UAST: Uncertainty-Aware Siamese Tracking

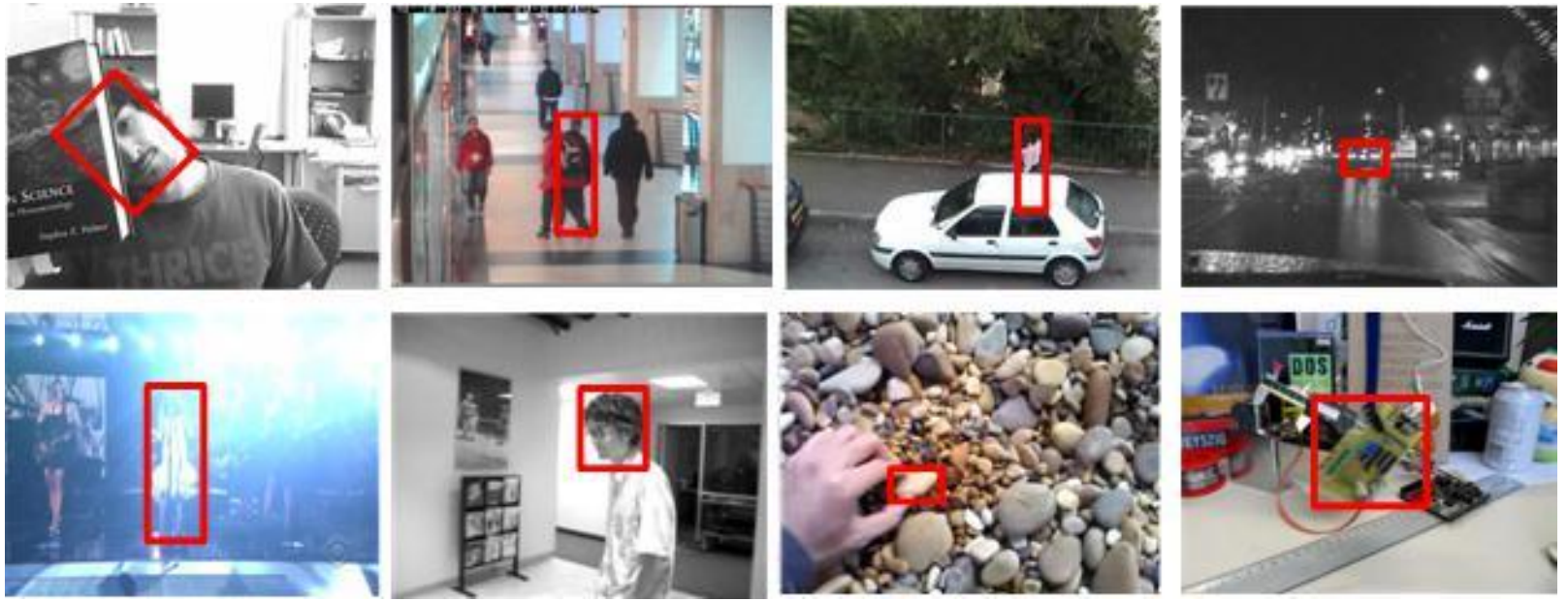
Dawei Zhang[1], Yanwei Fu[2], Zhonglong Zheng[1]

