一种基于生成对抗架构的目标检测增强算法

张 昀 黄 橙 施 健 张玉瑶 黄经纬 于舒娟 黄丽亚

(南京邮电大学电子与光学工程学院、柔性电子(未来技术)学院 南京 210000)

摘 要 目标检测网络的性能往往受制于特征提取网络的深度,而网络参数的大量增加只能带来检测系统性能的少量提升,同时需要引进许多额外的网络细节设计,这些都会导致训练难度的增加.本文提出了一种基于生成对抗训练的目标检测方法,它以减少特征分布的 EM 距离(Wasserstein 距离)为训练目标.具体来说,我们将检测网络从整个架构中提取出来,并对特征提取网络进行深入的对抗性训练.实验证明,本文提出的架构进一步提高了网络的特征提取能力,并且没有导致参数的增加.在 MS COCO 2017 数据集上,本文的架构将基于 ResNet101 的 CenterNet 网络性能从 36.1% *mAP* 提高到 37.2% *mAP*,将基于 Hourglass-104 的 *mAP* 从 42.2%提高到 43.0%.

关键词 计算机视觉;目标检测;生成对抗训练;特征提取;分类预测

中图法分类号 TP183 **DOI 号** 10.11897/SP.J.1016.2024.00647

An Enhanced Algorithm for Object Detection Based on Generative Adversarial Structure

ZHANG Yun HUANG Cheng SHI Jian ZHANG Yu-Yao HUANG Jing-Wei
YU Shu-Juan HUANG Li-Ya

(*College of Electronic and Optical Engineering & College of Flexible Electronics(Future Technology), Nanjing University of Posts and Telecommunications, Nanjing* 210000)

Abstract The performance of object detection networks is often limited by the depth of the feature extraction network. Increasing network parameters may yield limited improvements in the detection system's performance. Additional careful designs of network details are necessary, but they can significantly increase training difficulty. In this paper, generative adversarial networks are used as a method to further enhance the feature extraction capability of the network. In the normal architecture, by leveraging generative adversarial networks (GANs), it becomes possible to approximate the target distribution of a given task. This approach seeks a "near correct answer" by iteratively optimizing a non-convex game with continuous high-dimensional parameters. The generator and discriminator within the GAN framework strive to achieve a Nash equilibrium, resulting in an effective solution for the task at hand. In GANs, gradient descent is commonly employed to handle losses on both the generator and discriminator sides. In this paper, it is experimentally demonstrated that feature-highlighted images have similar feature distributions as unprocessed images, while the evolution of such feature distributions exhibits a continuous and learnable change as the degree of feature highlighting changes. Therefore, this paper introduces a new object detection method using generative adversarial training, which utilizes the ability of generative adversarial networks to fit feature distributions to enhance our object detection network.

收稿日期:2023-02-27;在线发布日期:2024-01-08.本课题得到国家自然科学基金(61977039)资助.**张 昀**,博士,副教授,主要研究领域为计算机视觉、软计算方法和通信信号处理.E-mail:y021001@njupt.edu.cn.**黄 橙**(通信作者),硕士,主要研究方向为计算机视觉、生成对抗网络.E-mail:1020020903@njupt.edu.cn.**施 健**,硕士,主要研究方向为计算机视觉.**张玉瑶**,硕士,主要研究方向为计算机视觉、生成对抗网络.**黄经纬**,硕士,主要研究方向为计算机视觉.**于舒娟**,硕士,教授,主要研究领域为计算机视觉、通信信号处理.**黄丽亚**,博士,教授,主要研究领域为计算机视觉、脑机接口技术.

Our approach focuses on minimizing the EM distance (Wasserstein distance) of the feature distribution, using features acquired with technically processed images as a benchmark to create a target distribution for the generative adversarial network. The features obtained from the original images will be considered as false information in generative adversarial, and the process of adversarial training will continuously improve the feature extraction capability of the network to obtain more realistic features, thus improving the target detection capability. Simultaneously, due to enhanced image features, the training of GAN (Generative Adversarial Network) yields a feature distribution that exceeds that of the original dataset, which allows additional gains to be obtained more easily than the usual training methods. A new loss function is also added during adversarial training to ensure steady improvement of the detector by constantly checking the object detection performance of the network. A comparative experiment conducted with the original CenterNet network on MS COCO (Microsoft Common Objects in COntext) 2017 reveals that the generative adversarial training method significantly improves the average precision for most of the examined backbone networks, while ensuring that there is no increase in the inference complexity of the network. Among the four backbone networks employed in the experiments, the mean improvement in network AP (Average Precision) values ranged from 0.3 to 0.9, demonstrating their success with minimal training efforts. Moreover, none of the four backbone networks experienced an increase in network parameters during inference. Experimental results indicate that the proposed architecture effectively enhances the network's feature extraction capability without compromising speed during inference.

Keywords computer vision; object detection; generate adversarial training; feature extraction; classification prediction

1 引 言

目标检测是计算机视觉的基本问题之一,它的衍生任务比如行人检测面部识别视频监控检测越来越受到人们的重视.近年来,随着机器性能的增强,应用领域和深度的增加,人们开始广泛采用复杂而深入的卷积层,以增强神经网络对图像细节的提取能力.然而,这种做法带来了严重的计算资源消耗问题,并且在训练过程中难以有效控制特征提取网络.针对这些问题,人们提出了一系列技术以便在提高网络性能的同时,避免网络参数的巨幅增加:基于锚点的模型[1-2]和无锚点模型[3-4]被用于应对目标建议算法进行多尺度检测时耗时长的问题;贪婪非极大值抑制和可学习的非极大值抑制[5-6]在不修改网络结构的情况下提高了目标检测的召回率;残差连接网络[7]、特征金字塔[8-10]提高了不同层级网络的利用率,减少了特征在传递过程中的损失,避免了网络参数的无效增加.生成对抗网络(GAN)通过学习从潜在空间到真实分布的映射,为目标检测任务提供了一种全新的方法.这种能力使得 GAN 能够以独特的方式处理目标检测问题,尤其是对小物体的检测[11-12]、低分辨率图片的检测[13-14].

对于生成对抗网络(Generative Adversarial Networks,GAN),一旦它获得一个任务的目标分布,就有可能学习到"接近正确答案"的匹配方式——即追求生成器和鉴别器间的一种具有连续高维参数的非凸博弈的纳什均衡.因此,相较于传统的目标检测从标注数据中学习拟合方式的训练方法,GAN 在目标检测中可以被用于生成特征,为在质量降低的图像上进行的目标检测提供鲁棒性.但由于 GAN 通常使用梯度下降法处理生成器和鉴别器的损失,这会导致它会停留在人为设计的损失函数的局部低值,而不是非凸博弈的纳什平衡点.为了解决这一问题,Salimans 等人[15]基于这种思路提出了用于生成图片,寻求纳什平衡的特征匹配方法和用于目标检测的半监督训练方法.它在 k 个目标类别中增加了一个生成图片的 $k+1$ 类用于训练目标检测器,检测器必须对真实样本进行分类,同时与生成器进行对抗.

但是一方面，鉴别器应以 1/2 的概率决定生成的数据是否为假，另一方面，检测器应合理地将生成的数据分类到真实的 k 类中，生成对抗网络所实现的识别假样本和预测标签两个任务的平衡难以作为网络训练的最优点.因此，检测和生成这两个任务很难在一个框架下应用.对于这一问题，在 Li 等人[16]的研究中，他们提出了一种名为“Triple-GAN”的三方博弈框架，其中引入了一个检测器作为第三方参与生成器和鉴别器之间的博弈.尽管该方法对对抗训练过程进行了仔细的设计(训练鉴别器时将生成器和检测器作为同一方，训练生成器和检测器时分别剔除后者和前者)，但是鉴别器对于检测器对真实图片的分类结果的否定，以及该方法尽力实现的生成器和检测器的强耦合会在一定程度上阻碍检测器的性能.因此，将两方博弈拓展到三方并不是一个有利于稳定训练的方法.我们认为，想利用生成对抗架构的强大生成能力增强特征质量，与其让检测器参与生成对抗的训练，不如使其游离在两方博弈之外，跟随两方博弈者一同成长.

在这个思路下，我们从一个训练完备的目标检测网络中提取出特征提取网络和检测网络.组合鉴别器和特征提取网络进行联合训练，以判断特征的品质，优化特征提取网络的性能.在完成训练后，我们重新回复原本的目标检测器，并对检测模块进行调整.

