Unsupervised Domain Adaptation for Nighttime Aerial Tracking

Junjie Ye[†], Changhong Fu[†,*], Guangze Zheng[†], Danda Pani Paudel[‡], and Guang Chen[†]
[†]Tongji University, China [‡]ETH Zürich, Switzerland
{ye.jun.jie, changhongfu, mmlp, guangchen}@tongji.edu.cn, paudel@vision.ee.ethz.ch

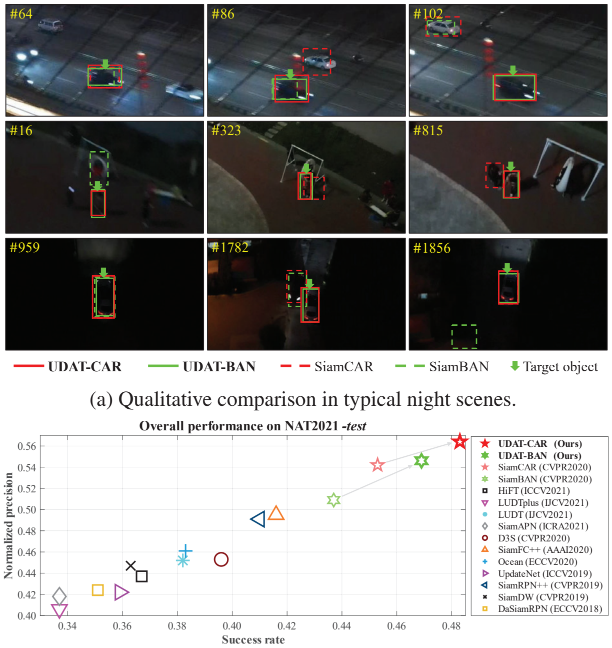
Abstract

Previous advances in object tracking mostly reported on favorable illumination circumstances while neglecting performance at nighttime, which significantly impeded the development of related aerial robot applications. This work instead develops a novel unsupervised domain adaptation framework for nighttime aerial tracking (named UDAT). Specifically, a unique object discovery approach is provided to generate training patches from raw nighttime tracking videos. To tackle the domain discrepancy, we employ a Transformer-based bridging layer post to the feature extractor to align image features from both domains. With a Transformer day/night feature discriminator, the daytime tracking model is adversarially trained to track at night. Moreover, we construct a pioneering benchmark namely NAT2021 for unsupervised domain adaptive nighttime tracking, which comprises a test set of 180 manually annotated tracking sequences and a train set of over 276k unlabelled nighttime tracking frames. Exhaustive experiments demonstrate the robustness and domain adaptability of the proposed framework in nighttime aerial tracking. The code and benchmark are available at https://github.com/vision4robotics/UDAT.

1. Introduction

Standing as one of the fundamental tasks in computer vision, object tracking has received widespread attention with multifarious aerial robot applications, *e.g.*, unmanned aerial vehicle (UAV) self-localization [49], target following [25], and aerial cinematography [2]. Driven by large-scale datasets [10,17,32] with the supervision of meticulous manual annotations, emerging deep trackers [4, 8, 14, 22] keep setting state-of-the-arts (SOTAs) in recent years.

Despite the advances, whether current benchmarks or approaches are proposed for object tracking under favorable illumination conditions. In contrast to daytime, images captured at night have low contrast, brightness, and signal-to-noise ratio (SNR). These differences cause the discrepancy in feature distribution of day/night images. Due to the cross-domain discrepancy, current SOTA trackers generalize badly to nighttime scenes [48, 50], which severely impedes the broadening of relevant aerial robot applications.



(a) Qualitative comparison in typical night scenes.

(b) Overall performance comparison on NAT2021-*test*.

Figure 1. (a) Qualitative comparison of the proposed unsupervised domain adaptive trackers (*i.e.*, UDAT-CAR and UDAT-BAN) and their baselines [8, 14]. (b) Overall performance of SOTA approaches on the constructed NAT2021-*test* benchmark. The proposed UDAT effectively adapts general trackers to nighttime aerial tracking scenes and yields favorable performance.

Regarding such domain gap and the performance drop, this work aims to address the cross-domain object tracking problem. In particular, we target adapting SOTA tracking models in daytime general conditions to nighttime aerial perspectives. One possible straightforward solution is to collect and annotate adequate target domain data for training. Nevertheless, such a non-trivial workload is expensive and time-consuming, since backbones' pre-training alone generally takes millions of high-quality images [9]. We

[*] Corresponding author

consequently consider the problem as an unsupervised domain adaptation task, where training data in the source domain is with well-annotated bounding boxes while that in the target domain has no manually annotated labels. Therefore, an **u**nsupervised **d**omain **a**daptive **t**racking framework, referred to as UDAT, is proposed for nighttime aerial tracking. To generate training patches of the target domain, we develop an object discovery strategy to explore potential objects in the unlabelled nighttime data. Besides, a bridging layer is proposed to bridge the gap of domain discrepancy for the extracted features.

Furthermore, the feature domain is distinguished by virtue of a discriminator during adversarial learning. Drawing lessons from the huge potential of the Transformer [43] in feature representation, both the bridging layer and the discriminator utilize a Transformer structure. Figure 1 exhibits some qualitative comparisons of trackers adopting UDAT and the corresponding baselines. UDAT raises baselines' nighttime aerial tracking performance substantially. Apart from methodology, we construct NAT2021, a benchmark comprising a *test* set of 180 fully annotated video sequences and a *train* set of over 276k unlabelled nighttime tracking frames, which serves as the first benchmark for unsupervised domain adaptive nighttime tracking. The main contributions of this work are fourfold:

- An unsupervised domain adaptive tracking framework, namely UDAT, is proposed for nighttime aerial tracking. To the best of our knowledge, the proposed UDAT is the first unsupervised adaptation framework for object tracking.
- A bridging layer and a day/night discriminator with Transformer structures are incorporated to align extracted features from different domains and narrow the domain gap between daytime and nighttime.
- A pioneering benchmark namely NAT2021, consisting of a fully annotated *test* set and an unlabelled *train* set, is constructed for domain adaptive nighttime tracking. An object discovery strategy is introduced for the unlabelled *train* set preprocessing.
- Extensive experiments on NAT2021-$test$ and the recent public UAVDark70 [21] benchmark verify the effectiveness and domain adaptability of the proposed UDAT in nighttime aerial tracking.

