Learning the Distribution-Based Temporal Knowledge with Low Rank Response Reasoning for UAV Visual Tracking

Guoxia Xu, *Member, IEEE,* Hao Wang, *Senior Member, IEEE,* Meng Zhao, *Member, IEEE,* Marius Pedersen, *Member, IEEE,* and Hu Zhu, *Member, IEEE,*

***Abstract*—In recent years, the constraint based correlation filter has shown good performance in unmanned aerial vehicle (UAV) tracking, which gains a lot popularity in many intelligence transportation applications. In this work, a distribution-based temporal knowledge driven method is proposed to leverage the temporal translation property in UAV tracking. Instead of focusing on the traditional issues in the correlation filter, we provide a new method of learning parametric distribution on temporal knowledge by Wasserstein distance which is successfully embedded to solve the problem of temporal degeneration in learning process of tracking. Furthermore, we approximate optimal response reasoning with low-rank constraint over response consistency. Furthermore, the proposed method is solved by a simple iterative scheme with alternating direction multiplication ADMM algorithm. We demonstrate the superior tracking performance in several public standard UAV tracking benchmarks compared with state-of-the-art algorithms.**

***Index Terms*—Visual Tracking, Low Rank Constraint, Wasserstein Distance, ADMM**

I. Introduction

VIDEO target tracking is an important research direction in the field of internet of vehicles, which is widely used in video surveillance [1], human-computer interaction [2], intelligent transportation [3] and other fields. Based on the requirement of intelligent transportation, it needs more powerful technical support in terms of traffic flow control, vehicle detection, and border control [4] to ensure safety and effectively improve the level of intelligence in traffic. Intelligent transportation infrastructure connects the internet of vehicles based on the collected information to adjust facility parameters according to the actual situation. It will inevitably bring challenges to data collection and its performance. With the continuous development of machine learning, the deep integration of UAV technology [5] and intelligent transportation has become the general trend. However, the main challenge of UAV tracking in the traffic domain is how to adapt fast change in the appearance of the target.

(Corresponding author: Hao Wang) (E-mail: hawa@ntnu.no)

Guoxia Xu, Hao Wang and Marius Pedersen are with Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway. Meng Zhao is with the Engineering Research Center of Learning-Based Intelligent System, Ministry of Education, Tianjing 300384, China, and also with the Key Laboratory of Computer Vision and System of Ministry of Education, School of Computer Science and Technology, Tianjing University of Technology, Tianjing 30084, China. Hu Zhu is with Jiangsu Province Key Lab on Image Processing and Image Communication, Nanjing University of Posts and Telecommunications, Nanjing 210003, China.

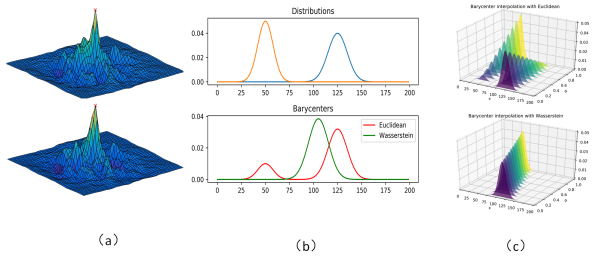


Fig. 1. a). Two response maps from UAV tracking shows the variants and these response maps can directly described as two distributions. b). Here, we show the predefined two distributions and calculate corresponding barycenter of Euclidean and Wasserstein. c). Upper and lower are barycenter interpolation with Euclidean and Wasserstein distance.

Even if the initial frame of an unknown scene is given, the main performance of predicting the target state of each frame will be limited by several appearance variants. Moreover, because the traditional target tracking background is basically fixed, the tracker only considers the problem of the target itself in the tracking process. However, due to the movement of traffic vehicles, the tracking process needs to take into account the complex scene variants and unpredicted interference. The characteristic of UAV target tracking [5], [6] is that both the target and the background are in motion, which is hard to well solve the difficulty of target tracking in the traffic scenarios and achieve good intelligent traffic video monitoring effect. In addition, due to the problems of mechanical vibration, object motion, target occlusion, background clutters [7], target tracking has brought greater challenges.

Benefiting from its easy implementation and fast prediction of discriminative correlation filter (DCF), DCF has attracted a lot attention in UAV tracking. Until now, there are three main research directions in UAV tracking: *spatial regularization, temporal smoothing,* and *robust feature representation*. To solve the first problem, spatial regularization DCF: SRDCF [8] are proposed based on the spatial penalty. This work also inspired other research work on spatial regularization [9], [10]. In [3], they proposed a new DCF tracker by suppressing the constraint of spatial boundary effect with spatial feature selection. Moreover, the spatial reliability enhanced DCF [7] [11] had proposed to indicate the reliability of background. However, these methods do not adaptively depress the background and consider the temporal information. To solve the second problem, a temporal regularization is introduced by [10] and [12] to realize the joint spatial-temporal solution and

obtain the better performance. For the third question, with the development of robustness image feature extraction method on deep neural network, the performance of DCF-based trackers has been greatly improved performance and solves the problem to some extent.

Recently, to combine temporal information, some latest models used a transformer to combine spatial and temporal information. STARK [13] had not used any proposals, anchors, and post-processing steps (such as cosine window or bounding box regression), which greatly simplified the visual tracking model. [14] developed a feature fusion network based on a self-context augmentation module with self-attention and a cross-feature augmentation module with cross-attention. Compared with correlation-based feature fusion, self-attention-based methods adaptively focus on useful information, such as edges and similar objects, and establish associations between distant features, enabling the tracker to obtain better classification and regression result. However, the response of redundant information in the global response will affect the accuracy. AutoTrack [12] automatically updated the hyper-parameters to accommodate the change of each frame with the global response. To achieve the better performance, the spatial constraints with content-aware [15] and bilateral regression ranking model [16] and other different hybrid response mining [17] [18] based methods had been proposed. While online learning of tracking has made good progress, there are still many problems in the temporal-based tracking framework. These existing methods only discover the reliability of spatial or temporal or background or response, the reliability of the temporal knowledge transfer is also deserved to investigate to avoid temporal degeneration. In existing temporal knowledge transfer based on the DCF tracking framework, euclidean distance is commonly used to measure the similarity of the targets of the two adjacent filters within a closed appearance [10], [12], [19]. Here, we recall a new concept about online temporal learning in visual tracking (*probability measurement*). This problem is unnoticed by the above methods. It also brings some questions over here: *what can we measure in online learning: probabilistic temporal fitting or direct temporal interpolation?*

In temporal-based framework, most methods assume that the target context between two frames is a component with minor changes, and the change of two adjacent target distributions can be kept only by interpolation. However, this is difficult to appear in reality. In UAV tracking, it is obvious that there is no such assumption that the tracked target has obvious occlusion or deformation. The noise drift in the temporal domain will also lead to tracking distortion inevitably. A more reasonable solution is to replace the measurement or transformation here. Regardless of temporal regularization or response mining methods, they are all looking for a transformation such that the representation of the updated frame is matched with historical information. The well-known class of transformation can be expressed in [20], in which the Kullback Leibler divergence was used in a deep neural network for visual tracking [21]. However, there is no closed-form solution that can express the similarity measurement.

