Rethinking Occlusion Modeling for UAV Tracking

Jian Zhang Xincheng Yu Yi Lin*
National Key Laboratory of Fundamental Science on Synthetic Vision,
College of Computer Science, Sichuan University, Chengdu, 610000, China
ywjianjian@163.com, xinchengyu@alu.scu.edu.cn, yilin@scu.edu.cn

Abstract

Occlusion remains one of the major challenges in UAV tracking, where dynamic viewpoints and complex environments often cause partial or complete visibility loss. Existing transformer-based trackers typically regard occlusion as random information dropout, overlooking its structured and spatially correlated nature in real-world scenes. We rethink occlusion modeling in UAV tracking as a structured process governed by spatial dependencies. Based on this insight, we introduce Clustered Occlusion Modeling (COM) to generate realistic, density-adaptive occlusion patterns that enhance feature robustness under partial visibility. Furthermore, we design Cost-Aware Depth Bias (CADB), which employs a depth-dependent prior to adjust inference depth, yielding better efficiency while maintaining competitive accuracy. Integrating COM and CADB into a unified single-stream transformer framework, termed OCTrack, our tracker achieves robust and efficient UAV tracking in occlusion-prone environments. Extensive experiments on multiple UAV benchmarks validate its effectiveness and demonstrate state-of-the-art performance.

1. Introduction

Unmanned aerial vehicles (UAVs) are increasingly deployed in vision-critical missions such as search and rescue [43], powerline inspection [52], and autonomous delivery [26], where reliable perception and precise control are essential under dynamic and uncertain conditions. With moving viewpoints and long-range observations, UAV visual tracking operates in rapidly changing and cluttered scenes [4, 8, 51], placing stringent demands on the robustness, adaptability, and real-time performance of onboard vision systems. Tracking becomes particularly challenging as targets are small, move abruptly, and are often occluded, breaking spatial-temporal continuity when visual evidence is partially or completely missing. Meanwhile, limited onboard resources make it challenging to balance tracking accuracy and efficiency.

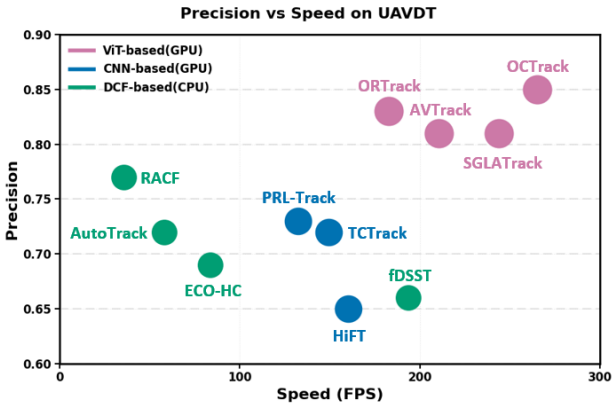

Figure 1. Comparison with state-of-the-art UAV tracking pipelines on the UAVDT dataset. OCTrack sets a new benchmark with 0.85 precision and real-time efficiency at around 265 FPS.

Over the past decade, visual tracking has progressed from discriminative filters to deep learning, culminating in transformer-based architectures that unify feature extraction and interaction. The introduction of transformers has reshaped the representational paradigm of visual tracking [9, 55], enabling joint optimization of feature extraction and interaction within a unified architecture. Single-stream models such as MixFormer [10], OSTrack [59], and DropMAE [49], through the synergy of global dependency modeling and large-scale pretraining, achieve substantial gains in generalization and robustness. Recent studies [30, 50, 54] have extended single-stream architectures to UAV tracking, achieving efficient real-time inference through adaptive token scheduling and dynamic depth control under resource constraints. While achieving notable progress in structural compression and computational efficiency, they still face challenges in handling dynamic occlusion, which often disrupts visual continuity and degrades tracking stability in complex UAV environments. This challenge motivates a rethinking of occlusion modeling in UAV tracking toward a

*Corresponding author

more systematic representation of visibility dynamics.

