MVDiffusion：通过对应感知扩散实现整体多视图图像生成

王延逸凡

2024 年 12 月 10 日

摘要

本文介绍了 MVDiffusion，一种通过给定像素到像素对应关系的文本提示生成一致的多视图图像的简单而有效的生成方法，。（例如，从全景图或给定语义的多视图图像中进行透视裁剪贴图和姿势）。本文基于 MVDiffusion 将生成任务从室内全景图迁移至室外全景图上，并且由于室外环境的复杂性，引入了更加强的控制信息，例如：素描图、语义图、轮廓图。鉴于现在缺少室外全景图的公开数据集，本文提出了一个高质量的街景数据集。经过改进的 MVDiffusion 在经过训练后，能在室外街景全景图的生成任务上达到最先进的效果。

关键词：扩散模型；多视角图像生成；条件控制

1 引言

使用扩散模型创建 360 ° 全景图像是计算机视觉中一个新兴但关键的前沿领域，在社会需求和应用上有着巨大的潜力，例如环境照明 [1,3]，VR/AR [21,22]，自动驾驶 [20] 和视觉导航 [6]。然而生成全景图像和正常图像有所不同，其重要面临两类困难。首先全景图像和正常图像存在几何信息和域信息的不同，比如全景图的纵横比一般是 2：1，而正常图像则是一个正方形。其次，相较于正常图像生成，全景图生成缺少相应的数据集，全景数据的稀缺以及其对应的文本则更难获取。尽管现有方法在室内场景全景图生成上已经有了很好的效果，如 [16,23]，但是对于室外场景，由于背景更复杂、环境因素多样、物体数量和种类等原因，室外场景生成难度要大得多。

据我们所知，现如今没有生成街景的工作。简单的方法是直接采用现有的室内场景生成方法，并将其应用于街景生成任务。我们发现简单的修改现有模型在街景生成上表现不佳。由于上述原因，街景的生成更加复杂。因此本文提出了一种新的架构，通过引入并利用轮廓信息来确保生成图像的稳定性和一致性。

在本文中，我们首先提出了一种带有结构提示的多视角稳定扩散模型，以应对街景生成任务中的挑战。此外，为了缓解在街景全景图领域中的数据缺失，我们还为生成任务构建并提出了一个大型多视图街景数据集 Street360。我们将我们的贡献总结如下：

- 提出了一种新的框架，该框架能偶利用结构信息，生成更加可控的街景图像

- 提出了一个室外街景数据集 Street360，该数据集是第一个专门针对街景数据的全景数据的大规模数据集。
- 我们的方法在街景数据集上取得了最先进的效果。

2 相关工作

在这一部分，我们将回顾与本文课题内容相关的先前工作，其中涵盖了得分基础生成模型、扩散模型及其改进、以及其他生成模型和技术的研究。

2.1 扩散模型

近年来，扩散模型 [4] 在图像生成领域掀起了风暴，因为它们变得更快 [1, 7] 并且在图像质量和分辨率方面更有能力 [12]。这一成功促进了扩散模型的各种应用的发展，例如文本到图像 [12]、图像条件生成 [3]、内绘（in-painting）[10, 11] 和主题驱动生成 [14]。这些应用程序中的大多数试图利用预先训练的扩散模型的先验知识来缓解任务特定数据的稀缺性，通过使用 LoRA [5] 等技术进行微调，或引入辅助模块来提取知识。我们还采用相同的原理来利用预先训练的潜在扩散模型 [13] 的能力来生成全景图像。

