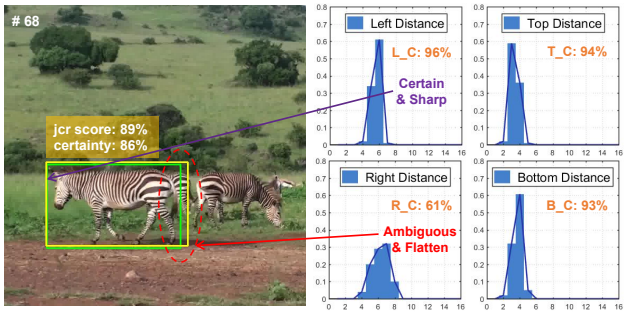
UAST: Uncertainty-Aware Siamese Tracking

Dawei Zhang [1] **Yanwei Fu** [2] **Zhonglong Zheng** [1 3]

Abstract

Visual object tracking is basically formulated as target classification and bounding box estimation. Recent anchor-free Siamese trackers rely on predicting the distances to four sides for efficient regression but fail to estimate accurate bounding box in complex scenes. We argue that these approaches lack a clear probabilistic explanation, so it is desirable to model the uncertainty and ambiguity representation of target estimation. To address this issue, this paper presents an Uncertainty-Aware Siamese Tracker (UAST) by developing a novel distribution-based regression formulation with localization uncertainty. We exploit regression vectors to directly represent the discretized probability distribution for four offsets of boxes, which is general, flexible and informative. Based on the resulting distributed representation, our method is able to provide a probabilistic value of uncertainty. Furthermore, considering the high correlation between the uncertainty and regression accuracy, we propose to learn a joint representation head of classification and localization quality for reliable tracking, which also avoids the inconsistency of classification and quality estimation between training and inference. Extensive experiments on several challenging tracking benchmarks demonstrate the effectiveness of UAST and its superiority over other Siamese trackers.

(a) Different approaches for target box estimation



(b) A qualitative case of uncertainty-aware tracking

Figure 1. (a) Comparison of different regression methods in visual tracking: anchor-based (such as SiamRPN, SiamRPN++), anchor-free (such as SiamFC++, Ocean), and our distribution-based UAST. (b) A representative example of the proposed uncertainty-aware tracking. Due to occlusion and similar objects, the ground-truth (green) may be not explainable enough, and many trackers are limited by such issues. Instead, distribution-based regression (yellow) can reflect the uncertainty information of localization prediction, where a flatten distribution depicts an uncertain and ambiguous boundary, and vice versa. Notably, UAST further provides the estimated certainty with respect to 4-directions (L_C, T_C, R_C and B_C), and the whole certainty value of the predicted box, while jcr score denotes our joint confidence representation score.

1. Introduction

Visual tracking is a fundamental yet challenging research topic in computer vision. It has a wide range of applications, such as surveillance system, UAV-based monitoring, human-computer interaction, and so on. Given only an arbitrary target annotation in the initial frame, object trackers aim at predicting its location and scale in subsequent frames of the video sequence. In the most general form, there is no prior knowledge of the object category and its surrounding environment (Huang et al., 2019). Although much progress has been achieved in recent years, accurate tracking is still a challenging task due to occlusion, motion blur, geometric deformation, scale and appearance variations.

In general, visual tracking can be formulated as a combination of classification and localization sub-tasks. The former aims to robustly predict the coarse location of the target, while the latter is designed to estimate precise bounding

[1]College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua, China [2]School of Data Science, Fudan University, Shanghai, China [3]Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China. Correspondence to: Zhonglong Zheng <zhonglong@zjnu.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

boxes. To enable accurate tracking, regression branch is of great importance as it is responsible for target box estimation. Based on this aspect, previous anchor-based Siamese trackers (Li et al., 2018; 2019) introduce region proposal networks (Ren et al., 2016) to perform bounding box regression. Recent anchor-free Siamese trackers (Xu et al., 2020; Zhang et al., 2020) that become more popular owning to its concise, elegant design and no anchor prior knowledge, directly regress distances to four sides of the box using fully convolutional networks. From a distributional perspective of view, these aforementioned regression methods can be regarded as a simple Dirac delta distribution since the goal of regression is to fit a single value for each output of the target box. Despite significant progress has been achieved, existing trackers do not consider to estimate the uncertainty of box coordinates. In other words, a single prediction of target boundary has no clear probabilistic interpretation due to lacking of extra localization representation information. Therefore, the resulting boxes are prone to inaccuracy or failures in some complex scenes. It is essential to model and estimate the uncertainty of bounding box representation.

Our main motivation is to explore the uncertainty of tracking. Although recent work (Danelljan et al., 2020) tries to exploit Gaussian distribution to model probabilistic representation of bounding boxes, it is not capable to completely and flexibly reflect the underlying distribution of object bounds. In fact, the real distribution is not necessarily symmetric like Gaussian, and even can be more arbitrary (Jiang et al., 2018). So can we model general distribution of bounding boxes to estimate the uncertainty for accurate object tracking?

Following the above analysis, we propose a novel general distribution-based regression formulation to learn localization uncertainty representation of bounding boxes for accurate tracking, inspired by the success of GFL (Li et al., 2020) in object detection. To be consistent with the existing anchor-free Siamese trackers (Xu et al., 2020; Zhang et al., 2020; Guo et al., 2021), the goal of our regression branch is also to predict the relative offsets of the spatial position to the four sides of bounding boxes. Differently, the proposed tracking framework can additionally model the uncertainty and ambiguity representation via learning discretized probability distribution along each of four directions over its continuous domain, without any extra prior knowledge. As shown in Figure 1, the learned distributions obviously reflect the underlying information by its shape. Impressively, the predicted distributions are usually sharp when boundaries are clear and certain, and is flatten when the right border is ambiguous. More than that, our tracker enables to inform which direction of the box boundary is uncertain using a quantitative value. Benefiting from this elegant solution for localization uncertainty reasoning, more accurate bounding boxes can be obtained owing to aware of the potential distributions of target boundaries.

Another limitation of most existing tracking methods is the misalignment between classification and regression. Namely, the position with high classification score may not correspond high regression accuracy, and vice versa, leading to a poor tracking performance. Recent anchor-free trackers (Xu et al., 2020) apply a quality estimation branch to assist the classification branch for final predictions. Nevertheless, the independent optimization of them also brings inconsistency between training and test. To this end, we present a simple yet effective joint representation head of classification and localization quality, which can be trained end-to-end and used directly during tracking. Furthermore, considering the strong correlation between the estimated uncertainty and regression accuracy, we exploit the learned distributions to design a task alignment sub-network for facilitating the learning of our joint representation head. In this way, it eliminates the misalignment and unsolved training-test inconsistency. Notice that our method almost does not degrade the training/inference time of basic trackers due to negligible additional computation cost.

We integrate our general distribution and joint representation into the recent state-of-the-art anchor-free trackers, termed as Uncertainty-Aware Siamese Tracking, UAST. In summary, our main contributions are as follows:

- We propose a novel distributional regression paradigm by learning general representation of bounding boxes for single object tracking, which is capable of flexibly capturing more informative target boundaries for accurate localization, and explicitly estimating the certainty value of each direction in a probabilistic way.
- Based on the learned distributions of bounding box, we propose a simple yet effective joint representation head of classification and localization quality by leveraging the estimated uncertainty and a lightweight task alignment sub-network, which bridges the gap between training and inference. Notably, it is almost cost-free.
- The proposed UAST achieves state-of-the-art performance on five public tracking benchmarks, including GOT-10k, LaSOT, OTB-100, VOT-2019 and UAV-123, demonstrating its effectiveness and tracking efficiency.

