IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 7, 2022, accepted March 22, 2022, date of publication March 25, 2022, date of current version March 31, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3162200

Orthogonal Single-Target Tracking

YOUJIN KIM AND JUNSEOK KWON, (Member, IEEE)
School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Junseok Kwon (jskwon@cau.ac.kr)

This work was supported in part by the Chung-Ang University Graduate Research Scholarship Grants in 2021, and in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2020R1C1C1004907.

ABSTRACT In this study, we propose a novel Wasserstein distributional tracking method that can balance approximation with accuracy in terms of Monte Carlo estimation. To achieve this goal, we present three different systems: sliced Wasserstein-based (SWT), projected Wasserstein-based (PWT), and orthogonal coupled Wasserstein-based (OCWT) visual tracking systems. Sliced Wasserstein-based visual trackers can find accurate target configurations using the optimal transport plan, which minimizes the discrepancy between appearance distributions described by the estimated and ground truth configurations. Because this plan involves a finite number of probability distributions, the computation costs can be considerably reduced. Projected Wasserstein-based and orthogonal coupled Wasserstein-based visual trackers further enhance the accuracy of visual trackers using bijective mapping functions and orthogonal Monte Carlo, respectively. Experimental results demonstrate that our approach can balance computational efficiency with accuracy, and the proposed visual trackers outperform other state-of-the-art visual trackers on several benchmark visual tracking datasets.

INDEX TERMS Computer vision, distance measurement, probability distribution.

I. INTRODUCTION

Visual tracking is a fundamental technique that can be used to predict target object (*e.g.*, vehicle) trajectories. Recently, visual tracking has enhanced its performance by defining visual tracking problems in the Wasserstein space. This Wasserstein space enables the accurate measurement of the distance between probability distributions. Because it can handle probability distributions, the Wasserstein distance has been used in various computer vision applications (*e.g.*, classification [1], detection [2], visual tracking [3], and 3D representation [4]) and has been applied to several machine learning tasks (*e.g.*, semi-supervised learning [5], adversarial learning [6], meta learning [7], reinforcement learning [8], and metric learning [9]).

Conventional visual tracking typically adopts the matching metrics in the Euclidean space, *e.g.*, l_1 and l_2 norms, Kullback Leibler divergence, and Jensen-Shannon divergence, while having several limitations under real-world visual-tracking environments. For example, l_1 and l_2 norms cannot accurately measure the discrepancy between the distributions. Kullback Leibler divergence is asymmetric, whereas Jensen-Shannon divergence is discontinuous and is not proportional to the discrepancy between the distributions. Thus, a new matching metric is required in the Wasserstein space, which has been rarely explored in visual tracking. In particular, the Wasserstein distance can measure the discrepancy between probability distributions of the reference appearance and the current target appearance at the estimated state. Because visual trackers explicitly consider the discrepancy of probability distributions, they can encode the uncertainty in measuring the distance from the distributional perspective.

However, calculating the Wasserstein distance requires high computational costs and is intractable in real-world settings with limited resources. To alleviate this problem, the following methods attempt to approximate the Wasserstein distance: For example, Kolouir *et al.* [10] projected the Wasserstein distance into one-dimensional spaces and presented the sliced Wasserstein distance. Cuturi *et al.* [11] transformed the optimal transport problems into maximum-entropy problems to speed up the computation and introduced the Sinkhorn distance. Genevaay *et al.* [12] proposed a stochastic optimization method for dealing with large-scale optimal transport problems. While these methods have made the distance computation tractable, they inevitably degrade the Wasserstein distance accuracy.

Thus, it is important to balance the approximation with accuracy in the computation of the Wasserstein distance.

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

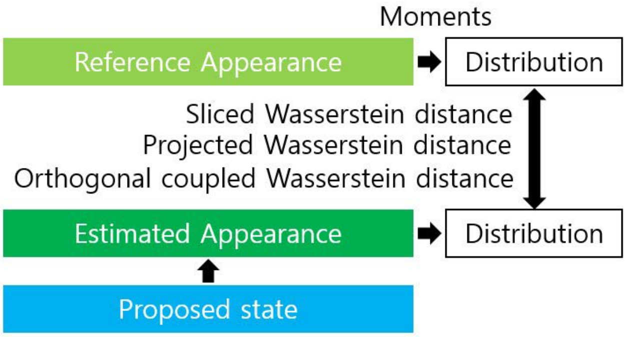

Moments
Reference Appearance
Distribution
Sliced Wasserstein distance
Projected Wasserstein distance
Orthogonal coupled Wasserstein distance
Estimated Appearance
Distribution
Proposed state

FIGURE 1. **Framework of the proposed visual tracking system. The proposed visual tracker proposes a new state at each time and estimate the target appearance. Then, our visual tracker compares the reference target appearance with the estimated target appearance from the distributional perspective using three Wasserstein-based distances, which are Sliced Wasserstein distance, Projected Wasserstein distance, and Orthogonal coupled Wasserstein distance.**

For this purpose, we adopt a variant of the sliced Wasserstein distance augmented by orthogonal coupling in the course of Monte Carlo simulation on the Wasserstein distance [13], called orthogonal coupled Wasserstein (OCW). Our OCW method can preserve the distance information in high-dimensional space, although the method approximates the Wasserstein distance to reduce computational cost.

In this study, we aim to solve visual tracking problems using the proposed OCW. The proposed visual tracking method represents a target appearance vector as a target appearance distribution to cope with ambiguities in the appearance representation. Subsequently, the OCW accurately and efficiently minimizes the discrepancy between the estimated and ground-truth target appearance distributions to obtain an accurate target configuration.

The contributions of the proposed method are as follows:

- We develop a novel sliced Wasserstein-based visual tracking system (SWT), in which two appearance distributions described by estimated configurations and ground truth configurations become similar via the optimal transport plan. This plan can be conducted using a finite number of probability distributions; thus, the computational costs can be considerably reduced.
- We present a novel projected Wasserstein-based visual tracking system (PWT), in which the discrepancy between the aforementioned sliced Wasserstein distance and true Wasserstein distance can be minimized using bijective mapping functions.
- We propose a novel orthogonal coupled Wasserstein-based visual tracking system (OCWT), in which the aforementioned projected distance can induce accurate projection directions using orthogonal Monte Carlo.

Figure 1 describes the framework of the visual tracking system.

The remainder of this paper is organized as follows. Section II relates the proposed method to the existing methods. Sections III, IV, and V propose a visual tracking method based on the sliced Wasserstein, projected Wasserstein, and orthogonal coupled Wasserstein distances, respectively. Section VI-A describes the experimental settings used in this study. We compare the proposed visual tracker with other state-of-the-art methods using the object tracking benchmark (OTB) and large-scale single object tracking (LaSOT) datasets in Sections VI-C and VI-D, respectively. Section VI-B analyzes our proposed visual trackers in depth. We conclude the study in Section VII.

II. RELATED WORK

While visual tracking has a long history, in this section, we discuss the methods most relevant to our study, which can be categorized into three groups: Wasserstein distributional visual tracking, visual tracking via projection, and deep learning-based visual tracking.

