2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

DaDiff: Domain-aware Diffusion Model for Nighttime UAV Tracking

Haobo Zuo[1], Changhong Fu[2,*], Guangze Zheng[1], Liangliang Yao[2], Kunhan Lu[2], and Jia Pan[1]

***Abstract*— Domain adaptation is an inspiring solution to the misalignment issue of day/night image features for nighttime UAV tracking. However, the one-step adaptation paradigm is inadequate in addressing the prevalent difficulties posed by low-resolution (LR) objects when viewed from the UAVs at night, owing to the blurry edge contour and limited detail information. Moreover, these approaches struggle to perceive LR objects disturbed by nighttime noise. To address these challenges, this work proposes a novel progressive alignment paradigm, named domain-aware diffusion model (DaDiff), aligning nighttime LR object features to the daytime by virtue of progressive and stable generations. The proposed DaDiff includes an alignment encoder to enhance the detail information of nighttime LR objects, a tracking-oriented layer designed to achieve close collaboration with tracking tasks, and a successive distribution discriminator presented to distinguish different feature distributions at each diffusion timestep successively. Furthermore, an elaborate nighttime UAV tracking benchmark is constructed for LR objects, namely NUT-LR, consisting of 100 annotated sequences. Exhaustive experiments have demonstrated the robustness and feature alignment ability of the proposed DaDiff. The source code and video demo are available at https://github.com/vision4robotics/DaDiff.**

I. Introduction

Vision-based UAV tracking has been widely applied for intelligent robot applications, *e.g.*, motion object analysis [1], geographical survey [2], and visual localization [3]. With high-quality daytime tracking datasets [4]–[6], the state-of-the-art (SOTA) trackers [7], [8] have achieved superior performance. However, these trackers perform poorly in night scenes because of the decreased illumination, signal-to-noise ratio, and contrast of nighttime images compared to daytime ones [9], [10]. The above differences between day and night images or image features cause the distribution discrepancy, spawning an extremely challenging application, *i.e.*, nighttime UAV tracking [11].

In literature, the SOTA methods [9], [10] construct tracking-oriented low-light enhancers with cutting-edge trackers to realize nighttime UAV tracking. Nevertheless, this kind of plug-and-play method generally focuses on the image level and can scarcely learn to minimize the distribution gap at the feature level, which is insufficient in providing discriminative image features required for high-accuracy tracking. Although the one-step adaption paradigm [12] is researched to achieve image feature alignment with end-to-end training, such kind of method performs unstably when facing common low-resolution (LR) object challenges from UAV perspectives due to the following two reasons: *1)* LR objects are hard to be identified from the background in one step due to the limited detail information and nighttime noise interference [14]; *2)* aligning the features of nighttime LR objects in one step is unstable due to the mismatch between the receptive field on LR features and the object sizes [15]. ***Therefore, how to align the nighttime LR object features to the daytime effectively and stably for robust nighttime UAV tracking is an urgent problem.***

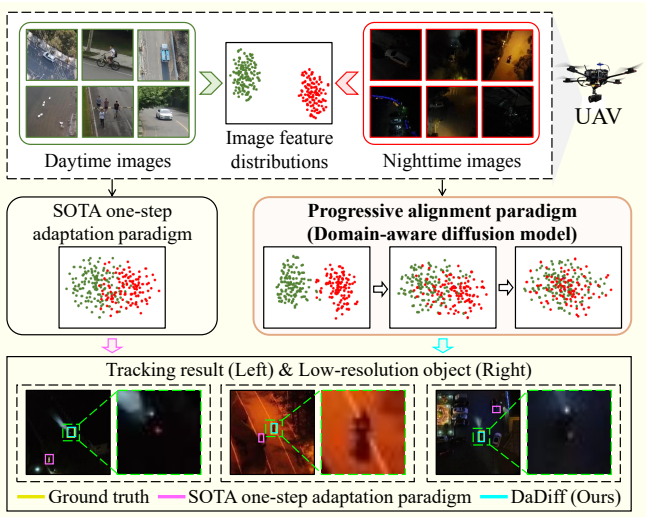

Fig. 1. Comparison of the one-step adaptation paradigm and the proposed domain-aware diffusion model, *i.e.*, DaDiff, for nighttime UAV tracking. The feature distributions are visualized through t-SNE [13]. Green and red indicate the daytime and nighttime image feature distributions, respectively. The scattergrams depict day/night feature distributions from different feature alignment methods. DaDiff successively and steadily narrows feature distribution discrepancy, thereby achieving superior tracking results, especially for low-resolution (LR) objects.

Diffusion models [16]–[18] have achieved superior performance in reconstructing the object information for LR images [19]–[22]. Diffusion-based methods can be regarded as a sort of variable model that uses a Markov chain to convert noise into data distribution. The generation ability of these models is typically derived from the step-by-step closing to the data distribution with the U-Net neural network. Additionally, such formulation allows for a guiding mechanism to control the image generation process with stability [23]. Compared to one-step generation methods [12], [24], multi-step diffusion models [18] offer significant advantages. They can avoid issues such as high-frequency information loss, excessive smoothness, mode collapse, and effect instability [20]. By progressively enhancing detail information and sharpening edge contours of LR objects, diffusion models

*Corresponding author
[1]Haobo Zuo, Guangze Zheng, and Jia Pan are with the Department of Computer Science, University of Hong Kong, and also with the Centre for Transformative Garment Production, Hong Kong 999077, China.
[2]Changhong Fu, Liangliang Yao, and Kunhan Lu are with the School of Mechanical Engineering, Tongji University, Shanghai 201804, China. Email: changhongfu@tongji.edu.cn

show promise for addressing feature alignment issues in the common LR object challenges of nighttime UAV tracking. Despite their potential, diffusion models have not yet been explored for nighttime UAV tracking. Furthermore, diffusion models are trained independently and cannot be seamlessly integrated into downstream tracking tasks. Therefore, bridging strategies are needed to leverage these models for nighttime UAV tracking effectively.

