Accepted in CVPR 2025 Embodied AI Workshop

用于边缘设备图像修复的多步引导扩散：走向具身人工智能的轻量级感知

Aditya Chakravarty
Independent Research
San Francisco, CA
chakravarty.aditya28@gmail.com

1. 引言

扩散模型已成为解决逆问题的强大工具，而无需针对特定任务的重新训练。像扩散后验采样（DPS） [4] 和 FreeDoM [17] 这样的方法通过使用外部定义的目标来引导生成过程，从而实现灵活且模块化的推理。相关的多调用方法 [1, 7, 11] 通过迭代引导扩展了这些理念，但通常需要重新训练、复杂的调度或特定领域的调优。

流形保持引导扩散（MPGD） [6] 最近被提出作为一种无需训练的替代方法，它将引导限制在学习的图像流形的切空间，从而在恢复过程中提高稳定性和真实性。然而，MPGD 和类似方法通常每次去噪步骤仅应用单次梯度更新，尚未探索更深层次的多步条件。此外，MPGD 主要是在以人脸为中心的数据集上开发和评估的，这对其在更多样化或不在分布内的内容上的鲁棒性提出了疑问。

在这项工作中，我们通过多步优化的视角重新审视了 MPGD：在每个去噪时间步中应用若干次梯度下降更新。受 RePaint [1] 和 LGD [11] 中先前观察到的重复更新可以提高保真度的启发，我们进行了一项实证研究，以探讨质量、多样性和推理成本之间的权衡。我们发现，增加指导步骤的数量显著提升了感知质量（LPIPS）和像素级准确度（PSNR），同时增强了对退化或分布外输入的鲁棒性。值得注意的是，我们展示了，虽然 MPGD 是在面部数据集上训练的，但通过多步条件化，它可以有效地恢复通用的非面部自然图像。

我们的实验专注于两个经典的逆问题：4× 超分辨率和高斯去模糊。我们使用 Jetson Orin Nano 一款紧凑的边缘 GPU 平台，在实际受限的具身 AI 系统中进行了各种优化深度的评估。ImageNet 和 UAV123 航拍图像上的结果表明，MPGD 结合多步优化是一个可行的轻量解决方案，适用于机器人和移动 AI 代理在无限制环境中进行实时视觉恢复。

2. 方法和实验设置

我们考虑 4× 超分辨率（双三次下采样）和高斯去模糊（核大小 61，强度 3.0），两者都带有加性噪声 $\sigma = 0.05$，按照 [6, 16]。我们使用像素空间 MPGD 实现与 DDIM 采样 [10] 和预训练的 FFHQ 自动编码器 [12]。我们为 1000 张 ImageNet [5] 测试图像生成输出，遍历步骤 $\in \{1, 3, 7, 15, 20\}$，时间步 $\in \{20, 50, 100\}$ 和指导尺度 $\in \{4, 7.5, 17.5\}$。对于评估，我们考虑 LPIPS [18]，SSIM [13]，推理时间和 PSNR。所有实验在单个 NVIDIA Jetson Orin Nano（8 GB VRAM）上运行。

应用案例：无人机 123 的空中检查 为了在实体人工智能环境中评估其实际可行性，我们在来自 UAV123 数据集的降质航拍视频中评估 MPGD [9]。该数据集包含无人机在建筑物、公路和工业场所上空的视频序列，代表了视觉检查场景。

我们提取了 300 个横跨多个场景的不同帧。这些帧使用与主要 ImageNet 基准相同的 MPGD 配置进行处理，无需进一步训练或调整。输出结果在 LPIPS、PSNR、SSIM 和推理时间上进行比较。

3. 结果与讨论

在这两个任务中，我们观察到性能随着优化步骤的增加和适度的指导尺度而提高，在大约 15 步时趋于饱和。重建结果（图 1）从 1 步时的通用人脸形状演变到 15 步时的结构良好的输出，即使对于分布外的图像也是如此。在 Jetson Orin Nano 上，每张图像的推理延迟在 50–100 毫秒之间（表 1），验证了 MPGD 在实时嵌入式感知中的适用性。

在 UAV123 图像上，MPGD 实现了较强的感知和像素级性能，尽管场景变化多端且存在噪声。表 2 显示，MPGD（15 步）在 LPIPS 和 PSNR 方面超越了 NAFNet [2] 和 Uformer [14]，同时在 Jetson Orin Nano 上保持实时吞吐量。这表明 MPGD 作为轻量级插件模块在航空基础设施检查中进行实时图像增强的实用性。

Method	LPIPS ↓	SSIM ↑	PSNR ↑	Time (ms) ↓
MPGD (15 steps)	0.32	0.90	20.91	80
NAFNet [2]	0.36	0.86	20.13	35
Uformer [14]	0.34	0.87	19.65	58

Table 1. MPGD 与基线模型的比较（ImageNet）。

跨越两个任务，我们观察到性能随着更多的优化步骤和适度的引导比例而提高，在大约 15 步时趋于饱和。定性分析表明，即使对于超出分布的图像，重构结果也从 1 步时的通用面部样式演变为 15 步时的结构良好的

