arXiv:2301.10227v2 [eess.IV] 8 Aug 2023

用于生成逼真且完整标注的显微镜图像数据集的去噪扩散概率模型

Dennis Eschweiler[1*]、R¨uveyda Yilmaz[1]、、Matisse Baumann[1]、、Ina Laube[1]、、Rijo Roy[1]、、Abin Jose[1]、、Daniel Br¨uckner[1] 和 Johannes Stegmaier[1*]

[1]亚琛工业大学成像与计算机视觉研究所，德国亚琛。

*通讯作者。电子邮箱：dennis.eschweiler@lfb.rwth-aachen.de；johannes.stegmaier@lfb.rwth-aachen.de；贡献作者：rueveyda.yilmaz@lfb.rwth-aachen.de；matisse.baumann@lfb.rwth-aachen.de；ina.laube@lfb.rwth-aachen.de；rijo.roy@lfb.rwth-aachen.de；abin.jose@lfb.rwth-aachen.de；daniel.brueckner@lfb.rwth-aachen.de；

摘要

计算机视觉领域的最新进展显著推动了逼真图像数据的生成，其中去噪扩散概率模型被证明是一种尤为有效的方法。本研究展示了扩散模型能够通过一种无监督且直观的方法，有效生成完全标注的显微镜图像数据集；该方法以所需结构的粗略草图为起点。所提出的流程有助于减少训练基于深度学习的分割方法时对人工标注的依赖，并且无需人工标注即可对多样化的数据集进行分割。正如我们通过涉及不同生物体和细胞类型的多种实际实验示例所展示的那样，该方法在简化数据生成过程，以及实现更高效、更具可扩展性的分割模型训练方面前景广阔。

2 用于生成显微镜图像数据的扩散模型

实现荧光显微镜图像数据中细胞的自动分割，是支持生物医学专家开展各类实验的重要一步 [1, 2]。实验设置的多样性也反映在图像数据的外观上，这给基于通常种类有限的已标注图像数据训练的分割方法带来了挑战。要克服这一挑战，就必须获取成本高昂且耗时费力的人工标注，这成为充分发挥基于学习的方法潜力的瓶颈，也限制了它们在实践中的应用。自动数据增强方法[3–5] 以及经过调整的分割流程 [6, 7], 能够减轻这一挑战，减少标注工作，但仍需要少量完全标注的图像数据作为基础。另一种方法是通过自动模拟方法，在任意数量的图像数据中复现细胞结构所需的特征 [8–13] 理想情况下，这种方法可以彻底取代人工标注。

近期，去噪扩散概率模型（DDPM） [14] 在生成逼真的图像数据方面展现出巨大潜力，而且无需已标注训练数据，也不需要像常用的生成对抗网络（GAN） [9, 10, 13]那样采用对抗式训练概念。我们提出了一种流程，利用 DDPM 自动生成完整标注的图像数据，随后用这些数据训练分割方法（图1 a）。在 DDPM 中，渐进式前向过程被定义为马尔可夫链：通过迭代地向真实图像 $\mathbf{x}_0$ 中添加少量噪声，直至得到纯噪声 $\mathbf{x}_T$ [14]。与之对应的反向过程则训练神经网络迭代地逆转前向过程，从而由 $\mathbf{x}_T$ [15]生成逼真的图像数据 $\hat{\mathbf{x}}_0$ 。然而，从纯噪声 $\mathbf{x}_T$开始无法生成完整标注的数据集，因为生成结构缺乏可控性，也没有相应的标注。为解决这些问题，我们对 DDPM 的应用进行了两项调整。首先，在 $t_{\text{start}} < T$ 处提前启动反向过程，确保 $\mathbf{x}_{t,\ \text{start}}$ 中的很大一部分结构指示尚未被噪声完全遮蔽，并能在生成逼真的图像数据时得以保留 [16]。其次，为了直观地建模和控制生成图像数据中的细胞结构，我们以草图取代真实图像数据，作为生成 $\mathbf{x}_{t,\ \text{start}}$的前向过程的起点。这些草图可以模拟生成，并提供细胞位置、形状及粗略结构特征的指示，从而简要勾勒出所需生成的场景。由于学习得到的反向过程仅使用真实图像数据进行训练，因此将其应用于 $\mathbf{x}_{t,\ \text{start}}$ 后，便会生成相应的逼真图像数据（图1b）。最终，借助草图中已知的细胞轮廓和位置，该流程能够以自动、无监督的方式生成逼真的完整标注图像数据。然而，为了在反向过程中保持高度真实性，来自草图域 $\mathcal{M}$的噪声样本 $\mathbf{x}_{t,\ \text{start}}$ 的数据分布必须与来自

