# TinaFace: Strong but Simple Baseline for Face Detection

Yanjia Zhu[*], Hongxiang Cai[*], Shuhan Zhang[†], Chenhao Wang[†], Yichao Xiong[‡]
Media Intelligence Technology Co.,Ltd
{yanjia.zhu, hongxiang.cai, shuhan.zhang, chenhao.wang, yichao.xiong}@media-smart.cn

## Abstract

*Face detection has received intensive attention in recent years. Many works present lots of special methods for face detection from different perspectives like model architecture, data augmentation, label assignment and etc., which make the overall algorithm and system become more and more complex. In this paper, we point out that* ***there is no gap between face detection and generic object detection****. Then we provide a strong but simple baseline method to deal with face detection named TinaFace. We use ResNet-50 [11] as backbone, and all modules and techniques in TinaFace are constructed on existing modules, easily implemented and based on generic object detection. On the hard test set of the most popular and challenging face detection benchmark WIDER FACE [48], with single-model and single-scale, our TinaFace achieves 92.1% average precision (AP), which exceeds most of the recent face detectors with larger backbone. And after using test time augmentation (TTA), our TinaFace outperforms the current state-of-the-art method and achieves 92.4% AP. The code is available at `https://github.com/Media-Smart/vedadet/tree/main/configs/trainval/tinaface`.*

## 1. Introduction

Face detection becomes a very important task in computer vision, since it is the first and fundamental step of most tasks and applications about faces, such as face recognition, verification, tracking, alignment, expression analysis etc.. Therefore, so many methods are presented in this field from different perspectives recently. Some works [6, 7, 49] introduce annotated landmarks information as extra supervision signal, and some of others [51, 57, 37, 17, 26, 25, 58] pay more attention to the design of network. Besides, some new loss designs [51, 57, 16] and data augmentation methods [17, 37] are presented. What's more, a few works [23, 58] begin to redesign the matching strategy and label assignment process. Obviously, face detection seems to be gradually separated out from generic object detection and forms a new field.

Intuitively, face detection is actually an application of generic object detection. To some degree, face is an object. So naturally there are a series of questions to be asked, *"what is the difference between face detection and generic object detection?"*, *"Why not using generic object detection techniques to deal with face detection?"*, and *"is it necessary to additionally design special methods for handling face detection?"*.

First, from the perspective of data, the properties that faces own also exist in objects, like pose, scale, occlusion, illumination, blur and etc.. And the unique properties in faces like expression and makeup can also correspond to distortion and color in objects. Then from the perspective of challenges encountered by face detection like multi-scale, small faces and dense scenes, they all exist in generic object detection. Thus, face detection seems to be just a subproblem of generic object detection. To better and further answer above questions, we provide a simple baseline method based on generic object detection to outperform the current state-of-the-art methods on the hard test set of WIDER FACE [48].

The main contributions of this work can be summarized as:

- Indicating that face detection is actually a one class generic object detection problem and can be handled by techniques in generic object detection.

- Providing a strong but simple baseline method for face detection named TinaFace. All ideas and modules used in TinaFace are based on generic object detection.

- With single-scale and single-model, we achieve 92.1% average precision(AP) in hard settings on the test subset of WIDER FACE, which already exceed most of recent methods with larger backbone and Test Time Augmentation (TTA). Our final model gets 92.4% AP in hard settings on the test subset and outperforms current state-of-the-art methods for face detection.

[*] Equal contribution.
[†] Data analysis.
[‡] Corresponding author.

