Hindawi
Discrete Dynamics in Nature and Society
Volume 2024, Article ID 9764828, 1 page
https://doi.org/10.1155/2024/9764828

Retraction

Retracted: FH-YOLOv4 with Constrained Aspect Ratio Loss for Video Face Detection and Public Safety

Discrete Dynamics in Nature and Society

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

In addition, our investigation has also shown that one or more of the following human-subject reporting requirements has not been met in this article: ethical approval by an Institutional Review Board (IRB) committee or equivalent, patient/participant consent to participate, and/or agreement to publish patient/participant details (where relevant).

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Wang, L. Hong, D. Gu, and P. Fu, "FH-YOLOv4 with Constrained Aspect Ratio Loss for Video Face Detection and Public Safety," *Discrete Dynamics in Nature and Society*, vol. 2022, Article ID 8355174, 10 pages, 2022.

Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 8355174, 10 pages
https://doi.org/10.1155/2022/8355174

Research Article

FH-YOLOv4 with Constrained Aspect Ratio Loss for Video Face Detection and Public Safety

Yue Wang, Liang Hong, Dewen Gu, and Pingping Fu

School of Management, Heilongjiang University of Science and Technology, Harbin 150001, China

Correspondence should be addressed to Pingping Fu; 2005801533@usth.edu.cn

Received 1 April 2022; Revised 16 May 2022; Accepted 15 June 2022; Published 9 August 2022

Academic Editor: Wei Zhang

Video face detection is a crucial first step in many facial recognition and face analysis systems. It should serve postprocessing steps as much as possible while satisfying high-accuracy real-time detection. In this paper, we first introduce the constrained aspect ratio loss (CARLoss) for better facial boxes regression and incorporate it into the modified FH-YOLOv4, then the IoU Tracker-based video face image deduplication algorithm is proposed on the detection level. Extensive experiments and comparative tests show the effectiveness of our method.

1. Introduction

Face detection aims at estimating face bounding boxes in a digital image without any scale and position priors, and its landing applications (long-distance automatic body temperature measuring device, digital camera auto-focus function, etc.) have affected all aspects of people's daily life. Importantly, face detection is also a prerequisite for tasks such as facial identify recognition, facial attribute analysis, face editing, and face tracking, and its performance directly affects the effectiveness of these tasks [1, 2]. Therefore, whether it is to satisfy the user's experience or serve the postprocessing steps, higher requirements are placed on the accuracy and real-time video face detection.

Benefiting from the development of generic object detection methods, face detection has also made significant progress. The idea of early face detection methods was to first extract hand-crafted features from sliding windows on the image and then feed them into a classifier to detect possible face regions. One of the most iconic works is Histogram of Oriented Gradients (HOG) followed by SVM [3]. After all, the accuracy of such methods is limited. With the improvement of computing and storage capabilities, deep learning-based face detection methods have surpassed traditional methods in terms of speed, accuracy, and portability.

Existing object detectors can be broadly divided into two categories: two-stage and one-stage methods. For two-stage detectors, R-CNN series [4–7] generate object proposals in the first stage for classification as well as bounding box refinement in the second stage. In particular, the Faster R-CNN [6] architecture uses a regional proposal network (RPN) rather than the selective search method to propose bounding boxes, making object detection much faster. Mask R-CNN [7] can generate high-quality segmentation masks for each instance while efficiently performing object detection. Unlike two-stage proposal-classification detectors, YOLO [8] (you look only once) is a one-step regression method proposed by Redmon et al., whose main contribution is real-time detection of full images and webcams. The YOLO pipeline first divides the input image into $S \times S$ non-overlapping grid cells, then each cell is responsible for detecting those objects whose center points fall within that cell. YOLO network runs at 45 frames per second with no batch processing on a Titan X GPU as compared to Faster R-CNN at 7 fps. However, the experiments showed that YOLO was not good at accurate localization. Soon, several follow-up works [9–14] adopt a series of design decisions from past works with novel concepts to improve YOLOs' speed and precision. For instance, the mature detection framework YOLOv4 [14] uses the architecture of

CSPDarkNet53 with an SPP layer as backbone, PANet as Neck and YOLO detection head. The detection of occluded targets, small targets, etc. has been significantly improved. Another classic one-stage detection framework, SSD [15], introduces a detection method based on pyramidal feature hierarchy to predict objects on feature maps of different receptive fields. Furthermore, CornerNet [16] directly detects an object bounding box as pairs of keypoints, i.e. top-left corner and the bottom-right corner, which triggers the emergence of series of anchor-free detection methods [17, 18].

As well, there are also algorithms [19–24] that are specially tailored for face detection. MTCNN [19] cascades the three networks of P-Net, R-Net, and O-Net, which can simultaneously detect faces and five facial landmarks. Also, Cai at el. [20] still adopt a multitask cascaded CNN-based framework for simultaneous face detection, dense face alignment and fine head pose estimation. The lightweight anchor-free face detector CenterFace [21] can run in real time on a single CPU core and 200 fps using NVIDIA 2080TI for VGA-resolution images but produces more false positives. RetinaFace [22], a generalized face localization method, its architecture consists of three main parts: feature pyramid network, context module, and cascade regression. Also, it utilizes a multi-task learning strategy that combines extra-supervision and self-supervision to achieve stable face detection, accurate 2D face alignment, and robust 3D face reconstruction.

In this paper, we focus on improving a state-of-the-art object detection method to make it suitable for the real-world video face detection task. More and more scenarios in today's society involve video face detection, but there are still some inherent challenges in the development of this technology. On the one hand, not only high precision but also real-time speed is required in video processing. On the other hand, there is a lot of redundant information between adjacent frames in the video data. If all detected faces are output, there will be multiple face images of the same person, resulting in a lot of repetitive work in postprocessing stage. So face deduplication at the detection level is also necessary. In summary, our key contributions are

(i) Due to smaller variations in the facial boxes' aspect ratio, CARLoss (constrained aspect ratio loss) is proposed as a new feedback mechanism.