2. Related Work

In this section, we briefly review recent single object trackers from the aspect of target state estimation, and introduce uncertainty estimation in computer vision, as well as discuss localization quality estimation of anchor-free methods.

2.1. Visual Object Tracking

Comparing with early popular correlation filters based trackers (Bolme et al., 2010; Henriques et al., 2014), Siamese

network based methods have achieved great progress in tracking community since its good balance of performance and speed. As a pioneering work, SiamFC (Bertinetto et al., 2016) applied multi-scale testing to obtain the target box, which is inefficient and inaccurate. This strategy is severely limited since no specific scale estimation is designed.

For the another popular category, ATOM (Danelljan et al., 2019) presents a customized IoU prediction network (Jiang et al., 2018) for target estimation. Nevertheless, it aggravates the computation burden and many hyper-parameters since multiple initial boxes need to be iteratively refined.

More recent advanced Siamese trackers consider performing classification and regression simultaneously, leading to superior tracking performance. SiamRPN tracker family (Li et al., 2018; 2019) introduce region proposal networks to regress the shift of position and scale between pre-defined anchor boxes and ground truth. Inspired by FCOS (Tian et al., 2019) in object detection, numerous anchor-free trackers (Guo et al., 2020; Chen et al., 2020; Zhang et al., 2020; Peng et al., 2021) have emerged to avoid relying on the prior of candidate boxes, and become more popular due to its simplicity in design. To be specific, SiamFC++ (Xu et al., 2020), SiamCAR (Guo et al., 2020) and SiamBAN (Chen et al., 2020) directly regress the offsets to box borders in a per-pixel-prediction manner. To alleviate the misalignment of classification and regression, Ocean (Zhang et al., 2020) uses a feature alignment module to obtain object-aware predictions for penalizing the classification branch. However, the two branches are trained separately but combined during tracking. Furthermore, SiamRCR (Peng et al., 2021) presents reciprocal links for making training and inference more consistent. Different from them, we devise a joint confidence representation head to tackle this issue.

2.2. Uncertainty Estimation in Computer Vision

Existing object detectors (Ren et al., 2016; Tian et al., 2019) and trackers (Li et al., 2019; Xu et al., 2020) apply Dirac delta distribution to govern the bounding box representation, learning a single prediction for each side of target boxes. Recently, in the object detection field, to model the localization uncertainty, Gaussian YOLOv3 (Choi et al., 2019) and KL-Loss (He et al., 2019) adopt Gaussian assumption to predict the variance of four edges. When the variance is larger, the distribution is flatter, indicating that the prediction is uncertain; the smaller the variance, the sharper the distribution, indicating that the predicted box is confident at the mean position. Nevertheless, these representations are either too simplified or too rigid, which can not reflect the underlying distribution in practice. Furthermore, GFL (Li et al., 2020) relaxes the assumption and directly learns a more flexible general distribution of boxes.

Naturally, the uncertainty estimation of targets also can be applied to visual tracking. PrDiMP (Danelljan et al., 2020) learns to predict the conditional probability density using a probabilistic regression model, which is trained by minimizing the KL divergence between the prediction and label distribution. (Zhong et al., 2021) uses KL to learn policy from teacher for distraction-robust active object tracking. UATracker (Zhou et al., 2021) estimates the uncertainty of IoU prediction, and exploits it to filter out unreliable samples for online learning based discriminative classifier in DiMP (Bhat et al., 2019). In contrast to them, we consider learning the discrete probability distribution of each side of bounding boxes for localization uncertainty estimation, which is more flexible and informative. Meanwhile, our approach still benefits from the advanced IoU-based loss due to compatible with anchor-free trackers. In addition, the certainty can be depicted by an explicit probability value.

2.3. Localization Quality Estimation

SiamFC++ estimates the localization quality based on centerness proposed in FCOS (Tian et al., 2019). However, centerness can not fully account for localization quality. Intersection-over-Union (IoU) between predicted boxes and ground-truth is also explored and proved to be effective in IoUNet (Jiang et al., 2018). After that, Ocean introduces an object-aware branch with predicted boxes, while SiamRCR (Peng et al., 2021) assigns dynamic weights in classification loss based on predicted IoU score. Differently, we exploit distance-IoU score (Zheng et al., 2020) as the label of our joint head, which is more comprehensive and suitable for object tracking. Recent advance (Li et al., 2021) suggests that the bounding box distribution with a sharp peak usually corresponds to accurate localization, and vice versa. Benefiting from the proposed distributed regression, we further utilize the estimated uncertainty representation of localization to weight the classification branch for high-quality examples.

3. Uncertainty-Aware Siamese Tracking

In this section, we describe the proposed UAST in detail. As shown in Figure 2, UAST has a similar structure with existing anchor-free trackers. Nevertheless, our approach is not only capable of learning a discrete probability distribution of four directions for describing the uncertainty of bounding boxes, but also models a joint representation head of classification and localization quality by leveraging the estimated uncertainty in box distributions. To our best knowledge, UAST is the first attempt to explore the power of uncertainty estimation for anchor-free tracking.

3.1. Anchor-Free Tracking

Different from RPN-based trackers, Anchor-free tracking methods directly classify and regress the target bounding box at per-pixel spatial location. Following the paradigm

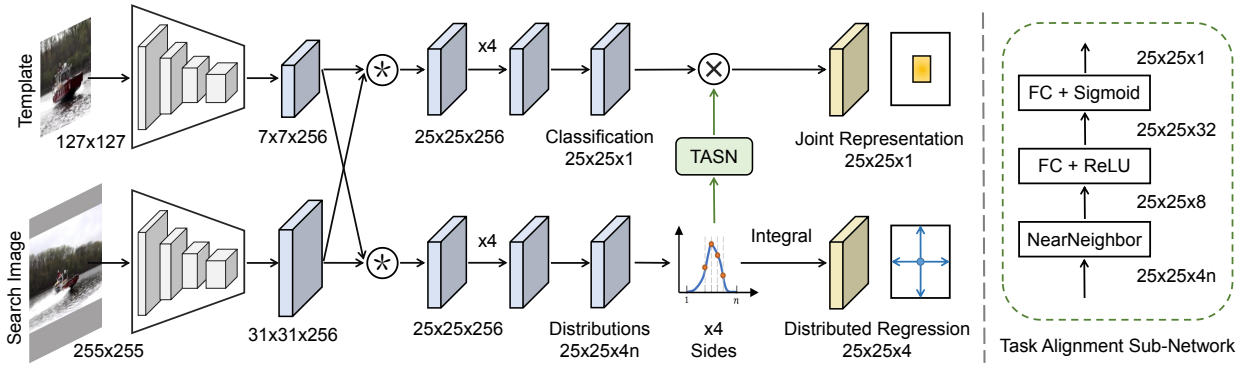


Figure 2. The main structure of the proposed Uncertainty-Aware Siamese Tracking framework. It consists of a backbone network for feature extraction, a feature matching module, an anchor-free head with distributional regression and joint representation, and a task alignment sub-network. Note that $\star$ and $\times$ mean depth-wise cross-correlation and element-wise multiplication operations, respectively.

of FCN (Long et al., 2015), for each position (i, j) in the feature map, we can map it to the search region for obtaining corresponding coordinates $(s/2 + is, s/2 + js)$ (s denotes the total stride of the network) in the original image. Specifically, the output of classification head represents the foreground and background scores of the corresponding locations in the input, while regression head with a 4D vector $T = (l, t, r, b)$ predicts the distances from corresponding locations to four sides of the ground-truth box. Let (x_0, y_0) and (x_1, y_1) denote the left-top and right-bottom corner of the ground truth, so the regression targets (left, top, right, bottom) of the location (i, j) can be calculated as:

$$\begin{aligned} l^* &= i - x_0, \quad t^* = j - y_0 \\ r^* &= x_1 - i, \quad b^* = y_1 - j \end{aligned} \tag{1}$$

Consequently, it allows to predict distances from the location (i, j) to four sides of the box. However, it has no a clear probabilistic explanation of bounding boxes due to lacking the uncertainty representation of target coordinates. It has insufficient information for accurate tracking, and is inflexible to deal with object variations in complex scenes.