2. Related work

2.1. Object tracking

Generally, recent object tracking approaches can be categorized as the discriminative correlation filter (DCF)-based approaches [12, 16, 19, 27] and the Siamese network-based approaches [4, 8, 14, 22]. Due to the complicated online learning procedure, end-to-end training can be hardly realized on DCF-based trackers. Therefore, restricted to inferior handcrafted features or inappropriate deep feature extractors pre-trained for classification, DCF-based trackers lose their effectiveness in adverse conditions.

Benefiting from considerable training data and end-to-end learning, Siamese network-based trackers have achieved robust tracking performance. This line of approaches is pioneered by SINT [41] and SiamFC [1], which regard object tracking as a similarity learning problem and train Siamese networks with large-scale image pairs. Drawing lessons from object detection, B. Li *et al*. [23] introduce the region proposal network (RPN) [33] into the Siamese framework. SiamRPN++ [22] further adopts a deeper backbone and feature aggregation architecture to improve tracking accuracy. To alleviate increasing hyperparameters along with the introduction of RPN, the anchor-free approaches [8, 14, 47] adopt the per-pixel regression to directly predict four offsets on each pixel. Recently, Transformer [43] is incorporated into the Siamese framework [4,6,45] to model global information and further boost tracking performance.

Despite the great progress, object tracking in adverse conditions, for instance, nighttime aerial scenarios, lacks thorough study so far. In [21], a DCF framework integrated with a low-light enhancer is constructed while lacking transferability and being restricted to handcrafted features. Some studies [48, 50] design tracking-related low-light enhancers for data preprocessing in the tracking pipeline. However, such a paradigm suffers from weak collaboration with the tracking model and the cascade structure can hardly learn to narrow the domain gap at the feature level.

2.2. Domain adaptation

Towards narrowing the domain discrepancy and transferring knowledge from the source domain to the target domain, domain adaptation attracts considerable attention and is widely adopted in image classification [3, 26, 40]. Beyond classification, Y. Chen *et al*. [7] design a domain adaptive object detection framework and tackle the domain shift on both image-level and instance-level. In [18], an image transfer model is trained to perform day-to-night transformation for data augmentation before learning a detection model. Y. Sasagawa and H. Nagahara [37] propose to merge a low-light image enhancement model and an object detection model to realize nighttime object detection. For the task of semantic segmentation, C. Sakaridis *et al*. [36] formulate a curriculum framework to adapt semantic segmentation models from day to night through an intermediate twilight domain. X. Wu *et al*. [46] employ an adversarial learning manner to train a domain adaptation network for nighttime semantic segmentation. S. Saha *et al*. [35]

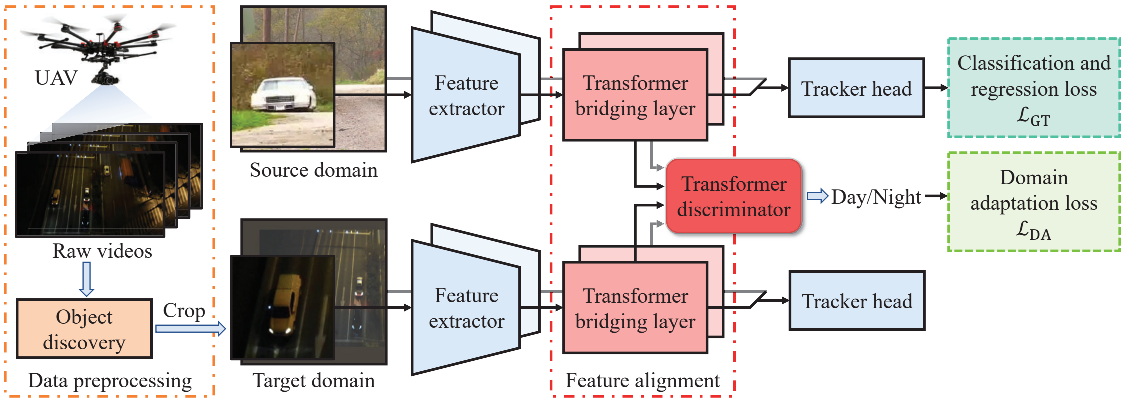

UAV
Raw videos
Object discovery
Crop
Data preprocessing
Source domain
Target domain
Feature extractor
Feature extractor
Transformer bridging layer
Transformer bridging layer
Transformer discriminator
Feature alignment
Day/Night
Tracker head
Tracker head
Classification and regression loss $\mathcal{L}_{\text{GT}}$
Domain adaptation loss $\mathcal{L}_{\text{DA}}$

Figure 2. Illustration of the proposed unsupervised domain adaptation framework for nighttime aerial tracking. The object discovery module is employed to find potential objects in raw videos for training patch generation. Features extracted from different domains are aligned via the Transformer bridging layer. A Transformer day/night discriminator is trained to distinguish features between the source domain and the target domain.

mine cross-task relationships and build a multi-task learning framework for semantic segmentation and depth estimation in the unsupervised domain adaptation setting. Despite the rapid development in other vision tasks, domain adaptation for object tracking has not been investigated yet. Therefore, an effective unsupervised domain adaptation framework for object tracking is urgently needed.

3. Proposed method

The paradigm of the proposed UDAT framework is illustrated in Fig. 2. For data preprocessing of the unlabelled target domain, we employ a saliency detection-based strategy to locate potential objects and crop paired training patches. In the training pipeline, features generated by the feature extractor are modulated by the bridging layer. In this process, adversarial learning facilitates the reduction of feature distribution discrepancy between the source and target domains. Through this simple yet effective process, trackers can achieve pleasant efficiency and robustness for night scenes comparable to daytime tracking.

3.1. Data preprocessing

Since deep trackers take training patches as input in each training step, we develop an object discovery strategy for data preprocessing on the unlabelled train set. Figure 3 illustrates the preprocessing pipeline. The object discovery strategy involves three stages, *i.e.*, low-light enhancement, salient object detection, and dynamic programming. Given the low visibility of nighttime images, original images are first lighted up by a low-light enhancer [24]. Afterward, enhanced images are fed into the video saliency detection model [52]. Candidate boxes are then obtained by building the minimum bounding rectangle of detected salient regions. To generate a box sequence that locates the same object across the timeline, motivated by [55], we adopt dynamic programming to filter noisy boxes. Assuming two boxes from the j-th frame and the k-th frame as $[x_{j,m}, y_{j,m}, w_{j,m}, h_{j,m}]$ and $[x_{k,n}, y_{k,n}, w_{k,n}, h_{k,n}]$, where m and n indicate the box indexes, and (x, y), w, h denote the top-left coordinate, width, and height of the box, respectively, the normalized distance D_{norm} is obtained as:

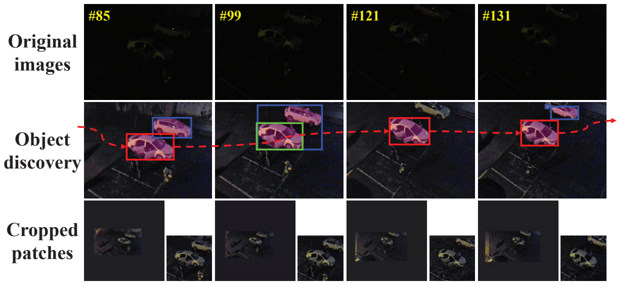


Figure 3. Illustration of object discovery, which contains low-light enhancement, salient object detection, and dynamic programming. The pink masks indicate detected salient regions, while the boxes are circumscribed rectangles of these regions. Dynamic programming selects red boxes and filters blue ones. The green box is obtained by linear interpolation between two adjacent frames. Note that the cropped patches are enhanced for visualization, original patches are utilized in the practical training process instead.

$$\begin{aligned} D_{\text{norm}} =& \left(\frac{x_{j,m} - x_{k,n}}{w_{k,n}}\right)^2 + \left(\frac{y_{j,m} - y_{k,n}}{h_{k,n}}\right)^2 \\ &+ \left(\log\left(\frac{w_{j,m}}{w_{k,n}}\right)\right)^2 + \left(\log\left(\frac{h_{j,m}}{h_{k,n}}\right)\right)^2 \quad . \end{aligned} \tag{1}$$

In dynamic programming, the normalized distance of candidate boxes between frames serves as a smooth reward, while adding a box in a frame to the box sequence means an incremental reward. For frames where none of the boxes is selected by dynamic programming, linear interpolation is adopted between the two closest frames to generate a new box. Ultimately, paired training patches are cropped from original images according to the obtained box sequence.

3.2. Network architecture

Feature extractor. Feature extraction of Siamese networks generally consists of two branches, *i.e*., the template branch and the search branch. Both branches generate feature maps from the template patch $\mathbf{Z}$ and the search patch $\mathbf{X}$, namely $\varphi(\mathbf{Z})$ and $\varphi(\mathbf{X})$, by adopting an identical backbone network φ. Generally, trackers adopt the last block or blocks of features for subsequent classification and regression, which can be represented as follows:

$$\begin{aligned}\varphi(\mathbf{X}) &= \mathrm{Concat}(\mathcal{F}_{N-i}(\mathbf{X}), ..., \mathcal{F}_{N-1}(\mathbf{X}), \mathcal{F}_{N}(\mathbf{X}))\ ,\\ \varphi(\mathbf{Z}) &= \mathrm{Concat}(\mathcal{F}_{N-i}(\mathbf{Z}), ..., \mathcal{F}_{N-1}(\mathbf{Z}), \mathcal{F}_{N}(\mathbf{Z}))\ ,\end{aligned} \tag{2}$$

where $\mathcal{F}_{*}(\cdot)$ indicates features extracted from the $*$-th block of a backbone with N blocks in total, and Concat denotes channel-wise concatenation. Since both $\varphi(\mathbf{X})$ and $\varphi(\mathbf{Z})$ will pass through the following Transformer bridging layer and discriminator, we take the instance of $\varphi(\mathbf{X})$ in the following introduction for clarity.

Transformer bridging layer. Features extracted from daytime and nighttime images are with a huge gap, such domain discrepancy leads to inferior tracking performance at night. Before feeding the obtained features to the tracker head for object localization, we propose to bridge the gap between the feature distributions through a bridging layer. In consideration of the strong modeling capability of the Transformer [43] for long-range inter-independencies, we design the bridging layer with a Transformer structure. Taking the search branch as instance, positional encodings $\mathbf{P}$ are first added to the input feature $\varphi(\mathbf{X}) \in \mathbb{R}^{N\times H\times W}$. Next, the summation is flattened to $(\mathbf{P} + \varphi(\mathbf{X})) \in \mathbb{R}^{HW\times N}$. Multi-head self-attention (MSA) is then conducted as:

$$\begin{aligned}\widehat{\varphi(\mathbf{X})}' &= \mathrm{MSA}(\mathbf{P} + \varphi(\mathbf{X})) + \mathbf{P} + \varphi(\mathbf{X}) \quad ,\\ \widehat{\varphi(\mathbf{X})} &= \mathrm{LN}(\mathrm{FFN}(\mathrm{Mod}(\mathrm{LN}(\widehat{\varphi(\mathbf{X})}'))) + \widehat{\varphi(\mathbf{X})}') \quad ,\end{aligned} \tag{3}$$

where $\widehat{\varphi(\mathbf{X})}'$ is an intermediate variable and LN indicates layer normalization. Moreover, FFN denotes the fully connected feed-forward network, which consists of two linear layers with a ReLU in between. Mod is a modulation layer in [4] to fully explore internal spatial information. The final output is flattened back to $N \times H \times W$. For each head of MSA, the attention function can be formulated as:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}})\mathbf{V} \quad . \tag{4}$$

In our case, the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are equal to the product of $(\mathbf{P} + \varphi(\mathbf{X}))$ and the corresponding projection matrix. By virtue of superior information integration of self-attention, the proposed Transformer bridging layer is adequate to modulate the nighttime object features output by the backbone for effective similarity maps.

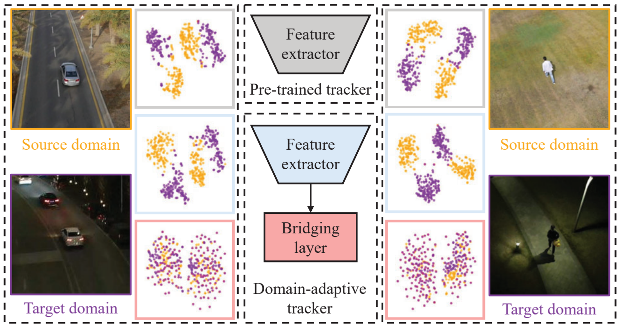


Figure 4. Feature visualization by t-SNE [42] of day/night similar scenes. Gold and purple indicate source and target domains, respectively. The scattergrams from top to down depict day/night features from feature extractor in the pre-trained tracker, feature extractor in the domain-adaptive tracker, and the bridging layer. The proposed Transformer bridging layer effectively narrows domain discrepancy.