This work introduces the diffusion models into nighttime UAV tracking for the first time, proposing a novel domain-aware diffusion model, *i.e.*, DaDiff. Specifically, the alignment encoder is developed to obtain valid domain-aware information of LR objects in negative light conditions. The tracking-oriented layer is presented to achieve close collaboration with tracking. To ensure the stability of alignment, the successive distribution discriminator is applied for identifying the different feature distributions at each diffusion timestep. The aligned result comparison of DaDiff and the one-step adaptation paradigm is exhibited in Fig. 1. DaDiff raises the tracking performance through successive and stable feature alignment. Besides, NUT-LR, an elaborate nighttime tracking benchmark, is constructed including 100 annotated sequences as the first LR object benchmark for nighttime UAV tracking. It focuses on the LR object challenges of nighttime UAV tracking, aiming at promoting the research on nighttime tracking to a broader field. The main contributions of this work are as follows:

- A novel progressive alignment paradigm, *i.e.*, DaDiff, is proposed for nighttime UAV tracking. According to our knowledge, this work first applies diffusion models for nighttime UAV tracking.
- An alignment encoder is developed to strengthen the detail information of nighttime LR objects. A tracking-oriented layer and a successive distribution discriminator are included to closely connect with the tracking tasks and gradually narrow the feature distribution gap between daytime and nighttime.
- A pioneering benchmark namely NUT-LR, comprising 100 annotated sequences with LR object challenges, is constructed for evaluation of LR object nighttime tracking under the UAV perspective.
- Comprehensive evaluation on NUT-LR, NUT-L [25], and UAVDark70 [11] benchmarks demonstrate the effectiveness and feature alignment ability of the proposed DaDiff for nighttime UAV tracking.

II. Related work

A. Nighttime UAV tracking

Nighttime UAV tracking has been applied for numerous practical applications, raising broad attention recently. At first, the SOTA approaches [9], [10] develop tracking-oriented low-light enhancers for nighttime UAV tracking, using leading-edge Siamese trackers [7], [26], [27]. However, this kind of approach has a limited connection with tracking tasks, and straightforward insertion tracking models hardly learn to reduce the distribution gap at the feature level. Due to the ability to reduce domain disparity and transfer knowledge from the source domain to the target domain, domain adaptation has been employed for various vision tasks [28], [29]. UDAT [12] brings unsupervised domain adaptation in nighttime UAV tracking for the first time, improving the tracker performance. Nevertheless, the one-step adaptation paradigm performs unstably when facing common LR object challenges in nighttime UAV tracking. The negative illumination conditions seriously weaken the detail information of the LR object, blurring its edge contour. Additionally, these adverse light conditions exacerbate nighttime noise interference. It is hard for the one-step adaptation paradigm to perceive and extract low-resolution object features directly.

B. Diffusion models

As a pioneering work, DDPM [16] represents a unique class of variable models that leverage a Markov chain to transition from a noise distribution to a data distribution. Based on it, DDIM [18] adopts smaller sampling steps to speed up the generation process, with the characteristic of generating deterministic samples from random noise. Recently, diffusion probabilistic models have achieved SOTA performance in reconstructing the object information for LR images [19]–[22]. S. Gao *et al.* [19] propose an implicit diffusion model for high-fidelity continuous LR image information enhancement. H. Li *et al.* [20] introduce the diffusion probabilistic model into image super-resolution, handling the over smoothness and model collapse. Furthermore, Z. Yue *et al.* [21] reduce the number of diffusion steps and eliminate the need for post-acceleration during inference, thereby realizing efficient LR image information recovery with the diffusion model. C. Saharia *et al.* [22] utilize the denoising diffusion probabilistic models to strengthen the detail information of LR images via repeated refinement. Despite significant development, diffusion models for nighttime UAV tracking have not been researched. Moreover, because diffusion models are trained independently, seamless integration into downstream tracking tasks remains elusive. Therefore, an effective diffusion model-based alignment framework for nighttime UAV tracking is urgently required.

III. Proposed method

In this section, the detailed structure of this work is described, as shown in Fig. 2. Throughout the training process, the proposed DaDiff aligns the features produced by the tracker backbone. In this procedure, adversarial learning successively reduces the gap between daytime and nighttime feature distributions at each diffusion timestep. By using this simple but effective alignment method, trackers can attain equal levels of stability and accuracy for night situations as they can during the daytime.

A. Feature extraction network

Siamese network feature extraction typically consists of two branches, the search branch and the template branch. By using the same backbone network, both branches extract

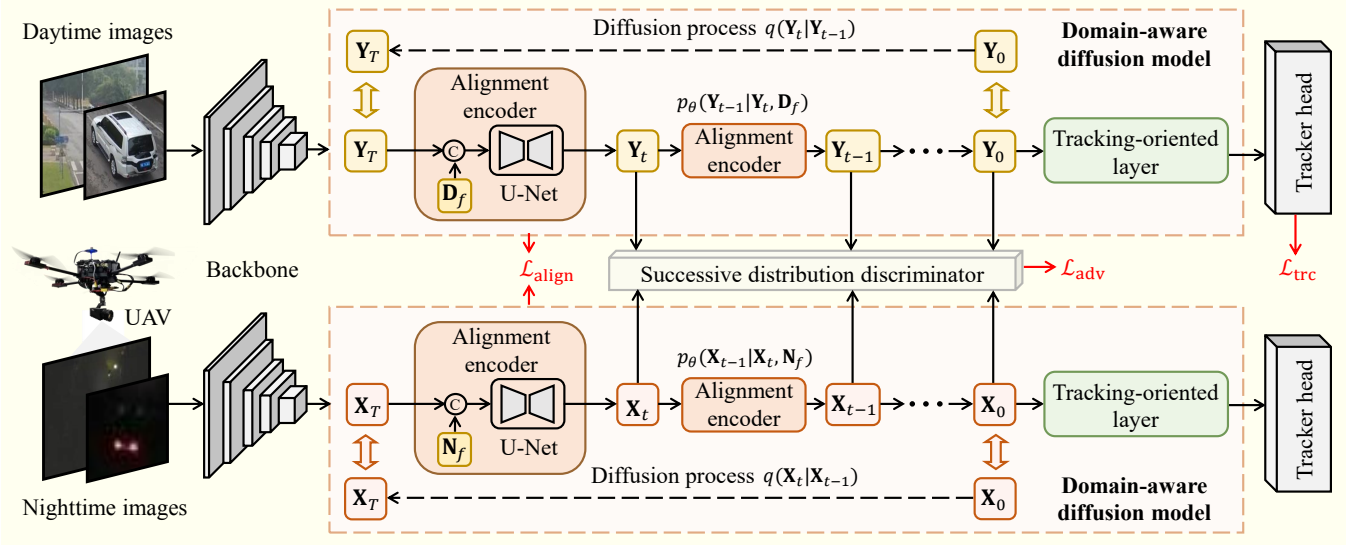

Fig. 2. Overview of the proposed DaDiff. *Domain-aware diffusion model* with *alignment encoder* is employed to narrow feature distribution discrepancy successively, achieving the feature alignment for nighttime UAV tracking. *Tracking-oriented layer* is developed to closely connect with the tracking tasks. *Successive distribution discriminator* is trained to distinguish features between the daytime and the nighttime gradually. Best viewed in color.

feature maps from the template patch $\mathbf{T}$ and the search patch $\mathbf{S}$, namely $\mathcal{F}(\mathbf{T})$ and $\mathcal{F}(\mathbf{S})$, by adopting an identical backbone network. Typically, trackers use the features of the last block or blocks for classification and regression.

