中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-16

论文引用格式： JI Yingrui, Wang Chenhao, Chen Jingbo, Yue Anzhi, Xi Zhihao, Chen Jiansheng. Parameter-Efficient Diffusion Model Adaptation and Spectral Consistency Learning for Controllable Multispectral Remote Sensing Image Generation[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260089. (纪璎芮, 王晨昊, 陈静波, 岳安志, 席智浩, 陈建胜. 多光谱遥感图像可控生成的扩散模型参数高效适配与光谱一致性学习[J/OL]. 中国图象图形学报,XXXX:1-16. DOI: 10.11834/jig.260089.)[DOI:10.11834/jig.260089]

多光谱遥感图像可控生成的扩散模型参数高效适配与光谱一致性学习

纪璎芮[1,2],王晨昊[1,2],陈静波[1,2],岳安志[1],席智浩[1],陈建胜[1]

1. 中国科学院空天信息创新研究院　国家遥感应用工程技术研发中心,北京　100094; 2. 中国科学院大学　电子电气与通信工程学院,北京　100049

摘　要： 针对深度学习在多光谱遥感应用中面临的数据获取困难与标注成本高昂问题,以及现有基础生成模型难以直接适配多光谱数据且从零训练计算开销巨大的现状,提出了一种面向多光谱遥感图像生成的参数高效适配扩散模型。该方法采用参数高效微调策略,通过在冻结的预训练扩散模型中嵌入各种低参数微调模块,不同于通用可控生成方法仅以数据驱动方式建模图像,本文在适配训练中显式引入遥感光谱物理约束,并针对地物语义-空间映射设计了文本感知编码机制。实现了从RGB图像域向四波段(RGB+NIR)图像域的低成本迁移,不同微调模块综合了光谱与空间纹理适配。在此基础上,引入基于归一化植被指数(Normalized Difference Vegetation Index, NDVI)和归一化水体指数(Normalized Difference Water Index, NDWI)的物理一致性损失,强制约束红光与近红外波段间的光谱相关性。此外,提出文本感知空间语义编码机制,利用语义分割掩膜实现对地物空间布局的精确控制。在FLAIR、Five-Billion-Pixels及IRSAMap等数据集上的实验表明,与ControlNet、T2I-Adapter等主流方法相比,本文方法在光谱保真度与语义对齐度上均有所提升,生成的近红外波段具备明确的物理意义。此外,利用生成数据辅助训练在下游开放词汇分割任务上取得了一定的精度提升,验证了该方法作为数据增强手段的可行性。本框架有效解决了RGB基础模型向多光谱遥感领域迁移时的通道不匹配与物理特征丢失问题,实现了低资源消耗下的高质量、可控多光谱数据生成,为缓解遥感解译任务中的数据稀缺问题提供了有效的数据增强方案。

关键词： 图像生成;扩散模型;多光谱数据;近红外;低秩参数微调

Parameter-Efficient Diffusion Model Adaptation and Spectral Consistency Learning for Controllable Multispectral Remote Sensing Image Generation

JI Yingrui[1,2], Wang Chenhao[1,2], Chen Jingbo[1,2], Yue Anzhi[1], Xi Zhihao[1], Chen Jiansheng[1]

1. National Engineering Research Center for Geoinformatics, Aerospace Information Research Institute, Beijing 100094, China; 2. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The acquisition of high-quality multispectral remote sensing imagery is a fundamental pathway for understanding earth surface physical properties and monitoring environmental dynamics. Multispectral data, particularly involving Near-Infrared (NIR) bands, reflects the unique physical attributes of land surfaces, which are crucial for gaining deeper insights

收稿日期: 2026-02-08; **修回日期:** 2026-03-11

* **通信作者:** 陈建胜,通信作者,男,高级工程师,主要研究方向为遥感图像智能解译、数据治理。E-mail: chenjs@aircas.ac.cn; 陈建胜 Email: chenjs@aircas.ac.cn

基金项目: 中国科学院战略性先导科技专项(XDA0360303)

Supported by: Strategic Priority Research Program of the Chinese Academy of Sciences (XDA0360303)

into vegetation health assessment, land-cover classification, and complex environmental analysis. Such data provides essential scientific support for regional ecological monitoring and territorial spatial planning. Deep learning has become an important research direction in remote sensing image interpretation. The core objective of these deep learning applications relies heavily on large-scale, accurately annotated datasets to train robust predictive models. Most existing generative studies focus on standard RGB image synthesis to alleviate data scarcity. Although substantial progress has been achieved in terms of generation quality and diversity using latent diffusion models (LDMs), directly applying RGB-pretrained foundation models remains inherently limited in their ability to model multispectral physical dynamics. In real-world remote sensing scenarios, imagery often exhibits complex spectral correlations and domain-specific spatial patterns. Relying solely on RGB-pretrained weights makes it difficult to faithfully characterize the physical constraints between specific wavelengths, such as the Red and NIR bands, and limits the ability to uncover latent geospatial features. Moreover, since foundation models lack multispectral representations, simple channel expansion approaches are highly susceptible to distribution mismatch and severe generation degradation. Consequently, exploring efficient adaptation of diffusion models to better capture the multispectral domain has become an important research direction. However, from a computational and structural perspective, training large-scale multispectral diffusion models from scratch remains computationally prohibitive and highly resource-intensive. Most widely used controllable generation methods are still limited to treating spatial conditions as simple structural constraints via shallow convolutional encoding. As a result, they are insufficient for modeling complex geospatial layouts and fail to adequately capture the inherent high-dimensional textural or spectral features of specific geographical categories, particularly in scenarios requiring precise semantic alignment. To address these limitations, this paper constructs a parameter-efficient Multispectral Controllable Diffusion Model framework, with the aim of providing robust data support for multispectral image interpretation tasks. The proposed method efficiently adapts RGB-based models to the four-band (RGB+NIR) remote sensing domain through a decoupled two-stage training strategy, which ensures convergence stability without full network retraining. We employ a parameter-efficient fine-tuning (PEFT) strategy that injects lightweight adapters and Low-Rank Adaptation (LoRA) modules into a frozen pre-trained backbone. To capture complex geospatial patterns, multi-scale depthwise convolutions and modulation mechanisms (MSM and MONA) are integrated into the architecture. Crucially, we introduce a spectral consistency loss based on the Normalized Difference Vegetation Index (NDVI) to enforce physically plausible correlations between the Red and NIR bands. This loss is constrained by a brightness threshold to exclude noise from low-reflectance regions, ensuring spectral fidelity. Additionally, a text-aware spatial-semantic encoding (TASSE) mechanism is designed to enable precise control over image layout using semantic segmentation masks, explicitly mapping category-specific text embeddings to spatial features. This design ensures the synthesized data's practical applicability and physical relevance, providing robust support for research on diverse land-cover categories and the development of downstream segmentation algorithms. Systematic experiments conducted on standard datasets, including FLAIR, Five-Billion-Pixels, and IRSAMap, reveal performance differences among generative models and validate the usability and inherent challenges of the proposed framework in real-world data augmentation tasks. Within a unified experimental framework, we compare our method against state-of-the-art approaches such as ControlNet, T2I-Adapter, CRS-Diff, and EarthSynth. By evaluating metrics like NDVI-RMSE, CLIP Score, and LPIPS, we assess the influence of physical constraints and spatial encoding on the model's ability to synthesize high-fidelity multi-band data. Experimental results demonstrate that the physical and spatial information integrated into the model plays a fundamental role in generation tasks, allowing the framework to successfully generate valid NIR bands, achieving an average NDVI-RMSE of 0.3173, and outperform baselines in both spectral fidelity and semantic alignment. Furthermore, utilizing the synthesized multi-band data for training open-vocabulary segmentation models yields mIoU gains of approximately 0.3% to 1.0%. Overall, the construction of this framework effectively resolves the channel mismatch and physical feature loss, providing a resource-efficient data foundation for advancing remote sensing image interpretation from data-scarce environments toward comprehensive, high-quality multispectral modeling. In future work, the proposed parameter-efficient adaptation framework will be continuously expanded to cover a broader range of remote sensing modalities, such as extending the generative capabilities to hyperspectral and Synthetic Aperture Radar (SAR) imagery. This expansion aims to better capture diverse physical properties and complex scattering mechanisms under different environmental conditions. In addition, we plan to

progressively incorporate temporal consistency constraint mechanisms to facilitate the generation of continuous remote sensing image sequences, thereby providing robust data foundations for long-term, stage-wise land surface change detection tasks. Beyond spatial and spectral enhancements, advanced large language models (LLMs) will also be integrated into the architecture to improve the framework's comprehension of complex geospatial instructions, ultimately achieving more fine-grained, interactive, and user-driven remote sensing image synthesis.

Key words: Image Generation; Diffusion Model; Multispectral Data; Near-Infrared; Low-Rank Adaptation.

