光电技术及应用

DOI：10.16818/j.issn1001－5868.2021092601

基于无监督级联的亚像素单应矩阵估计

吴柔莞[1,2]，徐智勇[1*]，张建林[1]

（1. 中国科学院光电技术研究所，成都 610209；2. 中国科学院大学，北京 100049）

摘　要： 为提高单应性估计的准确性和解决真实标注难获取的问题，提出一种具有修正功能的无监督单应性估计算法。该算法采用级联结构，其思想类似于迭代，其中每一级网络都保持相同的层数和参数量，下一级网络输出的单应性矩阵为真实矩阵与之前输出单应矩阵和的残差。考虑到模型复杂度和实时性的需求，文章采用两级网络级联。通过在COCO数据集中的5 000张图片上进行验证，结果表明，相比传统方法和其他基于深度学习的方法，所设计的级联无监督算法具有更准确的估计能力，其在测试集中的平均像素误差为0.54，较传统方法下降95.38%，运行速度达到95 f/s。

关键词： 单应性估计；无监督学习；级联结构；回归网络；深度学习

中图分类号： TP391　**文章编号：** 1001－5868(2022)01－0158－06

Sub-pixel Homography Matrix Estimation Based on Unsupervised Cascade

WU Rouwan[1,2], XU Zhiyong[1], ZHANG Jianlin[1]

（1. The Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, CHN;
2. University of Chinese Academy of Sciences, Beijing 100049, CHN）

Abstract: In order to improve the accuracy of homography estimation and solve the problem that it is difficult to obtain real labels, an unsupervised homography estimation algorithm with correction function is proposed. The algorithm adopts a cascade structure, and its idea is similar to iteration, in which each level network maintains the same number of layers and parameters, and the homography of the output of the next level network is estimated as the residual of the sum of the real matrix and the previous output homography matrix. Considering the requirements of model complexity and real-time, a two-level network cascade is adopted in this paper. Through the verification on 5 000 images in the COCO dataset, it shows that the cascade unsupervised algorithm designed in this paper has more accurate estimation ability than traditional methods and other methods based on deep learning. Its average pixel error in the test set is 0.54, which is 95.38% lower than that of traditional methods, and the running speed reaches 95 f/s.

Key words: homography estimation; unsupervised learning; cascade structure; regression network; deep learning

0　引言

单应性估计代表不同角度平面的两个图像之间的映射，它在计算机视觉应用中发挥着重要作用，比如图像拼接[1]、单目SLAM[2]、图像超分辨重建[3]、视频稳像[4]、图像去噪[5]等。

采用传统方法估计单应性矩阵普遍需要计算和匹配特征点，常采用的特征点有SIFT（Scale-Invariant Feature Transform）[6]、ORB（Oriented FAST and Rotated BRIEF）[7]等。具体步骤为：先

收稿日期：2021－09－26.

*通信作者：徐智勇　E-mail：xzy158@163.com

获得尽量多组的匹配特征点，接着通过直接线性变换[8]（Direct Linear Transform，DLT）和 RANSAC[9]相结合的方法来剔除异常点，并估计单应性矩阵。虽然基于特征点的方法通常可以获得良好的性能，但是它们高度依赖图像的质量。除此以外，在某些特定场景的图像会存在低亮度、无纹理区域的情况，这会导致特征点匹配误差大和个数分布不均匀，最终得出不准确的单应性估计。因此，对于这些非常规场景，采用基于特征点的方法得出单应性估计是具有挑战性的。

近些年，随着深度学习的快速发展，采用深度学习进行单应性估计的方法被相继提出。HomographyNet[10]是第一个基于深度学习进行单应性估计的方法，它采用有监督的形式对网络进行训练，虽然在特定场景的鲁棒性要强于传统方法，但是其误差依旧很大。随后，Nowruzi 等[11]提出采用分层卷积的方法，虽然减小了矩阵估计误差，但是其基于监督的学习模式对标签数据的需求限制其应用的普遍性和进行实际应用的开销代价。之后，Nguyen 等[12]为解决这一难题，提出采用无监督的形式对网络进行训练，虽然解决了数据标签问题，但是其网络的误差与有监督的方法相比，并没有很显著的下降。国内此类研究较少，其中郭梦夏等[13]采用 HomographyNet 的方法对视频序列进行单应性估计，即计算视频前后两帧的单应性矩阵，其网络估计准确度也是该文献未来需要解决的问题。