In this work, we rethink occlusion modeling for UAV tracking. We view occlusion not merely as missing visual evidence, but as a structured and statistical process with spatial dependencies. Recent work [51] introduced occlusion modeling based on spatial Cox processes, which improved robustness to some extent, yet they typically assume continuous intensity fields and fail to capture the spatial clustering widely observed in real-world aerial scenes. In UAV perspectives, occlusions often arise from groups of buildings, vegetation, or moving objects, exhibiting strong local correlation and scale continuity. These observations highlight the importance of modeling occlusion in a way that better reflects its structured behavior in UAV scenes.

Motivated by this observation, we propose a novel occlusion modeling strategy, termed Clustered Occlusion Modeling (COM), to enhance the robustness of UAV trackers against structured and spatially correlated occlusions. COM aims to expose the model to more realistic visibility variations by generating clustered and nonuniform occlusion patterns during training. These patterns simulate the aggregated and locally correlated nature of real-world occlusions, enabling the model to learn semantic consistency under partial visibility. In practice, this is implemented through an Adaptive Cluster Parameterization (ACP) mechanism that adjusts the number, density, and spatial spread of occlusion clusters according to the feature-map size, masking ratio, and occlusion style. Moreover, we design Cost-Aware Depth Bias (CADB), a simple yet effective mechanism to improve inference efficiency in ViT-based tracking. Recognizing that deeper layers incur disproportionately high computational cost while offering diminishing marginal gains, CADB introduces a fixed depth-dependent prior that softly modulates the routing logits according to layer cost. This prior encourages the model to favor shallower, more economical layers when deeper reasoning is not required, effectively encoding a computation-depth trade-off into the inference process. Extensive experiments on multiple UAV tracking benchmarks demonstrate that OCTrack achieves state-of-the-art performance. As shown in Fig. 1, our proposed tracker sets a new record on UAVDT dataset with a precision of 0.85 and a speed of around 265 FPS.

In summary, the main contributions of this work are summarized as follows:

- We propose Clustered Occlusion Modeling (COM), a novel approach that enhances feature robustness in UAV tracking. It captures spatial correlations and simulates realistic occlusion patterns through density-adaptive clustered masking.
- We design Cost-Aware Depth Bias (CADB), an effective mechanism that introduces a depth-dependent prior to regulate inference depth, reducing unnecessary computation while maintaining competitive accuracy.
- We present OCTrack, a unified UAV tracking framework that integrates COM and CADB for robust occlusion modeling and efficient inference, achieving state-of-the-art performance across multiple UAV benchmarks.

2. Related Work

2.1. UAV Visual Tracking

UAV visual tracking builds upon the progress of general object tracking [2, 9–11, 13, 27, 59] and has become a core capability for aerial perception and autonomous navigation. Yet the aerial platform introduces additional challenges caused by camera motion, drastic scale variation, and frequent occlusion. Early UAV trackers [12, 32] based on efficient DCF frameworks achieved real-time speed but often failed in cluttered or dynamic environments. The emergence of deep networks shifted UAV tracking toward feature learning and end-to-end optimization. CNN-based trackers including HiFT [3] and AutoTrack [63] improved robustness and representation capability, but their heavy computational cost restricts onboard deployment. Transformer-based UAV trackers [30, 50, 54] further enhance global feature modeling and efficiency through adaptive token computation or model compression. These architectures unify feature extraction and relation reasoning within a single framework, marking a key step toward generalizable UAV tracking. However, they remain vulnerable to target disappearance or severe occlusion. Most existing UAV trackers [31, 47] still handle occlusion reactively through re-detection or template updates rather than explicitly modeling its spatial correlation. This persistent trade-off underscores the need for principled occlusion modeling to achieve both robust and efficient UAV tracking.