3.2. Distributional Regression Representation

From a distribution perspective of view, the existing anchor-based and anchor-free trackers can be considered as a simple Dirac delta distribution $\delta(x - \xi)$ since the regression target is to fit a single label value ξ for each output of the box. It satisfies $\int_{-\infty}^{+\infty} \delta(x - \xi)\mathrm{d}x = 1$, and the integral form to resume ξ can be presented as the following equation:

$$\xi = \int_{-\infty}^{+\infty} \delta(x - \xi)x \, \mathrm{d}x \tag{2}$$

To address the limitation of Dirac delta, we propose to directly model a general distribution $P(x)$ without other priors. Given a range of label ξ ($\xi_0 \leq \xi \leq \xi_n, n \in \mathbb{N}^+$) with minimum ξ_0 and maximum ξ_n, we can obtain the prediction $\bar{\xi}$ of each side via calculating its integral:

$$\bar{\xi} = \int_{-\infty}^{+\infty} P(x)x \, \mathrm{d}x = \int_{\xi_0}^{\xi_n} P(x)x \, \mathrm{d}x \tag{3}$$

For this general distribution, a problem that needs to be solved is that it is difficult to model an arbitrary and continuous probability distribution with a small number of parameters in neural networks. To this end, we consider a discrete representation to fit this distribution. Specifically, the range $[\xi_0, \xi_n]$ can be divided into a set $[\xi_0, \xi_1, \xi_2, ..., \xi_{n-1}, \xi_n]$ with even interval. Hence, our regression branch has $n + 1$ predicted values for each edge of bounding boxes, which can represent probabilities through a softmax layer. Based on the discrete distribution property $\sum_{i=0}^{n} P(\xi_i) = 1$, the estimated regression value $\bar{\xi}$ can be calculated as $\bar{\xi} = \sum_{i=0}^{n} P(\xi_i)\xi_i$. Therefore, the proposed distributional regression formulation can also use previous loss objectives like IoU Loss in anchor-free trackers to train $\bar{\xi}$.

Although the regression target can be obtained according to Equation 3, we expect that the learned distributions are as certain or compact as possible for interpretability since the same integral result may correspond to different arbitrary distributions. In order to explicitly focus on the values (ξ_i and ξ_{i+1}) that are close to the label ξ, we further consider to optimize the shape of distributions using Distribution Focal Loss, DFL proposed in (Li et al., 2020):

$$\mathcal{L}_{dfl} = -\left((\xi_{i+1} - \xi)\log(\mathcal{P}_i) + (\xi - \xi_i)\log(\mathcal{P}_{i+1})\right) \tag{4}$$

where $\mathcal{P}_i$ and $\mathcal{P}_{i+1}$ denote $P(\xi_i)$ and $P(\xi_{i+1})$ respectively. Intuitively, DFL enlarges the probabilities of ξ_i and ξ_{i+1}.

3.3. Joint Confidence Representation

Recent research suggests that localization quality also needs to be considered with the classification score for final predictions during online tracking, but existing trackers (Guo et al.,

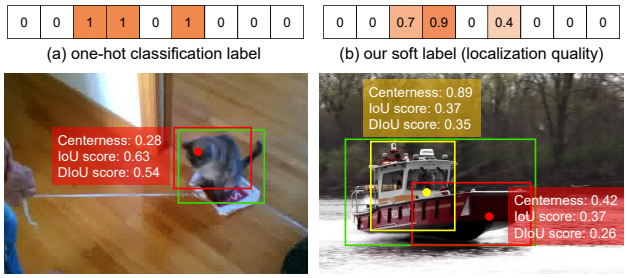


Figure 3. An illustration of our joint confidence representation. Instead of the fore/background label (a), we merge the target confidence and localization quality as the supervision of our jcr (b). (c) Comparisons of different localization quality targets including centerness, IoU and Distance-IoU scores applied in UAST.

2020; Zhang et al., 2020) exist an inconsistent problem between training and inference phases. To this end, we present a simple yet effective joint confidence representation head by leveraging the information from both classification and regression branches. To be specific, given the classification vector $\mathbf{V}_{cls}$ and localization quality vector $\mathbf{V}_{lq}$, our joint confidence representation $\mathbf{V}_{jcr}$ can be formulated as:

$$\mathbf{V}_{jcr} = \mathbf{V}_{cls} \times \mathbf{V}_{lq} \tag{5}$$

which can be trained end-to-end and directly utilized during tracking, because we explicitly optimize the final joint formulation (i.e., $\mathbf{V}_{jcr}$). In contrast to a standard binary classification label in Siamese tracking, we redefine the supervision for our joint representation head. To be specific, negative samples are still supervised by 0, while the supervision of positives is determined by the localization quality label. As shown in Figure 3, the on-hot label is replaced by our soft label for joint confidence representation. Namely, $\mathbf{V}_{jcr}$ where its value at the center range of ground-truth box directly learns its corresponding localization quality.

3.3.1. Localization Quality Label

Current trackers utilize centerness (Xu et al., 2020) or standard IoU score (Zhang et al., 2020) to supervise localization quality. Unfortunately, centerness mainly emphasize the center of target box, while IoU may lead to slow convergence and inaccuracy. Differently, Distance-IoU (Zheng et al., 2020) between the predicted bounding boxes and its ground-truth is applied as the label of positive samples in our joint head, which is a dynamic value being $[0, 1]$.

$$D\text{-}IoU = IoU - \frac{\rho^2\left(b, b^{gt}\right)}{c^2} \tag{6}$$

where $\rho\left(b, b^{gt}\right)$ denotes the Euclidean distance between the central points of predicted box and target box, and c is the diagonal length of the smallest enclosing box covering the two boxes. Notice that DIoU incorporates both normalized center distance and IoU score, which is more suitable for visual tracking task (see examples in Figure 3). Because most evaluation metrics are actually the center distance error (precision) and the average overlap rate (AUC score).

3.3.2. Task Alignment Sub-Network

Benefiting from distributional regression, instead of convolutional features, we can exploit the uncertainty information in box distributions to perform task alignment, facilitating the learning of our joint confidence representation. Specifically, considering that the learned distributions are highly related to the quality of regressed boxes, we construct a lightweight task alignment sub-network from the regression branch to generate high-quality estimation. As shown in Figure 2, we firstly select two nearneighbor values of prediction in each distribution $P(x)$, and concatenate them as the initial localization quality features $\mathbf{F} \in \mathbb{R}^{4\times 2}$:

$$\mathbf{F} = Concat\left(\{Neighbor\left(P(x)\right) \mid x \in \{l, r, t, b\}\}\right) \tag{7}$$

where $Neighbor(\cdot)$ feature can basically reflect the flatness of each distribution, and is robust to object scales.

Based on $\mathbf{F}$ from the regression branch, the localization quality vector $\mathbf{V}_{lq}$ can be obtained by the task alignment sub-network with two Fully-Connected (FC) layers, which are followed by ReLU and Sigmoid, respectively.

$$\mathbf{V}_{lq} = Sigmoid\left(\mathbf{W}_2(ReLU\left(\mathbf{W}_1\mathbf{F}\right))\right) \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{32\times 8}$ and $\mathbf{W}_2 \in \mathbb{R}^{1\times 32}$ represent two FC layers, respectively. It is worth noting that our TASN is very lightweight, and also has little computation overhead.

