cells

MDPI

Article

ContransGAN: Convolutional Neural Network Coupling Global Swin-Transformer Network for High-Resolution Quantitative Phase Imaging with Unpaired Data

Hao Ding [1], Fajing Li [1], Xiang Chen [1], Jun Ma [2], Shouping Nie [1], Ran Ye [1] and Caojin Yuan [1,*]

[1] Key Laboratory for Opto-Electronic Technology of Jiangsu Province, Nanjing Normal University, Nanjing 210023, China; 191035004@njnu.edu.cn (H.D.); 191002029@njnu.edu.cn (F.L.); 211002001@njnu.edu.cn (X.C.); nieshouping@njnu.edu.cn (S.N.); ran.ye@njnu.edu.cn (R.Y.)

[2] School of Electronic Engineering and Optoelectronic Techniques, Nanjing University of Science and Technology, Nanjing 210094, China; majun@njust.edu.cn

* Correspondence: yuancj@njnu.edu.cn

Abstract: Optical quantitative phase imaging (QPI) is a frequently used technique to recover biological cells with high contrast in biology and life science for cell detection and analysis. However, the quantitative phase information is difficult to directly obtain with traditional optical microscopy. In addition, there are trade-offs between the parameters of traditional optical microscopes. Generally, a higher resolution results in a smaller field of view (FOV) and narrower depth of field (DOF). To overcome these drawbacks, we report a novel semi-supervised deep learning-based hybrid network framework, termed ContransGAN, which can be used in traditional optical microscopes with different magnifications to obtain high-quality quantitative phase images. This network framework uses a combination of convolutional operation and multiheaded self-attention mechanism to improve feature extraction, and only needs a few unpaired microscopic images to train. The ContransGAN retains the ability of the convolutional neural network (CNN) to extract local features and borrows the ability of the Swin-Transformer network to extract global features. The trained network can output the quantitative phase images, which are similar to those restored by the transport of intensity equation (TIE) under high-power microscopes, according to the amplitude images obtained by low-power microscopes. Biological and abiotic specimens were tested. The experiments show that the proposed deep learning algorithm is suitable for microscopic images with different resolutions and FOVs. Accurate and quick reconstruction of the corresponding high-resolution (HR) phase images from low-resolution (LR) bright-field microscopic intensity images was realized, which were obtained under traditional optical microscopes with different magnifications.

Keywords: super-resolution; transformer; CNN; quantitative phase imaging; transport of intensity equation

Citation: Ding, H.; Li, F.; Chen, X.; Ma, J.; Nie, S.; Ye, R.; Yuan, C. ContransGAN: Convolutional Neural Network Coupling Global Swin-Transformer Network for High-Resolution Quantitative Phase Imaging with Unpaired Data. *Cells* **2022**, *11*, 2394. https://doi.org/10.3390/cells11152394

Academic Editors: An Pan, Baoli Yao, Chao Zuo, Fei Liu, Jiamiao Yang and Liangcai Cao

Received: 12 June 2022
Accepted: 31 July 2022
Published: 3 August 2022

1. Introduction

Recent advances in microscopy have allowed imaging of biological processes with a higher level of quality [1–4]. However, these microscopy techniques are usually limited by sophisticated setups and special experimental conditions. In addition, the resolution of optical microscopic imaging is limited by the numerical aperture (NA) of the microscope, the wavelength of illuminating light, and the pixel spacing of the imaging recording device. As a result, researchers in related fields are usually committed to dealing with the above trade-off and improving the imaging efficiency [5].

Since the huge volume of data and the continuous improvement of the computing power, artificial intelligence (AI) has developed rapidly. Over the past few years, deep learning technology has been leading the development of AI and is widely used in computer vision [6], natural language processing [7], speech recognition [8], and other fields.

Deep learning uses multiple-layer neural networks to automatically analyze signals or data, which has unique advantages in solving "inverse problems" and nonlinear problems. The unique advantages of deep learning technology have also aroused the interest of optical imaging scholars. Deep learning technology has been used to solve problems that lack effective solutions in optical computational imaging, such as optical tomography [9,10], optical fiber imaging [11], ghost imaging [12], scattering imaging [13], and low-light environment imaging [14]. At the same time, deep learning technology has also been widely used in various important directions of QPI, such as phase retrieval [14–18], super-resolution [19–23], phase unwrapping [24–27], label-free detection [28], and various other image enhancement techniques [29–32]. There is no exception for the use of deep learning in breaking through the trade-off between the resolution and the field of view in microscopy [33,34]. The general idea is to use a network to learn the relationship between LR and HR image pairs. The input LR image will be converted into an HR image but the same field of view as an LR image is maintained using the built network model. However, the required pairs of LR-HR data are obtained by mechanically switching between the low NA and high-power objectives in the experiment and numerical registration in pretreatment, which is a very tedious process. In order to replace the time-consuming image registration processes, Zhang [35] et al. degraded the captured HR microscopic images to simulate the LR microscopic image, However, it is necessary to adjust the parameters repeatedly to ensure that it is similar to the LR images, which are obtained by the experiment. In addition, the simulated LR microscopic images are not consistent with the real image degeneration process, which usually includes blurring, noise, and other defects. Further, some works [36–38] proposed the use of unsupervised methods for super-resolution imaging, in order to reduce the amount of training data. The feature extraction results of unsupervised methods are random to some extent, so noise or blurring is easily amplified. Meanwhile, there are some works that train the network using unpaired datasets [39,40]. This kind of semi-supervised deep learning network framework greatly reduces the difficulty of datasets acquisition. However, these works only involve conversions between image styles (such as intensity images are converted to phase images) and do not simultaneously achieve image resolution enhancement due to the traditional CNN network showing a poor performance in multiple tasks. For example, it is difficult to achieve the expected imaging effect when performing super-resolution and quantitative phase imaging at the same time [41].

Researchers have attempted to improve the network performance by adding the residual module, feedback mechanism, or attention mechanism [42–44] to CNN. However, these frameworks still have some fundamental limitations. On the one hand, the convolution operation of CNN is good at extracting local features but poot at extracting global features. The lack of a global understanding ability results in a loss of rich information in LR microscopic. On the other hand, the weights of the convolution network are fixed and it cannot adapt to the change in the input dynamically.

In this paper, we propose an end-to-end deep learning framework, which is termed ContransGAN, coupling CNN with Vision Transformer (ViT) [45] to obtain HR phase information from LR intensity information. This framework retains the advantage of CNN in extracting the details of local features, and enhances its ability to capture the global feature. Furthermore, the network framework can be trained with unpaired images, which ease the experiment and data preparation. The large FOVs HR quantitative phase images can be reconstructed from the LR intensity image using the trained ContransGAN. We verified the effectiveness of the ContransGAN algorithm and its generalization performance by acquiring LR microscopic images of different samples under an inverted microscope.

2. Methods

2.1. Imaging Hardware and TIE Phase Extraction

TIE is an ideal candidate for phase imaging with partially coherent illuminations [46]. In relation to its closed-form solution, TIE offers a cost-efficient way of measuring the phase by a single step. The phase recovery by TIE is standard and is briefly explained for the

completeness of this paper. In the paraxial approximation, TIE can be expressed as the following format

$$-k\frac{\partial I(r)}{\partial z} = \nabla\cdot[I(r)\nabla\varphi(r)] \tag{1}$$

where $k = 2\pi/\lambda$ is the wave number, r represents the transverse spatial coordinates perpendicular to the optical axis, φ is the phase distribution of the sample, ∇ is the two-dimensional gradient with respect to r, and $I(r)$ represents the actual intensity distribution along the optical axis in the plane of $z = 0$. The left-hand side of the Equation (1) represents the axial differentiation of the intensity distribution. Solving the TIE requires the axial differentiation to be obtained in advance. Specifically, the axial differentiation of the intensity distribution is obtained by acquiring two slightly out-of-focus images with an equal off-focus distance and opposite direction with respect to the center-focused image, using the central finite difference method estimation [47]. As expressed in Equation (2), the derivative of the intensity distribution is approximated by the finite difference as

$$\frac{\partial I(r)}{\partial z} \approx \frac{I(r,\Delta z) - I(r,-\Delta z)}{2\Delta z} \tag{2}$$

and the extracted phase can be expressed as the Equation (3)

$$\varphi(r) \approx -k\nabla^{-2}\nabla\cdot[I^{-1}(r)\nabla\nabla^{-2}\frac{I(r,\Delta z) - I(r,-\Delta z)}{2\Delta z} \tag{3}$$

where ∇^{-2} is the inverse Laplacian operator. In our work, we solved the TIE using the fast Fourier transform (FFT) algorithm [48] under the homogeneous Neumann boundary condition [49].

As shown in Figure 1a, an inverted microscope (Nikon ECLIPSE Ti-S) was used for the experimental setup. The illumination light source was a halogen lamp, and the central wavelength of the filtered illumination light source was 550 nm. The beam passed through the specimen and carried the specimen's information, and then is focused by an objective. The microscopic images were captured by a CCD camera through a tube lens. Under the high-power microscope objective ($40\times$ /0.65 NA), we captured the microscopic images with the same out-of-focus depths ($\Delta z = 3$ μm) on both sides of the focal plane of the specimen and the corresponding in-focus microscopic images, and then extracted the corresponding HR phase information with the TIE algorithm [50–53]. Under the low-power microscope objective ($4\times$ /0.1 NA, $10\times$ /0.25 NA, and $20\times$ /0.4 NA), we captured LR microscopic images with different resolutions as the input of the ContransGAN.

In our experiment, the spatial coherence of the illumination light source is determined by the size of the aperture diaphragm [54]. The spatial coherence is represented by the coherence parameter S, which is the ratio of the condenser aperture to the objective NA. As shown in Figure 1b, the contrast and resolution of the images that are recorded under the corresponding aperture diaphragm are different. Adjusting the aperture of the diaphragm to $S \approx 0.3$ ensures that the captured microscopic images have fine contrast, so that quantitative phase information can be calculated by TIE [46,55]. Meanwhile, adjusting the aperture of the diaphragm to $S \approx 0.6$ captures the LR microscopic images. Figure 1c,f show the microscopic image obtained in the experiment and the corresponding LR phase image recovered by TIE respectively. Figure 1d,e show the HR intensity image and the HR quantitative phase image reconstructed by TIE respectively. The proposed deep learning framework finally generates the HR phase images consistent with the Figure 1f.

