第 43 卷 第 14 期/2023 年 7 月/光学学报 研究论文

光学学报

基于密集视点插值的实时视频拼接方法

衡玮[1,2]，俞健[1,2*]，达飞鹏[1,2,3**]

[1]东南大学自动化学院，江苏 南京 210096；

[2]东南大学复杂工程系统测量与控制教育部重点实验室，江苏 南京 210096；

[3]东南大学深圳研究院，广东 深圳 518063

摘要 针对宽基线场景下的拼接因视差导致伪影和瑕疵的问题，提出了一种基于密集视点插值的实时视频拼接方法。该方法采用在左右相机的基线上补充密集中间视点的方式，为拼接的重叠区域合成平滑过渡的插值视图，以更好地对齐多个输入。为生成该插值视图，利用立体匹配中的匹配代价，设计了网络用来预测在原视图中采样的像素位移场。所提方法在没有插值视图真值的情形下，利用视点间的空间变换关系，指引网络学习视图生成规则。实验结果表明，所提方法能提升视频图像拼接后的视觉观感，并可以达到实时性能，满足实际场景中的应用需求。

关键词 机器视觉；视频拼接；宽基线；深度学习；视图插值

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/AOS230509

1 引 言

视频拼接技术可以方便获取更宽广的视野范围，在安防监控、智能驾驶、虚拟现实、视频会议等多种应用领域发挥着作用[1-3]。尽管图像和视频拼接的研究有着悠久历史，但目前已有方法的表现尚不完美，算法计算效率、宽基线大视差场景、低光照低纹理的挑战性场景，给拼接任务带来了许多挑战。对于视频图像拼接，最大的挑战之一是视差问题，在相机光心完全重合时，不会受到视差的影响，此时可以轻易地合成无伪影无瑕疵的图像。然而在实际应用中，相机光心完全重合的条件难以达成，并且在一些场景下相机也会分散排列，如车载全景系统、广视野安防监控系统，因此宽基线下的拼接问题具有研究的必要性。

传统的视频图像拼接方法中，AutoStitch是较为经典的方法，该方法使用一个全局单应性矩阵进行配准对齐[4]。然而这种方法没有视差处理能力，在相机光心不重合或场景深度变化大时，拼接结果会有明显瑕疵。为减轻这种问题，有研究人员提出了多单应性方法，例如使用两个单应性矩阵分别对齐背景和前景的方法[5]、划分网格估计局部单应性的APAP算法[6]、分割成超像素再估计最优单应性的方法[7]，这些方法进一步提升了图像对齐的质量。除此之外，缝合线[8-9]也是常用的处理方法：He等[10]将视频拼接分为初始模板计算和缝合线更新两个阶段，实现监控场景下的在线视频拼接；Lo等[11]使用缝合线优化双鱼眼相机的拼接问题。另外，有的方法在视频去抖的基础上实现视频拼接[12]，还有对于推扫式相机的影像进行拼接的方法[13]。视频拼接的产品应用方面，有NVIDIA的VRWorks SDK、Google Jump[14]和Rich360[15]，这些产品也在不断迭代升级，优化拼接效果。

近年来，基于深度学习的视频图像拼接技术引起了研究人员的关注，这为解决此类问题提供了全新的维度。传统的图像特征提取方法是手工设计的，在低光照、低纹理等挑战性场景下容易失败。在拼接中利用卷积神经网络进行图像特征提取，可以在这些场景下表现出更好的鲁棒性[16-17]。有的方法在拼接中使用无视差的合成数据集[18-19]，但这与实际应用场景往往不相符，宽基线和大视差场景下的拼接是主要的难点，众多研究围绕此问题展开。其中，有许多方法结合了其他视觉任务的方法，来辅助处理拼接问题：Lai等[20]结合光流法，实现适用于线性相机阵列的宽基线视频拼接网络，但只能应用于特定的相机配置；Li等[21]通过语义对齐的方法辅助图像拼接，图像语义信息丰富时能够得到较好的结果；Dai等[22]将拼接的合成阶段视为图像融合问题，提出边缘引导合成网络以计算融合权重，但视差太大时融合结果可能存在结构不一致。此外，有的方法使用基于深度学习的多单应性估计方法：Song等[23]利用多单应性估计，实现端到端的图像拼接网络，但同样只适用于特定的相机配置；Nie等[24]

收稿日期：2023-02-06；**修回日期**：2023-03-07；**录用日期**：2023-04-06；**网络首发日期**：2023-05-08

基金项目：国家自然科学基金(51475092)、江苏省前沿引领技术基础研究专项(BK20192004C)

通信作者：*yujian@seu.edu.cn；**dafp@seu.edu.cn

设计了网格局部单应性预测网络，需要设定网格大小以对齐图像区域。除此之外，还有方法利用图像变形与重建的思想，如：Nie 等[25]提出了包含低分辨率变形和高分辨率优化两个分支的无监督图像拼接框架，但只从图像特征的角度实现伪影消除；Kweon 等[26]提出了由像素变形模块和图像生成模块组成的深度学习框架，但处理大视差下的拼接问题，依赖像素匹配的准确性。这些方法带来许多新思路，值得研究人员进一步探索研究。

针对宽基线和大视差场景下的拼接问题，本文提出了一种基于密集视点插值的视频拼接方法，处理视差影响成像质量的难题。该方法重点处理拼接重叠区域的对齐问题，通过在左右相机的基线上补充密集中间视点的方式，提升重叠区域的对齐质量并获得平滑的视觉观感。本文设计用于预测生成该插值视图的网络，分为特征提取、相关性计算、高分辨率优化模块，预测的插值视图与视图的非重叠区域组合得到拼接结果。在没有插值视图真值的情形下，考虑将生成的插值视图变换到原视点，使其与输入视图保持一致，用来间接构造损失函数。本文方法的优势在于能以实时性能实现视觉观感良好的宽基线拼接，能够适应各种基线宽度，对于相机的配置变化具有鲁棒性，以及良好的应用价值。

2 系统设计

考虑宽基线的相机布置下，视差对成像结果的影响。如图 1(a)所示，假设圆形和三角形分别为三维空间中的前后两个物体，相机光心不重合时，在两个成像视图中物体会因视差导致不能对齐，甚至出现错位。如图 1(b)所示，相机光心完全重合时，不存在视差，此时在成像视图中物体能够对齐，可以很容易地融合成无伪影的视图。

