Fast 3D Cell Volume Reconstruction from Multi-focus Images by Cross-Modality Knowledge Distillation

Imam Khairi Lubis[1], Ken'ichi Morooka[1][0000−0002−3308−6803], Rin Matsukado[1], Yukako Nakamae[1], and Hajime Nagahara[2][0000−0003−1579−8767]

[1] Kumamoto University, Kumamoto, Japan
{khairi.lubis}@outlook.com
{morooka,nakamae}@cs.kumamoto-u.ac.jp
{mrinppp53}@gmail.com
[2] Osaka University, Osaka, Japan
nagahara@ids.osaka-u.ac.jp

Abstract. 3D cell volumes provide richer structural information than 2D images of cells. Therefore, the use of the 3D cell volumes enables stable and reliable cytological diagnosis. Previous work has introduced a Deep Image Prior (DIP)-based method for reconstructing 3D cell volumes from multi-focus image sequences. Although the method does not require a large-scale training dataset, it requires a long time to reconstruct 3D volumes. In this paper, we propose an efficient deep learning-based approach for reconstructing 3D cell volumes from multi-focus image sequences via cross-modality knowledge distillation, termed Deep Cell Volume Reconstruction (DeepCVR). To begin with, using 3D cell volumes, we construct a U-Net-based variational autoencoder, called CellVAE, which learns an optimal latent space for representing the structural information of 3D cell volumes. Next, using the encoder part of the trained CellVAE as a teacher model, we construct a student model, called MF-CellEncoder, which reproduces the teacher's latent variables from multi-focus images. Finally, DeepCVR is constructed by combining the MF-CellEncoder and CellVAE decoder. DeepCVR reconstructs 3D cell volumes in a single forward pass. This avoids the per-sample optimization required by the DIP-based method. Experimental results show that DeepCVR greatly reduces the reconstruction time while approximating DIP-based reconstructions with comparable quality. Code available at: https://github.com/imamkhairi/DeepCVR.

Keywords: 3D cell volume reconstruction · Multi-focus images · Cross-modality knowledge distillation

1 Introduction

In medical diagnosis and examination, 3D structural information of target objects provides more comprehensive morphological insight than conventional methods using 2D images. In cytology, imaging modalities such as confocal microscopy

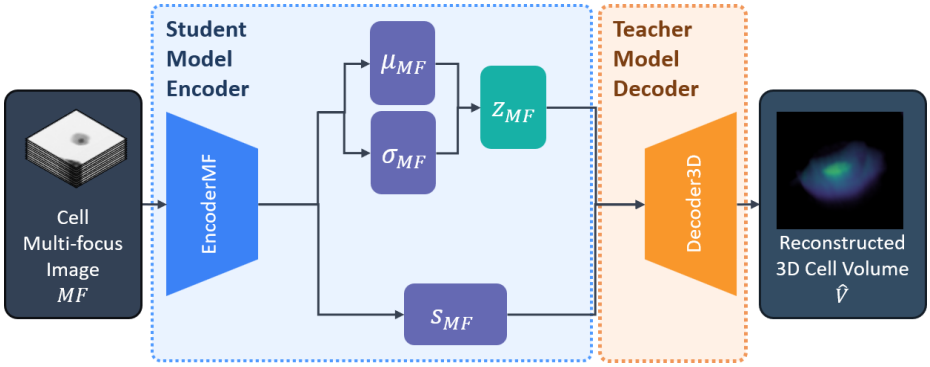


Fig. 1: Overview of the proposed method with student encoder and teacher decoder.

and electron microscopy are used to acquire 3D structures of cells [1, 6, 16]. However, these modalities require specialized equipment and complex sample preparation procedures, limiting their routine use in standard cytology.

In contrast, bright-field microscopy is widely used in cytological studies. Recently, bright-field Whole Slide Imaging (WSI) has become a powerful tool for cytological examination and analysis [3, 10, 13]. WSI can acquire a multi-focus image sequence by changing the focal position along the optical axis without additional hardware modifications. Each image can be regarded as an x-y plane observed at a different focal depth z.

Multi-focus image sequences can be used to estimate 3D cell structures. One computational approach is shape-from-focus (SFF), which estimates depth from focus measures across the image sequence [5, 12]. However, SFF-based approaches are often difficult to apply to translucent cells because underlying layers affect the observed intensity, making the focus cues ambiguous.