本文结合无监督学习的优点，提出了一种简单可扩展的级联结构进行图像视频序列间的单应性矩阵的高精度估计，通过在 MS-COCO[14]（Microsoft Common Objects in Context）数据集上的实验，表明级联框架可以通过上一级网络的输出来补偿下一级网络的输入，达到修正的目的，最后将所有阶段网络输出相加得到准确单应性矩阵，估计精度达亚像素级。除此以外，所提出的模型在满足准确性的同时，其简单的网络结构具有好的高效实时性，能广泛应用于各种视觉任务的预处理操作。

1　单应性估计

1.1　单应性矩阵引入

单应性矩阵对视频序列中两帧图像相对成像系统的视角变化进行描述。假设两张图像的对应点对的齐次坐标分别为 $\boldsymbol{x}=(u,v,1)^{\mathrm{T}}$ 和 $\boldsymbol{x}'=(u',v',1)^{\mathrm{T}}$，单应性矩阵 $\boldsymbol{H}$ 如式(1)所示，则有两坐标点与单应性矩阵的关系如式(2)：

$$\boldsymbol{H}=\begin{bmatrix}h_{11} & h_{12} & h_{13}\\ h_{21} & h_{22} & h_{23}\\ h_{31} & h_{32} & h_{33}\end{bmatrix} \tag{1}$$

$$\begin{bmatrix}u'\\ v'\\ 1\end{bmatrix}=\begin{bmatrix}h_{11} & h_{12} & h_{13}\\ h_{21} & h_{22} & h_{23}\\ h_{31} & h_{32} & h_{33}\end{bmatrix}\begin{bmatrix}u\\ v\\ 1\end{bmatrix} \tag{2}$$

即序列中某点的坐标在其对应变化帧中的坐标为该点对应式(2)的单应变化，经计算可以得其坐标如式(3)：

$$\begin{cases}u'=\dfrac{h_{11}u+h_{12}v+h_{13}}{h_{31}u+h_{32}v+h_{33}}\\ v'=\dfrac{h_{21}u+h_{22}v+h_{23}}{h_{31}u+h_{32}v+h_{33}}\end{cases} \tag{3}$$

对单应性矩阵的 9 个参数进行归一化处理，添加 $\boldsymbol{H}$ 矩阵模为 1 的约束条件，如式(4)所示：

$$h_{11}^2+h_{12}^2+h_{13}^2+h_{21}^2+h_{22}^2+h_{23}^2+h_{31}^2+h_{32}^2+h_{33}^2=1 \tag{4}$$

假如我们得到两幅图像中对应的 N 个点对，可以得到如式(5)所示的矩阵形式，矩阵 $\boldsymbol{A}$ 的维度由配对点个数决定[15]，$\boldsymbol{h}$ 为向量。

$$\boldsymbol{A}\boldsymbol{h}=0 \tag{5}$$

因为对单应性矩阵进行了归一化约束，所以在求解 8 个自由度的 $\boldsymbol{H}$ 矩阵时，至少需要 4 组匹配点对。但是，在实际应用中，往往会因为图像质量不高而出现特征点误匹配、定位点误差过高等。因此，为了得到准确的结果，一般会使用高于 4 组点对，并结合 RANSAC 方法进行单应性矩阵计算。

1.2　基于传统的单应性估计

基于传统方法计算单应性矩阵，最常使用的是基于特征点匹配的方法，一般有以下 5 个步骤：

(1)提取每张图像的 SIFT/ORB 等特征点；

(2)计算特征点的描述子；

(3)通过特征点描述子建立匹配关系；

(4)使用 RANSAC 方法对匹配后的特征点进行误匹配剔除操作；

(5)根据步骤(4)所得特征点，求解式(3)和式(4)组成的方程组，计算单应矩阵。

1.3　基于深度学习的单应性估计

虽然深度学习在计算机视觉领域已经取得了巨大成功，但是很多方向的第一次深度学习体验都是从分类网络开始，单应性估计也不例外[16]。

HomographyNet 是第一个将深度学习应用于单应矩阵估计的网络，其网络结构采用 VGG 风格[17]的形式，均采用 3×3 的卷积核，共 8 个卷积层，2 个全连接层。文献[11]的核心网络结构与 HomographyNet 类似，只是其更深，更复杂。文献[12]也延续了相同的网络结构，只是由有监督学习改进为无监督学习。