Actually, the distorted appearance in UAV tracking challenges the spatial or temporal based DCF methods. The above discussion motivates us to mitigate the problem of overfitting and omit the impact of unpredicted appearance. Fortunately, the Wasserstein distance with a common Lagrangian formulation and alleviates the need for a common space. In [22], they proposed a novel approach to learn domain invariant feature representations. Wasserstein generative adversarial network (GAN) [23] learned a more reasonable and efficient approximation method and cured the main training problem of GAN. Thus, we leverage a probability temporal fitting method motivated by the Wasserstein distance. To improve the anti-noise performance of the tracking, we use the Wasserstein distance to measure the similarity of the filter distribution instead of previous linear interpolation method for estimation of temporal filter. From our own observation on preliminary experiment shown in Fig. 1, two time-varying distributions with corresponding means and barycenter under euclidean metric and Wasserstein metric are respectively shown in (b). The barycenter map of 2D and 3D maps can be clearly shown the state-transition truth of the similarity of two distributions. It is noticeable that this property is desirable for UAV tracking to overcome several appearance variants.

With the development of target tracking, research on low rank has made great progress and achieved good results. He et al. [24] had been successfully used in object tracking by exploiting low-rank constraints to capture the underlying structure of candidate particles. To mitigate this issue, we further investigate the low rank reasoning over the temporal response. Therefore, we propose a new model (ATGT) as shown in the Fig. 2. The main contributions of our ATGT method include:

- A novel Wasserstein distance regularization method for measuring the temporal transition is proposed. By adaptively incorporating the probability temporal fitting manner, the filter is enabled to mitigate the problem of temporal degeneration.
- Different from inducing the representation, the low rank constraint is conducted on the temporal response to achieve beyond response consistency for improving tracking robustness and overcoming the appearance variants.
- The iterative process is solved by ADMM algorithm. A comprehensive evaluation of ATGT, including UVA123@10fps, DTB70, OTB100, UAVDT-M, and UAVDT-S. The results demonstrate the advantages of the ATGT, as well as its advantages over the most advanced trackers.

The main structure of the paper is as follows. Section I is the introduction of this paper. Section II introduces the related works. And Section III introduces the new method proposed in this paper. Then Section IV introduces the related experimental results. Finally, Section IV provides a brief summary of our work.

II. Related Work

A. Visual Tracking based on Correlation Filter and Low-rank Constraint

Deep learning methods have begun to make inroads in

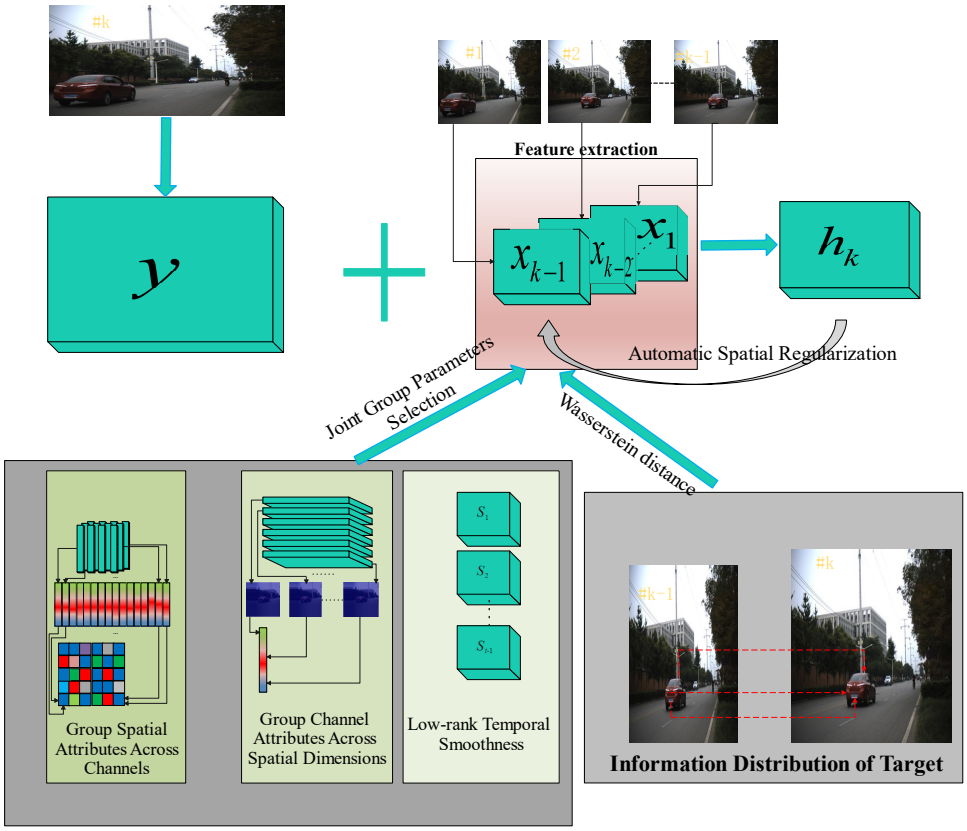


Fig. 2. Our proposed method adopts low rank temporal response constraint and group feature selection to improve the stability of correlation filter. In addition, the distribution-based knowledge is discovered from the wasserstein distance for probability temporal fitting

the field of target tracking, and have gradually surpassed traditional methods in terms of performance, resulting in great breakthroughs. However, the correlation filter-based tracking has still attracted a lot attention and played an important role in UAV application.

In the field of visual recognition and target tracking, how to extract essential data representation from data and learn useful information is a key problem in these fields. Low rank is an important property to describe the data structure, which is suitable for extracting the essential features of data. A low rank constraint from [19] is defined to achieve the temporal smooth presented as follows:

$$rank(W_t) - rank(W_{t-1}) \tag{1}$$

where $W_t = [\mathbf{vec}(w_1), \cdots, \mathbf{vec}(w_t)] \in \mathbb{R}^{N^2C \times t}$ is a matrix. Here, the constraint Equ. (1) imposes a low-rank property across frames because it impacts on the selection process. However, it is inefficient to calculate rank (W_t) in long-term videos. Therefore, its sufficient condition as a substitute is used:

$$d(w_t - u_{t-1}) \tag{2}$$

where $u_{t-1} = \sum_{k=1}^{t-1} w_k/(t-1)$ is the average of all the filters learned before, d is a distance metric. Therefore, the regularization term is used to adaptively strengthen the time low-rank attribute as follows:

$$\boldsymbol{R}_T = \sum_{k=1}^{C} \left\| \mathbf{W}_t^k - \Re(\mathbf{W}_{t-1}^k) \right\|_F^2 \tag{3}$$

where F norm in the Equ. (3) represents d in the Equ. (2), that is, F norm distance measure. $\Re$ in the Equ. (3) represents u_{t-1} in the Equ. (3).

However, the correlation of the temporal parameters S in the temporal mining model has not been fully studied. Therefore, the difference between each frame with low rank property is limited to avoid the mutation of the abnormal temporal parameters. As shown in the following formula:

$$S_t - S_{t-1} \tag{4}$$

In addition, to avoid redundant feature information leading to break the temporal consistency of the model, we adopt low-rank processing for the temporal parameter S_t as follows:

$$Rank(S_t) - Rank(S_{t-1}) \tag{5}$$

Based on Equ. (1), (2), (3) on the above inspiration , we can get the following formula:

$$Rank(S_t) - Rank(S_{t-1}) = \sum_{k=1}^{C} \left\| S_t^k - \Re(S_{t-1}^k) \right\|_F^2 \tag{6}$$

where $\Re(S_{t-1}^k) = \sum_{k=1}^{t-1} w_k/(t-1)$ is the average of all temporal parameters learned before.

B. Wasserstein Distance

In mathematical probability and mathematical statistics, it is a common way to measure distance by Wasserstein distance. In traditional target tracking, there is usually no big difference between two frames of targets. Thus, traditional target tracking

basically adopts Euclidean distance to measure the similarity of correlation filters of two frames of targets. However, when the target has obvious occlusion or deformation, the drift of temporal noise will lead to the distortion of target tracking.

