北京大學

本科生毕业论文

题目：三维模型最优二维视图生成方法研究

Synthesizing best 2D views of 3D models

姓　　名：黄道吉

学　　号：1600017857

院　　系：元培学院

本科专业：计算机科学与技术

指导老师：连宙辉

二〇二〇　年　五　月

摘要

生成三维模型的最优视图任务要求给定一个三维模型，我们能够选取出合适的视角，并且在这个视角下渲染出具有真实性的图片。随着三维建模方法的不断完善，三维模型的使用在近几年正变的越来越广泛。大规模三维模型数据库的出现更加便捷了有关三维模型的研究的进展。而三维模型不同于二维图片的特性使得它需要经过渲染才能显示在屏幕上。生成符合人类感知的三维模型的最优视图，将大大方便三维模型库的检索。近年来神经网络在二维视图生成工作的进展飞速，这也使得借助神经网络来生成三维模型的最优视图成为可能，尤其是生成对抗网络和变分自编码器在条件生成和非条件生成领域取得了很好的效果。本文回顾了以往在这个方向上研究者的工作，包括最优视角选择和新视角生成的算法，并提出了新的生成三维模型视图的方法。我们的方法首先能够根据一张导引图片提供的材质信息来渲染给定的三维模型在某一个视角下的视图，也能够通过在隐空间中采样来无条件的生成多样的三维模型的视图。定性和定量的实验结果表明，我们的新算法能够产生更加准确、更具真实性的三维模型图像，也在选择视角方面相对其他算法有自己的优势。最后，我们总结了本文的成果并展望了未来的工作。

关键词: 最优视角选择 新视角生成 材质迁移

Abstract

Synthesizing the best view of 3D models aims to choose to best viewpoint and rendering a realistic image given a 3D model. With the development of 3D modeling, the use of 3D models has been more and more prevalent. The emergent of 3D databases further advocates the development of 3D model related research. Unlike 2D images, 3D models have to be rendered to be displayed on a screen. Synthesizing the best view of 3D models in line with human perception will contribute to the retrieval of 3D databases. Thanks to the rapid development of neural networks' application on synthesizing 2D images, it has been possible in recent years to synthesize the best view of 3D models using neural networks, especially GAN and VAE, which yields great results in conditional and unconditional image synthesis. This paper reviews past research on this field, including best view synthesis and novel view synthesis, and propose a novel method to synthesize views of 3D models. First, our model can render an image of a given 3D model under a specific viewpoint while being consistent in style to a reference image. Second, our method can render various views of 3D models by sampling in latent space. Our new method is proved to be able to yield more accurate and realistic images through qualitative and quantitative experiments and outperforms previous best view selection methods. Finally, we summarize our contribution and proposed future work on this topic.

Key Words: best view selection novel view synthesis style transfer

全文目录

第一章 引言

1. 研究背景

三维模型是图形学和计算机视觉方向的研究重点。近年来，三维模型的应用变得越来越广泛，从游戏界和工业界的 CAD 模型，到前沿领域的自动驾驶，使用三维模型正大大便利着业界。RGB-D 传感器的应用也使得产生三维模型更加容易。在学术界，三维模型也有着广泛的应用：三维模型的分割 ([1,2])、重建 ([3,4])，以及利用三维模型强化对图片的理解 ([5])。这些因素都催生了大规模三维模型库的产生和广泛使用 (如 Shapenet [6], Pascal3D+ [7], ModelNet [8])。

在如此多的精力投入利用数据集解决问题的同时，相对少的精力投入到利用数据驱动的方法方便数据集的可视化和检索上。不同于二维图片便于观看、容易生成缩略图的特性，三维模型在不同视角下会产生不同的视图，并且需要材质信息才能渲染出一张具有真实性的图片。这使得检索三维模型的数据库是一件费事的工作。ShapeNet 数据集 [6] 将每一个类别的模型对齐到同一个朝向，并在固定的方向渲染了 8 张缩略图，对大部分模型提供了材质信息，ModelNet 数据集 [8] 只提供了三维模型而没有对应的材质信息，这些方法并不能提供一个便捷的检索三维模型的方案。现有的处理三维模型的软件 (如 MeshLab [9])，提供用户一个拖拽视角的界面，让用户寻找最好的视角。如果能设计出生成最优视图的算法，将会便利检索三维模型数据库。

我们认为生成三维模型的最优视图至少包括两个部分，一个部分是选定最优的视角，另一部分是在这个选定的视角下渲染出带有材质的二位视图。第一个部分以往工作主要从图形学入手，通过在三维模型的顶点或是在二维视图上定义信息 (熵)，取熵最大的视角作为最优视角。渲染材质的工作则集中利用了基于神经网络的生成模型，将材质生成问题定义为有条件的图片到图片翻译的问题。我们借鉴了这两方面方法的核心思想，并提出了新的生成三维模型最优视图的算法。

