Release the Power of Online-Training for Robust Visual Tracking

Yifan Yang,[1] **Guorong Li,**[*1,3] **Yuankai Qi,**[2] **Qingming Huang,**[*1,3,4]
[1]School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China.
[2]Harbin Institute of Technology, Weihai, China.
[3]Key Laboratory of Big Data Mining and Knowledge Management, CAS, Beijing, China
[4]Key Laboratory of Intell. Info. Process. (IIP), Inst. of Computi. Tech., CAS, China.
yangyifan@yeah.net, {liguorong, qmhuang}@ucas.ac.cn, qykshr@gmail.com

Abstract

Convolutional neural networks (CNNs) have been widely adopted in the visual tracking community, significantly improving the state-of-the-art. However, most of them ignore the important cues lying in the distribution of training data and high-level features that are tightly coupled with the target/background classification. In this paper, we propose to improve the tracking accuracy via online training. On the one hand, we squeeze redundant training data by analyzing the dataset distribution in low-level feature space. On the other hand, we design statistic-based losses to increase the inter-class distance while decreasing the intra-class variance of high-level semantic features. We demonstrate the effectiveness on top of two high-performance tracking methods: MDNet and DAT. Experimental results on the challenging large-scale OTB2015 and UAVDT demonstrate the outstanding performance of our tracking method.

Introduction

Visual tracking aims at predicting the trajectory of a target in an image sequence. Although much effort has been devoted in the past decades, it is still challenging to develop an efficient tracking method in the face of interfering factors such as heavy occlusions, out-of-plane rotations, and fast motion, etc.

Recent CNN-based trackers (Yan et al. 2019; Bertinetto et al. 2016; Nam and Han 2015; Qi et al. 2016; Pu et al. 2018; Bhat et al. 2018; Song et al. 2018; Qi et al. 2019) have shown the great success of hierarchical features learned by deep convolutional neural networks. Most of the existing trackers first pre-train a classification model using off-line data. Then the tracker is initialized via finetuning the learned models with the given target in the first frame. In the following frames, it is further finetuned with the collected samples, which is known as online training. However, the online training data is usually dominated by easy samples, and thus the model is prone to overfitting. Besides, the structure of the learned feature space is not always preserved. While the targets and the backgrounds lay close in the feature space, it is difficult to train a classifier to identify the targets.

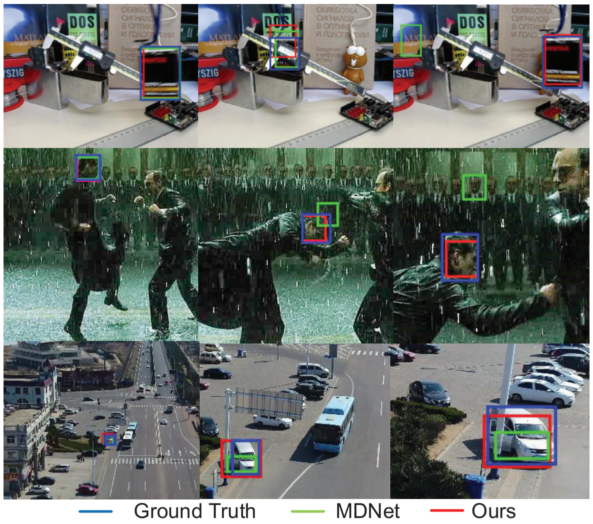


Figure 1: A comparison of our approach with the baseline MDNet on three example sequences. Trackers face four visual tracking challenges: large occlusion (top row), fast motion and illumination variation (middle row), and scale variations (bottom row). Through improved online-training, our approach achieves more accurate tracking results than MDNet.

To address the problems mentioned above, we propose a novel discriminative tracking method, which improves the online training method from two perspectives, i.e., delivering more compact and informative training data, and introducing statistic-based losses to obtain more discriminative features. To finetune tracker with a high-quality online training set, we propose a sample squeezing method to eliminate redundant samples. Different from existing approaches, the proposed method eliminates redundant samples by estimating the similarity of both old and new samples with novel proximity. The proposed proximity is more sensitive to the sample location in the feature space and is more capable of identifying the redundancy from a global perspective. As the proximity is updated efficiently, the sample squeezing method improves the diversity of the online-training set

[*]Corresponding author.

without much time-consuming. On the other hand, we propose statistic-based losses to obtain more stable and discriminative semantic features. The proposed statistic-based losses increase the inter-class distance of both classes, as well as decrease the intra-class variances. Through improved online-training, more discriminative features are learned for the classifier, and the tracker achieves more accurate tracking performance.

The main contributions of our method are summarized as follows:

- We propose a sample squeezing method to maintain the online-training set. The method delivers a more informative and compact training set by squeezing the redundant samples via a global perspective.
- We introduce statistic-based losses to preserve the structure of semantic feature space and obtain more discriminative features for the classifier.
- Extensive experimental results on both UAVDT and OTB100 datasets demonstrate the effectiveness of the proposed method.