[1]Department of Computer Science, Zhejiang Normal University, China

[2]School of Data Science, Fudan University, Shanghai, China

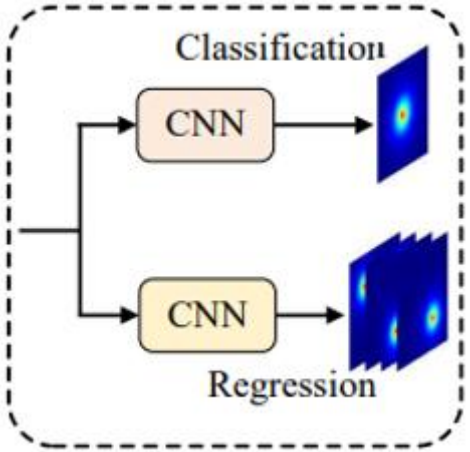
Introduction

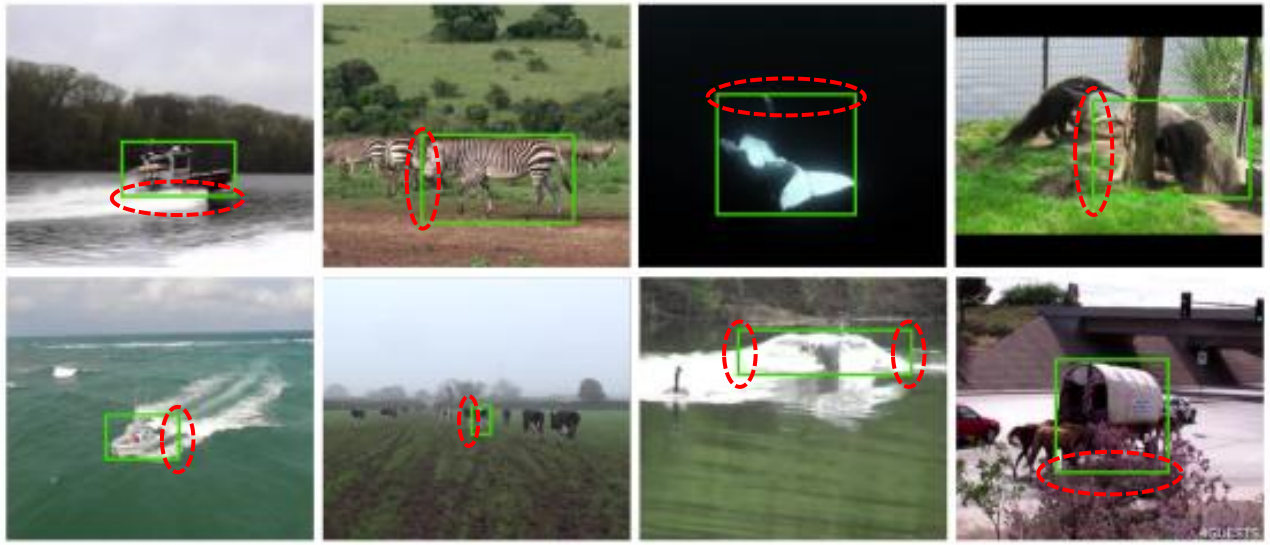
- **Single Object Tracking**
  - **Goal**: Track an arbitrary target in a video given the initial annotation.
  - **Challenges**:
    - Occlusion, Illumination variation, background clutter, etc.
    - Appearance changes, geometric deformation, scale variations.

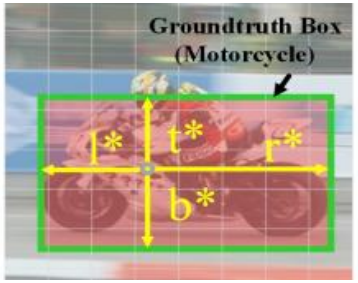


Motivations

- Visual object tracking is basically formulated as classification and bounding box regression.
- Recent popular anchor-free Siamese trackers rely on predicting the distances to four sides for efficient regression, but fail to estimate accurate box in complex scenes due to lacking of the uncertainty representation of bounding boxes.
- Another limitation of existing tracking methods is the misalignment between classification and regression (high classification score may not correspond high regression box).



