional feature vectors. Finally, the last convolution layer combines the feature maps to produce the final HR image. Despite its simple architecture, SRCNN surpassed all classical SR techniques becoming a real breakthrough in the field.

2.2 Modified 3D Residual-in-Residual Dense Block (m3DRRDB)

The 3D residual-in-residual block model (3DRRDB) [8] was introduced for remote sensing SR and proposes a method to combine several LR images to generate a HR image. This scheme can be recast to estimate a HR multi-band (RGB) image from one LR RGB image by just substituting the 3D convolutions of its pipeline with 2D convolutions. We denote this modified model m3DRRDB.

The key values of this architecture are: (1) The usage of dense blocks that stack large amounts of 2D feature maps and establish maximum information flow between blocks, and (2) fusion of global and local residual connections with residual scaling that solves the problem of vanishing gradient and stabilize training. As shown in Fig. 2, the pipeline of m3DRRDB is composed of eight Residual Dense Blocks (RDBs) [25] with a convolution layer (CONV) with kernel of 3×3 . The model has a global residual connection that connects the input to the output of the network before the upsampling layer. The importance of the global residual connection is to pass the information from the input to the output of the last RDB to avoid any gradient losses. In turn, as illustrated in lower part of Fig. 2, each RDB block consists of a dense block and a local residual connections that connect input of the dense block to its output after multiplying it by a residual scaling factor. Moreover, each dense block is composed of five convolutions (the first four convolutions are followed by ReLU) where the output of each CONV layer is fed as input to all subsequent CONV layers.

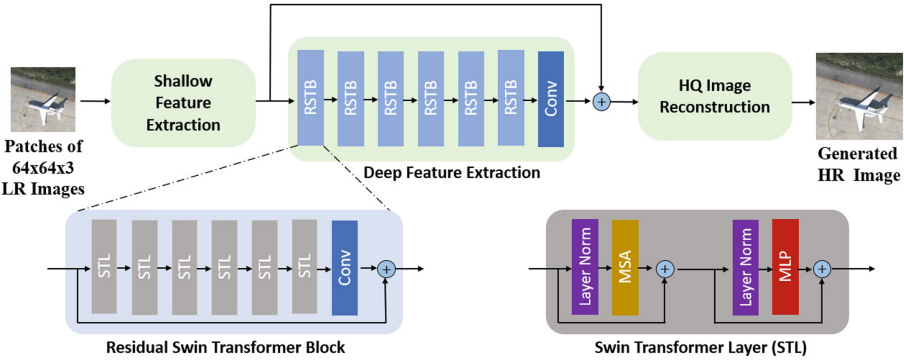


Fig. 3. SwinIR transformer architecture with illustration of RSTB block.

2.3 SwinIR Transformer

ViTs-based SwinIR [12] offers various advantages over CNN-based image restoration models which are: (1) Interactions between image content and attention weights that can be interpreted as spatially variable convolution which allows global interaction between contexts, and (2) the shifted window mechanism in SwinIR allows long-term dependencies modelling to avoid local interaction between image patches that happens in transformers.

As shown in Fig. 3, the pipeline of SwinIR starts with shallow feature extraction, followed by deep feature extraction, and ends with High-Quality (HQ) image reconstruction. The model has a skip connection that concatenates the inputs to deep features extraction phase to its output. In the shallow feature extraction phase, features are extracted from LR patches of size $64 \times 64 \times 3$ with a convolution layer, $CONV$, with kernel size 3×3 as shown in Eq. 1.

$$F_{SF} = CONV(LR), \tag{1}$$

where F_{SF} is the output of shallow feature phase.

For the deep feature extraction phase the Residual Swin Transformer blocks (RSTB) are used. This phase output, F_{DF} , is the output of six RSTB followed by $CONV$ layer with 3×3 kernel where intermediate feature is presented as F_1, F_2, \dots, F_6 . Equation 2 represents the complete scenario of deep feature extraction phase.

$$\begin{aligned} F_i &= RSTB_i(F_{i-1}), \quad i = 1, 2, \dots, 6, \\ F_{DF} &= CONV(F_6), \end{aligned} \tag{2}$$

where $RSTB_i$ is the i^{th} RSTB.

