nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-020-00283-x

Global voxel transformer networks for augmented microscopy

Zhengyang Wang[1,2], Yaochen Xie[1,2] and Shuiwang Ji[1]✉

Advances in deep learning have led to remarkable success in augmented microscopy, enabling us to obtain high-quality microscope images without using expensive microscopy hardware and sample preparation techniques. Current deep learning models for augmented microscopy are mostly U-Net-based neural networks, thus sharing certain drawbacks that limit the performance. In particular, U-Nets are composed of local operators only and lack dynamic non-local information aggregation. In this work, we introduce global voxel transformer networks (GVTNets), a deep learning tool for augmented microscopy that overcomes intrinsic limitations of the current U-Net-based models and achieves improved performance. GVTNets are built on global voxel transformer operators, which are able to aggregate global information, as opposed to local operators like convolutions. We apply the proposed methods on existing datasets for three different augmented microscopy tasks under various settings.

In modern biology and life science, augmented microscopy attempts to improve the quality of microscope images to extract more information, such as introducing fluorescent labels, increasing the signal-to-noise ratio (SNR) and performing super-resolution. Previous advances in microscopy have allowed the imaging of biological processes with higher and higher quality[1–8]. However, these advanced augmented-microscopy techniques usually lead to high costs in terms of the microscopy hardware and experimental conditions, resulting in many practical limitations. In addition, there are specific concerns when recording processes in live cells, tissues and organisms: the imaging process should neither notably affect the biological processes nor substantially harm the sample's health. For example, assessing phototoxicity is a major problem in live fluorescence imaging[9,10]. With these restrictions, high-quality microscope images are hard, expensive and slow to obtain. While some microscope images, like transmitted-light images[11], can be collected at relatively low cost, they are not sufficient to provide accurate statistics and correct insights without augmentation. As a result, modern biologists and life scientists usually have to deal with the trade-offs between the quality of microscope images and the restrictions in the process of collecting them[12–14].

In recent years, the development of deep learning[15] has pushed the boundaries of such trade-offs by enabling fast and inexpensive microscopy augmentation using computational approaches[16–18]. The augmented microscopy task is formulated as a biological image transformation problem in deep learning. Specifically, models composed of multi-layer artificial neural networks take low-quality microscope images as inputs and transform them into high-quality ones through computational processes. Deep learning has led to success in various augmented microscopy applications, such as prediction of fluorescence signals from transmitted-light images[8,19–25], virtual refocusing of fluorescence images[26], content-aware image restoration[27], fluorescence image super-resolution[28] and axial under-sampling mitigation[29].

Among these successful applications of deep learning, U-Net-based neural networks have been the mainstream models. U-Net was first proposed for 2D electron microscopy image segmentation[30] and later extended to other biological image transformation tasks, including cell detection and quantification[31]. In the field of augmented microscopy, most deep learning models directly apply U-Net-based neural networks by only changing the loss functions for training[21–23,25,27]. In general, the U-Net is an encoder–decoder framework of neural network architectures for image transformation. It consists of a down-sampling path to capture multi-scale and multi-resolution contextual information, and a corresponding up-sampling path to enable precise voxel-wise predictions. Recent studies have enhanced the U-Net by incorporating residual blocks[32–35] and supporting 3D image transformation[36].

Despite the success of these U-Net-based neural networks for augmented microscopy, we observe three intrinsic limitations caused by the fact that they implement the encoder–decoder path by stacking local operators like convolutions and transposed convolutions with small kernels. First, in local operators, the receptive field (RF) size of an output unit, determined by the kernel size, is usually small and does not aggregate information from the entire input (Fig. 1a). While stacking these local operators increases the RF size for the final output units[37], the RF size is still fixed given a specific neural network architecture. Each output unit follows a local path through the network and only has access to the information within its RF on the input image. Given a large input image, the network has to go deeper with more down-sampling and up-sampling operators to ensure each output unit received information from the entire input image. Such an approach is not efficient in terms of the amount of training parameters and computational expenses. In addition, the local path tends to focus on local dependencies among units and fails to capture long-range dependencies[38,39], which are crucial for accuracy and consistency in biological image transformation. Second, the fixed-size RF limits the model's inference performance as well. The U-Net is usually trained with small patches of paired images (Fig. 1b), where cutting large images into small patches increases the amount of training data and stabilizes the training process by allowing large batch sizes[40]. As the U-Net produces the output of the same spatial size as the input, it is common to feed in the entire image or patches of much larger spatial sizes than the training patches during the prediction procedure (Fig. 1b, Supplementary Fig. 1), in order to speed

[1]Texas A&M University, Department of Computer Science and Engineering, College Station, TX, USA. [2]These authors contributed equally: Zhengyang Wang, Yaochen Xie. ✉e-mail: sji@tamu.edu

up the inference[25,27,30,31]. However, with the fixed-size RF, the model fails to take advantage of the knowledge from the entire input if the spatial size of the input is larger than that of RF, preventing potential inference performance boost. Third, all the local operators work with kernels whose weights are fixed after the training process, which means the importance of an input unit to an output unit is determined and not input dependent during the inference stage (Fig. 1a). This property is helpful in detecting and extracting local patterns[15]. However, the model is supposed to be able to selectively use or ignore extracted information when transforming different input images, raising the need of operators that support input-dependent weights.

In this work, we argue that all three limitations can be addressed by introducing the attention operator[38] into U-Net-based neural networks. To demonstrate this point, we compare the attention operator with a typical local operator: convolution (Fig. 1a). There are essential differences between the convolution and the attention operator. On one hand, the convolution has a local RF determined by its kernel, where each output unit receives information from a local area of input units. Meanwhile, note that the kernel weights are fixed after training. In other words, the weights do not depend on inputs during the inference. On the other hand, the attention operator computes each output unit as a weighted sum of all input units, where the weights are obtained through interactions between different representations of the inputs (Methods). As a result, the attention operator is a non-local operator with a global receptive field, which can potentially overcome the first two limitations. In addition, the weights in the attention operator are input dependent, addressing the third limitation.

Based on this insight, we build a family of non-local operators upon the attention operator, namely global voxel transformer operators (GVTOs) (Fig. 1c, Methods). GVTOs organically combine local and non-local operators (Supplementary Fig. 11) and can capture both local and long-range dependencies. In particular, GVTOs extend the attention operator to serve as flexible building blocks in the U-Net framework. Specifically, we develop GVTOs to support not only size-preserving, but also down-sampling and up-sampling tensor processing, which covers all kinds of operators in the U-Net framework. It is worth noting that, while GVTOs are designed for the U-Net framework, they can also be used in other kinds of networks as well.

With GVTOs, we propose GVTNets (Fig. 1c, Methods), an advanced deep learning tool for augmented microscopy, to address the limitations and improve current U-Net-based neural networks. GVTNets follow the same encoder–decoder framework as the U-Net while using GVTOs instead of local operators only. To be concrete, we force GVTNets to connect the down-sampling and up-sampling paths using the size-preserving GVTO at the bottom level, which separates GVTNets from the U-Net. In addition, we allow users to flexibly use more GVTOs to replace local operators in the U-Net framework.

In the following, we (1) demonstrate the power of the basic GVTNets where only one size-preserving GVTO at the bottom level is applied, (2) show the effectiveness of employing more GVTOs in GVTNets and point out how GVTNets improve the inference performance, (3) explore the use of GVTOs in more complex and composite models, and (4) investigate the generalization ability of GVTNets. All the experiments are conducted on publicly available datasets for augmented microscopy[18,25,27].

Results

GVTNets training and inference. GVTNets are trained end-to-end under a supervised learning setting through back-propagation[15] (Methods). While the model aims at augmenting microscopy computationally, it still requires a relatively small amount of augmented microscopy images to be collected for training. Specifically, the training data are registered pairs of biological images before and after augmentation. Once trained, the model can be used to augment microscope images in silico, without involving expensive microscopy hardware or techniques. Following previous studies, we crop the training images into patches of smaller spatial size to train GVTNets. However, during the inference procedure, we feed in the entire image for prediction (Fig. 1b). Note that GVTNets are able to handle inputs of any spatial size, and in particular, tend to perform better given larger inputs due to the ability of utilizing global information from the entire input. The power of GVTNets come from the use of GVTOs, which are inherently different from local operators as well as the fully connected (FC) layers in deep learning (Methods).

Label-free prediction of 3D fluorescence images from transmitted-light microscopy. First, we ask whether basic GVTNets achieve improved performance over U-Net-based neural networks. A basic GVTNet differs from the U-Net only at its bottom level, by using a size-preserving GVTO instead of convolutions. This replacement is crucial, giving each output unit access to information from the entire input image, regardless of the spatial size. We apply a basic GVTNet on the public dataset from Ounkomol et al.[25], where the task is label-free prediction of 3D fluorescence images from transmitted-light microscopy (Fig. 2a).