0 引 言

多光谱影像作为对地观测的基础数据，涵盖了近红外(Near-Infrared, NIR)等特定波段，能够捕捉地表独特的物理属性，在植被健康监测、地物分类及环境分析等任务中发挥着关键作用。然而，训练高性能深度学习模型依赖于大规模、高质量的数据集，与普通自然图像相比，多光谱数据的获取与精细标注成本极高。因此，利用生成模型合成逼真的遥感影像以实现数据增强，已成为当前研究的热点(马愈卓等，2025，王耀领等，2026)。

在计算机视觉领域，去噪扩散概率模型(Denoising Diffusion Probabilistic Model, DDPM)(Ho 等，2020)和潜在扩散模型(Latent Diffusion Model, LDM)(Rombach 等，2022)展现了卓越的生成能力，Stable Diffusion 等基础模型已能生成多样且高保真的图像。尽管研究者尝试将其应用于遥感领域以缓解数据匮乏问题，但RGB自然图像与多光谱遥感影像之间存在显著的域差异(domain gap)。主流基础模型通常预训练于海量三通道RGB数据，缺乏处理遥感光谱维度的能力。直接将预训练权重迁移至四波段(RGB+NIR)数据面临巨大挑战：简单的通道扩展会破坏原有权重的分布对齐，导致生成质量下降；而从零训练多光谱扩散模型则面临高昂的计算成本。因此，如何高效地将RGB基础模型的强大先验知识迁移至多光谱域，是当前亟需解决的关键问题。

在图像合成领域，深度生成模型经历了从 GAN 到扩散模型的范式演进。早期研究主要基于 GAN (Goodfellow 等，2014，谭明奎等，2021)，尽管在视觉精度上表现优异，但常面临训练不稳定和模式崩塌问题(刘安安等，2024，谭明奎等，2026)。DDPM(Ho 等，2020)通过迭代去噪实现了更稳定的高保真合成，LDM(Rombach 等，2022)将扩散过程转移至压缩隐空间，大幅降低了计算开销，成为当前主流的生成框架。在遥感领域，DiffusionSat(Khanna 等，2024)将扩散模型扩展至大规模卫星数据集实现可控图像合成，EarthSynth(Pan J 等，2025)和 Text2Earth(Liu C 等，2025)利用大规模图文对构建了面向全球尺度的遥感生成基础模型。然而，上述基础模型大多仅适用于 RGB 图像。为生成多光谱数据，HyperLDM (Liu L 等，2023)利用条件 VQGAN 将光谱数据映射至低维空间实现合成，HSIGene(Pang 等，2024)通过超分辨率模块增强光谱细节，UnmixDiff(Shen 等，2025)将光谱解混融入扩散过程以保证物理可解释性。但这些方法通常需从零训练专用编码网络，计算成本较高，且缺乏灵活的空间控制能力。

在语义引导的可控生成方面，ControlNet(Zhang L 等，2023)和 T2I-Adapter(Mou 等，2024)通过引入可训练的复制分支或轻量级适配器，将空间条件注入冻结的预训练模型，为可控生成提供了稳健范式。在遥感领域，早期研究多依赖基于 GAN 的框架(如 Pix2Pix、SPADE)将语义分割图转换为卫星影像，但受限于模式覆盖不足。近期，CRS-Diff(Tang 等，2024)通过整合多种条件引导特征融合，GeoSynth (Sastry 等，2024)利用 OpenStreetMap 数据约束场景语义，证明了生成式数据增强在提升稀有类别分割性能方面的潜力。然而，多数方法将分割掩码仅视为空间约束，通过简单的卷积层编码，难以捕捉特定地物类别的高维语义信息，导致生成结果虽在空间上与掩码对齐，但缺乏目标类别固有的纹理或光谱特征。近期亦有研究借助视觉语言模型生成的文本描述，驱动遥感影像的跨时空领域自适应语义分割(陶超等，2025)，表明将类别文本语义引入地物特征建模具有一定的迁移潜力。针对多光谱数据的基于语义分割掩码的精确布局控制(Xu 等，2025)仍处于探索阶段。

针对上述挑战，本文提出了一种多光谱可控扩散模型。该方法旨在通过参数高效微调(Parameter-Efficient Fine-Tuning, PEFT)策略，将预训练RGB模

型高效适配至四波段遥感域，避免全网络重训练。具体而言，本文在UNet（Ronneberger等，2015）和变分自编码器（Variational Autoencoder，VAE）（Kingma等，2013）的关键组件中引入了低秩适应（Low-Rank Adaptation，LoRA）（Hu等，2021）和Adapter（Houlsby等，2019）等轻量级模块。此外，为更好地捕捉遥感数据的空间模式，模型集成了调制适配模块（Modulation-based Adaptation，Mona）（Yin等，2024）。该设计在冻结大部分原始参数以保持预训练生成能力的同时，实现了多光谱特征的学习。

除网络结构外，确保光谱准确性同样至关重要。常规生成模型往往独立处理各通道，忽视了遥感数据的物理约束。为此，本文引入了基于归一化植被指数（Normalized Difference Vegetation Index，NDVI）的光谱一致性损失，强制模型保持红光与近红外波段间正确的物理关联，确保生成植被的光谱合理性。最后，为实现对图像内容的精确控制，本文提出了一种融合文本嵌入的空间语义编码机制。通过将语义分割掩码融入生成过程，模型能够合成严格遵循指定空间布局与文本语义类别的多光谱影像。为全面评估生成质量，本文从光谱保真度（NDVI-RMSE、NDWI-RMSE）、语义对齐度（CLIP Score）、感知质量（LPIPS）及下游应用价值（mIoU）四个维度进行定量评估，具体指标定义详见第2.2节。在FLAIR、Five-Billion-Pixels及IRSAMap等数据集上的广泛实验表明，本文方法在光谱保真度、语义一致性及感知质量上均优于ControlNet、T2I-Adapter及CRS-Diff等主流方法，且利用生成数据辅助训练可提升下游语义分割精度。

综上所述，本文的主要贡献如下：

1）面向多光谱域迁移的解耦两阶段适配策略：提出了一种参数高效的两阶段训练策略，将光谱域适配与空间控制训练分阶段进行，避免了两类任务在联合训练时的梯度干扰，在不重训练基础模型的前提下实现了RGB预训练模型向四波段遥感域的稳定适配。

2）物理机制引导的光谱一致性学习：在去噪训练中引入了基于NDVI和NDWI的光谱一致性损失，分别对红光与近红外、绿光与近红外波段之间的物理关联进行显式约束，并通过梯度停止与亮度阈值掩码保证优化稳定性，使模型在生成近红外波段时显式遵循不同地物类型的光谱辐射规律，而非单纯拟合像素分布。

3）跨模态协同的文本感知空间语义编码：通过可学习线性注意力将类别文本嵌入映射至对应的空间掩膜位置，在分割掩膜的空间约束基础上引入了类别语义信息，弥补了现有方法仅以浅层卷积编码掩膜的不足。

1 研究方法

1.1 网络整体架构

本文提出了一种面向多光谱遥感影像可控生成的扩散模型高效适配框架，旨在将预训练Stable Diffusion模型有效迁移至四波段（RGB+NIR）遥感影像生成任务中，在保持原模型语义泛化能力的同时，增强其对遥感影像光谱特性、空间结构与物理一致性的建模能力。整体框架如图1所示，基于潜空间扩散模型（Latent Diffusion Model，LDM）构建，并在参数适配、损失设计与条件编码三个层面进行系统性扩展。

在基础扩散框架中，输入影像首先经由VAE Encoder映射至潜空间表示，并在正向扩散过程中逐步加入高斯噪声；UNet在文本条件与时间步约束下，对噪声残差进行监督训练。对于引入ControlNet的可控生成场景，结构条件经独立分支编码后注入主干UNet，以约束去噪过程中的空间布局。在预测阶段，模型从高斯噪声出发，在文本与结构条件的共同引导下迭代去噪得到潜空间表示，并通过VAE Decoder将去噪后的潜变量解码为最终影像。

在本文提出的框架中，整体训练与预测流程与上述潜空间扩散模型保持一致。不同之处在于，为适配四波段（RGB+NIR）遥感影像的输出需求，额外微调VAE编解码器。本文在冻结预训练主干参数的前提下，对UNet与VAE编解码部分联合引入低参数适配机制，使潜空间建模与输出影像在光谱分布及物理特性上保持一致。

在空间控制层面，引入了文本感知空间语义编码（Text-Aware Spatial Semantic Encoding，TASSE）模块来处理语义分割掩膜，将类别特定的文本嵌入映射至空间布局，并通过ControlNet分支引导UNet的特征生成，从而实现对地物分布的精确控制。

1.2 光谱感知参数高效适配结构

在多光谱遥感数据上对大型扩散模型进行全量

微调计算成本高昂，且在规模有限的遥感数据集上容易出现过拟合。为此，本文采用参数高效微调策略，基本原则是冻结预训练主干以保留 RGB 域的生成先验，仅在特定位置引入少量可训练参数以学习域间差异。该策略能够有效工作，在于近红外波段与可见光波段在空间纹理层面存在一定的物理相关性，使得预训练 RGB 模型的生成先验对近红外波段的学习具有参考价值（梅少辉等，2021）；同时，引入模块的参数量相对较小，对预训练特征空间的扰动有限。需要指出的是，当目标域与 RGB 域的成像机制差异较大时，如合成孔径雷达（Synthetic Aperture Radar，SAR）图像或热红外波段，上述相关性不再成立，此时单纯依赖 PEFT 进行跨域适配的效果将受到限制。

该策略在保持大部分预训练权重冻结的同时，仅向 UNet 和 VAE 的特定组件引入轻量级可训练模块。与通用 PEFT 方法中模块配置相对统一不同，本文针对多光谱遥感适配的两类核心挑战进行了差异化设计：一是 RGB 语义先验向多光谱特征空间的映射问题，由 LoRA 与 Adapter 分别从注意力层与前馈网络两个维度协同处理；二是遥感图像中多尺度地物结构的空间建模问题，由 MSM 与 MONA 分别从局部纹理与全局上下文两个层面加以应对。更重要的是，上述模块的训练过程受 2.3 节中 NDVI 物理约束的协同约束，使参数优化具有明确的遥感物理导向，而非纯粹的视觉数据拟合。本文的具体方法应用了四种适配机制：针对注意力层的低秩自适应（LoRA）（Hu 等，2021）、针对前馈网络的适配器（Adapter）（Houlsby 等，2019）、用于空间特征增强的多尺度调制模块（Multi-Scale Modulation，MSM）以及 Mona 模块（Yin 等，2024）。

