Soft labeling with quasi-Gaussian structure for training samples of deep classification trackers

Yan Peng¹, Jiantao Gao¹, Chang Liu¹, Xiaomao Li¹, Baojie Fan², Jiahong Chen¹, Jun Luo¹, Shaorong Xie¹ and Huayan Pu¹

Abstract
Deep classification tracking aims at classifying the candidate samples into target or background by a classifier generally trained with a binary label. However, the binary label merely distinguishes samples of different classes, while inadvertently ignoring the distinction among the samples belonging to the same class, which weakens the classification and locating ability. To cope with this problem, this article proposes a soft labeling with quasi-Gaussian structure instead of the binary labeling, which distinguishes the samples belonging to different classes and the same class simultaneously. Like as the binary label, the signs of labels for target and background samples are set to be plus and minus respectively to distinguish samples of different classes. Further, to exploit the difference among samples in the same class, the label values of samples in the same class are designed as a monotonically decreasing quasi-Gaussian function about Intersection over Union. Therefore, the corresponding response function is a two-piecewise monotonically increasing quasi-Gaussian combination function about Intersection over Union. Due to such response function, deep classification tracking trained with this proposed soft labeling achieves better classification and location performance. To validate this, the proposed soft labeling is integrated into the pipeline of the deep classification tracker SiamFC. Experimental results on OTB-2015 and VOT benchmark show that our variant achieves significant improvement to the baseline tracker while maintaining real-time tracking speed and acquires comparable accuracy as recent state-of-the-art trackers.

Keywords
Object tracking, deep classification tracking, soft labeling, Intersection over Union (IoU)

Introduction
Visual object tracking, one of the most important tasks in many robot applications, has been widely used in many fields such as intelligent manufacturing, human–computer interaction, video surveillance, and robotics.¹ ³ It is an indispensable part of robots¹ ⁴ serving as the “eye” for robots to communicate with the world as the Figure 1 shown.

Visual object tracking mainly contains five components: motion model, feature extractor, observation model, model updater, and ensemble post-processor, where feature

¹ Research Institute of USV Engineering, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
² Automation College, Nanjing University of Posts and Telecommunications, Nanjing, China

Corresponding author:
Xiaomao Li, Research Institute of USV Engineering, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China.
Email: lixiaomaosia@163.com
extractor has the greatest impact on visual tracking performance.\textsuperscript{6} Due to the powerful representation of deep networks, deep object tracking\textsuperscript{7–12} has become the research hotspot and the state-of-the-art algorithm in the field of visual tracking. Generally, deep object tracking can be divided into two main categories: deep regression tracking\textsuperscript{7,8,13–15} and deep classification tracking\textsuperscript{9–11,16}. Deep regression tracking outputs a response map through a regressor that learns a mapping between input deep features and the soft label. Deep classification tracking treats object tracking as an object and background two-category problem based on deep features, which classifies the samples into target or background through a classifier usually trained with the binary labeling. Recently, with the development of deep classification tracking, it has been able to achieve the real-time while ensuring certain tracking performance.

Other than Gaussian soft label of deep regression tracking, the label used in deep classification tracking is the binary label $\{-1, +1\}$ or $(1,0),(0,1)$. The samples with Intersection over Union (IoU) values greater than the threshold are considered as target samples, whose labels are set to $+1$ or $(1,0)$. The other ones are considered as background samples, whose labels are set to $-1$ or $(0,1)$. Although such binary label has the ability to distinguish the samples of different classes but inadvertently overlooks the difference among samples in the same class. This drawback makes the response map of deep classification tracking difficult to accurately reflect the target location. As shown in Figure 2(a), Deep classification tracker SiamFC\textsuperscript{11} trained with the binary labeling can discriminate between target and background samples, but the maximum position of its response map does not correspond to the target position accurately, which results in the target drift problem. As the tracking phase advances, the drift will accumulate and affect the subsequent frames. What’s more, neglecting the difference among samples within the same class in the training phase will reduce the classification ability of the tracker. As shown in Figure 2(b), SiamFC trained with the binary labeling misjudges the target and background samples due to such information neglect.

How to design a special labeling to solve the above problem? As we know, IoU characterizes the overlap rate between samples and the target, which can represent the probability of samples as the target to some extent. Inspired by this, this article uses the IoU values as the design criteria...
and proposes a novel soft labeling with quasi-Gaussian structure instead of the binary labeling to distinguish samples belonging to different classes and the same class simultaneously. Thus deep classification tracking trained with this proposed soft labeling performs better classification and locating ability as shown in Figure 2.

In the rest of this article, related work is introduced in the second section. The third section describes the proposed soft labeling with quasi-Gaussian structure and applies it to the deep classification tracker SiamFC. Then we compare and analyze the variant with its baseline tracker and the state-of-the-art trackers on popular tracking benchmarks: OTB-2015 and VOT, in the fourth section. Lastly, we conclude this article in the fifth section.

