/TemplateVersion (2026.1)

用于遥感图像生成的目标保真扩散

Ziqi Ye[1,2*], Shuran Ma[3*], Jie Yang[2], Xiaoyi Yang[1], Ziyang Gong[4], Xue Yang[4†‡], Haipeng Wang[1†]

[1] Fudan University [2]Shanghai Innovation Institute [3]Xidian University [4]Shanghai Jiao Tong University
yezq24@m.fudan.edu.cn, shrma@stu.xidian.edu.cn https://github.com/conquer997/OF-Diff

Abstract

高精度可控的遥感图像生成既有意义又具有挑战性。现有的扩散模型由于无法充分捕捉形态细节，通常产生低保真度的图像，这可能影响目标检测模型的鲁棒性和可靠性。为了提高遥感中生成目标的准确性和保真度，本文提出了对象保真度扩散（OF-Diff），有效地提高了生成目标的保真度。具体来说，我们首次基于布局在遥感中为扩散模型提取目标的先验形状。然后，我们引入一个具有扩散一致性损失的双分支扩散模型，该模型可以在采样阶段不提供真实图像的情况下生成高保真的遥感图像。此外，我们介绍了 DDPO 来微调扩散过程，使生成的遥感图像更加多样化且语义一致。综合实验表明，OF-Diff 在遥感领域的关键质量指标上优于最先进的方法。值得注意的是，几类多态和小目标的性能显示出显著提升。例如，飞机、船舶和车辆的 mAP 分别增加了 8.3 %、7.7 % 和 4.0 %。

介绍

合成高保真、空间可控的遥感 (RS) 图像是克服数据限制的关键前沿，这些限制阻碍了诸如物体检测 (Yang et al. 2021; Zhang et al. 2020; Yang et al. 2019) 等后续感知任务。然而，当前的 RS 生成方法通常依赖于模糊的文本提示 (Khanna et al. 2023; Sebaq and ElHelw 2024) 或辅助条件，如语义地图 (Sebaq and ElHelw 2024; Tang et al. 2024; Gong et al. 2024; Hu et al. 2025; Jia et al. 2025) 。虽然在视觉上似乎合理，但这种指导基本上与实例级地面真相脱节，未能提供有效数据增强所需的精确控制。

相比之下，基于对象边界框的布局到图像（L2I）生成在精确空间控制方面提供了一种更为稳健的解决方案。尽管这种方法在自然图像领域得到了广泛的发展，例如将其视为一种多模态融合任务的 LayoutDiffusion (Zheng et al. 2023) 、通过额外控制实现开放世界生成的 GLIGEN (Li et al. 2023) ，以及通过解耦对象来提高可控性的 ODGen (Zhu et al. 2024) ，其在遥感 (RS) 中的直接应用并非易事。这是由于 RS 图像的独特挑战，例如广阔的背景、任意的目标方向和密集的场景。

在 RS 布局到图像生成中，现有的方法如 AeroGen (Tang et al. 2025) 和 CC-Diff (Zhang et al. 2024) 采取不同的方法。AeroGen 是一种粗略的布局条件模型，空间和形状控制能力有限。相比之下，像 CC-Diff 这样的实例级方法通过引用真实实例实现更高的可控性和逼真度，但这导致对真实数据的质量和数量的严重依赖，限制了泛化和灵活性。我们总结了一些常见的失败模式（见图 1），包括控制泄漏、结构扭曲、密集生成崩溃和特征级不匹配。

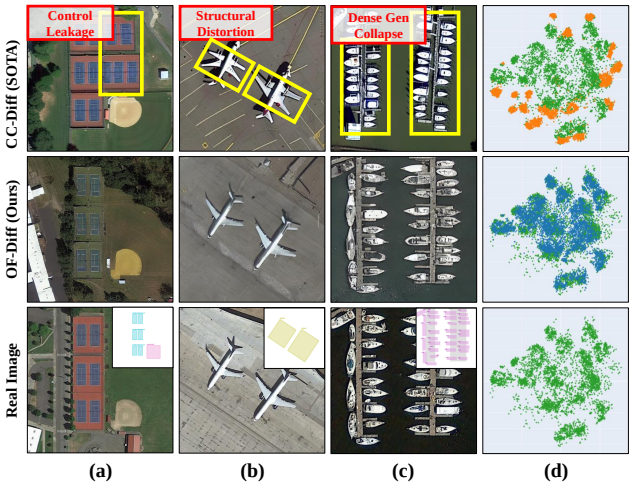


Figure 1: 在最先进的 (SOTA) 方法 (CC-Diff) 中存在四种关键故障模式：(a) 控制泄露：内容超出指定的布局；(b) 结构变形：导致畸形的对象形态；(c) 密集生成崩溃：在人群场景中失去控制；以及 (d) 特征级不匹配：从实际数据中发生的分布漂移，用 t-SNE 可视化。我们的 OF-Diff(第二行) 有效地解决了这些问题。

这些缺陷显著降低了物体检测任务的性能，限制了它们在智能遥感（RS）解释中的实际应用。在本文中，我们引入了对象保真扩散模型（OF-Diff）。其旨在提高 RS 图像中对象生成的形状保真度和布局一致性。如图 2 所示，现有的 L2I 方法主要分为两类。第一类是布局条件的基准，如图 2 (a) 所示，比如 Aerogen 和 LayoutDiffusion。第二类是具有基于实例模块的方法，如图 2 (b) 所示，比如 CC-Diff。然而，这些方法在采样阶段需要真实的实例和图像作为参考，以生成高质量的合成图像。相比之下，OF-Diff 只需要提取前景的形状就能生成高保真的 RS 图像，如图 2 (c) 所示。此外，它通过 DDPO 微调扩散，有效地提升了生成图像对下

*Equal contribution.
†Corresponding author. ‡Project lead.

