Continuous digital zooming using generative adversarial networks for dual camera system

Yifan Yang  |  Qi Li  |  Yongyi Yu  |  Zhuang He  |  Huajun Feng  |  Zhihai Xu  |  Yueting Chen

State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou 310000, China

Abstract
This paper presents a generative adversarial network (GAN) with patch match algorithm to realize a high-quality digital zooming using two camera modules with different focal lengths. In dual camera system, shorter focal length module produces the wide-view image with the low resolution. On the other hand, the longer focal length module produces the tele-view image via optical zooming. The long-focal image contains more details than short-focal image and can be used to guide short-focal image to reconstruct high frequency part. Firstly, a feature extraction block (FEB) is advanced to extract feature of long-focal image and short focal-image to reconstruct a wide-view image with different resolutions. Next, a patch match algorithm is integrated into convolution neural networks (CNN) to fuse information of long-focal with short-focal image and generate a new fused image. Finally, the fused image and short-focal image are merged with a feature fusion block (FFB) to predict high-resolution images. In addition, generative adversarial networks are used for filtering information integrated by previous network and output the zoomed image. Extensive experiments on benchmark datasets show that our algorithm achieves favorable performance against state-of-the-art methods.

1 | INTRODUCTION

It is well-known that high-resolution (HR) image can offer more details than its counterpart low-resolution (LR) image. These details are critical in various fields, such as remote sensing [1], medical diagnostic [2], intelligent surveillance [3], and so on. Continuous zooming system can continuously change the spatial resolution of an acquired image. Optical zooming system change the spatial resolution by moving the lens back and forth along the optical axis. As an economical alternative, digital zooming has a wide range of application areas in many imaging devices [4]. However, digital image processing algorithms such as image interpolation used in the digital zooming system generate jagging and blurring artifacts that significantly degrade the quality of resulting images. To solve this problem, a number of modified algorithms have been proposed in the past few decades [5, 6].

Existing digital zooming methods were proposed to increase the spatial resolution of the input image using either interpolation or super-resolution. Interpolation-based restoration approaches [7] aim to search the connections among the neighboring pixels and fill the missing pixels via a base function or an interpolation kernel. Though, it has fast processing time at low-computational complexity, the step-by-step forward approach fails to guarantee precision of the estimation, especially when there is noise.

To break the limitations of interpolation-based approaches, super-resolution (SR) methods [8–11] developed the image priors and learned the mapping relationships by sampling from a database containing a large number of HR and LR image pairs. But it requires an additional database and reconstruct the patch to build the HR image.

A dual camera system consists of two fixed-focus cameras with a common optical axis or a parallel optical axis. The two optical systems have different optical parameters, such as imaging focal length, exposure time, and sensor resolution. Therefore, images of different resolutions are acquired in the dual...
camera system. Figure 1 shows a classic dual camera system, in which imaging device one and imaging device two have the same optical axis. Light is split by dichroic prism and imaging into two imaging devices.

In a digital zooming system, long-focal image and short-focal image is used to enlarge image. To fuse the details of long-focal image and the structure of short-focal image, it is essential to patch two images. Figure 2 shows a conceptually simplified process of the digital zooming system under the assumption that two images have the same geometric parameters except for the focal length. To register two images, the short-focal image is first cropped into suitable size, and then up-sampled to the same size as short-focal image. However, it is difficult for neural networks to register two different images which are important for multi-image super resolution. Hence a patch match algorithm is advanced to match two images and transport them into neural networks. Then the long-focal image and the up-sampled image will be transported to digital zooming system and generate a high-resolution image which contents the details of long-focal image and the structure of short-focal image.

Actually, if just concatenate long-focal image and short-focal image together and input into convolution neural networks (CNN), the network will not make use of information of long-focal image to restore short-focal image, and the whole network is like a single-image super-resolution network, because two images are not aligned, and the network cannot handle it. Previous multi-frame super-resolution CNN used a registration module to align images. But for digital zooming system, two images are not similar in image scene, so registration cannot use a long-focal image to restore short-focal image scene which not in long-focal image. Patch match algorithm is advanced to divide the whole image into many patches and find the most similar patch in another image. In this way, this paper construct an reference image aligned to the short-focal image by long-focal image and achieved good results.

Although the experiment is based on the digital zooming system of a dual-camera with the same optical axis, it can be implemented on any dual-camera system. The implementation depends on the patch match algorithm which is for registering the image. In other words, a similar patch is searched for in the whole image, and the most similar patch is taken as the reference.

Generally, the short focal length is $\alpha_S$, and the long focal length is $\alpha_L$. The target magnification ratio is $\alpha$, $1 < \alpha \leq \alpha_L / \alpha_S$.