The main contributions of this article are summarized below:

1. A novel soft labeling with quasi-Gaussian structure is proposed to replace the binary labeling to enhance the classification and locating ability of deep classification tracking. The proposed labeling further solves the shortcomings of the binary labeling that ignores the distinction among the samples belonging to the same class.

2. A tracking algorithm is proposed to incorporate the soft labeling with quasi-Gaussian structure with the tracker SiamFC. Compared with the baseline tracker SiamFC, the proposed method achieves the significant improvement in terms of both accuracy and robustness on many existing popular benchmarks.

3. Extensive experiments on OTB-2015 and VOT benchmark against many state-of-the-art trackers are performed, and tracking results demonstrate the superiority and efficiency of our proposed tracking algorithm.

Related work

In 2012, AlexNet\textsuperscript{17} won the ILSVRC-2012\textsuperscript{18} competition and showed the powerful representational capabilities of deep features to the world. Since then, deep object tracking\textsuperscript{7,9,15,16,19} has emerged, which makes the field of visual object tracking a leap. Deep object tracking replaces manual features\textsuperscript{20,21} with the more powerful deep features as representation and achieves more remarkable performance than traditional object tracking\textsuperscript{22–27} According to the different nature, deep object tracking can be classified into two main categories: deep regression tracking and deep classification tracking.

Deep classification tracking

Deep classification tracking treats object tracking as a target and background two-category problem. It classifies the samples into target or background through a classifier usually trained with a binary label. Deep classification tracking mainly includes SVM-based deep classification trackers\textsuperscript{16}, deep classification trackers based on multi-domain convolutional neural networks\textsuperscript{9,10,33,34} and deep classification trackers based on the Siamese networks\textsuperscript{11,12,35–38} SVM-based deep classification trackers directly adopt R-CNN\textsuperscript{49} a convolutional neural network pre-trained on the multi-classification task dataset as the feature extractor and classify the samples into the target and background through the binary classifier SVM. Different from SVM-based deep classification trackers that can hardly benefit from end-to-end training, deep classification trackers based on multi-domain convolutional neural networks utilize the multi-domain convolutional neural networks as features extractor and binary classifier simultaneously to process the tracking task, which makes the end-to-end training possible. But to acquire the information about specific target and scenarios, they need to fine-tune the network online, which makes it difficult to achieve the real-time. Other than online fine-tuning, deep classification trackers based on Siamese networks obtain the specific information through
the Siamese networks. Deep classification trackers based on Siamese networks utilize the Siamese networks to convert the target and samples to the same embedding space and then classify samples into target or background by similarity comparison. The early deep classification tracker based on Siamese networks SINT\(^{15}\) has excellent tracking performance, but it is still far from being real-time due to the full connection layer and online update. Distinct from SINT, SiamFC\(^{11}\) adopts a fully convolutional Siamese network and no longer update the neural network online so that its real-time (86.5 Fps) reaches the first place in the deep classification trackers at that time while simultaneously guaranteeing a certain tracking accuracy. Therefore, recently more and more deep classification trackers\(^{12,36–38}\) have been improved on SiamFC so as to achieve high real-time while ensuring the certain tracking accuracy. In general, with the development of the deep classification tracking, it has been able to achieve a good balance between tracking performance and the real-time, and have achieved the start-of-state results.

However, we note that the binary labeling for deep classification trackers distinguishes the difference among samples in different classes but inadvertently elides the difference among samples within the same class. The neglect of the difference among the target samples makes the response values of the target samples difficult to accurately reflect the target position and causes the target drift problem. What’s more, due to such information neglect in the training phase, the classification ability of the deep classification tracking weakens and the misjudgment arises. To cope with problems of the binary labeling in deep classification tracking, this article proposes a soft labeling with quasi-Gaussian structure instead of the binary labeling to enhance the classification and locating ability of the deep classification trackers. Compared with the binary labeling, the soft labeling with quasi-Gaussian structure adds more information about the difference among samples within the same class into the training phase while considering the difference among the samples in different classes simultaneously.

### Soft labeling with quasi-Gaussian structure for deep classification tracking

We firstly describe the problems of the binary labeling and then propose a soft labeling with quasi-Gaussian structure for deep classification tracking. Lastly, we integrate the soft labeling into the pipeline of the deep classification tracker SiamFC to validate it.

### Problems in the binary labeling for deep classification tracking

There are two kinds of binary labels for deep classification tracking, namely \{-1, +1\} and \{(1,0),(0,1)\}. Deep classification trackers only outputting positive scores of samples\(^{11,12,35–38}\) generally adopt the \{-1, +1\} binary label while those outputting 2-D binary classification score\(^{9,10,33,34}\) adopt the \{(1,0),(0,1)\} binary label, which is shown in the Figure 3(a) and (b). Moreover, as the Figure 3(c) shows, these two kinds of binary labels are essentially the same. For simplicity, we adopt the \{-1, +1\} binary label as representation for the problem description.