For the HQ image reconstruction phase can be noted as shown in Eq. 3

$$I_{HR} = H_{REC}(F_{SF} + F_{DF}), \tag{3}$$

where I_{HR} is the super-resolved image, and H_{REC} is the reconstruction module.

As shown in Fig. 3, the RSTB consists of six swin transformer layers (STL) followed by a CONV layer. A residual connection connects the input of the first STL to output of CONV layer to avoid gradient losses. Each STL is composed of 2 blocks. The first block is composed of a normalization Layer followed by a multi-headed self-attention (MSA) layer, with a residual connection between the input to the normalization layer and the output of the MSA layer. STLs are alternating between having a shifted window MSA or a regular window MSA. The second block is composed of a normalization layer followed by a multi-layer perceptron (MLP), with a residual connection between the input to normalization layer and the output of MLP.

3 Scene Classification

CNNs are widely used DL models for extracting high-quality representations and delivering robust results on different classification tasks [1, 3]. From the pioneering Alexnet [10], several CNN models have been proposed for general object classification. Among them, VGG16 [17], and ResNet-101 [7] have been the most used CNNs for classification and also as a backbone for other tasks. For this reason, we have adopted these two models in our experiments on scene classification.

The VGG16 [17] architecture showed good performance, and this has been attributed to its small kernel size (3×3) and its number of trainable parameters, which aid in increasing the CNN’s depth to extract more features [1]. The VGG16 model is made up of five groups. The first two groups consist of two convolution layers with ReLU activation function, and a single max pooling layer after the activation function. The last three groups include three convolution layers with ReLU activation function and a single max pooling layer. Finally three fully connected layers that also use a ReLU activation function. The last layer of the model is a softmax.

ResNet-101 [7] was proposed to solve the problem of vanishing gradients in previous deep CNNs. It is composed of 101 residual blocks, where each block consists of a pipeline of three convolutions. First, it starts with a convolution with multiple kernels, each of size 1×1 , followed by ReLU activation function. Second, a convolution layer with multiple kernels, each of size 3×3 , followed by ReLU activation function. Third, a convolution layer with multiple kernels, each of size 3×3 . Finally, the input to the block is added to output of the third convolution and passes through ReLU activation function. The model ends with a fully connected layer for classification.

4 Experiments

To assess and quantify the impact of SR on aerial scene classification, we conducted two experiments applying different SR models before evaluating two CNN classifiers, namely VGG16 and ResNet-101, on three standard datasets of aerial images. In this section, we provide a general overview of the two experiments, explain the datasets used, and list the experimental settings.

Experiment 1: Ranking of SR Models. The goal of this experiment was to rank the different SR methods used in the classification experiments (see Experiment 2) and have an evaluation of them in terms of image quality. The models evaluated are the bicubic method (used as baseline in most SR works in the literature), SRCNN [5] (the first CNN proposed for SR), m3DRRDB [8] (a CNN model based on the idea of residual dense blocks), and SwinIR [12] (based on ViTs and the current state-of-the-art for SR). These models are evaluated using the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) which are the quantitative metrics mainly used to assess image quality in SR trials [22].

Experiment 2: Impact of SR on Aerial Scene Classification. In this experiment we assess the impact of the evaluated super-resolution (SR) methods on aerial scene classification across various low-resolution (LR) image scales and with different scene classification methodologies.

We utilized two well-known CNN-based classifiers, namely VGG16 [17] and ResNet-101 [7], and compared their performance on LR images at different scales to their performance on super-resolved images generated by different SR techniques. We also computed the classification results on HR original versions of the images to obtain the theoretical best possible result from the classifiers. A summary of the scenarios evaluated in this experiment is depicted in Fig. 4. The models were evaluated using four quantitative metrics [9]: accuracy, precision, recall, and F1-score. These metrics were computed for each class in each dataset and averaged across all classes.