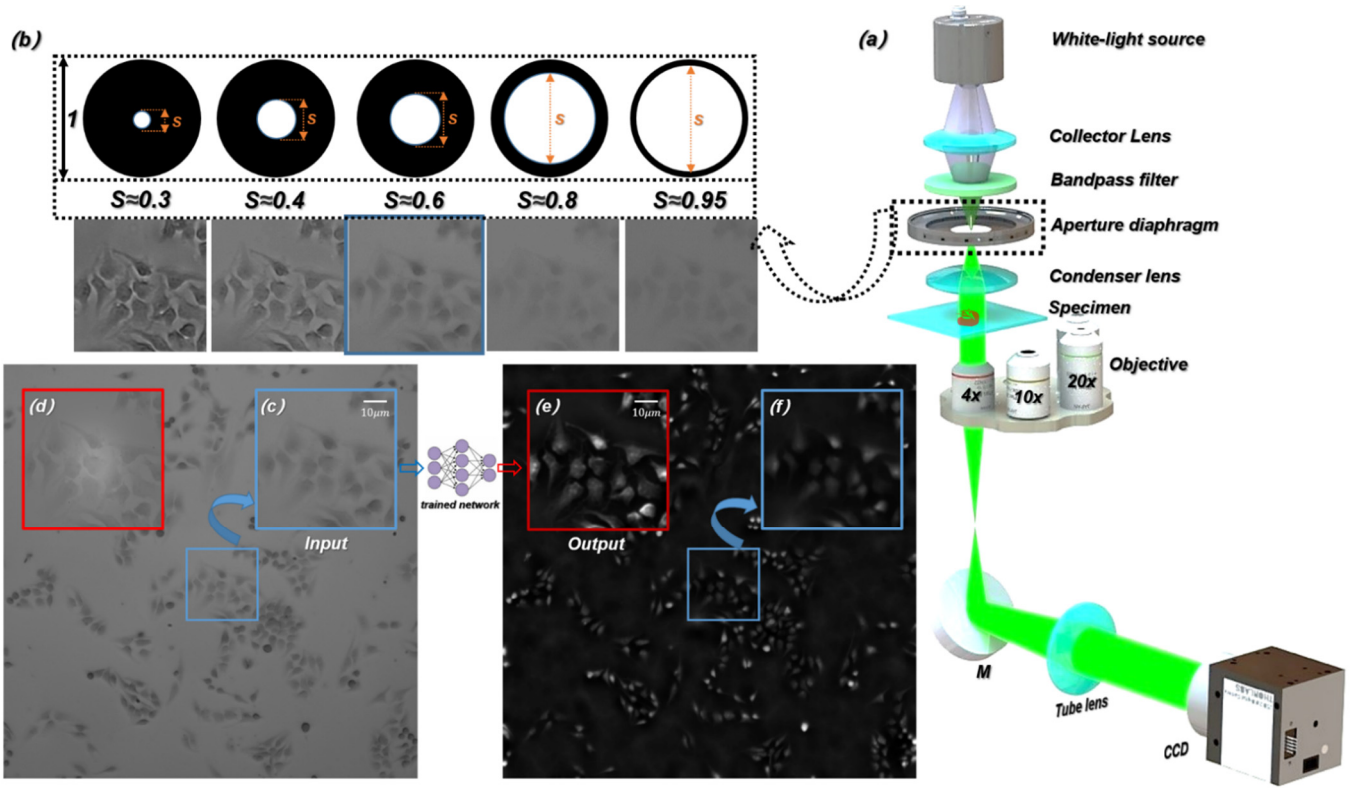

Figure 1. Experimental setup and acquisition process of the original images. (**a**) Imaging hardware setup. (**b**) Microscopic images obtained under a different aperture diaphragm. (**c**) The LR microscopic image captured directly in the experiment. (**d**) The HR microscopic image captured directly in the experiment. (**e**) The reconstructed HR quantitative phase image by TIE. (**f**) The reconstructed LR quantitative phase image by TIE.

2.2. Creation of Datasets and Networks Training Details

The ContransGAN proposed in our work was designed based on the CycleGAN architecture [56]. The network framework is essentially composed of two symmetrical generative adversarial networks (GAN). The flow chart of the entire training is shown in Figure 2a. The framework includes the generator G_{AB} and generator G_{BA} for performing conversion between images. Correspondingly, the discriminator D_A and discriminator D_B are responsible for judging whether the images generated by the generators are close to the reality. The training dataset consists of R_A (input, LR microscopic images) and R_B (ground truth, HR quantitative phase images) respectively. During the process of training, the LR microscopic image in R_A is input into the generator G_{AB} to obtain F_B, and then F_B is input into the discriminator D_B to extract eigenvalues, which are used to calculate $Loss_{DB}$. At the same time, F_B is also input into G_{BA} to generate RE_A. The training process of R_B is consistent with that of R_A. As expressed in Equation (4), the overall loss function can be written as

$$Loss = [Loss_{GAN}] + \lambda\left\{Loss_{cycle}\right\} = [Loss_{DA} + Loss_{DB}] + \lambda\left\{Loss_{cycleABA} + Loss_{cycleBAB}\right\} \quad (4)$$

where λ is used to adjust the proportion of $Loss_{cycle}$, and the value is set to 10. The main function of $Loss_{GAN}$ is to mutually promote the performance of the generators and discriminators. The overall loss function enables the generators to produce images approximating well to real ones; the main function of $Loss_{cycle}$ is to ensure that the output

images of the generators are different from the input images in style but consistent in content. Specifically, as expressed in Equation (5), $Loss_{GAN}$ can be written as

$$\begin{aligned} Loss_{GAN} &= Loss_{DA} + Loss_{DB} \\ &= E_b\left[(D_B(b)-1)^2\right] + E_a\left[(1-D_B(G_{AB}(a)))^2\right] \\ &+ E_b\left[(D_A(a)-1)^2\right] + E_b\left[(1-D_A(G_{BA}(b)))^2\right] \end{aligned} \tag{5}$$

where $E_{[\cdot]}$ represents the expected value of the random variable in square brackets; a and b represent the images in dataset R_A versus R_B respectively. $Loss_{cycle}$ is expressed in Equation (6), which used to further optimize the model. It can be written as

$$\begin{aligned} Loss_{cycle} &= Loss_{cycleABA} + Loss_{cycleBAB} \\ &= E_a[\| G_{BA}(G_{AB}(a)) - a \|_1] + E_b[\| G_{AB}(G_{BA}(b)) - b \|_1] \end{aligned} \tag{6}$$

where $\| \cdot \|_1$ represents the norm L1.

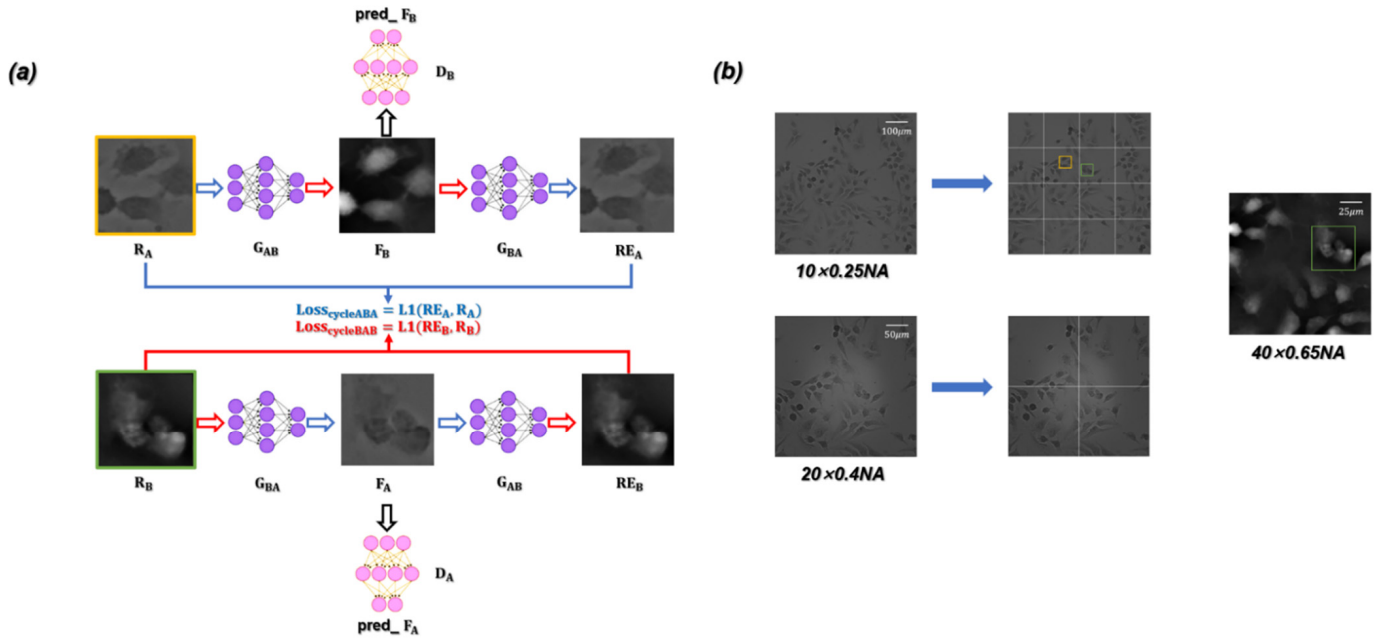

Figure 2. (**a**) Training flow diagram. (**b**) By segmenting the LR microscopic images with different resolutions, the sub-images with an FOV approximately equal to the HR quantitative phase images are obtained for training the ContransGAN. Orange and green rectangles are Region of Interest (ROI).

In this paper, unstained Hela cells and polystyrene microspheres (PSMs) are used as the experimental specimens. R_A and R_B consist of 3500 unpaired LR microscopic images and 3500 HR quantitative phase images respectively. It is worth noting that by segmenting the original LR microscopic images, we obtain the LR microscopic images, which are approximately equal to the FOVs of the HR quantitative phase images. As shown in Figure 2b, the LR microscopic image that was captured by the 10 × /0.25 NA objective is equally divided into 16 sub-images, so that the field-of-view range of each sub-image is approximately equal to that of the 40 × /0.65 NA objective. The FOV of each sub-image is approximately equal to the HR quantitative phase image reconstructed by the corresponding microscopic images captured under the 40 × /0.65 NA objective. In the process of model building, in order to enhance the network generalization ability and improve the training efficiency and precision, we cropped or scaled the input LR microscopic images by random image interpolation [57]. Among all the datasets, 85% are used for the training dataset and the remaining 15% for the testing dataset. The ContransGAN is implemented by python 3.6.8 based Pytorch 1.3.1 and the network training

and testing on a PC with double Intel Aeon Gold 5117 CPU @ 2.00 GHz and 128 GB RAM, using NVIDIA Forced RTX 2080 Ti GPU. The training process takes ~50 h for 80 epochs (in a batch size of 2). Finally, the imaging speed of the trained ContransGAN for a phase image can reach ~0.06 s.