Related Work

Our work is related to the previous approaches in two aspects. First, we summarize the sample management methods of online training. Then we overview the methods that improve the discriminative ability of trackers.

Online Training Sample Management

Existing tracking approaches employ a generative or a discriminative manner to track targets. In the training stage, the trackers learn embedding spaces for the objects, while in the reference stage, the trackers use various methods to maintain the embedding spaces.

The generative trackers do not online finetune the embedding spaces in the reference stage, and they linearly combine the known observations to formulate a target model. The DCF based trackers (Danelljan et al. 2016; Henriques et al. 2015; Valmadre et al. 2017a; Bertinetto et al. 2016) add newly collected target samples to the target models with a fixed and tiny step. Such updating strategy based on a strong assumption, i.e. the appearance is changing with a uniform speed. While there are heavy occlusions, out-of-plane rotations, or fast motion, the strong assumption degrades the discriminative ability of the target model and leads to the tracking failures. Several attempts are proposed to improve the online updating strategy. Bolme etc. (Bolme et al. 2010) propose a method to reject new samples based on the Peak-to-Sidelobe Ratio (PSR) criterion. This updating strategy evaluates the new samples rather than the whole training set and refuses the new target samples with significantly changed appearances. ECO (Danelljan et al. 2017) weights the collected target samples by their Euclid distances and the time tags then integrates the samples with a Mixed Gaussian Model. Although ECO considers all the training samples, as the Euclid distance only depicts the similarity of two samples, ECO still unable to describe the distribution of the whole training set. Furthermore, the ECO only formulates one target appearance model. The lack of negative training patches leads to over-fitting of the learned model, significantly affecting the performance in cases e.g., target deformations.

In the reference stage, the discriminative trackers finetune the embedding spaces to adapt to a specific video. Most discriminative trackers collect positive and negative samples along the tracking process. MDNet (Nam and Han 2015) employs a hard-negative mining technique to excavate valuable background samples. It also uniformly collects the training samples and keeps massive redundant ones.

In this paper, to facilitate discriminative trackers, we propose a sample squeezing method with neighbour proximity, which is more sensitive to the sample location in the feature space and capable to evaluate the distribution of the whole dataset.

Methods to Enhance the Discriminative Ability

Existing tracking methods employ several strategies to enhance the discriminative ability of trackers.

In the reference stage, the generative trackers obtain more representative features without finetuning the embedding spaces. SRDCF (Danelljan et al. 2015) weights the appearance models with kernels to diminish the influence of background and the unwanted boundary effects. However, it assumes the data are sampled correctly without spatial displacements and introducing noises to the representations. Danelljan et al. (Danelljan et al. 2017) integrate several kinds of features to improve the discriminative ability, but these features are not learned end-to-end and unable to adapt to the arbitrary object class. DSLT (Lu et al. 2018) and SiamFC-tri (Dong and Shen 2018) introduce the shrinkage loss and the triplet loss to the off-line training. Through online training, TADT (Li et al. 2019) selects the filters, which are more sensitive to the scale change. However, these trackers are still limited by the pre-trained model and lack the generalization ability to an unknown sequence.

Discriminative trackers use online training to learn more discriminative features and classifiers. MDNet (Nam and Han 2015) uses a multi-domain learning method to adapt the tracker to a specific video. DAT (Pu et al. 2018) generates spital attention maps to restrain the background features in the semantic domain, yet, needs abundant samples to obtain the attention maps. Vital (Song et al. 2018) erases the less significant area of the semantic features to promote the robustness of the classifier.

In this paper, we propose novel statistic-based losses to preserve the structure of the feature domain space and obtain more discriminative semantic features.

The Proposed Method

The framework of our proposed method is shown in Figure 2. The deep network first extracts the low-level features of input images. While the network goes deeper, we obtain the semantic features and deliver them to the classifier to tell the targets from backgrounds.

In the low-level feature space, we propose a sample squeezing method to enlarge the diversity of training set with novel proximity which is sensitive to the sample location.

