Multi-View Silhouette and Depth Decomposition for High Resolution 3D Object Representation

Xinze LI

11 May, 2024

Background

Background

- 物体的三维形状是由无数物理元素的组合构成的，这些元素的尺度从**粗略的结构和拓扑到每个表面的微小纹理**都有所不同。
- 深度生成模型（到2018年）捕捉了它们的整体结构和粗糙纹理。这些第一代模型利用了**体素**表示，其随分辨率立方体地缩放，使得在典型硬件上训练仅限于64^3形状。
- 许多最近的论文开始提出更好的缩放的高分辨率三维形状表示，例如基于网格、点云或八叉树，但这些通常需要更困难的训练流程和定制的网络架构。**（都是古典的图形学方法，彼时 NeRF 和 3DGS 还没有出来，大佬们批判性接受）**
- NIPS2018

Motivation

- Inspired By: canonical views （规范视图）即：一个三维物体的形状可以通过来自多个视点的一组二维图像。
- DL方法用于2D图像识别和生成可以轻松地扩展到高分辨率。
- 提出：多视角表示(multi-view representation) 是否可以有效地应用到 DL 方法中？

Pre Knowledge: What is canonical views?

- **规范视图**是指从**多个不同视角观察物体时所得到的一组标准化的二维图像。**这些视角可以涵盖物体的各个方面，从而完整地捕捉物体的形状和结构。
- 新问题：什么是标准化的二维图像?
- 标准化的二维图像指的是经过一定处理或调整，使它们符合某种特定标准或规范的图像。在规范视图的情况下，这些二维图像通常被调整为具有相同的尺寸、分辨率和视角，以便它们能够在同一框架下被直接比较或组合。这种标准化可以包括调整图像的大小、对齐物体的位置、统一光照条件等。
- 利用许多二维正交投影来捕捉形状是规范视图的一种方法。（本文的方法）

Novelty

- 提出了一种新颖的深度形状解释方法，通过**在二维中将物体的规范视图建模为深度图来捕捉其结构**，称之为多视图分解（MVD）框架。
- 通过利用许多**二维正交投影**来捕捉形状，以这种方式表示的模型可以通过在二维空间中执行**语义超分辨率**来扩展到高分辨率，更高分辨率的深度图最终被合并为高分三维物体

Key Components

- Two synergistic deep networks 分解表示物体深度的任务:
  - 一个输出轮廓 - 捕捉粗略结构
  - 另一个产生深度的局部变化 - 捕捉细节
- 优势：允许轮廓预测形成锐利的边缘来解决通常在最小化重建错误时发生的模糊图像问题。