2.3. Vision Transformer and Self-Attention Mechanism

Before introducing the generator and the discriminator, it is necessary to introduce the relevant theories of the Transformer in detail. Transformer [58] is a classic model for natural language processing (NLP) proposed by Google in 2017. It uses the self-attention mechanism instead of the sequential structure of the recurrent neural network (RNN) [59] so that the model can be trained in parallel and has global information. Recently, the Transformer structure has been used in ViT. Figure 3a shows the part used for feature extraction in ViT, which constructs a series of marker sequences by dividing each image into Patch with position embedding, and then uses the Transformer module to extract parametric vectors as visual representations. Position embedding records the sequence correlation between sequence data. Compared with the characteristics of the RNN sequential input, the method based on Transformer can directly input data in parallel and store the position relationship between data, which greatly improves the computing speed and reduces the storage space. In addition, with the increase in the number of network layers, the distribution of the data will continue to change. In order to ensure the stability of the data feature distribution, a layer of regularization [60] is introduced to reduce information loss.

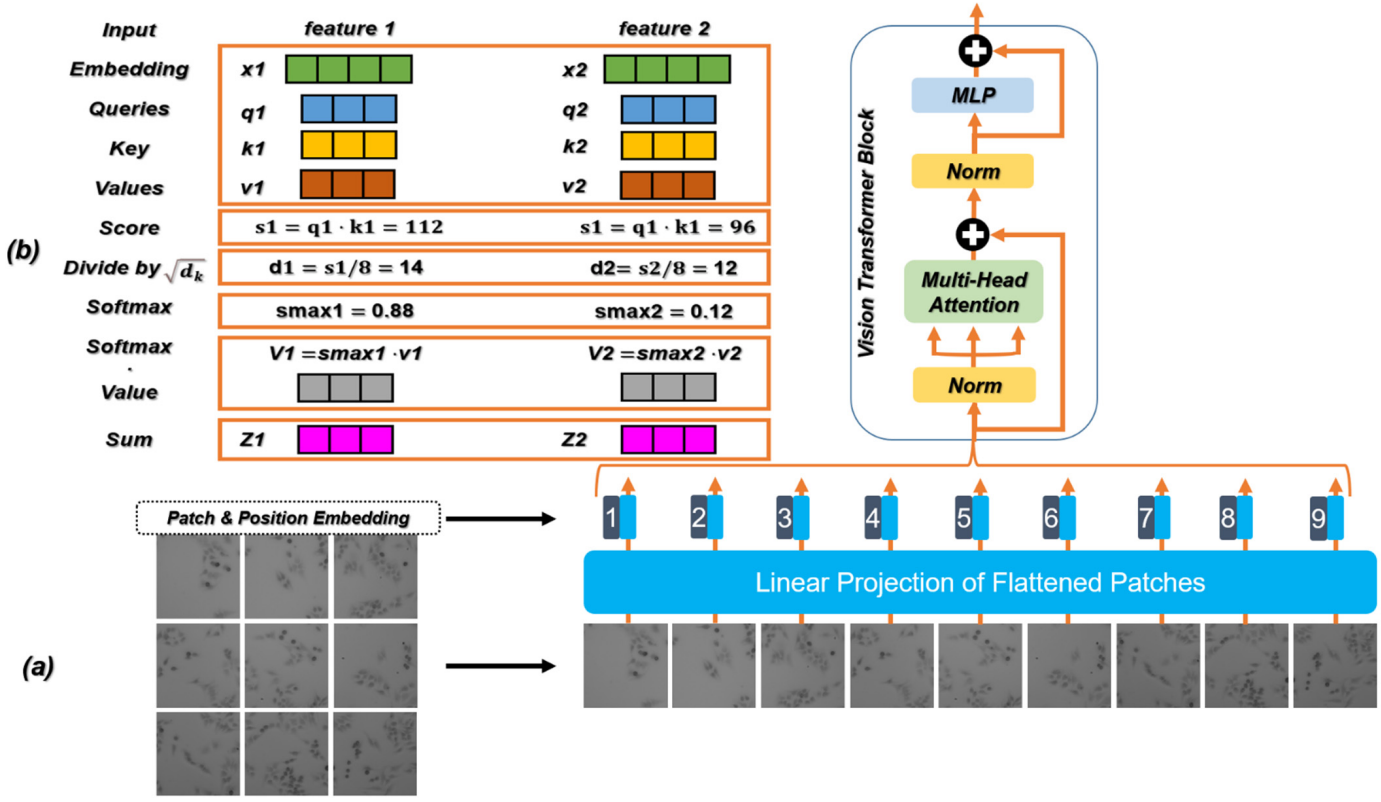

Figure 3. (**a**) The feature extraction process and the specific structure of ViT. (**b**) Calculation flow chart of the self-attention mechanism in ViT.

The attention mechanism imitates the internal process of biological observation behavior and enhances the fineness of observation in some areas. Since it can quickly extract the important features of sparse data, the attention mechanism is widely used in machine translation, speech recognition [61], image processing [62], and other fields. The attention mechanism has become an important concept in the field of neural networks. It is an

advanced algorithm for multitasking, which is widely used to improve the interpretation of neural networks, and helps to overcome some challenges in RNN, such as performance degradation with the increase in the input length and computational inefficiency caused by an unreasonable input sequence. The self-attention mechanism is the improvement of the attention mechanism, which reduces the dependence of the network on external information and is better at capturing the internal relevance of data or features. Transformer introduces the self-attention mechanism to avoid the use of recursion in the neural network, and completely relies on the self-attention mechanism to draw the global dependence between the input and output. In the calculation, the input needs to be linearly transformed to obtain the matrices: Query (Q), Key (K), and Value (V). As expressed in Equation (7), the calculation formula can be written as

$$\text{Attention}(A, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

where d_k is the number of columns of the matrix Q and K.

The calculation process of the self-attention mechanism is shown in Figure 3b and its steps are as follows:

Step 1: Create three vectors. The input feature map is linearly projected into three different spaces, resulting in three new vectors, namely Q, K, and V.

Step 2: Calculate the score.

Step 3: Divide by the scaling factor. The score in Step 2 divided by the scaling factor square $\sqrt{d_k}$ (the square root of the dimension of K), where the raw attention values are all clustered around the highest scoring value. This step can play the role of scaling and distraction.

Step 4: Normalization by the softmax [63]. The correlation between the current feature vector and each feature vector in the feature graph is obtained by the softmax.

Step 5: Multiply each V vector by the softmax. Reduce the concern of uncorrelated feature vectors.

Step 6: The accumulated weighted value vector generates an updated feature map as output.

Here, since each location has information about other features in the same image, the dependencies between long-distance interval features in space can be obtained. On this basis, the essence of the multi-head self-attention mechanism used in ViT is to split the three parameters Q, K, and V multiple times while the total number of parameters is constant, and each group of split parameters is mapped to different subspaces of high-dimensional space to calculate the attention weight to focus on different parts of the input. After several parallel calculations, the attention information in all subspaces is merged. Due to the different distribution of attention in different subspaces, multi-head self-attention is actually looking for the correlation between the input data from different angles, so that multiple relationships and subtle differences can be encoded. Multiple independent heads pay attention to different information (such as global information and local information) to extract more comprehensive and rich features.

2.4. Generator and Discriminator

Due to the introduction of the self-attention mechanism and multilayer perceptron (MLP) structure [64], ViT can reflect complex spatial transformation and long-distance feature dependence, thus obtaining global feature representation. However, ViT ignores local feature details, which reduces the distinguishability between high-frequency information and low-frequency information.

In our work, the Contrans was proposed as the generator, which uses two sampling channels to combine local features based on CNN and global representation based on Transformer to enhance representation learning. As shown in Figure 4a, the Contrans consists of an improved ViT module branch (termed Swin-Transformer [65]) and a CNN branch. In the process of training, ViT calculates the global self-attention of the feature maps. However,

Swin-Transformer is a process in which the window is enlarged, and then the calculation of self-attention is calculated in terms of the window, which is equivalent to introducing the information of local aggregation. This process is very similar to the convolution in CNN, just like the step size and convolution kernel size of CNN, so that the window is not coincident. The difference is that CNN performs convolution calculation in each window, and obtains a new window composed of eigenvalues, which represents the characteristics of this window, while Swin-transformer calculates the self-attention value of each window to obtain an updated window, then merges the windows through the operation of Patch Merging, and continues to calculate the self-attention of the merged window (this process is termed W-MSA), which can also reduce the computational complexity. As shown in Figure 5a, the size of the input is 224 × 224 and the window size is 7 × 7, which is composed of 7 × 7 Patch. A box in Figure 5a represents a window. The size of the Patch changes with the operation of Patch Merging. For example, the Patch of the initial feature map is 4 × 4. By splicing the Patch of the four surrounding windows, the Patch of the feature map of the next layer becomes 8 × 8. By a series of operations, the Swin-Transformer downsampling obtains the feature map with only 1 window and 49 Patch with a size of 32 × 32.