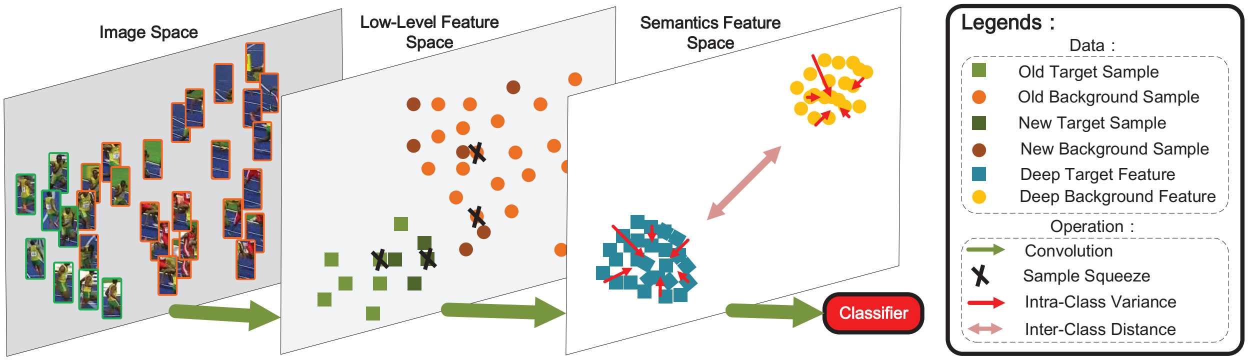


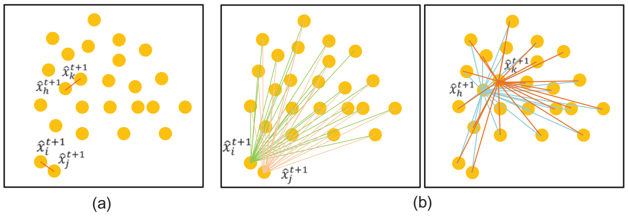
Figure 2: An overview of online-training stage of proposed tracker. We evaluate the training samples in the low-level feature space via the proposed neighbour proximity and eliminate the redundant samples to obtain more informative training set. In semantic feather space, we propose statistic-based losses to obtain more discriminative features for the classifier.

The proximity evaluates the distances among the samples in the low-level feature space, and the squeezing method eliminates the redundant ones.

To obtain more accurate classification results, we need more discriminative semantic features. Thus, in the semantic feature space, we propose the statistic-based losses to preserve the structure of the feature domain and employ the centers of each class to depict the feature distribution. The proposed losses diminish the intra-class distances among the samples and the centers to obtain more compact groups and enlarge the inter-class distance between groups at the same time. We elaborate on the proposed method in the following subsections.

Sample Squeezing

The moment of appearance change is unpredictable, and the distribution of the training samples is hard to estimate. Most existing methods assume that around the class center, the training samples distribute uniformly. Thus they finetune the network without evaluating the samples. However, the distributions are unbalanced in most cases. As a result, the sample centers stay close to a large number of similar samples and are insensitive to appearance changes.



(a)

(b)

Figure 3: Comparing of two proximities in a toy example. In Figure a), we evaluate the samples with the Euclid distance. In Figure b), we evaluate the samples with the neighbor proximity.

A toy example on the 2-dimensional plane is shown in Figure 3a, where a cluster of samples is the observations of the target. Sample pairs $\{\widehat{x}_i^{t+1}, \widehat{x}_j^{t+1}\}$ and $\{\widehat{x}_h^{t+1}, \widehat{x}_k^{t+1}\}$ are the redundant samples in different position of the cluster. Among the cluster, most samples have a similar appearance to the initial samples and are less informative. We name these samples global redundant samples. On the other hand, the rare samples at the edge of the cluster are more valuable observations. We propose the sample squeezing method to decide which sample is removed first. It is crucial for preserving valuable observations and enrich the diversity of the dataset.

Similarity Measurement To estimate the distribution of the dataset, we evaluate the similarities between the samples and propose novel proximity.

We denote the target and background training set in frame t as $\mathbf{T}^t$ and $\mathbf{B}^t$ respectively, where $\mathbf{T}^t = \{x_1^t, x_2^t, \cdots, x_n^t\}$ and $\mathbf{B}^t = \{s_1^t, s_2^t, \cdots, s_k^t\}$. Take the target training set for example, after the tracker predicts the target location in frame $t+1$, we add the newly collected samples of target $\{\widehat{x}_1^{t+1}, \widehat{x}_2^{t+1}, \cdots, \widehat{x}_m^{t+1}\}$ to the $\mathbf{T}^t$, and establish a temporary target set with $m+n$ elements, formulated as $\widehat{\mathbf{T}}^{t+1}$. In the following stage, we squeeze a selected redundant subset $\mathbf{R}^{t+1}$ with m elements to obtain $\mathbf{T}^{t+1}$. Before the prediction of next frame, we finetune the network with the $\mathbf{T}^{t+1}$ and $\mathbf{B}^{t+1}$.

To select $\mathbf{R}^{t+1}$, we first evaluate the similarities of samples in the Euclid space:

$$d^{\mathrm{I}}_{ij} = \|\widehat{x}_i^{t+1} - \widehat{x}_j^{t+1}\|_2^2. \tag{1}$$

The Euclid distance only reflects the similarities between the two samples and cannot depict the distribution of the whole dataset. To address this problem, we further evaluate the transmission distances of a sample to its neighbours, and introduce a transmit-vector:

$$t_i = (d^{\mathrm{I}}_{i1}, d^{\mathrm{I}}_{i2}, \cdots, d^{\mathrm{I}}_{ii-1}, 0, d^{\mathrm{I}}_{ii+1}, \cdots, d^{\mathrm{I}}_{il}). \tag{2}$$

In the semantic feature space, the transmit-vector reveals the relative location of a sample in the dataset.

To estimate the similarity between the transmit-vectors of the training samples, we propose neighbour proximity and formulate it as:

$$d^{\mathrm{II}}_{ij} = |t_i^j - t_j^i|, \tag{3}$$

where t_i^j denotes vector t_i with the j-th element set to 0.