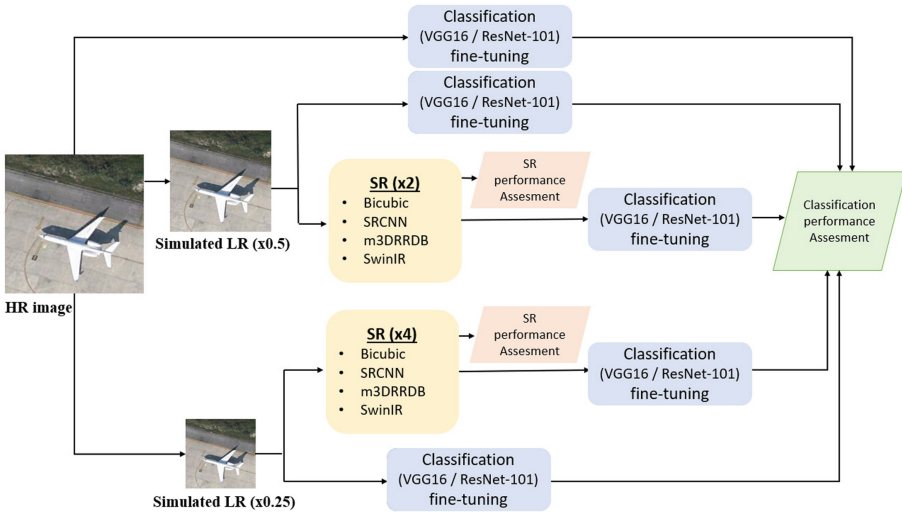


Fig. 4. Different experimental scenarios proposed on the three benchmark datasets.

4.1 Datasets

Experiments are conducted on three well-known benchmark aerial datasets: AID [23], NWPU-RESISC45 [3], and RSSCN7 [27]. Dataset details are listed in Table 1. The datasets are downsampled using the bicubic method to simulate LR images at scales of $\times 0.5$ and $\times 0.25$ for SR and scene classification experiments. The downsampling occurred by adding blur to the HR images and after that downsampling them to the required scale.

Table 1. Comparison between the used datasets, AID [23] dataset, NWPU-RESISC45 [3] dataset, and RSSCN7 [27] dataset.

Datasets	Total Images	Scene Classes	No. of Scenes/Class	Spatial Resolution (m/pixel)	Image Sizes	Year
AID [23]	10,000	30	~ 220 to 420	8 to 0.5	600 \times 600	2017
NWPU [3]	31,500	45	700	30 to 0.2	256 \times 256	2017
RSSCN7 [27]	2,800	7	400	unspecified	400 \times 400	2015

4.2 Experimental Settings

All the experiments were run on machines with a 3.80 GHz Core i7 processor, 32 GB of RAM, and an NVIDIA RTX 3090 with 24 GB of bandwidth. Moreover, the models are implemented using Python language and PyTorch DL framework. Each dataset used in the SR and classification experiments is divided into three portions as 78% for training, 2% for validation and 20% for testing.

Settings for Super-Resolution Methods. The training dataset is cropped to patches of size 64 and feed to the SR networks as batches of 16. The training data is augmented (on the fly) using random rotation, horizontal flip, and vertical flip before feeding it to SR networks. The SR models are trained for 600,000 iterations with learning rate of 0.0003 using a scheduler that decreases the learning rate to half at [250K, 400K, 450K, 475K] to avoid models overfitting during training. Adam optimizer is used with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$.

Settings for Classification Methods. The training data is batched into 32 samples and undergoes on-the-fly augmentation, including random horizontal and vertical flips, random rotations, and normalization before being fed into the classification network. Models are trained for 200 epochs with a learning rate of 0.001 using a cosine annealing scheduler and Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. For VGG16, we employ a pretrained model from the Imagenet dataset and fine-tune the last three groups and three fully connected layers while freezing the first two groups. For ResNet-101, we utilize a pretrained ResNet-101 model from the Imagenet dataset and fine-tune the last 50 residual blocks while freezing the first 51 residual groups.