Remark 1: Since both $\mathcal{F}(\mathbf{T})$ and $\mathcal{F}(\mathbf{S})$ of daytime and nighttime will pass through the weight-share DaDiff and the discriminator, the following introduction uses the nighttime features $\mathbf{N}_f$ as an example for clarity.

B. Domain-aware diffusion model

Alignment encoder. Diffusion models [18] are probabilistic models designed to learn a data distribution $p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t)$ by gradually denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov chain of length T. The features extracted by the feature extraction network are input into the diffusion models to generate the version of the corresponding daytime distribution. Specifically, in the forward diffusion process, the noise is gradually added to the data $\mathbf{X}_t \sim q(\mathbf{X}_t|\mathbf{X}_{t-1})$ in T steps with pre-defined value schedule α_t:

$$q(\mathbf{X}_t|\mathbf{X}_{t-1}) = \mathcal{N}(\mathbf{X}_t; \sqrt{1-\beta_t}\mathbf{X}_{t-1}, \beta_t\mathbf{I}) \quad , \tag{1}$$

where $\beta_t = 1-\alpha_t/\alpha_{t-1}$. A notable characteristic of diffusion models is that $\mathbf{X}_t$ at an arbitrary time-step t can be sampled from $\mathbf{X}_0$ as:

$$\mathbf{X}_t = \sqrt{\alpha_t}\ \mathbf{X}_0 + \sqrt{1-\alpha_t}\epsilon \quad , \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is a noise variable. While the reverse process is a process of noise removal. This process starts from random noise and gradually denoises to generate a real sample $p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t)$ according to the true distribution of each step of the reverse process. The proposed DaDiff utilizes the denoising process to achieve the day/night feature alignment, successively enhancing the diminished object information due to adverse illumination conditions. Thereby

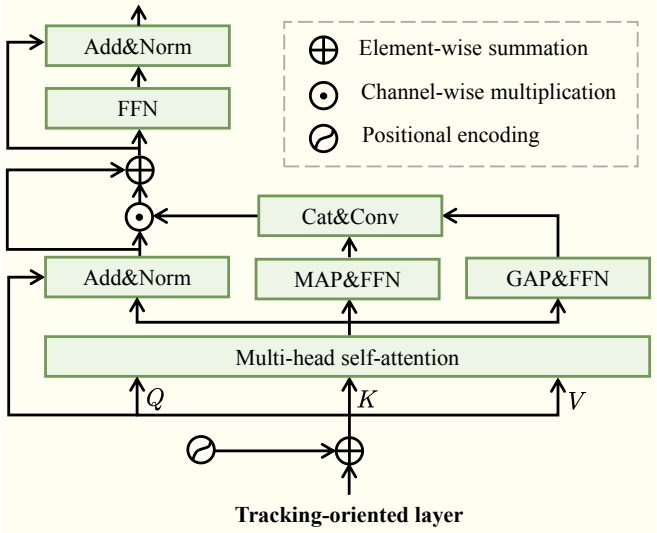

Fig. 3. Detailed workflow of *Tracking-oriented layer*. With the powerful information integration ability of Transformer [30] and the internal information exploration, *Tracking-oriented layer* can integrate the effective domain-aware information of aligned LR object features, closely collaborating with the tracking tasks.

the reverse process is also the process of generating data:

$$\begin{aligned} p_\theta(\mathbf{X}_{0:T}) &= \mathcal{N}(\mathbf{X}_0; 0, \mathbf{I}) \prod_{t=1}^{T} p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t) \quad , \\ p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t) &= \mathcal{N}(\mathbf{X}_{t-1}; \mu_\theta, \sigma_\theta^2\mathbf{I}) \quad , \end{aligned} \tag{3}$$

where μ_θ and σ_θ are parameters of the Gaussian distribution predicted by model p_θ. This process can be further interpreted as an equally weighted sequence of auto-encoders $\epsilon_\theta(\mathbf{X}_t, t); t = 1, ..., T$, which are trained to predict a denoised variant of their input $\mathbf{X}_t$, where $\mathbf{X}_t$ is a noisy version of $\mathbf{X}_0$. The process to obtain $\mathbf{X}_{t-1}$ can be expressed by:

$$\begin{aligned} \hat{\mathbf{X}}_0 &= \frac{\mathbf{X}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{X}_t, t))}{\sqrt{\alpha_t}} \quad , \\ \mathbf{X}_{t-1} &= \sqrt{\alpha_{t-1}}\ \hat{\mathbf{X}}_0 + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\mathbf{X}_t, t) \quad , \end{aligned} \tag{4}$$

then we can gradually get the desired data distribution by eliminating the noise predicted in each step. The corresponding objective $\mathcal{L}_{\text{dm}}$ can be simplified to:

$$\mathcal{L}_{\text{dm}} = \mathbb{E}_{\mathbf{X}_0,\epsilon\sim\mathcal{N},t}[\|\epsilon - \epsilon_\theta(\mathbf{X}_t, t)\|_2^2] \quad , \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is a noise variable and t is uniformly sampled from $\{1, \ldots, T\}$.

