第 58 卷 第 14 期/2021 年 7 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于深度学习的单视图彩色三维重建

朱育正，张亚萍*，冯乔生

云南师范大学信息学院，云南 昆明 650500

摘要 从单个图像中同时恢复 3D 形状及其表面颜色的任务极具挑战性，为此提出一个端到端的网络模型来解决这一难题，该模型采用编码器与解码器结构。以单张图像作为输入，首先通过编码器提取特征，再将其同时送入形状生成器和颜色生成器中，得到形状估计以及与其对应的表面颜色，最后通过可微渲染框架渲染生成最终的彩色三维模型。为了保证重构三维模型的细节，在网络的编码器中引入注意力机制以进一步提高重建效果。实验结果表明，与三维重建网络模型相比，所设计的模型在真实三维模型交并比上分别提高 10% 和 3%；与开源项目相比，所设计的模型在结构相似性上提高了 3%，在均方误差上降低了 1.2%。

关键词 深度学习；彩色三维重建；单视图；可微渲染器；注意力机制

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.1410010

Colorful 3D Reconstruction from Single Image Based on Deep Learning

Zhu Yuzheng, Zhang Yaping*, Feng Qiaosheng

School of Information Science and Technology, Yunnan Normal University, Kunming, Yunnan 650500, *China*

Abstract The task of recovering the 3D shape and its surface color from a single image at the same time is extremely challenging. For this reason, an end-to-end network model is proposed to solve this problem, which uses an encoder and decoder structure. Taking a single image as input, first extract the features through the encoder, and then send them to the shape generator and the color generator at the same time to get the shape estimation and the corresponding surface color, and finally through the differentiable rendering framework to generate the fianl color three-dimensional model. In order to ensure the details of the reconstructed 3D model, an attention mechanism is introduced into the network encoder to further improve the reconstruction effect. The experimental results show that compared with the 3D reconstruction network models, the designed model has a 10% and 3% increase in the real 3D model intersection ratio; compared with the open source project, the structural similarity of the designed model is improved by 3%, and the mean square error is reduced by 1.2%.

Key words deep learning; colorful three-dimensional reconstruction; single image; differentiable renderer; attention mechanism

OCIS codes 150.1135; 110.3010; 110.4155

1 引 言

计算机视觉研究的主要目标之一是从二维图像中重构出三维模型，目前已经有许多三维重建方法被提出[1-3]。对于单张图像的三维重建，因其具有特殊的输入形式，所以逐渐吸引了大量研究人员的关注与研究。根据重建方式的不同，其可以分为基于先验知识或约束的重建方法和基于深度学习的重建方法两类。传统重建方法往往利用先验知识或引入合适的约束对图像进行重建，例如在三维人体重建的背景下，研究人员提出了与 SiCloPe(Silhouette-based Clothed People)[3] 和 SMPL(Skinned Multi-

收稿日期：2020-10-10；**修回日期**：2020-11-12；**录用日期**：2020-11-14

基金项目：国家自然科学基金(61863037)、云南省“万人计划”青年拔尖人才专项

通信作者：* zhangyp@ynnu.edu.cn

Person Linear)[4]类似的人类特定模型，但这种方法多是针对特定场景而设计的，难以扩展到其他类别的物体上，而且该模型的泛化能力较弱以及前置条件较为苛刻，实际应用并不理想。随着深度学习技术的发展，基于单张图像的三维重建取得了显著进展[5]。对于三维形状的估计，Wu 等[6]建立的 3D ShapeNets 是基于体素表示的三维重建网络，通过扩展深度卷积置信网络(DBN)[7]来模拟体积形状；Choy 等[8]使用了具有长短期记忆(LSTM)的 3D-R2N2(3D Recurrent Reconstruction Neural Network)对单视图与多视图进行重建；Wu 等[9]提出的 MarrNet 模型是将 2.5D 估计组件作为中间组件，并引入了重投影一致性损失函数；Kato 等[10]提出了使用一种自定义函数来逼近光栅化的后向梯度，从而实现可微渲染，进而完成基于单张图像的三维网格重建任务。对于三维形状的重建，现有的工作已经取得一系列成果，而且可以获得较好的形状估计，然而只恢复三维形状是不够的，因为世界不仅仅只有形状，还包含丰富多彩的颜色。

目前，单视图图像的彩色 3D 模型重建仍是一个极具挑战的问题。Sun 等[11]提出了彩色体素网络(CVN)，在单个二维图像输入的条件下分别学习 3D 对象的形状与表面颜色，该过程是通过两套独立的编码器-解码器网络来实现的。具体来说，对于形状的重建，该过程最终是生成一个形状体，其中体素的状态表示对应的占有情况；对于表面颜色的恢复，颜色体有三个通道，而且与形状体具有相同的尺寸，最终的 3D 纹理模型是由形状体与颜色体合成而来。另一项工作，Kanazawa 等[12]利用带注释的图像集进行训练无需依赖真实的 3D 模型或多视图监督即可学习可变形模型，通过 SFM(Structure From Motion)方法学习的平均形状与按实例预测的形变对平均形状进行参数化处理，可以得到最终的形状预测，其贴图方法参考文献[13]。通过对输入图像进行双线性插值以获取纹理图，再通过球坐标将其映射到一个球体上，最后将该球体进行变形以得到最终的纹理模型，其限制在于预测的形状和形变都是镜像对称的，使用的应用场景存在一定限制。

