基于视图感知的单视图三维重建算法

王 年[①] 胡旭阳[①] 朱 凡[②] 唐 俊*[①]

[①](安徽大学电子信息工程学院 合肥 230031)

[②](起源人工智能研究院 阿布扎比 51133)

摘 要：尽管由于丢弃维度将3维(3D)形状投影到2维(2D)视图看似是不可逆的，但是从可视化到计算机辅助几何设计，各个垂直行业对3维重建技术的兴趣正迅速增长。传统基于物体深度图或者RGB图的3维重建算法虽然可以在一些方面达到令人满意的效果，但是它们仍然面临若干问题：(1)粗鲁的学习2D视图与3D形状之间的映射；(2)无法解决物体不同视角下外观差异所带来的的影响；(3)要求物体多个观察视角下的图像。该文提出一个端到端的视图感知3维(VA3D)重建网络解决了上述问题。具体而言，VA3D包含多邻近视图合成子网络和3D重建子网络。多邻近视图合成子网络基于物体源视图生成多个邻近视角图像，且引入自适应融合模块解决了视角转换过程中出现的模糊或扭曲等问题。3D重建子网络使用循环神经网络从合成的多视图序列中恢复物体3D形状。通过在ShapeNet数据集上大量定性和定量的实验表明，VA3D有效提升了基于单视图的3维重建结果。

关键词：视图感知；3维重建；视角转换；端到端神经网络；自适应融合

中图分类号：TN911.73; TP301.6 **文献标识码**：A **文章编号**：1009-5896(2020)12-3053-08

DOI: 10.11999/JEIT190986

Single-view 3D Reconstruction Algorithm Based on View-aware

WANG Nian[①] HU Xuyang[①] ZHU Fan[②] TANG Jun[①]

[①](*School of Electronic Information Engineering*, *Anhui University*, *Hefei* 230031, *China*)

[②](*Inception Institute of Artificial Intelligence*, *Abu Dhabi* 51133, *United Arab Emirates*)

Abstract: While projecting 3D shapes to 2D images is irreversible due to the abandoned dimension amid the projection process, there are rapidly growing interests across various vertical industries for 3D reconstruction techniques, from visualization purposes to computer aided geometric design. The traditional 3D reconstruction approaches based on depth map or RGB image can synthesize visually satisfactory 3D objects, while they generally suffer from several problems: (1)The 2D to 3D learning strategy is brutal-force; (2)Unable to solve the effects of differences in appearance from different viewpoints of objects; (3)Multiple images from distinctly different viewpoints are required. In this paper, an end-to-end View-Aware 3D (VA3D) reconstruction network is proposed to address the above problems. In particular, the VA3D includes a multi-neighbor-view synthesis sub-network and a 3D reconstruction sub-network. The multi-neighbor-view synthesis sub-network generates multiple neighboring viewpoint images based on the object source view, while the adaptive fusional module is added to resolve the blurry and distortion issues in viewpoint translation. The 3D reconstruction sub-network introduces a recurrent neural network to recover the object 3D shape from multi-view sequence. Extensive qualitative and quantitative experiments on the ShapeNet dataset show that the VA3D effectively improves the 3D reconstruction results based on single-view.

Key words: View-aware; 3D reconstruction; Viewpoint translation; End-to-end neural network; Adaptive fusional

收稿日期：2019-12-09；改回日期：2020-05-26；网络出版：2020-06-22

*通信作者： 唐俊 tangjunahu@163.com

基金项目：国家自然科学基金(61772032)

Foundation Item: The National Nature Science Foundation of China (61772032)

1 引言

单视图3维重建旨在从单张RGB图像中恢复物体的3D形状，这是一个极具挑战的任务，因为物体单视图受观察视角影响，往往会丢失大量的空间信息。为了解决这一挑战，众多已存在的3维重建方法都尝试在深度信息的辅助下恢复3D形状[1,2]。但是深度图的制作需要大量人力、物力的支撑且难以保证深度信息的准确性。与此同时，一些基于学习非线性转换的方法[3,4]，虽然在视觉上达到了令人满意的结果，但是仍然存在若干问题：(1)忽略了物体不同视角下的外观差异，仅依赖网络粗鲁的学习2D视图与3D形状之间的映射关系；(2)同一物体不同视角下的视图，3维重建结果存在较大的方差；(3)需要从多个视角观察物体以获得完整的空间结构。为了解决上述问题，文献[5]提出一种3D循环重建神经网络(3D Recurrent Reconstruction Neural Network, 3D-R2N2)学习从任意数量的2D视图中恢复物体的3D形状，但是该方法在物体仅有单视图作为网络输入时，重建效果并不理想。文献[6]在3D-R2N2基础上提出基于结构感知的3维重建(Structure-Aware 3D reconstruction, SA3D)方法，旨在通过深度生成模型学习特定视角下2D交叉视图间的映射关系，以此获得多视图，进而通过3D-R2N2重建物体形状，但是SA3D需要建立复杂的交叉视图映射，且输入只能是特定视角下的物体图像。

