基于 VFNet 与数据增强的海上目标检测算法

孙科[1]　杨文斌[1]　何宇亨[1]

摘要 随着深度卷积神经网络的快速发展，其在计算机视觉、自然语言处理、语音识别等多个领域得到广泛应用。尤其是计算机视觉领域，近年来在图像分类、目标检测、姿态估计、图像分割和人脸识别等任务有突破性进展。本文的主要研究内容就是基于目标检测网络对海上船舶目标进行识别与定位，并详细介绍了本文所使用的目标检测模型的网络结构与原理。此外，我们还提出并详细介绍了基于语义分割的海上目标数据增强方法，并进一步提高了目标检测模型的性能。除此之外，我们还对目标检测、数据增强等方法的相关研究做了简单介绍，并在文章的最后对实验结果进行了介绍与分析。

关键词 目标检测，数据增强，船舶目标

Marine object detection based on VFNet and data augmentation

Sun Ke[1]　Yong Winbin[2]　He Yuheng[1]

Abstract With the development of deep convolutional neural network, it has been widely used in many fields such as computer vision, natural language processing and speech recognition. Especially in the field of computer vision, breakthroughs have been made in tasks such as image classification, object detection, pose estimation, image segmentation and face recognition in recent years. The main content of this paper is to classify and locate marine boat objects based on the object detection network, and introduces the network structure and principle of the object detection model used in this paper in detail. In addition, we also proposed and introduced a augmentation method based on semantic segmentation for the data set of boat objects, and further improved the performance of the object detection model. Finally, we made a brief introduction to the related research of object detection and data augmentation, and analyzed the experimental results at the end of the paper.

Key words object detection,data augmentation, boat object

1　引言

近年来，深度卷积神经网络的迅速发展使得目标检测受益极大，尤其是在视觉图像方面，此类的目标检测算法更是层出不穷。随着技术的成熟，其已经被广泛的应用到人脸识别、无人汽车驾驶等领域。如果能够将目标检测算法应用于海洋场景下船舶目标检测，可以为海上航行人员的决策提供辅助，同时减少人力的投入与航行成本。

在此次竞赛中，我们尝试了诸多目标检测模型，诸如 FasterRCNN[6]、CascadeRCNN[26]、DCN[27]以及最近提出的 DetectoRS[28]、VFNet[1]等，并在给定的 9800 张船舶图像数据集上进行训练，最终在 VFNET 上取得了较好的实验结果。但众所周知，基于深度卷积神经网络的目标检测模型需要足够的标注数据作为训练集。而本文所使用的训练集由于部分类别的实例数量较少或多样性过于丰富使得其最终的检测结果并不理想。基于此，我们提出基于语义分割的海上目标数据增强方法，对指定类别的船舶目标进行数据扩充，并进一步提高了目标检测网络的性能。

1.1

2　相关工作

2.1 目标检测

目标检测是计算机视觉领域的一个主要方向，其被广泛应用于工业检测、人脸识别、智能视频监控、车牌识别等多个领域。传统的目标检测算法主要依赖于人工选取的特征来对物体进行检测。人工提取的特征主要针对某些特定对象,比如有的特征适合做边缘检测,有的适合做纹理检测,不具有普遍性。近年来，随着深度卷积神经网络技术的发展，基于深度神经网络的目标检测算法成为了理论和应用的研究热点。

目前比较流行的目标检测模型可以根据是否使用 anchor box 分为两大类。

首先，基于 anchor 的目标检测模型在特征图的每个位置设置 anchor box，并预测每个 anchor box 中存在目标的概率，通过调整 anchor box 的大小以匹配目标。这类网络又可以细分为两大类：One-Stage 目标检测算法与 Two-Stage 目标检测算法。其中，One-Stage 目标检测算法不需要候选区域（Region Proposal）阶段，其只需要一个阶段直接产生目标物体的类别概率和位置坐标值，比较典型的算法有 SSD[7]、CornerNet[2] 以及 YOLO[3] 系列等。而 Two-Stage 目标检测算法将检测问题划分为两个阶段：第一个阶段首先产生 Region Proposals，包含目标大概的位置信息；第二个阶段对候选区域进行分类和位置精修。这类算法主要包括 R-CNN[4]、Fast R-CNN[5]、Faster R-CNN[6]等经典模型

