文章编号：1005-5630(2021)06-0026-06 **DOI**: 10.3969/j.issn.1005-5630.2021.06.005

基于可变形卷积的单帧图像眼球定位追踪

王 鉴，张荣福
（1. 上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：针对目前眼球定位追踪算法存在的眼球定位精准度不高问题，以及为了改进眼球追踪算法的精准度并保证一定的图片处理速度，将可变形卷积网络应用于 YOLO 网络，对特征分布提取层面进行改进。利用可变形卷积的形变建模能力对卷积核中的各个采样点的位置增加一定的偏移变量，从而从原始单帧图像中提取更具有表征特征的信息，并与先进眼球定位追踪检测网络进行了实验对比。研究表明，可变形卷积 YOLO 网络的精准度可以达到 0.685，平均处理图片刷新率达 42 帧/s，优于原 YOLO 网络以及其他眼球定位追踪检测网络。

关键词：可变形卷积；YOLO 网络；眼球定位；形变建模
中图分类号：TP 751 **文献标志码**：A

Single-frame image eyeball tracking based on deformable convolution

WANG Jian，ZHANG Rongfu
（1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China）

Abstract: In order to improve the accuracy of the eye tracking algorithm and ensure a certain image processing speed, this paper proposes to combine the deformable convolution method to improve the feature distribution extraction level. The fixed-size sampling in the standard convolution makes it difficult for the learning network to adapt to the geometric deformation of the image. In order to solve this limitation, the deformation modeling ability of deformable convolution is used to add a certain offset variable to the position of each sampling point in the convolution kernel. So as to achieve the extraction of potential features, the single frame of the original image is described. According to the current research, the deformable convolution has made preliminary applications in the field of computer vision. After comparing with the advanced eyeball positioning tracking detection network experiment, the accuracy of the deformable convolutional YOLO network can reach 0.685, and the average image processing speed can reach 42 frames per second, which is better than the original YOLO network and the advanced eyeball and location tracking detection network.

收稿日期：2021-03-12
作者简介：王 鉴 (1996—)，男，硕士研究生，研究方向为图像处理、眼球追踪。E-mail：1005032910@qq.com
通信作者：张荣福 (1971—)，男，教授，研究方向为图像处理、工程光学。E-mail：zrf@usst.edu.cn

Keywords: deformable convolution; YOLO network; eyeball positioning; deformation modeling

引　言

眼球定位追踪的研究就是指研究跟踪视频中眼球的运动轨迹，该研究广受人们的关注。目前眼球定位追踪主要应用于电子设备、人机交互和虚拟现实中[1-2]。近年来，随着深度学习网络在计算机视觉取得重大突破，基于深度学习网络的眼球追踪技术也逐渐成为主要研究方向[3]。基于深度学习的眼球追踪主要分为单帧的目标图像检测任务以及基于视频帧的目标追踪任务[4]。本文主要对单帧的目标图像检测进行研究，以解决定位精确度不足的问题。

目前，单帧的目标图像检测已经取得了重大进展，但是仍然是一项具有挑战性的研究[5]，例如平衡检测算法的实时性和精准性。快速区域卷积神经网络（fast region convolutional neural network）[6]是利用提取相应候选区进行眼球的定位跟踪，该网络在眼球区域位置的定位精准度方面较为优异，但是通过数个卷积层计算处理，会使网络在整体分类速度上处于劣势，从而导致检测算法的实时性不佳。2016 年提出的 YOLO（you only look once）[7]检测算法，将单帧的目标图像检测任务转换为目标回归任务，通过对网格进行系统性的划分，将图像中快速检测出的目标类别通过边框回归的方式进行眼球追踪定位。然而 YOLO 算法网络的精准度不佳，普遍低于主流的神经网络算法[8]。在 YOLO 算法基础上，本文结合可变形卷积的相关算法对 YOLO网络进行改进，在保证实时性的同时，进一步提升整体网络的精准度。