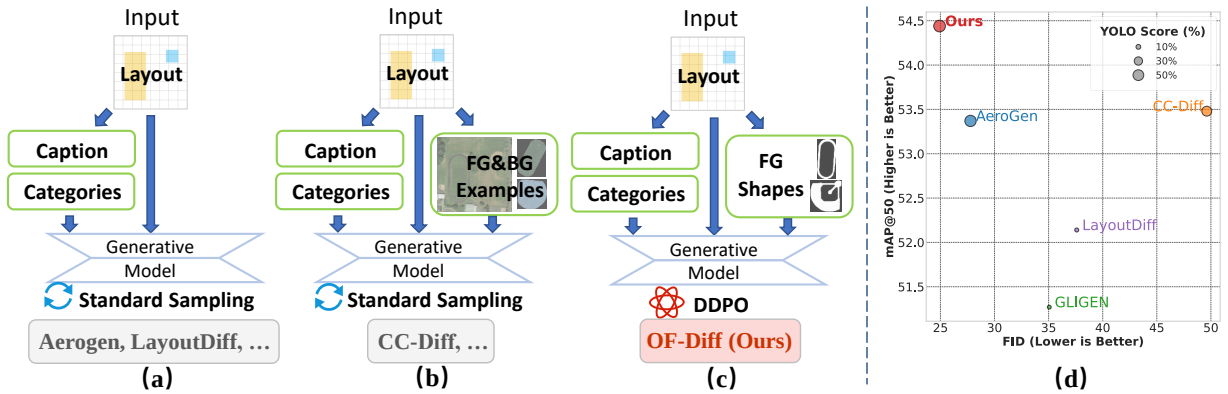


Figure 2: 与主流 L2I 方法相比，OF-Diff. FG/BG 代表前景/背景。(a) 布局条件基线。(b) 添加基于实例的模块，受限于来自真实数据的图块质量/数量。(c) OF-Diff 通过形状提取和 DDPO 增强真实度，而不依赖于图块。(d) 结果表明其优越性。

游任务的性能。图 2 (d) 的结果展示了 OF-Diff 相较于其他方法的优越性。我们的贡献总结如下：

- 我们引入了 OF-Diff，这是一种具有先验形状提取功能的双分支可控扩散模型，它在提高生成保真度的同时减少了对真实图像的依赖，提高了实际应用性。
- 我们提出了一种可控生成流程，使用 DDPO 微调扩散模型以处理遥感图像，进一步提高保真度和多样性。
- 大量实验表明，OF-Diff 能够生成高保真、布局和形状一致的密集物体图像，并作为物体检测任务的有效增强工具。

相关工作

图像生成的进展

扩散模型 (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Kingma et al. 2021) 在图像合成任务中越来越多地取代了生成对抗网络 (GANs) (Goodfellow et al. 2014; Karras et al. 2021) 和变分自编码器 (VAEs) (Kingma, Welling et al. 2013; Rezende, Mohamed, and Wierstra 2014)，因为它们具有训练稳定性和优质的输出质量。最近在高效采样器方面的进展，如 DDIM (Song, Meng, and Ermon 2021)，Euler (Karras et al. 2022) 和 DPM-Solver (Lu et al. 2022)，进一步提高了实用性。潜在扩散模型 (LDMs) (Rombach et al. 2022a) 在低维潜在空间中运行，显著降低了计算成本，同时保持视觉逼真度。像 DALL · E2 (Ramesh et al. 2022) 和 Imagen (Saharia et al. 2022) 这样的模型的成功展示了这种范式如何支持在庞大的互联网规模数据集上进行训练。因此，基于扩散的方法现在为高质量图像生成提供了坚实的基础。

布局到图像生成

可控图像合成主要包括文本到图像 (T2I) 和布局到图像 (L2I) 生成。虽然 T2I 模型 (Nichol et al. 2022; Ramesh et al. 2022) 通过文本提示实现语义对齐，L2I 方法提供更好的空间控制。最近的工作通过布局作为模态设计 (Zheng et al. 2023)、门控注意力 (Li et al. 2023) 和实例生成 (Wang et al. 2024; Zhou et al. 2024) 增强布局条件。然而，这些方法仅依赖于粗略的布局输入（例如边界框），缺乏对于合成形态复杂物体至关重要的精细形状信息。合成高保真训练数据对于推进遥感 (RS) 目标检测至关重要，该领域对于众多应用至关重要，但通常受到广泛注释数据集稀缺的阻碍。尽管其必要性，RS 图像的大多数生成模型，如 DiffusionSat 和 RSDiff，仍然依赖于粗略的语义指导。而其他方法利用多样化的控制信号，如 OpenStreetMaps，但它们通常未针对目标检测中心的边界框格式进行优化。这自然而然地推动了包括 AeroGen 和 CC-Diff 在内的 L2I 方法，这些方法通过布局掩码注意力和 FG/BG 双重重采样器提高了空间精度和上下文一致性。但它们仍然存在可控性有限和对真实数据高度依赖的问题。

方法

初步

扩散模型 (Song, Meng, and Ermon 2021) 的目标是通过从初始标准正态分布中采样得到的噪声表示中迭代重建数据，从而捕捉基础数据分布 $p(x)$。

去噪扩散概率模型 (Ho, Jain, and Abbeel 2020) 将模型参数化为函数 $\epsilon_\theta(x_t, t)$，以在任何时间步 t 预测样本 x_t 的噪声成分。训练目标是最小化实际噪声 ϵ 与预测噪声 $\epsilon_\theta(x_t, t)$ 之间的均方误差 (MSE) 损失：

$$\mathcal{L} = \mathbb{E}_{x_t, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[\left\| \epsilon_\theta \left(x_t, t \right) - \epsilon \right\|^2 \right]. \quad (1)$$

稳定扩散 (Rombach et al. 2022a; Qiu et al. 2025) 利用预训练的 VQ-VAE (Van Den Oord, Vinyals et al. 2017) 将图像编码到低维的潜在空间，在潜在表示 z_0 上进行训练。在条件生成的背景下，给定文本提示 c_t 和任务特定条件 c_f，在时间步长 t 时的扩散训练损失可以表示为：

$$\mathcal{L} = \mathbb{E}_{z_t, t, c_t, c_f, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[\left\| \epsilon - \epsilon_\theta \left(z_t, t, c_t, c_f \right) \right\|^2 \right]. \quad (2)$$

其中，$\mathcal{L}$ 代表完整扩散模型的整体学习目标。这个目标函数在与 ControlNet (Zhang, Rao, and Agrawala 2023) 结合微调扩散模型时显式应用。