以上算法的训练皆采用 MS-COCO 数据集，在其上选取任意的图像，对该图像经过随机变换和相应扰动，与原始图像组成图像对，构成网络的训练集。

2 无监督级联网络

本节将详细介绍所提出的网络结构，并展示如何从一对图像中计算单应性矩阵。

2.1 无监督单应估计学习

虽然基于监督的深度单应估计方法已经取得了不错的效果，但在实际应用中，真实标注是难以获取的，因此会出现模型无法落地或迁移精度差等问题。针对此问题，文献[12]提出了一种无监督的训练方式，该网络不再直接输出单应矩阵的 8 个参数，而是输出 4 个点的扰动量。采用直接线性变换的方法，计算初始点与加上扰动量后的初始点的单应矩阵，以此单应矩阵来作为前后序列的单应性估计。本文与文献[12]一样皆采用边界点的扰动量作为网络输出，以平均 L_1 灰度差值作为损失函数，如式(6)所示，其中 I_A 和 I_B 代表一对图像，$\boldsymbol{H}$ 代表单应性矩阵，齐次坐标点 $x_i=(u_i,v_i,1)^T$。

$$L=\frac{1}{|x_i|}\sum_{x_i}|I_A(H(x_i))-I_B(x_i)| \quad (6)$$

2.2 单级无监督网络结构

与目前基于深度学习的方法一样，本文依然采用 VGG 风格的网络作为核心结构，其单级无监督网络结构如图(1)所示，网络输入为一对待求取单应性估计的图像块。

首先对缩放后的 I_A 在随机位置 p 处裁剪得到

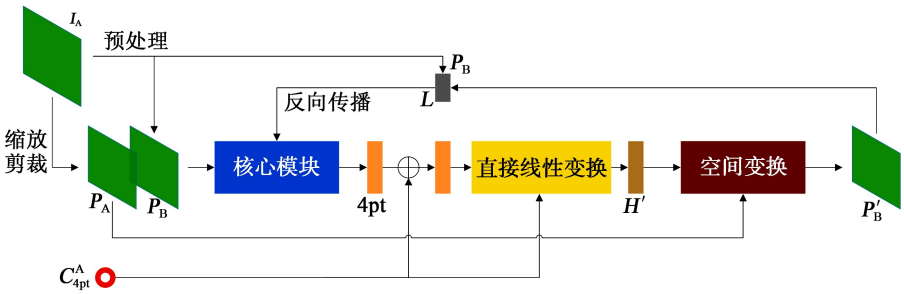


图 1 单级无监督网络结构

大小为 128×128 的方块 P_A，对方块 P_A 的四个边框点 C^A_{4pt} 进行$[-s,s]$之间的随机扰动，s 的最大值为 32，通过四个边框点与扰动后的四个边框点进行直接线性变换得到真实单应性矩阵 $\boldsymbol{H}$。接着对 I_A 进行单应性变换得到 I_B，在 I_B 的位置 p 裁剪相同大小的方块得到 P_B。P_A 和 P_B 在颜色通道上进行拼接输入核心模块，输出 4 个边框点的 8 个扰动值，将原始边框点与预测出的扰动值相加，得到 P_B 中四个边框点坐标的预测值，再通过直接线性变换得到预测的单应性矩阵 $\boldsymbol{H}'$。最后将 $\boldsymbol{H}'$ 与 P_A 进行空间变换，其思想类似于空间变换网络[19](Spatial Transformer Networks，STN)，得到 P'_B。

其中核心模块的结构如表 1 所示。

表 1 核心模块网络结构

网络层	输出通道	运算形式
Conv1	64	3×3 conv+ReLU+BN
	64	3×3 conv+ReLU+BN
Pooling1	64	2×2 max pooling，stride=2
Conv2	64	3×3 conv+ReLU+BN
	128	3×3 conv+ReLU+BN
Pooling2	128	2×2 max pooling，stride=2
Conv3	128	3×3 conv+ReLU+BN
	128	3×3 conv+ReLU+BN
Pooling3	128	2×2 max pooling，stride=2
Conv4	128	3×3 conv+ReLU+BN
	128	3×3 conv+ReLU+BN
Regression	1 024	fully connection+ReLU+dropout
	8	fully connection+ReLU