To reconstruct 3D cell volumes from bright-field microscopic images, Yamaguchi et al. [15] proposed a physics-based imaging framework that models the process of capturing multi-focus images of translucent cells using bright-field microscopy. In their method, the reconstructed 3D volume consists of a set of voxels, each of which has the transmittance value at that location. However, this method is time-consuming because of estimating 3D volume by an iterative least squares method.

Extending this framework, Azevedo et al. [2] introduced a Deep Image Prior (DIP) [14] to reconstruct 3D cell volumes from multi-focus images to improve the quality of 3D volumes. DIP solves the inverse image reconstruction problem by optimizing network parameters for each individual input without large-scale training data. However, because the optimization is performed independently for each input, the reconstruction process remains computationally expensive.

In this paper, we propose DeepCVR, a deep learning-based method for fast 3D cell volume reconstruction from multi-focus image sequences (Fig. 1). Deep-

CVR is based on the combination of knowledge distillation applied to image reconstruction tasks [7] and intermediate representations for cross-modality knowledge transfer [17]. This design separates the reconstruction problem into two steps. First, a variational autoencoder (VAE) [4], called CellVAE, is trained to learn a latent representation of 3D cell volumes. Using the encoder of the trained CellVAE as a teacher model, a student network, MF-CellEncoder, is trained to reproduce the latent and intermediate representations of 3D cell volumes from multi-focus inputs via cross-modality knowledge distillation. This avoids learning the 3D representation and cross-modal mapping simultaneously. Finally, DeepCVR is constructed by combining MF-CellEncoder and the decoder of CellVAE. In the inference phase, when a sequence of multi-focus images is provided, DeepCVR outputs the 3D volume of cells in the sequence.

The main contribution of this study is to eliminate the need for per-sample optimization in DIP-based 3D reconstruction by transferring the reconstruction process into a feed-forward model. DeepCVR uses teacher-student distillation to learn the reconstruction knowledge obtained from DIP-reconstructed volumes. This enables 3D reconstruction that approximates DIP-based reconstructions while significantly reducing reconstruction time.

2 Methodology

DeepCVR was trained using paired data consisting of multi-focus image sequences and their corresponding 3D volumes. Each sequence contains 11 focal slices with a resolution of 384×384 pixels. The corresponding 3D volumes are generated using the DIP-based method [2].

2.1 CellVAE

CellVAE is a 3D U-Net [11]-based VAE [4] that learns a latent representation $\boldsymbol{z}_{3D}$ and a set of intermediate features $\boldsymbol{s}_{3D}^{(i)}$ $(i = 1, \ldots, K)$ from a 3D cell volume $\boldsymbol{V}$, as shown in Fig. 2.

The input volume is padded to $12 \times 384 \times 384$, making the depth dimension compatible with the downsampling and upsampling operations in the network. Downsampling along the z-axis is performed only once, thereby preserving sufficient depth resolution in the latent representation.

The additional padded depth is used only during the internal network computation. After decoding, the output volume is cropped back to the original depth of 11. Therefore, the padded depth is excluded from both the reconstruction loss and the evaluation metrics.

Based on the VAE framework, the encoder models the approximate posterior distribution in the latent space as follows:

$$q(\boldsymbol{z}_{3D}|\boldsymbol{V}) = \mathcal{N}(\boldsymbol{\mu}_{3D}, \boldsymbol{\sigma}_{3D}^2), \tag{1}$$

where $\boldsymbol{\mu}_{3D}$ and $\boldsymbol{\sigma}_{3D}^2$ denote the mean and variance of the approximate posterior distribution, respectively. During training, the latent vector $\boldsymbol{z}_{3D}$ is sampled using the reparameterization trick.

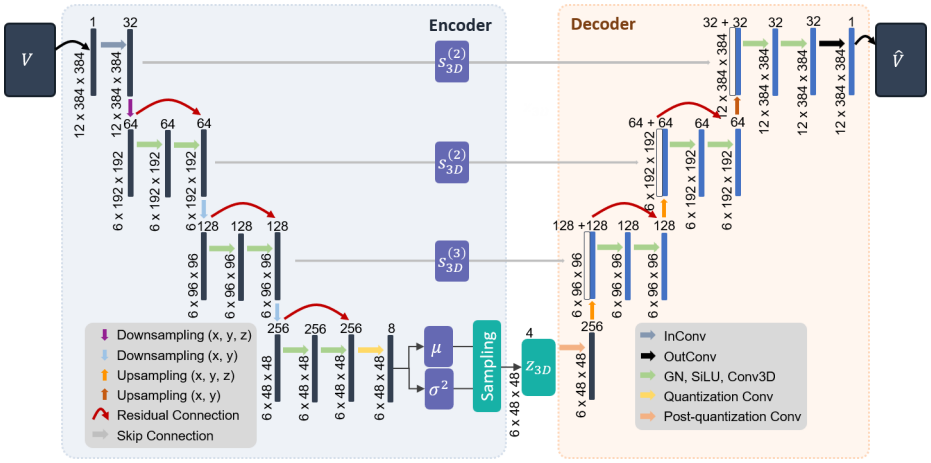


Fig. 2: CellVAE architecture.

