Progress In Electromagnetics Research, Vol. 171, 185–199, 2021

Deep Neural Networks for Image Super-Resolution in Optical Microscopy by Using Modified Hybrid Task Cascade U-Net

Dawei Gong[1, 2], Tengfei Ma[1, 3], Julian Evans[1], and Sailing He[1, 2, 3, *]

Abstract—Due to the optical diffraction limit, the resolution of a wide-field (WF) microscope cannot easily go below a few hundred nanometers. Super-resolution microscopy has the disadvantages of high cost, complex optical equipment, and high experimental environment requirements. Deep-learning-based super-resolution (DLSR) has the advantages of simple operation and low cost, and has attracted much attention recently. Here we propose a novel DLSR model named Modified Hybrid Task Cascade U-Net (MHTCUN) for image super-resolution in optical microscopy using the public biological image dataset BioSR. The MHTCUN has three stages, and we introduce a novel module named Feature Refinement Module (FRM) to extract deeper features in each stage. In each FRM, a U-Net is introduced to refine the features, and the Fourier Channel Attention Block (FCAB) is introduced in the U-Net to learn the high-level representation of the high-frequency information of different feature maps. Compared with six state-of-the-art DLSR models used for single-image super-resolution (SISR), our MHTCUN achieves the highest signal-to-noise ratio (PSNR) of 26.87 and structural similarity (SSIM) of 0.746, demonstrating that our MHTCUN has achieved the state-of-the-art in DLSR. Compared with the DLSR model DFCAN used for image super-resolution in optical microscopy specially, MHTCUN has a significant improvement in PSNR and a slight improvement in SSIM on BioSR. Finally, we fine-tune the trained MHTCUN on the other biological images. MHTCUN also shows good performance on denoising, contrast enhancement, and resolution enhancement.

1. INTRODUCTION

With the rapid development of computer technology and artificial intelligence, deep neural networks have a wide range of applications in various fields, such as object detection [1], semantic segmentation [2], instance segmentation [3], single-image super-resolution (SISR) [4–6], etc. SISR refers to obtaining a high-resolution (HR) image from a low-resolution (LR) image through a non-linear mapping. There are two main types of SISR methods. One is super-resolution based on traditional algorithms such as bilinear interpolation [7], and the other is deep-learning-based super-resolution (DLSR) [8].

The SRCNN [9] proposed by Dong et al. is the first SISR model used for SISR which surpassed traditional upsampling algorithms such as bilinear interpolation. SRCNN is composed of three convolutional layers. The SRCNN structure is very simple; however, the performance can be improved with higher complexity. More and more researchers use deep convolutional neural networks to approach the SISR problem. The main problems encountered by deep convolutional neural networks are gradient disappearance and gradient explosion. The traditional solutions are data initialization (e.g., normalized initialization) and regularization (e.g., batch normalization), which solves the gradient problem and deepens the depth. However, this can degrade network performance since greater depth increases

Received 9 November 2021, Accepted 14 December 2021, Scheduled 16 December 2021
* Corresponding author: Sailing He (sailing@zju.edu.cn).
[1] National Engineering Research Center for Optical Instruments, Centre for Optical and Electromagnetic Research, Zhejiang University, Hangzhou 310058, China. [2] Shanghai Institute for Advanced Study of Zhejiang University, Shanghai 200000, China. [3] Ningbo Research Institute, Zhejiang University, Ningbo 315100, China.

the error rate. He et al. proposes ResNet [10] to solve the degradation problem and the gradient problem simultaneously. The Enhanced Deep Super-resolution network (EDSR) [11] proposed by Lim et al. removes some unnecessary modules in the residual structure, such as the BN layer. Lim et al. also proposes a multi-scale model in which most of the parameters are shared at different scales. Such a model can have a good effect in dealing with each single-scale super-resolution. EDSR is constructed by simply stacking residual blocks and uses ReLU for activation. When a very large gradient flows through a ReLU neuron, this neuron will no longer be active. If the learning rate is very large, then it is very likely that 40% of the neurons in the network are "dead". Residual Channel Attention Network (RCAN) [12] proposed by Zhang et al. is to stack a very deep network using the residual in residual (RIR) structure. The RIR consists of several residual groups with long skip connections and collects rich low frequency information. Each residual group contains some residual blocks with short skip connections. Bypassing multiple skip connections, the main network is formed to focus on learning high-frequency information. In addition, Zhang et al. propose a residual channel attention block (RCAB) to adaptively adjust channel characteristics by considering the interdependence between channels. RCAN has more than 800 layers and huge parameters, which makes training very slow.

Unlike ResNet, Dense block [13] proposed by Huang et al. does not connect layers through summation, but connects features through concatenating. Residual Dense Network (RDN) [14] proposed by Zhang et al. makes improvements to the traditional Dense block in three aspects. They are contiguous memory (CM) mechanism, Local feature fusion (LFF) and Local residual learning (LRL). However, when simulating LR images, it uses three degradation models. The first is to use Bicubic Downsampling (BD) to process the HR image; the second is to use Gaussian Blur to blur the HR image; and the third is to downsample the HR image by BD first, and then adding Gaussian noise. For the Gaussian blur processing method, the sampling method, the shape and size of the Gaussian kernel and the trimming problem in the Gaussian blur calculation result in a large deviation between the obtained LR and the ground truth. Super-resolution Feedback Network (SRFBN) [15] proposed by Li et al. includes a feedback module to return super-resolution results many times. The advantage of such a return is that no additional parameters are added, and multiple returns are equivalent to deepening the network and continuously refining the generated super-resolution image. The output of the last feedback is re-input to the feedback module along with the input of the entire network, and is continuously returned. However, the curriculum learning (CL) strategy is used for training, which requires sorting the images of the training set according to the degree of difficulty. CL is difficult to implement because measuring difficulty and scheduling the training are complex problems, which require expert human input. Since the decision boundaries of human and machine models are not necessarily the same, the samples that humans consider easy are not necessarily easy for the model.

U-Net [16] proposed by Ronneberger et al. is used for biomedical image segmentation, and it consists of repeated two 3×3 convolutions, followed by a ReLU, and a max pooling for downsampling. The feature channel was halved in each downsampling. The expansion path includes an upsampling, which will halve the feature channel, followed by a feature map of the corresponding contraction path and two 3×3 convolutions. Finally, a 1×1 convolution is used to map the feature vector with 64 elements to a class label. U-Net has a total of 23 convolutional layers and is a fully convolutional neural network where there is no fully connected layer. A new framework Hybrid Task Cascade (HTC) [17] proposed by Chen et al. improves information flow by combining cascading and multitasking at each stage in instance segmentation tasks. HTC is an improvement on Cascade-Mask-RCNN [18]. Chen proposes Interleaved Execution, that is, the box branch is executed first, and the returned box is then handed over to the mask branch to predict the mask in each stage. In Cascade-Mask-RCNN, the mask branch between different stages does not have any direct information flow. Chen adds a connection between the mask branches of adjacent stages to provide the information flow of the mask branches. In addition, Chen introduces a semantic information branch into the framework to obtain a better spatial context. Since semantic information involves fine pixel-level classification of the whole image, it enables strong discrimination of foreground and background. By fusing the semantic information of this branch into the box and mask branches, the performance of these two branches can be greatly improved.