Remark 1: Figure 4 displays the t-SNE [42] visualizations of features from feature extractor in the baseline, feature extractor in the domain-adaptive tracker, and the bridging layer. From which we can observe that features extracted by backbones have a clear discrepancy, while those modified by the bridging layer show a coincidence in distribution.

Transformer discriminator. The proposed UDAT framework is trained in an adversarial learning manner. A day/night feature discriminator D is designed to facilitate aligning the source and target domain features, which consists of a gradient reverse layer (GRL) [13] and two Transformer layers. Given the modulated feature map $\widehat{\varphi(\mathbf{X})}$, the softmax function is performed and followed by a GRL:

$$\mathbf{F} = \mathrm{GRL}(\mathrm{Softmax}(\widehat{\varphi(\mathbf{X})})) \quad . \tag{5}$$

The intermediate feature $\mathbf{F} \in \mathbb{R}^{N\times H\times W}$ is then passed through a convolution layer with a kernel size of 4×4 and stride of 4 for patch embedding. $\mathbf{F}$ is then flattened to $(\frac{H}{4} \times \frac{W}{4}) \times N$ and concatenated with a classification token $\mathbf{c}$ as:

$$\mathbf{F}' = \mathrm{Concat}(\mathbf{c}, \mathrm{Flat}(\mathrm{Conv}(\mathbf{F}))) \quad . \tag{6}$$

Afterward, $\mathbf{F}'$ is input to two Transformer layers. Ultimately, the classification token $\mathbf{c}$ is regarded as the final predicted results. In the adversarial learning process, the discriminator is optimized to distinguish whether the features are from the source domain or the target domain correctly.

Tracker head. After the Transformer bridging layer, cross-correlation operation is conducted on the modulated features $\widehat{\varphi(\mathbf{X})}$ and $\widehat{\varphi(\mathbf{Z})}$ to generate a similarity map. Finally, the tracker head performs the classification and regression process to predict the object position.

3.3. Objective functions

Classification and regression loss. In the source domain training line, the classification and regression loss $\mathcal{L}_{\mathrm{GT}}$ between the ground truth and the predicted results are applied to ensure the normal tracking ability of trackers. We adopt tracking loss consistent with the baseline trackers without modification.

Domain adaptation loss. In adversarial learning, the least-square loss function [30] is introduced to train the generator G, aiming at generating source domain-like features from target domain images to fool the discriminator D while frozen. Here the generator G can be regarded as the feature extractor along with the Transformer bridging layer. Considering both the template and search features, the adversarial loss is described as follows:

$$\mathcal{L}_{\mathrm{adv}} = (\mathrm{D}(\widehat{\varphi(\mathbf{X}_{\mathrm{t}})}) - \ell_{\mathrm{s}})^2 + (\mathrm{D}(\widehat{\varphi(\mathbf{Z}_{\mathrm{t}})})) - \ell_{\mathrm{s}})^2 \quad , \tag{7}$$

where s and t refer to the source and the target domains, respectively. Besides, ℓ_{s} denotes the label for the source domain, which has the same size as the output of D. In summary, the total training loss for the tracking network is defined as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{GT}} + \lambda \mathcal{L}_{\mathrm{adv}} \quad , \tag{8}$$

where λ is a weight to balance the loss terms. We set λ as 0.01 in implementation.

During the training process, the tracking network and discriminator D are optimized alternatively. We define the loss function of D as:

$$L_{\mathrm{D}} = \sum_{d=\mathrm{s,t}} (\mathrm{D}(\widehat{\varphi(\mathbf{X}_d)}) - \ell_d)^2 + (\mathrm{D}(\widehat{\varphi(\mathbf{Z}_d)}) - \ell_d)^2 \quad . \tag{9}$$

Trained with true domain labels of input features, D learns to discriminate feature domains efficiently.

4. NAT2021 benchmark

The nighttime aerial tracking benchmark, namely NAT2021, is developed to give a comprehensive performance evaluation of nighttime aerial tracking and provide adequate unlabelled nighttime tracking videos for unsupervised training. Compared to the existing nighttime tracking benchmark [21] in literature, NAT2021 stands a two times larger *test* set, an unlabelled *train* set, and novel illumination-oriented attributes.

4.1. Sequence collection

Images in NAT2021 are captured in diverse nighttime scenes (*e.g.*, roads, urban landscapes, and campus) by a DJI Mavic Air 2 UAV[1] in a frame rate of 30 frames/s. Sequence categories consist of a wide variety of targets (*e.g.*, cars, trucks, persons, groups, buses, buildings, and motorcycles) or activities (*e.g.*, cycling, skating, running, and ball games). Consequently, the *test* set contains 180 nighttime aerial tracking sequences with more than 140k frames in total, namely NAT2021-*test*. Figure 5 displays some first frames of selected sequences. In order to provide an evaluation of long-term tracking performance, we further build a long-term tracking subset namely NAT2021-*L*-*test* consisting of 23 sequences that are longer than 1400 frames. Moreover, the training set involves 1400 unlabelled sequences with over 276k frames totally, which is adequate for the domain adaptive tracking task. A statistical summary of NAT2021 is presented in Tab. 1.

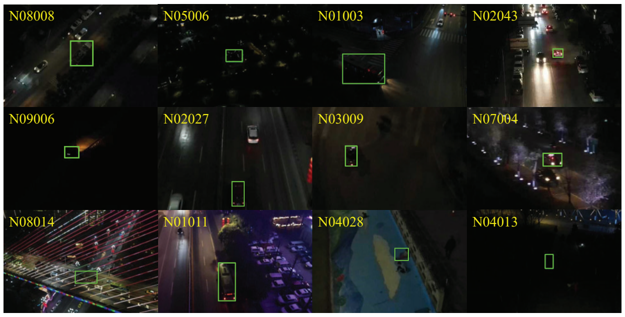


Figure 5. First frames of selected sequences from NAT2021-*test*. The green boxes mark the tracking objects, while the top-left corner of the images display sequence names.

Remark 2: Sequences in NAT2021 are recorded by ourselves using UAVs with permission. No personally identifiable information or offensive content is involved.

4.2. Annotation

The frames in NAT2021-*test* and NAT2021-*L*-*test* are all manually annotated by annotators familiar with object tracking. For accuracy, the annotation process is conducted on images enhanced by a low-light enhancement approach [24]. Afterward, visual inspection by experts and bounding box refinement by annotators are conducted iteratively for several rounds to ensure high-quality annotation.