本文提出的模型有两个创新点：1)在三维重建的任务中引入注意力机制，在网络的编码器中应用 CBAM(Convolutional Block Attention Module)，有助于编码器更好地从输入图像中提取特征，从而提升模型的重建效果；2)引入可微渲染框架 SoftRas，其能够使用可微函数直接渲染给定的网格，因此可以采用与 GPU(Graphic Processing Units)渲染框架相同的数据表示形式，即网格提供几何信息，纹理图像提供相应的颜色信息，从而实现彩色三维模型重建任务的模块化。

2 网络结构

实验的目的是提供一个有效的网络模型，实现单张图像的彩色三维模型重建。以端到端的三维重建网络模型[8]为基础，提出一个改进的算法模型，模型结构如图 1 所示，该模型采用了编码器-解码器体系结构，编码器用于特征提取。颜色和形状生成器共享同一个特征提取器。形状生成器由三个全连接层组成，其输出每个顶点的位移矢量，该矢量将模板网格变为目标模型。受 Sun 等[11]工作的启发，颜色生成器包含两个分支：一个对输入图像中的代表性颜色进行采样，可以构建调色板；另一个使用选择网络模块将调色板中的颜色组合起来，用于对采样点进行纹理处理。通过将颜色选择与学习得到的调色板相乘，可以获取完整的颜色预测。最后将预测的模型形状与表面颜色送入 SoftRas 渲染器[14]中，可以渲染得到彩色三维模型。

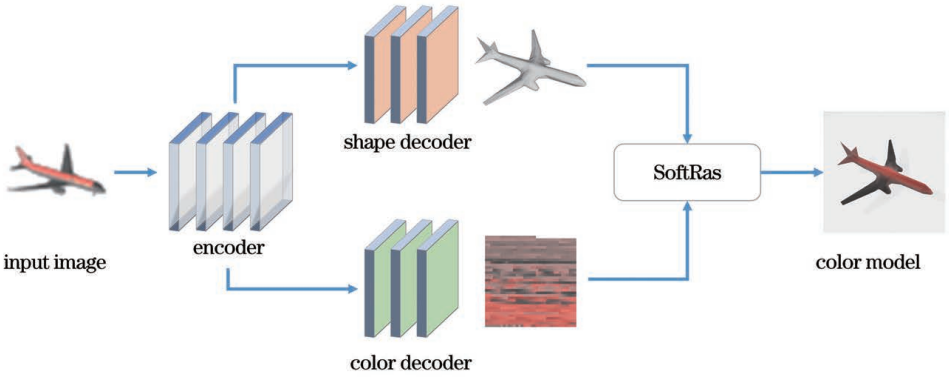

图 1 模型结构

Fig. 1 Model structure

2.1　形状重建

2D 到 3D 形状的重建已经被广泛研究，目前大多数模型都是基于卷积编码-解码网络架构实现的，三维形状的重建网络如图 2 所示。与基于深度学习的网格重建方法[14-16]类似，本文通过变形模板网格来合成三维模型，而不是直接生成模型，对比直接从 2D 图像生成网格，编码器能够更好地预测残差，而不是结构化输出，为此可以进一步提高重建细节。具体而言，使用具有 642 个顶点的各向同性球体作为基础，使用编码器来学习每个顶点的位移矢量，再通过解码器将模板网格变为目标模型，最后通过渲染器来生成预测轮廓图像 $\hat{\boldsymbol{I}}_s$，轮廓损失 L_s 可表示为

$$L_s=1-\frac{\|\hat{\boldsymbol{I}}_s\otimes\boldsymbol{I}_s\|_1}{\|\hat{\boldsymbol{I}}_s\oplus\boldsymbol{I}_s-\hat{\boldsymbol{I}}_s\otimes\boldsymbol{I}_s\|_1},\tag{1}$$

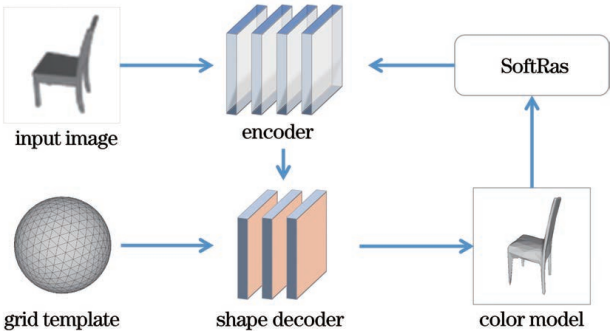

图 2　三维形状的重建网络

Fig. 2　Reconstruction network of 3D shape

式中：$\boldsymbol{I}_s$ 为真实轮廓图像；$\otimes$ 和 $\oplus$ 分别为两个矩阵间的直积运算与直和运算。

2.1.1　注意力机制

当运行现有论文的开源代码时，发现输出的部分模型缺乏细节信息，将这一问题归结为单视图重建任务的病态性和现有网络的片面性，而这些细节可以在输入图像中获取。为了捕捉这些细节，引入注意力机制。注意力机制是人工智能领域中模拟人类视觉模式的方法之一，人类通过快速扫描全局图像可以获得需要重点关注的目标区域，进而对这一区域投入更多的注意力，从而获取更多所需关注目标的细节信息。将这个原理应用到计算机视觉中，即在相应的任务中获得的数据是存在冗余的，因此不需要对所有的信息进行平均化处理，只需选择与目标任务最相关的信息。