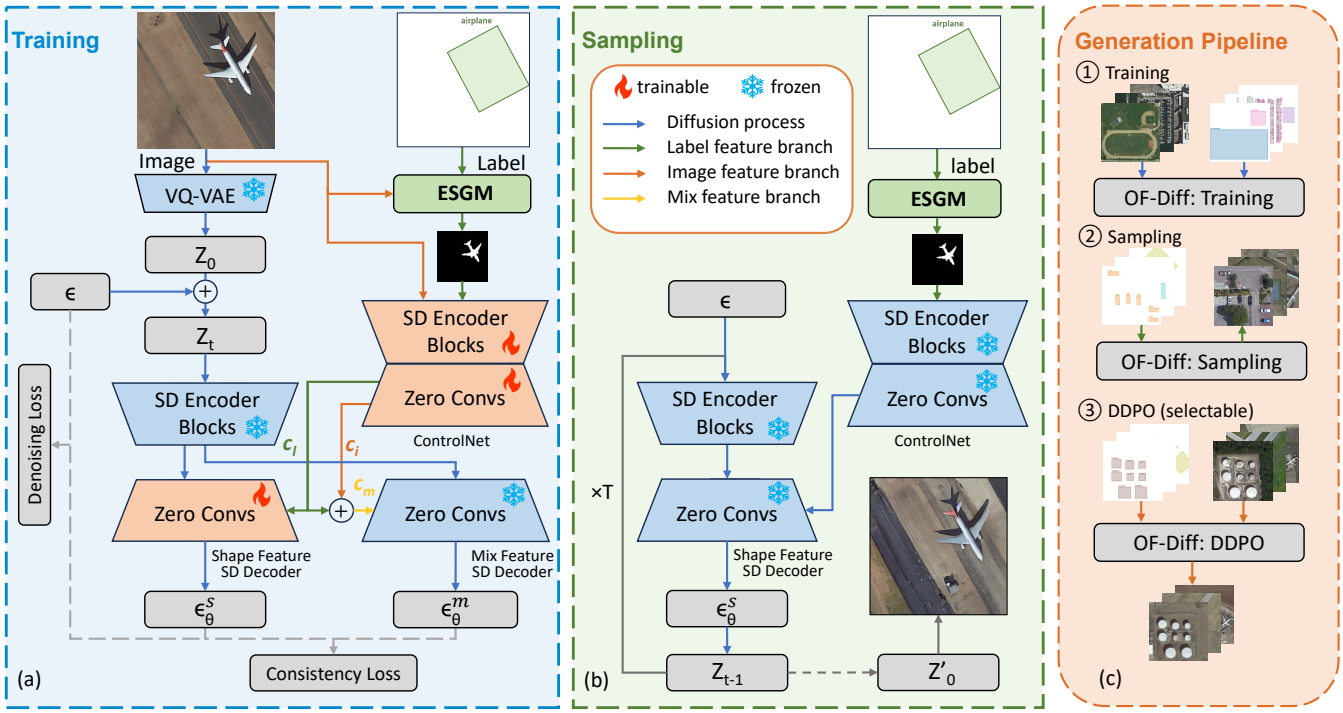


Figure 3: OF-Diff 的整体架构。(a) 在训练阶段，通过 ESGM 和图像获得的对象形状特征由 Control 进行特征提取，参数更新到双分支稳定扩散。在潜在空间中的噪声图像 Z_t 通过形状特征稳定扩散和混合特征稳定扩散处理，生成噪声预测 ε_θ^s 和 ε_θ^m 。扩散一致性损失通过 ε_θ^s 和 ε_θ^m 计算，去噪损失由 ε_θ^s 和 ε 计算。(b) 在采样阶段，仅使用标签和形状特征稳定扩散生成合成图像。(c) OF-Diff 的流程。如果期望生成的数据更符合特定数据集的分布，DDPO 是可选的。

OF-Diff 的架构

如图 XAMATHX_1_0 (a) 所示，在 OF-Diff 的训练阶段，扩散训练需要真实图像及其对应的标签。首先，它们被输入到增强形状生成模块 (ESGM) 进行形状提取，从而生成对象掩码。随后，真实图像及其对象掩码被输入到 ControlNet 进行特征提取，得到图像特征 XAMATHX_0 和标签特征 XAMATHX_1。第二，输入图像首先通过预训练的 VQ-VAE 压缩成潜在空间特征 XAMATHX_2。然后，它与随机高斯噪声 XAMATHX_3 连接形成 XAMATHX_4。在注入稳定扩散编码器块之后，特征 XAMATHX_5 被输入到双分支解码器架构。该架构的两个组成部分分别被指定为形状特征稳定扩散解码器和混合特征稳定扩散解码器。第三，对 XAMATHX_6 和停止梯度 XAMATHX_7 进行加权求和，得到 XAMATHX_8。XAMATHX_9 定义为：

$$c_{\mathrm{m}} = \frac{n}{N} \cdot c_i + \mathrm{sg}\left[c_l\right], \tag{3}$$

，其中 n 表示当前迭代次数，N 表示总迭代次数。为了使混合特征稳定扩散的预测能够作为锚点，引导形状稳定扩散的收敛轨迹，从而提高图像的形态保真度，我们在计算 c_m 时采用对 c_l 的显著梯度停止策略 (Chen and He 2021) 。形状特征稳定扩散和混合特征稳定扩散的损失 $\mathcal{L}_s$ 和 $\mathcal{L}_m$ 被定义为：

$$\mathcal{L}_s = \mathbb{E}\left[\left\|\epsilon_\theta^s - \epsilon\right\|^2\right], \epsilon_\theta^s = \epsilon_\theta\left(z_t, t, c_t, c_s\right), \tag{4}$$

$$\mathcal{L}_m = \mathbb{E}\left[\left\|\epsilon_\theta^m - \epsilon\right\|^2\right], \epsilon_\theta^m = \epsilon_\theta\left(z_t, t, c_t, c_m\right), \tag{5}$$

为了帮助形状特征的稳定扩散摆脱低保真度的局部极小值，并迁移到参数空间中高保真度的区域，我们引入了一种扩散一致性损失：

$$\mathcal{L}_c = \mathbb{E}\left[\left\|\epsilon_\theta^m - \mathrm{sg}\left[\epsilon_{\theta'}^{\mathrm{mix}}\right]\right\|^2\right], \tag{6}$$