3.4. Training Objective

We optimize the overall training objective as follows:

$$\mathcal{L} = \mathcal{L}_{jcr} + \lambda_1\mathcal{L}_{reg} + \lambda_2\mathcal{L}_{dfl} \tag{9}$$

where $\mathcal{L}_{jcr}$ is the binary cross entropy loss to train the joint representation head. We only consider positive samples for regression objective. $\mathcal{L}_{reg}$ is the IoU Loss for bounding box regression, while $\mathcal{L}_{dfl}$ forces the model to focus on learning the probabilities of values neighbored with the target box, leading to a reasonable distribution. In our experiments, λ_1 (2 as default) and λ_2 (1/4, averaged over four directions) are the hyper-parameters for balancing these three losses.

4. Experiments

4.1. Implementation Details

4.1.1. Framework

Like Ocean (Zhang et al., 2020), we employ a modified ResNet-50 (He et al., 2016) that only contains the first four

Algorithm 1 Uncertainty-Aware Siamese Tracking

Input: Frames $\{I_k\}_1^K$, initial target box B_1
Output: Target box $\{B_k\}_2^K$, certainty value $\{C_k\}_2^K$
for $k = 2$ **to** K **do**
  Perform feature extraction and matching;
  Model distributed representation $\{D_k^l, D_k^t, D_k^r, D_k^b\}$;
  Obtain 4 offsets $\{L_k, T_k, R_k, B_k\}$ by Eq. 3;
  Extract feature $\mathbf{V}_{lq}$ according to Eq. 7 and Eq. 8;
  Calculate the joint confidence score $\mathbf{V}_{jcr}$;
  Select the highest jcr and corresponding box B_k;
  Compute $\{C_k^l, C_k^t, C_k^r, C_k^b\}$ for 4 sides of box B_k;
  Average them and achieve the whole certainty C_k.
  if $C_k < 0.5$ **then**
    Warning: Uncertain Tracking Result!
  end if
end for

stages as our backbone. To be a fair comparison, the depth-wise correlation is utilized to generate the fused features for subsequent anchor-free head. Differently, we remove the separate quality assessment branch owing to our joint confidence representation of classification and quality. The last layer of our regression head for each side has $n + 1$ outputs instead of 1, incurring negligible computing cost.

4.1.2. Training Phase

The backbone is pre-trained on ImageNet (Russakovsky et al., 2015). The training image pairs are sampled by ImageNet VID and DET (Russakovsky et al., 2015), COCO (Lin et al., 2014), Youtube-BB (Real et al., 2017), GOT-10K and LaSOT (Fan et al., 2019). Template image is 127×127 pixels, while search region is 255×255 pixels. We totally train the network using synchronized stochastic gradient descent (SGD) with a batch size of 128 on 4 GPUs for 20 epochs, and employ warm-up in the first 5 epochs, and a learning rate exponentially decayed from 5e-3 to 1e-6 in the last 15 epochs. We freeze the backbone in the first 10 epochs, and fine-tune it in the remaining epochs. The weight decay and momentum are set as 1e-5 and 0.9, respectively.

4.1.3. Tracking Phase

The intuitive outputs of UAST are a set of distance probabilities and jcr score. We can easily predict the bounding boxes by calculating the integral of each distribution. Following (Li et al., 2018), the score map is also penalized by cosine window and scale change for motion smoothness. The corresponding box of the location with best jcr score is selected and updates the target state by linear interpolation. Meanwhile, UAST takes the summation of two adjacent probability in each border as the certainty values for four directions, and the mean of them as the overall reliably of tracking. Algorithm 1 shows the procedure in details.

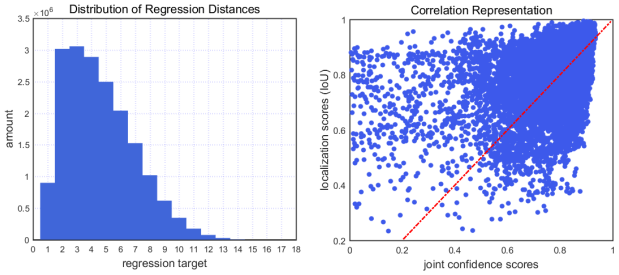


Figure 4. The histogram of regression targets in anchor-free tracking over 120000 training samples on GOT-10k train set, and the scatter diagram represents the correlation between IoU and the joint confidence scores for some randomly sampled instances.

Table 1. Ablation experiments of different variants of UAST on GOT-10K test set, baseline is Ocean without object-aware branch.

	Components	AO	$SR_{0.5}$	$SR_{0.75}$
O	Ocean	0.592	0.695	0.465
I	Baseline	0.572	0.674	0.435
II	+ General Dist.	0.584	0.687	0.446
III	+ Dist. FL	0.596	0.705	0.462
IV	+ Joint Rep.	0.614	0.723	0.485
V	+ Task Align.	0.635	0.741	0.514

UAST with a speed of 65 fps is implemented by PyTorch 1.1. Our experiments are conducted on a server with Intel Xeon (R) Gold 5118 CPU, and a Tesla V-100 16 GB GPU.

4.2. Ablation Study

4.2.1. Component-wise Analysis.

To verify the influence of the proposed approach, we perform a component-wise study on GOT-10k, as presented in Table 1. The offline version of Ocean (O) achieves 0.592 AO score. The baseline (I) denotes Ocean (Zhang et al., 2020) with a classification head (without localization quality branch) and an anchor-free regression head, so that only obtaining an AO score of 0.572. We replace the regression module with our general distributions, which yields an AO gain of 1.2 point, confirming that the proposed distribution based method (II) performs better than single prediction of simple Dirac delta distribution. In Line 3 (III), DFL also brings an improvement of 1.2% in terms of AO due to focus on nearby values of the ground-truth, which is helpful to accurate target estimation. Furthermore, adding the joint representation head of classification and localization quality (IV) can improves the AO of 1.8% and the $SR_{0.75}$ of 2.3%, since it benefits from our distance-IoU guided predictions. Finally, the uncertainty-aware quality feature generated by the proposed task alignment sub-network (V) brings a significant improvement of 2.1 point, showing the effectiveness of our uncertainty representation. Therefore, those different components all contribute to accurate tracking.

Table 2. Comparisons of different localization quality estimation.

LQE	NONE	CENTER	IOU	D-IOU	JCR-DIOU
AO	0.572	0.587	0.591	0.596	0.605

Table 3. Performances of various popular anchor-free Siamese trackers integrated by the proposed UAST on LaSOT test set.

TRACKER	DIS. REP.	JCR	SUCCESS	FPS
SIAMCAR	×	×	0.507	52
SIAMCAR + UAST	√	√	0.543	52
SIAMBAN	×	×	0.514	40
SIAMBAN + UAST	√	√	0.548	40
SIAMGAT	×	×	0.539	70
SIAMGAT + UAST	√	√	0.567	70
OCEAN	×	×	0.526	68
OCEAN + UAST	√	√	0.571	68

4.2.2. DISCUSSION ON DISTRIBUTED REGRESSION

To determine a reasonable range of n, we illustrate the distribution of bounding box regression targets in Figure 4. According to the statistical histogram over large training samples, the recommended value is preferably greater than or equal 14, and we set it to 16. In Table 1, we find that the general distribution can achieve better results, and DFL further boosts its performance. A representative case with its distributions and uncertainty over four directions is depicted in Figure 1, showing that the proposed distributed regression method can effectively represent the prediction confidence with respect to four sides of the target bounding box by its shape and the estimated certainty value. Notably, the right distance of zebra is ambiguous due to partly occlusion.

