基于孪生网络的快速视频目标分割

付利华,赵 宇,孙晓威,卢中山,王 丹,杨寒雪
(北京工业大学信息学部,北京 100124)

摘 要: 视频目标分割是计算机视觉领域中的一个研究热点,传统基于深度学习的视频目标分割方法在线微调深度网络,导致分割耗时长,难以满足实时的需求. 本文提出一种快速的视频目标分割方法. 首先,参数共享的孪生编码器子网将参考流和目标流映射到相同的特征空间,使得相同的目标具有相似的特征. 然后,全局特征提取子网在特征空间中匹配给定目标相似的特征,定位目标对象. 最后,解码器子网将目标特征还原,并通过连接目标流的低阶特征,提供边缘信息,最终输出目标的分割掩码. 在公开基准数据集上的实验表明,本文方法的分割速度有大幅度提升,同时具有较好的分割效果.

关键词: 视频目标分割; 计算机视觉; 深度学习; 孪生网络; 特征空间
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112 (2020)04-0625-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.04.001

Fast Video Object Segmentation Based on Siamese Networks

FU Li-hua, ZHAO Yu, SUN Xiao-wei, LU Zhong-shan, WANG Dan, YANG Han-xue
(*Faculty of Information Technology, Beijing University of Technology, Beijing* 100124, *China*)

Abstract: Video object segmentation (VOS) is a research hotspot in the field of computer vision. Traditional VOS based on deep learning fine-tunes the deep network online, which leads to long time-consuming segmentation and is difficult to meet real-time requirements. Therefore, we propose a fast VOS method. First, the weight-shared siamese encoder subnet maps the reference stream and the target stream to the same feature space; so that the same objects have similar features. Then, the global feature extraction subnet matches the features similar to the given object to locate the object. Finally, the decoder subnet restores the object features and gets edge information by connecting the low-level features of target stream to output the mask. Experiments on public benchmark datasets show that our method improves the speed significantly and achieves good performance.

Key words: video object segmentation; computer vision; deep learning; siamese network; feature space

1 引言

视频目标分割(Video Object Segmentation, VOS)是在不知道视频帧语义的情况下,自动计算视频帧序列中前景对象的像素级掩码. 根据第一帧给定的目标分割掩码,在视频后续帧中分割出标注的特定目标,即为半监督视频目标分割. 半监督视频目标分割广泛应用于基于视频理解的精确对象跟踪、交互式视频编辑和增强现实等领域.

目前,关于半监督视频目标分割的方法大多基于两种主流的方法:基于目标检测[1~7]的单次视频目标分割法(One-Shot Video Object Segmentation, OSVOS)[1]和基于掩码传播[8~14]的掩码跟踪法(MaskTrack)[7].

OSVOS等基于目标检测的视频目标分割方法[1~6],主要通过匹配第一帧给定目标的外观特征,实现视频目标的分割. 这类方法能够有效解决目标遮挡等问题,在处理目标外观稳定的视频时,能得到较好的分割结果. 但由于其没有考虑帧间的时序信息,因此,当目标的外观发生较大变化时,其分割精度会大幅下降.

MaskTrack等基于掩码传播的视频目标分割方法[7~11],主要通过传播前一帧的分割结果,从而为当前帧的分割位置提供指导. 这类方法考虑了连续帧间的时序信息,因此能够很好地适应目标复杂的外观变化. 但遮挡和快速运动等会影响目标分割掩码的传播过

收稿日期:2018-12-19;修回日期:2019-05-08;责任编辑:李勇锋
基金项目:北京市自然科学基金(No. 4173072)

程,并且多个相似目标重叠容易造成跟踪飘移,导致方法的目标分割性能下降.

OSVOS 和 MaskTrack 都是采用在线微调的方式进行视频目标分割. 这种方式根据视频第一帧图像与给定的目标掩码对网络进行在线微调,使得网络模型具有记忆给定目标外观的能力. 但由于在线微调需要对网络进行多次迭代训练,会大大增加分割的时间.