由于额外的图像先验控制，混合特征稳定扩散的预测噪声 ϵ_θ^m 比形状特征稳定扩散预测的 ϵ_θ^s 更为准确。然而，在采样过程中，我们并未使用包含真实图像的输入。因此，我们使用停止梯度操作，使得更为准确的 ϵ_θ^m 可以作为锚点，引导被遮掩扩散的收敛轨迹朝向参数空间中的高保真局部最小值。最终，用于训练 OF-Diff 的损失定义为：

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_c, \tag{7}$$

在采样阶段，如图 3 (b) 所示，仅使用冻结的 ControlNet 和形状特征稳定扩散与任意标签先验控制来合成 RS 图像。

在自然图像中，透视和比例的变化使得大多数对象没有唯一的几何模型。相反，遥感对象显示出准不变的形

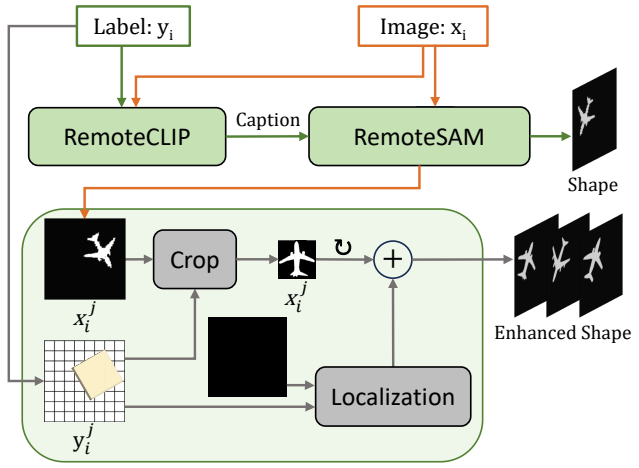


Figure 4: 增强形状生成模块（ESGM）的架构。

状。例如，球场是矩形的，烟囱和油罐是圆形的，飞机是双边对称的，具有明显的机头和机尾。这种形状的一致性使得可以使用掩码在遥感图像合成中实现强可控性。为了更好地利用类别标签进行对象形状提取，我们引入了增强形状生成模块（ESGM，参见图 4 ）。在训练阶段，ESGM 使用配对图像和标签生成精确的对象掩码。在采样时，它利用学习到的形状先验合成具有多样化对象形状的掩码。

对于给定的图像 x_i 及其对应于类别 j ($j \in [1, N]$) 的边界框 y_i^j ，我们首先利用 RemoteCLIP (Liu et al. 2024) 来生成边界框内对象的文本描述。利用这个描述和原始图像 x_i ，RemoteSAM (Yao et al. 2025) 随后生成相应的形状掩码 x_i^j 。

在形状增强阶段，我们使用其边界框 y_i^j 裁剪 x_i^j ，以提取仅包含目标对象的形状遮罩 x_i^j 。随后，我们对 x_i^j 应用随机旋转，并使用原始边界框坐标 y_i^j 将旋转后的遮罩重新定位到空白画布上，从而生成形状增强的遮罩。在训练阶段，ESGM 可以直接获得真实图像的形状。在采样时，ESGM 从训练期间或以后收集的遮罩池中选择增强形状，这是常见的数据高效实践，其中分割遮罩被重复使用或合成，对泛化影响最小。在我们的设置中，我们使用训练期间生成的遮罩。

为了增强微调模型生成的数据分布的多样性，并更好地与真实图像的分布保持一致，消噪扩散策略优化（DDPO）应用于 OF-Diff 的后期训练中。DDPO 将扩散模型的消噪过程视为一个多步马尔可夫决策过程（MDP）。映射关系定义如下：

$$\pi(a_t \mid s_t) \triangleq p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, c) \tag{8}$$

$$P(s_{t+1} \mid s_t, a_t) \triangleq (\delta_c, \delta_{t-1}, \delta_{\boldsymbol{x}_{t-1}}) \tag{9}$$

$$\rho_0(s_0) \triangleq \big(p(c), \delta_T, \mathcal{N}(0, I)\big) \tag{10}$$

$$R(s_t, a_t) \triangleq \begin{cases} r(\boldsymbol{x}_0, c), & \text{if } t = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

其中 δ_z 表示狄拉克 分布，其概率密度在除 z 以外的任何地方为零。符号 s_t 和 a_t 分别表示时间 t 的状态和动作。具体来说，s_t 被定义为由条件 c 、时间步长 t 以及该时间点的噪声图像 $\boldsymbol{x}_t$ 组成的元组，而 a_t 被定义为前一个时间步长的噪声图像 $\boldsymbol{x}_{t-1}$ 。策略用 $\pi(a_t \mid s_t)$ 表示，转移核用 $P(s_{t+1} \mid s_t, a_t)$ 表示，初始状态分布用 $\rho_0(s_0)$ 表示，奖励函数用 $R(s_t, a_t)$ 表示。为了使策略 $\pi(a_t \mid s_t)$ 最优化以最大化累积奖励 $\mathbb{E}_{\tau \sim p(,\cdot|\pi)}!\Big[\sum_{t=0}^{T} R(\boldsymbol{s}_t, \boldsymbol{a}_t)\Big]$ ，梯度 $\hat{g}$ 被计算如下：

$$\begin{aligned} \hat{g} = &\mathbb{E}\Bigg[\sum_{t=0}^{T} \frac{p_\theta(\boldsymbol{x}_{t-1} \mid c,\ t,\ \boldsymbol{x}_t)}{p_{\theta'}(\boldsymbol{x}_{t-1} \mid c,\ t,\ \boldsymbol{x}_t)} \\ &\cdot r(\boldsymbol{x}_0, c) \cdot \nabla_\theta \log p_\theta(\boldsymbol{x}_{t-1} \mid c,\ t,\ \boldsymbol{x}_t)\Bigg] \end{aligned} \tag{12}$$

$$r(\boldsymbol{x}_0, c) = \big(KNN(\boldsymbol{x}_0, \boldsymbol{x}_0) - \omega KL(\boldsymbol{x}_0, \boldsymbol{x}_0')\big) \tag{13}$$

