文章编号:1007-5321(2019)06-0105-06 **DOI**:10.13190/j.jbupt.2019-141

SA-Siam + +:基于双分支孪生网络的目标跟踪算法

田 朗, 黄平牧, 吕铁军
(可信分布式计算与服务教育部重点实验室(北京邮电大学), 北京 100876)

摘要: 为了解决 SiamFC 在目标快速移动、背景与前景相似、光照强烈等复杂场景下鲁棒性低的问题,提出了一种新的基于语义和外观双分支孪生网络的跟踪方法 SA-Siam + +,包括通过沙漏-通道注意力机制提取语义信息的语义分支和通过 SiamFC 提取外观信息的外观分支. 此外,将 AlexNet 网络更换为经过改进的 VGG-16 网络能显著增加特征提取能力. 在 OTB-2013、OTB-2015、UAV123 和 VOT2018 等目标跟踪标准数据集上进行了实验. 实验结果表明,所提算法获得的测试结果相比现有主流算法有较大提高,平均帧率为 49 帧/s, 满足实时性要求.

关 键 词: 孪生网络; 目标跟踪; 语义分支; 复杂场景; 通道注意力
中图分类号: TN929.53 **文献标志码**: A **OSID 码**:

SA-Siam + +: the Two-Branch Siamese Network-Based Object Tracking Algorithm

TIAN Lang, HUANG Ping-mu, LÜ Tie-jun
(Key Laboratory of Trustworthy Distributed Computing and Service (Beijing University of Posts and Telecommunications), Ministry of Education, Beijing 100876, China)

Abstract: To deal with the problem of low robustness of SiamFC in complex scenarios in which the object is moving fast, the background is similar to the foreground, and the illumination is strong, a new tracking method called SA-Siam + + was proposed based on two-branch siamese network, including semantic branch which is used to extract semantic information through the hourglass-channel attention mechanism and the appearance branch which is used to extract appearance information through SiamFC. In addition, replacing the AlexNet network with an improved VGG-16 network can significantly increase the feature extraction capabilities. Finally, experiments were carried out on OTB-2013, OTB-2015, UAV123 and VOT2018 which are standard object tracking datasets. It is shown show that the obtained with the proposed algorithm are greatly improved compared with the existing mainstream algorithms, and the average frame rate reaches 49 FPS, that can meet the real-time requirements.

Key words: siamese network; object tracking; semantic branch; complex scenario; channel attention

目标跟踪一直都是计算机视觉中具有挑战性的任务之一,是一个计算机视觉、模式识别等多学科交叉的研究领域. Luca Bertinetto 提出了一种基于相似度学习的全卷积孪生网络(SiamFC[1], fully convolutional siamese network). 基于跟踪的相关滤波器网络(CFNet[2], correlation filter based tracking)在浅

收稿日期: 2019-07-09
基金项目: 国家自然科学基金项目(61671072)
作者简介: 田 朗(1995—), 男, 硕士生.
通信作者: 吕铁军(1956—), 男, 教授, 博士生导师, E-mail: lvtiejun@tsinghua.org.cn.

层特征引入相关滤波器. 孪生实例跟踪网络(SINT[3], siamese instance search for tracking)在孪生网络基础上结合了光流信息实现了更好的性能. 由于SiamFC仅仅关注目标的颜色相似度,而且网络采用AlexNet[4],仅仅能够获取较为浅层的特征. 因此本文提出的SA-Siam + +在原有孪生网络的基础上加入了提取语义信息的语义分支. 并将AlexNet网络更换为经过改进的VGG-16 visual geometry group-16网络[5]. 最后在目标跟踪基准库(OTB, object tracking benchmark)[6]中的OTB-2013、OTB-2015、无人机目标跟踪数据集(UAV123, unmanned aerial vehicle 123)[7]和视频目标跟踪数据集(VOT2018, visual object tracking 2018)上进行测试,跟踪性能相比SiamFC等主流算法获得了较大的提高,在上述复杂场景下具有更强的鲁棒性.

