专题综述

计算机视觉中拥抱 Transformer 的五个理由

微软亚洲研究院 胡瀚

“统一性”是很多学科共同追求的目标，例如在物理学领域，科学家们追求的大统一，就是希望用单独一种理论来解释力与力之间的相互作用。人工智能领域自然也存在着关于“统一性”的目标。在深度学习的浪潮中，人工智能领域已经朝着统一性的目标前进了一大步。例如对于一个新的任务，基本都会遵循同样的流程对新数据进行预测：收集数据，做标注，定义网络结构，训练网络参数。但是，在人工智能的不同子领域中，基本建模的方式各种各样，并不统一，例如：自然语言处理目前的主导建模网络是Transformer；计算机视觉很长一段时间的主导网络是卷积神经网络（CNN）；社交网络目前的主导网络则是图网络等。

尽管如此，从去年年底开始，Transformer 还是在 CV 领域中展现了革命性的性能提升。这就表明 CV 和 NLP 有望统一在 Transformer 结构之下。这一趋势对于两个领域的发展来说有很多好处：（1）使视觉和语言的联合建模更容易；（2）两个领域的建模和学习经验能深度共享，从而加快各自领域的进展。

一、Transformer在视觉任务中的优异性能

视觉 Transformer 的先驱工作是谷歌在 ICLR2021 上发表的 ViT[1]，该工作把图像分成多个图像块（例如 16x16 像素大小），并把这些图像块比作 NLP 中的 token。然后，直接将 NLP 中的标准 Transformer 编码器应用于这些“token”，并据此进行图像分类。该工作结合了海量的预训练数据，例如谷歌内部 3 亿图片分类训练库 JFT-300M，在 ImageNet-1K 的 validation 评测集上取得了 88.55%的 top-1 准确率，刷新了该榜单上的记录。

ViT 应用 Transformer 比较简单直接，因为其没有仔细考虑视觉信号本身的特点。所以，它主要适应于图像分类任务，对于区域级别和像素级别的任务并不是很友好，例如物体检测和语义分割等。为此，学术界展开了大量的改进工作。其中，Swin Transformer 骨干网络[2]在物体检测和语义分割任务中大幅刷新了此前的记录，让学界更加确信 Transformer 结构将会成为视觉建模的新主流。具体而言，在物体检测的重要评测集 COCO 上，Swin Transformer 取得了单模型 58.7 的 box mAP 和 51.1 的 mask mAP，分别比此前最好的没有扩充数据的单模型方法高出了+2.7 个点和+2.6 个点。此后，通过改进检测框架以及更好地利用数据，网络的性能进一步取得了 61.3 的 box mAP 和 53.0 的 mask mAP，累计提升达+5.3 box mAP 和+5.5 mask mAP。在语义分割的重要评测数据集 ADE20K 上，Swin Transformer 也取得了显著的性能提升，达到 53.5 mIoU，比此前最好的方法高出 3.2 mIoU，此后随着分割框架和训练方法的进一步改进，目前已达到 57.0 mIoU 的性能。

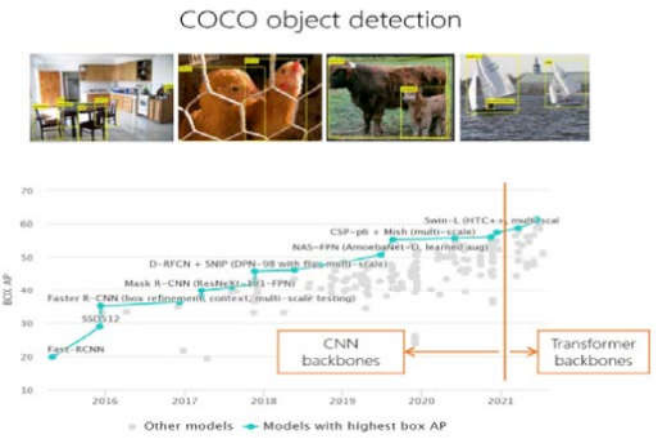

图 1　历年 COCO 物体检测评测集上的纪录

除了在物体检测和语义分割任务上表现优异外，基于 Swin Transformer 骨干网络的方法在众多视觉任务中也取得了优异的成绩，例如视频动作识别[3]、视觉自监督学习[4]、图像复原[5]、行人 Re-ID、医疗图像分割[6]等。

Swin Transformer 的主要思想也比较简单直接，就是将具有很强建模能力的 Transformer 结构和重要的视觉信号先验结合起来。这些先验主要有层次性（Hierarchy），局部性（locality）以及平移不变性的特点（translation invariance）。Swin Transformer 的一个重要设计是移位的不重叠窗口（shifted windows）。不同于传统的滑动窗，不重叠窗口的设计对硬件实现更友好，从而具有更快的实际运行速度。如下图左所示，在滑动窗口设计中，不同的点采用了不同的邻域窗口来计算相互关系，这种计算对硬件不太友好。如下图右所示，Swin Transformer 使用的不重叠窗口中，统一窗口内的点将采用相同的邻域来进行计算，对速度更友好。实际测试表明，非重叠窗口方法的速度比滑动窗口方法快 2 倍左右。在两个连续的层中，还做了移位的操作。在 L 层中，窗口分区从图像的左上角开始；在 L+1 层中，窗口划分则往右下移动了半个窗口。这样的设计保证了不重叠的窗口间可以有信息的交换。

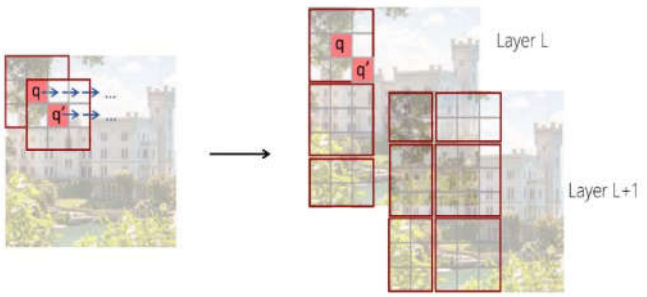

图 2　传统的滑动窗口方法（左），由于不同的查询所用到的关键字集合不同，其对存储的访问不太友好，实际运行速度较慢。移位的不重叠窗口方法（右），由于不同的查询共享关键字集合，实际运行速度更快，从而更实用