相比于基于 anchor 的方式，不使用

anchor 的检测算法不需要具体设计适合检测目标对象的 anchor box，这在近期引起了大众的广泛关注。这类方法也可以细分为两类。其中一种将目标检测问题定义为关键点或语义点检测问题包括 CenterNet[22]、ExtremeNet[23]等。另一类与单阶段使用 anchor 检测的方式类似，但并不使用 anchor box，而是将特征金字塔上的每个点分类为前景或背景，直接预测前景点到真实包围盒四周的距离，从而进行目标检测。如 DenseBox[24]引入了一个全卷积神经网络来大幅提高效率；FCOS[25]将目标边界框内的所有点都当作正样本，并检测所有正样本点以及从任一正样本点到边界框的距离；而本文所使用的 VFNet[1]正是基于 ATSS 版本的 FCOS 所搭建的。

2.2 数据增强

数据增强算法可以有效增加现有数据集的规模及多样性，同时能够减少数据标注的工作量并降低成本。对于目标检测模型本身来说，数据增强也可以增加模型的鲁棒性，减少模型的过拟合。

传统的数据增强方法通常是对图像进行几何变换，比如水平翻转[6]、多尺度策略[7]、补丁裁剪[8]等操作，或者是对图像进行颜色变化，以改变原始图像的几何空间或颜色空间。之后，提出了更复杂的图像变换，比如对图像中的内容进行随机擦除[9]、遮挡图像中的一部分或混合两个甚至多个图像[10,11]。还有一部分工作借助于合成图像来扩充数据集，比如借助于生成对抗网络所得到的生成图像[12,13,14]或在 2D、3D 场景中得到的实例渲染图像[15,16,17]等，但由这些方法所得到的数据与真实数据存在明显的域间隙，使得基于这些数据训练的模型很难推广到真实数据上，因此有效应用于目标检测。最近，基于剪切粘贴的一系列方法[18，19，20，21]被提出，其通过剪切目标实例并将其粘贴到适当的背景图像上来增强训练数据集。而本文所提出的数据增强方法正是基于该类方法进行的改进。

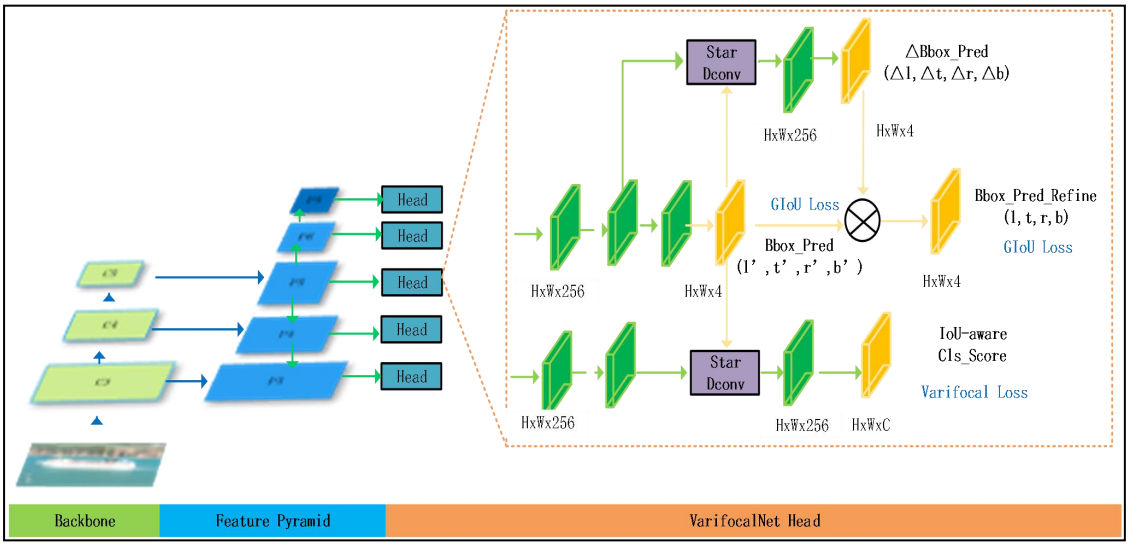

图 1　VFNet 网络结构

3　模型和方法描述

3.1 模型和方法原理

本文主要使用的目标检测模型是基于 ResNeXt101 为 backbone 的 VarifocalNet（简称 VFNet）网络，同时使用了可变形卷积网络（Deformable Convolutional Networks，简称 DCN）。以下是 VFNet 网络结构与原理

的简单介绍。

在目标检测任务中，非极大值抑制（Non-Maximum Suppression，简称 NMS）需要依据候选检测目标的排序来进行筛选框，因此这个排序的可靠性就非常重要。之前的工作主要采用 IOU 分支（IOU-Net）与 Centerness 得分（FCOS）来作为大量候选检测的排序依据。VFNet 提出了一个可以同时表示目标存在和定位精度感知的分类得分 IACS，实验证明了这是一个更优的候选框排序依据。其提出了新的 Varifocal loss 函数，来训练密集物体检测器使 IACS 回归，并设计了一种新的高效星形边界框特征表示法，用于预测 IACS 得分并改进边界框。最后，VFNet 还提出了一种基于 FCOS 架构的新型密集目标检测器，以利用 IACS 的优势。