In this study, CellVAE is used as a VAE-based model to learn a compact and regularized latent representation of 3D cell volumes. The probabilistic formulation regularizes the latent space during training. During reconstruction, we use the latent mean instead of random sampling to obtain deterministic outputs. To reduce the potential loss of spatial information in the latent space, intermediate features $\boldsymbol{s}_{3D}^{(i)}$ are extracted at the i-th encoder level and passed to the decoder through skip connections.

CellVAE is trained using an objective function consisting of a reconstruction loss and a Kullback–Leibler (KL) divergence regularization term:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + D_{\text{KL}}\left(q(\boldsymbol{z}_{3D}|\boldsymbol{V}) \,\|\, p(\boldsymbol{z})\right), \tag{2}$$

where $\mathcal{L}_{\text{rec}} = \|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_2^2$ denotes the mean squared error (MSE) between the input 3D volume $\boldsymbol{V}$ and its reconstructed volume $\hat{\boldsymbol{V}}$ produced by CellVAE. The KL term $D_{\text{KL}}\left(q(\boldsymbol{z}_{3D}|\boldsymbol{V}) \,\|\, p(\boldsymbol{z})\right)$ regularizes the approximate posterior distribution $q(\boldsymbol{z}_{3D}|\boldsymbol{V})$ toward a standard normal prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

2.2 MF-CellEncoder

As shown in Fig. 3, the inputs of CellVAE and MF-CellEncoder are different. CellVAE receives a 3D volume, whereas MF-CellEncoder receives a multi-focus image sequence. Therefore, we train MF-CellEncoder using teacher-student learning so that it can reproduce the 3D representation learned by the CellVAE encoder from multi-focus image inputs.

In this framework, the trained CellVAE encoder is used as the teacher model. Given a 3D volume, the teacher outputs the latent distribution parameters $\boldsymbol{\mu}_{3D}$ and $\boldsymbol{\sigma}_{3D}^2$, as well as intermediate features $\boldsymbol{s}_{3D}$. The MF-CellEncoder, which has

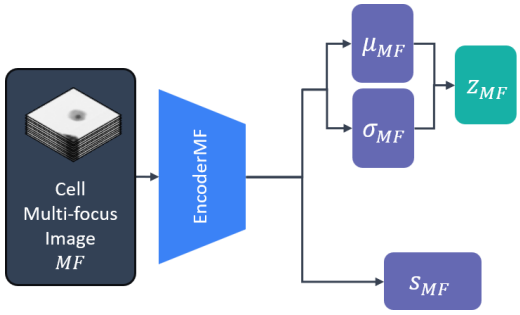


Fig. 3: MF-CellEncoder architecture.

the same architecture as the CellVAE encoder, is used as the student model. Given the corresponding multi-focus image sequence, MF-CellEncoder outputs the latent distribution parameters $\boldsymbol{\mu}_{MF}$ and $\boldsymbol{\sigma}^2_{MF}$, as well as intermediate features $\boldsymbol{s}_{MF}$.

The MF-CellEncoder is trained to reproduce the latent distribution parameters and intermediate features obtained from the CellVAE encoder. After training, 3D reconstruction is performed by feeding the output of MF-CellEncoder into the trained CellVAE decoder in a single forward pass.

MF-CellEncoder is trained using a distillation loss defined as follows:

$$\mathcal{L}_{\text{distill}} = \|\boldsymbol{\mu}_{MF} - \boldsymbol{\mu}_{3D}\|_2^2 + \left\|\log \boldsymbol{\sigma}^2_{MF} - \log \boldsymbol{\sigma}^2_{3D}\right\|_2^2 + \sum_{i=1}^{K} \left\|\boldsymbol{s}^{(i)}_{MF} - \boldsymbol{s}^{(i)}_{3D}\right\|_2^2, \quad (3)$$

where $\boldsymbol{\mu}_{MF}$ and $\boldsymbol{\sigma}^2_{MF}$ denote the mean and variance of the latent distribution produced by MF-CellEncoder, whereas $\boldsymbol{\mu}_{3D}$ and $\boldsymbol{\sigma}^2_{3D}$ are produced by the CellVAE encoder. Moreover, $\boldsymbol{s}^{(i)}_{MF}$ and $\boldsymbol{s}^{(i)}_{3D}$ represent the intermediate feature maps at the i-th $(i = 1, \ldots, K)$ encoder level. In this study, we set $K = 3$.