来张效果图

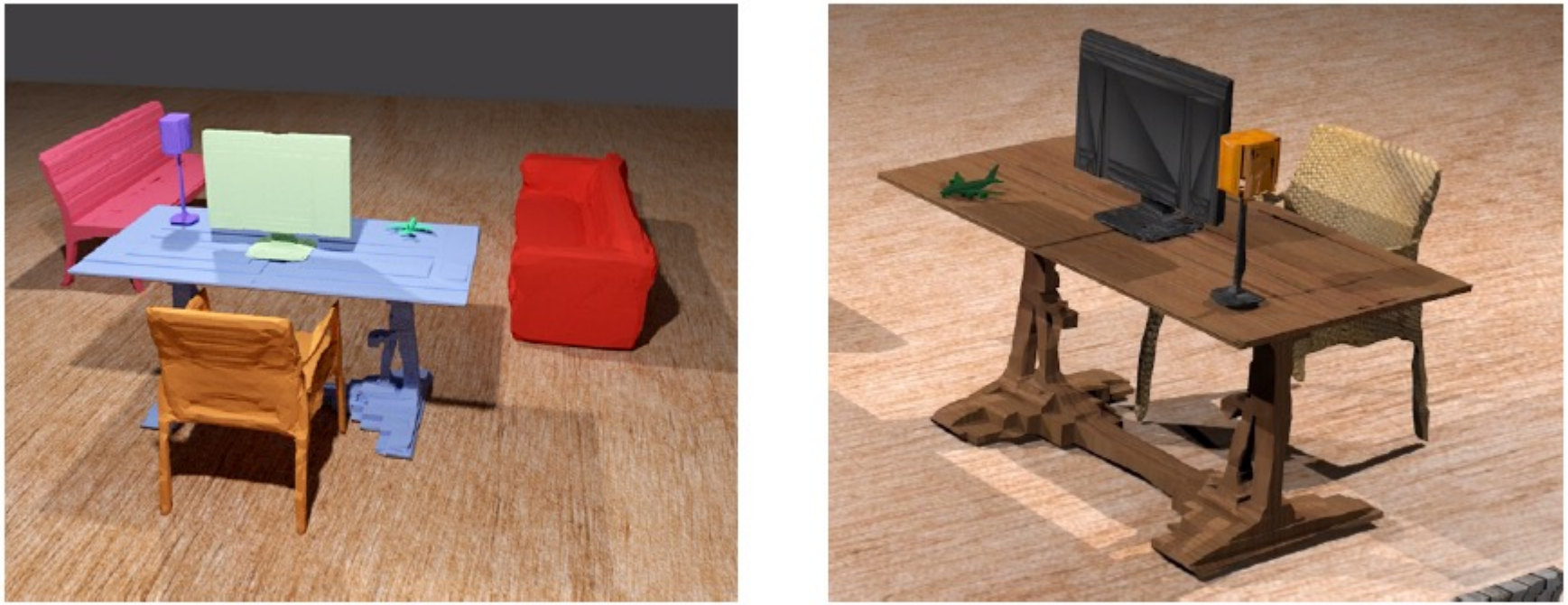
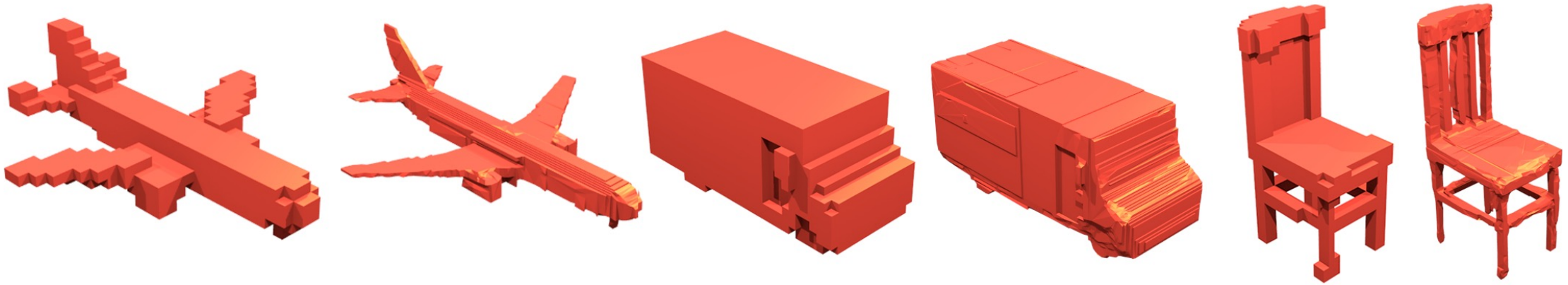


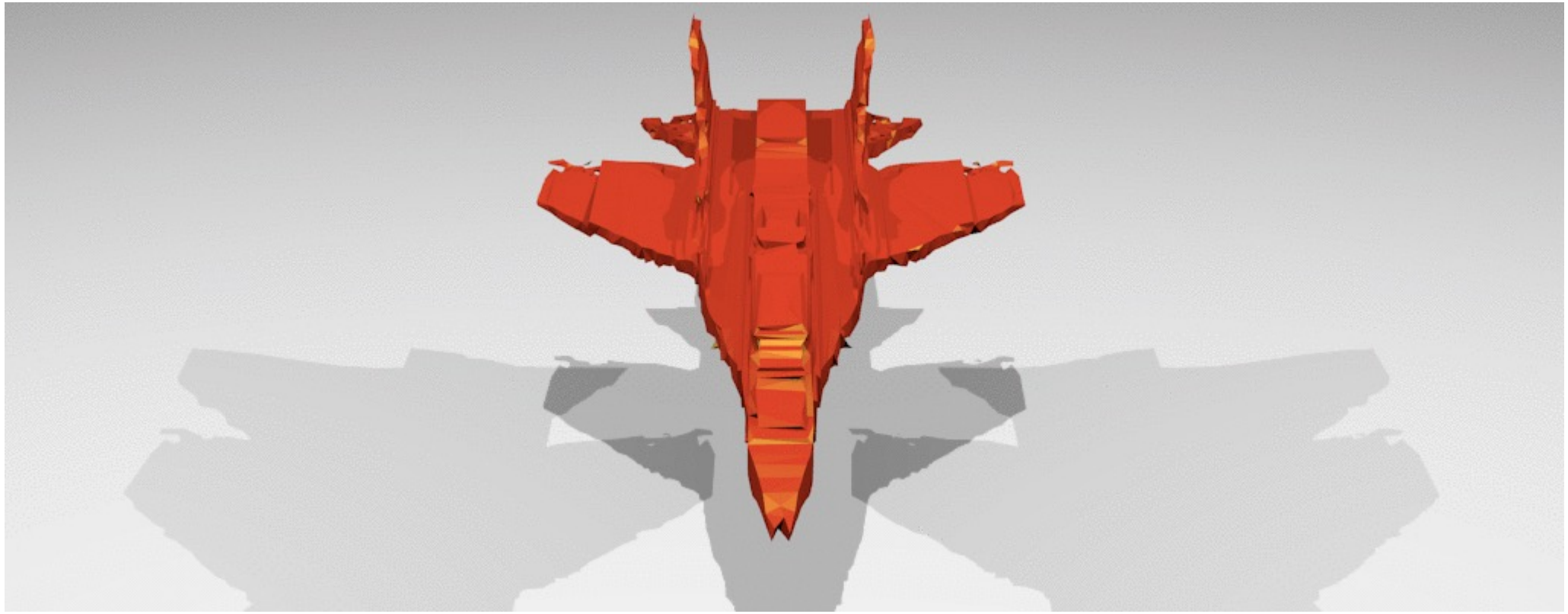
Figure 1: Scene created from objects reconstructed by our method from RGB images at 256^3 resolution. See the supplementary video for better viewing https://sites.google.com/site/mvdnips2018.