To meet the generation needs of specific tasks, the conditional mechanism is introduced into diffusion models [31]. Similar to other types of generative models [24], diffusion models are in principle capable of modeling conditional distributions of the form $p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t, \mathbf{N}_f)$. This can be implemented with a conditional alignment encoder $\epsilon_\theta(\mathbf{X}_t, \mathbf{N}_f, t)$ and paves the way to controlling the synthesis process through inputs $\mathbf{N}_f$. Therefore, the proposed DaDiff concatenates $\mathbf{X}_t$ with the flexible latent condition $\mathbf{N}_f$ to augment the generation capabilities of tracking-specific distribution. Generally, $\epsilon_\theta(\mathbf{X}_t, \mathbf{N}_f, t)$ and T timesteps are trained by a simplified objective $\mathcal{L}_{\text{align}}$:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{\mathbf{X}_0,\mathbf{N}_f,\epsilon\sim\mathcal{N},t}[\|\epsilon - \epsilon_\theta(\mathbf{X}_t, \mathbf{N}_f, t)\|_2^2] \quad , \tag{6}$$

where $\mathbf{X}_t$ is a linear combination of data $\mathbf{X}_0$ and noise ϵ by diffusion process. While during inference, the corresponding reverse generative Markov chain produces the expected data distribution $p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t, \mathbf{N}_f)$ by denoising process. Afterward, the denoised image features are input into the tracking-oriented layer to integrate the object information, closely connecting with the tracking tasks.

Remark 2: Through the successive denoising of nighttime features $\mathbf{N}_f$, aligned image features can be generated stably and controllably. Thereby it is able to handle the common LR object challenges in the night scenes effectively, especially in the interference of adverse illumination conditions.

Tracking-oriented layer. Diffusion models are difficult to directly collaborate with the tracking task due to their fixed training paradigm. Therefore, this work develops a tracking-oriented layer to integrate the domain-aware information of LR objects, bridging the aligned feature generation and the tracking process. In consideration of the strong modeling capability of the Transformer [30] for long-range inter-independencies, the tracking-oriented layer applies a Transformer structure, as shown in Fig. 3. The aligned features $\mathbf{N}_f^a$ are obtained after this layer. Specifically, the denoised results $\mathbf{X}_0$ are reshaped to $\mathbf{X}_0^a$ before encoding. Subsequently, the input of this layer $\mathbf{X}_0^b$ can be obtained by supplementing with a learnable positional encoding. The subsequent process can be expressed by:

$$\begin{aligned}
\mathbf{X}_0^c &= \text{Norm}(\text{mAtt}(\mathbf{X}_0^b) + \mathbf{X}_0^b) \quad , \\
\mathcal{W} &= \text{Conv}(\text{Cat}(\text{GAP}(\mathbf{X}_0^c), \text{MAP}(\mathbf{X}_0^c)) \quad , \\
\mathbf{X}_0^d &= \mathbf{X}_0^c + \gamma_1 * \mathcal{W} * \mathbf{X}_0^c \quad , \\
\mathbf{N}_f^a &= \text{Norm}(\text{FFN}(\mathbf{X}_0^d) + \mathbf{X}_0^d) \quad ,
\end{aligned} \tag{7}$$

where mAtt shows the mult-head self-attention. $\mathbf{X}_0^c$, $\mathbf{X}_0^d$ are intermediate variables and $\mathcal{W}$ is a weight matrix. GAP and MAP represent the global average pooling and max average pooling, thoroughly investigating the latent spatial information. Norm indicates layer normalization. In addition, FFN denotes the fully connected feed-forward network, which comprises of two linear layers separated by a ReLU. Besides, γ_1 and $*$ represent a learning weight and the channel-wise multiplication respectively. The final output is reshaped to the original size.

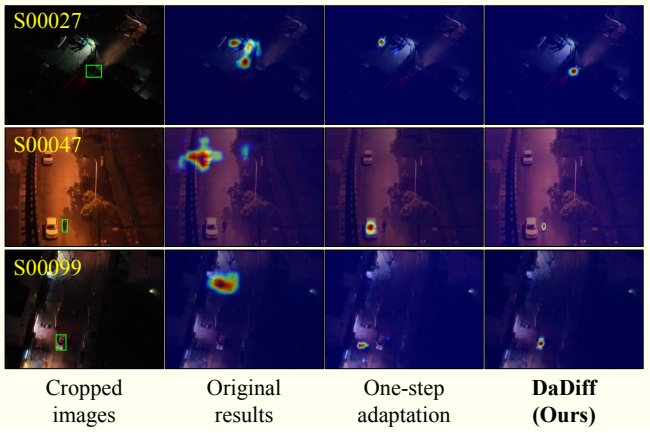

Fig. 4. Visual comparison of confidence maps generated by the Baseline, the one-step adaptation paradigm, and the proposed DaDiff. Target objects are marked by green boxes. The Baseline and the one-step adaptation paradigm struggle to extract robust LR object features in the interference of adverse illumination conditions. DaDiff stably and controllably aligns the image features by day/night domain awareness and applying the successive alignment strategy.

Remark 3: By virtue of superior information integration of Transformer, the proposed tracking-oriented layer is adequate to integrate the effective domain-aware information of aligned LR object features, thereby closely connecting the diffusion models with the tracking tasks.

Successive distribution discriminator. The proposed DaDiff framework is trained in an adversarial learning manner. A successive distribution discriminator [12] is applied to distinguish the different feature distributions at each diffusion timestep. Thereby DaDiff can step-by-step align the nighttime features with the daytime, thus handling the LR object challenges for nighttime UAV tracking, especially in the interference of adverse illumination conditions. In every diffusion process, the successive distribution discriminator D judges whether the features are from day or night. The adversarial optimization objective can be described as follows:

$$\mathcal{L}_{\text{adv}} = \sum_{t=1}^{T} (\text{D}(\mathbf{X}_t) - l_d)^2 \quad , \tag{8}$$

where t refers to the diffusion timestep. Besides, l_d denotes the label for the daytime features, which has the same size as the output of D.

Remark 4: DaDiff adopts the successive alignment strategy to generate the aligned features, more stable and controllable than the one-step adaptation paradigm. Thereby it can perceive and extract the robust LR object features in nighttime UAV scenes through gradual denoising. The superior performance of the proposed framework has been shown in Fig. 4, using Grad-Cam [32]. Moreover, Algorithm 1 displays the complete inference procedure of DaDiff.

