NOVA3D：用于单图像到三维生成的法线对齐视频扩散模型

1st Yuxiao Yang * , 2nd Peihao Li * , 3rd Yuhong Zhang, 4th Junzhe Lu
Tsinghua University , Tsinghua University , Tsinghua University , Tsinghua University

5th Xianglong He, 6th Minghan Qin, 7th Weitao Wang, 8th Haoqian Wang †
Tsinghua University , Tsinghua University , Tsinghua University , Tsinghua University

***Abstract*—生成的 3D AI 内容（AIGC）使得任何人都能更容易成为 3D 内容创作者。尽管最近的方法利用得分蒸馏取样从预训练的图像扩散模型中蒸馏 3D 对象，但它们通常因缺乏充分的 3D 先验知识而导致多视图不一致。在这项工作中，我们介绍了一种创新的单图像转 3D 生成框架 NOVA3D。我们的关键见解在于利用预训练的视频扩散模型中强大的 3D 先验，并在多视图视频微调期间整合几何信息。为了促进颜色和几何域之间的信息交换，我们提出了几何-时间对齐（GTA）注意机制，从而提高泛化能力和多视图一致性。此外，我们引入了解决冲突的几何融合算法，通过解决多视图不准确性和姿态对齐差异来改善纹理保真度。大量实验验证了 NOVA3D 相对于现有基准的优越性。**

从单视图图像提示创建 3D 对象对于视频游戏、虚拟现实和增强现实中的广泛应用至关重要。然而，这项任务非常不适定，并且面临显著的挑战。由于高质量 3D 对象数据的采集困难，3D 生成模型在现实性和泛化性方面落后于其 2D 对应模型。因此，利用相关任务中的先验信息，如文本到图像生成，成为增强生成的 3D 对象的现实性和多视图一致性的一个有前景的方法。越来越多的研究 [1], [2] 依赖于通过评分蒸馏采样（SDS）[1] 从预训练的文本到图像模型中提取 3D 表示。为了同时增强多视图一致性和效率，另一种方法 [3]–[6] 利用在 3D 数据集上微调的图像扩散模型 [7] 进行多视图图像生成，然后通过重建过程来获得 3D 对象。尽管这些方法缓解了对大量高质量 3D 数据的需求，但它们依然存在后视模糊纹理、泛化能力不足和 3D 一致性有限的问题。相比之下，人类主要从动态观察中（例如视频）获取 3D 先验，这使得能够从单幅图像推断出 3D 结构。受这种能力的启发，有很大的潜力去探索和利用嵌入在大规模预训练视频模型中的 3D 先验，以提升单图生成 3D 的能力。视频扩散模型在最近的进展中 [8], [9] 由于其生成复杂场景和动态变化，并拥有出色的跨帧一致性的能力，引起了相当大的关注。尽管一些研究已使用微调的视频扩散模型生成多视图图像 [10]–[12] ，但视频扩散模型在捕捉和理解 3D 几何方面的潜力仍未被充分探索。因此，这些方法通常难以建模详细的几何结构并生成高保真纹理细节。

为了增强多视图一致性并充分利用预训练视频扩散模型的几何先验，本文介绍了一种新颖的框架 NOVA3D，它利用嵌入在预训练视频扩散模型中的 3D 先验，从单视图图像生成高质量的纹理网格。我们的关键见解在于将几何信息作为辅助监督进行整合，这增强了预训练视频扩散模型中 3D 先验的激活。这种改进使得视频扩散模型能够预测多视图图像和相应的法线图，从而促进高保真纹理网格的重建。此外，我们在潜在视频扩散模型 (LVDM) 架构中引入了创新的几何-时间对齐 (GTA) 注意机制，该机制使 RGB 图像和法线图的生成对齐，从而无需修改预训练模型即可将 RGB 视频域的泛化能力转移到几何域。为了解决生成的姿态与预定义姿态之间的差异，以及细微的跨视图不一致性，我们提出了解冲突几何融合算法。该算法结合隐式冲突建模和姿态优化技术，确保可靠且一致的纹理网格生成。我们在 Google 扫描对象数据集 [13] 和分布外输入上的评估显示了 NOVA3D 的有效性，定量结果表明，与基线方法相比，其具备更高的保真度和泛化能力。