在过去大半年中，学术界视觉 Transformer 还涌现了大量变种，包括 DeiT[7]，LocalViT[8]，Twins[9]，PvT[10]，T2T-ViT[11]，ViL[12]，CvT[13]，CSwin[14]，Focal Transformer[15]，Shuffle Transformer[16]等。

二、拥抱Transformer的五个理由

除了刷新很多视觉任务的性能记录以外，视觉 Transformer 还拥有诸多好处。事实上，过去 4 年学术界不断挖掘出 Transformer 建模的各种优点，可以总结为如下图所示的五个方面。

图 3　过去 4 年学界不断挖掘出的 Transformer 建模的五个优点

2.1. 通用的建模能力

Transformer 的通用建模能力来自于两个方面：一方面 Transformer 可以看作是一种图建模方法。图是全连接的，节点之间的关系通过数据驱动的方式来学习。由于任意概念（无论具体或抽象）可以用图中的节点来表示，且概念之间的关系可以用图上的边来刻画，因此 Transformer 建模具有很强的通用性。

另一方面，Transformer 通过验证的哲学来建立图节点之间的关系，具有较好的通用性：无论节点多么异构，它们之间的关系都可以通过投影到一个可以比较的空间里面计算相似度来建立。如下图右所示，节点可以是不同尺度的图像块，也可以是“运动员”的文本输入，Transformer 均可以刻画这些异构的节点之间的关系。

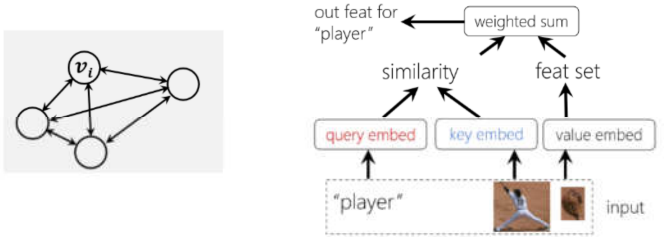

图 4　促成 Transformer 通用建模能力的两大原因：图建模（左）和验证哲学（右）

正是因为具备这样的通用建模能力，Transformer 中的注意力单元能被应用到各种各样的视觉任务中。具体而言，计算机视觉处理的对象主要涉及两个层次的基本元素：像素和物体。而计算机视觉所涉及到的任务主要就囊括了这些基本元素之间的关系，包括像素-像素，物体-像素和物体-物体的关系建模。此前，前两种关系

建模分别主要由卷积和 RoIAlign 来实现的，最后一种关系通常没有很好的建模方法。但是，Transformer 中的注意力单元由于其通用的建模能力，能被应用到所有这些基本关系建模中。近些年，在这个领域中已经出现了很多代表性的工作，例如：（1）非局部网络[17]。王小龙等人将注意力单元用于建模像素-像素的关系，并证明能帮助视频动作分类和物体检测等任务。元玉慧等人将其应用于语义分割问题，也取得了显著的性能提[18]。（2）物体关系网络[19]。注意力单元用于物体检测中的物体关系建模，这一模块也被广泛应用于视频物体分析中[20, 21, 22]。(3) 物体和像素的关系建模，典型的工作包括 DETR[23]，LearnRegionFeat[24]，以及 RelationNet++ [25]等。

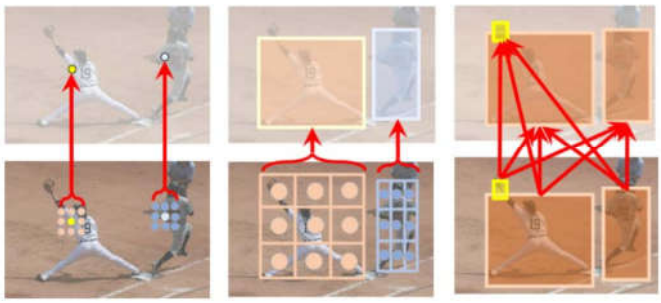

图 5　Transformer 能被应用于各种视觉基本元素之间的关系建模，包括像素-像素（左），物体-像素（中），物体-物体（右）

2.2. 和卷积形成互补

卷积是一种局部操作，一个卷积层通常只会建模邻域像素之间的关系。Transformer 是全局操作，一个 Transformer 层能建模所有像素之间的关系，它们能很好的互补。最早将这种互补性联系起来的是非局部网络[17]，在这个工作中，少量 Transformer 自注意单元被插入原始网络的几个地方，作为卷积网络的补充，并被证明在物体检测、语义分割和视频动作识别等问题中广泛有效。

此后，也有工作发现非局部网络在视觉中很难真正学到像素和像素之间的二阶关系[26]，为此，有研究员提出了一些针对这一模型的改进，例如解耦非局部网络[27]。

2.3. 更强的建模能力

卷积可以看作是一种模板匹配，图像中不同位置采用相同的模板进行滤波。Transformer 中的注意力单元则是一种自适应滤波，模板权重由两个像素的可组合性来决定，这种自适应计算模块具有更强的建模能力。

最早将 Transformer 这样一种自适应计算模块应用于视觉骨干网络建模的方法是局部关系网络 LR-Net[28]和 SASA[29]，它们都将自注意的计算限制在一个局部的滑动窗口内，在相同理论计算复杂度情况下取得了相比于 ResNet 更好的性能。然而，虽然理论上与 ResNet 的计算复杂度相同，但在实际使用中却要慢得多。一个主要原因是，不同的查询（query）使用不同的关键字（key）集合，如下图左所示，这对内存访问不太友好。