Here, we suppose that $d(x,y)$ is treated as the probability temporal fitting, and $f(x)$ and $g(x)$ are the probability density function of learned filter in our UAV tracking task. $h(x,y)$ is an arbitrary joint distribution, and its edge function is two probability density functions: $\int h(x,y)dx = g(y)$, $\int h(x,y)dy = g(x)$. Then for $p > 0$, the Wasserstein distance d_p^w is:

$$d_p^w(f,g) = \inf_h \sqrt[p]{\int d(x,y)^p h(x,y)dxdy} \tag{7}$$

The $\inf_h$ represents the lower bound of all possible joint probability functions, p usually takes 1 and 2. Actually, filter h_t is definitely treated as the a multi-variable distribution. Therefore, the Wasserstein distance can be used to measure the similarity of the filter with geometry of the underlying space even the two distributions without overlap. But it is noticeable that Equ. (7) can not be used in practice, because joint distribution function of two distribution is not available. Here comes the empirical distribution function:

$$\delta_x(x) = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x) \tag{8}$$

Then the Wasserstein distance of X and Y is:

$$d_p^w(X,Y) = d_p^w(\delta(X),\delta(Y)) \tag{9}$$

Then the final Wasserstein distance [25] can be presented as follows:

$$d_p^w(X,Y) = \inf_h \sqrt[p]{\sum_{i=1}^{n}\sum_{j=1}^{m} C_{i,j}d(x_i,y_j)^p} \tag{10}$$

where $C = \{C_{i,j}\}$ is the transfer Matrix.

III. The Proposed ATGT Model

In this section, we will introduce the model building process and solution method of ATGT Algorithm.

A. ATGT Model

The final ATGT model can be separated into innovation part to discover the inherent space of parameters S_t: $(R_s(S_t))$, $(R_c(S_t)$ and $(R_T(S_t))$.

$$\varepsilon(\mathbf{H}_t, S_t) = \frac{1}{2}\left\|Y - \sum_{k=1}^{K} X_t^k \circledast H_t^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\tilde{\mathbf{u}} \odot \mathbf{H}_t^k\right\|_2^2 + \frac{S_t}{2}\sum_{k=1}^{K}\left\|\mathbf{H}_t^k - \mathbf{H}_{t-1}^k\right\|_2^2 + \lambda_1 R_s(S_t) + \lambda_2 R_c(S_t) + \lambda_3 R_T(S_t) \tag{11}$$

1) *Spatial*: $\lambda_1 R_s(S_t)$ is in order to obtain the grouping attribute of each spatial location $(S_t)_{i,j:}$ by using l_2 norm,

Algorithm 1 Solution of the ATGT model with ADMM algorithm

1: *Input:* y ,γ, λ_1,λ_2,λ_3,$\mathbf{v}_t$,$\mathbf{g}_t$,N
2: *Initialization:*
$\mathbf{v}_t^0 = \mathbf{g}_t^0 = h^(0) = 0$, $i = 0$;
3: *Iteration:*
While $(i \leq N)$ do
(1) Update $\mathbf{H}^{t*}$ by solving Equ. (21), $t = 1, 2, \cdots K$;
(2) Update S_t by solving Equ. (22);
(3) Update $\mathbf{S_t^{'}}$ by solving the third sub-equation of Equ. (23);
(4) Update $\widehat{G}_t$ by solving Equ. (26);
(5) Lagrangian multiplier update $\widehat{V}^i$ by solving Equ. (27);
(6) Lagrangian multiplier update Γ by solving Equ. (28);
(7) $i = i + 1$;
end while
4: *Output:*
$h^{(i+1)}$

2) *Channel*: $\lambda_2 R_c(S_t)$ is using the Frobenius norm to obtain the grouping attributes for channels $\left\{S_t^k\right\}_{k=1}^{C}$,
3) *Temporal*: $\lambda_3 R_T(S_t)$ is the temporal regularization term between S_t and S_{t-1}, which is subject to low rank constraints to promote time consistency in the temporal parameters.
4) *Hyperparameters*: $\lambda_1, \lambda_2, \lambda_3$ is corresponding parameter.

Correspondingly,

1) *Spatial*: $R_s(S_t) = \sum_{i=1}^{N}\sum_{j=1}^{M}\|(S_t)_{ij:}\|$
2) *Channel*: $R_c(S_t) = \sum_{k=1}^{K}\left\|S_t^k\right\|_F$
3) *Temporal*: $R_T(S_t) = Rank(S_t) - Rank(S_{t-1})$

It is noticeable that the final ATGT is presented as follows:

$$\varepsilon(\mathbf{H}_t, S_t) = \frac{1}{2}\left\|Y - \sum_{k=1}^{K} X_t^k \circledast H_t^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\tilde{\mathbf{u}} \odot \mathbf{H}_t^k\right\|_2^2 + \frac{S_t}{2}\sum_{k=1}^{K}\left\|\mathbf{H}_t^k - \mathbf{H}_{t-1}^k\right\|_2^2 + \frac{\lambda_1}{2}(\sum_{i=1}^{N}\sum_{j=1}^{M}\|(S_t)_{ij:}\| + \lambda_2\sum_{k=1}^{K}\left\|S_t^k\right\|_F + \lambda_3\sum_{k=1}^{K}\left\|S_t^k - \Re(S_{t-1}^k)\right\|_F^2) \tag{12}$$

Furthermore, the Wasserstein distance is similar to the information distribution, which can weaken its influence of temporal degeneration. Thus, the treatment of $\mathbf{h}_t$ with Wasserstein distance is conducted to achieve the following model: From the Equ. (10), $W(N(H_t^k, H_{t-1}^k))$ the probability temporal fitting model with Wasserstein distance is presented as follows

$$\min_C \sqrt[p]{\sum_{i=1}^{|H_t^k|}\sum_{j=1}^{|H_{t-1}^k|} C_{i,j}\left\|H_t^k - H_{t-1}^k\right\|^p} \tag{13}$$

Then we get :

$$\varepsilon(\mathbf{H}_t, S_t) = \frac{1}{2}\left\|Y - \sum_{k=1}^{K} X_t^k \circledast H_t^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\tilde{\mathbf{u}} \odot \mathbf{H}_t^k\right\|_2^2 + \sqrt[p]{\sum_{i=1}^{|H_t^k|}\sum_{j=1}^{|H_{t-1}^k|} C_{i,j}\left\|H_t^k - H_{t-1}^k\right\|^p} + \lambda_2\sum_{k=1}^{K}\left\|S_t^k\right\|_F + \frac{\lambda_1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\|(S_t)_{ij:}\| + \sum_{k=1}^{K}\left\|S_t^k - \Re(S_{t-1}^k)\right\|_F^2 \tag{14}$$

In Equ. (14), where C is the transfer Matrix, in which $C_{i,j} > 0, \sum_{i=1}^{h_t^k} C_{i,j} = \frac{1}{|h_t^k|}$.