在此，认为注意力机制的引入应起到锦上添花的作用，而不会对现有模型的重建效果产生负优化，而且不能使模型的计算量过多的增加。所提方法是在编码器中应用 CBAM[17]，如图 3 所示，相比于 SENet(Squeeze-and-Excitation Networks)[18] 只关注通道的注意力机制可以取得更好的效果，其中 $\boldsymbol{Z}'$ 为处理后得到的特征。当 CBAM 在靠前的层中使用时，其更关注于输入图像的细节，当 CBAM 在靠后的层中使用时，其更关注于图像的结构特征。基于上述认识，应用 CBAM 可以使模型更好地完成彩色三维重建任务。

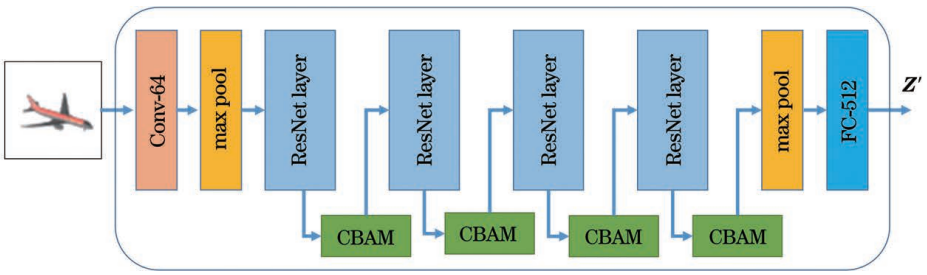

图 3　加入 CBAM 模块的编码器结构

Fig. 3　Encoder structure added to CBAM module

CBAM 是一种结合空间与通道的注意力机制，给定输入特征 $\boldsymbol{Z}\in\mathbb{R}^{C\times H\times W}$，依次推导一维通道注意力图 $\boldsymbol{M}_C\in\mathbb{R}^{C\times1\times1}$ 和二维空间注意力图 $\boldsymbol{M}_S\in\mathbb{R}^{1\times H\times W}$，其中 C 为通道数，H 和 W 分别为特征图的高和宽，完整的注意力机制定义为

$$\begin{cases}\boldsymbol{Z}_1=\boldsymbol{M}_C(\boldsymbol{Z})\otimes\boldsymbol{Z}\\ \boldsymbol{Z}_2=\boldsymbol{M}_S(\boldsymbol{Z}_1)\otimes\boldsymbol{Z}_1\end{cases},\tag{2}$$

式中：$\boldsymbol{Z}_1$ 和 $\boldsymbol{Z}_2$ 分别为通道注意力信息和空间注意力信息。通道注意力机制的计算过程如下：首先通过最大池化与平均池化操作来聚合输入特征的空间信息，生成两个不同的空间上下文，即 $\boldsymbol{Z}^C_{avg}$(输入特征经过平均池化后得到的特征)和 $\boldsymbol{Z}^C_{max}$(输入特征经过最大池化后得到的特征)，并送入一个隐藏激活大小为 $\mathbb{R}^{C/r\times1\times1}$ 的多层感知机(MLP)中，从而得到最终的通道注意力图 $\boldsymbol{M}_C\in\mathbb{R}^{C\times1\times1}$，其中 r 为压缩比。$\boldsymbol{M}_C$ 具体定义为

$$\boldsymbol{M}_{\mathrm{C}}(\boldsymbol{Z})=\mathrm{sigmoid}\left\{W_1\left[W_0\left(\boldsymbol{Z}_{\mathrm{avg}}^{\mathrm{C}}\right)\right]+W_1\left[W_0\left(\boldsymbol{Z}_{\mathrm{max}}^{\mathrm{C}}\right)\right]\right\},W_1\in\mathbb{R}^{C\times C/r},W_0\in\mathbb{R}^{C/r\times C},\tag{3}$$

式中：W_1 和 W_0 为 MLP 的权重。空间注意力机制的过程计算如下：通过最大池化与平均池化操作来聚合上一层输出特征图的通道信息，最终生成 $\boldsymbol{Z}_{\mathrm{avg}}^{\mathrm{S}}\in\mathbb{R}^{1\times H\times W}$ 和 $\boldsymbol{Z}_{\mathrm{max}}^{\mathrm{S}}\in\mathbb{R}^{1\times H\times W}$，再通过一个滤波器 $f^{7\times7}$，即由一个 7×7 标准的卷积层生成最终的二维空间注意力图，具体定义为

$$\boldsymbol{M}_{\mathrm{S}}(\boldsymbol{Z})=\mathrm{sigmoid}\left[f^{7\times7}\left[\boldsymbol{Z}_{\mathrm{avg}}^{\mathrm{S}};\boldsymbol{Z}_{\mathrm{max}}^{\mathrm{S}}\right]\right]。\tag{4}$$

