中期进展报告 - 基于扩散模型的高效灾害遥感图像文本控制生成和应用

陈振源

June 20, 2025

Abstract

遥感图像对灾害应急响应、调度、管理和决策具有重大作用。而灾害发生具有随机性和稀有性，同时，不同地区面临各种灾害的实际情况差异大，在它地的实践经验难以有效的迁移。针对高分辨率时序遥感图像数据集稀少且时空分布不均的问题，本研究创新性的提出一种遥感图像合成方法，使用生成式模型（扩散模型）进行灾害遥感图像基于文本和灾前影像的可控生成，以弥补世界范围内灾害影像数据缺少的问题。

1 背景

遥感图像在灾害应急响应、调度、管理和决策中发挥着至关重要的作用。通过对灾害发生区域的遥感监测，可以为应急部门提供及时、准确的信息支持，提升灾害响应的效率和科学性。然而，灾害事件具有高度的随机性和稀有性，这使得相关遥感数据的获取变得十分困难。许多灾害发生时无法及时获取高质量的遥感影像，导致数据积累缓慢，难以满足研究和实际应用的需求。此外，不同地区面临的灾害类型和实际情况存在较大差异，导致某一地区的经验和数据难以直接迁移和应用到其他地区。这种区域差异性进一步加剧了数据利用的局限性。目前，高分辨率时序遥感图像数据集不仅数量稀少，而且在时空分布上极为不均衡。部分地区数据丰富，而大多数受灾频发区域却缺乏足够的遥感影像资料。因此，亟需弥补全球范围内灾害遥感影像数据的缺口，推动相关数据的生成与共享，以支持灾害管理和科学研究的深入发展。

传统的遥感图像分析方法在灾害场景下往往难以满足需求。无论是人工还是基于模型的方法，都难以应对灾害的多样性和复杂性；文本引导的编辑能够实现自适应、用户驱动的图像修改，但现有方法大多针对自然场景设计，难以生成真实且语义一致的遥感影像。为此，本工作聚焦于利用高分辨率、双时相 RSCC 卫星数据，实现灾害场景下的遥感图像文本引导编辑。我们的目标是通过用户指令实现灵活、精准的图像编辑，支持场景推演、快速制图、灾害损失评估和变化检测等灾害管理任务。然而，在遥感领域实现高质量、可控的图像编辑面临技术难题，主要包括领域泛化能力不足以及现有框架的高计算资源需求。当前主流框架在灾害任务中缺乏真实感和语义对齐能力，传统架构计

算开销大，难以灵活部署于实际场景。针对上述挑战，我们提出了一种新颖的视觉模型及适应策略。主要贡献包括：面向灾害场景和文本指令的扩散 Transformer 架构，实现任务引导的遥感图像编辑（该架构专为灾害场景和文本化用户指令定制）；高效的模型微调策略，支持大规模适应（显著减少新任务所需的可训练参数，提高适应效率）；构建了全面的基准和新的评测指标，专用于文本引导的遥感图像编辑任务（填补了该领域的评测空白）；并且在高分辨率、双时相 RSCC 灾害影像上，相较现有方法取得了实证性提升（在极具挑战性的场景下展现出显著性能提升）。

2 相关工作

2.1 遥感领域的文本到图像生成

近年来，遥感领域的文本到图像生成方法主要基于扩散模型，通常通过对预训练模型进行微调来实现。在数据有限的场景下，已有研究探索了高效的微调策略 (Ou et al. 2023)。CRS-Diff (Tang et al. 2024) 通过引入 ControlNet (L. Zhang, Rao, and Agrawala 2023) 实现了多模态控制，DiffusionSat (Khanna et al. 2024) 则进一步提出了 3DControlNet 统一框架，支持时序生成、超分辨率和修复等多种下游任务。Text2Earth (C. Liu et al. 2025) 以文本到图像生成为主，同时支持图像编辑和修复等辅助任务。然而，这些方法大多依赖于容量较小的图像主干网络——通常为约 8 亿参数的预训练 UNet（如 Stable Diffusion），与新兴的大型扩散 Transformer(Diffusion Transformer, DiT) 模型相比，其表达能力受到限制。此外，这些方法的文本编码器通常沿用 Stable Diffusion 的 CLIP 变体，对于复杂任务或指令的深层语义理解能力有限。

2.2 基于扩散模型的图像编辑

图像编辑是计算机视觉中的基础任务，旨在根据用户意图对给定图像进行修改，同时保持图像的真实感和语义一致性。早期的指令引导图像编辑方法主要依赖于基于 UNet 的扩散模型。值得注意的是，InstructPix2Pix (Brooks, Holynski, and Efros 2023) 首次提出了基于用户指令的编辑新范式，MagicBrush (K. Zhang et al. 2023) 进一步采用该框架，并通过增强数据集提升了性能。