孪生网络是一种度量学习方法,其通过神经网络将两个输入映射到同一特征空间内,将同类物体不断拉近,不同类物体不断远离,以此获得两个输入间的相似程度. 孪生网络广泛应用于视频目标跟踪[12,13]、人脸验证[14]、图像检索[15,16]等任务.

为了解决上述问题,本文采用孪生网络提取第一帧中给定目标对象与后续帧之间共同的外观特征,以此在后续帧中检测给定目标,代替主流视频目标分割方法的在线微调方式,从而有效地减少分割的时间,同时为了保留视频的时序信息,使用前一帧的目标分割掩码为当前帧提供位置指导.

2 基于孪生网络的快速视频目标分割

本文提出一个基于孪生网络的快速视频目标分割方法,采用深度 Xception 网络[22]作为框架,将参考帧和给定的目标分割掩码组成参考流,当前帧与前一帧的目标分割掩码组成目标流,共同作为网络的输入. 设计孪生编码器子网,将参考流和目标流映射到相同的特征空间,使得相同的目标具有相似的特征;全局特征提取子网在特征空间中匹配与给定目标相似的特征,定位目标对象;最后,解码器子网将目标特征还原,最终输出目标的分割掩码. 本文的视频目标分割方法主要包括三部分:参数共享的孪生编码器子网、基于扩张卷积的全局特征提取子网和解码器子网,其主体结构如图 1 所示.

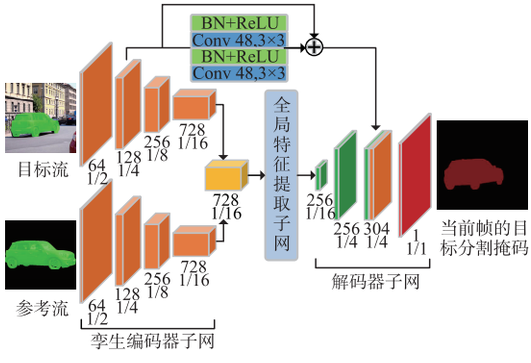

图1 基于孪生网络的快速视频目标分割网络结构图

2.1 参数共享的孪生编码器子网

孪生网络将两个输入映射到同一特征空间中,可以有效提取输入间的相似特征. 视频目标分割的目的是在后续帧中匹配第一帧给定的目标对象. 因此,本文设计参数共享的孪生编码器子网,代替在线微调方式,从而有效地减少目标分割的时间.

参数共享的孪生编码器子网的输入由两部分组成:参考流与目标流. 如图 2 所示,首先,使用可分离卷积残差块,将具有相似外观特征的向量不断拉近,不同外观特征的向量不断远离,以此逐步建立高维特征空间. 然后,将目标流和参考流的特征图进行特征融合,得到 728 维特征向量作为孪生编码器子网的输出.

在图 2 中,黄色的特征图即为孪生编码器子网的输出,本文使用 t-SNE(t-Distributed Stochastic Neighbor Embedding)[24]算法对其进行降维显示. 由结果图可以看出,在高维特征空间,参考流中的给定目标与目标流中的待分割目标具有相似的特征,而目标流中的背景特征与这两者分离.

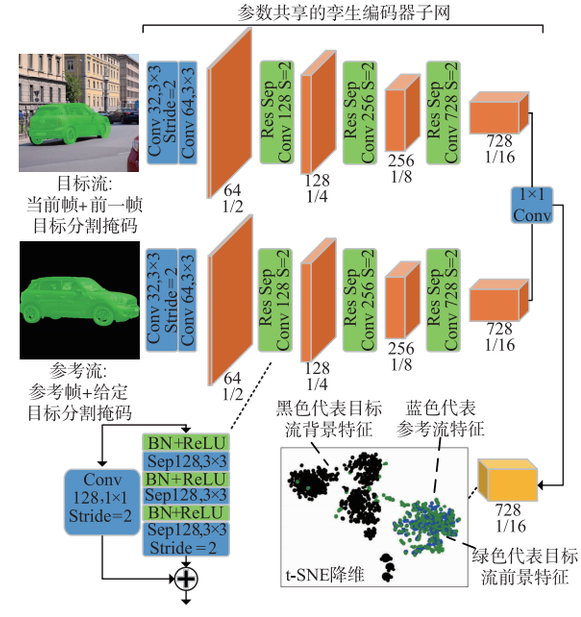

图2 参数共享的孪生编码器子网

2.2 基于扩张卷积的全局特征提取子网

基于扩张卷积的全局特征提取子网用于进一步提取参考流和目标流共同的全局特征,进而提取更抽象、更鲁棒的内在语义特征,并在特征空间中搜索与给定目标相似的特征,以定位待分割目标.