Major contribution of the proposed method includes:

- The paper proposes a generative adversarial network with a patch match algorithm (MatchGAN) for high-quality image digital zooming tasks. The patch match algorithm is integrated into networks which use the detailed information of the long-focal image to repair the corresponding area of the short-focus image and the structure of the neural network to reduce the image size to speed up the matching.
- The paper proposes feature extraction (FEB) and fusion block (FFB) to extract image feature for patch matching and fuse hierarchical features at different layer of our network. In order to prevent image information lost in deep network, feature fusion combine the shallow features and deep features together.
- Extensive experiments on several benchmark datasets demonstrate that MatchGAN achieves superior results compared with other state-of-art algorithms.

## 2 RELATED WORKS

At present, the continuous zoom scheme of the dual camera system is based on traditional pixel interpolation techniques,
such as bicubic, or single-frame image super-resolution based on machine learning or convolutional neural networks.

2.1 | Image super-resolution

Image super-resolution aims to reconstruct high-resolution images from low-resolution images. Mathematically, single-image super-resolution reconstruction is a serious ill-posed inverse problem, mainly due to the pixel number of a desired HR image larger than the number of a degraded LR image. Inspired by the manifold learning methods, Chang et al. [12] proposed a super-resolution method based on embedding-based scheme. This approach assumes that the LR and HR image patches share similar local geometry in two distinct feature spaces. Pandey et al. [13] classified super-resolution methods based on the priors used and they discussed image priors and regularization in detailed. But those methods are based on traditional ways, the results are worse than methods based on deep learning on visual and objective evaluation indicators.

Recently, deep learning based methods have achieved dramatic advantages against conventional methods. In 2014, Dong et al. [14] proposed the first end-to-end neural network SRCNN, but the network only has three layers, so it is difficult to learn the complex structures. The low convergence rate and working on a single scale are the other limitations of SRCNN. They established a mapping between the LR images and HR images. In 2016, VDSR [15] was proposed by Kim et al.; they used stack small filters and residual learning to accelerate the training and improved the performance of network. It is one of the first methods that utilize the residual learning to train the VDSR network. Because of using the residual image (residue between the ground-truth and bicubic interpolated image) and relatively large depth in the learning process, VDSR achieves superior results over SRCNN. However, with the depth of network growth, gradient disappear and gradient explosion will be the main problem to hinder the performance of the network. This problem was solved when He et al. [16] proposed residual net (ResNet). Additionally, the efficient sub-pixel convolution layer was proposed by ESPCN [17] to upscale the final LR feature maps into HR output to solve the problem of over-smooths and blurs in the original LR image. However, these methods tend to produce a blur result because of the loss function.

To solve this problem, Goodfellow [18] proposed generative adversarial network (GAN). They trained a generator to generator result and a discriminator to distinguish the result is real or not. As a result, the quality of the image improved greatly and image details were enriched. After that, GANs have been increasingly applied to various image processing fields [19–21].

2.2 | Patch match

In 2003, Criminisi et al. [22] presented a classical method to remove a large object according to the computation of priority and similarity of patches, while preserving important information of texture and structure. Afterward, many patch match methods have been proposed. Sun et al. [23] proposed a global structure propagation method which connected the geometry structure of the whole image manually and then propagated the known texture to an unknown region. Based on this, Li and Zhao [24] proposed an automatic structure completion method which avoided manual intervention and enhanced the efficiency of the algorithm. Wong et al. [25] proposed a non-local mean by utilizing multiple samples in an image to obtain more similar source patches. In 2009, Barnes et al. [26] proposed a seminal work of patch match, introduced an efficient way to find dense correspondences across images for structural editing. They used a large number of random samples and get good guesses. Due to its effectiveness in pruning out the search space, many researchers [27–30] extended this approach and applied in multiple domains. Korman and Avidan [30] incorporated the idea of image coherence into Locality Sensitivity Hashing and significantly improved the speed. He and Sun [29] combined KD-trees with patch match to perform more efficient neighbor matching. Especially, ref. [28] exploits patch match to overcome the infeasibility of searching over continuous output space.

2.3 | Application in dual camera system

To obtain a high-quality image, dual camera system is developed for many applications such as image denoising, super-resolution, object tracking and so on.

In 2016, Kim et al. [31] proposed an object tracking method using asymmetric dual camera system. They estimated the depth map between the wide- and tele-view image and used it to calculate the velocity of the object in the region of interest. Hicks et al. [32] presented an image denoising method in low-light condition with dual camera system. They used the first camera to acquire a monochrome image with high-sensitivity, spatial-resolution, and dynamic range. Using the second camera composed of RGB and near infrared pixels by replacing the green pixel of Bayer pattern with near infrared pixel to provide the chrominance information.