2.1.2　编码器-解码器

特征提取器即为编码器，由 ResNet 和 CBAM 共同组成。形状解码器由三个全连接层组成，隐藏层的大小分别为 1024 和 2048。

2.2　表面颜色学习

从单张图像中推断 3D 模型的表面颜色是一个不适定的问题，因为被遮挡物体的表面颜色是未知的，但输入图像中可用的语义信息和对象的一般属性通常可以为人类合理推测被遮挡物体的表面颜色提供线索。例如，如果从台灯的一侧看到灯罩是白色的，尽管有可能存在另一侧为另一种颜色，但通常会猜测其另一侧同样为白色。因此，表面颜色学习的目标是从给定的视图中对所捕获对象的信息进行完整的表面颜色估计。

首先，将颜色重建定义为一个分类问题，通过对输入图像进行多重采样来学习得到表面颜色，颜色重建网络如图 4 所示。颜色生成器由颜色采样网络与选择网络组成，具体而言，输入图像传递至编码器后，提取的特征被馈入颜色采样网络和颜色选择网络中，前者对用于构建调色板的代表颜色进行采样，后者将调色板中的颜色组合起来，用于对采样点进行纹理处理，通过将颜色选择与学习的调色板相乘来获得最终的颜色预测。颜色损失 L_{c} 由预测图像

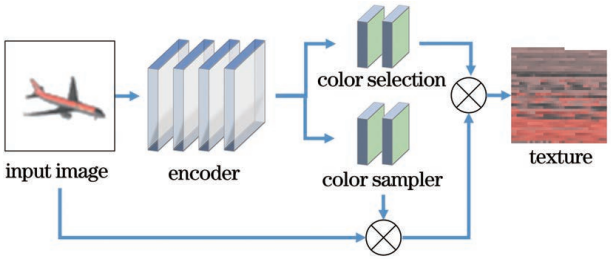

图 4　颜色重建网络

Fig. 4　Color reconstruction network

与输入图像之间的 L_1 范数来衡量，即最小化预测图像 $\hat{\boldsymbol{I}}_{\mathrm{c}}$ 与输入图像 $\boldsymbol{I}_{\mathrm{c}}$ 的绝对差值的总和，表达式为

$$L_{\mathrm{c}}=\left\|\hat{\boldsymbol{I}}_{\mathrm{c}}-\boldsymbol{I}_{\mathrm{c}}\right\|_1。\tag{5}$$

2.3　渲染器

渲染是指从三维世界中生成二维图像的过程。针对三维重建，显然会想到反转这一过程，从而实现从二维图像来推断三维信息，进而完成三维重建任务。然而标准的图形渲染器涉及一个称为光栅化的离散步骤，其是不可微分的，这阻碍了反向传播。采用深度学习来解决这一问题，Kato 等[10]提出了一种近似的光栅化梯度，其可以将渲染模块集成到神经网络中，但其存在正向传播和反向传播的不一致性，这可能会导致优化行为的异常，限制了对其他三维推理任务的泛化能力。

SoftRas 渲染器是所设计模型的核心结构，其能够使用可微函数直接渲染给定的网格，且前后传播具备一致性，能够提供高质量的梯度流，从而实现反转渲染这一过程。SoftRas 渲染器如图 5 所示，其中 $\boldsymbol{D}$ 为概率图。SoftRas 同时考虑外部条件（相机 $\boldsymbol{P}$ 和光照 $\boldsymbol{L}$），以及模型的内在属性（三角形网格 $\boldsymbol{M}$ 和每个顶点外观 $\boldsymbol{A}$，即颜色和材质等），通过转换基于相机 $\boldsymbol{P}$ 的输入三角网格 $\boldsymbol{M}$，可以得到网格法线 $\boldsymbol{N}$、像空间坐标 $\boldsymbol{U}$ 和深度 $\boldsymbol{Z}$，通过 $\{\boldsymbol{A},\boldsymbol{N},\boldsymbol{L}\}$ 来计算颜色 $\boldsymbol{C}$。

