EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems

Research Article EAI.EU

Enhancing Single-Image Super-Resolution using Patch-Mosaic Data Augmentation Method on Lightweight Bimodal Network

Nguyen Quoc Toan[1,*], Tang Quang Hieu[2,*]

[1]Department of Electronic and Electrical Engineering, Hongik University, Seoul, South Korea
[2]Department of Information Technology, FPT University, Ho Chi Minh City, Vietnam

Abstract

With the advancement of deep learning, single-image super-resolution (SISR) has made significant strides. However, most current SISR methods are challenging to employ in real-world applications because they are doubtlessly employed by substantial computational and memory costs caused by complex operations. Furthermore, an efficient dataset is a key factor for bettering model training. The hybrid models of CNN and Vision Transformer can be more efficient in the SISR task. Nevertheless, they require substantial or extremely high-quality datasets for training that could be unavailable from time to time. To tackle these issues, a solution combined by applying a Lightweight Bimodal Network (LBNet) and Patch-Mosaic data augmentation method which is the enhancement of CutMix and YOCO is proposed in this research. With patch-oriented Mosaic data augmentation, an efficient Symmetric CNN is utilized for local feature extraction and coarse image restoration. Plus, a Recursive Transformer aids in fully grasping the long-term dependence of images, enabling the global information to be fully used to refine texture details. Extensive experiments have shown that LBNet with the proposed data augmentation with zero-free additional parameters method outperforms the original LBNet and other state-of-the-art techniques in which image-level data augmentation is applied.

Received on 13 October 2022; accepted on 07 May 2023; published on 25 May 2023

Keywords: single-image super-resolution, data augmentation, vision transformer, CNN

doi:10.4108/eetinis.v10i2.2774

1. Introduction

The main objective of single image super-resolution (SISR) is to restore the corresponding high-resolution (HR) image with rich details and improved visual quality from a degraded low-resolution (LR) image. Because of their powerful feature extraction ability, convolutional neural networks (CNN)-based SISR methods have recently surpassed traditional methods. [1], for example, developed the Super-Resolution Convolutional Neural Network (SRCNN). Afterward, with the introduction of ResNet [2] and DenseNet [3], a plethora of CNN-based SISR models, such as VDSR [4], EDSR [5], and RCAN [6], were proposed. All these methodologies demonstrate that the deeper the network, the more effectively it performs. Nonetheless, due to limited storage and computing capabilities, these methods are challenging to implement in real-world environments. Therefore, research into a model that can acquire better performance while maintaining the network's lightweight has become promising. A recursive mechanism, such as DRCN [7] or DRRN [8], is one of the most common approaches. The other is to establish lightweight structures, such as CARN [9], FDIWN [10], and PFFN [11]. Although these models lower the number of model parameters to some extent through multiple techniques and structures, they also degrade performance, making it difficult to reconstruct high-quality images with rich details.

Vision Transformer has recently become a hot research topic. It can model the image's long-term

[★]Enhancing Single-Image Super-Resolution using Patch-Mosaic Data Augmentation Method on Lightweight Bimodal Network
[*]Corresponding authors. Email: (1) nqtoan@g.hongik.ac.kr and (2) hieutq10@fpt.edu.vn

dependence, and this powerful representation ability can aid in the restoration of the image's texture details. Most methods, however, use Transformer to substitute all original CNN structures which is inappropriate because CNN's capability to extract local features is indispensable. These features can retain their stability under various viewing angles, it is effective for image comprehension and reconstruction. Hence, merging CNN and Transformer to maximize the effectiveness of both is highly recommended for efficient image reconstruction.

Furthermore, neural networks for SISR, without question, require massive datasets to attain acceptable performance. It can sometimes be difficult to gather or improve the quality or quantity of qualified data samples. Most standard data augmentations are conducted at the image-level resulting in overall performance gains for both generality and robustness. Image-level augmentation typically preserves semantics globally, following human cognitive intuition. Humans, on the other hand, can recognize objects based on only a portion of the information. Patches, or image insider information, are powerful natural signals. Several low-level vision [12][13] and high-level vision [14][15][16] works used patches prior to the deep learning era. A major element of Vision Transformer (ViT) recently has been the splitting of a single image into multiple non-overlapping patches as input for neural networks [17]. Notwithstanding, research on how to implement non-image-level data augmentations (i.e., patch-oriented) is rare.

The proposed Patch-Mosaic could play a vital role in augmenting datasets for training computer vision models by providing a zero-parameter yet efficient method, it is hypothesized that it is suitable to launch a strategy for conducting augmentations out of the image-level. An image can be viewed as a patchwork of several patches. We can perform a specific augmentation on these pieces individually and then combine the transformed randomly pieces back into a single image via the Mosaic data augmentation method introduced in YOLOv4 [18], an improvement of CutMix [19], which combines four training images in specific ratios into one. Mosaic accomplishes this by resizing each of the four images, putting them together, and then selecting a random cutout from the stitched images to create the final Mosaic image. This strategy increases diversity at both the local region and overall image-levels, and it may also encourage neural networks to attain the same cognitive ability as humans in recognizing objects from partial information. As a result, it aids in boosting the dataset for significantly better training and model performance. In summary, the main contributions are as follows

- Patch-Mosaic data augmentation is proposed as a data augmentation method for improving SISR model performance.
- Using Patch-Mosaic in conjunction with Lightweight Bimodal Network (LBNet) [20] to improve model capabilities in popular test sets in scenarios of the low computational cost required application.

