

Dynamic Orthogonal Projection Constrained Discriminative Tracking

Bin Yu^{ID}, Ming Tang^{ID}, Member, IEEE, Guibo Zhu, Jinqiao Wang^{ID}, and Hanqing Lu, Senior Member, IEEE

Abstract—Due to the end-to-end feature learning with convolutional neural networks (CNNs), modern discriminative trackers improve the state of the art significantly. To achieve a strong discrimination, the learned features are usually high-dimensional, resulting in a massive number of parameters contained in the discriminative model and the increase of risk of over-fitting in the online tracking. In this letter, we try to alleviate the risk of over-fitting by means of the adaptive dimensionality reduction (DR) through CNNs. Specifically, an orthogonality constrained ridge regression model is proposed to reduce the dimensionality of features, and a dynamic sub-network (DOPNet) is designed to learn to perform DR. After trained with an orthogonality loss and a regression one, DOPNet generates a set of orthogonal bases (i.e., weights in FC layers) dynamically to reduce the feature dimensionality for a discriminative model in the online tracking. Based on the novel discriminative model and DOPNet, an effective and efficient tracker, DOPTracker, is developed. DOPTracker achieves the state-of-the-art results on four benchmarks, OTB-2015, VOT-2018, NFS, and GOT-10 k while running at 30 FPS.

Index Terms—Visual tracking, dimensionality reduction, discriminative model.

I. INTRODUCTION

GENERIC visual tracking is a long-standing topic in the field of computer vision and has attracted increasing attention over the last decades. Despite significant progress in recent years [2], [4], [10], [21], [30]–[34], visual tracking remains challenging due to numerous factors such as very limited online training samples, large appearance variation, and heavy background clutters.

Manuscript received December 10, 2021; revised February 2, 2022; accepted February 7, 2022. Date of publication February 14, 2022; date of current version March 3, 2022. This work was supported in part by the Key-Areas Research and Development Program of Guangdong Province under Grant 2020B010165001; in part by the National Natural Science Foundation of China under Grants 61772527, 61976210, 62076235, and 62002356; and in part by the Open Research Projects of Zhejiang Lab under Grant 2021KH0AB07. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Le Lu.

Bin Yu, Ming Tang, Guibo Zhu, and Hanqing Lu are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100049, China (e-mail: bin.yu@nlpr.ia.ac.cn; tangm@nlpr.ia.ac.cn; gbjzhu@nlpr.ia.ac.cn; luhq@nlpr.ia.ac.cn).

Jinqiao Wang is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100049, China, and also with the ObjectEye Inc., Beijing 100049, China (e-mail: jqwang@nlpr.ia.ac.cn).

This letter has supplementary downloadable material available at <https://doi.org/10.1109/LSP.2022.3150984>, provided by the authors.

Digital Object Identifier 10.1109/LSP.2022.3150984

Due to the end-to-end feature learning with convolutional neural networks (CNNs), discriminative trackers [4], [25], [33] have improved the state of the art significantly. By integrating the solver of the discriminative model, i.e., ridge regression, into the offline training, these trackers learn feature embeddings that are tightly coupled with their tracking schemes. The dimensionality of the learned feature embeddings, however, has to be high in these trackers to achieve the strong discrimination ability. On the other hand, the employment of high-dimensional features will increase the amount of trainable parameters of the discriminative models, even beyond the number of training samples. It is well-known [15], [17], [22] that over-parameterized models will increase the risk of over-fitting and degenerate generalization ability.

In general, the stronger discrimination ability of features comes from their higher dimensionality, and the stronger generalization ability comes from the lower dimensionality of model in the case of scarcity of online training samples. Further, the lower dimensionality of model means the lower dimensionality of features in our case. Thus, a reasonable way to address the above dilemma is to resort to the dimensionality reduction (DR) to find out a proper dimensionality of features, achieving both strong discrimination and generalization abilities. However, existing DR methods mainly focus on efficiency, e.g., principal component analysis (PCA) like in [8], [9] where they can not ensure the projection for DR is optimal for their discriminative models. ECO [7] learned discriminative filters and the projection matrix jointly to achieve good discrimination and generalization abilities. However, since its projection matrix is not orthogonal, ECO is not able to construct the feature of lower dimensionality to improve its performance more greatly.

