激光与光电子学进展 先进成像

基于自注意力机制的多视图三维重建方法

朱光照，韦博*，杨阿峰，徐欣
杭州电子科技大学通信工程学院，浙江 杭州 310037

摘要 多视图立体匹配是计算机视觉领域的一大研究热点，针对目前多视图立体重建完整性差、无法处理高分辨率图像和GPU内存消耗巨大、运行时间长等问题，提出一种基于自注意力机制的深度学习网络(SA-PatchmatchNet)。首先通过特征提取模块提取图像特征，再将其送入可学习的Patchmatch模块中，得到深度图，并对深度图进行优化，生成最终的深度图。为了捕捉深度推理任务中的重要信息，将自注意力机制融入到特征提取模块，提高了网络的特征提取能力。实验结果表明，SA-PatchmatchNet在Technical University of Denmark(DTU)数据集上进行测试，与PatchmatchNet相比，重建的完整性提升5.8%，整体性提升2.3%，与其他的state-of-the-art(SOTA)方法相比，完整性和整体性都得到了较大的提升。

关键词 深度学习；三维重建；多视图立体；自注意力机制
中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP222692

Multi-View 3D Reconstruction Method Based on Self-Attention Mechanism

Zhu Guangzhao, Wei Bo*, Yang Afeng, Xu Xin
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310037, *Zhejiang, China*

Abstract Multi-view stereo matching is a major hotspot in the field of computer vision. We propose a self-attention-based deep learning network (SA-PatchmatchNet) to address the issues of poor completeness of multi-view stereo reconstruction, inability to process high-resolution images, huge GPU memory consumption, and long running time. First, the feature extraction module extracted the image features and sent them to the learnable Patchmatch module to obtain the depth map, and then the depth map was optimized to generate the final depth map. Moreover, the self-attention mechanism was integrated into the feature extraction module to capture the important information in the deep reasoning task, thereby enhancing the network feature extraction ability. The experimental results show that the reconstruction completeness is improved by 5.8% and the entirety is improved by 2.3% compared with that of the PatchmatchNet when the SA-PatchmatchNet is tested on the Technical University of Denmark (DTU) dataset. The completeness and entirety of the proposed network are significantly improved compared with that of the other state-of-the-art (SOTA) methods.

Key words deep learning; 3D reconstruction; multi-view stereo; self-attention mechanism

1 引 言

多视图立体(MVS)匹配[1-2]是一种用校准后的摄像机根据在多个视角下获取到的图像来恢复3D场景的技术。作为摄影测量和计算机视觉的核心问题，多视图立体匹配已经被研究了好多年，传统的MVS方法在理想的环境中(朗伯曲面和丰富的纹理)取得了巨大成功，但它仍然是个挑战，因为在实际中存在着弱纹理、非朗伯曲面、光度变化等情况，导致重建效果不佳，故而传统MVS方法仍然有很大的进步空间。随着卷积神经网络(CNN)在一些领域获得巨大成功，例如场景理解[3]、语义分割[4]和立体匹配[5]，一些学者将CNN用于MVS方法，得到了出色的结果。基于学习的三维重建方法能够加入全局语义信息，例如反射先验和镜面反射，最后能得到鲁棒性更强的匹配，解决了传统方法无法克服的困难。

现阶段基于学习的MVS方法[6]主要有4种：基于体素的方法、基于网格的方法、基于点云的方法、基于

收稿日期：2022-10-08；**修回日期**：2022-11-13；**录用日期**：2022-11-24；**网络首发日期**：2023-01-04
通信作者：*weibo@hdu.edu.cn

深度图的方法。其中基于深度图的方法是该领域最热门的方向且是表现最好的。该类方法的流程：深度图初始化、匹配代价计算、代价聚合、深度估计、深度图优化，关键步骤是构建像素级匹配成本体积。提到基于深度图的学习方法，就不得不提香港科技大学权龙团队提出的MVSNet[7]，后续的一些网络都从此网络创新而来。它在深度范围内利用相机建立视锥体，基于相机视锥体进行成本体积的构建，然后使用3D卷积将成本体积正则化来进行深度图的预测。此网络的突出贡献是利用可微单应性扭曲（warp）将相机几何嵌入到网络中，帮助实现端到端，缺点是显存消耗过高，适用于估算低分辨率深度图。随后该团队改进MVSNet，提出R-MVSNet[8]，它使用门控循环单元（GRU）顺序正则化代价体，适用于大场景三维重建，但牺牲了运行时间。Yang等[9]提出的CVP-MVSNet输入的是金字塔式的图片，使用由粗到细的策略构建金字塔代价体，虽然减少了内存，但同样增加了运行时间。Wang等[10]提出PatchmatchNet，引入多尺度Patchmatch并改进核心算法，内存提升较大，但是其特征提取模块无法捕获深度推理任务的重要信息。

综上所述，近些年各种基于深度学习的三维重建state-of-the-art（SOTA）方法在实际应用中仍然面临着一些问题。为了解决上述问题，本文提出深度学习网络（SA-PatchmatchNet）。在PatchmatchNet基础上，在特征提取模块嵌入自注意力层，获得重要的深度信息，另外借助由粗到细的深度推断策略，可以有效解决大场景应用、重建完整性低、内存消耗高和运行时间长等问题。

2　网络模型

具体阐述SA-PatchmatchNet结构，它以一幅参考图像和多幅源图像作为输入，专注于每次为一个参考图像生成深度图。结构创新点在于将自注意力层引入特征提取网络，通过捕获全局信息获得更大的感受野和上下文信息，提高了特征提取能力，获得了更优的深度图质量。所提网络模型主要包括的模块有：自注意力机制的多尺度特征提取；从粗到细的可学习的Patchmatch，以实现匹配代价计算、成本聚合和深度回归；深度图优化。网络的具体结构如图1所示。