基于 LoRA 的注意力层适配，将可训练的低秩分解矩阵注入冻结的预训练层中。具体而言，在 UNet 和 VAE 的注意力机制中，本文在查询（Q）、值（V）和输出（O）投影矩阵中引入 LoRA 模块。对于权重矩阵 $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$，其更新公式为：

$$\boldsymbol{h} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{A}\boldsymbol{x} \tag{1}$$

式中，$\boldsymbol{B} \in \mathbb{R}^{d_{\text{out}} \times r_L}$ 和 $\boldsymbol{A} \in \mathbb{R}^{r_L \times d_{\text{in}}}$ 为秩 $r_L = 4$ 的低秩矩阵。在初始化方面，矩阵 $\boldsymbol{A}$ 采用 Kaiming 均匀分布初始化，矩阵 $\boldsymbol{B}$ 初始化为零，以确保训练过程从预训练状态平稳启动。

在 UNet 的 Transformer 模块中，本文在前馈神经网络（Feed-Forward Network，FFN）旁并行插入 Adapter 模块。该适配器采用“瓶颈”结构，由降维投影、高斯误差线性单元（Gaussian Error Linear Unit，GELU）激活函数和升维投影组成：

$$\text{Adapter}(\boldsymbol{x}) = \boldsymbol{W}_{\text{up}} \cdot \text{GELU}\left(\boldsymbol{W}_{\text{down}} \cdot \boldsymbol{x}\right) \tag{2}$$

式中，$\boldsymbol{W}_{\text{down}} \in \mathbb{R}^{d_{\text{in}} \times r_A}$ 和 $\boldsymbol{W}_{\text{up}} \in \mathbb{R}^{r_A \times d_{\text{out}}}$ 为可学习参数，瓶颈维度设定为 $r_A = 16$。该设计使得模型能够以极少的参数量学习领域特定的特征变换。

为增强模型对遥感影像中复杂空间结构的多尺度特征适配能力，本文引入多尺度特征增强机制，分别从局部纹理建模与全局上下文建模两个层面进行设计。

一方面，针对 VAE 解码器以及 UNet 上采样阶段 ResNet 模块中的卷积层，本文在其后插入所设计的多尺度调制模块（Multi-Scale Modulation，MSM），如图 2 所示。该模块利用不同尺寸的深度卷积核聚合特征：

$$\boldsymbol{y} = \boldsymbol{x} + s \cdot \text{SiLU}\left(\frac{1}{|\mathcal{K}|}\sum\nolimits_{k \in \mathcal{K}} \text{DWConv}_k(\boldsymbol{x})\right) \tag{3}$$

式中，对于上采样模块，卷积核集合设定为 $\mathcal{K} = \{3, 5, 7\}$；对于中间模块，设定为 $\mathcal{K} = \{3, 5\}$。为可学习的缩放因子。

另一方面，在 UNet 中间层的 Transformer 模块中，引入 Mona 模块（Yin 等，2024）以增强全局上下文建模能力。Mona 通过在注意力机制中引入轻量级的多尺度卷积算子与可学习缩放调制，将局部多尺度感受野显式注入全局建模过程，从而同时增强长程依赖建模能力与细粒度空间结构表达能力。

1.3 物理引导的光谱一致性学习

常规生成模型通常独立处理图像各个通道，从而忽略了遥感数据中内在的光谱相关性。针对这一问题，本文引入了基于归一化植被指数（NDVI）的物理约束。NDVI 是表征植被特征的常用指标，同时，NDVI 并非仅是一种经验性遥感指数，其红光-近红外辐射关系受地表物理特性约束，在植被覆盖区域具有较强的跨场景稳定性。将其引入生成模型的训练过程，本质上是在优化目标中嵌入一个与地表辐射机理相符的约束，引导模型学习两个波段之间正确的物理关联，而非独立拟合各通道的像素分布。NVDI 定义基于红光（Red）与近红外（NIR）波段反射率的差异：

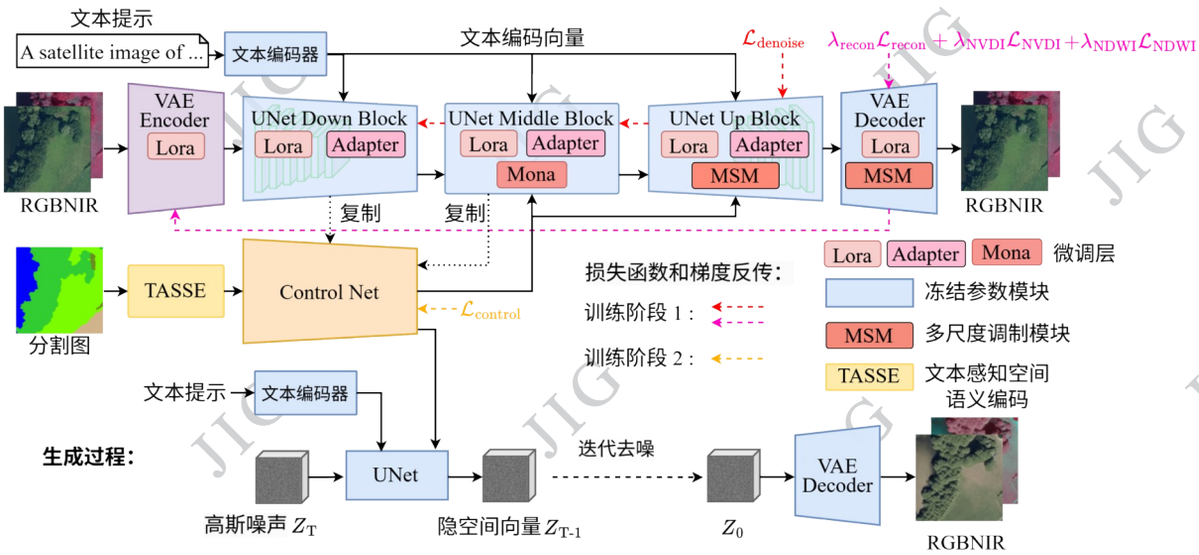

图 1　模型总体结构

Fig. 1　The overall of model

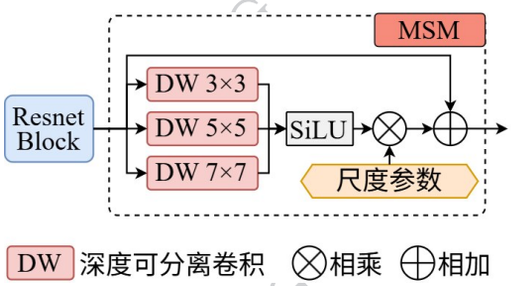

图 2　MSM 模型总体结构

Fig. 2　The overall of MSM

$$\mathrm{NDVI}(R,N)=\frac{N-R}{N+R+\varepsilon} \tag{4}$$

式中，R 和 N 分别代表红光与近红外波段的反射率，ε 为保证数值稳定性的微小常数。植被在红光波段具有较低的反射率而在近红外波段具有较高的反射率，这种波段响应差异是 NDVI 能够稳定表征植被状态的物理基础。需要指出的是，在水体、深阴影或裸土等非植被区域，红光与近红外的辐射关系与植被区域存在差异，NDVI 在此类区域的物理解释能力较弱，直接对全图施加该约束会引入无效的优化信号。本文利用植被区域的物理先验在 NDVI 特征空间内约束生成影像与真实影像的一致性，并通过亮度阈值掩膜将约束范围限定在具有明确光谱响应的像素上，以降低上述区域带来的计算不稳定问题。令 $\boldsymbol{x}\in[0,1]^{B\times 4\times H\times W}$ 表示真值影像，$\hat{\boldsymbol{x}}$ 表示 VAE 解码器输出的生成影像。将真值的红光与近红外通道记为 R 和 N，生成影像的对应通道记为 $\hat{R}$ 和 $\hat{N}$。在地表反射率较低的区域（如水体或深阴影），$R+N$ 的数值较小。这会导致 NDVI 计算数值不稳定，且对噪声极度敏感。为规避该问题，本文仅对满足亮度阈值的像素计算一致性损失。有效像素集合 $\boldsymbol{\Omega}$。定义如下：

$$\boldsymbol{\Omega}=\left\{(b,h,w)\mid R_{b,h,w}+N_{b,h,w}>\tau\right\} \tag{5}$$

式中，τ 为经验阈值，取为 0.1。该阈值的设定主要基于遥感图像的辐射特征与模型训练稳定性。在实际多光谱影像中，水体或深阴影等非植被区域的红光与近红外反射率普遍较低。若直接对全图计算 NDVI 一致性损失，由于此类区域的计算分母 $(R+N)$ 较小，NDVI 对特定波段的偏导数，如 $\frac{\partial \mathrm{NDVI}}{\partial N}=\frac{2R}{(N+R+\epsilon)^2}$ 会被显著放大。这不仅会导致计算结果对微小噪声极其敏感，还容易在反向传播过程中引发梯度爆炸。通过设定 $\tau=0.1$ 作为掩膜条件，可以有效滤除低信噪比像素，使物理约束集中作用于具有明确光谱响应的地物区域，从而保证网络优化的平稳进行。

