



A multivariate intersection over union of SiamRPN network for visual tracking

Zhihui Huang¹ · Huimin Zhao¹ · Jin Zhan¹ · Huakang Li¹

Accepted: 23 April 2021 / Published online: 13 May 2021
© The Author(s) 2021

Abstract

SiamPRN algorithm performs well in visual tracking, but it is easy to drift under occlusion and fast motion scenes because it uses ℓ_1 -smooth loss function to measure the regression location of bounding box. In this paper, we propose a multivariate intersection over union (MIOU) loss in SiamRPN tracking framework. Firstly, MIOU loss includes three geometric factors in regression: the overlap area ratio, the center distance ratio, and the aspect ratio, which can better reflect the coincidence degree of target box and prediction box. Secondly, we improve the definition of aspect ratio loss to avoid gradient explosion, improve the optimization performance of prediction box. Finally, based on SiamPRN tracker, we compared the tracking performance of ℓ_1 -smooth loss, IOU loss, GIOU loss, DIOU loss, and MIOU loss. Experimental results show that the MIOU loss has better target location regression than other loss functions on the OTB2015 and VOT2016 benchmark, especially for the challenges of occlusion, illumination change and fast motion.

Keywords Visual tracking · Multivariate intersection over union · Scale invariance · SiamRPN

1 Introduction

Visual target tracking is a subtask of computer vision, and many advanced methods have been explored in this research area. It has numerous applications in many domains, including visual navigation, intelligent video surveillance system, intelligent human–computer interaction, medical diagnosis. Deep learning demonstrates powerfulness in extracting and processing semantic features, and can model the appearance of object by learning multimedia information. Inspired by this, many successful applications of deep learning have been achieved in computer vision, such as image segmentation, object detection, image classification, target tracking, and so on.

Since 2013, the deep learning framework represented by SAE (stack auto-encoding) [1,2], CNN (convolution neural network) [3,4] and Siamese [5,6] has become the main backbone network of tracking algorithm. Deep learning has been showing great success on object tracking. DLT [1] for the first time introduced deep network to break the bottleneck of traditional tracking model. After that, CNN has been brought to enhance the target learning capability of tracker [3], owing to the invariance principle in nonlinear changes such as translation, scale change and rotation. With the continuous research of depth structure, Tao et al. [5] successfully applied Siamese network as the backbone network of tracking algorithm and made greatly progress in speed. However, it is weakness in many practical applications due to challenges such as illumination changes, partial occlusion, motion blur and low resolution, which obstruct the robust of tracking model. In recent years, the optimization trend of visual target tracking focuses on deepening neural network and improving feature extraction strategy, but ignoring the key role of loss function in model optimization. In computer vision such as target detection, recognition and semantic segmentation, the loss function can measure the performance of training model by comparing the difference between predictive value and actual data.

✉ Jin Zhan
gszhanjin@gpnu.edu.cn
Zhihui Huang
zhihuihuanggd@foxmail.com
Huimin Zhao
zhaohuimin@gpnu.edu.cn
Huakang Li
lihuakang2020@163.com

¹ School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

In this paper, we take advantage of the recent progress in bounding box (BBox) regression loss and to propose a novel multivariate intersection over union (MIOU) loss in SiamRPN [7] tracking framework. The proposed method can deal with the non-overlapping case between target box and prediction box, and speed up the convergence rate of the training model. In summary, this work has the following steps. Firstly, MIOU regression includes three important geometric factors in BBox regression: overlapping area ratio, center distance ratio and the aspect ratio, which can better reflect the coincidence degree of the target box and prediction box. Secondly, we improve the definition of aspect ratio loss to avoid gradient explosion and improve the optimization performance of prediction box. Finally, extensive experiments on OTB2015 and VOT2016 benchmark are carried out to validate our method effectiveness.

2 Related work

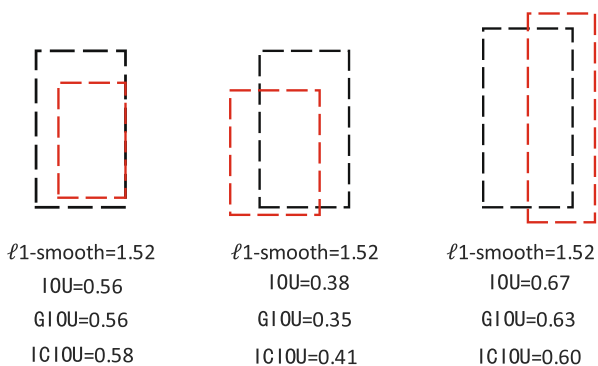
It is difficult to design a tracking model with both strong robustness and high precision. Therefore, many theoretical methods have been introduced to solve the tracking problem, such as classifier [8–10], sparse representation [11–13], saliency detection [14,15], feature selection [16–19] and deep learning [20,21]. Based on off-line training and online fine-tuning, prior depth trackers achieve better results than traditional methods, and the online fine-tuning timely adjustment parameters to adapt the change of target better. However, despite the favorable performance of deep learning on object tracking, it is still limited by many difficulties, including insufficient training samples, the foreground-background class imbalance, and high computational complexity in terms of time and space. Therefore, online depth methods are hardly meeting the requirements of real-time tracking.

In recent years, Siamese network has been introduced to solve the tracking problem. As an end-to-end off-line training network, Siamese network learns the matching function from external data and finds the candidate patch matching the target in the subsequent frame search area. It can achieve real-time tracking without model updating or online fine-tuning. SiamFC [22] uses the Siamese structure and makes full convolution matching in the detection frame according to the template frame. The tracking speed reaches 86 fps, which has aroused widespread concern and accelerated the application of Siamese network in object tracking. In order to address the weakness of model robustness, SiamFC++ [23] proposed four guidelines: decomposition of classification and state estimation, non-ambiguous scoring, prior knowledge-free and estimation quality assessment, which effectively improved the generalization of the tracker. Graph convolutional tracking (GCT) [24] constructed a graph convolution tracking framework base on the Siamese structure, which acquired more sufficient and stable characteristic from

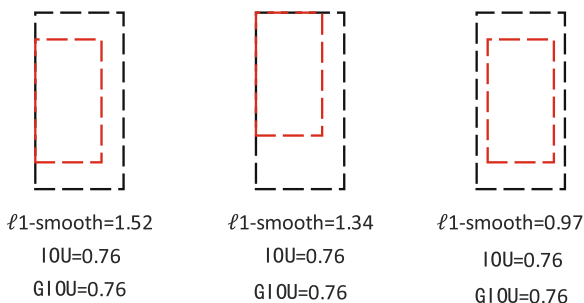
detection frame by combining the temporal and spatial context information, and the experimental results showed that the accuracy is improved greatly.