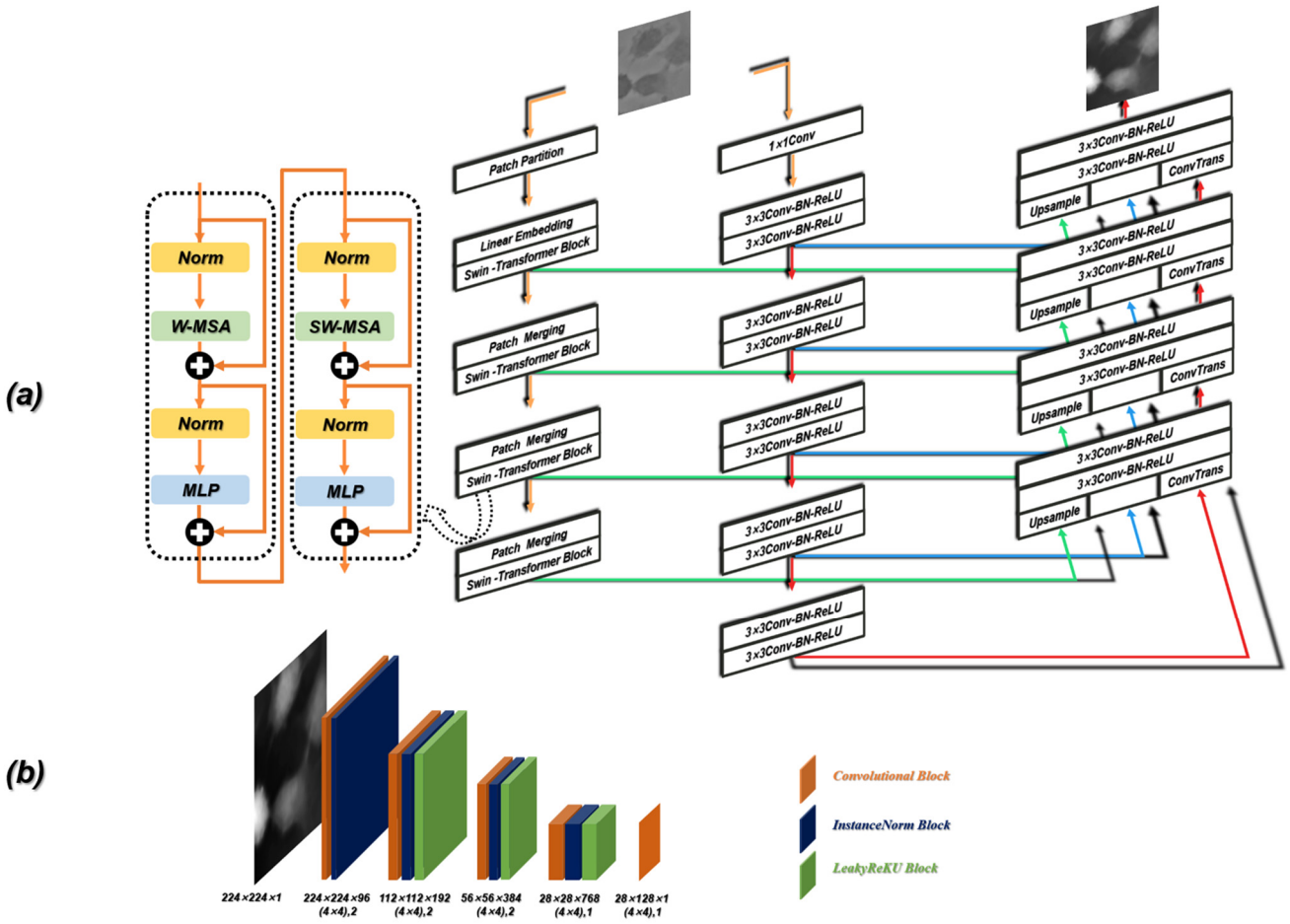

Figure 4. Detailed schematic of the ContransGAN architecture. (**a**) The schematic of the generator. (**b**) The schematic of the discriminator.

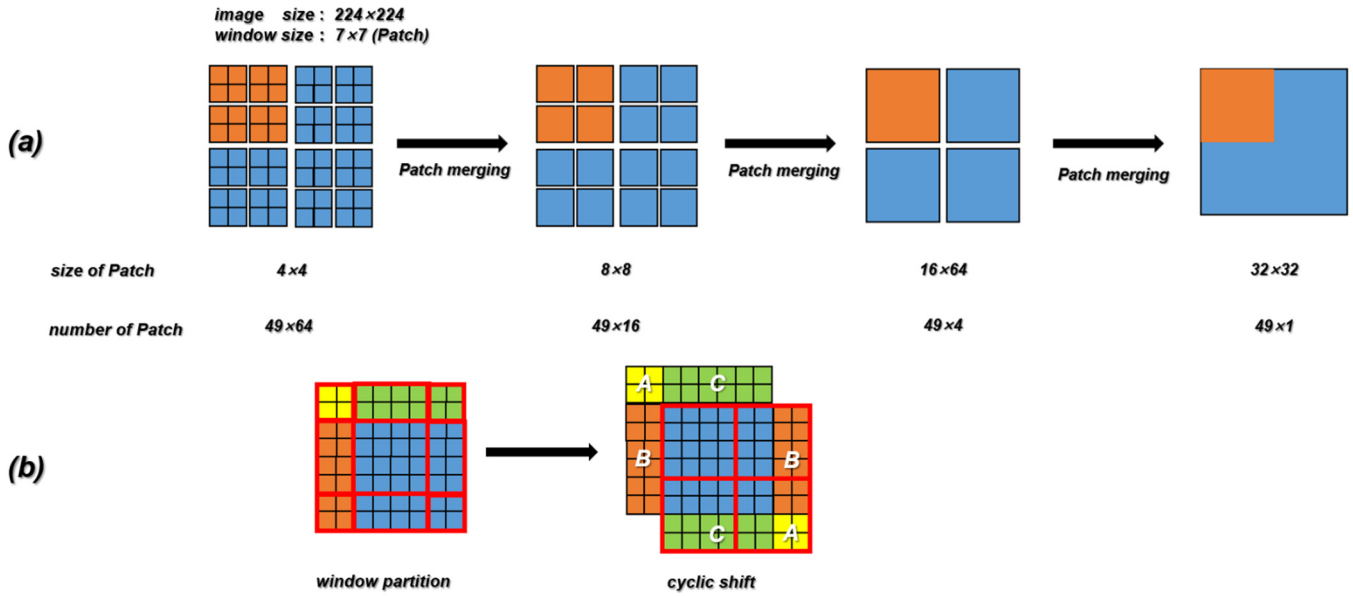

Figure 5. (**a**) The downsampling process of Swin-Transformer. (**b**) Schematic diagram of the SW-MAS operation flow. A, B and C are the regions which need to be moved in the previous feature map.

W-MSA operation reduces complexity but brings new problems, that is, a lack of information exchange between windows that are not coincident. In order to exchange information between windows, the region of the feature map can be divided and then moved and spliced. As shown in Figure 5b, the initial feature map is divided into nine regions. We move the upper left region (regions A, B, and C) to the lower right, and then divide the spliced feature map into four equal regions, so that the information between each window can be exchanged. After the downsampling of the Swin-Transformer and CNN branches, the deconvolution operation (transpose convolution) is performed on the feature map obtained by the CNN branch, and the result is spliced with the feature map generated by the corresponding feature layer in the downsampling process. In the process of stitching, to make the size of the three parts of the feature image consistent, the corresponding feature map of the Swin-Transformer needs to be magnified four times by the Upsample operation.

The discriminator in ContransGAN is the PatchGAN [66] structure. As shown in Figure 4b, the input image passes through a $\times 4$ kernel size convolution with stride 2 and the LeakyReLu [67] activation function. The result is the input of the next part. The next part consists of three repeating stages of a convolution layer, a normalization module and a LeakyReLu module. The discriminator divides the input image into overlapping regions, discriminates on each region, and averages the results. The local region of the image is distinguished by the designed discriminator, which improves the ability to model the high-frequency components, so the quality of the image is higher than that of the original GAN's discriminator.

3. Results and Discussion

3.1. Results of the Proposed Network

According to the formula of resolution [68], the theoretical resolution of the $4 \times /0.1$ NA objective is 2.75 µm. In order to directly reflect the super-resolution effect of the trained network, we first used PSMs with a diameter of 3 µm as specimens. As shown in Figure 6a, the resolution of the reconstructed phase image can be gradually improved by converting a microscope objective with a larger NA. When using the microscopic images captured by the $40 \times /0.65$ NA objective, the quantitative phase images of PSMs with HR and accurate surface morphology can be recovered. We used the microscopic images obtained by the four different NA objectives as the training dataset to create R_A. The trained network is termed ContransGAN-All. In order to quantitatively evaluate the test results, we used the

scale-invariant feature transform (SIFT) [69] algorithm to obtain the ground truth labels matching the output by the original HR quantitative phase images, and then calculated the structural similarity (SSIM) [70] values and peak signal-to-noise ratio (PSNR) [71] between the output images and the ground truth labels. The test results of the network are shown in Figure 6b and the results show that ContransGAN-All can accurately reconstruct the corresponding HR quantitative phase images for different resolution microscopic images, and the SSIM values between the output images and the corresponding ground truth labels are more than 0.90 and the PSNR values are greater than 31 dB. It preliminarily proves that the proposed network framework can directly generate the corresponding high-quality HR quantitative phase images through the LR microscopic images. Moreover, with the improvement of the resolution of the input images, the quality of the output images gradually improves. This is mainly because for ContransGAN-All, the higher the resolution of the input images, the richer the detailed structure information they contain, and the more features can be extracted to establish the mapping relationship between the LR microscopic images and the HR phase images to constrain the network to achieve better results. In addition, to prove that the ContransGAN is quantitative, we randomly calculated the phase heights of 50 PSMs generated by the ContransGAN-All. As shown in Figure 7, the result shows that the phase heights of these generated PSMs are all in the range of 2.8 μm–3.2 μm, corresponding to a median value of 3.03 μm, which is consistent with expectations (the average relative error is less than 6%).

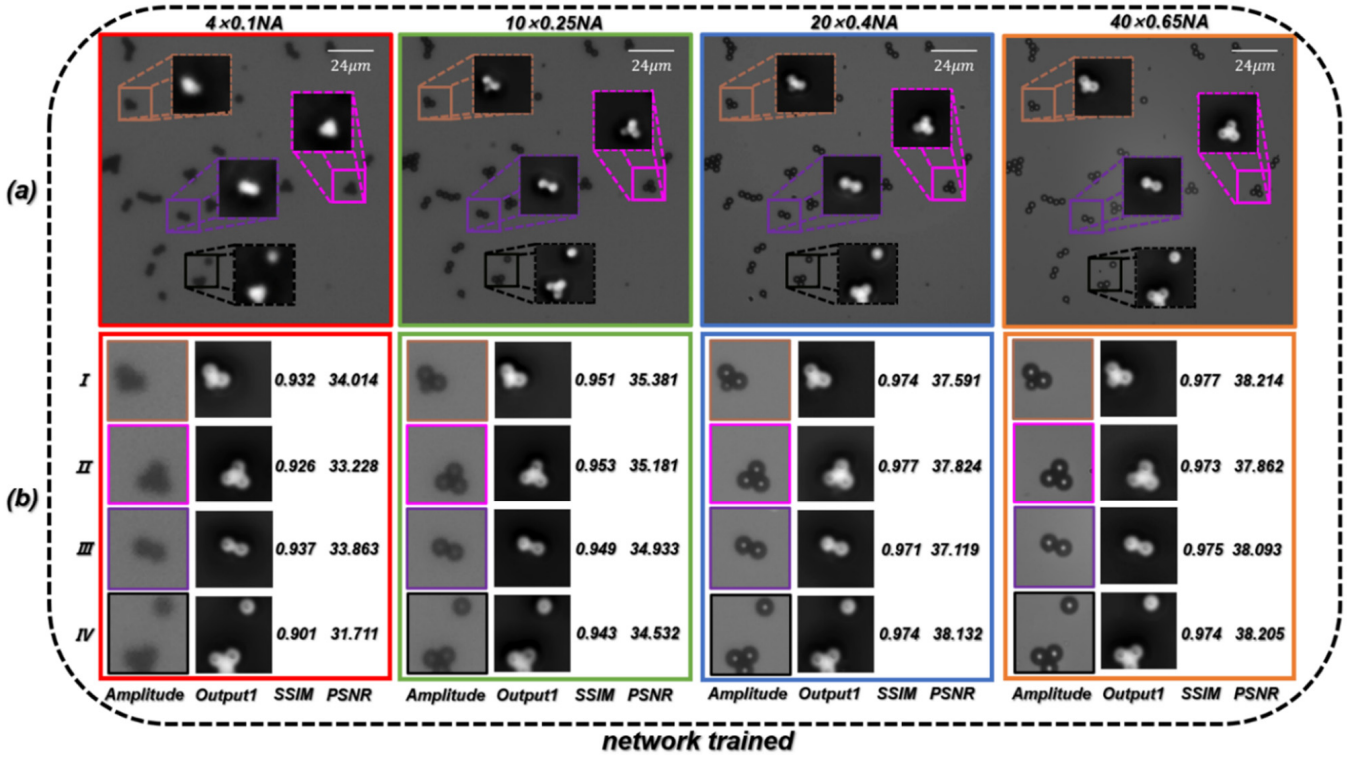

Figure 6. Test results of PSMs by the ContransGAN-All. (**a**) Microscopic images of the same FOV under a different NA objective and the quantitative phase images reconstructed by TIE. (**b**) Results for the corresponding region. Ground truth labels are the quantitative phase images under the 40 × /0.65 NA objective in (**a**); SSIM values and PSNR reflect the quantitative relationship between ground truth labels and Output1.