4.3. Attributes

To give an in-depth analysis of trackers, we annotate the test sequences into 12 different attributes, including aspect ratio change (ARC), background clutter (BC), camera motion (CM), fast motion (FM), partial occlusion (OCC), full

Table 1. Statistics of NAT2021. *test*: test set; *L*-*test*: long-term tracking test set; *train*: train set.

	NAT2021-*test*	NAT2021-*L*-*test*	NAT2021-*train*
Videos	180	23	1,400
Total frames	140,815	53,564	276,081
Min frames	81	1,425	30
Max frames	1,795	3,866	345
Avg. frames	782	2,329	197
Manual annotation	✓	✓	

[1] More information of the UAV can be found at https://www.dji.com/cn/mavic-air-2.

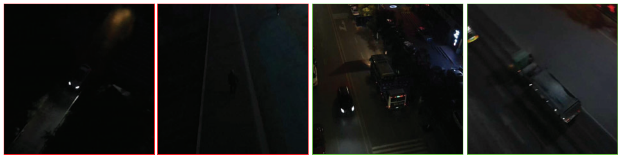

Ambient intensity: 7 Ambient intensity: 12 Ambient intensity: 26 Ambient intensity: 36

Figure 6. Ambient intensity of some scenarios. With an average ambient intensity of less than 20, objects are hard to distinguish with naked eyes. Such sequences are annotated with the low ambient intensity attribute.

occlusion (FOC), out-of-view (OV), scale variation (SV), similar object (SOB), viewpoint change (VC), illumination variation (IV), and low ambient intensity (LAI). In particular, to take a closer look at how illumination influences trackers, we rethink and redesign the illumination-related attributes. Concretely, the average pixel intensity of the local region centered at the object is computed and regarded as the illuminance intensity of the current frame. The average illuminance level of a sequence is considered the ambient intensity of the tracking scene. Sequences with different ambient intensities are displayed in Fig. 6, we observe that objects are hard to distinguish with naked eyes with an ambient intensity of less than 20. Therefore, such sequences are labelled with the LAI attribute.

Remark 3: In contrast to annotating the attribute of IV intuitively as previous tracking benchmarks do, this work judges IV according to the maximum difference of the illuminance intensity across a tracking sequence. More details of the attributes can be found in *supplementary material*.

Moreover, we evaluate current top-ranked trackers on the proposed benchmark (see Sec. 5.2), the results show that SOTA trackers can hardly yield satisfactory performance at a nighttime aerial view as in daytime benchmarks.

5. Experiments

5.1. Implementation details

We implement our UDAT framework using PyTorch on an NVIDIA RTX A6000 GPU. The discriminator is trained using the Adam [20] optimizer. The base learning rate is set to 0.005 and is decayed following the poly learning rate policy with a power of 0.8. The bridging layer adopts a base learning rate of 0.005 and is optimized with the baseline tracker. The whole training process lasts 20 epochs. The top-ranked trackers [8, 14] in the proposed benchmark are adopted as baselines. To achieve faster convergence, tracking models pre-trained on general datasets [10,17,28,32,34] are served as the baseline models. For fairness, tracking datasets [17, 32] that the pre-trained models learned on are adopted and no new datasets are introduced in the source domain. We adopt the one-pass evaluation (OPE) and rank performances using success rate, precision, and normalized precision. Evaluation metric definitions and more experiments can be found in the *supplementary material*.

5.2. Evaluation results

To give an exhaustive analysis of trackers in nighttime aerial tracking and facilitate future research, 20 SOTA trackers [1,4,5,8,11,14,15,22,29,39,44,47,51,53,54,56] are evaluated on NAT2021-$test$, along with the proposed UDAT. For clarity, two trackers further trained by UDAT are named UDAT-BAN and UDAT-CAR, respectively. Moreover, UAVDark70 [21] contains 70 nighttime tracking sequences with 66k frames in total, which can also serve as an evaluation benchmark.

5.2.1 Overall performance

NAT2021-*test*. As shown in Fig. 7 (a), the proposed UDAT-BAN and UDAT-CAR rank first two places with a large margin compared to their baselines. A performance comparison of UDAT and baseline trackers is reported in Tab. 2. Specifically, UDAT promotes SiamBAN over **7**% on all three metrics. In success rate, UDAT-BAN (0.469) and UDAT-CAR (0.483) raise the original SiamBAN (0.437) and SiamCAR (0.453) by **7.32**% and **6.62**%, respectively.

UAVDark70. Results in Fig. 7 (b) demonstrate that the performance of existing trackers is still unsatisfactory. UDAT trackers raise the performance of their baselines by ∼**4**%. Note that the data distribution in UAVDark70 is fairly different from that in NAT2021, while UDAT can still bring favorable performance gains, which demonstrate its generalization ability in variant nighttime conditions.

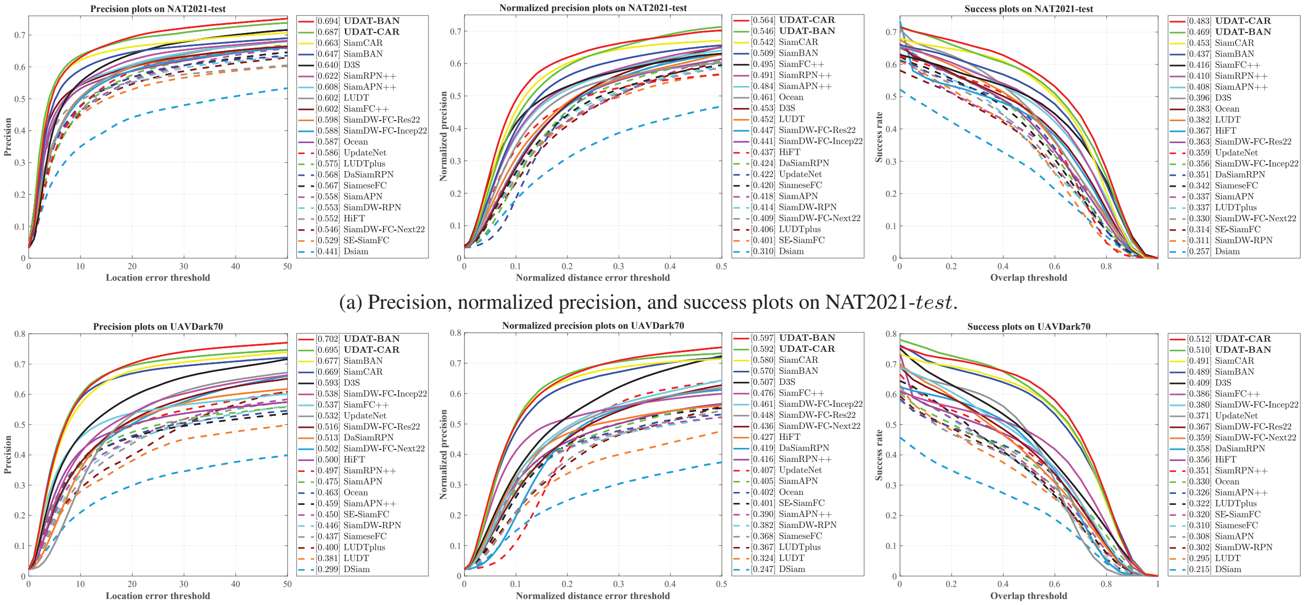
Gains brought by UDAT for different trackers on different benchmarks verify the effectiveness and transferability of the proposed domain adaptation framework.