As shown in Figure 3a, the Euclid distance is insensitive to the locations of the samples, and shows equal penalties while $\widehat{x}_i^{t+1}, \widehat{x}_h^{t+1}$ move towards $\widehat{x}_j^{t+1}, \widehat{x}_k^{t+1}$ respectively. However, moving from $\widehat{x}_i^{t+1}$ to $\widehat{x}_j^{t+1}$ decreases more neighbor proximity than moving from $\widehat{x}_h^{t+1}$ to $\widehat{x}_k^{t+1}$. Therefore, the neighbour proximity is more sensitive to the sample locations and capable of seeking redundancy from a global perspective.

Squeeze Redundancy With the the neighbor proximities among samples, we formulate the distance matrix $\mathbf{D}$ as:

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12}^{\mathrm{II}} & \cdots & d_{1l}^{\mathrm{II}} \\ d_{21}^{\mathrm{II}} & 0 & \cdots & d_{2l}^{\mathrm{II}} \\ \cdots & \cdots & \cdots & \cdots \\ d_{l1}^{\mathrm{II}} & d_{l2}^{\mathrm{II}} & \cdots & 0 \end{bmatrix}$$

The non-diagonal element with the smallest value indicates the most similar sample pair. To enrich the diversity of the training set, we eliminate the older sample in this pair. We repeat this operation m times until the size of the sample set equals to n again.

While squeezing the redundant samples from the data set, the data distribution varies correspondingly. While $\widehat{x}_j^{t+1}$ is eliminated, we update the $d_{ih}^{\mathrm{II}(k)}$ as:

$$d_{ih}^{\mathrm{II}(k)} = d_{ih}^{\mathrm{II}(k-1)} - |d_{ij}^{I} - d_{hj}^{I}|, \tag{4}$$

here, k denotes the k-th update. Then we delete the jth column and jth line of $\mathbf{D}^{(k)}$.

When the new samples arrive, the distance matrix expands on the contrary way. Thus, the sample squeezing method improves the diversity of training set without much time-consuming.

Statistic-based Losses

In visual tracking, the targets can be arbitrary objects. As such, the pre-trained deep features are less effective in distinguishing these targets, as the distribution of the target and background is also arbitrary in the domain of the deep feature. To depict the structure of the feature domain, we employ the centers of the target and background. We denote the center as $c = \frac{1}{n}\sum_{i=1}^{n} x_i$, and mark the centers of the target and background as c_t and c_b respectively. It's worth mentioning that in the training process, the centers are differentiable and updated with the randomly selected samples of training batches.

To achieve more discriminative features, we propose a loss term that favors larger distance between two centers:

$$L_d(c_b, c_t) = \frac{1}{|c_b - c_t|}. \tag{5}$$

We name this loss term as inter-class distance loss.

As the tracker is online-trained with sample batches, if we only increase the inter-class distance, it would tear apart the clusters. To make the clusters of samples to be compact, we propose a intra-class variance loss:

$$L_v(x_i, s_j, c_t, c_b) = \frac{1}{2m}\sum_{i=1}^{m} \|x_i - c_t\|_2^2 + \frac{1}{2n}\sum_{j=1}^{n} \|s_j - c_b\|_2^2, \tag{6}$$

here, x_i and s_j are the semantic features of a target sample and a background sample respectively.

The tracking-by-detection framework defines the target object as a positive class and the background as a negative class to train a binary classifier. We equip the binary classification layer with the softmax cross-entropy loss as the primary loss function. The cross-entropy loss function is presented as follows:

$$L_{ce}(p, y) = -(y * log(p) + (1-y)log(1-p)). \tag{7}$$

Where $y \in \{0, 1\}$ are the class labels, $p \in [0, 1]$ denotes the estimated probability for a class with label $y = 1$. Meanwhile, we define the probability for a class with label $y = 0$ as $1 - p$.

Then we add the inter-class distance and the intra-class variance losses to the cross-entropy loss with two scaler parameters λ_1 and λ_2, and the loss function is formulated as follows:

$$L = L_{ce} + \lambda_1 L_d + \lambda_2 L_v. \tag{8}$$

During training, the center of target set is updated as:

$$\varDelta c_t = \frac{1}{n}\sum_{i=1}^{n}(c_t - x_i). \tag{9}$$

With both the statistic-based losses, the online-training process forces the semantic features to be more compact and discriminative.

Experiments

In this section, we first present the implementation details. Then we evaluate our method on two standard benchmarks: OTB-2015 (Wu, Lim, and Yang 2015) dataset and UAVDT (Du et al. 2018) dataset. To analyze the effectiveness of each opponent in our method, we conduct ablation studies from four perspectives: 1) We evaluate our proposed sample squeezing method with three comparing experiments. 2) We assess the improvement coming up with the statistic-based losses. 3) We value our proposed online training method with another tracking-by-detection tracker. 4) We visualize the influences of the proposed losses on deep feature distribution.