In 2017, Yu et al. [33] presented a theoretical basis to realize a high-quality digital zooming using two camera modules with different focal lengths. They fused wide- and tele-view images by performing variational-based image restoration using the estimated PSF. Moon et al. [34] proposed a super-resolution method using dual camera system. They fused wide- and tele-view images by using the ideal high-frequency component estimated from the optically zoomed image. However, the boundary region of the zoomed wide-view image cannot be restored which have no same scene with tele-view image. Ma et al. [35] proposed to use a band-inpainting algorithm to fuse wide- and tele-view images in frequency domain. They reconstructed band information with super-resolution methods, and introduced an image inpainting algorithm to reconstruct poorly super-resolved texture areas to produce photo-realistic digital zooming images. But these methods are based on traditional image restoration algorithms and they cannot restore finer structures compared with the method based on neural networks.
He et al. [36] proposed a digital zooming method using patch match algorithm for wide- and tele-view images fusion. They used Markov random field (MRF) models to patch different features of every layer in convolutional neural networks and continuously iteratively optimized. This algorithm cost too much time on features match and the region not in long-focal image cannot be repaired by this way. The results of the experiment did not improve much on objective evaluation index and visuals. In our method, MatchGAN just match image features only once and construct an reference image for whole short-focal image region.

3 GENERATIVE ADVERSARIAL NETWORKS WITH PATCH MATCH ALGORITHM

Our task is to fuse the image of long focal length with short focal length and generate a new high-resolution image with middle focal length which keeps important appearance details of the identity. It is a challenge to implement image fusion into an end-to-end model. In this paper, an approach is proposed to address this generative adversarial network with patch match algorithm (MatchGAN). Firstly, a feature extraction block (FEB) is advanced to extract image features for matching. Secondly, in the middle of the net, a patch match algorithm is presented to match the short-focal features and long-focal features. Lastly, a feature fusion block (FFB) is proposed for image restructuring and output the final image. In addition, generative adversarial networks (GAN) is used to train network, in order to generate more realistic image. Another reason this paper use GAN is that it can be interpreted as patch correspondences [37], so, in MatchGAN, a discriminator is used to improve the performance of long-focal features and short-focal features matching and increase more image details. The overall framework of the proposed MatchGAN is shown in Figure 3.

It is difficult to match image as well as achieve image super-resolution in an end-to-end network. So MatchGAN combine the traditional algorithm with neural network to solve this problem. In this paper, a short-focus length image $S$ with a long-focus length image $L$ are integrated to generate a coarse result $I$ that match the structure of $S$ and the details of $L$ in target image $I$.

3.1 Network structure

As shown in Figure 3, MatchGAN mainly consists three parts: feature extraction, image match and image restoration stage. Compared with previous CNN-based method, our CNN should both extract low-level and high-level features and combine them for path match, and feature extraction is proposed to concentrate multi-level features of different images. In addition, short-focal features are fused with patch-match features for further restoration stage, which is different from single-image super resolution, and an image restoration net structure is presented. Especially, before inputting into the network, wide-view image is enlarged to $\alpha$ times using bicubic interpolation and cropped into desired size. Tele-view image is resized to desired size by interpolation directly. Two FEBs are used to extract the feature of image and decrease image size, and further conduct patch match. There are two branches of feature extraction stage for extracting features of long-focal and
short-focal image which shares the same architecture and the same set of weights. The structure of short-focal image is preserved and then added to image restoration layer. The shallow layer extracts the structure of image, and deep layer extracts the details of image. MatchGAN have already extracted details of long-focus length image, as well as the structure of short-focal image by passing through two FEBs. Then, the details of long-focus image are matched to short-focal image. And two FFBs and several convolutions are used to restore high-resolution middle-focal image.

3.1.1 Feature extraction

As shown in Figure 4, our FEB contains local feature extraction and local residual learning and down sample layer, leading to contiguous memory mechanism.

Feature extraction is applied on inputs to extract features. As analyzed above, the first convolution layer extracts feature \( F_1 \) from input by

\[
F_1 = \sigma(Conv_1(F_{pre})),
\]

where \( Conv_1(\cdot) \) denotes convolution operation with stride 1, \( \sigma \) denotes ReLU [38] activation, \( F_{pre} \) is the output of the previous layer. \( F \) is used for the further layer to extract feature. So

\[
F_n = \sigma(Conv_1(F_{pre})),
\]

where \( n \) denotes \( n^{th} \) convolution operation with stride one.

Local residual learning is applied after feature extraction. As there are several convolution layers before residual learning. The final output of the residual learning is

\[
F_{dL} = F_{d1} + F_{d2} + \cdots + F_{dres},
\]

where \( d \) denotes \( n^{th} \) level feature extraction.

Input layer is then pass through one convolution layer to keep the same dimension with \( F_{dL} \), then add on \( F_{dL} \)

\[
F_{dL} = \sigma(Conv_1(F_{pre})) + F_{dL}.
\]

Down sampling is used to reduce the image size, so that MatchGAN can decrease the time of patch math algorithm. Convolution with stride two is used to reduce image size. So

\[
F_{d} = Conv_2(F_{dL}),
\]

where \( Conv_2(\cdot) \) denotes convolution operation with stride 2, and \( F_{d} \) denotes the output of FEB. All convolution is our FEB is 3 × 3. In MatchGAN, image features are extracted by two FEB and then applied to patch match.