此外，直接优化 NDVI 可能会干扰可见光波段（RGB）的纹理建模。考虑到红光波段的辐射特性已在预训练 RGB 模型中得到充分学习，而近红外波段是本文新引入的生成目标，NDVI 一致性损失的优

化重点应落在近红外分支上。为防止此类干扰，本文在计算损失时对生成的红光通道应用梯度停止(stop-gradient)操作。该操作确保 NDVI 损失主要用于更新近红外分支，而不会影响红光通道。具体的输入形式为：

$$\mathrm{NDVI}\left(\hat{R}^{*},\hat{N}\right),\quad \hat{R}^{*}=\mathrm{stopgrad}\left(\hat{R}\right) \tag{6}$$

最终的 NDVI 一致性损失采用 Smooth L1(Huber)损失函数定义：

$$\mathcal{L}_{\mathrm{NDVI}}=\lambda_{\mathrm{ndvi}}\cdot\frac{1}{|\boldsymbol{\Omega}|}\sum_{(b,h,w)\in\boldsymbol{\Omega}}\mathcal{H}\left(\mathrm{NDVI}\left(\hat{R}^{*},\hat{N}\right)-\mathrm{NDVI}(R,N)\right) \tag{7}$$

式中，$\mathcal{H}(\cdot)$表示 Smooth L1 损失函数，λ_{ndvi}为损失权重。

除植被区域外，水体在多光谱影像中同样具有典型的波段响应特征：绿光波段反射率高于近红外波段，与植被的辐射规律相反。为此，本文进一步引入归一化水体指数(Normalized Difference Water Index, NDWI)约束，对绿光与近红外波段之间的物理关联进行显式建模：

$$\mathrm{NDWI}(G,N)=\frac{G-N}{G+N+\varepsilon} \tag{8}$$

式中，G和N分别代表绿光与近红外波段的反射率。与NDVI约束的处理方式一致，同样采用亮度阈值掩膜过滤低信噪比像素，有效像素集合定义为：

$$\boldsymbol{\Omega}'=\{(b,h,w)\,|\,G_{b,h,w}+N_{b,h,w}>\tau\} \tag{9}$$

对生成的绿光通道施加梯度停止操作，使NDWI损失的梯度仅回传至近红外分支：

$$\mathrm{NDWI}(\hat{G}^{*},\hat{N}),\hat{G}^{*}=\mathrm{stopgrad}(\hat{G}) \tag{10}$$

NDWI一致性损失定义为：

$$\mathcal{L}_{NDWI}=\lambda_{ndwi}\cdot\frac{1}{|\boldsymbol{\Omega}'|}\sum_{(b,h,w)\in\boldsymbol{\Omega}'}\mathcal{H}(\mathrm{NDWI}(\hat{G}^{*},\hat{N})-\mathrm{NDWI}(G,N)) \tag{11}$$

第一阶段总损失函数式同步更新为：

$$\mathcal{L}_{\mathrm{Stage1}}=\mathcal{L}_{\mathrm{denoise}}+\lambda_{\mathrm{recon}}\mathcal{L}_{\mathrm{recon}}+\lambda_{\mathrm{ndvi}}\mathcal{L}_{\mathrm{NDVI}}+\lambda_{\mathrm{ndwi}}\mathcal{L}_{\mathrm{NDWI}} \tag{12}$$

1.4 文本感知空间语义编码机制

为有效利用分割图中的高层语义信息并确保其与文本提示对齐，本文提出了文本感知空间语义编码机制。该方法利用分割掩膜作为结构约束，将特定类别的文本嵌入显式映射至空间特征图，如图3所示。这一机制为扩散模型提供了空间一致的语义条件。

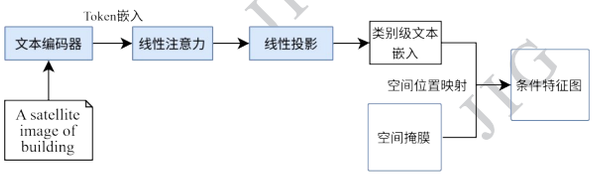

图3 TASSE模型总体结构

Fig. 3 The overall of TASSE

文本语义表征。对于每个语义类别c，首先利用预训练文本编码器处理其对应的文本短语，以获取Token级的隐含表征：

$$\boldsymbol{H}_{c}=\left[\boldsymbol{h}_{c,1},\boldsymbol{h}_{c,2},\cdots,\boldsymbol{h}_{c,T}\right],\boldsymbol{h}_{c,t}\in\mathbb{R}^{D_{\mathrm{text}}} \tag{8}$$

式中，T表示文本Token的长度。D_{text}为预训练文本编码器固有的输出特征维度。为从这些Token表征中获取单一的类别级嵌入，本文引入了一个可学习的线性注意力层。该层对每个Token的重要性进行建模，并沿序列维度进行聚合。每个Token的权重计算如下：

$$\alpha_{c,t}=\frac{\exp\left(\boldsymbol{w}^{\top}\boldsymbol{h}_{c,t}\right)}{\sum_{k=1}^{T}\exp\left(\boldsymbol{w}^{\top}\boldsymbol{h}_{c,k}\right)} \tag{9}$$

式中，$\boldsymbol{w}\in\mathbb{R}^{D_{\mathrm{text}}}$为可学习参数。最终的类别级文本嵌入定义为加权和：

$$\boldsymbol{e}_{c}=\sum_{t=1}^{T}\alpha_{c,t}\boldsymbol{h}_{c,t},\boldsymbol{e}_{c}\in\mathbb{R}^{D_{\mathrm{text}}} \tag{10}$$

随后，利用线性投影将类别嵌入映射至ControlNet主干网络的通道维度：

$$\boldsymbol{e}_{c}'=\boldsymbol{W}_{\mathrm{p}}\boldsymbol{e}_{c},\boldsymbol{W}_{\mathrm{p}}\in\mathbf{R}^{D_{\mathrm{cond}}\times D_{\mathrm{text}}} \tag{11}$$

式中，D_{cond}为ControlNet接收外部条件输入的固定通道维度。该投影操作确保了生成的语义特征能够与预训练主干网络的特征空间严格对齐。给定语义分割图$\boldsymbol{S}\in\mathbb{R}^{1\times H\times W}$，式中每个像素值对应一个类别ID，首先将其下采样以匹配潜在空间的分辨率(H',W')。针对每个类别c，构建一个二进制空间掩膜：

$$\boldsymbol{M}_{c}(x,y)=\mathbb{I}\left(\boldsymbol{S}(x,y)=c\right) \tag{12}$$

式中，$\mathbb{I}(\cdot)$为指示函数。通过将投影后的类别级文本嵌入$\tilde{\boldsymbol{e}}_{c}$扩展至对应的空间位置，生成最终的条件特征图：

$$\boldsymbol{F}(x,y)=\tilde{\boldsymbol{e}}_{c}\quad \text{if}\ \boldsymbol{M}_{c}(x,y)=1 \tag{13}$$

该过程生成了一个显式编码空间语义的条件张量：

$$\boldsymbol{F} \in \mathbb{R}^{320 \times H' \times W'} \tag{14}$$

1.5 训练策略

本文采用两阶段训练策略，旨在逐步实现模型向多光谱领域的适配，并随后赋予其空间控制能力。这种解耦机制有效保障了模型的收敛稳定性，并避免了光谱适配与空间对齐任务间的相互干扰。

第一阶段：多光谱适配训练

本阶段旨在将预训练 RGB 模型适配至四波段遥感数据。为保留 Stable Diffusion 模型原有的生成先验，本文冻结其原始参数，仅对 UNet 中引入的适配模块以及 VAE 的输入输出层参数进行更新。目标函数由三部分构成：去噪损失、重建损失及物理一致性损失。式中，去噪损失 $\mathcal{L}_{\text{denoise}}$ 用于优化 UNet 内部的可训练参数，其衡量的是添加噪声与模型预测噪声之间的差异：

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{z_0,\boldsymbol{c},t,\epsilon}\left[\left(\epsilon - \epsilon_\theta\left(z_t,t,\boldsymbol{c}\right)\right)_2^2\right] \tag{15}$$

式中，z_t 表示时间步 t 的潜在编码，$\boldsymbol{c}$ 为文本条件，ϵ 为高斯噪声，ϵ_θ 代表去噪网络。为实现四通道数据处理，本文对 VAE 的首尾层进行微调。重建损失 $\mathcal{L}_{\text{recon}}$ 计算输入图像与重建图像之间的均方误差（Mean Squared Error，MSE）：

$$\mathcal{L}_{\text{recon}} = \left|\boldsymbol{x} - \hat{\boldsymbol{x}}\right|_2^2 \tag{16}$$

式中，$\boldsymbol{x}$ 为多光谱影像真值，$\hat{\boldsymbol{x}}$ 为 VAE 解码器的输出。为确保光谱有效性，本文引入了式(8)定义的 NDVI 一致性损失 $\mathcal{L}_{\text{NDVI}}$。第一阶段的总损失函数为：

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{ndvi}}\mathcal{L}_{\text{NDVI}} \tag{17}$$