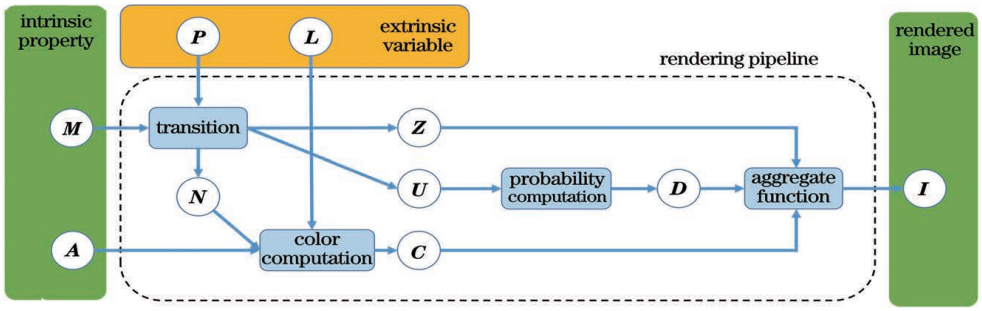

图 5　SoftRas 渲染器

Fig. 5　SoftRas rendering structure

SoftRas 使用概率图 D_j 对每个像素在特定三角形 f_j 内的概率进行建模，(i,j)为像素点坐标。为了估计像素 p_i 处 D_j 的概率，需要同时考虑 p_i 和 f_j 之间的相对位置与距离，故 D_j^i 定义为

$$D_j^i=\mathrm{sigmoid}\left[\delta_j^i\cdot\frac{d^2(i,j)}{\sigma}\right],\tag{6}$$

式中：σ 为控制概率分布的锐度的正标量；δ_j^i 为符号表示符，当 p_i 位于三角形 f_j 时，值为 1，反之值为 −1；$d(i,j)$为 p_i 至 f_j 的边缘最近的距离。

对于每个网格三角形 f_j，其颜色图 C_j 通过重心坐标插值的顶点颜色来获得，最后使用聚合函数 A_S 合并颜色图，获得基于概率图和深度 Z 的渲染输出 $\boldsymbol{I}$，进而建立起三维模型与二维图像的联系。聚合函数定义为

$$\begin{cases}I^i=A_S(\{C_j\})=\sum\limits_j w_j^iC_j^i+w_b^iC_b\\ w_j^i=\dfrac{D_j^i\exp(z_j^i/\gamma)}{\sum\limits_k D_k^i\exp(z_k^i/\gamma)+\exp(\varepsilon/\gamma)}\end{cases},\tag{7}$$

式中：C_b 为背景颜色；w 为权重，满足 $\sum\limits_j w_j^i+w_b^i=1$；$z_j^i$ 为三角形 f_i 上 2D 投影为 p_i 的 3D 点的归一化逆深度；ε 为一个常数；γ 为控制聚合函数的清晰度，默认设置为 1×10^{-4}；$k=1,2,\cdots\cdots,j$。

2.4　损　失

重建网络损失由三部分组成：轮廓损失 L_s、颜色损失 L_c 和几何损失 L_g。为了进一步提升重建效果，保证变形期间相邻顶点间的相对位置，防止网格相互交叉，这里引入 L_g 损失对形状与颜色进行预测拉普拉斯正则化，所以最终的重建网络损失为三种损失的加权总和为

$$L=L_s+\lambda L_c+\mu L_g,\tag{8}$$

式中：λ 和 μ 为超参数。

3　实　　验

3.1　评价指标

在评价彩色三维重建效果方面，以目前单视图重建效果最优的三个主流模型[OccNet(Occupancy Networks)[19]、3D-R2N2 和 SoftRas]作为基准，分别对重建形状和表面颜色进行评估。由于 Kato 等[10]提供的数据集仅包含真实三维模型的体素，故仅采用交并比(IOU)作为 3D 形状重建的标准评估指标，即对于一个 3D 形状及其预测，IOU 计算的是两者的交集和并集之间的比。通过测量不同视角下真实三维模型的视图与对应预测视图的图像结构相似性指标(SSIM)与均方误差(MSE)，可以共同评估表面颜色重建的效果。SSIM 指标越大，MSE 越小，代表表面颜色重建效果越好。SSIM 具体定义为

$$S_{\mathrm{SSIM}}(x,y)=\frac{(2\mu_x\mu_y+c_1)(\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)},\tag{9}$$

式中：μ_x 和 μ_y 分别为两张图像 x 和 y 的平均值；σ_x 和 σ_y 分别为两张图像 x 和 y 的标准差；σ_{xy} 为两者的协方差；c_1、c_2 和 c_3 为常数。

3.2　数据集

使用 Kato 等[10]提供的 ShapeNet 数据集进行网络训练和测试，其包含了飞机、椅子和桌子等 13 种类别，因为这些类别包含丰富的纹理信息，同时还提供多种形状，能够有效地检测出单视图彩色三维重建的效果。具体而言，每个对象在 24 个不同的视角下进行渲染，图像的分辨率为 64 pixel×64 pixel。

3.3　实验设置

使用设置为 $\alpha=0.0001$，$\beta_1=0.9$，$\beta_2=0.999$ 的 Adam 优化器进行优化，批处理大小设为 8。λ 值和 μ 值默认设为 1 和 0.001。在训练迭代 2000 轮内，学习率 $s_v=0.0001$，在 2000 轮以外，$s_v=0.00003$。