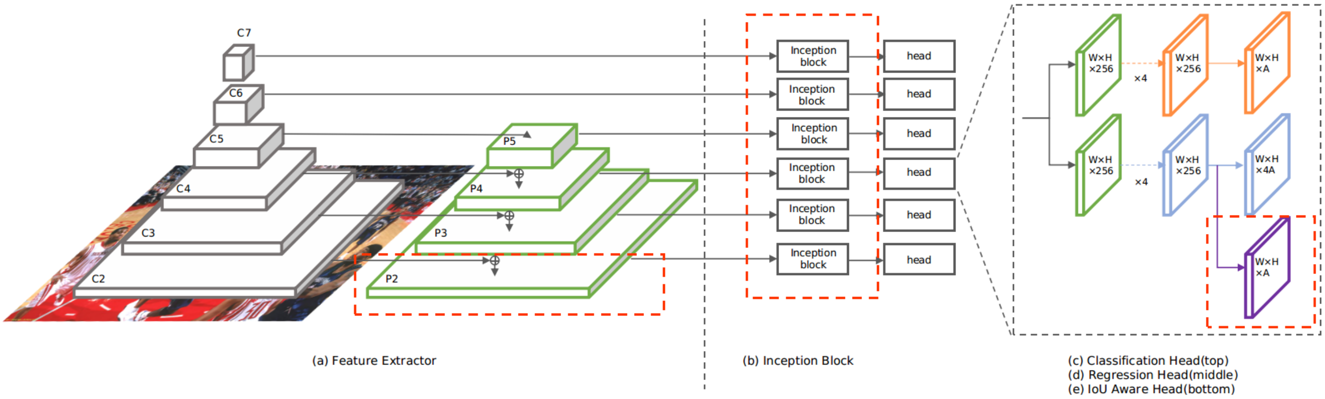

Figure 1: The model architecture of TinaFace. (a) Feature Extractor: ResNet-50 [11] and 6 level Feature Pyramid Network [18] to extract the multi-scale features of input image. (b) Inception block to enhance receptive field. (c) Classification Head: 5 layers FCN for classification of anchors. (d) Regression Head: 5 layers FCN for regression of anchors to ground-truth objects boxes. (e) IoU Aware Head: a single convolutional layer for IoU prediction.

## 2. Related Work

**Generic Object Detection.** Generic object detection aims at locating and classifying the existing objects in the given picture. Before the booming of deep learning, generic object detection is mainly based on the hand-crafted feature descriptors like SIFT [24] and HOG [5]. And the most successful methods like DPM [8] combine multi-scale hand-crafted features, sliding window, deformable part and SVM classifier to form a generic object detector.

With AlexNet [15] winning the championship of Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) by a large gap, the era of deep learning is coming, and generic object detection has been quickly dominated by deep learning methods. Two-stage methods start from R-CNN [10] and Fast R-CNN [9]. And soon Faster R-CNN [31] proposes RPN network to replace the selective search to generate proposals by pre-define anchors, which becomes the most classical anchor-based generic object detection method. Based on Faster R-CNN [31], there are so many new methods presented like FPN [18], Mask R-CNN [12], Cascade R-CNN [1] and etc.. In order to overcome the high latency of two-stage methods, many one-stage methods are presented like series of YOLO [30, 28, 29], SSD [22] and RetinaNet [19]. To handling the multiple scale or small objects problem, YOLOs [30, 28, 29] present novel anchor matching strategy including consideration of feedback of proposals and one ground-truth vs. one anchor, and also reweight the regression of width and height of objects. Then SSD [22] uses a hierarchy of backbone features, while FPN [18] presents feature pyramids. Besides, the series of SNIP [34] and SNIPER [35], multi-scale training and multi-scale testing can also deal with the multiple scale problem.

In addition to the new method proposed in generic object detection, developments in other fields, like normalization methods and deep convolutional networks, also promote generic object detection. Batch normalization (BN) [14] normalizes features within a batch along channel dimension, which can help models converge and enable models to train. In order to handle the dependency with batch size of BN, group normalization (GN) [44] divides the channels into groups and computes within each group the mean and variance for normalization. Then for deep convolutional networks, after AlexNet [15], VGG [33] increases depth using an architecture with very small $3 \times 3$ convolution filters, GoogLeNet [36] introduces Inception modules to use different numbers of small filters in parallel to form features of different receptive fields and help model to capture objects as well as context at multiple scales, and ResNet [11] demonstrates the importance of the original information flow and presents skip connection to handle the degradation with deeper networks.

