**研究背景**

这篇文章介绍了DiffusionSat，一个为卫星影像设计的生成基础模型。文章指出，扩散模型（如Stable Diffusion）在图像、语音和视频生成方面取得了显著成果，但尚未针对遥感数据进行优化。遥感数据在环境监测和作物产量预测等重要应用中广泛使用，具有多光谱、时间不规则采样等特点，与自然图像有显著差异。为了填补这一空白，作者提出了DiffusionSat，结合卫星影像的元数据（如地理位置、时间戳等），实现了多种生成任务，如时间生成、超分辨率和图像修复。

**预处理和微调数据集**

**预处理数据集1.Function Map of the World (fMoW):** 包含全球高分辨率（GSD 0.3m-1.5m）MAXAR卫星图像，每张图像属于62个类别之一。用于训练单图像生成任务，模型接收图像的文本描述和元数据作为输入，生成高分辨率图像。

**2.Satlas:** 包含NAIP和Sentinel-2卫星图像。本文使用Satlas-small中的NAIP图像（NAIP图像的分辨率约为1米），并使用与fMoW相同的元数据。用于增强模型在多任务学习中的表现，包括图像分割和目标检测。

**3.SpaceNet:** 包含用于目标检测、语义分割和道路网络映射的卫星图像数据集。本文使用SpaceNet v1、v2和v5子集，元数据与前述数据集一致。用于训练和验证模型在不同任务（如目标检测、语义分割等）中的表现。

**数据预处理**

**1.裁剪和缩放：**将图像裁剪为512x512像素，以便与模型的输入尺寸匹配。**2.标准化：**将图像像素值归一化，以适应模型的输入要求。**3.元数据处理：**将元数据（如经纬度、GSD、云覆盖率、年份、月份和日期）进行编码，使其可以作为模型的条件输入。

**模型架构调整1.文本和元数据嵌入：**在DiffusionSat中，使用CLIP模型对文本进行编码，并使用MLP对元数据进行编码。元数据编码采用与扩散时间步长相同的正弦嵌入方法。**2.扩散过程：**模型在VAE的潜在空间中进行扩散，生成潜在表示。通过在潜在表示中添加高斯噪声，生成不同时间步长的潜在表示。**3.去噪网络：**使用去噪UNet来预测添加的噪声，最终通过VAE解码器生成高分辨率图像。

**微调过程1.初始化：**使用Stable Diffusion 2.1的预训练权重初始化编码器E、解码器D、CLIP文本编码器Tθ和去噪UNet ϵθ。**2.训练：**仅更新去噪UNet ϵθ和元数据、时间步长嵌入fθj，以加快收敛速度。**3.随机归零：**训练过程中，随机将元数据向量m归零，以允许模型在元数据不可用或不准确时生成图像。

**微调数据集**文章通过使用已有的Stable Diffusion 2.1模型权重，并在其基础上进行微调，包括UNet去噪网络、元数据和时间步长嵌入。训练过程中随机将元数据向量归零，以便模型在元数据不可用或不准确时仍能生成图像。

**模型架构**

DiffusionSat采用了Stable Diffusion的变分自编码器（VAE）架构，并在此基础上进行了改进：**单图像生成：**模型接收文本和元数据作为输入，利用VAE对图像进行编码，通过扩散过程加入噪声，然后通过去噪网络生成高质量图像。**控制信号条件生成：**针对多光谱输入的超分辨率、时间生成和时间插值任务，模型扩展了3D版本的ControlNet，能够接受图像序列并结合其元数据进行生成任务。**超分辨率和时间插值：**通过控制信号和目标元数据，模型可以生成高分辨率图像或预测未来的图像。