Swin Transformer 提出了一种新的局部窗口设计，称为移位窗口（shifted windows）。这一局部窗口方法将图像划分成不重叠的窗口，这样在同一个窗口内部，不同查询使用的关键字集合将是相同的，从而拥有更好的实际计算速度。在下一层中，窗口的配置会往右下移动半个窗口，从而构造了前一层中不同窗口像素间的联系。

2.4. 对大模型和大数据的可扩展性

在 NLP 领域，Transformer 模型在大模型和大数据方面展示了强大的可扩展性。下图中，蓝色曲线显示近年来 NLP 的模型大小迅速增加。大家都见证了大模型的惊人能力，例如微软的 Turing 模型、谷歌的 T5 模型以及 OpenAI 的 GPT-3 模型。

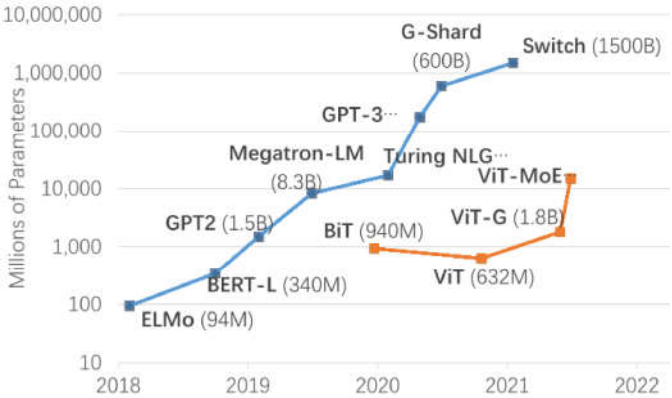

图 6　NLP 领域和计算机视觉领域模型大小的变迁

视觉 Transformer 的出现，也为视觉模型的扩大提供了重要的基础，目前最大的视觉模型是谷歌的 150 亿参数 ViT-MoE 模型[30]，这些大模型在 ImageNet-1K

分类上刷新了新的记录。

2.5. 更好的连接视觉和语言

在以前的视觉问题中，科研人员通常只会处理几十类或几百类物体类别。例如 COCO 检测任务中包含了 80 个物体类别，而 ADE20K 语义分割任务包含了 150 个类别。视觉 Transformer 模型的发明和发展，使视觉领域和 NLP 领域的模型趋同，这有利于联合视觉和 NLP 建模，从而将视觉任务与其所有概念联系起来。这方面的先驱性工作主要有 OpenAI 的 CLIP[31]和 DALL-E 模型[32]。

考虑这诸多优点，相信视觉 Transformer 将开启计算机视觉建模的新时代，我们也期待学界和业界能共同努力，进一步挖掘和探索这一新的建模方法给视觉领域带来的机遇和挑战。

责任编委 魏秀参

参考文献

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV 2021.

[3] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, Han Hu. Video Swin Transformer. Tech report 2021.

[4] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, Han Hu. Self-Supervised Learning with Swin Transformers. Tech report 2021.

[5] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, Jianfeng Gao. Efficient Self-supervised Vision Transformers for Representation Learning. Tech report 2021.

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. Tech report 2021.

[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou. Training data-efficient image transformers & distillation through attention. Tech report 2021.

[8] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, Luc Van Gool. LocalViT: Bringing Locality to Vision Transformers. Tech report 2021.

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, Chunhua Shen. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. Tech report 2021.

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. ICCV 2021.

[11] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, Shuicheng Yan. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. Tech report 2021.

[12] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, Jianfeng Gao. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. Tech report 2021.

[13] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. ICCV 2021.

[14] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, Baining Guo. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. Tech report 2021.

[15] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, Jianfeng Gao. Focal Self-attention for Local-Global Interactions in Vision Transformers. Tech report 2021.

[16] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, Bin Fu. Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer. Tech report 2021.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He. Non-local Neural Networks. CVPR 2018.
[18] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, Jingdong Wang. OCNet: Object Context for Semantic Segmentation. IJCV 2021.
[19] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, Yichen Wei. Relation Networks for Object Detection. CVPR 2018.
[20] Jiarui Xu, Yue Cao, Zheng Zhang, Han Hu. Spatial-Temporal Relation Networks for Multi-Object Tracking. ICCV 2019.
[21] Yihong Chen, Yue Cao, Han Hu, Liwei Wang. Memory Enhanced Global-Local Aggregation for Video Object Detection. CVPR 2020.
[22] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. ICCV 2019.
[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko. End-to-End Object Detection with Transformers. ECCV 2020.
[24] Jiayuan Gu, Han Hu, Liwei Wang, Yichen Wei, Jifeng Dai. Learning Region Features for Object Detection. ECCV 2018.
[25] Cheng Chi, Fangyun Wei, Han Hu. RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder. NeurIPS 2020.
[26] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, Han Hu. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. ICCV workshop 2019.
[27] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, Han Hu. Disentangled Non-Local Neural Networks. ECCV 2020.
[28] Han Hu, Zheng Zhang, Zhenda Xie, Stephen Lin. Local Relation Networks for Image Recognition. ICCV 2019.
[29] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, Jonathon Shlens. Stand-Alone Self-Attention in Vision Models. NeurIPS 2019.
[30] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, Neil Houlsby. Scaling Vision with Sparse Mixture of Experts. Tech report 2021.
[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. Learning Transferable Visual Models from Natural Language Supervision. Tech report 2021.
[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever. Zero-Shot Text-to-Image Generation. Tech report 2021.