The optical diffraction limit restricts the acquisition of high-frequency information in optical microscopy and has hindered the progress of scientific research. Super-resolution restoration of wide-field (WF) microscopy images has become increasingly important. In recent years, some hardware

methods such as Stimulated Emission Depletion Microscopy (STED) [19], Photoactivation Localization Microscopy (PALM) [20], stochastic optical reconstruction microscopy (STORM) [21], structured illumination microscopy (SIM) [22], and other super-resolution microscopies have been proposed, breaking the diffraction limit and accessing nanoscale information. However, the hardware methods have the disadvantages of cost and high environmental requirements, and DLSR is increasingly used in optical microscopy. A content-aware image restoration network (CARE) [23] proposed by Weigert et al. could push the limits of fluorescence microscopy. CARE could restore images with 60-fold fewer photons used and recover the image with up to tenfold downsampling along the axial direction. Wang et al. propose CMGAN [24] to realize the transformation of WF images into ground truth images. CMGAN is a GAN-based image transformation framework which can be used across different fluorescence microscopy modalities. Qiao et al. use the frequency difference between different features in the Fourier domain to improve the ability of the DLSR model to learn hierarchical representations of high-frequency information, instead of using structural features in the spatial domain. Therefore, the Fourier Channel Attention Block (FCAB) [25] is proposed, and Deep Fourier Channel Attention Network (DFCAN) is developed to make the super-resolution results of microscopic images more refined. At the same time, the first dataset BioSR for image super-resolution in optical microscopy was released for public.

Motivated by DFCAN and HTC, this paper proposes a new DLSR model named Modified Hybrid Task Cascade U-Net (MHTCUN) for image super-resolution in optical microscopy using the public biological image dataset BioSR [25]. Modified Hybrid Task Cascade (MHTC) has three stages, and information flow is introduced by feeding the pixel features of the previous pixel branch to the current pixel branch. A spatial contexts branch is introduced to predict high-level semantic features. To achieve better pixel prediction, we combine the semantic features after spatial contexts branch and pixel features in pixel branches. In each stage of MHTC, The Feature Refinement Module (FRM) is introduced to refine the feature maps. The FRM consists a U-Net, a convolution layer and a channel-wise concatenation. To learn the high-level representation of the high-frequency information of the feature maps, we use FCAB in each U-Net. We use auxiliary loss between outputs of each pixel branch and the ground-truth after blurring for supervised learning. We use the $4\times$ blurred images to supervise the first stage, the $2\times$ blurred images to supervise the second stage, and ground truth images to supervise the last stage. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are introduced as evaluation metrics, and both reach the highest scores of 26.87 and 0.746 using our best trained model. It indicates that MHTCUN has achieved state-of-the-art among many state-of-the-art DLSR models in SISR. Compared with DFCAN, MHTCUN has a significant improvement in PSNR and a slight improvement in SSIM. We also fine-tune MHTCUN on the other biological structures, which all showed good robustness in super-resolution.

The organization of this paper is as follows. In Section 2, we describe the proposed DLSR model MHTCUN in detail. In Section 3, we describe the loss function and evaluation metrics. The experimental results, ablation study, and specific details of the experiment are described in Section 4.

2. PROPOSED DLSR MODEL

Motivated by the superior performance of Hybrid Task Cascade (HTC) [17] in semantic segmentation, we combine Modified Hybrid Task Cascade (MHTC) with U-Net for SISR. Modified Hybrid Task Cascade U-Net (MHTCUN) consists of three stages, and each stage can output the super-resolution results from coarse to fine, as shown in Fig. 1.

In Fig. 1, a set of input training low resolution (LR) images are first upsampled by a factor of two and then flow a 3×3 convolution layer to generate the feature maps x. In MHTC, the feature maps x flow through a U-Net to extract deep features, then the deep features flow into three Feature Refinement Modules (FRMs) to refine the features. U-Net is integrated in each FRM, and x will be added to each FRM. Next, the deep features from the FRM flow into the pixel branches to predict super-resolution results. x flows into the spatial contexts branch and is then added in each pixel branch for better prediction. We use the outputs of stage three as final super-resolution results. Next, we introduce the details of the structure of MHTCUN.

Beginning with the general HTC architecture, we modify the original HTC architecture to form

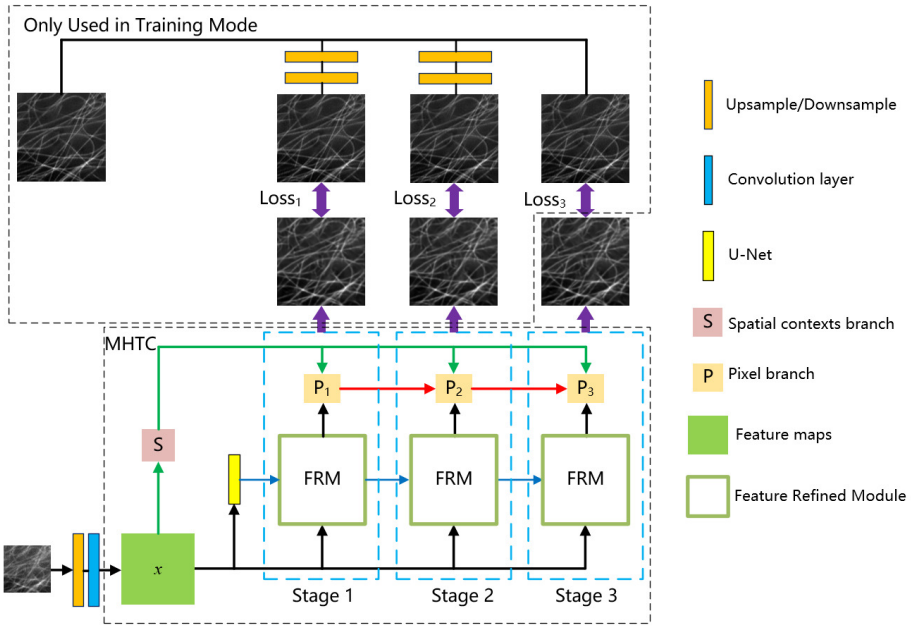

Figure 1. Architecture of the proposed Modified Hybrid Task Cascade U-Net (MHTCUN).

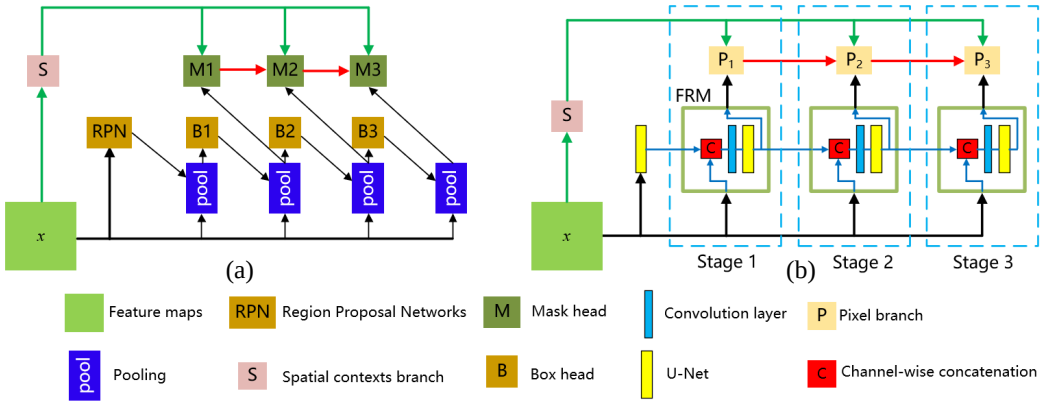

Figure 2. Comparison chart of Hybrid Task Cascade (HTC) and Modified Hybrid Task Cascade (MHTC) used in our MHTCUN. (a) HTC architecture. (b) MHTC architecture.