Implementation Details

We utilize the same network architecture of MDNet to construct our baseline tracker. MDNet appends three fully connected layers to three trained convolution layers as the classifier. The network is pre-trained on sequences collected from VOT datasets (Felsberg et al. 2015), excluding the videos included in OTB2015. During tracking, the weights of the first three convolutional layers are fixed to deliver stable features. Features of the third layer are used to estimate the global data redundancy, and the statistic-based losses are added to the fifth layer. We significantly reduce the scale

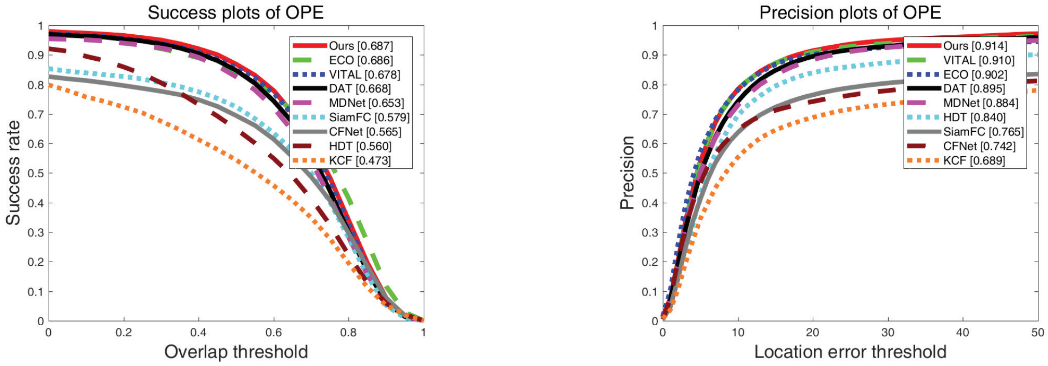


Figure 4: Overlap success and distance precision plots using the one-pass evaluation on the OTB-2015 datasets.

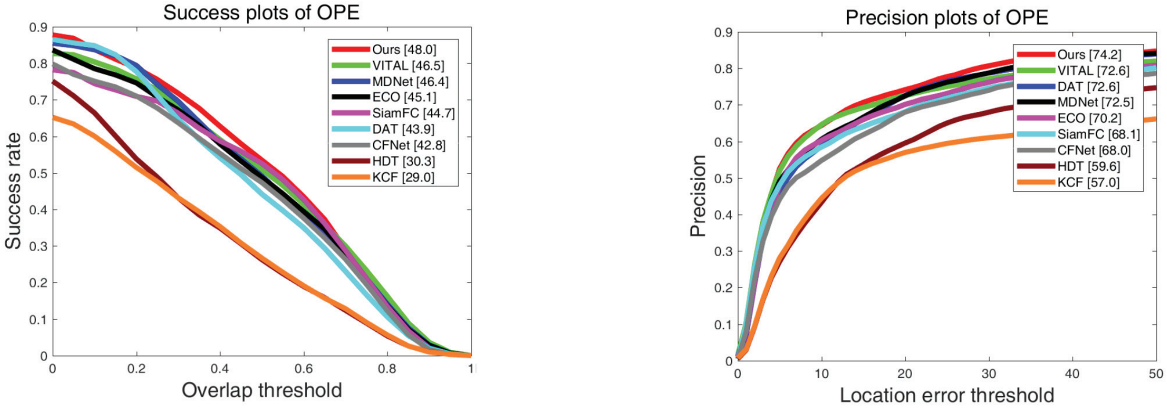


Figure 5: Overlap success and distance precision plots on the UAVDT dataset.

of training set: MDNet possesses 3000 online-training samples; we maintain only 150 samples. The learning rate of class centers is 2e-2, λ_1 is set to 1e-3, and λ_2 is set to 1e-2. We implement our tracker in Python using Pytorch (Paszke, Gross, and Lerer 2017) library. The implementation runs on an Intel Core i7-6700 3.4GHz CPU with 12GB of RAM and a GIGABYTE GTX 1080 Ti GPU with 11GB of VRAM, and the average speed is 1.0 FPS.

Evaluations on OTB2015

OTB is a popular tracking benchmark that contains 100 fully annotated videos with substantial variations. The evaluation is based on two metrics: center location error and bounding box overlap ratio. The one-pass evaluation (OPE) is employed to compare our algorithm with the eight state-of-the-art trackers including MDNet (Nam and Han 2015), ECO (Danelljan et al. 2017), DAT (Pu et al. 2018), Vital (Song et al. 2018), SiameseFC (Bertinetto et al. 2016), CFNet (Valmadre et al. 2017b), HDT (Qi et al. 2016) and KCF (Henriques et al. 2015).

We evaluate the tracking performance on OTB-2015 in both metrics and show the results in Figure 4. Our tracker outperforms all of them and shows a remarkable performance. Specifically, our method improves the baseline MDNet by a large margin, which is not off-line trained on auxiliary sequences for a fair comparison. The overall favourable performance of our tracker can be explained by the fact that the proposed algorithm strengthens the discriminative power of the tracker.