(ii) Adding a prediction head on the original YOLOv4 for tiny faces, so the modified FH-YOLOv4 is more suitable for large-scale variations.

(iii) A video face image deduplication algorithm based on IoU Tracker is proposed to serve postprocessing tasks.

The remainder of this paper is organized as follows: Section 2 reviews several popular loss functions for bounding box regression. In Section 3, we propose CAR-Losses and four-head architecture FH-YOLOv4 for better face detection. In Section 4, we propose a video face image deduplication algorithm based on IoU Tracker. Extensive experiments are conducted in Section 5. Section 6 presents our conclusions.

2. Analysis of Traditional Bounding Box Regression Loss Functions

Generally, the loss function of the object detection task consists of two components, classification loss and bounding box regression loss. In this section, we focus on the evolution of bounding box regression loss, and analyze several of the most representative loss functions. There are many formats for the labeling of the bounding box, the common ones include Pascal VOC, COCO, and YOLO, which can be converted into each other. We follow the label format of the YOLO dataset and the box is parameterized by the coordinate of its center point, the width and the height. For convenience, we denote the region proposal and ground truth as $B = (x, y, w, h)$ and $B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$, respectively.

2.1. l_n-Norm Loss. The l_n-norm loss functions are widely employed in bounding box regression. l_1 loss function stands for least absolute deviations, and l_2 loss function stands for least square errors. When the gradient descent algorithm is applied, the derivative function of l_1 loss is piecewise constant, so l_1 loss is insensitive to outliers but tends to fluctuate near the stable value in the later training period. l_2 loss is continuously differentiable, but due to its amplification effect on outliers, it is easy to cause gradient explosion in the early training stage. l_1-smooth loss [5] is exactly the integration of the advantages of l_1 loss and l_2 loss. However, the general representation of the location loss based on the l_n-norm is as follows:

$$L_{l_n} = \sum_{i \in x,y,w,h} l_n\left(B_i^{gt} - B_i\right), \tag{1}$$

which ignores the correlation between the four parameters of the bounding box.

2.2. IoU and GIoU Loss. Intersection over union (IoU) of B^{gt} and B,

$$IoU = \frac{\left|B^{gt} \cap B\right|}{\left|B^{gt} \cup B\right|}. \tag{2}$$

The evaluation criterion of positioning accuracy is used to de-redundant region proposals or determine positive and negative samples. In [25], for the first time, it was used as a measure of the distance between the candidate box and the ground truth to construct the loss function.

$$\begin{aligned} L_{IoU} &= -\ln IoU, \\ L_{IoU} &= 1 - IoU. \end{aligned} \tag{3}$$

IoU loss not only treats the bounding box as a unit, but the metric is also scale-invariant.

If the anchor box and the target box have no overlapping area, IoU loss can neither reflect how far apart the two boxes are nor guide the movement of the anchor box. To address this issue, GIoU loss is proposed in [26],

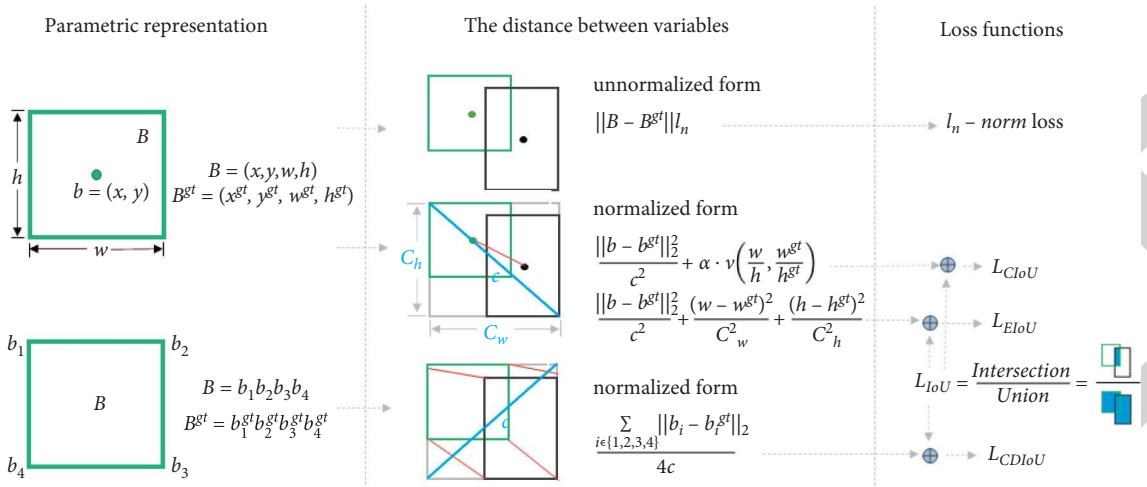

Figure 1: Illustrations of several loss functions that take into account the overlap area of two boxes and the normalized distance between parameters. Green denotes the anchor box. Black denotes the target box. Gray denotes the smallest enclosing box covering the two boxes.

$$GIoU = IoU - \frac{\left|E \setminus \left(B^{gt} \cup B\right)\right|}{|E|},$$
$$L_{GIoU} = 1 - GIoU, \tag{4}$$

where E is the smallest rectangular enclosing both B^{gt} and B. Empirically, this generalization tends to first increase the size of the proposal box to make it overlaps with the target box, and then degenerate into the IoU evaluation-feedback mechanism [27]. There are still problems such as slow convergence speed and inaccurate alignment. Similar to GIoU loss, DIoU loss ([28]) can still provide the direction of movement for the bounding box when it does not overlap with the target box.