从直观上来看，生成图片和真实图片之间具有足够的相似性，可以通过降低它们之间的距离来拟合特征分布.相关研究[15]利用此策略成功构建了一种能接近达到纳什均衡点的人工图像生成的模型.以此为基础，我们进一步设想：即便是由于特征抽取网络的性能差别而导致的特征提取结果的不一致，也能产生相似的分布结果(详细的实验证据参阅第 3.1 节).根据这个推测，我们把性能较低的特征抽取网络同低质的图像关联在一起：使用特征提取网络提取人工设计的具有丰富表征信息的图片，我们创建了一个“真实数据”分布.同样，我们也用那些缺乏清晰表达信息的图像与同样的特征抽取网络相连，形成了另一个“噪声”特征分布.这使得我们有可能以一种可控的方式得到更好的特征分布，并将它运用到生成对抗网络中去.

为了验证我们的想法，我们基于无锚点模型 CenterNet 网络设计了一种生成对抗式的目标检测方法，将表征信息不同的两类图片作为输入进行 GAN 的训练.

对此，本文工作的主要贡献展示如下：

(1)训练完备的目标检测网络存在难以继续提升的问题，对此我们提出了一种基于生成对抗架构的目标检测方法，该方法利用生成对抗架构的潜空间拟和能力增强特征质量，使得特征提取网络以较小的训练代价获得额外的提升.

(2)我们通过特殊设计的图片处理技术提升生成器的特征提取能力上限，使其不受制于数据集的图片质量.我们通过不断给予检测器新的训练梯度提升其检测性能.

(3)我们的架构在两个常用的目标检测数据集 PASCAL VOC 2007 和 MS COCO 2017 上进行了广泛的实验.实验结果表明，多种检测器在很短的时间内获得了可观的提升，在小目标和少参数的骨干网络上的提升更为明显.另外，我们的方法几乎“即插即用”，检测时的算法复杂度不会增加.

2　相关工作

2.1　目标检测任务

目标检测任务是本文工作的主要目标，它的特征对于研究人员获得更好的分类精度和使用更少的计算资源至关重要.为了提升目标检测过程中所获得的图片特征质量，人们进行了一系列新的尝试.

如网络中一些短步长、局部连接、参数共享的卷积核的感受野较小，能够捕捉边缘和局部信息，具有平移相等性.而网络的深层结构中的卷积核感受野更大，能够学习高级语义信息，具有平移不变性，利于目标分类[17].针对这两个特性人们提出了结合深层特征和浅层特征的特征融合技术.根据融合与预测的顺序，一种名为早融合的技术将多层网络的特征输出直接连接[18-19]或是映射到复数域[20-21]再进行预测.相对的，晚融合技术将不同层的特征分别预测，结合多个检测结果以实现高级语义信息和局部信息的兼容[2,10].

在一项研究中，研究者认为特征金字塔结构的关键在于多尺度特征融合和分治思想，因此该研究使用单入单出的 SiSo 编码器代替 FPN 的 MiMo，使用扩张编码器融合多尺度感受野，在保证了精度的情况下，大大减少了计算量[22].

与这些方法相比，我们的方法不是为特定的网络构建而修改的，而是使用一种类似“后处理”的技术.它以优化一个训练完备的网为目标，这使得它可以与其它改进特征质量的技术共存.

2.2 无锚点检测器

无锚点检测器不预设锚点框或候选区域，而是直接从特征图中推断出目标的位置和类别. 在本文的工作中，无锚点检测器由于全卷积和无锚点的特点能够被更方便地拆分，应用生成对抗架构就不需要复杂的训练微调和超参数设置，因此本文将无锚点检测器中的 CenterNet 架构作为主要的检测方式.

CenterNet 通过生成关键点热图获取目标位置和类别[23]. 该架构通过对从一张图片中抽取出三个特征矩阵：类别和目标位置的热力图、特征图的中心点偏置，目标的包围框尺寸. CenterNet 在对所有特征逐个扫描的过程中实现目标识别，这有效防止了由于锚点设定不当所引发的可能遗漏情况. 此外，CenterNet 并未受到非极大值抑制(NMS)的影响，原因在于其产生的候选项数量始终小于全部像素的一半，这一现象得益于高置信度的点能对其周边的其他点施加抑制效应. 所以，CenterNet 不依赖非极大值抑制去降低过多的候选项，进而缓解了计算压力. 在我们的实验中，我们进行了基于 CenterNet 网络的深度设计，因为它的无锚结构和基于卷积网络的分类器非常适合我们的框架.

2.3 生成对抗网络

本文的主要改进方法基于生成对抗架构(Generative Adversarial Networks，GAN)，它是一种用于无监督或半监督学习的强大网络，用于构建两个任务之间的最大最小值博弈，网络的目标函数如下所示[24]：

$$L_{\mathrm{GAN}}=V(D,G)=E_{x\sim p_{\mathrm{data}}(x)}[\log(D(x))]+E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

式(1)表明，生成对抗式网络有两个损失函数：$\log(D(x))$和$\log(1-D(G(z)))$是鉴别器的优化函数，$\log(1-D(G(z)))$是生成器的优化函数. 当任意一方训练时，另一方的网络参数都会被固定住. GAN 的训练极其不稳定，存在诸多问题：纳什平衡问题[25]、模式崩溃[26-27]、梯度消失[28]、训练不平衡[29]等. 近年来生成对抗网络在小目标检测和人脸检测任务上都有了长远的发展. 在人脸检测方面，一系列基于 GAN 的架构，如 InterFaceGAN、Image-2StyleGAN++、PGGAN、StyleGAN3 等都得到了应用[30-33].

在目标检测领域，GAN-DO 通过最大化两种特征之间的相似性来实现目标[14]：GAN 生成器产生的低质量图像特征和基线模型产出的原始高质量图像特征但是，基于低质量图像的训练很难超过作为基线的模型和高质量图像. 另外，在这个架构中使用的低品质图像可以被视为在网络的起始部分增加了一个无法计算梯度的平均池层. 这种结构的引入必然会导致一些边缘信息和细节的损失，进而降低了生成器的性能. 相对而言，我们使用具有表征信息丰富的图像代替利用更好的特征提取网络，通过一些图像处理技术，可以使生成器网络不断获得更好的训练目标. Posilović 等人[34]在材料缺陷检测技术中应用了生成对抗网络. 利用生成器提升训练样本以改善识别效果的方法，是“Object Detection+GAN”这一组合模式在目标检测领域的典型运用. 而我们只基于基础数据集进行少量的人工设计，在增加训练量的情况下持续为特征提取网络的训练提供合适的梯度. Sultana 等人[35]移动对象切割领域利用了生成对抗网络. 他们构建了三种结构：生成器、检测器和特征提取网络. 对于移动对象切割任务，生成器负责处理影像中的动态目标，其他两部分辅助生成器来减小数据样本图像与特质空间的差距. 虽然模型有一层模块用于专门获取梯度信息，然而这个模块的表现依然受到原始图片特征分布的限制. 而我们的方法通过人工标注的数据集人为地创造了更优秀的梯度，生成器能够持续获得正确的学习目标.

3 基于生成对抗架构的检测方法

如第 1 节所述，为了使用生成对抗训练目标检测网络，不应直接将两方博弈扩展为三方博弈. 这可能会导致其中任何两方都难以达到纳什均衡，避免网络内部协变量转移问题的难度也会急剧增加. 因此，我们将已经完成训练的 CenterNet 按照特征提取网络和分类网络(用于产生关键点、关键点精度损失、包围盒尺寸)分离，构建一个全新的鉴别器与特征提取网络构成对抗结构. 在每个阶段的对抗训练完成后，我们将 CenterNet 的两个部件重新组合，固定住特征提取网络并对检测网络进行微调.