综上所述，我们的贡献可以总结如下：

- 我们介绍了 NOVA3D，这是一种新颖的方法，它从视频扩散模型中释放出几何 3D 先验，用于从输入图像生成高质量的纹理网格。
- 我们提出了几何-时间对齐注意机制，以促进纹理与几何潜变量之间的模式交换，有效地将泛化性能转移到几何域。
- 我们提出了一种去冲突几何融合算法，结合隐式冲突建模和姿态优化技术，提高了鲁棒性和纹理逼真度。

I. 相关工作

A. 用于 3D 生成的图像扩散模型

近年来，图像扩散模型 [14]–[16] 发展迅速。然而，3D 数据的相对稀缺限制了原生 3D 生成模型的性能。以往的研究尝试利用预训练的图像扩散模型来进行 3D 物体生成。例如，DreamFusion [1] 提出了用于文本到 3D 任务的 SDS 方法，通过文本提示优化神经辐射场 [17] 。虽然后续研究 [2], [18] 通过多阶段优化、增强蒸馏和加速蒸馏方法来改进 SDS，但耗时的每个物体优化和多面问题仍然阻碍了该方法在实际应用中的可行性。为了解决这个问题，最近的方法 [3], [5] 对 3D 数据集 [7] 进行图像扩散模型的微调，从而生成与输入一致的多视角图像。然而，由于缺乏 3D 先验，这些模型往往需要从头开始在 3D 数据集上训练跨视角注意力层，以确保多视角一致性，这也限制了它们生成高质量、密集视图图像的能力。

B. 用于三维生成的视频扩散模型

最近，对视频扩散模型 [8], [9] 的研究有了显著进展。在大量真实视频数据集上预训练生成模型提供了广泛的 3D 先验，包括物体交互、旋转和相机移动。一些近期研究尝试利用来自视频扩散模型 [11], [12] 的先验进行 3D 物体生成，将物体的多视图图像视为连续帧，并使用来自 3D 数

* Equal Contributuion.
† Corresponding Author. wanghaoqian@tsinghua.edu.cn

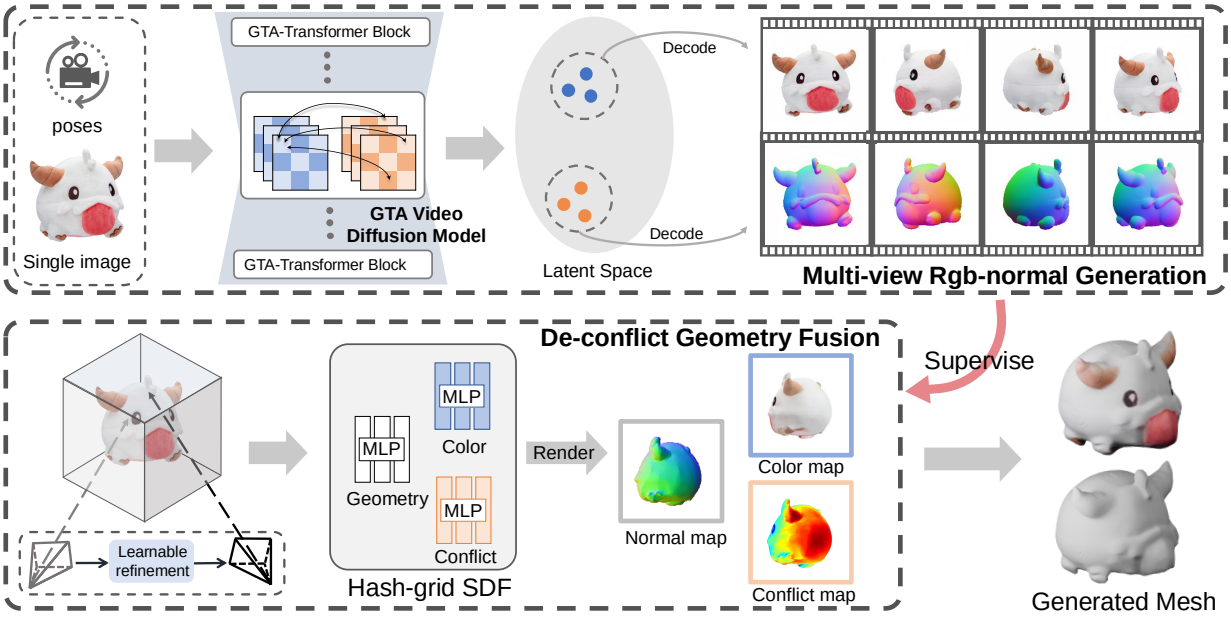

Fig. 1: NOVA3D 流程的概述。我们的方法首先利用结合 GTA 的视差视频扩散模型从单幅图像生成多视角图像及其对应的法线贴图。随后，这些结果通过一个去冲突的几何融合算法进行处理，以重建一个高保真度的纹理网格，从而准确捕捉细节。