2.3. Loss Functions That Consider Both IoU and Parametric Representation. As mentioned above, on the one hand, each box can be uniquely represented by a set of variables. For example, if the coordinates of two nonadjacent vertices are known, or the coordinates of the centroid and the width and height are given, then the rectangle can be positioned and drawn. On the other hand, IoU is an important indicator to judge the similarity of two boxes. Therefore, various positioning losses that take into account the overlap area of the two boxes and the normalized distance between the parameters are emerging one after another as illustrated in Figure 1.

In [27], the Complete-IoU (CIoU) loss is proposed,

$$L_{CIoU} = 1 - IoU + \frac{\left\|\mathbf{b} - \mathbf{b}^{gt}\right\|_2^2}{c^2} + \alpha \cdot v\left(\frac{w}{h}, \frac{w^{gt}}{h^{gt}}\right), \tag{5}$$

where $\mathbf{b} = (x, y)$ and $\mathbf{b}^{gt} = (x^{gt}, y^{gt})$ are the centroids of B and B^{gt}, respectively, c is the diagonal length of the smallest enclosing rectangle covering the two boxes, $v = 2(\arctan(w/h) - \arctan(w^{gt}/h^{gt}))^2/\pi^2$ measures the difference in the aspect ratio between the two boxes, and $\alpha = v/((1 - IoU) + v)$ is a control parameter. From formula (5), it is obvious that in the process of oriented boxes regression, the deep network first tries to pull the center point of the generating box towards the center point of the target box until the two boxes intersect, and then pay more attention to adjusting the aspect ratio later.

Efficient-IoU (EIoU) loss [29], a revised version of CIoU loss, directly minimizes the gap between the width and height of the two boxes instead of the aspect ratio. Its definition is as follows:

$$L_{EIoU} = 1 - IoU + \frac{\left\|\mathbf{b} - \mathbf{b}^{gt}\right\|_2^2}{c^2} + \frac{\left(w - w^{gt}\right)^2}{C_w^2} + \frac{\left(h - h^{gt}\right)^2}{C_h^2}, \tag{6}$$

where C_w and C_h are the width and height of the smallest enclosing box covering the two boxes. In addition, Control Distance-IoU (CDIoU) loss [30] considers the regression of the four vertices of the box. Starting from the upper left point of the rectangle, denote the four vertices of B and B^{gt} clockwise as $\mathbf{b}_i$ and $\mathbf{b}_i^{gt}$ $(i = 1, 2, 3, 4)$. The CDIoU loss is defined as follows:

$$L_{C\ DI\ oU} = 1 - IoU - \lambda\left(1 - \frac{\sum_{i\in\{1,2,3,4\}}\left\|\mathbf{b}_i - \mathbf{b}_i^{gt}\right\|_2}{4c}\right). \tag{7}$$

Compared with the previous loss functions, they have greatly improved the convergence speed and detection accuracy. Since the physical description of the distance between boxes is diverse, there is still a lot of room for optimization.

3. FH-YOLOv4 with Constrained Aspect Ratio Loss for Better Face Detection

As a special case of object detection, face detection is also featured by the limited aspect ratio of the facial box (ranging from 1 : 1 to 1 : 1.5) and large-scale variation (occupying several pixels or even thousands of pixels) [22]. These properties open up opportunities for us to adjust the loss