---

**Figure 3.** (a) Deep classification trackers only outputting positive scores of samples. (b) Deep classification trackers simultaneously outputting positive and negative scores. (c) The conversion method between the deep classification trackers trained with the \{-1, +1\} and \{(1,0),(0,1)\} binary label.
The logistic loss function corresponding to the \( \{-1,+1\} \) binary label is expressed as following:

\[
L(v_i) = \log \left( 1 + e^{-v_i y_i} \right)
\]

where \( y_i \) and \( v_i \) denotes the label value and the response value of the sample \( x_i \) respectively. Denoting \( y_i v_i \) as \( t_i \), then the logistic loss function \( L \) can be expressed as

\[
L(t_i) = \log(1 + e^{-t_i})
\]

Theoretical derivation and experiment (see Appendix 1 for details) indicate that \( t_i \) will approximately converge to a constant \( c \). Hence the response value \( v_i \) of target samples and background samples will converge to \( c \) and \( -c \), respectively. As shown in Figure 4, although the response value can distinguish target samples and background samples, the samples belonging to the same class cannot be distinguishable. Such disadvantage will result in the target drift problem and weakens classification ability in the tracking phase.

**Soft labeling with quasi-Gaussian structure for deep classification tracking**

To overcome the drawbacks of the binary labeling, we propose a soft labeling with quasi-Gaussian structure instead of the binary labeling to enhance the classification and locating ability of deep classification tracking. The proposed soft labeling takes into account the difference among samples belonging to the same and different classes simultaneously. Like as the binary label, to distinguish samples of different classes, the signs of labels for the positive and negative samples are set to be plus and minus respectively. Further, to exploit the difference among samples in the same class, the label values of different samples belonging to the same class are no longer the same but related to their IoU values.

As analyzed above, \( t_i = y_i v_i \) will converge to a constant \( c \) and the response values \( v_i \) are inversely proportional to the label values \( y_i \). In order to make the response value of the samples further representing the probability as target to distinguish samples in the same class, the label value should be designed to be inversely correlated with the probability as target. As we know, IoU characterizes the overlap rate between samples and the target, which can represent the probability as the target to some extent. Therefore, as shown in equation (3), the proposed soft labeling is designed as a two-piecewise continuous quasi-Gaussian combination function about IoU to distinguish samples belonging to different classes and the same class simultaneously

\[
y_i = \begin{cases} 
-\alpha_p \frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{(I_i - \mu_p)^2}{2\sigma_p^2}} + \lambda_p & 0 \leq I_i \leq 1 \\
-\alpha_n \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(I_i - \mu_n)^2}{2\sigma_n^2}} + \lambda_n & 0 \leq I_i < \theta 
\end{cases}
\]

where \( 0 \leq \theta \leq 1 \) is the IoU threshold for dividing positive and negative samples, \( i \) is the index of samples; \( p \) and \( n \) are the symbols for positive and negative samples; \( \mu \) and \( \sigma \) are the mean and standard variance of Gaussian distribution; \( \alpha \) and \( \lambda \) are the scale and bias factors. In addition, some constraints should be satisfied as shown in equation (4). The first one makes the label value of samples in the same class inversely correlated with the target probability, that is, IoU. The others are designed to make the absolute values of the label less than 1 to enlarge the difference between positive and negative samples.
The function curve of soft labeling with the quasi-Gaussian structure and its corresponding response function curve are shown in Figure 5. Like the binary labeling, the response values of target and background samples are always positive and negative respectively so that the difference between them is large enough to distinguish them well. However, different from the binary labeling, this difference between the response values of target and background samples becomes more significant, which will enhance the classification ability of deep classification trackers. More importantly, different from the binary labeling, the response values of samples belonging to the same class are no longer the same, but positively correlated with their IoU values, which makes the target location more accurate.

\[
\begin{align*}
\min \left( -\alpha_p \frac{1}{\sqrt{2\pi\sigma_p}} e^{\frac{(l_p - \mu_p)^2}{2\sigma_p^2}} \right) &< \lambda_p < 1 - \max \left( -\alpha_p \frac{1}{\sqrt{2\pi\sigma_p}} e^{\frac{(l_p - \mu_p)^2}{2\sigma_p^2}} \right) \\
-1 - \min \left( -\alpha_n \frac{1}{\sqrt{2\pi\sigma_n}} e^{\frac{(l_n - \mu_n)^2}{2\sigma_n^2}} \right) &< \lambda_n < \max \left( -\alpha_n \frac{1}{\sqrt{2\pi\sigma_n}} e^{\frac{(l_n - \mu_n)^2}{2\sigma_n^2}} \right)
\end{align*}
\]

Figure 5. Function curves of the soft labeling with quasi-Gaussian structure for deep classification tracking (a) and its corresponding response score (b). \( \Delta R_b \) denotes the difference between the response values of target and background samples corresponding to the binary labeling. \( \Delta R_s \) denotes the difference between the response values of target and background samples corresponding to the proposed soft labeling.