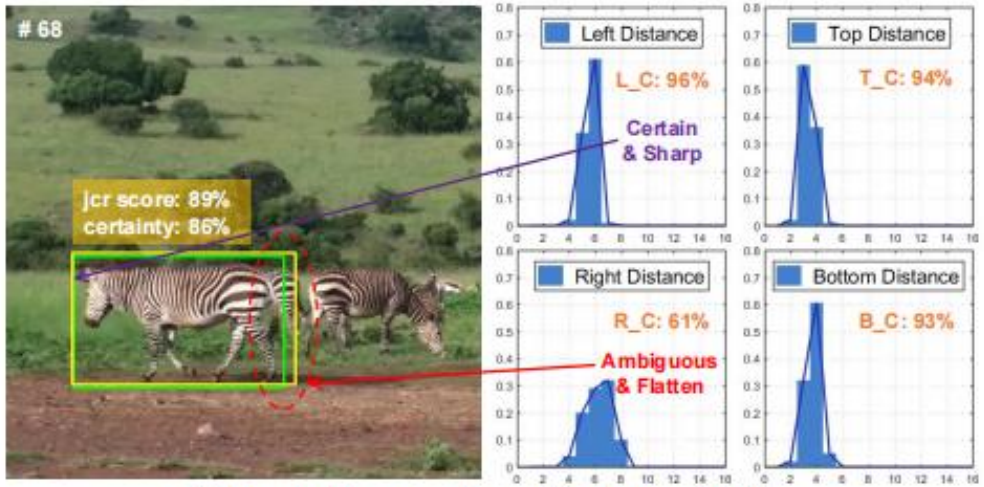




Contributions

- We propose a novel uncertainty-aware Siamese tracking method with a clear probabilistic explanation.
- We propose a novel distribution-based regression paradigm for visual tracking, which can flexibly capture more informative target boundaries, and provide the certainty value of each direction.
- Based on the learned distributions, we propose a simple yet effective joint representation head of classification and localization quality.
- UAST achieves state-of-the-art performance on five public tracking benchmarks, demonstrating its effectiveness and tracking efficiency.

(a) Different approaches for target box estimation



(b) A qualitative case of uncertainty-aware tracking

Uncertainty-Aware Siamese Tracking

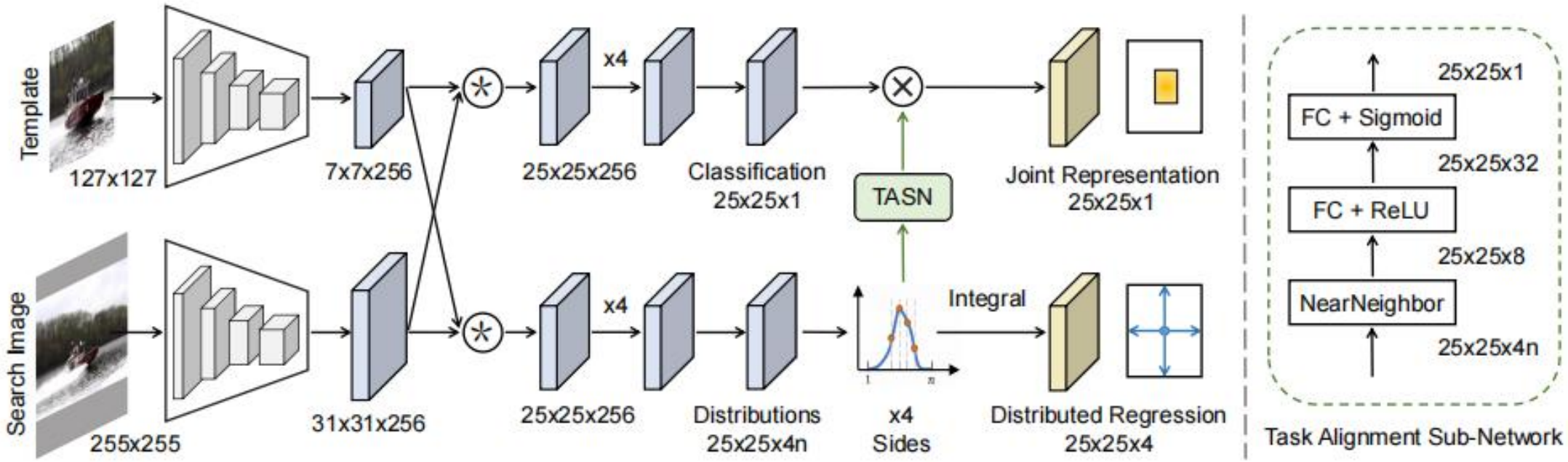


Figure 2. The main structure of the proposed Uncertainty-Aware Siamese Tracking framework. It consists of a backbone network for feature extraction, a feature matching module, an anchor-free head with distributional regression and joint representation, and a task alignment sub-network. Note that $\star$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