分别引入了基于 K-最近邻（KNN）和 KL 散度的奖励函数，以优化生成数据的多样性以及生成数据和真实数据之间的分布一致性。ω 是权重参数，$\boldsymbol{x}_0'$ 是数据集中的真实图像。

数据集。

实施细节。我们基于 Stable Diffusion 1.5 (Rombach et al. 2022b) 预训练模型，在每个数据集（DIOR/DOTA）上单独训练 OF-Diff。仅对 ControlNet 和形状特征 SD 解码器进行微调，而所有其他模块保持冻结状态。训练使用 AdamW 优化器，在配备 4 个 H100 GPU 的节点上进行，学习率为 1e-5。全局批量大小设置为 64，训练运行 100 个周期。

基准方法。我们将我们的方法与遥感领域（AeroGen (Tang et al. 2025) ，CC-Diff (Zhang et al. 2024) ）和自然图像（LayoutDiffusion (Zheng et al. 2023) , GLIGEN (Li et al. 2023) ）的最新 L2I 生成模型进行比较。为了公平比较，所有模型都使用我们的数据集设置重新训练，并分别遵循其官方的训练细节。

评估指标。为了更全面地评估 OF-Diff 的效果，我们采用了涵盖 4 个不同评估方面的 13 项指标。

- 生成保真度。我们使用 FID (Heusel et al. 2017) 和 KID (Bińkowski et al. 2018) 来评估感知质量，并使用 CMMD (Jayasumana et al. 2024) 来测量生成图像与真实图像之间的 CLIP 特征距离，以评估布局对齐。
- 布局一致性。我们使用预训练的分类器报告 CAS (Ravuri and Vinyals 2019) ，以评估对象的可识别性，并通过应用训练过的 Oriented R-CNN (Xie et al. 2021) (与 Swin Transformer 主干 (Liu et al. 2021) , MMRotate) 来生成图像以实现实例级别的一致性。
- 形状保真度。为了评估生成实例的几何质量，我们与真实形状进行成对比较。每对实例被裁剪，调整为 64 × 64 大小，并转换为边缘图。我们计算五个指标：交并比 (IoU)、Dice 系数、切割距离 (CD)、Hausdorff 距离 (HD) 和结构相似性指数 (SSIM) (Wang et al. 2004) 。
- 下游效用。我们在混合的真实和生成图像上训练检测器，并使用 Oriented R-CNN (Swin 骨干网络) 在 8 个 × NVIDIA 4090 GPU 上以 24 批量大小报告真实测试数据上的 mAP $_{50}$ 、mAP $_{75}$ 及总体 mAP。

Figure 5: DIOR 和 DOTA 的定性结果。与其他方法相比，OF-Diff 更具现实性和保真度。

定性结果

比较结果。图 5 将 OF-Diff 的生成结果与其他方法进行了对比。OF-Diff 不仅生成了更逼真的图像，还具有最佳的可控性。例如，在前两个案例中，OF-Diff 成功控制了生成对象的数量和布局信息。第三和第四个案例展示了 OF-Diff 在生成小目标时的准确性，这是其他算法未能准确做到的。最后一个案例展示了 OF-Diff 在生成具有复杂形状的物体（如飞机）方面相对其他算法的优越性。

多样性结果。图 6 显示 OF-Diff 生成的图像持续呈现出合理的纹理和逼真的物体形状。例如，在不同方向渲染的飞机与其周围环境保持连贯的语义关系。即使在小物体场景中（其中一些是来自 DOTA 数据集的灰度遥感图像），OF-Diff 仍能生成视觉上真实且几何上准确的结果。

量化结果

生成保真度与一致性。我们将 OF-Diff 与遥感领域的最新生成方法进行了比较，包括 layoutDiffusion (Zheng et al. 2023) 、GLIGEN (Li et al. 2023) 、AeroGen (Tang et al. 2025) 和 CC-Diff (Zhang et al. 2024) 。这些方法

Figure 6: 同一 OF-Diff 模型产生的不同结果的多样性。

的性能如表 ?? 所示。OF-Diff 在生成保真度指标（FID、KID、CMMD）和布局一致性指标方面几乎达到了最佳性能，特别是在 DOTA 数据集上。

目标检测的可训练性。根据 (Chen et al. 2023) 中的数据增强协议，我们使用 OF-Diff 增加一倍的训练样本，并使用扩展的数据集评估检测结果。如表 ?? 所示，OF-

Method	Dataset	IoU ↑	Dice ↑	CD ↓	HD ↓	SSIM ↑
LayoutDiff	DIOR	0.0497	0.0908	12.037	25.962	0.1667
GLIGEN		0.055	0.1002	12.257	25.850	0.1652
AeroGen		0.0855	0.153	8.209	20.314	0.2142
CC-Diff		0.0891	0.1582	8.0909	20.066	0.1963
Ours		0.1009	0.1763	7.6579	19.459	0.2691
LayoutDiff	DOTA	0.0402	0.0748	15.229	30.202	0.2194
GLIGEN		0.0645	0.1182	10.432	23.196	0.1967
AeroGen		0.0863	0.1536	8.1386	20.687	0.2261
CC-Diff		0.0692	0.1255	9.6226	21.247	0.2171
Ours		0.1205	0.2045	6.6317	17.311	0.2938

Table 1: Canny 边缘图上的对象形状保真度。我们通过计算 IoU、DICE、Chamfer 距离（CD）、Hausdorff 距离（HD）和 SSIM 来衡量生成实例与真实实例之间的形态相似性。

Method	Unknown Layout during Training					
	FID ↓	KID ↓	CMMD ↓	CAS ↑	YOLO Score ↑	mAP $_{50}$ ↑
LayoutDiff	44.58	0.018	0.539	29.34	10.37	53.07
GLIGEN	39.56	0.013	0.444	66.36	2.13	52.68
AeroGen	28.62	0.013	0.276	80.78	46.36	55.11
CC-Diff	49.92	0.024	0.513	78.01	51.74	53.72
Ours	24.18	0.012	0.271	83.34	49.59	56.65