4　实验结果与分析

4.1　注意力机制的实验

为了检测不同注意力机制对彩色三维重建效果的影响，通过在编码器中分别应用通道注意力机制 SENet 模块、CBAM 模块中的空间注意力模块和结合空间与通道两方面的 CBAM 模块，将使用模块后的效果进行对比。以开源项目 SoftRas 作为基准，这里将其命名为 SoftRas_O。应用不同注意力机制的各指标如表 1 所示，其中第一个为未引入注意力机制的 SoftRas_O，第二个为单独应用 CBAM 中的空间注意力模块(CBAM_S)的结果，第三个和第四个分别为应用 SENet 模块和完整 CBAM 模块的结果。从表1可以看到，在保持与应用SENet模

表 1　不同注意力机制的指标对比

Table 1　Comparison of indicators of different attention mechanism

Attention mechanism	IOU	SSIM	MSE
SoftRas_O	0.576	0.548	0.175
CBAM_S	0.589	**0.579**	0.161
SENet	**0.593**	0.563	0.186
CBAM	**0.593**	0.578	**0.163**

块相同的 IOU 指标的基础上，应用 CBAM 模块的整体效果最好。这里需要说明的是引入的注意力模块均在 ResNet 的每一层后添加。

为了进一步探究 CBAM 的应用以及其在模型中的位置对重建效果的影响，进行另一组对比实验，结果如表 2 所示，第一个为未添加 CBAM 模块的模型，第二个是在 ResNet 的前两层后添加的，第三个是在 ResNet 的后两层添加，最后一个是在 ResNet 的所有层后添加。从表 2 可以看到，所有引入注意力机制的模型的IOU值均大于未添加注意力模块的模型，总体来说关注所有层的模型效果更好。对比实验表明，CBAM 的引入对于提升三维重建效果是有效的，且应用在模型的所有层的效果最优。

表 2　不同 CBAM 与 SoftRas_O 的 IOU 指标对比

Table 2　Comparison of IOU indicators of different CBAMs and SoftRas_O

Category	IOU			
	SoftRas_O	CBAM 1-2	CBAM 3-4	CBAM 1-4
Airplane	0.603	0.646	**0.648**	**0.648**
Bench	0.444	**0.461**	0.456	0.454
Cabinet	0.616	0.629	0.613	**0.633**
Car	0.672	0.665	0.669	**0.675**
Chair	0.481	0.492	0.478	**0.493**
Display	0.556	**0.571**	0.554	0.563
Lamp	0.444	**0.463**	0.460	0.459
Loudspeaker	0.615	0.631	0.635	**0.636**
Rifle	0.643	0.677	**0.683**	0.674
Sofa	0.619	0.618	0.628	**0.630**
Table	0.453	**0.487**	0.476	0.470
Telephone	0.741	0.744	**0.776**	0.766
Vessel	0.589	0.611	0.610	**0.612**
Mean	0.575508	0.59192	0.59123	**0.59330**

4.2　三维形状重建

表 3 为现有主流模型与所设计的模型在 ShapeNet 数据集上的 IOU 对比。从表 3 可以看到，虽然所设计的模型在柜子与汽车上的 IOU 值略小于 OccNet，但所设计模型的预测结果和真实三维模型的交并比(IOU)是优于现有模型的，所设计的模型在真实三维模型交并比上分别提高 10% 和 3%。实际重建结果对比如图 6 所示。从图 6 可以

表 3　不同模型的 IOU 指标对比

Table 3　Comparison of IOU indicators of different models

Category	IOU			
	3D-R2N2	OccNet	SoftRas_O	Our model
Airplane	0.426	0.547	0.603	**0.648**
Bench	0.373	0.452	0.444	**0.454**
Cabinet	0.676	**0.732**	0.616	0.633
Car	0.661	**0.731**	0.672	0.675
Chair	0.439	0.502	0.481	**0.493**
Display	0.440	0.479	0.556	**0.563**
Lamp	0.281	0.370	0.444	**0.459**
Loudspeaker	0.611	0.653	0.615	**0.636**
Rifle	0.375	0.458	0.643	**0.674**
Sofa	0.626	0.671	0.619	**0.630**
Table	0.420	0.506	0.453	**0.470**
Telephone	0.611	0.709	0.741	**0.766**
Vessel	0.482	0.521	0.589	**0.612**
Mean	0.493	0.564	0.571	**0.593**