据集的渲染多视图图像微调视频扩散模型。然而，现有方法尚未充分利用视频扩散模型中的 3D 信息。因此，我们建议将几何信息作为监督与纹理信息相结合，以微调视频扩散模型，有效地激活视频扩散模型中的 3D 先验信息。

II. 方法

A. 问题表述。

给定一个物体的输入图像 y 和一系列预定义的相机姿态 $\pi_{i:m}$ ，存在一个关于颜色图像和法线图 m 视图的概率分布：

$$p_{ni}(i_{1:m}, n_{1:m}|y, \pi_{i:m}). \tag{1}$$

我们的目标首先是从 p_{ni} 的分布中采样 m 视图的多视图图像 $i_{i:m}$ 及其对应的法线图 $n_{1:m}$ ，然后执行去冲突几何融合算法以生成带纹理的网格。我们假设物体位于标准化 3D 立方体的中心，并采用一系列在 0 仰角均匀分布的相机姿态，从而无需输入仰角。具体来说，我们生成与以下分布匹配的多视图图像和法线图：

$$i_{1:m}, n_{1:m} = f(y, \pi_{1:m}) \sim p(i_{1:m}, n_{1:m}|y, \pi_{i:m}) \tag{2}$$

其中，f 是我们经过精调的视频扩散模型。

B. 释放视频扩散模型中的 3D 先验。

整体架构。通过引入时间维度、Conv3D 残差层和在每个空间层后面的时间注意力层，潜在的视频扩散模型 [9] 生成时间一致的图像序列。NOVA3D 采用了这个架构并从 SVD [8] 初始化权重，确保时间一致性并为多视角生成提供了强有力的先验。条件图像的 CLIP 嵌入随后被用作视频 UNet 中变压器模块的跨注意力层中的关键和价值。我们进行了几项调整以使预训练模型适合我们的任务：(a) 移除“运动桶 id”和“fps id”输入，因为它们与多视角生成无关；(b) 集成摄像机条件 π_i 和单热编码任务条件 t_i 。这些条件，作为标签嵌入到视频 U-net 中的每个序列中，代表查询姿势并决定生成外观或几何形状。

几何结合。对于多视图图像生成任务，之前的工作强调了在 3D 对象数据集中渲染的多视图图像上微调的视频扩散模型的泛化能力。为了在预训练的 SVD 中利用 3D 先验，我们在模型微调过程中整合几何信息。为此，直接的方法通常包括要么在 U-net 中将通道加倍，要么首先生成一系列图像，然后在条件下生成对应的法线贴图。然而，这两种方法都需要权重重新初始化，导致模型发生灾难性遗忘，并降低泛化性能。

与上述方法不同，我们的方法对模型的输出任务提供了控制，允许通过任务条件在颜色和几何域之间无缝过渡。这种增强不仅免除了重新初始化 U-Net 参数的需求，还利用几何信息作为额外的约束，从而增强了预训练期间嵌入的 3D 先验知识。此设计背后的理由是，多视角彩色图像通常缺乏足够的信息来准确反映物体的真实 3D 结构，特别是在无纹理表面上。因此，法线图的监督作为额外的约束，促进了视频扩散模型从视频生成任务到多视角生成任务的更顺畅适应。

C. 几何-时间对齐注意力机制

多任务去噪过程。我们在第 II-B 节中引入的增强功能，使得我们的模型能够生成多视图图像和法线贴图，而无需对网络进行显著修改。尽管 RGB 图像和法线贴图各自在视图之间确保了一致性，直接利用它们进行网格化可能导致纹理与几何的错位。此外，对象的颜色和几何之间存在相互联系。因此，同时考虑两者将有助于模型学习 3D 对象的真实分布。我们将去噪过程公式化如下：

$$p(i_{1:m}, n_{1:m}|\pi_{1:m}, y) = p(i^T_{1:m}, n^T_{1:m}|\pi_{1:m}, y) \cdot \prod_{t\in 1:T} p_\theta(i^{t-1}_{1:m}, n^{t-1}_{1:m}|i^t_{1:m}, n^t_{1:m}, \pi_{1:m}, y). \tag{3}$$