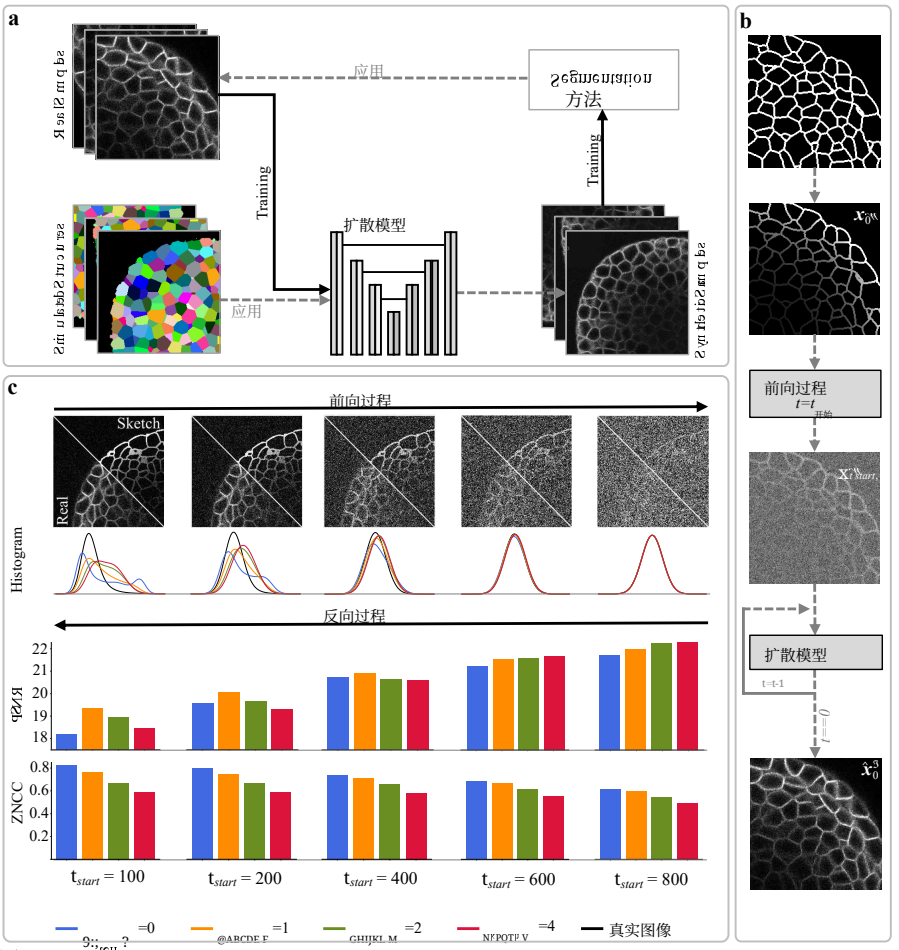


图 1 流程优化。(a) 整个流程包括在真实图像数据上训练扩散模型，并将其应用于模拟结构以生成完全标注的图像数据集；随后利用这些数据集训练模型，对真实数据进行分割。(b) 在应用去噪扩散概率模型时，标注会自动转换为粗略草图，以便后续应用前向过程，从而生成相应图像数据的逼真图像。(c) 由前向过程从真实图像或草图生成的含噪数据必须足够相似，才能在反向过程中生成逼真的图像数据。我们采用峰值信噪比 (PSNR) 和零归一化互相关 (ZNCC) 作为指标，评估从不同起点 t_{start} 和草图模糊因子 σ 生成的逼真图像数据。