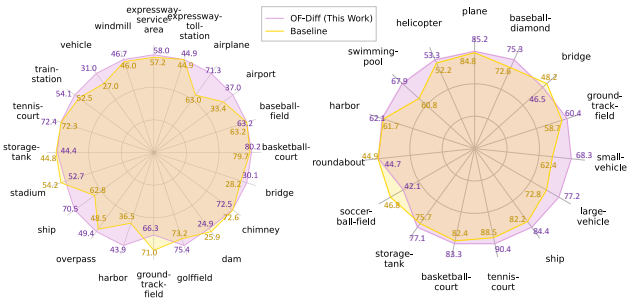
Table 2: 在训练过程中对未知布局数据集进行定量比较（DIOR Val）。

Diff 在 DIOR 和 DOTA 上的表现最佳。具体来说，与基线相比，OF-Diff 在 DIOR 上提高了 2.2 % 的 mAP，在 DOTA 上提高了 1.94 %。值得注意的是，几个多形态和小物体类别的性能有了显著提升。根据图 7 ，DIOR 上的飞机、船和车辆的 AP $_{50}$ 分别增加了 8.3 %、7.7 % 和 4.0 %，而在 DOTA 上的游泳池、小车辆和大车辆的 AP $_{50}$ 分别增加了 7.1 %、5.9 % 和 4.4 %。

物体-形状保真度。我们通过计算基于 Canny 边缘检测图的交并比（IoU）、DICE 系数、Chamfer 距离（CD）、Hausdorff 距离（HD）和结构相似性指数（SSIM）来测量生成实例与真实情况之间的形态相似性。如表 1 所示，结果表明 OF-Diff 在所有物体-形状保真度的评价指标中都达到了最先进的性能。具体来说，我们首先将旋转的边界框（R-Box）转换为水平边界框（H-Box），并使用 20 % 的边距裁剪实例，以确保完整捕捉到物体。然后将裁剪的图块调整到 64 × 64 像素，并使用 cv2.Canny 提取其形状。为了详细的定性比较，图 8 可视化了实例图块及其来自不同方法的对应边缘图。每组图像按以下顺序排列：真实情况、OF-Diff、AeroGen、CC-Diff、GLIGEN 和 LayoutDiff，展示了我们的方法在遵循物体形状上的优越能力。

未知布局的适应性。为了更好地展示这些方法的可扩展性，我们还在训练阶段生成基于未知布局的图像。根据表 2 ，对于未知布局来说，OF-Diff 在生成保真度、布局一致性和可训练性方面表现良好。在下游任务中，OF-Diff 仍然比第二好的方法提升了 1.54 % mAP。

下游的详细结果。表格 3 和 4 报告了在下游任务中竞



(a). Per-class Ap_{50} on DIOR Dataset. **(b).** Per-class Ap_{50} on DOTA Dataset.

Figure 7: 在 DIOR 和 DOTA 上的 AP $_{50}$ 。

争的生成方法在多个类别中获得的平均准确率（AP）。从表中 3 可以观察到，OF-Diff（我们的）在几个类别中取得了明显的优势。例如，OF-Diff 在飞机（71.3 %）、高尔夫球场（75.4 %）和船舶（70.5 %）上表现出色，相较于第二名方法有大约 5 % 到 10 % 的提升。在其他一些类别中，虽然 OF-Diff 未能达到最高 AP，但与最佳结果的差距依然很小。表格 4 显示，在 DOTA 数据集上，OF-Diff 在大约一半的类别中获得了最高的 AP，并且在小型车辆（68.3 %）、船舶（84.4 %）和游泳池（67.9 %）等类别中仍取得了显著的提升。

我们在表格 ?? 中通过 OF-Diff 评估了不同模块对图像生成语义一致性和下游可训练性的影响。我们发现字幕的输入会影响图像生成的逼真度。带有字幕生成的图像更符合语义一致性和人类审美，但这些图像的逼真度降低。这相当于数据分布偏离真实数据集，更倾向于预训练期间的数据分布。因此，每个模块的消融实验都是在没有字幕输入的基础上进行的。通过在具有双分支的扩散模型中加入额外的组件来评估每个模块对提高图像生成逼真度的贡献。DDPO 表示是否通过强化学习对训练的扩散模型进行微调。结果显示，增强形状生成模块（ESGM）、扩散一致性损失（DCLoss）以及基于 KNN 和 KL 散度的 DDPO 有效提高了性能指标。值得注意的是，ESGM 可以将 YOLOScore 大幅提高超过 10 %。

如图 ?? 所示，作为输入的额外标题的加入对图像生成的结果产生了显著影响。具体来说，合并标题提高了生成图像的美感，导致颜色构成更丰富、更具视觉吸引力。然而，这种改进是有代价的：类似于 CC-Diff，它导致生成的数据分布偏离原始真实数据的分布。相比之下，当没有作为输入提供额外标题时，尽管生成的图像在美学上可能不那么精致，但其数据分布更接近真实图像。此外，实验结果表明，这种方法在图像质量指标和下游任务有效性方面表现出更优的性能。

根据当前的实验结果，添加 DDPO 策略并不能在所有指标上同时优于先前的结果。使用强化学习策略确实可以改善下游任务的性能，但这并不意味着同时提高图像生成的质量。换句话说，强化学习策略也可以有目的地提高图像生成质量，但这可能会以不改善下游任务性能为代价。

局限性

提出的 OF-Diff 将从图像布局中提取的物体形状掩膜作为可控条件注入到扩散模型中，这有效地增强了物体的保真性并改善了小物体的生成。然而，这也使模型严

重依赖于提取的形状掩膜的质量。如果获得的掩膜已经失真，生成的图像同样会无法达到令人满意的质量。

结论

现有的图像生成方法在精确生成密集的小物体以及那些具有复杂形状的物体方面存在困难，例如遥感图像中的大量小型车辆和飞机。为了解决这个问题，我们引入了 OF-Diff，这是一种双分支可控扩散模型，具有先验形状提取和 DDPO。在训练阶段，我们提取物体的先验形状以增强可控性，并使用带有参数共享的双分支扩散来提高模型对真实图像的学习能力。因此，在采样阶段，OF-Diff 可以生成具有高保真度的图像，而无需真实图像作为参考。最后，我们通过结合 KNN 和 KL 散度的 DDPO 微调扩散，以使合成图像更加真实和一致。大量实验表明，OF-Diff 在生成复杂结构和密集场景中的小而难的物体方面的有效性和优越性。

References

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In International Conference on Learning Representations.