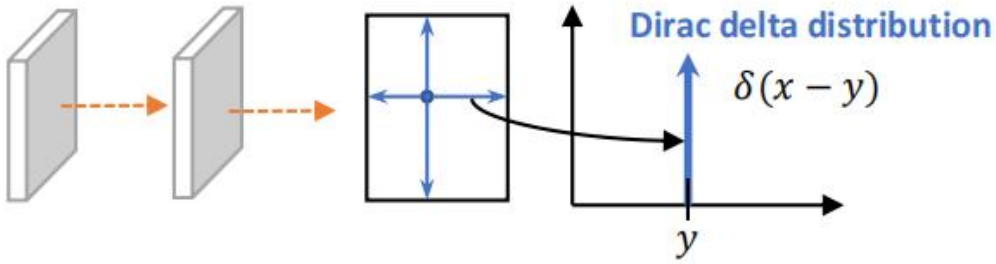
Distributed Regression

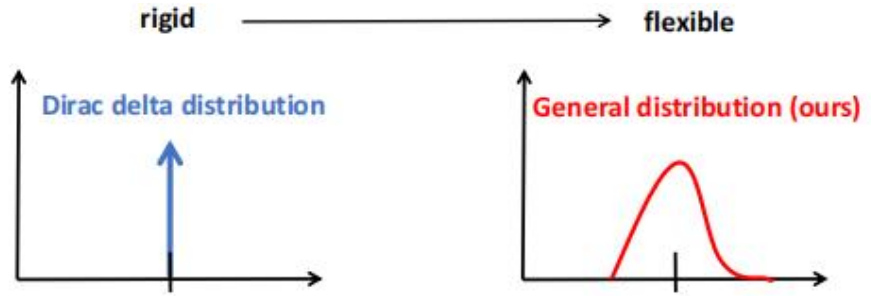
- From a distribution perspective of view, existing anchor-free trackers can be considered as a simple Dirac delta distribution, since the target is to fit a single label value.
- We propose to model a general distribution P(x), and calculate its integral for prediction:

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

- [ξ0, ξn] can be divided into a set [ξ0, ξ1, ξ2, ..., ξn−1, ξn] with even interval. We further consider to optimize the shape of distributions using DFL Loss.

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





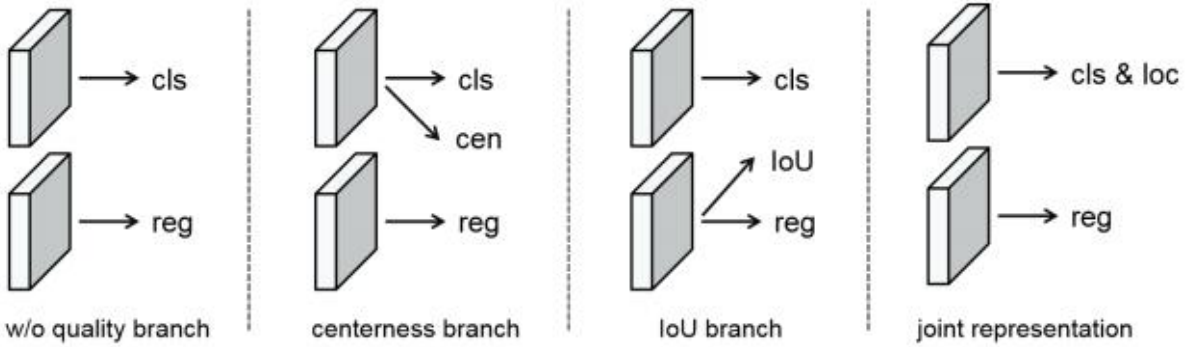
Joint Confidence Representation

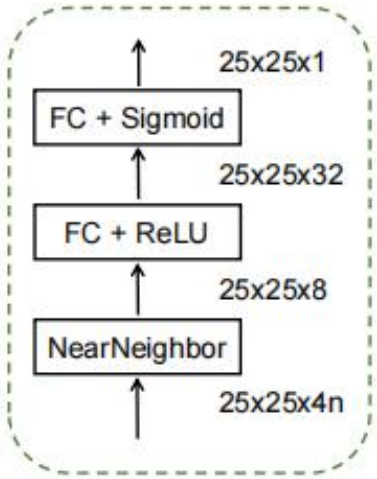
- Learning joint confidence representation of classification and localization quality.
- We further exploit the uncertainty in box distributions to perform task alignment, facilitating the learning of our joint confidence representation.