这表明在每个去噪步骤 t 中，我们的模型 f 作为噪声预测器，通过在带噪声的多视图彩色图像 $i^t_{1:m}$ 和相应的法线贴图 $n^t_{1:m}$ 上预测噪声，以共同获得去噪结果 $i^{t-1}_{1:m}$ 和 $n^{t-1}_{1:m}$ 。

GTA 注意模块。为了使模型能够生成对齐的颜色和法线贴图，并促进纹理和几何域之间的模式交换，我们提出了几何-时间对齐（GTA）注意机制。图 **??** 展示了 GTA 注意机制的操作动态。具体而言，在空间层面，GTA 注意机制使得同一视点中的 RGB 图像和法线贴图之间能够高效交互。同时，在时间层面，它确保在潜在特征图中对应位置的不同视点之间的对齐。这种简化的方法与嵌入在潜在视频扩散模型架构中的复杂信息处理模式相协调。更多实现细节请参见补充材料。

结合我们生成的法线贴图以辅助 3D 几何和纹理提取，我们在优化过程中采用隐式符号距离函数，从而简化法线贴图损失的计算。然而，我们的重建过程面临两个潜在的挑战：(a) 生成的姿态与查询姿态之间的细微偏差，以及 (b) 由于 16 视图生成的相对密集性导致重叠视图之间的轻微不一致。为了缓解这些挑战，我们引入了解决冲突的几何融合算法，接下来我们将对此进行讨论。

姿势精炼。为了解决生成姿势与预定义查询姿势之间的不对齐问题，我们引入了一种姿势精炼技术。相机姿势最初根据查询姿势设置，表示为每个视角的旋转矩阵和平移向量，在优化过程中进行精炼。具体来说，从 v_{th} 视角出发的每条光线通过一个可学习的精炼矩阵 M_v 进行偏移，该矩阵在同一视角内的所有光线上保持一致。此方法将姿势精炼到正确的角度，提升生成网格的质量。

冲突建模。相邻视角之间的微小冲突会导致优化不稳定，结果是几何和纹理模糊。为解决此问题，我们使用隐式连续函数 f_ψ 来建模重叠图像之间的冲突 h ：

$$h = f_\psi(f_c, f_g, d(v), l, x) \tag{4}$$

这里 f_c 、f_g 、$d(v)$ 、l 和 x 分别表示颜色 MLP 的输出、几何 MLP、光线方向、视图索引嵌入和坐标位置。在相机空间中的像素 p 处的冲突，表示为 H_p ，通过体积渲染将光线方向投影到二维像素平面上计算得出。H_p 量化了像素 p 与相邻图像的冲突。我们的解冲突颜色损失定义为：

$$\mathcal{L}_{color} = (1 - H_p)\|C_p - \hat{C}_p\|_2 + \lambda_0 H_p^2 \tag{5}$$

其中，C_p 和 $\hat{C}_p$ 分别为渲染的像素颜色和生成的图像颜色。在这个等式中，具有更高冲突值的像素被赋予较小的权重，从而减少重叠视图间不一致在重建过程中带来的负面影响。第二个等式作为正则化项，防止 H_p 过大，从而削弱颜色监督信号。

损失函数。

我们在优化过程的每次迭代中采样一批射线。给定射线上一点 k ，我们查询几何、颜色和冲突 MLPs，以沿射线方向渲染它，从而得出法线映射值 $h_k \in \mathbb{R}$ 、颜色值 c_k 和遮罩 m_k 。最终的优化目标整合了多个损失项：

$$\mathcal{L} = \mathcal{L}_{color} + \mathcal{L}_{normal} + \mathcal{L}_{mask} + \mathcal{R}_{eik} + \mathcal{R}_{sparse} + \mathcal{R}_{smooth} \tag{6}$$

