MFDN: Multiception Feature Distillation Network

Syed Sameen Ahmad Rizvi, Usneek Singh, Pratik Narang
Department of Computer Science and Information Systems
Birla Institute of Technology and Science
Pilani, Rajasthan, India
{p20190412,f20190127,pratik.narang}@pilani.bits-pilani.ac.in

***Abstract*—Image super-resolution refers to obtaining high-resolution images with refined details and enhanced visual quality from low-resolution inputs with coarse details. Because of its wide range of applications, image super-resolution has attracted the attention of researchers in the image processing and computer vision community. Even though CNN and GAN-based approaches are quite successful, it is unfeasible to use them in edge devices due to their heavy computational complexity. This work is our attempt in this direction to develop a lightweight image super-resolution model which uses multiception convolution layers for feature distillation. Our inspiration in this work was Residual Feature Distillation Network (RFDN), where we substantially decreased the parameters while maintaining the PSNR metric. The proposed approach was able to achieve image super-resolution with enhanced visual quality as compared to the baseline approach.**

***Index Terms*—Super-Resolution, Deep Learning, CNN.**

I. Introduction

Image super-resolution (SR) refers to the task of upscaling a given low-resolution (LR) image with coarse details to a high-resolution (HR) image with refined details and enhanced visual quality. Image super-resolution is often associated to other terminologies such as upsampling, interpolation, image scaling, zooming and enlargement. It has been proved that upsampling an image via super-resolution methods can largely refine the amount of available information and thus lead to accurate and robust vision-based machine learning systems [1]. As a result, super-resolution methods have achieved cross domain acceptability and enjoy a wide range of applications such as medical imaging, surveillance and security, aerial imaging, compressed image/video enhancement, action recognition, remote sensing , astronomical images, forensics, pose estimation, fingerprint and gait recognition and many more. Apart from improving the perceptual quality, it also helps in other deep learning based computer vision tasks such as object detection, image segmentation.

Super-resolution (SR) is a classical problem in the field of image processing and computer vision as it is an ill-posed inverse problem, i.e., instead of a unique solution there can be multiple solutions for the same low-resolution image. [2] Furthermore, the complexity of the problem is proportional to the upscaling factor i.e. at higher upscaling rates, the retrieval of finer details is even more complex and results in reproduction of wrong information. Image super resolution can be broadly classified into two categories, namely classical Image super-resolution and deep learning based super-resolution. In literature, a variety of classical SR methods have been proposed which include prediction based methods [3], [4], edge based methods [5], [6], statistical methods [7], [8], patch-based methods [5] and sparse representation methods [9]. With the dramatic increase in terms of data and computational capabilities, deep learning methods have become popular in a number of applications including image super-resolution. A variety of deep learning models have been applied in this regard ranging from Convolutional Neural Networks based SRCNN [10] to the most recent adversarial learning based SRGAN [11].

Even though recent advances in the field of image super-resolution have proved the supremacy of CNN and GAN based models, due to their heavy computational costs, it becomes unfeasible to use them in edge devices. This work is our attempt to further develop a light-weight image super-resolution model that reduces the parameters without affecting the performance of existing state of the art models. Our inspiration in this regard was the Residual Feature Distillation Network [12]. The major contributions of this work includes:

- We propose a super-resolution architecture with 4 stages – feature extraction block, feature distillation residual blocks, fusion block and reconstruction block. Feature distillation blocks use multiception convolution layers to improve parameter metric of the model.
- We propose pixel attention layer [13] after every upsampling layer in the reconstruction block to improve super-resolution results.
- We propose to train the network with smooth L1 loss instead of L1 loss as this loss function is more robust to outliers.

II. Related Work

With the evolution of deep learning paradigm, notable advances have been done in image super-resolution. Exhaustive surveys [2], [14], [15] have been done to summarize the prominent works done in deep learning based image super-resolution. In this section, we mention few notable works discussed in the literature .

A. Linear Networks

Linear networks consists of a single path for signal flow without any multiple branches. Several convolutional layers are queued one after the another, hence the input follows a sequential flow form initial to subsequent layers. Some of

the notable linear super-resolution networks are discussed as follows.