1 全卷积孪生网络SiamFC

全卷积孪生网络结构如图1所示,用z和x分别代表输入的样本图像和搜索图像,经过共享参数的卷积网络φ来提取特征,最后将输出的特征图进行互相关操作来计算相似度,互相关计算公式为

$$f(z,x)=g(\varphi(z),\varphi(x)) \tag{1}$$

其中:$(\cdot)$代表互相关计算的函数,φ代表卷积网络所表示的特征提取器. 样本图像z和搜索图像x经过特征提取器后,进行互相关操作,会得到一张分数图,在分数图上分数最大的就是目标的所在位置. 这种通过匹配图像计算相似度方法的好处就是不需要线上学习,只需要使用离线训练过的预训练模型即可,减少了线上更新的时间,因此可以达到很好的实时性效果. 而采用的全卷积网络也使得输入图像的大小不受限制,并且在测试中可以获得各个网格的子窗口的相似度.

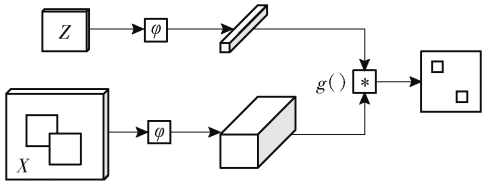


图1 SiamFC结构

SiamFC算法在图像输入网络前,对样本图像z和搜索图像x进行了预处理操作,分别将样本图像以及搜索图像裁剪为127×127和255×255尺寸的大小. 搜索图像x以样本图像z为中心,当一个窗格的延伸超过一个图像的尺寸时,就会用整个图像的平均像素值进行填充.

2 双分支孪生全卷积网络

2.1 问题描述

SiamFC是基于图像之间相似度进行判别跟踪的,因此当图像中的两种目标的颜色特别相似时,最终得到的相关分数也会特别大,SiamFC就可能会对跟踪目标做出错误判断. 此外,SiamFC采用的是AlexNet网络,很难提取到高层的外观特征. 这就导致了在目标快速移动,目标背景与前景相似以及光照强烈等复杂场景下鲁棒性低的问题. 因此SiamFC对于复杂场景是很敏感的. 笔者提出一种新的双分支孪生神经网络SA-Siam + +来解决这一问题. 新的网络包括了新提出的基于沙漏-通道注意力机制的语义分支和原有的由SiamFC组成的外观分支,这两个分支将在下文进行介绍.

2.2 网络框架

提出的一种新的网络框架使用了改进的VGG-16模型来增加特征提取器的深度,这样能提取到更加抽象的高层信息,此外引入一种新的沙漏-通道注意力机制来补充语义信息,这样网络不仅能学习到目标的外观信息还能学习到目标的语义信息,会使得目标跟踪在复杂场景下的鲁棒性更强. 具体的网络框架如图2所示. 整个网络分为外观分支和语义分支,外观分支负责捕获目标的外观信息,语义分支负责捕获目标的语义信息,并输出目标所属类别的权重. 在两个分支同时使用ImageNet[8]预先训练的经过改进的VGG-16网络作为用于提取特征的主干网络,使用VGG-16的原因在于在此络的卷积层之间没有填充操作,满足了前面提到的去填充要求. 增加填充层会使得在最后计算输出的相关得分图时

