DiffStain: Conditioned Diffusion-Based Semantic Virtual Staining with Mask Guidance

Yikai Han[1][0009−0008−9874−9517], Jimao Jiang[2][0000−0003−1463−4445], and Yuru Pei✉[2][0000−0001−8520−3509]

[1] School of Computer Science and Engineering, Beihang University, Beijing 100191, China

[2] School of Intelligence Science and Technology, Key Laboratory of Machine Perception (MOE), State Key Laboratory of General Artificial Intelligence, Peking University, Beijing 100871, China

peiyuru@cis.pku.edu.cn

Abstract. Fluorescent staining is crucial for studying the morphology and dynamics of subcellular structures in biological and medical research, though being slow, expensive, and causing phototoxicity in live cells. Existing methods use deep generative models for image-to-image translation to generate diverse fluorescent images of subcellular structures. However, the pixel-level image generation approaches struggle to preserve fine structural details during the reconstruction process. In this paper, we introduce DiffStain, a novel approach that leverages mask-guided diffusion models for semantic virtual staining. The goal is to generate fluorescent images based on a brightfield input image. Rather than relying on deliberately selected image filters for subcellular structure segmentation, our approach employs an unsupervised deep neural spectral clustering method to combat the noisy and ambiguous structural boundaries. We also integrate mask guidance into the reverse denoising process, which helps highlight the regions of the subcellular structures that require precise representation in the generated fluorescent images. The masks produced by the spectral clustering model provide valuable feedback, enabling iterative refinements of the fluorescent images. Experiments showcase that our DiffStain method achieves state-of-the-art virtual staining performances on public microscopy datasets. Code is available at: https://github.com/StrengthInNumber/DiffStain.

Keywords: Conditioned diffusion model · Fluorescent image · Mask guidance · Semantic virtual staining.

1 Introduction

Fluorescence microscopy is essential for monitoring the morphology and dynamics of subcellular structures in biological and medical image analysis [24, 3, 22, 17]. However, fluorescence staining is expensive, time-consuming, and poses risks of phototoxicity and photobleaching, particularly in live cells [10, 21]. In silico painting, which pioneered virtual staining, uses pixel-to-pixel translation

to convert label-free transmitted light microscopy images into organelle-specific fluorescence images [16, 7]. Virtual staining offers non-invasive, non-destructive imaging, enabling the analysis of subcellular structure shape, function, and physiological characteristics [5, 18, 1, 12]. Compared to traditional staining, virtual staining improves acquisition speed and multiplexing capabilities, supporting downstream tasks of organelle detection and registration, which are key for generating statistical models of subcellular structures [25, 2]. Deep learning techniques, such as CNNs [23], U-Nets [11], conditional GANs [9, 19], and transformers [26], have been applied to label-free virtual staining. Task-aware priors [30] and sparse view schemes [29] have been explored for 3D subcellular structure prediction [12]. Wieslander et al. [27] employed dense U-Nets and GANs for cell painting, integrating parallel virtual staining with the segmentation of nuclei. Given the diverse morphologies and dynamics of subcellular structures, powerful image generators are desirable to highlight the subcellular structures in the output fluorescence images.

Diffusion models have gained prominence for image generation due to their superior performance through iterative denoising and flexible distribution modeling. Unlike GANs, which can experience mode collapse and training instability, diffusion models are easier to train and can generate high-quality, diverse images probabilistically, starting from Gaussian noise. However, editing style codes in diffusion models to achieve fine-grained structure reconstruction remains challenging. Conditioned diffusion models have been applied to fluorescent image generation [8, 14], where models minimize the distance to the input [6] or denoise from noisy input images [13]. Class-guided denoising diffusion probabilistic models have been used for fluorescence image reconstruction, with carefully prepared class priors guiding the denoising process [8]. However, pixel-wise image generation often modifies the entire image to align with the target domain's distribution, which can lead to the loss of local structural details.