MHTC as shown in Fig. 2. Specifically, we remove the box head and RPN branch, replace the first pooling module with a U-Net and each other pooling module with an FRM, the mask head with a pixel branch. The FRM consists of three parts: a channel-wise concatenation, a 3×3 convolution layer for reducing dimension, and a U-Net.

A U-Net is integrated in each FRM for refining features, as shown in Fig. 3. In the encoder stage, the Fourier Channel Attention Block (FCAB) is introduced to extract high-level features of high-frequency information of feature maps, which are down-sampled to lower resolution using a convolution layer where the stride is set to 2. In the decoder stage, the lower resolution feature maps are processed by the FCAB. The feature maps are then upsampled using PixelShuffle [26], and a 3×3 convolution layer is embedded to reduce the dimensions of feature maps. Skip connection is added to fuse the features in different levels. Fig. 4 shows the details of FCAB. In FCAB, feature maps first flow to double depthwise separable convolution [27] layers and swish [28] activation. Fast Fourier Transforms (FFT) [25] are introduced to learn the attention of each channel and generate residual feature maps which are added to the input feature maps as the final output of FCAB.

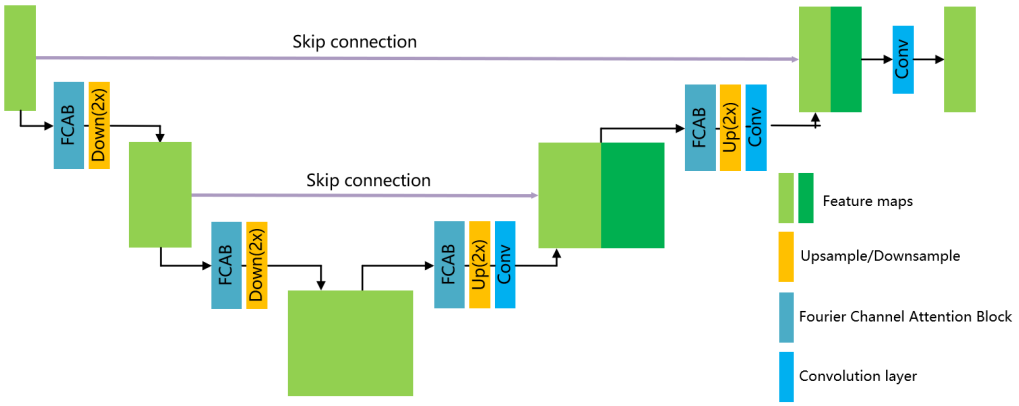

Figure 3. The detailed architecture of the U-Net used in the MHTC.

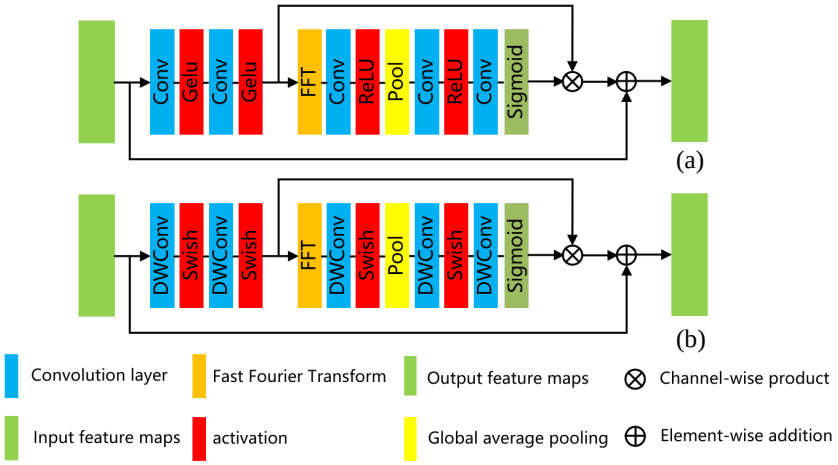

Figure 4. Details of the Fourier Channel Attention Block (FCAB) used in the U-Net shown in Fig. 3. (a) FCAB used in DFCAN [25]. (b) FCAB used in U-Net.

In MHTC (Fig. 2), the feature maps x first flow into a U-Net to extract deeper features. Next, we combine deeper features with x using channel-wise concatenation and use a convolution layer to reduce the dimension of feature maps. Then a U-Net refines the feature maps after the convolution layer. The results of the FRM flow to next stage and are added to the pixel branch for super-resolution prediction. We establish a feature information flow by feeding the pixel features of the previous pixel branch to the current pixel branch, as shown in Fig. 5.

In each pixel branch, the features after the FRM will be combined with pixel features of the previous pixel branch and semantic features from spatial contexts branch using element-wise summation. Then they flow through four depthwise separable convolution layers sequentially. The results after depthwise separable convolution layer are used as the residual of next pixel branch and flow into a reconstruction layer which is composed of a 3×3 convolution layer to generate super-resolution results. Spatial context features are added to each pixel branch for better prediction. The architecture of the spatial contexts branch is modeled on the Feature Pyramid [29], as shown in Fig. 6.

Figure 6 shows the architecture of spatial contexts branch. We use a 3×3 convolution layer where the stride is set to 2 to obtain the feature pyramid. To align in a common representation space, we use a 1×1 convolution layer on each level of the feature pyramid. The high-level feature maps are upsampled with transposed convolution. These transformed feature maps from different levels are subsequently fused by element-wise summation. Four depthwise separable convolutions are added sequentially, and a 1×1 convolution layer is added to obtain the spatial contexts features in the end.

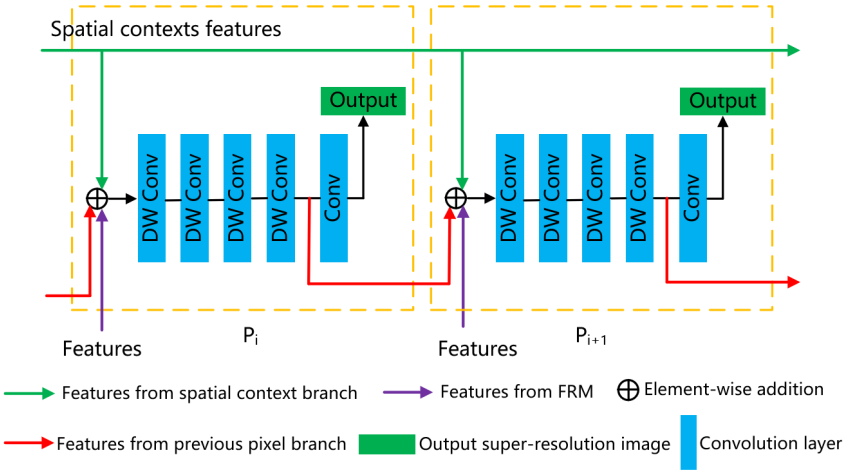

Figure 5. Details of the feature information flow used in the MHTC.