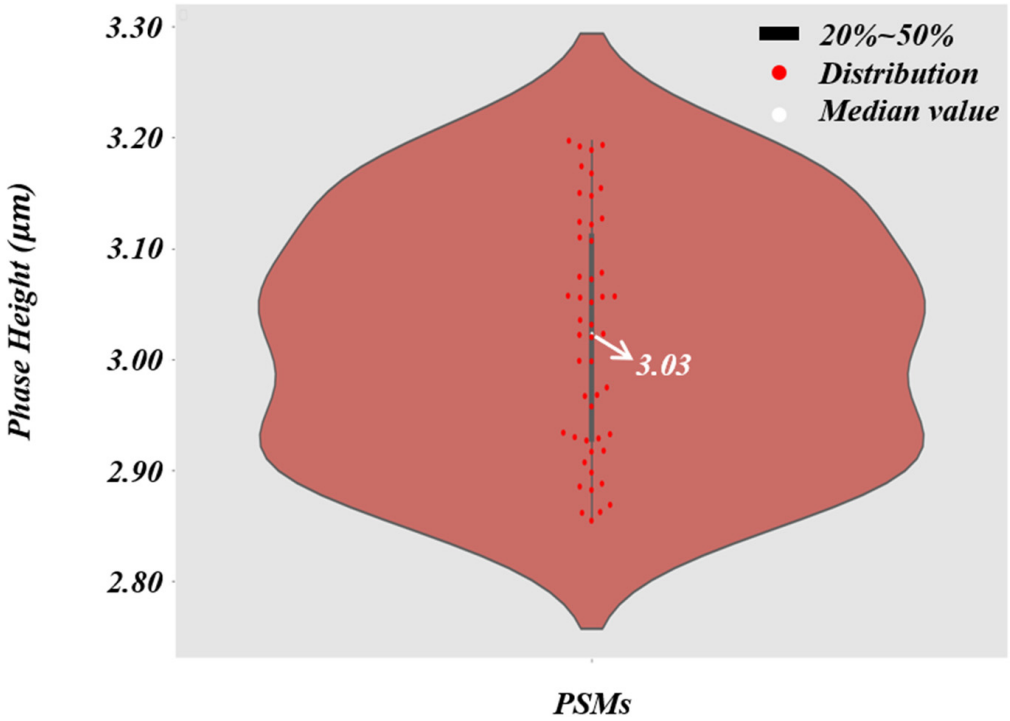

Figure 7. The violin plot of the PSMs' phase height. The thick black line in the middle indicates the interquartile range, the thin black line extending from it represents the 95% confidence interval, the white dot is the median value, and the red spots indicate the distribution of the PSMs' phase heights.

In order to test the HR quantitative phase images' generation quality of the network for biological samples in optical imaging tasks, we used the microscopic images of Hela cells by the 10 × /0.25 NA objective as the training dataset to create R_A. The trained network is termed ContransGAN-Hela. As shown in Figure 8, the SSIM values between the output images and the corresponding ground truth labels of the test results are all above 0.90 and the PSNR values are also greater than 31 dB. Comparing the amplified output images with the ground truth labels, it can be easily found that the ContransGAN-Hela can accurately perceive the high-frequency information in the LR intensity image, establish the mapping relationship between the microscopic images and the quantitative phase images, and give feedback on the output images. Therefore, the proposed ContransGAN is also robust for biological samples, which usually have a complex structure. In order to intuitively compare and analyze the image quality generated by the ContransGAN, we calculated the average SSIM value and PSNR between all generated HR quantitative phase images and the corresponding ground truth labels (Table 1). It can be concluded from the table that the imaging quality of PSMs with a relatively simple structure is better than that of complex biological samples. The main reason is that in the training process of the two networks, in order to ensure the consistency of the network training process, we did not change any parameters. Therefore, only the training images affect the final network performance, so the complexity of the images in the training data determines the quality of the network output image. The standard deviations (std) of SSIM and PSNR of different types of specimens indicate that the more complex the image information is, the more variables there are between the input and output, and the more difficult it is to establish the mapping relationship.

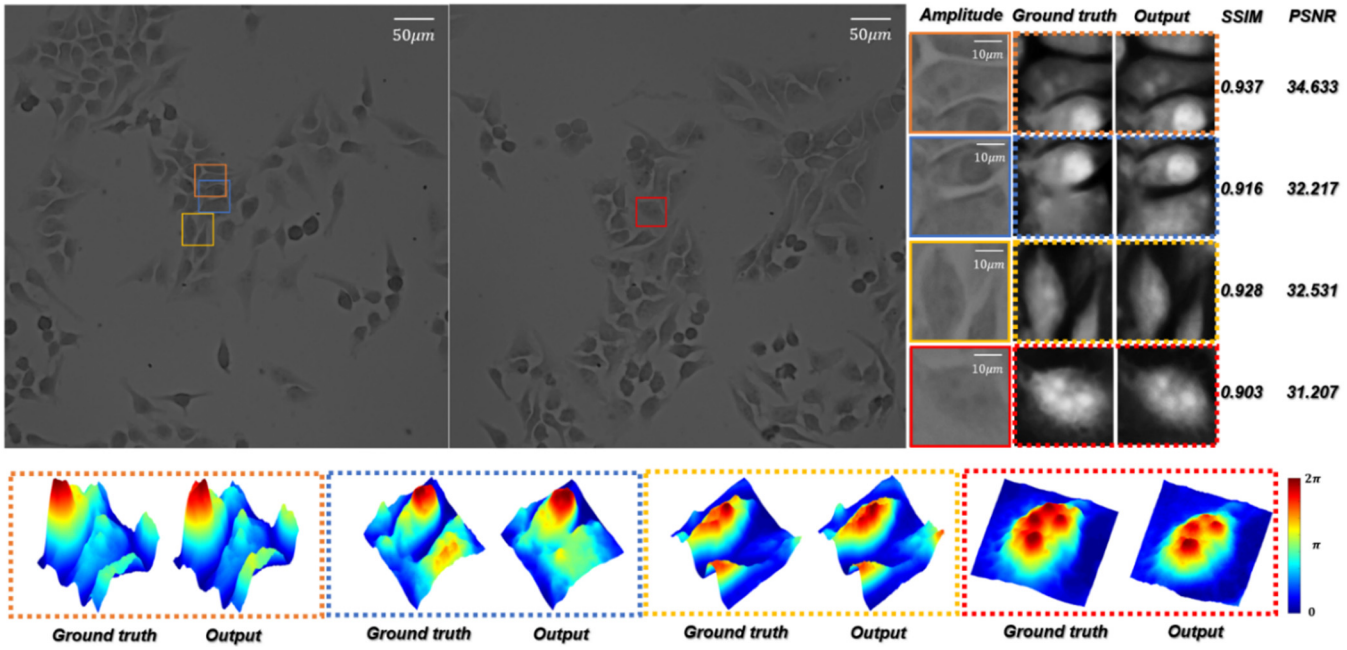

Figure 8. Test results of Hela cells by the ContransGAN-Hela. Amplitude represents the LR microscopic images; ground truth represents the HR quantitative phase images reconstructed by TIE; output represents the output images of the ContransGAN-Hela; SSIM and PSNR reflect the quantitative relationship between the ground truth and output; the dotted frame below is the three-dimensional visual phase distribution in the corresponding FOVs.

Table 1. Quality evaluation index of the test output image of different types of specimens.

Objective	4 × 0.1/NA		10 × 0.25/NA		20 × 0.4/NA		40 × 0.65/NA	
Index	SSIM (std)	PSNR (std)	SSIM (std)	PSNR (std)	SSIM (std)	PSNR (std)	SSIM (std)	PSNR (std)
PSMs	0.933 (±0.0107)	34.032 (±1.0231)	0.952 (±0.0113)	35.214 (±0.9892)	0.975 (±0.0118)	38.963 (±0.9153)	0.976 (±0.0097)	39.331 (±0.8322)
Hela cells	0.898 (±0.0172)	31.751 (±1.2048)	0.922 (±0.0151)	32.171 (±1.1920)	0.939 (±0.0135)	34.751 (±1.0723)	0.943 (±0.0129)	35.380 (±1.0133)

3.2. Comparison of Network Performance

In this paper, compared with the CycleGAN, which uses U-Net [34] as the generator, the difference between the proposed ContransGAN and the CycleGAN is that in order to improve the feature extraction ability of the model, we propose a new generator, termed Contrans. To compare the performance of the modification, we trained the CycleGAN (U-Net as the generator) and S-Transformer (Swin-Transformer as the generator) with the same training dataset. The other hyperparameters, including the learning rate, learning epoch, and batch size, are the same as the ContransGAN-Hela. The results are shown in Figure 9. Compared with the SSIM values and PSNR in Figure 8, the quantitative phase images reconstructed by ContransGAN-Hela are more accurate and have a better image quality. Although CycleGAN and S-Transformer can output a phase image that looks similar in structure, their phase distribution is inaccurate and some areas of the image are distorted. Specifically, in terms of detail generation, ContransGAN-Hela can extract the features of the LR microscopic images as much as possible, so that the final generated quantitative phase images are close to the real distribution. However, CycleGAN and S-Transformer only use CNN or Transformer for feature extraction and cannot fully utilize the information in LR microscopic images, so the generated quantitative phase images lose many detailed features. For further comparison, we plotted the normalized phase distribution curve along the implementation part in the dashed box. It is obvious that the

yellow curve output by ContransGAN-Hela matches the purple curve of the ground-truth label image, although in Figure 9 IV, since there is low contrast in the LR microscopic images, the final result has some deviation from the real distribution, but the phase distribution is almost the same. Relatively speaking, there is a considerable error between the red curve and the curve output by CycleGAN and S-Transformer.