In this paper, we introduce DiffStain, a novel framework for generating subcellular structure-specific fluorescent images from brightfield images, as shown in Fig. 1. DiffStain uses a conditioned diffusion approach, where subcellular structure masks guide the iterative denoising process. Instead of relying on preselected image filters for subcellular structure segmentation, we propose a deep neural spectral clustering (NSC) module to extract masks from fluorescent images. By leveraging pre-trained DINOViT features [15] and k-means clustering in the spectral embedding space, our unsupervised NSC model effectively identifies subcellular structures from noisy or ambiguous fluorescent images. To enhance the fluorescence image generation process, we incorporate mask guidance during online inference. The NSC-generated masks are fed back into the denoiser, ensuring the iterative denoising process highlights the subcellular structures of interest. We evaluate the effectiveness of the proposed DiffStain through extensive experiments on public microscopy datasets, demonstrating superior performances over existing methods. The main contributions of this work are as follows:

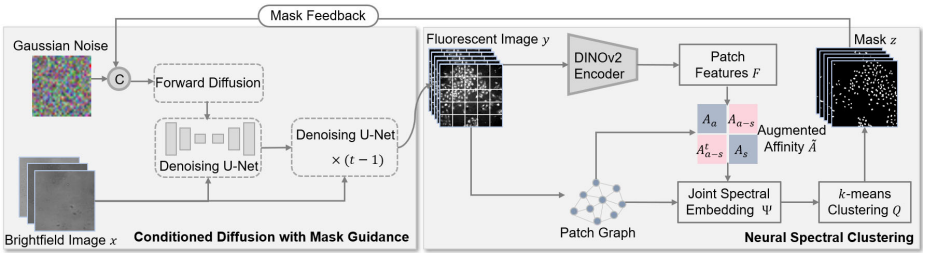

Fig. 1. Overview of our conditioned diffusion-based semantic virtual staining framework with mask guidance. The noisy brightfield images are used to condition the denoising process to generate multi-channel fluorescence images specific to various subcellular structures. We present an unsupervised deep neural spectral clustering (NSC) module to generate masks, which serves as a guide of the denoiser to highlight the structures of interest in the online inference of fluorescence images. NSC takes advantage of the pre-trained DINOViT features to build a patch graph and compute the joint spectral embedding using an augmented affinity matrix regarding an anchor image, followed by the k-means clustering in the spectral embedding space.

- We present DiffStain, an efficient mask-guided conditioned diffusion model for generating subcellular structure-specific fluorescence images from brightfield images.
- We introduce an unsupervised NSC-based masking scheme, which enables the efficient identification of subcellular structures, and incorporates the mask guidance into the denoising process to enhance online virtual staining by highlighting subcellular structures of interest.
- We validate DiffStain through extensive experiments, demonstrating its superiority over state-of-the-art methods on public microscopy datasets.

2 Method

Fig. 1 provides an overview of our proposed method, which is based on a diffusion-based image translation framework that follows the corrupt and reconstruct paradigm. Considering the diverse morphologies of subcellular structures, the iterative conditioned denoising process can lead to noisy and ambiguous structural boundaries in the fluorescent images. To preserve large amounts of fine-grained subcellular structures during the virtual staining process, we introduce a mask-guided denoising scheme by using structural masking to emphasize regions of interest. In each iteration, subcellular shapes embedded in the spatial masks are used to guide the reconstruction of fluorescence images, with these masked areas indicating various types of subcellular structures. We present an unsupervised deep neural structure clustering (NSC) model for subcellular structure segmentation in a low-dimensional spectral embedding space by leveraging the eigen-decomposition of a graph Laplacian matrix, which is robust to high-frequency noise and image artifacts. The estimated masks provide feedback during online inference for iterative refinement of fluorescent images.

2.1 Conditioned Diffusion Model for Virtual Staining

Diffusion models enable diverse image generation by progressively denoising random Gaussian noise to produce images that match a target distribution. We employ the conditional diffusion models [20] for the image-to-image translation from the brightfield image x to multi-channel fluorescent images y by denoising an input image with a form of p(y|x). The diffusion model consists of two main processes: the forward diffusion process and the reverse denoising process. In the forward process, Gaussian noise is iteratively added to the image in a Markovian fashion. The forward process is defined as: $q(y_t|y_{t-1}) := \mathcal{N}(y_t; \sqrt{1-\beta_t}y_{t-1}, \beta_t\mathbf{I})$, where β_t denotes the noise variance at step t. The image at time step y_t is conditioned on the original image y_0, and the transition is defined as: $q(y_t|y_0) = \mathcal{N}(y_t; \sqrt{\alpha_t}y_0, (1-\alpha_t)\mathbf{I})$, where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$ represents the cumulative effect of noise over time.