SiamRPN [7] tracker contains Siamese network and region proposal network (RPN). RPN subnetwork uses multi-dimensional features to quickly generate target recommendation area, and obtains K anchor points according to different preset aspect ratio. The introduction of RPN makes the network not affected by multi-scale regression calculation in target tracking, and improves the tracking speed and accuracy. However, SiamRPN is vulnerable to the case of object occlusion, background clutters and motion blur. SiamRPN++ [25] mainly improves the performance of feature extraction network, solves the problem that the network deepening destroys the translation invariance, and realizes the Siamese tracking driven by ResNet network. DaSiamRPN [26] generates semantic negative sample pairs in the training process and expands the training dataset to solve the problem of poor system recognition caused by unbalanced distribution of training data. A new interference awareness module is designed to capture targets by using context information and time information. The SiamMask [27] enhances loss monitoring by adding binary segmentation task, thus reducing the distance between target tracking and Vos (visual object segment). The trained learning model can achieve class independent object tracking and segmentation only depending on an initial boundary box. The deeper and wider SiamRPN [28] designs deeper and wider backbone network to improve the capability of Siamese tracker.

Although the trackers based on SiamRPN achieve good performance in many database evaluations, they use ℓ_1 -smooth [29] loss in location regression, which does not consider the correlation of the four corners of the bounding box, and multiple bounding boxes may have the same loss value. To alleviate the problem of class imbalance, Vital [30] adopts a high-order cost-sensitive loss to decrease the effect of easily negative samples successfully. Recently, the loss function of bounding box regression has been optimized. The n -normal form loss function represented by ℓ_1 -smooth [29] is very sensitive to the scale change of bounding boxes and cannot optimize the case of non-overlapping case, which is easy to cause the gradient to disappear. As shown in Fig. 1a, multiple detection boxes have the same ℓ_1 -smooth [29] loss value, but the IOU may vary greatly. In order to further improve the generalization performance of regression, scholars have proposed IOU loss [31] and GIOU loss [32]. When the prediction box and the target box do not intersect (non-overlapping), the IOU loss is 0. At this time, the loss function is not differentiable, so IOU loss cannot optimize the case of two boxes not intersecting. The GIOU loss can solve this problem, but because of the strong dependence on intersection over union, the convergence speed is slow. In reference [33], by directly minimizing the distance between the center points of two



(a) The comparison of three loss metrics under different bounding scales.



(b) The comparison of three loss metrics under the case when real bounding box surrounds the prediction box.

Fig. 1 Because $\ell_1\text{-smooth}$ is sensitive to the scale of bounding box, which cause failing to reflect the intersection information between real box (black) and prediction box (red) in the same value. Moreover, GIOU loss is transformed to IOU loss when real box (black) surrounds the prediction box (red), owing to the heavily relying on intersection over union (IOU)

bounding boxes, a distance intersection over union (DIOU) is proposed, which solves the problem of slow convergence. In addition, the authors also proposed the complete intersection (CIOU) loss of three important geometric variables: overlap area ratio, center distance ratio and aspect ratio. However, CIOU [33] uses the square of the angle difference of aspect ratio to measure the scale loss, so it has the problems of gradient explosion and non-co-directional optimization of the border, like $\frac{\partial \delta}{\partial w} = -\frac{h}{w} \frac{\partial \delta}{\partial h}$.

This paper analyzes the three factors that affect the location loss regression: the overlap area ratio, the center distance ratio and the aspect ratio of box height and width. We remove the square term of the angle difference corresponding to the aspect ratio, so as to avoid the gradient explosion problem and optimize the location regression performance better. The improved loss function (MIOU) is introduced into the regression branch of SiamRPN tracker, and achieves good performance.

The structure of the paper is as follows: firstly, the research background is sorted out in the introduction, and the related work is reviewed in the second part. Then, in the ‘‘Proposed method’’ section, we describe our method in detail, including the construction of network, the design of new geometric loss metrics, target class and target location. The experimental process and results are given in ‘‘Experimental results’’. Finally, in the conclusion and prospect part, the work of this paper is summarized and prospected.

3 Proposed method

The framework of SiamRPN tracker contains Siamese sub-network and region proposal subnetwork (RPN), where RPN network is constructed by two branches: classification loss and bounding box regression loss. We use ResNet50 [34] as the backbone instead of AlexNet in the original SiamRPN. In training, ResNet50 pays more attention to rich semantic information, breaks the space invariance limit of connected subnetwork, and helps the tracker to better adapt to the scene of target appearance change. In addition, we propose a multivariate intersection over union (MIOU) loss to replace the $\ell_1\text{-smooth}$ regression metric in the original RPN subnetwork and improve the tracking robustness.

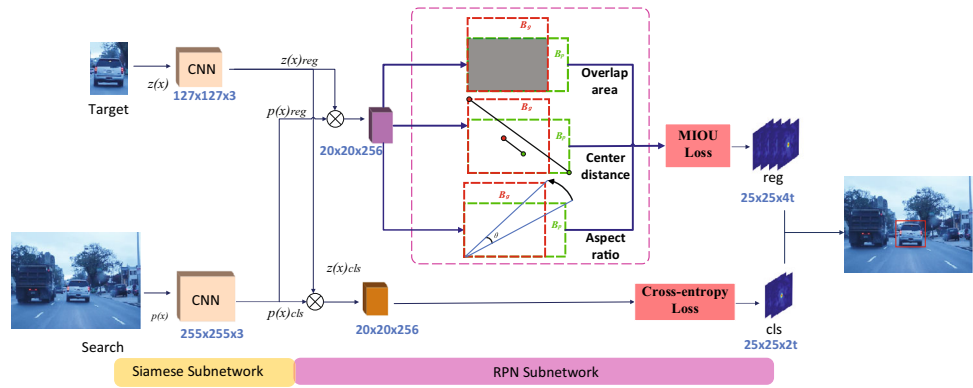
3.1 Network framework

As shown in Fig. 2, the target image $Z(x)$ and the search region $P(x)$ are input into two subnetworks of Siamese module, respectively. Meanwhile, they share the weights with the same structure during training. Considering the difference between classification and location, RPN is further divided into classification branch (*cls*) and regression branch (*reg*). In *cls* branch, we regard the classification problem as a qualitative output, and the regression problem is considered as a quantitative output in *reg* branch. So the outputs of Siamese subnetwork are fed into branch (*cls*) and (*reg*) individually. In detail, the classification branch convolutes $p(x)_{cls}$ with $z(x)_{cls}$ as convolution kernel, and the output channel number of $A_{w \times h \times 2t}^{cls}$ is $2t$, which indicates the positive and negative probability of candidate samples. Similarly, $z(x)_{reg}$ and $p(x)_{reg}$ produce the sensor $A_{w \times h \times 4t}^{reg}$ of $4t$ channels after correlation operation. We refer the regression result (d_x, d_y, d_w, d_h) as the four coordinates offsets of candidate targets. The specific operation process of the two tasks is as follows:

$$A_{w \times h \times 4t}^{cls} = z(x)_{cls} \otimes p(x)_{cls}, \tag{1}$$

$$A_{w \times h \times 4t}^{reg} = z(x)_{reg} \otimes p(x)_{reg}. \tag{2}$$

Fig. 2 Method network structure diagram



3.2 Classification loss

The classification loss of SiamRPN is cross-entropy (CE) loss. Cross-entropy method [35] is a unified method of reliability analysis and stochastic optimization design proposed by Rubinstein in 1997. Its essence is to transform the optimization problem into a small probability event estimation problem by using the optimal sampling probability density function instead of the original function of random variables based on Monte Carlo simulation. Cross-entropy can be directly used as the evaluation model of loss function, and the best training model is when the cross-entropy is minimum.

In the training, SiamRPN uses binary cross-entropy loss function for classification, assuming that the probability score of the i^{th} sample is p_i . The tag value is y_i ($y_i = 1$ means the sample is positive, otherwise, $y_i = 0$), and the calculation process is as follows:

$$L(y_i, p_i) = -(y_i \cdot \log(p_i)) + (1 - y_i)(1 - p_i) \tag{3}$$

If the total number of samples in class branch is N, then the classification loss is:

$$L_{cls} = \frac{1}{N} \sum_{i=1}^N L(y_i, p_i) \tag{4}$$

3.3 Multivariate intersection over union

There are three important geometric factors in border regression: overlap area, center distance and aspect ratio. DIOU [33] does not include aspect ratio factor, and the metric of aspect ratio in CIOU [33] loss measures the scale consistency by the square of the angle difference between the two bounding boxes, which is similar to the L_2 loss principle and has the problems of gradient explosion and instability. To solve the above problems, this paper redefines the loss formula to measure the aspect ratio, effectively avoids the gradient explosion, and improves the robustness.

In Fig. 3, let B_g and B_p represent the target box and prediction box, respectively, and the position of the box consisted of the coordinates of the two vertices in the lower left corner and the upper right corner, where $(\tilde{x}_1, \tilde{y}_1, \tilde{x}_2, \tilde{y}_2)$ denotes the position of B_g , and (x_1, y_1, x_2, y_2) is represented by B_p . In addition, b_g and b_p are the center points of B_g and B_p , and ρ represents the Euclidean distance of them. Noting that B_c denotes the smallest convex shape of B_g and B_p , c denotes the diagonal Euclidean distance of B_c .

The coordination $(X_{C1}, Y_{C1}, X_{C2}, Y_{C2})$ of B_c is as follows:

$$X_{C1} = \min(\tilde{x}_1, x_1), Y_{C1} = \min(\tilde{y}_1, y_1), \tag{5}$$

$$X_{C2} = \max(\tilde{x}_2, x_2), Y_{C2} = \max(\tilde{y}_2, y_2). \tag{6}$$

We use I to denote the maximum intersection box between B_g and B_p , and the coordinates $(X_{I1}, Y_{I1}, X_{I2}, Y_{I2})$ of I come from the following formula:

$$X_{I1} = \max(\tilde{x}_1, x_1), Y_{I1} = \max(\tilde{y}_1, y_1), \tag{7}$$

$$X_{I2} = \min(\tilde{x}_2, x_2), Y_{I2} = \min(\tilde{y}_2, y_2). \tag{8}$$

The IOU is the ratio of the intersection and union of the area of two rectangular boxes; the calculation process is as follows:

$$A_g = (\tilde{x}_2 - \tilde{x}_1) \times (\tilde{y}_2 - \tilde{y}_1), \tag{9}$$

$$A_p = (x_2 - x_1) \times (y_2 - y_1), \tag{10}$$

$$A_I = (X_{I2} - X_{I1}) \times (Y_{I2} - Y_{I1}), \tag{11}$$

$$A_u = A_g + A_p - A_I, \tag{12}$$

$$IOU = \frac{A_I}{A_u}, \tag{13}$$

where A_g and A_p are the area of B_g and B_p , respectively, A_I means the area of intersection between B_g and B_p , and A_u is the area formed by the union of two bounding boxes.

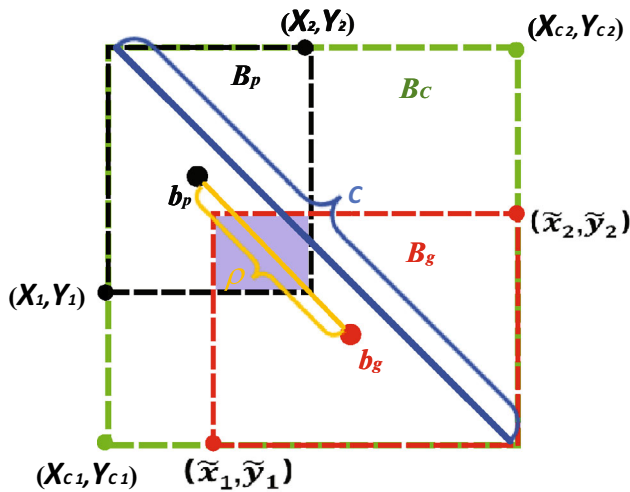


Fig. 3 Schematic diagram of target box and prediction box

The formula of center distance ratio of target box and prediction box is as follows:

$$R_{dis} = \frac{\rho^2}{c^2}, \tag{14}$$

where ρ is the Euclidean distance of the center point of two boxes, and c is the diagonal distance of the smallest external rectangle. R_{dis} is the penalty term of the center point distance.