3.1.2 Image match

With the features of two images, which may differ in appearance but have similar semantic structure, the mapping from long-focal features to short-focal features is needed to generate a new image based on the short-focal-image structures and long-focal-image details. By this way, any image areas can be mapped from long-focal images to short focal images, which means the input images can be taken on different axis. As long as the long-focal image have similar areas with short-focal image, the results will be improved.

Based on ref. [39], \( \phi \) represents a nearest-neighbor field (NNF) between two image features by

\[
\phi(p) = \arg \min_{\theta} \sum_{x \in N(p), y \in N(q)} \left( \| F_x(x) - F_y(y) \|_2^2 \right),
\]

where \( N(p) \) is the patch around \( p \). \( F(x) \) is a vector that represents all channels of feature layer at position \( x \). The patch size is set to 8 × 8. For each patch around pixel \( p \) in the short-focal features \( A \), its nearest neighbor position is found in long-focal features \( B \). Normalized features is used in patch similarity metric, and random search and propagation are adopt to support features maps.

For images, their NNFs are typically constant or linearly varying over large areas, so a random initialization and random search is used until convergence.

The \( N(p) \) location is initialized at random, most of the initial offsets are useless, but a certain number will be optimal or near-optimal. These are quickly diffused to the rest of the image in the propagation phase. Then the image is raster scanned top-down and left-to-right. MatchGAN use the nearby patch as an alternative estimates of current \( N(p) \), and select the best one.

To avoid getting trapped in bad local minima, random search is used when searching the most similar path. Besides it is time-consuming to search all the image path especially for high spatial resolution image. Therefore, MatchGAN adds a bi-dimensional random variable after every search. Experiments in ref. [26] show that typically patch match converges to a near-optimal NNF in less than 10 iteration. So this offset threshold is set as 1/10 of long-focal image size.
After the final NNFs $\phi$ is obtained, image $I_{\text{patch}}$ by patch aggregation can be reconstructed:

$$I_{\text{patch}}(p) = \frac{1}{n} \sum_{x \in N(p)} (B(\phi(x))) ,$$

(7)

where $n$ is the patch size of $N$.

Patch match is time-consuming, so in feature extraction stage, convolution with stride two is used to decrease image size. In image match stage, MatchGAN is faster because of small image size.

### 3.1.3 Image restoration

An image full filled with long-focal image details and short-focal image features are received from patch match. But the details of long-focal image cannot match the short-focal image perfectly, because short-focal image contains more areas than long-focal image. Long-focal image is used to repair the areas in short-focal image which resemble to long-focal image patch. So MatchGAN concatenate features of patch match and short-focal image features of the second FEB for further network to restore high-resolution image.

In image restoration stage, a feature fusion block (FFB) is proposed to reconstruct image. The structure of our FFB is shown in Figure 5. The last convolution filter is a de-convolution layer with stride two, this step can be formatted by

$$F_{\text{Deconv}} = \text{DeConv}_2(F_{\text{input}}) ,$$

(8)

where $\text{DeConv}_2(\cdot)$ denotes de-convolution operation with stride 2.

In FEB, convolution with stride 2 is used to speed up patch match. Part of image information which is essential in image restoration is lost as the same time of decreasing image size. To prevent information loss, MatchGAN add the output of FEB to layer after de-convolution, considering the whole network of match as a residual learning network. This operation helps us accelerate network convergence and reduce the difficulty of learning. Then local residual is used to restore image by

$$F_{f} = F_{f1} + F_{f2} + \cdots + F_{fn} ,$$

(9)

where $f$ denotes $f^{th}$ level feature fusion, $n$ denotes $n^{th}$ convolution operation with stride one.

In image restoration stage, two FFBs are used to restore image. After FFB, the zoomed image is generated through three convolution.

### 3.2 Discriminator

The discriminator network is used to discriminate real HR images from generated SR samples. The discriminator commonly composed of a series of convolution layer, ReLU [38] and batch-normalization [40] as shown in Figure 6.

Follow the architecture by Ledig et al. [41], fewer filters are used considering the weight of GAN loss is small compared to regular term. Discriminator model contains six convolutional layers, and the resulting one feature maps are followed by a final sigmoid activation function to obtain a probability for sample classification.

### 3.3 Loss function

$L_1$ norm is used as one part of loss function in MatchGAN, because $L_1$ loss function does not over-penalize larger errors, and is proved to be more powerful for performance and convergence [42].

$$\mathcal{L}_{l_1} = \|G(S, L) - I_{HR}\|_1 ,$$

(10)

where $G(S, L)$ is the generator with short-focus length image $S$ and long-focus length image $L$ input, and $I_{HR}$ denotes the real HR image.

But it is well known that $L_1$ loss produces blurry results on low-level problems. So GAN discriminator is used to make our zoomed image more realistic.