B. The Optimization of Proposed Model

Here, the model in Equ. (14) can be minimized with the alternating direction method of multipliers ADMM algorithm to obtain the global optimal solution. Equ. (14) has a convolution calculation, Pasival's theorem is used to facilitate the calculation by converting the problem into the frequency domain. For optimization, we use an auxiliary variable $\hat{g}_t$ by ordering $\hat{g}_t = \sqrt{T}\mathbf{F}\mathbf{h}_t(\hat{\mathbf{G}} = \left[\hat{g}_t^1, \hat{g}_t^2, \hat{g}_t^2, \cdots, \hat{g}_t^K\right])$ where $\mathbf{F} \in \mathbb{C}^{T\times T}$ denotes the orthonormal matrix and the symbol denotes the discrete Fourier transform (DFT) of a signal. Then we get the objective function in the frequency domain as follows:

$$\varepsilon(\mathbf{H}_t, S_t, \hat{\mathbf{G}}_t) = \frac{1}{2}\left\|Y - \sum_{k=1}^{K}\hat{X}_t^k \odot \hat{G}_t^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\tilde{\mathbf{u}} \odot \mathbf{H}_t^k\right\|_2^2 + \lambda_1\sum_{i=1}^{N}\sum_{j=1}^{M}\|S_{ijt}\| + \lambda_2\sum_{k=1}^{K}\left\|S_t^k\right\|_F + \lambda_3\sum_{k=1}^{K}\left\|S_t^k - \Re(S_{t-1}^k)\right\|_F^2 + \sqrt[p]{\sum_{i=1}^{|\hat{G}_t^k|}\sum_{j=1}^{|\hat{G}_{t-1}^k|} C_{i,j}\left\|\hat{G}_t^k - \hat{G}_{t-1}^k\right\|^p} \tag{15}$$

By minimizing Equ. (15), an optimal solution can be obtained by (ADMM) [26]. The Augmented Lagrange form of Equ. (15) can be formulated as:

$$\iota_t(\mathbf{H}_t, S_t, \widehat{\mathbf{G}}_t, \widehat{\mathbf{M}}_t) = \varepsilon(\mathbf{H}_t, S_t, \widehat{\mathbf{G}}_t) + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{G}_t^k - \sqrt{T}\mathbf{F}H_t^k\right\|_2^2 + \sum_{k=1}^{K}(\widehat{G}_t^k - \sqrt{T}\mathbf{F}H_t^k)^T\widehat{m}_t^k \tag{16}$$

where $\widehat{\mathbf{M}}_t = [\widehat{\mathbf{m}}_1, \widehat{\mathbf{m}}_2, \cdots, \widehat{\mathbf{m}}_K] \in \mathbb{R}^{T\times K}$ is the Fourier transform of the Lagrange multiplier and γ denotes the step size regularization parameter. By assigning $\mathbf{v}_t^k = m_t^k/\gamma\left(\mathbf{V}_t^k = \left[\mathbf{v}_t^1, \mathbf{v}_t^2, \cdots, \mathbf{v}_t^K\right]\right)$. Equ. (16) can be reformulated as:

$$\iota_t(\mathbf{H}_t, S_t, \widehat{\mathbf{G}}_t, \widehat{\mathbf{V}}_t) = \varepsilon(\mathbf{H}_t, S_t, \widehat{\mathbf{G}}_t) + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{G}_t^k - \sqrt{T}\mathbf{F}H_t^k + \widehat{\mathbf{v}}_t^k\right\|_2^2 \tag{17}$$

Due to the convexity of the proposed formulation, we apply the augmented Lagrange method [27] to optimize Equ. (17). Concretely, we introduce a slack variable $S^{'} = S$ and construct the following Lagrange function:

$$\varepsilon(\mathbf{H}_t, S_t, S_t^{'}, \hat{\mathbf{G}}_t) = \frac{1}{2}\left\|Y - \sum_{k=1}^{K}\hat{X}_t^k \odot \hat{G}_t^k\right\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\left\|\tilde{\mathbf{u}} \odot \mathbf{H}_t^k\right\|_2^2 + \lambda_1\sum_{k=1}^{K}\left\|S_t^{'k}\right\|_F + \lambda_2\sum_{i=1}^{N}\sum_{j=1}^{M}\left\|S_{ijt}^{'}\right\| + \lambda_3\sum_{k=1}^{K}\left\|S_t^k - \Re(S_{t-1}^k)\right\|_F^2 + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{G}_t^k - \sqrt{T}\mathbf{F}H_t^k + \widehat{\mathbf{V}}_t^k\right\|_2^2 + \frac{\mu}{2}\sum_{k=1}^{K}\left\|S_t^k - S_t^{'k} + \frac{\Gamma^k}{\mu}\right\|_F + W \tag{18}$$

where the $\sqrt[p]{\sum_{i=1}^{|\hat{G}_t^k|}\sum_{j=1}^{|\hat{G}_{t-1}^k|} C_{i,j}\left\|\hat{g}_t^k - \hat{G}_{t-1}^k\right\|^p}$ is W, Γ is the Lagrange multiplier sharing the same size as X, Γ^k is its k-th channel, and μ is the corresponding penalty. The augmented Lagrange function of the above is divided into main subproblems by the ADMM algorithm.

$$\begin{cases} H_t = \frac{1}{2}\sum_{k=1}^{K}\left\|\tilde{\mathbf{u}} \odot \mathbf{H}_t^k\right\|_2^2 + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{G}_t^k - \omega h_t^k + \widehat{\mathbf{v}}_t^k\right\|_2^2 \\ S_t = \lambda_3\sum_{k=1}^{K}\left\|S_t^k - \Re(S_{t-1}^k)\right\|_F^2 + \frac{\mu}{2}\sum_{k=1}^{K}\left\|S_t^k - S_t^{'k} + \frac{\Gamma^k}{\mu}\right\|_F \\ S_t^{'} = \lambda_1\sum_{k=1}^{K}\left\|S_t^{'k}\right\|_F + \lambda_2\sum_{i=1}^{N}\sum_{j=1}^{M}\left\|S_{ijt}^{'}\right\| + \frac{\mu}{2}\sum_{k=1}^{K}\left\|S_t^k - S_t^{'k} + \frac{\Gamma^k}{\mu}\right\|_F \\ \widehat{\mathbf{G}}_t = \left\|Y - \sum_{k=1}^{K}\hat{X}_t^k \odot \hat{G}_t^k\right\|_2^2 + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{G}_t^k - \omega H_t^k + \widehat{\mathbf{V}}_t^k\right\|_2^2 + W \end{cases} \tag{19}$$

For simplification, in the Equ. (19), where the $\sqrt{T}\mathbf{F}$ is ω , We detail the solution to each subproblem for the update as follows.

Update of $\mathbf{H}_t$: given $S_t, \widehat{\mathbf{G}}_\mathbf{t}, \widehat{\mathbf{V}}_\mathbf{t}, \mathbf{S}_\mathbf{t}^{'}$, we can optimize $\mathbf{H}_t$ by:

$$\min_{\mathbf{H}_t}\frac{1}{2}\sum_{k=1}^{K}\left\|\tilde{\mathbf{u}} \odot \mathbf{H}_t^k\right\|_2^2 + \frac{\gamma}{2}\sum_{k=1}^{K}\left\|\widehat{G}_t^k - \omega H_t^k + \widehat{\mathbf{V}}_t^k\right\|_2^2 \tag{20}$$

The closed-form solution of $\mathbf{h}_t$ can be written by:

$$\mathbf{H}_t^* = \left[\mathbf{U}^T\widetilde{\mathbf{U}} + \gamma T\mathbf{I}\right]^{-1}\gamma T(\mathbf{V}_t^k + \mathbf{G}_t^k) = \frac{\gamma T(\mathbf{V}_t^k + \mathbf{G}_t^k)}{(\tilde{\mathbf{u}} \odot \tilde{\mathbf{u}}) + \gamma T}, \tag{21}$$

where $\widetilde{\mathbf{U}} = diag(\tilde{\mathbf{u}}) \in \mathbb{R}^{T\times T}$ represents diagonal matrix.