SiamFC trained with the soft labeling

Deep classification tracker SiamFC is proposed by Luca Bertinet et al. in 2016. Due to the fully convolutional network and no online update, SiamFC becomes the most real-time deep classification tracker at that time. SiamFC transfers visual object tracking to a similarity problem in an embedding space through a fully convolutional Siamese network. It calculates the similarity between the target image patch and the candidate samples generated by the dense sampling and then tracks the object by regarding the sample with the highest similarity as the target. In the training phase, SiamFC adopts the \( \{-1, +1\} \) binary label and sets the label values of the samples according to the center.
distances between the samples and the searching region because the IoU values of the samples are negatively correlated to the center distance. The samples are considered to the positive samples if they are within the radius $R$ of the center as the Figure 7(a) shows.

In order to verify the effectiveness of the proposed soft labeling with quasi-Gaussian structure, we apply it to SiamFC, denoting the variant as SiamFC-label. Since the IoU value of the sample is negatively correlated to the center distance between the sample and the searching region in SiamFC, we set this relationship as $I_i = -\beta R_i + 1$, where $R_i$ and $\beta > 0$ denotes the center distance and negative correlation coefficient, respectively. Then the soft labeling of SiamFC-label is expressed as equation (5) and its visualization is shown intuitively in Figure 7(b)

\[
    y_i = \begin{cases} 
        -\alpha_p \frac{1}{\sqrt{2\pi\sigma_p}} e^{-\frac{(1 - \beta R_i - \mu_p)^2}{2\sigma_p^2}} + \lambda_p & 0 \leq R_i \leq R \\
        -\alpha_n \frac{1}{\sqrt{2\pi\sigma_n}} e^{-\frac{(1 - \beta R_i - \mu_n)^2}{2\sigma_n^2}} + \lambda_n & R_i > R 
    \end{cases}
\]

As the Figure 7(c) and (d) show, comparing with SiamFC, SiamFC-label has the following two advantages: (1) the response values for samples belonging to the same class are no longer the same but positively correlated with their IoU values; (2) the difference among samples of different classes is more significant. Due to such two advantages, SiamFC-label can locate the target more accurately and perform better classification ability in the online tracking phase as the Figure 2 shows. What’s more, only the parameter values of the pre-trained network are changed in the online tracking phase so that the amount of computation will not be affected. Therefore, SiamFC-label can perform significantly improved tracking accuracy while achieving high real-time performance.

**Experiments**

In order to evaluate the effectiveness of soft labeling with quasi-Gaussian structure, we compare the SiamFC-label with the baseline tracker and the state-of-the-art trackers on OTB-2015 and VOT-2016 benchmark datasets. In this section, we firstly introduce the implementation details. Next, we compare the variant SiamFC-label with the baseline tracker on the popular benchmark datasets. Then, we evaluate our proposed method on OTB-2015 and VOT benchmark datasets in comparison with the state-of-the-art trackers. Lastly, we present extensive attribute-based performance analysis to further illustrate the effectiveness of our proposed soft labeling with quasi-Gaussian structure for improving the locating precision and classification ability of the deep classification trackers.

**Implementation details**

In this article, the experiments are conducted on the popular OTB-2015 and VOT-2016 benchmarks. The OTB-2015 benchmark contains 100 challenging sequences, which includes various tracking scenarios and challenges. The OTB-2015 benchmark provides two evaluating indicators, overlap success rate, and distance precision (DP). The overlap success plot shows the rate of bounding boxes whose IoU score is larger than a given threshold. Area under curve (AUC) of the overlap success plot is applied
to rank the trackers. The DP plot shows the DP for different thresholds. Usually, the DP at 20 pixels is applied to rank the trackers. On the OTB-2015 benchmark, all trackers are evaluated with one-pass evaluation (OPE). The VOT-2016 benchmark is the fourth VOT challenge, which includes 60 sequences. The expected average overlap (EAO), accuracy, robustness, average overlap (AO), and equivalent filter operations (EFO) are used to evaluate trackers on VOT-2016. The main evaluating indicator, EAO, synthetically reflects the overall performance of the trackers.

Our tracker is implemented in Matlab using MatConvNet. SiamFC with three scales is selected as baseline tracker since this version runs faster than the one with five scales and only performs slightly lower. We set the parameters of soft labeling with quasi-Gaussian structure in equation (5) as Table 1. The means of Gaussian distribution are set their values as 1 to satisfy the first constraint in equation (4), which makes the response values of samples belonging to the same class are no longer the same but positively correlated with their IoU value. To make the difference between the response values of different samples belonging to the same class appropriate, we set the values of the standard variances of Gaussian distribution as 0.5 times the response map size. Then the values of the scale factors are set as $\sqrt{2\pi}$ times the value of the standard deviation, and the values of the bias factors are set as $1.215$ and $0$ separately to satisfy the last two constraints in equation (4), which makes the difference between target and background samples more significant compared with the binary label. Finally, the other parameters, such as the stride, the center distance, and negative correlation coefficient, are set to be same as that in SiamFC$^{11}$ for the comparisons with baseline trackers. We randomly sample from the dataset ILSVRC15$^{18}$ to train the parameters of the Siamese network by minimizing the loss with SGD using the deep learning toolbox MatConvNet. Our machine is equipped with a single NVIDIA GeForce 1080Ti and an