A. WASSERSTEIN DISTRIBUTIONAL VISUAL TRACKING

Yao *et al.* [14] transformed visual tracking problems into transportation problems via linear programming algorithms, where 1-Wasserstein distances (*i.e.*, earth mover's distances) were used as a distance metric. Danu *et al.* [15] employed the Wasserstein distance in a particle filter formulation to compare estimated multi-target states with ground truths in multi-sensor environments. Zeng *et al.* [3] measured the discrepancy between target-specific features using the 1-Wasserstein distance to accurately track vehicles. Danis *et al.* [16] used the Wasserstein distance to evaluate Bluetooth data via a sequential Monte Carlo method.

In contrast to these methods that use Wasserstein distributions to enhance the visual tracking accuracy, we use the orthogonal coupled Wasserstein distance to balance the accuracy with computational efficiency.

B. VISUAL TRACKING VIA PROJECTION

Xiao *et al.* [17] designed random projection matrices to find subspaces that make visual trackers robust to noise. Zhang *et al.* [18] transformed visual tracking problems into projection problems, in which a robust target representation model is learned via a projection onto the $l + p$ ball. Zhang *et al.* [19] proposed a visual tracker based on a structurally random projection for dimensionality reduction of the template space, in which the original distance was preserved with an efficient computation. Danelljan *et al.* [20] projected color names on an orthonormal basis of a 10-dimensional subspace to extract sophisticated color features for visual tracking.

In contrast to these methods that project the Euclidean space into the subspaces of the target appearance, we project the Wasserstein space and explicitly guide the projection direction for accurate visual tracking.

C. DEEP LEARNING-BASED VISUAL TRACKING

Li *et al.* [21] presented Siamese deep neural architectures combined with region proposal networks, which aimed to search for candidate regions for target objects. Valmadre *et al.* [22] proposed deep neural networks based on correlation filters that efficiently compared deep features with reference features. Zhang *et al.* [23] introduced very deep neural networks to extract representative features for accurate visual tracking. Bertinetto *et al.* [24] made Siamese networks fully convolutional for accurate and fast matching. Li *et al.* [25] applied meta information to deep neural networks for fast adaptation in different visual tracking environments and changes in target appearances. Zhu *et al.* [26] enhanced the discriminative power of deep neural networks using both negative and positive samples for target objects. Choi *et al.* [27] boosted the adaptive representation ability of deep neural networks using gradient information for visual tracking. Bhat *et al.* [28] used discriminative classifiers for deep neural networks, in which classifier weights were generated via a novel optimization technique. Guo *et al.* [29] presented dynamic Siamese network architectures that enable the update of target appearances online.

In contrast to these methods, we do not use complex deep neural architectures. Nevertheless, our proposed visual tracker exhibits state-of-the-art visual tracking performance, because target appearances are described by Wasserstein distributions; thus, several variations in target appearances can be covered during visual tracking.

D. OTHER VISUAL TRACKING

Li *et al.* [30] proposed a dual-regression framework for visual tracking, which combines discriminative fully convolutional module (for discriminative ability) and a fine-grained correlation filter (for accurate localization). Fan *et al.* [31] introduced a novel interactive learning framework for visual tracking, in which multiple convolutional filter models are interacted with each other and their responses are fused based on the confidence scores. Liu *et al.* developed robust visual trackers for thermal infrared objects based on multi-level similarity models under the Siamese framework [32], via the multi-task framework [33], and using the pretrained convolutional neural networks [34].

Muresan *et al.* [35] introduced a multi-object tracking method based on a affinity measurement function and a context aware descriptor for 3D objects. Karunasekera *et al.* [36] presented a multi-object visual tracking system using a new dissimilarity measure that considers object motion, appearance, structure, and size. Braso and Laura [37] proposed fully differentiable message passing networks for multi-object tracking, which is formulated as network flows.

In contrast to these methods, we presented a novel mathematical approach based on the Wasserstein distance to boost the visual tracking performance. Thus, this approach can be integrated into existing visual trackers to improve their performance. Please note that using the Wasserstein distance enables us to use many of useful mathematical properties.

III. SLICED WASSERSTEIN-BASED VISUAL TRACKING

A. SLICED WASSERSTEIN DISTANCE

The $p-$Wasserstein distance $\mathcal{W}_p$ measures the discrepancy between two probability distributions (*i.e.*, $\mu, \nu \in \mathscr{P}\left(\mathbb{R}^d\right)$), where $\mathscr{P}\left(\mathbb{R}^d\right)$ denotes the set of distributions defined on $\mathbb{R}^d$ and the p-th moment. We then define the $p-$Wasserstein distance as follows:

$$\mathcal{W}_p(\mu,\nu) = \left[\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} ||x-y||_2^p \gamma(dx,dy)\right]^{1/p}, \quad (1)$$

where $\Gamma(\mu,\nu)$ denotes the set of joint probability distributions defined on $\mathbb{R}^d \times \mathbb{R}^d$ (*i.e.*, $\Gamma(\mu,\nu) \subseteq \mathscr{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right)$). In (1), we can find the optimal transport plan γ between μ and ν, inducing $\mathcal{W}_p$.

The Wasserstein distance in (1) can directly consider probability distributions. However, it is difficult to define the set of joint probability distributions $\Gamma(\mu,\nu)$. Thus, conventional approaches [38] approximate ν as $\{\nu_m\}_{m=1}^M$ and $\mathcal{W}_p(\mu,\nu)$ as $\arg_\mu \min \sum_{m=1}^M w_m \mathcal{W}_p(\mu,\nu_m)$, where w_m denotes the m-th weight. As an alternative approach, μ and ν are assumed to have one-dimensional probability distributions (*i.e.*, $\mu,\nu \in \mathscr{P}\left(\mathbb{R}^1\right)$). Then, we can find the optimal transport plan γ using a finite number of probability distributions, which can considerably reduce the computational costs. This approach induces a sliced Wasserstein distance [13], [39].

To compute the sliced Wasserstein distance, we define the unit sphere S^{d-1} in $\mathbb{R}^d$. Subsequently, for a vector $s \in S^{d-1}$, we define the projection map *proj*, which transforms $x \in \mathbb{R}^d$ into $\prec s, x \succ \in \mathbb{R}^1$ (*i.e.*, $proj_s(x) = \prec s, x \succ$). We define the projection of the probability distribution μ as $proj_s^\#(\mu)$. Using $proj_s^\#(\mu)$, we can deal with one-dimensional probability distributions. Then, the sliced Wasserstein distance $\mathcal{W}_p^{slice}$ is defined as follows:

$$\mathcal{W}_p^{slice}(\mu,\nu) = \mathbb{E}_{s\in S^{d-1}} \mathcal{W}_p\left(proj_s^\#(\mu), proj_s^\#(\nu)\right). \quad (2)$$

In (2), $\mathbb{E}$ is implemented via a Monte Carlo simulation with N samples (*i.e.*, $s_1, \cdots, s_N \in S^{d-1}$) as (3).

$$\widetilde{\mathcal{W}}_p^{slice}(\mu,\nu) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{W}_p\left(proj_{s_n}^\#(\mu), proj_{s_n}^\#(\nu)\right). \quad (3)$$

B. VISUAL TRACKING

With $\widetilde{\mathcal{W}}_p^{slice}$, we present a novel sliced Wasserstein distance-based visual tracker. In the visual tracking context, μ and ν indicate the estimated and ground-truth target appearance distributions, respectively. We adopt empirical distributions for μ and ν, which are defined as follows:

$$\mu = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}(x_m),\ \nu = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(y_m). \quad (4)$$

In (4), $\mathbb{I}(x_m)$ denotes an indicator function (*i.e.*, $\mathbb{I}(x_m) = 1$, if $x = x_m$; otherwise, $\mathbb{I}(x_m) = 0$). We extract M appearance feature vectors using M moments.