Update of S_t: Given other variables in Equ. (19), the optimal solution of S_t can be determined as:

$$\min_{\mathbf{S}_t}\lambda_3\sum_{k=1}^{K}\left\|S_t^k - \Re(S_{t-1}^k)\right\|_F^2 + \frac{\mu}{2}\sum_{k=1}^{K}\left\|S_t^k - S_t^{'k} + \frac{\Gamma^k}{\mu}\right\|_F$$
$$\mathbf{S}_t^* = \frac{2\lambda_3}{2\lambda_3 + \mu}S_{t-1}^k + \frac{\mu}{2\lambda_3 + \mu}S_t^{'k} - \frac{\Gamma^k}{2\lambda_3 + \mu} \tag{22}$$

Update of $\mathbf{S}_\mathbf{t}^{'}$: Given other variables in Equ. (19), the optimal solution of $\mathbf{S}_\mathbf{t}^{'}$ is determined as:

$$\mathbf{S}_\mathbf{t}^{'*} = max\left(0, 1 - \frac{\lambda_1}{\mu\left\|\mathbf{P}^k\right\|_F} - \frac{\lambda_2}{\mu\left\|\mathbf{P}_{ij}\right\|_2}\mathbf{P}_{ij}^k\right) \tag{23}$$

where $\mathbf{P}_{ij}^k = S_{ij}^k + \gamma_{ij}^k/\mu$.

Update of $\widehat{\mathbf{G}}_\mathbf{t}$: It is very difficult to solve $\widehat{\mathbf{G}}_\mathbf{t}$ directly in Equ. (19) because of its complexity. we know that Parsavar's transformation of the $\widehat{\mathbf{G}}_\mathbf{t}$ subproblem in Equ. (23)

$$\min_{\widehat{\mathbf{G}}_\mathbf{t}} \left\| Y - \hat{X}_t^k \odot \hat{G}_t^k \right\|_2^2 + W + \frac{\gamma}{2} \left\| \widehat{G}_t^k - \omega H_t^k + \widehat{\mathbf{V}}_t^k \right\|_2^2 \tag{24}$$

Then we can get the solution of $\widehat{\mathbf{G}}_\mathbf{t}$ as :

$$\widehat{\mathbf{G}}_\mathbf{t}^* = \frac{\widehat{x}_t Y + \widehat{S}_t \widehat{G}_{t-1} + \gamma\omega H_t^k - \gamma \widehat{V}_t^k}{2\widehat{X}_t^T \widehat{X}_t + C\widehat{S}_t + \gamma} \tag{25}$$

After derivation using the Sherman Morrison formula, we can obtain its solution:

$$\widehat{\mathbf{G}}_\mathbf{t} = \frac{1}{C\widehat{S}_t + \gamma}(I - \frac{2\widehat{X}_t^T \widehat{x}_t}{2\widehat{X}_t^T \widehat{X}_t + C\widehat{S}_t + \gamma})\rho \tag{26}$$

where the vector ρ is $\widehat{X}_t Y + \widehat{S}_t \widehat{G}_{t-1} + \gamma\omega H_t^k - \gamma \widehat{V}_t^k$, γ takes the form$\gamma^{i+1} = min(\gamma_{max}, \rho\gamma^i)$.

Lagrangian multiplier update: after solving the four subproblems above, we can update the Lagrangian multipliers as

$$\widehat{V}^{i+1} = \widehat{V}^i + \gamma^i(\widehat{G}^{i+1} - \widehat{H}^{i+1}) \tag{27}$$

where i and $i+1$ denotes the iteration index and the step size regularization constant γ (initially equals to 1) takes the form of $\gamma(i+1) = min(\gamma_{max}, \beta y_i)$. ($\beta = 10, \gamma_{max} = 10000$)

$$\Gamma = \Gamma + \mu(S_t - S_t^{'}) \tag{28}$$

where Γ is the Lagrange multiplier sharing the same size as X, Γ^k is its $k-th$ channel, and μ is the corresponding penalty.

C. Computational Complexity Analysis

After the above analysis and derivation, because the optimization process is realized by the ADMM algorithm, the solution of each optimization sub-problem is closed. Therefore, it guarantees the global optimality of convergence to the Eckstein-Bertsekas condition. In addition, we set the number of iterations to 5. The detailed program is given as Algorithm 1. The convergence of Algorithm 1 can be guaranteed. Since the overall objective function in Equ. (19) is convex with a global optimal solution. In each iterative calculation of sub-problem , FFT and inverse FFT transformations are needed. Then the computational complexity is $O(DMNlog(MN))$. And the computational complexity of sub-problems h^t, $\wp_t$, $\wp_t^{'}$ and $\widehat{G}_t$ is $O(DMN)$. To this end, if the number of iterations is K, the total computational complexity of the model is $O(KDMN(log(MN)+4))$.

IV. Experiments and Results

A. Experiment Implements and Evaluation metrics

We implement our ATGT using MATLAB 2017a. The ATGT is implemented on a platform with one Intel(R) Core(TM) i5-4200M CPU processor(2.50GHz), 4GB RAM. We evaluate the performance of our ATGT and other trackers on six benchmark datasets, including DTB70 [28], UAVDT-S [29], OTB100 [30], UAVDT-M [31], and UVA123@10fps [32]

For quantitative comparison, we employ the precision plot [30] and the success plot [30]. The precision plot illustrates the percentage of frames whose tracked locations are within the given threshold distance to the ground truth. A representative precision score with the threshold equal to 20 pixels is used to rank the trackers. Meanwhile, the success plot is based on the overlap ratio that is as follows:

$$s = \frac{|r_t \bigcap r_0|}{|r_t \bigcup r_0|} \tag{29}$$

where r_t is tracker bounding box, and r_0 is the ground-truth bounding box, $\bigcap$ represents an overlapping area of the two, $\bigcup$ represents a total coverage area of the two, and the $||$ represents the acreage of an area.

In addition, the accuracy of the proposed tracking algorithm (precision graph) and the success graph (success graph) are used to represent the performance of the tracking algorithm. Thus, these two evaluation indicators also serve as the evaluation criteria of our experiment. Moreover, all tracking algorithms are sorted according to the region under the image. The execution speed of a tracking algorithm is given in frame rate.

B. Performance Analysis

*1) **Comparison with CPU-based trackers**:* The results are compared with 11 state-of-the-art trackers with both HOG feature based trackers and deep-based trackers, i.e, KCF [33], DSST [34], SAMF [35], SRDCF [8], STRCF [10], ECO-HC (with gray-scale) [36] , AutoTrackC [12], GFSDCF [19], ARCF-HC [28], HOG-LR, LADCF [37], ARCF-H [28].

Results on DTB70: As can be seen from part (a), (e) of Fig. (3), ATGT performs best on DTB70, We evaluate our tracker on a dataset DTB70. Fig. (3) shows the precision and success plots of all trackers. Among the existing methods, our ATGT has the best performance with the score of 0.492 and 0.714 on precision and success plot. Compared with the AutoTrack in second place on success plots which has the precision score of 0.472 and the success score of 0.699, our ATGT tracker has improved almost 2% and 1.5% respectively. By the way, compared with the GFSDCF which has the precision score of 0.448 and the success score of 0.672, our ATGT tracker has improved almost 4.4% and 5.7%. Because we add Wasserstein distance and low rank processing to AutoTrack, the scores of AutoTrack and ATGT are compared to better reflect the advantages of ATGT, which is shown in Fig. (3) to have the best performance of all trackers. This shows that our ATGT has better tracking than AutoTrack.

Results on OTB100: In part (b) of Fig. (3), ATGT also has the best success plot with a score of 0.673 , followed by STRCF and LADCF, and their scores were are 0.661 and 0.655 . In part (f) of Fig. (3), GFSFCF and AutoTrack rank behind them with an accuracy of 0.625 and 0.591. As for the precision plot, we can see the ATGT ranks the first with 0.879, exceeding GFSDCF and Autotrack by 6.5% and 8.6% respectively. Thus, we can say that our method performs best of all 11 methods. Because of the low rank smoothing of the temporal regularization term, we can clearly improve the information redundancy caused by the multi-channel feature, compared with other algorithms, our algorithm ATGT achieves good results in gray-scale and color videos target tracking.