4.2.3. DISCUSSION ON JOINT REPRESENTATION

In addition to classification, the measurement of localization quality is also important but ignored in the filed of tracking. Centerness is a pre-defined label that indicates the distances between locations and target center, while IoU scores reflect localization accuracy. We find that both of them can improve the AO more or less in Table 2. Nevertheless, DIoU performs better than them with an AO of 0.596 due to comprehensiveness. Figure 3 also shows that DIoU can depict the localization quality more accurately. To this end, we apply DIoU as the label of our joint head, and yields an obvious gain of 3.3%, demonstrating its effectiveness. More importantly, it can be trained end-to-end and directly utilized during tracking. Furthermore, we plot the scatter diagram between IoU scores and the predicted joint scores in Figure 4, leading to a more consistent correlation.

4.2.4. COMPATIBILITY FOR ANCHOR-FREE TRACKERS

We integrate the distributed regression and joint confidence representation in UAST to a series of recent anchor-free trackers, and make the minimal and necessary modifications to perform uncertainty-aware tracking. Based on the results in Table 3, UAST can consistently improve the success by 3 points or more on LaSOT, without loss of inference speed.

Table 4. State-of-the-art comparison on the GOT-10k test set in terms of average overlap (AO) and success rate (SR).

Trackers	AO	$SR_{0.5}$	$SR_{0.75}$
MDNet	0.299	0.303	0.099
ECO	0.316	0.309	0.111
SiamFC	0.374	0.404	0.144
SiamRPN++	0.517	0.616	0.325
ATOM	0.556	0.634	0.402
SiamCAR	0.569	0.670	0.415
SiamFC++	0.595	0.695	0.479
Ocean	0.592	0.695	0.473
D3S	0.597	0.676	0.462
DiMP50	0.611	0.717	0.492
LightTrack	0.623	0.726	-
RPT	0.624	0.730	0.504
SiamGAT	0.627	0.743	0.488
PrDiMP	0.634	0.738	0.543
UAST	0.635	0.741	0.514

4.3. Comparison with State-of-the-art Methods

We evaluate UAST with state-of-the-art methods on five tracking benchmarks including GOT-10k (Huang et al., 2019), VOT-2019 (Kristan et al., 2019), OTB-100 (Wu et al., 2015), UAV-123 (Mueller et al., 2016) and LaSOT (Fan et al., 2019). Without bells and whistles, UAST achieves the state-of-the-art performance, and experimental results are presented in detail in the following subsections.

4.3.1. GOT-10K BENCHMARK

GOT-10k (Huang et al., 2019) is a large-scale generic object tracking benchmark with 10000 video sequences, which includes 180 videos for testing. Note that it is zero-class-overlap between the train subset and test subset. Following the official protocol, we train UAST only with its training set, and evaluate it with 14 state-of-the-art tracking methods on the test set. As shown in Table 4, our UAST achieves 0.635 of AO, which is superior to other anchor-free trackers SiamGAT (Guo et al., 2021), RPT (Ma et al., 2020), Ocean (Zhang et al., 2020) and D3S (Lukezic et al., 2020). These results show the effectiveness of our localization uncertainty estimation. Moreover, UAST slightly performs better than the recent online learning based trackers PrDiMP (Danelljan et al., 2020), which further proves the generalization ability of the proposed tracker on some unseen target classes.

4.3.2. LASOT BENCHMARK

LaSOT (Fan et al., 2019) is a high-quality large-scale tracking benchmark with 280 long-term testing videos. We eval-

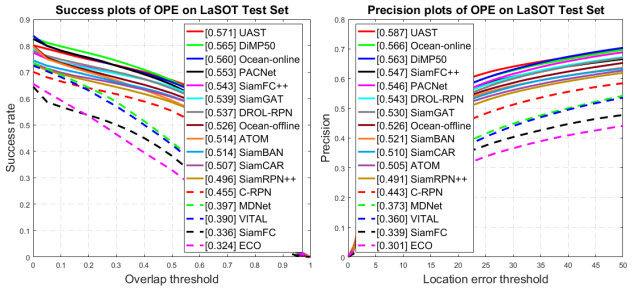


Figure 5. Precision and success plots of OPE on LaSOT test set.

Table 5. Comparison of tracking results on VOT-2019 Benchmark.

Trackers	EAO ↑	Accuracy ↑	Robustness ↓
SiamFCOS	0.223	0.561	0.788
SPM-Tracker	0.275	0.577	0.507
SiamMask	0.287	0.594	0.461
SiamRPN++	0.292	0.580	0.446
SiamDW	0.299	0.600	0.467
PACNet	0.300	0.573	0.401
ATOM	0.301	0.603	0.411
DiMP50	0.321	0.582	0.371
SiamBAN	0.327	0.602	0.396
Ocean	0.327	0.590	0.376
UAST	0.334	0.608	0.386

uate our tracker with DiMP-50 (Bhat et al., 2019), Ocean (Zhang et al., 2020), PACNet (Zhang et al., 2021), DROL-RPN (Zhou et al., 2020) and other 12 methods. Figure 5 shows that the proposed UAST achieves state-of-the-art performance with an AUC score of 0.571 and a precision of 0.587, performing better than other SOTA Siamese trackers. Impressively, our method obtains the best metrics among all trackers in comparison, and surpasses Ocean-online and DiMP-50 by a visible margin. It proves that UAST is also effective to reliably and accurately track long-term targets.

4.3.3. VOT-2019 Benchmark

We evaluate UAST on the Visual Object Tracking real-time challenge 2019 (Kristan et al., 2019). As shown in Table 5, our UAST achieves the performance on EAO criteria of 0.334, Robustness of 0.386 and Accuracy of 0.608, which is better than recent state-of-the-art trackers, such as Ocean, SiamBAN and DiMP. Note that UAST has an obvious advantage in terms of accuracy in all comparisons. It suggests that our tracker can accurately estimate the target box owing to the proposed distributional regression formulation. We further report the experimental results of EAO in Figure 6.

4.3.4. OTB-100 Benchmark

OTB-100 (Wu et al., 2015) is a classical benchmark in visual tracking, containing 100 short-term videos. We report the results on OTB-100 with SiamRPN++ (Li et al., 2019),

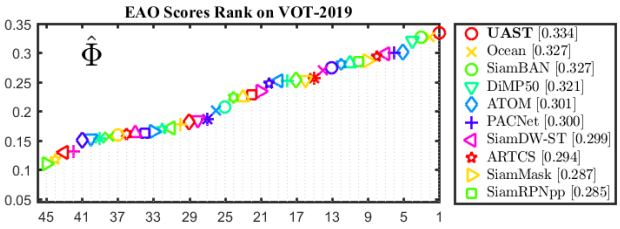


Figure 6. Expected averaged overlap result on VOT-2019.

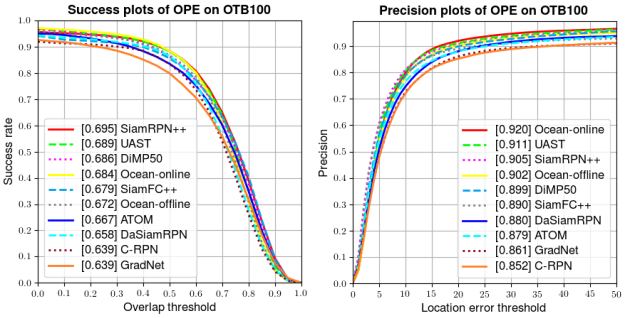


Figure 7. Precision and success plots of OPE on OTB100.

DiMP50 (Bhat et al., 2019), Ocean (Zhang et al., 2020), ATOM (Danelljan et al., 2019), SiamFC++ (Xu et al., 2020), etc. Figure 7 shows that UAST achieves a comparable performance with AUC of 0.689 and precision of 0.911, and obtains 2.3% and 0.9% improvements than Ocean.