2.3 级联无监督学习模型

为了得到更加准确的单应矩阵估计，本文提出

采用级联结构来修正各级网络的输出，其思想类似于迭代。其中第 n 个网络输出的单应性估计为真实值与前 $n-1$ 次单应性估计和的残差。通过多个相同网络输出的单应性矩阵相加达到修正的目的。如式(7)所示，其中 $\boldsymbol{H}^*$ 代表预测得到的单应性估计，$\boldsymbol{H}'_i$ 代表第 i 个网络输出的单应性矩阵。

$$\boldsymbol{H}^* = \sum_{i=1}^{n} \boldsymbol{H}'_i \tag{7}$$

单级网络的参数量为 35.2 M，考虑到模型复杂度和实时性的要求，本文采用两级网络级联，结构如图 2 所示。其中灰色梯形所表示的单级无监督网络结构如图 1 所示。

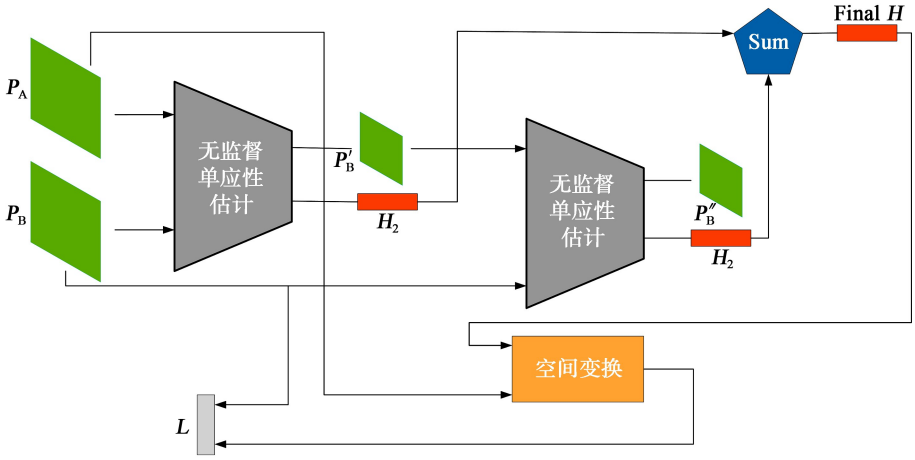


图 2　两层级联网络结构

3　实验与分析

本文使用 Ubuntu 20.04 搭载 Pycharm，编程框架选择 Pytorch[20]，电脑使用 Intel Core i7-10700 @3.80 GHz 16 RAM 和 NVIDIA GeForce RTX 2060 6 G GPU 的配置，算法执行的平均运行速度为 95 f/s。训练算法过程中，批次大小(Batchsize)为 32，学习率(Learning Rate)为 1×10^{-4}，优化器使用 Adam，丢失概率为 0.5，总迭代次数为 1.8×10^5，每经过 6×10^4 次的迭代，学习率缩小为原来的 1/10。

3.1　数据集的生成

采用 MS-COCO2017 数据集，首先将所有图像大小调整为 320×240 并转换为灰度图。I_A 为图像集中的一幅，在随机位置 p 处裁剪大小为 128×128 的正方形块 P_A，对 P_A 的边框点进行[$-s$，s]的随机扰动，s 最大值为 32。根据原始边框点和扰动后的边框点直接线性变换得到真实单应性矩阵 $\boldsymbol{H}$，I_A 与 $\boldsymbol{H}$ 做空间变换得到 I_B，在 I_B 的 p 位置裁剪相同大小的正方形块得到 P_B。最后将 P_A 和 P_B 在色彩通道维度上拼接得到一组输入图像对。共生成 118 287 组图像对作为训练集，5 000 组作为测试集。

3.2　实验结果与分析

像素误差是计算四个真实边框点与预测边框点之间的 L_2 距离和的平均。平均像素误差是测试集上像素误差和的平均。比较了本文方法和其他几种方法的平均像素误差，结果如表 2 所示。