真实图像域 $\mathcal{I}$的数据分布非常相似。这给参数 t_{start}最优取值的确定带来了优化方面的挑战。

优化 t_{start} 需要在生成精细细节与保留草图的结构相关性之间取得平衡。为了生成精细细节，前向过程后期需要大量噪声，但这可能会损害结构指示的保留；而要保留结构指示，则需要尽可能早地停止前向过程。我们使用一个公开的、完全标注的三维

显微镜图像数据集 [17], 对这一权衡的细节进行了评估。该数据集提供了经人工校正的标注，使我们能够精确评估所提流程的各个方面。为了分析前向过程，我们生成了 $\mathbf{x}_{\mathrm{t}}^{\mathcal{M}}$, start 和 $\mathbf{x}_{\mathrm{t}}^{\mathcal{I}}$, start 并通过构建数据分布直方图来测量它们的相似性。此外，为了评估学习到的反向过程，我们使用所有含噪样本生成了逼真的 $\hat{\mathbf{x}}_0$, ，并使用峰值信噪比（PSNR）定量评估纹理真实性，使用零归一化互相关（ZNCC）定量评估结构保真（图1c）。此外，我们发现，在应用前向过程之前对草图进行标准差为 σ 的高斯平滑，有助于避免边缘过于锐利，并更早地达到相似的数据分布，从而能够选择更早的 t_{start}。一般而言，t_{start} 的最佳值应尽可能早地设定，以确保草图与图像数据之间的结构相关性最大。此外，由于生成过程是迭代进行的，选择更早的 t_{start} 可直接缩短生成时间。我们通过实验证实，$t_{\text{start}} = 400$和 $\sigma = 1$ 能够在生成能力与结构保真之间实现良好的平衡。

虽然在参数优化过程中使用了手动标注结构，以便在不受结构差异影响的情况下测试流程的生成能力，但使用手动标注数据并不能代表生成完全标注图像数据集的实际场景。为评估流程中实现自动化的不同策略，我们采用粗略分割作为来源，为各种生物体和细胞类型生成草图。我们使用公开可用的通用分割方法 [22, 23]和公开可用的银标准注释 [19] ，为秀丽隐杆线虫 [19, 24], 、斑马鱼 [20],、小鼠干细胞 [19, 25], 和 HeLa细胞 [19, 21, 26] 中的细胞核获取细胞形状的粗略表示，并为斑马鱼 [20]中的细胞膜获取粗略表示。我们使用峰值信噪比（PSNR）作为指标，在有标注的区域评估图像质量；所有数据集的平均分数介于 19.58 dB 和 29.97 dB 之间。尽管标注中可能存在错误，从而影响报告的质量分数，但这些错误不会对所提出流程的应用产生显著影响。这是因为生成的图像数据与标注中的结构直接相关，因此能够创建无错误的图像-掩膜对。

依赖分割结果来收集逼真的结构，可能会限制流程的可扩展性，因为这取决于准确的通用分割方法是否可用。另一方面，模拟虽然情况更加复杂，但在泛化能力和可扩展性方面具有更大的潜力。主要挑战在于找到一种模拟技术，能够准确再现真实图像数据中可见的结构特征，因为任何不准确之处都可能导致真实图像数据与最终合成图像数据之间出现领域差距。尽管存在这些挑战，采用模拟方法开展实验有助于探索该流程的全部潜力和局限。由于缺少对应的真实图像数据，难以直接评估由模拟草图生成的图像数据的真实性，