The dataset is composed of 13 datasets corresponding to 13 different subcellular structures. All the images in the datasets are spatially registered and obtained from a database of images produced by the Allen Institute for Cell Science's microscopy pipeline[25]. The training and testing splits are provided by Ounkomol et al.[25] and available in our published code. For each structure, the training data are 30 spatially registered pairs of 3D transmitted-light images and ground-truth fluorescence images. The number of testing images is 18 for the cell membrane, 10 for the differential interference contrast (DIC) nuclear envelope and 20 for the others.

We use the model proposed by Ounkomol et al.[25], which is the current state-of-the-art model, as the baseline on the 13 datasets. The baseline model is a U-Net-based neural network of depth five containing 23,280,769 training parameters, while the basic GVTNet that we used is of depth four containing 6,172,225 training parameters (Supplementary Fig. 2). As a result, the basic GVTNet has only 26.5% of training parameters of the baseline model. In addition, the computation speed becomes faster; that is, the GVTNet takes 0.4 s to make a prediction for one 3D image while the U-Net takes 1 s (ref. [25]).

We quantify the model performance by computing the Pearson correlation coefficient on the testing data (Methods). On all of the 13 datasets, our basic GVTNet consistently outperforms the U-Net baseline. We perform one-tailed, paired *t*-tests and obtain *P* values smaller than 0.05 for all datasets, showing the improvements are statistically significant (Fig. 2b). The visualization of predictions indicates that the GVTNet captures more details than the U-Net baseline due to the access to more information, and is able to use global information to avoid local inconsistency (Fig. 2c). The quantitative testing results in terms of Pearson correlation coefficients are provided in Supplementary Table 1. Examples of predictions on testing images for all 13 structures can be found in Supplementary Fig. 3. These experimental results indicate the effectiveness of only one size-preserving GVTO and the resultant basic GVTNets.

We note that both GVTNets and the U-Net baselines perform poorly on the datasets corresponding to Golgi apparatus and desmosome subcellular structures. According to Ounkomol et al.[25], a possible explanation is that the correlations between the input transmitted-light microscope images and the target fluorescence images are weak in these two datasets. As most supervised deep learning methods models try to capture the correlations between inputs and outputs during training, the inference performance

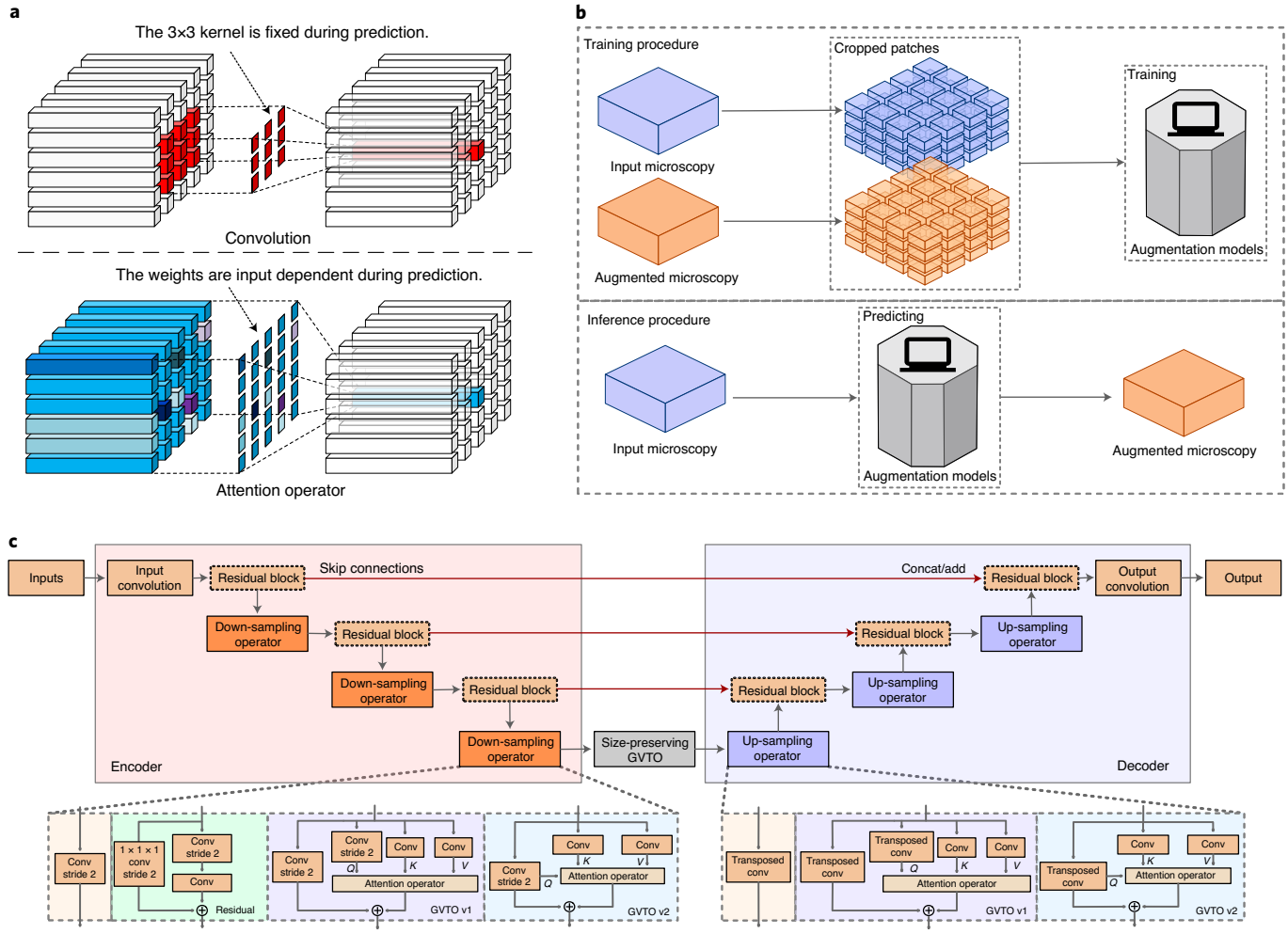

Fig. 1 | GVTNets architecture, training and inference. a, Comparisons between a 3×3 convolution and the attention operator in terms of RF and working mechanism during the inference procedure. The convolution, a typical local operator, has a fixed-size RF and fixed weights after training. On the contrary, the attention operator always allows a global RF and input-dependent weights during prediction. GVTOs, the key components of GVTNets, are built upon the attention operator. **b**, Training and inference of GVTNets. During the training procedure, registered pairs of microscope images before and after augmentation are collected and cut into small patches. During the inference procedure, the entire image is fed into the model to obtain the augmented output. **c**, Architecture of GVTNets. A GVTNet of depth four. The use of GVTOs differs from U-Net in GVTNet. The GVTNet fixes one size-preserving GVTO at the bottom level and allows optional GVTOs as down-sampling and up-sampling operators. Detailed descriptions of GVTNets and GVTOs are provided in the Methods.

could be poor if the correlations are weak. Nevertheless, a global view is helpful, as more correlations are considered when making the prediction for each voxel. And GVTNets indeed improve the performance on these datasets. In addition, we also find that the amount of improvements brought by GVTNets vary for different datasets. We suspect that this is due to the characteristics of different subcellular structures. For example, the shapes are sparser for actomyosin bundles and tight junctions (Supplementary Fig. 3). It may be more important to capture the correlations among distant voxels when predicting these subcellular structures, where GVTNets achieve more improvements.

Content-aware 3D image denoising. Next, we explore the potential of GVTNets by applying more GVTOs. Specifically, we apply GVTNets with both size-preserving and up-sampling GVTOs on two independent content-aware 3D image denoising tasks (Fig. 3a); namely, improving the SNR of live-cell imaging of planarian *Schmidtea mediterranea* and developing *Tribolium castaneum* embryos.