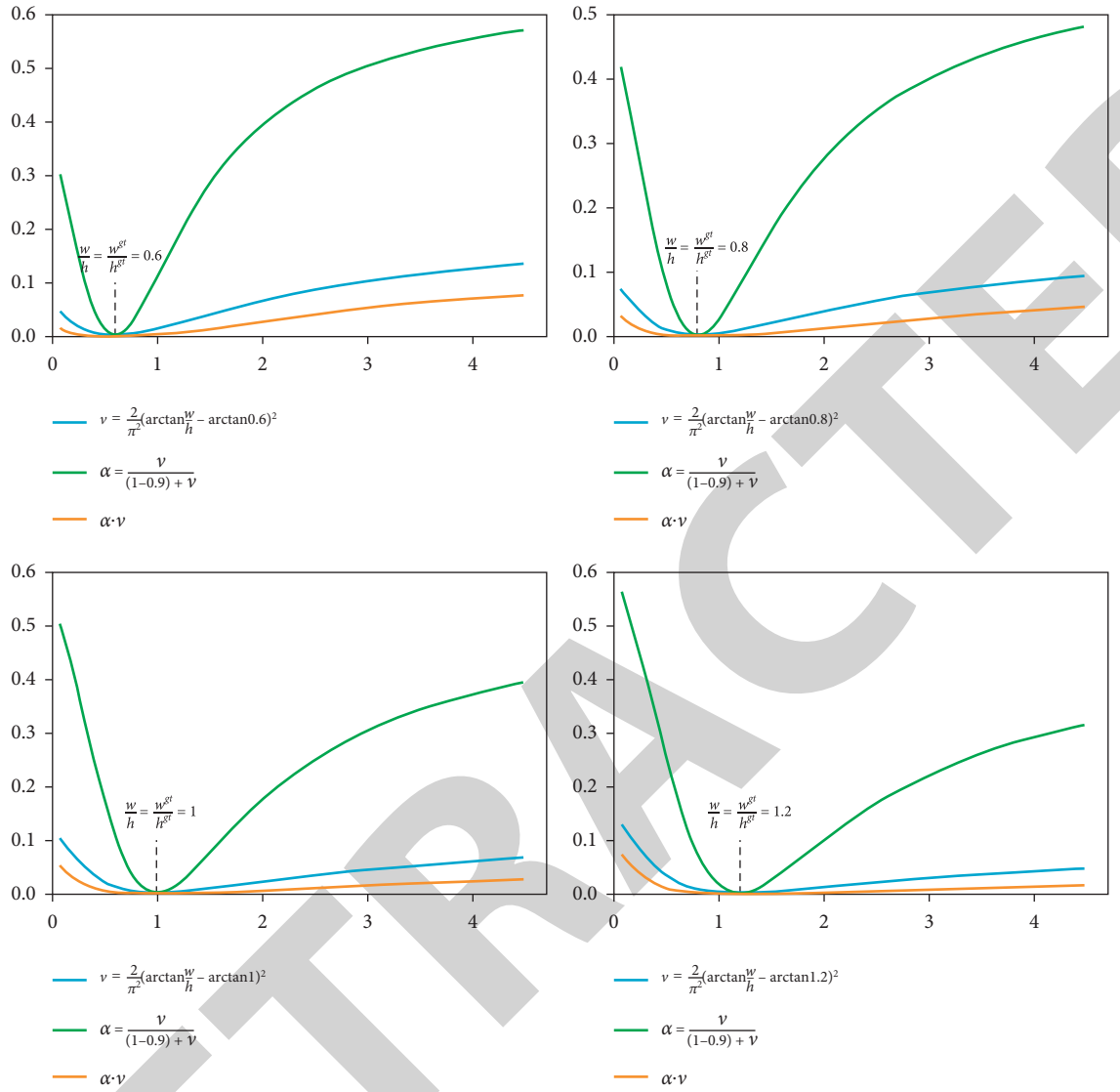

Figure 2: Exploration of the range of I_3 in CIoU loss. The x-axis label is w/h. Set the facial box width-to-height ratio w^{gt}/h^{gt} to 0.6, 0.8, 1.0, 1.2 in sequence, and fix IoU = 0.9. Then, the graphs of functions v, α, and $I_3 = \alpha \cdot v$ are shown.

function and network structure of advanced general object detection methods to yield more promising and faster facial box inference.

3.1. Constrained Aspect Ratio Losses for Better Facial Box Regressing

3.1.1. Limitation of CIoU Loss for Face Detection. Reviewing the definition of CIoU loss in formula (5), we might as well write its three key components as $I_1 = 1 - IoU$, $I_2 = \|\mathbf{b} - \mathbf{b}^{gt}\|_2^2/c^2$, and $I_3 = \alpha \cdot v(w/h, w^{gt}/h^{gt})$, where $I_1 \in [0,1]$, $I_2 \in [0,1)$. And then we analyze the range of I_3. Typically, the width-to-height ratio (of the facial box B^{gt}) $w^{gt}/h^{gt} \in [2/3, 1]$. In Figure 2, the graph of the function v is drawn in blue for the independent variable $w/h \in [0.08, 4.5]$ when w^{gt}/h^{gt} is fixed at 0.6, 0.8, 1.0, and 1.2, respectively. In addition, since the coefficient function α is monotonically increasing with respect to IoU, a larger IoU (fixed IoU = 0.9) is selected here to explore the upper bound of I_3, and the graph of α is plotted in green. So far, we can obtain the graph of $I_3 = \alpha \cdot v$ (orange curve) and find that the range of I_3 is about [0,0.1) in the process of facial box regression. Therefore, compared with I_1 and I_2, the contribution of I_3 to CIoU loss is very small.

3.1.2. Constrained Aspect Ratio Loss. To enhance the contribution of the shape parameter w/h to the facial box regression loss, we revise the CIoU loss and propose a series of more efficient versions, i.e., constrained aspect ratio losses (CARLosses). For the sake of brevity, CARLosses can be unified as expression (10):

$$L_{R_i}\left(B, B^{gt}\right) = 1 - IoU + \frac{\left\|\mathbf{b} - \mathbf{b}^{gt}\right\|_2^2}{c^2} + R_i\left(\frac{w}{h}, \frac{w^{gt}}{h^{gt}}\right). \quad (8)$$

Here, in view of the constraint that the aspect ratio of the facial box is limited, several penalty terms