表 2　各算法在测试集上结果

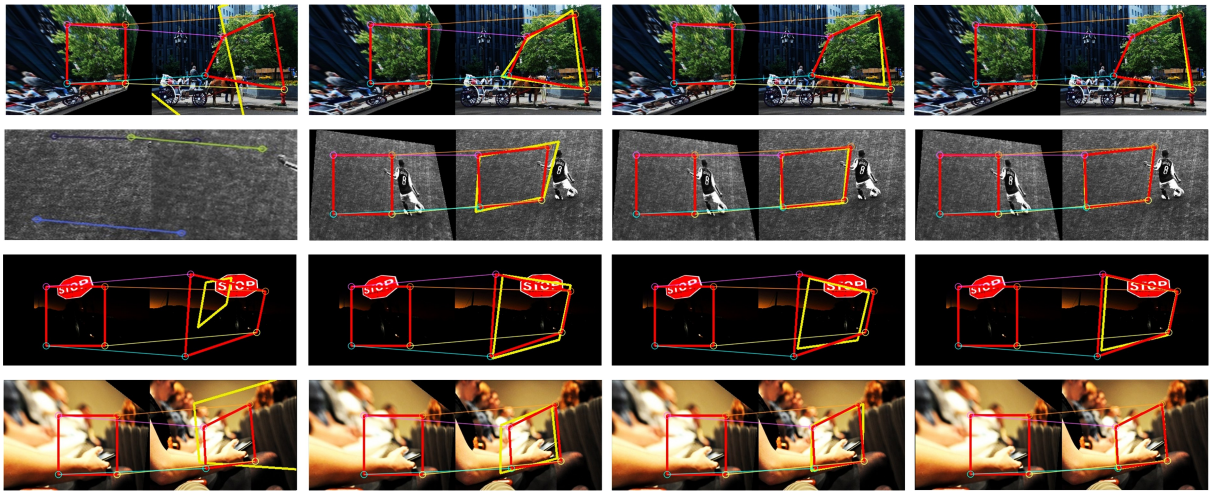
方法	平均像素误差	降低比例/%
ORB+RANSAC	11.7	—
HomographyNet[10]	9.20	21.37
分层网络[11]	3.91	66.58
无监督方法[12]	1.40	88.03
本文方法	0.54	95.38

在实验中，本文算法与三种基于深度学习的方法均采用 VGG 风格的特征提取网络，根据表 2 所示误差数据，与有监督方法相比，本文算法不仅解决了数据标注的问题，还大大提高了单应性估计的精度；与文献[12]的无监督方法相比，本文算法的误差更小，证明采用级联结构进行网络搭建可以对单应性估计进行有效修正；与传统方法相比，本文算法在精度上实现了质的飞跃。

如图 3 所示，对 COCO 数据集上四类不同质量和扭曲程度的单应性变换图像对进行可视化操作，其中右侧红色框代表真实边界框，黄色框代表预测边界框，左侧红色框内图像表示输入网络中的 P_B，传统方法若失败，则展示其特征点匹配图。

图 3(a)～(d)分别代表随机扰动大、纹理信息少、低亮度和图像模糊这四种情况。第一列结果展

示了基于传统方法进行单应性估计的可视化结果，后三列展示了基于深度学习进行单应性估计的可视化结果。通过几列可视化结果比较，不难发现，在非常规场景下，尤其是纹理信息少的情况下，基于深度学习的方法比传统方法具有更好的鲁棒性。基于特征点进行单应性估计的传统方法会因为无法获取足够匹配特征点对而出现单应性估计失败的问题。而采用基于深度学习的方法，只要训练集中有满足此类情况的数据，就会在一定程度上使网络具有单应性估计的能力。



（a） ORB+RANSAC （b） HomographyNet[10] （c） 无监督方法[12] （d） 本文方法

图 3 不同算法在不同情况下的单应性估计可视化结果

第二列和第三列结果相比较，可以说明无监督方法比有监督方法具有更好的单应性估计能力，一定程度上证明，使用无监督的训练方式对网络进行学习可以使其具有更好的泛化性能。第三列和第四列结果相比较，可以证明本文所提出的级联网络结构具有修正单应性矩阵的功能。不管是在哪一类情况下，本文所提算法预测出的边界框都与真实框更加接近，即单应性估计更加准确。因此，通过以上实验可以证明，基于无监督学习的级联网络不仅可以修正单应性估计，还可以得到更准确的单应性变换。