Super Resolution Convolutional Neural Network [10] popularly known as *SRCNN* was the first successful attempt that utilized only convolutional layers for image super-resolution task.SRCNN can be considered as a landmark work in deep learning based super-resolution that inspired several attempts in this direction. Fast Super-Resolution Convolutional Neural network [16]or *FSRCNN* builds upon SRCNN [17] to improve its speed and quality. The obejctive was to enhance the rate of computation to real-time (24 fps) as compared to SRCNN (1.3 fps). Very Deep Super-Resolution [18] *VDSR* is based on a deep CNN model originally proposed in VGG-net [19] and uses fixed-size convolutions (3×3) in all network layers.

B. Residual Networks

Unlike linear networks,residual learning used skip connections in the network architecture, to avoid gradients vanishing and makes it feasible to design a very deep net. In this paradigm algorithms learn what is called a residue i.e. the high frequencies between input and ground truth.

The Enhanced Deep Super-Resolution [20] *EDSR* builds upon the ResNet [21] architecture that was originally proposed for image classification to perform the image super-resolution task. Cascading residual network [22] *CARN* also levies ResNet [21] blocks to learn the relationship between low-resolution input and high-resolution output. Balanced Two-Stage Residual Network [23] *BTSRN* another residual learning based super resolution method uses two staged upsampling where feature maps are upsampled using a deconvolution followed by nearest neighbour upsampling

C. Densely Connected Networks

The DenseNet's [24] success in image classification inspired the researchers working in image super-resolution to develop algorithms based on densely connected CNN layers with improved performance.

SRDenseNet [25] the architecture uses dense connections, i.e. a layer directly operates on the output it receives from all previous layers. Information flows from low to high-level feature layers while avoiding the vanishing gradient problem, learning compact models and quickens the training process. Residual Dense Network *RDN* [26] integrates residual skip connections with dense connections. The idea is that the hierarchical feature representations should be completely used to learn local patterns.

D. GAN based Models

Generative Adversarial Networks (GAN) [27] uses a game theory based approach, which has two networks, namely a generator and discriminator network. The Generator tries to fool the discriminator by generating artificial high-resolution images. Finally when the algorithm converges the generator creates super-resolved images that a discriminator cannot distinguish as a real high-resolution image or an artificially super-resolved output. Accordingly, high-resolution images with better perceptual quality are generated.

SRGAN [11] used an adversarial objective function that promotes super-resolved outputs that are close to the manifold of natural images. *ESRGAN* [28] extends SRGAN [11] by incorporating dense blocks instead of batch normalization. Input of each dense block was also connected to the output of the respective block thereby making a residual connection over each dense block. Global residual connection was also used to enforce residual learning.

III. Methodology

A. The overall pipeline

The framework of the super-resolution technique proposed in this paper is based on the Residual Feature Distillation network (RFDN) [12]. Figure 1 explains the overall architecture of our proposed Multiception Feature Distillation Network (MFDN). The network consists of four parts: the first feature extraction convolution, multiple stacked residual feature distillation blocks (referred as RFDBs in [12]), the feature fusion part and the last reconstruction block. The first part of the network extracts coarse features from the input image; that can be stated by the equation:-

$$F_o = h(x) \tag{1}$$

where F_o is the extracted feature from input x using the h function. Then the stacked residual blocks gradually refine the features extracted earlier. It can be formulated as:-

$$F_k = H_k(F_{k-1}) \tag{2}$$

where F_k denotes the feature extracted from the k^{th} RFDB block using the F_{k-1} as the input. After refinement of features, all the features vectors are concatenated together in the fusion module. Final super resolved image is constructed using the reconstruction network. It can be stated as:-

$$y = R(F_{assemble} + F_o) \tag{3}$$

where y is the output image constructed using the R reconstruction function that takes input as F_o and $F_{assemble}$ where $F_{assemble}$ is the concatenated feature vector.

1) Feature distillation Block: The main function of feature distillation block (see Fig. 1) is channel splitting. It divides the input feature into two parts. One part is retained that will be concatenated later and the other part is fed into next distillation step. These steps can be depicted mathematically in the equation given below.

$$\begin{aligned} F_{distilled_1}, F_{coarse_1} &= DL_1(F_in), RL_1(F_in), \\ F_{distilled_2}, F_{coarse_2} &= DL_2(F_{coarse_1}), RL_2(F_{coarse_1}), \\ &\ldots\ldots \\ where, DL_1(F_in), RL_1(F_in) &= Split(L_1(F_in)), \\ DL_2(F_{coarse_1}), RL_2(F_{coarse_1)} &= Split_2(L_2(F_{coarse_1})) \\ &\ldots\ldots \end{aligned} \tag{4}$$