其中 $\mathcal{L}_{color}$ 是我们上面提到的消除冲突损失，$\mathcal{L}_{normal}$ 是 Wonder3D 中提出的几何感知法线损失 [5] ，其最大化生成法线与从 SDF 表示中提取的法线值的相似性，$\mathcal{L}_{mask}$ 是渲染遮罩 m_k 和生成遮罩 $\hat{m}_k$ 之间的 L2 损失，$\mathcal{R}_{eik}$ [19] , $\mathcal{R}_{sparse}$ [20] 和 R_{smooth} [5] 是旨在强制预测 SDF 具有单位 l_2 范数梯度、避免浮动物以及鼓励更平滑的预测 SDF 梯度的正则化项。

Methods	↑ PSNR	↑ SSIM	↓ LPIPS
Zero123 [4]	18.93	0.779	0.166
SyncDreamer [3]	20.05	0.798	0.146
V3D [11]	20.22	0.812	0.132
Envision3D [10]	20.55	0.852	0.130
SV3D [12]	20.88	0.897	0.112
Wonder3D [5]	23.25	0.822	0.104
w/o GTA	22.10	0.802	0.144
w/ cross-domain attn.	23.26	0.824	0.108
Ours	23.87	0.915	0.091

TABLE I: 新视图合成中的定量结果。

III. 实验

我们在 Objaverse 数据集的 LVIS 子集上进行训练，该数据集包含大约 30,000 个 3D 网格。RGB 图像和法线图在 16 个姿态下渲染，每个姿态的分辨率为 256×256 ，用于训练我们的模型。

我们的模型是从公开的 SVD [8] 微调的，使用 8 块 Nvidia A100 GPU 进行了 7 天，实际批次大小为 176。在 SDF 优化过程中，我们使用层次哈希网格 [21] 来用多层次细节编码三维位置，提高了效率。

A. 评估设置

基线。我们将我们的方法与几种单图像到 3D 的方法进行比较，包括 Zero123 [4] 、SyncDreamer [3] 、Wonder3D [5] ，以及最近的基于视频扩散模型的方法，例如 Envision3D [10] 、V3D [11] 和 SV3D [12] 。此外，我们还与各种前馈 3D 生成方法进行比较分析，包括 Shap-E [22] 、One-2-3-45 [23] 和 CRM [24] 。这次全面评估展示了我们的方法在不同基准和场景中的有效性和鲁棒性。

指标。按照之前的工作 [3], [5] ，我们在 Google 扫描对象 [13] 数据集上评估我们的方法，从日常物品到动物中选择 30 个对象。对于 NVS 任务，我们使用 PSNR、SSIM [25] 和 LPIPS [26] 指标来评估我们生成的多视图图像的质量。为了评估我们生成的纹理网格的质量，我们首先采用香农距离和体积 IoU 指标来进行几何评估。此外，我们在固定的 32 个姿势下重新渲染生成的网格，如 Envision3D [10] 所使用，以评估网格纹理的质量。

B. 新视角合成

我们评估了生成的多视图图像的质量，图 **??** 中展示了定性结果，而表格 I 中展示了定量结果。SyncDreamer [3] 的输出缺乏多视图一致性，并且表现出不现实的伪影。Wonder3D [5] 使用了一种多视图注意力机制，可以实现相对一致的多视图图像，但仅限于生成六个视图，这明显少于我们方法生成的十六个视图。虽然 SV3D [12] 通过使用仅 RGB 信息微调 SVD 来实现视图一致性，但生成对象的整体形状真实性和颜色细节仍然不够。相比之下，我们的模型在几何信息的辅助监督支持下，并利用预训练的 SVD 中的基本 3D 先验，擅长生成在视图间一致且语义上连贯的多视图图像。

我们评估并比较生成网格的几何和纹理质量，与当前最先进的方法进行对比。如表 **??** 所示，我们的方法在所有

Fig. 2: 对于生成的纹理网格与基线进行定性比较。

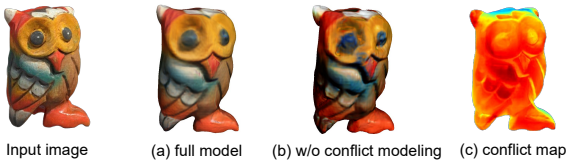

Fig. 3: 关于隐式冲突建模的消融研究。

指标上表现优异，突显出其生成具有丰富纹理细节的高保真 3D 内容的能力。图 2 展示了生成纹理网格的定性比较，进一步说明了我们的方法在网格几何、纹理和高层语义一致性方面显著优于基线。

