More Direct and stage-wise network for Face Super Resolution

Yohei Horiguchi*, Masaaki Ikehara* and Kei Shibasaki*
* University of Keio, Kanagawa, Japan
E-mail: horiguchi0722@keio.jp   Tel: +81-45-566-1530

Abstract—**In recent years, U-net-based models have demonstrated high performance in high-magnification face super resolution (FSR) tasks and are capable of outputting more vivid super resolution (SR) images. However, this model has some drawbacks, such as the possibility of unnecessary computations and uniform block selection at all stages of the decoder. Therefore, we propose a Direct and Stage-wise (DS) Net, which improves on CTCNet [1], a U-net based model with high qualitative results. This model outperforms previous networks by eliminating encoders to reduce computational waste, and by focusing on global feature extraction in the small tensor size stage.**

I. Introduction

Super-resolution (SR) is a traditional task in image processing that aims to reconstruct high resolution (HR) images from low resolution (LR) images. This task is an ill-posed problem since most LR images are affected by multiple degrading factors including aliasing, motion blur and out-of-focus and there are numerous possible SR results based on which degrading factors are dominant in each picture. To solve this task many machine learning-based methods are proposed.

Currently, various networks including RCAN [2] and SwinIR [3] have been proposed for low-magnification SR tasks such as 2×, and they can already output clear SR results that are not much different from ground truth (GT) images. However, research on SR tasks at high magnifications, such as 8×, has yet to output images like those of GT, and there is room for further development.

Face super resolution (FSR) is the major task in such high-magnification SR tasks. The reason why face images are often used in high-magnification SR is because face images have few high-frequency components. The skin area that occupies the majority of the face image can be represented with a certain degree of color, so there are fewer high-frequency components than in other types of images, such as building images. Since it is difficult to retrieve high-frequency components from LR images in high-magnification SR, face images with relatively few high-frequency components are considered suitable for this purpose. In addition, face image restoration can be utilized not only for entertainment applications but also for many tasks including verification, and analysis making it a task with high social demands.

In the field of FSR, many methods have been proposed. Recently, the mainstream approach involves up-sampling the LR image to HR size using bicubic interpolation and then passing it through a U-Net structured network, similar to image deblurring, to output the SR image, as seen in methods like SPARNet [4] and CTCNet [1].

However, there are two problems with these methods: First, the process of up-sampling the image size once and then down-sampling it again at the encoder part of the U-net structure is a two-step process, which increases the computational cost and may prevents meaningful feature extraction. The second problem is that in the decoder section, the same feature extraction block is used for all stages and feature extraction appropriate for each size stage is not performed.

To solve the above problems, we introduce Direct and Stage-wise (DS) Net which based on CTCNet [1], which is a Unet-based network that currently boasts high performance. This DSNet has 3 key aspects. First, we omitted the encoder part of the U-net and directly input LR images from the bottleneck part to reduce the computational cost. Second, we changed the Feature Refinement Module (FRM), the feature extraction block in the bottleneck part, to a Global-wise Feature Refinement Module (GFRM) that can extract more global information. Finally, the third improvement was achieved by removing the convolution neural networks (CNN) based Facial Structure Attention Unit (FSAU), which is good at extracting local information, from the Local-Global Feature Cooperation Module (LGCM) in the bottom layer of the decoder section, and replacing it with a transformer block only, thereby adjusting the number of blocks and improving accuracy. With these improvements, we were able to demonstrate superiority over CTCNet [1] and other major SR methods without a significant increase in computational load.

II. Related works

Since Baker and Kanade [5] proposed the concept of FSR, many methods were proposed for FSR tasks. Some FSR