近年来，Diffusion Transformer (Peebles and Xie 2023) 的出现为图像合成带来了比传统 UNet 扩散模型更具扩展性和灵活性的架构。在此基础上，StableFlow (Avrahami et al. 2025) 和 RF-Solver (Wang et al. 2024) 等工作提出了无需训练的编辑方法，分别通过关键层注入和注意力机制操控来利用 DiT 的内部结构，尽管 RF-Solver 并不依赖于标准的指令编辑方式。

与此同时，将多模态大语言模型集成到扩散框架中（如 Qwen2VL-Flux (StableKirito 2025)、Step1X-Edit (S. Liu et al. 2025) 和 SmartEdit (Huang et al. 2024)）使得基于复杂指令的图像编辑更加高效和多样化。另一方面，UltraEdit (Zhao et al. 2024) 和 Emu Edit (Sheynin et al. 2024) 仍然采用 UNet 架构，但结合了先进的文本编码器和大规模数据集，实现了更细粒度的指令编辑能力。

3 方法

本研究采用了与 ICEdit (Z. Zhang et al. 2025) 相同的框架。该方法基于大规模 DiT，通过上下文生成机制实现指令式图像编辑。具体而言，模型输入包括原始图像、编辑指令以及若干对参考编辑样例（即“上下文对”），模型通过学习这些上下文对之间的编辑关系，理解并执行用户给定的新编辑指令，从而生成符合要求的编辑结果。

在训练阶段，模型接收一组（原图、编辑指令、目标图）三元组，通过条件扩散过程学习从原图到目标图的编辑映射。推理时，用户只需提供原图和新的编辑指令，模型即可在无需额外微调的情况下，基于上下文推理能力完成相应的图像编辑任务。

该方法的核心优势在于：

- 通用性强：无需针对每种编辑任务单独训练，具备良好的泛化能力。
- 高效的上下文理解：通过上下文示例，模型能够理解复杂、多样的编辑指令。
- 端到端训练：整体框架可端到端优化，提升编辑质量和一致性。

因此，我们在本项目中直接采用了该框架，利用其强大的指令理解与图像编辑能力，完成了各类复杂的图像编辑任务。

4 数据集

本工作使用了 RSCC 数据集。该数据集面向灾害事件，包含大规模的遥感变化描述，专为灾害感知的双时相遥感影像理解设计。通过视觉推理模型 QvQ-Max，对 62,351 对灾前与灾后影像进行了详细的变化描述标注 (Chen et al. 2025)。

5 评估指标

本研究采用多种定量和定性评估指标，全面衡量遥感图像生成与编辑的质量与准确性：

- **CLIP** 图像相似度 (Radford et al. 2021; Hessel et al. 2021)：利用 CLIP 模型提取生成图像与目标图像的特征，通过计算特征余弦相似度，评估生成结果与真实目标在高层语义空间中的一致性。
- **DINO** 图像相似度 (Oquab et al. 2024)：采用 DINO 自监督视觉 Transformer 模型提取图像特征，计算生成图像与目标图像的特征相似度，进一步衡量结构和内容的一致性。
- **CLIP** 文图相似度 (Radford et al. 2021)：将用户编辑指令与生成图像分别输入 CLIP 模型，计算文本与图像之间的相似度分数，反映生成结果对指令的语义遵循程度。

- 基于多模态大语言模型的评价指标, 如 **VIEScore** (Ku et al. 2024)：引入 VIEScore 等多模态大语言模型，综合评估图像生成质量和编辑准确性。该类指标能够理解复杂的编辑指令，并对生成图像的语义、内容和编辑效果进行端到端的自动化评价，提升评测的客观性和全面性。

通过上述多维度指标，能够系统地评估模型在遥感灾害图像生成与编辑任务中的表现，确保生成结果在视觉质量、语义一致性和指令遵循性等方面均达到高水平。

6 预期成果

本项目预计将取得以下几方面的成果：

- 高质量的遥感灾害图像合成与编辑能力：基于大规模扩散 Transformer 和上下文编辑机制，能够根据用户文本指令和灾前影像，生成真实感强、语义一致的灾害遥感图像，满足灾害推演、损失评估和变化检测等多样化需求。
- 通用且高效的指令式图像编辑框架：实现无需针对每类编辑任务单独训练的通用编辑范式，显著提升模型在不同灾害类型、不同场景下的泛化能力和适应性。
- 高效的模型微调与适应策略：通过参数高效的微调方法，降低新任务适配的计算和数据成本，推动大模型在遥感领域的实际部署和应用。
- 完善的评测基准与指标体系：构建面向文本引导遥感图像编辑的评测基准和指标，填补该领域评测体系的空白，为后续相关研究提供标准化参考。
- 实证性性能提升：在高分辨率、双时相 RSCC 灾害影像上，模型在图像质量、编辑准确性和语义一致性等方面均优于现有主流方法，推动遥感灾害图像智能生成与编辑技术的发展。

通过上述成果，项目将为灾害应急管理、科学研究和实际应用提供强有力的数据和技术支撑，促进遥感智能生成与编辑方法在灾害场景下的落地与推广。

7 参考文献

References

Avrahami, Omri et al. (Nov. 2025). “Stable Flow: Vital Layers for Training-Free Image Editing”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. arXiv:2411.14430. DOI: 10.48550/arXiv.2411.14430. URL: http://arxiv.org/abs/2411.14430.