2. Related work

2.1. Single-image Super-resolution

CNN-based. SISR methods based on CNN have advanced significantly in recent years [21]. For example, SRCNN [1] was the first to employ CNN to SISR and achieve competitive performance at the time. EDSR [5] significantly improved model performance by incorporating residual blocks [2]. RCAN [6] developed an 800-layer network and introduced the channel attention mechanism. Several lightweight SISR models have been introduced in recent years, in addition to these deep networks. For example, [9] used the cascade mechanism to propose a lightweight Cascaded Residual Network (CARN). Using the distillation and selective fusion strategy, [22] proposed an Information MultiDistillation Network (IMDN). MADNet [23] improved multi-scale feature representation and learning by employing a dense lightweight network. [24] proposed a simple but effective deep lightweight SISR model capable of generating convolutional kernels adaptively from each position's local information. Nevertheless, the effectiveness of these lightweight models is low in quality as they prevent access to larger receptive fields and global information.

Transformer-based. Many Transformer-based methods for computer vision tasks have recently been proposed, promoting the development of SISR. [25] proposed a pre-trained Image Processing Transformer for image restoration. [26] proposed a SwinIR that produced excellent results by directing the Swin Transformer directly to the image restoration task. [27] proposed an Effective Super-resolution Transformer (ESRT) for SISR using a lightweight Transformer and feature separation strategy to decrease GPU memory consumption. Nevertheless, none of these models take into account the fusion of CNN and Transformer, making it difficult to strike the best balance of model size and performance.

2.2. Patches in data augmentation

Patches are commonly used as strong signals in traditional and learning-based vision tasks. Examples of applications include texture synthesis [28], bag-of-features-based classification [14][15][16], image denoising [13], super-resolution [29], image-to-image translation [30][31]. CNNs have recently adopted it and ViTs as input for classification networks [32][17].

Patches are also utilized for data augmentation. [33] only uses Gaussian Blur on a subset of the images. [34] creates large labeled instance datasets by cutting and pasting object instances onto backgrounds. [35] overlays 2D object images onto images of real-world environments. CutMix [19] integrates one image patch with a patch from another. PAA [36] extends the AutoAug [37] configuration by searching for augmentation policies in pre-defined patches. Through patch-based negative augmentation, [38] improves the robustness of ViTs. YOCO [39] applied patch data augmentation by cutting images into 2 pieces, then processing augmenting methods individually, then combining 2 images in the original one. Even so, no previous research has looked into how to perform the same augmentation on a non-image-level and Mosaic together.

3. Methodology

3.1. Lightweight Bimodal Network (LBNet)

Network Architecture. The Lightweight Bimodal Network (LBNet) [20], as shown in figure 1, is primarily composed of Symmetric CNN, Recursive Transformer, and reconstruction module. Symmetric CNN is conducted for local feature extraction, whereas Recursive Transformer is designed to learn image long-term dependence. I_{LR}, I_{SR}, and I_{HR} are the input LR image, reconstructed SR image, and corresponding HR image. A 3×3 convolutional layer is employed at the model's head for shallow extracting features.

$$F_{sf} = f_{sf}(I_{LR}), \tag{1}$$

F_{sf} is the extracted shallow features, while $f_{sf}()$ denotes the convolutional layer. The shallow features will then be transmitted to Symmetric CNN for local extracting the features.

$$F_{CNN} = f_{CNN}(F_{sf}), \tag{2}$$

where $f_{CNN}()$ is the Symmetric CNN and F_{CNN} is the extracted local features. One of the most important components of LBNet is symmetric CNN made up of several pairs of parameter-sharing Local Feature Fusion Modules (LFFMs) and channel attention modules. The following section will go over all of these modules.

Following that, the Recursive Transformer will be given all of these features for long-term dependence learning.

$$F_{RT} = f_{RT}(F_{CNN}), \tag{3}$$

where $f_{RT}()$ represents the Recursive Transformer and F_{RT} is the feature improved by global information. Finally, the refined F_{RT} and shallow F_{sf} features are combined and passed to the reconstruction module for the reconstruction of the SR image.

$$I_{SR} = f_{build}(F_{sf} + F_{RT}), \tag{4}$$

where $f_{build}()$ represents the reconstruction module constituted of a 3×3 convolutional layer and a pixel-shuffle layer.

LBNet is optimized with the L1 loss function [40] throughout training. The training dataset is given $\left\{I^i_{LR}, I^i_{HR}\right\}^N_{i=1}$,it is processed by

$$\theta^* = argmin_\theta \frac{1}{N}\sum_{i=1}^{n} |F(I^i_{LR} - I^i_{HR})|, \tag{5}$$

where θ denotes the LBNet parameter set, F$(I_{LR}) = I_{SR}$ denotes the reconstructed SR image, and N represents the number of training images.