In order to take the aspect ratio of prediction frame into account, we define θ reflects the difference of aspect ratio between B_g and B_p , as shown in Fig. 4. And θ_g denotes the inclination angle of the target box, while θ_p represents the prediction box inclination angle. Let $\theta_g = \arctan \frac{w_g}{h_g}$ and $\theta_p = \arctan \frac{w_p}{h_p}$, where w_g and h_g represent the width and height of the target box, w_p and h_p are taken from the prediction box. In order to achieve the aspect ratio alignment between B_g and B_p , we can see that $\theta < 0$ when $\theta_p < \theta_g$ in Eq.(15).

$$\theta = \theta_p - \theta_g \tag{15}$$

The formula of aspect ratio of target box and prediction box is as follows:

$$R_{asp} = \left(\frac{v}{1 - IOU + v}\right)\delta, \tag{16}$$

$$v = \frac{4\theta^2}{\pi^2}, \delta = \frac{8}{\pi^2} \times \theta \times (w_p \times h_p) \tag{17}$$

where δ is used to evaluate the aspect ratios alignment of the bounding box. When the value of δ is less than zero, it means prediction box B_p rotates counterclockwise during

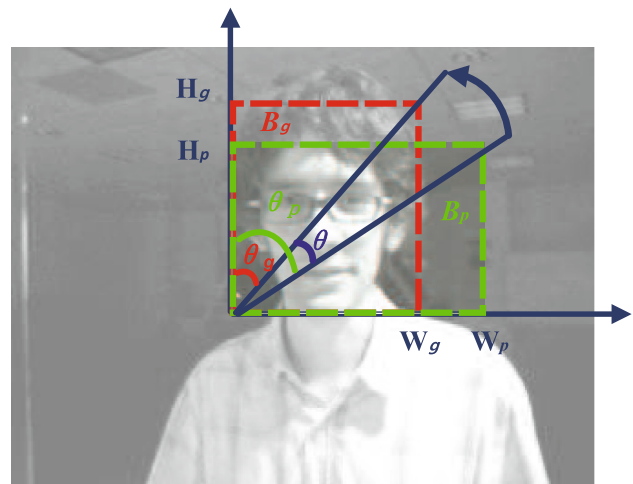


Fig. 4 Schematic diagram of optimization in bounding box shape

regression optimization. On the contrary, B_p rotates clockwise when $\theta > 0$, since $\theta_p > \theta_g$. This optimization process in bounding box shape can be visualized in Fig. 4.

In Eq. (17), δ is linearly related to angle difference θ and box area $w_p \times h_p$ to avoid gradient explosion. At the same time, in the process of scale optimization, there is no reverse relationship between the gradient $\frac{\partial \delta}{\partial w}$ and $\frac{\partial \delta}{\partial h}$. The gradient of δ w.r.t. w and h is as follows:

$$\frac{\partial \delta}{\partial w} = \frac{8}{\pi^2} \left(\arctan \frac{w_p}{h_p} - \arctan \frac{w_g}{h_g} - \frac{wh}{w^2 + h^2} \right) \tag{18}$$

$$\frac{\partial \delta}{\partial h} = \frac{8}{\pi^2} \left(\arctan \frac{w_p}{h_p} - \arctan \frac{w_g}{h_g} + \frac{wh}{w^2 + h^2} \right) \tag{19}$$

Then, our loss function based on multivariate intersection over union (MIOU) is defined as follows:

$$MIOU = IOU - R_{dis} - R_{asp} \tag{20}$$

The loss function of MIOU is defined as follows:

$$L_{MIOU} = 1 - IOU + R_{dis} + R_{asp} \tag{21}$$

$$= 1 - IOU + \frac{\rho^2}{c^2} + \left(\frac{v}{1 - IOU + v}\right)\delta \tag{22}$$

In the bounding box regression loss, the overlap area ratio and center distance ratio reflect the relative position relationship between the target box and the prediction box. According to these two loss functions, we can guide the regression of the prediction bounding box and accelerate the convergence speed in the training stage. In addition, the aspect ratio of boxes can avoid the invalid regression in the case of non-overlapping or the case that the target box completely contains the prediction box, which has good

Table 1 The distribution of 11 challenging attributes in the OTB2015 dataset: illumination variation (IV), scale variation (SV), occlusion (Occ), deformation (Def), motion blur (MB), in-plane rotation (IPR),

out-of-plane rotation (OPR), out of view (OV), background clutters (BC), low resolution (LR), and fast motion (FM)

	IV	SV	Occ	Def	MB	IPR	OPR	OV	BC	LR	FM
Sequences	38	65	49	44	29	51	64	14	31	9	39

Algorithm 1: Multivariate intersection over union metric as bounding box loss

Input: ground truth $B_p(\tilde{x}_1, \tilde{y}_1, \tilde{x}_2, \tilde{y}_2)$ and prediction $(B_p, x_1, y_1, x_2, y_2)$ bounding box

Output: L_{MIOU}

1. Ensuring B_p meets the condition: $x_2 > x_1, y_2 > y_1$:
 $x_1 = \min(x_1, x_2), x_2 = \max(x_1, x_2), y_1 = \min(y_1, y_2), y_2 = \max(y_1, y_2)$
 2. Calculating area of B_g and B_p in Eq.(9) and Eq.(10), getting A_g, A_p .
 3. Finding the coordinates of smallest enclosing box B_C in Eq.(5) and Eq.(6), getting $(X_{C1}, Y_{C1}, X_{C2}, Y_{C2})$ and $c^2 = (X_{C2} - X_{C1})^2 + (Y_{C2} - Y_{C1})^2$.
 4. Calculating the center point of B_g and $B_p, b_g = (x_{b_g}, y_{b_g}), b_p = (x_{b_p}, y_{b_p})$:
 $x_{b_g} = \frac{x_1 + x_2}{2}, y_{b_g} = \frac{y_1 + y_2}{2}, x_{b_p} = \frac{x_1 + x_2}{2}, y_{b_p} = \frac{y_1 + y_2}{2}$
 5. Calculating the Euclidean distance between b_g and $b_p, \rho^2 = (x_{b_p} - x_{b_g})^2 + (y_{b_p} - y_{b_g})^2$
 6. Calculating the center distance ratio R_{dis} : $R_{dis} = \frac{\rho^2}{c^2}$
 7. Finding the coordinates of intersection I between B_g and B_p in Eq.(7) and Eq.(8), getting:
 $(X_{I1}, Y_{I1}, X_{I2}, Y_{I2})$
 8. Calculating area of A_I, A_u in Eq.(11), Eq.(12), $IOU = \frac{A_I}{A_u}$.
 9. Calculating aspect ratio $R_{asp} = (\frac{v}{1-IOU+v})\delta$, where v and δ were calculated in Eq.(17).
 10. Calculating the MIOU loss: $L_{MIOU} = 1 - IOU + R_{dis} + R_{asp}$.
-

scale invariance. Our method uses these three factors to carry out the bounding box regression, which avoids the gradient explosion problem, improves the convergence speed of the model training, and enhances the robustness of the tracker.