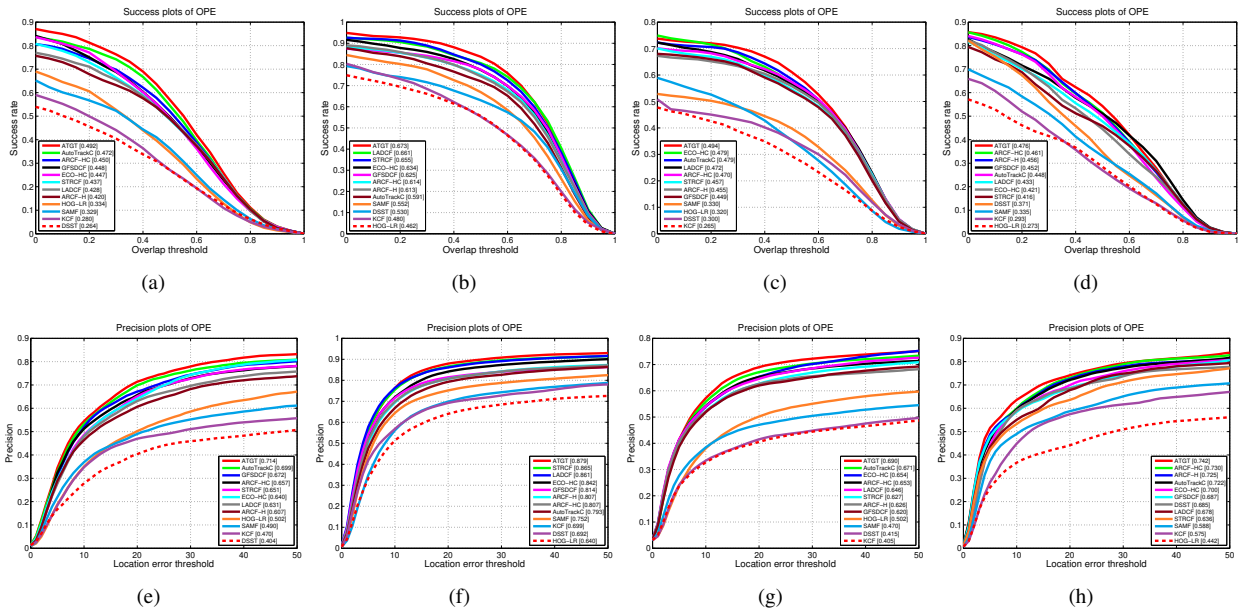


Fig. 3. Success and Precision plots of proposed methods as well as other trackers on the (a)DTB70, (b)OTB100, (c)UVA123@10fps, (d)UAVDT-S datasets.

Results on UVA123@10fps: As can be seen from part (c), (g) of Fig. (3), Overall, we can see that our method ATGT performs better than all the other state-of-the-art trackers in terms of success and precision. Compared with the second best tracker, ATGT achieves the improvement by 1.5% and 1.9% on UAV123, respectively. During tracking, because our Wasserstein distance reduces the influence of the target deformation, we can find that our ATGT algorithm has been improved obviously.

Results on UAVDT-S: In part (d),(h) of Fig. (3) show the precision and success plots of all trackers. Among the existing methods except ECO-HC, our ATGT gets the best performance with the score of 0.742 and 0.476 on precision and success plots respectively. Our ATGT performs better than AutoTrack with an AUC of 2.8%. In the process of target tracking, we adopt the Wasserstein distance of information distribution instead of the Euclidean distance from point to point, thus effectively reducing the influence of boundary effect, our Algorithm is a significant improvement over algorithms like AutoTrack.

Results on UAVDT-M: The colors red, blue and green represent the first, second, and third best. It is clear that our algorithm is better than the other 11 algorithms, in which the fraction of our algorithm in success and precision is 0.468 and 0.739, and are higher than ARCF-HC and ARCF-H. Thus, we can say that our algorithm is best performing in the UAVDT-M dataset compared to the other 11 algorithms. Combined with low rank smoothing and Wasserstein distance, ATGT based on AutoTrack provides excellent performance for state-of-the-art trackers. As shown in Fig (5), despite the interference of multiple targets in the video, our algorithm has achieved significant improvement and performance. To take the visualization clearly, we figure out the tracking results of ATGT (green wire frame) AutoTrack (blue wire frame) and GFSDCF (red wire frame) on 3 challenge video sequences for comparison, as shown in Fig. (4). As can be seen from Fig. (4), in these three video sequences, the difficulty of tracking is mainly caused by occlusion, fast movement and illumination changes respectively. Our method successfully captures the tracking target and keeps track of it all the time. The results show the accuracy and robustness of ATGT in the video sequences with challenge factors. It is clear to see that our tracker ATGT has a fps of between 1 and 3, which is roughly in line with Table I.

*2) **Comparison with Deep-based Trackers**:* We compared the trackers on the UAVDT-S dataset with deep-based feature trackers to better evaluate the performance of our ATGT. The results are compared with deep-based trackers (i.e, ADNet [38], MDNet [39], ASRCF [40], ECO [36], SiamFC [41], CFNet [42], MCPF [43], CCOT [44], CREST [45], HDT [46], FCNT [47], CF2 [48], SINT [49]) and cpu-based trackers(i.e, DSST [34], SAMF [35], STRCF [10], ECO [36] ARCF [28], HOG-LR, LADCF [37], AutoTrackc [12], GFSDCF [19], ARCF-H [28]).

As can be seen in Table (I), ATGT performs best on UAVDT-S. Table (I) shows the precision and success plots of all trackers. Among the existing methods, our ATGT wins the best performance with the score of 0.742 and 0.476 on precision and success plots respectively. Compared with the MDNet in second place on success plots which has the precision score of 0.725 and the success score of 0.464, our ATGT tracker has improved almost 1.7% and 1.2%, respectively. It is worth noting that the tracking speed of our algorithm is still relatively good performance on the UAVDT-S dataset with deep feature.

C. Tracking Process Analysis

Fig. (6) presents the overlap rate between the estimated and ground-truth bounding boxes at each frame of the BMX5 sequence from DTB70. In the sequence, when a person's posture changes so much that parts of themselves are obscured, AutoTrack does not adapt to this change, although it does