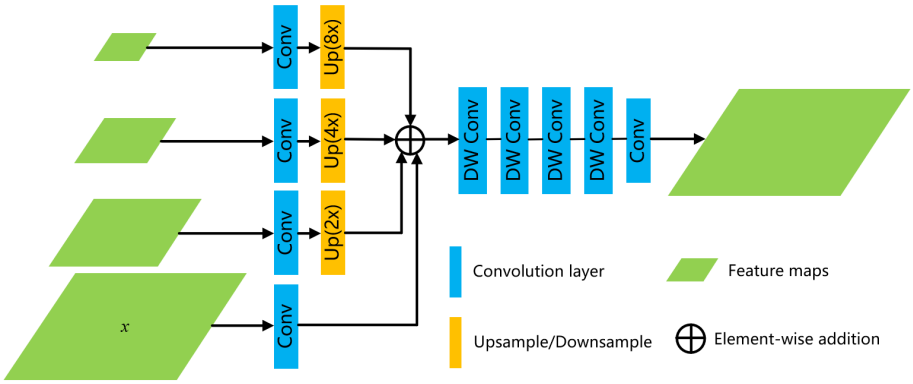

Figure 6. Details of spatial contexts branch used in the MHTC.

3. LOSS FUNCTIONS AND EVALUATION METRICS

We introduced a unified three-stage cascade framework to supervise the training process of MHTCUN so that the results of super-resolution could be better refined, as shown in Fig. 1. In the first stage, we down-sample the ground truth by a factor of four and upsample it to the original size using bilinear interpolation [7]. The loss is composed of Mean Square Error (MSE) [30] loss and structural similarity (SSIM) [31] loss, thus we calculate the loss between 4× blurred image and the output of first stage. Similarly, we calculate the loss between 2× blurred image and the output of the second stage. In the third stage, we calculate the loss between ground truth and the output of the third stage.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}||y(i,j) - \hat{y}(i,j)||^2 \tag{1}$$

The MSE between two monochrome images y and $\hat{y}$ can be calculated by formula (1), where the resolution is $m \times n$; y is the ground truth image; and $\hat{y}$ is the super-resolution image. The SSIM formula is based on three comparison measurements between the samples of y and $\hat{y}$: luminance (l), contrast (c), and structure (s). The individual comparison functions are Eqs. (2)–(4):

$$l(y,\hat{y}) = \frac{2\mu_y\mu_{\hat{y}} + C_1}{\mu_y^2 + \mu_{\hat{y}}^2 + C_1} \tag{2}$$

$$c(y,\hat{y}) = \frac{2\sigma_y\sigma_{\hat{y}} + C_2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2} \tag{3}$$

$$s(y,\hat{y}) = \frac{\sigma_{y\hat{y}} + C_3}{\sigma_y\sigma_{\hat{y}} + C_3} \tag{4}$$

where μ is the average of pixel values; σ is the variance of pixel values; $C_1,\ C_2,\ C_3$ are regularization constants. In order to simplify the expression, we set $C_3 = C_2/2$ [31]. Thus, the SSIM between y and $\hat{y}$ can be written as:

$$SSIM(y,\hat{y}) = l(y,\hat{y}) \cdot c(y,\hat{y}) \cdot s(y,\hat{y}) = \frac{(2\mu_y\mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2)} = l \cdot cs \tag{5}$$

where l is the luminance, and cs is the contrast sensitivity. The SSIM loss is:

$$L_{SSIM}(y,\hat{y}) = 1 - SSIM(y,\hat{y}) \tag{6}$$

Thus, the total loss of each stage is:

$$L_i = \alpha \cdot MSE(y,\hat{y}) + \beta \cdot L_{SSIM}(y,\hat{y}) \tag{7}$$

Through experiments, we know that MSE loss and SSIM loss are an order of magnitude different, and we set $\alpha = 1$, $\beta = 0.1$ in order to avoid the impact of gradient imbalance in our experiments. The total loss of MHTCUN is

$$L_{total} = a_1 \cdot L_1 + a_2 \cdot L_2 + a_3 \cdot L_3 \tag{8}$$

From the good performance of auxiliary loss in GoogLeNet [32], we set $a_1 = 0.3$, $a_2 = 0.3$, $a_3 = 1$ in our experiments.

We use peak signal-to-noise ratio (PSNR) and SSIM as evaluation metrics. PSNR can often be simply defined by MSE:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_y^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX_y}{\sqrt{MSE}}\right) \tag{9}$$

where MAX_y is the largest pixel value.

4. EXPERIMENTAL RESULTS

In this section, the results of the whole experiments are introduced. Firstly, we introduce the public dataset BioSR [25] and use F-actin filaments biological structure for our experiments. Then, we conduct ablation study on MHTCUN with different stages. In part 3, we compare the performance between MHTCUN and other DLSR models, such as SRCNN, EDSR, RCAN, RDN, and SRFBN. Next, we compare the performance between MHTCUN and the latest DLSR model DFCAN which is specially used for image super-resolution in optical microscopy. In part 5, we fine-tune the trained MHTCUN on other biological structures. Finally, the implementation details are discussed.

4.1. Dataset

This paper uses the public dataset BioSR [25] which is used for image super-resolution in optical microscopy specially. BioSR covers four biological structures, and they are clathrin-coated pits (CCPs), endoplasmic reticulum (ER), microtubules (MTs), and F-actin filaments (F-actin), as shown in Fig. 7.

The images were obtained by previous GI-SIM and TIRF-SIM setups with 3-phase $\times$3-orientation. We averaged each set of 3×3 raw images as diffraction-limited wide field (WF) images that were used as the input LR images. SIM images were used as the input ground truth (GT) images. We used an F-actin biology structure to train MHTCUN, and there were 51 cells with twelve signal levels. The size of input low resolution images was preprocessed to 128×128, and the corresponding GT image size was 256×256. In data preprocessing, we flipped the images along x axis, y axis, x and y axis for data augmentation. After preprocessing, the training set had 19,872 LR-GT image pairs, and the validation set had 1,920 image pairs.

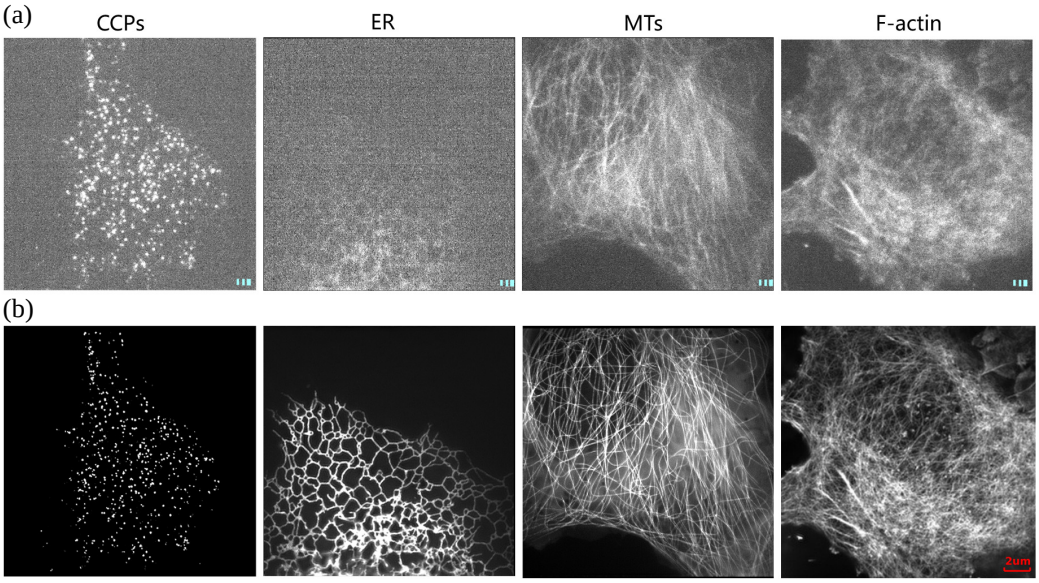

Figure 7. Representative images from the BioSR dataset. (a) Upper row shows the low resolution (LR) images. (b) Bottem row shows ground truth images corresponding to LR images. Different columns indicate different biological structures. Scale bar: 2 μm.