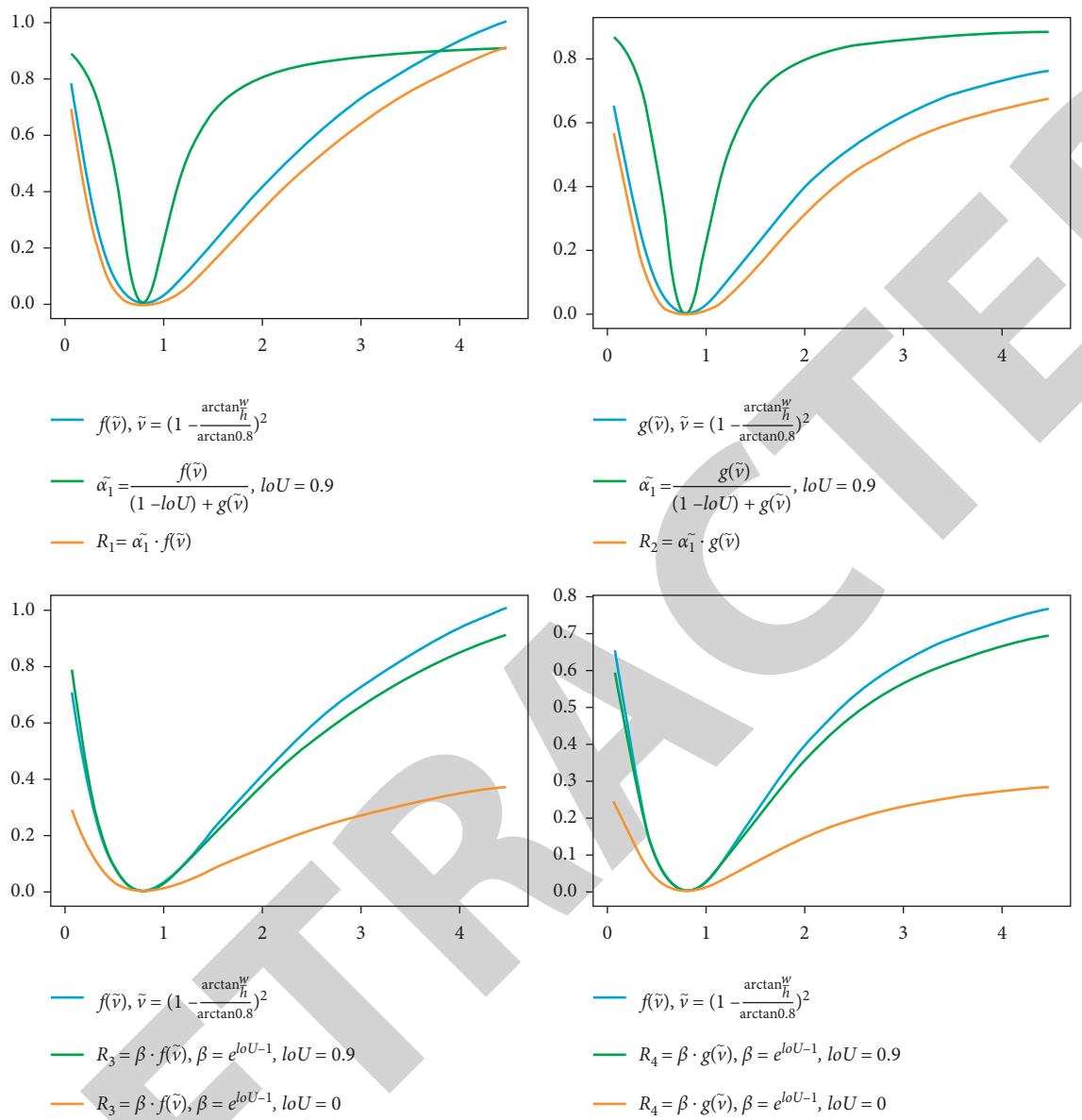

Figure 3: Description of the range of R_i $(i = 1, 2, 3, 4)$ in CARLosses. Set the facial box width-to-height ratio w^{gt}/h^{gt} to 0.8. The x-axis label is w/h.

$R_i(w/h, w^{gt}/h^{gt})$ $(i = 1, 2, 3, 4)$ are designed to improve the convergence speed and positioning accuracy.

Specifically, we define the function $\tilde{v} = (1 - \arctan(w/h)/\arctan(w^{gt}/h^{gt}))^2$, and introduce the piecewise function $f(\cdot)$ and the hyperbolic tangent function $g(\cdot)$:

$$f(x) = \begin{cases} 0, & \text{if } x \le 0 \\ x, & \text{if } 0 < x < 1, g(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}. \\ 1, & \text{if } x \ge 1 \end{cases} \tag{9}$$

Then, the following loss functions L_{R_i} are proposed.

Case 1.

$$L_{R_1} = 1 - IoU + \frac{\left\|\mathbf{b} - \mathbf{b}^{gt}\right\|_2^2}{c^2} + \tilde{\alpha_1} f(\tilde{v}), \tag{10}$$

where the trade-off parameter $\tilde{\alpha_1} = f(\tilde{v})/((1 - IoU) + f(\tilde{v}))$.

Case 2.

$$L_{R_2} = 1 - IoU + \frac{\left\|\mathbf{b} - \mathbf{b}^{gt}\right\|_2^2}{c^2} + \tilde{\alpha_2} g(\tilde{v}), \tag{11}$$

where the balance parameter $\tilde{\alpha_2} = g(\tilde{v})/((1 - IoU) + g(\tilde{v}))$.

Case 3.

$$L_{R_3} = 1 - IoU + \frac{\left\|\mathbf{b} - \mathbf{b}^{gt}\right\|_2^2}{c^2} + \beta f(\tilde{v}), \tag{12}$$