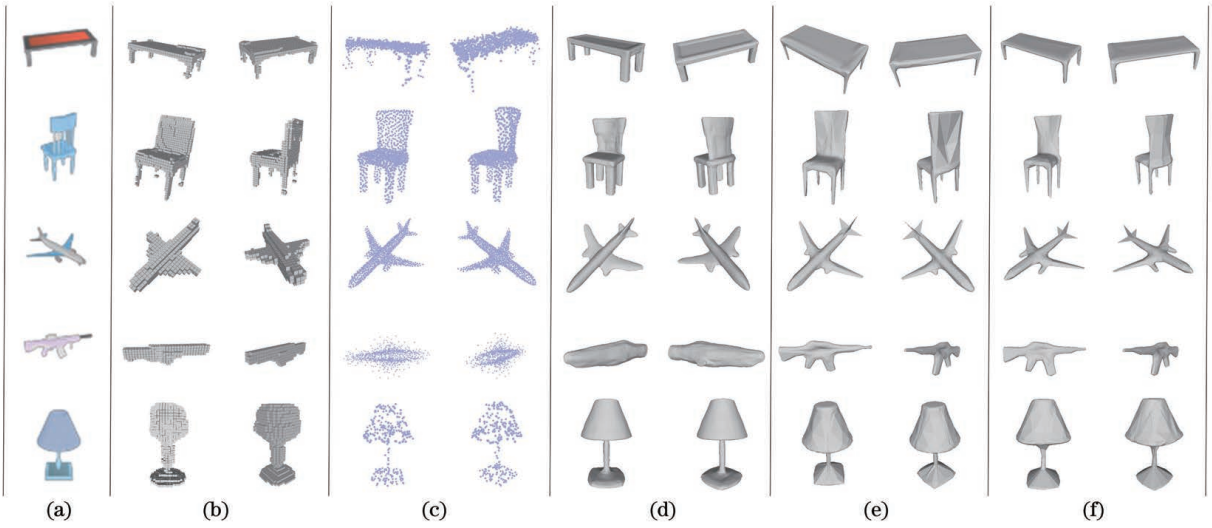

图 6　不同三维模型的预测对比图。(a)原始图像；(b) 3D-R2N2；(c) PSGN；(d) OccNet；(e) SoftRas_O；(f)所设计的模型

Fig. 6　Comparison of predictions of different 3D models. (a) Original images; (b) 3D-R2N2; (c) PSGN; (d) OccNet; (e) SoftRas_O; (f) designed model

看到，虽然所有的模型都能够完成重建任务，但 3D-R2N2 生成的模型表面十分粗糙，缺乏细节；PSGN (Point Set Generation Network)[20] 虽然输出了高质量的点云，但却缺乏连接性，必须进行额外的有损后续处理才能生成最终的网格；相比之下，OccNet 与所设计的模型都能够捕捉复杂的细节并生成平滑的三维模型，如 OccNet 重建的台灯灯罩，所设计的模型重建的桌子表面。但在某些情况下，与 OccNet 相比，所设计的模型可以生成更具吸引力的细节，如飞机的发动机和枪的表面。总体来说，所设计的模型的三维重建效果更好，更接近真实三维形状。

4.3　表面颜色重建

所设计的模型能够从输入图像中恢复模型的表面颜色，所设计的模型与 SoftRas_O 的 SSIM 与 MSE 对比结果如表 4 所示。从表 4 可以看到，所设计的模型的整体效果更好，相比于比 SoftRas_O，所设计的模型在 SSIM 上提高了 3%，在 MSE 上降低了 1.2%。所设计的模型从单个图像恢复的彩色三维模型如图 7 所示。从图 7 可以看到，虽然输入图像的分辨率仅为 64 pixel×64 pixel，但所设计的模型依然能够实现准确的色彩恢复与模型细节的还原，例如机身与机翼的颜色过渡。

表 4　所设计的模型与 SoftRas_O 的 SSIM 与 MSE 对比结果

Table 4　Comparison results of SSIM and MSE between designed model and SoftRas_O

Category	SSIM		MSE	
	SoftRas_O	Our model	SoftRas_O	Our model
Airplane	**0.627**	0.610	**0.084**	0.090
Bench	0.475	**0.493**	0.186	**0.183**
Cabinet	0.565	**0.587**	0.188	**0.182**
Car	0.584	**0.592**	0.197	**0.184**
Chair	0.442	**0.486**	0.240	**0.222**
Display	0.453	**0.488**	0.264	**0.250**
Lamp	0.537	**0.749**	0.156	**0.068**
Loudspeaker	0.569	**0.591**	**0.218**	0.219
Rifle	**0.592**	**0.592**	0.081	**0.070**
Sofa	0.489	**0.494**	0.230	**0.228**
Table	0.537	**0.548**	0.198	**0.197**
Telephone	0.654	**0.673**	0.136	**0.133**
Vessel	0.600	**0.617**	**0.096**	0.098
Mean	0.548	**0.578**	0.175	**0.163**