式中，λ_{recon} 与 λ_{ndvi} 为平衡各项权重的超参数。

第二阶段：空间控制训练

第二阶段侧重于利用语义分割掩膜实现精确的空间控制。在此阶段，本文冻结第一阶段优化的所有参数（含光谱适配模块），并初始化 ControlNet 分支及本文提出的文本感知空间语义编码模块。模型基于文本提示 $\boldsymbol{c}$ 和空间条件张量 $\boldsymbol{F}$ 进行生成训练。优化目标为噪声预测的重建损失：

$$\mathcal{L}_{\text{Stage2}} = \mathbb{E}_{z_0,\boldsymbol{c},\boldsymbol{F},t,\epsilon}\left[\ /\!/\ \epsilon - \epsilon_\theta\left(z_t,t,\boldsymbol{c},\boldsymbol{F}\right)\ /\!/\ _2^2\right] \tag{18}$$

通过保持主分支冻结，该阶段使模型能够在不损害第一阶段习得的多光谱生成能力的前提下，学习从语义布局到空间特征的映射关系。

2 实验与分析

2.1 实验数据集

为验证所提出方法的有效性，在三组公开数据集上实验。

1）FLAIR 数据集（Garioud 等，2022）：FLAIR 是一个大规模的法国土地覆盖制图基准数据集。该数据集由空间分辨率为 0.2 米的航空影像组成，包含与本研究密切相关的四个光谱波段：红光、绿光、蓝光及近红外（NIR）。原始高分辨率影像被裁剪为 512×512 像素的图块。在标注方面，数据集涵盖了 13 个语义类别。在数据划分上，共有 61,712 个图块用于训练和验证，另有两个独立的测试集，分别包含 15,700 和 16,050 个图块。

2）Five-Billion-Pixels 数据集（Tong 等，2022）：该数据集提供了分辨率为 4 米的高分二号（Gaofen-2）影像，旨在支持国家级尺度的土地覆盖制图任务。数据集包含 150 幅高分辨率影像，共标注了 24 个土地覆盖类别。为适应模型训练需求，原始影像裁剪为 512×512 像素的图块。经处理后，该数据集共提供 61,712 对训练样本以及 15,700 对测试样本。

3）IRSAMap 数据集（Meng 等，2025）：IRSAMap 是一个面向高分辨率土地覆盖矢量制图的全球性数据集。使用其吉林一号影像数据源的 RGBNIR 四波段影像，空间分辨率统一为 0.5 米。数据覆盖范围广泛，包含跨越六大洲的 79 个典型区域。影像被裁剪为 512×512 像素的图块。该数据集包含 10 个语义类别的标注信息，总计约 20,000 个图块，涉及超过 180 万个独立的地物实例。

2.2 实验评价指标

为了定量评估生成多光谱影像的质量，本文选取了 NDVI-RMSE、NDWI-RMSE、CLIP Score、LPIPS 和 mIoU 4 个评价指标，分别从光谱保真度、语义一致性、感知质量以及下游应用价值等方面进行综合考量。

NDVI-RMSE：多光谱生成任务要求保持波段间严格的物理关联。本文利用归一化植被指数（normalized difference vegetation index，NDVI）的均方根误差（root mean square error，RMSE）来量化红光波段与近红外波段（NIR）间的光谱一致性。NDVI 的计

算公式为 $\mathrm{NDVI}=\frac{\mathrm{NIR}-\mathrm{Red}}{\mathrm{NIR}+\mathrm{Red}+\epsilon}$。为增强评估的鲁棒性，本文设定反射率阈值（式6）以剔除阴影及水体噪声，仅对有效像素计算RMSE。NDVI-RMSE值越低，表明生成的植被光谱特征越接近真实分布，光谱保真度越高。

NDWI-RMSE：除植被外，水体是多光谱遥感影像中另一类具有典型光谱特征的地物，其绿光波段反射率高于近红外波段反射率，与植被的波段响应规律相反。本文引入归一化水体指数（normalized difference water index，NDWI）的均方根误差作为补充光谱评估指标，用于量化绿光波段与近红外波段之间的光谱一致性。NDWI的计算公式为 $\mathrm{NDWI}=\frac{\mathrm{Green}-\mathrm{NIR}}{\mathrm{Green}+\mathrm{NIR}+\varepsilon}$。与NDVI-RMSE的处理方式一致，同样设定反射率阈值以剔除低信噪比像素，仅对有效像素计算RMSE。NDWI-RMSE值越低，表明生成影像在水体等地物区域的波段关系越符合真实物理分布。

CLIP Score：用于评估生成图像与输入文本提示词之间的语义对齐度。本文并未直接使用在自然图像上预训练的标准CLIP模型，而是采用了适配遥感数据的MovSeg（Ji 等，2026）文本与图像编码器。该方法能够更准确地捕捉地物类别等领域特定语义特征。CLIP Score得分越高，意味着生成内容与文本描述的语义一致性越强。

LPIPS：利用学习感知图像块相似度（learned perceptual image patch similarity，LPIPS）（Zhang R 等，2018）评估生成图像的视觉真实感。该指标基于深度特征计算生成图像与真实影像（ground truth）之间的特征距离。LPIPS值越低，表明生成图像在感知质量上与真实数据越相似。

mIoU：为验证生成数据在下游任务中的应用价值，本文设计了开放词汇分割实验。具体而言，用已有的分割标注引导生成对应的多光谱影像，利用分割图像对训练分割模型，并使用平均交并比（mean intersection over union，mIoU）在真实测试集上评估其性能。实验中采用多光谱影像开放词汇分割模型MovSeg（Ji 等，2026）作微调训练。mIoU值越高，说明生成图像包含逼真的结构与语义模式，能够有效支持判别模型的训练。

2.3 实验设置

本方法基于PyTorch框架实现，所有实验均在NVIDIA RTX 3090 Ti GPU上完成。模型采用两阶段训练策略。第一阶段主要用于光谱适配，模型以遥感数据集上的DiffusionSat（Khanna 等，2023）预训练权重初始化，输入分辨率为512×512，训练12,000步，学习率设为 1×10^{-4}，批大小为8。该阶段引入低参数微调策略以降低计算开销。

第二阶段用于训练空间控制模块，在第一阶段权重基础上继续训练，学习率调整为 1×10^{-5}，批大小为8，训练12,000步，并启用ControlNet分支以学习空间约束。在两个训练阶段中，均采用bf16混合精度训练以降低显存占用。

本文方法框架中的文本提示由数据集的语义分割图中各类别的像素占比自动构造，并进一步加入影像类型及时间信息，以增强全局语义约束的稳定性与一致性。

2.4 生成多光谱影像的定量比较

表1展示了本文提出的方法与4种主流方法的定量对比结果。对比实验选取了两种通用视觉模型T2I-Adapter（Mou 等，2024）和ControlNet（Zhang L 等，2023）以及两种遥感专用模型CRS-Diff（Tang 等，2024）和EarthSynth（Pan J 等，2025）。评估指标包括NDVI-RMSE、CLIP Score和LPIPS。符号“-”表示基线方法仅支持RGB合成，无法生成近红外波段，因此NDVI-RMSE与NDWI-RMSE均无法计算。每个类别中的最佳结果以粗体显示。实验结果表明，本文方法在各项指标上均表现最优。具体而言，本文方法取得了最低的平均LPIPS=1.5452和最高的CLIP Score=0.1607，说明与其他方法相比，本文方法生成的图像具有更高的感知质量与更优的语义对齐度。此外，值得注意的是，本文方法是唯一能够生成有效近红外数据的框架，平均NDVI-RMSE为0.3173，平均NDWI-RMSE为0.2854，对比方法在这两项指标上均无法参与比较。

图2展示了本文方法与CRS-Diff及EarthSynth的定性对比结果。生成过程受输入语义分割掩码与描述场景特征的文本提示词（如“茂密的森林”、“有道路的城市区域”）共同约束。为评估光谱有效性，图中展示了各模型的RGB生成结果，并对比了本文方法的假彩色合成图（NIR-Red-Green）与真实影像（ground truth）。

在视觉质量与文本一致性方面，CRS-Diff生成的图像往往较为模糊，且在复杂的植被与城市区域缺乏细粒度纹理；EarthSynth生成的特征虽较为清晰，但常伴有视觉伪影及不自然的噪声模式。相比之下，本文方法生成的图像具有高保真度，且能准确响应文本提示词。以“茂密的森林”为例，本文方法能够生成与语义类别一致的丰富植被纹理，而对比方法通常生成较为通用的内容，难以完全契合具体的文本描述。

在空间一致性方面，本文方法与输入语义掩码保持了精确对齐。道路、建筑物及水体等不同地物的边界严格遵循布局约束。相比之下，对比模型偶尔出现结构失控，导致物体形状扭曲或与语义标签不匹配。此外，如假彩色合成图所示，本文方法能够正确合成近红外（NIR）波段。在NIR-Red-Green合成图中，植被区域呈现鲜红色，符合真实影像中的物理光谱特征。引入的光谱一致性损失约束了红光与近红外波段间的物理关联，确保了生成数据的光谱准确性。

表1　在不同数据集上，对所提出的方法与现有最先进方法进行了定量比较

Table 1　Quantitative comparison of state-of-the-art approaches on different datasets.