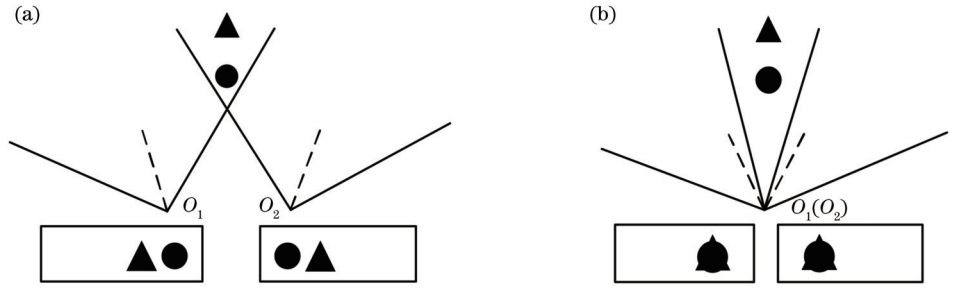

图 1 视差对成像结果的影响。(a)相机光心不重合时视图不能对齐；(b)相机光心完全重合时视图能够对齐

Fig. 1 Effect of parallax on imaging results. (a) Views cannot be aligned when cameras′ optical centers do not coincide; (b) views can be aligned when cameras′ optical centers are completely coincident

为减少视差对视频图像拼接质量的影响，考虑为重叠区域重新生成插值视图。如图 2 所示，阴影区域是两个相机的最大重叠区域，所提方法主要处理该区域，对非重叠区域做较少的额外处理。所提方法的目标是使生成的插值视图无伪影、视觉观感平滑，并与左图和右图的非重叠区域之间形成平滑过渡。

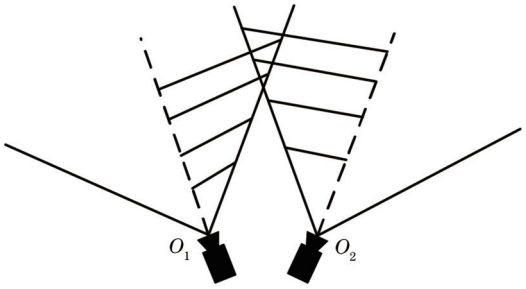

图 2 相机排列示意图

Fig. 2 Camera arrangement diagram

针对重叠区域的融合问题，考虑使用密集视点插值的方式。如图 3 所示，在左右两个相机的基线上均匀补充 k 个插值视点，每个视点划分了一定空间区域，在该区域范围内获取局部视图。当插值视点数量足够多时，可以形成逐渐的过渡。根据重叠区域视角的左边界和右边界，将视角均匀分成 $k+2$ 等份。设两个相机间的基线宽度为 b，则插值视点与原始左视点的距离可表示为

$$x_{k+1}^{(i)}=\frac{i}{k+1}b,\ 0\leqslant i\leqslant k+1。\tag{1}$$

设焦距为 f，生成的插值视图在其图像坐标系的横坐标为 x_o，在左右原视点图像中分别采样的图像横坐标为 x_1 和 x_2。那么，在 x_o 对应的插值视点处，视差 $d=x_1-x_2$。经过立体校正后，可忽略 y 轴坐标，此时得到的坐标关系示意图如图 4 所示，像素的三维坐标点 P 在 x 轴和 z 轴上的坐标可表示为

$$x=\frac{x_o}{f}z+x_{k+1}^{(i)},\ z=\frac{fb}{d}。\tag{2}$$

根据式(1)和式(2)，x_1 和 x_2 可表示为

$$x_1=\frac{f}{z}x=\frac{f}{z}\left(\frac{x_o}{f}z+x_{k+1}^{(i)}\right)=$$

$$x_o+\frac{f}{z}\cdot\frac{i}{k+1}\cdot b=x_o+\frac{i}{k+1}\cdot d,\tag{3}$$

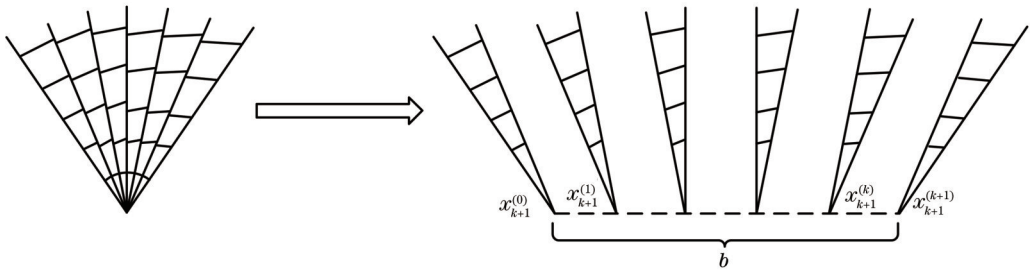

图 3　宽度为 b 的基线上补充 k 个中间视点

Fig. 3　Add k intermediate viewpoints on the baseline with width of b

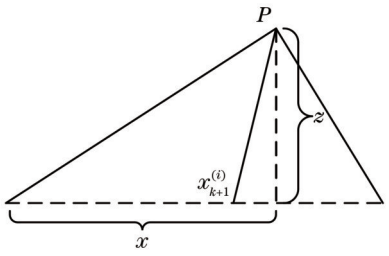

图 4　坐标点之间的关系

Fig. 4　Relation between coordinate points

$$x_2 = x_1 - d = x_o + \left(\frac{i}{k+1} - 1\right) \cdot d。 \tag{4}$$

设计网络模型以预测视差 d 的值为目标，此时根据 x_o 的值和对应的插值视点，即可计算得到 x_1 和 x_2。

进一步地，根据式(3)，用 x_1 和 x_2 表示出 x_o，可表示为

$$x_o = x_1 - \frac{i}{k+1}(x_1 - x_2)。 \tag{5}$$

当插值视点数量 k 固定不变时，基线长度的变化不会改变对应的插值视点，由式(5)可知，x_1 和 x_2 对应插值视图中唯一的图像坐标 x_o，基线长度的变化不会影响插值视图的生成。

总体而言，在实际应用的过程中，本文采用的方法可分为以下具体步骤：

1)进行双目相机标定，得到相机的内参数和两个相机间的旋转、平移变换矩阵，获取相机间的配置关系。

2)根据标定结果计算相机的畸变校正查找表、两个相机相对同一水平面的俯仰角、重叠区域视角范围。

3)根据相机俯仰角，得到变换到同一水平面的像素位移场，以便在水平方向上进行拼接。

4)根据重叠区域视角范围，分离出可能的最大重叠区域部分，通过立体校正的方式调整到共面且行对准，得到立体校正后的像素位移场，以便只在一个维度上处理图像数据。

5)根据畸变校正查找表对相机采集的原始视图去畸变，根据步骤3得到的像素位移场，把相机采集的视图调整到同一水平面上。

6)根据步骤4得到的像素位移场，分离得到共面且行对准的重叠区域部分，作为模型的输入。

7)利用模型预测像素位移场，分别在左右输入视图中采样生成图像，并线性加权融合后作为生成的重叠区域插值视图。

8)将生成的重叠区域插值视图与非重叠区域部分，共三个部分进行组合得到整个图像，并进行柱面投影处理，以对齐融合区的边界，由此得到最终拼接结果。

每次进行视频中一帧的拼接时，重复进行步骤5～8即可。

3　网络设计

深度学习方法在特征提取能力和运行速度上有优势，且可以自动学习到本文设计出的这种底层模式，获取端到端的预测结果。本文设计了网络来预测式(3)和式(4)中的视差 d，这代表在插值视图的像素点 x_o 与相对应的每个插值视点 $x_{k+1}^{(i)}$ 处，由左右原视图进行采样的坐标差值，与一般意义上在左右视点处的视差图存在区别。根据 d 得到两张图像的像素位移场后，从原视点图像中进行采样生成插值视图。网络分为特征提取模块、相关性计算模块、高分辨率优化模块，该网络从左右图像中提取特征，在低分辨率上计算匹配代价后，预测采样像素位置，然后上采样经过高分辨率优化模块预测最终结果。设计的网络整体架构如图5所示。