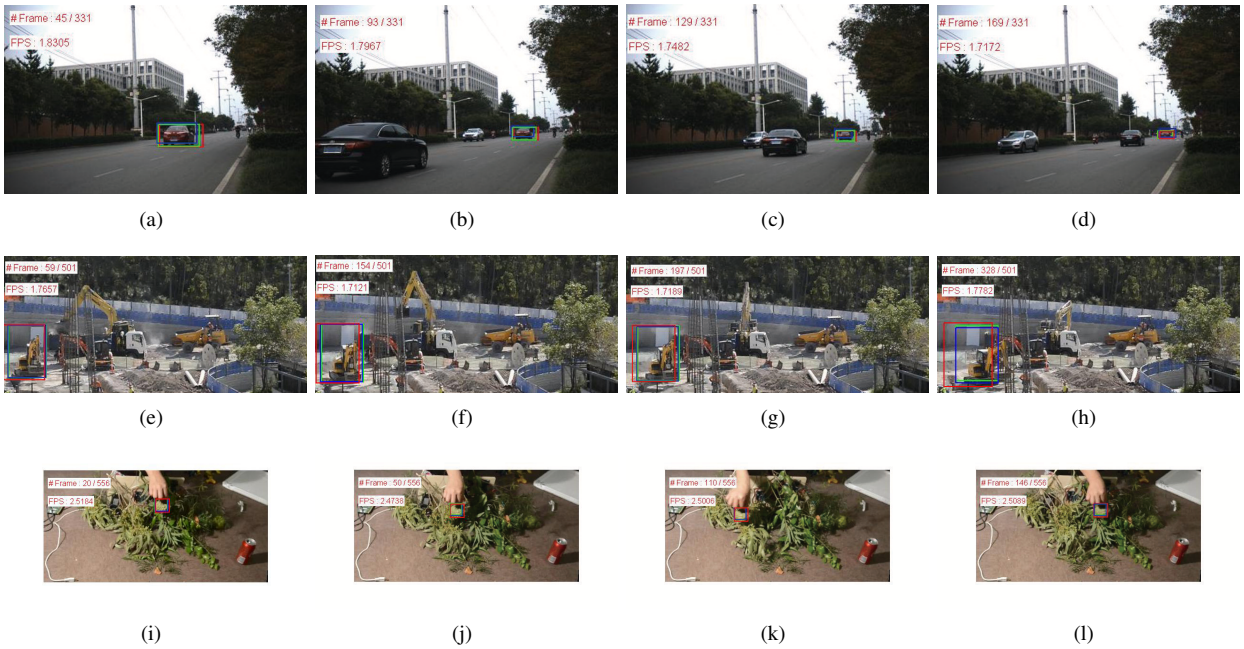


Fig. 4. The comparison of tracking for ATGT, AutoTrack and GFSDCF on 3 video sequences. Figure (a), (b), (c), (d) show how our method handles targets with fast movement. Figure (e), (f), (g), (h) show how our method handles targets with occlusion. Figure (i), (j), (k), (l) show how our method handles targets with lighting changes.

TABLE I
Comparison with the Deep Trackers and handcrafted feature trackers on UAVDT-S. The RED, GREEN, and BLUE fonts show the best three results, respectively. The superscript indicates the speed.

Tracker	Prec	Succ	Fps	Tracker	Prec	Succ	Fps	Tracker	Prec	Succ	Fps
AutoTrack	0.722	0.448	20.1	ECOHC	0.700	0.421	12.8	CCOT	0.659	0.409	0.9
GFSDCF	0.687	0.452	11.2	LADCF	0.678	0.433	11.0	CREST	0.649	0.396	4.3
ARCF	0.725	0.456	157.6	ADNet	0.683	0.429	7.5	HDT	0.596	0.303	9.0
STRCF	0.636	0.416	10.6	MDNet	0.725	0.464	1.0	FCNT	0.656	0.245	3.2
ARCF2	0.730	0.461	147.9	SINT	0.570	0.290	96.8	CF2	0.602	0.355	9.8
HOGLR	0.442	0.273	2.6	ASRCF	0.704	0.443	0.8	CFNet	0.680	0.428	41.0
DSST	0.685	0.371	44.7	ECO	0.702	0.452	16.5	MCPF	0.660	0.399	3.6
SAMF	0.588	0.335	4.3	SiamFC	0.681	0.447	37.9	ATGT	0.742	0.476	8.7

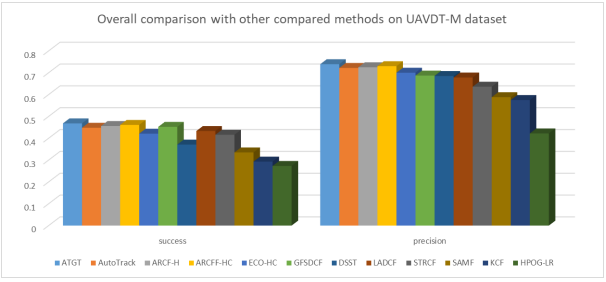


Fig. 5. Overall comparison with other compared methods on UAVDT-M dataet

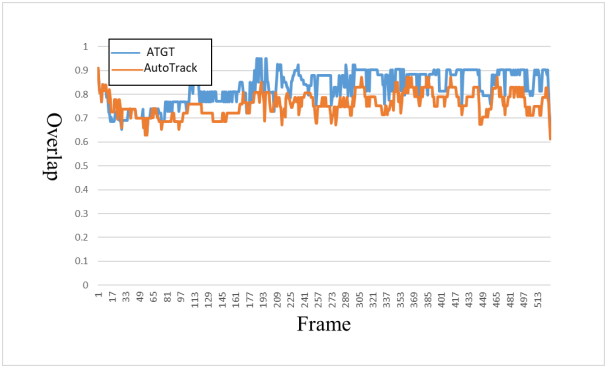
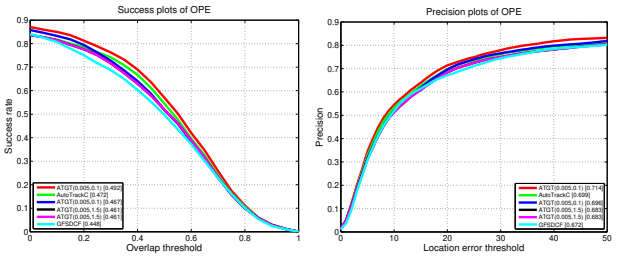


Fig. 6. Overlap of ATGT and AutoTrack on the sequence BMX5 from DTB70.

not completely lose its target as shown in Fig. (6)(before frame 54). The detection response of AutoTrack at frame #72 becomes less salient and contains more noise compared to ATGT, because AutoTrack is still affected by the redundancy of the history frame when tracking the next frame. Therefore, AutoTrack begins to drift gradually and finally fails to track the person (frame #72). As shown in Fig. (6), except for a few frames, the overlap rate of AutoTrack between the estimated and ground-truth bounding boxes remains around 0 after.

D. Ablation Analysis

In order to study the effect of the parameter setting on the tracking effect, we mainly analyze the parameter of the Wasserstein distance and the parameter of penalty term in UAVDT-S dataset, that is C and μ, and get the best parameter



(a) (b)

Fig. 7. Comparison of tracking results based on different parameters C.

setting of tracking effect. In this paper, we discuss the tracking performance of this method under different parameters by using the method of controlling variables.

In our experiment, we first fix the parameter μ to 0.005. As shown in Fig. (7), when the value of C is 0.1, both success and precision are higher 0.01 and 0.02 higher than Autotrack and GFSDCF. Then let's fix C to 0.1. From Fig 8 we can see that μ reaches its maximum value when μ is 0.005.

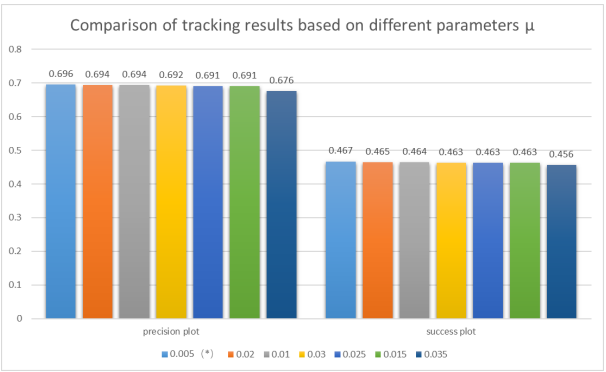


Fig. 8. Comparison of tracking results based on different parameters μ

V. Conclusion

In this paper, an UAV tracking method (ATGT) is proposed based on the distribution of temporal knowledge. The probability temporal fitting and low rank property are conducted based on the correlation filter model. To accurately and reasonably allocate the temporal parameters and reduce the influence of temporal degeneration, the Wasserstein distance is used to replace the Euclidian distance to describe the similarity of filters. The low rank constraint is used to achieve beyond response consistency. In the process of model optimization, the ADMM algorithm is used to solve the whole iterative process. Compared with state-of-the-art techniques, a large number of experiments have proved the performance of our model. However, this paper only improves the global response and the temporal domain update of the correlation filter. To improve its accuracy, it can be considered to improve the trade-off between spatial and temporal domain to improve the accuracy of target tracking.