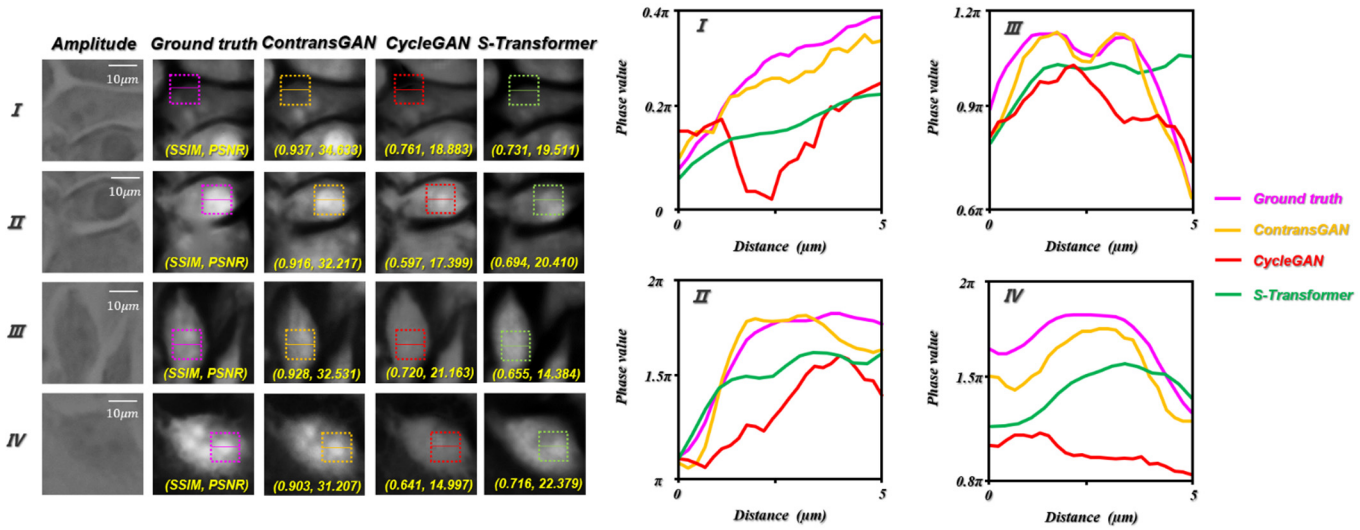

Figure 9. Test results of Hela cells by the CycleGAN-Hela and S-Transformer. SSIM and PSNR reflect the quantitative relationship between the ground truth and the network output phase images; the curve on the right is the phase value curve of the realization part in the dotted line box in the corresponding FOV. I, II, III and IV are different ROIs.

3.3. Generalization Capability and Accuracy Analysis

The above discussion is based on the microscopic images training network obtained under the objective containing different NA. In order to further test the generalization performance of the proposed ContransGAN, we used only the microscopic images of PSMs captured under the 4 × /0.1 NA objective, the microscopic images of PSMs captured under the 10 × /0.25 NA objective, and the microscopic images of Hela cells obtained under the 10 × /0.25 NA objective as training data to train the ContransGAN and obtained three corresponding trained networks. Their corresponding test results are shown in Figure 10. As shown in Figure 10a, the trained network of PSMs captured by the 4 × /0.1 NA objective was tested by the other microscopic images captured with different NA objectives. It is obvious that the proposed network can reconstruct high-quality quantitative phase images with good forward compatibility. However, there is no good performance of the network backward compatibility. As shown in Figure 10b,c, the network that was trained with the microscopic images captured by the 10 × /0.25 NA objective was tested by the microscopic images captured with the smaller NA objective (4 × /0.1 NA). The results show that the network cannot be backward compatible to generate high-quality HR quantitative phase images. It is not difficult to understand that when training with LR microscopic images, the features extracted by the network to establish the mapping relationship between images also exist in the corresponding HR microscopic images with richer information, so the network trained with LR microscopic images can be better reconstructed to generate HR quantitative microscopic images when using HR microscopic images as the network input. Conversely, the network trained by the HR microscopic images with richer image information cannot reflect the corresponding mapping relationship because of the lack of detail features in the LR microscopic images, so the generated images have only approximate morphological features. Especially when imaging biological samples with a relatively complex structure, the network trained by the HR microscopic images is used to generate quantitative phase images of LR microscopic images, which will also be affected by the

noise in the original LR microscopic images. In our work, the HR quantitative phase images corresponding to microscopic images with different resolutions can be quickly generated by the trained ContransGAN, which trained with the microscopic images captured under the 4 × /0.1 NA objective.

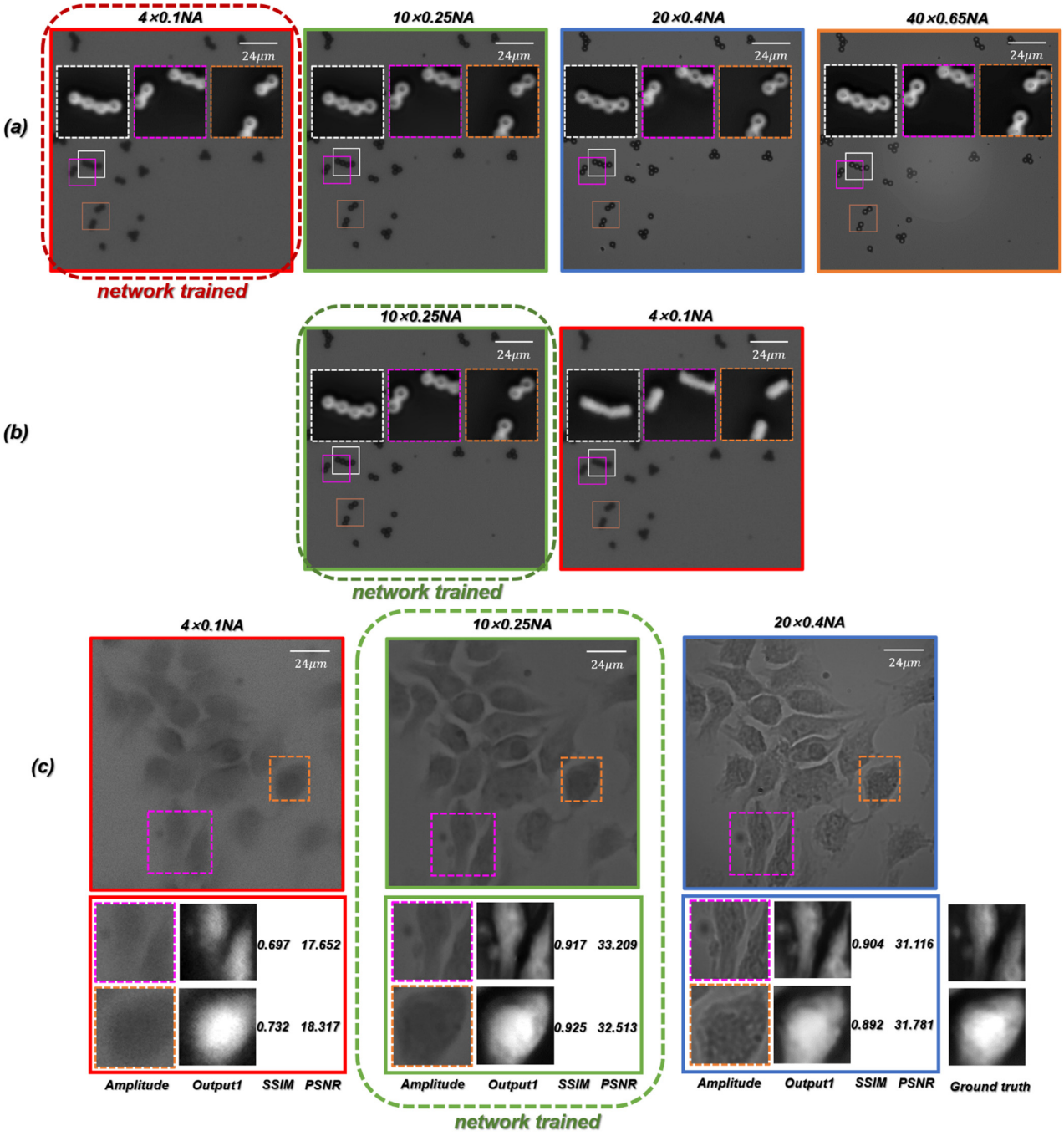

Figure 10. Test results of microscopic images captured by different NA objectives using the ContransGAN trained by microscopic images captured by a certain NA objective. (**a**) The network trained with the microscopic images under the 4 × /0.1 NA objective generated HR phase images of the microscopic images captured by the higher NA objective. (**b**) The network trained with the microscopic images under the 10 × /0.25 NA objective generated HR phase images of the microscopic images captured by the higher NA objective. (**c**) The network trained with the microscopic images under the 10 × /0.25 NA objective generated HR phase images of the microscopic images captured by the other NA objectives.

In practical optical imaging tasks, it is difficult to stay in-focus during long-term observation or imaging. We need to consider the performance of the network if an object is located at distances different from those in the training dataset. In order to further test the generalization performance of the ContransGAN, we trained the network with the out-of-focus microscopic images with an interval of 550 μm between −10 μm and 10 μm from the focal plane under the 10 × /0.25 NA objective. Then, we tested the trained network with the out-of-focus microscopic image captured at any distance from −10 μm to 10 μm under the same objective and compared the generated results with the ground truth labels. As shown in Figure 11a, the results indicated that the trained network is able to correctly obtain the mapping relationship between LR out-of-focus microscopic images and the corresponding in-focus HR quantitative phase images with the values of SSIM versus PSNR being above 0.94 versus 34 dB, respectively. This means that the proposed ContransGAN can perform auto-focusing, phase retrieval, and super-resolution imaging at the same time.

Since phase retrieval through TIE requires capturing of microscopic intensity images at the aperture of the diaphragm S ≈ 0.3, and the acquisition of the LR microscopic images requires constant switching of the aperture of the concentrator, it is natural to consider how well the network can perform if the test microscopic images are captured with different apertures of the concentrator. To test this, we trained the ContransGAN with the microscopic images with different contrast obtained by different apertures of the diaphragm under the 10 × /0.25 NA objective. We tested the trained network with the microscopic images captured under the same objective at any aperture of the concentrator, and also compared the generated results with the ground truth labels. As shown in Figure 11b, the results indicated that the trained network is able to correctly give the mapping relationship between LR microscopic images different contrast and the corresponding HR quantitative phase images, with the values of SSIM versus PSNR being above 0.94 versus 34 dB, respectively. This proves that even if the contrast of the LR microscopic intensity images is not systematic, the proposed ContransGAN is robust and can provide an accurate prediction.