In order to obtain a discriminative model which exploits low-dimensional features and is of both strong discrimination and generalization abilities, in this paper, we propose a novel ridge regression model subject to an orthogonal projection constraint, and prove that the ridge regression model can be analytically solved, given any set of orthogonal bases. The key of our idea is both to learn features of high dimensionality to obtain a strong discrimination ability in offline training stage, and to obtain strong generalization ability of the ridge regression model through the dimensionality reduction in the online tracking stage. Moreover, the dimensionality reduction with orthogonal matrices reserves more discrimination ability than with non-orthogonal ones of the same sizes.

Then, we design a lightweight sub-network, DOPNet, to learn to construct proper sets of orthogonal bases. To the best of our

knowledge, this work is the first to introduce the orthogonality of the weights (i. e., the projection matrix) into the widely used FC layers to improve the generalization ability of networks. It is worthy expecting that the generalization will also be improved with such or similar projections in FC layers in other visual tasks.

Finally, based on the trained DOPNet and the novel discriminative model, an effective and efficient discriminative tracker, DOPTTracker, is proposed. In DOPTTracker, similar to DiMP [4] and DCFST [33], the learning of feature embedding is tightly coupled with the discriminative model, resulting in the model of strong discrimination ability. On the other hand, in comparison to DiMP and DCFST, the number of trainable parameters of the discriminative model drops greatly from 1024 to 370 in DOPTTracker, leading to stronger generalization ability.

Extensive experiments on four popular benchmarks, OTB-2015 [28], VOT-2018 [20], NfS [19], and GOT-10k [14], show that our DOPTTracker achieves the state-of-the-art performance on all datasets, and outperforms the baseline DCFST on both accuracy and efficiency.

In summary, our contributions are in three folds.

1. We propose a novel ridge regression model subject to an orthogonal projection constraint.
2. We design a lightweight sub-network to learn to construct sets of orthogonal bases.
3. We develop an effective and efficient discriminative tracker DOPTTracker which achieves the state-of-the-art performance.

II. PROPOSED METHODOLOGY

A. Ridge Regression With Orthogonal Projection Constraint

In general, the optimization problem of standard ridge regression is formulated as follows,

$$\min_{\mathbf{w}} \|\mathbf{w}^\top \mathbf{X} - \mathbf{y}^\top\|_2^2 + \lambda \|\mathbf{w}\|_2^2, \quad (1)$$

where $\mathbf{w} \in \mathbb{R}^D$, $\mathbf{X} \in \mathbb{R}^{D \times N}$ consists of N D -dimensional training samples, $\mathbf{y} \in \mathbb{R}^{N \times 1}$ is the labels, and $\lambda \geq 0$ is the regularization parameter.

It is clear that the learned regression model will be of overfitting and degenerate the generalization ability if $D > N$. In order to find out a subspace of lower dimensionality than the original one to improve the generalization ability, we propose the ridge regression with orthogonal projection constraint as follows.

Let $\mathbf{w} = \mathbf{P}\alpha$, where projection matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M] \in \mathbb{R}^{D \times M}$, and $\alpha \in \mathbb{R}^{M \times 1}$, $M \ll N$. Then, our novel optimization problem of ridge regression is expressed as

$$\begin{aligned} \min_{\alpha, \mathbf{P}} \mathcal{F}(\alpha, \mathbf{P}) &\equiv \|\alpha^\top \mathbf{P}^\top \mathbf{X} - \mathbf{y}^\top\|_2^2 + \lambda \|\mathbf{P}\alpha\|_2^2, \\ \text{s.t. } &\|\mathbf{P}^\top \mathbf{P} - \mathbf{I}_M\|_F^2 = 0, \end{aligned} \quad (2)$$

where $\mathbf{I}_M \in \mathbb{R}^{M \times M}$ is an identity matrix. The dimensionality of training samples is reduced from D to M with \mathbf{P} if $M < D$. \mathbf{P} is constrained to be column orthogonal. Then, the sufficient

discriminative ability can be remained and the risk of overfitting can be reduced, by means of a proper dimensionality of optimal α^* .