The discriminator has only one output to classify the zoomed image is real or fake by

$$\min_{G} \max_{D} \mathcal{V}(D, G) = E_{S \sim p_{S}(S), L \sim p_{L}(L)}[\log(1 - D(G(S, L)))] + E_{I_{HR} \sim p_{\text{data}}(I_{HR})}[\log(D(I_{HR}))] ,$$

(11)
where $D$ denotes the discriminator which is used to identify whether the input image is a real image.

### 3.4 Implementation details

In MatchGAN, the size of all convolutional layers is set to $3 \times 3$ except the first layer in FEB whose kernel size is $5 \times 5$, and the number of filters is shown on figures. The convolution layers in feature extraction, image restoration stage have 16 filters, and the filter numbers in Figure 2 are 16, 32, 16 and 3 from left to right. In discriminator, the number of channels for all filters is 16 except the first one is 3 and the last is 1. The convolutional filters are initialized using the method of He et al. [43] For all the convolution layer, this paper pad zeros to each side of the input to keep size fixed. The final convolution layer has three output channels, as output high-resolution color image. The network can also process gray images if change that to one. The patch size in image match stage is set to $8 \times 8$.

Experiment is based on the dual camera with the same optical axis. In training and testing phase, MatchGAN use the center area of the image as long-focal image and down sample the whole image as the short-focal image. For the magnification less than 2, the center 1/2 area is used as long-focal image. For the magnification between 2 and 4, center 1/4 area is used as long-focal image.

DIV2K [44] is used as our training set. DIV2K [44] consists of 800 training images, 100 validation images, and 100 test images. All of our models are trained with 800 training images and 10 validation images are used in the training process. In each training batch, this paper randomly sample 16 patches with the size of $128 \times 128$. An epoch has 1000 iterations of back-propagation. LR image is generated by bicubic interpolation. This paper implement our MatchGAN with the tensorflow framework and update it with Adam optimizer [45]. The learning rate is initialized to $2e^{-4}$ for all layers and decreased by a factor of 0.96 for every 100 epochs. Followed by Ledig et al. [41], we alternate updates to the generator and discriminator network.

### 4 EXPERIMENTS AND ANALYSIS

In this section, we first analyze the contributions of different components of proposed network. We then compare our MatchGAN with state-of-the-art algorithms on four benchmark datasets.

| Algorithm     | SET5 PSNR/SSIM | SET14 PSNR/SSIM | BSD100 PSNR/SSIM | URBAN100 PSNR/SSIM |
|---------------|----------------|-----------------|-----------------|-------------------|
| Without GAN   | 36.56/0.9526   | 33.26/0.9167    | 32.28/0.8957    | 30.23/0.8432      |
| Without patch match | 36.17/0.9452 | 33.47/0.9158    | 31.47/0.8910    | 32.18/0.9019      |
| MatchGAN      | **37.78/0.9626** | **33.35/0.9170** | **32.36/0.8996** | **32.77/0.9182**  |
TABLE 2  Quantitative evaluation of state-of-the-art SR algorithms: Average PSNR/SSIM for scale factors 1.5, 2, 3 and 4