2. 相关工作

2.1 视角选择

三维模型的最优视角选择任务旨在对给定的三维模型给出符合人类认知的最优的视角。这并不是一个良定义的问题，以往的研究方向往往采用在三维顶点或是二维像素上定义某种函数而将其转换为最优化问题。传统上认为最佳视角是包含最多信息的视角，不同的方法对信息的定义各不相同。在三维模型的二维视图上定义信息的文章主要包括：视角熵 [10]，曲率熵 [11]，轮廓熵 [11]，在不同的视角的投影中取信息最大的投影作为最佳视角。[12] 文中对比了几种基于几何学的方法的结果和人为标注的最优视角的差别，文章得出 MeshSaliency [13] 和视角熵 [10] 的方法是效果最好的传统方法。Mesh Saliency [13] 通过在每一个三维定点定义与曲率有关的显著性，并将可见的显著性加和最大的视角定义为最佳视角。文章更加提出了一种在视角空间中类似梯度下降的方法寻找最优视角的方法，而不需在视角空间中方格搜索 (grid search) 最优视角。视角熵 [10] 的方法关注二维投影中可见的每一个三维面片的投影面积，并将投影面积构成的分布的熵最大的视角作为最优的视角。我们认为这些方法有时并不会产生令人满意的效果。他们破坏了同一类三维模型共享同一个最佳视角的规则，并且对三维模型的建模方式很敏感。本研究首先复现了经典的传统方法，在以后的行文中，采用 Mesh Saliency 和视角熵作为传统方法的代表，和我们提出的方法作比较。

2.2 视角生成

新视角生成 (novel view synthesis) 旨在给定一个三维模型在一个或多个视角下的视图来生成新视角下的视图。因为不同视角下可见的像素不同，这个任务本质上是一个非良定义的问题，而需要足够强的先验知识和正则化约束来得到可接受的结果。以往解决新视角生成任务的方法大致可以分为两大类：基于几何学的和基于学习的方法。几何学的方法能够从输入图片显式的估计三维模型的结构和材质信息。Multi-view stereo [14] 方法可以通过多个视角的输入图片直接重构出三维模型。Flynn et al. [15] 提出的深度神经网络能够在不同视角的图片中进行插值。几何学的方法主要缺点是作为训练数据的三维模型难以获得，并且缺失的像素会导致错误的破洞填补 (hole-filling)。基于学习的方法将新视角生成看作

图片生成任务，或采取预测从原图片到目标图片的流 [16–18] 的方式，或是采用某种正则化后直接生成每一个像素 [19–22]。不同的方法针对它非良定义的特性使用了不同的正则化方法，如感知损失函数 [18]，生成对抗网络的损失函数 [20] 和三维信息 [21] 的方式。本文提出的模型可以看作 [21] 的进一步拓展，将文中的三维信息进一步拆分为材质和内容信息，从而能够应用到材质迁移任务上。本文的方法在应用到新视角生成时，与流方法和像素生成的经典方法比较均有定性和定量结果的优势。

2.3 材质迁移

材质迁移旨在给定一张内容图片和一张材质图片 (或材质属性)，生成一张具有前者内容和后者的材质信息的图片。一种通用的思路是从输入图片中编码出表征内容和表征材质的向量，如 [23–26]。囿于成对训练数据缺失，材质迁移方法无法使用有监督的一范数或二范数损失函数，而需要采取独特的方法训练内容和材质的分离：如认为内容与材质信息分出 VGG 的不同层中 [27]，最小化循环损失 (cycle loss) [28] 或增加独特的判别器判定材质信息是否一致 [29, 30]。受 AdaIN [31] 启发，另一种融合材质和内容的方法是从内容或材质信息中回归参数来调整神经网络中间层激活量的大小和偏移，如 [32, 33]。在三维领域，将一个三维模型的材质迁移至另一三维模型或直接生成三维模型的材质，可以通过直接在三维空间中生成每一个点或面片的颜色 [34]，也可以将三维模型投影至二维空间进行，如 VON [35]。前者的方法生成结果较为模糊，并且因渲染过程不可导，需要采用近似渲染方法和可微分的渲染器 [36]。而后者能够借用图像领域材质迁移的成熟方法，取得更好的效果，并且作为这些方法中常用的深度图，也能够很好的在二维空间表征三维的信息。因此本文采用类似 VON 的模式，将三维模型渲染为深度图后，在深度图上渲染模型的材质信息。这个材质信息可以来自材质图片，也可以是真实图片，或者是随机采样的隐变量。定性和定量的结果表明我们的方法在生成图片质量上又一定的优势。

3. 研究内容

本文研究生成三维模型的最优二维视图的任务。它要求对给定的三维模型我们能够生成符合人类观感的图片，这包括选取出一个合适的视角和生成具有真实性的材质两部分，其中材质可以来自参考图片也可以非条件的生成。我们回

顾了在这个任务上前人的工作，综述并且实现了在视角选择和新视角生成领域经典的算法。在前人工作的基础上，我们提出了新的模型，将以往方法中的隐空间一分为二，分别编码三维模型的与视角无关的内容和材质信息，利用重构损失和对抗损失函数训练隐变量到图像空间的映射。我们认为最优视角是模型包含最多信息的视角，这种信息可以通过三维模型的某一视角下的图片重构不同视角下视图的精确度来衡量。在实验结果上，我们将我们的结果和复现的经典方法作比较，在定性结果和定量结果上，均体现出了优势。

4. 本文结构

第一章引言介绍了本文的研究背景，在回顾了相关工作的基础上阐述了本文的研究内容。第二章将会介绍本文的实验原理，包括生成对抗网络 (GAN) 和变分自编码器 (VAE) 的推导，以及应用在本文的情境下的公式。其后，第三章将介绍本文提出的模型结构。我们的方法和其他方法的结果比较将放在第四章中。最后，我们在第五章总结本文的工作，并且展望未来可能的进展方向。