胡瀚

微软亚洲研究院高级研究员。主要研究方向是：视觉表征学习以及视觉-语言联合表征学习。
Email: hanhu@microsoft.com

专题综述

全局和局部运动估计研究与展望

北京工业大学 毋立芳 电子科技大学 刘帅成 北京工业大学 相叶

一、引言

近年来自媒体、视频等可视媒体数据的爆炸式增长，对高质量的视频获取和智能化的视频分析提出了更高需求。运动是视频中的一个重要特征，它是获取视频的重要属性，也是视频对象行为的动态表达，在视觉关系表达、行为理解、视频稳像、视频对齐等应用中非常重要。

视频中的运动主要包括两类，一类是由相机运动导致的全局运动，如体育视频中的镜头运动、监控视频中外力影响导致的镜头晃动或者手持设备拍摄视频中的抖动等等；另一类是视频中的对象运动，称为局部运动，即视频中的运动主体发出的运动。另外还有一些与以上两类运动无关的区域如体育视频中的记分牌区域、LOGO 区域等。实际应用中，全局运动和局部运动可能存在关联性，如体育视频中的特定事件由特定运动员站位和动作组成，并且常用特定类型的相机表现手法，因而二者之间有一定关联。而在有些视频如自媒体视频或监控视频中，相机运动很多时候是干扰，不是关注重点。显然，有效估计全局和局部运动对于上述应用都非常重要。

目前大多数的运动估计方法如 PWC-Net[1]、循环全对场变换（RAFT）[2]等都是估计运动光流场，它包括全局运动和局部运动，我们称为混合运动。显然这种混合运动不能有效表达视频中的对象运动和对象行为，如图 1 所示。

除了光流场估计方法，也有一类专门估计全局运动的方法。包括传统方法[3]和基于深度学习的方法如深度单应性估计方法[4]、无监督深度学习方法[5] 、无监督深度单应性估计方法[6]等。基于深度学习的方法估计精度较高。然而这类方法无法得到视频中的局部运动。因此，有必要研究全局和局部运动估计方法。

图 1　视频图像和光流场，不同的颜色（灰度）代表不同的运动方向（幅度），不同对象颜色和亮度不同，说明其运动关联性较小。左图：篮球视频中的光流场包含全局运动、局部运动的混合运动以及静态区域；右图：相机静止，光流场表达局部运动

二、全局和局部运动分析

局部运动是视频中所有对象各自运动的总体表达，通常与对象行为或对象关系直接相关。然而作为不同的运动主体，大多数情况下，不同对象运动关联性较小，如图 1 所示。因此不同对象的运动幅度和运动方向不同，局部运动不具有移不变特性。

全局运动由相机运动产生，图像中不同位置的全局运动均服从于相同的相机运动，因此从系统的角度，全局运动具有空间移不变特性，它是一种线性移不变系统，可以用统一的系统函数来表达。在较远的场景如体育视频、监控视频中，场景较接近于一个平面，相机运动可

以用参数化模型来表达。结合全局运动的线性移不变特性，图像中全部像素点的坐标值变化均服从于统一的参数化全局运动模型，因此可以由局部区域的全局运动点来估计适用于整幅图像的全局运动参数。

相机运动包括平移(Translation)、旋转(Rotation)、缩放（Zoom in（out））、水平摇动（Pan）、垂直摇动（Tilt）。综合相应相机运动的参数化表达[8]，可以得到公式（1）。

$$\begin{cases} x' = (x+a) + (cx+d) + gx^2 + x + (\cos\theta x + \sin\theta y) \\ y' = (y+b) + (ey+f) + y + hy^2 + (-\sin\theta x + \cos\theta y) \end{cases} \quad (1)$$

进一步，可以表达为公式（2）。不同应用场景下，可以结合实际的相机运动进行简化。

$$\begin{cases} x' = m_2x^2 + m_1x + m_0 + m_3y \\ y' = n_2y^2 + n_1y + n_0 + n_3x \end{cases} \quad (2)$$

三、全局运动估计

全局运动估计的基本思路是由图像或者混合运动光流图估计相机运动参数或者全局运动图像，基本方法分为三类，分别是：传统方法、基于深度学习的方法和基于光流场分离的方法。

3.1. 传统方法

传统方法利用图像匹配技术，计算帧间变换模型，实现全局运动估计。常见的全局运动模型包括平移变换、仿射变换和单应性变换等。相较于平移、仿射，单应性变换具有更高的自由度。计算单应性变换通常需要检测和匹配特征点，比如 SIFT、SURF 等，然后通过鲁棒估计，如 RANSAC，剔除错误的匹配点，最后利用正确的匹配点拟合出单应性矩阵。

该类方法存在以下问题：（1）对于有重复纹理的场景 (比如很多建筑的窗户非常类似)，系统可能检测出很多特征点，但进行匹配时却不能有效一一对应；（2）弱纹理、无纹理。对于没有什么特征纹理的场景，系统很难在这些部分找出特征点；（3）大前景干扰。单应性变换只能拟合图像中的平面运动，非平面运动对应的特征点需要利用鲁棒估计加以排除。当图像中出现大前景干扰时，会对系统的鲁棒性造成很大挑战；（4）夜景、噪声干扰。在夜景、噪声干扰下，系统往往只能在一小块区域检测出特征点，然而用一小块区域来进行全局运动估计，效果往往不尽如人意。

3.2. 基于深度学习的方法

针对传统方法存在的问题，近年来一些研究者提出了深度学习的方法，实现更加鲁棒和准确的全局运动估计。DeTone 等人[4]通过给一个网络输入两张图像，可以直接得出单应性变换。该方法为有监督方法，模型的训练数据是人为对一张图像变形获得的，即随机产生一个单应性矩阵作为网络的监督，将该矩阵作用在任意一张图像上进行形变，形变前后的图像作为输入。然而，因为视差和运动物体的原因，真实世界不同图像之间除了角度变化还有内容上的差异。因此该方法在面对真实世界图像时效果不尽如人意。