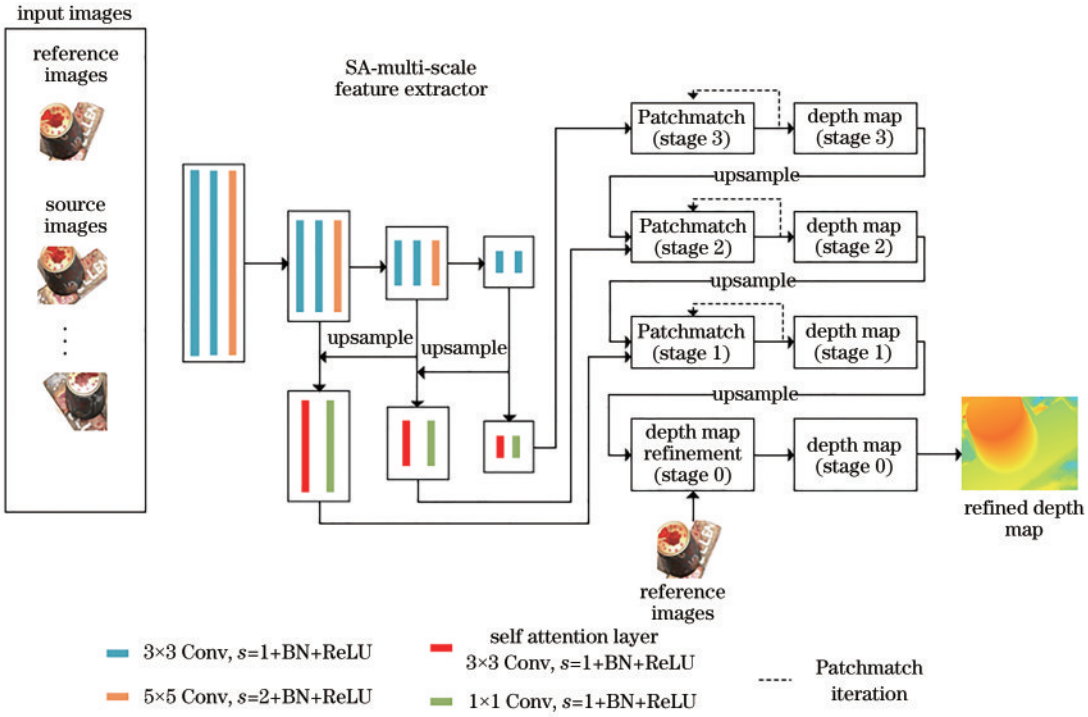

图1　SA-PatchmatchNet整体结构

Fig. 1　Overall structure of SA-PatchmatchNet

2.1　自注意力机制的多尺度特征提取

自注意力多尺度特征提取（SA-MsFe）模块如图1的中间部分所示，整体架构采用特征金字塔网络（FPN）[11]，由卷积层和自注意力层两部分组成。特征提取具体流程包括两步：从输入的参考图像$\boldsymbol{I}_0$和源图像$\{\boldsymbol{I}_i\}_{i=1}^{N-1}$以多分辨率提取像素级特征，对应图1中间部分的上半层；在最低分辨率级别，将像素特征输入到自注意力层，并在之后再次进行卷积，将最终的像素特征输入到可学习的Patchmatch模块对应阶段。期间上采样与上一层卷积后形成的特征进行融合，融合结果作为具有更高分辨率级别的自注意力层的输入，对应图1中间部分的下半层。

与其他MVS网络的特征提取网络不同的是，所提网络在FPN中嵌入了自注意力机制，获得了深度关键信息。自注意力（self-attention）[12]是一类可以把单个序列的各个位置联系起来，目的是计算该序列表示

形式的注意力机制。自注意力层结构如图 2 所示，它被定义为适用于单个语境而不是跨多个语境的注意力，可以直接对依赖项进行构造，而不用考虑其在输入或输出序列上的位置。注意力功能可以描述为将查询和一组键值对映射到输出，其中查询、键、值与输出都是向量。输出通过值的加权和计算得到，其中分配到每个值的权重通过查询和对应键的兼容性函数得到，具体描述为

$$\boldsymbol{y}_{uv}=\sum_{m,n\in B}\mathrm{Softmax}_{mn}\left(\boldsymbol{q}_{uv}^{\mathrm{T}}\boldsymbol{k}_{mn}\right)\boldsymbol{v}_{mn},\tag{1}$$

式中：$\boldsymbol{q}_{uv}=\boldsymbol{W}_{\mathrm{Q}}\boldsymbol{x}_{uv}$ 表示查询；$\boldsymbol{k}_{mn}=\boldsymbol{W}_{\mathrm{K}}\boldsymbol{x}_{mn}$ 表示键；$\boldsymbol{v}_{mn}=\boldsymbol{W}_{\mathrm{V}}\boldsymbol{x}_{mn}$ 表示值；矩阵 $\boldsymbol{W}_{p}\left(p=\mathrm{Q},\mathrm{K},\mathrm{V}\right)$ 是一个学习的权重矩阵，由学习参数组成；B 是用于卷积计算的图像块，与卷积核大小一样；$\boldsymbol{y}_{uv}$ 是输出。式(1)包括 3 步：

1)计算查询$\left(\boldsymbol{q}_{uv}\right)$、键$\left(\boldsymbol{k}_{mn}\right)$和值$\left(\boldsymbol{v}_{mn}\right)$；