The channel splitting equation is decoupled into two deconvolution layers RL and DL. $F_{distilled}$ is passed directly

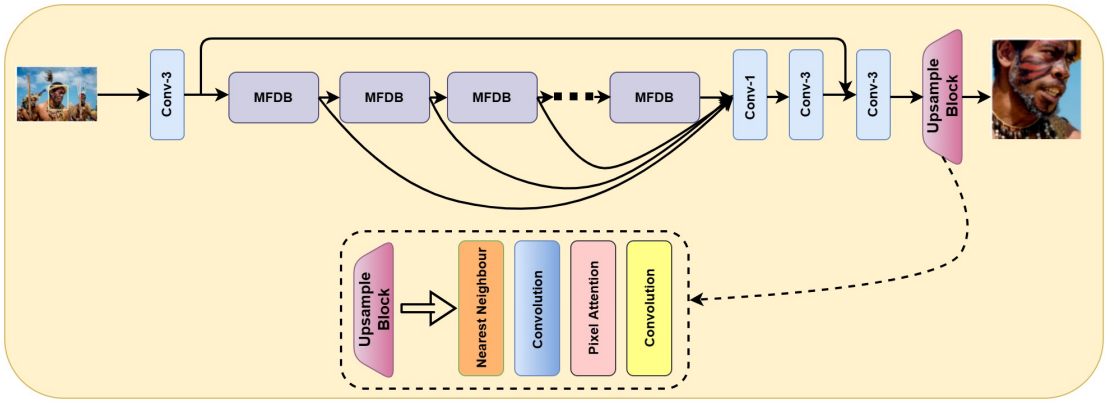

Fig. 1. The architecture of Multiception Feature Distillation Network (MFDN).

to concatenation block while F_{coarse} is passed to the next convolution block for further refinement. This block also uses the concept of shallow residual learning. For this purpose, a shallow residual block consists of a 3 x 3 convolution, an identity connection and a activation unit. The entire concept of RFDB ([12]) block is summarised in the Fig. 2.

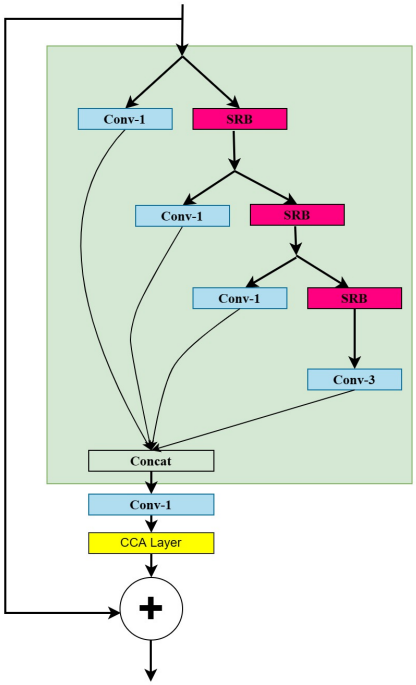

Fig. 2. Structure of RFDB block

2) Multiception layers: As described by [29], Multiception convolution significantly reduces the number of parameters of standard convolution-based models by 32.48% on average while still preserving accuracy.The idea of depthwise separable convolution originated from the group convolution used in AlexNet [30]. Depth-wise separable deconvolution decouples the standard convolution layer into two operations- the spatial convolution and point-wise convolution. The spatial convolution performs operation on individual channel and then point-wise convolution uses 1 x 1 kernel for cross channel convolution. It has fewer parameters and computation cost. The parameters in a standard convolution layer is given by:-

$$w_i = C * K^2 * N \tag{5}$$

where C and N are number of channels in input and output vectors and K is the kernel size. The modified parameters in a depthwise separable convolution is given by:-

$$w_i = C * (K^2 + N) \tag{6}$$

Hence in a depthwise convolution, a single convolution is applied to a single channel at a time unlike standard CNN's in which it is done for all the channels. However the accuracy and performance of model is decreased by using these convolution layers. Multiception convolution layers overcome this challenge of decreased accuracy without increasing parameters and computation cost by much amount.