**详细版：**

**1. 整体架构概览**DiffusionSat模型主要包含以下几个部分：变分自编码器（VAE）：用于对输入图像进行编码和解码。扩散过程：在潜在空间中对编码后的图像进行扩散和去噪。条件嵌入：使用文本和元数据作为条件信息来指导图像生成。去噪UNet：用于去噪扩散后的潜在表示，生成最终图像。**2. 变分自编码器（VAE）**编码器（Encoder E）：将输入图像编码为潜在表示。通过一个卷积神经网络（CNN）将图像x ∈ RC×H×W转换为潜在表示z ∈ RC′×H′×W′。解码器（Decoder D）：将去噪后的潜在表示解码回原始图像空间。解码器也是一个CNN，将潜在表示z转换回高分辨率图像x̂。**3. 扩散过程**噪声添加：在潜在空间中对编码后的图像加入高斯噪声，生成不同时间步长的噪声图像zt = αtz + σtϵ，其中ϵ是高斯噪声，αt和σt是与扩散时间步长t相关的参数。去噪过程：通过去噪UNet预测并去除噪声，生成去噪后的潜在表示ẑ。**4. 条件嵌入**文本嵌入（Text Embedding）：使用CLIP模型对文本描述进行编码，将文本τ转换为文本嵌入τ′。元数据嵌入（Metadata Embedding）：使用多层感知机（MLP）对元数据进行编码。每个元数据项（如经纬度、时间戳等）通过正弦嵌入进行编码，再通过MLP转换为嵌入向量。时间步长嵌入（Timestep Embedding）：使用正弦嵌入对扩散时间步长进行编码，生成时间步长嵌入向量。**5. 去噪UNet**输入层：接收噪声图像zt、文本嵌入τ′和元数据嵌入c作为输入。卷积层：通过一系列卷积层对输入进行处理，提取特征。跨注意力层（Cross-Attention Layers）：在每一层中通过跨注意力机制结合文本和元数据嵌入，指导图像生成过程。输出层：生成去噪后的潜在表示ẑ。**6. 训练过程**预训练权重初始化：使用Stable Diffusion 2.1的预训练权重初始化编码器、解码器、CLIP文本编码器和去噪UNet。微调：仅更新去噪UNet和元数据、时间步长嵌入，以加快收敛速度和提高生成效果。随机归零：训练过程中，随机将元数据向量归零，以允许模型在元数据不可用或不准确时生成图像。**7. 条件生成任务**超分辨率：通过低分辨率图像和元数据生成高分辨率图像。时间生成：通过一系列时间上不连续的图像预测某一目标时间的图像。图像修复：在图像中存在损坏或遮挡的情况下进行修复，生成完整图像。

**如何评估**

**文章通过以下评估指标评估模型的生成质量：单图像生成：**使用FID（Frechet Inception Distance）、IS（Inception Score）和CLIP-score评估图像的视觉质量。

**FID（Frechet Inception Distance）**：衡量生成图像与真实图像在Inception网络特征空间中的分布差异。FID越低，生成图像质量越高。

**IS（Inception Score）**：通过Inception网络对生成图像进行分类，评估生成图像的多样性和质量。IS越高，表示生成图像的质量和多样性越好。

**CLIP-score**：使用CLIP模型计算生成图像与对应文本描述的相似性，评估生成图像与文本描述的一致性。CLIP-score越高，表示生成图像与文本描述的一致性越好。

**条件生成：**在有参考真实图像的情况下，使用SSIM（结构相似性指数）、PSNR（峰值信噪比）、LPIPS（感知图像补丁相似性）和MSE（均方误差）评估图像质量。

**SSIM（结构相似性指数）**：衡量生成图像与真实图像在结构上的相似性。SSIM越高，表示生成图像的结构信息与真实图像越接近。

**PSNR（峰值信噪比）**：衡量生成图像与真实图像之间的误差。PSNR越高，表示生成图像的质量越好。

**LPIPS（感知图像补丁相似性）**：使用VGG网络特征计算生成图像与真实图像的感知相似性。LPIPS越低，表示生成图像的感知质量越高。

**MSE（均方误差）**：衡量生成图像与真实图像之间的平均误差。MSE越低，表示生成图像的质量越好。

实验结果显示，DiffusionSat在fMoW数据集上的单图像生成和超分辨率任务中均优于其他方法，证明了其在生成高质量卫星图像方面的有效性。