如图 1 所示，我们使用图像处理技术对从数据集 X 抽取出的样本进行特征增强，$\hat{x}=f(x)$，$x\in X$. 具体而言，高斯模糊被用于实现背景分离，锐化处理用以突出细节特征. 如图 2 所示，上面列出的三幅图像经过高斯模糊处理和锐化处理，以增强信息的表现力，下面的三幅图像是原始图像. 如果图像中包含多个目标，我们将多个目标看作整体，全部作为前景，只对目标之外的区域进行模糊. 此外，我们也再一次从数据集抽取出等量的图片，而这些图片并未

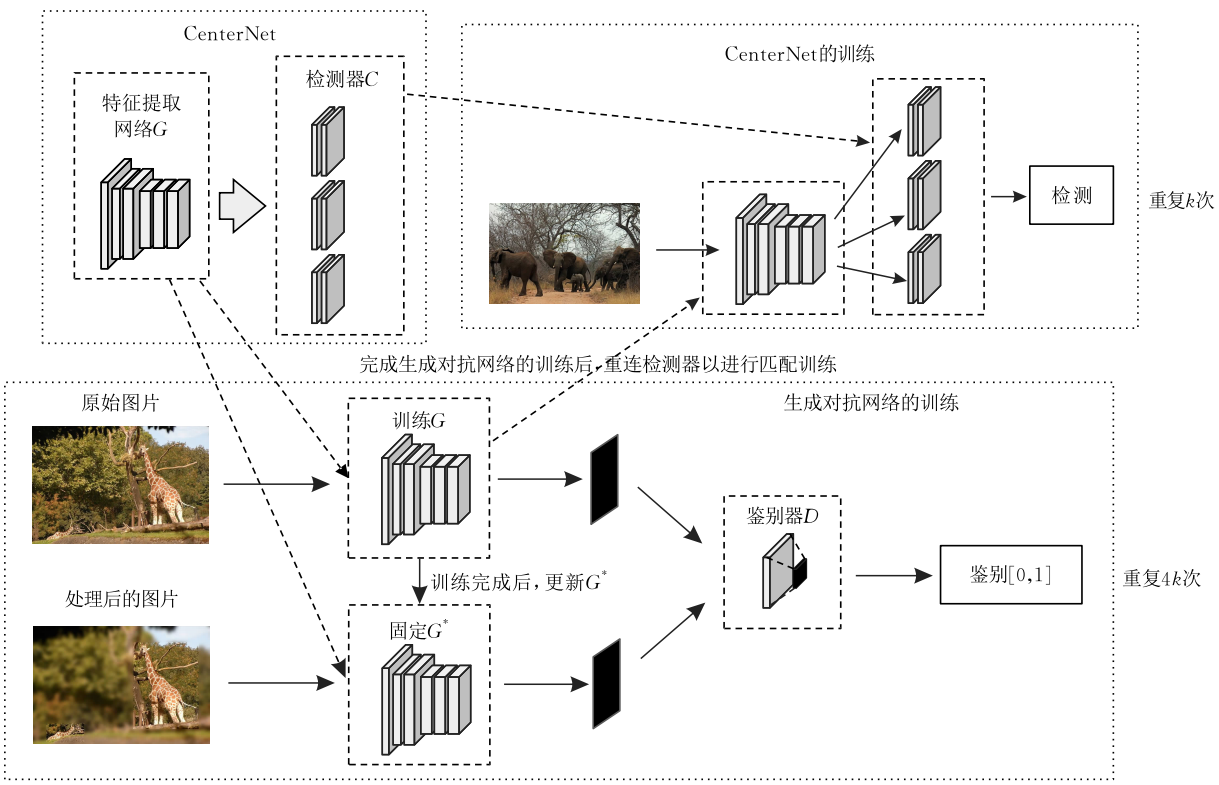


图 1　基于生成对抗训练的 CenterNet 目标检测网络的训练方式

图 2　不同信息表征的图片对比(为了增强表现力，示例图片使用的高斯模糊半径为 6 pixel，具体实验中不会使用这么大的参数)

接受过任何处理. 所以，我们创建出了两组数据：一批未经处理的图像 X 被视为“噪声”，已处理过的照片批次 $\hat{X}$，其提取出的特征作为“真实数据”. 我们认为，相较于未经处理的图像，经过处理的图像能够提取出更多丰富的特征信息，这种特征信息上的差异足以被鉴别器所检测到.

另外，只有在生成式对抗训练结束后，两个特征提取网络的权重才进行共享，提取“真实分布”的特征提取网络权重仅提供更优的目标分布，不参与训练. 因为对于生成对抗训练来说，目标分布作为真实数据分布应当是确定的.

整个目标检测架构由三种网络组成：来自于目标检测架构的检测网络 C、特征提取网络 G 和手动构建的鉴别器 D. 在训练过程中，固定参数的特征提取网络 G^* 处理特征增强的数据批次 $\hat{X}$，生成的高质量特征 $\hat{Y}$ 作为目标分布. 而原始图像 X 通过参与训练的特征提取网络 G 产生普通特征 Y. 鉴别器 D 的目标是分辨两种特征的来源，从而为特征提取网络

提供梯度.

总之,我们使用对抗训练调整性能难以获得提升的目标检测网络,从高质量的图像特征抽取出优秀的目标分布,训练增强网络的特征提取效率.第一节中假设的实验论证、损失函数、网络结构和训练细节将在下面的小节中给出.

3.1 假设的实验论证

在本节中,我们设定了一个前提:三种特征有相似的分布形式,并且后两者以相同的方向远离前者.它们分别是:(1)优秀的特性(保留了尽可能多的信息);(2)性能较弱的特征提取网络从正常图片获得的特征;(3)性能正常的特征提取网络从表征信息较少的图片获取的特征.

首先,我们需要考虑在不同性能下特征提取网络的数据分布.为了更方便地测量两个分布之间的距离,我们将图片特征的每一像素 u 增加 a 以保证增加后的特征值 u 为正值:

$$u \leftarrow u + a \tag{2}$$

接着,我们采用 box-cox 变换[36]处理特征分布:

$$u^{(\lambda)} = \begin{cases} \dfrac{u^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(u), & \lambda = 0 \end{cases} \tag{3}$$

通常来说,对于任意 λ,box-cox 转换后的值 $u^{(\lambda)}$ 是原始值 u 在允许范围内的单调函数,这一变换在 $u>0$ 时成立,获得转换值的关键在于求取使 u 到 $u^{(\lambda)}$ 的变换确立的 λ 参数族.

我们使用 $u^{(\lambda)} = V\beta + e$ 描述转换后的值 $u^{(\lambda)}$ 与回归自变量 V 之间的线性关系.在这个过程中,$u^{(\lambda)}$ 表示已知矩阵,即回归因变量,β 是与变换后的观测数据相关的未知参数向量,e 作为误差服从标准正态分布.对自变量和因变量进行回归的过程就是求取能够使这一线性关系成立的 (β,σ^2) 和 λ 最优解.在本实验中,$u^{(\lambda)}$ 表示特征图中的像素值分布,V 为对应分布的概率密度值.总之,转换后的观测值 $u^{(\lambda)}$ 满足完全正态理论假设,即 $u^{(\lambda)} \sim N(V\beta,\sigma^2\boldsymbol{I})$.由于 λ 的值依赖于 (β,σ^2),因此我们通过两次极大似然估计方法分别求取 (β,σ^2) 和 λ:

$$L(\beta,\sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \exp\left\{-\frac{(u^{(\lambda)}-V\beta)'(u^{(\lambda)}-V\beta)}{2\sigma^2}\right\} J(\lambda;u) \tag{4}$$

$$J(\lambda;u) = \prod_{i=1}^{n}\left|\frac{\mathrm{d}u_i^{(\lambda)}}{\mathrm{d}u_i}\right| = \prod_{i=1}^{n} u_i^{\lambda-1} \tag{5}$$

其中 $J(\lambda;u)$ 变换的 Jacobi 行列式.

获得 λ 的最优值后,转换得到的特征分布将符合完全正态理论的预期.

Wasserstein 距离同时满足衡量距离的三条特性:正定性、对称性和三角不等式,并且可以衡量任意两个分布的距离.但是,受限于计算方式,我们通常只能计算一、二阶的分布距离或是两个高斯分布的 Wasserstein 距离,这也是我们对数据使用 box-cox 变换的原因.两个高斯分布的 Wasserstein 距离计算公式如下:

$$W_2^2(X,Y) = \|m_1 - m_2\|^2 + \mathrm{tr}\left[\sum\nolimits_1 + \sum\nolimits_2 - 2\left(\sum\nolimits_1^{1/2}\sum\nolimits_2\sum\nolimits_1^{1/2}\right)^{1/2}\right] \tag{6}$$

$$t_X^Y(x) = m_2 + \sum_1^{-1/2}\left[\sum_1^{1/2}\sum_2\sum_1^{1/2}\right]^{1/2}\sum_1^{-1/2}(x-m) \tag{7}$$

其中 $t_X^Y(x)$ 表示从分布 X 到分布 Y 的最优映射.

我们使用训练周期为 30、60、90、120 时的特征提取网络抽取特征,使用 Wasserstein 距离衡量经过 box-cox 变换后的特征分布间的距离.实验过程中,我们将训练周期为 150 时的网络获得的分布作为近似的最优分布.表 1 展示了随着训练的进行,特征分布与最优分布之间的差异,而图 3 则呈现了在各种训练周期中特征分布的情况.