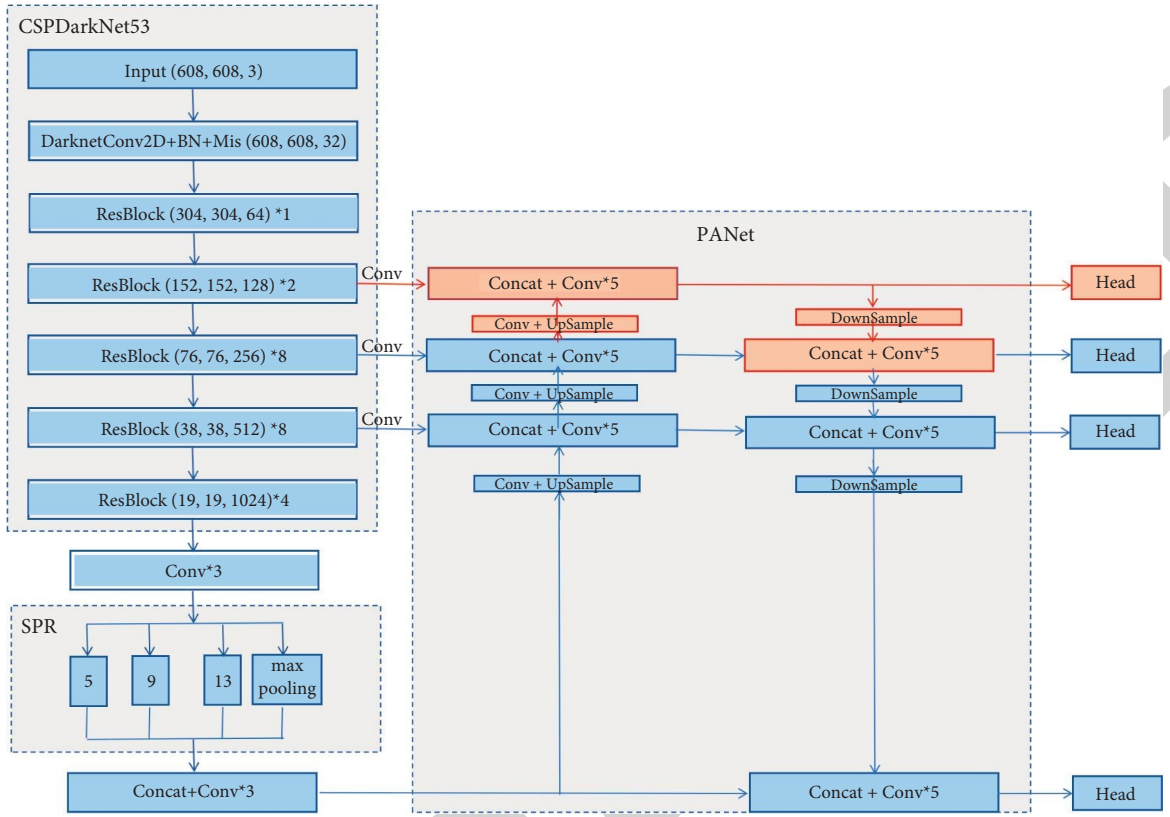

Figure 4: The architecture of the FH-YOLOv4. Compared with the original YOLOv4 structure, one more head (Red) is introduced for tiny face detection.

where the weighting coefficient $\beta = e^{IoU-1}$.

Case 4.

$$L_{R_4} = 1 - IoU + \frac{\left\|\mathbf{b} - \mathbf{b}^{gt}\right\|_2^2}{c^2} + \beta g(\vec{v}), \tag{13}$$

where the coefficient β is the same as the one in expression (12).

First, rough visualization of components R_i in L_{R_i} shows that the contribution of the shape parameter w/h to facial box regression loss is indeed greatly increased, see Figure 3. Furthermore, the performance evaluation of the proposed CARLosses on face detection is presented in Section 5.

3.2. Improved YOLOv4 by Adding a Prediction Head for Tiny Faces

3.2.1. Four-Head Structure (FH-YOLOv4) for Large-Scale Variations. Since the facial boxes in the WIDER FACE dataset vary dramatically in scale (from a few pixels to tens of thousands of pixels), and almost half of them are small instances (occupying less than 200 pixels), we add on the original YOLOv4 framework a prediction head to facilitates correct detection of tiny faces. As shown in Figure 4, the prediction head (the red branch) we add is generated from a low-level feature map with a small receptive field, which is more sensitive to tiny faces. Therefore, FH-YOLOv4 contains a total of 4 detection heads, which are used to detect tiny, small, medium and large faces, respectively.

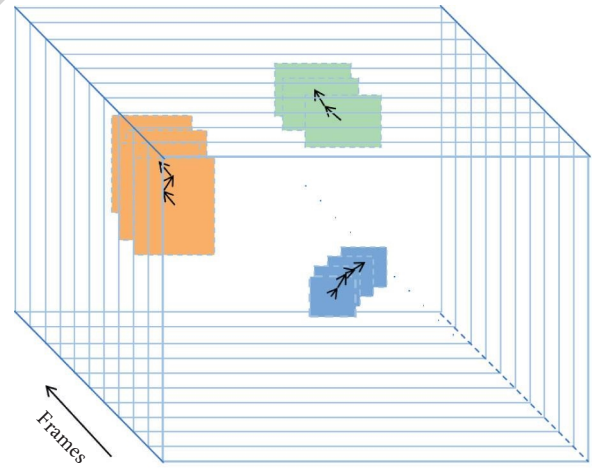

Figure 5: Illustration of the facial boxes in consecutive frames. Blue, Orange, and Green represent the facial boxes of different people, respectively.

3.2.2. Predefined Anchors. We still use k-means clustering to mine the facial bounding box priors. Cluster the width and height in the annotation information of the benchmark dataset WIDER FACE [31] for face detection. In the case

Input: $V = \{v_i\}_{i=1,2,\ldots,F}$, T, p
Initialize: flag = 0, temp = ∅, temp_loc = ∅
For v_i **in** V:
  if flag == T:
    Save the faces stored in the temp;
    flag = 0, temp = ∅, temp_loc = ∅;
  $D_i = \emptyset$, $L_i = \emptyset$;
  Perform face detection on image v_i,
  add the faces $\{d_{ij}\}_{j\in\mathbb{N}}$ to D_i,
  add the corresponding location $\{l_{ij}\}_{j\in\mathbb{N}}$ to L_i;
  if $D_i \neq \emptyset$:
    if temp == ∅:
      Add $\{d_{i,j}\}$ to temp, add $\{l_{i,j}\}$ to temp_loc;
      flag+ = 1, continue;
    for l_{ij} in L_i:
      [max_loc, max_IoU] = max IoU (l_{ij}, temp_loc);
      if max_$IoU < p$:
        Add $d_{i,j}$ to temp, add $l_{i,j}$ to temp_loc.
      else:
        Replace max_loc with $l_{i,j}$;
  flag+ = 1;
Output: The faces saved after deduplication operation.

Algorithm 1: The IoU tracker-based video face deduplication algorithm.

Table 1: Quantitative comparison of traditional YOLOv4 trained using L_{CIoU} (baseline) and proposed L_{R_i}.

	L_{CIoU}	L_{R_1}	L_{R_2}	L_{R_3}	L_{R_4}
Epoch 50th	66.14%	65.39%	66.46%	66.79%	66.58%
Epoch 100th	72.91%	72.07%	72.95%	73.30%	73.63%
Epoch 150th	74.08%	74.01%	74.55%	73.96%	74.26%
Max AP	74.55%	74.28%	**74.63%**	74.10%	**74.57%**

where the size of the input image is 608 × 608, we employ 12-mean clustering to get predefined anchors' size. The 12 clusters centroids are (2.36,3.96), (3.56,6.33), (4.75,8.75), (5.34,12.67), (6.53,10.68), (8.31,15.11), (10.69,20.77), (15.44,28.49), (22.56,40.95), (34.44,59.09), (59.38,97.92), and (139.53,208.18). Also, the prior settings in YOLOv4 are adjusted accordingly to make it more suitable for face detection tasks.