3.1　特征提取模块

特征提取模块用来分别提取左图和右图的图像特征，其中输出特征图的分辨率为输入图像的1/8，在低分辨率上进行后续处理。使用的特征提取网络在输入图像中共享权重，该模块由一系列3×3卷积层构成，降采样的过程中保持通道数为32。经过三次降采样后，得到1/8分辨率的特征图，作为后续模块的输入。网络使用该结构，优点是能够拥有较大的感受野，关注了图像的边缘纹理特征，并学习得到输入的紧凑向量表示。

3.2　相关性计算模块

本文采用立体匹配中常用的方法，构建3D cost volume [27-28]来计算图像间的相关性。将左右视图提取

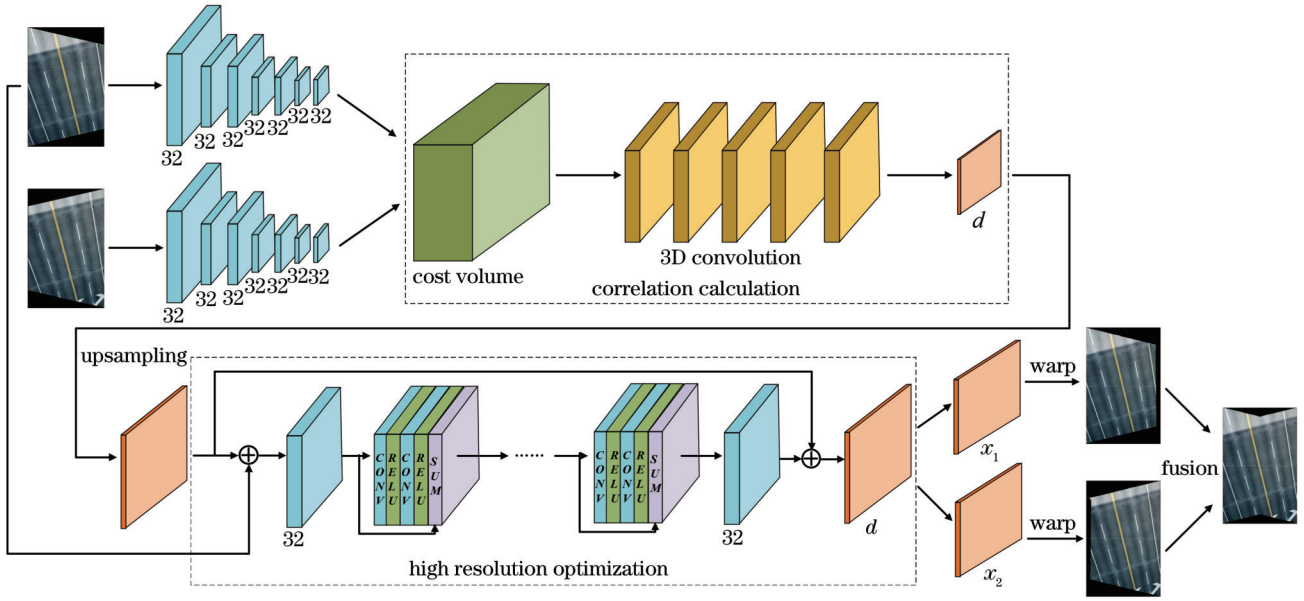

图 5　网络整体架构

Fig. 5　Overall architecture of the network

得到的 1/8 分辨率的特征图相减，根据式(3)和式(4)在特征图的像素上水平移动，以 difference 做差的方式[27]构建 cost volume，移动次数最大值 D_{max} 设置为 32。这种 cost volume 的构建方式比较轻量级，网络需要的参数数量少，计算效率很高。然后，通过 4 个滤波器数量为 32 和 1 个滤波器数量为 1 的 3D 卷积层预测输出结果，利用 softmax 函数把预测结果转化为概率 p，最后计算得到 d 的预测值，可表示为

$$d=\sum_{i=0}^{D_{max}-1} i\cdot p_i。\tag{6}$$

3.3　高分辨率优化模块

以上模块输出 1/8 分辨率的 d，为了得到原分辨率下的结果，本文设计了高分辨率优化模块。将 1/8 分辨率的 d 双线性上采样到原分辨率后，和原始 RGB 图像一起作为模块的输入，以指导 d 的值进行微调。该部分由卷积层和 6 个残差块构成，残差块中使用了膨胀卷积来扩大感受野，膨胀参数设置为[1,2,4,8,1,1]。模块输出的结果和输入的 d 相加，通过 ReLu 函数得到最终结果。这样实现的优点是可以引入原分辨率下的高频细节信息，进一步细化上采样的结果。

3.4　图像融合处理

根据式(3)和式(4)，由网络预测输出的 d 计算得到 x_1 和 x_2，分别得到左右视图采样的像素位移场 $f_{flow\,1}$ 和 $f_{flow\,2}$，然后对两个原始视图分别进行采样生成两张图像，进行简单的线性加权融合处理后，生成重叠区域部分的插值视图 I_o，可表示为

$$I_o=W\cdot\mathrm{warp}(I_1,f_{flow\,1})+(1-W)\cdot\mathrm{warp}(I_2,f_{flow\,2}),\tag{7}$$

式中：W 为设置的线性融合权重；warp 为变换函数。

3.5　损失函数

生成的插值视图是由每个中间视点在其空间范围内的局部视图组合而成，可以想到在补充的每个中间视点处放置相机，来捕获一系列图像以合成插值视图真值。但因插值视点密集排布，这种方式较为繁琐不易操作且可能误差较大，无法精确到亚像素级别，插值点的数量也不方便进行调整。本文考虑在没有插值视图真值的情形下，通过间接方式构造损失函数。通过原视点处的深度信息，计算得到原视点图像中像素点的三维坐标。由于插值视点把三维空间分为若干区域，通过二分查找的方式，可以快速查找到 3D 坐标点被划分到的空间区域，找到对应的插值视点，进而计算得到在插值视图中对应的亚像素坐标。根据原视图中像素和插值视图中亚像素的对应关系，可以将网络预测的插值视图，变换到原视点处。通过约束其和输入图像保持一致，来指引网络生成所需的插值视图。为准确获取所需的原视点深度信息，本文使用 Airsim 模拟器在虚拟环境中采集数据。