表 1 不同训练周期下特征分布与最优分布的差异

训练周期	λ	$\lvert\mu_i-\mu_{150}\rvert$	$\sigma^2/10^{-2}$	W(距离最优分布)$/10^{-2}$
30	0.267	0.061	3.063	21.017
60	0.253	0.008	1.638	9.432
90	0.193	0.007	1.742	9.316
120	0.185	0.000	2.161	0.480
150	0.182	—	2.190	—

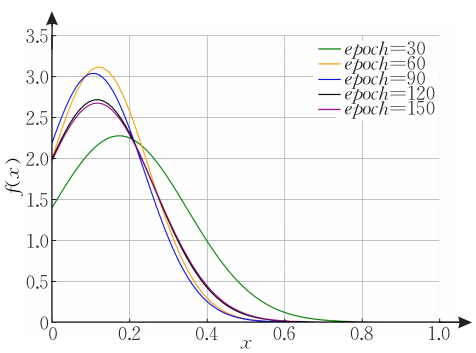


图 3 不同训练周期下的特征分布图

观察表 1 数据可知,除第 30 个训练周期的特征提取表现相对欠佳外,其他大部分训练阶段中,特征分布的方差 σ^2 及转换系数 λ 都相当接近,这证明了最优分布作为目标分布的可行性.随着训练进程推进,各样本点的平均值 μ_i 与其理想分布间的差异逐

步减小，而其Wasserstein 距离的变化与$|\mu_i-\mu_{150}|$保持一致. 上述结果表明，随着网络特征提取能力的增强，特征分布会趋向并接近平衡状态下的一种最佳分布模式.

我们在验证图像表示方式差异如何影响特征抽取时，选用了训练期为 30 的特征提取模型，并针对各种不同的程度的特征突出进行测试. 关于背景部分，我们使用了高斯模糊，其半径设置为 1.0、1.5、2.0 和 2.5 像素. 对于图像细节，我们使用拉普拉斯算子获取锐化效果，拉普拉斯算子的权重参数 α 用于获取不同程度的锐化. α 取值为 0.2、0.4、0.6、0.8. 实验结果如表 2 所示.

表 2　不同信息表征的图片特征分布与最优分布的差异

高斯模糊＋锐化	λ	$\lvert\mu_i-\mu_{150}\rvert$	$\sigma^2/10^{-2}$	W(距离最优分布)$/10^{-2}$
1.0 pixel＋0.2	0.216	0.018	2.346	3.263
1.5 pixel＋0.4	0.204	0.010	2.514	1.552
2.0 pixel＋0.6	0.195	0.008	2.231	0.990
2.5 pixel＋0.8	0.199	0.007	2.061	0.943

观察表 2 可知，当图像中的信息表达得更为清晰时，其特性分布和最佳分布的差异程度将会降低，即随着图像中所含的信息量的提升，其特性分布也逐步趋近于真实的分布状态. 另外，如果用常规的特征提取网络来处理那些信息量较低的图形并获取特征，那么它们的特性分布则更可能远离理想的分布模式.

将上面两个实验进行对照可以发现，随着特征提取网络的不断优化，网络特征提取能力越来越强，特征分布与最优分布的 Wasserstein 距离不断缩小，提升图片的信息表征能力而不对特征提取网络进行优化可以产生同样的效果. 总之，如果赋予生成对抗网络拟合更优特征分布的训练目标，那么特征提取网络将获得更优秀的训练梯度，使用具有更强信息表达能力的图片输出的特征分布作为训练目标是可行的.

3.2　损失函数

网络的损失包括两部分，生成对抗架构损失和检测模型损失.

(1) 生成对抗训练损失 $L_{\text{GAN-total}}$. 生成对抗损失主要来自特征提取网络和鉴别器的对抗训练. 为避免 GAN 模式失效及训练不均衡对检测网络产生负面影响，同时保证特征提取网络的稳定性，我们在训练损失中加入了约束 $L_{\text{CenterNet}}$，用于防止目标检测结构的退化：

$$L_{\text{GAN-total}}=L_{\text{GAN}}+\alpha L_{\text{CenterNet}} \tag{8}$$

其中 L_{GAN}是对抗损失，$L_{\text{CenterNet}}$表示目标检测模型的限制，α 是 CenterNet 网络损失的加权系数.

参考图 2，训练过程中的两组训练样本：表征信息能力更强的批次 $\hat{X}$ 和原始数据集合 X 分别送入固定参数和参与训练的两个特征提取网络中，同时以 $\hat{y}$ 和 y 的形式给出它们与 $\hat{x}$ 和 x 之间的映射关系：

$$\hat{y}=G^*(\hat{x}),\ P(a\leqslant\hat{Y}\leqslant b)=\int_{\hat{y}} f_{\hat{Y}}(y)\mathrm{d}y$$

$$y=G(x),\quad P(a\leqslant Y\leqslant b)=\int_{y} f_Y(y)\mathrm{d}y$$

$$\text{其中 } \hat{y}\in[a,b],\ y\in[a,b] \tag{9}$$

其中 $G^*(\cdot)$表示未被用于训练的提取网络，该网络在每个训练周期结束后会与其对应的 $G(\cdot)$进行权重共享. $f_{\hat{Y}}(\cdot)$和 $f_Y(\cdot)$表示权重 $\hat{y}$ 和 y 的概率密度函数，同时也是生成对抗训练所追求的目标分布及需要适应的训练样本.

对于鉴别器，使用 Wasserstein 距离作为评估数据分布距离的标准，并在 GAN 的训练过程当中采用 WGAN-GP 以实施梯度惩罚来增强对抗训练的稳健性[37]. 加入了梯度惩罚后，生成对抗网络的训练损失则表现为以下形式：

$$L_{\text{GAN}}=E_{\hat{y}\sim f_{\hat{Y}}}[D(\hat{y})]-E_{x\sim f_X}[D(G(x))]+\lambda_{\text{penalty}}E_{x\sim f_{\text{penalty}}}[(\|\nabla_x D(x)\|-1)^2] \tag{10}$$

$E_{\hat{y}\sim f_{\hat{Y}}}[D(\hat{y})]-E_{x\sim f_X}[D(G(x))]$为生成对抗的标准训练损失，$\lambda_{\text{penalty}}E_{x\sim f_{\text{penalty}}}[(\|\nabla_x D(x)\|-1)^2]$表示输入梯度 L2 范数的双边约束，其中 f_{penalty}是通过从 f_Y和 $f_{\hat{Y}}$分布之间的区域上随机抽样得到的分布，λ_{penalty}代表了双边约束的强度，实验中设置$\lambda_{\text{penalty}}=10$.

(2) 检测器的训练损失 L_C：检测器损失 L_C由 centernet 网络的输出矩阵(热力图、目标尺寸和中心点偏置)构建而成[23]：

$$L_C=L_k+\lambda_{\text{size}}L_{\text{size}}+\lambda_{\text{off}}L_{\text{off}} \tag{11}$$

其中 L_k，L_{size}，L_{off}分别是关键点的逻辑回归损失、边界盒大小的 L1 损失和关键点偏移的 L1 损失，λ_{size}和 λ_{off}是对应损失的权重系数. 根据我们的实验结果，CenterNet 原文中使用的这两个权重因子的值为 0.1 和 1，但现在我们已经将其调整至更低的值：0.05 和 0.8. 因为特征提取网络的训练对目标识别中心的捕获能力下降，导致其他两个模块对损失的影响增加. 在 MS COCO 2017 数据集上帮助探测器的性能提前 2(ResNet-101)到 4(ResNet-18)个训练批次达到最优值.