C. 讨论

几何-时间对齐 (GTA) 注意力。为了验证所提出的 GTA 注意力机制的有效性，我们进行了不同模型配置的实验：(a) 微调结合 GTA 注意力模块的视频扩散模型，(b) 利用 Wonder3D [5] 提出的跨域注意力模块进行微调，以及 (c) 一个不包含 GTA 或跨域注意力模块的变体模型。图 **??** 显示了可视化效果，而表 I 和 **??** 呈现了定量结果。

在图 **??** 中的 (a) 和 (b) 进行比较时，(b) 中缺乏 GTA 注意力阻碍了帧之间的信息交换，导致几何法线未能理解物体的整体形状，并与生成的纹理信息对齐。同样地，如图 **??** 的 (a) 和 (c) 中所示，颜色和法线特征之间的缺乏交互阻碍了从纹理域到几何域的泛化性转移。相比之下，如 (a) 中所示，视频扩散模型架构中 GTA 模块的整合能够生成更清晰、更准确的法线贴图，同时确保颜色和法线特征之间的卓越一致性。这突出了我们方法在弥合多任务和多视图依赖性方面的有效性。

解冲突几何融合算法。我们通过表格 **??** 中的定量比较和图 3 中展示的定性可视化评估了冲突建模和姿势优化方法的有效性。如图 3 的 (a) 和 (b) 所示，冲突建模方法有效缓解了多视图生成结果中的细微不一致，生成的网格具有更真实的纹理。如图 **??** 的 (b) 和 (c) 可视化的那样，冲突图中值较高的区域对应于 (b) 中的过饱和和模糊区域。这表明我们的冲突图有效捕捉了重叠视图中的不一致性，从而能够生成具有高保真纹理的网格。

IV. 结论

在本文中，我们介绍了 NOVA3D，这是一种创新的方法，可以在视频扩散模型中释放三维先验，从任何单张图像生成高质量的纹理网格。通过将几何信息作为辅助监督信号，并采用几何-时间对齐注意机制，我们的微调视频扩散模型可以生成密集视图对齐的图像和法线贴图。此外，去冲突几何融合算法有效地解决了生成图像中的细微多视图冲突，并解决了生成姿态与预定义姿态之间的错位。实验结果验证了我们的方法具有稳健且广泛适用的性能，显著优于现有基线。

References

[1] Ben Poole, Ajay Jain, et al., "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[2] Chen-Hsuan Lin, Jun Gao, et al., "Magic3d: High-resolution text-to-3d content creation," in *CVPR*, 2023, pp. 300–309.

[3] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," *arXiv preprint arXiv:2309.03453*, 2023.

[4] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *ICCV*, 2023, pp. 9298–9309.

[5] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al., "Wonder3d: Single image to 3d using cross-domain diffusion," *arXiv preprint arXiv:2310.15008*, 2023.

[6] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang, "Mvdream: Multi-view diffusion for 3d generation," *arXiv preprint arXiv:2308.16512*, 2023.

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi, "Objaverse: A universe of annotated 3d objects," in *CVPR*, 2023, pp. 13142–13153.

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[9] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *CVPR*, 2023, pp. 22563–22575.

[10] Yatian Pang, Tanghui Jia, et al., "Envision3d: One image to 3d with anchor views interpolation," *arXiv preprint arXiv:2403.08902*, 2024.

[11] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu, "V3d: Video diffusion models are effective 3d generators," *arXiv preprint arXiv:2403.06738*, 2024.

[12] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani, "Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion," *arXiv preprint arXiv:2403.12008*, 2024.

[13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2553–2560.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.
[17] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
[18] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *NeurIPS*, vol. 36, 2024.
[19] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman, "Implicit geometric regularization for learning shapes," *arXiv preprint arXiv:2002.10099*, 2020.
[20] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang, "Sparseneus: Fast generalizable neural surface reconstruction from sparse views," in *European Conference on Computer Vision*. Springer, 2022, pp. 210–227.
[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.
[22] Heewoo Jun and Alex Nichol, "Shap-e: Generating conditional 3d implicit functions," *arXiv preprint arXiv:2305.02463*, 2023.
[23] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *NeurIPS*, vol. 36, 2024.
[24] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu, "Crm: Single image to 3d textured mesh with convolutional reconstruction model," *arXiv preprint arXiv:2403.05034*, 2024.
[25] Zhou Wang, Alan C Bovik, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