2.2. Occlusion Feature Representation

Occlusion remains a persistent challenge in UAV tracking, where partial visibility and background interference disrupt spatial continuity and weaken feature discrimination, making occlusion representation essential for stable aerial perception. Early studies [6, 24, 36] addressed occlusion using handcrafted descriptors, motion cues, or sensor fusion, but these methods lacked the structural flexibility required to generalize in complex environments. Deep convolutional architectures have been widely explored for learning representations that maintain semantic consistency under partial occlusion [5, 38, 40, 42]. Vision Transformers (ViTs) model global dependencies through self-attention [14, 45], demonstrating potential resilience to occlusion. Masked modeling [20] further showed that ViTs can reconstruct missing content from sparse visible patches, indicating promise for occlusion-aware representation learning. ORTrack [51] introduced Spatial Cox Processes to model occlusion as a stochastic spatial process, capturing the spatially varying

occlusion patterns commonly observed in UAV tracking. Although effective, this approach assumes a fixed statistical distribution of occlusion, limiting its adaptability to dynamic and clustered patterns in UAV tracking. Inspired by real clustered and dynamic occlusions, we rethink occlusion modeling as an adaptive clustered process, enabling the learning of structurally consistent and semantically coherent representations for UAV tracking.

2.3. Efficient Vision Transformers

Vision Transformers (ViTs) [14] have achieved remarkable success in visual recognition by modeling global context through self-attention. However, their quadratic computational complexity and high memory consumption hinder real-time deployment, particularly on resource-constrained platforms. To improve efficiency, many studies have explored lightweight ViTs [17, 45] based on low-rank approximation [46], model compression [7, 44], and hybrid CNN-ViT architectures [48, 53, 58]. Although these designs reduce computational cost, they often compromise representational quality or limit input flexibility. Conditional computation has recently emerged as an efficient inference strategy, allowing ViTs to adapt computational cost to input complexity [41]. Achieving real-time performance in UAV tracking remains challenging because of limited onboard computation and power budgets. Recent efforts have explored efficient inference strategies for ViT-based tracking, including dynamic token pruning [56], adaptive layer activation [54], and background-aware token selection [30] to balance accuracy and speed in UAV tracking. While these strategies effectively reduce computational cost, their dependence on confidence-based control and unstructured computation often results in unstable latency and limited reliability in real-time scenarios. We introduce a cost-aware depth bias that adaptively favors economical layers when additional depth provides diminishing returns, enabling efficient inference for real-time UAV tracking.

3. Method

In this section, we first revisit occlusion modeling in UAV tracking and analyze the limitations of existing masking-based strategies. Then, we rethink occlusion as a structured spatial process and propose the Clustered Occlusion Modeling for robust representation learning. Furthermore, we introduce the Cost-Aware Depth Bias to improve inference efficiency by encouraging depth selection when appropriate. Finally, we present OCTrack, an efficient UAV tracker built upon a ViT-based one-stream architecture, which unifies occlusion modeling and adaptive inference.

3.1. Revisiting Occlusion in UAV Tracking

Masking strategies critically influence how models learn to handle occlusion during training. To evaluate their ability to simulate realistic occlusion patterns, we revisit two representative approaches: random patch masking ($\mathcal{M}_r$) from MAE [20] and spatial Cox masking ($\mathcal{M}_c$), a spatial point process-based strategy [23, 51].

Let the template feature map Z be a tensor of dimension $\mathbb{R}^{C\times H\times W}$, which is partitioned into a grid of $h \times w$ non-overlapping patches. A random score matrix $\mathbf{R} \in [0,1]^{h\times w}$ is then generated, where each element R_{ij} is independently sampled from a uniform distribution. A binary mask $\mathbf{B} \in \{0,1\}^{h\times w}$ is constructed by setting exactly $K = \lfloor \rho\, hw \rfloor$ locations to zero (0 = occluded, 1 = visible), where ρ denotes the masking ratio. The masked template is then defined as

$$\mathcal{M}_r(Z) = Z \odot \mathrm{Up}(\mathbf{B}), \tag{1}$$

where $\odot$ denotes element-wise multiplication and $\mathrm{Up}(\cdot)$ aligns the $h \times w$ binary mask to the $H \times W$ feature resolution through nearest-neighbor upsampling.