2) 通过计算查询和键的内积$\left(\boldsymbol{q}_{uv}^{\mathrm{T}}\boldsymbol{k}_{mn}\right)$来测量它们的相似性，然后使用 Softmax 操作将相似性映射到(0,1)；

3) 对步骤 2)的相似性值进行加权，并对 B 中的每个像素重复所有步骤，然后将所有输出相加。

矩阵 $\boldsymbol{W}_{\mathrm{Q}}$ 用于沿着 $\boldsymbol{x}_{mn}$ 周围的所有通道提取像素信息。矩阵 $\boldsymbol{W}_{\mathrm{K}}$ 沿所有通道在 $\boldsymbol{x}_{mn}$ 提取信息，因此矩阵 $\boldsymbol{W}_{\mathrm{Q}}$ 和矩阵 $\boldsymbol{W}_{\mathrm{K}}$ 用于相似性度量。矩阵 $\boldsymbol{W}_{\mathrm{V}}$ 用作线性变换，将 $\boldsymbol{x}_{mn}$ 从输入通道映射到输出通道。

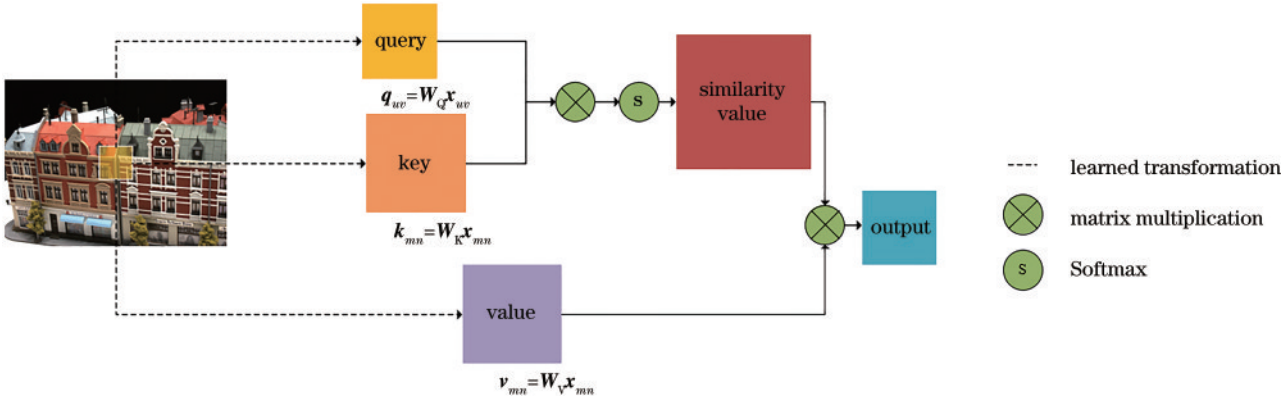

图 2　自注意力层内部结构

Fig. 2　Internal structure of self-attention layer

2.2　可学习的 Patchmatch

传统 Patchmatch 算法[13]能快速查找图像块之间的近似最近邻匹配，通过初始化、代价传播、随机搜索等迭代步骤快速找到匹配像素块。继传统的 Patchmatch 算法，可学习的 Patchmatch 内部结构如图 3 所示，仍然包括 3 个主要的步骤：

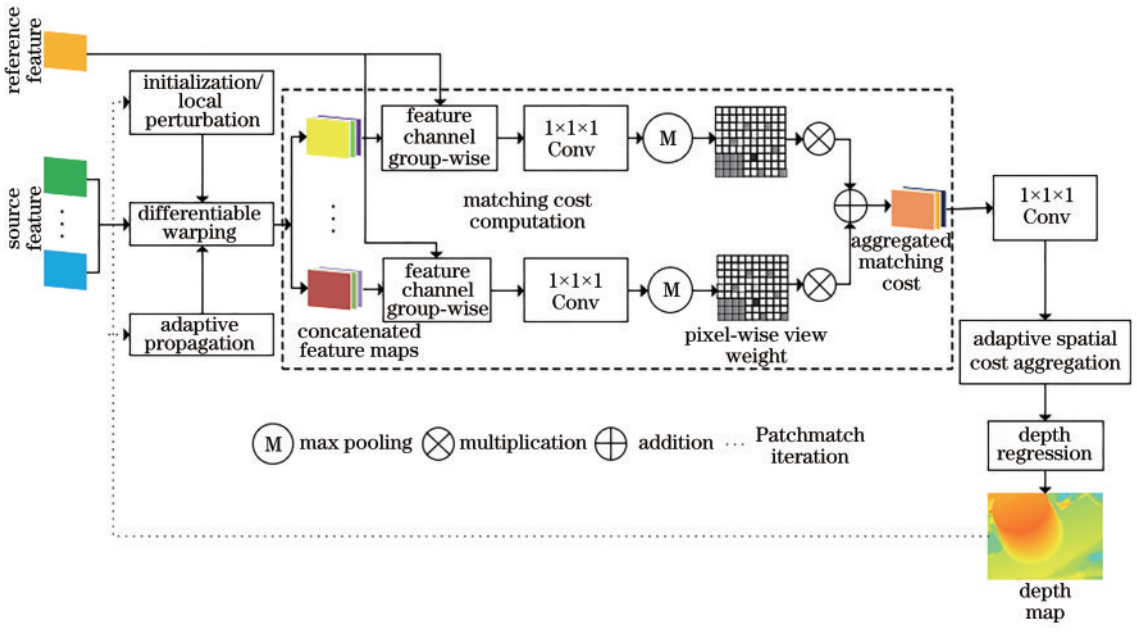

图 3　可学习的 PatchmatchNet 结构

Fig. 3　Learnable PatchmatchNet structure

1)深度初始化,生成随机的深度假设;