2.2 生成全景图

全景图像生成包含不同的设置，包括全景外绘和文本到全景生成。全景轮廓绘制 [18,19,23] 关注于从部分输入图像生成 360 度全景。各种方法，如 StyleLight 和 BIPS，已经解决了特定的用例，重点是 HDR 环境照明和机器人引导场景。最近的工作 [19,23] 改进了扩散模型的真实性，但往往缺乏对来自预训练模型的丰富先验信息的利用，限制了推广。另一方面，生成模型的最新发展已经在从文本输入合成沉浸式视觉内容方面开辟了新的前沿 [15, 17, 19, 23]。作为一种基于图像的表示方法，从文本中生成全景图的方法受到了广泛的关注。Text2Light [3] 采用 VQGAN 结构从文本合成 HDR 全景图像。为了使用预先训练的扩散模型以任意分辨率生成，DiffCollage [24]、MultiDiffusion [2] 和 SyncDiffusion [8] 建议融合扩散路径，而 PanoGen [6] 通过迭代修补来解决。Lu 等人 [9] 采用了自回归框架，但存在效率低下的问题。相比之下，本文介绍了 MVDiffusion，这是一种多视图文本到图像生成架构，只需对标准的预训练文本到图像扩散模型进行最少的更改，即可在两个多视图图像生成任务上实现最先进的性能。

3 本文方法

3.1 本文方法概述

MVDiffusion 通过运行稳定扩散模型的多个副本/分支来同时生成多个图像，该模型具有新颖的分支间“对应感知注意力”（CAA）机制，以促进多视图一致性。图 2显示了多分支 UNet 和 CAA 设计的概述。该系统适用于当图像之间存在像素到像素的对应关系时，特别是对于以下情况：1）生成全景或将透视图像外推到全景。全景由共享相机中心的透视图像组成，通过平面断层摄影获得像素到像素的对应关系，以及 2）给定几何结构的纹理映射，其中任意相机姿态的多个图像通过基于深度的反投影和投影建立像素到像素的对应关系。

整体流程如图 1所示：

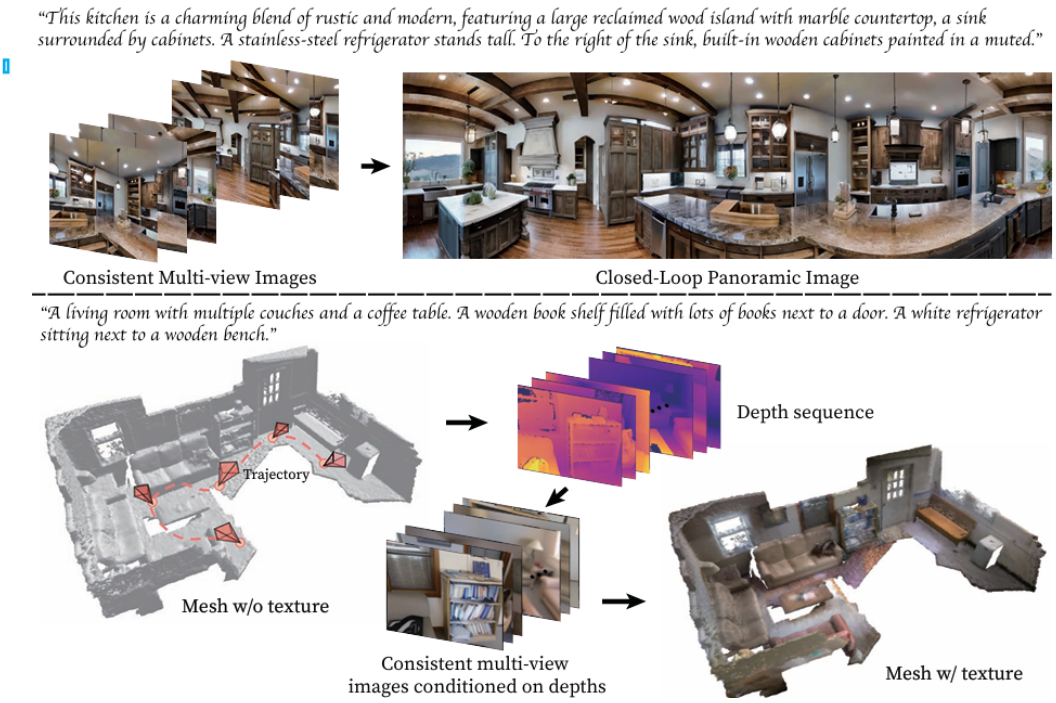

图 1. 整体方法示意图

3.2 全景图生成

在 MVDiffusion 中，全景图是通过生成八个透视图来实现的，每个透视图都拥有 90 ° 的水平视野，重叠 45 °。为了实现这一点，我们使用冻结的预训练稳定扩散模型，通过生成模块生成 8 张 512 × 512 图像。