本文旨在提出一种基于视图感知的3维重建(View-Aware 3D reconstruction, VA3D)神经网络，它具有下列优点：(1)VA3D是一种可端到端训练的网络，模型输入为物体的单视图，输出为通过模型重建后的3D形状；(2)输入的单视图不受视角限制更加自由；(3)在重建过程中，可生成物体的多视图图像用于提升3维重建质量。具体而言，VA3D由多邻近视图合成子网络(Multiple-neighboring-view Synthesis sub-Network, MSN)和3D重建子网络级联而成。其中，MSN被用来建立从输入源视图到其邻近视角下图像的映射，在邻近视角图像生成的过程中，创新性地引入自适应融合模块通过学习不同输出策略间的掩膜来平衡输出图像存在的图像模糊和边缘扭曲等问题。而3D重建子网络在3D-R2N2的基础上修改了3D解码模块的跳跃连接方式，其被用于从合成的多视图序列中恢复物体的3D形状。

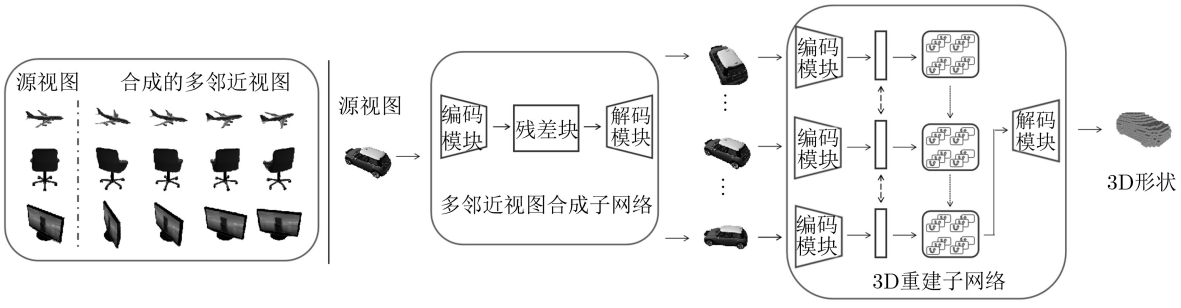
2 基于视图感知的3维重建网络

VA3D首先通过MSN生成与源视图视角邻近的多视图图像，进而通过3D重建子网络从合成的多视图序列中恢复物体的3D形状。图1(a)显示了通过MSN生成的多邻近视图序列，可以看到输出的多视图序列可以极大地恢复源视图受观察视角影响所丢失的物体空间信息。图1(b)显示了VA3D的网络流程。

2.1 问题描述

VA3D输入的RGB源视图被定义成$\boldsymbol{I}_s \in \mathbb{R}^{H\times W\times 3}$，$H$和$W$分别表示源视图的高和宽，该源视图可在物体任意视角下观察得到。令$\{\boldsymbol{I}_n\}^C=\{\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2},\boldsymbol{I}_{n_3},\boldsymbol{I}_{n_4}\}$表示相应同一物体下真实的多邻近视图集合。这些多邻近视图与源视图具有固定的视角差，例如$\boldsymbol{I}_{n_1}$和$\boldsymbol{I}_{n_2}$与源视图$\boldsymbol{I}_s$的角度差分别为顺时针$2\theta°$和$\theta°$，而$\boldsymbol{I}_{n_3}$和$\boldsymbol{I}_{n_4}$与源视图$\boldsymbol{I}_s$的角度差分别为逆时针$\theta°$和$2\theta°$，θ为预定义的投影偏转角度差值。

MSN为生成对抗网络(Generative Neural Network, GAN)[7–9]框架，由生成器G和判别器D组成。其中，生成器G旨在通过$\boldsymbol{I}_s$生成其相应的邻近视图集合$\left\{\tilde{\boldsymbol{I}}_n\right\}^C=\left\{\tilde{\boldsymbol{I}}_{n_1},\tilde{\boldsymbol{I}}_{n_2},\tilde{\boldsymbol{I}}_{n_3},\tilde{\boldsymbol{I}}_{n_4}\right\}$，而判别器$D$用于区分真实的多视图序列$\{\boldsymbol{I}\}^C=\{\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2},\boldsymbol{I}_s,\boldsymbol{I}_{n_3},\boldsymbol{I}_{n_4}\}$和“虚假”的多视图序列$\left\{\tilde{\boldsymbol{I}}\right\}^C=\left\{\tilde{\boldsymbol{I}}_{n_1},\tilde{\boldsymbol{I}}_{n_2},\boldsymbol{I}_s,\tilde{\boldsymbol{I}}_{n_3},\tilde{\boldsymbol{I}}_{n_4}\right\}$。然后，将生成的多视图序列输入进3D重建子网络中，得到实例化的重建体素结构$\boldsymbol{V}$。本文使用三元组$\{\boldsymbol{I}_s^m,\boldsymbol{I}_n^m,\boldsymbol{V}^m\}_{m=1}^M$训练VA3D模型，$M$表示训练集中样本的数量。注意，VA3D不需要任何源视图的类别标签或视角信息等作为先验条件。