2)传播,将深度假设传播给邻域像素;

3)评估,计算所有深度假设的匹配代价,选择最佳解决方案。

此结构是网络的核心,经过特征提取后,输入特征图,经过深度图初始化、匹配代价计算、代价聚合、深度估计后生成深度图。

2.2.1 深度初始化

由于最初没有深度图,所以在预设的深度范围[$d_{\min}$, $d_{\max}$]的逆深度范围,每个像素随机初始化48个深度层,并在该范围内划分48个区间,为保证每个区间有一个假设覆盖,每个区间随机采样,就相当于有这么多深度假设值。在有了初始深度图后,对于每个阶段上的后续迭代,在一个归一化的逆深度范围内,均匀地给每个像素再假设几个深度值来进行局部扰动,在接下来更精细的阶段减小上述逆深度范围,这样可以局部优化结果并纠正错误的估计值。

2.2.2 自适应传播

由于邻域像素的深度假设值可能比当前像素深度假设值更好,因此考虑邻域像素的深度信息。将邻域像素的深度值传播到当前像素,也作为当前像素的假设值,这样使得估计出的深度值更加精确。但是邻域像素和当前像素可能不处于同一平面,邻域像素深度假设值传播过来可能会起到反作用,所以采用自适应传播。基于一个2D可形变卷积[14],以参考特征图为输入,计算当前像素同一表面的邻近点的像素坐标与当前像素坐标的额外偏移量,具体为

$$D_p(\boldsymbol{p})=\left\{D\left[\boldsymbol{p}+\boldsymbol{o}_k+\Delta o_k(\boldsymbol{p})\right]\right\}_{k=1}^{K_e}, \tag{2}$$

式中:$\boldsymbol{o}_k$是固定偏移量;$\Delta\boldsymbol{o}_k$是学习的额外偏移量;K_e是邻域像素数;$\boldsymbol{D}$是来自上一次迭代的深度图;$D_p(\boldsymbol{p})$是得到的深度假设。

2.2.3 匹配代价体的构建

利用可微单应性扭曲,将源图像特征图扭曲到参考视图对应的深度层下,利用参考视图和源视图的相机参数$\boldsymbol{K}$、$\boldsymbol{R}$、$\boldsymbol{t}$,计算源特征图上像素$\boldsymbol{p}$所对应的像素,进而得到warp后的源特征图,然后计算匹配代价,具体为

$$\boldsymbol{p}_{i,j}=\boldsymbol{K}_i\cdot\left[\boldsymbol{R}_{0,i}\cdot\left(\boldsymbol{K}_0^{-1}\cdot\boldsymbol{p}\cdot\boldsymbol{d}_j\right)+\boldsymbol{t}_{0,i}\right], \tag{3}$$

式中:$\boldsymbol{K}$是相机内参;$\boldsymbol{R}$是旋转矩阵;$\boldsymbol{t}$是平移向量;$\boldsymbol{R}_{0,i}$是相对旋转;$\boldsymbol{t}_{0,i}$是相对平移;$\boldsymbol{d}_j$是深度假设;$\boldsymbol{p}_{i,j}$是第i个源视图第j个深度假设扭曲后的像素。

通过上述变换,所有源视图深度特征扭曲为参考图像深度特征的坐标。匹配代价的计算必须将来自任意数量的源视图的信息集成到单个像素和深度假设的匹配代价。考虑到GPU内存消耗问题,使用分组内积的方式计算每个深度假设的代价。通过将特征通道分为G组,减少内存。第g组的相似性为

$$S_i(\boldsymbol{p},j)^g=\frac{1}{C_1/G}\left\langle F_0(\boldsymbol{p})^g,F_i(\boldsymbol{p}_{i,j})^g\right\rangle, \tag{4}$$

式中:$F_0(\boldsymbol{p})$是参考特征图特征;$F_i(\boldsymbol{p}_{i,j})$是扭曲后的源特征图特征;$\langle,\rangle$表示内积;$C_1$是特征通道数;$S_i(\boldsymbol{p},j)^g$是第$g$组的相似性。

使用像素视图权重对$N-1$个视图进行聚合,由于Patchmatch第3阶段第1次迭代深度假设丰富,所以在此计算权重并固定。采用初始相似度集,利用Sigmoid激活函数输出每个像素0到1之间的数字,计算加权平均值,具体为

$$\bar{S}(\boldsymbol{p},j)=\frac{\sum_{i=1}^{N-1}w_i(\boldsymbol{p})\cdot S_i(\boldsymbol{p},j)}{\sum_{i=1}^{N-1}w_i(\boldsymbol{p})}, \tag{5}$$

式中:$w_i(\boldsymbol{p})$是像素$\boldsymbol{p}$在其所有深度假设中最大的可见性权重;$S_i(\boldsymbol{p},j)$是每组相似度;$\bar{S}(\boldsymbol{p},j)$是像素$\boldsymbol{p}$和第$j$个假设的最终加权平均后的每组相似性。将所有像素合成$\bar{S}(\boldsymbol{p},j)$通过$1\times1\times1$卷积核的3D卷积,获得每像素和深度假设的单个成本。