3 Experiments

3.1 Overview

We evaluated the proposed DeepCVR using multi-focus image sequences extracted from WSIs. Owing to privacy restrictions on clinical data, the full dataset cannot be publicly released. However, several sample images are available in the Cervical Cancer Cell Image Database: Multi-focus Cytology Dataset [9].

CellVAE and MF-CellEncoder were trained using 3D cell volumes reconstructed by the DIP-based method [2], generated with 500 optimization epochs. The dataset contains 320 pairs of multi-focus image sequences and their corresponding 3D reconstructions, with 80 samples from each of the NILM, LSIL, HSIL, and SCC cytological categories, which are defined according to the Bethesda

Table 1: Quantitative reconstruction results measured by PSNR, SSIM, and average reconstruction time.

Category	PSNR↑ [dB]	SSIM↑	Time↓ [msec]	DIP Time↓ [$\times 10^4$ msec]
NILM	45.34	0.9981	99.634	111.929
LSIL	41.72	0.9977	99.642	112.192
HSIL	43.88	0.9983	99.663	112.224
SCC	39.83	0.9954	99.683	111.941
Overall Average	42.99	0.9974	99.654	112.066

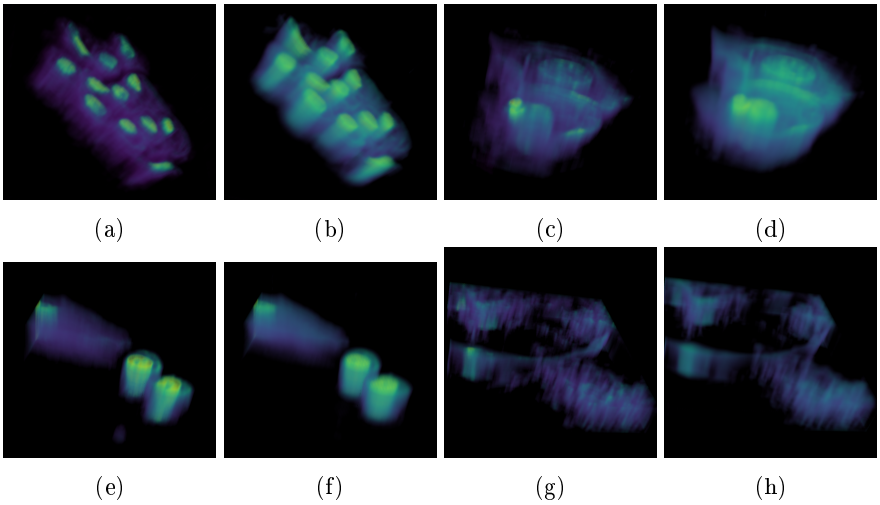
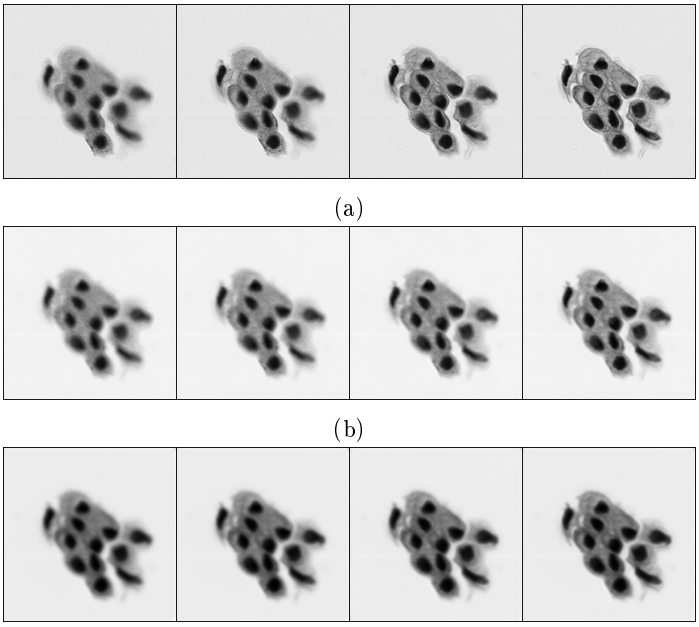


Fig. 4: Reconstruction results using the DIP-based method and DeepCVR. (a), (b) NILM; (c), (d) LSIL; (e), (f) HSIL; and (g), (h) SCC samples, where the left image in each pair shows the DIP-based reconstruction and the right image shows DeepCVR.