生成模块。该模块可生成 8 幅 512 × 512 的图像。它通过同步降噪过程来实现这一点。这个过程包括将每个噪声潜势馈送到一个共享的 UNet 架构（称为多分支 UNet）中，以同时预测噪声。为了保证多视图的一致性，在每个 UNet 块之后引入了对应感知注意（CAA）块。CAA 块位于最终 ResNet 块之后，负责接收多视图要素并将其融合在一起。

对应感知注意（CAA）。CAA 的架构如图 3所示。CAA 块并发地对 N 个特征图进行操作，如图 2所示。对于表示为 F 的第 i 个源特征图，它与表示为 F1 的剩余（N-1）个目标特征图执行交叉关注。对于位于源特征图中的位置处的标记，我们基于具有局部邻域的目标特征图 F1（不一定在整数坐标处）中的对应像素 tl 来计算消息。具体地，对于每个目标像素 tl，我们通过将整数位移（dx/dy）添加到（x/y）坐标来考虑 K ×K 邻域 N（tl），其中 |DX| < K/2，|dy| < K/2。实际上，我们使用 K = 3，邻域为 9 个点。

$$\mathbf{M} = \sum_{l} \sum_{t_*^l \in \mathcal{N}(\mathbf{t}^l)} \text{SoftMax}\left(\left[\mathbf{W_Q}\bar{\mathbf{F}}(\mathbf{s})\right] \cdot \left[\mathbf{W_K}\bar{\mathbf{F}}^l(t_*^l)\right]\right) \mathbf{W_V}\bar{\mathbf{F}}^l(t_*^l)$$

$$\bar{\mathbf{F}}(\mathbf{s}) = \mathbf{F}(\mathbf{s}) + \gamma(0), \quad \bar{\mathbf{F}}^l(t_*^l) = \mathbf{F}^l(t_*^l) + \gamma(\mathbf{s}_*^l - \mathbf{s}).$$

消息 M 计算遵循标准注意力机制，该标准注意力机制将信息从目标特征像素 t1 聚集到源。WQ、WK 和 WV 是查询、键和值矩阵。关键区别在于基于目标特征 Fl（tl）在源图像中的

对应位置 sl（t）与 s 之间的 2D 位移，将位置编码 (·) 添加到目标特征 Fl（tl）。位移提供了当地社区的相对位置。注意，位移是一个 2D 向量，我们对 x 和 y 坐标中的位移应用标准频率编码，然后连接。目标特征 Fl（tl）不处于整数位置并且通过双线性插值获得。为了保留稳定扩散模型的固有功能，我们将 Transformer 块的最终线性层和残差块的最终卷积层初始化为零，如 ControlNet 中所建议的那样。这种初始化策略确保我们的修改不会破坏稳定扩散模型的原始功能。