第二章 实验原理

本章会详细论证我们的方法的统计原理。在介绍了我们方法中主要使用的统计模型：变分自编码器 (VAE) 和生成对抗网络 (GAN) 之后，我们将推导出本文的方法的统计基础，为下一章介绍我们的方法的具体实现做好铺垫。

生成模型旨在学习一个分布 $X \sim P(X)$，其中 X 可能是来自一个高维空间的数据点，例如我们实验的图像数据来自 $128 \times 128 \times 3$ 维的分布。根据实际应用的需要，生成模型更关心的是从 $P(X)$ 中采样，而不是数值的学习到对应于每一个 X 取值的 $P(X)$，后者并不一定对前者有帮助。变分自编码器和生成对抗网络是成功的生成模型的两个例子，他们成功的利用了神经网络来拟合 $P(X)$，并且对数据分布并没有很强的先验假设，因而成为广泛应用的生成模型。

1. 变分自编码器

变分自编码器 (VAE) 主要可以用来无监督的学习复杂的数据分布。通过将直接采样 X 的任务分解为先采样一个隐变量 z，再通过学习一个映射将 z 映射到 X 的空间中，VAE 将一个复杂的采样任务分解成一个从隐空间采样和一个学习映射的任务。这样做的优点有

- 因为 z 从属的空间往往是简单的空间，例如大多数方法中使用的标准正态分布，隐空间中的采样任务很容易实现
- 映射往往用神经网络来实现，这样容易通过随机梯度下降的方法优化，并且计算量也不会十分巨大

公式化的话，VAE 期望优化训练集中的每一个数据点的概率值，$P(X) = \int P(X|z;\theta)P(z)dz$，其中由 z 生成 X 的分布是由 θ 参数化的神经网络。为方便后续推导，我们假定 $P(X|z;\theta) = N(X|f(z;\theta), \sigma^2 I)$。图 1 给出了 VAE 的图模型，对每一个训练集中的数据点 X，由 θ 参数化的神经网络将一个从正太分布中采样出的隐变量 z 映射到 X。如果直接对上式采样，将意味着很大的计算负担，并且对于大多数的 z，$P(X|z)$ 都接近于 0，很多样本是无意义的。因此，变分自编

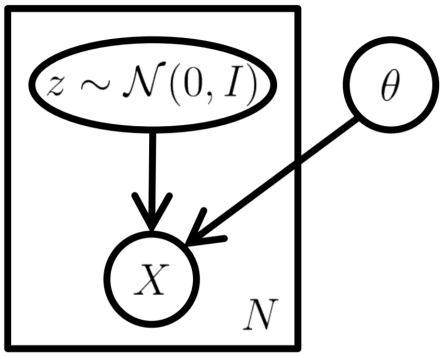

图 1 VAE 的图模型

码器需要另一个函数 $Q(z|X)$ 来拟合真实的 $P(z|X)$ 分布，这也正是名字中变分一词的来历。其中 Q 分布通常也假定为正态分布，由另一个模型估计出来。这样，通过观察下式

$$\begin{aligned} KL(Q(z|X)||P(z|X)) &= E_{z\sim Q}[\log Q(z|X) - \log P(z|X)] \\ &= E_{z\sim Q}[\log Q(z|X) - \log P(X|z) - \log P(z)] + \log P(X) \end{aligned} \tag{2.1}$$

我们移项得到

$$\begin{aligned} \log P(X) &= E_{z\sim Q}[\log P(X|z)] - KL(Q(z|X)|P(z)) + KL(Q(z|X)||P(z|X)) \\ &\geq E_{z\sim Q}[\log P(X|z)] - KL(Q(z|X)|P(z)) \end{aligned} \tag{2.2}$$

上式左边正是我们想要优化的：最大化训练集的概率。而上式右边两项又可以通过神经网络求出：两项分别代表重构图片的概率，在高斯假定的情况下近似为 L2 误差，和两个正态分布之间的 KL 散度。我们通过优化右边这个变分下界，来近似的最大化训练集的概率。

从表示学习 (representation learning) 的角度看，VAE 学习到了高维空间图像的一个低维表示。在表示学习中，一个重要的任务是学习到一个分离 (disentangle) 的表示，这种表示既存储了高维空间向量的全部信息，本身又有很强的可解释性。例如对于手写字符图像任务，一种可行的分离表示可以是 n 维隐变量的第一个维度表示字符的类别，第二/三个表示字符的斜度/粗细，每一个维度表征数据集中变化的一个因子（factor of variation）。这种表示意味着我们可以通过调整一个维度来调整生成图像的某一个特征，对风格迁移任务有很大的帮助。在传统的 VAE 基础之上，研究者提出了不同的 VAE 变种来提升 VAE 编码的分离能力，

如 [37,38]。他们可以看作在传统的 VAE 基础上加入 $\int Q(z|X)P(X)dX$，用不同的方式实现这个后验分布，优化模型的分离能力。

上述的分离学习方法专注于在隐变量的每一个维度都分离的控制不同的变化因子。在细粒度的分离学习之外，我们也可以将 VAE 的隐变量分成几组，每组控制到不同的特征。[39] 文认为图像是由形状 (shape) 和外表 (appearance) 共同控制生成，文中提出了一个条件的 U-Net，用形状和 RGB 图片生成新的图片。[26] 文用无监督的方法分离图像中的结构和风格信息。本文同样以 VAE 为理论基础，专注于分离图片中几何信息和风格信息。我们在理论推导方面受到了上两文的启发，但本文选取了深度图表示结构信息，而不是隐变量 [39] 或关键点 [26]，另外在模型实现方面也和上两文有所不同。此外，应对材质迁移后的图片训练集中无监督的情况，我们引入了生成对抗网络来优化生成结果，这些在本章的下一节中会详细论述。