Nguyen 等人[5]提出了一种无监督深度学习方法，通过优化图像对之间的损失，在真实数据上训练，从而克服了上述合成数据的局限。但该方法利用全图进行估算，没能有效排除图像中的运动区域和非平面区域。为了应对上述问题，Zhang 等人[6]采用无监督方法，对于输入的两张图像，提取深度特征的同时，估算一个 mask，其功能可类比为 NN RANSAC，从而剔除掉干扰区域，更鲁棒地回归单应性矩阵。

3.3. 基于光流场分离的方法

光流场是全局运动、局部运动以及场景无关区域的混合。根据线性移不变特性，可以从光流场中提取具有线性移不变性质的全局运动特征参数，再由这些特征估计出完整的全局运动。主要方法包括：

（a）基于统计分析的全局运动估计方法[7]

假设视频中不存在相机的扫换和旋转运动，则公式（2）可以简化为公式（3）。

$$\begin{aligned} x' &= m_1x + m_0 \\ y' &= n_1y + n_0 \end{aligned} \quad (3)$$

由公式（3）可以得到以下结论：（1）运动场包含水平和垂直方向两个通道，水平分量和垂直分量相互独立；（2）在 X（Y）方向的运动场分量中，运动幅度分布与

点的 Y（X）坐标无关，X（Y）坐标相同的点具有相同幅值。Y（X）坐标相同的点，运动幅度与 X（Y）坐标之间呈线性关系。

基于统计分析的全局运动估计算法基于以下常识—大多数情况下，视频边缘区域只包含全局运动。统计得到第 1 列（行）和最后一列（行）的水平（垂直）方向的运动幅值，以此为基础计算得到公式 (3) 中的参数，实现运动参数估计。统计分析法的优点是速度快，但是当视频边缘区域存在运动对象时，估计的全局运动模型参数会存在较大误差。

（b）基于迭代优化的全局运动估计[8]

体育视频转播过程中，常用到除旋转以外的相机运动，因此公式（2）简化为

$$\begin{cases} x' = m_2 x^2 + m_1 x + m_0 \\ y' = n_2 y^2 + n_1 y + n_0 \end{cases} \quad (4)$$

上式中 x 和 y 相互独立，因此可以分别估计其运动参数。将非全局运动视为异常点，利用光流中的全局点拟合相机运动模型，通过计算拟合误差逐步识别并舍弃数据空间中的异常点，提升全局运动估计结果的准确性，如图 2 所示。

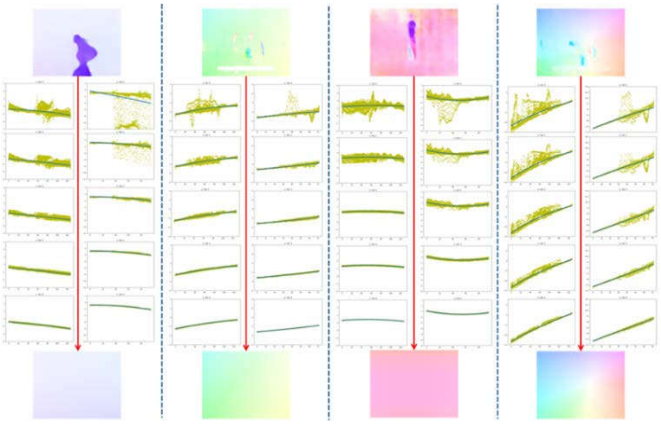

图 2　基于迭代优化的全局运动估计示例

（c）GLM-Net

迭代优化方法中 x 和 y 方向的运动估计相互独立，当图像中存在旋转时，该方法性能下降。针对这一问题，Yang 等人[9]提出了 GLM-Net，可以同时估计全局和局部运动的深度框架，如图 3 所示。设计 Mask Auto-encoder 框架实现全局运动估计的训练。将光流平铺为一维向量输入到网络中，考虑到相机模型表达最全参数为 8 个，设计编码网络最小降维到 8。其次通过解码网络将该向量解码为完整的全局运动。训练过程中，由于缺少完整的真实全局运动作为监督信号，因此，将光流中局部点的位置作为 mask 屏蔽掉，利用剩余的全局运动对网络输出进行约束。在验证阶段，网络无需监督即可从光流中估计完整的全局运动。

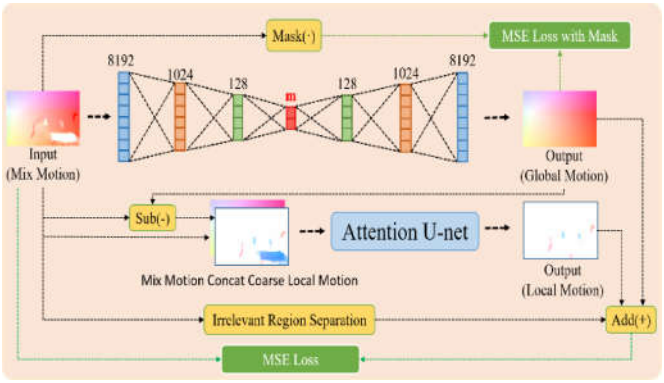

图 3　GLM-Net 网络框架

四、局部运动估计

最简单的局部运动估计方法就是由混合运动光流减去全局运动，但是当视频中存在运动无关区域如记分牌区域时，则该方法会引入运动无关区域而失效。考虑到这类运动无关区域多为静止区域，Wu 等人[7]提出了一种基于时空域阈值的局部运动估计算法。综合考虑并抑制连续 T 帧运动幅度低于设定阈值的像素点，从而有效去除场景无关区域。该方法的缺点在于需要连续 T 帧光流参与计算，无法利用单帧光流实现局部运动估计。GLM-Net（如图 3 所示）框架中，采用 Attention U-net 作为局部运动估计网络，将粗糙的局部运动和原始光流场拼接后作为输入，自动学习网络去除运动无关区域，输出优化后的局部运动区域。

五、应用

本节介绍全局和局部运动估计应用，图像数据库包括图像对齐数据库 DHE[6]和篮球比赛群体行为识别数据库 NCAA[10]以及个体行为识别数据库 UCF-101[11]。