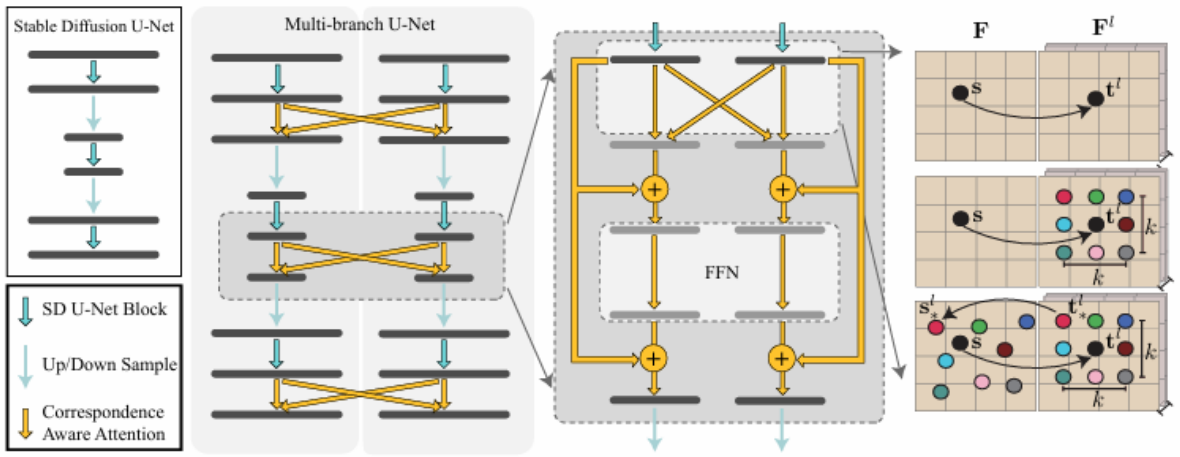

图 2. CCA 模块示意图

全景式外扩。目标是基于单个透视图像（一个条件图像）和每个视图的文本提示生成完整的 360 度全景视图（七个目标图像）。我们使用 SD 的修补模型作为基础模型，因为它采用一个条件图像。与生成模型类似，初始化为零的 CAA 块被插入到 UNet 中，并在我们的数据集上进行训练。

在生成过程中，该模型将目标图像和条件图像的潜在噪声从标准高斯模型中恢复出来。在条件图像的 UNet 分支中，我们将一个 1 的掩码连接到图像（总共 4 个通道）。然后，这个连接的图像作为修复模型的输入，这确保了条件图像的内容保持不变。相反，在目标图像的 UNet 分支中，我们将黑色图像（像素值为零）与零掩码连接起来作为输入，从而要求修复模型基于文本条件和与条件图像的对应关系生成全新的图像。

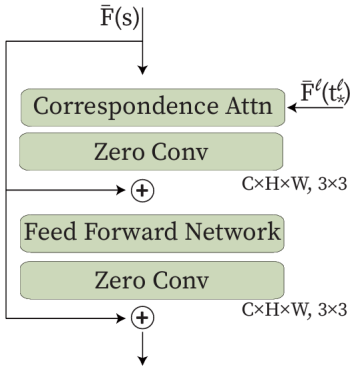

图 3. CAA 模块架构

Method	FID↓	IS↑	CS↑
Impainting	42.13	7.08	29.05
Text2light	48.71	5.41	25.98
SD(Pano)	23.02	6.58	28.63
SD(Perspective)	25.59	7.29	30.25
MVDiffusion(Ours)	21.44	7.32	30.04

表 1. 定量对比

3.3 损失函数定义

训练。我们将 CAA 块插入到预训练的稳定扩散 Unet 或稳定扩散修复 Unet 中，以确保多视图一致性。当我们使用以下损失来训练 CAA 块时，预训练的网络被冻结：

$$L_{\text{MVDiffision}} := \mathbb{E}_{\left\{\mathbf{z}_i^i=\mathcal{E}(\mathbf{x}^i)\right\}_{i=1}^N,\{\epsilon^i\sim\mathcal{N}(0,I)\}_{i=1}^N,\mathbf{y},t}\left[\sum_{i=1}^N \|\epsilon^i-\epsilon_\theta^i(\left\{\mathbf{Z}_t^i\right\},t,\tau_\theta(\mathbf{y}))\|_2^2\right]$$

4 复现细节

4.1 与已有开源代码对比

源代码在 https://github.com/Tangshitao/MVDiffusion。在 MVDiffuison 的基础上增加了条件输入，并且将网络进行更改以适应输入的改变。将 Oneformer 部署到本地，编写脚本文件获取场景的粗略语义图。其次，我将改进后的使用轮廓信息的框架在更大型的室外场景数据集上进行训练，最终得到了比原模型好的效果。