本文对 YOLO 算法进行改进，利用可变形卷积的形变建模特性对网络的采样方式进行进一步的改进[9]。传统卷积（CNN）采用的是基于单一滑动窗口的区域采样策略，没有目标针对性，因此存在窗口冗余较大及时间复杂度较高的问题。传统卷积对未知大型形状变换目标的建模存在固有缺陷[10]，此缺陷源于卷积模块是基于单一几何结构设计。卷积模块对输入的特征图进行固定位置的采样，在池化层方面同样以固定的比例池化。该特性对算法整体性能有较大的影响，例如，在同一层级的卷积核中，所含激活单元的感受野相同，但是各个采样点的位置存在对应着不同尺度或者变形的物体情况。因此，对感受野大小或者尺度变化进行自适应建模是精确定位的重要条件。研究证明，标准卷积中的固定规格采样难以适应目标区域的几何形变[10]。为了解决这个问题，本文使用可变形卷积以及相应的可变形感兴趣区域池化，增强对目标多尺度形变的建模能力。这两种处理模块是基于相同平行网络学习偏移量（偏移），使得卷积核在输入的特征图中的采样点发生定量的偏移，使网络能较集中于目标区域或者感兴趣区域。经过多次实验，证明本文的方法在精准度方面与未改进 YOLO 网络相比提升了 4.7%，并可以实现网络的完整端到端训练。

1　可变形卷积网络

1.1　可变形卷积核

可变形卷积网络主要是处理稠密空间图像信息的算法网络，有着简单、高效以及可进行端到端网络学习的优势。

可变形卷积和标准卷积都是基于二维空间操作，且都是在相同的通道上进行。标准的卷积操作通常可以分为两部分：（1）在输入的特征图上使用标准固定网格进行采样；（2）对各个采样点的数值进行加权运算。

特征图的标准卷积[10]可表示为

$$y(P_0)=\sum_{P_n\in R} w(P_n)\cdot x(P_0+P_n) \tag{1}$$

式中：P_0 为特征图的原始位置；P_n 包含采样点中所列位置；R 为每个分块的索引编号；$w(P_n)$ 为权重；$x(P_0+P_n)$ 为原始图。由式（1）可知，标准卷积操作只是对输入的图像作相应的采样加权处理，缺少形变建模的能力。而可变形卷积引入

了偏移量的概念，通过在标准采样网格中增加一个偏移量进行形变。因此同样的特征图位置 P_0 可表示为

$$y(P_0)=\sum_{P_n\in R} w(P_n)\cdot x(P_0+P_n+\Delta P_n) \tag{2}$$

式中：ΔP_n 为偏移量；$x(P_0+P_n+\Delta P_n)$ 和 $y(P_0)$ 是原始图和经过卷积采集后的特征图的映射关系。通过设计网络对偏移量的学习，可以将固定的采样点位置改进为不规则的采样位置，如图 1 所示。

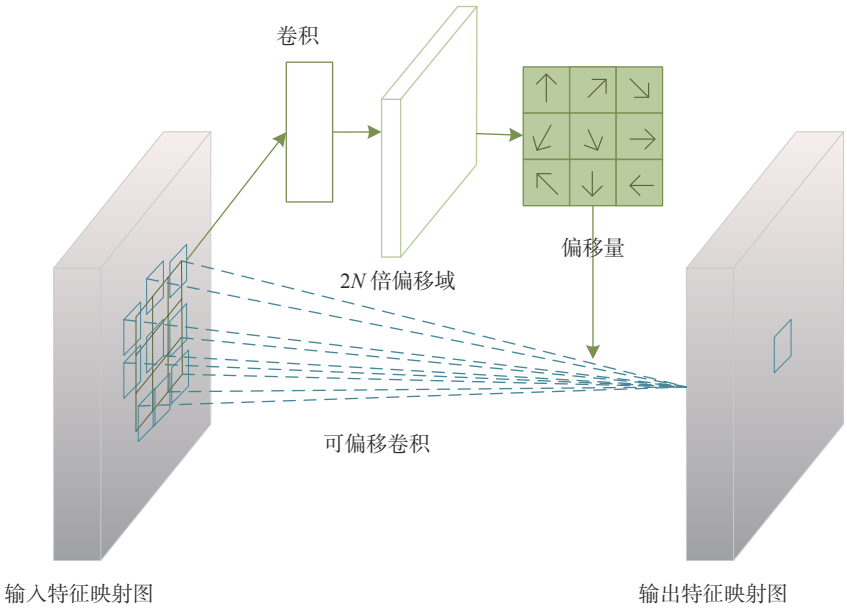

图 1　可变形卷积示意图

Fig. 1　Schematic diagram of deformable convolution

偏移量 ΔP_n 的获取是通过在相同的输入特征映射上使用标准卷积层计算获得，如图 1 所示。卷积核的尺寸与当前标准卷积层尺寸相同，例如图 1 中的卷积核尺寸为 3*3。偏移域的输出值与输入特征映射具有相同的空间尺寸，通道维数为 $2N$ 对应 N 维的 2D 偏移量。在网络训练阶段，可同时学习输出特征的标准卷积核和可变形卷积偏移量。为了学习偏移量可以反向传播误差，使用双线性运算计算反向传播。