Symmetric CNN and Recursive Transformer. Symmetric CNN is a type of CNN that is specifically designed for extracting local features. It consists primarily of paired parameter-sharing Local Feature Fusion Modules (LFFMs) and Channel Attention (CA) modules. The parameter-sharing of every two symmetrical modules can enhance parameter and performance balance. Moreover, each pair of parameter-sharing modules will be merged via the channel attention module to fully utilize the extracted features.

Symmetric CNN is a dual-branch network, as illustrated in figure 1. The shallow feature F_{sf} will be transferred to the top branch first, and the outputs of each LFFM in the top branch will be used as input to the corresponding LFFM in the down branch. The entire operation can be formulated as having:

$$F^{T,1}_{LFFM} = f^{T,1}_{LFFM}(F_{sf}), i = 1, \tag{6}$$

$$F^{T,i}_{LFFM} = f^{T,i}_{LFFM}(F^{T,i-1}_{LFFM}), i = 2, ..., n, \tag{7}$$

$$F^{D,i}_{LFFM} = f^{D,i}_{LFFM}(F^{D,i-1}_{LFFM} + f^{i}_{CA}(F^{T,i}_{LFFM}), i = 2, ..., n, \tag{8}$$

where $f^{T,i}_{LFFM}()$ and $f^{D,i}_{LFFM}()$ denote the $i-th$ LFFM in the top and bottom branches, respectively. The $i-th$ channel attention module is denoted by $f^i_{CA}()$. It is important to highlight that when $i = 1$, $F^{D,1}_{LFFM} = f^{D,1}_{LFFM}(F^{T,n}_{LFFM} + f^1_{CA}(F^{T,1}_{LFFM}))$. Furthermore, the weight-sharing method is applied to the paired modules, resulting in $f^{D,i}_{LFFM}() = f^{T,i}_{LFFM}()$. Finally, the values of all of these LFFMs are concatenated, and a 1×1 convolutional layer is employed for feature fusion and

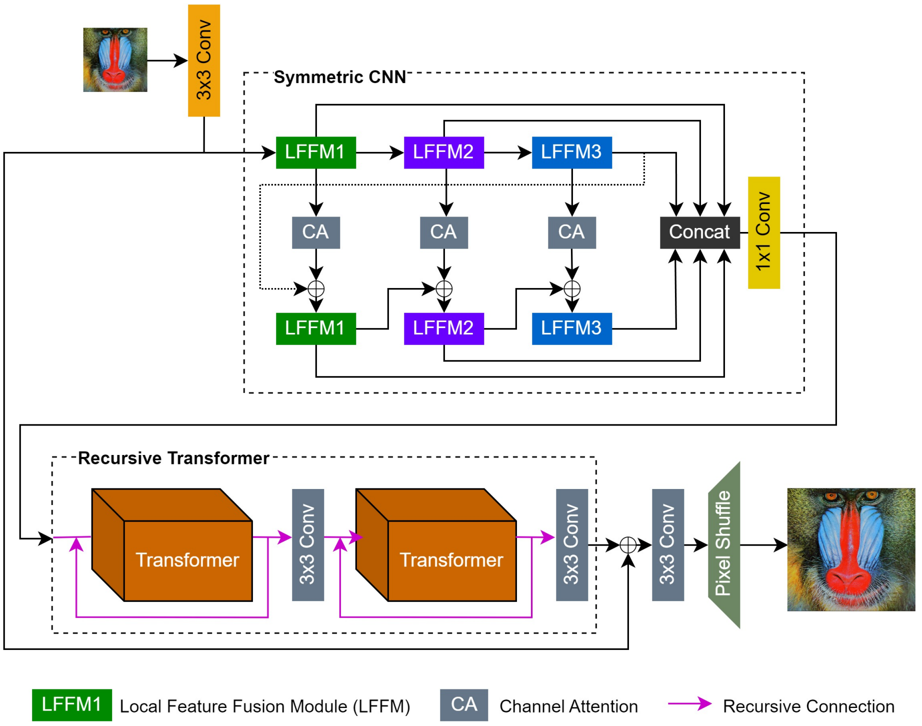

Figure 1. Structure of Lightweight Bimodal Network (LBNet)

compression. Therefore as result, the most efficient features extracted at various levels will be forwarded to the following section to learn the long-term dependence of images.