Algorithm 1 Domain-aware diffusion model

Input: nighttime features $\mathbf{N}_f$

$\mathbf{X}_t \sim q(\mathbf{X}_t|\mathbf{X}_{t-1})$

for $t = T, ..., 1$ **do**

$\hat{\mathbf{X}}_0 = \frac{\mathbf{X}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{X}_t, \mathbf{N}_f, t))}{\sqrt{\alpha_t}}$

$\mathbf{X}_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{\mathbf{X}}_0 + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\mathbf{X}_t, \mathbf{N}_f, t)$

end for

Send $\mathbf{X}_0$ into tracking-oriented layer to generate $\mathbf{N}_f^a$

Output: aligned features $\mathbf{N}_f^a$

C. Tracker head

Following the feature alignment, the tracker head predicts the tracked object's location using classification and regression. In the daytime training phase, the classification and regression loss $\mathcal{L}_{\text{trc}}$ are applied to connect DaDiff with the tracking task, assuring the trackers' normal tracking capacity. The applied tracking loss is commensurate with the baseline trackers without change. In conclusion, the total training loss for the proposed framework is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{trc}} + \lambda_2\mathcal{L}_{\text{adv}} + \lambda_3\mathcal{L}_{\text{align}} \quad , \tag{9}$$

where λ_1, λ_2, and λ_3 are the coefficients to balance the contributions of each loss, respectively.

IV. NUT-LR Benchmark

This work develops a nighttime UAV tracking dataset, namely NUT-LR, to evaluate the nighttime tracking performance comprehensively, especially for LR objects. Compared with the nighttime UAV tracking benchmarks [11], [25], NUT-LR provides a dedicated dataset covering LR objects in various adverse light scenes highly related to the practical applications, as shown in Fig. 5.

Remark 5: Referring to the authoritative public dataset [5] and actual UAV tracking, the LR object is defined that the size of the target is less than 25×25 in NUT-LR.

A. Data collection

A classical UAV platform is applied to photograph images of NUT-LR in diverse evening views at 30 frames/s, such as highways, squares, bridges, and universities. The UAV tracks LR objects from an UAV perspective of more than 100 meters. Sequence categories include various objectives, *e.g.*, cars, persons, groups, bikes, and motorcycles. Moreover, the proposed benchmark NUT-LR contains 100 nighttime UAV tracking sequences in total.

B. Attributes

The test sequences of NUT-LR are categorized into 10 various attributes to provide a thorough study of trackers, including aspect ratio change (ARC), background clutter (BC), camera motion (CM), fast motion (FM), occlusion (OCC), scale variation (SV), similar object (SOB), viewpoint change (VC), illumination variation (IV), and low ambient intensity (LAI) [12]. Tracking LR objects under these attributes is more challenging than general objects. Due to the few pixels in the nighttime images, the detail information of the LR object is seriously weakened by adverse illumination conditions, such as IV and LAI. Therefore, NUT-LR can promote the tracker designs with additional challenges.

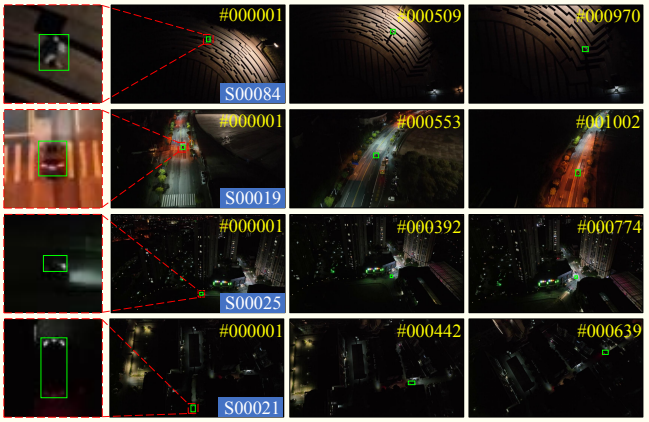

Fig. 5. Typically frames of selected sequences from NUT-LR. The green boxes mark the tracked objects and the red dotted boxes are the enlarged target areas for a clear view of the tracked LR objects. While the bottom-right corner of the image displays the sequence name and the top-right one shows the frame number.

Remark 6: SOTA trackers are evaluated on the proposed benchmark and the results demonstrate that existing trackers and the one-step adaptation paradigm hardly provide sufficient performance when confronting LR objects in nighttime UAV tracking.

V. Experiments

A. Implementation details

The proposed DaDiff framework is implemented on an NVIDIA A100 Tensor Core GPU with PyTorch. The base learning rate of the successive distribution discriminator is set at 0.005 and decays according to the poly learning rate policy with a 0.8 power. While DaDiff adopts a base learning rate of 0.0015 and is optimized with the baseline tracker. There are 50 epochs throughout the whole training procedure. The SOTA trackers [7], [8] are adopted as Baselines. Pre-trained tracking models on generic datasets [4]–[6] are used as the baseline models to accelerate convergence. For the sake of fairness, in the daytime training branch, only the tracking datasets [4], [6] on which the pre-trained models were trained are used, and no additional daytime datasets are added. While in the nighttime training branch, the unlabeled benchmark NAT2021-train [12] is applied for alignment training.

B. Evaluation metrics

In one-pass evaluation metrics, precision, normalized precision, and success rate are key factors for evaluating tracker performance [33]. The success rate is calculated by considering the intersection over union (IoU) between the actual and predicted bounding boxes. The success plot represents the fraction of frames where the IoU exceeds a preset threshold. Precision, on the other hand, is determined by measuring the center location error (CLE) between the predicted and actual locations. The precision plot visualizes the share of frames where the CLE falls within a specific range. Furthermore,