| Parameter value | Parameter value |
|-----------------|-----------------|
| $\mu_p$         | $\mu_n$         | $\sigma_p$ | $\sigma_n$ | $\alpha_p$ | $\alpha_n$ | $\lambda_p$ | $\lambda_n$ | $k$ | $R$ | $\beta$ |
| 0.5 times the response map size, i.e. $\frac{12}{2}$ | 0.5 times the response map size, i.e. $\frac{12}{2}$ | $\sqrt{2\pi\sigma_p}$, i.e. $\frac{11\sqrt{2\pi}}{2}$ | $\sqrt{2\pi\sigma_n}$, i.e. $\frac{11\sqrt{2\pi}}{2}$ | 1.215 | 0 | Same as that in SiamFC$^{11}$, i.e. 8 | Same as that in SiamFC$^{11}$, i.e. 16 | | | | | |

**Figure 7.** (a) The binary labeling for SiamFC. (b) The soft labeling with quasi-Gaussian structure for SiamFC. (c) Response map of SiamFC trained with the binary labeling. (d) Response map of SiamFC trained with the soft labeling.
Intel Xeon E5-2650 at 2.20 GHz, and our software platform is Ubuntu 16.04 + Matlab 2017a + CUDA 8.0 + cudnn v6.0. The maximum graphics memory used for the simulation is 2 GB.

Comparisons with baseline trackers

For a more comprehensive validity evaluation of our proposed soft labeling with quasi-Gaussian structure, we compare the SiamFC-label with its baseline tracker on OTB-2015 and VOT-2016 benchmarks. Note that, SiamFC\textsuperscript{11} provides two tracking models, denoted by SiamFC-color and SiamFC-colorgray in this article. The difference between these two trackers is that SiamFC-colorgray converts 25% of the pairs to grayscale in training phase to handle the gray videos. We replace the binary labeling of these two trackers with the proposed soft labeling in the training phase, denoting the variants as SiamFC-label-color and SiamFC-label-colorgray respectively.

For SiamFC-label-color, only its label is different from SiamFC-color while all other hyper-parameters are the same as SiamFC-color. Experiment results shown in Figure 8 indicate that SiamFC-label-color achieves overall 1.8% and 1.9% improvement to SiamFC-color in terms of precision and success metric on OTB-2015 benchmark. What’s more, SiamFC-label-color performs better than SiamFC-colorgray on the precision and success metric, even without the trick for handling the gray videos.

To maximize the improvement caused by our proposed soft labeling, we make appropriate adjustments to the hyper-parameters and adapt the trick of handling the gray videos for SiamFC-label-colorgray. (1) Hyper-parameters: As described in the “Soft labeling with quasi-Gaussian structure for deep classification tracking” section, the soft...
labeling makes the difference between response values for different classes more significant, which is more conducive to classifying samples but slows the convergence process. Thus, compared with training over 50 epochs in SiamFC, we train two more epochs, a total of 52 epochs. For 52 epochs training, the learning rate of the first 50 epochs is decayed geometrically after epoch from $10^{-5}$ to $10^{-7}$, which is consistent with SiamFC, while the learning rates of the last 2 epochs are $9.3260 \times 10^{-6}$ and $8.1113 \times 10^{-6}$, respectively. 

(2) The trick for handling the gray videos: We adopt the trick of re-training a special gray network with all grayscale pairs in SiamFC-tri instead of the trick in SiamFC to handle the gray videos. For the special gray network, we only convert all pairs to grayscale while the other hyper-parameters in the training phase are all consistent with the color network. As shown in Figure 8, comparing with SiamFC-color, SiamFC-label-colorgray achieves 3.5% and 2.7% improvement on precision and success metric, respectively. Further, SiamFC-label-colorgray achieves overall 2% and 0.8% improvement of precision and success metric respectively in comparison with SiamFC-colorgray.

In addition, we take SiamFC-label-color as the representation of SiamFC-label to compare with the baseline tracker SiamFC on VOT-2016 benchmark. As shown in the Table 2 and Figure 9, compared to the baseline tracker, SiamFC-label-(color) performs more favorably
in terms of EAO, accuracy, robustness, and AO, while operating at almost the same frame-rate with SiamFC (86.3 Fps vs. 86.5 Fps).