在每一次 GAN 网络训练结束之后，会进行特征提取网络 $G(\cdot)$与检测网络 $C(\cdot)$的联合训练. 在

训练过程中，只有检测器进行小幅度的调整. 三个损失的定义如下：

$$L_k=\frac{-1}{N}\sum_{xyc}\begin{cases}[1-C_k(G^*(x))]^{\alpha}\cdot & \\ \log[C_k(G^*(x))], & Y_{xyc}=1\\ (1-Y_{xyc})^{\beta}[C_k(G^*(x))]^{\alpha}\cdot & \\ \log[1-C_k(G^*(x))], & \text{其他}\end{cases} \tag{12}$$

$$L_{\text{off}}=\frac{1}{N}\sum_{p}\left|C_{\text{off}}(G^*(x))-\left(\frac{p}{R}-\tilde{p}\right)\right| \tag{13}$$

$$L_{\text{size}}=\frac{1}{N}\sum_{k=1}^{N}|C_{\text{size}}(G^*(x))-s_k| \tag{14}$$

其中 $G^*(\cdot)$ 代表不参与更新的特征提取网络. $C_k(\cdot)$，$C_{\text{off}}(\cdot)$ 和 $C_{\text{size}}(\cdot)$ 分别表示检测网络的热力图、中心点偏置和目标尺寸的 L1 约束. Y_{xyz} 表示热力图中对应坐标的标签值，c 为目标类别，R 为缩放系数，p 和 $\tilde{p}$ 分别表示缩放前后的中心点坐标，s_k 为原始尺寸.

3.3 基于对抗收敛性的改进

生成对抗网络的加入往往会带来训练收敛性的不稳定，尤其是我们的网络中加入了与生成器性能差异较大的鉴别器. 尽管鉴别器与生成器经过了足够时间共同训练，但网络的复杂度增加可能会引入额外的优化困难.

受到 CBNetv2[38] 的启发，我们使用生成器训练阶段中固定参数的特征提取网络 G^* 对训练目标 G 进行辅助监督.

如图 4 所示，一个新的网络分支被添加到了生成器的训练过程中. 为了保证两个特征提取网络的耦合作用，固定参数的特征提取网络 G^* 的输出通过级联的方式加入到训练目标 G 中. 进行训练的网络 G 中的输入来自于上层网络的输出和特征提取网络 G^* 对应层级之后的固定层数（图 4 中表示为 n）的网络输出：

$$x_l=F_l\left(x_{l-1}+\lambda_l\sum_{i=l}^{l+n}u_i^l(x_i^*)\right) \tag{15}$$

其中 x_l 代表特征提取网络 G 第 l 层网络输出，n 为辅助网络层数，$u_i^l(\cdot)$ 为第 i 层输出到第 l 层输出的上采样，$F(\cdot)$ 为卷积变换，λ_l 是权重参数，在实验中被设置为 0.05.

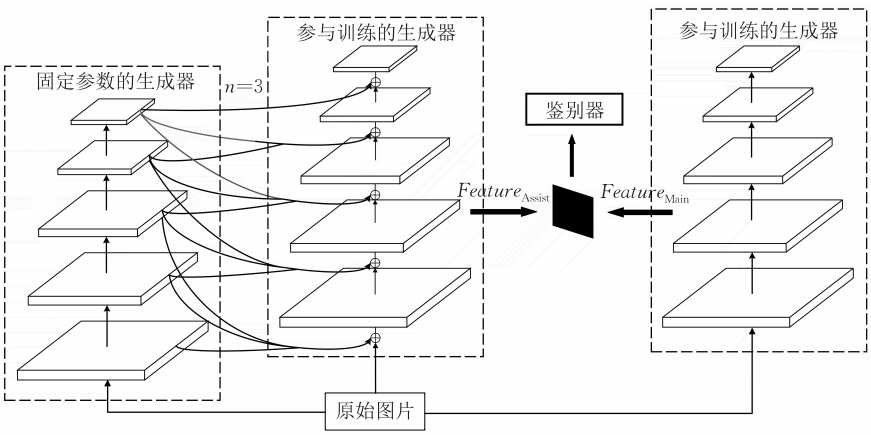


图 4 改进后的生成器训练过程

这一结构所产生的损失被用作辅助监督，提供额外的正则化. 另外，这一网络分支不会参与鉴别器和分类器的训练过程.

3.4 骨干网络

为证明本节的方法能有效地提高已经过深度训练的目标检测模型的性能，我们将采用 Resnet-18、Resnet-101[39]、DLA-34[40] 和 Hourglass-104[23,41] 四种骨干网络测试 CenterNet 的性能及对抗训练的表现.

3.5 训练细节

在生成对抗网络的训练过程中，生成器和鉴别器的性能差异是导致网络不平衡的重要原因. 而在我们的对抗网络中，生成器来自于完成了 150 个训练周期的 CenterNet 的特征提取网络，而鉴别器是全新建立的. 为了避免 GAN 的不平衡训练，我们在所有训练开始前先固定住生成器，基于 COCO 数据集[42] 对鉴别器进行 50 个周期的训练，然后再进入生成对抗网络的交替训练.

完成多次生成对抗训练后，我们对未参加对抗训练的检测器进行优化. 因为特征提取网络的参数数量远远超过了检测网络，所以它的优化会更加谨慎，需要更多的迭代次数. 通过实际实验验证，生成对抗学习和检测学习的交替训练次数为 4∶1. 具体的训练流程可以参考算法 1.

算法 1.　基于生成对抗训练的目标检测算法.

输入：原始图片 x，高质量图片 $\hat{x}$，对抗训练学习率 lr_{gan}，鉴别器学习率 lr_c，训练批次 m，CenterNet 损失权重 α，鉴别器预训练批次 $n_{\text{early-D}}$，对抗训练鉴别器训练批次 $n_{\text{training-D}}$，对抗网络迭代次数 $n_{\text{training-GAN}}$，训练周期 n

模型：特征提取网络 $G(\cdot)$，不参与训练的特征提取网络 $\hat{G}(\cdot)$，鉴别器 $D(\cdot)$，检测器 $C(\cdot)$，生成对抗和辅助监督网络 $G'(\cdot)$，星号“*”表示在训练中被固定参数的网络类型

1. FOR $i=1,2,\cdots,n_{\text{early-D}}$ DO
2. 　鉴别器 $D(\cdot)$ 预训练
3. END FOR
4. FOR $i=1,2,\cdots,n$ DO
5. 　FOR $j=1,2,\cdots,n_{\text{training-GAN}}$ DO
6. 　　FOR $k=1,2,\cdots,n_{\text{training-D}}$ DO
7. 　　$D_\theta \leftarrow \nabla_\theta\left[\frac{1}{m}\sum_{k=1}^{m}D_\theta(\hat{G}(\hat{x}))-\frac{1}{m}\sum_{k=1}^{m}D_\theta(G_\theta^*(x))\right]$.
8. 　　$\theta_d \leftarrow \theta_d + lr_{\text{gan}}\cdot\text{Adam}(\theta, D_\theta)$.
9. 　　END FOR
10. 　　$G_\theta \leftarrow -\nabla_\theta\left[\frac{1}{m}\sum_{k=1}^{m}D_\theta^*(G_\theta'(x))-\alpha\cdot\frac{1}{m}\sum_{k=1}^{m}C_\theta^*(G_\theta(x))\right]$.
11. 　　$\theta_g \leftarrow \theta_g - lr_{\text{gan}}\cdot\text{Adam}(\theta, G_\theta)$.
12. 　　$\hat{G}_\theta \leftarrow G_\theta$
13. 　END FOR
14. 　$C_\theta \leftarrow \nabla_\theta\frac{1}{m}\sum_{k=1}^{m}C_\theta(G_\theta^*(x))$.
15. 　$\theta_c \leftarrow \theta_c - \text{lr}_c\cdot\text{Adam}(\theta, C_\theta)$.
16. END FOR

输出：特征提取网络 $G(\cdot)$

4　实　验

我们在 4 个骨干网络上对 CenterNet 网络进行了对照实验：ResNet-18，ResNet-101，DLA-34，Hourglass-104. 对照实验是在完整训练了 150 个周期的原生 CenterNet 和训练完成后进行了对抗训练的 CenterNet 之间进行的，实验主要基于 MS COCO 数据集进行. 与 CenterNet 的设置[23] 相同，我们采用随机翻转、随机缩放（在 0.6 到 1.3 之间）、裁剪和颜色抖动进行数据增强，使用 Adam[43] 优化总体目标，初始学习率为 4×10^{-4}. 使用不同 *IOU* 值下的平均准确度：AP，AP_{50} 和 AP_{75} 作为主要的性能指标，同时使用 AP_S，AP_L 和 AP_M 来测试目标检测网络对不同尺寸的目标的检测能力. 表 3、表 4 的实验结果均来自 4 次重复实验的结果，以保证数据的可靠性和准确性. PASCAL VOC 2007 数据集用作附加实验，其类别和图片的数量相对较少.