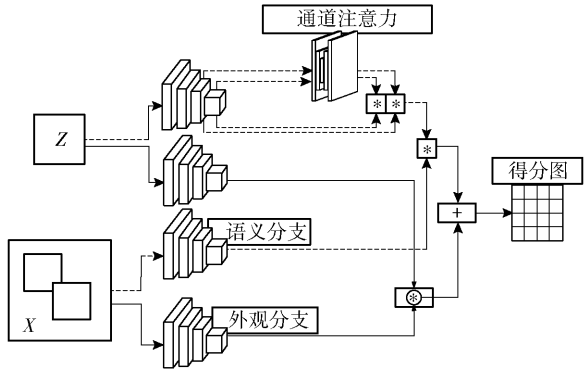


图2 SA-Siam + +网络框架

引入新的噪声，对于反馈损失值会造成干扰. 如果使用深度更大的残差网络 ResNet[9]，网络会在层与层之间的连接处增加一个捷径连接单元用以解决梯度爆炸的问题，而如果使用更宽的 GoogLeNet[10] 也会有 Inception 模块. 以上这 2 种更加先进的网络都会引入填充操作，所以本文舍弃了这两种网络，选择迁移性更好的 VGG-16 网络. Dual Cross-Entropy Loss[11] 提出一种新的损失函数，解决了损失值在训练过程中接近零时梯度消失的问题. 通道最大池化(CMP[12], channel max pooling)提出了一种通道最大池化层，减少了神经网络的参数数量. 在未来工作中可以考虑进来. 在调用 VGG-16 模型时，对其进行改进，只使用 VGG-16 模型的前 11 层卷积，将前 10 层的参数冻结，令其等于预训练模型对应的参数，仅训练第 11 层卷积，为了使输入沙漏-通道注意力模块的通道数一致，将第四卷积块的输出通道数量通过 1×1 卷积核修改为与第三卷积块一致. 最后两个分支各自的搜索图像与样本图像作互相关计算，得到相关输出图尺寸为 17×17，方便与 SiamFC 基线系统进行对比. 使用预训练模型可加快网络的训练，减少对跟踪速度的影响.

2.2.1　外观分支

在外观分支部分采用和 SiamFC 相似的结构. 将样本图像 z 和搜索图像 x 输入网络后，经过特征提取后，计算互相关得分，计算公式为

$$h_{\mathrm{a}}(z,x)=\mathrm{corr}(f_{\mathrm{a}}(z),f_{\mathrm{a}}(x)) \tag{2}$$

其中 $f_{\mathrm{a}}(\cdot)$ 代表特征提取器. 在计算互相关分数后，采用如下所示的损失函数进行反馈重传：

$$\arg\min\frac{1}{N}\sum_{i=1}^{N}\{L(h_{\mathrm{a}}(z_i,x_i;\theta_{\mathrm{a}}),Y_i\} \tag{3}$$

其中 θ_{a} 代表网络中需要优化的参数. N 代表输入的样本数量，z_i 代表第 i 个样本图像，x_i 代表第 i 个搜索图像，$Y_i\in\{-1,+1\}$ 代表给定的标签响应图，L 代表交叉熵损失函数. $h_{\mathrm{a}}(\cdot)$ 为式(2).

2.2.2　语义分支

在语义分支部分，仍然采用与 VGG-16 相似的网络结构. 不同的是，语义网络的输出是目标的所属类别，所以在 SiamFC 的基础上，增加了如图 3 所示的沙漏-通道注意力机制. 对于一幅图像，不同的通道对于不同的目标具有不同的敏感性，某些通道只对特定目标感兴趣. 为了减少计算量以及提高跟踪的准确性，引入通道注意力机制来输出各个通道对于不同目标的权重，再让输出的权重矩阵与 VGG 网络输出的各个通道相乘，即可把重要的特征增强，不重要的特征减弱，使提取的特征指向性更强. 在神经网络中，网络的任务由最后一层决定，当最后一层为 Sigmoid 函数时，网络输出为对应通道权重；当网络的最后一层为卷积层时，即输出对应的特征图. 同时，在语义分支部分，为了提取到不同层次的语义特征，借鉴了 DenseNet[13] 与 SENet[14] 的思想，将第四卷积块和第五卷积块的特征分别通过通道注意力模块，然后将其融合，这里也将损失函数定义成与式(3)相似的形式：

$$\arg\min\frac{1}{N}\sum_{i=1}^{N}\{L(h_{\mathrm{s}}(z_i,x_i;\theta_{\mathrm{s}}),Y_i)\} \tag{4}$$