Symmetric CNN relies mainly on LFFM. Figure 2 shows that LFFM is essentially an improvement of DenseBlock [3]. In contrast to DenseBlock, (1) FRDAB is applied to substitute for the original convolutional layer to strengthen feature extraction; (2) a 1×1 group convolutional layer is used before each FRDAB to reduce the dimension; and (3) local residual learning is created to better information transmission. The full operation of LFFM is as follows:

$$F^{1}_{FRB} = f^{1}_{FRB}(F^{m-1}_{LFFM}), \tag{9}$$

$$F^{2}_{FRB} = f^{2}_{FRB}(f^{1}_{gc}([F^{m-1}_{LFFM}, F^{1}_{FRB}])), \tag{10}$$

$$F^{3}_{FRB} = f^{3}_{FRB}(f^{2}_{gc}([F^{m-1}_{LFFM}, F^{1}_{FRB}, F^{2}_{FRB}])), \tag{11}$$

$$F^{m}_{LFFM} = f^{m-1}_{LFFM} + f_{1\times1}([F^{m-1}_{LFFM}, F^{1}_{FRB}, F^{2}_{FRB}, F^{3}_{FRB}]), \tag{12}$$

where F^{i}_{FRB} is the output of the $i-th$ (i = 1,2,3) FRDAB module in LFFM. $f^{i}_{gc}()$ represents the $j-th$ (j = 1,2) group convolutional layer followed by FRDAB. The input and output of the $m-th$ LFFM module are denoted by F^{m-1}_{LFFM} and F^{m}_{LFFM}, respectively.

FRDAB, as shown in 2, is a dual-attention block that is specifically designed for feature improvement. The multi-branch structure is created specifically for feature extraction and utilization. The feature will be passed to two branches in this section, and each branch will use a different value of convolutional layers to modify the size of the receptive field to acquire different scaled features. Following that, channel attention is used to extract channel statistics for re-weighting in the channel dimension, and spatial attention is applied to re-weight the pixel based on the feature map's spatial context relationship. At last, the addition operation combines the results of these two attention operations. The final features obtained using this method will have a firmer suppression of the smooth areas of the input image.

Symmetric CNN is intended for the extraction of local features. Notwithstanding, this is far from sufficient to reconstruct high-quality images because the depth of the lightweight network makes having a large enough receptive field to achieve global information difficultly. To tackle this issue, a Recursive Transformer and Transformer to learn the long-term dependence of images (RT) are employed. Unlike previous methods, a recursive mechanism is applied to fully train the Transformer without significantly increasing GPU memory requirement or model parameters. Figure 1 depicts that RT comes before the reconstruction module, which is made up of two Transformer Modules

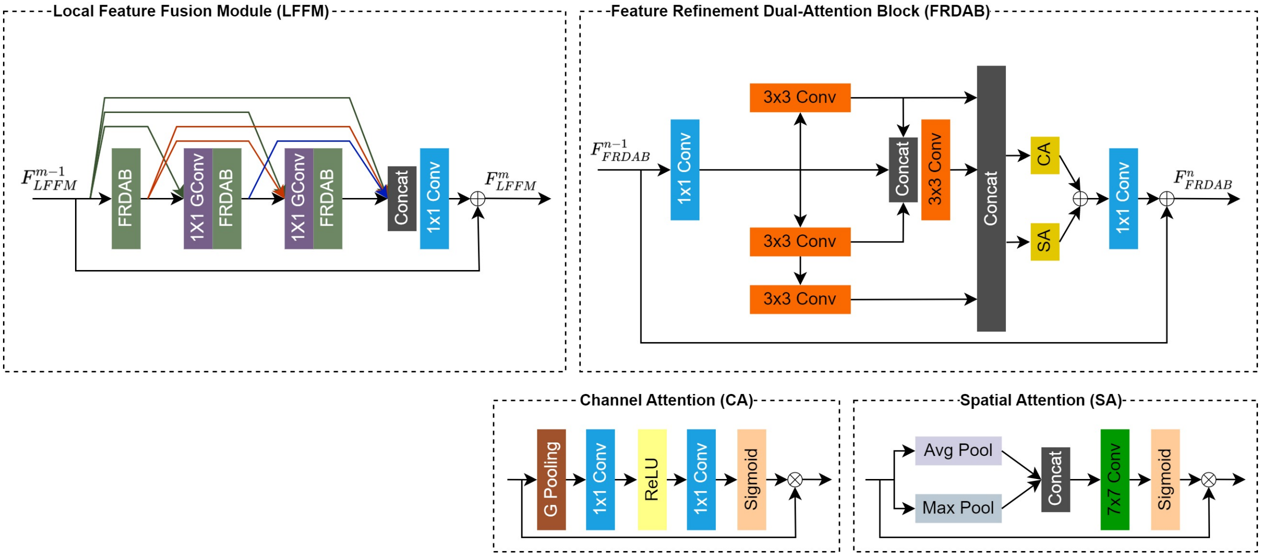

Figure 2. Local Feature Fusion Module (LFFM) and Feature Refinement Dual-Attention Block (FRDAB) architecture