如图 3 所示,全局特征提取子网输入为孪生编码器子网输出. 首先,以更深的层次结构提取更丰富的全局特征;然后,利用扩张卷积,增加卷积操作的感受野,更好地表达目标内在的语义信息;最后,采用扩张空间金字塔池化操作[23](Atrous Spatial Pyramid Pooling, ASPP),使用不同扩张率的扩张卷积,得到具有不同感受野的特征图,以多尺度的方式将不同感受野的特征图

融合,生成前景特征.

同样使用 t-SNE 算法对输出的 256 维特征向量进行降维显示,从结果图可以看出,高阶特征已经将待分割目标的特征与背景特征很好地分离,表明全局特征提取子网能有效地在特征空间中搜索与给定目标相似的特征,以定位待分割目标对象.

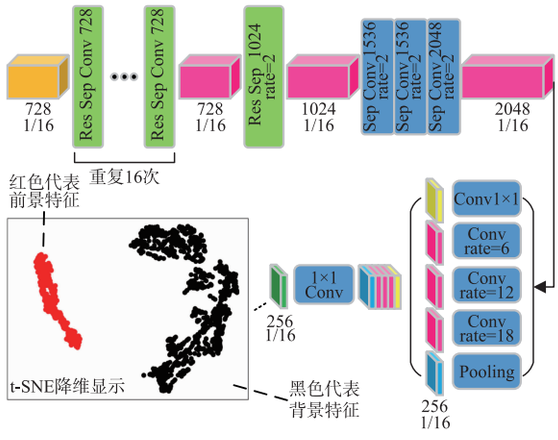

图3　基于扩张卷积的全局特征提取子网

2.3　解码器子网

解码器子网将特征空间中抽象的目标特征还原,并通过连接目标流的低阶特征,获得目标的边缘信息,最终输出目标的分割掩码,其结构图如图 4 所示. 因为高阶特征更能表达图像的内在语义信息,所以在融合低阶特征图时,使用 1×1 卷积降低低阶特征图的通道数,使得高阶特征图的通道数占较大比重.

由于在目标流的输入中加入了第四通道——前一帧的分割掩码,这会导致在低阶特征中,前一帧的分割掩码影响目标的边缘细节信息,使目标分割结果的细节与边缘并不理想. 为解决该问题,在本文的解码器子网中,利用残差结构保留有效信息,去除冗余信息的特点,将低阶特征经过一个残差块后输入到解码器子网中,能更好地过滤冗余低阶特征.

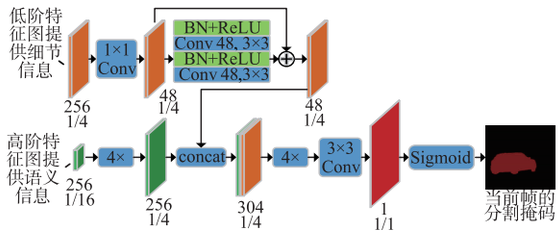

图4　解码器子网

3　实验结果与分析

为评价提出算法的有效性,本文在三个大型公开的基准数据集:DAVIS-2016[18]、DAVIS-2017[19]和 YouTube-VOS[20]进行实验. DAVIS-2016 用于单目标分割验证,DAVIS-2017 用于多目标分割验证. YouTube-VOS 数据集是 2018 年 9 月 ECCV(European Conference on Computer Vision)最新推出的公开基准数据集,包含 4000 个视频数据.