We define a target object configuration at time t as $O_t = \{o_1, o_2, o_3\}$, where o_1, o_2, and o_3 denote x-axis position, y-axis position, and scale of the target in an image, respectively. Given the best target configuration at time $t-1$, $\hat{O}_{t-1}$, our goal of visual tracking is to find the best target configuration at time t, $\hat{O}_t$. For this purpose, we randomly

Algorithm 1 Sliced Wasserstein Distance-Based Tracker (SWT)

```
Input: $\hat{O}_{t-1}$
Output: $\hat{O}_t$
$\left\{O_t^{(c)}\right\}_{c=1}^{C} \sim \mathcal{N}\left(\hat{O}_{t-1}, \Sigma^2\right)$
for $c = 1$ to $C$ do
    $\widetilde{\mathcal{W}}_p^{slice}(\mu^{(c)}, \nu) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{W}_p\left(proj_{s_n}^{\#}(\mu^{(c)}), proj_{s_n}^{\#}(\nu)\right)$
end for
$c^* = \arg\min_c \widetilde{\mathcal{W}}_p^{slice}\left(\mu^{(c)}, \nu\right)$ for $c = 1, \cdots, C$
$\hat{O}_t = O_t^{(c^*)}$
```

Algorithm 2 Projected Wasserstein Distance-Based Tracker (PWT)

```
Input: $\hat{O}_{t-1}$
Output: $\hat{O}_t$
$\left\{O_t^{(c)}\right\}_{c=1}^{C} \sim \mathcal{N}\left(\hat{O}_{t-1}, \Sigma^2\right)$
for $c = 1$ to $C$ do
    $\mathcal{W}_p^{proj}(\mu^{(c)}, \nu) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M} \left|\left|proj_{s_n^{new}}(x_m) - proj_{s_n^{new}}(b(y_m))\right|\right|_2^p$
end for
$c^* = \arg\min_c \mathcal{W}_p^{proj}\left(\mu^{(c)}, \nu\right)$ for $c = 1, \cdots, C$
$\hat{O}_t = O_t^{(c^*)}$
```

Algorithm 3 Orthogonal Coupled Wasserstein Distance-Based Tracker (OCWT)

```
Input: $\hat{O}_{t-1}$
Output: $\hat{O}_t$
$\left\{O_t^{(c)}\right\}_{c=1}^{C} \sim \mathcal{N}\left(\hat{O}_{t-1}, \Sigma^2\right)$
for $c = 1$ to $C$ do
    $\widetilde{\mathcal{W}}_p^{ort}(\mu^{(c)}, \nu) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{W}_p\left(proj_{s_n^{ort}}^{\#}(\mu^{(c)}), proj_{s_n^{ort}}^{\#}(\nu)\right)$
end for
$c^* = \arg\min_c \widetilde{\mathcal{W}}_p^{ort}\left(\mu^{(c)}, \nu\right)$ for $c = 1, \cdots, C$
$\hat{O}_t = O_t^{(c^*)}$
```

search for candidate configurations around $\hat{O}_{t-1}$. Thus, our motion model is based on a normal distribution, as follows.

$$\left\{O_t^{(c)}\right\}_{c=1}^{C} \sim \mathcal{N}\left(\hat{O}_{t-1}, \Sigma^2\right). \tag{5}$$

In (5), $O_t^{(c)}$ denotes the c-th candidate configuration that is proposed based on a normal distribution with center $\hat{O}_{t-1}$ and standard deviation Σ. Subsequently, we measure the sliced Wasserstein distance $\widetilde{\mathcal{W}}_p^{slice}$ between appearance distributions described by candidate configuration $O_t^{(c)}$ and ground truth configuration O_t^{GT}, which are $\mu^{(c)}$ and ν, respectively. Our objective is to find the best index c^*, in which the corresponding appearance distribution $\mu^{(c)}$ described by candidate configuration $O_t^{(c)}$ can minimize the distance:

$$c^* = \arg\min_c \widetilde{\mathcal{W}}_p^{slice}\left(\mu^{(c)}, \nu\right) \quad \text{for } c = 1, \cdots, C. \tag{6}$$

In (6), the best target configuration at time t is $\hat{O}_t = O_t^{(c^*)}$. Algorithm 1 shows the entire pipeline of the proposed visual tracker based on the sliced Wasserstein distance.

IV. PROJECTED WASSERSTEIN-BASED VISUAL TRACKING

A. PROJECTED WASSERSTEIN DISTANCE

Using the sliced Wasserstein distance, we can considerably reduce the computational cost, but can obtain erroneous results, because there exists discrepancy between sliced Wasserstein distance and true Wasserstein distance. In particular, according to s in (3), the projected vector $proj_s(x)$ can be biased [40]. In particular, $proj_s(x) < proj_s(x')$ does not make $proj_s(y) < proj_s(y')$. To solve this problem, bijective mapping has been introduced to measure the sliced Wasserstein distance [13]. Bijective mapping induces

$$proj_s\big(b(y)\big) < proj_s\big(b(y')\big), \text{ if } proj_s(x) < proj_s(x'). \tag{7}$$

In (7), the bijective mapping $b(\cdot)$ can be implemented by sorting $\{y_m\}_{m=1}^{M}$, which results in $\left\{y_m^{sort}\right\}_{m=1}^{M}$, and selecting $y_{\texttt{argsort}(x_m)}^{sort}$ for x_m, where `argsort` returns indices that sort $\{x_m\}_{m=1}^{M}$ and $\texttt{argsort}(x_m)$ returns the index of x_m.

Subsequently, the projection is conducted using a new projection vector $s^{new} \in \mathbb{S}^{d-1}$, which is different from s in (7). Using s^{new}, we can prevent the aforementioned projection from being biased. The projected Wasserstein distance is then defined as (8).

$$\mathcal{W}_p^{proj}(\mu, \nu) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\left|\left|proj_{s_n^{new}}(x_m) - proj_{s_n^{new}}(b(y_m))\right|\right|_2^p, \tag{8}$$

where $x_m \sim \mu$ and $y_m \sim \nu$ as in (4).

B. VISUAL TRACKING

Our objective is to find the best index c^*, in which the corresponding appearance distribution $\mu^{(c)}$ described by candidate configuration $O_t^{(c)}$ can minimize the distance as (9).

$$c^* = \arg\min_c \mathcal{W}_p^{proj}\left(\mu^{(c)}, \nu\right) \quad \text{for } c = 1, \cdots, C, \tag{9}$$

where the best target configuration at time t is $\hat{O}_t = O_t^{(c^*)}$. Algorithm 2 shows the entire pipeline of the proposed visual tracker based on the projected Wasserstein distance.

V. ORTHOGONAL COUPLED WASSERSTEIN-BASED TRACKING

A. ORTHOGONAL COUPLED WASSERSTEIN DISTANCE

Using the projected Wasserstein distance, we can reduce the discrepancy between the sliced Wasserstein distance and the true Wasserstein distance. However, the projection direction of s in $proj_s$ is crucial for the success of the projected Wasserstein distance, as mentioned in [41]. In this context, we use

TABLE 1. Quantitative comparison of the proposed methods. The best results are written in boldface.