4 Experimental results

4.1 Experimental design

Since the training of Siamese network only needs image pairs, we use ILSVRC-VID dataset to train model and use OTB2015 dataset [36] to test model. ILSVRC-VID is the target detection dataset in ImageNet Large Scale Visual Recognition Competition. It includes 3862 snippets for training, 555 snippets for verification and 937 snippets for testing. Each snippet consists of 56,458 images. The ILSVRC-VID dataset has 30 categories, which are carefully selected, taking into account different scene factors, such as motion, video background interference, illumination changes, and so on. The size of the ILSVRC-VID dataset is 85GB, and the training time on the remote supercomputing server is about 3 days (CPU is 2*12 cores, Intel Xeon E5-2692 V2, 64GB memory, 1T disk storage).

The OTB2015 dataset is one of the standard datasets for target tracking, which consists of 100 fully annotated videos with 11 challenging attributes, including illumination variation (IV), scale variation (SV), occlusion (Occ), deformation

(Def), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane (OPR), out-of-view (OV), background clutters (BC) and low resolution (LR). Table 1 shows the distribution of various challenge attributes in the OTB2015 dataset. Among them, the video test scenarios covered by SV and OPR attributes are relatively wide, accounting for more than half of the total dataset. Secondly, Occ, IPR and Def account for a relatively high proportion, indicating that OTB2015 pays more attention to the test of the target's own deformation.

During training, for each video sequence, the target that comes from the first frame is regarded as template frame and the subsequent frame is put into search branch. Among them, the template branch adjusts the input image block size to $127 * 127$ by using the convoluted operation, while the uniform scale of image block in searching branch is $255 * 255$. Finally, according to the overlap area ratio calculation results, when iou value of candidate patches is greater than 0.6, it is judged as positive sample, while the iou value of negative samples is set to be no more than 0.3. The learning rate is initially set to $5 * 10^{-3}$ and the number of anchors is 5. Since the target deformation difference is not obvious in the tracking process, the anchor aspect ratios are set to (0.33, 0.5, 1, 2, 3), while the anchor area is constant. Finally, a total of 20 epochs are performed.

4.2 Experimental analysis

Quantitative analysis: In the performance evaluation, we mainly compare our method against the four state-of-the-art metrics including ℓ_1 -smooth loss, IOU loss, GIOU loss and DIOU loss simultaneously. Firstly, We choose the average center location error as evaluation standard to quantify the performance of the methods. When the effective of tracker is better, the error value is lower, otherwise, the higher. To quickly validate the effectiveness of our proposed method, we only select 10 videos sequences from OTB2015 dataset. Table 2 shows the center error values in 10 videos, in which bold represents the best verification results. According to the results, our method performs better than ℓ_1 -smooth, IOU, GIOU and DIOU.

In Table 2, the scene attributes contained in the video sequences are labeled: illumination variation (IV), scale variation (SV), occlusion (Occ), deformation (Def), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane (OPR), out-of-view (OV), background clutters (BC) and low resolution (LR).

In order to further verify the influence of three variables (overlap area ratio, center distance ratio and aspect ratio) on boundary regression, we compared the tracking effect and training iterations under different regression variable combinations on SiamRPN tracker with 100 sequences of OTB2015, as shown in Table 3, where IOU , R_{dis} and R_{asp} represent the loss functions of overlap area ratio, center distance ratio and aspect ratio, respectively, and epoch is the number of iterations with the best performance during training. The smaller the number of iterations, the faster the convergence speed. Otherwise, the regression optimization takes a long time. It can be seen that the performance of SiamRPN+(1-IOU+ R_{dis}) is better than that of the single one, and the number of iterations is reduced by one. The performance of SiamRPN+(1-IOU+ R_{asp}) is improved in average precision, but the average success and epoch are not improved. The performance of SiamRPN+(1-IOU+ R_{dis} + R_{asp}) (ours) is better than other models. This shows that R_{dis} plays a major role in accelerating convergence, and R_{asp} plays a major role in improving tracking accuracy, and the comprehensive performance of loss measure of three geometric variables is the best.

To further enhance the quantitative analysis, we use the vot2016 dataset to compare the tracker performance of our method with other algorithms. VOT2016 is a benchmark containing 60 video sequences. The common evaluation of VOT2016 includes accuracy (average overlap while tracking successfully), robustness (failure times) and expected average overlap (EAO). EAO is used to evaluated the overall performance which takes account of both accuracy and robustness. The bold in Table 4 means the better performance of method. In EAO and accuracy, the higher the value, the

Table 2 The average center location error (aCLE) of different regression losses on SiamRPN tracker framework

Sequence	Mhyang (IV,OPR, DEF,BC)	Vase (SV,FM,IPR)	Subway (Occ,Def,BC)	Trellis (IV,OPR, SV,IPR,BC)	Jumping (MB,FM)	Deer(MB, FM,IPR,BC)	Biker(OPR, SV,Occ, MB,FM, OV,LR)	Car4 (IS,SV)	David2 (OPR,IPR)	BlurFace (MB,FM, IPR)
SiamRPN + ℓ_1 -smooth [29]	5.01	19.11	133.74	6.48	6.03	28.29	2.83	2.22	5.16	17.64
SiamRPN + L_{IOU} [31]	5.20	18.24	95.80	6.73	6.19	88.63	2.52	2.08	4.54	19.56
SiamRPN + L_{GIOU} [32]	7.43	14.23	3.82	8.22	5.93	29.06	67.88	2.45	4.46	16.89
SiamRPN + L_{DIOU} [33]	5.14	17.57	3.99	6.77	6.00	22.62	2.57	2.34	5.02	18.93
SiamRPN + L_{MIU} (ours)	3.99	13.71	3.07	5.90	5.66	14.11	2.27	1.87	4.07	13.98

Bold data represent the best results of single video tasks

Table 3 The average success, average precision and epoch of different regression losses in SiamRPN tracker framework on 100 sequences of OTB2015