References

[1] J. J. Pantrigo, J. Hernández, and A. Sánchez, "Multiple and variable target visual tracking for video-surveillance applications," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1577–1590, 2010.

[2] Y. Wu and T. S. Huang, "Nonstationary color tracking for vision-based human-computer interaction," *IEEE transactions on Neural Networks*, vol. 13, no. 4, pp. 948–960, 2002.

[3] Z. Lv, Y. Li, H. Feng, and H. Lv, "Deep learning for security in digital twins of cooperative intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2021, doi=10.1109/TITS.2021.3113779.

[4] S. Islam, P. X. Liu, and A. El Saddik, "Robust control of four-rotor unmanned aerial vehicle with disturbance uncertainty," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 3, pp. 1563–1571, 2014.

[5] Z. Lv, D. Chen, H. Feng, H. Zhu, and H. Lv, "Digital twins in unmanned aerial vehicles for rapid medical resource delivery in epidemics," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–9, 2021, doi=10.1109/TITS.2021.3113787.

[6] W. Wang, X. Li, L. Xie, H. Lv, and Z. Lv, "Unmanned aircraft system airspace structure and safety measures based on spatial digital twins," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2021, doi=10.1109/TITS.2021.3108995.

[7] C. Fu, J. Jin, F. Ding, Y. Li, and G. Lu, "Spatial reliability enhanced correlation filter: An efficient approach for real-time uav tracking," *IEEE Transactions on Multimedia*, pp. 1–1, 2021, doi=10.1109/TMM.2021.3118891.

[8] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4310–4318.

[9] ——, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4904–4913.

[11] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6309–6318.

[12] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 923–11 932.

[13] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 448–10 457.

[14] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8126–8135.

[15] R. Han, W. Feng, and S. Wang, "Fast learning of spatially regularized and content aware correlation filter for visual tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 7128–7140, 2020.

[16] H. Zhu, H. Peng, G. Xu, L. Deng, Y. Cheng, and A. Song, "Bilateral weighted regression ranking model with spatial-temporal correlation filter for visual tracking," *IEEE Transactions on Multimedia*, 2021.

[17] F. Li, C. Fu, F. Lin, Y. Li, and P. Lu, "Training-set distillation for real-time uav object tracking," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9715–9721.

[18] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time uav tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2891–2900.

[19] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Joint group feature selection and discriminative filter learning for robust visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7950–7960.

[20] Y. Li, J. Zhu, S. C. Hoi, W. Song, Z. Wang, and H. Liu, "Robust estimation of similarity transformation for visual object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8666–8673.

[21] M. Danelljan, L. V. Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] T. Zhang, K. Jia, C. Xu, Y. Ma, and N. Ahuja, "Partial occlusion handling for visual tracking via robust part matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1258–1265.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 214–223.

[24] Y. He, M. Li, J. Zhang, and J. Yao, "Infrared target tracking based on robust low-rank sparse learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 2, pp. 232–236, 2015.

[25] F.-Y. Wang, "Probability distance inequalities on riemannian manifolds and path spaces," *Journal of Functional Analysis*, vol. 206, no. 1, pp. 167–190, 2004.

[26] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011, vol. 3, no. 1.

[27] M. Huang, S. Ma, and L. Lai, "Robust low-rank matrix completion via an alternating manifold proximal gradient continuation method," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2639–2652, 2021.

[28] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time uav tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2891–2900.

[29] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.

[30] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2411–2418.

[31] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[33] M. Tang and J. Feng, "Multi-kernel correlation filter for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3038–3046.

[34] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*. Bmva Press, 2014, pp. 1–11.

[35] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 254–265.

[36] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6638–6646.

[37] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5596–5609, 2019.

[38] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Young Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2711–2720.

[39] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4293–4302.

[40] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4670–4679.

[41] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 850–865.

[42] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2805–2813.

[43] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4335–4343.

[44] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 472–488.

[45] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. Lau, and M.-H. Yang, "Crest: Convolutional residual learning for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2555–2564.

[46] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4303–4311.

[47] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3119–3127.

[48] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3074–3082.

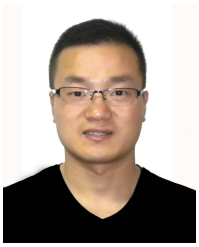
[49] C. Jiang, J. Xiao, Y. Xie, T. Tillo, and K. Huang, "Siamese network ensemble for visual tracking," *Neurocomputing*, vol. 275, pp. 2892–2903, 2018.

Guoxia Xu (Member) received the B.S. degree in information and computer science from Yancheng Teachers University, Jiangsu Yancheng, China in 2015, and the M.S. degree in computer science and technology from Hohai University, Nanjing, China in 2018. He was a research assistant in City University of Hong Kong and Chinese University of Hong Kong. Now, he is pursing his Ph.D. degree in Department of Computer Science, Norwegian University of Science and Technology, Gjøvik Norway. His research interest includes pattern recognition, image processing, and computer vision. E-mail: gxxu.re@gmail.com

Hao Wang (Senior Member) received the B.Eng. and Ph.D. degrees, both in computer science and engineering, from South China University of Technology, China in 2000 and 2006, respectively. He is an Associate Professor in the Department of Computer Science, Norwegian University of Science & Technology, Norway. His research interests include big data analytics, industrial internet of things, high performance computing, and safety critical systems. He has published 170+ papers in reputable international journals and conferences including IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Industrial Electronics, Transactions on Industrial Informatics, Internet of Things Journal, Transactions on Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on Geoscience and Remote Sensing and ACM Computing Surveys. He serves as the Editorial Board Member and Guest Editors for several international journals. He served as a TPC Co-Chair for IEEE Transactions on Computational Social Systems 2020, IEEE International Conference on Computer and Information Technology 2017, International Conference on Big Data Intelligence and Computing 2015, and Reviewers for many prestigious journals and conferences. He is a Senior Member of IEEE and a Member of ACM. He is the Chair for Sub TC on Healthcare in IEEE IES Technical Committee on Industrial Informatics. E-mail: hawa@ntnu.no

Meng Zhao (Member) received the Ph.D. degree from Tianjin University, in 2016. She is currently an Associate Professor with the Tianjin University of Technology. She was granted the Alain Bensoussan Fellowship by European Research Consortium for Informatics and Mathematics, in 2020. Her research interests include medical image processing and computer vision.

Marius Pedersen (Member) received the B.Sc. degrees in computer engineering and media technology (MiT) from the Gjøvik University College, Norway, in 2006 and 2007, and the Ph.D. degree in color imaging from the University of Oslo, Norway, in 2011, sponsored by Océ. He is currently a Full Professor with the Department of Computer Science, NTNU, Gjøvik, Norway. He is also the Director of The Norwegian Colour and Visual Computing Laboratory (Colourlab). His research interest includes subjective and objective image quality.



Hu Zhu (Member) received the B.S. degree in mathematics and applied mathematics from Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and the M.S. and Ph.D. degrees in computational mathematics and pattern recognition and intelligent systems from Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2013, respectively. Now, he is professor in the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include pattern recognition, image processing, and computer vision. E-mail: peter.hu.zhu@gmail.com