	SWT	PWT	OCWT
AUC	0.512	0.517	**0.523**
Precision	0.529	0.533	**0.535**
Normalized precision	0.602	0.606	**0.607**

TABLE 2. Analysis of the proposed OCWT according to different values of *N* (Monte Carlo samples) in (3). The best results are written in boldface.

	50	100	200
AUC	0.518	**0.523**	**0.523**
Precision	0.533	0.535	**0.536**
Normalized precision	0.606	0.607	**0.609**

orthogonal directions, because orthogonal directions of projection vectors guarantee the improvement of estimator variance for the projected Wasserstein distance, as proven in [13]. To sample mutually orthogonal vectors $s_1^{ort}, \cdots, s_N^{ort} \in S^{d-1}$ (*i.e.*, $\prec s_i^{ort}, s_j^{ort} \succ = 0$ for $i \neq j$), we employ orthogonal Monte Carlo (OMC) techniques in [42]. Using the OMC, mutually orthogonal vectors can be efficiently obtained form the unit sphere S^{d-1} in $\mathbb{R}^d$.

Let $\mathbb{G}$ be a d-dimensional Givens rotation [43]. Then, $\mathbb{G}$ is an an orthogonal matrix in S^{d-1}, which is parameterized with two indices $i, j \in \{1, \cdots d\}$ and an angle $\theta \in [0, 2\pi)$, as (10).

$$\mathbb{G}[i, j, \theta]_{k,l} = \begin{cases} \cos(\theta) & \text{if } k = l \in \{i, j\} \\ -\sin(\theta) & \text{if } k = i, l = j \\ \sin(\theta) & \text{if } k = j, l = i \\ 1 & \text{if } k = l \notin \{i, j\} \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where all coordinates of $\mathbb{R}^d$ are fixed except i and j, and the two-dimensional subspace is spanned using the rotation of θ. Using $\mathbb{G}[i, j, \theta]$, we can sample orthogonal vectors via Kac's random walk on the Markov chain $\mathbb{K}_t|_{t=1}^{\infty}$.

$$\mathbb{K}_{1:T} = \prod_{t=1}^{T} \mathbb{G}[i_t, j_t, \theta_t]. \tag{11}$$

In (11), the sequence of $\mathbb{K}_t \times s_n^{ort}$ is a Markov chain on S^{d-1} [44]. Then, the orthogonal coupled Wasserstein distance $\mathcal{W}_p^{ort}$ is defined as follows.

$$\mathcal{W}_p^{ort}(\mu, \nu) = \mathbb{E}_{s \in S^{d-1}} \mathcal{W}_p \left(proj_{s^{ort}}^{\#}(\mu), proj_{s^{ort}}^{\#}(\nu) \right). \tag{12}$$

In (12), $\mathbb{E}$ is implemented via a Monte Carlo simulation with N samples (*i.e.*, $s_1^{ort}, \cdots, s_N^{ort} \in S^{d-1}$) as (13).

$$\widetilde{\mathcal{W}}_p^{ort}(\mu, \nu) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{W}_p \left(proj_{s_n^{ort}}^{\#}(\mu), proj_{s_n^{ort}}^{\#}(\nu) \right). \tag{13}$$

B. VISUAL TRACKING

Our objective is to find the best index c^*, in which the corresponding appearance distribution $\mu^{(c)}$ described by candidate configuration $O_t^{(c)}$ can minimize the distance as (14):

$$c^* = \arg\min_c \widetilde{\mathcal{W}}_p^{ort} \left(\mu^{(c)}, \nu \right) \quad \text{for } c = 1, \cdots, C, \tag{14}$$

TABLE 3. Analysis of the proposed OCWT according to different values of *C* (candidate configurations) in (5). The best results are written in boldface.

	5	10	20
AUC	0.521	**0.523**	0.522
Precision	0.532	**0.535**	0.534
Normalized precision	0.605	**0.607**	**0.607**

TABLE 4. Analysis of the proposed OCWT according to different values of *M* (moment statistics) in (4). The best results are written in boldface.

	1	4	6
AUC	0.518	**0.523**	**0.523**
Precision	0.529	0.535	**0.536**
Normalized precision	0.599	**0.607**	**0.607**

where the best target configuration at time t is $\hat{O}_t = O_t^{(c^*)}$. Algorithm 3 shows the entire pipeline of our visual tracker based on the orthogonal coupled Wasserstein distance.

VI. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) OTB DATASET

To demonstrate the effectiveness of the proposed methods, we compared three proposed visual trackers (*i.e.*, **SWT, PWT**, and **OSWT**) with 9 recent deep learning-based visual trackers (*i.e.*, ECO-HC [45], TADT [46], SiamRPN++ [21], SINT-op [47], C-COT [48], DAT [49], ECO [45], SiamDW [23], and SINT [47]) using the OTB dataset [50]. This dataset includes various attributes for visual tracking environments, including out-of-view, out-of-plane rotation, deformations, motion blur, scale variation, illumination change, fast motion, background clutter, in-plane rotation, low resolution, and occlusions. To evaluate the visual tracking methods, precision and success plots, and the area under the curve (AUC) were used, in which the precision plot computed the ratio of frames such that the discrepancy between the estimated and ground-truth configurations of the targets is less than a specific threshold. The success plot computed the percentage of frames such that the intersection of the union between the estimated and ground-truth bounding boxes is greater than a specific threshold. AUC was used to compute the area under the success plot.

2) LaSOT DATASET

We also compared our visual trackers with visual trackers (*e.g.*, StructSiam [51], DASiam [26], GlobalTrack [52], SiamRPN++ [21], ATOM [53], ECO [45], CFNet [22], and SPLT [54]) including state-of-the-art correlation filter-based trackers (*e.g.*, GFSDCF [55], ASRCF [56], STRCF [57], and BACF [58]) using the LaSOT dataset [59]. This dataset contains 1, 400 test sequences, in which the average length is greater than 2, 512 frames. To evaluate the visual tracking methods, precision, normalized precision, and area under the curve were used.

3) VOT DATASET

In addition, we compared visual trackers (*e.g.*, CFCF [60], LSART [61], CFWCR [62], and ECO [45]) using the

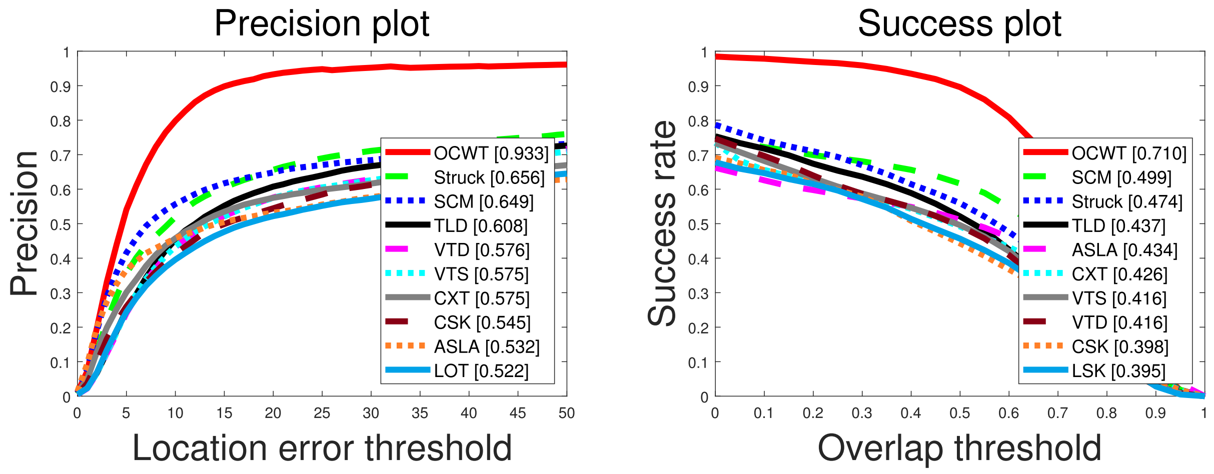


FIGURE 2. Quantitative comparison with non-deep-learning visual trackers using the OTB dataset.