Evaluations on UAVDT

UAVDT is a benchmark for 3 foundational visual tasks including detection, multi-object-tracking, and single-object-tracking. The single-object-tracking dataset consists of 50 videos (40k frames) captured by the unmanned aerial vehicle (UAV) platform from complex scenarios. We evaluate eight state-of-the-art trackers, including MDNet, ECO, DAT, Vital, SiameseFC, CFNet, HDT, and KCF to compare with our algorithm.

As shown in Figure 5, our tracker outperforms all of them and shows a remarkable performance. UAV captures videos in relatively high altitudes, and the targets tend to be small. Thus, comparing with the tracking results on the OTB benchmark, the overlap success and the distance precision scores dramatically decrease. Our tracker achieves superior performance.

We show five example sequences in Figure 6. The five sequences contain different challenges coming up with UAV videos. The targets in these sequences are tiny, and the background is cluttered. It can be seen that the proposed tracker

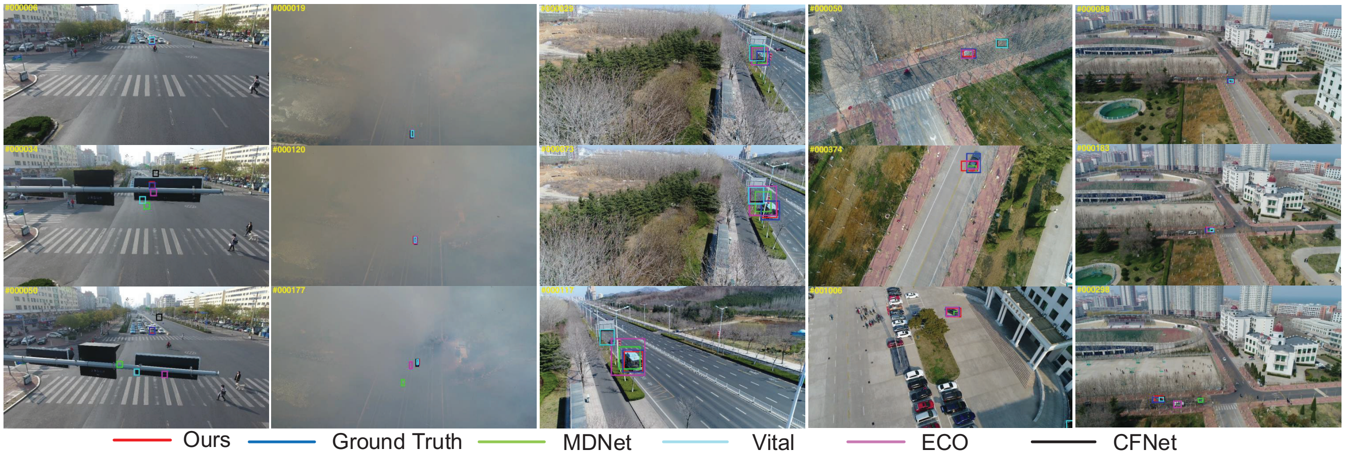


Figure 6: Qualitative results of the proposed method on UAVDT sequences. Trackers face five typical UAV challenges: long time occlusion (first column), unclear vision (second column), cluster background (third column), camera rotation (fourth column) and similar objects (fifth column).

Table 1: Evaluation results on OTB-2013. The comparing trackers are the baseline tracker MDNet, and the modified baseline trackers which are equipped with the sample squeezing method and the data augmentation respectively. ED stands for Euclid Distance, Np stand s for Neighbor Proximity, DA stands for Data Augmentation, while SLoss stans for Statistic-based Loss. Besides, DP stands for distance precise, and AUC stands for average overlap ratio score.

	MDNet	MDNet +ED	MDNet +NP	MDNet +DA	MDNet +SLoss +NP	MDNet +SLoss w/o NP
DP	0.924	0.925	**0.933**	0.921	**0.941**	0.935
AUC	0.682	0.680	**0.685**	0.660	**0.702**	0.690

is more robust than the other trackers. This phenomenon demonstrates that our online training can generate more discriminative features.

Ablation Study

In this section, we investigate how the proposed algorithm improves the tracking performance. We adopt MDNet and DAT as baseline trackers.

Evaluate Sample Squeezing Method We first evaluate our proposed neighbour proximity. We equip the squeezing method with the Euclid distance as a comparing method, then evaluate the performances of both squeezing approaches and show the results in Table 1. It can be seen that the squeezing method with the Euclid distance receives a performance degradation, due to it squeezes the dataset without increasing the diversity. As the neighbour proximity evaluate the dataset distribution from a global perspective, it enable the squeezing method to eliminate the global redundancy.

In a classification task, data augmentation is a frequently-used technique to enrich the diversity of limited training data. Thus, we first compare our proposed sample squeezing method with the data augmentation technique. We conduct data augmentation in the initial training. 30% of samples were randomly selected for the flip, rotation, and blur operation, and 10% samples were selected for dropout. We report the evaluation results on OTB-2013 (Wu, Lim, and Yang 2013) in Table 1. It can be seen that the MDNet tracker with the data augmentation receives a performance degradation. On the contrary, our sample squeezing method improves the baseline. Moreover, we achieve improvement with a small training set. The data augmentation does not simulate the target changes and introduces noise to the training set. This phenomenon affirms our method can enrich the diversity of the training set and is more suitable for visual tracking.