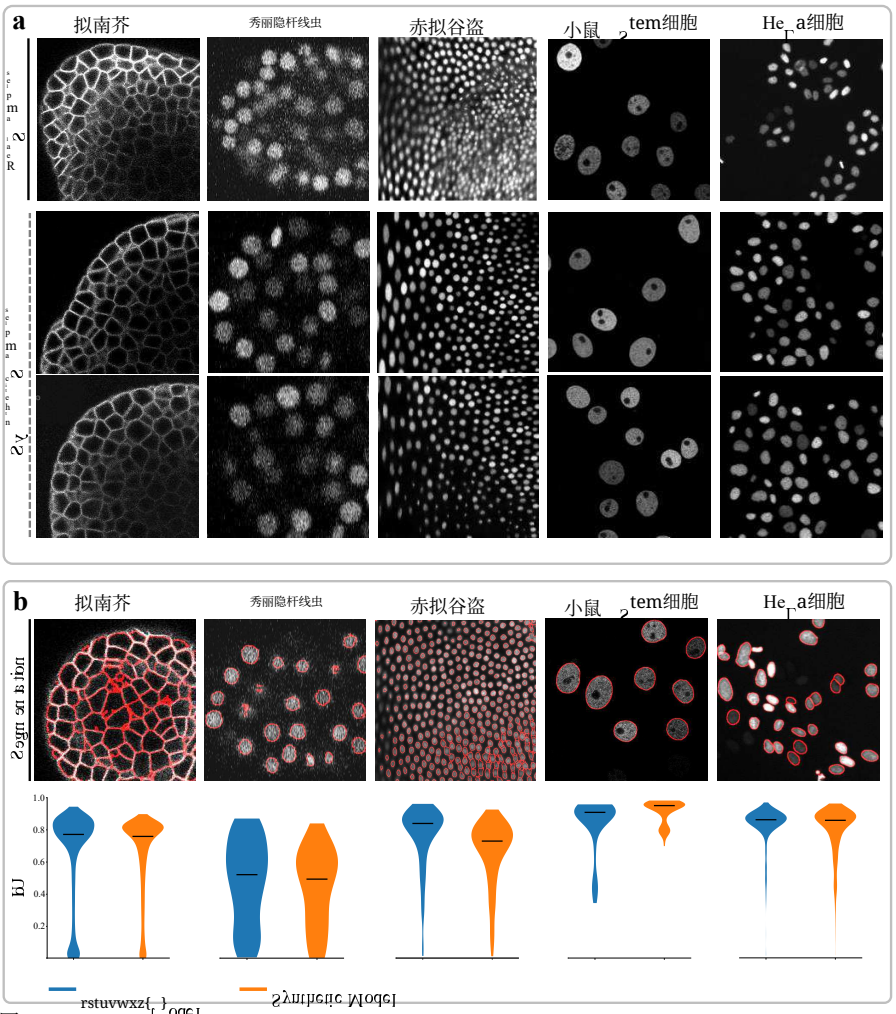


图2 应用 (a) 使用模拟结构的扩散模型生成的真实图像样本和全合成图像样本。(b) Cellpose 分割方法 [18] 在合成数据集上训练，并应用于真实图像数据，无需人工生成的标注即可生成结果（红色叠加层）。图中以小提琴图展示了交并比（IoU）得分：蓝色表示在大量人工标注图像数据上训练的公开通用模型，橙色表示在合成数据上训练的模型；黑色横线表示中位数。所有数据集均可从以下地址公开获取 [10, 17, 19–21]。

因此，我们采用了一种更具实践性的评估方式。相反，我们重点考察生成的数据作为训练数据的可用性：从头开始训练 Cellpose 方法 [18] ，随后将其应用于真实图像数据。通过将分割结果的准确性作为替代指标，我们间接评估了生成数据的真实性。此外，我们还将分割结果与公开可用的预训练 Cellpose 模型所获得的结果进行了比较；该模型既作为基线，也为在大型、多样且完全标注的图像数据集上训练的模型提供了参考。模拟结果针对五种