由图 3 可知，划分出来的空间区域中，有一些空间区域被忽略，在像素点上计算出的三维坐标点中，一些点不存在对应的插值视点。对这些像素点不予考虑，这样变换到原视点处的图像有许多空洞，将其作为二进制掩码 M，从输入中过滤提取图像，此过程的示例图片如图 6 所示。根据所提的方法，将两张插值视图分别变换到左右原视点处，与掩码过滤后的输入图像进行比较，构造内容损失和感知层损失。

内容损失 L_c 是计算该变换图像和原视点图像的 L1 Loss，表示为

$$L_c=\|\mathrm{warp}(I_o)-I_i\cdot M\|_1,\tag{8}$$

式中：I_i 为原视点图像；I_o 为生成的插值视图。

感知层损失 L_p 的作用是使图像的特征保持一致性，本文利用预训练的 VGG-19 特征提取网络[29]中的

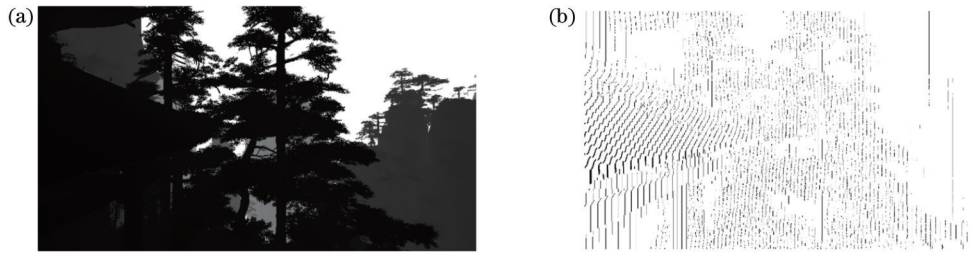

图 6　根据深度图生成二进制掩码。(a)深度图;(b)生成的二进制掩码($k=100$)

Fig. 6　Binary mask generated from depth map. (a) Depth map; (b) generated binary mask ($k=100$)

conv5_3层提取高级语义特征,在该层上计算MSE Loss,表示为

$$L_{\mathrm{p}}=\|F\left[\mathrm{warp}(I_{\mathrm{o}})\right]-F(I_{\mathrm{i}}\cdot M)\|_2, \tag{9}$$

式中:$F(\cdot)$用于输出预训练网络提取的特征图。

由此将损失函数表示为

$$L=(L_{\mathrm{c}}^1+L_{\mathrm{c}}^2)+\lambda(L_{\mathrm{p}}^1+L_{\mathrm{p}}^2), \tag{10}$$

式中:λ为设置的比例系数,本文设置为0.1;L_{c}^1和L_{c}^2分别为左右图像的内容损失;L_{p}^1和L_{p}^2分别为左右图像的感知层损失。

4　实　　验

4.1　网络实现细节

由于所提方法是将重叠区域映射到共面且行对准,因此为适应各种相机配置,无需采用复杂的数据采集方式,可以直接通过平行放置的双目相机来构造数据集。由于在训练时,需要预先输入相机的配置信息,生成预处理的模板和构造Loss所需的信息,并考虑相机视野较大的情形,因此本文使用Airsim模拟器,在虚拟环境中配置平行放置的双目相机,使用固定的基线宽度,在场景地图下合成2600组图像来构造训练数据集。每组数据由4张图像组成,包括左相机图像和深度图、右相机图像和深度图。本文使用PyTorch框架实现网络模型,模型由Adam optimizer训练,初始学习率设置为0.001,权重衰减参数设置为0.0001。在具体训练过程中,首先设置特定的相机配置方式,如重叠区域视角范围设置为50°,以便从数据集中提取该视角范围的有效区域,作为模型的输入,并根据视角信息构造损失函数,使网络能够快速收敛,此过程一共训练了150个epoch。然后,再在多个相机配置方式上微调模型,再训练50个epoch,进一步提升模型的鲁棒性和泛化能力。

4.2　插值点数量的影响

当插值点数量k分别取10和100时,测试生成的插值视图,并与多波段融合方法进行比较,示例图片如图7所示。可以看出:当k取值过小时,生成的图像可能在插值视点所属区域的交界处有一定割裂感,影响视觉观感;当k取值较大时,视觉观感能够更加平滑。所提方法是由局部视图合成的插值图像,与物理光学系统不匹配,实际选择k值时,可以按照重叠区域水平像素宽度与k的比值在10以内的方式。比值小于1时,插值视点在视图中包含的水平像素数量不大于1个单位,此时继续增加k值的意义不大。

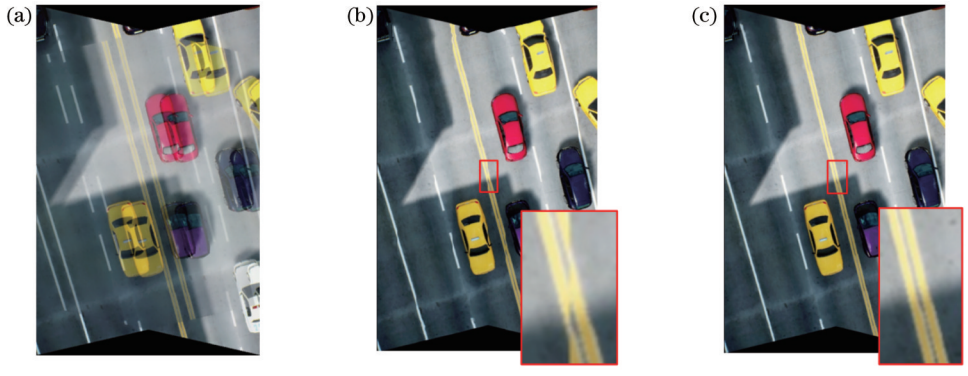

图 7　图像融合结果。(a)多波段融合;(b)所提方法,$k=10$;(c)所提方法,$k=100$

Fig. 7　Results of image fusion. (a) Multiband blending; (b) proposed, $k=10$; (c) proposed, $k=100$

4.3　拼接结果分析

在现实场景中,使用两个相机来测试实际的拼接效果,相机采样的分辨率为1280×720,使用双目相机标定来获取当前的相机配置,再进行后续的拼接,图8显示了对于示例图像进行拼接的具体过程。首先根据相机标定结果,对图像去畸变,投影调整到同一水平面;然后提取出重叠区域,并投影到共面且行对准,输入到模型中生成插值视图;最后,该视图与原来的非重叠区域组合,经柱面投影后得到拼接结果。