In order to solve for (α, \mathbf{P}) in Problem (2), let the Lagrangian of Problem (2) be

$$\begin{aligned} \mathcal{L}(\alpha, \mathbf{P}, \xi) &= \|\alpha^\top \mathbf{P}^\top \mathbf{X} - \mathbf{y}^\top\|_2^2 + \lambda \|\mathbf{P}\alpha\|_2^2 \\ &\quad - \xi \|\mathbf{P}^\top \mathbf{P} - \mathbf{I}_M\|_F^2. \end{aligned} \quad (3)$$

Then, the Karush-Kuhn-Tucker (KKT) conditions [5] of problem (2) are

$$\begin{cases} \nabla_{\alpha, \mathbf{P}} \mathcal{L}(\alpha, \mathbf{P}, \xi) = \mathbf{0}, \\ \|\mathbf{P}^\top \mathbf{P} - \mathbf{I}_M\|_F^2 = 0, \end{cases} \quad (4)$$

Suppose λ is large enough. It can be derived from (4) that

$$\mathbf{P}\alpha = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_D)^{-1} \mathbf{X}\mathbf{y}. \quad (5)$$

It is easy to see that the minimum of Problem (2) can be achieved by replacing $\mathbf{P}\alpha$ with (5) because $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_D)^{-1} \mathbf{X}\mathbf{y}$ is the optimal solution of Problem (1), $\mathbf{w} = \mathbf{P}\alpha$, and $\mathcal{F}(\alpha, \mathbf{P})$ is obtained by substituting \mathbf{w} with $\mathbf{P}\alpha$ in (1). Therefore, the KKT conditions are *necessary and sufficient* for Problem (2).

Given any column orthogonal matrix \mathbf{P}^* , according to (4), we have

$$\alpha^* = (\mathbf{P}^{*\top} \mathbf{X}\mathbf{X}^\top \mathbf{P}^* + \lambda \mathbf{I}_M)^{-1} \mathbf{P}^{*\top} \mathbf{X}\mathbf{y} \quad (6)$$

for a large enough λ . (α^*, \mathbf{P}^*) satisfies (4), being an optimal solution of Problem (2). Consequently, an optimal solution to Problem (2) can be obtained through using (6) with any column orthogonal matrix.

Nevertheless, different (α^*, \mathbf{P}^*) does not mean the same *generalization ability* of the model. Therefore, we have to find out a proper \mathbf{P}^* and its corresponding α^* to achieve strong generalization ability in the online tracking task. The approach to acquiring such (\mathbf{P}^*, α^*) will be stated in the next subsection.

B. Dynamic Sub-Network for Dimensionality Reduction

We propose to generate the proper \mathbf{P}^* with the dynamic sub-network DOPNet, i. e., $\mathbf{P}^* = \varphi_\rho(\mathbf{U})$, where $\varphi_\rho(\cdot)$ is the DOPNet with \mathbf{U} as its input, and ρ contains the learnable parameters. The target features \mathbf{U} and the features of *test samples* \mathbf{Z} are used to train $\varphi_\rho(\cdot)$ in an end-to-end way by minimizing the regression loss subject to the orthogonality constraint, obtaining the strong generalization ability of the model. The optimization problem of DOPNet is formulated as follows,

$$\begin{aligned} \min_{\rho} & \|\hat{\mathbf{y}}^\top - \mathbf{y}^\top\|_2^2, \\ \text{s.t. } & \|\varphi_\rho^\top(\mathbf{U})\varphi_\rho(\mathbf{U}) - \mathbf{I}_M\|_F^2 = 0, \end{aligned} \quad (7)$$

where

$$\hat{\mathbf{y}} = \mathbf{Z}^\top \varphi_\rho(\mathbf{U})\alpha^*, \quad (8)$$