不同数据集中的细胞膜，包括拟南芥，以及秀丽隐杆线虫 [19, 24]，赤拟谷盗 [19]，小鼠干细胞 [19, 25] 和 HeLa细胞 [19, 26]（图2a）。图2b所示结果表明，尽管仅使用200个生成样本训练的合成图像数据模型，其训练数据仅包含合成图像数据，但两个模型在各数据集上的表现都同样出色。值得注意的是，合成模型无需任何人工生成的标注，就取得了相当的分数，无需像通用模型那样依赖大量已标注图像数据[18]。这一能力使分割模型有望应用于完全不同的数据集和结构；对于这些数据集和结构，通用分割模型通常会遇到挑战或局限。由于可用于评估的真实标注数据十分有限，我们进行了 Wilcoxon秩和检验 [27] ，以确定模型之间的分割准确率是否存在潜在差异。对于展示小鼠干细胞的数据集，最大 p值达到0.0002；检验结果证实，两个模型在所有数据集上的表现相当。这表明，生成的数据还可用于系统评估和基准测试，从而无需人工标注即可评估分割模型在特定数据集上的表现。

总的来说，合成图像数据始终保持较高的峰值信噪比，且分割结果可与使用大型已标注数据集训练的最先进方法相媲美，这凸显了生成数据的真实性，并证明了其实际可用性。然而，该方法存在一项局限：前向过程中引入的噪声会使生成非常暗的细胞颇具挑战，因为添加的噪声会掩盖结构信息。因此，弱光照区域难以准确分割，因为使用合成数据训练的模型可能无法充分应对真实数据带来的所有挑战。尽管存在这一局限，所得分割得分仍表明，利用直观的无监督方法生成的数据，可以全自动地训练专用模型，并取得最先进的分割结果。该流程解决了应用分割方法所需的人工标注问题，并提出应将重点转向确定适合创建细胞结构的配置。生成这些结构可以通过模拟方法实现，也可以利用 Cellpose 等公开可用的通用分割方法 [18]或 SAM [28]。因此，对于标注有限或完全缺失的数据集，基于深度学习的分割方法将更易于应用。为推动这一目标，所有完全标注的合成图像数据集均已公开，网址为https://osf.io/dnp65/；训练和应用代码可在 https://github.com/stegmaierj/ DiffusionModelsForImageSynthesis获取。

参考文献

[1] Meijering, E., Carpenter, A.E., Peng, H., Hamprecht, F.A., Olivo-Marin, J.-C.：展望生物图像分析的未来。《自然·生物技术》34(12), 1250–1255（2016）[2] Meijering, E.：生物图像分析中的深度学习纵览。《计算与结构生物技术杂志》18，2312（2020）[3] Shorten, C., Khoshgoftaar, T.M.：深度学习图像数据增强综述。《大数据杂志》6(1), 1–48（2019）[4]Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.：利用学习到的变换进行数据增强，实现单样本医学图像分割。载于：计算机视觉与模式识别会议，第8543–8553页（2019）[5] Eschweiler, D., Schock, J., Stegmaier, J.：通过概率图像多样化提升三维显微镜图像数据中的分割效果。载于：医学成像模拟与合成国际研讨会（SASHIMI），第24–33页（2022）[6] Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.：Models Genesis。《医学图像分析》67, 101840（2021）[7]Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.：nnU-Net：一种用于基于深度学习的生物医学图像分割的自配置方法。《自然·方法》18(2)，203–211（2021）[8]Svoboda, D., Ulman, V.：MitoGen：用于荧光显微镜下细胞群体三维合成延时序列生成的框架。《IEEE医学成像汇刊》36(1)，310–321（2016）[9] B¨ohland, M., Scherr, T., Bartschat, A., Mikut, R., Reischl, M.：合成标签图像对象属性对生成对抗网络支持的分割流程的影响。载于：第29届计算智能研讨会，第289–305页（2019）

[10]Eschweiler, D., Rethwisch, M., Jarchow, M., Koppers, S., Stegmaier, J.: 用于分割和基准测试的三维荧光显微镜数据合成。《PLOS One》16(12), 0260509(2021)