再来张



Example super-resolution results, when increasing from 32^3resolution to 256^3 resolution.

还有。就是喜欢看图



Related Work

平常是懒得讲的，但感觉最近自己提问别人的太细了，自己还是讲讲好，**免得被喷**

Deep Learning with 3D Data

- 从最初的使用**3D卷积**进行对象分类的架构开始。当应用于3D生成时，这些方法通常使用**自编码器网络 (Auto Encoder)**，其中解码器由3D反卷积层组成。该解码器接收3D形状的潜在表示，并在3D体素空间中的每个离散位置产生一个占用概率。这种方法已经**与生成对抗性方法 (GAN) 结合起来**，用于生成新颖的3D物体，但分辨率受到限制。
- 确实是老了，估计后面拿这些老方法改 Diffusion 的文章一大堆。

古典 2D Super-Resolution

- 2D图像的超分辨率是一个研究较多的问题。传统上，图像超分辨率使用字典式方法，将图像的补丁匹配到更高分辨率的对应物上。这方面的研究也扩展到了深度图超分辨率。现代的超分辨率方法建立在深度卷积网络和生成对抗网络之上，后者使用对抗性损失来想象RGB图像中的高分辨率细节。
- 最近的 2D 超分，CNN 架构的都已经 Old-Fashion （对于任何问题任何模型，当然，超大核或者可变核大小的除外）。利用生成式的都在弄 Diffusion。本人相中的是 U-Net 架构，不过都还没有用在 3D 上。估计是内存爆炸了。

现代 2D Super-Resolution with diffusion

Solving Diffusion ODEs with Optimal Boundary Conditions for Better Image Super-Resolution (Ma et al., 2023)	It details a method to improve super-resolution image quality by solving diffusion ODEs with optimal boundary conditions, effectively reducing the randomness in the image generation process.
Waving Goodbye to Low-Res: A Diffusion-Wavelet Approach for Image Super-Resolution (Not listed, 2023)	Combines denoising diffusion probabilistic models with discrete wavelet transformations to enhance super-resolution, outperforming existing diffusion-based methods.
Spatio-Angular Convolutions for Super-resolution in Diffusion MRI (Lyon et al., 2023)	Presents a novel convolutional framework for enhancing the resolution of diffusion MRI images, showcasing a significant reduction in required parameters and improved performance in clinical analyses.