2. 生成对抗网络

生成对抗网络（GAN）是另一种常用的生成模型，与 VAE 不同，它并不直接估计数据的概率分布，而是间接的近似 $P(X)$。GAN 的模型分成两个部分：生成器和判别器，前者能采样出和训练集数据同分布的样本，而判别器能够判别一个给定样本是否来自于训练集的分布，二者通常是以神经网络参数化的模型。GAN 的训练通过上述两方的博弈，最终得到一个纳什均衡，即生成器不能通过微调参数生成更加以假乱真的图片，判别器也不能通过微调参数提升自己判别真假的能力。和 VAE 相比，GAN 的优化目标并没有用到变分下界，并且神经网络的优秀拟合能力使得 GAN 理论上是能够拟合任何分布的。另外，由于 GAN 的优化目标是判断训练集的分布和生成器生成的分布是否拟合，对每一个样本从整体上判断生成质量，经验上它所生成的样本往往更具真实性。而 VAE 由于损失函数中的重构损失往往由 L1/2 损失函数估计，这种像素级别的损失函数偏向于生成一个平均图像，导致生成图像往往边界不清晰，图像模糊。

公式化论述的话，我们遵从 GAN 原文的记法，训练集 $x \sim p_{data}(x)$，生成器和判别器分别记为 G, D，是两个神经网络，前者接受一个随机变量 z 作为输入，它通常来自一个标准正态分布，力图输出符合分布 p_{data} 的样本，后者类似一个分类器，接受一个样本，输出 $[0, 1]$ 之间的数值，力图对训练集中样本输出 1，而对生成器生成的样本输出 0。按前文的说明，GAN 要优化的是下式

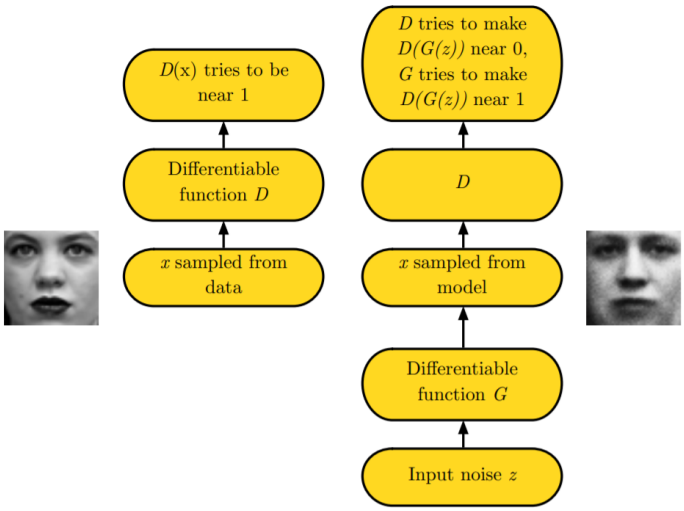

图 2 GAN 的框架

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (2.3)$$

GAN 原文中证明了上式事实上等价于优化生成器的分布和训练集的 Jensen-Shannon 散度，亦即对称的 KL 散度，并且也在理论上论述了最优解的存在性。

在材质迁移的任务中，由于融合了两张图像的材质和内容的图像不可得，因而无法使用 L1/2 损失函数，在这种情况下，往往使用 GAN 来判断生成图片的质量。而仅仅判断要求生成图片属于训练集的分布是不足以要求生成图片融合了不同的风格和内容的。一种可行的改进是使用判别器接受成对的图片判别是否拥有同一材质或内容 [30]，这要求合理的利用训练数据构造真和假的数据对。图像生成任务中，另一种对生成对抗网络的改进是使用多尺度的判别器，这种方法利用多个判别器分别接受原尺度和按不同比例缩小的图片，使得生成图片在各个尺度上都能拟合训练集的特征。利用 ProGAN [40] 的角度来看，我们也可以认为多尺度的判别器事实上为生成器提供了更多的信息，使得它能够从小尺度（也就是全局）的特征开始学习，最后去拟合大尺度（细节）的特征。本文的方法使用了上述两种对生成对抗网络的改进，使生成图像更加真实。

3. 本文实验原理

我们首先正式的阐述本文研究的问题，并且明确后文中会用到的记号。三维模型的最优视图生成任务即针对一个给定的三维模型 $m \in M$，我们的模型能够生成一张图片 $i \in I = R^{m\times n\times 3}$，在后文中除非特殊说明 $m = n = 128$。我们将这

个任务拆解为给定一个视角 v，将三维模型投影为深度图和给定一个材质 s，将深度图渲染为 RGB 图片的两个问题，如（2.4）式。其中 $proj(.)$ 为投影函数，它能够给定一个三维模型和视角，生成一个深度图，即每一个像素的取值和模型上对应的点距相机的距离有关的图片。我们实验中的视角模型只考虑方位角（az）、仰角（el）的变化，其他的参数都预设为 Blender 中的默认值。

$$\begin{aligned} i_d &= proj(m, v) \in R^{m \times n}, vp = [az, el, dist] \\ i_{out} &= f(i_d, s) \end{aligned} \tag{2.4}$$