Brooks, Tim, Aleksander Holynski, and Alexei A. Efros (2023). "InstructPix2Pix: Learning To Follow Image Editing Instructions". en. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18392–18402. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Brooks_InstructPix2Pix_Learning_To_Follow_Image_Editing_Instructions_CVPR_2023_paper.html.

Chen, Zhenyuan et al. (2025). *RSCC: A Large-Scale Remote Sensing Change Caption Dataset for Disaster Events*. Pre-published.

Hessel, Jack et al. (Nov. 2021). "CLIPScore: A Reference-free Evaluation Metric for Image Captioning". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2021. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7514–7528. DOI: 10.18653/v1/2021.emnlp-main.595. URL: https://aclanthology.org/2021.emnlp-main.595 (visited on 07/18/2024).

Huang, Yuzhou et al. (2024). "SmartEdit: Exploring Complex Instruction-based Image Editing with Multimodal Large Language Models". en. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8362–8371. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Huang_SmartEdit_Exploring_Complex_Instruction-based_Image_Editing_with_Multimodal_Large_Language_CVPR_2024_paper.html.

Khanna, Samar et al. (Oct. 2024). "DiffusionSat: A Generative Foundation Model for Satellite Imagery". In: *The Twelfth International Conference on Learning Representations*. (Visited on 06/29/2024).

Ku, Max et al. (Aug. 2024). "VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2024. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 12268–12290. DOI: 10.18653/v1/2024.acl-long.663. URL: https://aclanthology.org/2024.acl-long.663/ (visited on 05/06/2025).

Liu, Chenyang et al. (2025). "Text2Earth: Unlocking Text-Driven Remote Sensing Image Generation with a Global-Scale Dataset and a Foundation Model". In: *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–23. ISSN: 2168-6831. DOI: 10.1109/MGRS.2025.3560455. (Visited on 05/10/2025).

Liu, Shiyu et al. (May 2025). "Step1X-Edit: A Practical Framework for General Image Editing". In: arXiv:2504.17761. arXiv:2504.17761 [cs]. DOI: 10.48550/arXiv.2504.17761. URL: http://arxiv.org/abs/2504.17761.

Oquab, Maxime et al. (2024). "DINOv2: Learning Robust Visual Features without Supervision". In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=a68SUt6zFt (visited on 07/19/2024).

Ou, Ruizhe et al. (Oct. 2023). "A Method of Efficient Synthesizing Post-disaster Remote Sensing Image with Diffusion Model and LLM". In: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1549–1555. DOI: 10.1109/APSIPAASC58517.2023.10317383. (Visited on 08/21/2024).

Peebles, William and Saining Xie (2023). "Scalable Diffusion Models with Transformers". en. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205. URL: https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Radford, Alec et al. (July 1, 2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html (visited on 07/07/2024).

Sheynin, Shelly et al. (2024). "Emu Edit: Precise Image Editing via Recognition and Generation Tasks". In: pp. 8871–8879. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Sheynin_Emu_Edit_Precise_Image_Editing_via_Recognition_and_Generation_Tasks_CVPR_2024_paper.html.

StableKirito (May 2025). *erwold/qwen2vl-flux*. Python. URL: https://github.com/erwold/qwen2vl-flux.

Tang, Datao et al. (2024). "CRS-Diff: Controllable Remote Sensing Image Generation With Diffusion Model". In: *IEEE Transactions on Geoscience and Remote Sensing* 62, pp. 1–14. ISSN: 1558-0644. DOI: 10.1109/TGRS.2024.3453414. (Visited on 04/29/2025).

Wang, Jiangshan et al. (Nov. 2024). "Taming Rectified Flow for Inversion and Editing". In: arXiv:2411.04746. arXiv:2411.04746. DOI: 10.48550/arXiv.2411.04746. URL: http://arxiv.org/abs/2411.04746.

Zhang, Kai et al. (Dec. 2023). "MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing". en. In: *Advances in Neural Information Processing Systems*. Vol. 36, pp. 31428–31449. URL: https://papers.nips.cc/paper_files/paper/2023/hash/64008fa30cba9b4d1ab1bd3bd3d57d61-Abstract-Datasets_and_Benchmarks.html.

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala (Oct. 2023). "Adding Conditional Control to Text-to-Image Diffusion Models". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. DOI: 10.1109/ICCV51070.2023.00355. URL: https://ieeexplore.ieee.org/document/10377881.

Zhang, Zechuan et al. (Apr. 29, 2025). *In-Context Edit: Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer*. DOI: 10.48550/arXiv.2504.20690. arXiv: 2504.20690 [cs]. URL: http://arxiv.org/abs/2504.20690 (visited on 05/02/2025). Pre-published.

Zhao, Haozhe et al. (Nov. 2024). “UltraEdit: Instruction-based Fine-Grained Image Editing at Scale”. en. In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=9ZDdlgH6O8#discussion.