4. The Application of FH-YOLOv4 for Video Face Detection: IoU Tracker-Based Face Image Deduplication

Because the detected faces may be within the range of video surveillance for a long time, a large number of repeated detection faces are filled between consecutive frames. It is obviously unreasonable to do a follow-up 1: N face recognition for all repeating faces. So, in this subsection, we study the face deduplication algorithm, which deduplicates the detected faces to reduce the number of recognitions of the same person.

Inspired by the idea of IoU Tracker [32] in target tracking, we simplify and apply it to video face deduplication tasks. Different from multiobjective tracking [33], the deduplication algorithm proposed in this paper does not need to store a large amount of historical information, which reduces the storage cost.

When the time interval between frames is short and the person moves slowly, analyzing the position of all facial boxes in the current frame and the previous frame will find that the facial boxes of the same person will partially overlap (please see Figure 5). Thus, we can use the IoU as a measurement indicator for face deduplication. A detailed description of the IoU Tracker-based video face deduplication algorithm is shown in Algorithm 1, where V represents the test video, a continuous image sequence containing F frames in total. Period T and IoU threshold p need to be set empirically, flag is a counter, temp, and temp_loc are used to store faces and their corresponding locations, respectively. The detections at frame i are recorded in D_i and L_i, d_{ij} is the j^{th} face at frame i, l_{ij} is the location of d_{ij}. IoU (l_{ij}, temp_loc) means calculating the IoU between l_{ij} and all the facial boxes in the temp_loc, and the element in the temp_loc that makes the max_IoU valid is assigned to max_loc.

Because no visual information about the frames is used, the overall complexity of the method is very low. So, it can be thought of as a simple filtering process at the detection level.

5. Experiments

5.1. Effect of CARLosses. To evaluate the performance of the proposed CARLosses on face detection, we train the traditional YOLOv4 using L_{CIoU} (baseline) and proposed L_{R_i} ($i = 1, 2, 3, 4$) on the WIDER FACE training set. The input image size is set to be 608 × 608. Also, the AP at epochs 50th, 100th, 150th and maximum AP on the WIDER FACE validation set are recorded in Table 1. The results show that

Table 2: The performance when incorporating the CIoU loss and CARLosses with FH-YOLOv4.

Methods	YOLOv4+ L_{CIoU}	FH-YOLOv4+ L_{CIoU}	FH-YOLOv4+ L_{R_1}	FH-YOLOv4+ L_{R_2}	FH-YOLOv4+ L_{R_3}	FPS+ L_{R_4}
AP	74.55%	78.89%	78.91%	78.38%	79.06%	79.20%

Table 3: The performance of FH-YOLOv4 on test videos.

Test videos	Frames	Video length	Detection time	Total run time	FPS
video1.mp4	7783	00:05:18	00:05:55	00:07:27	21.9
video2.mp4	6283	00:04:18	00:04:40	00:05:56	22.4
video3.mp4	7783	00:05:18	00:06:21	00:07:55	20.4
video4.mp4	7783	00:05:18	00:06:06	00:07:38	21.3
video5.mp4	7783	00:05:18	00:05:58	00:07:33	21.7
video6.mp4	7783	00:05:18	00:05:57	00:07:34	21.8
video7.mp4	7783	00:05:18	00:03:59	00:04:19	32.6

Table 4: Effectiveness of the IoU tracker-based face deduplication algorithm.

Test videos	Detected faces	Time for detection	Reserved faces	Detection + deduplication	Deduplication ratio (%)
video1.mp4	2660	00:00:51	59	00:00:54	97.782
video2.mp4	4276	00:00:49	86	00:00:52	97.989
video3.mp4	672	00:00:50	48	00:00:52	92.857
video4.mp4	1750	00:00:51	65	00:00:53	96.286
video5.mp4	1573	00:00:51	56	00:00:54	96.440

training YOLOv4 using our L_{R_2} and L_{R_4} can not only promote the fast convergence of the model but also improve its performance compared to CIoU loss.

5.2. Ablation Studies. Aims to verify the effectiveness of FH-YOLOv4 and CARLoss in face detection, some experiments are carried out, and the results are recorded in Table 2.

First, cooperating the CIoU loss (L_{CIoU}) with traditional YOLOv4 and FH-YOLOv4, respectively, one can find that the added prediction head for tiny faces brings an astonishing performance gain of 4.34%. Second, when FH-YOLOv4 is trained with our proposed CARLosses, L_{R_1}, L_{R_3}, and L_{R_4} can all improve its performance compared to CIoU loss. It is worth noting that, the results in Tables 1 and 2 show consistent improvements in the performance of YOLOv4 and FH-YOLOv4 when they are trained using L_{R_4}. Thus, in the following text, we will abbreviate the proposed method FH-YOLOv4+ L_{R_4} as FH-YOLOv4.

5.3. Video Face Detection Speed. We apply FH-YOLOv4 to perform face detection on 7 test videos. The basic information of the videos (number of frames, duration), the time spent in the detection process, and the total running time are shown in Table 3. Among them, the total running time refers to the time consumption of all links including the reading of video frames, face detection, and result saving. Also, FPS refers to the number of video frames detected per second.

With the exception of video7.mp4, the FSP of other videos is around 21, which does not show a considerable speed advantage. On the one hand, the newly added prediction head for tiny faces in FH-YOLOv4 increases the number of network layers, which inevitably leads to an increase in the amount of computation. However, it is worth sacrificing a small computational cost in exchange for a big boost to the AP. On the other hand, the above experiment is a frame-by-frame detection of video. In practical application scenarios, the restricted random sampling (RRS) method [34] is usually used to randomly sample the video frames first, and then only the sampled video frames are processed. This not only improves the efficiency of the program but also allows more time for subsequent face recognition tasks.