VFNet 的网络结构如图 1 所示，它的骨干网络与 FPN 和 FCOS 相同。不同之处在于探测器的头部结构，VFNet 头由两个子网组成。其中，定位子网进行边界框回归和细化。它从 FPN 的每一层获取特征映射作为输入，首先通过三个带有 ReLU 激活的 3*3 卷积层，生成具有 256 个通道的特征图。定位子网的一个分支再次卷积特征映射，并输出表示初始边界框每个空间位置的 4 维距离向量。给定初始盒和特征映射，另一个分支对 9 个特征采样点使用星型可变形卷积，产生距离缩放因子，将缩放因子乘以距离向量，生成精细的边界框。

另一子网负责预测 IACS，其具有与用于细化操作的分支相似架构。该子网输出每个空间位置带有分类数元素的向量，每个元素的值联合表示目标的置信度和定位精度。

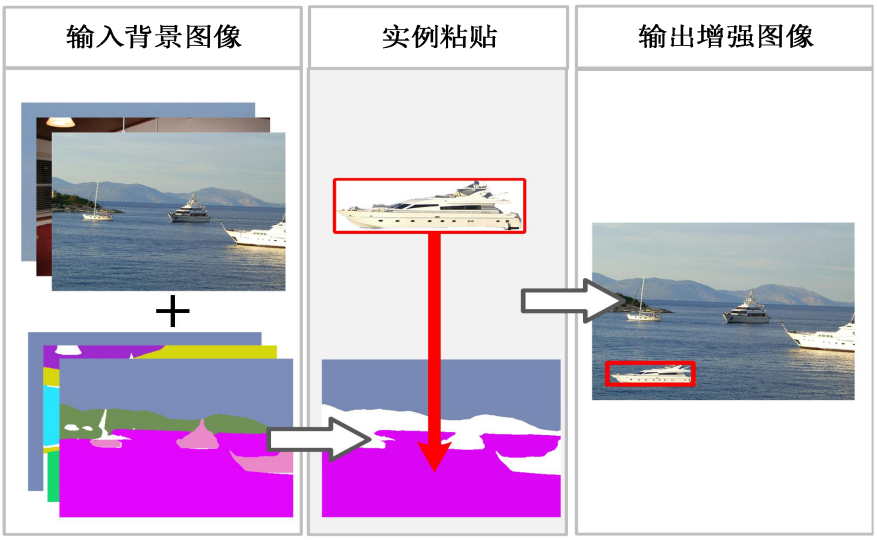

图 2 数据增强示意图

3.2 基于语义分割的数据增强方法

基于以上模型，我们提出了基于语义分割的船舶目标数据增强方法，如图 2 所示。我们将训练集中的各类实例分割标注并剪切出来，构建实例库，并对部分图像进行语义分割，将海面语义分割出来，并将其作为背景图像，然后将实例库中的实例粘贴于任意背景图像中的海面上，从而生成新的图像样本，并将其用于以上模型的微调。

数据分析。在此之前，为更有目的性地进行增强，我们首先将训练集分为训练集与验证测试集，以分析各类别的检测结果，同时对实例数量进行分析，分析结果如表 1 所示。如表所示，我们发现其中 island reef、sailboat、othership 等三个类虽然实例数量居

多但检测结果却较低，而 container ship、liner 的虽然实例数量较少但检测结果却都在 0.750 以上。对于后者，我们认为依然有增强的必要，以进一步提高其检测结果。对于前者，我们通过数据对比发现，这三类目标较差的检测结果可能是由于实例目标过小以及多样性过于丰富导致的，进而使得目标检测模型难以学习其特征。而本文所提出的基于语义分割的海上目标数据增强方法能够有效增加目标实例的小尺度目标以及实例的位置、尺度等多样性。因此，我们也对这三类数据进行了增强。

表 1 训练集各类实例的数量及准确率分析

	liner	container	carrier	island	sailboat	other ship
Number	1447	871	3121	5910	3350	8843
AP	0.783	0.821	0.719	0.383	0.657	0.462

分割标注。在对数据集分析过后，我们分别对 island reef 、 sailboat 、 othership 以及 container ship、liner 等类别进行增强。在此之前，我们还需对实例图像进行实例分割标注以及对背景图像进行语义分割标注，如图 3 所示。我们共对 389 张背景图像进行语义分割，并分割实例 island reef 250 个、other ship 247 个、sailboat 114 个、container ship 100 个、liner 100 个。