在 DAVIS 数据集进行测试时,本文首先采用 MS-COCO 数据集[21]进行预训练,然后在 DAVIS-2017 数据集中选取 60 个视频继续训练,余下的 30 个视频作为验证集. 在 Intel E5-2620 CPU、NVIDIA 1080Ti GPU 和 Win10 64 位操作系统下,基于 TensorFlow 开源框架,采用随机梯度下降算法(Stochastic Gradient Descent,SGD)训练模型,batch 大小为 4,momentum 为 0.9,第一阶段预训练学习率为 $1e^{-5}$,训练 50 万步,第二阶段训练学习率 $1e^{-6}$,训练 5 万步.

在 YouTube-VOS 数据集进行测试时,由于其数据量较大,本文直接在 YouTube-VOS 数据集上进行训练. 采用 Adam 算法(Adaptive Moment Estimation)训练模型,batch 大小为 4,设置初始学习率为 $1e^{-4}$,训练 100 个 epoch.

3.1　主流方法测评

本文使用 DAVIS 数据集[18,19]提供的基准代码计算预测的分割掩码与标注掩码之间的区域相似度 J(Region Similarity)、轮廓精确度 F(Contour Accuracy)以及对应的运行速度.

本文方法与当前几种较流行的方法进行对比实验,其中基于非深度学习的方法为:OFL[8],基于深度学习的方法 OSVOS[1],OnAVOS[3],OSMN[4],OSVOS-S[6],MSK[7],FAVOS[10].

3.1.1　DAVIS-2016

在 DAVIS-2016 数据集上,本文方法与对比方法的性能评估结果如表 1 所示.

表 1　不同视频目标分割方法在 DAVIS-2016 数据集上的性能评估结果(%)

方法	在线微调	J mean	F mean	速度(FPS)
OnAVOS	√	**86.1**	**84.9**	0.076
MSK	√	79.7	75.4	0.083
OSVOS	√	79.8	80.6	0.110
FAVOS		77.9	76.0	4.367
OSMN		74.0	69.0	7.142
OFL		68.0	63.4	0.016
本文方法		79.3	78.7	**7.693**

从表 1 可以看出:

(1)基于在线微调的方法可以取得较好的分割效果,但是在线微调非常耗时,且不能很好地满足场景的快速变化. OnAVOS[3]、OSVOS[1]、MSK[7]等方法采用了在线微调方式. 由表 1 可知这些方法分割速度都在 0.126FPS 以下. 本文采用孪生网络匹配参考流与目标

流中具有相似外观特征的目标对象,以代替在线微调,大大降低了分割时间.

(2)相比其它目标分割方法,OSMN[4]、FAVOS[10]分割速度最快.本文方法采用深度 Xception[22]网络作为模型的主体网络,具有更深的网络层次和较少的模型参数,因此本文方法的分割精度和速度都优于 OSMN[4]和 FAVOS[10].