表 3　本文提出的架构基于不同骨干网络的检测精度和时间

骨干网络	AP			AP_{50}			AP_{75}			AP_{mask}			$AP_{\text{box}}^{\text{test-dev}}$	FPS			*epoch*
	N.A.	F	MS	N.A.	F	MS	N.A.	F	MS	N.A.	F	MS	N.A.	N.A.	F	MS	
ResNet-18	29.6	31.8	34.1	45.6	47.9	52.9	30.6	33.1	36.6	26.1	26.9	28.5	30.0	**144**	**72.0**	**12.0**	**13**
ResNet-101	35.6	36.9	40.0	53.5	55.2	58.4	37.6	39.4	42.6	31.7	33.2	35.9	36.4	46	27.0	4.0	23
Hourglass-104	**41.1**	**43.2**	**45.7**	**59.8**	**60.9**	**63.8**	**43.9**	**45.4**	**49.8**	**35.8**	**37.7**	**40.1**	**41.6**	14	7.7	1.6	30
DLA-34	37.9	40.1	42.8	55.3	57.2	59.8	41.6	43.3	45.9	33.2	36.6	37.8	38.5	52	27.0	4.0	26

表 4　本文架构与原始网络的检测精度对比

	骨干网络	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	$AP_{\text{box}}^{\text{test-dev}}$	参数量
CenterNet	Res-18	144	28.0/30.2/33.5	47.2/51.3	31.0/35.4	10.1/14.5	30.2/31.6	47.5/48.1	28.2	13.2 M
C-Net+GAN	Res-18	144	**29.6/31.8/34.1**	**47.9/52.9**	**33.1/36.6**	**12.7/15.9**	**31.1/33.6**	**48.5/49.5**	**30.1**	34.2 M
CenterNet	Res-101	46	34.2/36.1/39.3	54.6/58.5	38.7/42.1	17.7/20.1	41.5/43.2	50.2/52.9	35.0	49.2 M
C-Net+GAN	Res-101	46	**35.6/37.2/40.5**	**55.6**/58.5	**39.7/42.4**	**19.5/22.6**	**42.6/44.6**	**50.3/53.8**	**36.4**	121.8 M
CenterNet	H-104	14	40.3/42.2/44.5	**61.3**/63.2	**46.0**/49.1	24.0/26.2	45.1/47.0	52.3/56.8	40.8	109.2 M
C-Net+GAN	H-104	14	**41.1/43.0/46.2**	61.2/**63.6**	45.9/**49.5**	**24.9/27.0**	**46.3/47.6**	**53.1/56.9**	**41.6**	262.1 M
CenterNet	DLA-34	52	37.2/39.0/41.6	56.5/59.4	42.5/45.0	19.9/21.0	42.5/43.7	51.8/54.9	37.7	50.5 M
C-Net+GAN	DLA-34	52	**37.9/40.3/42.5**	**57.4/59.8**	**43.6/45.9**	**21.6/22.8**	**43.4/44.3**	**52.8/55.3**	**38.5**	126.3 M

实验的硬件设备是 AMD Ryzen 7 5800X，训练使用 8 台 GeForce RTX 3090，测试使用 1 台 GeForce RTX 3080. 软件环境基于 PyTorch 1.10.2、PyTorch-Lightning 1.5.10 和 CUDA 11.3，使用官方预训练权重.

4.1　网络框架性能

表 3 展示了本文的方法在 COCO 数据集上的效果. 实验指标包括网络在不同 *IOU* 值下的平均准确度、平均推理时间、每秒处理的图片数量（FPS）以及达到最佳性能所需的迭代次数（*epoch*）. 我们还对

原始数据集(N.A.)、翻转(F)和多尺度增强(MS)的方法进行了测试,并最终通过对 4 次重复实验取均值得到了最终的数据.实验结果表明,使用 Hourglass-104 骨干网构建的 CenterNet 检测器的性能最为优秀,尤其是在使用了多尺度增强(MS)之后,它的平均检测准确率(AP)达到 45.7%,同时它的推理速度最低,只有 1.6 FPS.而利用 ResNet-18 构建的检测网络则具备最高的工作效率,达到 12 FPS,但是,它的AP(MS)值为 34.1%.另外,ResNet-101 及 DLA-34 所构成的检测器在性能和检测准确率上都处于上述两种情况之间的位置,而在此两者之中,DLA-34 的表现要比 ResNet-101 更胜一筹.此外,从训练时间的角度来看,小型且性能较差的网络能够更快地达到生成对抗训练的最优性能,而进一步提高已经相对优异的架构的性能则更为困难.

表 4 为使用对抗训练前后的对比数据,包括了 4 个不同骨干网络和 IOU 指标下的对比结果.其中,Res、C-Net 和 H 分别代表 ResNet、CenterNet 和 Hourglass 的缩写;FPS 表示在没有进行数据增强时的检测速度.将生成对抗训练方法应用于四类主干架构之后,大多数情况下的平均精确度都得到了显著改善.以具有最少网络参数的 ResNet-18 为例,它的 AP 值提高了 1.9%.至于拥有更多参数的 ResNet-101、Hourglass-104 和 DLA-34,它们的 AP 值分别增长了 1.4%、0.8%和 0.8%.由此可见,针对那些拥有较小参数的主干网络,额外引入的生成对抗训练能有效地增强其实际表现.此外,针对大尺寸目标,四个骨干网络的 AP 平均增益为 0.73%/0.7%(F/MS),低于小尺寸目标的 1.75%/1.63%.虽然我们的模型在处理小规模目标和小网络架构方面表现更为突出,但对于参数较多的主要网络和大规模目标,生成对抗训练方法依旧能够产生一定的效果.训练时我们在特征提取网络和检测头的基础上增加了鉴别器和辅助骨干网络,网络的训练参数量增加.但由于在检测时被重新组合进目标检测网络,鉴别器不参与检测,因此,网络的修改对检测速度没有产生负面影响,FPS 指标的前后一致证明了这一点.

图 5 展示了在生成对抗训练阶段和检测器训练期间,未经增强的训练和验证数据集与最优分布之间的 Wasserstein 距离.Wasserstein 的计算方法与第 3.1 节相同.我们只在检测器的训练阶段使用验证数据集进行验证,所以我们不使用训练损失和验证损失.相反,根据相关研究[40],Wasserstein 距离也可以评估生成对抗网络的收敛状态.左半部分:随着对抗过程的进行,当前分布与最优分布之间的 Wasserstein 距离逐渐减小.右半部分:检测器能够快速超越原有的性能与 GAN 部分.由收敛曲线可知,在加入了生成对抗架构后,检测器可以在 15 个训练批次内快速达到收敛,这证明了对抗架构的可行性.

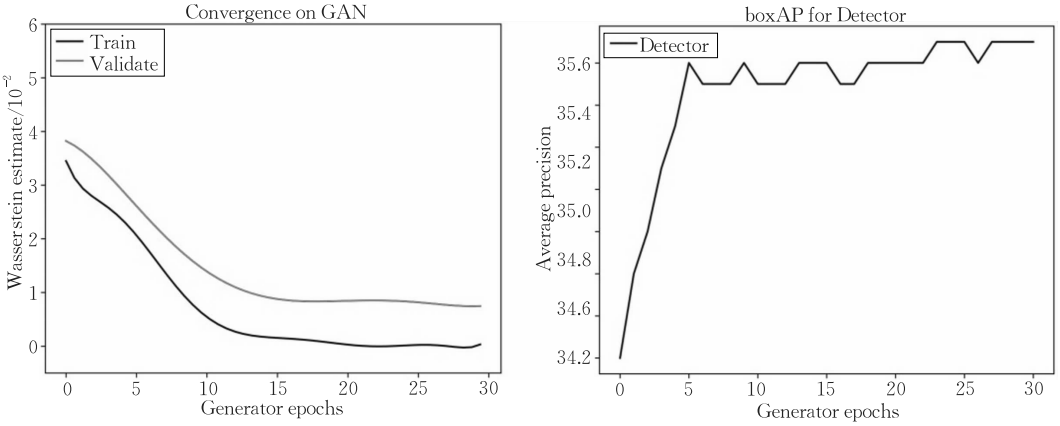


图 5 不同训练阶段生成器和检测器性能曲线

4.2 高斯模糊和锐化程度对性能的影响

依据第 3.1 节中的实验数据,我们可以证实:通过调整不同的模糊与锐化的水平,可以有效地辅助特征提取网络的提取特征,进而推进生成对抗训练.为了更直观地衡量这些图像处理方法对网络效能的具体影响,我们在[1,3]范围内以 0.5 像素为步长设置高斯模糊半径,并在[0.2,1.0]范围内以 0.2 为步长设置锐化程度的加权系数,实验结果如表 5 和表 6 所示.这些实验都是基于 ResNet-101 的对抗训练,并且没有使用测试增强技术.