1.2　可变形感兴趣区域池化

感兴趣区域池化模块是目标检测中常用的池化策略，是基于目标检测方法中的目标区域。在标准区域池化中，通常将任意输入大小的区域调整为固定尺寸大小的特征图。设给定的输入特征图为 x，待池化区域尺寸为 $w*h$，初始分块区域为 P_0，临近分块区域为 P，感兴趣区域池化将目标区域划分为 $k*k$ 个小区块并记为 b_{in}，同时经过处理后输出一个尺寸同样为 $k*k$ 的特征图。该特征图可表示为[9]

$$y(i,j)=\sum_{P\in b_{\text{in}}(i,j)} x(P_0+P)/n_{ij} \tag{3}$$

式中 n_{ij} 为 b_{in} 区块中的像素数。

通过以上标准池化层，可以类比得到可变形池化，即

$$y(i,j)=\sum_{P\in b_{\text{in}}(i,j)} x(P_0+P+\Delta P_{ij})/n_{ij} \tag{4}$$

相较于标准的感兴趣池化操作，同样对各个池化点增加相应的偏移量。首先，通过标准的感兴趣区域得到该输入对于位置的特征图。然后，通过该特征图加上全连接层计算生成每个对应区域的归一化偏移量 $\Delta\hat{P}_{ij}$。最后，根据感兴趣区域的高度和宽度尺寸进行元素对转换为 ΔP_{ij}。为了