将所提方法与多波段融合、APAP[6]、LPC[30]、LB-

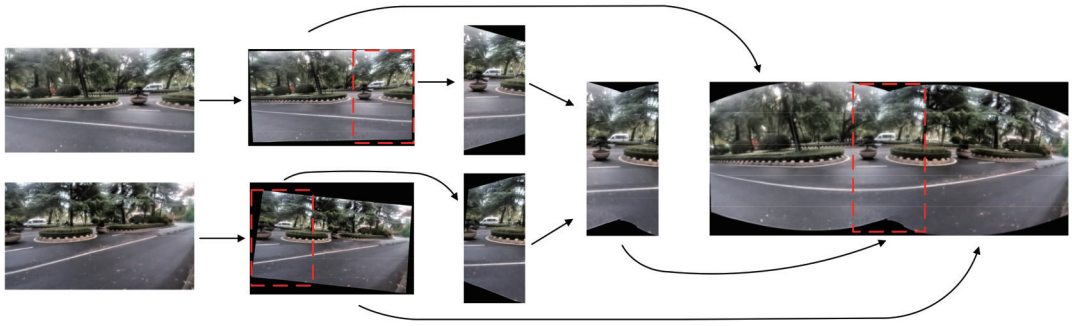

图 8　拼接具体过程

Fig. 8　Specific process of stitching

UDHN[25]方法进行对比，不同场景下的拼接结果如图 9 所示。进一步地，从 KITTI 原始数据中获取有视差的视频帧，对视频的连续三帧进行拼接，与上述方法的对比结果如图 10 所示。从图 9 和图 10 可以看出：多波段融合方法在视差的影响下会出现伪影；APAP 为代表的网格优化方法估计了局部单应性，虽然部分情况下也可消除伪影，但在网格变形后非重叠区域可能有较大的形状失真；LPC 是最新的基于传统方法的多单应性方法，能够更好地对齐，但仍然可能存在形状失真；LB-UDHN 是基于深度学习的拼接方法，但有时也会出现瑕疵。本文所提方法可以消除伪影，平滑对齐输入且形状失真很小，拥有良好的视觉效果。

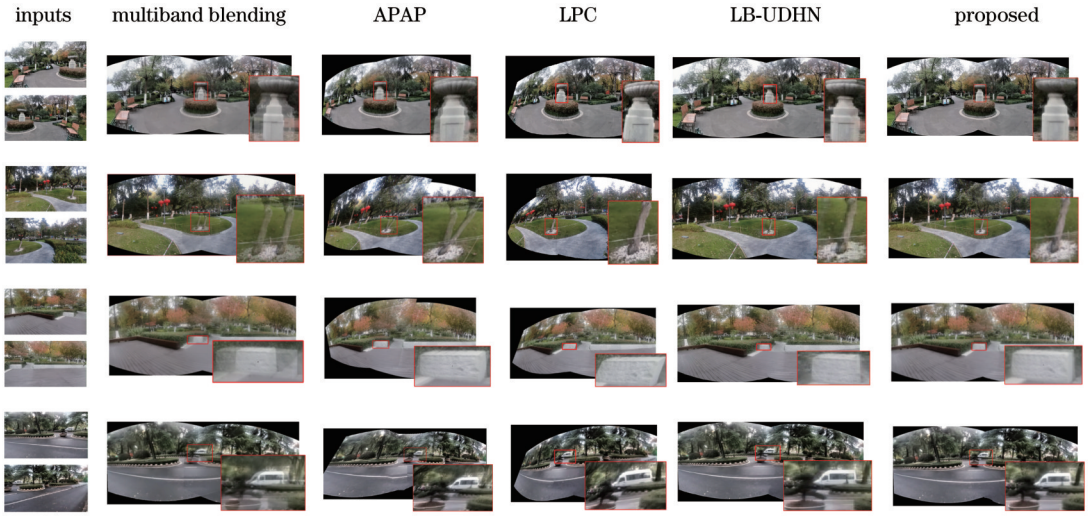

图 9　不同场景下的拼接结果对比

Fig. 9　Comparison of stitching results under different scenarios

4.4　对齐质量评估

从定量角度分析，使用常用的图像质量评价指标结构相似性（SSIM）和峰值信噪比（PSNR），评估重叠区域的对齐质量。由于所提方法预先将重叠区域投影到共面且行对准，因此可以使用公开数据集 FlyThings3D 和 KITTI 2015 的双目图像进行测试，图片的特点是已初步对齐但存在视差，且场景的视差变化很大。同时，测试多波段融合、APAP[6]、LPC[30]、LB-UDHN[25]和 MGDH[24]方法在数据集下的对齐指标，结果如表 1 所示。可以看出，以 APAP、LPC 为代表的传统特征检测和匹配的方法，表现还不够好，因为测试数据集有很多挑战性场景，传统方法在这些场景下还不够鲁棒，而且视差变化大也进一步影响了结果。基于学习的单应性方法表现一般，因为这些方法没有利用相机的配置信息，只从图像特征的角度进行处理。所提方法利用相机标定获取相机配置信息，进行预处理后再对齐，并针对处理了视差，因此可以获得更好的对齐质量。在消融实验中，验证高分辨率优化模块的作用。在去除了高分辨率优化模块后，按照所提方法重新训练网络，得到模型 v1，测试对齐的指标，并与包含该模块的运行结果进行对比。可以看出：使用高分辨率优化模块后，能够进一步提升图像对齐的质量，因为该模块引入了更多高频细节信息。

4.5　模型大小和运行时间

测试所提方法的模型大小和运行时间，测试平台选择 RTX 2060 GPU，在输入图像的分辨率为 1280×720 的情况下，与基于深度学习的 LB-UDHN[25]和 MGDH[24]方法进行对比，结果如表 2 所示。同时，