5.2.2 Long-term tracking evaluation

As one of the most common scenes in aerial tracking, long-term tracking involves multiple challenging attributes. We further assess trackers on NAT2021-L-$test$. Top-10 performances are reported in Tab. 3. Results show that UDAT realizes competitive long-term tracking performances, considerably arousing the performance upon baseline trackers.

Table 2. Performance comparison of UDAT and baseline trackers. Δ denotes gains of percentages brought by UDAT.

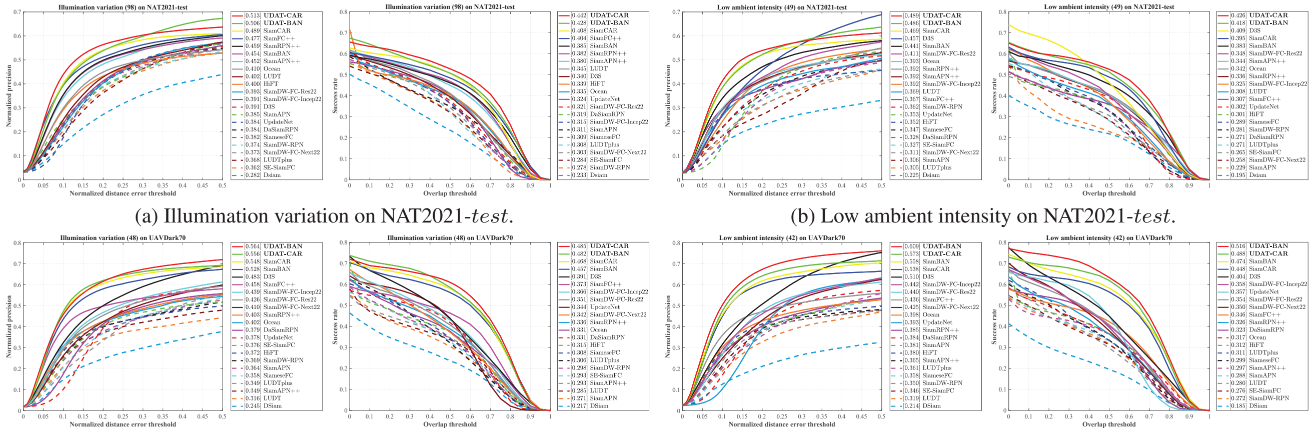
	NAT2021-$test$			UAVDark70		
Benchmarks	Prec.	Norm. Prec.	Succ.	Prec.	Norm. Prec.	Succ.
SiamCAR	0.663	0.542	0.453	0.669	0.580	0.491
UDAT-CAR	0.687	0.564	0.483	0.695	0.592	0.512
Δ_{CAR} (%)	**3.62**	**4.06**	**6.62**	**3.89**	**2.07**	**4.28**
SiamBAN	0.647	0.509	0.437	0.677	0.570	0.489
UDAT-BAN	0.694	0.546	0.469	0.702	0.597	0.510
Δ_{BAN} (%)	**7.26**	**7.27**	**7.32**	**3.69**	**4.74**	**4.29**



(a) Precision, normalized precision, and success plots on NAT2021-*test*.

(b) Precision, normalized precision, and success plots on UAVDark70.

Figure 7. Overall performance of SOTA trackers and UDAT on nighttime aerial tracking benchmarks. The results show that the proposed UDAT trackers realize top-ranked performance and improve baseline trackers favorably.



(a) Illumination variation on NAT2021-*test*.

(b) Low ambient intensity on NAT2021-*test*.

(c) Illumination variation on UAVDark70.

(d) Low ambient intensity on UAVDark70.

Figure 8. Normalized precision plots and success plots of illumination-related attributes on NAT2021-*test* and UAVDark70.

Table 3. Performance of top-10 trackers on NAT2021-*L-test*. Δ represents the percentages of UDAT trackers exceeding the corresponding baselines. The top-2 performance is emphasized with bold font. UDAT trackers yield competitive long-term tracking performance.

Trackers	HiFT [4]	SiamFC++ [47]	Ocean [54]	SiamRPN++ [22]	UpdateNet [51]	D3S [29]	SiamBAN [8]	SiamCAR [14]	**UDAT-BAN**	**UDAT-CAR**	$\Delta_{\mathrm{BAN}}(\%)$	$\Delta_{\mathrm{CAR}}(\%)$
Prec.	0.433	0.425	0.454	0.431	0.434	0.492	0.464	0.477	**0.496**	**0.506**	6.94	5.99
Norm. Prec.	0.316	0.344	0.370	0.342	0.314	0.364	0.366	0.375	**0.406**	**0.413**	11.01	9.96
Succ.	0.287	0.297	0.315	0.299	0.275	0.332	0.316	0.330	**0.352**	**0.376**	11.51	14.25

5.2.3 Illumination-oriented evaluation

Since the greatest difference between daytime and nighttime tracking is illumination intensity, we perform an in-depth illumination-oriented evaluation for a better analysis of illumination influence on trackers. The results are shown in Fig. 8. Note that we additionally annotate sequences in UAVDark70 with the proposed LAI attribute. The results show that existing trackers suffer from illumination-related attributes. For the IV challenge, the best success rates of existing trackers are 0.408 on NAT2021-*test* and 0.468 on UAVDark70. Assisted by the proposed domain adaptive training, UDAT-CAR realizes a success rate of 0.442 and 0.485, respectively, which fairly improve the existing best performance. As for LAI, UDAT-BAN raises the normalized precision of its baseline SiamBAN by over **9%** on both benchmarks. From the comparison, we can see that trackers' illumination-related performance remains a large room for improvement and the adoption of domain adaptation in adverse illumination scenes is effective and crucial.