In the reverse denoising process, the model aims to reconstruct the original image by training a neural network ϵ_θ to predict the noise at each step. The reverse process is described by the conditional probabilistic distribution: $p_\theta(y_{t-1}|y_t) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, x, t), \beta_t\mathbf{I})$, where $\mu_\theta(y_t, x, t)$ is the learnable mean of the distribution. Using Langevin dynamics to estimate the gradient of the data log-likelihood, the iterative reverse denoising process can be written as:

$$y_{t-1} = \frac{1}{\sqrt{\beta_t}}\left(y_t - \frac{1-\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x, y_t, \alpha_t)\right) + \sqrt{1-\beta_t}\epsilon_t. \tag{1}$$

The neural network ϵ_θ is trained by minimizing the distance between the actual noise and the predicted noise at each time step, and the loss function $\mathcal{L}_{dm} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,(x,y),\alpha}\left\|\epsilon_\theta^{(t)}(x, y_t, \alpha) - \epsilon\right\|_1$. The loss function corresponds to maximizing the likelihood with respect to the weighted variational lower bound. By minimizing the loss function, both $\mu_\theta(y_t, x, t)$ and $\epsilon_\theta^{(t)}(x, y_t, \alpha)$ are optimized to conduct the image translation and produce high-fidelity fluorescent images.

2.2 Neural Spectral Clustering

We introduce NSC, an unsupervised method for subcellular structure segmentation from fluorescence images in the spectral embedding space. Unlike traditional interactive image filters and operators, NSC leverages a pre-trained DINOViT features [15] and spectral clustering to combat high-frequency perturbations in identifying fine-grained structures. The workflow of NSC is shown in Fig. 1. Given a fluorescence image $y \in \mathbb{R}^{m\times n\times l}$, the goal is to estimate a subcellular structure mask $z \in \{0,1\}^{m\times n\times l}$, where l represents the channel number corresponding to various subcellular structures.

First, we build a patch graph from the fluorescence image using pre-trained DINOViT features, which capture long-range relationships between repetitive fine-grained structures through self-attention. To account for the fine granularity of subcellular structures, the image is subdivided into small fields of view (FOVs) image $y_s \in \mathbb{R}^{q\times q}$, where the patch size is comparable to subcellular structures.

Next, we perform eigendecomposition on the graph Laplacian matrix of the patch graph. Rather than conducting independent spectral clustering on each small FOV image, we apply joint spectral embedding to avoid spectral distortion and inconsistent cluster assignments across the small FOV image. This is achieved by augmenting the affinity matrix with information from an anchor image y_a, ensuring consistent subcellular structure identification across images. The augmented affinity matrix is defined as: $\tilde{A} = \begin{bmatrix} A_a & A_{a-s} \\ A_{a-s}^T & A_s \end{bmatrix}$, which accounts for both inter- and intra-image patch-wise relationships. A_a and A_s denote the affinity matrix of the anchor and the small FOV images. A_{a-s} denote the patch-wise affinity between y_a and y_s. The augmented affinity matrix $\tilde{A} \in \mathbb{R}^{2n_p \times 2n_p}$ is calculated using cosine similarity, and $\tilde{A}_{ij} = \frac{F_i F_j^T}{\|F_i\|_2 \|F_j\|_2} \odot (F_i F_j^T > 0)$, where F_i and F_j represent the DINOViT features of patches i and j. n_p denotes the patch number of a small FOV image. The normalized Laplacian matrix $L = I - D^{-1/2}\tilde{A}D^{-1/2}$, where D is the degree matrix with $D_{ii} = \sum_j \tilde{A}_{ij}$, is used for spectral clustering. The eigenvectors $\Psi \in \mathbb{R}^{2n_p \times r}$ corresponding to the first r non-zero eigenvalues provide the spectral embedding of the fluorescence images.

Finally, we apply k-means clustering to the spectral embedding and generate a clustering assignment matrix $Q \in \{0,1\}^{2n_p \times k}$. Using the shared anchor image, we synchronize cluster assignments across small FOV images through a cross-image label mapping function $\theta_{i,j} : \{1, \ldots, k\} \to \{1, \ldots, k\}$ from image i to image j, and $\theta_{i,j}(u) = v$ when $v = \arg\max_{v^*} \lfloor Q_i \rfloor_u \cdot \lfloor Q_j \rfloor_{v^*}$. operator $\lfloor \cdot \rfloor_u$ returns the first n_p dimensional vector of column u. k-means clustering on Ψ is robust to high-frequency noise in the fluorescence images, ensuring reliable segmentation. Moreover, the joint spectral embedding via an anchor image ensures consistent clustering label assignments, avoiding additional cross-image cluster synchronization.