$\mathbf{Z} \in \mathbb{R}^{D \times N}$ contains the D -dimensional features of N test samples, and

$$\alpha^* = (\varphi_\rho^\top(\mathbf{U})\mathbf{X}\mathbf{X}^\top \varphi_\rho(\mathbf{U}) + \lambda \mathbf{I}_M)^{-1} \varphi_\rho^\top(\mathbf{U})\mathbf{X}\mathbf{y}. \quad (9)$$

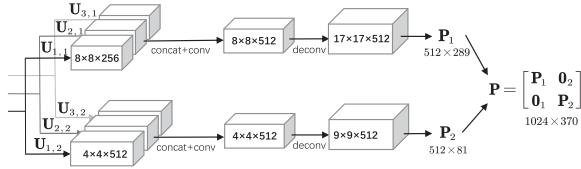


Fig. 1. The architecture of DOPNet. All convs are 3×3 , deconvs are 3×3 with stride 2, and we use BatchNorm [16] and PReLU [12] after each conv or deconv layer.

After training $\varphi_\rho(\cdot)$ on large-scale tracking datasets, the optimal solution (α^*, \mathbf{P}^*) that generalizes best can be efficiently obtained by using $\varphi_\rho(\cdot)$ and (9).

It is worth noticing that although theoretically there are $DM + M$ parameters to optimize in Problem (2), \mathbf{P}^* can be directly produced by the trained DOPNet without online optimization. Therefore, only M parameters of α need to learn in the online tracking.

Network Architecture: We use the same feature extraction network as that in DCFST. To obtain the target features in each frame, we extract two feature maps, $\mathbf{U}_{i,1}$ and $\mathbf{U}_{i,2}$, of the target ROI with the fixed size of $8 \times 8 \times 256$ and $4 \times 4 \times 512$, respectively, after the PrPool layers [18] of the feature extraction network, where i is the frame number. Three different frames, named base frames, are randomly sampled from a sequence. Then, target features from base frames are used to train DOPNet, that is $\mathbf{U} = \{(\mathbf{U}_{i,1}, \mathbf{U}_{i,2})\}_{i=1}^3$. Note that we use three base frames for a good balance between accuracy and efficiency. To obtain enough bases efficiently and effectively, we propose to predict one base at each spatial location of features. Fig. 1 shows the architecture of DOPNet where the parameters of two branches are not shared. Concretely, first, the three feature maps in each branch are concatenated along the channel dimension and then fed into the following convolutional layer to output the fused feature map whose number of channels is the same as the dimensionality of samples. Second, two deconvolutional layers expand the spatial sizes of two feature maps to 17×17 and 9×9 , respectively, generating $17 \times 17 + 9 \times 9$ bases. Two groups of the bases are then reshaped to form projection matrices \mathbf{P}_1 and \mathbf{P}_2 , respectively. The final projection matrix is expressed as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0}_2 \\ \mathbf{0}_1 & \mathbf{P}_2 \end{bmatrix}_{1024 \times 370}, \quad (10)$$

where $\mathbf{0}_1 \in \mathbb{R}^{512 \times 289}$ and $\mathbf{0}_2 \in \mathbb{R}^{512 \times 81}$ are two zero matrices. By using \mathbf{P} , the number of parameters of our discriminative model is reduced significantly from 1024 to 370. More complex architectures of the projection head may have the potential to improve tracking performance further, but are not the focus of this letter.

C. Loss Function

We convert the constraint of Problem (7) to the orthogonality loss of DOPNet in the way similar to that in [1], [3], [24], [27], [29] as follows.

$$\mathcal{L}_{\text{orth}} = \|\varphi_\rho^\top(\mathbf{U})\varphi_\rho(\mathbf{U}) - \mathbf{I}_M\|_F^2. \quad (11)$$

We use this method instead of directly solving the ‘hard orthogonal constraint’ [3], [27] mainly due to the following two reasons. First, the orthogonality penalty is differentiable and requires no SVD, thus computationally cheaper [3]. Second, although theoretically the orthogonality penalty can only lead to the approximate orthogonality of projection matrices and further a suboptimal solution to Problem (2), the solution is still experimentally satisfactory in the localization performance.