Multiception depthwise convolution simultaneously uses multiple kernels (see Fig. 3) of varying sizes (eg:- 1 x 1, 3 x 3, 5 x 5, 7 x 7) in spatial convolution part and then concatenates all these intermediate inputs before passing to the point-wise convolution. It enables the network to learn multi-representations of input channels instead of just its single representation. The smaller filters are more sensitive to local information while the larger filters provide large perception on input feature vector. They can easily replace the standard convolution layers in the neural network. The number of parameters in a multiception layer using 3 kernels is given by the equation:-

$$w_i = 3C * (\sum_{j=1}^{3} K_j^2 + N) \tag{7}$$

where K_j is the size of kernel, C and N are number of input and output channels respectively.

The change is brought in RFDB block where 3 x 3 convolution of SRB sub-block is replaced by Multiception depthwise (see Fig. 4) layer that has kernels of size (1 x 1, 3 x 3 , 5 x 5).

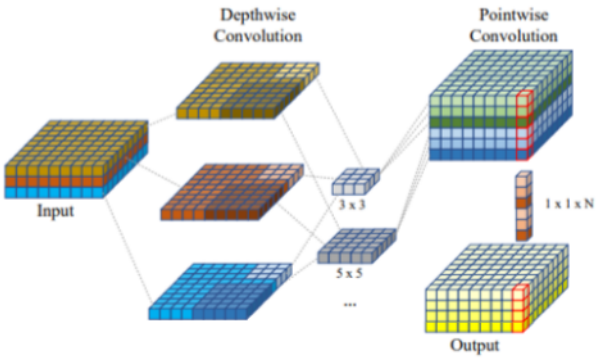

Fig. 3. Multiception depthwise convolution

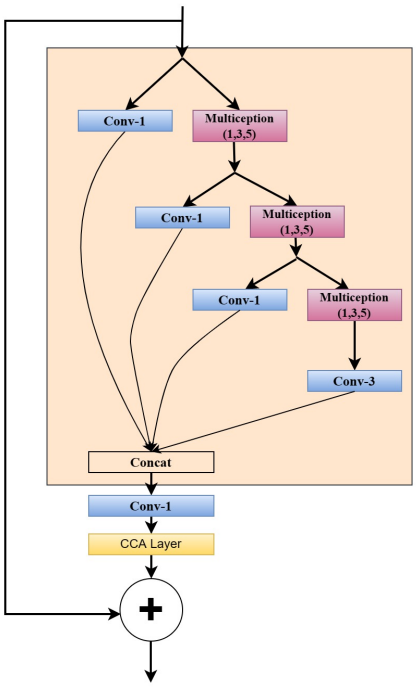

Fig. 4. The Multiception Feature Distillation Block (MFDB)

3) Upsampling block: The model replaces the standard Pixel-Shuffle reconstruction block in the network by the Pixel attention block [13]. The general idea in reconstruction block uses upsampling and convolution layers (refer Fig. 1). RFDN uses PixelShuffle upsampling layer but this is computationally expensive. Our upsampling block uses Nearest neighbour interpolation in the upsampling procedure instead of PixelShuffle. We introduce Pixel attention block after every upsampling layer. Nearest neighbour interpolation layer reduces the previous parameters of upsampling layers by nearly half.

The pixel attention layers generates a 3D matrix of attention features (C x H x W) where C is the number of channels, H and W are the dimensions of the input vector. It is different from channel attention (generated 1D attention vector- C x 1 x 1) and spatial attention (generates 2D attention vector- 1 x H x W).

We conducted experiment to study the change in PSNR with the above described new upsampling block. In this experiment ,we trained two models which are entirely same in structure except for the last upsampling block. We tested our models on random images sampled from Flickr2K dataset that were not used while training. The results are described in the given table (I)

TABLE I
STUDY THE EFFECT OF NEW UPSAMPLING BLOCK ON PSNR

Component	PSNR
PixelShuffle	27.46
Nearest neighbour+Pixel attention	27.75

IV. EXPERIMENTS AND RESULTS

A. Training Strategy

In this model, patches of size 256 x 256 are randomly cropped from the LR images as input for each training mini-batch of size 64. Training data is augmented with random horizontal flips and 90 rotations. Model is trained with ADAM optimiser with initial learning rate set to 5 x10^{-4}. The learning rate is halved after every 200 iterations. The model is trained for 1000 epochs. Multiception Feature Distillation network (MFDN) uses a channel number of 52 to achieve a better reconstruction quality. The networks are implemented by using PyTorch framework on NVIDIA RTX 3090 GPU . DIV-2K [31] dataset is used as training data. Flickr-2K [32] is used as validation data. Smooth L1 loss is used for training. It is a combination of L1 loss and L2 loss. It perfectly avoids the flaws of L1 and l2 loss. Smooth L1 Loss is less sensitive to outliers than L2 Loss, or more robust. The magnitude of the gradient can be controlled.The loss function of our network is given by the following equation:-