2.3 Mask Guided Denoising

In pixel-wise image translation for virtual staining of multi-channel fluorescence images, it is crucial to preserve fine-grained subcellular structures during the image generation process. We introduce a mask guidance scheme with respect to various subcellular structures, which enhances the iterative denoising process by put focus on subcellular shapes embedded in the NSC-based masking. We add noise to the mask obtained by NSC through the forward process of the diffusion model, and take the noised mask as y_t for the reverse denoising. The mask-guided process is defined as:

$$y_t = \sqrt{\alpha_t} \cdot z + \sqrt{1 - \alpha_t} \cdot \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I). \tag{2}$$

The mask z obtained by NSC can filter the high-frequency noise in the subcellular structure image generated at the early stages, and the low-frequency information such as the general contours and semantic features can be retained in the forward process. The mask guidance enhances pixel generation in the regions of interest, allowing for semantic-aware denoising and further plausible

generation of subcellular structural details. The semantically aware denoising approach ensures that the final output highlights subcellular structures by modeling both the distribution of the input brightfield image and the desired features of the target fluorescence images.

3 Experiments

Datasets and Metrics. We evaluated the proposed approach using the publicly available JUMP Cell Painting dataset (cpg0000) from the Cell Painting Gallery on the Registry of Open Data on AWS [4]. We use ten plates with each representing different biological phenotypes. Each plate contains 2,000 images with five fluorescent image channels, including nucleus (DNA), endoplasmic reticulum (ER), cytoplasmic RNA (RNA), Actin, Golgi, plasma membrane (AGP), and mitochondria (Mito), as well as three-channel light field images. Specifically, the images were corrected using the appropriate illumination correction function, after which they were normalized and re-sampled to a resolution of 512×512. A maximum pixel intensity cutoff was applied to exclude extreme outliers. The dataset was randomly split, using nine plates for training and one for testing.

For evaluation, we used four metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), and Structural Similarity Index Measure (SSIM), to assess the consistency between the generated fluorescence images and the ground truth.

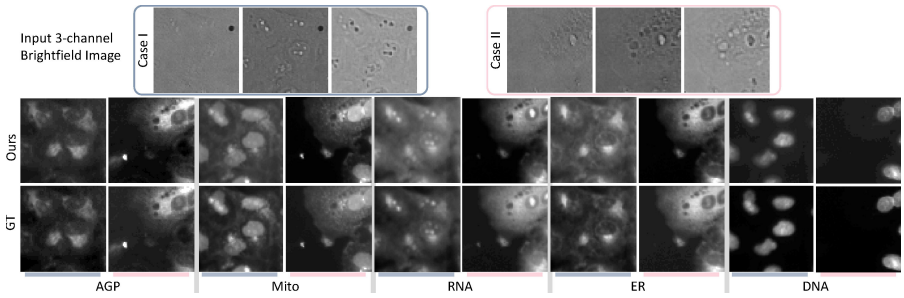

Fig. 2. Two sampled cases of fluorescence image prediction regarding five types of subcellular structures. The ground truth is shown side-by-side.

Implementation Details. The proposed approach was implemented on a machine with an NVIDIA 3090 GPU, utilizing the PyTorch framework. The batch size and the learning rate are set to 1 and $5e^{-4}$. We use a patch graph with n_p=4096 nodes. The resolution of the small FOV image is set to 224×224, as the pre-trained DINOv2 model [15] and q=224. The patch size is set to 8×8. The eigenvector number r and cluster number k are set to 5. The training process required approximately 200 iterations and 8 hours. During inference, the

denoising step was set to 2,000 iterations. The average inference of a 512×512 five-channel fluorescence image requires 3 minutes.

Results. We first demonstrate the efficacy of the proposed approach on virtual staining on brightfield images. Fig. 2 shows the side-by-side comparison of the predicted fluorescence images and the ground truth. Our method exhibits consistency on all five channels with the ground truth with an average PCC of 0.862 and an SSIM of 0.636 (Table 1). Note that the proposed DiffStain is feasible to identify fine-grained subcellular structures with a large variety of morphologies and orientations from the input brightfield images. Moreover, our method retains the structural shape and distributions during generating fluorescence images.