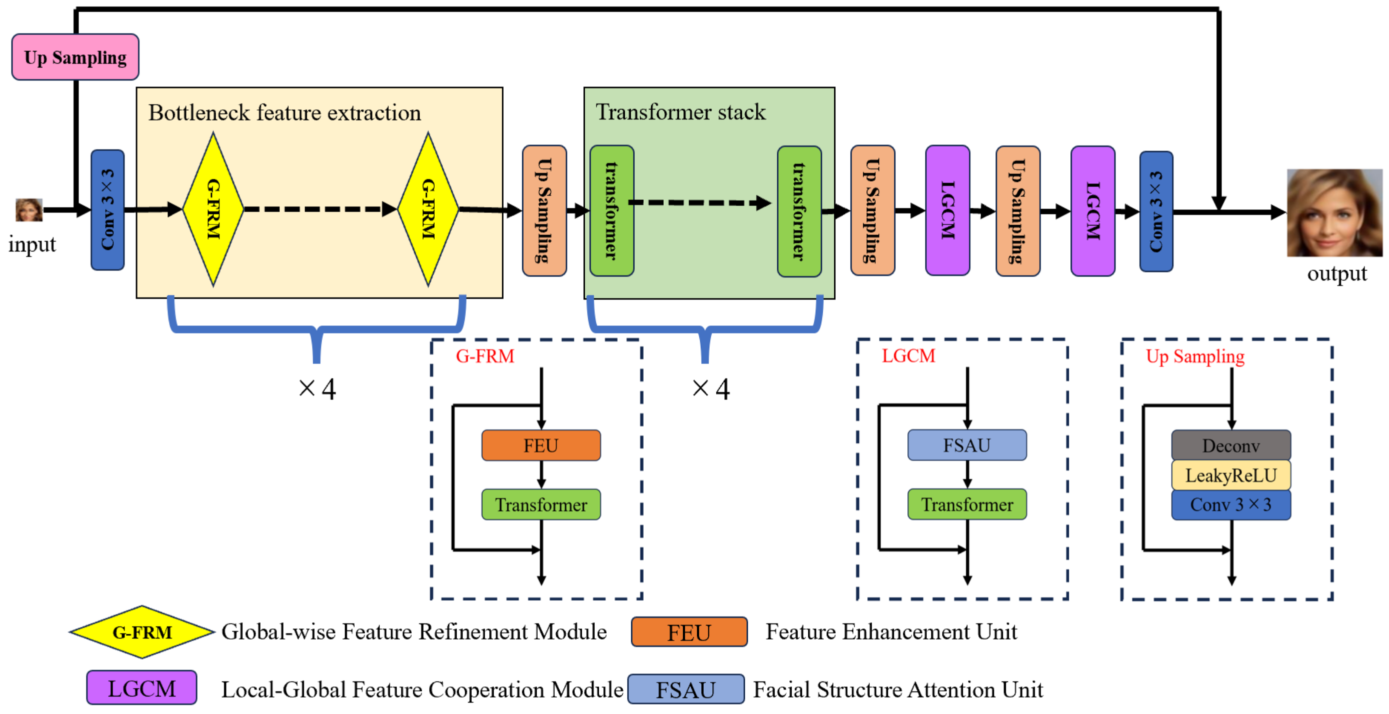

Fig. 1 The complete structure of the proposed Direct and Stage-wise (DS) network

methods utilize prior-guided training. These methods succeed in utilizing multiple information such as facial landmarks and action units to reconstruct SR images but it requires an additional labeled dataset [6]. Recently, the diffusion model [7] has contributed to the improvement of SR results however these models often require high computation capacity and therefore, implementation in end devices is still a long way off. CTCNet [1], which we will use as our base network model, can be trained on a dataset of HR-LR pairs and achieves SR with relatively low computational cost with high accuracy.

CTCNet [1] is composed of three parts: an encoder part, a bottleneck part, and a decoder part. Each part performs feature extraction using CNN-based blocks (FSAU and Feature Enhancement Unit (FEU)) and transformer blocks to enable both global and local feature extraction. The contents of each block will be lightly touched upon in the proposed method, but for details, please refer to this paper on CTCNet [1]. However, there are two problems with this network: The first is the wastefulness of the encoder part. In CTCNet [1], the LR image is up-sampled by bicubic sampling and then down-sampled while extracting features in the encoder part, but this requires a large computational load because of the feature extraction of the large-size image. In addition, it is questionable to what extent applying the high-dimensional information extraction characteristic of the U-net structure will contribute to SR results.

The second is that the same feature extraction block is used in each stage of the decoder section. The LGCM feature extraction block in the decoder section consists of a CNN-based FSAU block suitable for local information extraction and a transformer block suitable for global information extraction [1]. However, since the size of the tensor differs at each stage, it will be essential to adjust the size of the tensor at each stage to determine whether the emphasis should be placed on global or local information. Specifically, since information on fine details of a face image cannot be obtained when the tensor is small, we believe that the focus should be on global information extraction for the entire face and that the weight of local information extraction should be increased at the stage where a large tensor is processed.

III. PROPOSED METHOD

A. Overall architecture

DSNet has the structure shown in the Fig.1. In DSNet, we input LR images directly to the network and repeat feature extraction and up-sampling until we get images that are equivalent to the size of HR images. By this direct and simple network structure, we could drastically reduce unnecessary computation cost while preserving SR performance.

To better demonstrate the model, we define I_{LR}, I_{SR}, and I_{HR} as the LR input image, the recovered SR image, and the ground-truth HR image, respectively. Also, the size of the LR image is set to 16×16.