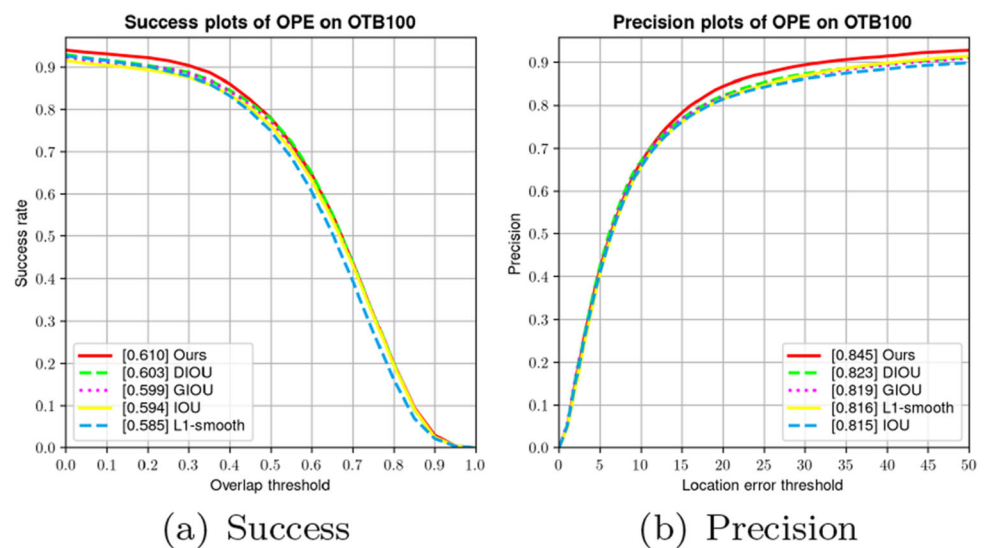
Method	Average success	Average precision	Epoch
SiamRPN+(ℓ_1 -smooth)	0.585	0.816	15
SiamRPN+(1-IOU)	0.594	0.815	13
SiamRPN+(1-IOU+ R_{dis})	0.603	0.823	12
SiamRPN+(1-IOU+ R_{asp})	0.600	0.824	13
SiamRPN+(1-IOU+ R_{dis} + R_{asp})(ours)	0.610	0.845	12

Bold data represent the best results

Table 4 Comparison with the various method in terms of expected average overlap (EAO), robustness (failure rate), and accuracy on the VOT2016 benchmark

Method	ℓ_1 -smooth [29]	IOU [31]	GIOU [32]	DIOU [33]	Ours
EAO	0.311	0.343	0.338	0.329	0.358
Accuracy	0.568	0.577	0.571	0.581	0.578
Robustness	0.322	0.252	0.270	0.280	0.238

Fig. 5 The comprehensive success plots (left) and precision plots (right) of comparison loss functions on 100 sequences in OTB2015



better the performance of the algorithm. On the contrary, the lower the robustness is, the less time the tracking fails, which means the method is more robust. Table 4 shows that our method (MIOU) is able to outperform the trackers in robustness and EAO.

Figure 5 shows the overall tracking success plots and precision plots for all 5 loss functions on 100 sequences in OTB2015. The success score and precision score of our approach are 0.610 and 0.845. The curves of these five methods are very close, but our method is 0.7% and 2.2% higher than the second method in success score and precision score, respectively. We set the error threshold of 20 pixels in precision plots, and the area under curve values of success plots represents the overlap rate between the prediction box and target bounding box.

Figure 6 shows success plot of different algorithms on 9 challenging attributes, including fast motion, occlusion, scale variation, motion blur, illumination variation, low resolution, deformation, out-of-view and out-of-plane rotation.

Our method outperforms the other metrics trackers significantly in terms of 9 challenges, especially in occlusion, scale variation and motion blur, owing to provide regression direction in distance and shape of bounding box. Since GIOU loss and ℓ_1 -smooth loss have strong laziness on intersection over union calculation, it shows slow convergence and easy divergence of training. However, our method is less sensitive to the 9 challenges, which performs more generalization and robustness.

Qualitative analysis:

We illustrate the qualitative results in five different methods on a subset of 6 sequences in Fig. 7. The sequences of Football and Subway contain serious Occ (Occlusion), Def (deformation) and BC (Background Clutters), ℓ_1 -smooth [29] and DIOU [33] occur tracking drift in #109 of Football, but our method, IOU [31] and GIOU [32] keep tracking successfully in the end. Sequences Surfer is typical of target SV (scale variation) and FM (fast motion), as we can see that GIOU has the problem of serious scale tracking fail-