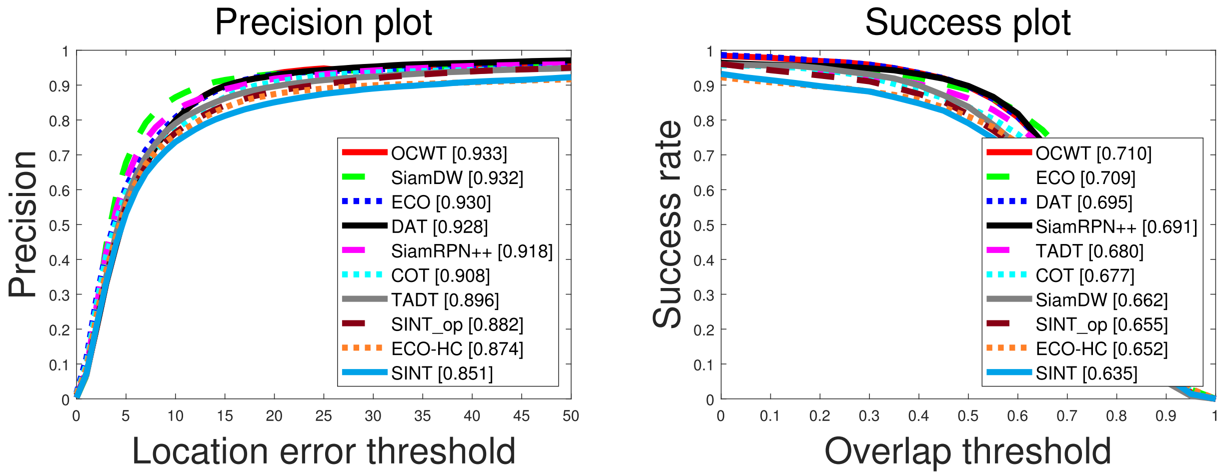


FIGURE 3. Quantitative comparison with deep-learning visual trackers using the OTB dataset.

TABLE 5. Analysis of the proposed OCWT according to different values of *T* (frames) in (11). The best results are written in boldface.

	30	90	120
AUC	0.522	0.523	**0.524**
Precision	0.533	0.535	**0.536**
Normalized precision	0.606	**0.607**	**0.607**

VOT2017 dataset, which contains 60 videos with diverse attributes. To evaluate the visual trackers, accuracy and robustness metrics were used.

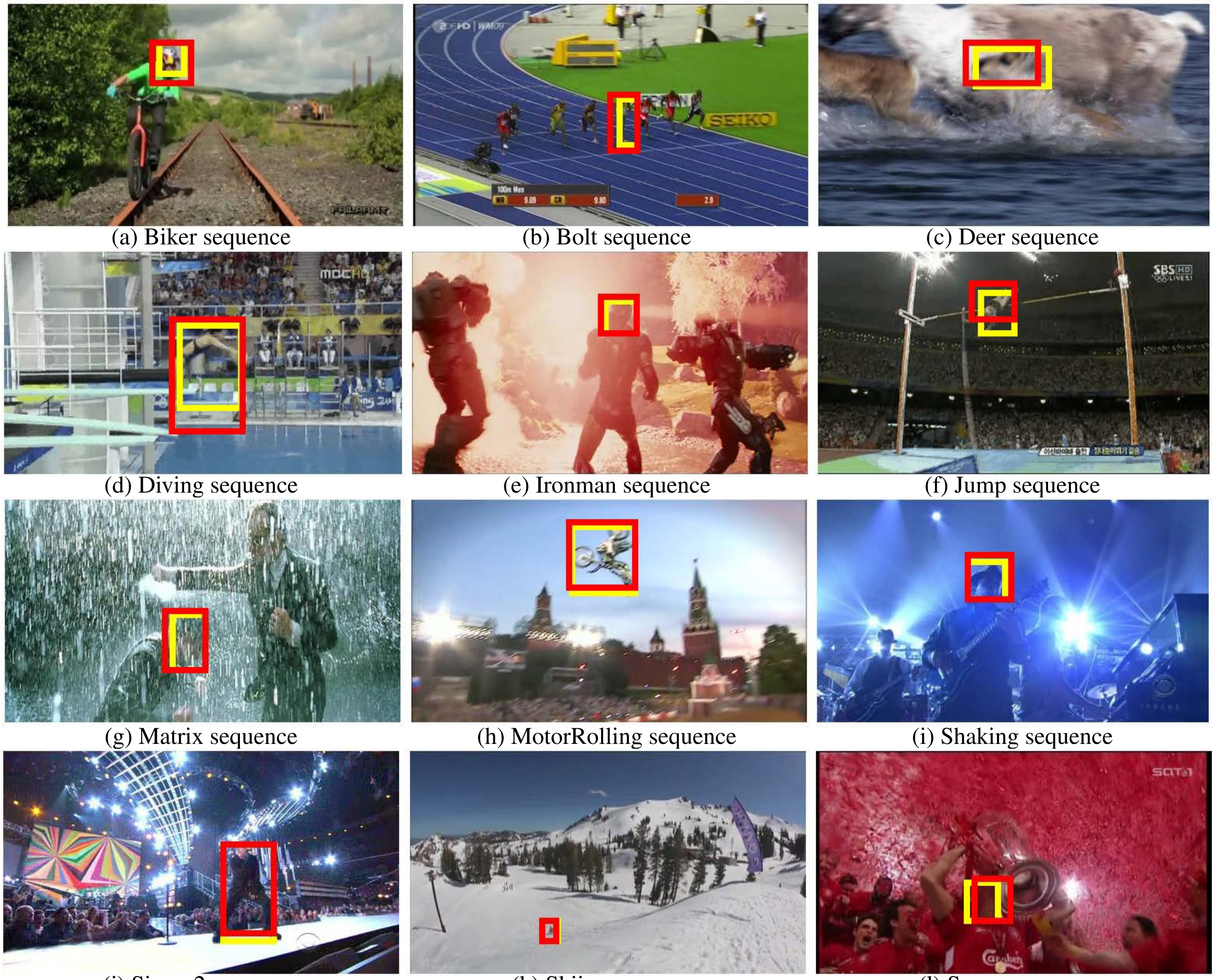
4) HYPERPARAMETERS

For the experiments, we used $N = 100$ Monte Carlo samples in (3), $M = 4$ moment statistics (*i.e.*, mean, variance, skewness, and kurtosis) in (4), $C = 10$ candidate configurations in (5), $\Sigma = \{0.1, 0.1, 0.001\}$ in (5), and $T = 90$ frames in (11).