5.2.4 Visualization

As shown in Fig. 9, we visualized some confidence maps of UDAT and its baseline using Grad-Cam [38]. The baseline model fails to concentrate on objects in adverse illuminance, while UDAT substantially enhances the baseline's nighttime perception ability, thus yielding satisfying nighttime tracking performance.

5.2.5 Source domain evaluation

Apart from favorable performance at nighttime, we expect that trackers do not suffer degradation at the source domain during adaptation. Evaluation on a daytime tracking benchmark UAV123 [31] is shown in Tab. 4. The results show that UDAT brings slight performance fluctuation within 2% in success rate and 0.5% in precision.

5.3. Empirical study

To demonstrate the effectiveness of proposed modules, *i.e.*, domain adaptive training (DA), object discovery preprocessing (OD), and bridging layer (BL), this subsection provides empirical studies of UDAT. Concretely, we first ablate BL and substitute OD with random cropping to adopt naive DA on the baseline tracker. The results on the second row of Tab. 5 show that DA slightly promotes nighttime tracking, with a slight upgrade in success rate. However, adopting random cropping as preprocessing leads to abundant meaningless training samples, the model therefore can hardly learn the data distribution on the target domain. In that case, further activation of BL only makes a limited difference. As shown in the fourth row of Tab. 5, when employing OD instead of random cropping, performance on the target domain obtains a 3.24% boost in success rate, which verifies the effectiveness of the proposed saliency detection-based data preprocessing. Further, BL doubles the promotion brought by OD, complete UDAT realizes a precision of 0.687 and a success rate of 0.483, achieving favorable nighttime tracking performance. The results verify that the proposed bridging layer fairly enables the tracker to generate discriminative features from nighttime images.

Table 4. Evaluation on the source domain. The results show the adaptation only brings slight performance fluctuation on the source domain.

Trackers	SiamBAN	UDAT-BAN	SiamCAR	UDAT-CAR
Succ.	0.603	$0.591_{1.96\%\downarrow}$	0.601	$0.592_{1.58\%\downarrow}$
Prec.	0.788	$0.784_{0.52\%\downarrow}$	0.793	$0.793_{0.04\%\downarrow}$

Table 5. Empirical Study of the proposed UDAT on NAT2021-*test*. DA, OD, and BL denote domain adaptive training, object discovery preprocessing, and bridging layer, respectively.

DA	OD	BL	Prec.	Norm. Prec.	Succ.
			0.663	0.542	0.453
✓			$0.662_{0.19\%\downarrow}$	$0.540_{0.33\%\downarrow}$	$0.459_{1.33\%\uparrow}$
✓		✓	$0.664_{0.16\%\uparrow}$	$0.547_{1.04\%\uparrow}$	$0.464_{2.45\%\uparrow}$
✓	✓		$0.676_{1.95\%\uparrow}$	$0.549_{1.42\%\uparrow}$	$0.467_{3.24\%\uparrow}$
✓	✓	✓	$\mathbf{0.687}_{3.62\%\uparrow}$	$\mathbf{0.564}_{4.17\%\uparrow}$	$\mathbf{0.483}_{6.82\%\uparrow}$

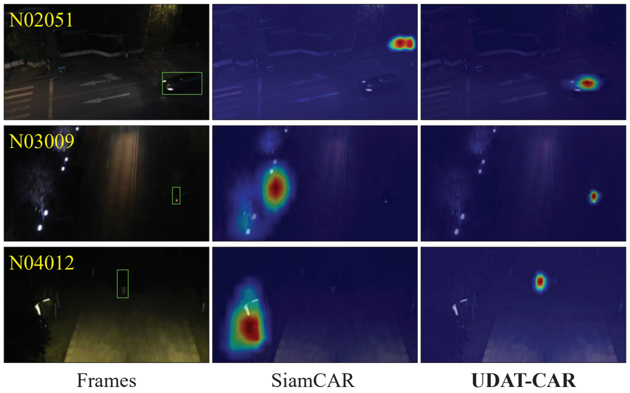


Figure 9. Visual comparison of confidence maps generated by the baseline and the proposed UDAT. Target objects are marked by green boxes. The baseline struggles to extract discriminable features in dim light. UDAT substantially raises the perception ability of baseline in adverse illuminance.

6. Conclusion

In this work, a simple but effective unsupervised domain adaptive tracking framework, namely UDAT, is proposed for nighttime aerial tracking. In our UDAT, an object discovery strategy is introduced for unlabelled data preprocessing. The Transformer bridging layer is adopted to narrow the gap of image features between daytime and nighttime. Optimized through adversarial learning with a Transformer discriminator, the learned model substantially improves nighttime tracking performance upon SOTA approaches. We also construct NAT2021, a pioneering benchmark for unsupervised domain adaptive nighttime tracking. Detailed evaluation on nighttime tracking benchmarks shows the effectiveness and domain adaptability of UDAT. The limitation of this work lies in the absence of pseudo supervision in the target domain. Future work will focus on reliable pseudo supervision, with which we believe the performance of nighttime tracking can be further improved. To sum up, we are convinced that the UDAT framework along with the NAT2021 benchmark can facilitate research on visual tracking at nighttime and in other adverse conditions.

Acknowledgement: This work was supported in part by the National Natural Science Foundation of China under Grant 62173249 and in part by the Natural Science Foundation of Shanghai under Grant 20ZR1460100.

References

[1] Luca Bertinetto, Jack Valmadre, João F. Henriques, Vedaldi Andrea, and Philip H. S. Torr. Fully-Convolutional Siamese Networks for Object Tracking. In *ECCVW*, pages 850–865, 2016. 2, 6

[2] Rogerio Bonatti, Cherie Ho, Wenshan Wang, Sanjiban Choudhury, and Sebastian Scherer. Towards a Robust Aerial Cinematography Platform: Localizing and Tracking Moving Targets in Unstructured Environments. In *IROS*, pages 229–236, 2019. 1

[3] Pau Panareda Busto and Juergen Gall. Open Set Domain Adaptation. In *ICCV*, pages 754–763, 2017. 2

[4] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. HiFT: Hierarchical Feature Transformer for Aerial Tracking. In *ICCV*, pages 15437–15446, 2021. 1, 2, 4, 6, 7

[5] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. SiamAPN++: Siamese Attentional Aggregation Network for Real-Time UAV Tracking. In *IROS*, pages 3086–3092, 2021. 6

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer Tracking. In *CVPR*, pages 8126–8135, 2021. 2

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In *CVPR*, pages 3339–3348, 2018. 2