**Comparisons with state-of-the-art trackers**

We compare the trackers SiamFC-label-color and SiamFC-label-colorgray with the state-of-the-art trackers using OPE with DP and overlap success metrics as proposed in OTB-2015 benchmark datasets, which mainly includes LCT, KCF, SRDCF, SAMF, DSST, MEEM, and CFNet. As shown in Figure 10, SiamFC-label-colorgray and SiamFC-label-color respectively achieve the first and fourth best DP (79.1% and 77.4%) while the second and third best performance in success metric (59.0% and 58.2%). Although SiamFC-label-colorgray and SiamFC-label-color rank slightly lower than SRDCF in terms of success metrics, their real-time (85.7 Fps and 86.3 Fps) is much faster than SRDCF (5 Fps) as shown in Table 3.

Furthermore, qualitative experiments on VOT-2016 benchmark against the state-of-the-art tracker are performed, which mainly includes MDNet_N, DPT, SiamFC, deepMKCF, DAT, KCF, SAMF, DSST. As shown in Figure 9 and Table 2, SiamFC-label(-color) behaves comparably with the state-of-the-art tracker in terms of EAO, ranking the second on VOT-2016 benchmark. Especially, SiamFC-label(-color) achieves the best accuracy among all these compared trackers.

**Attribute-based performance analysis**

Extensive performance analysis on the locating precision and classification ability is presented to further illustrate the effectiveness of the proposed soft labeling. As with the experiments in the “Comparisons with state-of-the-art trackers” section, we select SiamFC-color as the baseline tracker and compare SiamFC-label-color with SiamFC-color and SiamFC-colorgray on the OTB-2015 dataset to rule out other interference factors.

**Locating ability improvement:** We selected 2, 4, 6, 8, 10 pixels instead of 20 pixels as the threshold of precision metric, and then compared the overall performance of SiamFC-label-color, SiamFC-color, and SiamFC-colorgray. Figure 11 presents the locating precision improvement percentage of SiamFC-label-color in comparison to SiamFC-color and SiamFC-colorgray at different thresholds. Experimental results indicate that the smaller threshold value (i.e. the higher locating precision) is, the larger locating precision improvement percentage SiamFC-label-color achieves. This further illustrates the proposed soft labeling can enhance locating ability of the deep classification trackers.

More specifically, experiments on Car4 sequence are presented in Figure 12 to intuitively demonstrate the locating ability improvement caused by the soft labeling. Note that the location error of SiamFC-label-color is less than that of SiamFC-color and SiamFC-colorgray overall. This clearly proves that SiamFC-label-color locates the target more accurately than SiamFC-color and SiamFC-colorgray.

**Classification ability improvement:** Besides the locating ability, the proposed quasi-Gaussian combination soft label can also enhance the classification ability because the important information about the difference among samples in the same class is added in the training phase. Qualitative results on four sequences are presented in Figure 13 where SiamFC-color and SiamFC-colorgray both fail to track when the targets undergoing large appearance changes, whereas SiamFC-label-color can locate them robustly.
Conclusions

In this article, we revisit the binary labeling for deep classification trackers and indicate the problems in binary labeling through theoretical and experimental analysis. To solve such problems, we propose a soft labeling with quasi-Gaussian structure instead of the binary labeling to enhance the classification and locating ability of deep classification tracking, which takes into account the difference among the samples of the same and different classes simultaneously. To verify the effectiveness of our proposed soft labeling, we apply it to improve the deep classification tracker SiamFC, and then compare the variant with its base-line tracker and the state-of-the-art trackers on OTB-2015 and VOT benchmark datasets. Further, we present extensive attribute-based performance analysis to further illustrate the validity of our proposed soft labeling. More than SiamFC, our proposed soft labeling with quasi-Gaussian structure works on other deep classification tracking algorithms, which is our further work. Moreover, in various real-world applications such as robots, unmanned surface vessel (USV), and so on, our proposed method can achieve more precise and robust tracking performance.