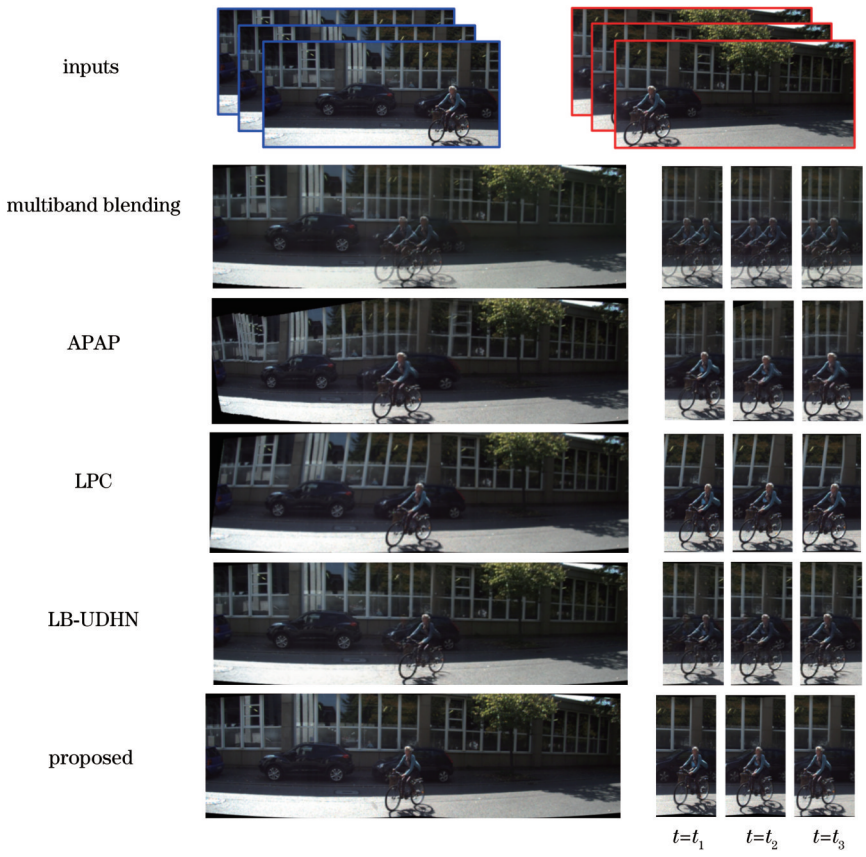

图 10　视频帧的拼接结果对比

Fig. 10　Comparison of stitching of video frames

表 1　SSIM 及 PSNR 指标对比

Table 1　Comparison of SSIM and PSNR metrics

Method	FlyThings3D		KITTI 2015	
	SSIM	PSNR	SSIM	PSNR
Multiband blending	0.555	15.78	0.416	14.51
APAP	0.672	16.42	0.523	16.03
LPC	0.648	16.33	0.568	15.89
LB-UDHN	0.718	20.41	0.654	18.59
MGDH	0.764	20.42	0.616	18.65
Proposed_v1	0.807	21.94	0.693	20.15
Proposed	0.826	23.31	0.708	20.54

测试每一个模块的平均运行时间，时间占比如图 11 所示。因为所提方法经过相机标定处理，能够初步对齐图像，且采用轻量级的 cost volume 构建方法，所以在模型尺寸和运行速度方面都有优势，处理帧率可以达到 30 frame/s 以上，满足视频实时在线拼接的需求。

表 2　模型尺寸和速度的对比

Table 2　Comparison of model size and the speed

Method	Model size /MB	Runtime /ms
LB-UDHN	643	245
MGDH	187	97
Proposed	2.4	24

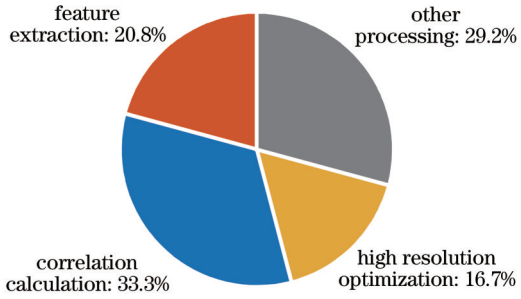

图 11　不同模块的运行时间占比

Fig. 11　Proportion of running time of different modules

4.6　基线变化的分析

从式(5)的理论分析可知，当插值视点数量 k 固定不变时，基线宽度的变化不会影响所提方法生成平滑过渡的插值视图，即使基线的变化会导致图像内容的变化。实验过程中，在 Airsim 模拟器中设置不同的相机基线宽度，获取测试图片集，融合的示例图片如图 12 所示。同时，在测试集中计算 SSIM 和 PSNR 指标，得到的结果如表 3 所示，可以看出，所提方法在不同基线宽度下都能够较好地对齐，获得较高的指标提升。所提方法对于基线宽度的变化是鲁棒的，虽然数据集来源于固定基线，但预测结果可以适应于不同的基线宽度。

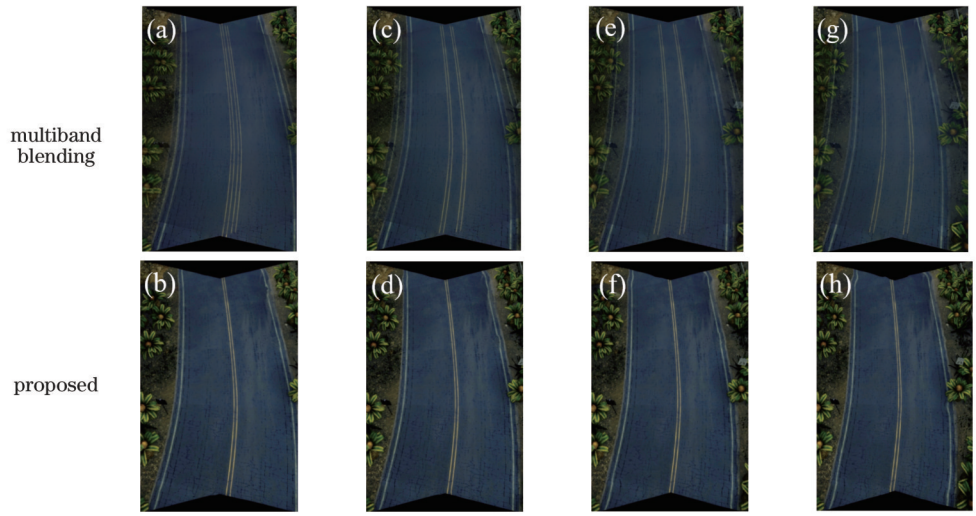

图 12　不同基线宽度下的融合结果。(a)(b) 0.5 m;(c)(d) 1 m;(e)(f) 1.5 m;(g)(h) 2 m

Fig. 12　Fusion results under different baseline widths. (a)(b) 0.5 m; (c)(d) 1 m; (e)(f) 1.5 m; (g)(h) 2 m

表 3　不同基线宽度下的 SSIM 及 PSNR 指标

Table 3　SSIM and PSNR metrics under different baseline widths

Baseline width /m	SSIM		PSNR	
	Multiband blending	Proposed	Multiband blending	Proposed
0.5	0.591	0.818	19.88	24.48
1.0	0.581	0.808	19.67	24.23
1.5	0.569	0.805	19.34	24.16
2.0	0.552	0.796	18.87	23.83