2.2.4 自适应空间代价聚合

与自适应传播类似,同样考虑到在聚合成本时邻域像素与中心像素位于不同表面的问题,为了防止跨表面边界聚合,使用一个2D CNN计算同一平面邻域像素的坐标偏移量,然后计算邻域像素的坐标和深度权重,使用一个3D CNN计算邻域像素与中心像素的特征相似性,具体为

$$\bar{C}(\boldsymbol{p},j)=\frac{\sum_{k=1}^{K_e}w_k d_k C(\boldsymbol{p}+\boldsymbol{p}_k+\Delta\boldsymbol{p}_k,j)}{\sum_{k=1}^{K_e}w_k d_k}, \tag{6}$$

式中:w_k是特征相似性权重;d_k是深度相似性权重;$\boldsymbol{p}_k$是固定的像素偏移,$\Delta\boldsymbol{p}_k$是额外的偏移;C是匹配代价;$\bar{C}(\boldsymbol{p},j)$是聚合后的代价。

2.2.5 深度回归

利用Softmax将代价体转换为概率体,并基于每个像素的概率分布,计算期望实现深度回归,具体为

$$D(\boldsymbol{p})=\sum_{j=0}^{M-1}\boldsymbol{d}_j\cdot P(\boldsymbol{p},j), \tag{7}$$

式中:$D(\boldsymbol{p})$是深度值;$P(\boldsymbol{p},j)$是像素$\boldsymbol{p}$在第j个深度假设下的概率;M是像素$\boldsymbol{p}$的深度假设数。

2.3 深度图优化

从低分辨率深度图进行上采样时,深度边界可能会出现过度平滑现象。由于输入的参考图像含有边界信息,因此可以利用参考图像引导并优化Patchmatch输出的深度图。所以,在该模块中,使用一种深度残差网络(ResNet),残差网络学习并输出残差,然后将残差值与Patchmatch输出的深度图相加,最后得到优化后的更优的深度图。

2.4 损失函数

损失函数衡量 Patchmatch 每个阶段估计出的深度图和 Technical University of Denmark(DTU)数据集官方提供的真值(GT)之间的损失以及经过深度图优化后和 GT 之间的损失，采用 smooth L1 损失，具体为

$$L=\sum_{k'=1}^{3}\sum_{i'=1}^{n_{k'}}L_{i'}^{k'}+L_{\mathrm{ref}}^{0}, \tag{8}$$

即在阶段 $k'\left(k'=1,2,3\right)$ 上的 Patchmatch 第 i' 次迭代输出的深度图与 GT 之间的损失 $L_{i'}^{k'}$ 和经过优化后输出的深度图与 GT 之间的损失 L_{ref}^{0} 之和为最终的损失。

3 实验内容

3.1 数据集

使用的数据集为 DTU 数据集[15]，是丹麦理工大学提供的处理多视图立体视觉任务的数据集。此数据集包含各不相同的 124 个场景，每个场景都包括范围广泛的对象，每个场景有 49 或 64 个位置，RGB 图像的数量为 49 或 64 张，而且有 7 种不同强度的光照情况。数据集中图像的像素分辨率是 1600×1200，相机内参和外参通过 Matlab 校准工具箱得到。与其他深度学习一样，本文将此数据集分为训练集、验证集、测试集。验证集有 18 个场景，具体为[3, 5, 17, 21, 28, 35, 37, 38, 40, 43, 56, 59, 66, 67, 82, 86, 106, 117]，测试集有 22 个场景，具体为[1, 4, 9, 10, 11, 12, 13, 15, 23, 24, 29, 32, 33, 34, 48, 49, 62, 75, 77, 110, 114, 118]，训练集为剩下的 78 个场景。

3.2 实验环境

硬件环境：CPU 为 AMD EPYC、内存 64G、8 核。显卡为 NVIDIA GeForce RTX 3090×1，显存 24G。软件环境：Ubuntu20.04.3 LTS、Python3.8、PyTorch1.10.1、CUDA11.1、cuDNN8.0.5。

3.3 实现细节

在训练阶段使用 640×512 的图像分辨率，测试阶段采用 1600×1200 的图像分辨率。考虑到 GPU 消耗，测试阶段将 batch size 设为 1，输入 5 张图片，即每个 batch size 有 1 个参考图像和 4 个源图像。根据 DTU 数据集提供的图像数据，一共训练了 27097(49×7×79)个图像(一个场景 49 个位置，7 种不同的光照强度，总共 79 个场景)，测试 7546(49×7×22)个图像。根据 DTU 数据集，预设的深度假设范围是[425 mm, 935 mm]。第 3、2、1 阶段 Patchmatch 迭代次数分别设置为 2、2、1，随机扰动的个数在阶段 3、2、1 设为 16、8、8，自适应传播的邻域数设置为 16、8、0。自适应空间代价聚合的所有阶段自适应匹配成本聚合的邻域数设置为 9。网络学习率初始值设为 0.001，待训练到 10、12、14 个 epoch 时，对学习率分别进行减半操作，共训练 16 个 epoch。使用 Adam 优化器[16]，β_1=0.9，β_2=0.999。

3.4 实验结果

为了证明所提方法的有效性，对所提方法与传统方法 Camp[17]、Furu[18]、Tola[19]、Gipuma[20]，基于学习的三维重建方法 MVSNet、R-MVSNet、P-MVSNet[21]、Point-MVSNet[22]、Fast-MVSNet[23]、CasMVSNet[24]、CVP-MVSNet、M3VSNet[25]、PatchmatchNet 进行定量结果对比。实验采用的同样是 DTU 官方提供的 Matlab 版本的评估代码，实验是在 Matlab R2018a 版本上进行的。采用准确性(Acc)指标、完整性(Comp)指标和整体性(Overall)指标进行定量测试，Overall 是三个指标中最重要的，是 Acc 和 Comp 的总和平均值，代表着重建整体性误差。对比结果如表 1 所示，这 3 个评价指标数值越低表示重建效果越好。