Note that different from previous works [1], [3], [24], [27], [29] which aimed at learning orthogonal weight matrices that are unchanged during inference, our DOPNet is devised to generate dynamic orthogonal projection matrices as the input changes during the online tracking. Besides, we use the same regression loss \mathcal{L}_{reg} as that in DCFST [33] to replace the original one in Problem (7) to optimize DOPNet and the feature extraction network more effectively.

Finally, the total loss of DOPNet is given as follows.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \eta \mathcal{L}_{\text{orth}}, \quad (12)$$

where η is the tradeoff hyper-parameter.

D. Online Tracking With DOPTracker

Updating: For robustness, we update the projection matrix by changing the input in every frame as follows.

$$\begin{aligned} \tilde{\mathbf{U}}_T &= \mathbf{U}, & T = 1, \\ \tilde{\mathbf{U}}_T &= (1 - \gamma)\tilde{\mathbf{U}}_{T-1} + \gamma\mathbf{U}, & T > 1, \end{aligned} \quad (13)$$

where $\mathbf{U} = \{(\mathbf{U}_{i,1}, \mathbf{U}_{i,2}), (\mathbf{U}_{i,1}, \mathbf{U}_{i,2}), (\mathbf{U}_{i,1}, \mathbf{U}_{i,2})\}$, and γ is a weight parameter. Note that, different from the training stage of DOPNet, three pairs of feature maps are equal. Besides, during tracking, we update the regression model by using a moving average method as that in DCFST every 10 frames.

Detection and Refinement of Target Object: Given the solution to Problem (2) by using (6) and $\varphi_\rho(\cdot)$, and the test data matrix \mathbf{Z} in a new frame, we obtain the predicted location response $\hat{\mathbf{y}}$ by using (8). The element of $\hat{\mathbf{y}}$ which takes the maximal value is accepted as the predicted location of the target object in the new frame. Then, we refine the bounding box of the predicted target with ATOM [6], as done in [4], [33].

E. Discussion

Here, we explicitly discuss the difference between this work and the baseline DCFST. 1) The focuses are different. DCFST focuses on integrating the discriminative model solvers into the offline training to learn optimal features for the discriminative model. DOPTracker focuses on alleviating the risk of over-fitting by means of the adaptive dimensionality reduction through CNNs. 2) The discriminative models are different. DOPTracker uses ridge regression model with orthogonal projection constraint (Problem (2)) while DCFST uses standard ridge regression model (Problem (1)). 3) The networks are different. DCFST contains a feature extraction network and a model solver. DOPTracker contains an additional dynamic sub-network for dimensionality reduction. 4) The loss functions are different. DCFST only employs a regression loss to train the feature

TABLE I
ABLATION STUDIES OF DOPTRACKER ON NFS AND THE TEST SET OF GOT-10 K

Change from DOPTracker:		NfS	GOT-10k	
Changed	From → To	AUC	AUC	FPS
Loss	Baseline (DCFST)	0.641	0.638	25
	Orth+Reg→Reg	0.622	0.598	30
	Orth+Reg→L1+Reg	0.630	0.610	30
Base numbers	Orth+Reg→L2+Reg	0.633	0.617	30
	289+81 → 100+64	0.617	0.605	33
	289+81 → 256+100	0.660	0.650	30
—	289+81 → 400+100	0.650	0.645	28
	DOPTracker (ours)	0.660	0.653	30

extraction network while DOPTracker uses the orthogonality loss and the regression loss to train our networks.

III. EXPERIMENTS

A. Implementation Details

Offline Training: DOPNet and the feature extraction network are trained jointly and offline in our approach, and the gradient in DOPNet will not back-propagate to the input features. We use the training splits of the TrackingNet [23], LaSOT [11], GOT-10k [14] datasets. Our proposed networks with ResNet-50 [13] as the backbone network receive a tuple of frames, i. e., 3 base frames for extracting target features, a training frame and a test one, as the input in the offline training. The hyper-parameter η in (12) is set to 10^{-8} and the other settings are the same as those in [33]. See more details in supplementary materials.