（2.4）式确定的两个问题，前者可以确定性的通过 Blender 渲染，关于视角选择的问题会在后文中阐述。而后者我们应用了 VAE 和 GAN 来指导生成，在这里我们假定每一张 RGB 图片 $i \in I = R^{m \times n \times 3}$ 都由内容（c）、材质（s）和视角（v）三个独立的变量指导生成，其中内容和材质是需要推断的隐变量，而视角在训练时给出，是已知的变量。而每一张深度图 $i_d \in R^{m \times n}$ 只由内容和视角两个互相独立的变量指导生成，而视角信息在训练时给出，因此可以粗略的认为深度图表征内容信息。公式化表示的话：

$$i_{RGB} \sim P(i|c, s, v), i_d \sim P(i|c, v) \tag{2.5}$$

我们训练时先从真实图片中提取出内容和材质隐变量，再通过这两个隐变量重组出输出图片。类似前文对 VAE 的推导，我们得到训练过程中优化训练集的概率，如（2.6）式。其中第一、四步推导应用了贝叶斯公式，第二步的推导基于我们对材质隐变量的高斯分布假设 $p(c|i) = N(f(i, \theta), I)$，因而 $-\log p(c|i) = \frac{1}{2}||c - f(i, \theta)||^2 + \frac{N}{2}\log(2\pi) \geq 0$，第三步利用对数函数的凸性，是在积分意义下的 Jensen 不等式。

$$\begin{aligned} \log p(i) &= \log p(c) + \log p(i|c) - \log p(c|i) \\ &\geq \log p(c) + \log \int p(i, s|c) ds \\ &\geq \log p(c) + E_q \log \frac{p(i, s|c)}{q(s|i, c)} \\ &= \log p(c) + E_q \log \frac{p(i|c, s)p(s)}{q(s|i)} \\ &= \log p(c) + E_q \log p(i|c, s) + KL(q(s)||p(s)) \end{aligned} \tag{2.6}$$

参考 CVAE 的推导，两边同时对视角取条件，得到最终的优化目标

$$\log p(i|v) = \log p(c) + E_q \log p(i|c,s,v) + KL(q(s)||p(s)) \tag{2.7}$$

其中第一项代表我们推导出的内容隐变量的概率，它越表征内容信息，这个概率值应越大，在实验中我们通过网络重构出的深度图和真实深度图之间的 L1 损失衡量，注意与 [26] 提出的在关键点上的弱监督不同，我们在这里加的是一个很强的限制条件。第二项即为重构出的图片的真实性，我们遵从 VAE 的惯例，用重构出的图片和输入图片之间的 L1 损失实现。第三项为推导出的材质隐变量的 KL 散度损失，用两个正态分布之间的 KL 散度实现。

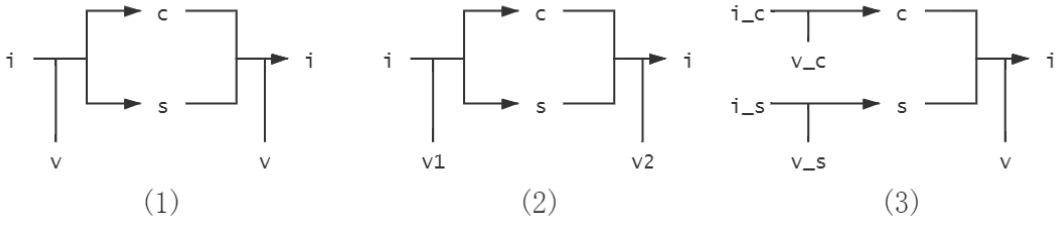

图 3 我们设计的三个实验的流程

上述损失函数只描述了重构图片时的损失，单独实现这三个损失函数只保证了我们模型重构图片和采样的精度，但并不能保证我们假定的隐变量之间互相独立。为此，除了图像重构之外，我们另外设计了两个实验：新视角生成和材质迁移。三个实验的流程见图 3。注意视角变量 v 在训练时给出，我们在从图片中提取出材质和内容变量的过程中也输入已知的视角信息，来提升提取信息的准确性。这种做法也有助于提取到视角无关的材质和内容信息。

为保证视角和其他隐变量相互独立，我们在训练过程中给出所有图片的视角，并且设计了新视角生成的实验：从输入图片中提取内容和材质隐变量，在另一个视角下生成输出图片。在这个实验中，视角 v^1, v^2 均在训练时给出，损失函数的推导同上，损失函数的实现也完全相同。

为保证内容和材质变量相独立，我们设计了交换材质迁移的实验。在这个实验中，材质和内容隐变量分别从不同的图片中提取出，融合得到生成的图片。我们希望生成的图片能够兼具两张图片的内容和材质。但和之前叙述过的实验不同，这个实验并没有监督数据，因此不能用 L1/2 损失来估计优化目标中 $E_q \log p(i|c,s,v)$ 一项，在这里我们应用生成对抗网络来估计这一项。如前文叙述，生成对抗网络的损失函数只在分布的意义上要求生成图片属于训练集的图片的分布，这个损失函数本身只要求生成图片的真实性，并没有要求生成图片融

合两张图片的内容和材质。但在实验中，我们发现只应用这个损失函数，生成的图片也足以达成这个目标。为更进一步要求这个目标，我们在训练时加入另一个判别器，它接受两张图片作为输入，判别这两张图片是否具有同一材质。对内容隐变量不需加入这个约束项，因为作为输入的深度图本身编码了足够多的内容信息也有 L1/2 损失函数约束，然而材质信息是形式上没有其他约束的。