(a) 源视图与合成的多邻近视图　　(b) 基于视图感知的3维重建网络流程

图 1 视图感知3维重建

2.2　多邻近视图合成子网络

2.2.1　网络结构

MSN的生成器通过将输入$\boldsymbol{I}_s$下采样到低分辨率，然后对高低两个分辨率下的图像进行卷积，该操作也在文献[10]中被证明可以提升生成图像的质量。此外，为了从多个邻近视角合成图像，在生成器中设计了4个输出分支，每个分支输出与源视图$\boldsymbol{I}_s$具有固定角度差的邻近视图，不同分支间共享编码模块和残差模块，但是每个分支都拥有独立的解码模块。图2显示了生成器的结构，其中Conv $n\times n$表示核尺寸为$n\times n$的卷积，IN表示实例标准化(Instance Normalization, IN)[11]，LReLU表示Leakey ReLU激活函数[12]。每个输出分支的最后有两个输出层，一个输出层用于直接回归目标图像$\left\{\tilde{\boldsymbol{I}}_r\right\}^{\mathrm{C}}=\left\{\tilde{\boldsymbol{I}}_{r_1},\tilde{\boldsymbol{I}}_{r_2},\tilde{\boldsymbol{I}}_{r_3},\tilde{\boldsymbol{I}}_{r_4}\right\}$，而另一个输出层则回归密集流场的偏移矢量$\left\{\tilde{\boldsymbol{F}}\right\}^{\mathrm{C}}=\left\{\tilde{\boldsymbol{F}}_1,\tilde{\boldsymbol{F}}_2,\tilde{\boldsymbol{F}}_3,\tilde{\boldsymbol{F}}_4\right\}$，该偏移矢量中每个位置的值表示相应位置目标图像像素点从源视图$\boldsymbol{I}_s$中采样的坐标位置，然后使用双线性采样核得到目标图像在各空间位置上的像素值。式(1)显示了目标图像$\tilde{\boldsymbol{I}}_f$在空间位置(p,q)上像素值的计算过程，其中$\boldsymbol{F}_x^{p,q}$和$\boldsymbol{F}_y^{p,q}$分别表示(p,q)位置的采样坐标x和y，N表示$(\boldsymbol{F}_y^{p,q},\boldsymbol{F}_x^{p,q})$的4像素领域。采样后的视图图像集合被定义为$\left\{\tilde{\boldsymbol{I}}_f\right\}^{\mathrm{C}}=\left\{\tilde{\boldsymbol{I}}_{f_1},\tilde{\boldsymbol{I}}_{f_2},\tilde{\boldsymbol{I}}_{f_3},\tilde{\boldsymbol{I}}_{f_4}\right\}$

$$\tilde{\boldsymbol{I}}_f^{p,q}=\sum_{(h,w)\in N}\boldsymbol{I}_s^{h,w}\max\left(0,1-\left|\boldsymbol{F}_y^{p,q}-h\right|\right)\cdot\max\left(0,1-\left|\boldsymbol{F}_x^{p,q}-w\right|\right)\tag{1}$$

自适应融合模块通过学习掩膜以平衡回归偏移矢量和直接生成图像两种输出策略的不足。具体而言，将通过两个输出分支得到的图像$\tilde{\boldsymbol{I}}_{r_i}$和$\tilde{\boldsymbol{I}}_{f_i}$沿着深度方向拼接在一起，作为该模块的输入，然后通过两个连续的Conv-IN-LReLU层，以及一个Conv-Sigmoid层得到预测的权重图$\tilde{\boldsymbol{I}}_{m_i}$，该权重图各个像素位置上的值表示相应位置$\tilde{\boldsymbol{I}}_{r_i}$和$\tilde{\boldsymbol{I}}_{f_i}$的权重，所有位置值的大小均在$(0,1)$之间。最后的输出目标视图$\tilde{\boldsymbol{I}}_{n_i}$由网络直接生成图像$\tilde{\boldsymbol{I}}_{r_i}$和通过回归偏移矢量采样得到的图像$\tilde{\boldsymbol{I}}_{f_i}$通过线性计算得到，其计算过程如式(2)所示，$\otimes$表示两矩阵逐元素相乘

$$\tilde{\boldsymbol{I}}_{n_i}=\tilde{\boldsymbol{I}}_{f_i}\otimes\tilde{\boldsymbol{I}}_{m_i}+\tilde{\boldsymbol{I}}_{r_i}\otimes\left(1-\tilde{\boldsymbol{I}}_{m_i}\right)\tag{2}$$

图3(a)显示了权重图的计算过程。直接生成图像的方法由于输出维度过高，网络很难捕捉到物体全部的细节变化，因此输出图像会比较模糊。而通过回归偏移矢量从源视图中进行双线性采样的方法，虽然得到的图像十分清晰但是在物体边缘处会产生严重的扭曲现象。基于这些观察，MSN在每个输出分支最后增加了自适应融合模块，以通过网络主动地学习自适应掩膜，最后通过线性插值计算得到目标邻近视角图像，图3(b)显示了该计算过程。在图3(b)中，可以看到最终的输出图像，无论是在边缘还是在清晰度上都要明显优于融合进程之前的输出图像，为后续基于合成多视图的3维重建提供了保障。