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

- Selecting nearneighbor values of predictions, and concatenate them as initial features, then using two FC ayers to obtain localization quality vector.

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

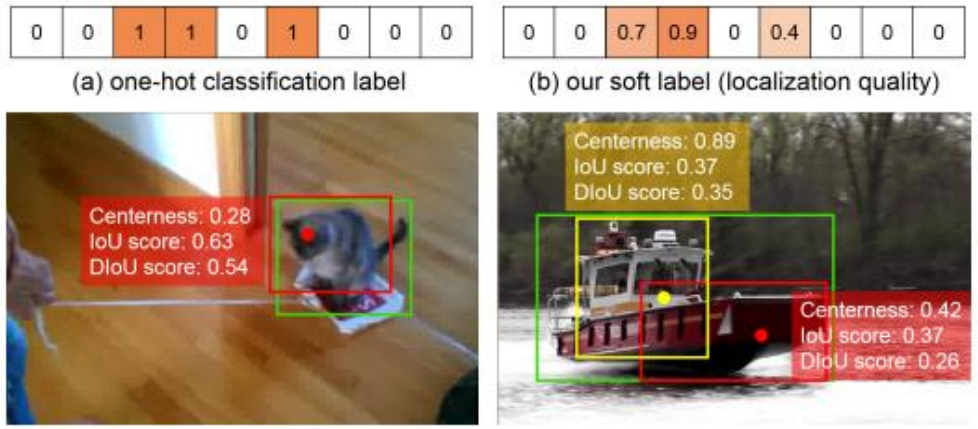




Task Alignment Sub-Network

Training Objective

For JCR, negative samples are still supervised by 0, while the supervision of positives is determined by the localization quality label (Distance-IoU)



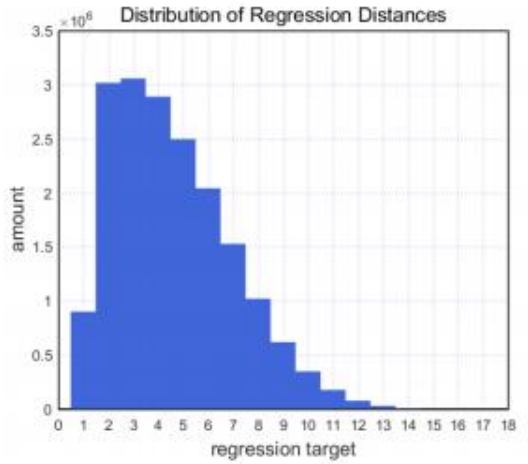
$$Y_+ = argmax(DIoU, 0),\ DIoU = IoU - \frac{\rho^2\left(b, b^{gt}\right)}{c^2}$$

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{dfl}$$

Algorithm 1 Uncertainty-Aware Siamese Tracking

Input: Frames $\{I_k\}_1^K$, initial target box B_1
Output: Target state $\{B_k\}_2^K$, 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

Ablation Study



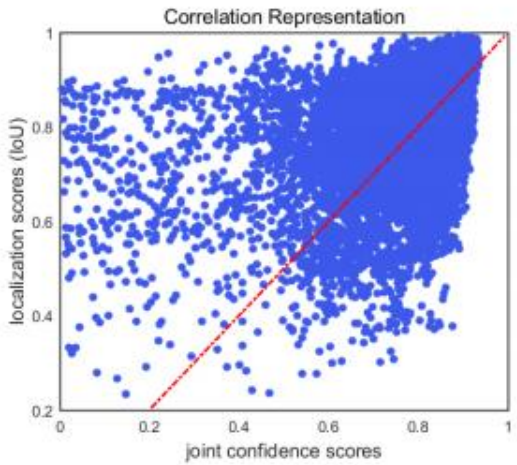


Table 1. Ablation experiments of different variants of UAST on GOT-10K test set, baseline is Ocean without object-aware branch.