System [8]. This dataset was split into 256 training samples, 32 validation samples, and 32 test samples, ensuring that each category was equally represented in each split.

DeepCVR was evaluated on the test set using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and average reconstruction time. PSNR and SSIM were computed between the volumes reconstructed by DeepCVR and those reconstructed by the DIP-based method, using only the original 11 depths after removing the padded depth. These metrics assess the reconstruction quality of DeepCVR relative to the DIP-based method, whereas average reconstruction time assesses the computational efficiency of the proposed method.



(a)

(b)

(c)

Fig. 5: Multi-focus image sequence for a NILM sample: (a) original input, (b) simulated multi-focus images from the DIP-based reconstruction, and (c) simulated multi-focus images from the DeepCVR reconstruction.

3.2 Results

Table 1 shows that the proposed method achieves high PSNR and SSIM values with respect to the DIP-based reconstructions, indicating that DeepCVR produces volumes that closely approximate the DIP-based results. Moreover, the proposed method reduces the average reconstruction time per sample to 99.654 [msec], whereas the DIP-based method requires 112.066 $\times 10^4$ [msec].

Fig. 4 shows examples of reconstructed 3D cell volumes produced by DeepCVR and the DIP-based method. From the figures, the proposed method produces volumetric structures with cell shapes and structural details comparable to those obtained using the DIP-based method. Although mild stretching artifacts were observed in some cases, overall morphological consistency was preserved.

Multi-focus images from the reconstructed volumes simulated using the imaging model [15] are shown in Fig. 5. Since the imaging model predicts grayscale images, the simulated multi-focus images are shown in grayscale. They maintain similar focus transitions across depth while exhibiting slight blurring and intensity differences compared with the original input images.

3.3 Discussion

The qualitative and quantitative results indicate that the proposed method can reconstruct 3D volumes from a multi-focus image sequence with quality comparable to DIP-based reconstructions, while drastically reducing the computational cost of reconstruction. A key limitation of the DIP-based method is the need for per-sample iterative optimization. When optimized for 500 epochs per sample, DIP requires approximately 112.066 $\times 10^4$ [msec] per volume. In contrast, the proposed method performs reconstruction in approximately 99.654 [msec] with a single forward pass, achieving an approximately 1.1×10^4-fold speedup.

Fig. 4 (b) shows that the proposed method sometimes contains minor stretching artifacts in the reconstructed 3D volume. However, when examining the multi-focus images simulated from the reconstructed volumes, as shown in Fig. 5, the simulated images maintain focus transitions across depth similar to those of the original input images. Nevertheless, slight blurring was observed in some reconstructed structures. This effect mainly arises from the VAE-based reconstruction process, in which probabilistic latent representations introduce smoothing during decoding. While this smoothing improves reconstruction stability, it may reduce the sharpness of fine structural details compared with the DIP-based reconstruction. One solution for this limitation is to introduce sharper reconstruction constraints, such as edge-aware loss functions or hybrid architectures that better preserve structural features.

A limitation of this study is that we did not compare DeepCVR with a direct end-to-end reconstruction model. This study focuses on distilling DIP-based reconstruction into a fast feed-forward model. A comparison with end-to-end baselines will be included in our future work.