其中：θ_{s} 是需要训练的参数，N 代表样本数量. z_i 代表第 i 个样本图像，x_i 代表第 i 个搜索图像，$Y_i\in\{-1,+1\}$ 代表给定的标签响应图，L 代表交叉熵损失函数. $h_{\mathrm{s}}(\cdot)$ 为式(2). 为了解决第三卷积块和第四卷积块输出通道的数量不匹配的问题，将第三卷积块的最后一层卷积输出通道数修改与第四卷积块最后一层一致. 最后输出为互相关得分图，其计算公式如下：

$$h_{\mathrm{s}}(z,x)=\mathrm{corr}(g(\xi f_{\mathrm{s}}(z)),g(f_{\mathrm{s}}(x))) \tag{5}$$

其中：$(\cdot)$ 代表融合操作，z 代表输入的样本图像，ξ 代表通道注意力模块输出的权重，x 代表输入的搜索图像. $\mathrm{corr}(\cdot)$ 代表互相关操作. 在语义分支，为了不影响算法的实时性，在特征提取部分采用了预训练的 VGG-16 模型，仅训练融合模块和通道注意力模块.

2.2.3　沙漏-通道注意力机制

注意力机制起源于人类对于图像感知的研究. 在一幅图像中，人们往往只会关注整张图像的一小部分，这就是注意力机制. 对于一副原始图像，只有(R,G,B)三个通道，经过卷积计算后，通道数会增加. 而神经网络也只会关注其感兴趣的通道，而忽略其他不重要的通道，这就是通道注意力机制.

传统的通道注意力机制采用类似于多层感知机的结构，即若干个全连接层彼此连接，在最后一层通过 Sigmoid 函数输出对应通道权重. 但是全连接网络所带来的计算量是巨大的，所以导致了实时性的降低. SA-Siam + + 创新性地采用了沙漏式的全连接网络，能够大量减少计算中的参数.

图 3 为笔者提出的沙漏-通道注意力机制的结构图，首先将通道 i 输出的特征图划分为 3×3 个网

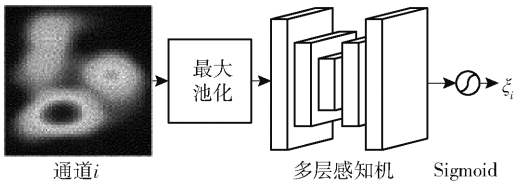


图3 沙漏-通道注意力机制

格,在每个小网格上采用最大池化,然后经过由多层感知机组成的沙漏式的全连接层,沙漏式的全连接层的神经元以9→5→4→5→9的顺序排列. 最后通过Sigmoid函数将输出限制到[0,1]的范围,并将这个值与主干网络的输出相乘来增强整个网络的特征表示能力.

采用全连接网络的原因是为了利用通道之间的相关性来表达目标与背景之间的语义信息. 此外,将神经元以沙漏的形状分布,这样不仅能够减少计算量,而且不会忽略掉一些重要的特征信息.

3 实验

网络的外观分支和语义分支都以VGG-16网络为主干. 卷积层后带有Relu[15]激活函数. 外观分支经过特征提取后直接计算互相关得分图,语义分支的特征提取器的第三卷积块与第四卷积块首先经过通道注意力模块,然后通过融合计算互相关得分图. 独立训练2个分支,只在测试阶段将2个分支联合测试.