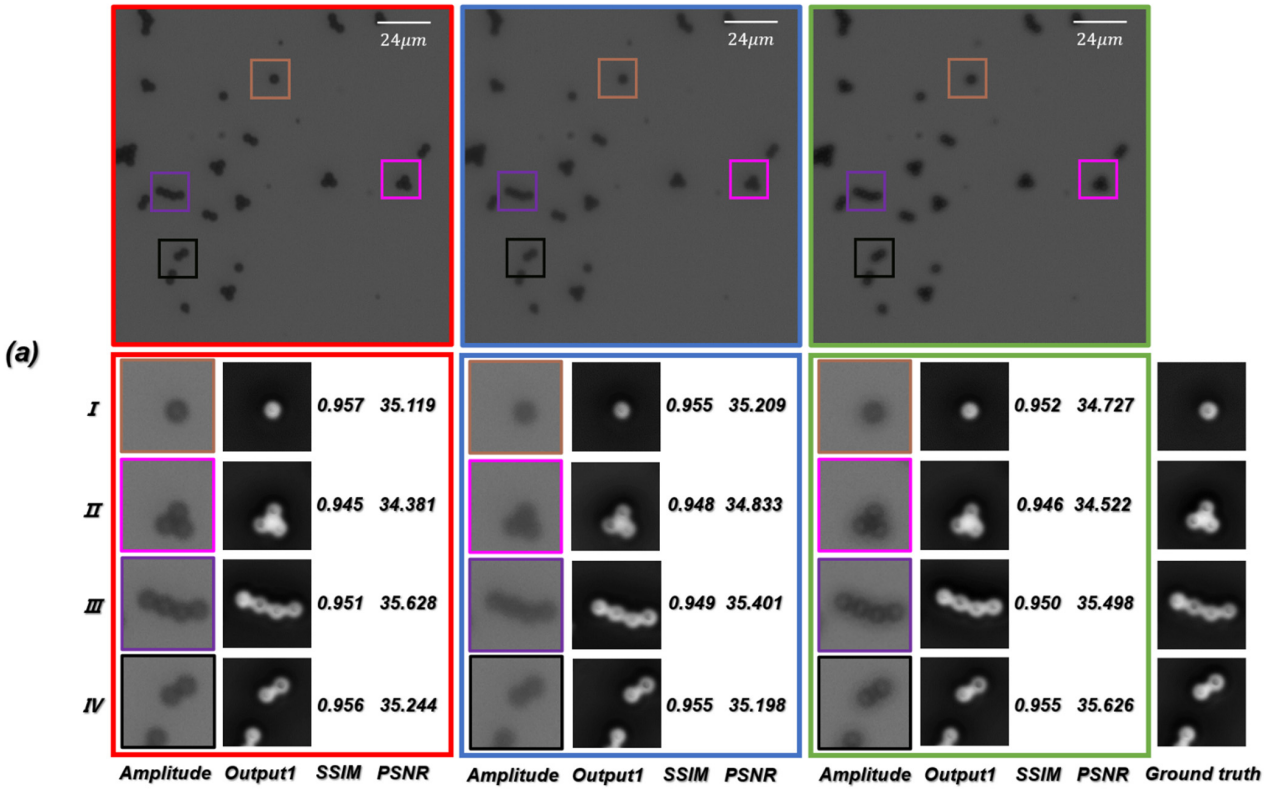

Figure 11. *Cont.*

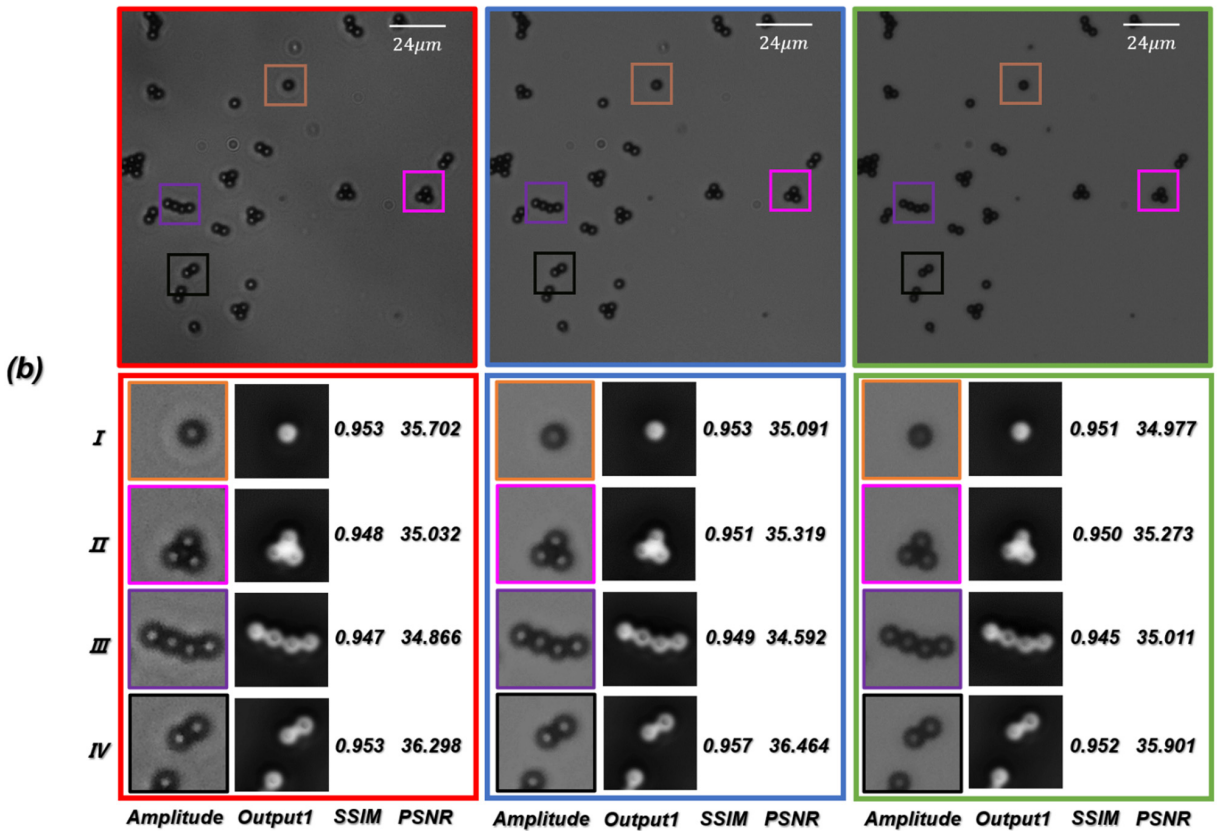

Figure 11. (**a**) Test results of the LR out-of-focus microscopic images. (**b**) Test results of the LR microscopic images with different contrast captured at different apertures of the concentrator.

4. Conclusions

In summary, we introduced a novel end-to-end deep learning-based network framework for super-resolution QPI. It can recover the corresponding HR quantitative phase image from an LR microscopic intensity image captured by a commercial microscope. The framework does not need to train with paired data. Using the proposed Contrans as the generator, the feature extraction ability of the network is greatly enhanced and the information in the LR microscopic images can be fully utilized. After training, the HR quantitative phase information of the object can be quickly extracted from a single LR microscopic intensity image with different resolutions. The feasibility of the proposed framework for QPI was quantitatively proved by experiments. The framework can adapt to various problems in optical microscopic imaging, such as defocus, different resolution, and different contrast, and has strong robustness.

Author Contributions: Conceptualization, H.D.; methodology, H.D.; software, H.D. and C.Y.; validation, F.L., X.C. and S.N.; formal analysis, J.M. and F.L.; investigation, H.D.; resources, C.Y.; data curation, H.D. and R.Y.; visualization, H.D.; writing—original draft, H.D.; writing—review and editing, H.D. and C.Y.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) under Grant 62175112, Grant 61975081 and Grant 62105156; Postgraduate Research &Practice Innovation Program of Jiangsu Province (No. KYCX22_1540).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Acknowledgments: We would like to express our gratitude to all members of our laboratories for the helpful discussions and support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Betzig, E.; Patterson, G.H.; Sougrat, R.; Lindwasser, O.W.; Olenych, S.; Bonifacino, J.S.; Michael, W.D.; Jennifer, L.S.; Hess, H.F. Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science* **2006**, *313*, 1642–1645. [CrossRef] [PubMed]
2. Heintzmann, R.; Gustafsson, M.G.L. Subdiffraction resolution in continuous samples. *Nat. Photonics* **2009**, *3*, 362–364. [CrossRef]
3. Gao, P.; Yuan, C. Resolution enhancement of digital holographic microscopy via synthetic aperture: A review. *Light Adv. Manuf.* **2022**, *3*, 105–120. [CrossRef]
4. Meng, Z.; Pedrini, G.; Lv, X.; Ma, J.; Nie, S.; Yuan, C. DL-SI-DHM: A deep network generating the high-resolution phase and amplitude images from wide-field images. *Opt. Express* **2021**, *29*, 19247–19261. [CrossRef]
5. Wang, Z.; Xie, Y.; Ji, S. Global voxel transformer networks for augmented microscopy. *Nat. Mach. Intell.* **2021**, *3*, 161–171. [CrossRef]
6. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef]
7. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
8. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
9. Chitchian, S.; Mayer, M.A.; Boretsky, A.; Van Kuijk, F.J.; Motamedi, M. Retinal optical coherence tomography image enhancement via shrinkage denoising using double-density dual-tree complex wavelet transform. *J. Biomed. Opt.* **2012**, *17*, 116009. [CrossRef]
10. Huang, Y.; Lu, Z.; Shao, Z.; Ran, M.; Zhou, J.; Fang, L.; Zhang, Y. Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network. *Opt. Express* **2019**, *27*, 12289–12307. [CrossRef]
11. Rahmani, B.; Loterie, D.; Konstantinou, G.; Psaltis, D.; Moser, C. Multimode optical fiber transmission with a deep learning network. *Light Sci. Appl.* **2018**, *7*, 1–11. [CrossRef]
12. He, Y.; Wang, G.; Dong, G.; Psaltis, D.; Moser, C. Ghost imaging based on deep learning. *Sci. Rep.* **2018**, *8*, 1–7. [CrossRef]
13. Li, Y.; Xue, Y.; Tian, L. Deep speckle correlation: A deep learning approach toward scalable imaging through scattering media. *Optica* **2018**, *5*, 1181–1190. [CrossRef]
14. Goy, A.; Arthur, K.; Li, S.; Barbastathis, G. Low photon count phase retrieval using deep learning. *Phys. Rev. Lett.* **2018**, *121*, 243902. [CrossRef]
15. Rivenson, Y.; Zhang, Y.; Günaydın, H.; Teng, D.; Ozcan, A. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light Sci. Appl.* **2018**, *7*, 17141. [CrossRef]
16. Wu, Y.; Rivenson, Y.; Zhang, Y.; Wei, Z.; Günaydin, H.; Lin, X.; Ozcan, A. Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery. *Optica* **2018**, *5*, 704–710. [CrossRef]
17. Sinha, A.; Lee, J.; Li, S.; Barbastathis, G. Lensless computational imaging through deep learning. *Optica* **2017**, *4*, 1117–1125. [CrossRef]
18. Castaneda, R.; Trujillo, C.; Doblas, A. Video-Rate Quantitative Phase Imaging Using a Digital Holographic Microscope and a Generative Adversarial Network. *Sensors* **2021**, *21*, 8021. [CrossRef]
19. Liu, T.; De Haan, K.; Rivenson, Y.; Wei, Z.; Zeng, X.; Zhang, Y.; Ozcan, A. Deep learning-based super-resolution in coherent imaging systems. *Sci. Rep.* **2019**, *9*, 1–13. [CrossRef]
20. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.H.; Liao, Q. Deep learning for single image super-resolution: A brief review. *IEEE Trans. Multimed.* **2019**, *21*, 3106–3121. [CrossRef]
21. Wang, H.; Rivenson, Y.; Jin, Y.; Wei, Z.; Gao, R.; Günaydın, H.; Bentolila, L.A.; Kural, C.; Ozcan, A. Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nat. Methods* **2019**, *16*, 103–110. [CrossRef]
22. Jin, L.; Liu, B.; Zhao, F.; Hahn, S.; Dong, B.; Song, R.; Elston, T.C.; Xu, Y.; Hahn, K.M. Deep learning enables structured illumination microscopy with low light levels and enhanced speed. *Nat. Commun.* **2020**, *11*, 1–7. [CrossRef]
23. Xypakis, E.; Gosti, G.; Giordani, T.; Santagati, R.; Ruocco, G.; Leonetti, M. Deep learning for blind structured illumination microscopy. *Sci. Rep.* **2022**, *12*, 8623. [CrossRef]
24. Dardikman, G.; Shaked, N.T. Phase unwrapping using residual neural networks. In *Computational Optical Sensing and Imaging*; Optical Society of America: Orlando, FL, USA, 2018.
25. Wang, K.; Li, Y.; Kemao, Q.; Di, J.; Zhao, J. One-step robust deep learning phase unwrapping. *Opt. Express* **2019**, *27*, 15100–15115. [CrossRef]
26. Yin, W.; Chen, Q.; Feng, S.; Tao, T.; Huang, L.; Trusiak, M.; Asundi, A.; Zuo, C. Temporal phase unwrapping using deep learning. *Sci. Rep.* **2019**, *9*, 1–12. [CrossRef]
27. Huang, W.; Mei, X.; Wang, Y.; Fan, Z.; Chen, C.; Jiang, G. Two-dimensional phase unwrapping by a high-resolution deep learning network. *Measurement* **2022**, 111566. [CrossRef]