表 5 高斯模糊的影响

高斯模糊	AP/%
1.0	+0.4
1.5	+1.1
2.0	+0.9
2.5	+0.2
3.0	−0.7

表 6　锐化程度的影响

锐化系数	AP/%
0.2	+0.5
0.4	+0.6
0.6	+0.6
0.8	+0.6
1.0	+0.6

在表 5 中，我们设置了不同像素的高斯模糊半径，包括 1.0、1.5、2.0 和 2.5. 结果显示，在对抗性训练中，无论使用何种半径，都能为原始目标检测网络带来收益. 当半径设定为 1.5 个像素时，识别效果最为显著，相较于未采用高斯模糊的情况，整体精确度提升了 1.1%. 当半径超出 3.0 像素后，由于背景中的缺失的上下文信息开始产生负面影响，抵消了突出目标特性的正向作用，导致网络性能的下降. 通过比较表 2 的数据，虽然增大高斯模糊的半径能令实验数据更接近最优分布，但并不能为对抗性训练创造更优的梯度，反而抑制了网络学习目标语境的行为.

表 6 显示，微小的边缘锐化变化对网络性能增益影响不大，大致在 0.5%到 0.6%之间. 即便是较低的 0.2 锐化程度，也能产生足够的效果来突出物体的细节.

4.3　对抗训练的效率

我们的对抗训练建立在一个经过充分训练的目标检测网络基础上. 按照传统的训练方法，一个训练完备的网络很难在额外的训练下获得可观的增益. 为了证明我们提出的训练方法可以以更小的代价显著提升这类网络的性能，我们对经过 150 个训练周期的 CenterNet 网络分别采用标准训练及本文提到的对抗式训练进行对比. 表 7 展示了对抗训练的收益. 在实验中，我们使用了一台 GeForce RTX 3080 主机进行基于 ResNet-101 的对抗训练，没有使用测试增强方法.

表 7　常规训练和对抗训练的收益以及消耗的计算机资源

常规训练		对抗训练	
时间/min	AP/%	时间/min	AP/%
0	34.2	0	34.2
480	34.3	60	34.8
960	34.0	120	35.3
1440	34.2	180	35.6
1920	34.3	240	35.5
2400	34.2	360	35.6
2880	34.3	420	35.6
		480	35.6

在表 7 中，我们对经过充分训练的网络采用了不同的训练方式进行进一步训练. 我们通过传统的训练方法，在 48 小时内完成了 12 个训练批次，平均精度在 34.2%至 34.3%之间波动，最后精度提升了 0.1%. 在使用生成对抗的训练方式时，6 个小时的训练就使得网络的性能达到了稳定. 实验证明，我们的架构能够快速而有效地提升一个经过充分训练的目标检测网络的性能.

4.4　附加实验

为了验证模型的性能，我们使用了 PASCAL VOC 2007 数据集[44]. 基于之前提到的 4 种骨干网络，在输入图片分辨率为 384×384 和 512×512 的情况下，使用 IOU=50%时的 mAP 作为评估标准. 实验结果如表 8 所示，它与第 4.1 节所获得的结论是一致的.

表 8　基于 PASCAL VOC 2007 数据集的训练效果

	Backbone	Resolution	AP_{50}
CenterNet	ResNet-18	384×384	72.4
Our method	ResNet-18	384×384	**74.1**
CenterNet	ResNet-18	512×512	75.8
Our method	ResNet-18	512×512	**77.2**
CenterNet	ResNet-101	384×384	77.4
Our method	ResNet-101	384×384	**78.5**
CenterNet	ResNet-101	512×512	78.6
Our method	ResNet-101	512×512	**79.7**
CenterNet	DLA-34	384×384	79.7
Our method	DLA-34	384×384	**80.6**
CenterNet	DLA-34	512×512	80.8
Our method	DLA-34	512×512	**81.4**

我们同样在其他几个出色的目标检测架构上进行了实验，包括 YOLOv4[45]、YOLOv7[46] 以及使用 Swin Transformer[47] 构建的 Cascade Mask R-CNN[48]. 考虑到同样采用了数据增强方法，我们还将采用了 AutoAugment 策略的 RetinaNet 检测器纳入了对比实验. 实验结果如表 9～表 11 所示.

针对 YOLOv4 架构，我们将网络从 Neck 部分和检测头部分实施了拆分，作为特征抽取网络和检测器. 与 CenterNet 相比，YOLOv4 的热力图卷积参数更为丰富，因此执行了更多的训练周期，对抗训练和检测训练的比例为 3∶1，实验在 COCO 数据集上进行. 在训练 YOLOv7 的过程中，我们的特征增强技术应用于 mosaic 增强之后，否则 mosaic 增强的随机缩放和截取应用在我们处理过的图片会导致图片细节信息的过多丢失.

对于采用 Swin Transformer 构建的 Cascade Mask R-CNN 结构，我们的策略是实施多尺度的训练，并且对图像的长宽尺寸进行了约束，短边在 480 到 800 像素之间，长边不超过 1000 像素. 我们使用 AdamW[49]

表 9 基于 YOLO 系列网络和 COCO 数据集的训练效果

	Resolution	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv4	320×320	89	39.2	**58.6**	40.9	16.9	44.1	**59.2**
YOLOv4+GAN	320×320	89	**40.6**	58.5	**41.3**	**19.4**	**45.6**	58.8
YOLOv4	416×416	84	41.7	61.4	42.1	22.0	46.6	57.5
YOLOv4+GAN	416×416	83	**42.4**	**61.5**	**42.7**	**24.9**	**48.6**	**57.7**
YOLOv7	640×640	157	50.8	**69.4**	55.3	35.1	56.0	66.7
YOLOv7+GAN	640×640	157	**51.1**	69.3	**55.5**	**35.8**	**56.2**	**67.1**
YOLOv7-X	640×640	112	52.8	**71.0**	57.2	36.5	57.6	68.7
YOLOv7-X+GAN	640×640	112	**53.0**	70.9	**57.4**	**37.0**	**57.9**	**68.8**

表 10 基于 Swin Transformer 的 Cascade Mask R-CNN 架构和 COCO 数据集的训练效果

	AP	AP_{50}	AP_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	参数
Cascade Mask R-CNN	47.8	66.2	**52.5**	41.4	**63.6**	45.1	80 M
Cascade Mask R-CNN+GAN	**48.3**	**66.4**	52.4	**41.9**	63.5	**45.5**	194 M
S-T+Cascade Mask R-CNN	50.5	69.3	54.6	43.8	66.5	47.2	82 M
S-T+Cascade Mask R-CNN+GAN	**50.8**	**69.5**	**54.7**	**44.1**	**66.9**	**47.8**	201 M

注:S-T 为 Swin Transformer 的缩写.

表 11 基于 ResNet101 骨干网络和 RetinaNet 检测器的数据增强方法效果对比

	Backbone	AP	AP_{50}	AP_{75}
RetinaNet	ResNet101	39.2	57.0	42.7
RetinaNet+AutoAugment		40.6	58.9	43.4
RetinaNet+GAN		40.8	58.6	43.9
RetinaNet+AutoAugment+GAN		**41.4**	**59.6**	**44.3**

作为优化工具,同时与 ResNet101 模型进行了比较实验.对抗训练及分类训练的交替执行频率为 4∶1.实验在 COCO 数据集上进行.在 AutoAugment 方法中,我们在每个大小为 5k 的 COCO 训练集批次上搜索了增强策略.较优的策略包括旋转、Y 轴翻转和包围盒 Y 轴翻转.

从表 9 的数据中可以看出,使用生成对抗训练方法后的 YOLOv4,其识别能力有明显增强.针对 320×320 及 416×416 两种分辨率的图片,其识别精度的均值提升分别为 1.4%与 0.7%.表 10 则详细描述了 Cascade Mask R-CNN 网络在 ResNet-101 和 Swin Transformer 中的平均准确率和掩膜平均准确率(mask AP).虽然鉴别器的加入导致了网络参数和训练时间的增加,但这并不会影响检测网络的推理时间.根据表 11 的数据,我们的方法在 RetinaNet 检测器上获得了与 AutoAugment 相似的增益,但由于我们的方法可以应用于正常的训练之后,因此将两种方法结合起来可以进一步提升效果.