(TM) and 2 convolutional layers. The finished operation of RT is represented as

$$F_{RT} = f_{3\times3}(f^{\circlearrowright}_{TM2}(f_{3\times3}(f^{\circlearrowright}_{TM1}(F_{CNN})))), \tag{13}$$

where $f_{3\times3}()$ and $f_{TM}()$ signify the convolutional layer and TM, respectively. The recurrent connection is denoted by $\circlearrowright$ indicating the output of TM will be functioned as its new input and looped S times. Only the encoding part of the standard Transformer structure is utilized which is inspired by ESRT [27] for the TM. As illustrated in 3, TM is primarily made up of two layer normalization layers, one Multi-Head Attention (MHA), and a Multi-Layer Perception (MLP) (MLP). Defining the input embeddings as F_{in} and the output embeddings F_{out} can be formulated by

$$F_{mid} = F_{in} + f_{MHA}(f_{norm}(F_{in})), \tag{14}$$

$$F_{out} = F_1 + f_{MLP}(f_{norm}(F_{mid})), \tag{15}$$

The layer normalization operation is represented by $f_{norm}()$. The MHA and MLP modules are depicted by $f_{MHA}()$ and $f_{MLP}()$. MHA's input feature map, like ESRT's [27], is projected into Q, K, and V via a linear layer to reduce GPU memory consumption. In the meantime, a feature reduction strategy is used to reduce the Transformer's memory consumption further. According to [41], each MHA head may undertake a scaled dot product attention, concatenate all outputs, and then conduct a linear transformation to result in the output. The scaled dot product attention is demonstrated as

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{16}$$

The Transformer can be fully trained and utilized using this strategy without increasing the model's parameters or GPU memory consumption.

3.2. Patch-Mosaic

This section outlines the proposed Patch-Mosaic method. Let $X \in \mathbb{R}^{C\times H\times W}$ represent an image and $a()$ denotes data augmentations, with $a() : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{C\times H\times W}$, $X^* = a(X)$. Here, X^* is the augmented image, and $a()$ may be any data augmentation or a combination of data augmentations. Unlike image-level augmentation, which applies the augmentation directly to the image $X^* = a(X)$, Patch-Mosaic initially cuts the image into four equally sized patches, with equal probability in either the height or width domain, as

$$[X_{1,2,3,4}] = cut_{H,W}(X), X_i \in R^{C\times\frac{H}{2}\times\frac{W}{2}} \tag{17}$$

Afterward, within each patch, $a()$ is processed separately, and the augmented pieces are integrated back together using the Mosaic augmentation method as

$$X^* = Mosaic[a_1(X_1), a_2(X_2), a_3(X_3), a_4(X_4)] \tag{18}$$

Randomness determines augmentations including random probabilities of being applied, random implementation, and random magnitudes. $a1()$, $a2()$, $a3()$, and $a4()$ are all instances of a(), but despite using the same data augmentation a(), they may differ significantly, enhancing diversity at both the local region and holistic image-levels.

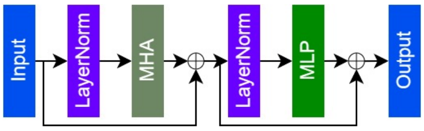

Figure 3. Transformer Module (TM)

In YOLOv4[18] the Mosaic data augmentation algorithm refers to the CutMix[19] data augmentation technique, which is a further improvement of CutMix. General methods of data augmentation are divided into 5 categories (1) Geometric transformation (Horizontal flip and Vertical flip), (2) Photometric transformation (Color jitter and Gaussian blur), (3) Information dropping (Random erasing and Cutout), (4) Search-based (AutoAug and RandAug), and (5) Mixbased (Mixup and CutMix). CutMix splices two images and sends the spliced images to the neural network for training. The Mosaic data augmentation algorithm utilized four images for Mosaics to produce a combined image comprising four original images that improved model training efficiency. Simultaneously, the Mosaic data augmentation algorithm has different options in a single original image. It can train multiple objects in the same composite image, which the original data enhancement algorithm cannot do. The workflow from conducting the Patch-Mosaic data augmentation algorithm on the DIV2K [5] dataset is depicted in figure 4.

4. Experiment

The machine with an AMD Ryzen Threadripper 2950X @ 4.40 GHz, RAM 64GB DDR4 2666MHz, NVIDIA GeForce GTX 2080 Ti 12GB x 2, and CUDA are used to train all the networks. The operating system is Ubuntu 20.04.4. Overall, the model has 32 input and output channels, three (n = 3) LFFMs, and the Transformer module recurses twice (S = 2).

4.1. Dataset

DIV2K [5] is used as the training set. Five benchmark test datasets are used for evaluation to validate the effectiveness of the proposed Patch-Mosaic with LBNet. Set5 [42], Set14 [43], BSDS100 [44], Urban100 [45], and Manga109 [46] are five benchmark test datasets used to validate the effectiveness of the proposed Patch-Mosaic with LBNet. Table 1 summarises the dataset and augmentation information for training models.