Table 1. Quantitative results on virtual staining by compared methods. *per* and *tar* denote using labels of perturbations and targets respectively. † denotes using masks from adaptive thresholding.

	Average				RNA				ER			
	MSE↓	MAE↓	PCC↑	SSIM↑	MSE↓	MAE↓	PCC↑	SSIM↑	MSE↓	MAE↓	PCC↑	SSIM↑
U-Net [11]	0.016	0.095	0.818	0.511	0.015	0.081	0.847	0.560	0.016	0.096	0.830	0.514
DoDNet [28]	0.015	0.091	0.820	0.550	0.013	0.080	0.864	0.610	0.015	0.097	0.824	0.546
Class$_{per}$ [8]	0.029	0.131	0.782	0.409	0.019	0.119	0.868	0.510	0.020	0.123	0.820	0.482
Class$_{tar}$ [8]	0.030	0.135	0.780	0.401	0.019	0.119	0.860	0.519	0.019	0.119	0.824	0.479
Palette [20]	0.025	0.129	0.801	0.465	0.019	0.111	0.868	0.526	0.020	0.119	0.832	0.485
Palette† [20]	0.015	0.089	0.831	0.568	**0.011**	0.078	0.898	0.649	0.015	0.095	0.851	0.600
DiffStain	**0.014**	**0.088**	**0.862**	**0.636**	**0.011**	**0.076**	**0.914**	**0.696**	**0.014**	**0.093**	**0.881**	**0.655**

Comparison. We summarize the main comparison results of virtual staining on fluorescence images regarding five subcellular structures in Table 1 and Fig. 3. Our DiffStain is compared with state-of-the-art baseline approaches, including U-Net [11], DoDNet [28], Palette [20], and class-guided diffusion [8]. DiffStain benefits from the mask guidance and is effective in generating fluorescence images with emphasis on structures of interest, yielding sharper inter-structure boundaries and consistently outperforming compared methods across all reported metrics. We notice that the images generated using convolutional encoder-decoder models are limited to account for fine-grained structures [11, 28], where the U-Net and task-coded decoder learning tend to produce smooth and blurry structural contours. The task coding concerning the various subcellular classes is feasible to relieve model complexity by learning one decoder to generate multi-channel fluorescence images, though the simple concatenation of task codes with the feature embedding is limited in generating high-quality images.

The diffusion model is feasible to produce high-quality and natural fluorescence images conditioned on input brightfield data which is able to generate structures in alignment with the ground truth. However, it is not a trivial task to capture the structures of interest via the pixel-level generation. We found phantoms in the conditioned diffusion model Palette [8], where the generated images bear irrelevant structures not exist in the input brightfield images. The class la-

bels provide additional information for fluorescence image generation, though the class label serves as a global prior and cannot ensure to produce local fine-grained structural details [8]. Moreover, we observe performance gains for the conditioned diffusion model with mask guidance. For instance, the Palette† with the mask guidance achieves an SSIM gain of 0.013. By incorporating rich semantic information into the denoising process, our method outperforms the compared diffusion model-based staining methods with PCC gains of 0.031 (Palette†), 0.061 (Palette), 0.082 (Class$_{tar}$), and 0.080 (Class$_{per}$). The proposed DiffStain exploits the conditioned diffusion models and mask feedback from the unsupervised NSC model to ensure accurate and detailed subcellular structure identification and fluorescence image generation.

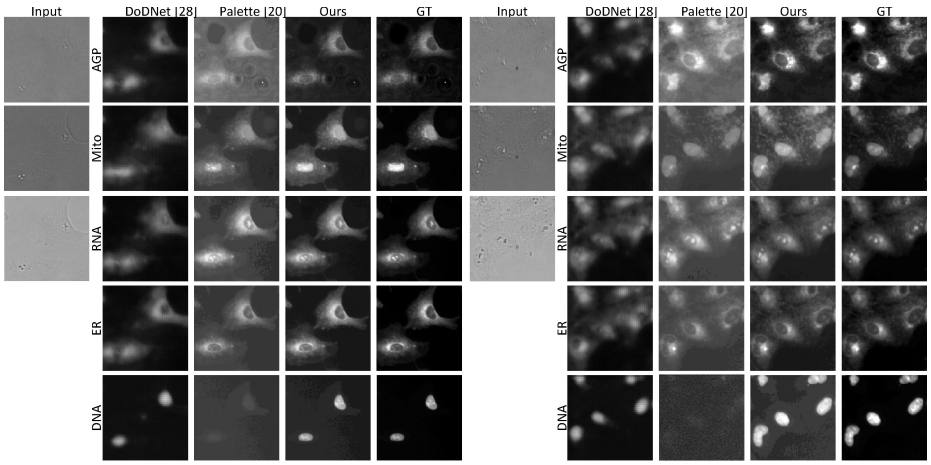

Fig. 3. Qualitative comparison for five-channel fluorescent images prediction from brightfield images.