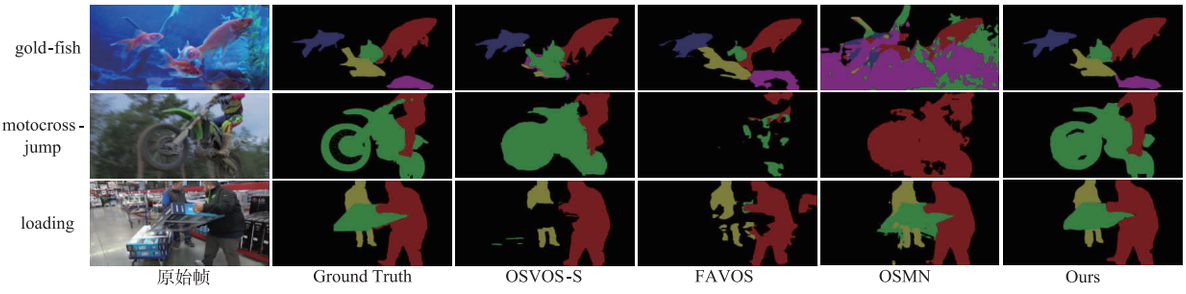

图5 本文方法与对比方法在DAVIS-2017数据集上的部分分割结果比较

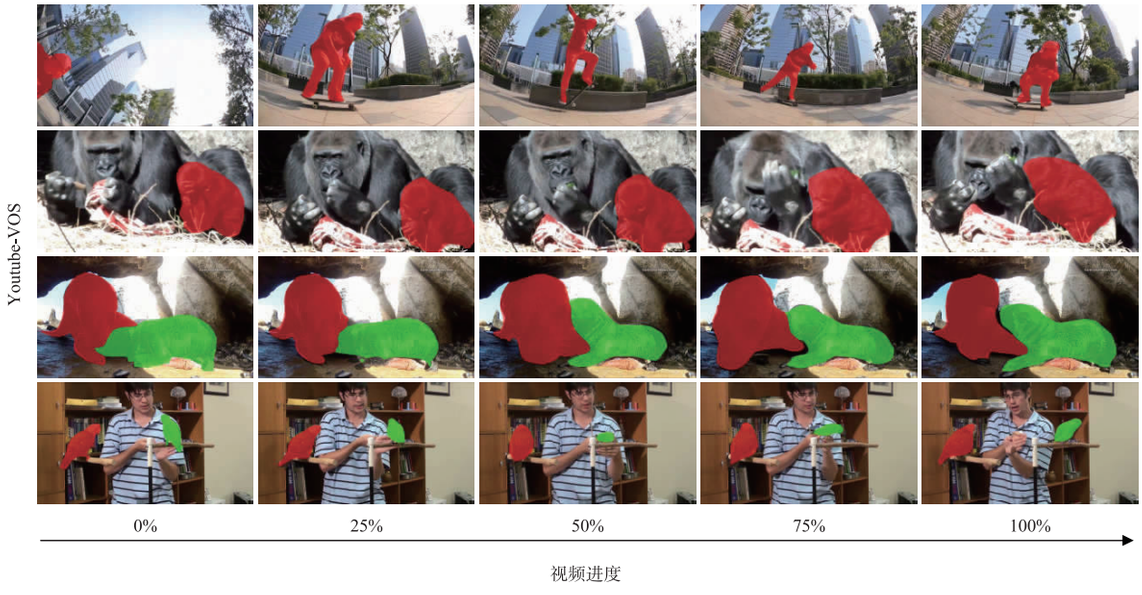

图6 本文方法在YouTube-VOS验证集上的部分分割结果

3.1.2 DAVIS-2017

本文方法与对比方法在 DAVIS-2017 数据集上的部分分割结果如图 5 所示,性能评估结果如表 2 所示.

表 2 不同视频目标分割方法在 DAVIS-2017 数据集上的性能评估结果(%)

方法	J mean	J recall	F mean	F recall	速度(FPS)
OSVOS-S	55.1	60.2	**62.1**	**71.3**	0.500
MSK	51.2	59.7	57.3	65.5	0.083
OSMN	52.5	60.9	57.1	66.1	7.142
FAVOS	54.6	61.1	61.8	70.5	4.367
本文方法	**56.5**	**64.2**	56.3	70.2	**7.693**

从图 5 和表 2 中可以看出:

(1)OSVOS-S[6]基于语义实例分割且单独处理每一帧,没有考虑视频帧间的时序信息.如 loading,由于语义分割模块没有学习到“纸箱”这一语义类别,所以出现分割丢失.本文方法基于孪生结构,可以有效地获取参考帧与目标帧之间的共同外观特征,实现目标分割,不受特定语义的限制.

(2)FAVOS[10]基于目标跟踪,依赖时序稳定性.如 motocross-jump,视频中目标运动剧烈,产生跟踪飘移,导致分割结果不连续.本文在目标流中加入前一帧的目标分割掩码,当目标外观变化剧烈或存在多个相似目标时,可获得更好的分割结果.

(3)OSMN[4]受网络层次的限制.如 gold-fish,目标与背景颜色相似,内容较为复杂.由于网络本身不能很好地描述内在的语义特征,从而导致将目标与背景混淆.本文以深度 Xception 网络[22]作为模型的主体网络,具有更深的网络层次和较少的模型参数,可以提取更加丰富且稳定的特征,因此在处理复杂场景时可以获得较好的分割结果.