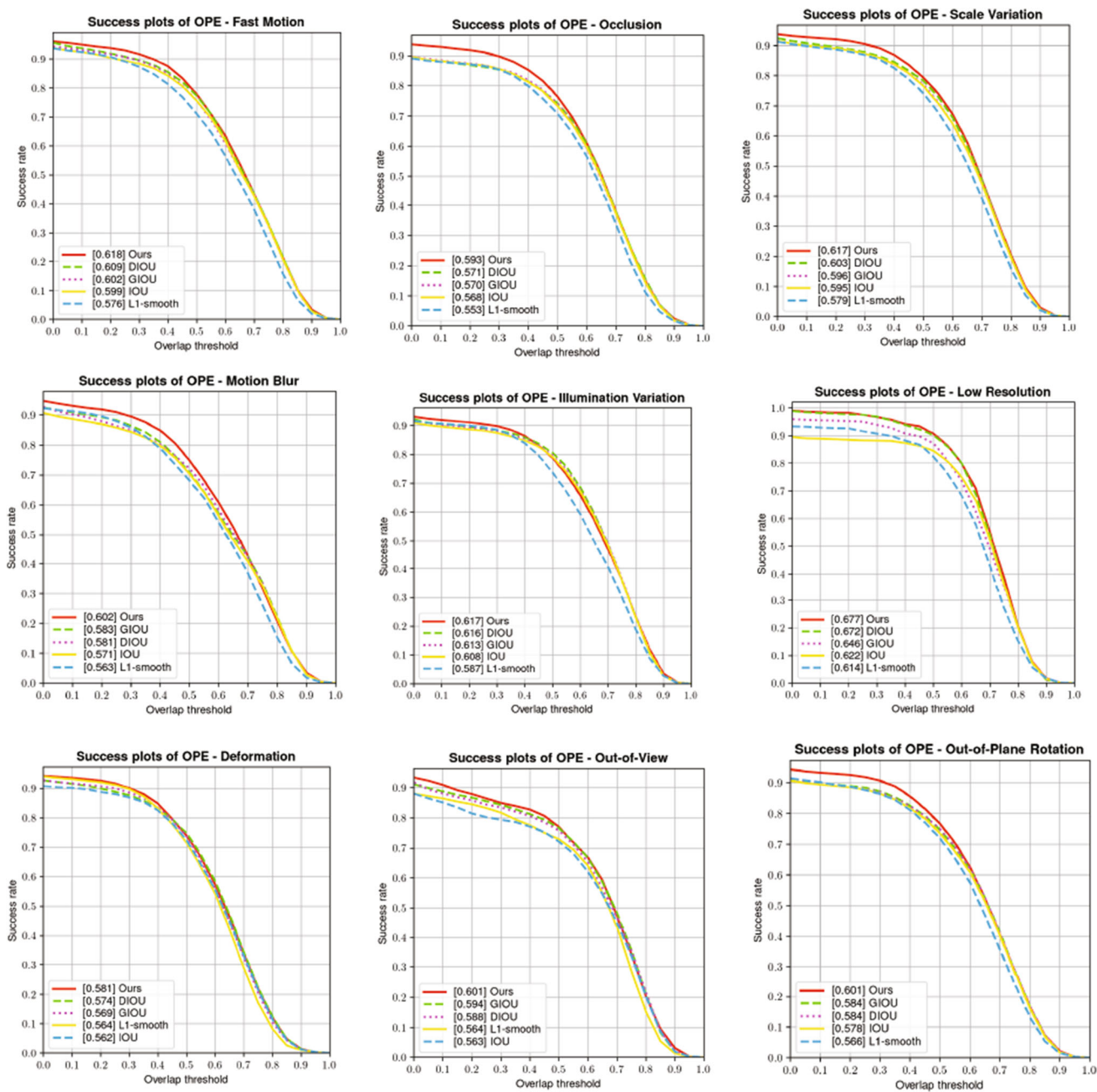
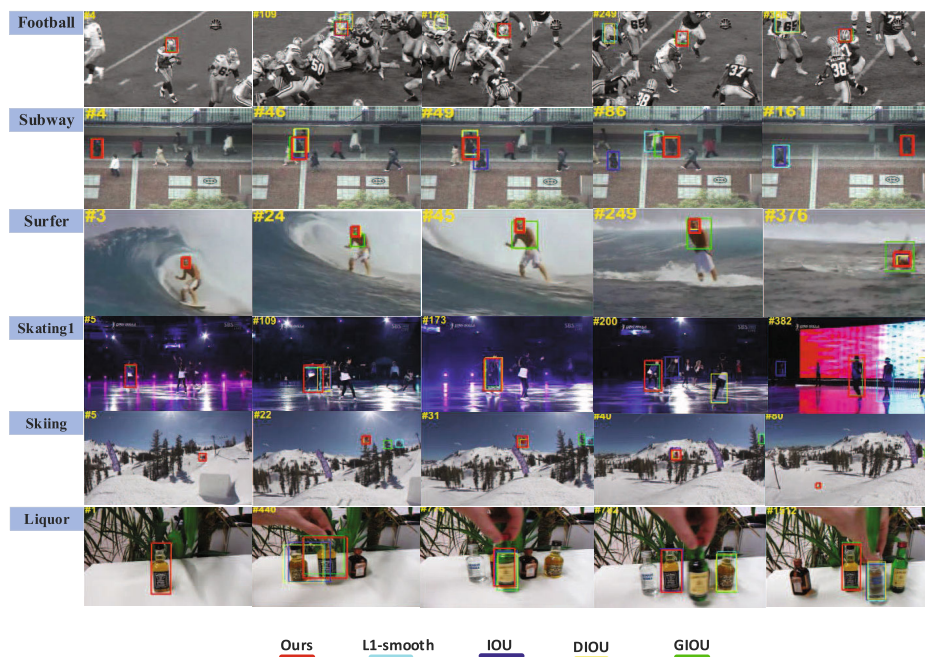


Fig. 6 The success plots between different algorithms in 9 sequence challenges, including fast motion, occlusion, scale variation, motion blur, illumination variation, low resolution, deformation, out-of-view, out-of-plane rotation

ure, while other trackers perform well in these challenges. In sequences Skating 1, many trackers suffer from short-term occlusion in $\#173$; however, our method and GIOU can effectively deal with the non-overlap case and reposition target in $\#200$. Moreover, our approach can identify the target in obvious illumination change. When IV (illumination change) and SV (scale variation) occur in skiing and liquor simultaneously, ℓ_1 -smooth [29], IOU [31], GIOU [32] and DIOU [33] are seriously affected by the susceptibility of scale, and leading to tracking failed. But our method has a good track-

ing effect in these cases and maintain long-term tracking. In general, the results clearly show that using our method as the bounding box regression loss performs consistently better in videos, while some failure cases occurred in ℓ_1 -smooth loss, IOU loss, GIOU loss and DIOU loss.

Fig. 7 Qualitative results of the proposed method (red), ℓ_1 -smooth loss (blue), IOU loss (cyan), GIOU loss (green) and DIOU loss (yellow) (football, subway, surfer, Skating1, skiing, liquor) on OTB2015



5 Conclusion and future work

SiamPRN tracking algorithm has real-time and excellent tracking performance, but it is prone to drift in the case of occlusion and non-overlapping, which is due to the ℓ_1 -smooth [29] loss of its bounding box regression branch. Therefore, this paper analyzes the tracking effects of IOU [31], DIOU and CIOU [33] regression loss functions in the SiamRPN framework, and proposes a multivariate intersection over union (MIOU) regression loss measurement method. MIOU uses the overlap area ratio, the center distance ratio and the aspect ratio of the boundary box, which has scale invariance and speeds up the convergence speed in the training process. On the other hand, we improve the definition of aspect ratio factor to adjust the scale alignment of the bounding box. Therefore, MIOU loss solves the problem of ℓ_1 -smooth loss localization failure in the case of non-overlapping, and can maintain long-term tracking. On the OTB2015 dataset and VOT2016 benchmark, experimental results show that the MIOU loss has better target location regression than other loss functions, especially for the challenges of occlusion, illumination change and fast motion.

We will further study this work. Firstly, similar objects are more sensitive to the interference of targets. We plan to combine the structural information of the target to improve the ability to identify distracters. Secondly, we find that the shape of RPN anchor has a great influence on the effectiveness of the model, so we will introduce adaptive feature fusion to refine the features based on the underlying anchor shapes.