在判别器中，将集合$\left\{\tilde{\boldsymbol{I}}\right\}^{\mathrm{C}}$中的元素依序沿着深度通道拼接在一起作为判别器的输入。判别器可以引导生成器尽可能地输出与真实视图语义一致的目标图像，同时让真实视图序列和合成视图序列保持一致，既在单幅图像上加以约束，也在整个多视图序列上加以约束。此外，在判别器中还使用了空洞卷积，在参数量保持不变的同时，提升了网络的感受野，增加了生成器的上下文感知能力[13]。

2.2.2　损失函数

MSN联合多个损失函数对模型参数进行优

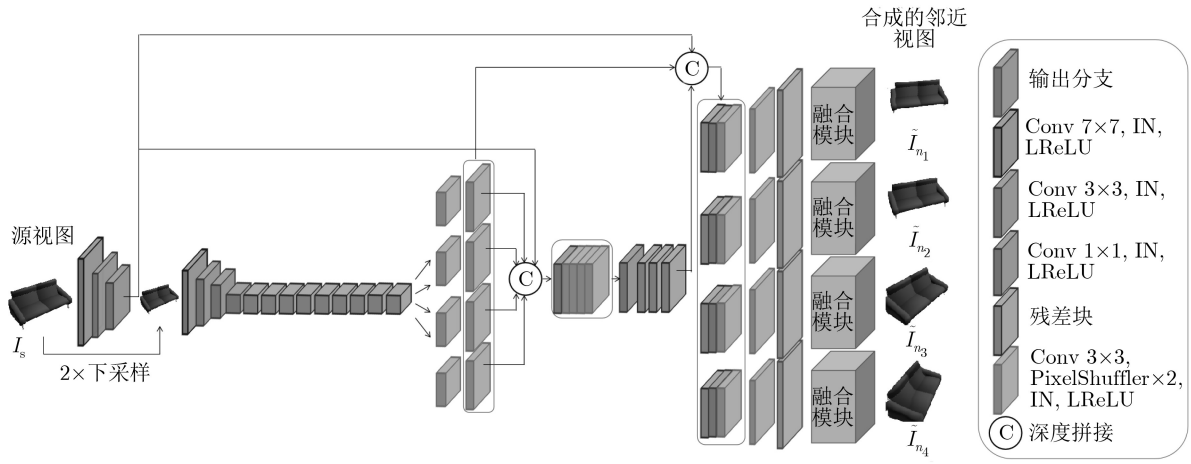
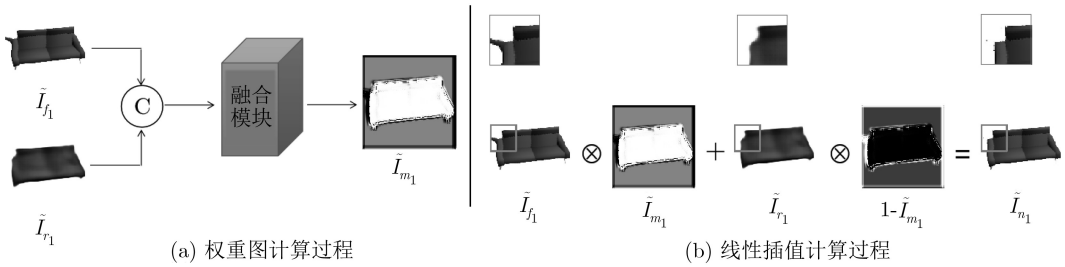


图 2 MSN生成器结构



(a) 权重图计算过程 (b) 线性插值计算过程

图 3 自适应融合

化，这些损失函数可以分为两大类，一类是像素级损失，另一类则是语义级损失。像素级的huber损失平衡了均方误差和平均绝对误差，其收敛速度更快且更加鲁棒。计算过程如式(3)所示。其中，超参数δ为平衡系数，在本文实验中均设置$\delta=1$

$$\mathcal{L}_{\text{rec}}=\mathbb{E}_{\boldsymbol{I}_{n_i}\in\{\boldsymbol{I}_n\}^{\text{C}},\tilde{\boldsymbol{I}}_{n_i}\in\{\tilde{\boldsymbol{I}}_n\}^{\text{C}}}\begin{cases}\dfrac{1}{2}\left(\boldsymbol{I}_{n_i}-\tilde{\boldsymbol{I}}_{n_i}\right)^2, & \left|\boldsymbol{I}_{n_i}-\tilde{\boldsymbol{I}}_{n_i}\right|\le\delta\\ \delta\left|\boldsymbol{I}_{n_i}-\tilde{\boldsymbol{I}}_{n_i}\right|-\dfrac{1}{2}\delta^2, & \left|\boldsymbol{I}_{n_i}-\tilde{\boldsymbol{I}}_{n_i}\right|>\delta\end{cases}\tag{3}$$