3.1.3 YouTube-VOS

YouTube-VOS[20]的验证集包含 91 个目标类别.为

了评估算法对分割目标的泛化能力,验证集中有 65 个是训练集包含的目标类别,称为已知类别(seen),26 个是训练集不包含的目标类别,称为未知类别(unseen). G overall 代表四个评估指标的平均值. 本文方法与对比方法在 YouTube-VOS 验证集上的性能评估结果如表 3 所示,在验证集上部分的分割结果如图 6 所示.

表 3　不同视频目标分割方法在 YouTube-VOS 验证集上的性能评估结果(%)

方法	在线微调	G	J seen	J unseen	F seen	F unseen	FPS
OSVOS	√	58.8	59.8	54.2	60.5	60.7	0.1
OnAVOS	√	55.2	60.1	46.6	62.7	51.4	0.076
MSK	√	53.1	59.9	45.0	59.5	47.9	0.083
OSMN		51.2	60.0	40.6	60.1	44.0	7.142
本文方法		54.5	59.2	50.8	58.6	49.3	**7.693**

从表 3 中可以看出,本文将视频目标分割看作一种特征匹配问题,利用孪生结构,有效地提取第一帧给定目标与当前帧共同外观特征,以定位并分割目标对象,所以本文方法不需要预先学习目标类别,对已知类别对象和未知类别对象均可获得较高的分割精度. 从图 6 中可以看出,随着视频序列的播放,分割效果可以保持较高的鲁棒性.

3.2　本文算法分阶段的效果对比

为了验证本文算法各阶段的有效性,从网络的输入:参考流和目标流两个方面,在 DAVIS-2017 数据集上进行实验分析. 评估实验结果如表 4 和图 7 所示.

表 4　本文算法分阶段效果的定量分析(%)

	阶段	mIou	ΔmIOU
网络输入	- Reference	33.1	-23.4
	- Previous	47.4	-9.1
完整算法	-	56.5	-

图7　本文算法分阶段的分割效果对比

为了评估目标检测的有效性,将参考流的输入设置为空图,此网络结构命名为“-Reference”. 此时网络仅依赖视频帧间的时序稳定性,当目标被遮挡时,则会丢失分割目标. 其次,通过掩码的传播,错误的分割会不断地传播给后续帧,造成错误叠加.

为了评估掩码传播的有效性,将目标流中前一帧的目标分割掩码设置为空图,此网络命名为“-Previous”. 此时网络仅依赖参考流中给定的目标外观信息. 随着视频的播放,目标的外观逐步发生改变,使得网络难以适应较大的外观变化,造成分割精度下降.

4　结论

本文提出了一个基于孪生网络的快速视频目标分割方法,能快速、有效地处理单目标和多目标的视频目标分割问题. 针对传统基于深度学习的视频目标分割方法采用在线微调网络导致分割速度慢的问题,本文方法设计了参数共享的孪生编码器子网,将输入的参考流和目标流映射到同一特征空间,并提取给定目标与后续帧间的共同外观特征,以此检测给定目标. 同时,本文采用深度 Xception 网络为主体网络,利用其网络层次深、感受野大以及模型参数少等特点,获得良好的分割精度. 实验结果表明本文方法能有效地解决目标遮挡、大幅度外观变化等问题,能快速、有效地分割出视频中的目标.

参考文献

[1] Caelles S, Maninis K K, Pont-Tuset J, et al. One-shot video object segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolulu, Hawaii, USA: IEEE, 2017. 5320-5329.

[2] Shin Yoon J, Rameau F, Kim J, et al. Pixel-level matching for video object segmentation using convolutional neural networks[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Venice, Italy: IEEE, 2017. 2167-2176.

[3] Voigtlaender P, Leibe B. Online adaptation of convolutional neural networks for video object segmentation[A]. Proceedings of the British Machine Vision Conference[C]. London, UK: BMVA, 2017. 1942-1958.

[4] Yang L, Wang Y, Xiong X, et al. Efficient video object segmentation via network modulation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, USA: IEEE, 2018. 6499-6507.