4 Conclusion

We proposed DeepCVR, an efficient method for 3D cell volume reconstruction from multi-focus image sequences. By training a student encoder to reproduce latent and intermediate representations learned from DIP-reconstructed volumes, the proposed method distills DIP-based reconstruction into a fast feed-forward model. Experimental results showed that DeepCVR approximates DIP-based reconstructions with comparable quality while reducing reconstruction time from 112.066 $\times 10^4$ [msec] to approximately 99 [msec] per sample. Future work will focus on further analysis of the blurring effects observed in simulated multi-focus images to improve reconstruction fidelity and integration with downstream cytological analysis tasks.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number JP25K24609.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allgeier, S., Bartschat, A., Bohn, S., Peschel, S., Reichert, K.M., Sperlich, K., Walckling, M., Hagenmeyer, V., Mikut, R., Stachs, O., Köhler, B.: 3d confocal laser-scanning microscopy for large-area imaging of the corneal subbasal nerve plexus. Scientific Reports **8** (05 2018). https://doi.org/10.1038/s41598-018-25915-6
2. Azevedo, C., Santra, S., Kumawat, S., Nagahara, H., Morooka, K.: Deep volume reconstruction from multi-focus microscopic images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 47–57. Springer (2024)
3. Jones, N., Nazarian, R., Duncan, L., Kamionek, M., Lauwers, G., Tambouret, R., Wu, C.L., Nielsen, G., Brachtel, E., Mark, E., Sadow, P., Grabbe, J., Wilbur, D.: Interinstitutional whole slide imaging teleconsultation service development: Assessment using internal training and clinical consultation cases. Archives of pathology & laboratory medicine **139**, 627–635 (11 2014). https://doi.org/10.5858/arpa.2014-0133-OA
4. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014)
5. Lewis, D., Chatoux, H., Mansouri, A.: Sff-rti: an active multi-light approach to shape from focus. The Visual Computer **40**, 1–13 (07 2023). https://doi.org/10.1007/s00371-023-02902-1
6. Liu, S., Weaver, D.L., Taatjes, D.J.: Three-dimensional reconstruction by confocal laser scanning microscopy in routine pathologic specimens of benign and malignant lesions of the human breast. Histochemistry and Cell Biology **107**, 267–278 (1997), https://api.semanticscholar.org/CorpusID:10939356
7. Matcha, N., Ramanarayanan, S., Al Fahim, M., GS, R., Ram, K., Sivaprakasam, M.: Sft-kd-recon: Learning a student-friendly teacher for knowledge distillation in magnetic resonance image reconstruction. In: Medical imaging with deep learning. pp. 1423–1440. PMLR (2024)
8. Nayar, R., Wilbur, D.C. (eds.): The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes. Springer, Cham, 3 edn. (2015). https://doi.org/10.1007/978-3-319-11074-5
9. Ohno, E., Ohno, S., Miyamoto, T., Yakushiji, H., Osawa, Y., Onishi, T., Nishimori, M., Shibahara, K.: Cervical Cancer Cell Image Database: Multi-focus Cytology Dataset (CCCID). Zenodo (2026). https://doi.org/10.5281/zenodo.18904734, https://doi.org/10.5281/zenodo.18904734
10. Pantanowitz, L., Sinard, J., Henricks, W., Fatheree, L., Carter, A., Contis, L., Beckwith, B., Evans, A., Otis, C., Lal, A., Parwani, A.: Validating whole slide imaging for diagnostic purposes in pathology guideline from the college of american pathologists pathology and laboratory quality center. Archives of pathology & laboratory medicine **137**, 1710–1722 (05 2013). https://doi.org/10.5858/arpa.2013-0093-CP
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

12. Subbarao, M., Choi, T.: Accurate recovery of three-dimensional shape from image focus. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(3), 266–274 (1995). https://doi.org/10.1109/34.368191
13. Tabata, K., Mori, I., Sasaki, T., Itoh, T., Shiraishi, T., Yoshimi, N., Maeda, I., Harada, O., Taniyama, K., Taniyama, D., Watanabe, M., Mikami, Y., Sato, S., Kashima, Y., Fujimura, S., Fukuoka, J.: Whole-slide imaging at primary pathological diagnosis: Validation of whole-slide imaging-based primary pathological diagnosis at twelve japanese academic institutes. Pathology International **67**(11), 547–554 (2017). https://doi.org/10.1111/pin.12590
14. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Deep image prior. International Journal of Computer Vision **128**, 1867 – 1888 (2017), https://api.semanticscholar.org/CorpusID:4531078
15. Yamaguchi, T., Nagahara, H., Morooka, K., Nakashima, Y., Uranishi, Y., Miyauchi, S., Kurazume, R.: 3d image reconstruction from multi-focus microscopic images. In: Dabrowski, J.J., Rahman, A., Paul, M. (eds.) Image and Video Technology. pp. 73–85. Springer (2020)
16. Zechmann, B., Möstl, S., Zellnig, G.: Volumetric 3d reconstruction of plant leaf cells using sem, ion milling, tem, and serial sectioning. Planta **255** (06 2022). https://doi.org/10.1007/s00425-022-03905-3
17. Zhao, L., Peng, X., Chen, Y., Kapadia, M., Metaxas, D.N.: Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6528–6537 (2020)