4.2. Ablation Study

We compare the effects of MHTCUN with different stages on the experimental results. Single-stage MHTCUN refers to using the output of stage 1 as the final super-resolution image. At this time, we calculate the loss between the GT image and the output of stage 1. Two-stage MHTCUN refers to using the output of stage 2 as the final super-resolution image. At this time, we calculate the loss between the GT image and the output of stage 2, and the loss between the 2× blurred image and the output of stage 1. Three-stage MHTCUN refers to using the output of stage 3 as the final super-resolution image. At this time, we calculate the loss between the GT image and the output of stage 3, the loss between the 2× blurred image and the output of stage 2, and the loss between the 4× blurred image and the output of stage 1. The training process is shown in Fig. 8.

It can be seen from Fig. 8 that the MHTCUN with 3 stages has the fastest convergence speed and

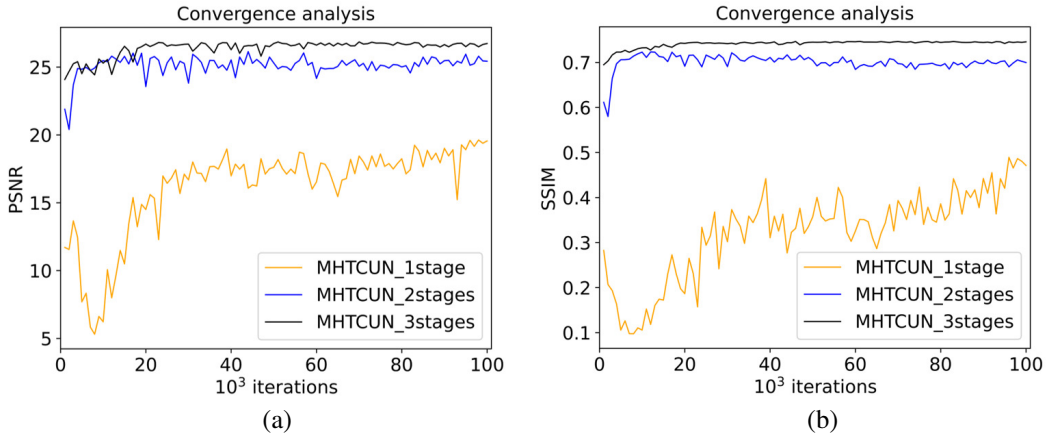

Figure 8. Training convergence analysis among different stages of MHTCUN. (a) Peak signal-to-noise ratio (PSNR). (b) Structural similarity (SSIM).

can reach a very high level in both PSNR and SSIM metrics. We choose MHTCUN with 3 stages for training.

4.3. Compared with State-of-the-Art DLSR Models for SISR

We compare MHTCUN with state-of-the-art DLSR models SRCNN, EDSR, RCAN, RDN, SRFBN in terms of both quantitative results and visual results. For quantitative comparison, we compare the PSNR and SSIM values of different models, as shown in Table 1. In Table 1, bold means the best, and our MHTCUN with about 26.40M parameters yields the best performance in PSNR of 26.87 and SSIM of 0.746. The performance of RCAN is second only to MHTCUN, but the number of its parameters is more than twice that of MHTCUN. For quality comparison, we provide the visual results for MHTCUN and other state-of-the-art DLSR models, as shown in Fig. 9. Training convergence analysis can be seen from Fig. 10.

Figure 10 shows the convergence of various DLSR model training iterations for 100,000 times. It can be seen from the convergence of PNSR and SSIM that MHTCUN has a faster convergence rate than other state-of-the-art DLSR models. Although RCAN can achieve PNSR and SSIM comparable to MHTCUN, it has more parameters. In the third section, we discussed that SSIM can finally be simplified

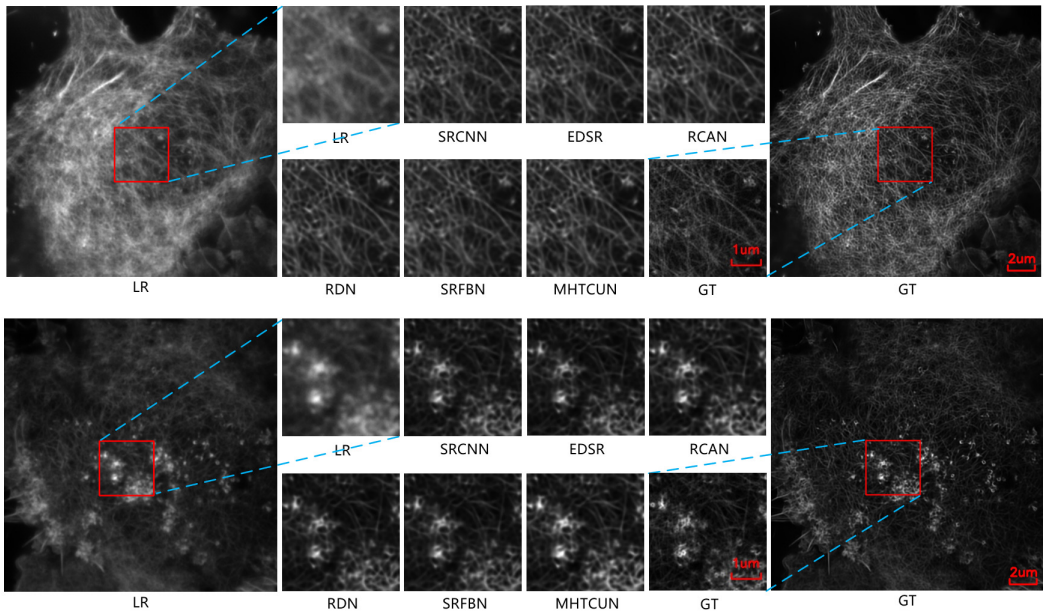

Figure 9. Super-resolution results using different DLSR models. The red box is the enlarged part. Scale bar: 2 µm or 1 µm (for magnified images).

Table 1. Performance comparison of various DLSR models on the F-actin biology structure.