表 1 不同方法在 DTU 数据集上的测试结果

Table 1 Test results of different methods on the DTU dataset

Method	Acc /mm	Comp /mm	Overall /mm
Camp	0.835	0.554	0.695
Furu	0.613	0.941	0.777
Tola	0.342	1.190	0.766
Gipuma	0.283	0.873	0.578
MVSNet	0.396	0.527	0.462
R-MVSNet	0.383	0.452	0.417
P-MVSNet	0.406	0.434	0.420
Point-MVSNet	0.342	0.411	0.376
Fast-MVSNet	0.336	0.403	0.370
CasMVSNet	0.325	0.385	0.355
CVP-MVSNet	0.296	0.406	0.351
M3VSNet	0.636	0.531	0.583
PatchmatchNet	0.427	0.277	0.352
Proposed method	0.427	0.261	0.344

由表 1 数据可以得知：与一些传统 MVS 方法和 MVS 学习方法相比，所提方法的准确性数值较高，原因可能是方法在传播过程中没有对所有可能的深度值进行验证；所提方法的完整性和整体性数值都是最低的，与 PatchmatchNet 相比，完整性提升 5.8%，整体性提升 2.3%，证明了所提方法的有效性。完整性和整体性的提高主要得益于注意力机制，其用于捕获深度推理任务的重要信息，提高了特征提取能力，从而使模型预测出更精确的深度值。

所提方法在 DTU 数据集的三组基准上的测试结果如图 4 所示，其中图 4(a)为深度图，图 4(b)为置信度图，图 4(c)为重建的点云结果。

同时，所提方法在 DTU 数据集上与其他基于学习的多视图立体三维重建的 SOTA 方法进行比较，整体性误差和 GPU 内存、运行时间的关系如图 5 所示，其中图 5(a)为整体性误差和内存消耗的关系，图 5(b)为整体性误差和运行时间的关系。根据文献[10]的

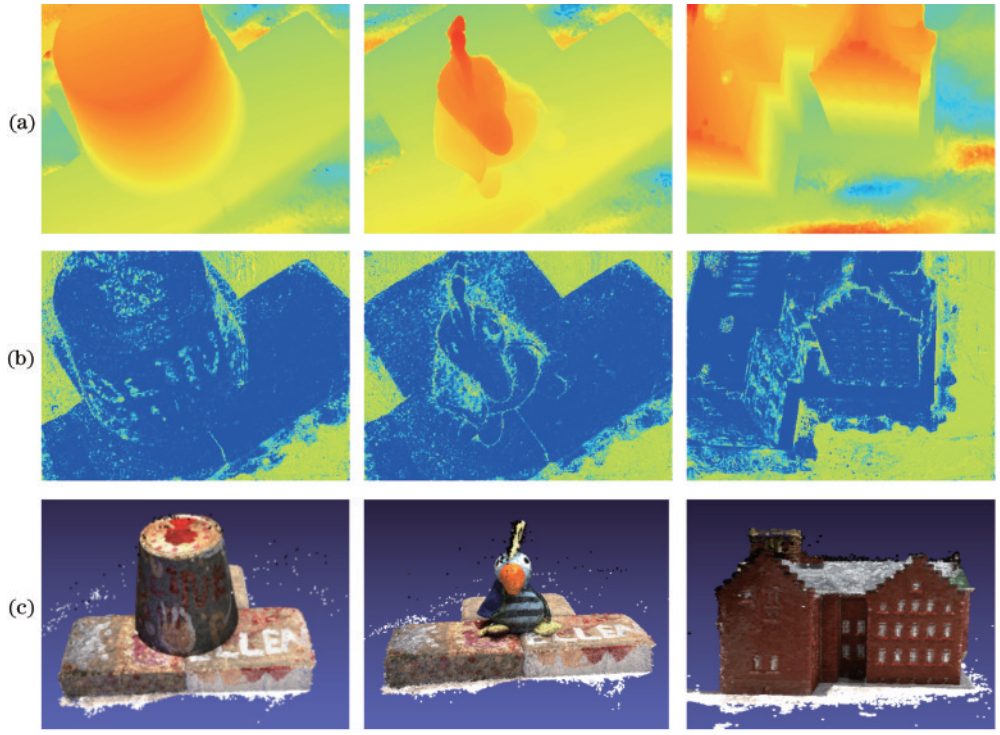

图 4　DTU 数据集基准测试结果。(a)深度图;(b)置信度图;(c)点云结果图

Fig. 4　DTU dataset benchmark results. (a) Depth map; (b) confidence map; (c) point cloud result

实验数据,为了与其他方法进行对比,也将图像大小设置为 1152×864。从图 5 可知:CVP-MVSNet 得益于以图像金字塔构建的代价体金字塔,估计深度图与每一层的深度残差叠加得到最终的深度图,整体性误差最低,但网络模型复杂,运行时间较长;而所提方法在 PatchmatchNet 基础上加入了自注意力层,整体性误差较之更低且兼顾了运行时间,表现出较强的竞争力。