4.3.5. UAV-123 Benchmark

UAV123 (Mueller et al., 2016) consists of 123 sequences captured by low-altitude UAVs. It can be used to evaluate whether the tracker is suitable for deployment in aerial scenarios. To this end, we compare the proposed method with 9 state-of-art trackers. Figure 8 shows the results in detail. UAST outperforms most previous Siamese trackers, and obtains a close auc score with SiamGAT (Guo et al., 2021). For precision, our tracker obtains the top rank of 0.860, which is superior than SiamGAT, DiMP50 and ATOM. It demonstrates the effectiveness of our uncertainty-aware tracker.

4.4. Discussion

Both GFL (Li et al., 2020) and our method directly learn the joint representation. However, UAST is designed especially for visual tracking since only one object should be tracked. On the other hand, GFL mainly develops Focal loss (Lin et al., 2017) for data imbalance problem in object detection, while UAST aims at exploring uncertainty for tracking. In addition to the shape of distributions, UAST further estimates the uncertainty by a quantitative value, which is instructive and potentially influential in the field of tracking (see more related discussions in the appendix). It is expected that the estimated uncertainty can be utilized as crucial information for safety-critical vision systems.

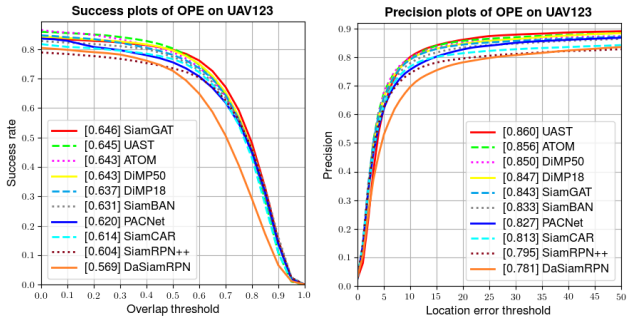


Figure 8. Precision and success plots of OPE on UAV-123.

5. Conclusion

In the paper, we propose to learn a distribution based regression formulation for accurate visual tracking, which models localization uncertainty representation. It is an entirely new perspective in tracking community, since our method has an explicit probabilistic interpretation with highly flexible discretized distributions. Furthermore, we address the task misalignment of anchor-free trackers by learning a joint representation of classification and quality estimation. Experiments show that UAST outperforms previous state-of-the-arts on several tracking benchmarks. We hope our work could inspire the research of uncertainty in object tracking.

Acknowledgements

This work was supported by the Natural Science Foundation of China under Grant No. 11871438, and the Key Projects of Natural Science Foundation of Zhejiang Province under Grant No. LZ22F020010.

References

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. S. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, pp. 850–865, 2016.

Bhat, G., Danelljan, M., Gool, L. V., and Timofte, R. Learning discriminative model prediction for tracking. In *IEEE International Conference on Computer Vision*, pp. 6182–6191, 2019.

Bolme, D. S., Beveridge, J. R., Draper, B. A., and Lui, Y. M. Visual object tracking using adaptive correlation filters. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2544–2550, 2010.

Chen, Z., Zhong, B., Li, G., Zhang, S., and Ji, R. Siamese box adaptive network for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6668–6677, 2020.

Choi, J., Chun, D., Kim, H., and Lee, H.-J. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *IEEE International Conference on Computer Vision*, pp. 502–511, 2019.

Danelljan, M., Bhat, G., Khan, F. S., and Felsberg, M. Atom: Accurate tracking by overlap maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4660–4669, June 2019.

Danelljan, M., Gool, L. V., and Timofte, R. Probabilistic regression for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7183–7192, 2020.

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., and Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5374–5383, June 2019.

Guo, D., Wang, J., Cui, Y., Wang, Z., and Chen, S. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6269–6277, 2020.

Guo, D., Shao, Y., Cui, Y., Wang, Z., Zhang, L., and Shen, C. Graph attention tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9543–9552, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, Y., Zhu, C., Wang, J., Savvides, M., and Zhang, X. Bounding box regression with uncertainty for accurate object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2888–2897, 2019.

Henriques, J. F., Caseiro, R., Martins, P., and Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2014.

Huang, L., Zhao, X., and Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1562–1577, 2019.

Jiang, B., Luo, R., Mao, J., Xiao, T., and Jiang, Y. Acquisition of localization confidence for accurate object detection. In *European Conference on Computer Vision*, pp. 784–799, 2018.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J.-K., Cehovin Zajc, L., et al.

The seventh visual object tracking vot2019 challenge results. In *IEEE International Conference on Computer Vision Workshops*, pp. 2206–2241, 2019.

Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X. High performance visual tracking with siamese region proposal network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8971–8980, June 2018.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., and Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4282–4291, June 2019.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., and Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, 2020.

Li, X., Wang, W., Hu, X., Li, J., Tang, J., and Yang, J. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11632–11641, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pp. 2999–3007, Oct 2017.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

Lukezic, A., Matas, J., and Kristan, M. D3s-a discriminative single shot segmentation tracker. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7133–7142, 2020.

Ma, Z., Wang, L., Zhang, H., Lu, W., and Yin, J. Rpt: Learning point set representation for siamese visual tracking. In *European Conference on Computer Vision*, pp. 653–665, 2020.

Mueller, M., Smith, N., and Ghanem, B. A benchmark and simulator for uav tracking. In *European Conference on Computer Vision*, pp. 445–461, 2016.

Peng, J., Jiang, Z., Gu, Y., Wu, Y., Wang, Y., Tai, Y., Wang, C., and Weiyao, L. Siamrcr: Reciprocal classification and regression for visual object tracking. In *International Joint Conference on Artificial Intelligence*, pp. 952–958, 2021.

Real, E., Shlens, J., Mazzocchi, S., Pan, X., and Vanhoucke, V. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7464–7473, July 2017.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Tian, Z., Shen, C., Chen, H., and He, T. Fcos: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision*, pp. 9627–9636, 2019.

Wu, Y., Lim, J., and Yang, M.-H. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015.

Xu, Y., Wang, Z., Li, Z., Yuan, Y., and Yu, G. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI Conference on Artificial Intelligence*, pp. 12549–12556, 2020.

Zhang, D., Zheng, Z., Jia, R., and Li, M. Visual tracking via hierarchical deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pp. 3315–3323, 2021.

Zhang, Z., Peng, H., Fu, J., Li, B., and Hu, W. Ocean: Object-aware anchor-free tracking. In *European Conference on Computer Vision*, pp. 771–787, 2020.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI Conference on Artificial Intelligence*, pp. 12993–13000, 2020.

Zhong, F., Sun, P., Luo, W., Yan, T., and Wang, Y. Towards distraction-robust active visual tracking. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12782–12792, 2021.

Zhou, J., Wang, P., and Sun, H. Discriminative and robust online learning for siamese visual tracking. In *AAAI Conference on Artificial Intelligence*, pp. 13017–13024, 2020.

Zhou, L., Ledent, A., Hu, Q., Liu, T., Zhang, J., and Kloft, M. Model uncertainty guides visual object tracking. In *AAAI Conference on Artificial Intelligence*, pp. 3581–3589, 2021.

A. More Discussions about Distributed Regression.

Figure 9 shows some examples of noisy, incorrect or ambiguous ground truth bounding box annotations from GOT-10K (Huang et al., 2019). However, the previous bounding box regression methods (i.e., SiamRPN (Li et al., 2018), SiamFC++ (Xu et al., 2020)) do not take such the ambiguities of the ground truth bounding boxes into account. As a result, the learning is unstable, and the loss is relatively large in these cases. To address this issue, we propose a novel bounding box regression formulation with general distribution. The learned probability distribution is interpretable, since it can reflect the level of uncertainty of bounding box predictions.