The datasets, published by Weigert et al.[27], contain pairs of 3D low-SNR images and ground-truth high-SNR images for training and testing. The training data are provided in the form of 17,005 and 14,725 small, cropped patches of size 64×64×16 for planaria and *Tribolium* datasets (hereafter 'Planaria' and 'Tribolium' datasets), while the testing data are 20 testing images of size 1,024×1,024×95 and 6 testing images of average size around 700×700×45 for the two datasets, respectively. In addition, the testing data come with three image conditions referring to three different SNR levels, leading to three degrees of denoising difficulty (Fig. 3b). Here, the image conditions refer to the laser power and exposure time during image collection[27]. Generally, low laser power and short exposure time lead to low SNR levels. Concretely, in the Planaria dataset, four different laser-power/exposure-time conditions are used: GT (ground truth) and C1–C3; specifically 2.31 mW, 30 ms (GT); 0.12 mW, 20 ms (C1); 0.12 mW, 10 ms (C2); and 0.05 mW, 10 ms (C3). Similarly, in the Tribolium dataset, four different laser-power imaging conditions are used: GT and C1–C3; specifically, 20 mW (GT); 0.5 mW (C1); 0.2 mW (C2); and 0.1 mW (C3). As a result, each ground-truth

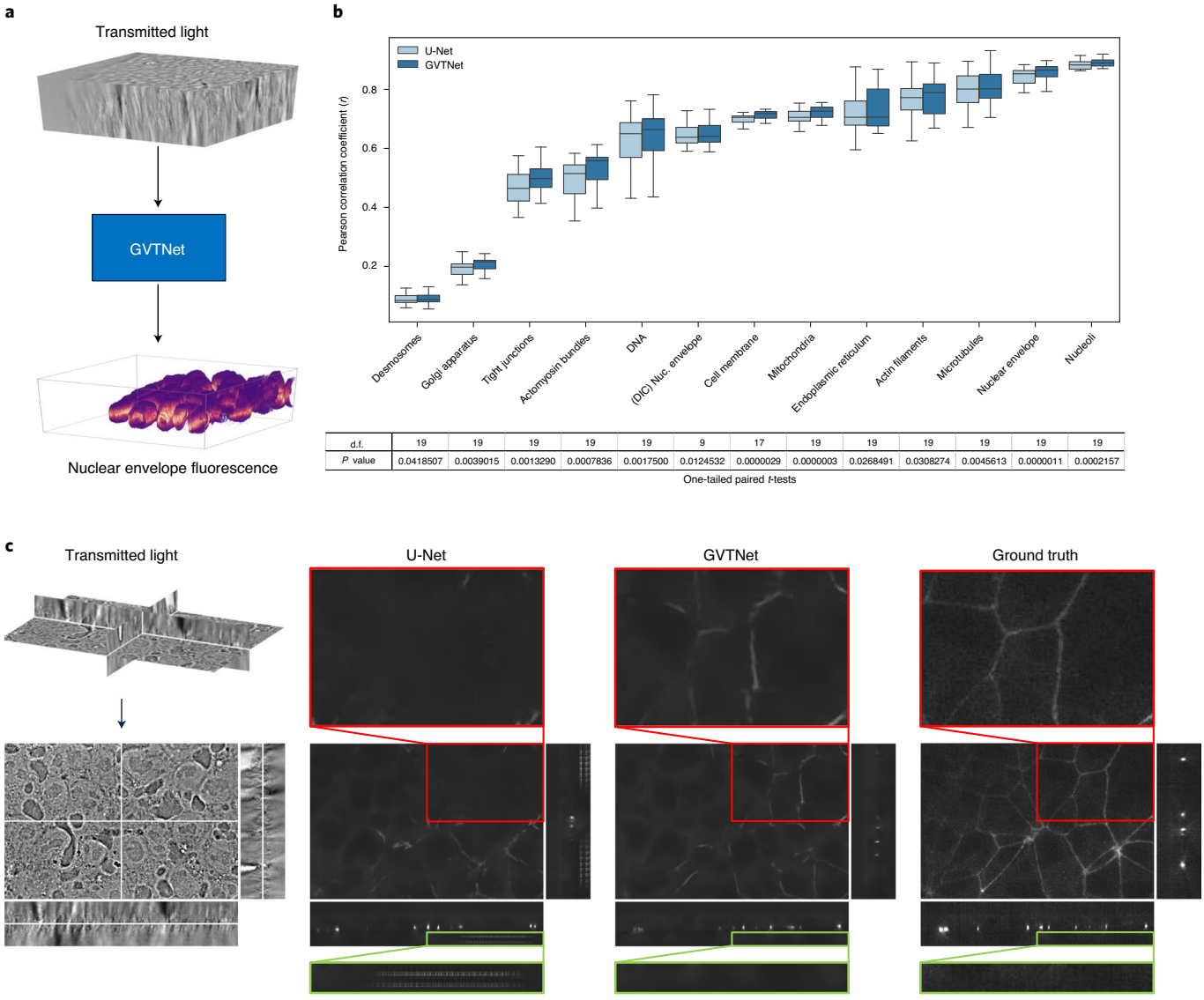

Fig. 2 | GVTNets on label-free prediction of 3D fluorescence images from transmitted-light microscopy. a, Task overview. This augmented microscopy task is to predict the fluorescence images of subcellular structures from inexpensive transmitted-light images without fluorescent labels. **b**, Prediction performance for different subcellular structures (higher is better). Top, distributions of the Pearson correlation coefficient (*r*) between the ground-truth images and the predicted images of the GVTNet on the testing datasets for 13 different subcellular structures. Each pair of images leads to a point in the distribution. In the box and whisker plots, the 25th, 50th and 75th percentile points are marked by the box, and whiskers indicate the minimum and maximum. The number of testing images is 18 for the cell membrane, 10 for the differential interference contrast (DIC) nuclear envelope and 20 for the others. Bottom, one-tailed paired *t*-test results on the performance of the GVTNet and the U-Net baseline. The degrees of freedom is the number of testing images minus one. All the *P* values are smaller than 0.05. **c**, Example of predicting fluorescence images of tight junctions from transmitted-light images. From left to right: transmitted-light input images, the predicted fluorescence images using the U-Net baseline, the predicted fluorescence images using the GVTNet and the ground-truth fluorescence images. We visualize the centre *z*-, *y*- and *x*-slices for 3D images. Clearly, the GVTNet captures more details. In addition, the GVTNet avoids artefacts caused by local operators.

high-SNR image in the testing dataset has three corresponding low-SNR images.

The baseline models in these experiments are the content-aware image restoration (CARE) networks[27], which are based on the 3D U-Net[36]. The U-Net-based CARE networks achieve the current best performance on these two datasets, serving as a strong baseline. We build a GVTNet by replacing the bottom convolutions and up-sampling operators with corresponding size-preserving and up-sampling GVTOs (Supplementary Fig. 4).

In order to quantify the model performance, we compute two evaluation metrics: the structural similarity index (SSIM)[41] and normalized root-mean-square error (NRMSE) (Methods). The models are evaluated individually under three SNR levels. The visualization results demonstrate that the GVTNet can take advantage of long-range dependencies to recover more details in areas with weak signals than the U-Net (Fig. 3c). The quantitative results also indicate substantial and consistent improvements of the GVTNet over the U-Net-based CARE under all image conditions on both datasets, revealing the advantages of GVTNets with more GVTOs (Supplementary Fig. 5, Supplementary Table 2). More examples of predictions on testing images can be found in Supplementary Figs. 7.

To provide insights on how GVTNets improve the inference performance by utilizing global information, we conduct extra experiments by varying the spatial sizes of input images during the