Ablation Study. We assess the effectiveness of the mask guidance in Table 1. We remove the mask guidance from the conditioned diffusion model as the Palette, where the brightfield image is used as the input to generate the fluorescence images. Removing the mask guidance leads to a drop of 0.171 in SSIM and 0.061 in PCC, highlighting the importance of the mask guidance. Note that our approach does not require the manual selection of image operators for mask generation. In contrast, the proposed NSC method effectively identifies various subcellular structures by leveraging joint spectral embedding, demonstrating robustness to high-frequency noise in fluorescence images. When compared to Palette†, which uses masks derived through adaptive thresholding, the proposed DiffStain with NSC-masking achieves SSIM improvements of 0.047 and 0.055 for RNA and ER-related fluorescence images, respectively, demonstrating the merit of NSC-masking to guide the denoising process.

4 Conclusion

We presented DiffStain, a novel diffusion model designed for the virtual staining of brightfield images. Our method incorporates mask guidance with rich semantic information into the iterative denoising process, improving the identification of subcellular structures and enabling the generation of high-quality multi-channel fluorescence images. Specifically, our work is inspired by recent advancements in diffusion model-based virtual staining [8]. We augment these methods with the learnable NSC masking scheme, allowing plausible mask inference insensitive to high-frequency perturbation, as well as efficient mask feedback to the denoising process. Extensive experiments on microscopy datasets demonstrate the superiority of our method over previous approaches in virtual staining tasks.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under Grant 62272011 and Beijing Natural Science Foundation 7232337.

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

References

1. Bai, B., Yang, X., Li, Y., Zhang, Y., Pillar, N., Ozcan, A.: Deep learning-enabled virtual histological staining of biological samples. Light, Science & Applications **12** (2022)
2. Cai, Y., Cai, Y., Hossain, M.J., Hériché, J.K., Politi, A.Z., Politi, A.Z., Walther, N., Koch, B., Koch, B., Wachsmuth, M., Nijmeijer, B., Kueblbeck, M., Martinic-Kavur, M., Ladurner, R., Ladurner, R., Alexander, S., Peters, J.M., Ellenberg, J.: Experimental and computational framework for a dynamic protein atlas of human cell division. Nature **561**, 411 – 415 (2018)
3. Carlton, J.G., Jones, H., Eggert, U.S.: Membrane and organelle dynamics during cell division. Nature Reviews Molecular Cell Biology **21**, 151–166 (2020)
4. Chandrasekaran, S.N., Cimini, B.A., Goodale, A., Miller, L., Kost-Alimova, M., Jamali, N., Doench, J.G., Fritchman, B., Skepner, A., Melanson, M., Arevalo, J., Caicedo, J.C., Kuhn, D., Hernandez, D., Berstler, J., Shafqat-Abbasi, H., Root, D.E., Swalley, S., Singh, S., Carpenter, A.E.: Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nature Methods **21**, 1114 – 1121 (2022)
5. Cheng, S., Fu, S., Kim, Y.M., Song, W., Li, Y., Xue, Y., Yi, J., Tian, L.: Single-cell cytometry via multiplexed fluorescence prediction by label-free reflectance microscopy. Science Advances **7** (2020)
6. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 14347–14356 (2021)
7. Christiansen, E.M., Yang, S.J., Ando, D.M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O¡¯Neil, A., Shah, K., Lee, A.K., et al.: In silico labeling: predicting fluorescent labels in unlabeled images. Cell **173**, 792–803.e19 (2018)