4.2. Evaluation metrics

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) are used as performance indicators for SR images on the Y channel of the YCbCr color space. These evaluations were demonstrated using two images. They are referred to as image 1 and image 2. Image 1 is the original degraded image from the test dataset, while image 2 is a reconstruction of image 1. SSIM is a method for determining how similar two images are. The SSIM values range from -1 to 1. PSNR is used to make comparisons of image 1's quality to image 2's quality calculated using the mean squared error (MSE). The mean square error (MSE) of an estimator represents the average of the squares of the estimation errors, or the discrepancy between the estimate and what is evaluated, lower MSE is better performance. Unlike, the higher the value of PSNR and SSIM, the more high-quality the reconstructed image is, as determined by equation 19, 20 and 21. The ground truth (reference) is represented by Y, and the reconstructed images are described by Y^*:

$$MSE = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}(Y^*(i,j) - Y(i,j))^2 \tag{19}$$

$$PSNR(Y, Y^*) = 10log_{10}\frac{255^2}{MSE} \tag{20}$$

The SSIM assessment between images P_{Y^*} and P_Y is summarized as follows:

$$SSIM(P_{Y^*}, P_Y) = \frac{(2\mu_{P_{Y^*}}\mu_{P_Y} + c_1)(2\sigma_{Y^*}\sigma_{P_Y} + c_2)}{(\mu^2_{P_{Y^*}} + \mu^2_{P_Y} + c_1)(\sigma^2_{P_{Y^*}} + \sigma^2_{P_Y} + c_2)} \tag{21}$$

where $\sigma_{P_Y}(\sigma_{P_{Y^*}})$ and $\mu_{P_{Y^*}}(\mu_{P_Y})$ denotes the knowledge and standard deviation of patch $P_{Y^*}(P_Y)$. Minor constants are defined as $c1$ and $c2$. Therefore, SSIM's average score patch-based over the image is SSIM (Y^*,Y).

4.3. Result

Compared to SISR SOTA models using image-level data augmentation methods, LBNet with the proposed Patch-Mosaic achieved better performance. A detailed comparison of the top 5 most common datasets with a 4x scale (Set5 [42], Set14 [43], BSDS100 [44], Urban100 [45], and Manga109 [46]) with LBNet [20], SwinIR [26], and ESRT [27] is shown in table 2. The results clearly show that LBNet (trained on DIV2K) with Patch-Mosaic acquired competitive results using the same parameters as LBNet. These results demonstrate the efficacy of the proposed Patch-Mosaic.

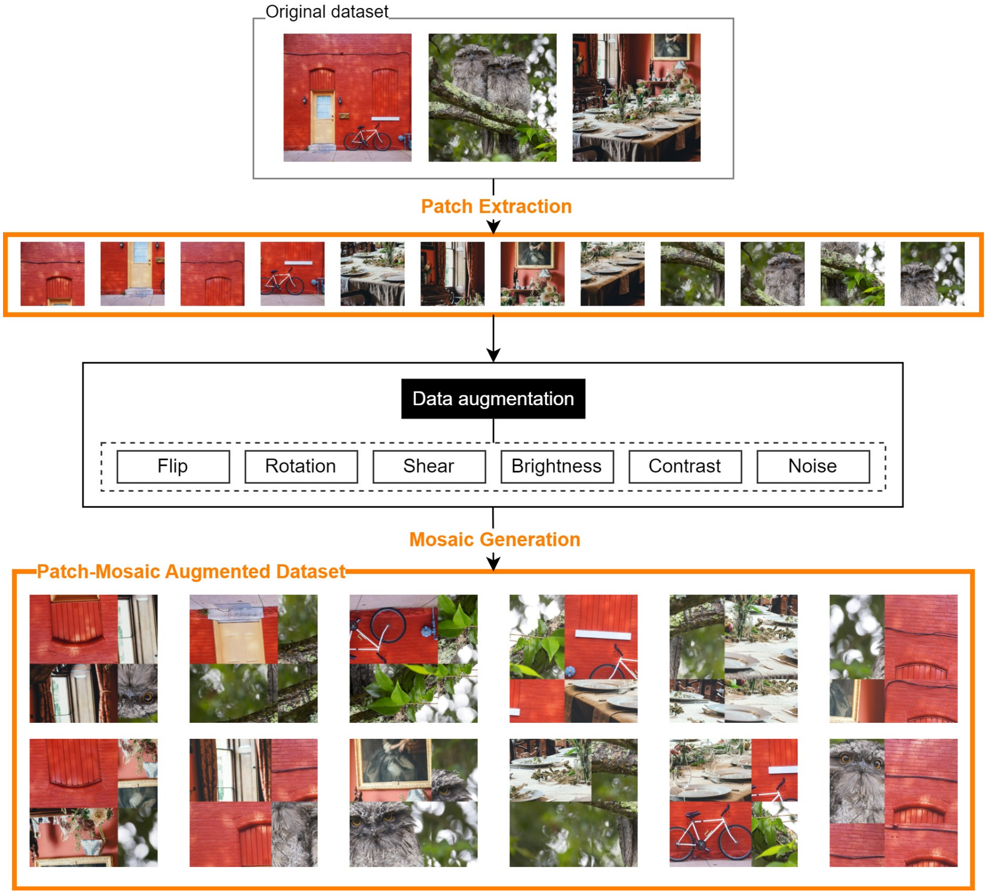

Figure 4. Illustration of the proposed Patch-Mosaic data augmentation

Table 1. Summary of dataset and augmentation method

Method + Dataset and data augmentation method	
SwinIR [26]	DIV2K [5], Flickr2K [47] + Image-level
ESRT [27]	DIV2K [5] + Image-level
LBNet [20]	DIV2K [5] + Image-level
LBNet [20](Patch-Mosaic)	**DIV2K [5] + Patch-Mosaic**