V. 附录

VI. 训练详情

我们从建立在 EDM 框架上的稳定视频扩散 (SVD) 模型开始。我们在 EDM 框架中采用预条件函数:

$$c_{skip}(\sigma) = \frac{\sigma_{data}^2}{\sigma^2 + \sigma_{data}^2} \tag{7}$$

$$c_{out}(\sigma) = \frac{\sigma\sigma_{data}}{\sqrt{\sigma^2 + \sigma_{data}^2}} \tag{8}$$

$$c_{in}(\sigma) = \frac{1}{\sqrt{\sigma^2 + \sigma_{data}^2}} \tag{9}$$

$$c_{noise}(\sigma) = \frac{1}{4} ln(\sigma). \tag{10}$$

我们还采用了噪声分布和加权函数:

$$\log\sigma \sim \mathcal{N}(P_{mean}, P_{std}^2) \tag{11}$$

$$\lambda(\sigma) = (1 + \sigma^2)\sigma^{-2} \tag{12}$$

在训练过程中，我们设定为 $\sigma_{\text{data}} = 1$ ，并逐渐将噪声分布向更高的水平偏移，这对于高质量视频生成是必要的。具体来说，从使用 $P_{\text{mean}} = 1.0$ 和 $P_{\text{std}} = 1.6$ 的 SVD 预训练配置开始，我们在分别 8,000、16,000 和 24,000 个全局步骤时将噪声参数调整为 $\{P_{\text{mean}}, P_{\text{std}}\} = \{1.8, 1.6\}$ 、$\{2.2, 1.8\}$ 和 $\{2.5, 2.0\}$ 。此外，与 Wonder3D [5] 的多阶段训练策略不同，我们在将 SVD 与 GTA 注意机制结合后采用单阶段训练方法。这确保了在整个训练过程中 RGB 和法线贴图之间的持续信息交换，从而最大化保留来自 SVD 的 3D 先验。我们的模型是在分辨率为 256×256 的渲染多视图数据集上进行训练的，使用学习率为 2×10^{-5} 并结合衰减率为 0.9999 的指数移动平均进行 AdamW 优化器，训练大约 30,000 步。

VII. GTA 模块的实现细节

为了说明所提出的 GTA 注意力机制，我们在算法 1 中详细介绍了基本 RGB 法线对齐注意力的实现。此外，我们在算法 2 中提供了详细的 GTA 融合视频变压器块的描述。

Algorithm 1 对齐注意力

Input: z // (nv 2 b) d c
$query, key, value \leftarrow W^q(z), W^k(z), W^v(z)$
// 分解 rgb 批处理和正常批处理
$key\_rgb, key\_norm \leftarrow$ torch.chunk(key)
$value\_rgb, value\_norm \leftarrow$ torch.chunk($value$)
// 在标记长度维度上连接 rgb 和 normal 潜在变量
$key \leftarrow$ torch.cat($[key\_rgb, key\_norm], dim = 1$)
$value \leftarrow$ torch.cat($[value\_rgb, value\_norm], dim = 1$)
$z \leftarrow attention(key, vaule, query)$
return z

Algorithm 2 GTA 注入视频变换器块

Input: z , *embeddings* for cross-attn
// 空间层
$z \leftarrow \text{ResBlock}(z)$
$z \leftarrow \text{SelfAttention}(z)$
// 帧对齐注意力
$z \leftarrow \text{AlignmentAttention}(z)$
$z \leftarrow \text{CrossAttention}(z)$
$z \leftarrow \text{Conv3D}(z)$
// 时间层
$z \leftarrow rearrange(z, (nv2b)chw \rightarrow (2bhw)nvc)$
$z \leftarrow \text{ResBlock}(z)$
$z \leftarrow \text{SelfAttention}(z)$
// 时间维度对齐注意力
$z \leftarrow \text{AlignmentAttention}(z)$
$z \leftarrow \text{CrossAttention}(z)$
$z \leftarrow rearrange(z, (2bhw)nvc \rightarrow (nv2b)chw)$
return z