B. ANALYSIS OF THE PROPOSED METHOD

To examine the effectiveness of each proposed technique in Table 1, we compared the proposed SWT with its extensions, PWT and OCWT. As shown in the table, describing multiple appearances of the target using Wasserstein distributions is helpful for accurate visual tracking, where our simple SWT-based visual tracker outperforms state-of-the-art visual trackers including GlobalTrack in terms of normalized precision (as shown in Table 6).

We also examined the robustness of the proposed method against hyperparameter settings. Table 2 shows that the proposed OCWT is not sensitive to different settings for the number of Monte Carlo samples. Although the OCWT exhibited more accurate results with more samples at the cost of computational time, it still shows accurate visual tracking performance even with 50 samples. Table 3 includes the visual tracking results of the proposed OCWT according to the different number of candidate configurations (C in (5)). If we consider a large number of candidate regions for the target, we have more chances of getting trapped in local minima; thus, visual tracking accuracy decreased when we used 20 candidate regions. In contrast, if we consider a very small number of candidate regions for the target, the visual tracking accuracy can decrease because search areas are not sufficient to find the target. However, in any case, our tracker is not sensitive to the number of candidate configurations.



(a) Biker sequence (b) Bolt sequence (c) Deer sequence

(d) Diving sequence (e) Ironman sequence (f) Jump sequence

(g) Matrix sequence (h) MotorRolling sequence (i) Shaking sequence

(j) Singer2 sequence (k) Skiing sequence (l) Soccer sequence

FIGURE 4. Qualitative evaluation using the OTB dataset. The yellow and red boxes denote the tracking results of ground truths and the proposed visual tracker, respectively.

Table 4 lists the visual tracking results of the proposed OCWT according to the different numbers of moment statistics (M in (4)). As shown in the table, using a single moment statistic to describe the target appearance was not sufficient to accurately track the target. If we use more than four moment statistics, the visual tracking performance converges, where our visual tracker can successfully track the target. Table 5 shows that the proposed OCWT is not sensitive to different settings with respect to the number of frames (T in (11)). Although we could obtain more accurate orthogonal vectors with a large number of frames, the performance improvement was not significant. Even though the orthogonal vectors are not accurate, using them is crucial for robust visual tracking. It should be noted that the proposed OCWT with orthogonal vectors considerably outperforms the PWT without orthogonal vectors.

C. COMPARISONS ON THE OTB DATASET

Our method was quantitatively compared with non-deep-learning visual trackers. As shown in Figure 2, the proposed method considerably surpassed existing non-deep-learning visual trackers in all evaluation metrics (*i.e.*, precision plot, success plot, and AUC). While the second-best methods are Struck and SCM for the precision and success plots, respectively, the proposed method outperformed these methods by a large margin. Empirically, we argue that accurate visual tracking results of our method are induced by precisely measuring the discrepancy between two distributions of estimated and ground-truth appearances via advanced Wasserstein-based techniques. Our method was also compared with recent deep-learning visual trackers, as shown in Figure 3. The method exhibited state-of-the-art performance in all evaluation metrics, although our method also adopted no complex deep neural network architecture. In contrast, SiamDW showed the second-best performance in terms of the precision plot, even though it employed a deeper and wider neural network architecture for visual tracking. Thus, this quantitative comparison verified the effectiveness of our Wasserstein distributional tracking, in which the discrepancy between the two appearance distributions is efficiently minimized.

TABLE 6. Quantitative comparison using the LaSOT dataset. The best results are written in boldface.

	GlobalTrack	ATOM	SiamRPN++	DASiam	SPLT	StructSiam	CFNet	ECO	OCWT
AUC	0.521	0.518	0.496	0.448	0.426	0.335	0.275	0.324	**0.523**
Precision	0.529	0.506	0.491	0.427	0.396	0.333	0.259	0.301	**0.535**
Normalized precision	0.599	0.576	0.569	-	0.494	0.418	0.312	0.338	**0.607**

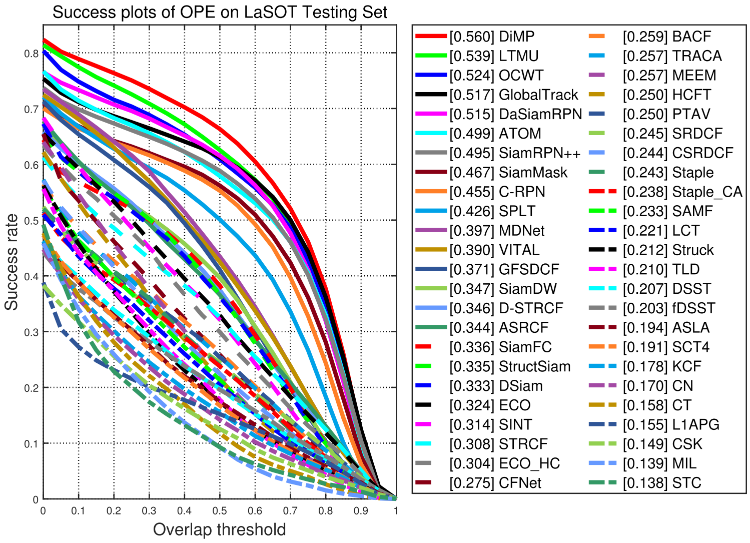


FIGURE 5. Success plot of visual trackers using the LaSOT dataset.

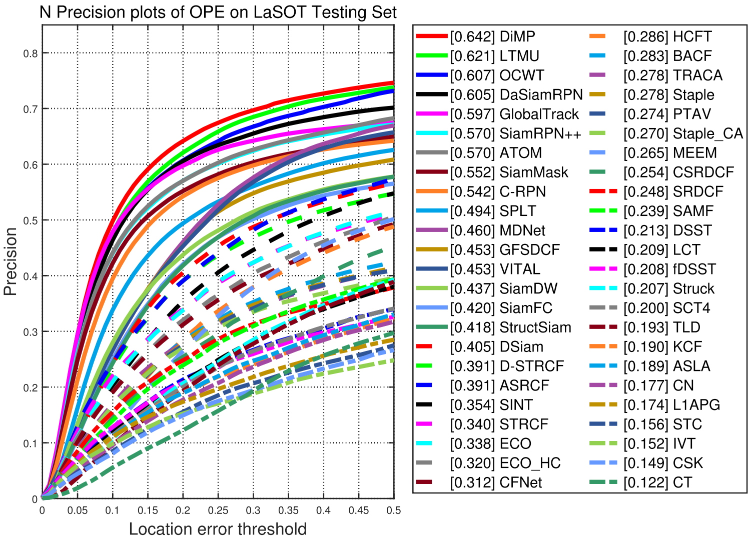


FIGURE 6. Normalized precision plot of visual trackers using the LaSOT dataset.

It is noteworthy that we present a novel appearance model for visual tracking based on the Wasserstein distribution; thus the proposed technique can be plugged into existing visual trackers to improve their visual tracking accuracy.