4.7　训练方式的影响

本文测试不同训练方式对插值视图生成质量的影响，重叠区域视角大小与插值视图的生成方式存在相关性。比较在特定的重叠区域视角(50°)下进行训练，和多个重叠区域视角下训练微调后的结果，从低重叠率到高重叠率，在测试集上计算 PSNR 指标，如图 13 所示。可以看出，在特定的重叠区域视角下进行训练时，视角大小改变时，PSNR 指标有略微的下降，对于相机配置变化的鲁棒性还不够。进一步地，经过多个视角参数下的训练微调后，PSNR 指标在多个不同的视角上均得到提升，网络鲁棒性和泛化能力得以提升，可以得到更好的结果。

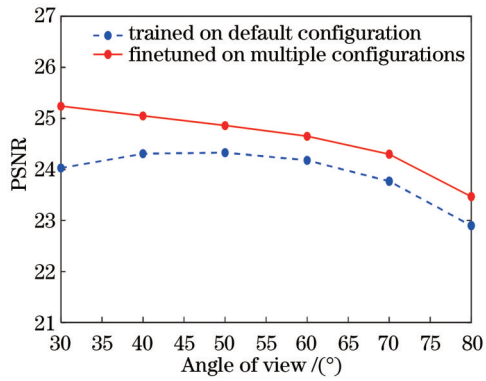

图 13　不同视角参数下的 PSNR 指标

Fig. 13　PSNR metrics under different angles of view

5　结　　论

本文提出了一种基于密集视点插值的实时视频拼接方法，以处理在宽基线和大视差场景下的拼接问题，在提升拼接图像质量的同时提高拼接效率。在所提方法中，通过在左右相机的基线上补充密集中间视点，每个视点划分一定空间区域，在该区域范围内获取局部视图，为拼接的重叠区域合成平滑过渡的插值视图。然后，提出了用于生成该插值视图的网络，网络分为特征提取、相关性计算、高分辨率优化模块，预测插值视图在原视图中的采样位置，生成的插值视图与非重叠区域组合后得到拼接结果。所提方法在没有插值视图真值的情形下，在虚拟环境中计算原视点处的三维信息，通过二分法查找到对应的插值视点空间区域，将插值视图变换到原视点下构造损失函数，指引网络学习视图生成规则。通过各种实验，验证了所提方法能够提升视频图像拼接后的视觉观感，适应于各种基线宽度，具有良好的泛化能力，且可以达到实时拼接的性能，满足实际应用中的在线拼接需求。在未来的研究中，需要进一步保证系统在各种复杂环境下的鲁棒性，在满足实时性能的前提下，进一步提升视频图像拼接的质量。

参　考　文　献

[1]　宋聪聪，高策，张艳超，等. 基于三维球面模型的全景视频实时拼接方法[J]. 光学学报，2023，43(10): 1010002.
Song C C, Gao C, Zhang Y C, et al. Real-time stitching method of panoramic video based on three-dimensional spherical model[J]. Acta Optica Sinica, 2023, 43(10): 1010002.

[2]　陈浩，杨恺伦，胡伟健，等. 基于全景环带成像的语义视觉里程计[J]. 光学学报，2021，41(22): 2215002.

Chen H, Yang K L, Hu W J, et al. Semantic visual odometry based on panoramic annular imaging[J]. Acta Optica Sinica, 2021, 41(22): 2215002.

[3] Kim H G, Lim H T, Ro Y M. Deep virtual reality image quality assessment with human perception guider for omnidirectional image[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(4): 917-928.

[4] Brown M, Lowe D G. Recognising panoramas[C]//Proceedings Ninth IEEE International Conference on Computer Vision, October 13-16, 2003, Nice, France. New York: IEEE Press, 2008: 1218-1225.

[5] Gao J H, Kim S J, Brown M S. Constructing image panoramas using dual-homography warping[C]//CVPR, June 20-25, 2011, Colorado Springs, CO, USA. New York: IEEE Press, 2011: 49-56.

[6] Zaragoza J, Chin T J, Brown M S, et al. As-projective-as-possible image stitching with moving DLT[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 2339-2346.

[7] Lee K Y, Sim J Y. Warping residual based image stitching for large parallax[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 8195-8203.

[8] Gao J, Li Y, Chin T J, et al. Seam-driven image stitching[C]//34th Annual Conference of the European Association for Computer Graphics, May 6-10, 2013, Girona, Spain. London: Eurographics Association, 2013: 45-48.

[9] Lin K M, Jiang N J, Cheong L F, et al. SEAGULL: seam-guided local alignment for parallax-tolerant image stitching[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9907: 370-385.

[10] He B T, Yu S H. Parallax-robust surveillance video stitching[J]. Sensors, 2015, 16(1): 7.

[11] Lo I C, Shih K T, Chen H H. Efficient and accurate stitching for 360° dual-fisheye images and videos[J]. IEEE Transactions on Image Processing, 2022, 31: 251-262.

[12] Nie Y W, Su T, Zhang Z S, et al. Dynamic video stitching via shakiness removing[J]. IEEE Transactions on Image Processing, 2018, 27(1): 164-178.

[13] 黎荆梅, 范永祥, 王宁, 等. 航空推扫式外拼接成像系统子视场相对定向方法[J]. 光学学报, 2021, 41(18): 1811001.
Li J M, Fan Y X, Wang N, et al. Relative orientation method for airborne pushbroom combined imaging system[J]. Acta Optica Sinica, 2021, 41(18): 1811001.

[14] Anderson R, Gallup D, Barron J T, et al. Jump: virtual reality video[J]. ACM Transactions on Graphics, 2016, 35(6): 1-13.

[15] Lee J, Kim B, Kim K, et al. Rich360: optimized spherical representation from structured panoramic camera arrays[J]. ACM Transactions on Graphics, 2016, 35(4): 1-11.

[16] Shi Z F, Li H, Cao Q J, et al. An image mosaic method based on convolutional neural network semantic features extraction[J]. Journal of Signal Processing Systems, 2020, 92(4): 435-444.

[17] Hoang V D, Tran D P, Nhu N G, et al. Deep feature extraction for panoramic image stitching[M]//Nguyen N T, Jearanaitanakij K, Selamat A, et al. Intelligent Information and Database Systems. Lecture notes in computer science. Cham: Springer, 2020, 12034: 141-151.

[18] Zhao Q, Ma Y K, Zhu C, et al. Image stitching via deep homography estimation[J]. Neurocomputing, 2021, 450: 219-229.

[19] Nie L, Lin C Y, Liao K, et al. A view-free image stitching network based on global homography[J]. Journal of Visual Communication and Image Representation, 2020, 73: 102950.

[20] Lai W S, Gallo O, Gu J, et al. Video stitching for linear camera arrays[C]//30th British Machine Vision Conference (BMVC 2019), September 9-12, 2019, Cardiff, Wales, UK. London: British Machine Vision Association, 2019: 1-12.