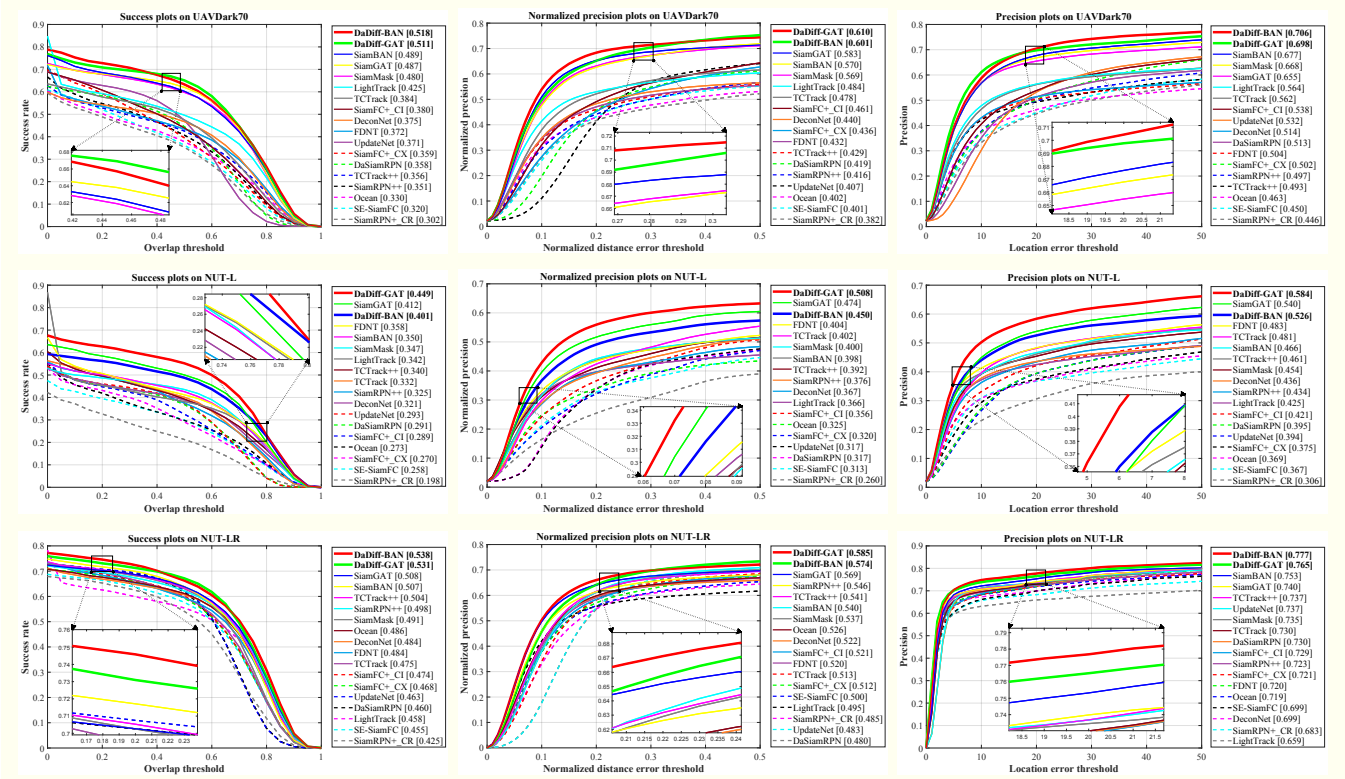

Fig. 6. Overall performance of SOTA trackers and DaDiff on UAVDark70 [11], NUT-L [25], and NUT-LR benchmarks. The evaluation results indicate that the proposed method improves the tracking performance on all benchmarks.

normalized precision is obtained by normalizing precision across different sizes of the ground truth bounding box, aiming to eliminate the impact of varying object sizes on precision. The normalized precision plot is evaluated by calculating the area under the curve.

C. Evaluation results

16 SOTA trackers [7], [8], [26], [27], [34]–[42] are tested on NUT-LR, together with the proposed DaDiff, to provide an extensive evaluation of trackers for nighttime UAV tracking and aid further research. For clarity, two trackers further trained by DaDiff are named DaDiff-GAT and DaDiff-BAN, respectively. Moreover, two challenging and authoritative public datasets, *i.e.*, NUT-L [25] and UAVDark70 [11] are also served as the evaluation benchmarks.

Remark 7: For the justice, every compared tracker adopts the tracking model from the official code and all evaluation experiments are completed on the same platform.

TABLE I

PERFORMANCE COMPARISON BETWEEN BASELINE TRACKERS AND DADIFF. Δ INDICATES PERCENTAGE INCREASES BROUGHT BY DADIFF. PREC. AND SUCC. MEAN THE PRECISION AND THE SUCCESS RATE. DADIFF HAS IMPROVED NIGHTTIME UAV TRACKING PERFORMANCE SIGNIFICANTLY.

Benchmark	NUT-LR			NUT-L			UAVDark70		
Metric	Succ.	Norm. Prec.	Prec.	Succ.	Norm. Prec.	Prec.	Succ.	Norm. Prec.	Prec.
SiamGAT	0.508	0.569	0.740	0.412	0.474	0.540	0.487	0.583	0.655
DaDiff-GAT	0.531	0.585	0.765	0.449	0.508	0.584	0.511	0.610	0.698
$\Delta_{\mathrm{GAT}}(\%)$	**+4.5**	**+2.8**	**+3.4**	**+9.0**	**+7.2**	**+8.1**	**+4.9**	**+4.6**	**+6.6**
SiamBAN	0.507	0.540	0.753	0.350	0.398	0.466	0.489	0.570	0.677
DaDiff-BAN	0.538	0.574	0.777	0.401	0.450	0.526	0.518	0.601	0.706
$\Delta_{\mathrm{BAN}}(\%)$	**+6.1**	**+6.3**	**+3.2**	**+14.6**	**+13.1**	**+12.9**	**+5.9**	**+5.4**	**+4.3**

TABLE II

COMPARISON OF THE ONE-STEP ADAPTATION PARADIGM AND DADIFF. NORM., PREC., SUCC., DA, AND Δ INDICATE THE NORMALIZATION, THE PRECISION, THE SUCCESS RATE, THE DOMAIN ADAPTATION, AND THE PERCENTAGE INCREASE, RESPECTIVELY. THE TRACKER WITH DADIFF ACHIEVES SUPERIOR TRACKING PERFORMANCE IN ALL NIGHTTIME UAV TRACKING BENCHMARKS.

Benchmark	NUT-LR			NUT-L			UAVDark70		
Metric	Succ.	Norm. Prec.	Prec.	Succ.	Norm. Prec.	Prec.	Succ.	Norm. Prec.	Prec.
SOTA one-step DA [12]	0.517	0.562	0.764	0.377	0.434	0.498	0.510	0.597	0.702
DaDiff-BAN	**0.538**	**0.574**	**0.777**	**0.401**	**0.450**	**0.526**	**0.518**	**0.601**	**0.706**
$\Delta(\%)$	**+4.1**	**+2.1**	**+1.7**	**+6.4**	**+3.7**	**+5.6**	**+1.6**	**+0.7**	**+0.6**

1) Overall performance: **UAVDark70.** As illustrated in the top row of Fig. 6, DaDiff trackers raise the performance of SiamBAN (0.489) and SiamGAT (0.487) by 5.9% and 4.9%. It can be demonstrated that the proposed method has improved the tracking performance of the baseline trackers against different nighttime tracking challenges.

NUT-L. Results in the second row of Fig. 6 show the proposed DaDiff-BAN and DaDiff-GAT consistently achieve satisfactory results. Apart from LR objects, the proposed method can improve the tracking performance in various long-term nighttime scenes significantly. In success rate, DaDiff improves the baseline trackers by over 9%.