5.1. 图像对齐

用估计得到的全局运动对第一张图像进行变换并与第二张图像进行对齐。DHE 数据库中，利用对应的两

张图像中人工标注的匹配全局点与图像对齐后的实际位置计算误差，评价图像对齐效果。对比实验结果如表 1 所示。可以看出 GLM-Net 能够得到与已有全局运动估计方法可比的结果。

表 1　不同方法的匹配点平均误差

	RE	LL	SF	LF
基于监督的方法	7.12	6.86	7.83	4.46
基于非监督的方法	1.88	2.27	1.93	1.97
SIFT + RANSAC	1.72	4.97	1.82	1.84
SIFT + MAGSAC	**1.71**	4.91	1.88	1.79
ORB + RANSAC	1.85	2.56	2.00	2.29
ORB + MAGSAC	2.02	2.78	1.92	2.25
LIFT + RANSAC	1.76	2.14	1.82	1.92
LIFT + MAGSAC	1.73	2.10	1.79	1.79
SOSNet + RANSAC	1.72	4.58	1.84	1.83
SOSNet + MAGSAC	1.73	4.39	1.76	**1.72**
CAU	1.81	**1.94**	**1.75**	1.77
GLM-Net	1.81	1.95	1.97	2.07

5.2. 相机运动估计

基于光流场分离的全局运动估计算法给出全局运动的可视化结果，进一步，由全局运动估计相机运动，不同算法的结果对比如图 4 所示，a 到 f 列分别为原始图像、原始光流、基于统计分析方法的全局运动估计结果、基于 RANSAC 方法的结果、基于迭代优化方法的结果以及基于 GLM-Net 方法的结果，每张图对应的两个数据是由全局运动估计的相机平移和缩放运动。可以看出，当视频边缘区域存在局部运动时，统计分析方法和 RANSAC 算法均存在较大误差。前者由于采用了边缘像素点，后者基于随机采样的点进行全局运动拟合，因此当局部运动在光流中占比较高时存在较大误差。第三行图像底部区域有运动无关区域，统计分析方法、RANSAC 算法和迭代优化算法的结果均不理想。由于该算法将底部垂直方向的局部运动拟合为全局运动，因此估计的全局运动中保留了部分局部运动。GLM-Net 算法在上述情况下均得到较好结果。

5.3. 行为识别

分别以原始混合光流和不同方法估计得到的局部运动作为输入，利用 3D 卷积网络 (C3D、R3D、P3D、I3D) 提取局部运动的时空特征进行行为识别。对比实验结果如表 2 所示。可以看出，基于局部运动的行为识别结果优于基于混合运动的结果，基于 GLM-Net 的结果略好于基于迭代优化的全局运动估计和基于时空域阈值的局部运动估计的结果。

图 4　不同方法估计的全局运动和相机运动参数对比。(a) 原始图像 (b) 原始光流 (c) 基于统计分析方法估计的全局运动 (d) 基于 RANSAC 方法估计的全局运动 (e) 基于迭代优化方法估计的全局运动 (f) 基于 GLM-Net 方法估计的全局运动

表 2　行为识别结果对比

	UCF-101			NCAA		
	混合光流	迭代优化+时空域阈值	GLM-Net	混合光流	迭代优化+时空域阈值	GLM-Net
C3D	0.631	0.654	**0.684**	0.650	0.688	**0.690**
R3D	0.746	0.763	**0.775**	0.652	0.693	**0.702**
P3D	0.808	0.825	**0.839**	0.668	0.701	**0.711**
I3D	0.823	0.843	**0.859**	0.675	0.722	**0.731**

六、总结与展望

本文介绍了全局和局部运动估计方法及其应用，通过对混合运动光流场进行分离，能获得准确的全局运动和局部运动。实验结果表明，有效的全局和局部运动估计方法对于估计相机运动、图像对齐以及提升行为识别性能都有很大帮助。目前的研究还比较初步，后续有诸多改进点和探索点：在目前的光流分离中，如何有效的引入全局运动的物理意义、如何由相邻两帧图像直接估计全局和局部运动、如何利用相机参数对运动估计进行强约束、如何运用场景的深度信息进行引导等，都值得进一步研究。

责任编委　储珺

参考文献

[1] Sun D, Yang X, Liu M, Kautz J, Ieee. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)2018. p. 8934-43.

[2] Teed Z, Deng J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow; proceedings of the Computer Vision – ECCV 2020, pp. 402-419.

[3] Harltey A and Zisserman A. Multiple view geometry in computer vision (2. ed.). Cambridge University Press.2006.

[4] DeTone D, Malisiewicz T, Rabinovich A. Deep Image Homography Estimation [J]. arXiv e-prints, 2016, arXiv:1606.03798.

[5] Nguyen T, Chen S, Shivakumar SS, Taylor C, Kumar V. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model [J]. IEEE Robotics and Automation Letters, 2018, 3(3): 2346-53.

[6] Zhang J, Wang C, Liu S, Jia L, Ye N, Wang J, Zhou J, Sun J. Content-Aware Unsupervised Deep Homography Estimation; proceedings of the Computer Vision – ECCV 2020, pp. 653-669.
数据库链接：https://github.com/JirongZhang/DeepHomography

[7] Wu L, Yang Z, Wang Q, Jian M, Zhao B, Yan J, Chen C. Fusing motion patterns and key visual information for semantic event recognition in basketball videos. Neurocomputing. 2020;413:217-29.