方法	IRSAMap				FLAIR				Five-Billion-Pixels				平均			
	NDVI-RMSE ↓	NDWI-RMSE ↓	CLIP ↑	LPIPS ↓	NDVI-RMSE ↓	NDWI-RMSE ↓	CLIP ↑	LPIPS ↓	NDVI-RMSE ↓	NDWI-RMSE ↓	CLIP ↑	LPIPS ↓	NDVI-RMSE ↓	NDWI-RMSE ↓	CLIP ↑	LPIPS ↓
T2I-Adapter	–	–	0.0604	1.6713	–	–	0.148	1.4966	–	–	0.1120	1.6377	–	–	0.1068	1.6019
ControlNet	–	–	0.0725	1.6900	–	–	0.158	1.4500	–	–	0.1160	1.6400	–	–	0.1155	1.5933
CRS-Diff	–	–	0.0803	1.6732	–	–	0.166	**1.3916**	–	–	0.1194	1.6242	–	–	0.1219	1.563
EarthSynth	–	–	0.0711	1.7259	–	–	0.1637	1.4962	–	–	0.1119	1.6613	–	–	0.1156	1.6278
所提方法	**0.3008**	**0.2714**	**0.1350**	**1.5444**	**0.2507**	**0.2231**	**0.1794**	1.5178	**0.4005**	**0.3618**	**0.1676**	**1.5734**	**0.3173**	**0.2854**	**0.1607**	**1.5452**

2.5　生成多光谱影像的定性分析

图3直观展示了所提框架在跨季节场景生成中的时序泛化能力。实验通过固定语义分割布局作为空间约束，利用差异化的文本提示词（将原始提示文本中月份和季节替换）驱动模型生成，旨在从视觉真实感和光谱一致性两个维度综合评估模型性能。

真彩色影像（第1、3行）展示了生成图像符合自然规律的地表物候演变特征。植被区域呈现出显著的季节性更替：从春夏季节茂密的深绿色，逐渐过渡至秋季植被衰退期的枯黄与褐色调；冬季场景则通过积雪覆盖及低太阳高度角下的独特光照条件，展现了典型的寒冷气候特征。值得注意的是，在城市建筑与道路等不透水面区域，模型展现了较强的几何结构保持能力，将季节性纹理变化限定在植被与自然地表范围内，而未改变人工地物的拓扑形态，证明了空间语义控制的精确性。

标准假彩色合成图（NIR-Red-Green，第2、4行）进一步验证了NDVI一致性损失对光谱维度的有效约束。在生长旺盛的春夏季，健康植被因其叶肉细胞结构对近红外波段的高反射特性，在假彩色合成图中呈现高饱和度的亮红色；而在植被凋落或被积雪覆盖的冬季，以及水体和不透水面区域，由于近红外反射率显著降低，图像呈现青色或暗沉色调。这种视觉与光谱特征的高度对齐，充分证实了本文提出的物理约束机制能够确保生成波段具备辐射物理意义上的合理性，同时参数高效适配策略成功保留了预训练基础模型的高保真生成先验，实现了在复杂多光谱场景下的高质量可控合成。

2.6　利用生成影像的分割模型训练效果分析

表2展示了在Five-Billion-Pixels、FLAIR和IRSAMap 3个数据集上的开放词汇分割定量结果。实验中采用多光谱影像开放词汇分割模型MovSeg（Ji等，2026）作微调训练。实验主要对比了3种训练数据设定的模型测试性能：仅利用真实数据训练的基准模型（记为GT）、在真实数据基础上增加本文方法生成影像的模型（记为GT+TASSE Gen），以及增加标准RGB编码生成影像的模型（记为GT+RGB Gen）。

在 Five-Billion-Pixels 数据集上，GT+TASSE Gen 方法取得了最高的 mIoU(63.95%)，优于基准模型(62.99%)及 GT+RGB Gen 方法(63.47%)。在 FLAIR 数据集上，本文方法同样实现了最优性能(59.13%)；相比之下，GT+RGB Gen 的结果(58.65%)甚至低于基线模型(58.97%)，这表明标准的 RGB 生成方法在此类数据上产生了负面影响。在 IRSAMap 数据集上，两种生成式增强方法均优于基线得分(91.04%)；其中本文方法略微领先(91.39%)。综合对比结果表明，引入类别文本嵌入的空间编码方式相比直接使用 RGB 掩膜编码，在多数数据集上可为下游分割模型带来一定的性能提升，但提升幅度因数据集特性而有所差异，在 IRSAMap 上的增益相对有限。

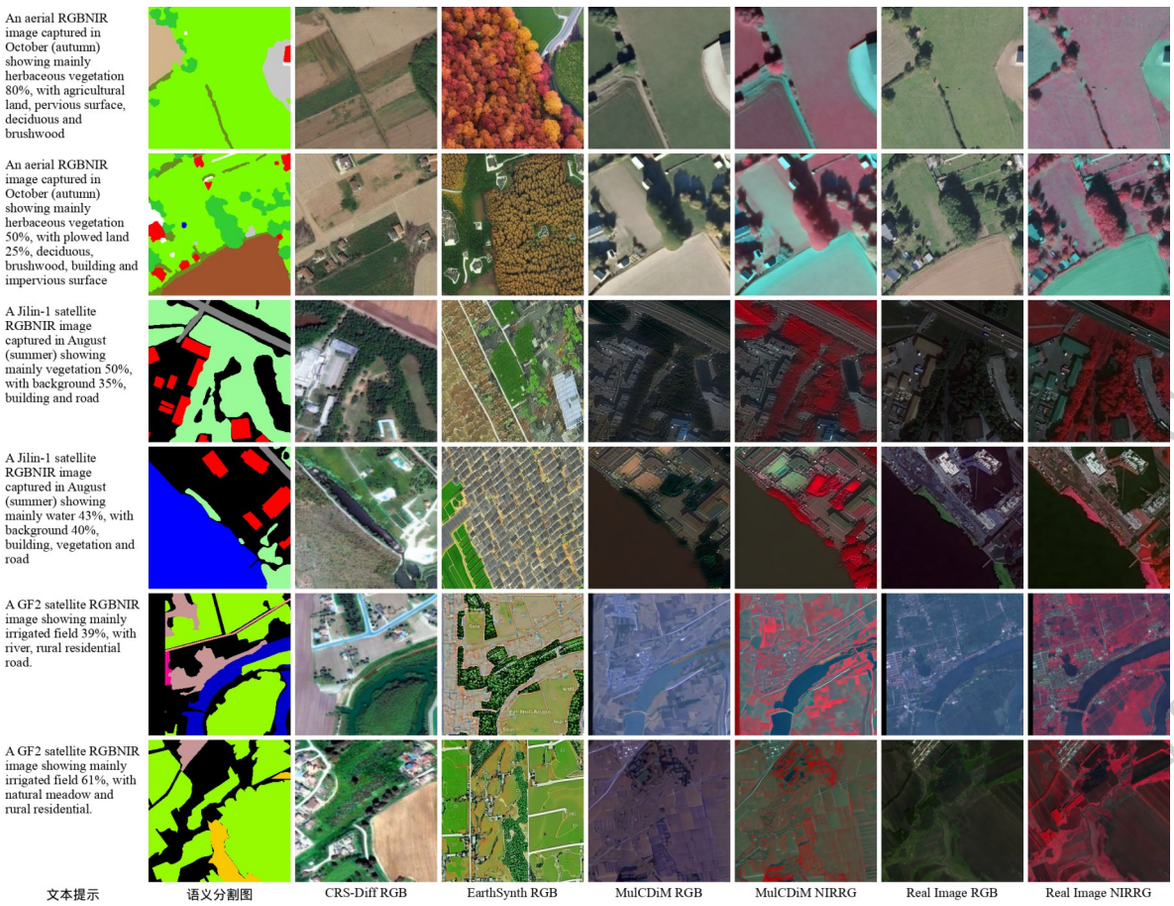

图4　将所提出的方法在不同数据集上与基线方法进行可视化比较。

Fig. 4　Visual comparison of the proposed method against baseline approaches on different datasets

2.7　消融实验

为了验证所提框架中各关键组件在四波段遥感图像生成任务中的有效性，本文采用逐步叠加的策略在统一基线模型上进行了消融实验。表3展示了在训练第一阶段，各模块引入后对 NDVI-RMSE、CLIP Score 及 LPIPS 指标的定量影响。

基线模型性能：实验以仅微调 VAE 输入输出卷积层的模型为基准。该设置虽然实现了从 RGB 到 RGB-NIR 四波段输出的维度适配，但由于缺乏对遥感领域特征的深层适配，生成质量有限。

LoRA 模块的作用：在引入 LoRA 对注意力模块进行低秩微调后，CLIP Score 提升了 0.018。这一结果与 LoRA 在注意力层引入低秩可训练参数、增强跨模态特征对齐的设计预期一致。参数高效的注意力调整增强了文本条件与生成影像间的跨模态对齐能力，验证了高层语义建模在遥感生成任务中的关键作用。

Adapter 与 MONA 的联合增益：进一步集成 Adapter 与 MONA 模块后，三项指标均呈现持续改善趋势。Adapter 通过在前馈网络中嵌入轻量级可学习映射，提升了特征变换的灵活性；而 MONA 利用显式调制机制增强了通道与空间的自适应响应能力，

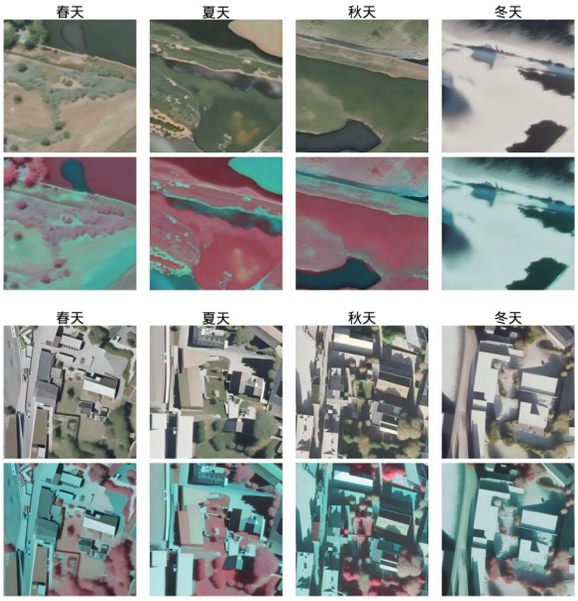

图5　季节性泛化能力的可视化

Fig. 5　Visualization of seasonal generalization capabilities.