4. 模型实现细节

本节将讨论我们的方法的具体实现，包括各个模型的输入输出、参数设置和训练过程。上一节的讨论可以得到我们需要实现的映射包括

$$c = f(i, v),\ i_d = g(c, v), s = h(i, v), i = m(c, s, v) \tag{2.8}$$

以及为保证对抗训练的判别器，下文中会依次详细的介绍每一部分的实现和参数，以及我们应用到的两个模块 SPADE 和 CoordConv 的细节。我们的代码参考了 pix2pix-cyclegan，spade，coordconv 的实现细节，在整合这些模块的基础上编写了自己的模型和训练代码，实现细节开源在 github.com/DanDoge/FYP 中。后文中我们会使用形如图？的图来描述我们的模型的具体结构，其中每一层的含义和参数意义与它们在 pytorch 中的定义一致，如 $conv(inputnc = 1, outputnc = 3, kernelsize = 3, stride = 1, padding = 1)$ 代表输入一通道、输出三通道、卷积核大小 3×3、步长为 1、外衬一像素的卷积层。

4.1 生成器

我们期望编码器能够从二维图像中提取出三维信息，并且通过解码器将三维信息映射回二维平面。对此，二维图像中的坐标信息是十分重要的，例如输入一辆车的侧面时，有关车轮的信息更有可能在图像的下面获取到而不是上面，解码三维信息时也同理。而卷积层由于自身的空间不变性并不能实现这种需求。为解决这种矛盾，我们将生成器中的所有卷积层替换成坐标卷积（coordconv），这种卷积现在输入变量后附加两个通道，分别包含每一个位置的 x,y 坐标，显式的让卷积层获得位置信息，在不同位置生成不同的输出特征。

在解码器整合内容信息和材质信息时，我们采用了 SPADE 模块来帮助生成。我们采用的 SPADE 模块由 InstanceNorm 和一个 MLP 组成，MLP 负责从深度图和材质特征中估计出用来正则化特征的 γ 和 β。通过将输入特征的均值和方

差替换成估计出的统计量，我们试图在特征维度对齐生成图片的分布和训练图片的分布。不同的 SAPDE 模块 MLP 参数不同，这样我们就能利用连续的几个 SAPDE 模块逐渐将材质信息和内容信息整合。

4.2 编码器

编码器用来实现 $s = h(i, v)$，从图像中提取出和视角无关的材质特征。编码器在几个卷积模块和全局池化层之后，利用两个全连接层估计材质隐变量的均值和方差。其中一个卷积模块由 coordconv 和激活层组合而成，全局平均池化层将各个位置提取到的材质信息整合成全局的材质特征。在我们的实验中，材质隐变量的维度设定为 64 维，这足以表征材质信息。视角信息只在最开始拼接在图像上，使得输入编码器的变量通道数为 $3 + 2 = 5$ 维。

4.3 判别器

判别器采用卷积层和非线性层叠加的方式实现。每一次卷积都使用步长为 2 的卷积，并且倍增输出的通道数以实现在下采样的同时，不丢失原图中信息的效果，通道数最大值设定为 1032。最后一层卷积输出一通道的结果，判别输入向量的真实性。在我们的实验中，下采样的次数为 4 次，对于输入的 128×128 的图像最终输出长宽为 8×8 的向量，每一个位置表示它所对应的原图中 16×16 的部分的真实性。对于后文会叙述的一次输入两张图片的情况，每一个位置表示它对应的 16×16 的部分的两张图片是否属于同一个材质。在原尺度的判别器之外，我们还增加了一个小尺度的判别器，它的输入是原尺度判别器输入 $\frac{1}{2}$ 下采样一次的结果。它的网络结构与原尺度判别器类似，只是通道数减半，以保证每一个通道的信息量大致和原尺度相同。它的输出的每一个位置同样对应它的输入的 16×16 的部分，也就是原图的 32×32 的部分。加入小尺度的判别器保证了输出图像在多尺度上都和目标图像有相同的统计特征，为生成器提供更多的指导。优化过程中我们对两个尺度的判别器的输出都做同样的优化：对真样本。下文中提到的判别器都指的是原尺度和小尺度判别器的组合，图中显示了两个尺度的判别器组合的网络结构。

判别器的损失函数采用最小二乘优化（LSGAN），原文中证明了它等价于优化皮尔森卡方散度，训练更加稳定。我们尝试了其他 GAN 损失函数的变体，包括 WGAN，对生成图像的结果没有很大影响。判别器的损失函数优化目标是使得生成器的生成结果贴近于某一个分布，而这一个分布是通过训练数据指定的。

不同的训练目标对应着不同的指导分布，也就对应着不同的输入数据。在我们的训练过程中，我们希望生成器的生成结果达到下列两个目标：

- 生成高质量的图片。为此我们使用生成图片和真实图片分别作为假/真数据的代表输入判别器，这个损失函数使得生成图片能更具真实性。
- 生成具有特定材质/内容的图片。为此我们引入另一种判别器，它的输入是两张图片的拼接，即输入通道数加倍。训练时，对于 RGB 图片的生成任务，我们要求生成图片的材质相同，采用同一个物体在不同视角下的 RGB 图片作为真样本，用生成器输出的结果和一张真实图片拼接作为假样本训练；对于深度图的生成任务，我们对称的引入一个要求生成深度图的内容一致的判别器。