	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

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

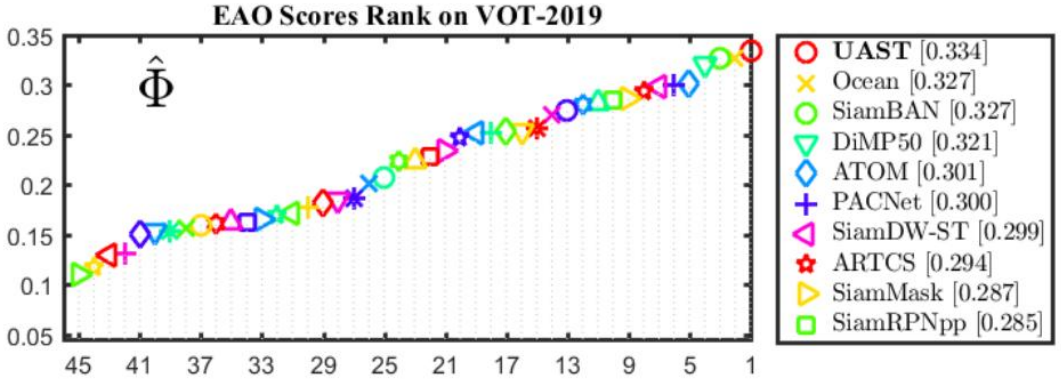
Table 3. Performances of various popular anchor-free Siamese trackers integrated by the proposed UAST on LaSOT test set.

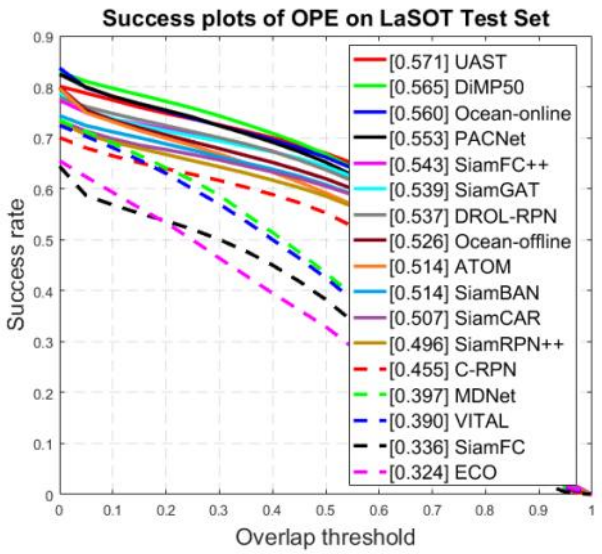
Tracker	Dis. Rep.	JCR	Success	Fps
SiamCAR	×	×	0.507	52
SiamCAR + UAST	√	√	0.543	52
SiamBAN	×	×	0.514	40
SiamBAN + UAST	√	√	0.548	40
SiamGAT	×	×	0.539	70
SiamGAT + UAST	√	√	0.567	70
Ocean	×	×	0.526	68
Ocean + UAST	√	√	0.571	68