Table 2. Comparison with other SOTA methods (x4). With the Patch-Mosaic technique, LBNet can achieve better results with the same parameters as the original network

Method	Params	Set5	Set14	BSD100	Urban100	Manga109
SwinIR [26]	897K	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980	30.92/0.9151
ESRT [27]	751K	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	30.75/0.9100
LBNet [20]	742K	32.29/0.8960	28.68/0.7832	27.62/0.7382	26.27/0.7906	30.76/0.9111
LBNet [20](Patch-Mosaic)	**742K**	**32.98/0.9070**	**29.25/0.7919**	**28.20/0.7502**	**27.07/0.8101**	**31.55/0.9256**

5. Conclusion

In this study, a novel approach to data augmentation called Patch-Mosaic was introduced as a means to enhance the performance of single-image super-resolution (SISR) models. This technique was specifically applied to the Lightweight Bimodal Network (LBNet), which utilizes a Symmetric CNN for local feature extraction and a Recursive Transformer for capturing long-term image dependencies. The Symmetric CNN employs a Local Feature Fusion Module (LFFM) and a Feature Refinement Dual-Attention Block (FRDAB) to optimize feature extraction and utilization.

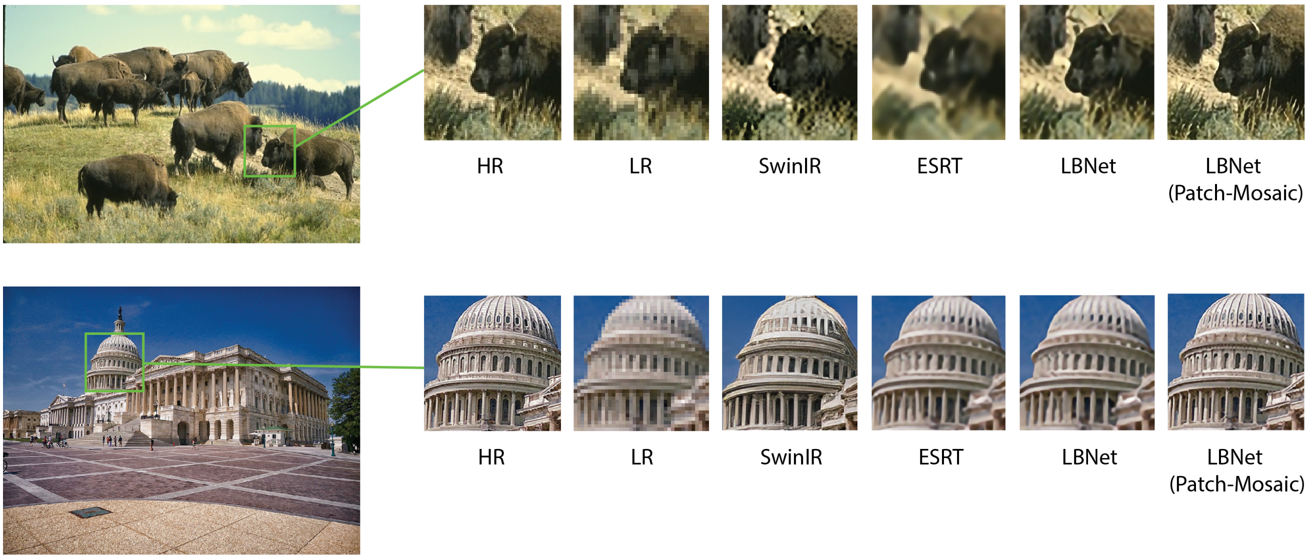

Figure 5. Visual comparison between LBNet trained with Patch-Mosaic versus SwinIR, ESRT, and original LBNet with image-level data augmentation method

The Recursive Transformer, on the other hand, trains the Transformer fully through its recursive mechanism, thereby enabling the Transformer to learn global constants and enhance features. In short, the Patch-Mosaic method effectively combines patch-level image augmentation with mosaic-generating data techniques to enhance the performance of lightweight SISR networks while reducing computational costs. The results of this study highlight the significance of data augmentation in improving the performance of super-resolution models. Notwithstanding, this research only conducted Patch-Mosaic with 4 patches (an image extracted for 4 patches, then a combined image includes 4 mosaic-augmented patches). In future research, working with more patches could be employed, it may have the potential to better the task of data augmentation for SISR.

References

[1] Dong C, Loy CC, He K, Tang X. Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. Springer; 2014. p. 184-99.

[2] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

[3] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700-8.

[4] Kim J, Lee JK, Lee KM. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 1646-54.