- ICLR 2024
- No submission reported
- No submission reported (But NIPS template)

现代 2D Super-Resolution with U-Net

S.No.	Paper Title (Author, Year)	Insight from Abstract	Citations
1.	Super-Resolution of Brain MRI via U-Net Architecture (Kalluvila, 2023)	Demonstrates the effectiveness of U-Net for super-resolving MRI scans, emphasizing its potential to reduce costs and enhance diagnostic capabilities.	Not available
2.	MLE^{2}A^{2}U-Net: Image Super-Resolution via Multi-Level Edge Embedding and Aggregated Attentive Upsampler Network (Mehta et al., 2023)	Introduces an innovative U-Net architecture with attention mechanisms for improving the balance between contextual and spatial details in image super-resolution.	1
3.	UArch: A Super-Resolution Processor With Heterogeneous Triple-Core Architecture for Workloads of U-Net Networks (Duan et al., 2023)	Proposes a novel processor architecture designed to optimize U-Net performance for medical imaging, enhancing throughput and reducing latency.	Not available
4.	D^{2}UNet: Dual Decoder U-Net for Seismic Image Super-Resolution Reconstruction (Not listed, 2023)	Describes a dual decoder U-Net that significantly improves the resolution and fidelity of seismic images by focusing on edge and detail reconstruction.	1

古典 Efficient 3D Representations

- 八叉树方法 (Octree):利用非均匀离散化的体素空间，通过在局部基于形状适应离散化级别，以高效地捕捉3D物体。
- 次表面预测（HSP）是一种八叉树式方法，它将体素空间划分为自由空间、占用空间和边界空间。对象以不同的分辨率尺度生成，其中占用空间以非常粗糙的分辨率生成，而边界空间以非常细微的分辨率生成。
- 八叉树生成网络（OGN）使用卷积网络直接在八叉树上操作，而不是在体素空间中。这些方法仅展示了分辨率为256^3的新型生成结果。我们的方法在这种分辨率下实现了更高的准确性，并且能够高效地生成分辨率达到512^3的新对象。

现代 Efficient 3D Representations

- 3D Gaussian Splatting
- NeRF
- …

Method

他们说的自己的 Novelty

- 该方法利用了六个轴对齐的正交深度图（ODM），以在不直接与体素交互的情况下，有效地将3D对象缩放到高分辨率。
- 正交深度图（Orthographic Depth Map，ODM）是一种用于表示物体表面深度的图像，其中每个像素的值表示物体沿着观察方向的表面距离。在正交深度图中，物体的深度信息被投影到一个平面上，这个平面与观察方向垂直，因此被称为“正交”。这意味着在正交深度图中，每个像素的值代表了沿着观察方向上物体的深度，而不受视角和透视效果的影响。
- 来看看怎么个事儿

Overview

- 因为是 Multi-view，模型接下来对当前的 Multi-view 进行操作
- 这对网络将超分辨率任务分解为两个部分：一部分是从低分辨率 ODM（六轴对齐的正交深度图）预测轮廓，另一部分是预测相对深度。

Architecture Overview

Low-resolution Object Reconstruction Result as input (see appendix B). Original data is RGB image.