When an LR image is input to the network, shallow features are first extracted by 3 × 3 convolution, and then both global and local features are extracted using a G-FRM block in the Bottleneck feature extraction section. Next, the up-sampling block consisting of a 6 × 6 transposed convolutional layer with stride 2, a LeakyReLU activation function, and a 3 × 3 convolution with stride 1 is used for ×2 up-sampling to obtain 32 × 32 size feature map. After up-sampling to 32 × 32 size, features are extracted with emphasis on global information through the Transformer stack.

After that, the process of 2× up-sampling by the up-sampling block and feature extraction by the LGCM block is repeated twice, and finally, I_{out} is made using 3×3 convolution. Finally, the output image is given by $I_{SR} = I_{out} + I_{LR}^{8\uparrow}$. Meanwhile, $I_{LR}^{8\uparrow}$ is made by conducting ×8 bicubic up-sampling of I_{LR}. Given a training dataset $\{I_{LR}^{i}, I_{HR}^{i}\}_{i=1}^{N}$, DSNet utilize pixel-level loss function:

$$\mathcal{L}(\Theta) = 1/N \, \Sigma_{i=1}^{N} \left\| F_{DSNet}(I_{LR}^{i}, \Theta) - I_{HR}^{i} \right\|, \qquad (1)$$

Meanwhile, N, $F_{DSNet}(\cdot)$ and Θ denote the number of training datasets, DSNet and parameter set of DSNet, respectively.

B. *Bottleneck feature extraction*

In the bottleneck part that processes the 16×16 tensor, CTCNet [1] use the FRM block which consists of two feature extraction blocks, FSAU and FEU. Since both of these feature extraction blocks are CNN based ones, they are suited for local feature extraction. Especially, the FEU uses a double-branch structure to extract features from the original scale tensor and the down-sampled tensor, and finally fuse them to reduce the computational cost [1].

However, since feature extraction by FRM is performed on extremely small tensors in the bottleneck part, it is more important to extract global information than to extract local information.

Therefore, in DSNet we deleted FSAU blocks from FRM and inputted transformer blocks instead, naming it G-FRM block. The reason why we put transformer block instead of CNN based block is because transformer blocks can learn the relationships between elements of a sequence and thus suited for extracting global information [8]. In DSNet, we use transformer block that consist of Multi-Dconv head Transposed Attention (MDTA) and Gated-Dconv Feed-forward Network (GDFN) from [9]. MTDA is designed to address the computational challenges of conventional self-attention mechanisms. Instead of applying self-attention across the spatial dimensions, it operates across the channel dimensions, thus efficiently capturing global context with reduced computational cost. Additionally, it uses depth-wise convolutions to focus on local context before generating the overall attention map, ensuring a balance between local and global information processing. The GDFN modifies the traditional feed-forward network in two key ways. Firstly, it uses a gating mechanism, which involves two separate paths of linear transformations that are combined element-wise, with one path being activated by the GELU function. Secondly, it incorporates depth-wise convolutions to capture information from adjacent pixels, enhancing the network's ability to learn local image structures crucial for effective image restoration.

Thus, by combining the FEU, which can extract local information at low computational cost, and the transformer block, which can extract global information, it is possible to perform feature extraction suited to small tensors.

C. *Transformer stack*

In the bottom stage of encoding part, CTCNet use LGCM that use both CNN based FSAU block and transformer block in order to extract global and local features [1]. Similar to the changes made to the FRM in Bottleneck feature extraction, we decided to eliminate the FSAU and keep only the transformer, since this part also targets small 32 x 32 feature maps and we thought it would be better to put more emphasis on global information extraction. However, we thought that simply using a single transformer block would sacrifice pure feature extraction capability, so we decided to compensate for the lack of FSAU by connecting the transformers in series. After the experiment explained in ablation study, we identified that 4 that four transformers in series was optimal, and we named it the transformer stack.

IV. Experiment

A. *Dataset*

In this paper, we utilize 2 different datasets as training datasets in order to examine the efficiency of our method. The first dataset we use is the full dataset which consists of 18,000 samples of the CelebA [10] dataset for training, 200 samples for validating. This dataset is the same as the dataset which was used in the CTCNet paper [1], and used for comparing our method and other FSR methods. The second dataset we use is a small dataset which consists of 3,000 samples of the CelebA dataset for training, and 100 samples for validating. This small dataset is used for ablation study and since this small dataset required a shorter time to train, we used this dataset to mitigate the time we need for evaluation. Although the overall accuracy of the networks will be reduced due to the small number of samples that can be referenced in training, we believe that the relative advantage between networks can be confirmed even using a small dataset.