为了提供语义级别的约束，MSN使用了LSGAN(Least Squares GAN)[14]提出的对抗损失函数，并且对判别器进行梯度惩罚[15]以提升生成器与判别器训练的稳定性。对抗损失的计算过程如式(4)所示。此外，MSN还从判别器中提取多个中间层特征图以求特征匹配损失，计算过程如式(5)所示。其中，D^l表示提取判别器的第l层特征图

$$\mathcal{L}_{\text{adv}}^{G}=\mathbb{E}_{\{\tilde{\boldsymbol{I}}\}^{\text{C}}}\left[D\left(\left\{\tilde{\boldsymbol{I}}\right\}^{\text{C}}\right)-1\right]^2\tag{4}$$

$$\mathcal{L}_{\text{fm}}=\sum_l\left\|D^l\left(\{\boldsymbol{I}\}^{\text{C}}\right)-D^l\left(\left\{\tilde{\boldsymbol{I}}\right\}^{\text{C}}\right)\right\|_1\tag{5}$$

此外，MSN还使用感知损失[16]，即分别对合成图像和相应的真实图像提取预训练VGG19[17]的第2和第4激活函数ReLU层之后的特征图，并使用L_1计算两特征图间的距离，计算过程如式(6)所示。其中，VGG^l表示提取VGG19的第l层特征图

$$\mathcal{L}_{\text{per}}=\sum_l\mathbb{E}_{\boldsymbol{I}_{n_i}\in\{\boldsymbol{I}_n\}^{\text{C}},\tilde{\boldsymbol{I}}_{n_i}\in\{\tilde{\boldsymbol{I}}_n\}^{\text{C}}}\cdot\left\|\text{VGG}^l\left(\boldsymbol{I}_{n_i}\right)-\text{VGG}^l\left(\tilde{\boldsymbol{I}}_{n_i}\right)\right\|_1\tag{6}$$

最后，MSN总的损失函数为上述各损失函数的线性加权之和，如式(7)所示。其中，λ_1, λ_2, λ_3, λ_4和λ_5分别表示各项损失函数的系数，$\mathcal{L}_{\text{reg}}$表示对生成器参数求取的正则化损失

$$\mathcal{L}_{\text{total}}=\lambda_1\mathcal{L}_{\text{rec}}+\lambda_2\mathcal{L}_{\text{adv}}^{G}+\lambda_3\mathcal{L}_{\text{fm}}+\lambda_4\mathcal{L}_{\text{per}}+\lambda_5\mathcal{L}_{\text{reg}}\tag{7}$$

2.3 3D重建子网络

3D重建子网络由3个模块组成：2D编码，3D GRU(Gated Recurrent Unit)和3D解码模块，该网络在3D-R2N2的基础上优化了3D解码模块的跳跃连接方式。首先，将序列化的多视图图像输入进2D编码模块中，该编码模块通过共享参数的卷积层将输入的序列图像映射至统一的潜在特征空间中。然后，将编码模块的输出依次按照时间步送入3D GRU模块中，它可以记忆过去的信息并且随着新输入的特征而动态地更新内部记忆细胞的状态，有效地去除不同视角下图像间的冗余，减小图像生成过程中产生的噪声对3维重建的干扰。再将GRU最后的记忆细胞状态输入进3D解码模块中，通过解码得到物体的3维体素形状。

值得注意的是，本文的方法与3D-R2N2模型相比具有若干独特的优势：(1)充分利用合成多视图，克服单视图由于观察不足的内在限制；(2)降低了因视角差异对模型重建结果的影响；(3) MSN提供了一个灵活的接口，扩充了基于多视图3维重建算法的应用。

3 实验结果分析

3.1 模型训练细节

在所有实验中，模型输入均为128×128的RGB彩色图像，输出为$32\times32\times32$的3D体素形状。式(7)中的超参数被设置为$\lambda_1=1$, $\lambda_2=0.5$, $\lambda_3=1$, $\lambda_4=5\text{e}-4$及$\lambda_5=1\text{e}-6$。同时，为了端到端训练VA3D网络，在实验中使用了3个Adam优化器[18]，分别优化MSN生成器、判别器以及3D重建子网络，学习率分别被初始化为1e-4, 5e-5和1e-4，动量均为$\beta_1=0.5$, $\beta_2=0.999$。模型的迭代次数为30个epoch，批量大小(batch size)为12。所有实验均在NVIDIA P100 GPU上使用开源机器学习框架TensorFlow实现。

3.2 数据集和评价标准

ShapeNet数据集[19]是大型3D CAD模型存储库，已被广泛应用于3D重建[3–6]，3D姿态估计[20]和3D检索[21]等领域。所有实验均基于ShapeNet数据集的子集ShapeNetCore，共约51300个3D形状。本文获得了其中42105个3D形状，并按照4:1的比例随