3.1 模型训练

在离线训练阶段,使用视频目标检测数据集(ILSVRC-2015[16], imagenet large scale visual recognition challenge)训练,数据集包含了4 500个人工标注标记的视频序列,包含目标快速运动,背景与前景相似,光照强烈等各种复杂场景,是当前用于训练目标跟踪任务的常用数据集. 采用将外观分支与语义分支分开独立训练的方法,每次训练随机选取样本图像和搜索图像,迭代次数为30次. 每次训练6 400个样本对. 每次小批量训练样本数为8. 使用动量为0.9的随机梯度下降算法进行优化. 同时,为了使优化算法有更好的性能,学习率在每次迭代都缩减为上一次的0.5倍,在搜索图像上采用3个不同的缩放因子$\{q^s \mid q=1.0375,\ s=-1,0,1\}$.

所提出的模型使用Pytorch 0.4.0框架来搭建,且实验评估是在一台配置为E5-2628 CPU和显卡GTX1080Ti的主机上进行的,经测试,模型的平均帧率为49FPS.

3.2 模型测试

在模型测试期间,将2个分支联合测试,使用公式(6)来计算互相关得分图:

$$h(z,x)=\lambda h_a(z,x)+(1-\lambda)h_s(z,x) \tag{6}$$

λ为语义分支与外观分支的权重因子,需要在实验中不断测试. 经过测试,在λ取得不同的值时,模型会有不同的跟踪效果,具体数据见表1. Pre代表准确率,Suc代表成功率. 当λ等于0.1时,模型可以获得最佳效果. 从表1可以看出,增加的语义分支会使跟踪器表现更好,但是当λ等于0时,模型的效果并不好,说明外观分支的作用还是很大的,并不能单独地依靠语义分支进行跟踪. 这也验证了增加语义分支的正确性.

表1 不同λ值对比

λ	准确率	成功率
0	0.741	0.679
0.02	0.802	0.760
0.04	0.802	0.760
0.06	0.805	0.761
0.08	0.802	0.579
0.10	0.810	0.770
0.20	0.804	0.758
0.30	0.805	0.761
0.40	0.798	0.752
0.50	0.798	0.752
0.60	0.795	0.743

4 实验结果与分析

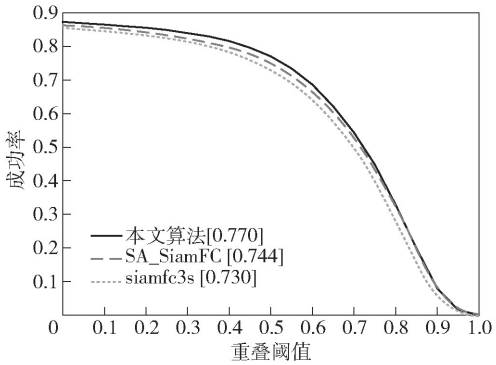
下面使用OTB-2013、OTB-2015、UAV123以及VOT2018数据集来进行测试,并且与SiamFC、SA-Siam[17]进行对比.

4.1 OTB

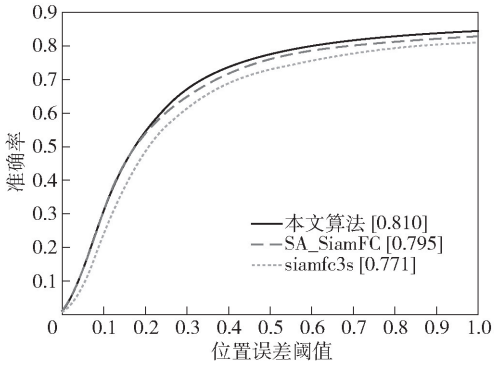
OTB-2015和OTB-2013是目标跟踪领域使用的基准库,分别包含了100个和50个人工标注的视频序列,包含了目标快速移动,背景与前景相似,光照强烈等各种复杂场景. 主要评判指标有成功率和准确率. 经过测试,OTB-2015对比结果如图4(a)所示,OTB-2013对比结果如图4(b)所示.