**Face Detection.** As an application of generic object detection, the history of face detection is almost the same. Before the era of deep learning, face detectors are also based on hand-crafted features like Haar [39]. After the most popular and challenging face detection benchmark WIDER FACE dataset [48] presented, face detection develops rapidly focusing on the extreme and real variation problem including scale, pose, occlusion, expression, makeup, illumination, blur and etc.. Almost all the recent face detection methods evolve from the existing generic object detection methods. Based on SSD [22], S$^3$FD [58] extends anchor-associated layers to C3 stage and proposes a scale compensation anchor matching strategy in order to cover the small faces, PyramidBox [37] proposes PyramidAnchors (PA), Low-level

Feature Pyramid Networks (LFPN), Context-sensitive Predict Module (CPM) to emphasize the importance of context and data-anchor-sampling augmentation to increase smaller faces, and DSFD [16] introduce a dual-shot detector using Improved Anchor Matching (IAM) and Progressive Anchor Loss (PAL). Then Based on RetinaNet [19], RetinaFace [6] manually annotates five facial landmarks on faces to serve as extra supervision signal, RefineFace [57] introduces five extra modules Selective Two-step Regression (STR), Selective Two-step Classification (STC), Scale-aware Margin Loss (SML), Feature Supervision Module (FSM) and Receptive Field Enhancement (RFE), and HAMBox [23] emphasize the strong regression ability of some unmatched anchors and present an Online High-quality Anchor Mining Strategy (HAMBox). Besides, ASFD [51] uses neural architecture search technique to automatically search the architecture for efficient multi-scale feature fusion and context enhancement.

To sum up, methods presented in face detection almost cover every part of deep learning training from data processing to loss designs. It is obvious that all of these methods focus on the challenge of small faces. However, actually there are so many methods in generic object detection, which we mention above, solving this problem. Therefore, based on some of these methods, we present TinaFace, a strong but simple baseline method for face detection.

# 3. TinaFace

Basically, we start from the one-stage detector RetinaNet [19] as some previous works do. The architecture of TinaFace is shown in Figure 1 where the red dashed boxes demonstrate the different parts from RetinaNet [19].

## 3.1. Deformable Convolution Networks

There is an inherent limitation in convolution operation, that is, we feed it with a strong prior about the sampling position which is fixed and rigid. Therefore, it is hard for networks to learn or encode complex geometric transformations, and the capability of models is limited. In order to further improve the capability of our model, we employ DCN [4] into the stage four and five of the backbone.

## 3.2. Inception Module

Multi-scale is always a challenge in generic object detection. The most common ways to deal with it are multi-scale training, FPN architecture and multi-scale testing. Besides, we employ inception module [36] in our model to further enhance this ability. The inception module uses different numbers of $3 \times 3$ convolutional layers in parallel to form features of different receptive fields and then combine them, which help model to capture objects as well as context at multiple scales.

## 3.3. IoU-aware Branch

IoU-aware [43] is an extremely simple and elegant method to relieve the mismatch problem between classification score and localization accuracy of a single-stage object detector, which can help resort the classification score and suppress the false positive detected boxes (high score but low IoU). The architecture of IoU-aware is shown in Figure 1, and the only difference is the purple part, a parallel head with a regression head to predict the IoU between the detected box and the corresponding ground-truth object. And this head only consists of a single $3 \times 3$ convolution layer, followed by a sigmoid activation layer. At the inference phase, the final detection confidence is computed by following equation,

$$score = p_i^{\alpha} IoU_i^{(1-\alpha)} \tag{1}$$

where $p_i$ and $IoU_i$ are the original classification score and predicted IoU of $i$th detected box, and $\alpha \in [0, 1]$ is the hyperparameter to control the contribution of the classification score and predicted IoU to the final detection confidence.

## 3.4. Distance-IoU Loss

The most common loss used in bbox regression is Smooth L1 Loss [9] , which regresses the parameterizations of the four coordinates (box's center and its width and height). However, these optimization targets are not consistent with the regression evaluation metric IoU, that is, lower loss is not equivalent with higher IoU. Therefore, we turn to different IoU losses presented in past few years, directly regressing the IoU metric, such as GIoU [32], DIoU and CIoU [61]. The reason we choose DIoU [61] as our regression loss is that small faces is the main challenge of face detection since there are about two thirds data in WIDER FACE [48] belong to small object and DIoU [61] is more friendly to small objects. Practically, DIoU gets better performance on APsmall of the validation set of MS COCO 2017 [20]. And theoretically, DIoU is defined as:

$$L_{DIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} \tag{2}$$

where $\mathbf{b}$ and $\mathbf{b}^{gt}$ denote the central points of predicted box and ground-truth box, $\rho(\cdot)$ is the Euclidean distance, and c is the diagonal length of the smallest enclosing box covering the two boxes. The extra penalty term $\frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2}$ proposes to minimize the normalized distance between central points of predicted box and ground-truth box. Compared to large objects, the same distance of central points in small objects will be penalized more, which help detectors learn more about small objects in regression.