Acknowledgements This research was funded by National Natural Science Foundation of China (Nos. 61772144, 62072122), and Education Dept. of Guangdong Province (No.2019KSYS009), Foreign Science and Technology Cooperation Plan Project of Guangzhou Science Technology and Innovation Commission (No. 201807010059).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Wang, N., Yeung, D. Y.: Learning a deep compact image representation for visual tracking. In: Proceedings of the Neural Information Processing Systems (NIPS), pp. 809-817 (2013)
2. Zhou, X., Xie, L., Zhang, P., Zhang, Y.: An ensemble of deep neural networks for object tracking. In: Proceedings of 2014 IEEE International Conference on Image Processing (ICIP), pp. 843-847 (2014)

3. Wang, N., Li, S., Gupta, A., Yeung, D.Y.: Transferring rich feature hierarchies for robust visual tracking. *arXiv2015* (2015)
4. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. *arXiv 2016*(2016)
5. Tao, R., Gavves, E., Smeulders, A.W.M.: Siamese Instance Search for Tracking. In: *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 850–865 (2016)
6. Xuan, S., Li, S., Zhao, Z., Kou, L., Zhou, Z., Xia, G.: siamese networks with distractor-reduction method for long-term visual object tracking. *Pattern Recognit.* **8**, (2020)
7. Li, B., Yan, J., Wu, W., Zhu Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: *Proceedings of IEEE conference on computer vision and pattern recognition*, pp. 8971–8980 (2018)
8. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised online boosting for robust tracking. In: *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 234–247 (2008)
9. Babenko, B., Yang, M.H., Belongie, S. Visual tracking with online multiple instance learning. In: *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 983–990 (2009)
10. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking–learning–detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(7), 1409–1422 (2012)
11. Mei, X., Ling, H.: Robust visual tracking using l1 minimization. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. pp. 1436–1443 (2009)
12. Wang, D., Lu, H., Yang, M.H.: Online object tracking with sparse prototypes. *IEEE Trans. Image Process (TIP)* **22**(1), 314–325 (2013)
13. Zhang, T., Liu, S., Xu, C., Yan S., Ghanem Be., Ahuja N., Yang, M.H.: Structural sparse tracking. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 150–158 (2015)
14. Wang, Z., Ren, J., Zhang, D., Sun, M., Jiang, J.: A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. *Neurocomputing* **287**, 68–83 (2018)
15. Yan, Y., Ren, J., Zhao, H., Sun, G., Wang, Z., Zheng, J., Marshall, S., Soraghan, J.: Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. *Cognit. Comput.* **10**(1), 94–104 (2017)
16. Han, J., Zhang, D., Cheng, G., Lei, G., Ren, J.: Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Trans. Geosci. Remote Sens.* **53**(6), 3325–3337 (2015)
17. Zabalza, J., Ren, J., Zheng, J., Zhao, H., Qing, C., Yang, Z., Du, P., Marshall, S.: Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing* **185**, 1–10 (2016)
18. Tschannerl, J., Ren, J., Yuen, P., Sun, G., Zhao, H., Yang, Z., Wang, Z., Marshall, S.: MIMR-DGSA: unsupervised hyperspectral band selection based on information theory and a modified discrete gravitational search algorithm. *Inf. Fusion* **51**, 189–200 (2019)
19. Xia, H., Zhang, Y., Yang, M., Zhao, Y.: Visual tracking via deep feature fusion and correlation filters. *Sensors* **20**(12), 3370 (2020)
20. Zhou, X., Xie, L., Zhang, P., et al.: An ensemble of deep neural networks for object tracking. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 843–847 (2014)
21. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: *Proceedings of Computer vision and pattern recognition(CVPR)*, pp. 4293–4302 (2016)
22. Bertinetto, L., Valmadre J., Henriques, J. F., Vedaldi, A., Torr, Philip, H.S.: Fully-Convolutional Siamese Networks for Object Tracking. *Proceedings of European Conference on Computer Vision (ECCV)*.pp.850-865(2016)
23. Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G.: SiamFC++: towards robust and accurate visual tracking with target estimation guidelines. In: *Proceedings of AAAI*, pp. 12549–12556 (2020)
24. Gao, J., Zhang, T., Xu, C.: Graph convolutional tracking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, pp. 4649–4659 (2019)
25. Li, B., Wu, W., Wang, Q., et al.: SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In: *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE (2020)
26. Zhu, Z., Wang, Q., Li, B., et al.: Distractor-aware Siamese Networks for Visual Object Tracking. In: *ECCV2018*. Springer, Cham (2018)
27. Wang, Q., Zhang, L., Bertinetto, L., et al.: Fast online object tracking and segmentation: A unifying approach[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019: 1328-1338
28. Zhang, Z., Peng, H.: Deeper and wider Siamese networks for real-time visual tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4591–4600 (2019)
29. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2015)
30. Song, Y., Ma, C., Wu, X., Gong L., Bao L., Zuo W., Shen C., Lau, R.W.H., Yang, M.H.: VITAL: Visual tracking via adversarial learning. In: *Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 8990–8999 (2018)
31. Yu, J., Jiang, Y., Wang, Z., et al.: UnitBox: an advanced object detection network. In: *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520 (2016)
32. Rezaatofghi, H., Tsoi, N., Gwak, J.Y., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 658–666 (2019)
33. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU Loss: faster and better learning for bounding box regression. In: *Proceedings of AAAI*, pp. 12993–13000 (2020)
34. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
35. Rubinstein, R.: The cross-entropy method for combinatorial and continuous optimization. *Methodol. Comput. Appl. Probab.* **1**(2), 127–190 (1999)
36. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1834–1848 (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Zhihui Huang received her B.S. degree in Computer Science and Technology from Guangdong Ocean University, Zhanjiang, in 2018. Currently, she is a Master student in the School of Computer Science, Guangdong Technical Normal University, Guangzhou. Her research interests include machine learning and computer vision.



Jin Zhan received the Master's and Doctor's degrees from Sun Yat-sen University, Guangzhou, China, in 2004 and 2015, respectively. She is currently an associate professor in the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou. Her research interests include image and video intelligent analysis, machine learning and computer vision.



Huimin Zhao was born in Shanxi, China, in 1966. He received the B.Sc. and the M.Sc. degrees in signal processing from Northwestern Polytechnical University, Xi'an, China, in 1992 and 1997, respectively, and the Ph.D. degree in electrical engineering from the Sun Yat-sen University, Guangzhou, China, 2001. He is currently a Professor and the Dean with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou. His research interests include

image, video, and information security technology.



Huakang Li received his B.S. degree in Electrical Engineering and Automation from Southern College of Sun Yat-sen University, Guangzhou, in 2019. Currently, he is a Master student in the School of Computer Science, Guangdong Technical Normal University, Guangzhou. His research interests include deep learning and gait recognition.