所以我们的网络中共有八个判别器，分别对应两个生成任务、两个目标和两种尺度的组合。

生成对抗网络的训练不稳定，可能在训练过程中出现网络崩溃的现象，这往往是判别器学习过快导致的。对此 kato 文在新视角生成任务种采用了一种弱化的判别器：即将原视角下重建的结果作为真样本，新视角下生成的结果作为假样本。这样做的初衷是训练初期重建的结果往往会好过新视角生成的结果，判别器能够学习到一些知识，但又不至于学习的过快。我们尝试过这种训练方式，但结果并不好。因为这样判别器学习到的分布特征其实完全来自于生成器，整个网络的监督信息依赖于对生成结果的 L1/2 损失函数。在我们的实验中，单独这个损失函数是不足够的，因此我们的训练过程完全采用真实图片输入。

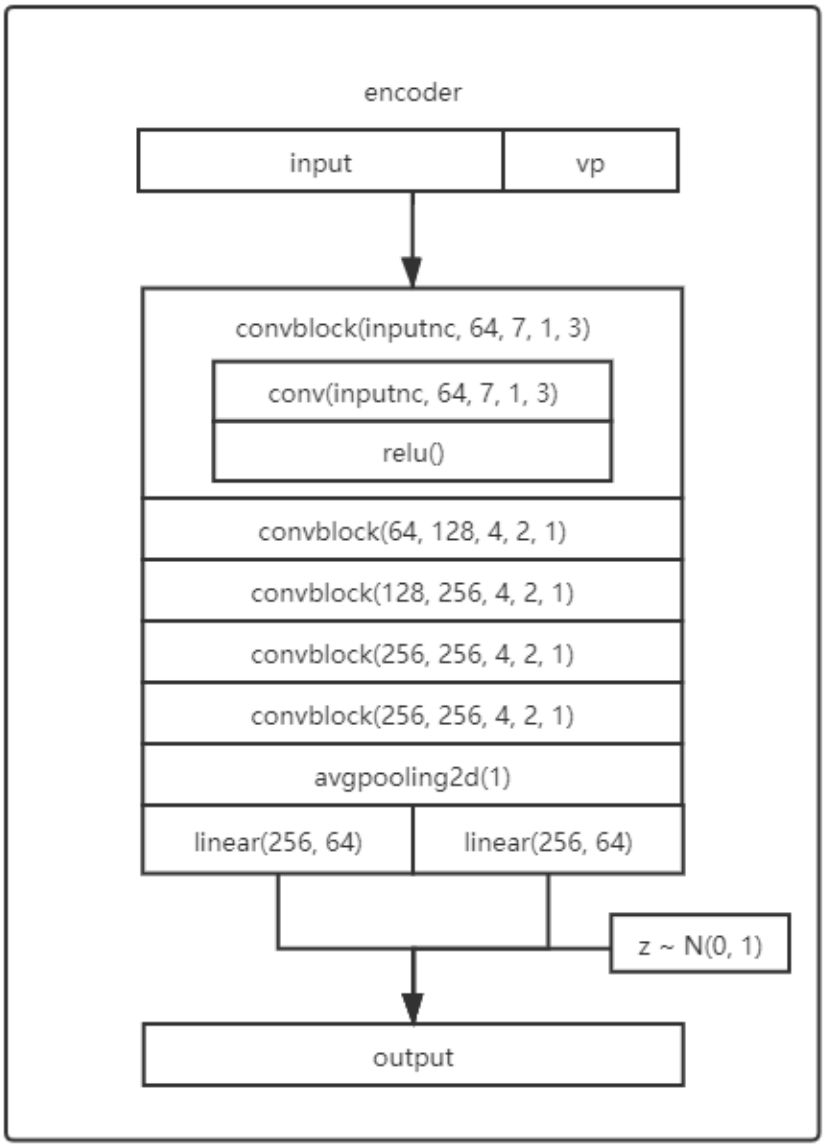
encoder
input
vp
convblock(inputnc, 64, 7, 1, 3)
conv(inputnc, 64, 7, 1, 3)
relu()
convblock(64, 128, 4, 2, 1)
convblock(128, 256, 4, 2, 1)
convblock(256, 256, 4, 2, 1)
convblock(256, 256, 4, 2, 1)
avgpooling2d(1)
linear(256, 64)
linear(256, 64)
z ~ N(0, 1)
output