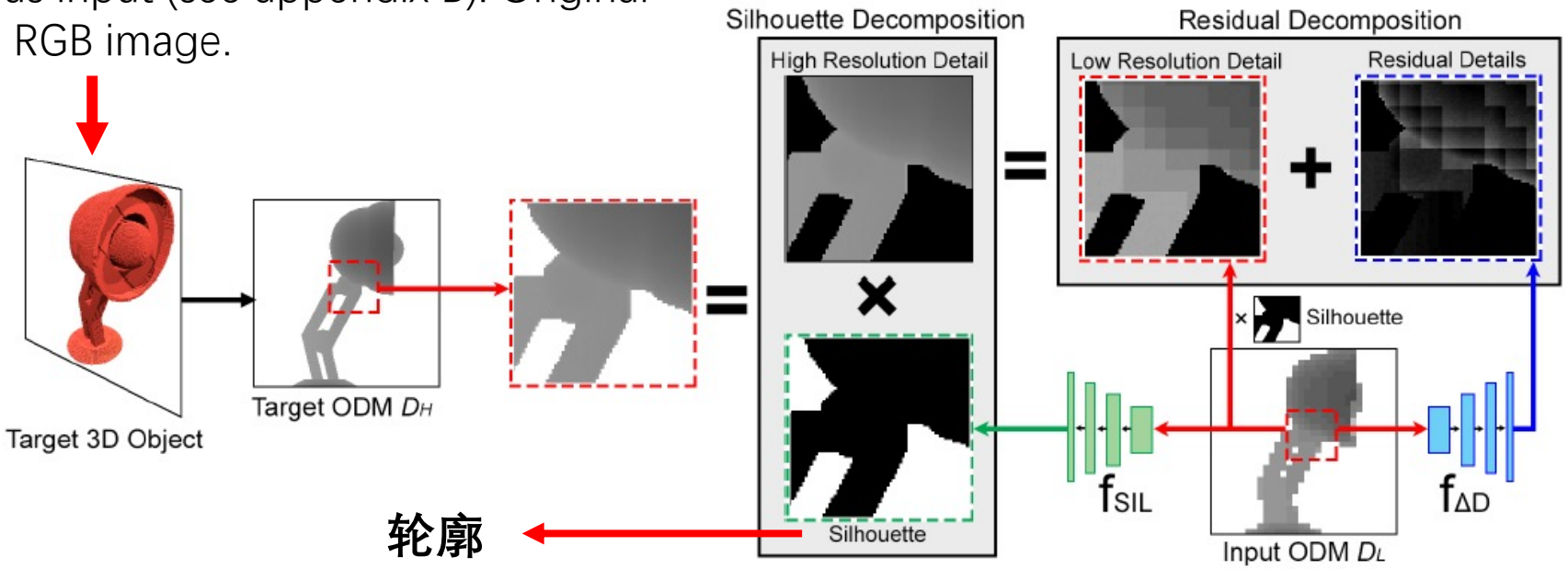


Figure 3: Our Multi-View Decomposition framework (MVD). Each ODM prediction task can be decomposed into a silhouette and detail prediction. We further simplify the detail prediction task by encoding only the residual details (change from the low resolution input), masked by the ground truth silhouette.

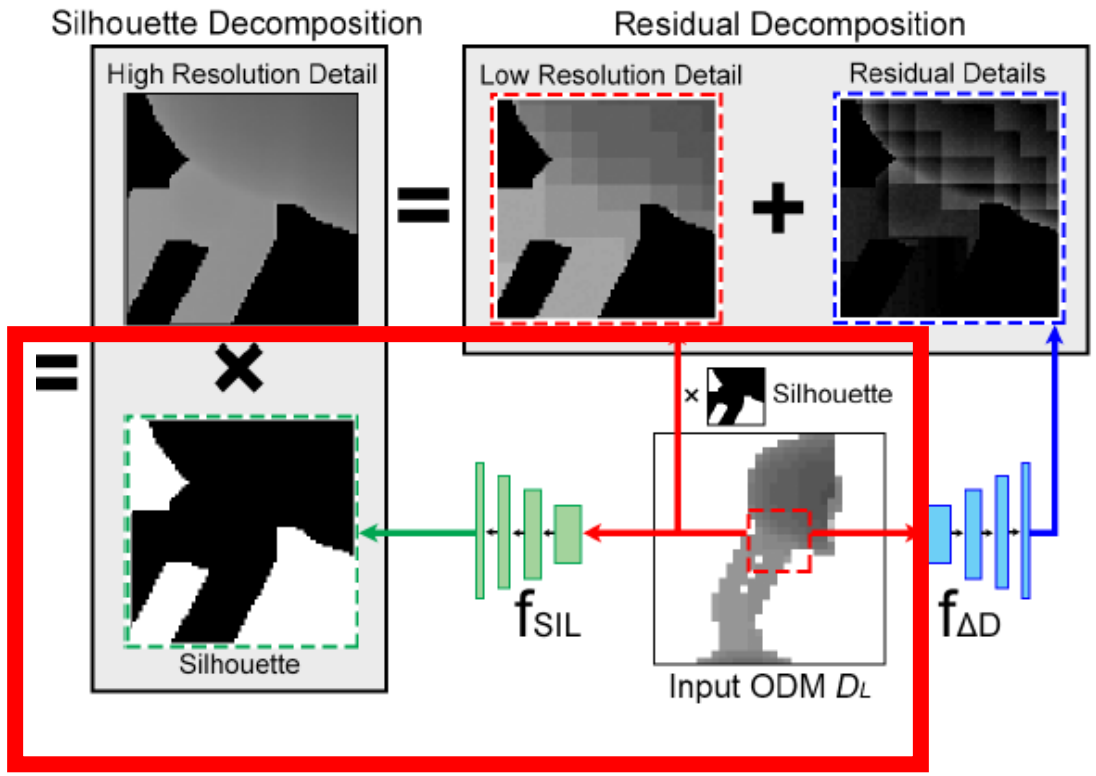
Orthographic Depth Map Super-Resolution

- 首先获得低分辨率3D对象的六个主视图的正交深度图 ODM。
- 这种投影可以通过对齐于轴的3D数组通过 z-clipping 快速且轻松地计算出来。然后直接在这些正交深度图上进行超分辨率处理，然后将其映射到低分辨率对象上以生成高分辨率对象。
- Z-clipping:用于将物体或场景从三维空间投影到二维屏幕上时，对超出指定深度范围的部分进行裁剪或剪除。