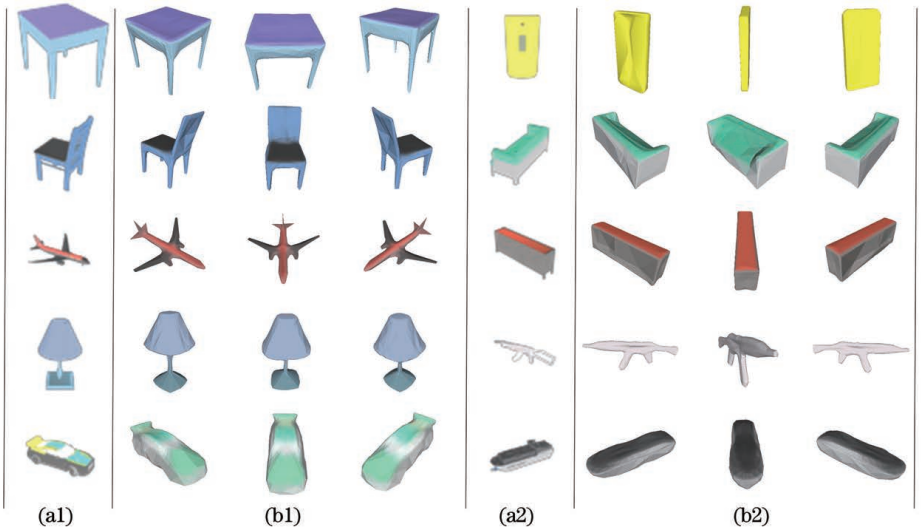

图 7　不同彩色三维模型的重建结果。(a1)(a2)原始图像；(b1)(b2)重建结果

Fig. 7　Reconstruction results of different color 3D models. (a1) (a2) Original images; (b1) (b2) results of reconstruction

5　结　　论

针对目前单视图三维重建的工作主要集中在形状恢复上，提出了一个基于深度学习的三维重构网络，并引入一个可以同时考虑外部变量与内在变量的可微渲染框架，其能够生成从图像像素到网格顶点及其属性(颜色和法线等)的有效梯度流；为了保证重构三维模型的细节，模型引入注意力机制以进一步提高重建效果。实现结果表明，所设计的模型不仅能够从单幅图像中重构三维模型的形状与表面颜色，而且能够较好地还原三维模型的细节。但由于三维模型的数据库仍不够完

善，且彩色三维模型不仅仅涉及形状恢复，还涉及表面颜色，模型复杂，训练难度较大。所设计的模型仅在简单的物体类别上进行了训练和验证，如何通过单张图像建立复杂的彩色三维模型是下一步需要解决的一个问题。

参 考 文 献

[1] Pang Z Y, Zhou Z F, Wang L D, et al. Improved three-dimensional reconstruction algorithm for point cloud data[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021102.
庞正雅, 周志峰, 王立端, 等. 改进的点云数据三维重建算法[J]. 激光与光电子学进展, 2020, 57(2): 021102.

[2] Zhang Z J, Cheng X J, Cao Y J, et al. Application of 3D reconstruction of relic sites combined with laser and vision point cloud[J]. Chinese Journal of Lasers, 2020, 47(11): 1110001.
张子健, 程效军, 曹宇杰, 等. 结合激光与视觉点云的古遗迹三维重建应用研究[J]. 中国激光, 2020, 47(11): 1110001.

[3] Natsume R, Saito S, Huang Z, et al. SiCloPe: silhouette-based clothed people[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4475-4485.

[4] Loper M, Mahmood N, Romero J, et al. SMPL: a skinned multi-person linear model[J]. ACM transactions on graphics, 2015, 34(6): 248.

[5] Chen J, Zhang Y Q, Song P, et al. Application of deep learning to 3D object reconstruction from a single image[J]. Acta Automatica Sinica, 2019, 45(4): 657-668.
陈加, 张玉麒, 宋鹏, 等. 深度学习在基于单幅图像的物体三维重建中的应用[J]. 自动化学报, 2019, 45(4): 657-668.

[6] Wu Z R, Song S R, Khosla A, et al. 3D ShapeNets: a deep representation for volumetric shapes[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1912-1920.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[8] Choy C B, Xu D F, Gwak J, et al. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 628-644.

[9] Wu J, Wang Y, Xue T, et al. Marrnet: 3D shape reconstruction via 2.5 D sketches[C]//Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. New York: Curran Associates, 2017: 540-550.

[10] Kato H, Ushiku Y, Harada T. Neural 3D mesh renderer[EB/OL]. (2017-11-20)[2020-10-09]. https://arxiv.org/abs/1711.07566v1.

[11] Sun Y, Liu Z, Wang Y, et al. Im2avatar: colorful 3d reconstruction from a single image[EB/OL]. (2018-04-17)[2020-10-09]. https://arxiv.org/abs/1804.06375.

[12] Kanazawa A, Tulsiani S, Efros A A, et al. Learning category-specific mesh reconstruction from image collections[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision - ECCV 2018. Lecture notes in computer science, Cham: Springer, 2018, 11219: 371-386.

[13] Zhou T H, Tulsiani S, Sun W L, et al. View synthesis by appearance flow[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 286-301.

[14] Liu S C, Chen W K, Li T Y, et al. Soft rasterizer: a differentiable renderer for image-based 3D reasoning[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea. New York: IEEE Press, 2019: 7707-7716.

[15] Wang N Y, Zhang Y D, Li Z W, et al. Pixel2Mesh: generating 3D mesh models from single RGB images[C]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[16] Kato H, Harada T. Learning view priors for single-view 3D reconstruction[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 9770-9779.

[17] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 3-19.

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-

23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[19] Mescheder L, Oechsle M, Niemeyer M, et al. Occupancy networks: learning 3D reconstruction in function space[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4455-4465.

[20] Fan H Q, Su H, Guibas L. A point set generation network for 3D object reconstruction from a single image[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2463-2471.