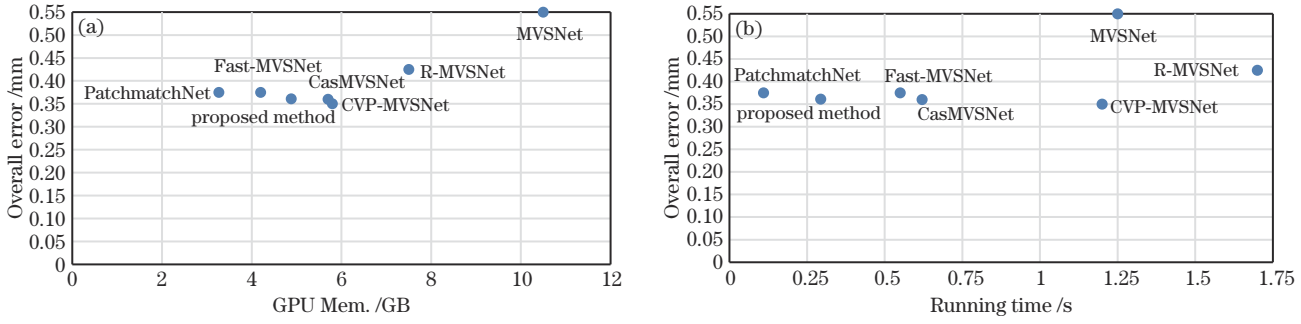

图 5　整体性误差和 GPU 内存、运行时间的关系,图像大小为 1152×864。(a) 整体性误差与 GPU 内存消耗的关系;(b) 整体性误差与运行时间的关系

Fig. 5　Relationship between the overall error and GPU memory and running time, the image size is 1152×864. (a) Relationship between overall error and GPU memory consumption; (b) relationship between overall error and running time

为了评估所提方法的泛化能力,在 Tanks and Temples 数据集[26]中的 advanced 数据集进行测试,它包含从内部和大型室外场景拍摄的室内场景,具有复杂的几何布局和相机轨迹。使用在 DTU 上训练的所提方法,没有任何微调,与其他 SOTA 方法的对比结果如表 2 所示。评估指标是 F-score,也是衡量准确性和完整性的总体表现,越高越好。正如表 2 所示,相比其他方法,所提方法在 Tanks and Temples 数据集上表现最好,与 PatchmatchNet 相比,F-score 提升了 1.3%,表现出了具有竞争力的泛化能力。

表 2　不同方法在 Tanks and Temples 数据集上的结果

Table 2　Results of different methods on Tanks and Temples dataset

Method	F-score
COLMAP	27.24
R-MVSNet	24.91
CasMVSNet	31.12
PatchmatchNet	32.31
Proposed method	32.72

3.5　消融实验

为了进一步证明自注意力层的有效性,进行消融实验,在 DTU 数据集上对单独使用自注意力层与不

使用自注意力层的结果进行对比，使用准确性、完整性、整体性、GPU内存消耗、运行时间、重构出的点数衡量重建的结果。为与文献[10]中的方法进行对比，输入图像尺寸保持一致，即实验所用的图片分辨率也设为1600×1200，消融实验定量结果对比如表3所示，深度图对比结果如图6所示，点云对比如图7所示，二者用的是DTU数据集中的场景15。

表3　消融实验定量结果对比

Table 3　Comparison of quantitative results of ablation experiment

Module	Acc / mm	Comp / mm	Overall / mm	GPU / GB	Running time /s	Vertices
MsFe	0.427	0.277	0.352	10.89	0.210	51527707
SA-MsFe	0.427	0.261	0.344	7.8	0.562	52289896

从表3可以看出，在原来的多尺度特征提取(MsFe)基础上引入了自注意力机制(SA-MsFe)后，可以有效提升重建的完整性和整体性，分别提升5.8%和2.3%，在Meshlab三维点云软件下，重构出的点数(Vertices)增加1.5%。这主要归功于自注意力机制在特征提取阶段由粗到细地对图像特征进行提取，提升了特征提取能力。由于在网络中增加了该层，导致运行时间增加，不过方法仍然能够用于大场景的重建。图6为DTU数据集场景15的深度图，能够看出所提方法得到的深度图在局部区域的深度值更加准确，也更加向GT方向优化，从侧面印证了表2的数据。图7为DTU数据集场景15的点云重建图的局部放大图，能够看出所提方法重建的字母更完整，并且更容易从三维重建结果中识别目标。

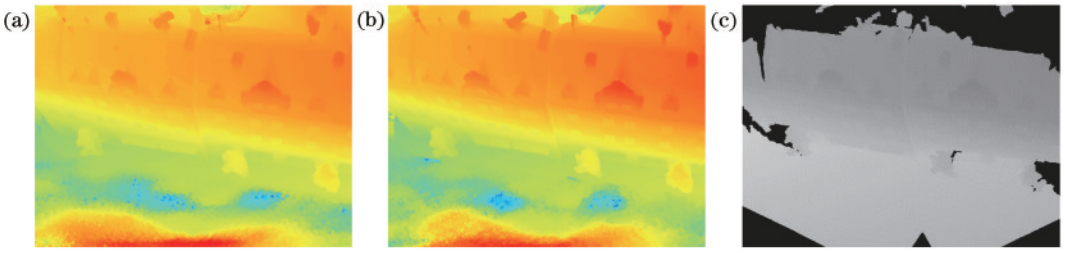

图6　消融实验深度图对比。(a)未改进的深度图；(b)改进后的深度图；(c) GT

Fig. 6　Depth map comparison of ablation experiment. (a) Depth map without improvement; (b) improved depth map; (c) GT