# TinaFace:强大但简单的人脸检测baseline

蔡鸿祥朱闫佳[*]张舒涵王晨好[†]熊逸超[‡]传媒智能科技有限公司

2023 年 1 月 16 日

**摘 要**

人脸检测近年来受到了广泛的关注。许多工作从模型架构、数据增强、标签分配等不同角度提出了大量的人脸检测专用方法,使得整个算法和系统变得越来越复杂。在本文中,我们指出了**人脸检测与一般目标检测之间没有差距**。在此基础上,我们提出了一种较强而简单的人脸检测baseline:TinaFace。我们的TinaFace以ResNet-50[11]为骨干网络,其中的所有模块和技术都是在现有模块上构造的,基于通用对象检测方法,易于实现。在最流行和最具挑战性的人脸检测基准**WIDER FACE**[48]的hard测试集上,单模型和单尺度上,我们的TinaFace达到了92.1%的平均精度,这超过了大多数最近的更大的人脸检测器的表现。在使用了TTA方法之后,我们的TinaFace比目前最先进的方法表现更好,达到了92.4%的AP。

## 1 介绍

人脸检测是计算机视觉中一个非常重要的任务,它是人脸识别、验证、跟踪、对齐、表情分析等大多数任务和应用的第一步。因此,近年来在这一领域出现了许多不同角度的方法。一些文献[6,7,49]将注释信息作为额外的监督信号,另外一些文献[51,57,37,17,26,25,58]更加注重网络的设计。此外,还提出了一些新的损失设计[51,57,16]和数据增强方法[17,37],还有一些工作开始重新设计匹配策略和标签分配流程。显然,人脸检测似乎逐渐从一般的目标检测中分离出来,形成了一个新的领域。

直观地说,人脸检测实际上是通用目标检测的一种应用。在某种程度上,脸就是一个检测对象。所以自然就会出现一系列问题:比如"人脸检测与一般对象检测有什么区别?""为什么不用一般对象检测技术来处理人脸检测?""是否有必要另外设计处理人脸检测的特殊方法?"

首先,从数据的角度来看,人脸拥有的属性也存在于物体中,比如姿态、比例、遮挡、光照、模糊等。像表情和化妆这种面部的独特属性,也可以对应物体的扭曲和颜色。人脸检测所遇到的多尺度、小人脸、密集场景等挑战都存在于一般的目标检测中。因此,人脸检测似乎只是一般对象检测的一个子问题。为了更好地进一步回答上述问题,我们提供了一种基于通用对象检测的baseline, 在WIDER FACE的hard测试集上胜过目前最先进的方法。

本文的主要贡献可以总结为:

- 说明人脸检测实际上是一类通用对象检测问题,可以通过通用对象检测技术进行处理。
- 提供了一种强大而简单的面部检测基线方法TinaFace。TinaFace中使用的所有思想和模块都是基于通用对象检测的。
- 在单尺度和单模型的情况下,我们在WIDER FACE测试子集下达到了92.1%的平均精度(AP),这已经超过了当前大多数具有较大主干网络并使用TTA方法的模型。我们的模型最终在测试子集的中获得92.4%的AP,优于当前最先进的人脸检测方法。

## 2 相关工作

### 2.1 通用目标检测

通用目标检测的目的是对给定图像中存在的目标进行定位和分类。在深度学习蓬勃发展之前, 一般的目标检测主要是基于手工制作的特征描述算子, 如SIFT[24]和HOG[5]。最成功的方法如DPM[8]将多尺度手工制作的特征、滑动窗口、可变形部件和SVM分类器相结合, 形成通用的目标检测器。随着AlexNet[15]获得2012年大规模视觉识别挑战赛(ILSVRC2012)冠军,深度学习时代到来,通用目标检测迅速被深度学习方法所主导。两阶段方法从R-CNN[10]和Fast R-CNN[9]开始, 很快, Faster R-CNN[31]就提出了RPN网络, 用预定义 anchors 的方法代替selective search算法,成为最经典的基于anchors的通用目标检测方法。基于Faster R-CNN[31],很多新方法出现了, 如FPN[18]、Mask R-CNN[12]、Cascade R-CNN[1]等。为了克服两阶段方法速度较慢的问题, 出现了许多单阶段方法, 如YOLO系列[30,28,29]、