[21] Li A C, Guo J, Guo Y W. Image stitching based on semantic planar region consensus[J]. IEEE Transactions on Image Processing, 2021, 30: 5545-5558.

[22] Dai Q Y, Fang F M, Li J C, et al. Edge-guided composition network for image stitching[J]. Pattern Recognition, 2021, 118: 108019.

[23] Song D Y, Um G M, Lee H K, et al. End-to-end image stitching network via multi-homography estimation[J]. IEEE Signal Processing Letters, 2021, 28: 763-767.

[24] Nie L, Lin C Y, Liao K, et al. Depth-aware multi-grid deep homography estimation with contextual correlation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4460-4472.

[25] Nie L, Lin C Y, Liao K, et al. Unsupervised deep image stitching: reconstructing stitched features to images[J]. IEEE Transactions on Image Processing, 2021, 30: 6184-6197.

[26] Kweon H, Kim H, Kang Y, et al. Pixel-wise deep image stitching[EB/OL]. (2021-12-12)[2022-02-09]. https://arxiv.org/abs/2112.06171.

[27] Khamis S, Fanello S, Rhemann C, et al. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11219: 596-613.

[28] Sun D Q, Yang X D, Liu M Y, et al. PWC-net: CNNs for optical flow using pyramid, warping, and cost volume[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: 8934-8943.

[29] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2022-02-19]. https://arxiv.org/abs/1409.1556

[30] Jia Q, Li Z J, Fan X, et al. Leveraging line-point consistence to preserve structures for wide parallax image stitching[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 12181-12190.

Real-Time Video Stitching Method Based on Dense Viewpoint Interpolation

Heng Wei[1,2], Yu Jian[1,2*], Da Feipeng[1,2,3**]

[1]*School of Automation, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[2]*Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[3]*Shenzhen Research Institute, Southeast University, Shenzhen* 518063, *Guangdong, China*

Abstract

Objective A video stitching method based on dense viewpoint interpolation is proposed to solve the problem of artifacts and defects caused by parallax when stitching under wide baseline scenes. Video stitching technology can facilitate access to a broader field of view and plays a vital role in security surveillance, intelligent driving, virtual reality, and video conferencing. One of the biggest challenges of the stitching task is the parallax. When the cameras′ optical centers perfectly coincide, they are unaffected by parallax and can easily synthesize perfect images. However, achieving the complete coincidence of camera optical centers in practical applications is not easy. The cameras are also scattered in some scenes, such as vehicle-mounted panoramic systems and wide field security surveillance systems. Therefore, it is important to study the problem of stitching in wide baseline scenes. A standard method uses a global homography matrix for alignment, but it has no parallax processing capability, which results in obvious flaws in wide baseline and large parallax scenes. In order to solve the above problems, many researchers have proposed corresponding solutions from the perspectives of multiple homography and mesh optimization. However, the mesh deformation may have significant shape distortion. Some deep learning methods combine vision tasks of optical flow, semantic alignment, image fusion, and image reconstruction to help deal with the stitching problem. However, the parameter information of cameras is not fully utilized, so the stitching results sometimes still show defects. Therefore, we wish to make full use of the parameter information of cameras and synthesize the smooth interpolated view by supplementing intermediate viewpoints between cameras to achieve better visual perception.

Methods The present study proposes a real-time video stitching method based on dense viewpoint interpolation. The method focuses on the overlapping regions of stitching and synthesizes the smooth interpolated view by supplementing dense intermediate viewpoints on the baseline of cameras, which can better align multiple inputs. In the first place, binocular camera calibration is performed to obtain internal parameters and the transformation matrix of the cameras. The original views acquired by cameras are de-distortioned and adjusted to the same horizontal plane for stitching in the horizontal direction. The maximum possible overlap regions are separated and adjusted to coplanarity and row alignment by stereo correction so that the image data can be processed in only one dimension. Subsequently, pixel-level displacement fields sampled in the original views for the overlapping regions are predicted by using the cost volume in stereo matching. Without the ground truth of the interpolated view, the network is guided to learn view generation rules by using spatial transformation relationships between viewpoints. Through the pixel-level displacement fields generated by the network, two images are sampled in the input views respectively and fused by linear weights to generate the interpolated view of the overlapping regions. Finally, the generated interpolated view is combined with non-overlapping regions of two views. The cylinder projection is performed to align the fusion boundaries of three regions and obtain the final stitching result.

Results and Discussions In this paper, the stitching results of the proposed method are compared with mainstream stitching methods. Multiband blending may show artifacts under the influence of parallax, while the method based on multiple homography and mesh optimization may have significant shape distortion in non-overlapping regions after mesh deformation. The proposed method can eliminate artifacts and smoothly align the inputs with little shape distortion, resulting in better visual perception (Fig. 9 and Fig. 10). Furthermore, we evaluate the alignment quality of the overlapping regions. The traditional methods only deal with stitching from the perspective of image features, and the alignment quality is relatively low in the case of large parallax variations. The proposed method combines camera calibration information for preprocessing and deals explicitly with the parallax problem to obtain better alignment quality (Table 1). Regarding model size and speed, the proposed method has advantages because it can initially align images after camera calibration and uses a lightweight construction method of cost volume. The processing frame rate of 720 p video can reach more than 30 fps to meet the demand for online video stitching (Table 2). In the analysis of the variation of baseline width, the proposed method can align well under different baseline widths (Fig. 12). In addition, all of them can obtain a high improvement of indicators (Table 3), which is robust to the variation of the baseline width. In conclusion, the

proposed method can improve the visual perception after stitching, eliminate artifacts, and smoothly align the inputs. It has high alignment quality, little shape distortion, and great application value because of its lightweight design and fast processing speed.

Conclusions Applying the proposed video stitching method based on dense viewpoint interpolation can effectively deal with the problem of stitching in wide baseline and large parallax scenes. The interpolated view with the smooth transition is synthesized for the overlapping regions of stitching by supplementing dense intermediate viewpoints on the baseline of the left and right cameras. A network for generating the interpolated view is proposed, which is divided into modules of feature extraction, correlation calculation, and high-resolution optimization to predict the sampling locations in the original views. The generated interpolated view is combined with the non-overlapping regions to obtain the stitching result. Moreover, the proposed method calculates the three-dimensional information at the original viewpoint in the virtual environment without the ground truth of the interpolated view. The corresponding spatial region of the interpolated viewpoint is searched by dichotomization. The interpolated view is transformed into the original viewpoint under the constructed loss function, which guides the network to learn the view generation rules. Various experiments have proved that the proposed method can improve the visual perception of video frames after stitching. It is adaptive for different baseline widths, has great generalization ability, and achieves real-time performance to meet the online stitching requirements in practical applications.

Key words machine vision; video stitching; wide baseline; deep learning; view interpolation