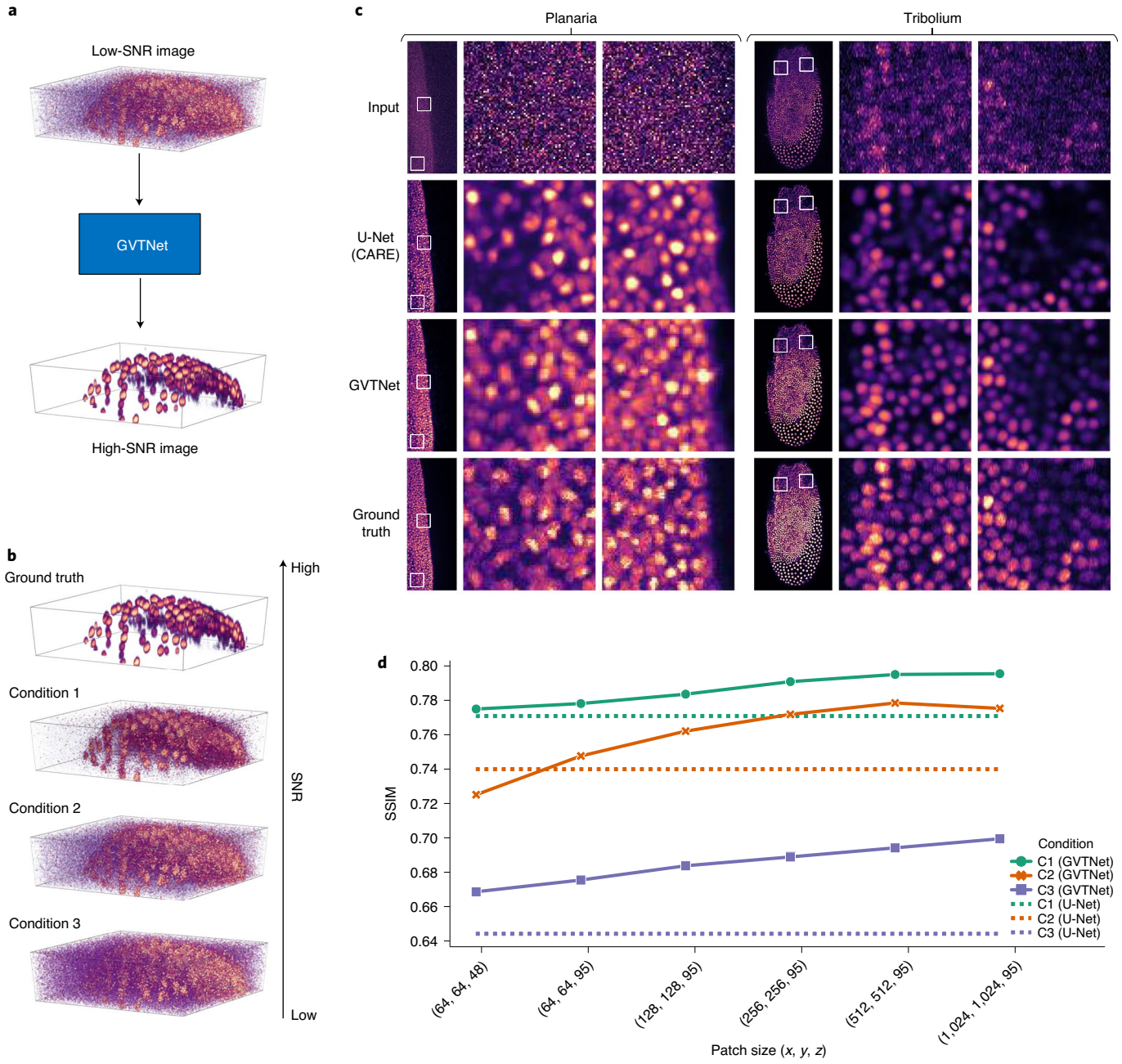

Fig. 3 | GVTNets on content-aware 3D image denoising. a, Task overview. This augmented microscopy task is to improve the SNR by removing the noise from the low-SNR images captured in poor imaging conditions. **b**, Image conditions. The ground-truth image captured with full exposure and laser power, along with three noised images captured in weaker conditions at three different levels (C1–C3). **c**, Examples of content-aware 3D image denoising. From top to bottom: the input noisy images, the predicted denoised images using the U-Net-based CARE, the predicted denoised images using the GVTNet and the ground-truth denoised images. On both Planaria and Tribolium datasets, the U-Net fails to capture details in input regions with weak signals. On the contrary, the GVTNet obtains more precise predictions with more detail in such regions. **d**, The inference performance in terms of SSIM over increasing prediction patch sizes on the Planaria dataset (higher is better). The number of testing images is 20. Dotted lines represent the U-Net and solid lines represent the GVTNet. The inference performance of the GVTNet increases with larger prediction patch sizes, showing its ability to utilize knowledge from the entire input.

inference process. To be specific, as both GVTNets and the U-Net are able to handle inputs of any spatial size, we can either feed the entire image directly into the model or crop the image into small prediction patches and reconstruct the entire augmented image after prediction (Supplementary Fig. 1). Theoretically, since the RF size in the U-Net is fixed and local, the prediction results will be the same as long as the size of the prediction patches is larger than that of RF. On the other hand, the RF size in GVTNets is dynamic and global, and always covers the entire input image. This property allows the use of more knowledge for better inference performance given large prediction patches, even if the training patches are much smaller. To verify this insight, we train the GVTNet and CARE on the Planaria dataset and compare prediction results in terms of SSIM when using prediction patches of sizes ranging from 64×64×48 to

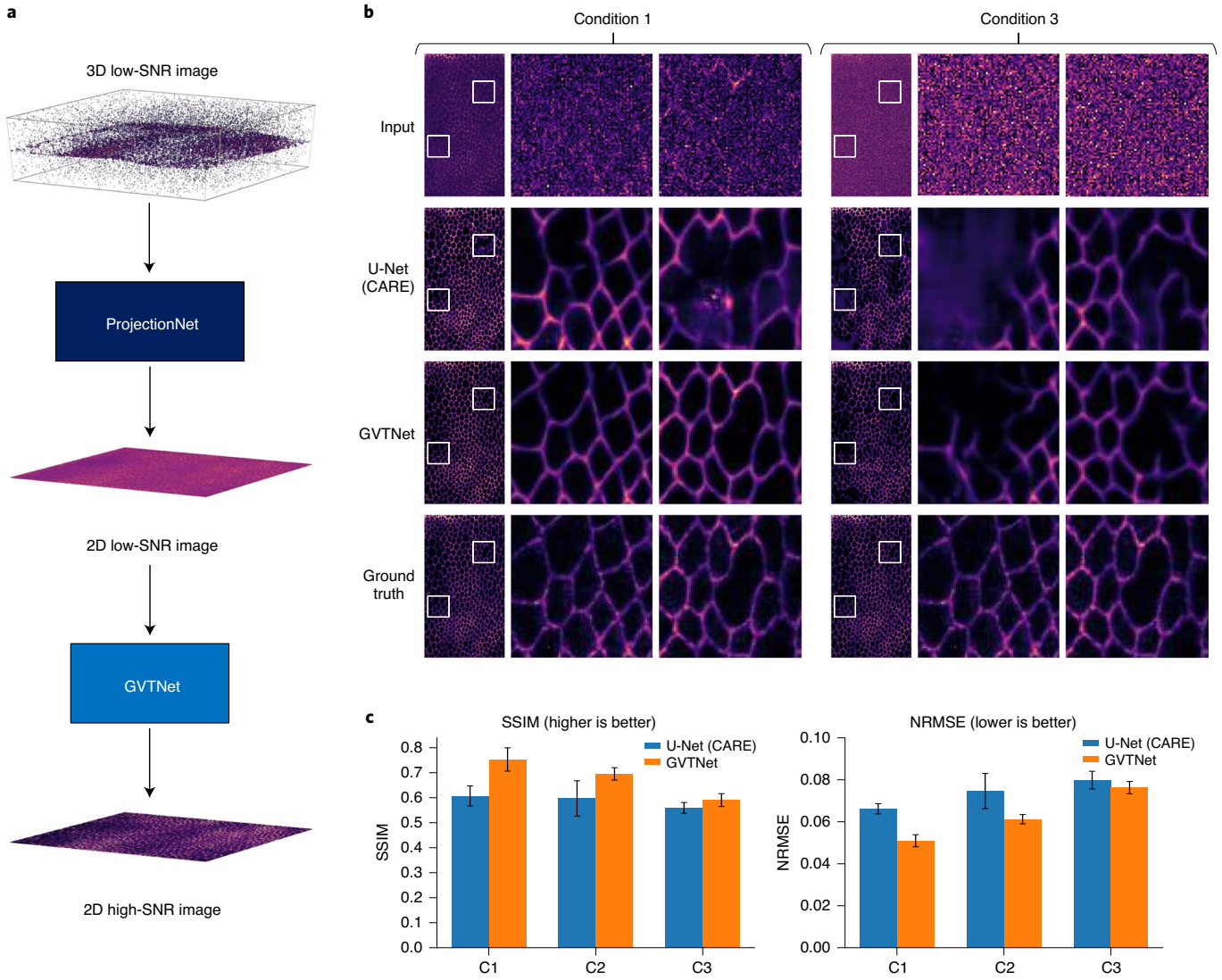

Fig. 4 | GVTNets on content-aware 3D-to-2D image projection. a, Task overview. This augmented microscopy task is to project a low-SNR 3D image into a high-SNR 2D surface. The model consists of a ProjectionNet that produces an intermediate 2D low-SNR image and a 2D GVTNet that outputs the high-SNR 2D image. **b**, Examples of content-aware 3D-to-2D image projection. From top to bottom: the predicted images using the U-Net-based CARE, the predicted images using the GVTNet and the ground-truth images. Visualization results indicate that the U-Net is more sensitive to the irregular input voxel values and collapses in surrounding areas. In addition, the U-Net tends to give ambiguous and blurred predictions where the input information is insufficient. On the contrary, the GVTNet is more robust to these cases. **c**, Prediction performance on the testing data of the Flywing dataset, in terms of SSIM and NRMSE under three imaging conditions. The number of testing images is 26. The 68% confidence intervals are marked by computing the standard deviation over testing images.

1,024×1,024×95 (entire image size). The results are summarized in Fig. 3d. Examples of predictions are provided in Supplementary Fig. 8 and detailed performance comparisons with error bars are provided in Supplementary Fig. 9. The prediction results of the U-Net remain the same when increasing prediction patch sizes, forming a horizontal line (Supplementary Fig. 1). On the contrary, important improvements can be observed for the GVTNet. These results show a unique advantage of GVTNets, that a performance boost can be achieved by using larger prediction patches. This advantage can be easily achieved without the need to retrain the model, which is expensive and time consuming. However, note that larger prediction patch sizes do not always lead to better performance. The correlations among voxels will get weaker and weaker with increasing distance. It is possible that more noise is aggregated than useful information when the prediction patch size reaches a certain point. For example, in Fig. 3d, we can observe a local maximum when the prediction patch size is 512×512×95 for C2.