Moreover, we online train a baseline tracker with the statistic based losses, yet without the squeezing method. As shown in Table 1, in the metric of AUC, the performance of the modified tracker increases 1.17%, comparing to the performance of baseline. However, it is lower than the performance of the tracker with both statistic-based losses and squeezing method. This phenomenon affirms that our squeezing method efficiently eliminates the redundant samples, and facilitates the statistic losses to be more sensitive to the appearance changes. As the proposed squeezing method efficiently increases the diversity of training data, it can facilitate the CF-based trackers to establish a more discriminative target representations.

Evaluate Statistic-based Losses Next, we evaluate the performance of the statistic-based losses and adopt MDNet with sample squeezing as the baseline tracker. In the experiment, we add the intra-class variance and the inter-class distance losses sequentially to the classification loss and report the tracking performance on the OTB-2013 dataset in Table 2. As shown, in overlap ratiometric, the intra-class variance loss contributes 0.6% improvement to the overall performance, and the inter-class distance loss contributes 1.1%. Our method achieves the best performance among the comparing trackers.

Table 2: Evaluation results on OTB-2013 dataset. We adopt MDNet with sample squeezing as baseline. IV stands for the Intra-class Variation loss, and ID is short for the Inter-class Distance loss.

	Baseline	Baseline + IV	Baseline + IV + ID(**Ours**)
DP	0.933	0.939	**0.941**
AUC	0.685	0.691	**0.702**

Visualize the Influence of the Statistic-based Losses In the following experiments, we utilize the t-SNE method (Der Maaten and Hinton 2008) to visualize the semantics feature space. First, We conduct three online-trainings, and successively add cross-entropy loss, intra-class variance loss, and inter-class distance loss to the loss function.

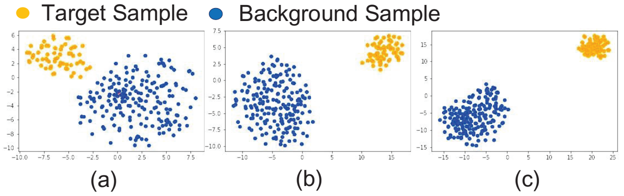


Figure 7: Visualisation of feature distribution in the semantic feature space. (a) depicts the feature distribution while tracker is trained by the original cross-entropy loss. (b) reveals the feature distribution when we add the intra-class variance to the loss function. (c) shows how the inter-class distance influences the distribution.

We use the 'Basketball' sequence in the OTB dataset as a test sequence, and visualize the feature spaces of the same training step in Figure 7. As shown in Figure 7a, the target and background samples remain in a pack while the network is trained without the statistic-based losses. As shown in Figure 7b, there is an apparent gap between target and background groups while the network is trained with the intra-class variance loss. As Shown in Figure 7c, the target and background samples are further separated by the two statistic-based losses. Comparing to the features learned by the softmax cross-entropy loss, the features learned by the proposed losses are more discriminative, which facilitate the classifier to be more robust.

Next, we use the 'DragonBaby' sequence as a test sequence to visualize the transformation of the feature space during online-training. We showed the results in Figure 8. In different iterations, the intra-class and the inter-class losses draw the samples progressively. As shown in Figure 8b, while the inter-class distance is increasing, the background cluster is temporally torn apart, and the intra-class variation rises correspondingly. After several training iterations, forcing by the intra-class variation loss, the cluster is compact again. These experiments not just affirm the effectiveness of proposed statistic-based losses, but also confirm that there is strong complementarity between the two proposed losses.

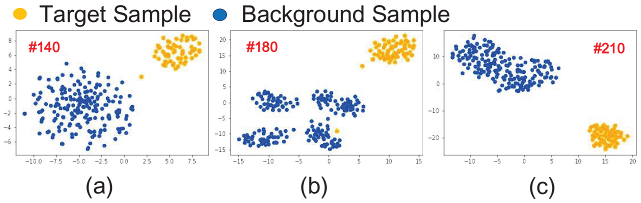


Figure 8: Visualisation of feature distribution during training. The red number in each subfigure denotes the training iteration. Online training is conducted from left to right.

Evaluate the Proposed Framework Then, we apply the improved online-training to DAT to affirm its efficiency. As evaluated in OTB-2013 and shown in Table 3b, the modified DAT achieves a slight increase in both metrics. As DAT trains an attention-aware classifier with sufficient training data, our compact training set unavoidably degrades the performance.

Table 3: Evaluation results on OTB-2013 dataset. The comparing trackers are DAT and DAT equipped with sample squeezing and statistic-based losses

	DAT	DAT + sample squeezing + statistic-based losses
DP	0.944	**0.952**
AUC	0.704	**0.705**

The experiments above confirm the efficiency of our proposed method and show that the squeezing method facilitates the statistic-based losses to obtain further improvement. Thus, it is essential for the online-training methods to maintain the distribution of both the training set and the semantic feature space.