[11]B¨ahr, D., Eschweiler, D., Bhattacharyya, A., Moreno-Andr´es, D., Antonin, W., Stegmaier, J.：CellCycleGAN：利用统计形状模型和条件生成对抗网络进行细胞群体的时空显微图像合成。载于：IEEE生物医学国际研讨会

成像（ISBI），第 15–19 页（2021）[12] Wiesner, D., Suk, J., Dummer, S., Svoboda, D., Wolterink, J.M.：用于活细胞形状生成式建模的隐式神经表示。载于：国际医学图像计算与计算机辅助干预会议，第 58–67 页（2022）[13] Bruch, R., Keller, F., B¨ohland, M., Vitacolonna, M., Klinger, L., Rudolf, R., Reischl, M.：合成大规模三维细胞培养物三维显微图像，用于训练和基准测试。《PloS One》18(3)，0283828（2023）[14] Ho, J., Jain, A., Abbeel, P.：去噪扩散概率模型。《神经信息处理系统进展》（NeurIPS）33，6840–6851（2020）[15] Nichol, A.Q., Dhariwal, P.：改进的去噪扩散概率模型。载于：国际机器学习会议（ICML），第 8162–8171 页（2021）[16] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., Ermon, S.：SDEdit：利用随机微分方程进行引导式图像合成与编辑。载于：国际学习表征会议（2021）[17] Willis, L., Refahi, Y., Wightman, R., Landrein, B., Teles, J., Huang, K.C., Meyerowitz, E.M., J¨onsson, H.：拟南芥顶端干细胞生态位中的细胞大小与生长调控。《美国国家科学院院刊》113(51)，8238–8246（2016）[18] Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.：Cellpose：一种通用细胞分割算法。《自然·方法》18(1)，100–106（2021）[19] Ulman, V., Maˇska, M., Magnusson, K.E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., 等：细胞跟踪算法的客观比较。《自然·方法》14(12)，1141–1152（2017）[20] Faure, E., Savy, T., Rizzi, B., Melani, C., Staˇsov´a, O., Fabr`eges, D., ˇSpir, ˇ R., Hammons, M., C´underl´ık,R., Recher, G., 等：处理发育中生物体三维时间显微图像并重建其细胞谱系的工作流程。《自然通讯》7(1)，1–10（2016）[21] Moreno-Andr´es, D., Bhattacharyya, A., Scheufen, A., Stegmaier, J.：Live-CellMiner：一种分析有丝分裂进程的新工具。《PloS One》17(7)，1–24（2022）

[22] Stegmaier, J., Otte, J.C., Kobitski, A., Bartschat, A., Garcia, A., Nienhaus, G.U., Str¨ahle, U., Mikut, R.：太字节级、时间分辨三维显微图像序列中染色细胞核的快速分割。《公共科学图书馆·综合》9(2)，1–11（2014）[23] Eschweiler, D., Stegmaier, J.：稳健的三维细胞分割：扩展 Cellpose 的视野。载于：IEEE 国际图像处理会议，第 191–195 页（2022）[24] Murray, J.I., Bao, Z., Boyle, T.J., Boeck, M.E., Mericle, B.L., Nicholas, T.J., Zhao, Z., Sandel, M.J., Waterston, R.H.：在秀丽隐杆线虫中以细胞分辨率自动分析胚胎基因表达。《自然·方法》5(8)，703–709（2008）[25] B´artov´a, E., ˇSust´aˇckov´a, G., Stixov´a, L., Kozubek, S., Legartov´a, S., Folt´ankov´a, V.：Oct4 蛋白向胚胎干细胞紫外线损伤染色质的募集。《公共科学图书馆·综合》6(12)，27281（2011）[26] Neumann, B., Walter, T., H´erich´e,J.-K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U. 等：通过延时显微镜对人类基因组进行表型分析，揭示细胞分裂相关基因。《自然》464(7289)，721–727（2010）[27] Mann, H.B., Whitney, D.R.：关于检验两个随机变量中一个是否在随机意义上大于另一个。《数学统计年刊》，50–60（1947）[28] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y. 等：分割一切。arXiv:2304.02643（2023）