Table 2. Quantitative Results for different SR models on AID [23] dataset, NWPU-RESISC45 [3] dataset, and RSSCN7 [27] dataset, in terms of PSNR and SSIM.

Method	Scale	AID		NWPU		RSSCN7	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	$\times 2$	32.17	0.8623	29.81	0.8173	29.80	0.7923
SRCNN	$\times 2$	34.46	0.9011	31.86	0.8697	31.56	0.8385
m3DRRDB	$\times 2$	35.41	0.9152	32.67	0.8871	32.21	0.8538
SwinIR	$\times 2$	36.91	0.9323	33.28	0.9217	33.18	0.8688
Bicubic	$\times 4$	28.72	0.7358	26.88	0.6686	26.82	0.6323
SRCNN	$\times 4$	29.22	0.7443	27.42	0.6912	27.48	0.6712
m3DRRDB	$\times 4$	30.38	0.7779	28.19	0.7332	28.12	0.6823
SwinIR	$\times 4$	31.49	0.8123	29.05	0.8992	28.92	0.6881

5 Results and Discussion

In this section, we present the results of the experiments described in Sect. 4.

5.1 Experiment 1: Ranking of Super-Resolution Methods

Table 2 lists the results of Experiment 1. It arranges the SR models according to their quantitative results on the three benchmark datasets [3, 23, 27]. The results are organized according to the two upscaling factors ($\times 2$ and $\times 4$). As can be seen in Table 2, the results obtained by the SR models agree with the previously reported results on other datasets [12]: SwinIR outperforms the other methods, obtaining the best PSNR and SSIM results at different scales on the three benchmark datasets, and the model based on residual dense blocks, m3DRRDB, overcomes SRCNN. The bicubic method obtained the lowest results in all the tested scenarios. Moreover, as could be expected, all the SR models get better results at scale ($\times 2$) than at scale ($\times 4$) in both PSNR and SSIM. Figure 5 shows qualitative results of the SR methods for $\times 4$ scale on AID [23] dataset.

5.2 Experiment 2: Impact of Super-Resolution on Aerial Scene Classification

The results of Experiment 2, highlighting the impact of SR on aerial scene classification, are summarized in Tables 3 (results for VGG16) and 4 (results for ResNet-101). In each table, the four metrics (i.e., accuracy, precision, recall, and F1-score) are given for each evaluated scenario. For each dataset, we present the results on the LR images downsampled at two different scales, $\times 0.5$ and $\times 0.25$, and the corresponding results of the SR methods, with upsampling $\times 2$ and $\times 4$. For each dataset, we provide the results on the original HR images (i.e., before downsampling) as the theoretical best result achievable.

From the results in the tables, we can state that, overall, SR enhances the results on aerial scene classification. In the experiments, all the SR methods consistently yielded better classification performance in all the metrics compared to the results on LR images. Such improvement is achieved across all scales, classification methods, and datasets. Additionally, the scale of upsampling has an