机划分训练集与测试集。同时，使用开源工具包ShapeNet-Viewer投影3D形状到相应的2D视图。在相同的摄像机参数设置下，所有的2D视图均从24个不同视角渲染得到，即1个3D形状对应24个不同视角下的物体图像，图像两两之间的物体旋转角度差θ为15°。

此外，本文使用IoU(Intersection-over-Union)和F-score作为实验结果的衡量指标。IoU表示网络重建的3D体素形状与真实体素形状的交并比，F-score则表示3D体素块预测的精确度与召回率的调和值。

3.3 定性结果比较

图4显示了若干定性比较的示例样本，所有样本均来自测试集。同时作为基准，图4还提供了基于源视图(仅1个图像)的3D-R2N2重建结果。从中可以看到，VA3D通过MSN生成高质量的多视图图像不仅具有更好的视觉感受，而且与基准相比拥有更接近GT(Ground-Truth)的3D体素形状。具体而言，通过3D-R2N2重建的形状表面十分粗糙且在局部缺乏细节表现。例如，图4中第2行所示的汽车，第6行所示的板凳支撑腿等，其他的示例样本均显示了类似的提升。

3.4 定量结果比较

表1显示了测试集上各类别的比较结果，其中3D-R2N2_1和3D-R2N2_5分别表示以单视图和5个视图作为输入训练得到的模型。为了公平地比较，本文重新训练了3D-R2N2，并设置相同的体素化阈值τ=0.4。

从表1中可以看出，VA3D相较于3D-R2N2_1在所有类别上均有提升，且平均IoU值提升了7%左右。此外，在与3D-R2N2_5的对比中可以看到，VA3D虽然使用合成的5个邻近视图进行训练，但其重建结果与使用5个真实视图训练的3D-R2N2_5差别很小，甚至在一些类别上超越了3D-R2N2_5。这是因为3D重建子网络的输入为序列化的多视角图像而不是杂乱视角下的多视图，序列化的多视图

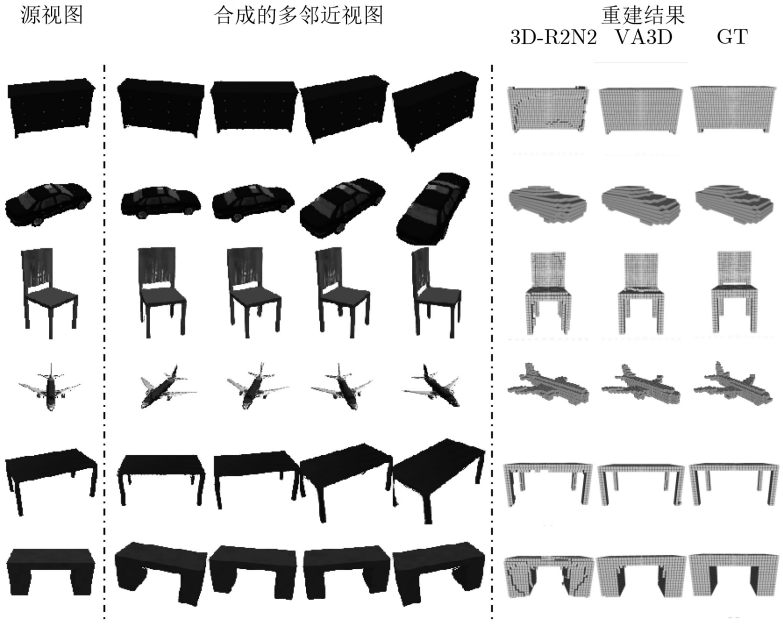


图 4 定性比较示例样本

表 1　定量比较结果

类别	IoU			*F*-score		
	3D-R2N2_1	3D-R2N2_5	VA3D	3D-R2N2_1	3D-R2N2_5	VA3D
柜子	0.7299	0.7839	**0.7915**	0.8267	0.8651	**0.8694**
汽车	0.8123	0.8551	0.8530	0.8923	0.9190	0.9178
椅子	0.4958	0.5802	0.5643	0.6404	0.7155	0.6995
飞机	0.5560	0.6228	**0.6385**	0.7006	0.7561	**0.7641**
桌子	0.5297	0.6061	**0.6128**	0.6717	0.7362	**0.7386**
长凳	0.4621	0.5566	0.5533	0.6115	0.6991	0.6936
平均	0.5976	0.6674	**0.6689**	0.7238	0.7818	0.7805

依据视角信息排列输入进3D重建子网络中，减小了杂乱视角下物体外观差异过大对模型的影响。

此外，本文还对比了另一种基于合成多视图的3维重建算法-SA3D。为了公平地比较，本文在与SA3D相同的类别上重新训练VA3D模型。从表2中可以看出，VA3D的平均IoU值高出6%左右。VA3D相较于SA3D对输入图像的视角要求更加灵活，无需特定视角图像作为网络输入，如图5所示。VA3D通过引入生成对抗机制，使得生成器输出的图像多了对抗损失的约束，不同于像素级别的损失，该损失从语义上对目标图像进行监督，从而能够在感知上提升输出图像的质量。同时，自适应融合模块的设计使得模型可以通过计算掩膜融合两种不同输出策略得到的图像，解决了目标图像中存在的图像模糊和边缘扭曲等问题。