5 结 论

在本文中,我们通过实验证明了具有优秀信息表征的图片在特征提取网络中能够获得更好的特征分布.借此,我们提出了一种新的目标检测框架,利用 GAN 网络的对抗训练,能够使训练完善的目标检测网络进一步提高性能.在生成对抗训练过程中,我们通过一些图像处理技术使生成器容易获得更好的学习目标.此外,在对抗性训练中不加入检测模块,既提高了训练的稳定性,又避免了三个模块的共同博弈.

实验证明,我们所构建的框架对于已经完成训练的目标检测器具有很好的适用性,其训练效率高且消耗的计算资源少,能以较少的资源获取较大的额外增益.它不会产生推理时间增加或者训练难度增大的负面效应,几乎可以说是一种即插即用的训练方法.然而,该体系结构在训练过程中存在一定的局限性,即对于大型目标或具有大量参数的网络增益较小.此外,由于我们在训练特征提取网络时强化了训练图像的特征,对于一些差异性较大的数据集,迁移学习将变得更加困难.针对上述局限性,我们未来的改进可以从两个方面入手:(1)针对不同的对象应用不同的图像处理技术和增强级别,改进特征突出方法;(2)改进模型结构,使检测模块的训练无需固定特征提取网络的权值.

参 考 文 献

[1] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149

[2] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot MultiBox detector//Proceedings of the Computer Vision—ECCV 2016: 14th European Conference. Part I 14. Amsterdam, The

Netherlands, 2016: 21-37

[3] Law H, Deng J. CornerNet: Detecting objects as paired keypoints//Proceedings of the Computer Vision—ECCV 2018: 15th European Conference. Munich, Germany, 2018: 734-750

[4] Tian Z, Shen C, Chen H, et al. FCOS: Fully convolutional one-stage object detection//Proceedings of the IEEE CVF International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 9627-9636

[5] Wan L, Eigen D, Fergus R. End-to-end integration of a convolution network, deformable parts model and non-maximum suppression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 851-859

[6] Hosang J, Benenson R, Schiele B. Learning non-maximum suppression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4507-4515

[7] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks//Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Part IV 14. Amsterdam, The Netherlands, 2016: 630-645

[8] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916

[9] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2881-2890

[10] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117-2125

[11] Li J, Liang X, Wei Y, et al. Perceptual generative adversarial networks for small object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1222-1230

[12] Bai Y, Zhang Y, Ding M, et al. SOD-MTGAN: Small object detection via multi-task generative adversarial network//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 206-221

[13] Ehsani K, Mottaghi R, Farhadi A. SeGAN: Segmenting and generating the invisible//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6144-6153

[14] Prakash C D, Karam L J. It GAN DO better: GAN-based detection of objects on images with varying quality. IEEE Transactions on Image Processing, 2021, 30(11): 9220-9230

[15] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs//Proceedings of the 30th Annual Conference on Neural Information Processing Systems 2016. Barcelona, Spain, 2016: 2234-2242

[16] Li C, Xu T, Zhu J, et al. Triple generative adversarial nets. Advances in Neural Information Processing Systems, 2017, 30: 4088-4098

[17] Goodfellow I, Bengio Y, Courville A. Deep Learning. Boston, USA: MIT Press, 2016

[18] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3431-3440

[19] Ray A, Kumar S, Reddy R, et al. Multi-level attention network using text, audio and video for depression prediction//Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. New York, USA, 2019: 81-88

[20] Qin Q, Hu W, Liu B. Feature projection for improved text classification//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Washington, USA, 2020: 8161-8171

[21] Kong T, Yao A, Chen Y, et al. HyperNet: Towards accurate region proposal generation and joint object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 845-853

[22] Chen Q, Wang Y, Yang T, et al. You only look one-level feature//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Montreal, Canada, 2021: 13039-13048

[23] Zhou X, Wang D, Krähenbühl P. Objects as points. arXiv preprint arXiv:1904.07850, 2019

[24] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014. Montreal, Canada, 2014: 2672-2680

[25] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 214-223

[26] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015

[27] Zhou P, Xie L, Ni B, et al. Omni-GAN: On the secrets of cGANs and beyond//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 14061-14071

[28] Wu Y L, Shuai H H, Tam Z R, et al. Gradient normalization for generative adversarial networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 6373-6382

[29] Salimans T, Kingma D P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks//Proceedings of the 30th Annual Conference on Neural Information Processing Systems 2016. Barcelona, Spain, 2016: 901-909

[30] Shen Y, Gu J, Tang X, et al. Interpreting the latent space of GANs for semantic face editing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9243-9252

[31] Abdal R, Qin Y, Wonka P. Image2StyleGAN: How to edit the embedded images?//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8296-8305

[32] Yang X, Li Y, Qi H, et al. Exposing GAN-synthesized faces using landmark locations//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. Paris, France, 2019: 113-118

[33] Karras T, Aittala M, Laine S, et al. Alias-free generative adversarial networks. Advances in Neural Information Processing Systems, 2021, 34: 852-863

[34] Posilović L, Medak D, Subašić M, et al. Generative adversarial network with object detector discriminator for enhanced defect detection on ultrasonic B-scans. Neurocomputing, 2021, 459(7): 361-369

[35] Sultana M, Mahmood A, Bouwmans T, et al. Moving objects segmentation using generative adversarial modeling. Neurocomputing, 2022, 506(8): 240-251

[36] Box G E P, Cox D R. An analysis of transformations. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1964, 26(2): 211-243

[37] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein GANs//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5767-5777

[38] Liang T, Chu X, Liu Y, et al. CBNet: A composite backbone network architecture for object detection. IEEE Transactions on Image Processing, 2022, 31: 6893-6906

[39] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[40] Yu F, Wang D, Shelhamer E, et al. Deep layer aggregation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2403-2412

[41] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation//Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Part VIII 14. Amsterdam, The Netherlands, 2016: 483-499

[42] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Part V 13. Zurich, Switzerland, 2014: 740-755

[43] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[44] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111(1): 98-136

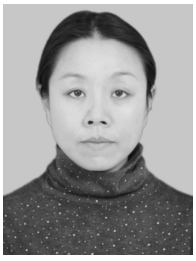
[45] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934, 2020

[46] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7464-7475

[47] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 10012-10022

[48] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6154-6162

[49] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017



ZHANG Yun, Ph. D., associate professor. Her research interests include computer vision, soft computing algorithm theory and communication signal processing.

HUANG Cheng, M. S. Her research interests include computer vision and generate adversarial network.

SHI Jian, M. S. His research interest is computer vision.

ZHANG Yu-Yao, M. S. Her research interests include computer vision and generate adversarial network.

HUANG Jing-Wei, M. S. His research interest is computer vision.

YU Shu-Juan, M. S., professor. Her research interests include computer vision and intelligent signal processing.

HUANG Li-Ya, Ph. D., professor. Her research interests include computer vision and brain computer interface technology.

Background

Object detection is one of the fundamental problems in computer vision, but with the increasing application areas and the improvement of computer performance, a large number of deep and complex convolutional layers are used to enhance the feature extraction capability of the network for dense details of images. It directly leads to a dramatic increase in computational resource consumption and difficulties in controlling the feature extraction network during training. The poor detection ability of object detection may be caused by the network's insufficient ability to extract local details or global features, or it may be due to unbalanced datasets and uneven positive and negative samples, but ultimately it is the performance of the feature extraction network that is the problem. How to improve the feature extraction ability of the network effectively and without losing efficiency is the main concern of the researchers.

For a well-trained object detection network, using conventional training methods to further improve the network capability is exponentially more expensive in terms of computational resources and time consumption. It is common to use architectures with higher utilization of network layers, use multi-scale features in classification, and avoid information loss in gradient transfer during conventional training. These approaches inevitably lead to more complex network structures and difficult to control training processes. Due to the learning ability of generative adversarial networks for potential spatial distributions, researchers have thought of using generative adversarial networks to enhance datasets or discriminators to strengthen target detectors, both of which create new problems of inefficient training and pattern collapse in object detection, respectively.

In this paper, we use a generative adversarial architecture to efficiently improve the feature extraction capability of the object detector, using intensively characterized pictures to consistently provide a better distribution for the generator. This approach is similar to a plug-and-play network optimizer, which effectively and efficiently improves network performance without changing the network architecture, i. e., without affecting the detection speed. Experiments on multiple target detection datasets, backbone networks, and target detection architectures show that our approach is outstanding for detection optimization of small networks, small datasets, and small targets. For other types of networks and detections, our enhanced approach can also improve to the limits of the network architecture with shorter training time compared to normal training.

This work is supported by the National Natural Science Foundation of China (61977039).