4.2 UAV123

UAV123数据集包含了来自无人机视角的123



(a) OTB-2015下各跟踪算法的成功率对比



(b) OTB-2013下各跟踪的算法准确率对比

图 4　OTB-2015、OTB-2013 下各跟踪算法性能对比

个视频帧序列,所有序列都使用矩形边框进行完全注释,包含了长时视频跟踪和短时视频跟踪两大类别. 实验结果对比如表 2 所示.

表 2　UAV123 数据集下各跟踪器对比

跟踪算法	本文算法	SA_SiamFC	siamfc3s
准确率	0. 719	0. 704	0. 700
成功率	0. 614	0. 606	0. 601

4. 3　VOT2018

VOT2018 数据集作为单目标跟踪领域的另一个标准数据集,包含了 60 个人工标注的彩色视频序列. 此外,对跟踪器的实时性提出更高要求. 表 3 所示为本文算法与其他跟踪器的对比实验结果.

表 4 所示为使用沙漏-通道注意力机制与使用传统通道注意力机制时的对比实验结果. 传统的通道注意力机制来自于 SENet[18],将多个相同尺寸的全连接层组合,最后通过 Sigmoid 函数输出对应通道权重.

此外,VGG-16 网络抓取目标高层信息的能力更强. 从图 5 各跟踪器的对比效果图可以看出,所提出的跟踪算法表现得更加鲁棒. 图中白色线代表所提算法.

表 3　VOT2018 实验对比

跟踪算法	期望平均覆盖率	准确率	鲁棒性
本文算法	0. 363	0. 584	0. 276
SA_SiamFC	0. 337	0. 566	0. 258
siamfc3	0. 187	0. 503	0. 585

表 4　沙漏-通道注意力与传统通道注意力对比

注意力类型	准确率	成功率	帧率
沙漏	0. 810	0. 770	49
传统	0. 795	0. 744	28

图 5　各跟踪器效果对比

5　结束语

在原有全卷积孪生网络的基础上提出了一种新的目标跟踪框架,创造性地引入沙漏-通道注意力机制以及更换特征提取网络来增强在目标快速移动,背景与前景相似,光照强烈等复杂场景下的模型鲁棒性. 同时又将用于提取特征的深度网络的主干部分更换为经过改进后的 VGG-16 网络,显著地增强了获取目标高层特征信息的能力. 经过在几个目标跟踪数据集上的实验验证,跟踪性能相比于 SiamFC 等主流算法有很大提高. 但是网络仍然还存在很多不足. 例如在前端预处理中只是粗糙地裁剪图像为指定大小;在目标发生形变、目标消失等场景下,跟踪效果还不太理想;在目标被遮挡时,跟踪有时会失败. 这都是未来需要进一步研究的工作重点.

参考文献:

[1] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 850-865.

[2] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE Press, 2017: 2805-2813.

[3] Tao Ran, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Press, 2016: 1420-1429.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv: 1409. 1556, 2014.

[6] Wu Yi, Lim J, Yang M H. Online object tracking: a benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE Press, 2013: 2411-2418.

[7] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[M]// Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 445-461.

[8] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE Press, 2009: 248-255.

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Press, 2016: 770-778.

[10] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE Press, 2015: 1-9.

[11] Li Xiaoxu, Yu Liyun, Chang Dongliang, et al. Dual cross-entropy loss for small-sample fine-grained vehicle classification[J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4204-4212.

[12] Ma Z, Chang D, Li X. Channel max pooling layer for fine-grained vehicle classification[J]. arXiv preprint arXiv: 1902. 11107, 2019.

[13] Huang Gao, Liu Zhuang, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE Press, 2017: 4700-4708.

[14] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 7132-7141.

[15] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]// Proceedings of the 27th International Conference on Machine Learning (ICML-10). Haifa, Israel: [s. n.], 2010: 807-814.

[16] Russakovsky O, Deng Jia, Su Hao, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[17] He Anfeng, Luo Chong, Tian Xinmei, et al. A twofold Siamese network for real-time object tracking[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 4834-4843.

[18] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 7132-7141.