| Algorithm | scale | SET5 PSNR/SSIM | SET14 PSNR/SSIM | BSD100 PSNR/SSIM | URBAN100 PSNR/SSIM |
|-----------|-------|----------------|----------------|------------------|-------------------|
| Bicubic   | 1.5   | 34.20/0.9318   | 31.51/0.8708   | 30.91/0.8599     | 27.23/0.8432      |
| SRCNN     | 1.5   | 36.60/0.9408   | 32.16/0.9057   | 32.11/0.8904     | 29.85/0.8978      |
| FSRCNN    | 1.5   | 37.21/0.9522   | 32.87/0.9075   | 32.06/0.8997     | 30.26/0.8973      |
| VDSR      | 1.5   | 38.43/0.9617   | 33.84/0.9170   | 33.23/0.9055     | 30.39/0.9115      |
| LapSRN    | 1.5   | 38.57/0.9638   | 33.86/0.9186   | 33.08/0.9020     | 31.36/0.9123      |
| IDN       | 1.5   | 38.57/0.9633   | 34.23/0.9228   | 33.20/0.9059     | 31.73/0.9163      |
| Ours      | 1.5   | **38.69/0.9698** | **34.31/0.9231** | **33.49/0.9048** | **32.71/0.9210**  |
| Bicubic   | 2     | 33.66/0.9299   | 30.24/0.8688   | 29.56/0.8431     | 26.88/0.8403      |
| SRCNN     | 2     | 36.66/0.9542   | 32.45/0.9067   | 31.36/0.8879     | 29.50/0.8946      |
| FSRCNN    | 2     | 37.00/0.9558   | 32.63/0.9088   | 31.01/0.8876     | 30.74/0.8976      |
| VDSR      | 2     | 37.53/0.9587   | 33.03/0.9124   | 31.90/0.8960     | 30.76/0.9140      |
| LapSRN    | 2     | 37.52/0.9591   | 32.99/0.9124   | 31.80/0.8952     | 30.91/0.9103      |
| IDN       | 2     | **37.83/0.9600** | 33.30/0.9148   | 32.08/0.8985     | 31.27/0.9196      |
| Ours      | 2     | **37.78/0.9626** | **33.35/0.9170** | **32.36/0.8996** | **32.77/0.9182**  |
| Bicubic   | 3     | 30.39/0.8682   | 27.55/0.7742   | 27.21/0.7385     | 24.46/0.7349      |
| SRCNN     | 3     | 32.75/0.9090   | 29.30/0.8215   | 28.41/0.7863     | 26.24/0.7989      |
| FSRCNN    | 3     | 33.16/0.9140   | 29.43/0.8242   | 28.13/0.7651     | 26.84/0.8011      |
| VDSR      | 3     | 33.66/0.9213   | 29.77/0.8314   | 28.82/0.7976     | 27.14/0.8279      |
| LapSRN    | 3     | 33.81/0.9220   | 29.79/0.8325   | 28.82/0.79804    | 27.07/0.8275      |
| IDN       | 3     | 34.11/0.9253   | 29.99/0.8354   | 28.95/0.8013     | 27.42/0.8359      |
| Ours      | 3     | **34.72/0.9313** | **30.19/0.8414** | **29.43/0.8201** | **28.92/0.8573**  |
| Bicubic   | 4     | 28.42/0.8104   | 26.00/0.7027   | 25.96/0.6675     | 23.14/0.6577      |
| SRCNN     | 4     | 30.48/0.8628   | 27.50/0.7513   | 26.90/0.7101     | 24.52/0.7221      |
| FSRCNN    | 4     | 30.71/0.8657   | 27.59/0.7535   | 26.81/0.7035     | 24.99/0.7048      |
| VDSR      | 4     | 31.35/0.8838   | 28.01/0.7674   | 27.29/0.7251     | 25.18/0.7524      |
| LapSRN    | 4     | 31.54/0.8852   | 28.09/0.7700   | 27.32/0.7275     | 25.21/0.7562      |
| IDN       | 4     | **31.82/0.8903** | **28.25/0.7730** | **27.41/0.7297** | **25.41/0.7632**  |
| Ours      | 4     | 31.63/0.8811   | **28.45/0.8365** | **28.87/0.7486** | **26.40/0.7705**  |

4.1  Model analysis

4.1.1  Patch match

To validate the effect of patch match structure, we remove the image match stage and double the filters of the last layer of feature extraction stage to use the same number of MatchGAN. In Figure 7, compared with the second picture, MatchGAN without patch math generate more artifacts, because it cannot find corresponding area of reconstructed patch and generate many artificiality as the result without patch match. In Table 1, the quantitative results show that the patch match in CNN leads to moderate performance improvement.

4.1.2  GAN

Generally, GAN is interpreted as more realistic details generation and lower PSNR and SSIM [46]. In ref. [37], it is developed as a patch correspondence and applied in many image inpainting programs. The result shown in Table 1 proved that GAN can improve the performance of patch match, as the consequence, our network with GAN achieved higher PSNR and SSIM. Based on the results of ref. [37] and our own experimental results, the mechanism of GAN kind of matches similar patch, but its specific principle needs further discussion and experiments. When we provide similar area in long-focal image, it will use it to reconstruct similar area in short-focal image. When just used as single-image super-resolution, it will use similar area of training set to reconstruct image. That is the reason why we achieved higher PSNR in Section 4.2. In Figure 7, we can find that MatchGAN generated more details than MatchGAN without GAN.

4.2  Experiment results

We compare the proposed algorithm with other six SR algorithms: bicubic, SRCNN [14], FSRCNN [47], VDSR [15],
LapSRN [48], IDN [49]. For digital zooming, we generate the training data by randomly sampling $128 \times 128$ patches from four different datasets: Set5 [50], SET14 [51], BSD100 [52], URBAN100 [53]. Among these datasets, SET5, SET14 and BSD100 consist of natural scenes; URBAN100 contains many images with details in different frequency. We compare the proposed MatchGAN with several state-of-the-art SR algorithms.

We evaluate the SR images with two commonly used image quality metrics: PSNR, SSIM [46]. All the experiments are implemented on GTX 1080 TI GPU, we train those algorithms with the same strategy. Table 2 shows quantitative comparisons for $1 \times 2 \times 3 \times 4 \times 5$ SR. Our MatchGAN performs favorably against existing methods on most datasets. In particular, MatchGAN achieves higher SSIM values even PSNR is not the best among those algorithms, which has shown that our feature fusion can fuse similar patch of different image region.