Content-aware 3D-to-2D image projection. While we use GVTOs to build GVTNets, GVTOs are a family of operators that support any size-preserving, down-sampling and up-sampling tensor processing and can be used outside GVTNets. Therefore, we further examine the proposed GVTOs on more complicated and composite models. In particular, we apply GVTOs and GVTNets on the 3D *Drosophila melanogaster* flywing surface projection task[42,43] (Fig. 4a).

The model for this task is supposed to take a noised 3D image as the input and project it into a denoised 2D surface image. The typical deep learning model involves two parts: a network for 3D-to-2D surface projection, followed by a network for 2D image denoising. For example, the current best model, CARE[27], uses a task-specific

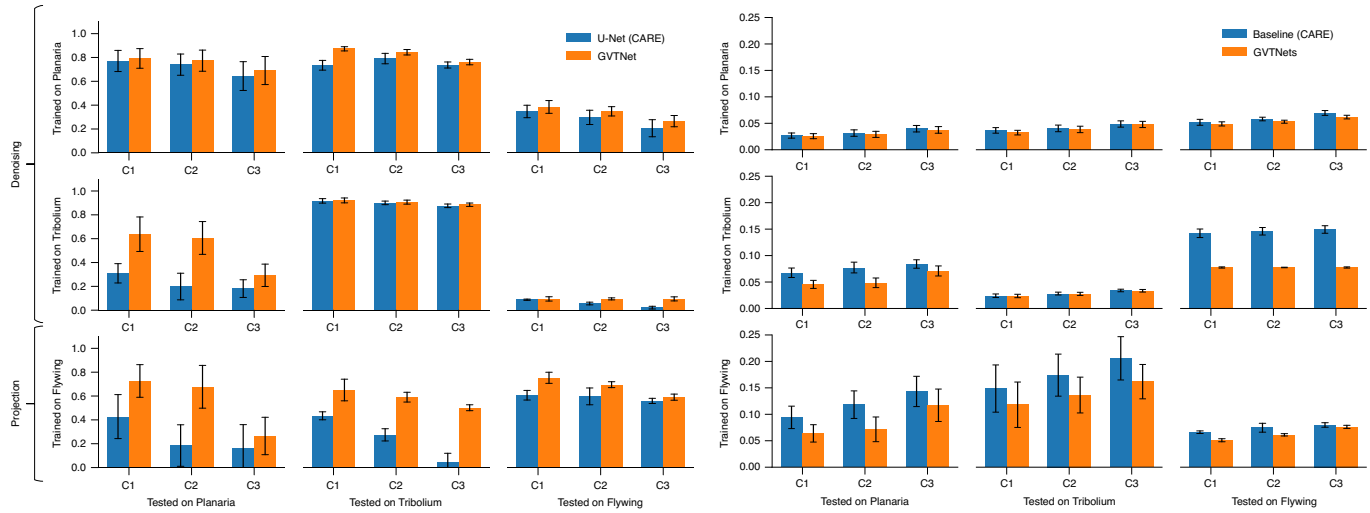

Fig. 5 | Generalization ability of GVTNets. Comparisons of the transfer learning performance between the U-Net base CARE and our GVTNets, in terms of SSIM (left, higher is better) and NRMSE (right, lower is better). Rows represent the dataset on which the models are trained and columns represent the dataset on which the models are tested. The first two rows correspond to the 3D denoising tasks and the third row corresponds to the 3D-to-2D projection tasks. The diagonal charts are the performance of models trained and tested on the same datasets. The GVTNets can achieve a promising performance under this simplest transfer learning setting, due to the input-dependent weights of GVTOs. The 68% confidence intervals are marked by computing the standard deviation over testing images.

convolutional neural network (CNN)[15] for projection and a 2D U-Net for denoising. The task-specific CNN is also composed of convolutions, down-sampling and up-sampling operators. We design our model based on CARE by applying GVTOs in the first CNN and replace the 2D U-Net with a 2D GVTNet (Supplementary Fig. 10). The resulting composite model employs size-preserving and up-sampling GVTOs in both parts.

We compare our model with CARE on the 'Flywing' dataset[27] in terms of SSIM and NRMSE. The dataset contains 16,891 pairs of small 3D noisy image patches and ground-truth 2D surface image patches for training, and 26 complete images for testing.

The quantitative results indicate that the composite model augmented by GVTOs achieves substantial improvements (Fig. 4c). We provide detailed quantitative results in Supplementary Table 3. The visualization results show that the GVTOs have a stronger capability to recognize non-noisy objects at regions of lower SNR within an image, where the original model tends to fail (Fig. 4b). This is because the global information is of great importance to the projection tasks, especially along the z-axis, where the projection happens. Specifically, for each (x, y) location in the 3D image, only one voxel along the z-axis will be projected to the 2D surface. This restriction is only available when the model has the global information along the z-axis. Therefore, plugging GVTOs into the projection process can effectively improve the overall performance. More examples of predictions on testing images can be found in Supplementary Fig. 11.

Transfer learning ability of GVTNets. We have shown the effectiveness of GVTNets for augmented microscopy applications under a supervised learning setting. In the following, we further investigate the generalization ability of GVTNets under a simple transfer learning setting[44], where we train GVTNets on one dataset and perform testing on other datasets for the same task. In this case, the inconsistencies between the training and testing data often lead to the collapse of models based on local operators, such as the U-Net. One reasonable explanation is that the weights of kernels in local operators are fixed after training and independent to the inputs[44]. This limits the ability to deal with the different data distributions in training and inference procedures.

As GVTOs achieve input-dependent weights, we hypothesize that GVTNets are more robust to such inconsistencies and have a better generalization ability. We conduct experiments to verify the hypothesis using the three datasets from Weigert et al.[27]: the Planaria, Tribolium and Flywing datasets. Note that all these datasets originally have 3D high-SNR ground-truth images for the 3D denoising task. By applying PreMosa[45] on the 3D ground-truth images, we can obtain 2D ground-truth images for the 3D-to-2D projection task. Therefore, these datasets can be used in either task for both training and testing. The baseline models are still the U-Net-based CARE networks in these experiments, and we use the same GVTNet as introduced above for comparison (Supplementary Fig. 4, Supplementary Fig. 10). In general, we train GVTNet and CARE on one of the three datasets, and compare their testing performance on the remaining two datasets, resulting in three sets of experiments. To be concrete, the first two experiments where either the Planaria or Tribolium dataset is used for training are doing the 3D denoising tasks. The third experiment where models are trained on the Flywing dataset is performing the 3D-to-2D projection task.

The comparison results in terms of SSIM and NRMSE are shown in Fig. 5. Detailed quantitative results can be found in Supplementary Table 4. GVTNet obtains a more promising transfer learning performance than CARE, indicating a better generalization ability.

Discussion

We have introduced GVTNets built on GVTOs, an advanced deep learning tool for augmented microscopy. Compared with U-Net, GVTNets are more powerful models that are capable of capturing long-range dependencies and selectively aggregating global information for inputs of any spatial size. With GVTNets, various augmented microscopy tasks can be performed with notably improved accuracy, such as predicting the fluorescence images of subcellular structures directly from transmitted-light images without using fluorescent labels, conducting content-aware image denoising and projecting a 3D microscope image to a 2D surface for analysis. We have demonstrated the superiority of GVTNets and GVTOs on several publicly available datasets for augmented microscopy[18,25,27]. In particular, we have provided examples where GVTNets achieve better inference performance with inputs of larger spatial sizes, indicating

the ability of utilizing global information. In addition, besides the supervised learning setting, GVTNets outperform the U-Net under a simple transfer learning setting, showing better generalization ability due to input-dependent weights.

We anticipate that our work would exert potential impacts on biological image analysis in general and augmented microscopy specifically. Image analysis plays an indispensable role in biological research, where machine learning methods and tools have been widely used and dramatically advanced biological research and discoveries. In particular, the past decade has witnessed revolutionary changes in machine learning with the rapid developments of deep learning[15]. Recent studies[22,24,25,27,46] have shown that deep learning allows biological research to transcend the limits imposed by imaging hardware, enabling discoveries at scales and resolutions that were previously impossible. We observe that most of these biological image analysis tasks can be formulated as biological image transformation problems[18]. In such tasks, U-Net[30,31] is the most popular and successful deep model, achieving state-of-the-art performances[18,24,25,27]. Our proposed GVTNets can be directly used to replace U-Net and boost the performance by addressing U-Nets' intrinsic limitations. Specifically, our experimental results have shown the superiority of GVTNets in various augmented microscopy tasks. These results are expected to have an immediate and strong impact on basic biology by enabling discoveries, observations and measurements that were previously unobtainable. In addition, since the limitations of U-Net are general and not task specific, we anticipate that GVTNets will improve on U-Net in other biological image transformation tasks and potentially benefit a wider range of biological research based on image analysis. Last but not least, from a practical perspective, deployment of solutions is as important as their development[18]. To make GVTNets easy to use in various biological image transformation tasks, we publish our code as an open-source tool with detailed instructions (Supplementary Note 2). Our code may greatly benefit both biology and computer science research communities.