[*] 相同贡献

[†] 数据分析

[‡] 通讯作者

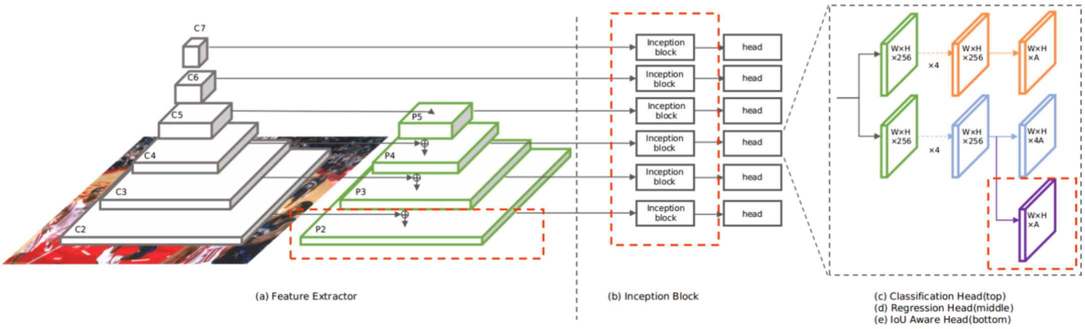

Figure 1: TinaFace的模型架构。(a)特征提取器:ResNet-50[11]和6级特征金字塔网络[18],提取输入图像的多尺度特征。(b)Inception模块增强接受野。(c)分类头:5层FCN用于锚框的分类。 (d)回归头:5层FCN,用于锚框回归到地真框。(e) IoU感知头:用于IoU预测的单个卷积层。

SSD[22]和RetinaNet[19]等。为了解决多尺度或小物体的问题, YOLOs[30,28,29]提出了新的锚点匹配策略,包括考虑 proposals 反馈和一个地真对一个锚点的策略, 并对物体宽度和高度的回归进行重估。SSD[22]使用一个主干特征的层次结构, 而FPN[18]使用特征金字塔。此外, 像SNIP[34]、SNIPER[35]系列这样的,多尺度训练、多尺度测试也可以应对多尺度问题。除了通用目标检测中提出的新方法外,其他领域的发展,如归一化方法和深度卷积网络,也促进了通用目标检测。批处理归一化(BN)[14]沿通道维对批处理内的特征进行归一化, 可以帮助模型收敛, 使模型能够训练。为了解决BN对batchsize的依赖问题,组处理归一化(GN)[44]将通道分成组,并在每组内计算归一化的平均值和方差。在深度卷积网络的发展中,AlexNet[15], VGG[33]使用架构和非常小的3*3卷积过滤器增加深度, GoogLeNet[36]介绍了 Inception 模块在多尺度检测中,使用不同数量的小过滤器并联形成的特性不同的接受域,帮助捕获对象以及上下文模型。 ResNet[11]展示了原始信息流的重要性, 并提出了 skip 连接来处理更深网络的退化。

## 2.2 人脸检测

人脸检测作为通用目标检测的一种应用,其发展历史几乎是相同的。在深度学习时代之前,人脸检测器也是基于Haar[39]等手工制作的特征。在最受欢迎和最具挑战性的人脸检测基准WIDER FACE 数据集提出之后, 人脸检测针对尺度、姿态、遮挡、表情、化妆、光照、模糊等极端和实际变化问题得到了快速发展。目前几乎所有的人脸检测方法都是从现有的通用目标检测方法发展而来的。基于SSD[22],$S^3$FD [58]为了小脸的检测将锚点相关层延伸至C3阶段, 提出了一种规模补偿锚匹配策略。 PyramidBox[37]提出PyramidAnchors (PA),低级特征金字塔网络(LFPN),上下文敏感的预测模块(CPM) 强调环境的重要性, 并使用数据锚框采样增强来增强小人脸的检测。 DSFD[16]引入了改进锚匹配(IAM)和渐进锚Loss(PAL)的双镜头检测器。 然后基于Retinanet网络[19],RetinaFace [6]人工标注人脸上的5个特征点作为额外的监督信号, 同时引入了选择性两步回归(Selective Two-step Regression, STR)、选择性两步分类(Selective Two-step Classification, STC)、尺度感知边缘损失(Scale-aware Margin Loss, SML)、特征监督模块(Feature supervision Module, FSM)和接受场增强(RFE) 5个额外模块。 HAMBox[23]强调了一些不匹配锚的强大回归能力,提出了一种在线高质量锚挖掘策略(HAMBox)。此外, ASFD[51]采用神经体系结构搜索技术自动搜索体系结构,实现高效的多尺度特征融合和上下文增强。