一个 challenge

- 通过一组深度图表示对象引入了一个具有挑战性的学习问题，需要**深度上的局部和全局一致性**。
- 此外，最小化均方误差会导致模糊的图像，缺乏锐利的边缘。(原因是因为：深度图需要是双峰的，具有深度的大变化来创建结构，而深度的小变化则用于创建纹理和细节。深度图为什么是双峰的呢？这是因为在一个场景中，物体的深度分布通常可以分为背景和物体，背景景深，物体景浅)
- 他们的 proposal:将学习问题分解为分别预测轮廓和深度图(预测整体形状与细节分开处理)

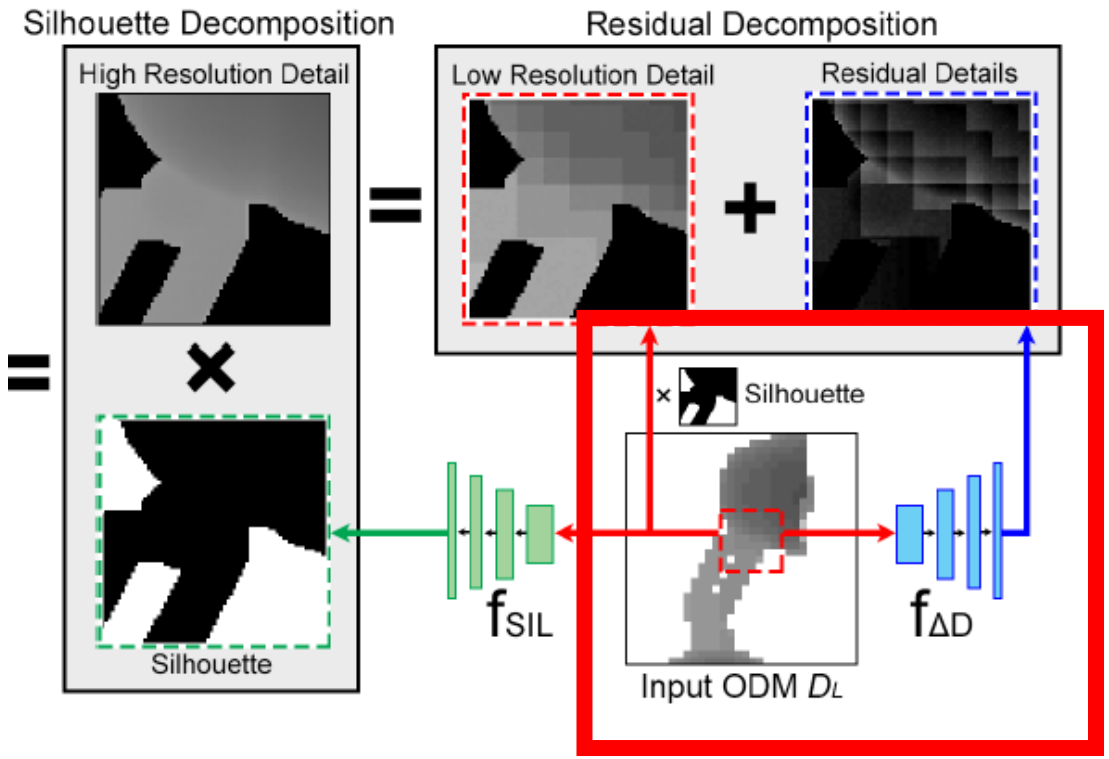
Architecture



- D_L 当前 view 的 ODM
- 用于预测高分辨率轮廓的深度卷积网络 f_{SIL}
- 网络输出每个像素被占用的概率

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \| f_{\text{SIL}}(D_L^{(i)}; \theta) - \mathbb{1}_{D_H^{(i)} \neq 0}(D_H^{(i)}) \|_2,$$

Architecture



- D_L 又被输入到第二个网络 $f_{\Delta D}$ 中，其最终输出经过 Sigmoid 函数处理，产生对于在固定范围 r 内的 ODM 变化的估计。
- 这个输出被添加到低分辨率深度图中，以产生我们对于受限高分辨率深度图 C_H 的预测。

$$C_H = r\sigma(f_{\Delta \mathrm{D}}(D_L;\phi)) + g(D_L),$$

where $g(\cdot)$ denotes up-sampling using nearest neighbor interpolation.