[5] Lim B, Son S, Kim H, Nah S, Mu Lee K. Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2017. p. 136-44.

[6] Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y. Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 286-301.

[7] Kim J, Lee JK, Lee KM. Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 1637-45.

[8] Tai Y, Yang J, Liu X. Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 3147-55.

[9] Ahn N, Kang B, Sohn KA. Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 252-68.

[10] Gao G, Li W, Li J, Wu F, Lu H, Yu Y. Feature distillation interaction weighting network for lightweight image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 661-9.

[11] Zhang D, Li C, Xie N, Wang G, Shao J. PFFN: Progressive Feature Fusion Network for Lightweight Image Super-Resolution. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021. p. 3682-90.

[12] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Toronto, Ontario: University of Toronto; 2009. 0.

[13] Kervrann C, Boulanger J. Optimal spatial adaptation for patch-based image denoising. IEEE Transactions on Image Processing. 2006;15(10):2866-78.

[14] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos. In: Computer Vision, IEEE International Conference on. vol. 3. IEEE Computer Society; 2003. p. 1470-0.

[15] Csurka G, Dance C, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. vol. 1. Prague; 2004. p. 1-2.

[16] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). vol. 2. IEEE; 2006. p. 2169-78.

[17] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.

[18] Bochkovskiy A, Wang CY, Liao HYM. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:200410934. 2020.

[19] Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 6023-32.

[20] Gao G, Wang Z, Li J, Li W, Yu Y, Zeng T. Lightweight Bimodal Network for Single-Image Super-Resolution via Symmetric CNN and Recursive Transformer. arXiv preprint arXiv:220413286. 2022.

[21] Li J, Pei Z, Zeng T. From beginner to master: A survey for deep learning-based single-image super-resolution. arXiv preprint arXiv:210914335. 2021.

[22] Hui Z, Gao X, Yang Y, Wang X. Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th acm international conference on multimedia; 2019. p. 2024-32.

[23] Lan R, Sun L, Liu Z, Lu H, Pang C, Luo X. MADNet: a fast and lightweight network for single-image super resolution. IEEE transactions on cybernetics. 2020;51(3):1443-53.

[24] Xiao J, Ye Q, Zhao R, Lam KM, Wan K. Self-feature learning: An efficient deep lightweight network for image super-resolution. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021. p. 4408-16.

[25] Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, et al. Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition;. p. 12299-310.

[26] Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R. Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 1833-44.

[27] Lu Z, Li J, Liu H, Huang C, Zhang L, Zeng T. Transformer for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 457-66.

[28] Efros AA, Leung TK. Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2. IEEE; 1999. p. 1033-8.

[29] Shocher A, Cohen N, Irani M. "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 3118-26.

[30] Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. Springer; 2020. p. 319-45.

[31] Han J, Shoeiby M, Petersson L, Armin MA. Dual contrastive learning for unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 746-55.

[32] Brendel W, Bethge M. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. arXiv preprint arXiv:190400760. 2019.

[33] Gontijo Lopes R, Yin D, Poole B, Gilmer J, Cubuk ED. Improving robustness without sacrificing accuracy with Patch Gaussian augmentation. arXiv e-prints. 2019:arXiv-1906.

[34] Dwibedi D, Misra I, Hebert M. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 1301-10.

[35] Georgakis G, Mousavian A, Berg AC, Kosecka J. Synthesizing training data for object detection in indoor scenes. arXiv preprint arXiv:170207836. 2017.

[36] Lin S, Yu T, Feng R, Li X, Jin X, Chen Z. Local patch autoaugment with multi-agent collaboration. arXiv preprint arXiv:210311099. 2021.

[37] Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:180509501. 2018.

[38] Qin Y, Zhang C, Chen T, Lakshminarayanan B, Beutel A, Wang X. Understanding and improving robustness of vision transformers through patch-based negative augmentation. arXiv preprint arXiv:211007858. 2021.

[39] Han J, Fang P, Li W, Hong J, Armin MA, Reid I, et al. You Only Cut Once: Boosting Data Augmentation with a Single Cut. arXiv preprint arXiv:220112078. 2022.

[40] Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. IEEE Transactions on computational imaging. 2016;3(1):47-57.

[41] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[42] Bevilacqua M, Roumy A, Guillemot C, Alberi Morel ML. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In: British Machine Vision Conference (BMVC). Guildford, Surrey, United Kingdom; 2012. Available from: https://hal.inria.fr/hal-00747054.

[43] Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations. In: International conference on curves and surfaces. Springer; 2010. p. 711-30.

[44] Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE

International Conference on Computer Vision. ICCV 2001. vol. 2. IEEE; 2001. p. 416-23.

[45] Huang JB, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 5197-206.

[46] Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, et al. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications. 2017;76(20):21811-38.

[47] Timofte R, Agustsson E, Van Gool L, Yang MH, Zhang L. Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2017. p. 114-25.

[48] Zhang X, Gao P, Liu S, Zhao K, Li G, Yin L, et al. Accurate and efficient image super-resolution via global-local adjusting dense network. IEEE Transactions on Multimedia. 2020;23:1924-37.