In Figure 8, we show visual comparison on Urban100 with a scale factor of 1.5. For comparison, we only show the other three best algorithms on the figures. Our method reconstructs more clear patterns on the stone brick. Besides our approach suppresses artifacts by matching image patch which usually occurs in GANs. As shown in Figure 7, our result also are less artificial than algorithm without patch. In Figure 9, we can find more clear broken lines on tiles. Generally, the more similar the scene in long-focal and short-focal image, the better the quality of reconstructed image achieved, because we can generate reference image to reconstruct image. In Figure 10, we achieved higher PSNR and SSIM compared with other methods, but in the center area (we use this area as long-focal

![Image of Figure 8](image8.jpg)

**Figure 8:** Urban100 with a scale factor of 1.5. Our method shows more clear patterns on the stone brick.

![Image of Figure 9](image9.jpg)

**Figure 9:** Clear broken lines on tiles. Our method achieves higher quality.

![Image of Figure 10](image10.jpg)

**Figure 10:** Higher PSNR and SSIM compared with other methods.

![Image of Figure 11](image11.jpg)

**Figure 11:** Results on real images ($\times 2$) in red frame.
4.3 Qualitative evaluations on real image

We further qualitatively evaluate the proposed algorithm against state-of-the-art methods on real image. To capture the realistic tele- and wide-view image pairs, we use iPhone X with two camera lens mounted on a tripod. We define images captured at 51 mm focal length as the wide image and images captured at 28 mm focal length as the tele-view image. Those two images have a certain displacement and parallax, and are not perfectly aligned. As we stated in Section 3.1.2, MatchGAN has no strict restrictions on input images. The results in Figure 11 show that the flower in the picture is more obvious in MatchGAN with the input of Figure 12. And Figure 13 shows that the state-of-the-art methods are still blurry with the input of Figure 14. In contrast, MatchGAN generates the images with clearer detailed structures, which demonstrate that the proposed algorithm generalizes well.

5 LIMITATIONS

While our method is capable of generating high-quality HR image on a certain scale, it is time-consuming. Patch math is a slow algorithm. We use some convolution layers with stride two as the basic structure of our model because we can decrease the time of patch match algorithm greatly, but compared with other networks, it is still slower and we are trying to fix it in the future. Additionally, we use long-focal image to guide short-focal image reconstruction. When the scale factor is too large like 8× and 16×, the long-focal image is too small compared with short-focal image to guide the reconstruction and the concatenate structure will make some confusion to original image, so our method is not suitable for large scale factor.

6 CONCLUSIONS

In this work, we proposed a deep convolutional network with patch match algorithm for continuous digital zooming. Our model progressively predicts high-frequency residuals guided by long-focal image. By using the feature structure, we decrease the time of patch match algorithm, and the matched image can concatenate to the original image for details reconstruction. Extensive evaluations on benchmark datasets demonstrate that the proposed model performs favorably against the state-of-the-art SR algorithms in terms of visual quality.

ACKNOWLEDGEMENTS

This work was supported by Basic Research Projects (2018110C081) and the Science and Technology Project of Jiangsu Province (BE2019063).

REFERENCES

1. Lu, H., et al.: Reference information based remote sensing image reconstruction with generalized nonconvex low-rank approximation. Remote Sens. 8(6), 499 (2016)
2. Greenspan, H.: Super-resolution in medical imaging. Comput. J. 52(1), 43–63 (2009)
3. Wang, N., et al.: A comprehensive survey to face hallucination. Int. J. Comput. Vis. 106(1), 9–30 (2014)
4. Lu, J., Healy, D.: Contrast enhancement via multiscale gradient transformation. In: Proceedings of 1st International Conference on Image Processing, vol. 2. pp. 482–486. IEEE Computer Society, Los Alamitos, (1994)
5. Battito, S., et al.: A locally adaptive zooming algorithm for digital images. Image Vision Comput. 20(11), 805–812 (2002)
6. Park, S.C., et al.: Super-resolution image reconstruction: A technical overview. IEEE Signal Process Mag. 20(3), 21–36 (2003)
7. Hou, H.S., Andrews, H.C.: Cubic splines for image interpolation and digital filtering. IEEE Trans. Acoust. Speech Signal Process. 26, 508–517 (1979)
8. Cui, Z., et al.: Deep network cascade for image super-resolution. In: Leit, D., Padilla, T., Schiele, B., Tuylstraers, T. (eds.) European Conference on Computer Vision, pp. 49–64. Springer International Publishing, Cham (2014)

9. Deng, C., et al.: Similarity constraints-based structured output regression machine: An approach to image super-resolution. IEEE Trans. Neural Netw. Learn. Syst. 27(12), 2472–2486 (2016)

10. Kumar, N., Sethi, A.: Fast learning-based single image super-resolution. IEEE Trans. Multimedia 18(8), 1504–1515 (2016)

11. Zeng, K., et al.: Coupled deep autoencoder for single image super-resolution. IEEE Trans. Cybern. 47(1), 27–37 (2017)

12. Chang, H., et al.: Super-resolution through neighbor embedding. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2004)

13. Pandey, G., Ghanebar, U.: Classification of priors and regularization techniques pertinent to single image super-resolution. Vis. Comp. 36, 1291–1304 (2020)

14. Chao, D., et al.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision. Springer, Cham (2014)

15. Kim, J., et al.: Accurate image super-resolution using very deep convolutional networks. CoRR (2015)

16. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Piscataway (2016)

17. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. arXiv:1609.05158, 2016

18. Goodfellow, I., et al.: Generative adversarial networks. Advances in Neural Information Processing Systems, vol. 3. MIT Press, Cambridge (2014)

19. Yuan, Q., et al.: Blind motion deblurring with cycle generative adversarial networks. Vis. Comp. 36, 1591–1601 (2020)

20. Pan, J., et al.: Physics-based generative adversarial models for image restoration and beyond. arXiv:1808.00605, 2018

21. Kupyn, O., et al.: DeblurGAN: Blind motion deblurring using conditional adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8183–8192. IEEE, Piscataway (2018)

22. Criminisi, A., et al.: Object removal by exemplar-based inpainting. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005). IEEE, Piscataway (2003)

23. Jian, S.: Image completion with structure propagation. ACM Trans. Graph. 24 (2005)

24. Li, S., Zhao, M.: Image inpainting with salient structure completion and texture propagation. Pattern Recognit. Lett. 32(9), 1256–1266 (2011)

25. Wong, A., Orchard, J.: A non-local-means approach to exemplar-based inpainting. In: 2008 15th IEEE International Conference on Image Processing. IEEE, Piscataway (2008)

26. Kuevratkupong, P., et al.: Patchmatch: A randomized correspondence algorithm for structural image editing (2009)

27. Barnes, C., et al.: The generalized patchmatch correspondence algorithm. In: European Conference on Computer Vision. Springer, Cham (2010)

28. Besse, F., et al.: Pmhp: Patchmatch belief propagation for correspondence field estimation. Int. J. Comput. Vision 110(1), 2–13 (2014)

29. Sun, J., He, K.: Computing nearest-neighbor fields via propagation-assisted kd-trees. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2012)

30. Korman, S., Avidan, S.: Coherency sensitive hashing. IEEE Trans. Pattern Anal. Mach. Intell. 38(6), 1099–1112 (2015)

31. Han, H.C.: Digital photographing apparatus and method of operating the same (2017)

32. Hicks, R.: Infrared and visible light dual sensor imaging system (2017)

33. Yu, S., et al.: Continuous digital zooming of asymmetric dual camera images using registration and variational image restoration. Multidimension. Syst. Signal Process. 29, 1959–1987 (2018)

34. Moon, B., et al.: Continuous digital zooming using local self-similarity-based super-resolution for an asymmetric dual camera system. J. Opt. Soc. Am. A Opt. Image Sci. Vis. 34(6), 991–1003 (2017)

35. Ma, H., et al.: Photo-realistic continuous digital zooming for an asymmetrical dual camera system. Opt. Laser Technol. 109, 110–122 (2019)

36. Gui-ran, H., et al.: Dual-focal camera continuous digital zoom based on cnr and feature extraction. J. Zhejiang Uni. (Eng. Sci.) 53(6), 1182–1189 (2019)

37. Fragoso, V., et al.: Patch correspondences for interpreting pixel-level CNNs. arXiv:1711.10683, 2017

38. Ghorou, X., et al.: Deep sparse rectifier neural networks. JMLR 15, 01 (2010)

39. Liao, J., et al.: Visual attribute transfer through deep image analogy. ACM Trans. Graphics 36(4), 120 (2017)

40. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. PMLR 37, 44–51 (2015)

41. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv:1609.04802, 2016

42. Lim, B., et al.: Enhanced deep residual networks for single image super-resolution. arXiv:1707.02921, 2017

43. He, K., et al.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. arXiv:1502.01852, 2015

44. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, Piscataway (2017)

45. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980, 2014

46. Wang, Z., et al.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13(4), (2004)

47. Dong, C., et al.: Accelerating the super-resolution convolutional neural network. In: eibe, B., Matas, J., Sebe, N., Welling, W. (eds) European Conference on Computer Vision, pp. 391–407. Springer International Publishing, Cham (2016)

48. Lai, W.S., et al.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5835–5843. IEEE, Piscataway (2017)

49. Zheng, H., et al.: Fast and accurate single image super-resolution via information distillation network. arXiv:1803.09454, 2018

50. Bevilacqua, M., et al.: Neighbor embedding based single-image super-resolution using semi-non-negative matrix factorization. In: IEEE International Conference on Acoustics, IEEE, Piscataway (2012)

51. Zeyde, R., et al.: On single image scale-up using sparse-representations. In: Boissonnat, J.-D., Chenin, P., Cohen, A., Gout, C., Lyche, T., Mazure, M.-L., Schumaker, L. (eds.) International Conference on Curves and Surfaces. Springer, Berlin Heidelberg (2012)

52. Martin, D., et al.: 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. vol. 2, pp. 416–423. IEEE, Piscataway (2001)

53. Huang, J.B., et al.: Single image super-resolution from transformed self-exemplars. In: Computer Vision and Pattern Recognition. IEEE, Piscataway (2015)