5.4. Face Deduplication. In the following experiments, the FH-YOLOv4 and IoU Tracker-based deduplication algorithms are combined to detect and deduplicate the faces in the video clips. Under the setting of the IoU threshold $p = 0.8$, the statistical results of the number of faces before and after deduplication are shown in Table 4. It can be seen that the IoU Tracker-based deduplication algorithm can effectively remove repeated faces without increasing a lot of computing time, which helps to relieve the pressure of subsequent face recognition.

6. Conclusion

In this paper, we first propose CARLosses for better facial boxes regression according to the fact that the width-to-height ratio of faces is roughly ranging from 1 : 1 and 1 : 1.5. Second, by clustering the width and height in the annotation information of the benchmark dataset WIDER FACE, we add a prediction head on the original YOLOv4 for tiny faces and obtain a modified network structure FH-YOLOv4. The

FH-YOLOv4 guided by CARLoss can achieve a significant improvement in AP compared to traditional YOLOv4. Third, we propose the IoU Tracker-based face image deduplication algorithm, and the deduplication rate is over 95 for all test videos. Experiments demonstrate that our method can achieve real-time speed and high accuracy, making it an ideal alternative for most face detection and recognition applications.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partially supported by the National Social Science Fund Project Research on the Mechanism and Countermeasures of the List of Powers to Promote the Modernization of Local Government Governance (15BZZ043); Practical Research on the Construction of Local Government Power List System to Promote Governance Modernization (17ZZE430); and Heilongjiang Province Philosophy and Social Science Research Planning Project Research on Development Management and Transformation Countermeasures of Energy Industry in Heilongjiang Province (16GLE03).

References

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics Recognition Using Deep Learning: A Survey," 2019, https://arxiv.org/abs/1912.00271.

[2] S. Minaee, P. Luo, Z. Lin, and K. Bowyer, "Going Deeper into Face Detection: A Survey," 2021, https://arxiv.org/abs/2103.14983.

[3] N. Dalal and B. Trings, "Histograms of oriented gradients for human detection 2005," in *Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR)*, pp. 886–893, San Diego, CA, USA, June 2005.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[5] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, December 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, Venice, Italy, October 2017.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[9] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[10] J. Redom and A. Farhadi, "Yolov3: An Incremental Improvement," 2018, https://arxiv.org/abs/1804.02767.

[11] C. Li, R. Wang, J. Fei, and L. Fei, "Face detection based on YOLOv3," *Recent Trends in Intelligent Computing Communication and Devices*, Springer, Berlin, Germany, pp. 277–284, 2020.

[12] S. H. Tuli, A. Liu, and W. Liu, "A novel face detector based on YOLOv3," in *Proceedings of the AI 2020: Advances in Artificial Intelligence*, pp. 55–68, Canberra,Australia, November 2020.

[13] L. Z. Chun, L. Dian, J. Y. Zhi, W. Zhang, and C. Zhang, "YOLOv3: face detection in complex environments," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 1153–1160, 2020.

[14] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal Speed and Accuracy of Object Detection," 2020, https://arxiv.org/abs/2004.10934.

[15] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot MultiBox detector," in *Proceedings of the Computer Vision-ECCV 2016*, pp. 21–37, Amsterdam, The Netherlands, October 2016.

[16] H. Law and J. Deng, "Cornernet: detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 734–750, Munich, Germany, September 2018.

[17] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: keypoint triplets for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6569–6578, Seoul, Republic of korea, November 2019.

[18] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, Seoul, Republic of korea, November 2019.

[19] K. Zhang, Z. Zhang, Z. Qiao, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 23, pp. 1499–1503, 2016.

[20] Z. Cai, Q. Liu, S. Wang, and B. Yang, "Joint head pose estimation with multi-task cascaded convolutional networks for face alignment," in *Proceedings of the 2018 24th International Conference on Pattern Recognition*, pp. 495–500, ICPR, Beijing, China, August 2018.

[21] Y. Xu, W. Yan, G. Yang, J. Luo, T. Li, and J. He, "CenterFace: Joint Face Detection and Alignment Using Face as point Scientific Programming," *Scientific Programming Towards a Smart World 2020*, vol. 2020, Article ID 7845384, 8 pages, 2020.

[22] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5203–5212, Seattle, WA, USA, June 2020.

[23] Y. Liu, X. Tang, X. Wu, J. Han, J. Liu, and E. Ding, "Hambox: Delving into Online High-Quality Anchors Mining for Detecting Outer Faces," 2019, https://arxiv.org/abs/1912.09231.

[24] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "Tinaface: Strong but Simple Baseline for Face Detection," 2020, https://arxiv.org/abs/2011.13183.

[25] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: an advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520, New York, NY, USA, October 2016.

[26] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, Beach, CA, USA, June 2019.

[27] Z. Zheng, P. Wang, D. Ren et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8574–8586, 2022.

[28] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ren, and D. Ren, "Distance-IoU loss: faster and better learning for bounding box regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, Article ID 12993, 2020.

[29] Y. F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and Efficient IOU Loss for Accurate Bounding Box Regression," 2021, https://arxiv.org/abs/2101.08158.

[30] D. Chen and D. Miao, "Control Distance IoU and Control Distance IoU Loss Function for Better Bounding Box Regression," 2021, https://arxiv.org/abs/2103.11696.

[31] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: a face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, Las Vegas, NV, USA, June 2016.

[32] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–6, AVSS, Lecce, Italy, August 2017.

[33] E. Bochinski, T. Senst, and T. Sikora, "Extending IOU based multi-object tracking by visual information," in *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–6, AVSS, Auckland, New Zealand, November 2018.

[34] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," in *Proceedings of the Computer Vision - ECCV 2016*, pp. 20–36, Springer, Amsterdam, The Netherlands, October2016.