[5] Chen Y, Pont-Tuset J, Montes A, et al. Blazingly fast video object segmentation with pixel-wise metric learning[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, USA: IEEE, 2018. 1189-1198.

[6] Maninis K K, Caelles S, Chen Y, et al. Video object segmentation without temporal information[J]. IEEE Transactions on PAMI, 2018, 41(6): 1515 – 1530.
[7] Perazzi F, Khoreva A, Benenson R, et al. Learning video object segmentation from static images[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolulu, Hawaii, USA: IEEE, 2017. 3491 – 3500.
[8] Tsai Y H, Yang M H, Black M J. Video segmentation via object flow[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Las Vegas, Nevada, USA: IEEE, 2016. 3899 – 3908.
[9] Jang W D, Kim C S. Online video object segmentation via convolutional trident network[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Honolulu, Hawaii, USA: IEEE, 2017. 7474 – 7483.
[10] Cheng J, Tsai Y H, Hung W C, et al. Fast and accurate online video object segmentation via tracking parts[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, USA: IEEE, 2018. 7415 – 7424.
[11] Wug Oh S, Lee J Y, Sunkavalli K, et al. Fast video object segmentation by reference-guided mask propagation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, USA: IEEE, 2018. 7376 – 7385.
[12] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[A]. European Conference on Computer Vision[C]. Amsterdam, Netherlands: Springer, 2016. 850 – 865.
[13] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking [A]. Proceedings of the IEEE Computer Vision and Pattern Recognition [C]. Honolulu, Hawaii, USA: IEEE, 2017. 5000 – 5008.
[14] Zhao J, Cheng Y, Xu Y, et al. Towards pose invariant face recognition in the wild[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, USA: IEEE, 2018. 2207 – 2216.
[15] Wang F, Kang L, Li Y. Sketch-based 3d shape retrieval using convolutional neural networks[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Boston, Massachusetts, USA: IEEE, 2015. 1875 – 1883.
[16] Yelamarthi S K, Reddy S K, Mishra A, et al. A zero-shot framework for sketch based image retrieval[A]. European Conference on Computer Vision[C]. Munich, Germany: Springer, 2018. 316 – 333.
[17] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[A]. Proceedings of the IEEE Computer Vision and Pattern Recognition [C]. Miami, FL, USA: IEEE, 2009. 248 – 255.
[18] Perazzi F, Pont-Tuset J, McWilliams B, et al. A benchmark dataset and evaluation methodology for video object segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Las Vegas, Nevada, USA: IEEE, 2016. 724 – 732.
[19] Pont-Tuset J, Perazzi F, Caelles S, et al. The 2017 davis challenge on video object segmentation [J]. arXiv preprint, 2017, arXiv: 1704. 00675.
[20] Xu N, Yang L, Fan Y, et al. Youtube-VOS: Sequence-to-sequence video object segmentation [A]. Proceedings of the European Conference on Computer Vision [C]. Munich, Germany: Springer, 2018. 585 – 601.
[21] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[A]. European Conference on Computer Vision [C]. Zurich, Switzerland: Springer, 2014. 740 – 755.
[22] Chollet F. Xception: Deep learning with depthwise separable convolutions[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolulu, Hawaii, USA: IEEE, 2017. 1800 – 1807.
[23] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs [J]. IEEE Transactions on PAMI, 2018, 40(4): 834 – 848.
[24] Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(10): 2579 – 2605.

作者简介

付利华 女, 1976 年 9 月出生, 四川安岳人. 2005 年在西北工业大学计算机学院获得工学博士学位. 现为北京工业大学信息学部副教授, 主要研究方向为智能信息处理、图像处理和计算机视觉.
E-mail: fulh@ bjut. edu. cn

赵宇(通信作者) 男, 1994 年 8 月出生, 河北唐山人. 2017 年在华东交通大学获得工学学士学位, 现为北京工业大学信息学部硕士研究生, 主要研究方向为图像处理和计算机视觉.
E-mail: zhaoyu2333@ emails. bjut. edu. cn