For the test data set, 1000 images from CelebA were used. In all data sets, the CelebA images were cropped to 128×128 size as HR, and LR images are made by conducting 8× bicubic down-sampling.

Table 1. Quantitative comparisons for ×8 SR on the CelebA test sets [1]. Best and second-best results are **bolded** and <u>underlined</u>.

Method	PSNR ↑	SSIM ↑
Bicubic	23.61	0.6779
SAN [12]	27.43	0.7826
RCAN [2]	27.45	0.7824
HAN [13]	27.47	0.7838
SwinIR [3]	27.88	0.7967
FSRNet [14]	27.05	0.7714
FACN [15]	27.22	0.7802
SPARNet [4]	27.73	0.7949
SISN [16]	27.91	0.7971
CTCNet [1]	<u>28.37</u>	<u>0.8115</u>
DSNet (ours)	**29.15**	**0.833**

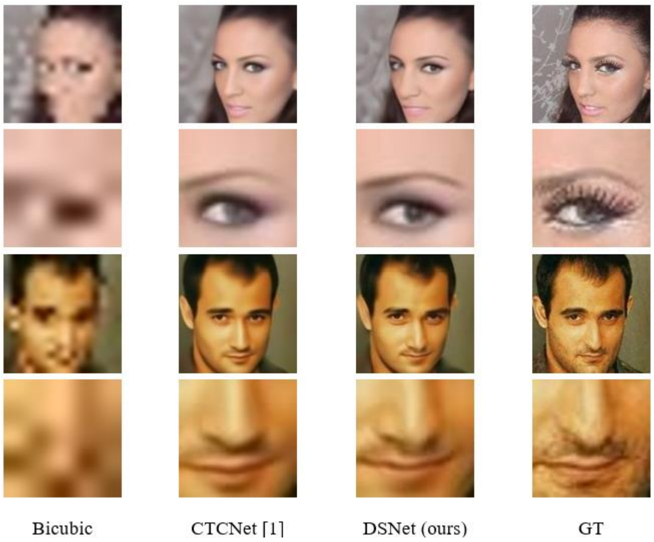

Fig. 2 Visual comparisons for ×8 SR on the CelebA test set.

B. Implementation details

Other training conditions are set as in CTCNet [1]. That is, implementation of our model is done by PyTorch framework, the optimizer is Adam and set β1 = 0.9 and β2 = 0.99, the initial learning rate is set to 2 × 10 -4, the batch size is 10 and the number of epochs is 100. Also, PSNR and SSIM [11] are used for qualitative comparisons.

C. Comparison with other methods

We compare DSNet with 9 major state-of-the-art methods. These methods include general SR methods SAN [12], RCAN [2], HAN [13], SwinIR [3], major FSR methods FSRNet [14], FACN [15], SPARNet [4], SISN [16], and CTCNet [1]. Since DSNet was experimented under the same conditions as CTCNet and When CTCNet was trained and tested under these conditions, the same results as reported in the CTCNet paper were obtained, for the qualitative evaluation values of the state-of-the-art methods, we referred to the result shown on CTCNet paper [1].

Table 2. Quantitative result of omitting encoder section. Better results are **bolded**.

Encoder	PSNR ↑	Params (M) ↓
✓	25.87	24.40
×	**26.04**	**17.67**

Table 3. Quantitative evaluation of contents in FRM block. Best and second-best results are **bolded** and <u>underlined</u>. Bottom contents are used in DSNet

FEU	FSAU	transformer	PSNR ↑	Params (M) ↓
✓	✓	×	26.17	19.87
×	×	✓	22.54	**11.43**
×	✓	✓	26.18	<u>16.63</u>
✓	✓	✓	<u>26.26</u>	22.62
✓	×	✓	**26.37**	19.34

1) Quantitative comparison:

When compared with other SOTA methods, we could see from Tab. 1 that results on DSNet outperforms other method in both PSNR and SSIM evaluation.

2) Qualitative comparison

In Fig.2 when we compare CTCNet [1] and DSNet, we could see, that DSNet can reconstruct SR images that are closer to GT images. For example, when comparing the results of the second male image, it can be seen that DSNet outputs results that are closer to GT in terms of nose shape, etc.

D. Ablation study

In this section, we conducted three ablation studies using the small dataset and examined the efficiency of omission of the encoder section, GFRM in Bottleneck feature extraction and transformer stack, respectively. Since we used small training dataset for Ablation study, quantitative results are decreased compared to results shown in tab. 1 however, we believe that even if the absolute amount of quantitative results has been reduced, the relative superiority relationship between different networks can still be confirmed.