Figure 13. Qualitative results comparing SiamFC-color with the baseline trackers on six challenging sequences in the OTB-2015 benchmark (from top to down: Dragonbaby, Box, Girl2, Bolt2, Human4, and Matrix_1, respectively).
Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project is supported by the Key Projects of the National Natural Science Foundation of China (No. 91648119), the National Nature Science Foundation of China (No. 61673254), and the National Nature Science Foundation of China (No. U1613226). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Yilmaz A, Javed O, and Shah M. Object tracking: a survey. ACM Comput Surv 2006; 38: 13.
2. Coppi D, Calderara S, and Cucchiara R. Transductive people tracking in unconstrained surveillance. IEEE Trans Circ Syst Video Technol 2015; 26: 762–775.
3. Liu S and Feng Y. Real-time fast moving object tracking in severely degraded videos captured by unmanned aerial vehicle. Int J Adv Robot Syst 2018; 15: 1729881418759108.
4. Okuno HG, Nakadai K, Lourens T, et al. Sound and visual tracking for humanoid robot. Appl Intell 2004; 20: 253–266.
5. Liu H, Luo J, Wu P, et al. People detection and tracking using RGB-D cameras for mobile robots. Int J Adv Robot Syst 2016; 13: 1729881416657746.
6. Wang N, Shi J, Yeung D-Y, et al. Understanding and diagnosing visual tracking systems. In: Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015, pp. 3101–3109.
7. Danelljan M, Robinson A, Khan FS, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking. In: European conference on computer vision, Amsterdam, The Netherlands, 8–16 October 2016, pp. 472–488. Springer International Publishing.
8. Danelljan M, Bhat G, Shahbaz Khan F, et al. ECO: efficient convolution operators for tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017, pp. 6638–6646.
9. Nam H and Han B. Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, Nevada, USA, 27–30 June 2016, pp. 4293–4302.
10. Jung I, Son J, Baek M, et al. Real-time MDNet. In: Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September, 2018, pp. 83–98.
11. Bertinetto L, Valmadre J, Henriques JF, et al. Fully-convolutional siamese networks for object tracking. In: European conference on computer vision, Amsterdam, The Netherlands, 8–16 October 2016, pp. 850–865. Springer International Publishing.
12. He A, Luo C, Tian X, et al. A twofold siamese network for real-time object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, Utah, USA, 19–21 June 2018, pp. 4834–4843.
13. Ma C, Huang J-B, Yang X, et al. Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015, pp. 3074–3082.
14. Wang L, Ouyang W, Wang X, et al. Visual tracking with fully convolutional networks. In: Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015, pp. 3119–3127.
15. Danelljan M, Hager G, Shahbaz Khan F, et al. Convolutional features for correlation filter based visual tracking. In: Proceedings of the IEEE international conference on computer vision workshops, Santiago, Chile, 7–13 December 2015, pp. 58–66.
16. Hong S, You T, Kwak S, et al. Online tracking by learning discriminative saliency map with convolutional neural network. In: International conference on machine learning, Lile, France, 6–11 July 2015, pp. 597–606.
17. Krizhevsky A, Sutskever I, and Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, Lake Tahoe, Nevada, USA, 3–8 December 2012, pp. 1097–1105.
18. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015; 115: 211–252.
19. Wang N and Yeung D-Y. Learning a deep compact image representation for visual tracking. In: Advances in neural information processing systems, Lake Tahoe, Nevada, USA, 5–8 December 2013, pp. 809–817.
20. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: International conference on computer vision & pattern recognition (CVPR’05), San Diego, CA, USA, 21–23 September 2005, pp. 886–893. IEEE Computer Society.
21. Danelljan M, Shahbaz Khan F, Felsberg M, et al. Adaptive color attributes for real-time visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, Ohio, USA, 24–27 June 2014, pp. 1090–1097.
22. Henriques JF, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels. In: European conference on computer vision, Florence, Italy 7–13 October 2012, pp. 702–715. Springer.
23. Fan B, Cong Y, Li X, et al. Structured low rank tracker with smoothed regularization. In: 2016 Visual communications and image processing (VCIP), Chengdu, China, 26–30 November 2016, pp. 1–4. IEEE.
24. Ma J, Luo H, Hui B, et al. Robust scale adaptive tracking by combining correlation filters with sequential Monte Carlo. Sensors 2017; 17: 512.
25. Fan B, Li X, Cong Y, et al. Structured and weighted multi-task low rank tracker. Patt Recogn 2018; 81: 528–544.
26. Wang Z and Cai B. A compressed multiple feature and
adaptive scale estimation method for correlation filter-based visual tracking. *Int J Adv Robot Syst* 2018; 15: 1729881417751511.
27. Yue F and Li X. Improved kernelized correlation filter algorithm and application in the optoelectronic tracking system. *Int J Adv Robot Syst* 2018; 15: 1729881418776582.
28. Qi Y, Zhang S, Qin L, et al. Hedged deep tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, Nevada, USA, 27–30 June 2016, pp. 4303–4311.
29. Wang L, Ouyang W, Wang X, et al. STCT: sequentially training convolutional networks for visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, Nevada, USA, 27–30 June 2016, pp. 1373–1381.
30. Song Y, Ma C, Gong L, et al. CREST: convolutional residual learning for visual tracking. In: Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017, pp. 2555–2564.
31. Held D, Thrun S, and Savarese S. Learning to track at 100 fps with deep regression networks. In: European conference on computer vision, Amsterdam, The Netherlands, 8–16 October 2016, pp. 749–765. Springer International Publishing.
32. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556* 2014.
33. Nam H, Baek M, and Han B. Modeling and propagating CNNs in a tree structure for visual tracking. *arXiv preprint arXiv:160807242* 2016.
34. Pu S, Song Y, Ma C, et al. Deep attentive tracking via reciprocal learning. In: *Advances in neural information processing systems*, Montreal, Canada, 3–8 December 2018, pp. 1931–1941.
35. Tao R, Gavves E, and Smeulders AW. Siamese instance search for tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, Nevada, USA, 27–30 June 2016, pp. 1420–1429.
36. Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017, pp. 2805–2813.
37. Wang Q, Gao J, Xing J, et al. DCFNet: discriminant correlation filters network for visual tracking. *arXiv preprint arXiv:170404057* 2017.
38. Dong X and Shen J. Triplet loss in siamese network for object tracking. In: *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, 8–14 September 2018, pp. 459–474.
39. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, Ohio, USA, 24–27 June 2014, pp. 580–587.
40. Wu Y, Lim J, and Yang MH. Object tracking benchmark. *IEEE Trans Patt Anal Mach Intell* 2015; 37: 1834–1848.
41. Kristan M, Leonardis A, Matas J, et al. The visual object tracking VOT2016 challenge results. In: *European Conference on Computer Vision*, Amsterdam, Netherlands, 8–16 October 2016.
42. Vedaldi A and Lenc K. MatConvNet: Convolutional neural networks for MATLAB. In: *Proceedings of the 23rd ACM international conference on multimedia*, Casablanca, Morocco, 20–23 December 2015, pp. 689–692. ACM.
43. Ma C, Yang X, Zhang C, et al. Long-term correlation tracking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, Massachusetts, USA, 8–10 June 2015, pp. 5388–5396.
44. Henriques JF, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters. *IEEE Trans Patt Anal Mach Intell* 2015; 37: 583–596.
45. Danelljan M, Hager G, Shahbaz Khan F, et al. Learning spatially regularized correlation filters for visual tracking. In: *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, 7–13 December 2015, pp. 4310–4318.
46. Li Y and Zhu J. A scale adaptive kernel correlation filter tracker with feature integration. In: *European conference on computer vision*, Zurich, Switzerland, 6–12 September 2014, pp. 254–265. Springer International Publishing.
47. Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking. In: *British machine vision conference*, Nottingham, UK, 1–5 September 2014. BMVA Press.
48. Zhang J, Ma S, and Sclaroff S. MEEM: robust tracking via multiple experts using entropy minimization. In: *European conference on computer vision*, Zurich, Switzerland, 6–12 September 2014, pp. 188–203. Springer International Publishing.
49. Lukežič A, Zajc LČ, and Kristan M. Deformable parts correlation filters for robust visual tracking. *IEEE Trans Cybern* 2017; 48: 1849–1861.
50. Tang M and Feng J. Multi-kernel correlation filter for visual tracking. In: *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, 7–13 December 2015, pp. 3038–3046.
51. Possegger H, Mauthner T, and Bischof H. In defense of color-based model-free tracking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, Massachusetts, USA, 8–10 June 2015, pp. 2113–2120.