使偏移量的输出与感兴趣区域大小保持不变，有必要对偏移量进行归一化。可变形池化的计算流程如图 2 所示。

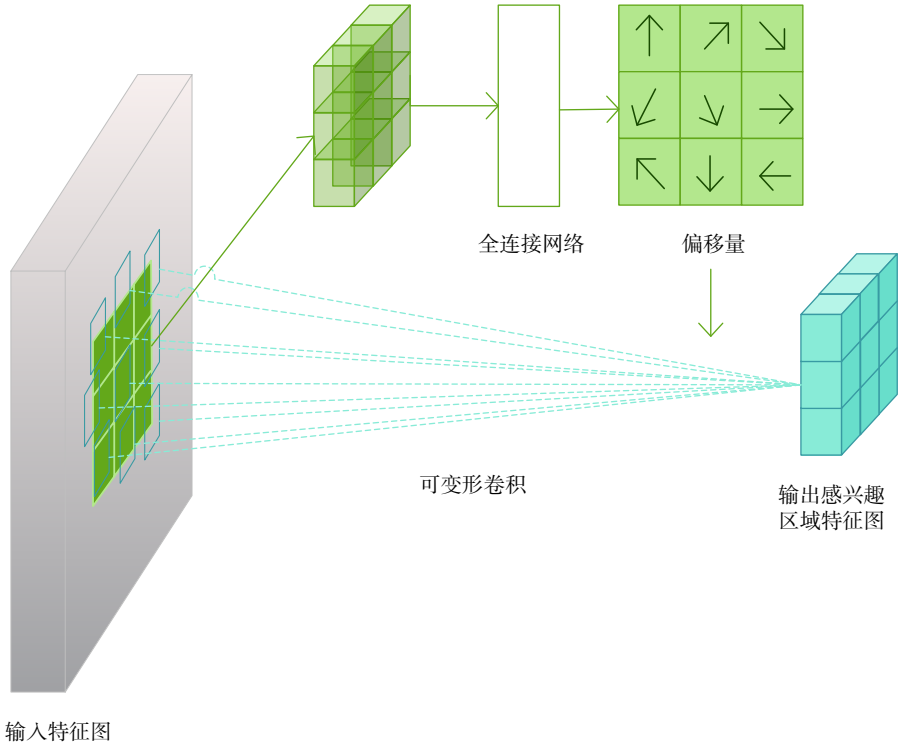

图 2　可变形池化示意图

Fig. 2　Schematic diagram of deformable pooling

1.3　可变形卷积网络在 YOLO 中的改进

YOLO 网络是近几年目标检测领域的创新算法，该算法舍弃通过复杂网络模型对目标物体进行分类和修改定位精度的主流目标检测思想，而是将一般目标检测问题转化成一个回归，能直接在待处理图像中的多个位置上回归分析出目标的边界框（bounding box）及其所属分类类别。对比其他目标检测算法，YOLO 算法的检测算法较快，标准版的 YOLO 算法在 Titan X 显卡上刷新率可以达到 45 帧/s，更快的 Fast -YOLO的刷新率更是达到 155 帧/s。并且可以很好地利用图像的整体信息，具有更好的泛化能力和迁移能力。但是 YOLO 网络对目标边界框会施加较高的空间限制，只能预测有限的目标类。因此，YOLO 网对物体检测的精度不是最优，较容易产生定位错误，尤其是在密集度高且物体偏小的情况，例如对人物面部眼球的定位。

因此，本文利用可变形卷积对 YOLO 网络的卷积方式进行改进，改变 YOLO 网络较高的空间限制，从而提高网络整体的分类精准度。图 3 为可变形卷积 YOLO 网络模型示意图。

2　实验结果与分析

2.1　YOLO 网络及其改进版对比

为检验本文的可变形卷积 YOLO 网络在目标检测精准度和处理速度上的变化，将本文网络与其他实时检测方法 Fast-YOLO 网络[7] 进行比较。实验使用 kaggle 中的 Fakefaces 数据作为训练集，该数据集包含 6 400 张人脸彩色图像，像素分辨率为 1 024*1 024。实验设备为 Tesla P100 显卡，Ubantu 操作系统。

YOLO 网络在原有的基础上已进行了多次改进，目前已经发展到 YOLO V3 版本。通过改变 YOLO 网络结构的复杂度，可以提高目标检测速度和目标检测精准度。虽然 YOLO V3 在 TitanX 上的处理速度可以达到 51 帧/s，最高精准度

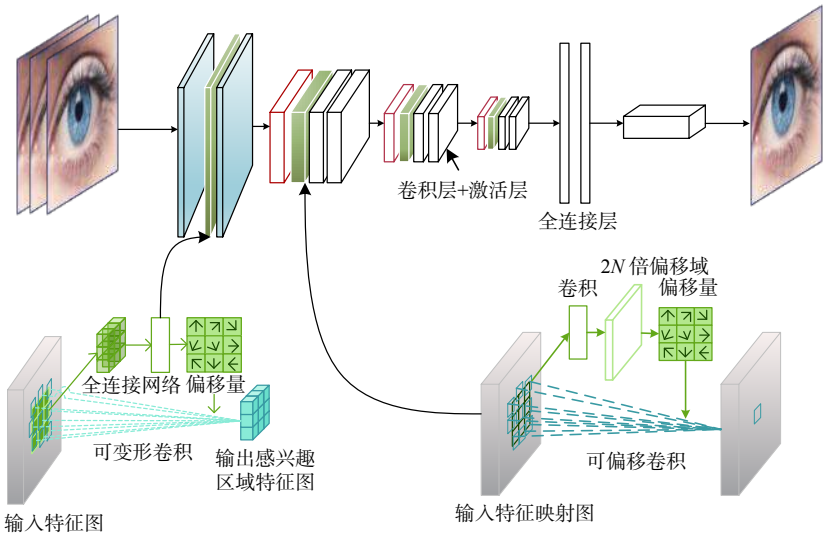

图 3　可变形卷积 YOLO 网络模型示意图

Fig. 3　Schematic diagram of deformable convolutional YOLO network model

达到 57.9%，但是仍有可提升的空间。可变形卷积 YOLO 网络与其他 YOLO 网络的对比如表 1 所示。

表 1　可变形卷积 YOLO 网络与其他 YOLO 网络对比表

Tab. 1　Comparison of deformable YOLO network and other YOLO networks

方法	网络框架	精准度	刷新率/(帧/s)
YOLO V1	ResNet-101	0.634	45
YOLO V3	ResNet-152	0.579	51
Fast-YOLO	Vgg16	0.527	155
可变形卷积YOLO	Vgg16	0.685	42