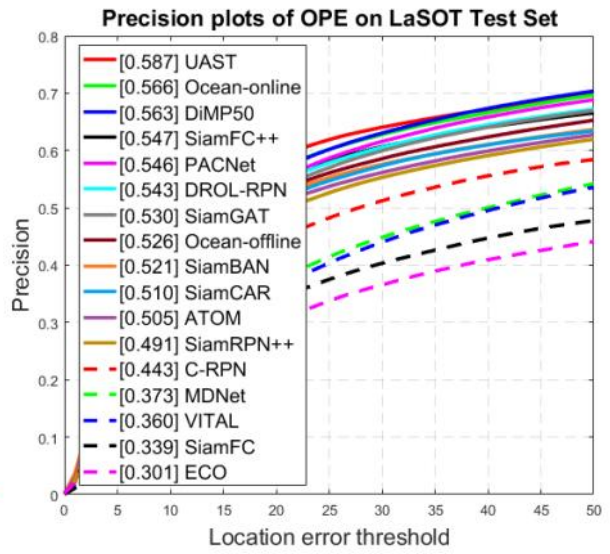
Comparison with State-of-the-arts

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

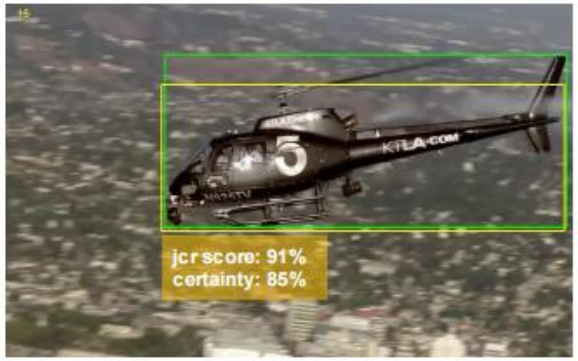
Trackers	AO	$SR_{0.5}$	$SR_{0.75}$
MDNet	0.299	0.303	0.099
ECO	0.316	0.309	0.111
SiamFC	0.374	0.404	0.144
SiamRPN++	0.517	0.616	0.325
ATOM	0.556	0.634	0.402
SiamCAR	0.569	0.670	0.415
SiamFC++	0.595	0.695	0.479
Ocean	0.592	0.695	0.473
D3S	0.597	0.676	0.462
DiMP50	0.611	0.717	0.492
LightTrack	0.623	0.726	-
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