In the literature, there exist many other studies that attempt to improve various aspects of U-Net[47–51]. Among them, some studies[47,48,51] explore a similar direction to our work, which is to allow U-Net to capture long-range dependencies or global context information. They can be divided into two main categories. One is to add modules composed of dilated convolutions, like Zhang et al.[47] and CE-NET[51]. Dilated convolutions can expand the RF of convolutions to capture longer-range dependencies. However, they are still local operators in essence, sharing similar limitations. For example, they cannot collect global information when inputs become larger than the RF. The other category is to apply global pooling to extract global information and use it to facilitate local operators, such as RSGU-Net[48]. However, important spatial information is lost during global pooling, which potentially limits performance. Differing from these two categories, we extend the attention operator to achieve the goal. To demonstrate the advantages of our method over previous methods, we compare GVTNets with representative models, RSGU-Net[48] and CE-NET[51], on content-aware 3D image denoising tasks (Supplementary Table 5). Our method outperforms both methods substantially, with similar computational cost.

Other studies[49,50] improve U-Net in orthogonal directions. Oktay et al.[49] propose adding the gate mechanism to the skip connections, filtering out irrelevant information. It is worth noting that the gate mechanism and the attention mechanism are essentially different in terms of computation, functionality and flexibility. The gate mechanism performs spatially element-wise filtering so that there is no explicit communication between spatial locations. On the contrary, the attention mechanism aggregates information from all spatial locations (Methods). Moreover, the gate mechanism can only be used for size-preserving tensor processing, while the attention mechanism can be extended for down-sampling and up-sampling tensor processing by our GVTOs. Zhou et al.[50] propose a nested U-Net architecture by adding dense skip connections. The nested architecture facilitates the training and yields better inference performance.

In terms of augmenting images with deep learning methods, generative adversarial network (GAN)[52] is a promising choice[18,22,46,53]. We point out that GAN-based methods are orthogonal to our GVTNets in the sense that they can be used together. Note that GAN is composed of a generator and a discriminator. In GAN-based image augmentation models, the generator is typically a U-Net[53], which we can improve with our GVTNets. We conduct experiments on content-aware 3D image denoising tasks. The results can be found in Supplementary Table 6. As indicated by the results, under the GAN framework, our GVTNets can improve on U-Net.

The key components of GVTNets are GVTOs. One concern about GVTOs is the efficiency. Given inputs of the same size, GVTOs usually require more time and take up more memory for computation than local operators like convolutions. This is due to the use of the self-attention operator. However, the high cost of GVTOs does not necessarily make GVTNets more expensive than U-Net. By taking advantage of the more powerful GVTOs, the overall network architecture can be simpler, improving the efficiency. For example, in the label-free fluorescence image prediction experiments, we have shown that a GVTNet can outperform a U-Net-based neural network with only 26.5% of training parameters and faster computation speed.

Another limitation of GVTNets is a shared disadvantage of current deep learning models[25,27]. Models trained on one biological image transformation dataset can hardly be used for another dataset. Therefore, high-quality training data must be collected for each task, which is expensive and time consuming. GVTNets have shown promising improvements under the simplest transfer learning setting without fine-tuning. We anticipate that the combination of GVTNets and recent advances of transfer learning[44] and meta learning[54] can greatly alleviate this limitation.

Methods

Network architecture. *General framework.* GVTNets follow the same encoder–decoder framework as U-Net[30,31,36], which represents a family of deep neural networks for biological image transformations. An encoder takes the image to be transformed as the input and computes feature maps of gradually reduced spatial sizes, which encode multi-scale and multi-resolution information from the input image. Then a corresponding decoder uses these feature maps to produce the transformed image, during which feature maps of gradually increased spatial sizes are computed. GVTNets support both 2D and 3D biological image transformations. We use the 3D case to describe the architecture in detail (Fig. 1c).

In our GVTNets, the encoder starts with an initial 3×3×3 convolution that transforms the input image into a chosen number of feature maps of the same spatial size, initializing the encoding. The encoding process is achieved by down-sampling operators interleaved with optional size-preserving operators. Each down-sampling operator halves the size along each spatial dimension of feature maps but doubles the channel dimension, that is, the number of feature maps. To be specific, given a $d\times h\times w\times c$ tensor representing c feature maps of the spatial size $d\times h\times w$ as inputs, a down-sampling operator will output an $d/2\times h/2\times w/2\times 2c$ tensor. Feature maps of the same spatial size are considered at the same level. As a result, the number of levels, also known as the depth of the network, is determined by the number of down-sampling operators in the encoder.

Correspondingly, the decoder is composed of the same number of up-sampling operators interleaved with optional size-preserving operators. The decoding process computes feature maps of increased spatial sizes in a level-by-level fashion, where each up-sampling operator doubles the size along each spatial dimension of feature maps but halves the channel dimension, as opposed to down-sampling operators. Therefore, there is a one-to-one correspondence between down-sampling and up-sampling operators. The decoder ends with an output convolution that outputs transformed image of the same spatial size as the input image.

The encoder and decoder are connected at each level. The bottom level contains the outputs of the encoder, which are feature maps of the smallest size in the U-Net framework. These feature maps, after optional size-preserving operators, serve as inputs to the decoder. In upper levels, there exist skip connections between the encoder and decoder. Concretely, the input feature maps to each

down-sampling operator are concatenated or added to the output feature maps of the corresponding up-sampling operator. The skip connections allow the decoder to take advantage of encoded multi-scale and multi-resolution information, which increases the capability of the framework and facilitates the training process[30,35].

Global voxel transformer networks. The major difference between our GVTNets and the original U-Net lies in the choices of the size-preserving, down-sampling and up-sampling operators. GVTNets are equipped with GVTOs, which can be flexibly used for size-preserving, down-sampling or up-sampling tensor processing. In particular, GVTNets fix the size-preserving operator at the bottom level to be the size-preserving GVTO, ensuring that global information is encoded and aggregated before going through the decoder. The other size-preserving operators are set to pre-activation residual blocks[33], consisting of two 3×3×3 convolutions with the ReLU activation function[55] (Supplementary Fig. 12a). Down-sampling and up-sampling GVTOs can be used as corresponding operators based on the datasets and tasks.

Global voxel transformer operators. As described above, the key components of our GVTNets are GVTOs, which are able to selectively use long-range information among input units. We take the 3D case to illustrate the size-preserving GVTO first, followed by the down-sampling and up-sampling GVTOs.

Size-preserving GVTO. Given the input third-order tensor $\mathscr{X} \in \mathbb{R}^{d\times h\times w\times c}$ representing c feature maps of the spatial size $d\times h\times w$, the size-preserving GVTO performs three independent 1×1×1 convolutions on $\mathscr{X}$ and obtains three tensors, namely the query ($\mathscr{Q}$), key ($\mathscr{K}$) and value ($\mathscr{V}$) tensor, where $\mathscr{Q}, \mathscr{K}, \mathscr{V} \in \mathbb{R}^{d\times h\times w\times c}$. Afterwards, $\mathscr{Q}$, $\mathscr{K}$ and $\mathscr{V}$ are unfolded along the channel dimension[56] into matrices $Q, K, V \in \mathbb{R}^{c\times dhw}$. These matrices go through the attention operator defined as

$$Y = V \cdot \mathrm{Normalize}(K^T Q) \in \mathbb{R}^{c\times dhw}$$

where Normalize(·) is a normalization function that normalizes each column of $Q^T K \in \mathbb{R}^{dhw\times dhw}$. Specifically, the size-preserving GVTO simply uses $1/dhw$ as the normalization function:

$$Y = V\frac{K^T Q}{dhw} = \frac{1}{dhw} VK^T Q \in \mathbb{R}^{c\times dhw}$$

where dhw is the second dimension of Q and subjected to corresponding changes in the down-sampling and up-sampling GVTOs. After the attention operator, the matrix Y is then folded back to a tensor $\mathscr{Y} \in \mathbb{R}^{d\times h\times w\times c}$. The final outputs of the size-preserving GVTO is the summation of $\mathscr{X}$ and $\mathscr{Y}$, which means a residual connection from the inputs to the outputs[32]. In particular, we use the pre-activation technique as well[33]. As a result, the size-preserving GVTO preserves the dimension of the inputs (Supplementary Fig. 13e).