增强数据。通过语义分割与实例分割标注，我们得到充足的背景图像库与实例库。我们可以通过将两者进行任意组合，得到大量生成的增强图像，如图 4 所示。

4 实验结果

本次比赛模型训练基于 mmdetection2.7 框架，具体的训练细节如下：首先，初始学习率 1e-2，训练 16 个 epoch 后衰减至 1e-3，20 个 epoch 后衰减至 1e-4，直至训练至 24 轮，其中在 21 轮得到最佳检测结果。因此，我们将模型在第 21 个 epoch 所得到的检查点上进行微调，而微调所使用的数据集则只使用上述增强的数据集。最后，我们使用训练命令进行训练，训练至 6 个 epoch 后，取得目前最优的 mAP 值 65.0。

图 3 语义分割标注与实例分割标注示意图

图 4 增强图像示意图

References

1. Zhang, Haoyang , et al. "VarifocalNet: An IoU-aware Dense Object Detector." (2020).
2. Hei Law, Jia Deng. ″CornerNet: Detecting Objects as Paired Keypoints.″ Int. J. Comput. Vis. 128(3): 642-656 (2020).
3. Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, Ali Farhadi. ″You Only Look Once: Unified, Real-Time Object Detection.″ CoRR abs/1506.02640 (2015).
4. Ross B. Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. ″Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.″ CVPR 2014: 580-587.
5. Girshick, Ross . "Fast R-CNN." Computer Science (2015).
6. *S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards real-time object detection with region proposal net-works." In NIPS. 2015.*
7. B. Singh, M. Najibi, and L. S. Davis. "SNIPER: Efficient multi-scale training." In NeurIPS. 2018.
8. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector." In ECCV. 2016.
9. Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. "Random erasing data augmentation." arXiv: 1708.04896. 2017.
10. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. "mixup: Beyond empirical risk minimization." In International Conference on Learning Representations. 2018.
11. J. Lemley, S. Bazrafkan, and P. Corcoran. "Smart augmentation learning an optimal data augmentation strategy." IEEE Access, 5:5858 – 5869. 2017.
12. S. Azadi, D. Pathak, S. Ebrahimi, and T. Darrell. "Compositional GAN: Learning conditional image composition." arXiv: 1807.07560. 2018.
13. L. Chongxuan, T. Xu, J. Zhu, and B. Zhang. "Triple generative adversarial nets." In Advances in Neural Information Processing Systems (NIPS). 2017.
14. T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid. "A bayesian data augmentation approach for learning deep models." In Advances in Neural Information Processing Systems(NIPS). 2017.
15. Karsch, K., Hedau, V., Forsyth, D., Hoiem, D. "Rendering synthetic objects into legacy photographs." ACM Transactions on Graphics (TOG)30(6) (2011) 157
16. Movshovitz-Attias, Y., Kanade, T., Sheikh, Y. "How useful is photo-realistic rendering for visual learning?" In: Proceedings of the European Conference on Computer Vision (ECCV). 2016.
17. Su, H., Qi, C.R., Li, Y., Guibas, L.J. "Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views." In: Proceedings of the International Conference on Computer Vision (ICCV). 2015.
18. N. Dvornik, J. Mairal, and C. Schmid. "Modeling visual context is key to augmenting object detection datasets." In ECCV. 2018.
19. "Cut, paste and learn:Surprisingly easy synthesis for instance detection." In ICCV. 2017.
20. Gupta, A., Vedaldi, A., Zisserman, A. "Synthetic data for

text localisation in natural images." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

21. Georgakis, G., Mousavian, A., Berg, A.C., Kosecka, J. "Synthesizing training data for object detection in indoor scenes." arXiv preprint arXiv: 1702.07836. 2017.
22. Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian. CenterNet. ″Keypoint Triplets for Object Detection.″ ICCV 2019: 6568-6577
23. Zhou, Xingyi , J. Zhuo , and P. Krahenbuhl . "Bottom-Up Object Detection by Grouping Extreme and Center Points." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2019.
24. Lichao Huang, Yi Yang, Yafeng Deng, Yinan Yu. "DenseBox: Unifying Landmark Localization with End to End Object Detection." CoRR abs/1509.04874 (2015).
25. Tian, Zhi , et al. "FCOS: Fully Convolutional One-Stage Object Detection." 2019 IEEE/CVF International Conference on Computer Vision (ICCV) IEEE, 2020.
26. Zhaowei Cai, Nuno Vasconcelos. "Cascade R-CNN: Delving into High Quality Object Detection." CoRR abs/1712.00726 (2017)
27. Dai, Jifeng , et al. "Deformable Convolutional Networks." (2017).
28. Qiao, Siyuan , L. C. Chen , and A. Yuille . "DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution." arXiv (2020).