Figure 9. In visual object tracking, the ground-truth bounding boxes have inherent ambiguities in some cases. The first row shows that the object boundary is unclear and ambiguous due to shadow or itself factor; the ambiguities of the second row are introduced by similar objects or background noises; and the last row are examples of occlusion. These aspects are modeled by our distributed representation.

As shown in Figure 10, it illustrates the representations of Dirac delta, and the proposed general distributions, where the assumption goes from rigid (Dirac delta) to flexible (General). A very significant advantage of our work is that the learned probability distributions ca reflect the uncertainty of bounding box predictions. We also list several key comparisons about these distributions in Table 6. The proposed distribution decouples the representation and loss objective of bounding box regression, making it compatible for existing anchor-free tracking methods, including both edge level for learning its probability representation and box level for learning bounding box regression.

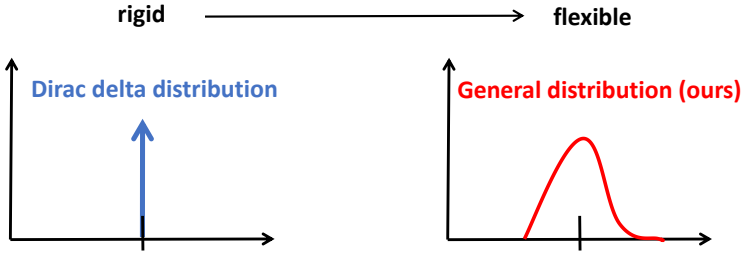


Figure 10. Illustrations of distributions of bounding box regression, from rigid (Dirac delta) to flexible (General). Existing trackers roots at a fixed point using Dirac delta, have limitations in modeling real data distribution. In contrast, our distribution is more flexible as its shape can reflect the uncertainty information of bounding box predictions.

Type	Dirac delta distribution		Ours distribution	
Probability Density	$\delta(x-\xi)$		$P(x)$	
Inference Target	x		$\int P(x)x\,\mathrm{d}x$	
Loss Objective	$\frac{(x-\xi)^2}{2}$	L_{IoU}	$\frac{(\int P(x)x\mathrm{d}x-\xi)^2}{2}$	L_{IoU}
Optimization Level	edge	box	edge	box

Table 6. Comparisons of different representation of distributions for bounding box regression. Edge level denotes optimization over four directions, while the box level means IoU-based Losses that consider bounding boxes as a whole objective.

B. Analysis of Different Localization Quality Label.

Centerness (Xu et al., 2020) mainly consider the center location of the target box, while IoU score (Zhang et al., 2020) may cause inaccurate quality label due to lacking of modeling center distance. To this end, our localization quality label applies Distance-IoU score (Zheng et al., 2020) between the predicted bounding box and its ground-truth, which incorporates both the normalized center distance and IoU score, which is more suitable and effective in visual tracking. As shown in Fig. 3, in the left case, the red point is a bit far from the target center, so the centerness is small. We discover the major problem of centerness is that its definition leads to unexpected small label value, which causes unstable training. In practice, its receptive field corresponds the head of the cat, so the predicted box is not bad, has a 0.63 IoU score. In contrast, the Distance IoU provides a suitable and reliable label vale. Another case in the right figure also shows that DIoU performs better than IoU and centerness with same IoU score but different spatial locations.

C. Discussion of Difference with Anchor-free Siamese Trackers.

- Existing trackers do not consider to estimate the uncertainty of box coordinates, so that it has no clear probabilistic interpretation. In contrast, we propose a novel formulation to learn general distribution of bounding box representation with uncertainty for visual object tracking, termed as distribution-based regression method.
- There is an usage inconsistent problem between the classification and quality estimation since the classification and localization quality are trained separately but combined during online tracking. Different from them, we devise a joint representation head to tackle this issue.
- Most existing tracking methods have a limitation of the task misalignment between classification and regression. Namely, the position with high classification score may not achieve high regression accuracy, and vice versa. Differently, benefiting from the proposed distributed regression framework, we propose to utilize the uncertainty information from box distributions to guide the learning of our joint representation head.

D. Discussion of Difference with GFL in Object Detection.

Both GFL (Li et al., 2020) and our method directly learn the joint representation of classification and localization quality. However, GFL is much more fit for object detection, while in visual tracking one and only one object should be tracked. Nevertheless, our work differs from GFL in four fundamental ways.

1). In GLF, the training samples of the classification and regression heads are identical. Both are sampled from the positions within the ground-truth boxes. The ambiguous matching between anchors and object severely hinders the robustness of tracker. Differently, our method is asymmetric which is tailored for visual tracking task. To be specific, the joint representation head only considers the pixels closing to the target center as positive samples, while the regression head considers all the pixels in the ground-truth box as training samples. This fine-grained sampling strategy guarantees the joint head can learn a robust similarity metric for localization, which is important for tracking.

2). GFL uses IoU score and Quality Focal Loss (QFL) to supervise the joint head. However, the IoU score may be not credible in some cases (see the figure 3), and QFL is also not suitable for single object tracking that belongs a binary classification problem. To this end, our supervision applies Distance-IoU (Zheng et al., 2020) between the predicted box and its ground-truth, which incorporates both normalized center distance and IoU score. This measurement is more suitable and effective in visual tracking. After that, we compare our loss function with others in Table 7 .

3). GFL qualitatively interprets the uncertainty according to the shape of distributions (e.g., sharp or flatten). However, more than that, our UAST can estimate the uncertainty using a quantitative values in $[0, 1]$. Specifically, as shown in Figure 1,

UAST provides the estimated certainty with respect to 4-directions (L_C, T_C, R_C and B_C), and the whole certainty value of the predicted box. For examples, UAST estimates lower right-directional certainty value (e.g., R_C: 61%) of the target due to the ambiguity caused by partly occlusion, confirming its effectiveness.

4). GFLV2 utilizes the statistics of bounding box distributions to perform localization quality estimation. To be specific, GFLV2 chooses the Top-k values along with the mean value of each distribution vector as the basic statistical feature, which is not representative and unstable in some bad cases. Differently, we select two nearneighbor values of prediction in each distribution as our initial localization quality features, providing a more simple, efficient yet effective method.

Loss type	$y > 0, P$	$y = 0, N$
FL	$-\alpha\|y-p\|^{\gamma}log(p)$	$-(1-\alpha)p^{\gamma}log(1-p)$
QFL	$\|y-p\|^{\gamma}\cdot\mathcal{L}_{BCE}$	$-p^{\gamma}log(1-p)$
QFLv2	$f(km)\cdot\|y-p\|^{\gamma}\cdot\mathcal{L}_{BCE}$	$-p^{\gamma}log(1-p)$
VFL	$y\cdot\mathcal{L}_{BCE}$	$-\alpha p^{\gamma}log(1-p)$
wBCE	$w^{+}\cdot\mathcal{L}_{BCE}$	$w^{-}\cdot\mathcal{L}_{BCE}$
Ours	$w^{+}\cdot f(nn)\cdot\mathcal{L}_{BCE}$	$w^{-}\cdot f(nn)\cdot\mathcal{L}_{BCE}$

Table 7. Comparison of different loss functions used in the classification or joint representation branch. y is target IoU between the predicted box and ground-truth. p denotes the predicted classification score, α is a weighting factor, and $\mathcal{L}_{BCE}$ means binary cross-entropy loss. $f(\cdot)$ denotes the different functions of localization quality estimation.