[8] Wu L, Yang Z, Jian M, Shen J, Yang Y, Lang X. Global motion estimation with iterative optimization-based independent univariate model for action recognition. Pattern Recognition [J]. 2021;116:107925. 代码链接：https://github.com/BJUT-VIP/Global-Motion-Estimation-with-iterative-optimization-based-Independent-Univariate-Model

[9] Yang Y, Xiang Y, Liu S, Wu L, Zhao B, Zeng B. GLM-Net: Global and Local Motion Estimation via Task-Oriented Encoder-Decoder Structure. ACM Multimedia (MM). 2021. 代码链接：https://github.com/BJUT-VIP/GLM-Net-Global-and-Local-Motion-Estimation-via-Task-Oriented-Encoder-Decoder-Structure

[10] Ramanathan V, Huang J, Abu-El-Haija S, Gorban A, Murphy K, Fei-Fei L. Detecting events and key actors in multi-person videos; proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), F 27-30 June 2016, 2016 [C]. 数据库链接：https://www.kaggle.com/ncaa/ncaa-basketball

[11] Soomro K, Roshan Zamir A, Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild [J]. arXiv e-prints, 2012, arXiv:1212.0402. 数据库链接：https://www.kaggle.com/pevogam/ucf101

毋立芳

北京工业大学信息学部教授，研究方向：视觉内容理解、智能 3D 打印、社会媒体计算。
Email: lfwu@bjut.edu.cn

刘帅成

电子科技大学信息与通信工程学院副教授，研究方向：图像/视频处理，底层计算机视觉。
Email: liushuaicheng@uestc.edu.cn

相叶

北京工业大学讲师，研究方向：视频群体行为识别，视频分析与处理等。
Email: xiangye@bjut.edu.cn

热点追踪

基于运动知识的视觉 SLAM 回环检测

中科院自动化研究所 刘秉熙 唐付林 傅禹杰 吴毅红

一、摘要

SLAM 系统在对未知环境的长期探索后，不可避免地产生轨迹预估误差和建图误差。视觉回环检测是对这一问题的公认解决方案，可以理解为一个在线的图像检索问题，要求实时、鲁棒的匹配当前地点与先前参观过的地点。基于局部特征的聚类技术广泛应用于回环检测，但不能很好地在移动平台上同时满足低时耗和高准确率。本文提出基于运动知识的视觉 SLAM 回环检测算法。这里的运动知识包括连续运动模型、基于网格的运动统计和运动状态区分。更进一步我们设计了一种灵活且有效的决策来决定局部特征和全局特征的使用。相关成果被 ICRA 2021 录取为口头报告。

二、引言

SLAM 系统在对未知环境的长期探索后，不可避免地产生轨迹预估误差和建图误差[1,2]。视觉回环检测是对这一问题的公认解决方案，可以理解为一个在线的图像检索问题，要求实时、鲁棒的匹配当前地点与先前参观过的地点，如图 1 所示。人工设计的全局特征计算较为快速，但易受光照、视角变化的影响。人工设计的局部特征鲁棒能解决视角问题，但计算比较耗时。局部特征的聚类技术被提出，其中基于无监督训练的词袋模型广泛应用于回环检测[1,2,3]。随着深度学习的发展，卷积神经网络在图像表达取得惊人的表现，同时逐渐被尝试应

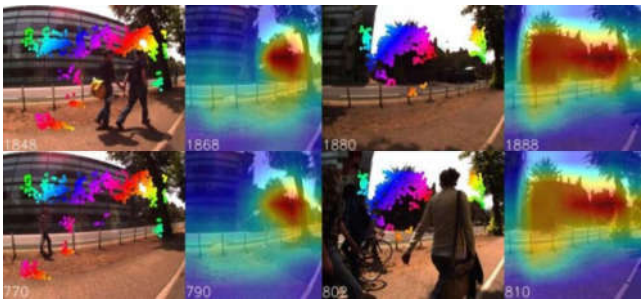

图 1　回环检测实例

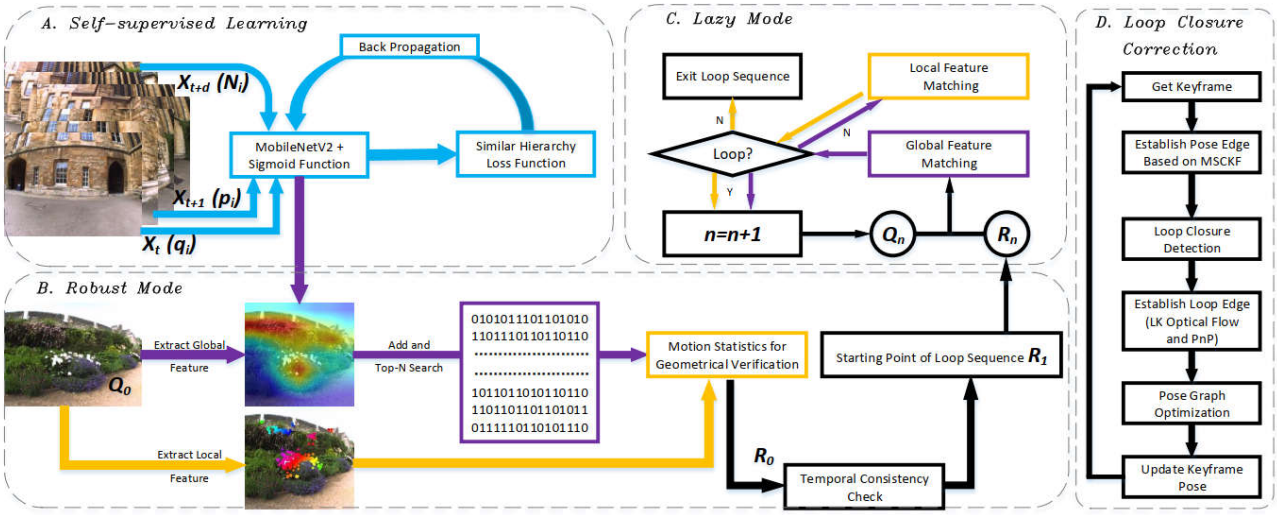

图 2　回环检测整体框架

用于位置识别和回环检测[4]。但最新提出的基于CNN的回环检测方法既没有考虑在移动平台上的实时运行表现，也没有充分融合运动学知识。因此，该工作着重研究一个与运动学知识紧密相关的 SLAM 回环检测和位姿优化系统，如图 2 所示。