表2　在不同数据集上开放词汇分割性能（mIoU）的定量比较

Table 2　Quantitative comparison of open-vocabulary segmentation performance (mIoU) on different datasets.

训练数据	数据集(mIoU)		
	Five-Billion-Pixels	Flair	IRSAMap
GT	62.992	58.974	91.044
GT+TASSE Gen	**63.956**	**59.131**	**91.388**
GT+RGB Gen	63.471	58.654	91.237

使模型在保持语义一致性的同时，有效提升了感知质量。

多尺度调制模块（MSM）通过多尺度深度卷积聚合局部纹理特征，以增强空间细节表达。实验结果显示，引入MSM后LPIPS指标降低了0.084，在单步改进中感知质量提升更高。在遥感图像生成中，多尺度空间建模对感知质量具有影响。MSM利用多尺度特征的显式调制，有效提升了模型对局部纹理、边界结构及空间层次关系的刻画能力，从而显著改善了生成影像的结构一致性与细粒度纹理表现。

NDVI一致性损失的光谱约束：在完整架构中引入NDVI一致性损失后，NDVI-RMSE进一步下降0.021，在此基础上，加入NDWI一致性损失对绿光与近红外波段施加额外约束后，NDWI-RMSE下降0.023。NDVI-RMSE也有小幅改善（-0.011），表明两类物理约束通过对近红外波段的协同优化存在一定的互补效果，而CLIP Score和LPIPS基本保持不变，引入物理约束未对视觉生成质量造成干扰。

综上所述，完整框架在三项指标上均取得最优性能，充分验证了各模块在语义对齐、空间感知与光谱保真度方面的有效性与互补性。

在训练第二阶段，本文基于第一阶段的最优权重进一步微调ControlNet分支，对比了将分割图转为RGB输入的基线方法与TASSE机制。TASSE在NDVI-RMSE和LPIPS上均有改善。实验结果显示，TASSE在CLIP Score上的提升较小（+0.001），原因在于该指标主要反映全局语义一致性，受文本提示词影响较大，而TASSE的优势更多体现在局部区域的语义与结构对齐上，这与其在NDVI-RMSE和LPIPS上的改进更为一致。这表明TASSE通过建立类别特定的文本嵌入与空间布局的映射关系，能够更精确地约束生成影像的局部结构与光谱一致性。

3　结　论

针对深度学习在多光谱遥感应用中面临的数据稀缺与标注成本高昂问题，以及现有RGB基础生成模型难以直接迁移至多光谱域且全量训练计算开销巨大的现状，本文提出了一种面向多光谱遥感图像生成的参数高效适配扩散模型。该框架通过低秩自适应微调与物理约束机制，在低资源消耗下实现了高质量、可控的多光谱数据生成。主要研究结论如下：

1）模型适配策略的有效性：本文提出的参数高效微调策略，通过在冻结的预训练RGB影像生成主干网络中嵌入低参数微调的光谱与空间纹理适配模块，有效解决了RGB像域向多光谱图像域迁移时的通道不匹配与特征分布差异问题。实验表明，该策略在大幅降低训练参数量与显存占用的同时，成功保留了基础模型强大的生成先验，实现了对四波段（RGB+NIR）遥感影像的高保真合成。但该策略的有效性建立在源域与目标域之间存在一定特征相关性的前提下，对于与RGB影像差异较大的模态，其适配能力存在相应的局限。

2）物理一致性与空间可控性：针对传统生成模型忽视光谱物理关联的缺陷，本文引入基于NDVI的光谱一致性损失，强制约束了红光与近红外波段

表3　消融实验表

Table 3　Ablation Study

训练阶段	方法	NDVI-RMSE ↓	NDWI-RMSE ↓	CLIP ↑	LPIPS ↓
阶段 1	微调输入输出卷积(基准)	0.413	0.351	0.111	1.866
	+LoRA	0.406 (−0.007)	0.344 (−0.007)	0.129 (**+0.018**)	1.804 (−0.062)
	+LoRA+Adapter	0.395 (−0.011)	0.336 (−0.008)	0.140 (+0.011)	1.747 (−0.057)
	+LoRA+Adapter+MONA	0.392 (−0.003)	0.332 (−0.004)	0.152 (+0.012)	1.693 (−0.054)
	+LoRA+Adapter+MONA+MSM	0.386 (−0.006)	0.327 (−0.005)	0.158 (+0.006)	1.609 (**−0.084**)
	+LoRA+Adapter+MONA+MSM+NDVI_loss	0.365 (**−0.021**)	0.318 (−0.009)	0.159 (+0.001)	1.594 (−0.015)
	+LoRA+Adapter+MONA+MSM+NDVI_loss+NDWI_loss	**0.354** (**−0.011**)	**0.295** (**−0.023**)	**0.160** (+0.001)	**1.582** (−0.012)
阶段 2	RGB (基准)	0.338	0.301	0.160	1.592
	TASSE	**0.317** (**−0.021**)	**0.283** (**−0.018**)	**0.161** (**+0.001**)	**1.545** (**−0.047**)

间的物理相关性，确保了生成植被的光谱合理性。同时，设计的文本感知空间语义编码机制，建立了语义分割掩膜与生成特征间的精确映射，解决了复杂地理场景下的布局控制问题，提升了生成内容与文本/掩膜的语义对齐度。

3）数据增广的实用价值：在FLAIR、Five-Billion-Pixels及IRSAMap等数据集上的实验结果显示，本文方法在光谱保真度、感知质量及语义一致性等指标上均优于ControlNet与T2I-Adapter等主流方法。更重要的是，在FLAIR、Five-Billion-Pixels及IRSAMap数据集上，利用生成的多光谱数据辅助训练使下游开放词汇分割的mIoU提升了0.3%~1.0%。这一提升幅度相对有限，但在训练数据固定的条件下仍具有一定的参考价值，说明生成数据在结构与语义层面与真实数据具有一定的一致性，可作为缓解数据稀缺问题的辅助手段之一。

未来展望：本文方法的适用前提是目标光谱域与RGB域之间存在一定的特征相关性，近红外波段满足这一条件，因而适配效果较好。对于高光谱数据，波段数量大幅增加，各波段之间的物理关联更为复杂，直接套用当前框架可能面临潜变量维度扩展和物理约束设计的挑战；对于SAR图像，其成像机制与光学影像存在根本差异，现有的PEFT策略和NDVI约束均不直接适用，需要针对散射机理重新设计约束项。此外，本文实验均基于0.2~4米分辨率的航空或卫星影像，在更低分辨率或跨传感器场景下，生成质量和下游任务的提升幅度可能有所下降，这也是后续需要验证的方向。

综上所述，本文提出框架为多光谱遥感数据的低成本生成提供了一套行之有效的解决方案，在补充稀缺样本、提升解译模型泛化能力方面具有重要的应用潜力。

参考文献(References)

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR: 214-223 [DOI:10.48550/arXiv.1701.07875]

Brock A, Donahue J and Simonyan K. 2019. Large scale GAN training for high fidelity natural image synthesis//Proceedings of the International Conference on Learning Representations. New Orleans, USA: OpenReview [DOI:10.48550/arXiv.1809.11096]

Dhariwal P and Nichol A. 2021. Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems, 34: 8780-8794 [DOI:10.48550/arXiv.2105.05233]

Esser P, Rombach R and Ommer B. 2021. Taming transformers for high-resolution image synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12873-12883 [DOI: 10.1109/CVPR46437.2021.01268]

Garioud A, Peillet S, Bookjans E, Giordano S and Wattrelos B. 2022. FLAIR #1: Semantic segmentation and domain adaptation dataset. arXiv preprint arXiv: 2211.12979 [DOI: 10.48550/arXiv.2211.12979]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, 27:

2672-2680 [DOI:10.48550/arXiv.1406.2661]

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S and Lerchner A. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework//Proceedings of the International Conference on Learning Representations. Toulon, France: OpenReview [DOI:10.48550/arXiv.1606.05579]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840-6851 [DOI:10.48550/arXiv.2006.11239]

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M and Gelly S. 2019. Parameter-efficient transfer learning for NLP//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 2790-2799 [DOI:10.48550/arXiv.1902.00751]

Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L and Chen W. 2022. LoRA: Low-rank adaptation of large language models//Proceedings of the International Conference on Learning Representations. Virtual: OpenReview [DOI: 10.48550/arXiv. 2106.09685]

Isola P, Zhu J Y, Zhou T and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5967-5976 [DOI:10.1109/CVPR.2017.632]

Ji Y, Wang C, Chen J, et al. 2026. MovSeg: Efficient adaptation of vision-language models for multispectral open-vocabulary segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1-12 [DOI: 10.1109/JSTARS. 2026.3658442]

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4401-4410 [DOI:10.1109/CVPR.2019.00453]