Model	#Params(M)	PSNR	SSIM
SRCNN	0.23	25.29	0.723
EDSR	10.01	25.59	0.737
RCAN	60.99	26.84	0.746
RDN	86.53	26.02	0.741
SRFBN	8.37	25.75	0.737
MHTCUN	26.40	**26.87**	**0.746**

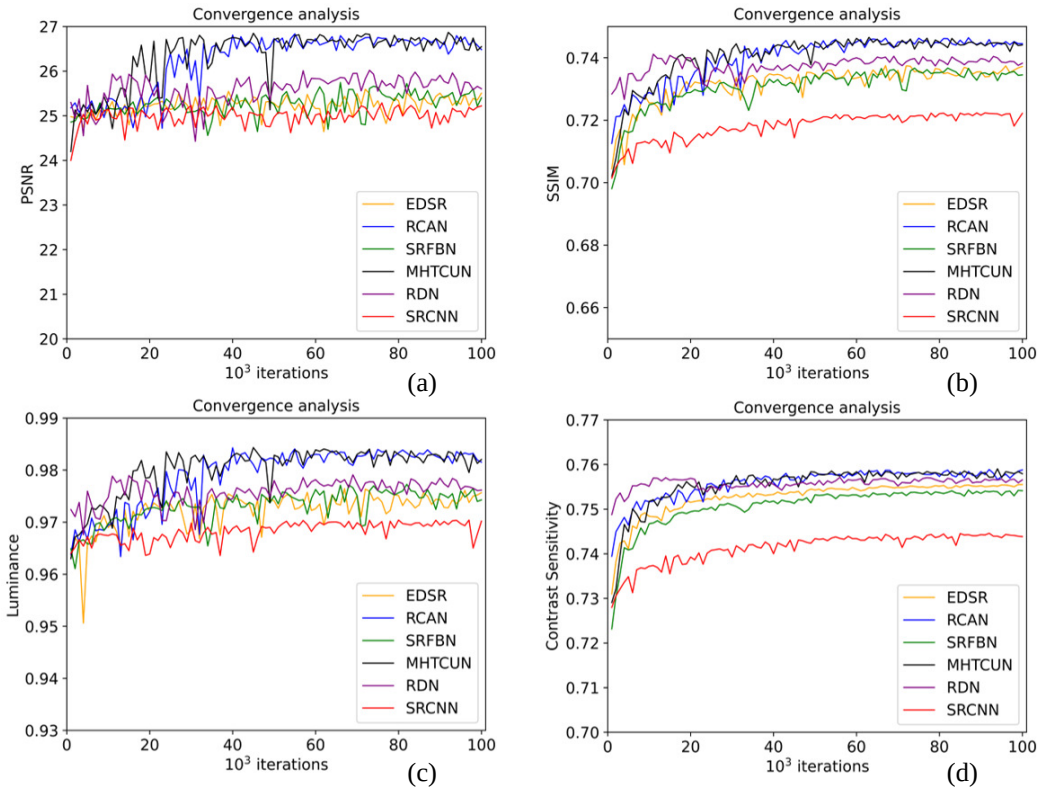

Figure 10. Training convergence analysis among different DLSR models. (a) Peak signal-to-noise ratio (PSNR). (b) Structural Similarity (SSIM). (c) Luminance. (d) Contrast Sensitivity.

as luminance and contrast sensitivity. Therefore, we compare the convergence speed of luminance and contrast sensitivity, as shown in the two figures below. It can be seen that the luminance and contrast sensitivity of MHTCUN also have a faster convergence speed.

4.4. Compared with DFCAN

DFCAN is the best reported model trained on F-actin filaments biological structure. Therefore, we use PSNR and SSIM values to compare DFCAN and MHTCUN in terms of both quantitative results and visual results. For quantitative comparison, we use Tukey box-and-whisker plots, as shown in Fig. 11.

From Fig. 11 we can see that the PSNR of MHTCUN reached higher value both in median and maximum than DFCAN. MHTCUN had a significant improvement in PSNR and a slight improvement in SSIM. For visual comparison, we provided the visual results between MHTCUN and DFCAN, as shown in Fig. 12. Fig. 12 shows that MHTCUN had better detail representation and more obvious contrast and brightness than DFCAN.

Intensity profile reflects the changes in the image structure. We scale each magnified image in Fig. 12 to fit the ground truth by minimizing the Mean Square Error (MSE) [30]. Compared with the ground truth, the MSEs of the MHTCUN and DFCAN images are 37.2563 and 51.5397, respectively. We draw in Fig. 13 the scaled intensity profiles along the lines in the magnified images of Fig. 12. It can be seen that the trend of the scaled intensity profile along the line of the MHTCUN image is much closer to the ground truth than that of the DFCAN image. In addition, we use the decorrelation analysis [33] to quantify the resolutions for the MHTCUN and DFCAN images. The decorrelation analysis is based on partial phase correlation and aims to find the highest frequency that contains sufficient signals standing out above the noise, as shown in Fig. 14. The resolution of the DFCAN image is 108.245 nm by using "ImageDecorrelationAnalysis" plugin [33] in software ImageJ [34], and the resolution of the MHTCUN image is 95.715 nm. Therefore, MHTCUN can produce a higher resolution image than DFCAN.

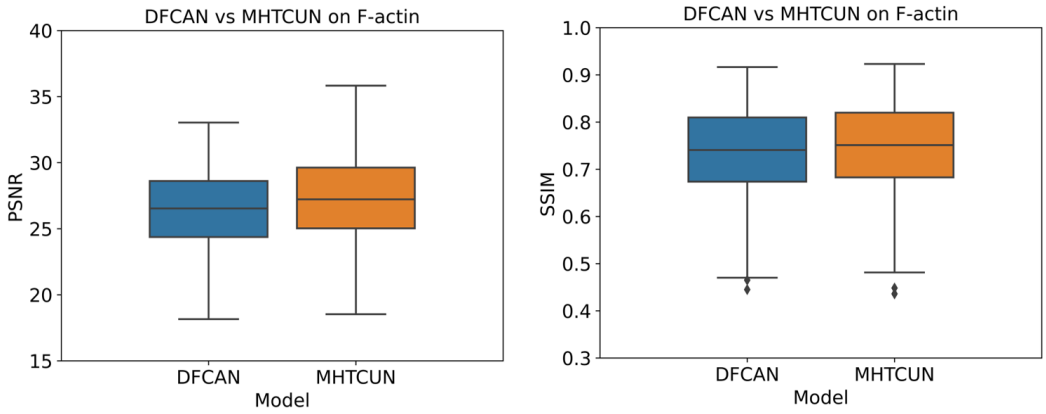

Figure 11. Model comparison between DFCAN and our MHTCUN on the dataset of F-actin flaments.

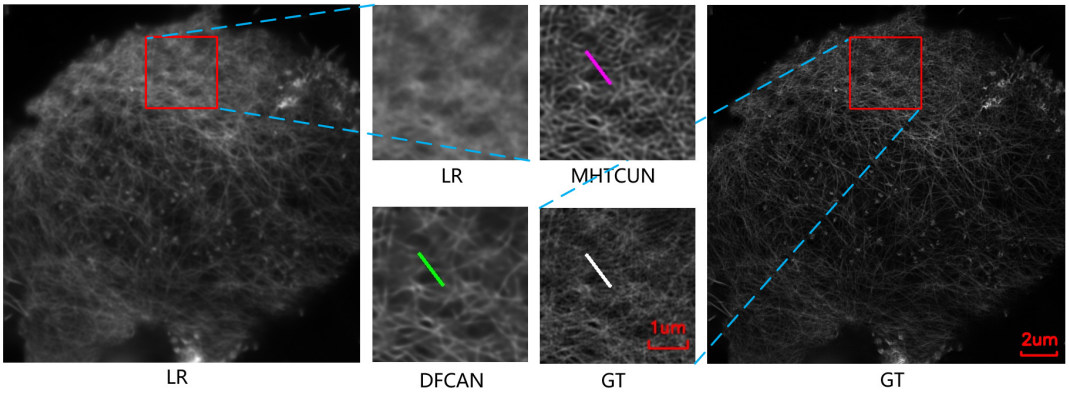

Figure 12. Comparison of super-resolution results on the dataset of F-actin filaments using DFCAN and our MHTCUN. Scale bar: 2 µm or 1 µm (for magnified images). Line colors in magnified images: magenta for MHTCUN; orange for DFCAN; white for GT.