**Appendix I**

The logistic loss function $L$ about $t_i = y_i \cdot v_i$ is expressed as following

$$L(t_i) = \log(1 + e^{-t_i})$$  \hspace{1cm} (1A)$$

We observed that this loss function has the following two significant characteristics:
The first derivative about $t_i$ is always less than 0, that is, $\frac{\partial L}{\partial t_i} = -\frac{1}{1 + \eta} < 0$, so the loss function monotonically decreases with respect to $t_i$.

The second derivative about $t_i$ is always greater than 0, that is, $\frac{\partial^2 L}{\partial t_i^2} = \eta > 0$, so the first derivative about $t_i$ is monotonically increasing with respect to $t_i$.

The function of the gradient descent is expressed as following

$$t_i^{n+1} = t_i^n - \alpha \frac{\partial L}{\partial t_i^n} \quad (1B)$$

where $\alpha > 0$ denotes the learning rate. And since the first derivative about $t_i$ is always less than 0, then

$$t_i^{n+1} > t_i^n \quad (1C)$$

Further, since the first derivative about $t_i$ is monotonically increasing with respect to $t_i$ and is always less than 0, then

$$\frac{\partial L}{\partial t_i^n} < \frac{\partial L}{\partial t_i^{n+1}} < 0 \quad (1D)$$

Thus, the absolute value of the first derivative about $t_i^{n+1}$ and $t_i^n$ satisfies the following

$$0 < \left| \frac{\partial L}{\partial t_i^{n+1}} \right| < \left| \frac{\partial L}{\partial t_i^n} \right| \quad (1E)$$

Substituting equation (1B) into equation (1E), then

$$0 < |t_i^{n+2} - t_i^{n+1}| < |t_i^{n+1} - t_i^n| \quad (1F)$$

Equation (1F) indicates that $t_i$ will gradually converge to a constant $c$ until $|t_i^{n+1} - t_i^n| < \xi$ where $\xi$ denotes an infinitesimal quantity.

What’s more, to further validate this theoretical derivation, we conduct the experiments on convergences of gradient descent for different initial values. As the Figure 1A shows, $t_i$ will converge to a constant $c = 7$ for different initial values when $|t_i^{n+1} - t_i^n| < \xi$ where $\xi$ has a value of $10^{-3}$ here. Each color denotes a convergence case with an initial value.

---

Figure 1A. Experiment on convergences of gradient descent for different initial values. $t_i$ will converge to a constant $c = 7$ for different initial values when $|t_i^{n+1} - t_i^n| < \xi$ where $\xi$ has a value of $10^{-3}$ here. Each color denotes a convergence case with an initial value.