NUT-LR. As indicated in the third row of Fig. 6, DaDiff-BAN and DaDiff-GAT rank first two places with a large margin compared to their Baselines. A performance comparison of DaDiff and baseline trackers is reported in TABLE I. In success rate, DaDiff-BAN (0.538) and DaDiff-GAT (0.531) raise the original SiamBAN (0.507) and SiamGAT (0.508) by 6.1% and 4.5%, respectively.

Remark 8: The improvement brought by DaDiff attests to the efficacy of the proposed diffusion models-based alignment framework, particularly for tracking LR objects.

2) Comparison with one-step adaptation paradigm: To prove the alignment effect and robustness of the proposed DaDiff compared with the one-step adaptation paradigm, the previous SOTA method [12] is used for evaluation. SiamBAN [8] is selected as the baseline tracker. As shown in TABLE II, DaDiff is superior in all benchmarks. The superior results prove that DaDiff is competent for feature alignment in nighttime UAV tracking, especially for tracking LR objects.

Remark 9: To be fair, the compared previous SOTA approach employs the official code's pre-trained model.

3) Attribute-based evaluation: To exhaustively evaluate DaDiff when facing LR objects with various challenges, attribute-based comparisons are conducted on NUT-LR, as shown in TABLE III. The trackers with DaDiff achieve superior performance in comparison with other top 4 trackers. Specifically, DaDiff significantly improves the performance of tracking LR objects in attributes of ARC, SV, and IV. The satisfactory results demonstrate that DaDiff can gradually upgrade the detail information of LR objects in adverse illumination conditions.

TABLE III

ATTRIBUTE-BASED EVALUATION OF TOP 6 TRACKERS ON NUT-LR. THE BEST TWO PERFORMANCES ARE RESPECTIVELY HIGHLIGHTED IN THE RED AND GREEN COLORS. THE TRACKERS WITH DADIFF HAVE IMPROVED THE TRACKING PERFORMANCE OF ORIGINAL TRACKERS IN DIFFERENT ATTRIBUTES.

Attributes	ARC		SV		IV	
Trackers	Succ.	Norm. Prec.	Succ.	Norm. Prec.	Succ.	Norm. Prec.
TCTrack++	0.451	0.472	0.538	0.582	0.502	0.542
SiamRPN++	0.478	0.512	0.543	0.598	0.499	0.550
SiamGAT	0.480	0.530	0.556	**0.632**	0.516	0.579
SiamBAN	0.483	0.512	0.547	0.585	0.508	0.541
DaDiff-GAT	**0.504**	**0.545**	**0.586**	**0.653**	**0.537**	**0.594**
DaDiff-BAN	**0.506**	**0.538**	**0.581**	0.623	**0.542**	**0.581**

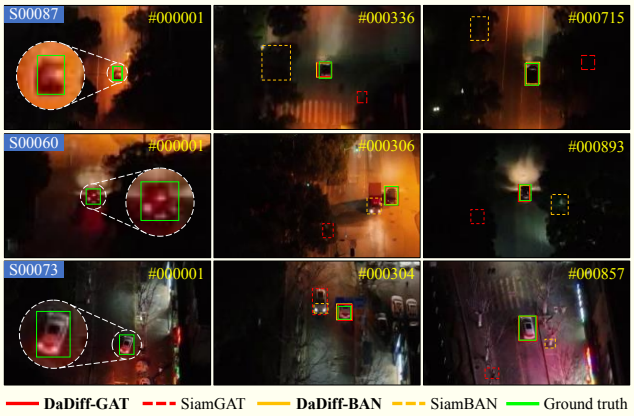

Fig. 7. Qualitative comparison of SOTA trackers with and without DaDiff on NUT-LR. The cropped original image frames are blurred due to LR objects. When the baseline trackers lose nighttime objects, DaDiff improves the perception ability of trackers to successfully track these LR objects.

4) Qualitative evaluation: Some typical nighttime tracking scenes and the performance of SOTA trackers in NUT-LR are displayed in Fig. 7. The baseline trackers fail to focus on LR objects in poor lighting, whereas DaDiff significantly improves the baseline trackers' capacity for nighttime perception, producing satisfactory nighttime tracking performance.

D. Ablation study

To verify the effectiveness of the proposed framework, comprehensive ablation studies are presented in this subsection. SiamBAN [8] and NUT-LR are chosen as the Baseline and the evaluation benchmark, respectively. For clarity, we first introduce various modules, including alignment encoder (AE), tracking-oriented layer (TL), and successive distribution discriminator (SD). The results on TABLE IV show that AE slightly promotes nighttime tracking, with a slight upgrade in success rate but degradation in precision. It indicates that a lack of TL for bridging the feature alignment with the tracking tasks can lead to task mismatch. Therefore, AE can hardly learn the data distribution suitable for nighttime UAV tracking. When employing TL, performance on the nighttime tracking scenes obtains a 3.8% boost in success rate, which verifies the effectiveness of the proposed tracking-oriented layer. Furthermore, SD increases the promotion brought by AE due to the successive feature distribution discrimination process. The results verify that the proposed various modules fairly assist the tracker in generating discriminative features from nighttime images, especially for LR object challenges.

TABLE IV

ABLATION STUDY OF VARIOUS PARTS OF THE PROPOSED FRAMEWORK ON NUT-LR. Δ SYMBOLIZES THE IMPROVEMENT OVER BASELINE. PREC. AND SUCC. REPRESENT THE PRECISION AND THE SUCCESS RATE RESPECTIVELY.

AE	TL	SD	Prec.	$\Delta_{prec}(\%)$	Succ.	$\Delta_{succ}(\%)$
-	-	-	0.640	-	0.426	-
✓	-	-	0.638	-0.3	0.431	+1.2
✓	-	✓	0.654	+2.2	0.439	+3.1
✓	✓	-	0.658	+2.8	0.442	+3.8
✓	✓	✓	**0.677**	**+5.8**	**0.452**	**+6.1**

VI. Conclusion

This work proposes a novel progressive alignment paradigm, named domain-aware diffusion model (DaDiff), for nighttime UAV tracking, especially handling LR object challenges. Specifically, an alignment encoder is developed to enhance the detail information of LR objects weakened by adverse illumination conditions. A tracking-oriented layer is developed to achieve close collaboration with the tracking tasks. To ensure the stability of the alignment effect, a successive distribution discriminator is applied for distinguishing the different feature distributions at each diffusion timestep. Detailed evaluation on nighttime tracking benchmarks shows the effectiveness and feature alignment ability of DaDiff. To summarize, we are confident that this work can contribute to the advancement of visual tracking at night and in other challenging environments, especially for LR objects.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 62173249), the Natural Science Foundation of Shanghai (No. 20ZR1460100), and the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative.