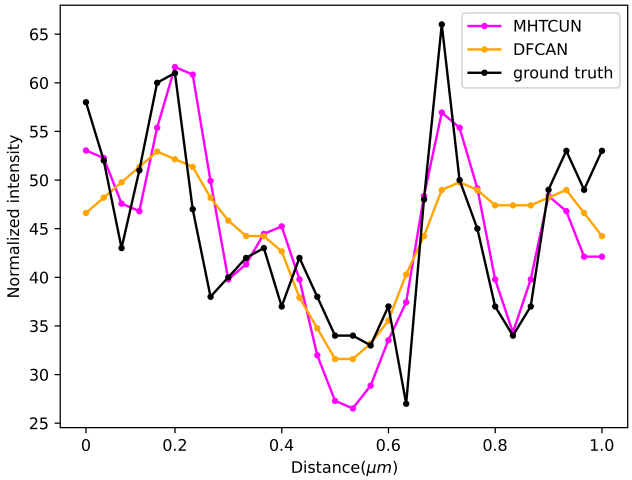

Figure 13. Scaled intensity profiles comparison along the lines in the magnified images of Fig. 12.

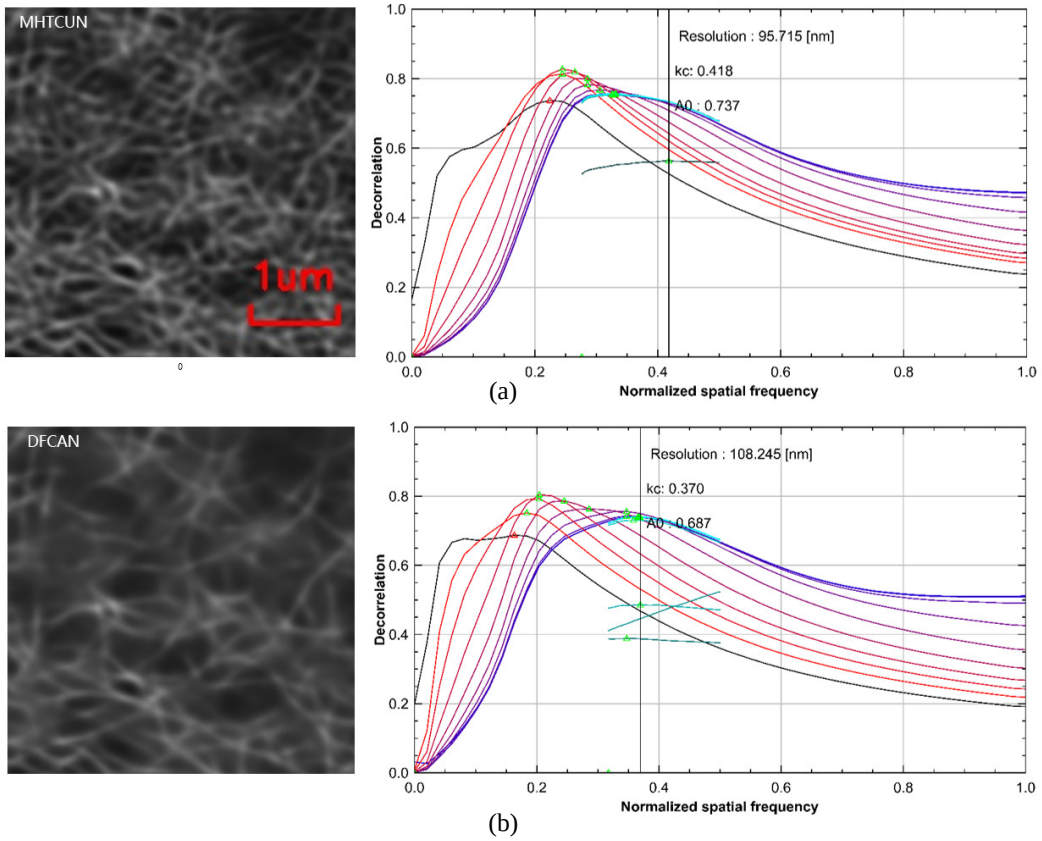

Figure 14. Comparison of resolution between the MHTCUN and DFCAN images in Fig. 12. (a) MHTCUN results. (b) DFCAN results. Black curve shows the decorrelation function without any high-pass filtering; blue curve is the decorrelation function of highest frequency peak; green dots are the local maxima of the decorrelation functions. Scale bar: 1 μm.

4.5. MHTCUN on Other Biology Structures

To demonstrate the robustness of MHTCUN, we fine-tuned trained MHTCUN on clathrin-coated pits (CCPs), endoplasmic reticulum (ER), and microtubules (MTs) biology structures. We fine-tuned 30,000 iterations on other biology structures and visualized the experimental results, as shown in Fig. 15. It can be seen from Fig. 15 that compared with low-resolution input images, our MHTCUN super-resolution results showed strong denoising, brightness, and contrast enhancement capabilities.

4.6. Other Implementation Details

Our experimental environment is as follows. The CPU is 32-core Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz; the memory is 128G, the graphics card is 4 GeForce RTX 2080Ti; the operating system is Ubuntu 16.04; the version of pytorch is 1.8.1; the version of cuda is 11.1, and the version of cudnn is 8.0.4.

For all DLSR models, we use MSE and SSIM as the loss function; the learning rate is set to 0.0001; the learning rate reduction method is step decay; and ADAM is used as the optimizer.

In our MHTCUN, the kernel size of all depthwise separable convolution layers is set to 3×3 and 1×1. The number of channels of input and output feature maps of the U-Net are set to 64. In the encoder step of U-Net, the numbers of channels of feature maps are set to 64, 128, 256. In the convolution operation, we pad zeros to fix the size of feature maps to the original size.