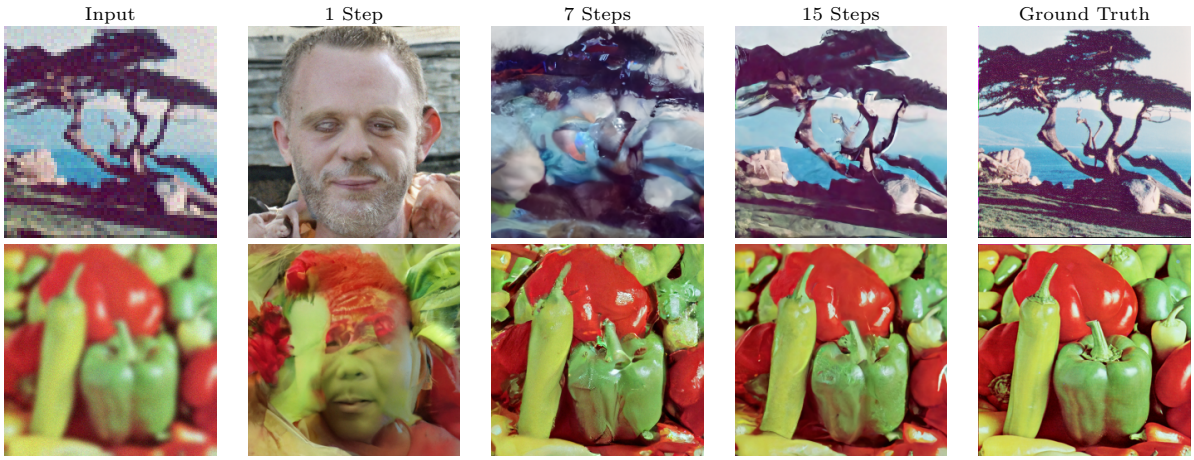

Figure 1. 在 1、7 和 15 步时的 SR 和去模糊结果比较。

Method	LPIPS ↓	SSIM ↑	PSNR ↑	Time (ms) ↓
MPGD (15 steps)	0.35	0.88	21.20	90
NAFNet	0.38	0.84	20.10	42
Uformer	0.37	0.85	19.90	65

Table 2. 在退化的 UAV123 帧（300 个样本）上的性能。

输出。推理延迟每张图像在 50–100 毫秒之间，这验证了 MPGD 在实时嵌入式感知中的适用性。

4.

结论和正在进行/未来的工作

我们提出了一种多步骤流形保持引导扩散（MPGD）方法，用于无需训练的图像修复，目标是在如 Jetson Orin Nano 这样的边缘设备上进行部署。我们在标准基准和 UAV123 数据集上的实验表明，MPGD 作为一种轻量级的、无需重新训练的视觉模块，对于如无人机和移动机器人等具体现智能体，在严格的电力和计算限制下进行操作是有用的。关键的是，我们展示了多步骤优化使得在面部数据上训练的模型能够令人惊讶地很好地泛化到自然的、非面部图像——这表明任务驱动的优化可以在实践中补偿不匹配的训练域。

这项工作是正在进行的努力的一部分，旨在将 MPGD 扩展到更广泛的具身感知挑战，包括弱光导航、基础设施检查和领域迁移下的视觉定位。近期在基于扩散的决策制定 [3, 8] 方面的进展表明，生成性先验可以在多样化的环境中支持稳健的视觉模块。在此基础上，我们计划探索推理期间的自适应优化深度 [11]，以及针对资源有限的边缘部署量身定制的轻量级测试时适应策略 [15]。我们还在研究 MPGD 在多模态和非线性逆问题上的扩展，包括在空间提示或语言线索上进行条件化。这些方向旨在将 MPGD 定位为具身 AI 系统中实时感知的可部署和适应性强的骨干。

References

[1] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2022. 1

[2] Liangyu Chen, Xiaojie Chu, and Xiangyu Zhang. Simple baselines for image restoration. In European Conference on Computer Vision (ECCV) , pages 186–202, 2022. 1

[3] Linxi Chi, Zichen Huang, Tao Yu, Ziyu Ma, Ankur Handa, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. arXiv preprint arXiv:2303.04137 , 2023. 2

[4] Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L.Klasky, and JongChul Ye. Diffusion posterior sampling for general noisy inverse problems. In Proc. International Conference on Learning Representations (ICLR) , 2023. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pages 248–255, 2009. 1

[6] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In Proc. International Conference on Learning Representations (ICLR) , 2024. poster. 1

[7] Jiachun Pan, Hanshu Yan, Jun Hao Liew, Jiashi Feng, and Vincent Y. F. Tan. Towards accurate guided diffusion sampling through symplectic adjoint method. arXiv preprint arXiv:2312.12030 , 2023. 1

[8] Jacky Liang, Zirui Yuan, Qiyang Wu, et al. Generative pretraining for decision making and control. arXiv preprint arXiv:2210.03029 , 2023. 2

[9] Matthias Mueller, Neil Smith, and Bernard Ghanem.

A benchmark and simulator for uav tracking. In ECCV , 2016. 1

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In International Conference on Learning Representations (ICLR) , 2021. 1

[11] Yang Song, Chenlin Meng, and Stefano Ermon. Loss-guided diffusion: Learning to denoise images conditioned on a loss function. In International Conference on Machine Learning (ICML) , 2023. 1, 2

[12] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , pages 17968–17977, 2021. 1

[13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing , 13(4):600–612, 2004. 1

[14] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Guihua Chen, Jin Gu, Jianzhuang Liu, and Chen Dong. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , pages 17683–17693, 2021. 1

[15] Zhen Wang, Qiang Wang, Kai Zhang, Mengyuan Xu, and Yiran Wang. Tta-lite: Memory-efficient test-time adaptation for deep models on the edge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2023. 2

[16] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 , 2022. 1

[17] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In Proc. IEEE/CVF International Conference on Computer Vision (ICCV) , 2023. 1

[18] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pages 586–595, 2018. 1