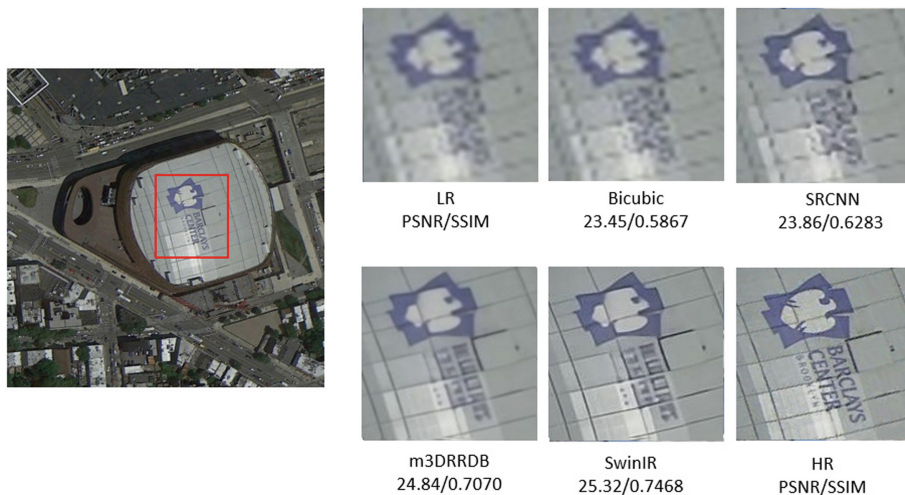


Fig. 5. Qualitative comparison of $\times 4$ SR on AID [23] dataset.

Table 3. VGG16 classification results for different SR models on AID [23] dataset, NWPU-RESISC45 [3] dataset, and RSSCN7 [27] dataset in terms of mean accuracy, mean precision, mean recall, and mean F1-score.

Dataset	Method	Scale	Classification Metric				
			Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	
AID	LR	$\times 0.5$	74.20	74.88	74.14	74.51	
	Bicubic	$\times 2$	75.45	76.28	75.23	75.75	
	SRCNN	$\times 2$	77.85	78.38	77.42	77.90	
	m3DRRDB	$\times 2$	78.40	78.86	77.89	78.37	
	SwinIR	$\times 2$	79.90	79.86	79.47	79.66	
	LR	$\times 0.25$	73.50	74.33	73.47	73.90	
	Bicubic	$\times 4$	74.90	75.78	74.74	75.25	
	SRCNN	$\times 4$	76.60	77.35	76.36	76.85	
	m3DRRDB	$\times 4$	77.20	77.83	76.90	77.36	
	SwinIR	$\times 4$	77.90	78.44	77.47	77.95	
	HR	–	81.25	81.03	80.78	80.91	
	NWPU-RESISC45	LR	$\times 0.5$	85.89	86.79	86.93	86.86
		Bicubic	$\times 2$	86.14	87.00	87.11	87.06
		SRCNN	$\times 2$	87.60	88.16	88.48	88.32
m3DRRDB		$\times 2$	88.30	88.52	89.02	88.77	
SwinIR		$\times 2$	89.73	89.79	90.18	89.98	
LR		$\times 0.25$	85.73	86.64	86.82	86.73	
Bicubic		$\times 4$	85.95	86.82	86.99	86.90	
SRCNN		$\times 4$	87.40	87.96	88.34	88.15	
m3DRRDB		$\times 4$	87.21	87.78	88.10	87.94	
SwinIR		$\times 4$	88.44	88.66	89.11	88.88	
HR		–	90.10	89.99	90.60	90.29	
RSSCN7		LR	$\times 0.5$	73.75	74.18	73.84	74.01
		Bicubic	$\times 2$	74.46	74.95	74.56	74.75
		SRCNN	$\times 2$	75.71	76.16	75.85	76.00
	m3DRRDB	$\times 2$	76.07	76.51	76.19	76.35	
	SwinIR	$\times 2$	76.79	77.23	76.88	77.05	
	LR	$\times 0.25$	71.96	72.74	72.11	72.42	
	Bicubic	$\times 4$	73.93	74.75	74.01	74.38	
	SRCNN	$\times 4$	74.82	76.04	74.92	75.48	
	m3DRRDB	$\times 4$	75.71	77.06	75.83	76.44	
	SwinIR [12]	$\times 4$	75.18	76.44	75.28	75.86	
	HR	–	78.21	79.28	78.27	78.77	

Table 4. ResNet-101 classification results for different SR models on AID [23] dataset, NWPU-RESISC45 [3] dataset, and RSSCN7 [27] dataset in terms of mean accuracy, mean precision, mean recall, and mean F1-score.