$$\begin{cases} L(\theta) = 0.5 * ||H_{network}(I^{LR}) - I^{HR}||^2/beta & if x \leq beta \\ L(\theta) = ||H_{network}(I^{LR}) - I^{HR}|| - 0.5 * beta & otherwise \end{cases} \quad (8)$$

where $x = ||H_{network}(I^{LR}) - I^{HR}||$, I^{HR} is high resolution image. I^{LR} is low resolution image, H is the mapping function from input to output. and beta is set parameter (we set $beta = 0.5$).

B. Model Complexity Analysis

The complexity of model is analysed in terms of parameters, floating point operations on input. The comparison between IMDN, RFDN and our model is presented in the table (II) below.

TABLE II
COMPARISON BETWEEN IMDN, RFDN AND OUR APPROACH ON PARAMETERS, FLOPS

	IMDN	RFDN	MFDN
Parameters	0.894 M	0.433 M	0.312 M
FLOPs	58.53	27.1	37.6

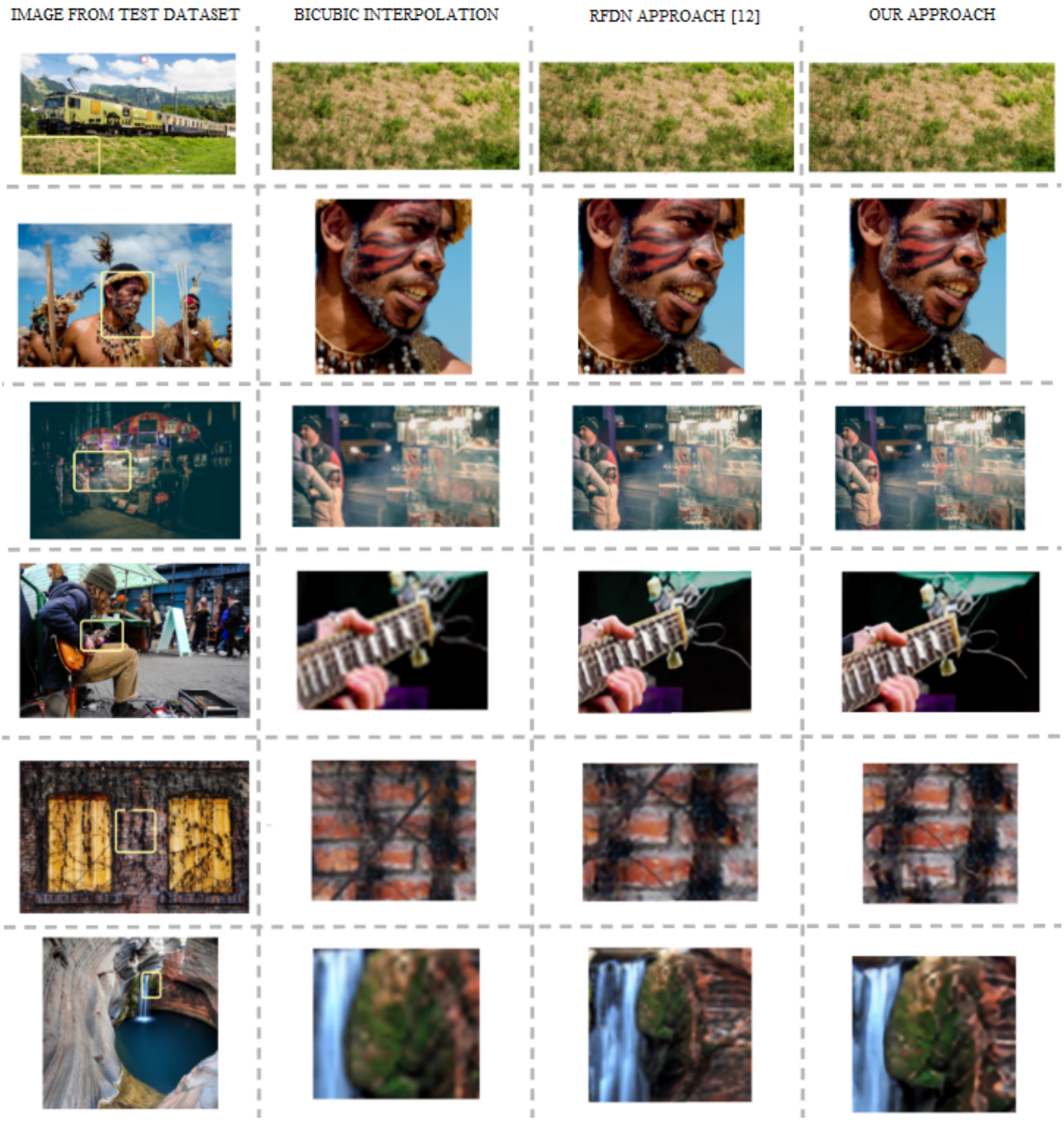

Fig. 5. Visual comparison of super resolution methods on test dataset of AIM 2020 challenge. Column 1 contains low resolution image from test dataset. Column 2, column 3 and column 4 represent super resolved results using bicubic interpolation, RFDN [12] and MFDN respectively on a cropped regions of LR image(shown by yellow rectangle) with up-scaling factor of 4.

C. Results

We trained both the models RFDN and MFDN with the same training details described above. The models were tested on validation and test images provided by the NTIRE [33], AIM 2022 Efficient super-resolution challenge. We present comparison of PSNR of both models on validation and test data. Results are summarised as follows (ref Table III):- We observed that MFDN had increase of PSNR by 0.02 despite of decrease in number of trainable parameters in the model as compared to the RFDN model. This observation shows that Multiception depthwise layers and reconstruction block using Pixel attention reduces the model parameters without changing the performance of system.We demonstrate the qualitative

TABLE III
PSNR SCORES ON VALIDATION AND TEST DATA OF AIM 2022 EFFICIENT SUPER RESOLUTION CHALLENGE, 2022

	Validation PSNR	Test PSNR
RFDN	28.36	28.14
MFDN	28.38	28.16

results (refer Fig. 5) of image super resolution in the given table. These results also compare the SR results from RFDN with our approach.

V. Conclusion

In this paper, we presented MFDN and gave a comprehensive analysis of feature distillation mechanism. We worked on reducing the parameters of our architecture without compromising its performance. We introduce two major changes in the network. First is the use of Multiception depthwise layers in the distillation block that significantly reduce the parameters in the network. Second is use of pixel attention and nearest neighbour interpolation in the reconstruction module of the network that improved PSNR results on the model. We train the network with Smooth L1 loss to capture the edge information better. Experiments conducted show that the proposed method reduce the parameters without sacrificing model performance both quantitatively and qualitatively. The reduced parameters in the model make it portable to mobile devices useful for practical applications.