Khanna S, Liu P, Zhou L, Meng C, Rombach R, Burke M, Lobell D B and Ermon S. 2024. DiffusionSat: A generative foundation model for satellite imagery//Proceedings of the International Conference on Learning Representations. Vienna, Austria: OpenReview [DOI: 10.48550/arXiv.2312.03606]

Kingma D P and Welling M. 2014. Auto-encoding variational Bayes//Proceedings of the International Conference on Learning Representations. Banff, Canada: OpenReview [DOI: 10.48550/arXiv. 1312.6114]

Liu C, Chen K, Zhao R, Zou Z and Shi Z. 2025. Text2Earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. IEEE Geoscience and Remote Sensing Magazine, 13(3): 238-259 [DOI: 10.1109/MGRS. 2025.3560455]

Liu L, Chen B, Chen H, Zou Z and Shi Z. 2023. Diverse hyperspectral remote sensing image synthesis with diffusion models. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-16 [DOI: 10.1109/TGRS.2023.3335975]

Liu Q, Zhou H, Xu Q, Liu X and Wang Y. 2021. PSGAN: A generative adversarial network for remote sensing image pan-sharpening. IEEE Transactions on Geoscience and Remote Sensing, 59(12): 10227-10242 [DOI:10.1109/TGRS.2020.3042974]

Meng Y, Deng L, Xi Z, 等, 2025. IRSAMap: Toward Large-Scale, High-Resolution Land Cover Map Vectorization [J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 63: 1-19). [DOI:10.1109/tgrs.2025.3600249]

MOU C, WANG X, XIE L, 等, 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models [C]//Proceedings of the AAAI conference on artificial intelligence: 卷 38. 4296-4304.

Pan H. 2021. Cloud removal for remote sensing imagery via spatial attention generative adversarial network. IEEE Transactions on Geoscience and Remote Sensing, 59(8): 6667-6681 [DOI: 10.1109/TGRS.2020.3029633]

Pan J, Lei S, Fu Y, et al. 2025. EarthSynth: Generating informative earth observation with diffusion models. arXiv preprint arXiv: 2505.12108 [DOI:10.48550/arXiv.2505.12108]

Pang L, Cao X, Tang D, et al. 2024. HSIGene: A foundation model for hyperspectral image generation. arXiv preprint arXiv: 2409.12470 [DOI:10.48550/arXiv.2409.12470]

Park T, Liu M Y, Wang T C and Zhu J Y. 2019. Semantic image synthesis with spatially-adaptive normalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2337-2346 [DOI:10.1109/CVPR.2019.00244]

Peebles W and Xie S. 2023. Scalable diffusion models with transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4195-4205 [DOI: 10.1109/ICCV51070.2023.00387]

Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, Penna J and Rombach R. 2023. SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv: 2307.01952 [DOI:10.48550/arXiv.2307.01952]

Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks//Proceedings of the International Conference on Learning Representations. San Juan, Puerto Rico: OpenReview [DOI: 10.48550/arXiv.1511.06434]

Razavi A, Van Den Oord A and Vinyals O. 2019. Generating diverse high-fidelity images with VQ-VAE-2. Advances in Neural Information Processing Systems, 32: 14866-14876 [DOI:10.48550/arXiv.1906.00446]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-

tern Recognition. New Orleans, USA: IEEE: 10684-10695 [DOI: 10.1109/CVPR52688.2022.01042]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: Convolutional networks for biomedical image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI:10.1007/978-3-319-24574-4_28]

Sastry S, Khanal S, Dhakal A and Aich A. 2024. GeoSynth: Contextually-aware high-resolution satellite image synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 460-470 [DOI: 10.1109/CVPR52733.2024.00050]

Sebaq A and ElHelw M. 2024. RSDiff: Remote sensing image generation from text using diffusion model. Neural Computing and Applications, 36 (36): 23103-23111 [DOI: 10.1007/s00521-024-10515-3]

Shen S, Pan B, Zhang Z, Chen Y and Zhang H. 2025. Hyperspectral image generation with unmixing guided diffusion model. arXiv preprint arXiv:2506.02601 [DOI:10.48550/arXiv.2506.02601]

Song J, Meng C and Ermon S. 2021. Denoising diffusion implicit models//Proceedings of the International Conference on Learning Representations. Virtual: OpenReview [DOI: 10.48550/arXiv. 2010.02502]

Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S and Poole B. 2021. Score-based generative modeling through stochastic differential equations//Proceedings of the International Conference on Learning Representations. Virtual: OpenReview [DOI: 10.48550/arXiv.2011.13456]

Tang D, Cao X, Hou X, Jiang Z and Dai Q. 2024. CRS-Diff: Controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-14 [DOI: 10.1109/TGRS.2024.3353493]

Tong X Y, Xia G S and Zhu X X. 2023. Enabling country-scale land cover mapping with meter-resolution satellite imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 196: 178-196 [DOI: 10.1016/j.isprsjprs.2022.12.011]

Van Den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning. Advances in Neural Information Processing Systems, 30: 6306-6315 [DOI:10.48550/arXiv.1711.00937]

Wang T C, Liu M Y, Zhu J Y, Tao A, Kautz J and Catanzaro B. 2018. High-resolution image synthesis and semantic manipulation with conditional GANs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8798-8807 [DOI:10.1109/CVPR.2018.00917]

Xu Y, Liu H, Yang R, Yu S and Li J. 2025. Remote sensing image semantic segmentation sample generation using a decoupled latent diffusion framework. Remote Sensing, 17(13): 2143 [DOI: 10.3390/rs17132143]

Yin D, Hu L, Li B and Xu C. 2024. Mona: Modulation-based adaptation for remote sensing image generation. arXiv preprint arXiv: 2408.08345 [DOI:10.48550/arXiv.2408.08345]

Zhang L, Rao A and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3836-3847 [DOI:10.1109/ICCV51070.2023.00355]

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 586-595 [DOI:10.1109/CVPR.2018.00068]

马愈卓,张永飞,贾伟,刘家瑛,甘甜,杨文瀚,卓君宝,刘武,马惠敏. 2025. 面向计算机视觉的数据生成与应用研究进展. 中国图象图形学报,1-81) [DOI: 10.11834/jig.250085]

Ma Yuzhuo, Zhang Yongfei, Jia Wei, Liu Jiaying, Gan Tian, Yang Wenhan, Zhuo Junbao, Liu Wu, Ma Huimin. 2025. Recent advances in data generation and its applications in computer vision. Journal of Image and Graphics, 30(6):1872-1952 DOI: 10.11834/jig.250085.

陶超,郭鑫,胡柯彦,沈羽翔,王昊. 2025. 以语言为媒介的遥感图像跨时空领域自适应 语义分割. 中国图象图形学报,30(9):3153-3170) [DOI: 10.11834/jig.240640]

Tao Chao, Guo Xin, Hu Keyan, Shen Yuxiang, Wang Hao. 2025. Language-guided cross-spatiotemporal domain adaptation for remote sensing image semantic segmentation. Journal of Image and Graphics, 30(9):3153-3170 DOI: 10.11834/jig.240640.

梅少辉, 张博威, 马明阳, 贾森. 近红外高光谱图像数据预测技术[J]. 中国图象图形学报, 2021, 26(8): 1786-1795). DOI: 10.11834/jig.210184

Shaohui Mei, Bowei Zhang, Mingyang Ma, Sen Jia. Predicting near-infrared hyperspectral images from visible hyperspectral images[J]. Journal of Image and Graphics, 2021, 26(8): 1786-1795. DOI: 10.11834/jig.210184.

王耀领, 王宏琦, 许滔. 2021. CGAN样本生成的遥感图像飞机识别. 中国图象图形学报, 26(3): 663-673) [DOI: 10.11834/jig.200001]

Yaoling Wang, Hongqi Wang, Tao Xu. Aircraft recognition of remote sensing image based on sample generated by CGAN[J]. Journal of Image and Graphics, 2021, 26(3): 663-673. DOI: 10.11834/jig.200001.

谭明奎, 许守恺, 张书海, 等. 2021. 深度对抗视觉生成综述. 中国图象图形学报, 26(12): 2751-2766) [DOI: 10.11834/jig.210252]

Mingkui Tan, Shoukai Xu, Shuhai Zhang, Qi Chen. A review on deep adversarial visual generation[J]. Journal of Image and Graphics, 2021, 26(12): 2751-2766. DOI: 10.11834/jig.210252.

刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报, 29(06):1535-1554) DOI: 10.

11834/jig.240003.

Liu Anan, Su Yuting, Wang Lanjun, Li Bin, Qian Zhenxing, Zhang Weiming, Zhou Linna, Zhang Xinpeng, Zhang Yongdong, Huang Jiwu, Yu Nenghai. 2024. Review on the progress of the AIGC visual content generation and traceability. Journal of Image and Graphics, 29(06):1535-1554 DOI: 10.11834/jig.240003.

作者简介

纪璎芮,1996年生,女,博士研究生,研究方向为遥感图像生成、语义分割。E-mail: jiyingrui22@mails.ucas.ac.cn

王晨昊,男,博士研究生,研究方向为遥感图像生成、语义分割。E-mail: wangchenhao22@mails.ucas.ac.cn

岳安志,男,研究员,要研究方向为数据工程和土地利用/土地覆盖分类。E-mail:yueaz@aircas.ac.cn

陈静波,男,研究员,主要研究方向为遥感图像智能处理、遥感大数据分析。E-mail: chenjb@aircas.ac.cn

席智浩,男,助理研究员,主要研究方向为计算机视觉、领域自适应、遥感图像解译。E-mail: xizh@aircas.ac.cn