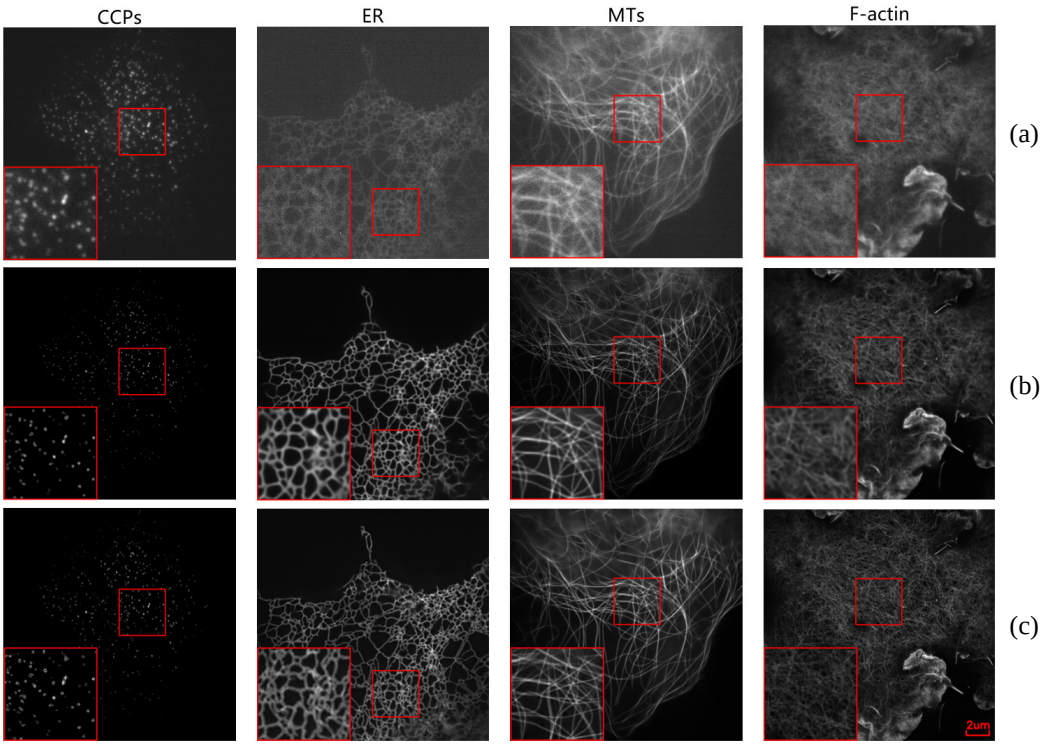

Figure 15. MHTCUN results for different biological structures. (a) The input low resolution images. (b) The super-resolution images. (c) The ground truth images. Scale bar: 2 µm.

5. CONCLUSION

This paper proposes a new DLSR model MHTCUN for image super-resolution in optical microscopy. In MHTC, we introduce three FRMs to refine the features and three pixel branches to generate super-resolution results from coarse to fine. We establish a feature information flow between different pixel branches, and the spatial context branch is introduced for better prediction. In the training process, we blur the ground truth images by a factor of 4 or 2 to calculate the loss between super-resolution results of different stages of MHTC. The experimental results show that MHTCUN has reach state-of-the-art performance among many state-of-the-art DLSR models both in PSNR and SSIM. Compared with DFCAN, MHTCUN has a significant improvement in PSNR. MHTCUN has shown superior performance in various aspects such as denoising, contrast enhancement, and resolution enhancement.

ACKNOWLEDGMENT

This work was partially supported by the National Key Research and Development Program of China (No. 2018YFC1407506), Special Development Fund of Shanghai Zhangjiang Science City, Ningbo Science and Technology Project (2020Z077 and 2020G012), Key Research and Development Program of Zhejiang Province (2021C03178), the National Natural Science Foundation of China (No. 11621101), and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).

REFERENCES

1. Zhao, Z.-Q., P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 11, 3212–3232, 2019.

2. Trajanovski, S., C. Shan, P. J. C. Weijtmans, S. G. B. de Koning, and T. J. M. Ruers, "Tongue tumor detection in hyperspectral images using deep learning semantic segmentation," *IEEE Transactions on Biomedical Engineering*, Vol. 68, No. 4, 1330–1340, 2020.
3. Zhao, S., D. M. Zhang, and H. W. Huang, "Deep learning-based image instance segmentation for moisture marks of shield tunnel lining," *Tunnelling and Underground Space Technology*, Vol. 95, 103156, 2020.
4. Yang, W., X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, Vol. 21, No. 12, 3106–3121, 2019.
5. Ledig, C., 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," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4681–4690, 2017.
6. Kim, K. I. 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, 1127–1133, 2010.
7. Kirkland, E. J., "Bilinear interpolation," *Advanced Computing in Electron Microscopy*, 261–263, Springer, 2010.
8. Liu, T., K. De Haan, Y. Rivenson, Z. Wei, X. Zeng, Y. Zhang, and A. Ozcan, "Deep learning-based super-resolution in coherent imaging systems," *Scientific Reports*, Vol. 9, No. 1, 1–13, 2019.
9. Dong, C., C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 2, 295–307, 2015.
10. He, K., X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778, 2016.
11. Lim, B., S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136–144, 2017.
12. Zhang, Y., K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image superresolution using very deep residual channel attention networks," *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301, 2018.
13. Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708, 2017.
14. Zhang, Y., Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2472–2481, 2018.
15. Li, Z., J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3867–3876, 2019.
16. Ronneberger, O., P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-assisted Intervention*, 234–241, Springer, 2015.
17. Chen, K., J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al., "Hybrid task cascade for instance segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4983, 2019.
18. Cai, Z. and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6154–6162, 2018.
19. Hell, S. W. and J. Wichmann, "Breaking the diffraction resolution limit by stimulated emission: Stimulated-emission-depletion fluorescence microscopy," *Optics Letters*, Vol. 19, No. 11, 780–782, 1994.

20. Hess, S. T., T. P. K. Girirajan, and M. D. Mason, "Ultra-high resolution imaging by fluorescence photoactivation localization microscopy," *Biophysical Journal*, Vol. 91, No. 11, 4258–4272, 2006.
21. Rust, M. J., M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)," *Nature Methods*, Vol. 3, No. 10, 793–796, 2006.
22. Gustafsson, M. G. L., "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy," *Journal of Microscopy*, Vol. 198, No. 2, 82–87, 2000.
23. Weigert, M., U. Schmidt, T. Boothe, A. Müller, A. Dibrov, A. Jain, B. Wilhelm, D. Schmidt, C. Broaddus, S. Culley, et al., "Content-aware image restoration: Pushing the limits of fluorescence microscopy," *Nature Methods*, Vol. 15, No. 12, 1090–1097, 2018.
24. Wang, H., Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydın, L. A. Bentolila, C. Kural, and A. Ozcan, "Deep learning enables cross-modality superresolution in fluorescence microscopy," *Nature Methods*, Vol. 16, No. 1, 103–110, 2019.
25. Qiao, C., D. Li, Y. Guo, C. Liu, T. Jiang, Q. Dai, and D. Li, "Evaluation and development of deep neural networks for image super-resolution in optical microscopy," *Nature Methods*, Vol. 18, No. 2, 194–202, 2021.
26. Shi, W., J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1874–1883, 2016.
27. Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
28. Ramachandran, P., B. Zoph, and Q. V. Le, "Searching for activation functions," arXiv preprint arXiv:1710.05941, 2017.
29. Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125, 2017.
30. Allen, D. M., "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, Vol. 13, No. 3, 469–475, 1971.
31. Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, Vol. 13, No. 4, 600–612, 2004.
32. Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9, 2015.
33. Descloux, A., K. S. Grußmayer, and A. Radenovic, "Parameter-free image resolution estimation based on decorrelation analysis," *Nature Methods*, Vol. 16, No. 9, 918–924, 2019.
34. Abràmoff, M. D., P. J. Magalhães, and S. J. Ram, "Image processing with imagej," *Biophotonics International*, Vol. 11, No. 7, 36–42, 2004.