为了评估我们模型的泛化能力，我们使用一个二维人工智能生成内容（AIGC）工具来创建自然环境中的图像，确保这些图像不包含在训练数据集中。如图 4 所示，NOVA3D 生成了 16 个视角的 RGB 图像和法线贴图，具有显著的跨视角一致性和连贯的整体形状，这展示了我们模型强大的泛化能力。

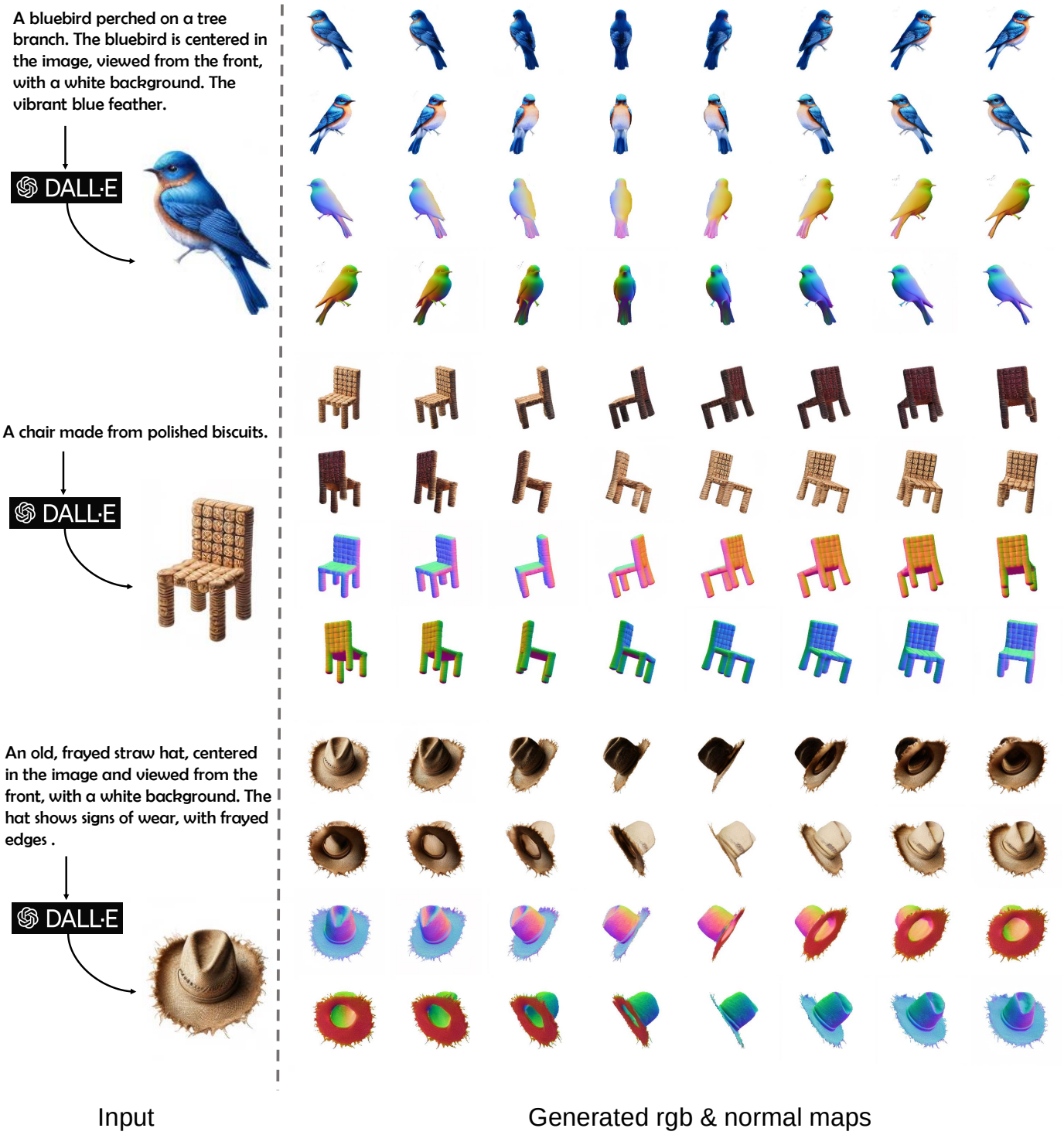

Fig. 4: NOVA3D 在生成图像和法线贴图上的定性结果是基于由现成的 AIGC 工具生成的野外图像而调整的。