Up-sampling using nearest neighbor interpolation

- 选择输入图像中最近的像素值来赋予输出图像中的像素值。这种方法计算效率高，但在处理具有显著细节的图像时，可能会引入块状伪影。
- 邻近插值通过将每个新像素赋值为其最近邻像素的值来实现图像的放大。虽然这种方法简单高效，但在处理细节丰富的图像时可能会出现较明显的块状效果。
- 对以上说法不负责任，以上说法来自 GPT-4o

Training Loss

- 最小化预测值与GT高分辨率深度图 $\mathbf{D}_H$ 之间的 MSE 来训练网络 $f_{\Delta D}$
- 在训练期间，我们仅使用GT轮廓对输出进行 mask，以便$f_{\Delta D}$有效地关注细节。

$$\mathcal{L}(\phi) = \sum_{i=1}^{N} \|(C_H^{(i)} \circ \mathbb{1}_{D_H^{(i)}(j,k)\neq 0}(D_H^{(i)})) - D_H^{(i)}\|_2 + \lambda V(C_H^{(i)}), \tag{3}$$

where $\circ$ is the Hadamard product. The total variation penalty is used as an edge-preserving denoising which smooths out irregularities in the output.

Result

- 受限深度图和轮廓网络的输出然后结合起来，产生对高分辨率ODM的完整预测。

$$\hat{D}_H = C_H \circ f_{\mathrm{SIL}}(D_L; \theta). \tag{4}$$

$\hat{D}_H$ denotes our predicted high resolution ODM which can then be mapped back onto the original low resolution object by model carving to produce a high resolution object. Each of the 6 high resolution ODMS are predicted using the same 2 network models, with the side information for each passed using a forth channel in the corresponding low resolution ODM passed to the networks.

- 不是很懂什么是 forth channel。哪位dalao能指点一下。

3D Model Carving （过于细节）

- 感觉是最近看的所有论文中把方法说的最细致的一篇。老实人没跑的了。
- 现在的步骤是：如何把六个ODM与低分辨率对象通过神经网络以后结合起来形成一个高分辨率对象。
- 可以理解为是：怎么处理全局一致性吧 @Huanrong LIU
- **以下内容都没有图，也没有代码看，非常抽象。**

3D Model Carving

- 首先，通过自适应平均滤波器进一步平滑上采样的ODM，该滤波器仅考虑小半径内的相邻像素。为了保留边缘，只包括中心像素值范围内的相邻像素。

3D Model Carving

- Model Carving 首先通过最近邻插值将低分辨率模型上采样到所需分辨率。
- 之后，对于当前ODM的模型输出结果，$\widehat{D}_H$ 来确定新的表面。
- 该过程分为两步：1. 结构 carving，对应的是轮廓的预测 f_{SIL} 2. 细节 carving，对应的是深度预测 C_H

3D Model Carving

- 对于结构 carving: 对于每个预测的 f_{SIL}，如果一个坐标上被预测为没有像素占用，那么所有垂直于该坐标的体素都会被标记以便被移除。
- 如果有两张 ODM 对应的同一个体素都要被移除，那么该体素才会被真的移除。
- 由于六个ODM观察到的表面区域存在大量重叠，因此要求轮廓的一致性以维护对象的结构。

3D Model Carving

- 对于细节 carving:上述过程同样适用。
- 然而，我们不需要在深度图的预测中需要两张ODM的一致。这是因为，与轮廓不同，深度图可能会在对象表面上引起或加深凹陷，这些凹陷可能在任何其他面上都不可见。要求深度图之间达成一致意见会消除它们影响这些凹陷的能力。因此，进行细节 carving 只需要移除每个ODM的每个坐标的垂直体素，直到达到预测的深度。

Result

Super-Resolution IoU Results against nearest neighbor baseline

Category	Baseline	Depth Variation ($f_{\Delta D}$)	Silhouette (f_{SIL})	MVD (Both)
Car	73.2	80.6	86.9	**89.9**
Chair	54.9	58.5	67.3	**68.5**
Plane	39.9	50.5	70.2	**71.1**

Table 1: Super-Resolution IoU Results against nearest neighbor baseline and an ablation over individual networks at 256^3 from 32^3 input.

3D Object Reconstruction IoU at 256^3.