References

[1] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, and M. Nixon, "Super-resolution for biometrics: A comprehensive survey," *Pattern Recognition*, vol. 78, pp. 23–42, 2018.

[2] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[3] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 206–221.

[4] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[5] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, pp. 1–11, 2011.

[6] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[7] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[8] Z. Xiong, X. Sun, and F. Wu, "Robust web image/video super-resolution," *IEEE transactions on image processing*, vol. 19, no. 8, pp. 2017–2028, 2010.

[9] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[11] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[12] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 41–55.

[13] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *European Conference on Computer Vision*. Springer, 2020, pp. 56–72.

[14] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[15] X. Li, Y. Wu, W. Zhang, R. Wang, and F. Hou, "Deep learning methods in real-time image super-resolution: a survey," *Journal of Real-Time Image Processing*, vol. 17, no. 6, pp. 1885–1909, 2020.

[16] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European conference on computer vision*. Springer, 2016, pp. 391–407.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[18] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition. 2015," *arXiv preprint arXiv:1409.1556*, 2015.

[20] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 252–268.

[23] Y. Fan, H. Shi, J. Yu, D. Liu, W. Han, H. Yu, Z. Wang, X. Wang, and T. S. Huang, "Balanced two-stage residual networks for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 161–168.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[25] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4799–4807.

[26] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[28] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[29] G. Bao, M. B. Graeber, and X. Wang, "Depthwise multiception convolution for reducing network parameters without sacrificing accuracy," in *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2020, pp. 747–752.

[30] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[31] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[32] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *IJCV*, vol. 123, no. 1, pp. 74–93, 2017.

[33] Y. Li, K. Zhang, L. V. Gool, R. Timofte *et al.*, "Ntire 2022 challenge on efficient super-resolution: Methods and results," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022.