References

[1] J. Shao, B. Du, C. Wu, and L. Zhang, "Tracking Objects From Satellite Videos: A Velocity Feature Based Correlation Filter," *IEEE TGRS*, vol. 57, no. 10, pp. 7860–7871, 2019.

[2] M. Thomas, C. Kambhamettu, and C. A. Geiger, "Motion Tracking of Discontinuous Sea Ice," *IEEE TGRS*, vol. 49, no. 12, pp. 5064–5079, 2011.

[3] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "AutoTrack: Towards High-Performance Visual Tracking for UAV With Automatic Spatio-Temporal Regularization," in *CVPR*, 2020, pp. 11 920–11 929.

[4] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild," *IEEE TPAMI*, vol. 43, no. 5, pp. 1562–1577, 2021.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014, pp. 740–755.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, pp. 211–252, 2015.

[7] D. Guo, Y. Shao, Y. Cui, Z. Wang, L. Zhang, and C. Shen, "Graph Attention Tracking," in *CVPR*, 2021, pp. 9538–9547.

[8] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese Box Adaptive Network for Visual Tracking," in *CVPR*, 2020, pp. 6667–6676.

[9] J. Ye, C. Fu, Z. Cao, S. An, G. Zheng, and B. Li, "Tracker Meets Night: A Transformer Enhancer for UAV Tracking," *IEEE RA-L*, vol. 7, no. 2, pp. 3866–3873, 2022.

[10] J. Ye, C. Fu, G. Zheng, Z. Cao, and B. Li, "DarkLighter: Light Up the Darkness for UAV Tracking," in *IROS*, 2021, pp. 3079–3085.

[11] B. Li, C. Fu, F. Ding, J. Ye, and F. Lin, "ADTrack: Target-Aware Dual Filter Learning for Real-Time Anti-Dark UAV Tracking," in *ICRA*, 2021, pp. 496–502.

[12] J. Ye, C. Fu, G. Zheng, D. P. Paudel, and G. Chen, "Unsupervised Domain Adaptation for Nighttime Aerial Tracking," in *CVPR*, 2022, pp. 8886–8895.

[13] L. Van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *JMLR*, vol. 9, no. 11, p. 2579–2605, 2008.

[14] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection," in *CVPR*, 2022, pp. 13 658–13 667.

[15] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-Aware Trident Networks for Object Detection," in *ICCV*, 2019, pp. 6053–6062.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *NeurIPS*, 2020, pp. 6840–6851.

[17] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *NeurIPS*, 2021, pp. 8780–8794.

[18] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," in *ICLR*, 2021, pp. 1–22.

[19] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit Diffusion Models for Continuous Super-Resolution," in *CVPR*, 2023, pp. 10 021–10 030.

[20] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models," *IJON*, vol. 479, pp. 47–59, 2022.

[21] Z. Yue, J. Wang, and C. C. Loy, "ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting," in *NeurIPS*, 2023, pp. 13 294–13 307.

[22] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image Super-Resolution via Iterative Refinement," *IEEE TPAMI*, vol. 45, no. 4, pp. 4713–4726, 2022.

[23] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion Models: A Comprehensive Survey of Methods and Applications," *CSUR*, vol. 56, no. 4, pp. 1–39, 2023.

[24] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models," *IEEE TPAMI*, vol. 44, no. 11, pp. 7327–7347, 2022.

[25] L. Yao, H. Zuo, G. Zheng, C. Fu, and J. Pan, "SAM-DA: UAV Tracks Anything at Night with SAM-Powered Domain Adaptation," *arXiv preprint arXiv:2307.01024*, pp. 1–8, 2023.

[26] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks," in *CVPR*, 2019, pp. 4277–4286.

[27] H. Zuo, C. Fu, S. Li, J. Ye, and G. Zheng, "DeconNet: End-to-End Decontaminated Network for Vision-Based Aerial Tracking," *IEEE TGRS*, vol. 60, pp. 1–12, 2022.

[28] S. Saha, A. Obukhov, D. P. Paudel, M. Kanakis, Y. Chen, S. Georgoulis, and L. Van Gool, "Learning to Relate Depth and Semantics for Unsupervised Domain Adaptation," in *CVPR*, 2021, pp. 8193–8203.

[29] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation," in *CVPR*, 2021, pp. 15 764–15 773.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *NeurIPS*, 2017, pp. 6000–6010.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *CVPR*, 2022, pp. 10 674–10 685.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *ICCV*, 2017, pp. 618–626.

[33] 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 *ECCV*, 2018, pp. 300–317.

[34] H. Zuo, C. Fu, S. Li, J. Ye, and G. Zheng, "End-to-End Feature Decontaminated Network for UAV Tracking," in *IROS*, 2022, pp. 12 130–12 137.

[35] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "Towards Real-World Visual Tracking With Temporal Contexts," *IEEE TPAMI*, vol. 45, no. 12, pp. 15 834–15 849, 2023.

[36] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *ECCVW*, 2016, pp. 850–865.

[37] B. Yan, H. Peng, K. Wu, D. Wang, J. Fu, and H. Lu, "LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search," in *CVPR*, 2021, pp. 15 180–15 189.

[38] I. Sosnovik, A. Moskalev, and A. Smeulders, "Scale Equivariance Improves Siamese Tracking," in *WACV*, 2021, pp. 2764–2773.

[39] W. Hu, Q. Wang, L. Zhang, L. Bertinetto, and P. H. Torr, "SiamMask: A Framework for Fast Online Object Tracking and Segmentation," *IEEE TPAMI*, vol. 45, no. 3, pp. 3072–3089, 2023.

[40] L. Zhang, A. Gonzalez-Garcia, J. V. D. Weijer, M. Danelljan, and F. S. Khan, "Learning the Model Update for Siamese Trackers," in *ICCV*, 2019, pp. 4009–4018.

[41] Z. Zhang and H. Peng, "Deeper and Wider Siamese Networks for Real-Time Visual Tracking," in *CVPR*, 2019, pp. 4586–4595.

[42] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-Aware Anchor-Free Tracking," in *ECCV*, 2020, pp. 771–787.