Down-sampling and up-sampling GVTOs. The extension from the size-preserving GVTO to the down-sampling and up-sampling GVTOs is achieved by changing the convolutions that compute $\mathscr{Q}$, $\mathscr{K}$, $\mathscr{V}$. We take the down-sampling GVTO as an example for illustration. Given the same input tensor $\mathscr{X} \in \mathbb{R}^{d\times h\times w\times c}$, we use a 3×3×3 convolution with stride 2 to obtain $\mathscr{Q} \in \mathbb{R}^{d/2\times h/2\times w/2\times 2c}$ and two independent 1×1×1 convolutions to generate $\mathscr{K} \in \mathbb{R}^{d\times h\times w\times 2c}$ and $\mathscr{V} \in \mathbb{R}^{d\times h\times w\times 2c}$. The following computation is the same; that is, $\mathscr{Q}$, $\mathscr{K}$, $\mathscr{V}$ are unfolded along the channel dimension into matrices $Q \in \mathbb{R}^{2c\times dhw/8}$ and $K, V \in \mathbb{R}^{2c\times dhw}$, which are fed into the same attention operator and output the matrix $Y \in \mathbb{R}^{2c\times dhw/8}$. Folding it back results in a tensor $\mathscr{Y} \in \mathbb{R}^{d/2\times h/2\times w/2\times 2c}$. Comparing the dimensions of $\mathscr{X}$ and $\mathscr{Y}$, we achieve a down-sampling process that halves the size along each spatial dimension of feature maps but doubles the channel dimension. We complete the down-sampling GVTO by adding the residual connection in two ways, corresponding to two versions of the down-sampling GVTO (Supplementary Fig. 13a,b). One is to perform an extra 3×3×3 convolution with stride 2 through the residual connection from $\mathscr{X}$ to $\mathscr{Y}$, in order to transform $\mathscr{X}$ to have the same dimension as $\mathscr{Y}$; the other is to directly add $\mathscr{Q}$ to $\mathscr{Y}$, based on the fact that $\mathscr{Q}$ is obtained from $\mathscr{X}$.

The up-sampling GVTO is dual to the down-sampling GVTO. Instead of using a convolution with stride 2, it uses a 3×3×3 transposed convolution with stride 2 to obtain $\mathscr{Q} \in \mathbb{R}^{2d\times 2h\times 2w\times c/2}$. In addition, the other two 1×1×1 convolutions generate $\mathscr{K} \in \mathbb{R}^{d\times h\times w\times c/2}$ and $\mathscr{V} \in \mathbb{R}^{d\times h\times w\times c/2}$. The up-sampling GVTO doubles the size along each spatial dimension of feature maps but halves the channel dimension and also has two versions corresponding to different residual connections (Supplementary Fig. 13c,d).

Advantages of GVTOs. It is noteworthy that each spatial location in the output tensor of GVTOs has access to all the information in the input tensor, and is able to selectively use or ignore information. We illustrate this point by regarding $\mathscr{X} \in \mathbb{R}^{d\times h\times w\times c}$ as $d\times h\times w$ c-dimensional vectors, where each vector represents the information in a spatial location. In this view, each vector has a one-to-one correspondence to each column in K and V in GVTOs, respectively. Revisiting the attention operator, each column in Y is a vector representation of each spatial location in the output tensor, and has a one-to-one correspondence to each column in Q. Moreover, each column in Y is computed as the weighted sum of columns in V, whose weights are determined by the interaction between the corresponding column in Q and all columns in K. The weights can be viewed as filters of the amount of information from each spatial location in the inputs to the outputs. In addition, as both Q and K are computed from the input tensor, the weights are input dependent. Therefore, GVTOs achieve the dynamic non-local information aggregation.

Comparisons with fully FC layers. It is important to note that the proposed GVTOs are different from FC layers in fundamental ways, although they both allow each output unit to use information from the entire input. Compared to FC layers, outputs in GVTOs are computed based on relations among inputs. Thus the weights are input dependent, rather than learned and fixed during prediction as in FC layers. The only trainable parameters in GVTOs are the convolutions to compute $\mathscr{Q}$, $\mathscr{K}$, $\mathscr{V}$, whose sizes are independent of input and output sizes. As a consequence, GVTOs allow variable-size inputs, and the positional correspondence between inputs and outputs is preserved in GVTOs. By contrast, FC layers require fixed-size inputs and positional correspondence is lost.

Training loss. GVTNets are trained in an end-to-end fashion with two options of the loss functions. One is the mean-squared error (MSE):

$$L_{\mathrm{MSE}}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y represents the ground-truth image, $\hat{y}$ represents the model's predicted image and N represents the total number of voxels in the image. The other is the mean absolute error (MAE):

$$L_{\mathrm{MAE}}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Both MSE and MAE measure the differences between the predicted image and the ground-truth image. The training process applies the Adam optimizer[57] with a user-chosen learning rate to minimize the loss.

Evaluation metrics. *Pearson correlation coefficient.* Pearson correlation coefficient (r) is computed as

$$r(y,\hat{y}) = \frac{\sum_{i=1}^{N}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \mu_y)^2 \sum_{i=1}^{N}(\hat{y}_i - \mu_{\hat{y}})^2}}$$

where μ_y and $\mu_{\hat{y}}$ are the mean of voxel intensities in y and $\hat{y}$, respectively.

Normalized root-mean-square error. The root-mean-square error (RMSE) is computed as

$$\mathrm{RMSE}(y,\hat{y}) = \sqrt{L_{\mathrm{MSE}}(y,\hat{y})}$$

The NRMSE simply adds a normalization function on y and $\hat{y}$, respectively. In our tools and experiments, we apply the same percentile-based normalization and transformation as in Weigert et al.[27]. Concretely, the NRMSE is defined by

$$\mathrm{NRMSE}(y,\hat{y}) = \sqrt{\min_{\phi} L_{\mathrm{MSE}}(\phi(\hat{y}), N(y, 0.1, 99.9))}$$

where

$$N(y, 0.1, 99.9) = \frac{y - \mathrm{percentile}(y, 0.1)}{\mathrm{percentile}(y, 99.9) - \mathrm{percentile}(y, 0.1)}$$

is the percentile-based normalization, and $\phi(\hat{y}) = \alpha\hat{y} + \beta$ denotes a transformation that scales and shifts $\hat{y}$. During the implementation, we let $\alpha = \frac{\mathrm{Cov}(y-\bar{y},\hat{y}-\bar{\hat{y}})}{\mathrm{Var}(\hat{y}-\bar{\hat{y}})}$ and $\beta = 0$ to obtain $\phi(\hat{y})$ so that the MSE is minimized.

Structural similarity index. The SSIM[41] is computed as

$$\mathrm{SSIM}(y,\hat{y}) = \frac{(2\mu_y\mu_{\hat{y}} + c_1)(2\sigma_{y\hat{y}} + c_2)}{(\mu_y{}^2 + \mu_{\hat{y}}{}^2 + c_1)(\sigma_y{}^2 + \sigma_{\hat{y}}{}^2 + c_2)}$$

where σ_y is the variance of y, $\sigma_{\hat{y}}$ is the variance of $\hat{y}$, $\sigma_{y\hat{y}}$ is the covariance of y and $\hat{y}$, and $c_1 = (0.01L)^2$, $c_2 = (0.03L)^2$ are two constant parameters of SSIM. Here, L represents the range of intensity values and is set to 1.

Task-specific configurations. The settings of our device are: GPU: Nvidia GeForce RTX 2080 Ti 11GB; CPU: Intel Xeon Silver 4116 2.10GHz; OS: Ubuntu 16.04.3 LTS.

Label-free prediction of 3D fluorescence images from transmitted-light microscopy. The basic GVTNet used in the experiments of label-free prediction of 3D fluorescence images is illustrated in Supplementary Fig. 2. The network has

depth four, where the skip connections add feature maps from the encoder to the decoder. In particular, the bottom block of the basic GVTNet is the size-preserving GVTO (Supplementary Fig. 11e). The number of feature maps after the initial convolution is set to 32. Batch normalization[58] with the momentum of 0.997 and epsilon of 0.00001 is applied before each ReLU activation function.

The 13 subtasks corresponding 13 different subcellular structures are performed separately and independently. To train the GVTNet, the 30 pairs of training images are randomly cropped into patches of size 64×64×32 and each training batch contains 16 pairs of patches. We minimize the MSE loss using the Adam optimizer with a learning rate of 0.001 for 70,000 to 100,000 minibatch iterations, depending on different subtasks. The training procedure lasts approximately 675 min to 945 min for each of the 13 datasets[25].

Context-aware 3D image denoising. The GVTNet used in the image denoising tasks is illustrated in Supplementary Fig. 4. It follows a 3D U-Net framework of depth three, that is, including two down-sampling and up-sampling operators, respectively. The skip connections merge feature maps from the encoder to the decoder by concatenation instead of addition. The bottom block is the size-preserving GVTO and two up-sampling operators are the up-sampling GVTOs v2 (Supplementary Fig. 11d). The number of feature maps after the initial convolution is set to 32. No batch normalization is applied.

We use the MAE loss with the Bayesian deep learning technique[59] (Supplementary Note 1) to train GVTNet. The training patch size is 64×64×16. We train the model with a batch size of 16 and a base learning rate of 0.0004 with a decay rate 0.7 for every 10,000 minibatch iterations. The training procedure takes 50 epochs and lasts about 345 min and 290 min for the Planaria and Tribolium datasets[27], respectively.