Dataset	Method	Scale	Classification Metric				
			Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	
AID	LR	×0.5	81.75	81.95	81.53	81.74	
	Bicubic	×2	82.25	82.55	82.04	82.30	
	SRCNN	×2	83.45	83.76	83.28	83.52	
	m3DRRDB	×2	84.60	84.75	84.35	84.55	
	SwinIR	×2	85.35	85.53	85.13	85.33	
	LR	×0.25	80.85	81.77	80.77	81.27	
	Bicubic	×4	81.45	81.98	81.72	81.85	
	SRCNN	×4	82.50	83.03	82.61	82.82	
	m3DRRDB	×4	83.20	83.47	83.05	83.26	
	SwinIR	×4	84.80	85.01	84.58	84.80	
	HR	–	88.65	88.69	88.34	88.51	
	NWPU-REISSIC45	LR	×0.5	86.68	86.42	87.10	86.76
		Bicubic	×2	87.49	87.27	87.81	87.54
		SRCNN	×2	88.71	88.46	89.08	88.77
m3DRRDB		×2	89.35	89.02	89.74	89.38	
SwinIR		×2	90.22	89.96	90.54	90.25	
LR		×0.25	85.89	85.78	86.44	86.10	
Bicubic		×4	86.30	86.11	86.71	86.41	
SRCNN		×4	88.05	87.79	88.49	88.14	
m3DRRDB		×4	87.38	87.15	87.75	87.45	
SwinIR		×4	88.86	88.55	89.15	88.85	
HR		–	92.13	91.95	92.27	92.11	
RSSCN7		LR	×0.5	82.68	83.21	82.68	82.94
		Bicubic	×2	83.75	84.40	83.75	84.07
		SRCNN	×2	84.11	84.64	84.11	84.37
	m3DRRDB	×2	84.82	85.21	84.82	85.02	
	SwinIR	×2	85.71	85.93	85.71	85.82	
	LR	×0.25	82.32	82.85	82.32	82.59	
	Bicubic	×4	83.21	83.83	83.21	83.52	
	SRCNN	×4	83.93	84.43	83.93	84.18	
	m3DRRDB	×4	83.39	84.02	83.39	83.7	
	SwinIR	×4	84.82	85.21	84.82	85.02	
	HR	–	87.32	87.65	87.32	87.49	

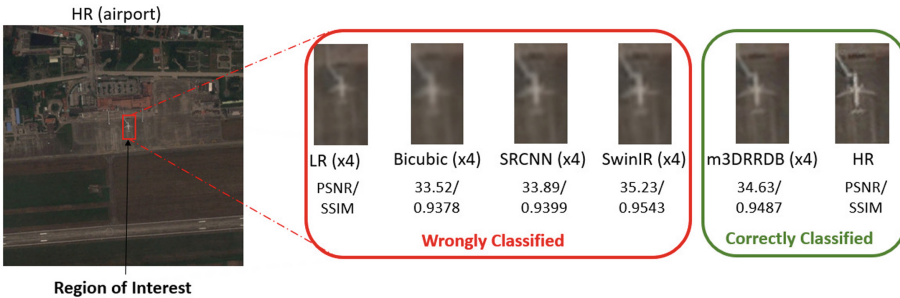


Fig. 6. Fine-tuned VGG16 missclassification example: Airport scene missclassified as Farm Land with illustration of degradation in RoI (note: PSNR/SSIM is for the full super-resolved image not the RoI region only).

impact on the classification metrics, with classifiers yielding better metrics when SR methods are applied to upsample at $\times 2$ scale compared to the corresponding results at $\times 4$ scale. This scale effect is consistent across all benchmark datasets.