Online Tracking: The search area is 5^2 times the target area. The weight parameter γ in (13) is 0.01. On a single TITAN X (Pascal) GPU, DOPTracker achieves a real-time speed of 30 FPS.

B. Ablation Studies

Loss Function We additionally train our networks without the orthogonality loss and the performance gets worse (see Table I). This confirms that orthogonality of the projection matrix plays an important role in achievement of strong discrimination ability when the feature dimensionality is reduced. Moreover, the results verify that the tracking performance cannot be improved if we only increase the number of parameters in the networks. As shown in Table I, when we use L1/L2 penalty for the projection matrix in the loss function instead of the orthogonality penalty, the results are both better than that when imposing no penalty for P but L1/L2 penalty can not benefit the solution to Problem (2) and thus is less effective.

Base Numbers: When the numbers of bases produced by DOPNet are reduced to 100+64, the discriminative model is prone to be of under-fitting and further leads to degenerated performance. In our experiments, base numbers around 370, e. g., 256+100, can also lead to similar results to ours (see Table I), showing the robustness of our model to the base numbers. When we try to generate much more bases by DOPNet, e. g., 400+100, the performance can not be further improved because the simple architecture in DOPNet is not able to generate more suitable

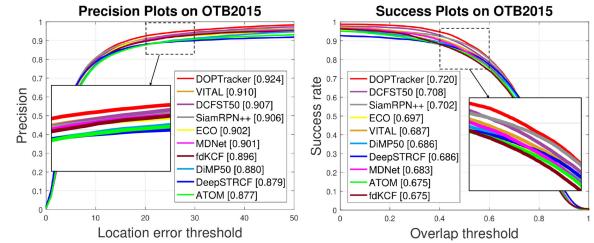


Fig. 2. The precision and success plots of the proposed DOPTracker and other state-of-the-art trackers on OTB-2015.

TABLE II
STATE-OF-THE-ART COMPARISONS ON THE VOT-2018

	SiamseRPN++ [21]	ATOM [6]	DiMP50 [4]	DCFST50 [33]	SiamRCNN [26]	DOPTracker
EAO ↑	0.414	0.401	0.440	0.452	0.408	0.471
Robustness ↓	0.234	0.204	0.153	0.136	0.220	0.122
Accuracy ↑	0.536	0.600	0.590	0.607	0.609	0.608

TABLE III
STATE-OF-THE-ART COMPARISONS ON GOT-10 K

	SiamseRPN++ [21]	ATOM [6]	DiMP50 [4]	DCFST50 [33]	SiamRCNN [26]	DOPTracker
AUC ↑	0.611	0.556	0.611	0.638	0.649	0.653

TABLE IV
STATE-OF-THE-ART COMPARISONS ON NFS

	SiamseRPN++ [21]	ATOM [6]	DiMP50 [4]	DCFST50 [33]	SiamRCNN [26]	DOPTracker
AUC ↑	0.620	0.590	0.620	0.641	0.639	0.660

bases for effective DR. Note that for simplicity we do not attempt to refine base numbers by cropping the feature maps in DOPNet. The final numbers of bases, i. e., 289+81, are totally dependent on the input sizes of feature maps and the convolutional and deconvolutional layers in DOPNet.

C. State-of-The-Art Comparisons

OTB-2015: Fig. 2 shows the precision and AUC scores of DOPTracker and other state-of-the-art trackers. DOPTracker obtains AUC and precision scores of 0.720 and 0.924, respectively, outperforming DCFST50 by 0.8% and 1.4%, respectively.

GOT-10k: In this experiment, to ensure a fair comparison, we retrain our network only with the training splits of GOT-10 k. DOPTracker obtains 0.653 in terms of AUC score, surpassing its baseline DCFST50 by 1.5% (see Table III).

VOT-2018: DOPTracker obtains the EAO score of 0.471, outperforming DCFST50 by 1.9% (see Table II). Although DOPTracker obtains a similar accuracy score to that of DCFST50, our tracker surpasses DCFST50 in terms of robustness with a gain of 1.4%, which indicates that DOPTracker has stronger generalization ability than DCFST50.