作为实时检测的早期网络，YOLO V1 网络的检测精准度高达 63.4%，同时仍保持较高的实时性，刷新率达到 45 帧/s。为全面对比 YOLO 网络的各个版本，本文使用 YOLO V3 和 Fast-YOLO 进行对比。Fast-YOLO 网络是目前最快的 YOLO 版本，刷新率达到 155 帧/s，但检测精准度明显低于 YOLO V1。而 YOLO V3 网络则更加均衡，在控制网络结构规模的情况下，处理速度有稳步的提升，但是精准度降低较大，与速度最高的 Fast-YOLO 网络相比也并没有较高精准度的提升，反而牺牲过多的处理速度。使用可变形卷积改进的 YOLO 网络在精准度方面有较大提升，可达到 0.685，而在图像处理速度方面几乎与最早版本的 YOLO V1 网络持平。综合以上情况，对 YOLO 网络进行可变形卷积的改进有助于目标检测网络的整体提升。

2.2　与其他先进检测网络对比

通过以上 YOLO 网络各个版本对比实验，可以得出，可变形卷积 YOLO 网络表现较佳。在此基础，本文通过与当前先进目标检测网络进行对比，进一步验证可变形卷积 YOLO 网络在目标检测领域中的表现。

在目标检测方面，本文选取可变形部件模型（deformable part model，DPM）[11] 和 Region-CNN（R-CNN）[12] 系列网络进行对比，实验结果如表 2 所示。

由表 2 对比可知：100 Hz DPM 模型的速度最高，刷新率达到 100 帧/s，但是相对的检测精准度也是最低的，只有 0.160；Fastest DPM 牺牲

表 2　可变形卷积 YOLO 网络与其他先进网络对比表

Tab. 2　Comparison of deformable YOLO network and other advanced networks

方法	网络框架	精准度	刷新率/(帧/s)
100 Hz DPM	SVM	0.160	100
30 Hz DPM	SVM	0.261	30
Fastest DPM	SVM	0.304	15
Fast R-CNN	Vgg-16	0.701	0.5
R-CNN Minus R	Vgg-32	0.535	6
YOLO Vgg-16	Vgg-16	0.654	18
可变形卷积 YOLO	Vgg-16	0.685	42

过多的检测处理速度，提高的精准度却相对有限；R-CNN 网络的检测精准度较高，尤其是 Fast R-CNN 的检测精准度最高，高达 0.701，但是处理速度过慢，无法用于实时检测。综上所示，使用可变形卷积改进的 YOLO 网络在检测速度和精准度上都取得较高的成绩，更加适用于眼球定位追踪任务中。

3　结　论

本文引用可变形卷积解决 YOLO 网络的空间限制问题，使 YOLO 网络在眼球定位追踪领域这类目标物体较密集且目标较小的检测中具有较好的精准度表现。通过对 YOLO 网络的改进，生成可变形卷积 YOLO 网络，该网络可以更好地实现眼球追踪定位的适用性，并在实时性和目标检测精准度上取得平衡。

实验结果表明：本文的可变形卷积 YOLO 网络可以用于快速重扫描眼球追踪检测，在较小地降低实时性的情况下可大幅提升目标检测的精准度，减少背景误报造成的误差，具有重要的应用价值。目前，该方法还有待进一步地扩大其应用范围，例如，在多帧视频中的应用，在保证视频处理的实时性的同时也能有较强的定位精准度，网络泛化能力的提升，等等。

参考文献：

[1] 天津电眼科技有限公司. 一种基于眼球追踪技术的智能眼镜控制方法: 中国, 201510489844.1[P]. 2017-05-24.

[2] 刘淼, 熊韬, 郭建咏, 等. 候车视觉搜索与公交车体外观特征的眼动跳视研究 [J].上海理工大学学报, 2016, 38(5): 472-478

[3] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436 – 444.

[4] 孙劲光, 孟凡宇. 基于深度神经网络的特征加权融合人脸识别方法 [J]. 计算机应用, 2016, 36(2): 437 – 443.

[5] PANG S C, DEL COZ J J, YU Z Z, et al. Combining deep learning and preference learning for object tracking[C]//Proceedings of the 23rd International Conference on Neural Information Processing. Kyoto, Japan: Springer, 2016.

[6] GIRSHICK R. Fast R-CNN[J]. Computer Science, 2015.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016.

[8] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv E-Prints， 2018, 108(4):625-633.

[9] ZHU X Z, HU H, LIN S, et al. Deformable ConvNets V2: more deformable, better results[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019.

[10] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[J]. Computer Vision and Pattern Recognition, 2017, 9(1):334-420.

[11] SADEGHI MA, FORSYTH D. 30 Hz object detection with DPM V5[C]//Proceedings of 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014.

[12] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J].Neural Information Processing Systems, 2015, 452(1):108-133.

（编辑：刘铁英）