Content-aware 3D-to-2D image projection. The model for surface projection is composed of a 3D-to-2D projection network and a 2D denoising network, as illustrated in Supplementary Fig. 10. The projection network predicts the probability of each voxel in the 3D input image belonging to the 2D surface, and uses summation weighted by the predicted probabilities along the z-axis to finish the projection. The probabilities are estimated by a GVTO-augmented CNN. The following 2D denoising network is simply a 2D version of the GVTNet used in the image denoising tasks.

During training, the 3D input patch size is 64×64×50 and the 2D ground-truth patch size is 64×64×1. The other training settings are the same as those in image denoising experiments, except that we do not use the Bayesian deep learning technique. The training procedure lasts 295 min for the Flywing dataset[27].

Data availability

Datasets for label-free prediction of 3D fluorescence images from transmitted-light microscopy[25] can be downloaded from https://downloads.allencell.org/publication-data/label-free-prediction/index.html. Datasets for context-aware 3D image denoising and 3D-to-2D image projection[27] can be downloaded from https://publications.mpi-cbg.de/publications-sites/7207.

Code availability

The code for GVTNets training, prediction and evaluation (in Python/TensorFlow) is publicly available at https://github.com/divelab/GVTNets and ref. [60].

Received: 20 February 2020; Accepted: 14 December 2020;
Published online: 25 January 2021

References

1. Gustafsson, M. G. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *J. Microsc.* **198**, 82–87 (2000).
2. Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J. & Stelzer, E. H. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science* **305**, 1007–1009 (2004).
3. Betzig, E. et al. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
4. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nat. Meth.* **3**, 793–796 (2006).
5. Heintzmann, R. & Gustafsson, M. G. Subdiffraction resolution in continuous samples. *Nat. Photon.* **3**, 362–364 (2009).
6. Tomer, R., Khairy, K., Amat, F. & Keller, P. J. Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *Nat. Meth.* **9**, 755–763 (2012).
7. Chen, B.-C. et al. Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. *Science* **346**, 1257998 (2014).
8. Belthangady, C. & Royer, L. A. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat. Meth.* **16**, 1215–1225 (2019).
9. Laissue, P. P., Alghamdi, R. A., Tomancak, P., Reynaud, E. G. & Shroff, H. Assessing phototoxicity in live fluorescence imaging. *Nat. Meth.* **14**, 657–661 (2017).
10. Icha, J., Weber, M., Waters, J. C. & Norden, C. Phototoxicity in live fluorescence microscopy, and how to avoid it. *Bioessays* **39**, 1700003 (2017).
11. Selinummi, J. et al. Bright field microscopy as an alternative to whole cell fluorescence in automated analysis of macrophage images. *PLoS ONE* **4**, e7497 (2009).
12. Pawley, J. B. in *Handbook of Biological Confocal Microscopy* (ed. Pawley, J. B.) 20-42 (Springer, 2006).
13. Scherf, N. & Huisken, J. The smart and gentle microscope. *Nat. Biotechnol.* **33**, 815–818 (2015).
14. Skylaki, S., Hilsenbeck, O. & Schroeder, T. Challenges in long-term imaging and quantification of single-cell dynamics. *Nat. Biotechnol.* **34**, 1137–1144 (2016).
15. LeCun, Y. et al. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
16. Sullivan, D. P. & Lundberg, E. Seeing more: a future of augmented microscopy. *Cell* **173**, 546–548 (2018).
17. Chen, P. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
18. Moen, E. et al. Deep learning for cellular image analysis. *Nat. Meth.* **16**, 1233–1246 (2019).
19. Johnson, G. R., Donovan-Maiye, R. M. & Maleckar, M. M. Building a 3D integrated cell. Preprint at https://doi.org/10.1101/238378 (2017).
20. Ounkomol, C. et al. Three dimensional cross-modal image inference: label-free methods for subcellular structure prediction. Preprint at https://doi.org/10.1101/216606 (2017).
21. Osokin, A., Chessel, A., Carazo Salas, R. E. & Vaggi, F. GANs for biological image synthesis. In *Proc. IEEE International Conference on Computer Vision* 2233-2242 (2017).
22. Yuan, H. et al. Computational modeling of cellular structures using conditional deep generative networks. *Bioinformatics* **35**, 2141–2149 (2019).
23. Johnson, G., Donovan-Maiye, R., Ounkomol, C. & Maleckar, M. M. Studying stem cell organization using 'label-free' methods and a novel generative adversarial model. *Biophys. J.* **114**, 43A (2018).
24. Christiansen, E. M. et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell* **173**, 792–803 (2018).
25. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Meth.* **15**, 917–920 (2018).
26. Wu, Y. et al. Three-dimensional virtual refocusing of fluorescence microscopy images using deep learning. *Nat. Meth.* **16**, 1323–1331 (2019).
27. Weigert, M. et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nat. Meth.* **15**, 1090–1097 (2018).
28. Wang, H. et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nat. Meth.* **16**, 103–110 (2019).
29. Rivenson, Y. et al. Deep learning microscopy. *Optica* **4**, 1437–1443 (2017).
30. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
31. Falk, T. et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Meth.* **16**, 67–70 (2019).
32. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
33. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision* 630–645 (Springer, 2016).
34. Fakhry, A., Zeng, T. & Ji, S. Residual deconvolutional networks for brain electron microscopy image segmentation. *IEEE Trans. Med. Imaging* **36**, 447–456 (2017).
35. Lee, K., Zung, J., Li, P., Jain, V. & Seung, H. S. Superhuman accuracy on the SNEMI3D connectomics challenge. Preprint at https://arxiv.org/abs/1706.00120 (2017).
36. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 424–432 (Springer, 2016).
37. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://arxiv.org/abs/1409.1556 (2014).
38. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* 5998–6008 (2017).
39. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 7794–7803 (2018).
40. Wilson, D. R. & Martinez, T. R. The general inefficiency of batch training for gradient descent learning. *Neural Networks* **16**, 1429–1451 (2003).
41. Wang, Z. et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
42. Aigouy, B. et al. Cell flow reorients the axis of planar polarity in the wing epithelium of drosophila. *Cell* **142**, 773–786 (2010).

43. Etournay, R. et al. Interplay of cell dynamics and epithelial tension during morphogenesis of the *Drosophila* pupal wing. *eLife* **4**, e07090 (2015).
44. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions Knowl. Data Eng.* **22**, 1345–1359 (2009).
45. Blasse, C. et al. PreMosa: extracting 2D surfaces from 3D microscopy mosaics. *Bioinformatics* **33**, 2563–2569 (2017).
46. Cai, L., Wang, Z., Gao, H., Shen, D. & Ji, S. Deep adversarial learning for multi-modality missing data completion. In *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1158–1166 (Association for Computing Machinery, 2018).
47. Zhang, Q., Cui, Z., Niu, X., Geng, S. & Qiao, Y. Image segmentation with pyramid dilated convolution based on ResNet and U-Net. In *International Conference on Neural Information Processing* 364–372 (Springer, 2017).
48. Huang, J. et al. Range scaling global U-Net for perceptual image enhancement on mobile devices. In *Proc. European Conference on Computer Vision (ECCV)* (Springer, 2018).
49. Oktay, O. et al. Attention U-Net: learning where to look for the pancreas. Preprint at https://arxiv.org/abs/1804.03999 (2018).
50. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: a nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 3–11 (Springer, 2018).
51. Gu, Z. et al. CE-Net: context encoder network for 2D medical image segmentation. *IEEE Trans. Med. Imaging* **38**, 2281–2292 (2019).
52. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680 (MIT Press, 2014).
53. Rivenson, Y. et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nat. Biomed. Eng.* **3**, 466–477 (2019).
54. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. 34th International Conference on Machine Learning* **70**, 1126–1135 (JMLR, 2017).
55. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105 (2012).
56. Kolda, T. G. & Bader, B. W. Tensor decompositions and applications. *SIAM Rev.* **51**, 455–500 (2009).
57. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations* (2015).
58. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* 448–456 (2015).
59. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* 5574–5584 (2017).
60. Wang, Z., Xie, Y. & Ji, S. zhengyang-wang/GVTNets: Code for "Global voxel transformer networks for augmented microscopy" (version v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.4285769 (2020).

Acknowledgements

We thank the teams at CARE and the Allen Institute for Cell Science for making their data and tools publicly available. This work was supported in part by National Science Foundation grants DBI-1922969, IIS-1908166 and IIS-1908220, National Institutes of Health grant 1R21NS102828 and Defense Advanced Research Projects Agency grant N66001-17-2-4031.

Author contributions

S.J. conceived and initiated the research. Z.W. and S.J. designed the methods. Z.W. and Y.X. implemented the training and validation methods. Z.W. and Y.X. designed and developed the software package. S.J. supervised the project. Z.W., Y.X. and S.J. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00283-x.

Correspondence and requests for materials should be addressed to S.J.

Peer review information *Nature Machine Intelligence* thanks Ruogu Fang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.