表 2 对比SA3D算法结果

算法	平均IoU
SA3D	0.6162
VA3D	**0.6741**

图 5 SA3D与VA3D生成的多视图对比

表 3 MSN中不同输出策略的影响

模型	SSIM	PSNR	IoU	*F*-score
仅使用$\{\tilde{\boldsymbol{I}}_r\}^{\mathrm{C}}$	0.8035	19.8042	0.6525	0.7649
仅使用$\{\tilde{\boldsymbol{I}}_f\}^{\mathrm{C}}$	0.8435	20.5273	0.6530	0.7646
自适应融合	**0.8488**	**20.6203**	**0.6554**	**0.7672**

3.5 消融研究

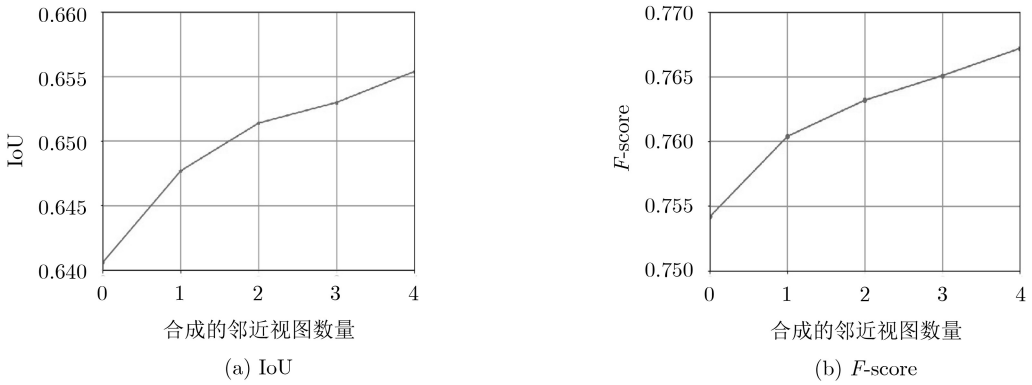
在2.2节中，介绍了使用掩膜融合两种不同输出策略获得视图的方法。为了验证该方法的效果，本节去除了自适应融合块，分别在两种不同输出策略中训练/测试模型，并使用峰值信噪比(Peak Signal to Noise Ratio, PSNR)和结构相似性(Structural SIMilarity, SSIM)两个指标来衡量合成视图的质量。实验结果如表3所示，可以看到所提出的自适应融合方法不仅提升了合成视图的质量，而且还提升了3维重建结果。

图6显示了不同数量的合成视图对重建性能的影响，可以看到随着合成视图数量的增加，3维重建结果也越来越好。由于受到计算机显存和计算时间等限制，本文最多只合成4个视图。表4显示了同一物体不同视图下重建结果的方差，可以看到合成视图能够减小同一物体不同视角下重建结果的差异，通过分析可以得出合成多视图能够有效降低物体视角或外观变化对模型重建所造成的影响。

表5列出了不同损失函数组合下模型的评估结果。从表5中可以看到，当去掉$\mathcal{L}_{\mathrm{rec}}$和$\mathcal{L}_{\mathrm{per}}$不仅会降低合成视图的质量，而且还会降低3维重建的结果。此外，去掉$\mathcal{L}_{\mathrm{adv}}$虽然在图像结果的评价标准上获得了提升，但由于缺乏语义级的图像约束，相应的3维重建结果也出现了降低。

4 结束语

本文提出了一种可端到端训练的视图感知3维重建网络，解决了传统3维重建方法对输入数据的严格要求。此外，还设计了多邻近视图合成子网络，该网络可以学习物体2D图像的视角转换，基于源视图并行的生成物体多个邻近视角下的图像。为了保证生成高质量的图像，多邻近视图合成子网络引入自适应融合模块以联合不同输出策略的优



(a) IoU (b) *F*-score

图 6 不同合成视图数量的IoU和*F*-score

表 4　重建结果的方差

模型	σ^2_{IoU}	$\sigma^2_{F-\text{score}}$
合成视图数量=0	0.0057	0.0061
合成视图数量=4	**0.0051**	**0.0054**

表 5　不同损失函数的组合

模型	SSIM	PSNR	IoU	*F*-score
无重建损失$\mathcal{L}_{\text{rec}}$	0.8462	20.2693	0.6540	0.7658
无对抗损失$\mathcal{L}_{\text{adv}}$	**0.8516**	**21.4385**	0.6539	0.7651
无感知损失$\mathcal{L}_{\text{per}}$	0.8416	20.3141	0.6525	0.7645
全部损失	0.8488	20.6203	**0.6554**	**0.7672**

点。最后，通过在ShapeNet数据集上进行广泛定性和定量的实验表明，本文所提出的方法提升了基于单视图的3维重建结果，且有效降低了物体图像因视角差异所带来的外观变化对3维重建过程的影响。

参 考 文 献

[1] EIGEN D, PUHRSCH C, and FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]. The 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2366–2374.