8. Cross-Zamirski, J.O., Anand, P., Williams, G.B., Mouchet, E., Wang, Y., Schönlieb, C.B.: Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) pp. 3802–3811 (2023)
9. Cross-Zamirski, J.O., Mouchet, E., Williams, G.B., Schönlieb, C.B., Turkki, R., Wang, Y.: Label-free prediction of cell painting from brightfield images. Scientific Reports **12** (2021)
10. Icha, J., Weber, M., Waters, J.C., Norden, C.: Phototoxicity in live fluorescence microscopy, and how to avoid it. BioEssays **39** (2017)
11. Imboden, S., Liu, X., Payne, M.C., Hsieh, C.J., Lin, N.Y.C.: Trustworthy in silico cell labeling via ensemble-based image translation. Biophysical Reports **3** (2023)
12. Jo, Y., Cho, H.S., Park, W.S., Kim, G., Ryu, D., Kim, Y.S., Lee, M., Park, S., Lee, M.J., Joo, H., Jo, H., Lee, S.G., Lee, S., Min, H.S., Heo, W.D., Park, Y.: Label-free multiplexed microtomography of endogenous subcellular dynamics using generalizable deep learning. Nature Cell Biology **23**, 1329 – 1337 (2021)
13. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2021)
14. Navidi, Z., Ma, J., Miglietta, E.A., Liu, L., Carpenter, A., Cimini, B.A., Haibe-Kains, B., Wang, B.: Morphodiff: Cellular morphology painting with diffusion models. bioRxiv (2024)
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. ArXiv **abs/2304.07193** (2023), `https://api.semanticscholar.org/CorpusID:258170077`
16. Ounkomol, C., Seshamani, S., Maleckar, M.M., Collman, F., Johnson, G.R.: Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Nature Methods **15**, 917–920 (2018)
17. Prinz, W.A., Toulmay, A., Balla, T.: The functional universe of membrane contact sites. Nature Reviews Molecular Cell Biology **21**, 7–24 (2020)
18. Rivenson, Y., Liu, T., Wei, Z., Zhang, Y., Ozcan, A.: Phasestain: the digital staining of label-free quantitative phase microscopy images using deep learning. Light, Science & Applications **8** (2018)
19. Rivenson, Y., Wang, H., Wei, Z., de Haan, K., Zhang, Y., Wu, Y., Günaydn, H., Zuckerman, J.E., Chong, T., Sisk, A.E., Westbrook, L., Wallace, W.D., Ozcan, A.: Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nature Biomedical Engineering **3**, 466 – 477 (2018)
20. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. ACM SIGGRAPH 2022 Conference Proceedings (2021)
21. Scherf, N., Huisken, J.: The smart and gentle microscope. Nature Biotechnology **33**, 815–818 (2015)
22. Scorrano, L., De Matteis, M.A., Emr, S., Giordano, F., Hajn¨®czky, G., Kornmann, B., Lackner, L.L., Levine, T.P., Pellegrini, L., Reinisch, K., et al.: Coming together to define membrane contact sites. Nature Communications **10**, 1287 (2019)
23. Valen, D.A.V., Kudo, T., Lane, K.M., Macklin, D.N., Quach, N.T., Defelice, M., Maayan, I., Tanouchi, Y., Ashley, E.A., Covert, M.W.: Deep learning automates

the quantitative analysis of individual cells in live-cell imaging experiments. PLoS Computational Biology **12** (2016)
24. Valm, A.M., Cohen, S., Legant, W.R., Melunis, J., Hershberg, U., Wait, E., Cohen, A.R., Davidson, M.W., Betzig, E., Lippincott-Schwartz, J.: Applying systems-level spectral imaging and analysis to reveal the organelle interactome. Nature **546**, 162–167 (2017)
25. Viana, M.P., et al.: Integrated intracellular organization and its variations in human ips cells. Nature **613**, 345 – 354 (2023)
26. Wang, Z., Xie, Y., Ji, S.: Global voxel transformer networks for augmented microscopy. Nature Machine Intelligence **3**, 161 – 171 (2020)
27. Wieslander, H., Gupta, A., Bergman, E., Hallström, E., Harrison, P.J.: Learning to see colours: Biologically relevant virtual staining for adipocyte cell images. PLoS ONE **16** (2021)
28. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1195–1204 (2020), `https://api.semanticscholar.org/CorpusID:227118834`
29. Zheng, J., Ding, Y., Liu, Q., Cao, Y., Hu, Y., Wang, Z.: Sparsessp: 3d subcellular structure prediction from sparse-view transmitted light images. ArXiv **abs/2407.02159** (2024)
30. Zhou, D., Gu, C., Xu, J., Liu, F., Wang, Q., Chen, G., Heng, P.A.: Repmode: Learning to re-parameterize diverse experts for subcellular structure prediction. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3312–3322 (2022)