28. Göröcs, Z.; Tamamitsu, M.; Bianco, V.; Wolf, P.; Roy, S.; Shindo, K.; Yanny, K.; Wu, Y.; Koydemir, H.C.; Rivenson, Y.; et al. A deep learning-enabled portable imaging flow cytometer for cost-effective, high-throughput, and label-free analysis of natural water samples. *Light Sci. Appl.* **2018**, *7*, 1–12. [CrossRef]
29. Rivenson, Y.; Liu, T.; Wei, Z.; Zhang, Y.; De Haan, K.; Ozcan, A. PhaseStain: The digital staining of label-free quantitative phase microscopy images using deep learning. *Light Sci. Appl.* **2019**, *8*, 1–11. [CrossRef]
30. Nygate, Y.N.; Levi, M.; Mirsky, S.K.; Turko, N.A.; Rubin, M.; Barnea, I.; Dardikman-Yoffe, G.; Haifler, M.; Shaked, N.T. Holographic virtual staining of individual biological cells. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9223–9231. [CrossRef]
31. Bian, Y.; Jiang, Y.; Deng, W.; Shen, R.; Shen, H.; Kuang, C. Deep learning virtual Zernike phase contrast imaging for singlet microscopy. *AIP Adv.* **2021**, *11*, 065311. [CrossRef]
32. Wu, Y.; Luo, Y.; Chaudhari, G.; Rivenson, Y.; Calis, A.; De Haan, K.; Ozcan, A. Bright-field holography: Cross-modality deep learning enables snapshot 3D imaging with bright-field contrast using a single hologram. *Light Sci. Appl.* **2019**, *8*, 1–7. [CrossRef] [PubMed]
33. Zomet, A.; Peleg, S. Multi-sensor super-resolution. In Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV 2002), Orlando, FL, USA, 3–4 December 2002; pp. 27–31.
34. Glasner, D.; Bagon, S.; Irani, M. Super-resolution from a single image. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 349–356.
35. Zhang, H.; Fang, C.; Xie, X.; Yang, Y.; Mei, W.; Jin, D.; Fei, P. High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network. *Biomed. Opt. Express* **2019**, *10*, 1044–1063. [CrossRef] [PubMed]
36. Chen, S.; Han, Z.; Dai, E.; Jia, X.; Liu, Z.; Xing, L.; Zou, X.; Xu, C.; Liu, J.; Tian, Q. Unsupervised image super-resolution with an indirect supervised path. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 19–25 June 2021; pp. 468–469.
37. Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; Lin, L. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 701–710.
38. Lugmayr, A.; Danelljan, M.; Timofte, R. Unsupervised learning for real-world super-resolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27 October 2019; pp. 3408–3416.
39. Terbe, D.; Orzó, L.; Zarándy, Á. Deep-learning-based bright-field image generation from a single hologram using an unpaired dataset. *Opt. Lett.* **2021**, *46*, 5567–5570. [CrossRef] [PubMed]
40. Zhang, Y.; Noack, M.A.; Vagovic, P.; Fezzaa, K.; Garcia-Moreno, F.; Ritschel, T.; Villanueva-Perez, P. PhaseGAN: A deep-learning phase-retrieval approach for unpaired datasets. *Opt. Express* **2021**, *29*, 19593–19604. [CrossRef]
41. Ding, H.; Li, F.; Meng, Z.; Feng, S.; Ma, J.; Nie, S.; Yuan, C. Auto-focusing and quantitative phase imaging using deep learning for the incoherent illumination microscopy system. *Opt. Express* **2021**, *29*, 26385–26403. [CrossRef]
42. Ptak, R. The frontoparietal attention network of the human brain: Action, saliency, and a priority map of the environment. *Neurosci.* **2012**, *18*, 502–515. [CrossRef]
43. Huang, L.; Pashler, H. A Boolean map theory of visual attention. *Psychol. Rev.* **2007**, *114*, 599. [CrossRef]
44. Chen, Y.; Liu, L.; Phonevilay, V.; Gu, K.; Xia, R.; Xie, J.; Zhang, Q.; Yang, K. Image super-resolution reconstruction based on feature map attention mechanism. *Appl. Intell.* **2021**, *51*, 4367–4380. [CrossRef]
45. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
46. Wang, K.; Di, J.; Li, Y.; Ren, Z.; Kemao, Q.; Zhao, J. Transport of intensity equation from a single intensity image via deep learning. *Opt. Lasers Eng.* **2020**, *134*, 106233. [CrossRef]
47. Paganin, D.; Nugent, K.A. Noninterferometric phase imaging with partially coherent light. *Phys. Rev. Lett.* **1998**, *80*, 2586. [CrossRef]
48. Gureyev, T.E.; Nugent, K.A. Rapid quantitative phase imaging using the transport of intensity equation. *Opt. Commun.* **1997**, *133*, 339–346. [CrossRef]
49. Allen, L.J.; Oxley, M.P. Phase retrieval from series of images obtained by defocus variation. *Opt. Commun.* **2001**, *199*, 65–75. [CrossRef]
50. Teague, M.R. Deterministic phase retrieval: A Green's function solution. *JOSA* **1983**, *73*, 1434–1441. [CrossRef]
51. Rong, L.; Wang, S.; Wang, D.; Tan, F.; Zhang, Y.; Zhao, J.; Wang, Y. Transport of intensity equation-based terahertz lensless full-field phase imaging. *Opt. Lett.* **2021**, *46*, 5846–5849. [CrossRef]
52. Zuo, C.; Li, J.; Sun, J.; Fan, Y.; Zhang, J.; Lu, L.; Zhang, R.; Wang, B.; Huang, L.; Chen, Q. Transport of intensity equation: A tutorial. *Opt. Lasers Eng.* **2020**, *135*, 106187. [CrossRef]
53. Zhang, J.; Chen, Q.; Sun, J.; Tian, L.; Zuo, C. On a universal solution to the transport-of-intensity equation. *Opt. Lett.* **2020**, *45*, 3649–3652. [CrossRef]
54. Zuo, C.; Sun, J.; Li, J.; Zhang, J.; Asundi, A.; Chen, Q. High-resolution transport-of-intensity quantitative phase microscopy with annular illumination. *Sci. Rep.* **2017**, *7*, 1–22. [CrossRef]
55. Zuo, C.; Chen, Q.; Qu, W.; Asundi, A. High-speed transport-of-intensity phase microscopy with an electrically tunable lens. *Opt. Express* **2013**, *21*, 24060–24075. [CrossRef]

56. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
57. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *60*, 84–90. [CrossRef]
58. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010. [CrossRef]
59. Rodriguez, P.; Wiles, J.; Elman, J.L. A recurrent neural network that learns to count. *Connect. Sci.* **1999**, *11*, 5–40. [CrossRef]
60. Girosi, F.; Jones, M.; Poggio, T. Regularization theory and neural networks architectures. *Neural Comput.* **1995**, *7*, 219–269. [CrossRef]
61. Tang, H.; Xue, J.; Han, J. A Method of Multi-Scale Forward Attention Model for Speech Recognition. *Acta Electonica Sin.* **2020**, *48*, 1255.
62. Wang, W.; Shen, J.; Yu, Y.; Ma, K.L. Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Trans. Vis. Comput. Graph.* **2016**, *23*, 2014–2027. [CrossRef]
63. Wang, M.; Lu, S.; Zhu, D.; Lin, J.; Wang, Z. A high-speed and low-complexity architecture for softmax function in deep learning. In Proceedings of the 2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Chengdu, China, 26–30 October 2018; pp. 223–226.
64. Gardner, M.W.; Dorling, S.R. Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmos. Environ.* **1998**, *32*, 2627–2636. [CrossRef]
65. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
66. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
67. Zhang, X.; Zou, Y.; Shi, W. Dilated convolution neural network with LeakyReLU for environmental sound classification. In Proceedings of the 2017 22nd International Conference on Digital Signal Processing (DSP), London, UK, 23–25 August 2017; pp. 1–5.
68. Heintzmann, R.; Ficz, G. Breaking the resolution limit in light microscopy. *Brief. Funct. Genom.* **2006**, *5*, 289–301. [CrossRef] [PubMed]
69. Lindeberg, T. Scale Invariant Feature Transform. *Scholarpedia* **2012**, *7*, 10491. [CrossRef]
70. Wang, Z.; Chen, J.; Hoi, S.C.H. Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3365–3387. [CrossRef] [PubMed]
71. Winkler, S.; Mohandas, P. The evolution of video quality measurement: From PSNR to hybrid metrics. *IEEE Trans. Broadcasting* **2008**, *54*, 660–668. [CrossRef]