Tables 3 and 4 also reveal that for the $\times 2$ scale, SwinIR demonstrates the best performance as a SR method in terms of producing super-resolved images which obtain the highest classification results across all three benchmark datasets. This result is also observed at the $\times 4$ scale for ResNet-101 (Table 4) for on the three datasets, and for VGG16 (Table 3) on the AID and NWPU-RESISC45 datasets. However, in the case of VGG16 on the RSSCN7 dataset, SwinIR super-resolved images do not achieve the best classification results, and m3DRRDB is the best SR method instead. Additionally, in the $\times 4$ scale, SRCNN super-resolved images slightly outperform m3DRRDB super-resolved images in classification for NWPU-RESISC45 dataset using VGG16 (Table 3), and also for both NWPU-RESISC45 and RSSCN7 datasets using ResNet-101 (Table 4). Hence, we can observe that in some cases the ranking of SR methods obtained in terms of PSNR and SSIM in Experiment 1 is not kept when we evaluate them according to the improvement they provide on classification. This highlights the limitations of PSNR and SSIM as metrics for predicting the impact of SR methods on other tasks. Thus, from our results, we can conclude that achieving the highest PSNR and SSIM results does not guarantee the highest performance in aerial scene classification. This could be due to the fact that these metrics evaluate overall scene clarity rather than a specific region of interest (RoI) within the scene, which sometimes can be more relevant for a given task.

This possibility is illustrated in Fig. 6 which shows a missclassification example where the image represents class “airport”. We hypothesize that inside the RoI (an aircraft that can be crucial to identify the class of the image) a lot of detail is lost in the LR image, and bicubic, SRCNN, and SwinIR can not recover it in the super-resolved images, which leads to wrongly classify the image as class “farm land”. However, it is classified correctly as class “airport” with m3DRRDB which obtains a super-resolved image where details of the RoI are much clearer and more similar to the HR image. In this example, the full scene scored highest PSNR/SSIM of 35.23/0.9543 using SwinIR but it is missclassified while it is correctly classified with m3DRRDB (second best PSNR/SSIM of 34.63/0.9487).

6 Conclusion

In this paper, we have studied the effect of different SR methods on aerial scene classification. We have first ranked different types of SR methods (traditional, CNN-based, and ViTs-based) in terms of image quality on two simulated LR image scales ($\times 0.5$ and $\times 0.25$) from three benchmark datasets on aerial scene classification. Then, we have assessed how pre-processing images with SR techniques affects aerial scene classification by two well-known CNN classifiers (VGG16 and ResNet-101). This second experiment included aerial scene classification comparisons between simulated LR images at different scales ($\times 0.5$, $\times 0.25$) and different super-resolved images obtained from the considered

SR methods. The extensive experimental work shows that SR methods consistently improve aerial scene classification compared to LR images for all the SR methods, all the scales, all the datasets, and all the classifiers tested.

Furthermore, we proved that the performance of a SR method in terms of PSNR and SSIM is not always directly related to the degree of improvement on aerial scene classification, especially when working with small LR scale images ($\times 0.25$). We draw the hypothesis that this result is due to the fact that PSNR and SSIM are metrics designed to measure the overall clarity of the image rather than that of specific RoIs, which can be determinants for a given task. In aerial scene classification, with the challenging inter-class similarity and intra-class diversity of aerial images, details of a certain RoI can be especially valuable for a correct classification. Moreover, we also hypothesize that sometimes SR of LR images with very small scales ($\times 0.25$) can amplify artifacts that affect the classification results. However, more experiments are needed to prove these hypothesis. As a future work, the presented study can be extended to test the effects of SR on other tasks such as image segmentation and object detection on specific datasets. Moreover, further studies on smaller scales ($\times 8$ and $\times 16$) are needed to test if the reported improvement on classification performance holds for lower resolutions, where the effects of artifacts and image degradation can be increasingly challenging.

Acknowledgment. This work is partially supported by Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC’s general activities, and the Departament de Recerca i Universitats from the Generalitat de Catalunya with reference 2021SGR01499.