图 4 编码器的框架

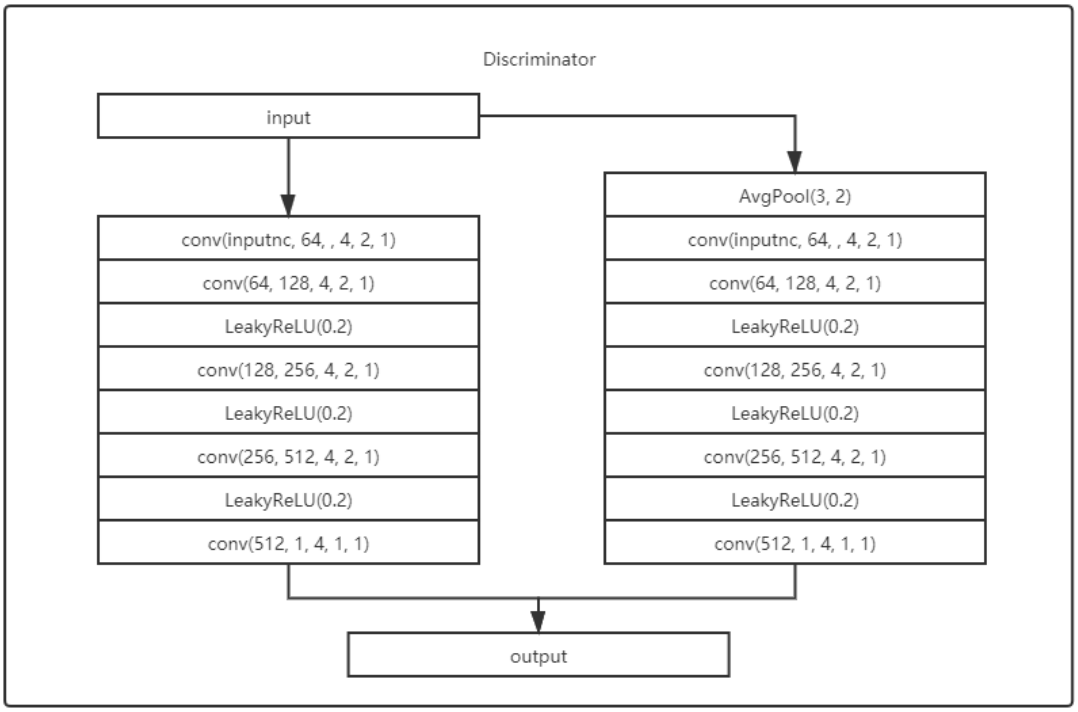
Discriminator
input
conv(inputnc, 64, , 4, 2, 1)
conv(64, 128, 4, 2, 1)
LeakyReLU(0.2)
conv(128, 256, 4, 2, 1)
LeakyReLU(0.2)
conv(256, 512, 4, 2, 1)
LeakyReLU(0.2)
conv(512, 1, 4, 1, 1)
AvgPool(3, 2)
conv(inputnc, 64, , 4, 2, 1)
conv(64, 128, 4, 2, 1)
LeakyReLU(0.2)
conv(128, 256, 4, 2, 1)
LeakyReLU(0.2)
conv(256, 512, 4, 2, 1)
LeakyReLU(0.2)
conv(512, 1, 4, 1, 1)
output

图 5 判别器的框架

参考文献

[1] Xiaobai Chen, Aleksey Golovinskiy, and Thomas A. Funkhouser. A benchmark for 3d mesh segmentation. In *SIGGRAPH 2009*, 2009.

[2] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, 2014.

[3] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. *ArXiv*, abs/1604.00449, 2016.

[4] Priyanka Mandikal, L. NavaneetK., Mayank Agarwal, and Venkatesh Babu Radhakrishnan. 3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. In *BMVC*, 2018.

[5] Christopher Bongsoo Choy, Michael Stark, Sam Corbett-Davies, and Silvio Savarese. Enriching object detection with 2d-3d registration and continuous viewpoint estimation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2512–2520, 2015.

[6] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, abs/1512.03012, 2015.

[7] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82, 2014.

[8] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2014.

[9] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. Meshlab: an open-source mesh processing tool. In *Eurographics Italian Chapter Conference*, 2008.

[10] Pere-Pau Vázquez, Miquel Feixas, Mateu Sbert, and Wolfgang Heidrich. Automatic view selection using viewpoint entropy and its applications to image-based modelling. *Comput. Graph. Forum*, 22:689–700, 2003.

[11] David L. Page, Andreas F. Koschan, Sreenivas R. Sukumar, Besma Roui-Abidi, and Mongi A. Abidi. Shape analysis algorithm based on information theory. *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, 1:I–229, 2003.

[12] Helin Dutagaci, Chun Pan Cheung, and Afzal Godil. A benchmark for best view selection of 3d objects. In *3DOR@MM*, 2010.

[13] Chang Ha Lee, Amitabh Varshney, and David W. Jacobs. Mesh saliency. In *SIGGRAPH 2005*, 2005.

[14] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Found. Trends. Comput. Graph. Vis.*, 9(1-2):1–148, June 2015.

[15] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deep stereo: Learning to predict new views from the world's imagery. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, June 2016.

[16] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jagannath Malik, and Alexei A. Efros. View synthesis by appearance flow. In *ECCV*, 2016.

[17] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *European Conference on Computer Vision*, 2018.

[18] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. *The IEEE International Conference on Computer Vision (ICCV)*, Nov 2019.

[19] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision (ECCV)*, 2016.

[20] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[21] Xiaogang Xu, Ying-Cong Chen, and Jiaya Jia. View independent generative adversarial network for novel view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[22] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–711, 2017.

[23] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. Beautyglow: On-demand makeup transfer framework with reversible generative network. pages 10034–10042, 06 2019.

[25] TingTing Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *MM '18*, 2018.

[26] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Disentangling content and style via unsupervised geometry distillation. *ArXiv*, abs/1905.04538, 2019.

[27] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[28] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation. *ArXiv*, abs/1705.09966, 2017.

[29] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter. M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 405–415, 2017.

[31] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[32] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization, 2019.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[34] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018.

[35] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and William T. Freeman. Visual object networks: Image generation with disentangled 3D representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[36] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[37] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $\$\beta\$ - vae. 2018.

[38] P. K. Rubenstein, B. Schölkopf, and I. Tolstikhin. Learning disentangled representations with wasserstein auto-encoders. In *Workshop at the 6th International Conference on Learning Representations (ICLR)*, May 2018.

[39] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. 2018.

[40] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2017.

本科期间的主要工作和成果

本科期间参加的主要科研项目
本研基金

1. 国家创新训练计划. 基金类型. 连宙辉. 2018-2019

致谢

感谢