最后，论文中的一些 baseline 模型没有给出代码，我使用代码将其结果进行评估。

4.2 创新点

使用了语义图作为提示，并且将生成场景从室内迁移到室外场景，在室外场景的生成上取得了比较好的效果。原有的文本提示词是根据 BLIP2 进行生成的，经过观察发现，由于全景图切分成视角图之后相邻视角之间的图像变化不大，而又因为 BLIP2 生成的文本比较简单，因此描述信息可能一致，为此我们采用 OVIS 进行生成详细的文本描述。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。表 1展示了在室外场景数据集上训练的结果，评价指标用 FID、PSNR、CS、IS。等，这些指标在原文中均有使用，我遵循原文的设置进行测试。可以明显地看出，在所有指标上，我们训练的室外模型明显由于之前的模型。从表 1中的各项量化指标中，可以看到训练是十分成功的。

下面是 baseline 模型和本文的模型的对比图：


A living room filled with furniture and a large mirror. A table with a lamp on it. A living room filled with furniture and a chandelier.
Ours
Inpainting
SD(Pano)
Text2light

图 4. 生成图片对比

6 总结与展望

本文介绍了 MVDiffusion，一种创新的方法，同时产生一致的多视图图像。我们的主要新奇是集成的对应感知注意力（CAA）机制，通过识别像素到像素的对应关系，确保跨视图的一致性。这种机制被纳入每个 UNet 块的稳定扩散。通过使用冻结的预训练稳定扩散模型，大量的实验表明，MVDiffusion 在全景图像生成和多视图深度到图像生成方面实现了最先进的性能，有效地减轻了以前方法的累积误差问题。此外，我们的高级想法有可能扩展到其他生成任务，如视频预测或 3D 对象生成，为更复杂和大规模场景的内容生成开辟了新的途径。

此外，使用了一个新的室外数据集并构建其边缘图像和文本提示。实验结果表明，新的 MVDiffusion 模型能够生成与原图更相似的图像，比原 MVDiffusion 模型生成效果要好。不足之处在于，使用 OVIS 方法对服饰进行描述太过于详细，无法充分利用其中的文本信息，由于资源的限制模型训练的数据集数量偏少，可能导致模型的泛化性能较低。

MVDiffusion 提供了多种可控的生成方式，使得用户可以更好的根据自己的需求来生成图像，如何充分利用多种可控条件和模型中的注意力机制来实现更细粒度的室外场景生成，并结合使用 OVIS 之类的图生文方法提取文本提示信息，是未来需要解决的关键问题，这有利于游戏或者艺术设计师们仅用文字描述和简单的条件控制（如边缘）就能够很快实现与期望相符的场景设计。

参考文献

[1] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11431–11440, 2022.

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

[3] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

[4] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[5] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[6] Mohit Bansal Jialu Li. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *arxiv*, 2023.

[7] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022.

[8] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *arXiv preprint arXiv:2306.05178*, 2023.

[9] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation, 2024.

[10] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022.

[11] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022.

[12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[13] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[14] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[15] Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Zhao Yang. Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6898–6906, New York, NY, USA, 2023. Association for Computing Machinery.

[16] Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint 2307.01097*, 2023.

[17] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4933–4943, 2024.

[18] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6811–6821, 2023.

[19] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *The Twelfth International Conference on Learning Representations*, 2023.

[20] Jianru Xue, Jianwu Fang, and Pu Zhang. A survey of scene understanding by event reasoning in autonomous driving. *International Journal of Automation and Computing*, 15:249 – 266, 2018.

[21] Bangbang Yang, Wenqi Dong, Lin Ma, Wenbo Hu, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. Dreamspace: Dreaming your room space with text-driven panoramic texture propagation. *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 650–660, 2023.

[22] Bangbang Yang, Yinda Zhang, Yijin Li, Zhaopeng Cui, Sean Fanello, Hujun Bao, and Guofeng Zhang. Neural rendering in a room: Amodal 3d understanding and free-viewpoint rendering for the closed scene composed of pre-captured objects. *ACM Trans. Graph.*, 41(4):101:1–101:10, July 2022.

[23] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360◦ panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[24] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023.