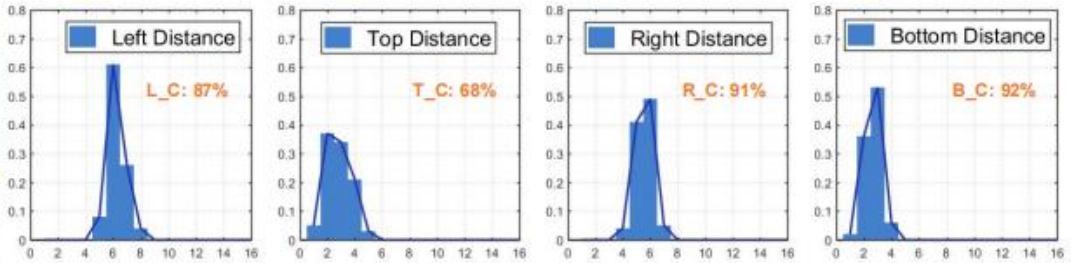


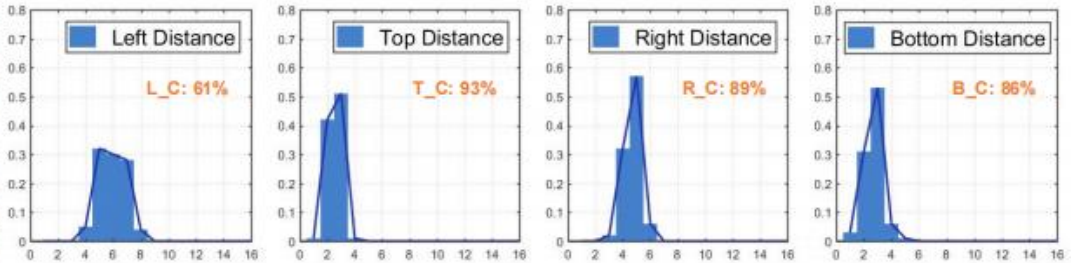


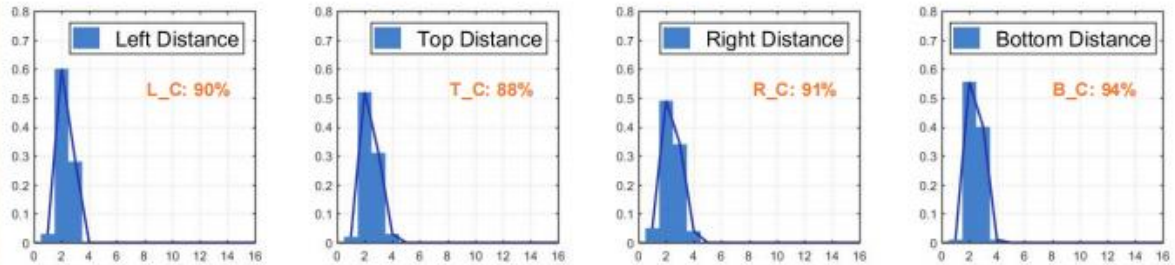


Qualitative Results

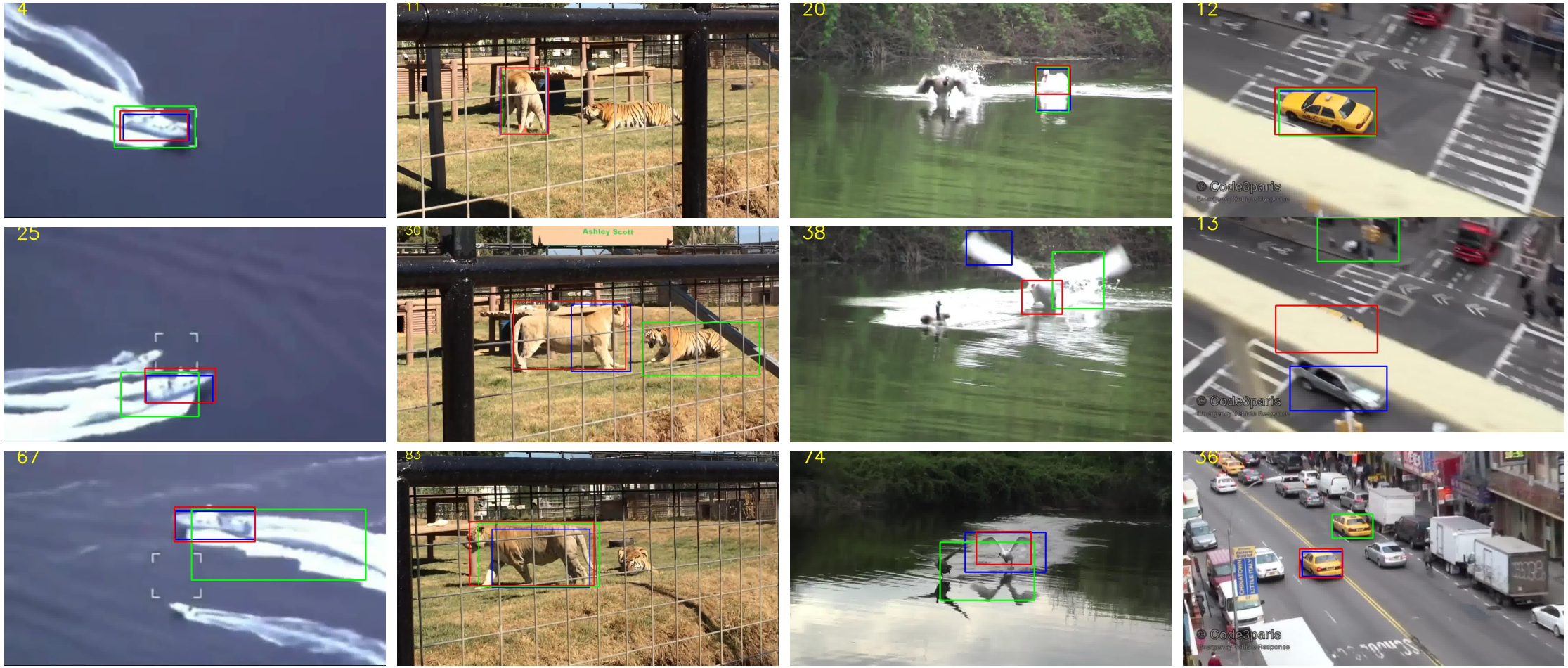








Qualitative Results



Conclusion

- In the paper, we propose to learn a distribution based regression for accurate tracking, which models localization uncertainty representation. It is an entirely new perspective in tracking community.
- 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.

Thanks !