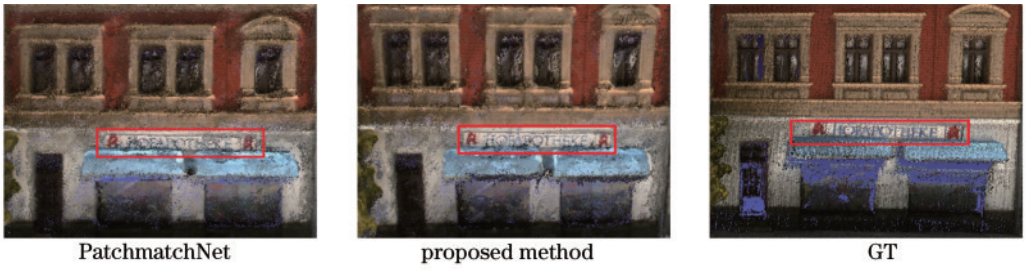

图7　消融实验点云局部放大对比

Fig. 7　Point cloud local magnification comparison map in the ablation experiment

4　结　　论

提出了一个基于自注意力机制的深度学习网络SA-PatchmatchNet。该网络通过在特征提取网络中嵌入自注意力层进行多尺度特征提取，捕获了深度推理任务的重要数据，改善了特征提取能力，很好地解决了由粗到细提取策略几乎不能很好抓取深度信息的问题，从而得到更优的深度图，兼顾了内存消耗与重建的完整性和整体性。在DTU数据集上的基准测试结果显示，所提网络的完整性和整体性指标都优于其他SOTA网络，并最终得到了更精准的深度图和三维点云。预测一张1600×1200像素分辨率的深度图，所提方法仅需要0.596 s，总GPU消耗7.88 GB，结合重建的整体性，更适合于高分辨率图像，也可以广泛应用于大场景的重建。另外所提方法在Tanks and Temples数据集上也表现出了较强的泛化能力。

参　考　文　献

[1] 苗兰芳. 一个基于多视图立体视觉的三维重建方法[J]. 浙江师范大学学报(自然科学版), 2013, 36(3): 241-246.
Miao L F. A 3-D object reconstruction method based on multi-view stereo[J]. Journal of Zhejiang Normal University (Natural Sciences), 2013, 36(3): 241-246.

[2] 张彦雯, 胡凯, 王鹏盛. 三维重建算法研究综述[J]. 南京信息工程大学学报(自然科学版), 2020, 12(5): 591-602.
Zhang Y W, Hu K, Wang P S. Review of 3D reconstruction algorithms[J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2020, 12(5): 591-602.

[3] Luo Y W, Zheng L, Guan T, et al. Taking a closer look at domain shift: category-level adversaries for semantics

consistent domain adaptation[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: 2502-2511.

[4] Luo Y W, Liu P, Guan T, et al. Significance-aware information bottleneck for domain adaptive semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 6777-6786.

[5] Chang J R, Chen Y S. Pyramid stereo matching network [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: 5410-5418.

[6] 廖杰. 基于多视角照片的复杂场景高精度三维重建研究[D]. 武汉: 武汉大学, 2021.
Liao J. Accurate 3D reconstruction for complex scenes based on multi-view images[D]. Wuhan: Wuhan University, 2021.

[7] Yao Y, Luo Z X, Li S W, et al. MVSNet: depth inference for unstructured multi-view stereo[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11212: 785-801.

[8] Yao Y, Luo Z X, Li S W, et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference[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: 5520-5529.

[9] Yang J Y, Mao W, Liu M M, et al. Cost volume pyramid based depth inference for multi-view stereo[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 136-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 4876-4885.

[10] Wang F J H, Galliani S, Vogel C, et al. PatchmatchNet: learned multi-view patchmatch stereo [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14189-14198.

[11] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You need[EB/OL]. (2017-06-12)[2022-08-06]. https://arxiv.org/abs/1706.03762.

[13] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: a randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphic, 2009, 28(3): 24.

[14] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 764-773.

[15] Aanæs H, Jensen R R, Vogiatzis G, et al. Large-scale data for multiple-view stereopsis[J]. International Journal of Computer Vision, 2016, 120(2): 153-168.

[16] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2022-02-05]. https://arxiv.org/abs/1412.6980.

[17] Campbell N D F, Vogiatzis G, Hernández C, et al. Using multiple hypotheses to improve depth-maps for multi-view stereo[M]//Forsyth D, Torr P, Zisserman Z. Computer vision-ECCV 2008. Lecture notes in computer science. Heidelberg: Springer, 2008, 5302: 766-779.

[18] Furukawa Y, Ponce J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362-1376.

[19] Tola E, Strecha C, Fua P. Efficient large-scale multi-view stereo for ultra high-resolution image sets[J]. Machine Vision and Applications, 2012, 23(5): 903-920.

[20] Galliani S, Lasinger K, Schindler K. Massively parallel multiview stereopsis by surface normal diffusion[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 873-881.

[21] Luo K Y, Guan T, Ju L L, et al. P-MVSNet: learning patch-wise matching confidence aggregation for multi-view stereo[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 10451-10460.

[22] Chen R, Han S F, Xu J, et al. Point-based multi-view stereo network[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 1538-1547.

[23] Yu Z H, Gao S H. Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and Gauss-Newton refinement[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1946-1955.

[24] Gu X D, Fan Z W, Zhu S Y, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 2492-2501.

[25] Huang B C, Yi H W, Huang C, et al. M3VSNET: unsupervised multi-metric multi-view stereo network[C]//2021 IEEE International Conference on Image Processing, September 19-22, 2021, Anchorage, AK, USA. New York: IEEE Press, 2021: 3163-3167.

[26] Knapitsch A, Park J, Zhou Q Y, et al. Tanks and temples: benchmarking large-scale scene reconstruction [J]. ACM Transactions on Graphics, 2017, 36(4): 78.