Chen, K.; Xie, E.; Chen, Z.; Wang, Y.; Hong, L.; Li, Z.; and Yeung, D.-Y. 2023. GeoDiffusion: Text-Prompted Geometric Control for Object Detection Data Generation.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34: 8780–8794.

Gong, Z.; Wei, Z.; Wang, D.; Ma, X.; Chen, H.; Jia, Y.; Deng, Y.; Ji, Z.; Zhu, X.; Yokoya, N.; et al. 2024. Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. arXiv preprint arXiv:2410.22629.

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.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33: 6840–6851.

Hu, X.; Gong, Z.; Wang, Y.; Jia, Y.; Luo, G.; and Yang, X. 2025. Earth-Adapter: Bridge the Geospatial Domain Gaps with Mixture of Frequency Adaptation. arXiv preprint arXiv:2504.06220.

Jayasumana, S.; Ramalingam, S.; Veit, A.; Glasner, D.; Chakrabarti, A.; and Kumar, S. 2024. Rethinking fid: Towards a better evaluation metric for image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9307–9315.

Jia, Y.; Marsocci, V.; Gong, Z.; Yang, X.; Vergauwen, M.; and Nascetti, A. 2025. Can Generative Geospatial Diffusion Models Excel as Discriminative Geospatial Foundation Models? arXiv preprint arXiv:2503.07890.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems, 35: 26565–26577.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. Advances in neural information processing systems, 34: 852–863.

Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D.; and Ermon, S. 2023. Diffusionsat: A generative foundation model for satellite imagery. arXiv preprint arXiv:2312.03606.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. Advances in neural information processing systems, 34: 21696–21707.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 22511–22521.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; and Zhou, J. 2024. Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62: 1–16.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, 10012–10022.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems, 35: 5775–5787.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In International Conference on Machine Learning, 16784–16804. PMLR.

Qiu, K.; Gao, Z.; Zhou, Z.; Sun, M.; and Guo, Y. 2025. Noise-Consistent Siamese-Diffusion for Medical Image Synthesis and Segmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference, 15672–15681.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2): 3.

Ravuri, S.; and Vinyals, O. 2019. Classification accuracy score for conditional generative models. Advances in neural information processing systems, 32.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In International conference on machine learning, 1278–1286. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-Resolution Image Synthesis with Latent Diffusion Models. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis With Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35: 36479–36494.

Sebaq, A.; and ElHelw, M. 2024. Rsdiff: Remote sensing image generation from text using diffusion model. Neural Computing and Applications, 36(36): 23103–23111.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In International Conference on Learning Representations.

Tang, D.; Cao, X.; Hou, X.; Jiang, Z.; Liu, J.; and Meng, D. 2024. Crs-diff: Controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing.

Tang, D.; Cao, X.; Wu, X.; Li, J.; Yao, J.; Bai, X.; Jiang, D.; Li, Y.; and Meng, D. 2025. AeroGen: Enhancing remote sensing object detection with diffusion-driven data generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, 3614–3624.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. Advances in neural information processing systems, 30.

Wang, X.; Darrell, T.; Rambhatla, S. S.; Girdhar, R.; and Misra, I. 2024. Instancediffusion: Instance-level control for image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6232–6242.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4): 600–612.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3520–3529.

Yang, X.; Yan, J.; Feng, Z.; and He, T. 2021. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI conference on artificial intelligence, volume 35, 3163–3171.

Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Yao, L.; Liu, F.; Chen, D.; Zhang, C.; Wang, Y.; Chen, Z.; Xu, W.; Di, S.; and Zheng, Y. 2025. RemoteSAM: Towards Segment Anything for Earth Observation. arXiv preprint arXiv:2505.18022.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, 3836–3847.

Zhang, M.; Liu, Y.; Liu, Y.; Zhao, Y.; and Ye, Q. 2024. CC-Diff: enhancing contextual coherence in remote sensing image synthesis. arXiv preprint arXiv:2412.08464.

Zhang, X.; Han, L.; Han, L.; and Zhu, L. 2020. How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery? Remote Sensing, 12(3): 417.

Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22490–22499.

Zhou, D.; Li, Y.; Ma, F.; Zhang, X.; and Yang, Y. 2024. MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6818–6828.

Zhu, J.; Li, S.; Liu, Y. A.; Yuan, J.; Huang, P.; Shan, J.; and Ma, H. 2024. Odgen: Domain-specific object detection data generation with diffusion models. Advances in Neural Information Processing Systems, 37: 63599–63633.