[8] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese Box Adaptive Network for Visual Tracking. In *CVPR*, pages 6667–6676, 2020. 1, 2, 6, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255, 2009. 1

[10] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, Yong Xu, Chunyuan Liao, Lin Yuan, and Haibin Ling. LaSOT: A High-quality Large-scale Single Object Tracking Benchmark. *IJCV*, 129:439–461, 2021. 1, 6

[11] Changhong Fu, Ziang Cao, Yiming Li, Junjie Ye, and Chen Feng. Siamese Anchor Proposal Network for High-Speed Aerial Tracking. In *ICRA*, pages 510–516, 2021. 6

[12] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning Background-Aware Correlation Filters for Visual Tracking. In *ICCV*, pages 1144–1152, 2017. 2

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, volume 37, pages 1180–1189, 2015. 4

[14] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. In *CVPR*, pages 6268–6276, 2020. 1, 2, 6, 7

[15] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. Learning Dynamic Siamese Network for Visual Object Tracking. In *ICCV*, pages 1781–1789, 2017. 6

[16] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-Speed Tracking with Kernelized Correlation Filters. *IEEE TPAMI*, 37(3):583–596, 2015. 2

[17] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE TPAMI*, 43(5):1562–1577, 2021. 1, 6

[18] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. AugGAN: Cross Domain Adaptation with GAN-based Data Augmentation. In *ECCV*, page 731–744, 2018. 2

[19] Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, and Peng Lu. Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking. In *ICCV*, pages 2891–2900, 2019. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, pages 1–11, 2015. 6

[21] Bowen Li, Changhong Fu, Fangqiang Ding, Junjie Ye, and Fuling Lin. ADTrack: Target-Aware Dual Filter Learning for Real-Time Anti-Dark UAV Tracking. In *ICRA*, pages 496–502, 2021. 2, 5, 6

[22] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *CVPR*, pages 4277–4286, 2019. 1, 2, 6, 7

[23] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High Performance Visual Tracking with Siamese Region Proposal Network. In *CVPR*, pages 8971–8980, 2018. 2

[24] Chongyi Li, Chunle Guo, and Change Loy Chen. Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation. *IEEE TPAMI*, pages 1–14, 2021. 3, 5

[25] Rui Li, Minjian Pang, Cong Zhao, Guyue Zhou, and Lu Fang. Monocular Long-Term Target Following on UAVs. In *CVPRW*, pages 29–37, 2016. 1

[26] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. Domain Generalization and Adaptation Using Low Rank Exemplar SVMs. *IEEE TPAMI*, 40(5):1114–1127, 2018. 2

[27] Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. AutoTrack: Towards High-Performance Visual Tracking for UAV With Automatic Spatio-Temporal Regularization. In *CVPR*, pages 11920–11929, 2020. 2

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755, 2014. 6

[29] Alan Lukežič, Jiří Matas, and Matej Kristan. D3S – A Discriminative Single Shot Segmentation Tracker. In *CVPR*, pages 7131–7140, 2020. 6, 7

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least Squares Generative Adversarial Networks. In *ICCV*, pages 2813–2821, 2017. 5

[31] Matthias Mueller, Neil Smith, and Bernard Ghanem. A Benchmark and Simulator for UAV Tracking. In *ECCV*, pages 445–461, 2016. 8

[32] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. In *CVPR*, pages 7464–7473, 2017. 1, 6

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE TPAMI*, 39(6):1137–1149, 2017. 2

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 6

[35] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning To Relate Depth and Semantics for Unsupervised Domain Adaptation. In *CVPR*, pages 8197–8207, 2021. 2

[36] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-Guided Curriculum Domain Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation. *IEEE TPAMI*, pages 1–15, 2020. 2

[37] Yukihiro Sasagawa and Hajime Nagahara. YOLO in the Dark - Domain Adaptation Method for Merging Multiple Models. In *ECCV*, pages 345–359, 2020. 2

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*, pages 618–626, 2017. 8

[39] Ivan Sosnovik, Artem Moskalev, and Arnold W.M. Smeulders. Scale Equivariance Improves Siamese Tracking. In *WACV*, pages 2765–2774, January 2021. 6

[40] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of Frustratingly Easy Domain Adaptation. In *AAAI*, pages 2058–2065, 2016. 2

[41] Ran Tao, Efstratios Gavves, and Arnold W. M. Smeulders. Siamese Instance Search for Tracking. In *CVPR*, pages 1420–1429, 2016. 2

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11):2579–2605, 2008. 4

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NeurIPS*, pages 6000–6010, 2017. 2, 4

[44] Ning Wang, Wengang Zhou, Yibing Song, Chao Ma, Wei Liu, and Houqiang Li. Unsupervised Deep Representation Learning for Real-Time Tracking. *IJCV*, 129(2):400–418, 2021. 6

[45] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. In *CVPR*, pages 1571–1580, 2021. 2

[46] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation. In *CVPR*, pages 15769–15778, 2021. 2

[47] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. In *AAAI*, pages 12549–12556, 2020. 2, 6, 7

[48] Junjie Ye, Changhong Fu, Ziang Cao, Shan An, Guangze Zheng, and Bowen Li. Tracker Meets Night: A Transformer Enhancer for UAV Tracking. *IEEE RA-L*, 7(2):3866–3873, 2022. 1, 2

[49] Junjie Ye, Changhong Fu, Fuling Lin, Fangqiang Ding, Shan An, and Geng Lu. Multi-Regularized Correlation Filter for UAV Tracking and Self-Localization. *IEEE TIE*, 69(6):6004–6014, 2022. 1

[50] Junjie Ye, Changhong Fu, Guangze Zheng, Ziang Cao, and Bowen Li. DarkLighter: Light Up the Darkness for UAV Tracking. In *IROS*, pages 3079–3085, 2021. 1, 2

[51] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Learning the Model Update for Siamese Trackers. In *ICCV*, pages 4009–4018, 2019. 6, 7

[52] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic Context-Sensitive Filtering Network for Video Salient Object Detection. In *ICCV*, pages 1533–1543, 2021. 3

[53] Zhipeng Zhang and Houwen Peng. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In *CVPR*, pages 4586–4595, 2019. 6

[54] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-Aware Anchor-Free Tracking. In *ECCV*, pages 771–787, 2020. 6, 7

[55] Jilai Zheng, Chao Ma, Houwen Peng, and Xiaokang Yang. Learning to Track Objects from Unlabeled Videos. In *ICCV*, pages 13526–13535, 2021. 3

[56] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware Siamese Networks for Visual Object Tracking. In *ECCV*, pages 103–119, 2018. 6