综上所述,人脸检测中的方法几乎涵盖了深度学习训练从数据处理到损失设计的每一个环节。很明显,所有这些方法都集中在小脸的挑战上。然而, 实际上在通用对象检测中有很多我们在前面提到过的方法可以解决这个问题。因此, 在这些方法的基础上, 我们提出了 TinaFace ：一种强大但简单的人脸检测baseline方法。

# 3 TinaFace

我们在单阶段检测器RetinaNet[19]之前的工作上进行改进。TinaFace的架构如图1所示,红色虚线框显示了与RetinaNet[19]不同的部分。

## 3.1 可变形卷积网络

卷积运算有其固有的局限性,即对采样位置的强先验是固定的、刚性的。因此,网络对复杂几何变换的学习和编码困难,模型的能力受到限制。为了进一步提高模型的性能, 我们将DCN[4]应用到主干网络的第四阶段和第五阶段。

## 3.2 Inception 模块

多尺度一直是通用目标检测中的一个难题。常用的处理方法有多尺度训练、FPN体系结构和多尺度测试。此外,我们在我们的模型中

使用了inception模块[36]来进一步增强这种能力。 inception模块使用不同数量的3 x 3卷积层并行形成不同接受域然后将它们组合在一起,帮助模型在多个尺度上捕捉检测对象和上下文。

### 3.3 IoU-aware 分支

IoU-aware[43]是一种非常简单优雅的可以缓解单级目标检测器分类分数与定位精度不匹配的问题的方法,可以利用分类分数,抑制误报检测框(高分低IoU)。 IoU-aware的架构如图1所示,唯一不同的是紫色部分,一个平行头和一个用来预测被检测盒与对应的地真对象之间的IoU的回归头。而这个头部只有一个3 x 3的卷积层,然后是一个sigmoid激活层。在推理阶段的最终检测置信度计算公式如下:

$$score = p_i^{\alpha} IoU_i^{(1-\alpha)} \tag{1}$$

其中 $p_i$ 和 $IoU_i$ 是第i个检测盒的原始分类分数和预测 IoU , $\alpha \in [0, 1]$是控制分类分数和预测IoU对最终检测置信度贡献的超参数。

### 3.4 Distance-IoU Loss

在bbox回归中最常用的损失是smooth L1损失[9],它回归四个坐标(box的中心及其宽度和高度)的参数。然而, 这些优化目标与回归评价指标IoU并不一致,即损失越低并不等于IoU越高。因此, 我们转向过去几年出现的不同IoU损失,直接回归IoU度量,如 GIoU [32]、DIoU 和 CIoU [61]。我们之所以选择DIoU[61]作为我们的回归损失,是因为小人脸是人脸检测的主要挑战, 在Widerface[48]中约有三分之二的数据属于小目标,而 DIoU [61]对小目标更友好。在实际应用中, DIoU 在MS COCO 2017[20]验证集APsmall上的性能较好。理论上,DIoU定义为:

$$L_{DIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b^{gt}})}{c^2} \tag{2}$$

其中 $boldb$ 和 $boldb_{gt}$ 表示预测框和地真框的中心点,$\rho()$ 是欧氏距离, $c$ 是覆盖两个盒的最小外接盒的对角线长度。额外罚款项 $\frac{\rho^2(\mathbf{b}, \mathbf{b^{gt}})}{c^2}$ 提出对预测盒中心点与地面真盒之间的归一化距离进行最小化。相对于大目标,同样距离的小目标中心点会受到更多的惩罚,这有助于检测器在回归过程中更好的学习小目标。