E. More Experimental Results on LaSOT.

In addition to the success and precision plots shown in the body part, we here provide the normalized precision plot over the LaSOT test set (Fan et al., 2019) containing 280 video sequences. The normalized precision score is computed as the percentage of frames where the normalized distance (relative to the target size) between the predicted and ground-truth target center location is less than a threshold D. It is plotted over a range of thresholds $D \in [0, 0.5]$. The trackers are ranked using the area under this curve, which is shown in the legend of the Figure 11. We compare with state-of-the-art trackers Ocean (Zhang et al., 2020), DiMP (Bhat et al., 2019), DROL (Zhou et al., 2020), SiamFC++ (Xu et al., 2020), SiamGAT (Guo et al., 2021), and etc. Our UAST outperforms previous state-of-the-art Siamese trackers. Compared to the ResNet-50 based Ocean-online, DiMP50 and SiamGAT, our approach achieves gains of 0.7%, 1.2% and 2.5% respectively.

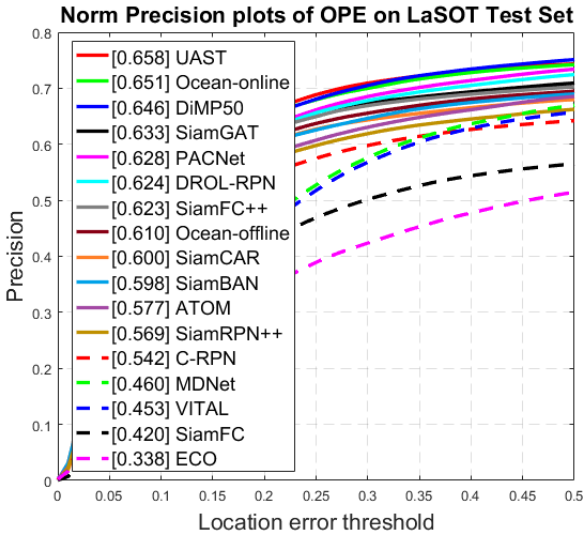


Figure 11. Normalized precision plot on the LaSOT test set. The average normalized precision is shown in the legend.

F. More Examples of Distributed Regression Representation.

We demonstrate more examples with our distributed bounding boxes predicted by UAST. As demonstrated in Figure 12, we show several cases with boundary ambiguities. In some cases, our model can produce more reasonable coordinates of bounding boxes. The predicted distributions are informative since its shape reflects the level of the certainty of the bounding boxes. The first three rows exist unclear boundaries, where distributions are flatten. The last row with clear boundaries and sharp distributions are shown, where is very confident to generate accurate bounding boxes.

Figure 12. Examples of distributed bounding box representation. The first three rows exist some boundary ambiguities and uncertainties, where the learned distributions may tend to be flatten. In some cases, we even observe a distribution with two peaks. Interestingly, they do correspond to ambiguous boundaries in the input image. For example, the top boundary of the airplane, the left boundary of the cat, and the left boundary of the deer. The last row has extremely clear boundaries, so that the learned distributions are relatively sharp and result in more reliable and accurate bounding box estimations. Predictions are marked yellow in images, while ground-truth boxes are green.

G. Quantitative Results

The representative quantitative results of our proposed UAST on the test set of GOT-10k dataset are shown in Figure 13. We also present the quantitative results of two representative state-of-the-art trackers: online learning based DiMP-50 and anchor-free SiamFC++ for a comparison. To be specific, Figure 2 demonstrates that DiMP50 and SiamFC++ may fail to track the targets in cases of partial occlusions, fast motion, scale variation and occlusion. In the third row sequence, DiMP50 and SiamFC++ drift from the moving animal in frame 38. Our UAST can locate the target accurately with more reasonable localization confidence thanks to our joint representation which integrate the classification and localization quality. In the firth sequences, UAST can quickly adapt to the great scale variations of the flying people.

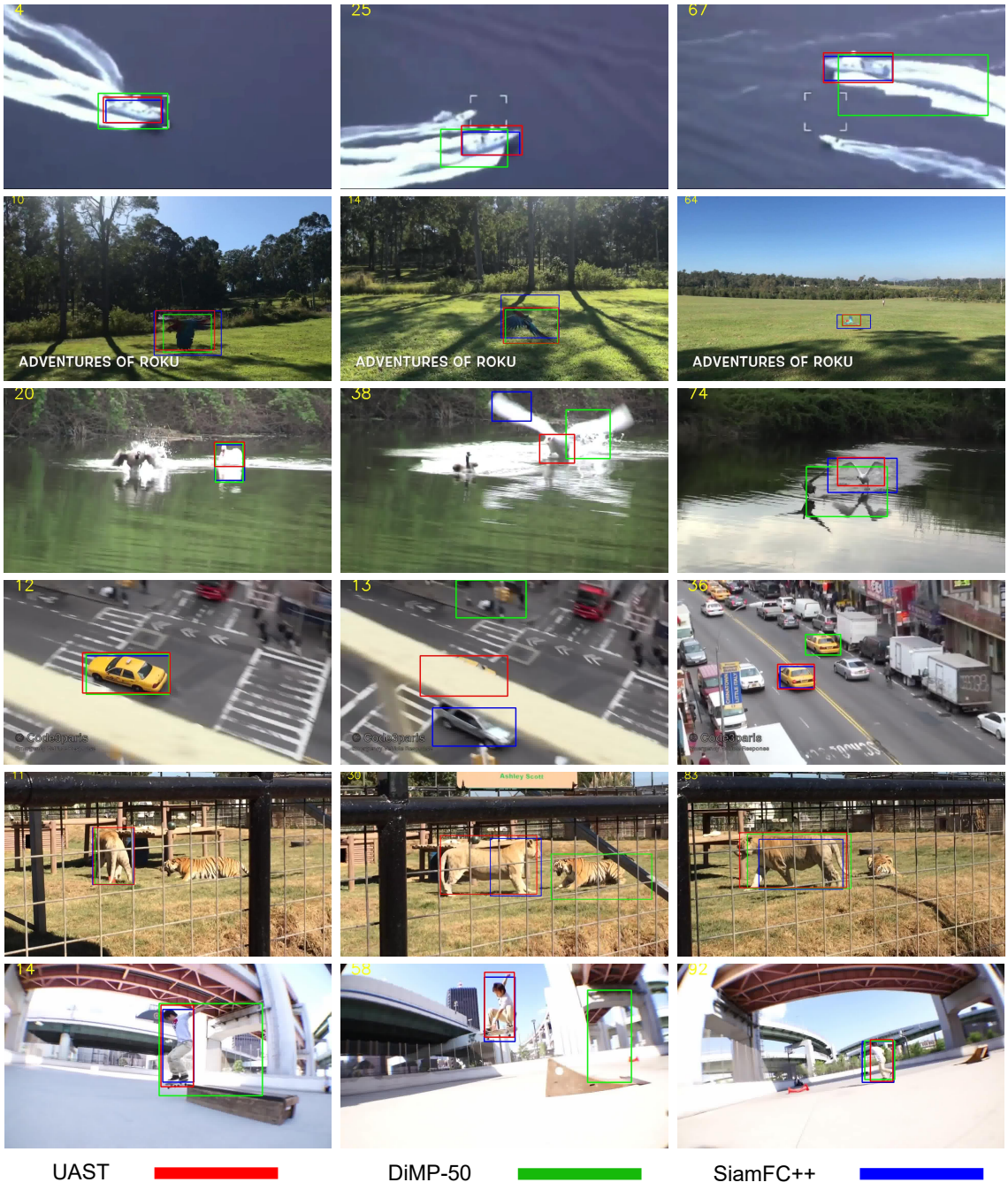


Figure 13. Quantitative result comparison of our UAST (red) with representative trackers DiMP-50 (green) and SiamFC++ (blue). Observed from the visualization results, UAST can estimate more precise bounding boxes when encountering circumstances of partial occlusions, deformation, scale changes and fast movement. This comparison demonstrates that the proposed distributed regression formulation is more effective because our method provides a clear interpretation of the boxes.