三、正文

首先，一种基于连续运动模型的自监督标签方法被提出，即固定某个时间戳的图像为当前帧，时间序列上越是靠近当前帧的图像应该是更加相似的，而与当前帧相隔时间越长的图像相似度越低。被训练的轻量化网络用于提取图像的全局特征，两个全局特征之间可以快速计算汉明距离且准确地度量场景相似性。

仅依赖全局特征的检索是不鲁棒的且无法计算位姿，所以回环候选帧被提取局部特征并对这些特征进行匹配。在本文的研究工作中，综合评价了多个技术方案，选取了基于网格的运动统计的局部特征匹配方法作为回环检测系统的几何一致性检验模块，高效地解决了视角变化和遮挡问题。

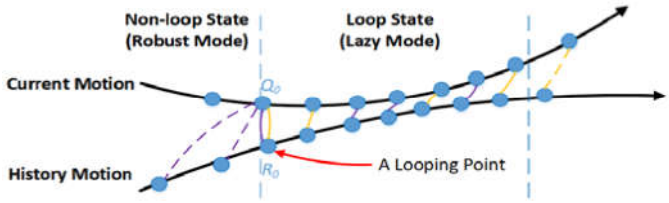

图 3　回环的抽象蓝图

最后，如图 2 所示，我们通过区分运动状态，设计了一个充分利用线性存储结构和有机融合全局和局部特征的检测策略。当一个运动系统到达回环点时，它在将接下来的一段时间处于一个回环路径。因此，我们区分运动状态为非回环状态和回环状态，如图 3 所示。假设查询图像Q_0检测到R_0，则后续帧Q_i会检测到R_i。被区分的两种运动状态分别对应回环检测系统中的两种模式：鲁棒模式和偷懒模式。鲁棒模式下，我们利用了全局特征检索、局部特征验证和时间一致性验证；偷懒模式下则是两种特征自适应交替使用，目的是提高检测速度的同时可以适应尺度或视角变化较大的场景。

表 1　位姿优化过程中的各项实验数据

Stages	Outdoor1	Outdoor2
Number of Keyframes	1101	432
Total Optimization Time (ms)	98.14	33.24
Mean Optimization Time (ms/keyframe)	0.029	0.077
Reprojection Error (pixel)	1.53	1.83

表 2　100%准确率下不同算法的召回率

	City Centre	New College	KITTI 00	KITTI 05
FAP-MAP 2.0[1]	40.11	52.63	61.22	48.51
DLoopDetector[2]	30.59	47.56	72.43	51.97
An et al.[3]	66.48	76.74	91.23	85.15
Tsintotas et al.[4]	52.44	16.30	**93.18**	**94.20**
Proposed	**86.01**	**91.21**	93.02	92.53

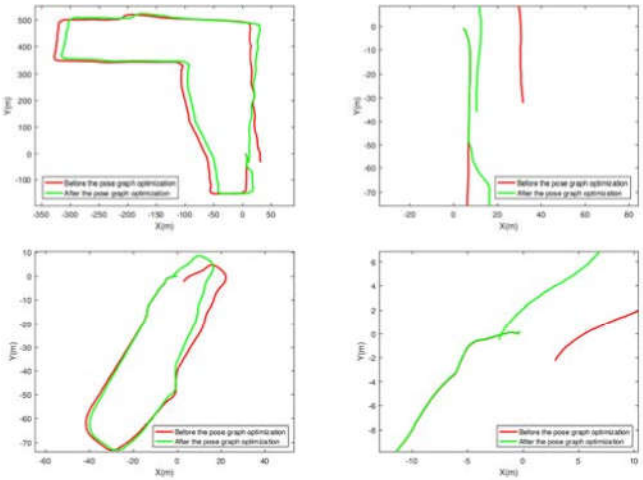

图 4　真实场景下的回环检测及误差矫正

利用上述回环检测系统和课题组的 VIO 系统，进一步设计了一个位姿优化模块用于纠正累计误差。我们提出的系统在多个公开数据集和真实场景数据下进行了大量实验，并与先进方法进行了结果对比。表 1 展示了被量化的平均优化时间和重投影误差。图 4 展示了真实场景下的位姿图优化前后的轨迹。表 2 展示我们的算法和其他先进算法在公开数据集下结果对比，评价指标是 100%准确率下的召回率。我们提出的算法在 New College 上比结果最好的算法要高出 14.47%。更重要的是，我们的结果在多个数据集下是比较稳定的。

责任编委　崔海楠

参考文献

[1] Cummins M, Newman P. Appearance-only SLAM at large scale with FAB-MAP 2.0[J]. The International Journal of Robotics Research, 2011, 30(9): 1100-1123.

[2] Gálvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics, 2012, 28(5): 1188-1197.

[3] Yue H, Miao J, Yu Y, et al. Robust Loop Closure Detection based on Bag of SuperPoints and Graph Verification[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2019: 3787-3793.

[4] Tsintotas K A, Bampis L, Gasteratos A. Probabilistic appearance-based place recognition through bag of tracked words[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 1737-1744.

刘秉熙

中科院自动化研究所硕士生。主要研究方向为视觉定位。
Email: bingxi.liu@nlpr.ia.ac.cn

唐付林

中科院自动化研究所助理研究员。主要研究方向为 SLAM。
Email: fulin.tang@nlpr.ia.ac.cn

傅禹杰

中科院自动化研究所博士生。主要研究方向为图像匹配。
Email: yujie.fu@nlpr.ia.ac.cn

吴毅红

中科院自动化研究所研究员。主要研究方向为相机定位与标定、三维重建、SLAM 等。
Email: yhwu@nlpr.ia.ac.cn