为了验证级联结构对单应性估计的修正作用，我们在硬件条件允许的情况下，做了单级结构和本文所用的两层级联结构的消融实验，二者在测试集上的平均像素误差分别为 1.40 和 0.54。

实验结果表明，通过简单可扩展的级联结构可以提高单应性估计的准确度，使用两级级联结构即可达到亚像素的精度。

4 总结

本文采用无监督学习的方式，并针对现有单应性网络估计误差大的问题，提出了一种级联结构，通过上一级网络的输出来补偿下一级网络的输入，达到修正单应性估计的目的。在 COCO 数据集上的实验结果验证了本文方法的有效性，其精度达到亚像素级别。虽然本文方法提高了单应性估计的准确性，但是在训练和测试中还是以局部运动来代替全局运动，如何通过引入注意力机制来完成对整幅图像有效信息的提取是以后研究工作需要解决的问题。

参考文献：

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

[2] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. IEEE Trans. on Robotics, 2015, 31(5): 1147-1163.

[3] 鲁 甜，刘 蓉，刘 明，等. 基于特征图注意力机制的图像超分辨率重建[J]. 计算机工程，2021，47(3)：261-268.
Lu Tian, Liu Rong, Liu Ming, et al. Image super-resolution reconstruction based on attention mechanism of feature map [J]. Computer Engineering, 2021, 47(3): 261-268.

[4] Wu R, Xu Z, Zhang J, et al. Robust global motion estimation for video stabilization based on improved *K*-means clustering and superpixel[J]. Sensors, 2021, 21(7): 2505.

[5] 张墨华，彭建华. 面向图像先验建模的可扩展高斯混合模型[J]. 计算机工程，2020，46(4)：220-227.
Zhang Mohua, Peng Jianhua. Extensible Gaussian mixture model for image prior modeling[J]. Computer Engineering, 2020, 46(4): 220-227.

[6] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. Inter. J. of Computer Vision, 2004, 60(2): 91-110.

[7] Rublee E, Rabaud V, Konolige K, et al. ORB: an efficient alternative to SIFT or SURF[C]// IEEE Inter. Conf. on Computer Vision, 2011.

[8] Szeliski R. Image alignment and stitching: A tutorial[J]. Foundations and Trends® in Computer Graphics and Vision, 2006, 2(1): 1-104.

[9] Fischler M A, Bolles R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography-ScienceDirect[J]. Readings in Computer Vision, 1987: 726-740.

[10] Detone D, Malisiewicz T, Rabinovich A. Deep image homography estimation[J]. RSS Workshop on Limits and Potentials of Deep Learning in Robotics, 2016: arXiv: 1606.03798.

[11] Japkowicz N, Nowruzi F E, Laganiere R. Homography estimation from image pairs with hierarchical convolutional networks[C]// The IEEE Inter. Conf. on Computer Vision (ICCV), 2017.

[12] Nguyen T, Chen S W, Shivakumar S S, et al. Unsupervised deep homography: A fast and robust homography estimation model[J]. IEEE Robotics and Automation Letters, 2017: arXiv: 1709.03966.

[13] 郭孟夏. 一种基于深度学习的视频单应性估计算法[J]. 现代计算机, 2021, 2(6): 5-10.
Guo Mengxia. A video homography estimation algorithm based on deep learning[J]. Modern Computer, 2021, 2(6): 5-10.

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]// European Conf. on Computer Vision, 2014.

[15] Andrew A M. Multiple View Geometry in Computer Vision [M]. Cambridge: Kybernetes, 2001.

[16] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014: arXiv: 1409.1556.

[18] Zhang J, Wang C, Liu S, et al. Content-aware unsupervised deep homography estimation[J]. Springer, 2020: arXiv: 1909.05983.

[19] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 2017-2025.

[20] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch[C]// 31st Conf. on Neural Information Proc. Systems, 2017.

作者简介：

吴柔莞(1997—)，女，四川省德阳市人，硕士研究生，主要研究方向为视频稳像中的运动估计；

徐智勇(1970—)，男，四川省成都市人，研究员，主要研究方向为图像处理及视频序列分析。