[2] WU Jiajun, WANG Yifan, XUE Tianfan, *et al*. Marrnet: 3D shape reconstruction via 2.5D sketches[C]. The 31st Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 540–550.

[3] WANG Nanyang, ZHANG Yinda, LI Zhuwen, *et al*. Pixel2mesh: Generating 3D mesh models from single RGB images[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 55–71. doi: 10.1007/978-3-030-01252-6_4.

[4] TANG Jiapeng, HAN Xiaoguang, PAN Junyi, *et al*. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single RGB images[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4536–4545. doi: 10.1109/cvpr.2019.00467.

[5] CHOY C B, XU Danfei, GWAK J Y, *et al*. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction[C]. The 14th European Conference on Computer Vision, Amsterdam, the Netherlands, 2016: 628–644. doi: 10.1007/978-3-319-46484-8_38.

[6] HU Xuyang, ZHU Fan, LIU Li, *et al*. Structure-aware 3D shape synthesis from single-view images[C]. 2018 British Machine Vision Conference, Newcastle, UK, 2018.

[7] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, *et al*. Generative adversarial nets[C]. The 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2672–2680.

[8] 张惊雷, 厚雅伟. 基于改进循环生成式对抗网络的图像风格迁移[J]. 电子与信息学报, 2020, 42(5): 1216–1222. doi: 10.11999/JEIT190407.

ZHANG Jinglei and HOU Yawei. Image-to-image translation based on improved cycle-consistent generative adversarial network[J]. *Journal of Electronics & Information Technology*, 2020, 42(5): 1216–1222. doi: 10.11999/JEIT190407.

[9] 陈莹, 陈湟康. 基于多模态生成对抗网络和三元组损失的说话人识别[J]. 电子与信息学报, 2020, 42(2): 379–385. doi: 10.11999/JEIT190154.

CHEN Ying and CHEN Huangkang. Speaker recognition based on multimodal generative adversarial nets with triplet-loss[J]. *Journal of Electronics & Information Technology*, 2020, 42(2): 379–385. doi: 10.11999/JEIT190154.

[10] WANG Tingchun, LIU Mingyu, ZHU Junyan, *et al*. High-resolution image synthesis and semantic manipulation with conditional gans[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8798–8807. doi: 10.1109/cvpr.2018.00917.

[11] ULYANOV D, VEDALDI A, and LEMPITSKY V. Instance normalization: The missing ingredient for fast stylization[EB/OL]. https://arxiv.org/abs/1607.08022, 2016.

[12] XU Bing, WANG Naiyan, CHEN Tianqi, *et al*. Empirical evaluation of rectified activations in convolutional network[EB/OL]. https://arxiv.org/abs/1505.00853, 2015.

[13] GOKASLAN A, RAMANUJAN V, RITCHIE D, *et al*. Improving shape deformation in unsupervised image-to-image translation[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 662–678. doi: 10.1007/978-3-030-01258-8_40.

[14] MAO Xudong, LI Qing, XIE Haoran, *et al*. Least squares generative adversarial networks[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2813–2821. doi: 10.1109/iccv.2017.304.

[15] GULRAJANI I, AHMED F, ARJOVSKY M, *et al*. Improved training of wasserstein GANs[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 5767–5777.

[16] LEDIG C, THEIS L, HUSZÁR F, *et al*. Photo-realistic single image super-resolution using a generative adversarial network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 105–114. doi: 10.1109/CVPR.2017.19.

[17] SIMONYAN K and ZISSERMAN A. Very deep

convolutional networks for large-scale image recognition[EB/OL]. https://arxiv.org/abs/1409.1556, 2014.

[18] KINGMA D P and BA J. Adam: A method for stochastic optimization[EB/OL]. https://arxiv.org/abs/1412.6980, 2014.

[19] CHANG A X, FUNKHOUSER T, GUIBAS L, *et al*. Shapenet: An information-rich 3D model repository[EB/OL]. https://arxiv.org/abs/1512.03012, 2015.

[20] GRABNER A, ROTH P M, and LEPETIT V. 3D pose estimation and 3D model retrieval for objects in the wild[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 3022–3031. doi: 10.1109/cvpr.2018.00319.

[21] HE Xinwei, ZHOU Yang, ZHOU Zhichao, *et al*. Triplet-center loss for multi-view 3D object retrieval[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1945–1954. doi: 10.1109/cvpr.2018.00208.

王 年：男，1966年生，教授，博士，主要从事模式识别与图像处理等方面的研究.

胡旭阳：男，1995年生，硕士生，研究方向为图像生成和3维重建.

朱 凡：男，1987年生，博士，主要从事计算机视觉方面的研究.

唐 俊：男，1977年生，教授，博士，主要从事模式识别与计算机视觉等方面的研究.

责任编辑：余 蓉