References

1. Alom, M.Z., et al.: The history began from AlexNet: a comprehensive survey on deep learning approaches. arXiv (2018). <http://arxiv.org/abs/1803.01164>
2. Anwar, S., Khan, S., Barnes, N.: A deep journey into super-resolution: a survey. arXiv (2020). <https://doi.org/10.48550/arXiv.1904.07523>
3. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: benchmark and state of the art. Proc. IEEE **105**(10), 1865–1883 (2017). <https://doi.org/10.1109/JPROC.2017.2675998>
4. Khoo, J.J.D., Lim, K.H., Phang, J.T.S.: A review on deep learning super resolution techniques. In: 2020 IEEE 8th Conference on Systems, Process and Control (ICSPC), pp. 134–139 (2020). <https://doi.org/10.1109/ICSPC50992.2020.9305806>
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
6. Hardy, P., Dasmahapatra, S., Kim, H.: Can super resolution improve human pose estimation in low resolution scenarios? In: 17th International Conference on Computer Vision Theory and Applications, pp. 494–501 (2022). www.scitepress.org/Link.aspx?doi=10.5220/0010863700003124

7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
8. Ibrahim, M.R., Benavente, R., Lumbreras, F., Ponsa, D.: 3DRRDB: super resolution of multiple remote sensing images using 3D residual in residual dense blocks. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 322–331 (2022). <https://doi.org/10.1109/CVPRW56347.2022.00047>
9. Ibrahim, M.R., Youssef, S.M., Fathalla, K.M.: Abnormality detection and intelligent severity assessment of human chest computed tomography scans using deep learning: a case study on SARS-COV-2 assessment. *J. Ambient. Intell. Humaniz. Comput.* **14**(5), 5665–5688 (2023). <https://doi.org/10.1007/s12652-021-03282-x>
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc. (2012). www.papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
11. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114 (2017). <https://doi.org/10.1109/CVPR.2017.19>
12. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: image restoration using swin transformer. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1833–1844 (2021). <https://doi.org/10.1109/ICCVW54120.2021.00210>
13. Lu, Z., et al.: Transformer for single image super-resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 456–465 (2022). <https://doi.org/10.1109/CVPRW56347.2022.00061>
14. Sajjadi, M.S.M., Scholkopf, B., Hirsch, M.: EnhanceNet: single image super-resolution through automated texture synthesis. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4501–4510 (2017). <https://doi.org/10.1109/ICCV.2017.481>
15. Salvetti, F., Mazzia, V., Khaliq, A., Chiaberge, M.: Multi-image super resolution of remotely sensed images using residual attention deep neural networks. *Remote Sens.* **12**(14), 2207 (2020). <https://doi.org/10.3390/rs12142207>
16. Shermeyer, J., Van Etten, A.: The effects of super-resolution on object detection performance in satellite imagery. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1432–1441 (2019). <https://doi.org/10.1109/CVPRW.2019.00184>
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015 (2015). arxiv.org/abs/1409.1556
18. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2790–2798 (2017). <https://doi.org/10.1109/CVPR.2017.298>
19. Tai, Y., Yang, J., Liu, X., Xu, C.: MemNet: a persistent memory network for image restoration. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4549–4557 (2017). <https://doi.org/10.1109/ICCV.2017.486>
20. Vidal, R.G., et al.: UG²: a video benchmark for assessing the impact of image restoration and enhancement on automatic visual recognition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1597–1606 (2018). <https://doi.org/10.1109/WACV.2018.00177>

21. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11133, pp. 63–79. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11021-5_5
22. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004). <https://doi.org/10.1109/TIP.2003.819861>
23. Xia, G.S., et al.: AID: a benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **55**(7), 3965–3981 (2017). <https://doi.org/10.1109/TGRS.2017.2685945>
24. Yang, W., et al.: Advancing image understanding in poor visibility environments: a collective benchmark study. *IEEE Trans. Image Process.* **29**, 5737–5752 (2020). <https://doi.org/10.1109/TIP.2020.2981922>
25. Zhang, Y., et al.: Residual dense network for image super-resolution. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018). <https://doi.org/10.1109/CVPR.2018.00262>
26. Zhou, L., Chen, G., Feng, M., Knoll, A.: Improving low-resolution image classification by super-resolution with enhancing high-frequency content. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 1972–1978 (2021). <https://doi.org/10.1109/ICPR48806.2021.9412876>
27. Zou, Q., Ni, L., Zhang, T., Wang, Q.: Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **12**(11), 2321–2325 (2015). <https://doi.org/10.1109/LGRS.2015.2475299>