NfS: We evaluate our approach on the 30 FPS version of the dataset, containing 100 challenging videos. Table IV shows that DOPTracker obtains an AUC score of 0.660, surpassing DCFST50 by 1.9%.

IV. CONCLUSION

In this letter, we propose a novel ridge regression model of an orthogonal projection constraint and devise a simple yet powerful dynamic sub-network DOPNet to learn to perform adaptive dimensionality reduction. Our DOPTracker achieves state-of-the-art performance on four popular benchmarks.

REFERENCES

- [1] J. Amjad, Z. Lyu, and M. R. Rodrigues, "Deep learning for inverse problems: Bounds and regularizers," 2019, *arXiv:1901.11352*.
- [2] Y. Bai and M. Tang, "Object tracking via robust multitask sparse representation," *IEEE Signal Process. Lett.*, vol. 21, no. 8, pp. 909–913, Aug. 2014.
- [3] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4261–4271.
- [4] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2019, pp. 6182–6191.
- [5] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [6] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4660–4669.
- [7] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6638–6646.
- [8] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.
- [9] M. Danelljan, F.S. Khan, M. Felsberg, and J. V. de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1090–1097.
- [10] C. Deng, S. He, Y. Han, and Z. Boya, "Learning dynamic spatial-temporal regularization for UAV object tracking," *IEEE Signal Process. Lett.* vol. 28, pp. 1230–1234, 2021, doi: [10.1109/LSP.2021.3086675](https://doi.org/10.1109/LSP.2021.3086675)
- [11] H. Fan *et al.*, "LaSot: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 5374–5383.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2015, pp. 1026–1034.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
- [14] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562–1577, May 2021.
- [15] N. Huang, D. W. Hogg, and S. Villar, "Dimensionality reduction, regularization, and generalization in overparameterized regressions," 2020, *arXiv:2011.11477*.
- [16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [17] T. Isomura and T. Toyoizumi, "Dimensionality reduction to maximize prediction generalization capability," 2020, *arXiv:2003.00470*.
- [18] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 784–799.
- [19] H.K. Galoogahi, A. C. Fagg, D. H. Ramanan, and S. Lucey, "Need for Speed: A benchmark for higher frame rate object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2017, pp. 1125–1134.
- [20] M. Kristan *et al.*, "The visual object tracking VOT2018 challenge results," in *Proc. IEEE Eur. Conf. Comput. Vision*, 2018, pp. 3–53.
- [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. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4282–4291.
- [22] S. Mosci, L. Rosasco, and A. Verri, "Dimensionality reduction and generalization," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 657–664.
- [23] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. IEEE Eur. Conf. Comput. Vision*, 2018, pp. 300–317.
- [24] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.
- [25] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2805–2813.
- [26] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 6577–6587.
- [27] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal, "On orthogonality and learning recurrent networks with long term dependencies," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3570–3578.
- [28] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
- [29] D. Xie, J. Xiong, and S. Pu, "All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6176–6185.
- [30] B. Yu *et al.*, "High-performance discriminative tracking with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2021, pp. 9856–9865.
- [31] L. Zheng, Y. Chen, M. Tang, J. Wang, and H. Lu, "Siamese deformable cross-correlation network for real-time visual tracking," *Neurocomputing*, vol. 401, pp. 36–47, 2020.
- [32] L. Zheng, L. Tang, and J. Wang, "Learning robust Gaussian process regression for visual tracking," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1219–1225.
- [33] L. Zheng, M. Tang, Y. Chen, J. Wang, and H. Lu, "Learning feature embeddings for discriminant model based tracking," in *Proc. Eur. Conf. Comput. Vision*, 2020, pp. 759–775.
- [34] H. Zhu, Y. Han, Y. Wang, and G. Yuan, "Hybrid cascade filter with complementary features for visual tracking," *IEEE Signal Process. Lett.*, vol. 28, pp. 86–90, 2021.