Conclusions

In this paper, we propose a novel discriminative tracker and improve tracking performance through online-training. To avoid tracker overfit to easy samples, we propose a sample squeezing method to deliver a more compact and informative online-training set. More specifically, we evaluate the global redundancy of training set with neighbour proximity and squeeze redundant samples with an efficient update model. Moreover, to preserve the structure of classes in semantics feature space, we introduce the class centers to the objective function. By increasing the inter-class distance and decreasing the intra-class variances of both classes, we control the impact of dynamic data distribution and improve the accuracy of the tracker. Experiments on two datasets demonstrate state-of-the-art performance. In future work, we will apply our proposed method to the CF based trackers.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China: 61620106009, 61772494, 61931008, U1636214,61836002, 61976069 and 61902092, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013, in part by Youth Innovation Promotion Association CAS, in part by the University of Chinese Academy of Sciences, in part by “the Fundamental Research Funds for the Central Universities” (Grant No.HIT.NSRIF.2020005).

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. S. 2016. Fully-Convolutional Siamese Networks for Object Tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bhat, G.; Johnander, J.; Danelljan, M.; Khan, F. S.; and Felsberg, M. 2018. Unveiling the Power of Deep Tracking. *IEEE Computer Society Conference on The European Conference on Computer Vision (ECCV)* 11206 LNCS:493–509.

Bolme, D. S.; Beveridge, J. R.; Draper, B. A.; and Lui, Y. M. 2010. Visual object tracking using adaptive correlation filters. *IEEE Computer Society Conference on The European Conference on Computer Vision (ECCV)* 2544–2550.

Danelljan, M.; Hager, G.; Khan, F. S.; and Felsberg, M. 2015. Learning spatially regularized correlation filters for visual tracking. *IEEE Computer Society Conference on International Conference on Computer Vision(ICCV)* 4310–4318.

Danelljan, M.; Robinson, A.; Khan, F. S.; and Felsberg, M. 2016. Beyond correlation filters: Learning continuous convolution operators for visual tracking. *IEEE Computer Society Conference on The European Conference on Computer Vision (ECCV)* 9909 LNCS:472–488.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2017. Eco: Efficient convolution operators for tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6931–6939.

Der Maaten, L. V., and Hinton, G. E. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9:2579–2605.

Dong, X., and Shen, J. 2018. Triplet loss in siamese network for object tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 472–488.

Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; and Tian, Q. 2018. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. *IEEE Computer Society Conference on The European Conference on Computer Vision (ECCV)* 1–17.

Felsberg, M.; Berg, A.; Hager, G.; Ahlberg, J.; Kristan, M.; Matas, J.; Leonardis, A.; Cehovin, L.; Fernandez, G.; Vojir, T.; et al. 2015. The thermal infrared visual object tracking vot-tir2015 challenge results. *IEEE Computer Society Conference on International Conference on Computer Vision(ICCV)* 639–651.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2015. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)* 37(3):583–596.

Li, X.; Ma, C.; Wu, B.; He, Z.; and Yang, M. 2019. Target-aware deep tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lu, X.; Ma, C.; Ni, B.; Yang, X.; Reid, I. D.; and Yang, M. 2018. Deep regression tracking with shrinkage loss. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 369–386.

Nam, H., and Han, B. 2015. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*.

Paszke, A.; Gross, S.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Pu, S.; Song, Y.; Zhang, H.; and Yang, M.-h. 2018. Deep Attentive Tracking via Reciprocative Learning. *Neural Information Processing Systems(NIPs)*.

Qi, Y.; Zhang, S.; Qin, L.; Yao, H.; Huang, Q.; Lim, J.; and Yang, M.-H. 2016. Hedged Deep Tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4303–4311.

Qi, Y.; Zhang, S.; Zhang, W.; Su, L.; Huang, Q.; and Yang, M. 2019. Learning Attribute-Specific Representations for Visual Tracking. *The Thirty-Third AAAI Conference on Artificial Intelligence(AAAI)* 8835–8842.

Song, Y.; Ma, C.; Wu, X.; Gong, L.; Bao, L.; Zuo, W.; Shen, C.; Lau, R.; and Yang, M.-H. 2018. VITAL: VIsual Tracking via Adversarial Learning. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)* 8990–8999.

Valmadre, J.; Bertinetto, L.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. S. 2017a. End-to-end representation learning for correlation filter based tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5000–5008.

Valmadre, J.; Bertinetto, L.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. S. 2017b. End-to-end representation learning for Correlation Filter based tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1(7).

Wu, Y.; Lim, J.; and Yang, M. H. 2013. Online object tracking: A benchmark. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)* 2411–2418.

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

Yan, B.; Zhao, H.; Wang, D.; Lu, H.; and Yang, X. 2019. ’skimming-perusal’ tracking: A framework for real-time and robust long-term tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.