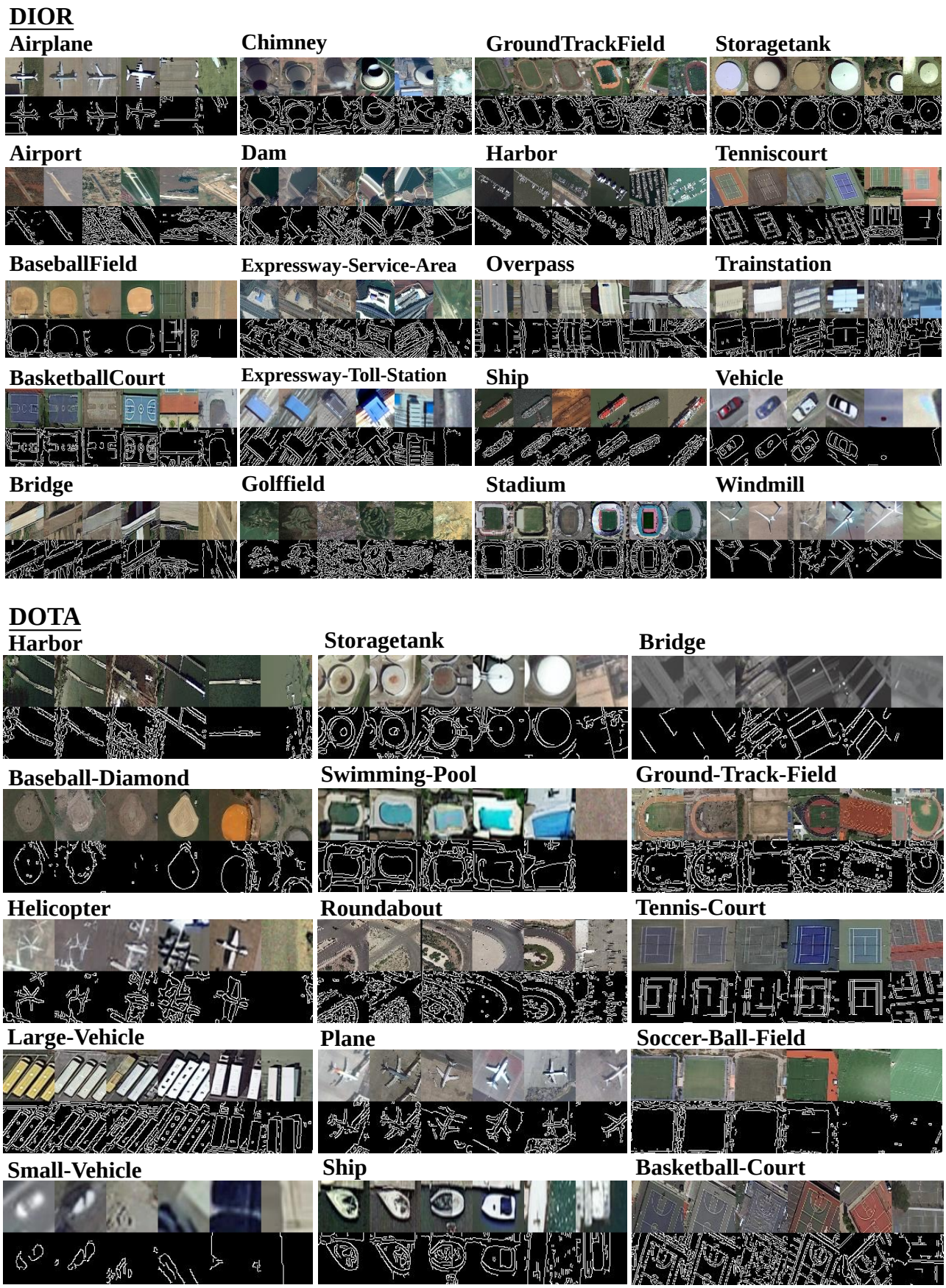


Figure 8: 针对 DIOR 和 DOTA 数据集在相同的 bbox 上生成实例补丁及其 Canny 边缘图的比较。每组图像的顺序如下：Ground Truth, OF-Diff, AeroGen, CC-Diff, GLIGEN, 和 LayoutDiff。

Method	DIOR Dataset									
	Expressway Service-area	Expressway toll-station	Airplane	Airport	Baseball field	Basketball court	Bridge	Chimney	Dam	Golf Field
LayoutDiff	53.1	44.8	62.8	29.4	63.2	79.6	25.9	72.6	22.4	69.3
GLIGEN	52.7	44.8	62.6	26.7	63.0	79.6	25.2	72.6	19.5	67.4
AeroGen	58.1	45.2	63.1	32.7	63.4	81.0	29.5	72.6	21.1	69.1
CC-Diff	53.5	45.1	62.9	38.4	63.3	79.9	29.3	72.7	27.6	70.5
Ours	58.0	44.9	71.3	37.0	63.2	80.2	30.1	72.5	24.9	75.4

Method	DIOR Dataset									
	Ground Track-field	Harbor	Overpass	Ship	Stadium	Storage Tank	Tennis Court	Trainstation	Vehicle	Windmill
LayoutDiff	71.2	32.8	43.9	62.9	59.0	52.5	72.4	52.1	26.9	46.0
GLIGEN	70.1	30.3	45.8	62.8	56.8	52.0	72.5	49.0	26.9	45.3
AeroGen	71.0	42.7	50.7	62.9	56.6	44.5	72.5	52.6	31.4	46.7
CC-Diff	64.6	43.1	49.0	63.0	61.7	44.7	72.4	54.4	27.0	46.5
Ours	66.3	43.9	49.4	70.5	52.7	44.4	72.4	54.1	31.0	46.7

Table 3: 在 DIOR 数据集上的详细下游可训练性结果（通过平均精度测量）。

Method	DOTA Dataset							
	Plane	Baseball-diamond	Bridge	Ground Track-field	Small-vehicle	Large-vehicle	Ship	Tennis-court
LayoutDiff	80.4	74.2	48.8	59.9	62.9	72.7	82.5	89.6
GLIGEN	87.0	72.9	47.3	56.4	63.7	73.1	82.7	90.1
AeroGen	86.1	77.3	48.6	58.6	64.5	78.1	82.5	83.3
CC-Diff	87.2	73.4	47.1	57.8	64.3	73.9	82.6	89.2
Ours	85.2	75.3	46.5	60.4	68.3	77.2	84.4	90.4

Method	DOTA Dataset						
	Basketball-court	Storage-tank	Soccer-ball-field	Roundabout	Harbor	Swimming-pool	Helicopter
LayoutDiff	78.9	76.3	46.3	47.6	60.9	62.1	57.9
GLIGEN	79.6	76.5	42.2	43.1	60.7	62.3	53.8
AeroGen	77.3	79.9	44.8	46.6	59.4	62.6	56.4
CC-Diff	79.0	82.7	42.7	43.1	58.9	62.7	52.9
Ours	83.3	77.1	42.1	44.7	62.1	67.9	53.3

Table 4: 在 DOTA 数据集上的详细下游可训练性结果（通过平均精度衡量）。

Figure 9: DIOR 和 DOTA 上的其他定性结果。结果表明，OF-Diff 在生成小物体方面具有一定的优越性和准确性，并且在生成物体形状方面也具有优势。例如，第三行中的飞机目标由 OF-Diff 生成得更为准确，结构更为逼真。第四行和第五行的小型车辆以及第六行的大型车辆也生成得更加准确。此外，第七行中的小型船只生成得更为精确。