1) Omission of encoder section

In order to assess efficiency of omitting encoder we compared CTCNet with/without encoder stage. Tab. 2 shows that the omission of encoder significantly reduces the number of parameters (reduces the parameters of the model to about 70% of the original) in the model, thus reduces the computational load, and also improves PSNR result.

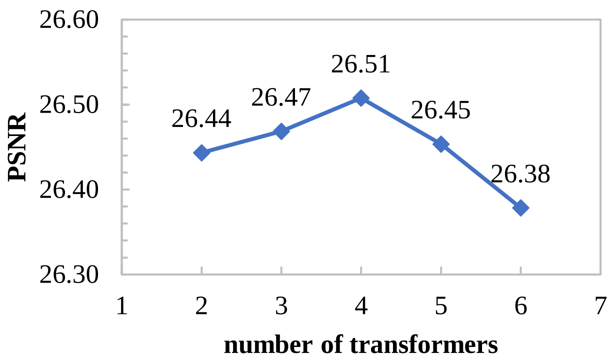

Fig. 3 Relationship between the number of transformers and PSNR in transformer stack.

2) GFRM in Bottleneck feature extraction

In order to verify the effectiveness of GFRM, we compared the qualitative results when a combination of FEU, FSAU, and transformer were introduced into the FRM for the feature extraction block at the bottleneck area. As shown in Tab. 3, it was found that the eliminating FSAU from the FRM and inserting transformer instead produced the best PSNR, and that the number of parameters could be kept at a reasonable value.

3) Number of transformer blocks in transformer stack

To determine the optimal number of transformers at the Transformer stack, we examined the PSNR values when varying the number of transformers. Fig. 4 shows that the best results were obtained when four transformers were used.

V. Conclusions

In this paper, we proposed DSNet, which is a significant improvement over CTCNet. In our study, we found that eliminating encoders in the U-net based SR models does not reduce the accuracy of SR and significantly reduces the computational cost. We also found that instead of using the same feature extraction block in all stages, it is better to use a global feature extraction-intensive network for feature extraction on small tensors. With these two modifications, DSNet was able to produce better quantitative results than other FSR methods.

There are two future prospects. Although the proposed DSNet is able to produce SR results that are somewhat closer to GT than existing methods, it still lacks the ability to reproduce the finest details. Therefore, we believe that a network that can extract more local information and restore details is required.

The second is the addition of more novel blocks. Although this study has succeeded in improving SR performance by making major modifications to CTCNet, the blocks themselves are not different from those used in CTCNet, so there is room to further improve SR performance by adding more novel blocks in the future.

References

[1] Gao, Guangwei, et al. "CTCNet: A CNN-transformer cooperation network for face image super-resolution." IEEE Transactions on Image Processing 32 (2023): 1978-1991.

[2] Zhang, Yulun, et al. "Image super-resolution using very deep residual channel attention networks." Proceedings of the European conference on computer vision (ECCV). 2018.

[3] Liang, Jingyun, et al. "Swinir: Image restoration using swin transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021

[4] Chen, Chaofeng, et al. "Learning spatial attention for face super-resolution." IEEE Transactions on Image Processing 30 (2020): 1219-1231.

[5] Baker, Simon, and Takeo Kanade. "Hallucinating faces." Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580). IEEE, 2000.

[6] Zhang, Chenggong, and Zhilei Liu. "Face super-resolution with progressive embedding of multi-scale face priors." 2022 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2022.

[7] Gao, Sicheng, et al. "Implicit diffusion models for continuous super-resolution." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[8] Khan, Salman, et al. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s (2022): 1-41.

[9] Zamir, Syed Waqas, et al. "Restormer: Efficient transformer for high-resolution image restoration." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[10] Liu, Ziwei, et al. "Deep learning face attributes in the wild." Proceedings of the IEEE international conference on computer vision. 2015.

[11] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing 13.4 (2004): 600-612.

[12] Dai, Tao, et al. "Second-order attention network for single image super-resolution." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[13] Niu, Ben, et al. "Single image super-resolution via a holistic attention network." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020.

[14] Chen, Yu, et al. "Fsrnet: End-to-end learning face super-resolution with facial priors." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[15] Xin, Jingwei, et al. "Facial attribute capsules for noise face super resolution." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

[16] Lu, Tao, et al. "Face hallucination via split-attention in split-attention network." Proceedings of the 29th ACM international conference on multimedia. 2021.