Category	AE	HSP [9]	MVD (Ours)
Car	55.2	70.1	**72.7**
Chair	36.4	37.8	**40.1**
Plane	28.9	56.1	**56.4**

(a) $Data_{HSP}$

Category	AE	OGN [44]	MVD (Ours)
Car	68.1	78.2	**80.7**
Chair	37.6	-	**43.3**
Plane	34.6	-	**58.9**

(b) $Data_{3D-R2N2}$

Table 2: 3D Object Reconstruction IoU at 256^3. Cells with a dash - indicate that the corresponding result was not reported by the original author.

3D object reconstruction surface sampling F1 scores

Category	3D-R2N2 [3]	PSG [6]	N3MR [14]	Pixel2Mesh [45]	MVD (Ours)
Plane	41.46	68.20	62.10	71.12	**87.34**
Bench	34.09	49.29	35.84	57.57	**69.92**
Cabinet	49.88	39.93	21.04	60.39	**65.87**
Car	37.80	50.70	36.66	**67.86**	**67.69**
Chair	40.22	41.60	30.25	54.38	**62.57**
Monitor	34.38	40.53	28.77	51.39	**57.48**
Lamp	32.35	41.40	27.97	**48.15**	**48.37**
Speaker	45.30	32.61	19.46	48.84	**53.88**
Firearm	28.34	69.96	52.22	73.20	**78.12**
Couch	40.01	36.59	25.04	51.90	**53.66**
Table	43.79	53.44	28.40	66.30	**68.06**
Cellphone	42.31	55.95	27.96	70.24	**86.00**
Watercraft	37.10	51.28	43.71	55.12	**64.07**
Mean	39.01	48.58	33.80	59.72	**66.39**

Table 3: 3D object reconstruction surface sampling F1 scores.

来点图: Super-resolution rendering results at 512^3

(a) (b)

Figure 4: Super-resolution rendering results. Each set shows, from left to right, the low resolution input and the results of MVD at 512^3. Sets in (b) additionally show the ground-truth 512^3 objects on the far right.

来点图: Super-resolution rendering results at 256^3

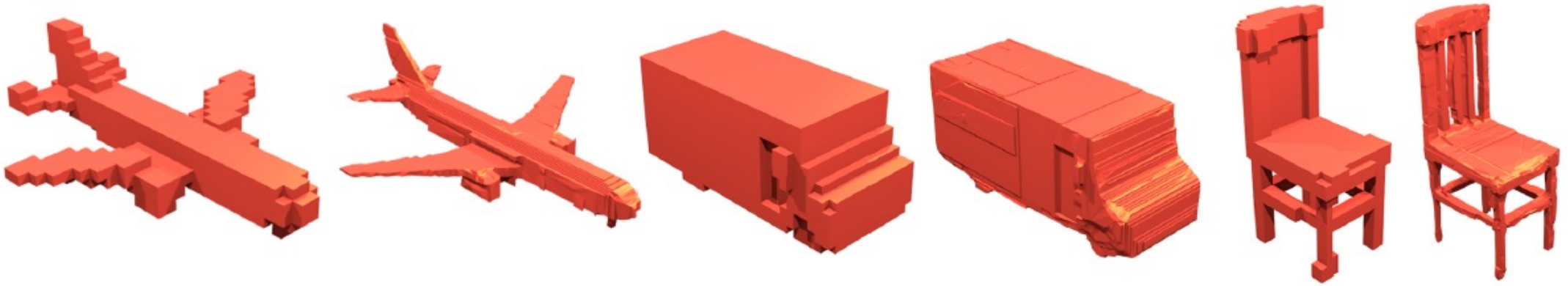


Figure 5: Super-resolution rendering results. Each pair shows the low resolution input (left) and the results of MVD at 256^3 resolution (right).

来点图: 3D Reconstruction result

Figure 6: 3D object reconstruction 256^3 rendering results from our method, MVD (bottom), of the 13 classes from ShapeNet, by interpreting 2D image input (top).

Conclusion

Conclusion

- 预测高分辨率对象结构时应用多视图表示法
- 多视图分解框架

Evaluation

- 2024 年来看这篇文章，感觉已经很古早了。~~配图都没有现在好看，故事也没有现在的精彩。~~
- 个人认为这篇最大的亮点在于，提出把这个轮廓和细节分成两个网络来做超分辨率。且其基础理论在于场景 depth map 为双峰函数，这样子相当于分别拟合两个峰。
- 此外，在生成重建结果的时候，也按照轮廓-细节分别处理。这是一个不错的做法。对于 3DGS 也有一定的可移植性。