Figure 4 shows the qualitative visual tracking results of our method for the OTB dataset. The test video sequences contain fast motions (*e.g.*, (a) Biker, (b) Bolt, and (c) Deer sequences), nonrigid deformation (*e.g.*, (d) Diving, (e) Ironman, and (f) Jump sequences), background clutter (*e.g.*, (g) Matrix, (h) MotorRolling, and (i) Shaking sequences), occlusions (*e.g.*, (g) Matrix and (l) Soccer sequences), illumination changes (*e.g.*, (e) Ironman, (g) Matrix, (i) Shaking, and (j) Singer2 sequences), and small objects (*e.g.*, (f) Jump and (k) Skiing sequences). Although these sequences are very challenging, our method accurately tracked the targets. This accurate visual tracking performance steps from the modeling of multiple appearances using the Wasserstein distributions.

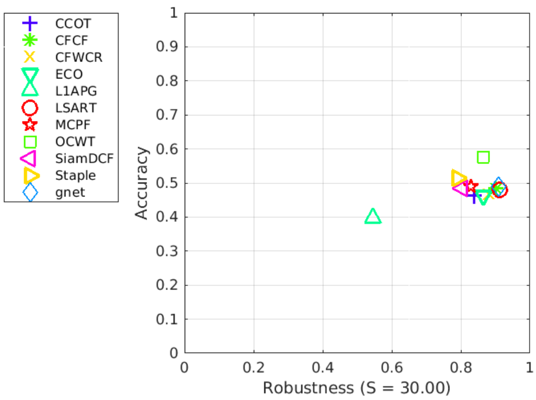


FIGURE 7. Quantitative comparison of visual trackers using the VOT dataset.

TABLE 7. Comparisons of speed in terms of frames per second (FPS). The best results are written in boldface.

	DaSiamRPN	TRACA	CACF	CFNet	SiamFC	OCWT
FPS	97	65	33	73	86	79

D. COMPARISONS ON THE LaSOT DATASET

Table 6 shows quantitative comparisons between the proposed OCWT and recent state-of-the-art visual trackers using the LaSOT dataset. As shown in the table, our method produces accurate visual tracking results and outperforms other visual trackers, where GlobalTrack shows the second-best visual tracking performance. However, GlobalTrack adopted a complex backbone network (ResNet) to extract representative features, while the proposed method used a small backbone network (VGG) to exhibit state-of-the-art performance with small computational costs. These experimental results demonstrate that the advantage of using Wasserstein distributions for the target appearances makes the proposed visual tracker robust to several variations in the target appearances, which can be caused by illumination changes, deformation, and background clutters.

E. COMPARISONS ON THE VOT DATASET

Figures 5 and 6 show the success and normalized precision plots of visual trackers using the LaSOT dataset, respectively. As shown in figures, the proposed visual tracker, OCWT, is comparable with recent state-of-the-art visual trackers such as DiMP and LTMU, while our method considerably outperforms state-of-the-art correlation filter-based trackers (*e.g.*, GFSDCF [55], ASRCF [56], STRCF [57], and BACF [58]).

Figure 7 demonstrates the effectiveness of the proposed method in the VOT dataset. The proposed visual tracker, OCWT, is the state-of-the-art visual tracker in terms of accuracy, while its robustness is also competitive to other methods. LSART exhibits the best performance in terms of robustness, but it inaccurately tracks target objects compared with the proposed method.

F. COMPARISONS OF SPEED

Table 7 reports speed in terms of FPS. Correlation filter-based visual trackers are fast, because mathematical operations are computationally efficient. The proposed method can also compute 79 frames per second, which is relatively faster than other non-correlation filter-based visual trackers. This indicates that the proposed orthogonal coupled Wasserstein distribution is useful for improving visual tracking accuracy with low computational costs.

VII. CONCLUSION

In this study, we propose a novel Wasserstein distributional tracking method that can balance approximation with accuracy in terms of Monte Carlo estimation. To achieve this goal, we present three different visual tracking systems: sliced Wasserstein-based, projected Wasserstein-based, and orthogonal coupled Wasserstein-based. Sliced Wasserstein-based visual trackers can find accurate target configurations using the optimal transport plan, which minimizes the discrepancy between appearance distributions described by the estimated and ground truth configurations. Because this plan involves a finite number of probability distributions, the computation costs can be considerably reduced. Projected Wasserstein-based and orthogonal coupled Wasserstein-based visual trackers further enhance the accuracy of visual trackers using bijective mapping functions and orthogonal Monte Carlo, respectively. Experimental results demonstrate that our approach can balance computational efficiency with accuracy and the proposed visual trackers outperform other state-of-the-art visual trackers on benchmark visual tracking datasets.

REFERENCES

[1] S. Kolouri, Y. Zou, and G. K. Rohde, “Sliced Wasserstein kernels for probability distributions,” in *Proc. CVPR*, 2016, pp. 5258–5267.

[2] Y. Han, X. Liu, Z. Sheng, Y. Ren, X. Han, J. You, R. Liu, and Z. Luo, “Wasserstein loss based deep object detection,” in *Proc. CVPRW*, 2020, pp. 4299–4305.

[3] Y. Zeng, X. Fu, L. Gao, J. Zhu, H. Li, and Y. Li, “Robust multivehicle tracking with Wasserstein association metric in surveillance videos,” *IEEE Access*, vol. 8, pp. 47863–47876, 2020.

[4] D. W. Shu, S. W. Park, and J. Kwon, “3D point cloud generative adversarial network based on tree structured graph convolutions,” in *Proc. ICCV*, 2019, pp. 3859–3868.

[5] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher, “Wasserstein propagation for semi-supervised learning,” in *Proc. ICML*, 2014, pp. 306–314.

[6] S. W. Park and J. Kwon, “SphereGAN: Sphere generative adversarial network based on geometric moment matching and its applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1566–1580, Mar. 2020.

[7] V. K. Verma, D. Brahma, and P. Rai, “Meta-learning for generalized zero-shot learning,” in *Proc. AAAI*, 2020, pp. 6062–6069.

[8] A. M. Metelli, A. Likmeta, and M. Restelli, “Propagating uncertainty in reinforcement learning via Wasserstein barycenters,” in *Proc. NeurIPS*, 2019, pp. 4333–4345.

[9] J. Xu, L. Luo, C. Deng, and H. Huang, “Multi-level metric learning via smoothed Wasserstein distance,” in *Proc. IJCAI*, 2018, pp. 2919–2925.

[10] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, “Generalized sliced Wasserstein distances,” in *Proc. NeurIPS*, 2019, pp. 261–272.

[11] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Proc. NIPS*, 2013, pp. 2292–2300.

[12] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, “Stochastic optimization for large-scale optimal transport,” in *Proc. NIPS*, 2016, pp. 3440–3448.

[13] M. Rowland, J. Hron, Y. Tang, K. Choromanski, T. Sarlos, and A. Weller, “Orthogonal estimation of Wasserstein distances,” in *Proc. Mach. Learn. Res.*, 2019, pp. 186–195.

[14] G. Yao and A. Dani, “Visual tracking using sparse coding and earth mover’s distance,” 2018, *arXiv:1804.02470*.

[15] D. Danu, T. Kirubarajan, and T. Lang, "Wasserstein distance for the fusion of multisensor multitarget particle filter clouds," in *Proc. ICIF*, 2009, pp. 25–32.

[16] F. S. Danis and A. T. Cemgil, "Model-based localization and tracking using Bluetooth low-energy beacons," *Sensors*, vol. 17, p. 2484, Nov. 2017.

[17] L. Xiao, H. Wang, and Z. Hu, "Visual tracking via adaptive random projection based on sub-regions," *IEEE Access*, vol. 6, pp. 41955–41965, 2018.

[18] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. CVPR*, 2012, pp. 2042–2049.

[19] S. Zhang, H. Zhou, F. Jiang, and X. Li, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1749–1760, Nov. 2015.

[20] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van De Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. CVPR*, 2014, pp. 1090–1097.

[21] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. CVPR*, 2018, pp. 4282–4291.

[22] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. CVPR*, 2017, pp. 2805–2813.

[23] Z. Zhang and H. Peng, "Deeper and wider Siamese networks for real-time visual tracking," in *Proc. CVPR*, Jun. 2019, pp. 4591–4600.

[24] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

[25] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, "GradNet: Gradient-guided network for visual object tracking," in *Proc. ICCV*, 2019, pp. 6162–6171.

[26] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware Siamese networks for visual object tracking," in *Proc. ECCV*, 2018, pp. 101–117.

[27] J. Choi, J. Kwon, and K. M. Lee, "Deep meta learning for real-time target-aware visual tracking," in *Proc. ICCV*, 2019, pp. 911–920.

[28] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. ICCV*, 2019, pp. 6182–6191.

[29] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic Siamese network for visual object tracking," in *Proc. ICCV*, 2017, pp. 1763–1771.

[30] X. Li, Q. Liu, N. Fan, Z. Zhou, Z. He, and X. Jing, "Dual-regression model for visual tracking," *Neural Netw.*, vol. 132, pp. 364–374, Oct. 2020.

[31] N. Fan, Q. Liu, X. Li, Z. Zhou, and Z. He, "Interactive convolutional learning for visual tracking," *Knowl. Based Syst.*, vol. 214, Oct. 2021, Art. no. 106724.

[32] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning deep multi-level similarity for thermal infrared object tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2114–2126, 2021.

[33] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, W. Liu, and Y. Liang, "Multi-task driven feature models for thermal infrared tracking," in *Proc. AAAI*, 2020, pp. 11604–11611.

[34] Q. Liu, X. Lu, Z. He, C. Zhang, and W. Chen, "Deep convolutional neural networks for thermal infrared object tracking," *Knowl. Based Syst.*, vol. 134, pp. 189–198, Jun. 2017.

[35] M. P. Muresan and S. Nedevschi, "Multi-object tracking of 3D cuboids using aggregated features," in *Proc. ICCP*, 2019, pp. 11–18.

[36] H. Karunasekera, H. Wang, and H. Zhang, "Multiple object tracking with attention to appearance, structure, motion and size," *IEEE Access*, vol. 7, pp. 104423–104434, 2019.

[37] G. Brasó and L. Leal-Taixé, "Learning a neural solver for multiple object tracking," in *Proc. CVPR*, 2020, pp. 6247–6257.

[38] M. Staib, J. M. Claici, S. amd Solomon, and S. Jegelka, "Parallel streaming Wasserstein barycenters," in *Proc. NIPS*, 2017, pp. 2292–2300.

[39] J. Rabin, G. Peyr, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," *Lect. Notes Comput. Sci.*, vol. 6667, pp. 435–446, Mar. 2011.

[40] H. V. Hasselt, "Double Q-learning," in *Proc. NIPS*, 2010, pp. 2613–2621.

[41] F. Pitié, A. C. Kokaram, and R. Dahyot, "Automated colour grading using colour distribution transfer," *CVIU*, vol. 107, nos. 1–2, pp. 123–137, 2007.

[42] K. Choromanski, M. Rowland, W. Chen, and A. Weller, "Unifying orthogonal Monte Carlo methods," in *Proc. ICML*, 2019, pp. 1203–1212.

[43] W. Givens, "Computation of plane unitary rotations transforming a general matrix to triangular form," *J. Soc. Ind. Appl. Math.*, vol. 6, no. 1, pp. 26–50, 1958.

[44] N. S. Pillai and A. Smith, "KAC's walk on N-sphere mixes in n log n steps," *Ann. Appl. Probab.*, vol. 27, no. 1, pp. 631–650, 2017.

[45] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. CVPR*, 2017, pp. 6638–6646.

[46] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. CVPR*, 2019, pp. 1369–1378.

[47] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese instance search for tracking," in *Proc. CVPR*, 2016, pp. 1420–1429.

[48] M. Danelljan, A. Robinson, F. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. ECCV*, 2016, pp. 472–488.

[49] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, "Deep attentive tracking via reciprocative learning," in *Proc. NIPS*, 2018, pp. 1–8.

[50] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. CVPR*, 2013, pp. 2411–2418.

[51] Y. Zhang, L. Wang, J. Qi, D. Wang, M. Feng, and H. Lu, "Structured Siamese network for real-time visual tracking," in *Proc. ECCV*, 2018, pp. 351–366.

[52] L. Huang, X. Zhao, and K. Huang, "GlobalTrack: A simple and strong baseline for long-term tracking," in *Proc. AAAI*, 2019, pp. 351–366.

[53] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proc. CVPR*, 2019, pp. 4660–4669.

[54] B. Yan, H. Zhao, D. Wang, H. Lu, and X. Yang, "'Skimming-perusal' tracking: A framework for real-time and robust long-term tracking," in *Proc. ICCV*, 2019, pp. 2385–2393.

[55] T. Xu, Z. Feng, X. Wu, and J. Kittler, "Joint group feature selection and discriminative filter learning for robust visual object tracking," in *Proc. ICCV*, 2019, pp. 7950–7960.

[56] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. CVPR*, 2019, pp. 4670–4679.

[57] F. Li, C. Tian, W. Zuo, L. Zhang, and M. H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. CVPR*, 2018, pp. 4904–4913.

[58] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. CVPR*, 2017, pp. 1135–1143.

[59] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "LaSOT: A high-quality benchmark for large-scale single object tracking," 2018, *arXiv:2009.03465*.

[60] E. Gundogdu and A. A. Alatan, "Good features to correlate for visual tracking," 2017, *arXiv:1704.06326*.

[61] C. Sun, D. Wang, H. Lu, and M.-H. Yang, "Learning spatial-aware regressions for visual tracking," in *Proc. CVPR*, 2018, pp. 8962–8970.

[62] Z. He, Y. Fan, J. Zhuang, Y. Dong, and H. Bai, "Correlation filters with weighted convolution responses," in *Proc. ICCVW*, 2017, pp. 1992–2000.

YOUJIN KIM received the B.S. degree in integrative engineering from Chung-Ang University, Seoul, South Korea, in 2021, where she is currently pursuing the M.S. degree in artificial intelligence. Her research interests include generative model, graph model, and deep neural networks.

JUNSEOK KWON (Member, IEEE) received the B.Sc. degree, the M.Sc. degree in the topic of object tracking (supervised by Prof. Kyoung Mu Lee), and the Ph.D. degree in electrical engineering and computer science from Seoul National University, South Korea, in 2006, 2008, and 2013, respectively. He was a Postdoctoral Researcher under Prof. Luc Van Gool with the Computer Vision Laboratory, ETH Zurich, from 2013 to 2014. He is currently an Associate Professor with the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea. He is working in the field of object tracking to capture the dynamics of cities. His research interests include visual tracking, visual surveillance, and Monte Carlo Sampling method and its variants.