To introduce spatial dependency, a finite Cox point process is defined over the template grid domain $\Omega = [1,h] \times [1,w]$. The process first constructs a spatially varying intensity field $\lambda(p)$ that specifies the expected density of occlusion across spatial locations, and then samples occlusion points accordingly. Each position $p \in \Omega$ is assigned a random intensity:

$$\lambda(p) = \lambda_0 \frac{\exp(-\|p-p_0\|^2/2\sigma^2)}{\sum_{q\in\Omega}\exp(-\|q-p_0\|^2/2\sigma^2)}, \tag{2}$$

where p_0 denotes the template center, σ controls the spatial spread, and $\lambda_0 \sim \mathrm{Poisson}(\eta)$ defines the expected number of occluded patches. Analogously, the masked template is obtained as $\mathcal{M}_c(Z) = Z \odot \mathrm{Up}(\mathbf{B}')$, where $\mathbf{B}'$ is sampled from the above Cox-process-based distribution. While both $\mathcal{M}_r$ and $\mathcal{M}_c$ rely on stochastic sampling to emulate occlusion, they fail to reflect the clustered and structured nature of real-world visibility loss, which often occurs in contiguous spatial regions. This limitation motivates a rethinking of occlusion modeling toward adaptive and physically grounded representations.

3.2. Rethinking Occlusion in UAV Tracking

Existing masking-based approaches rely on stochastic sampling to simulate occlusion, either independently across patches or through predefined spatial correlations. Although such randomization regularizes feature learning, it oversimplifies the physical nature of occlusion in aerial scenarios. This mismatch is especially pronounced in low-altitude, high-speed UAV tracking, where targets are typically small, background clutter is severe, and occlusions predominantly arise from large, static scene structures such as buildings, trees, or vehicles. Consequently, real-world visibility loss exhibits strong spatial coherence and semantic clustering, which are poorly captured by conventional

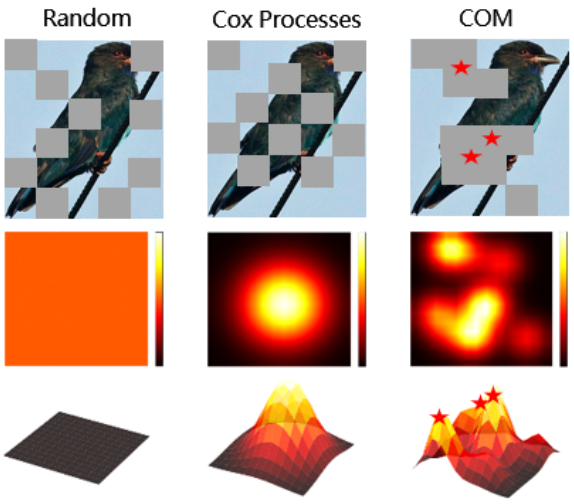

Figure 2. Comparison of occlusion modeling strategies. Random masking produces unstructured visibility loss, while Cox process based masking introduces limited spatial correlation. The proposed COM generates clustered and density-adaptive occlusion patterns, and the lower rows show the corresponding intensity and spatial distributions.

random masking schemes. As illustrated in Fig. 2, random masking produces unstructured visibility loss, while Cox process based masking introduces limited spatial correlation, both of which fail to capture the clustered nature of real-world occlusions. These structured occlusions typically occur in contiguous spatial regions, simultaneously disrupting both the target appearance and the surrounding contextual cues, which often leads to inconsistent feature representations and unstable localization. To overcome these limitations, occlusion modeling should evolve beyond handcrafted randomness toward an adaptive process that captures the spatial continuity and contextual dependencies inherent in real-world UAV scenes. This perspective motivates the development of a learnable occlusion mechanism capable of generating controllable, clustered, and semantically consistent masks, thereby bridging the gap between synthetic occlusion simulation and the physical characteristics of aerial visibility loss.

We propose a Clustered Occlusion Modeling (COM) strategy that simulates occlusion as a structured spatial process rather than independent random masking. Given a template feature map $Z \in \mathbb{R}^{C\times H\times W}$ and a masking ratio $\rho \in (0,1)$, the number of masked patches is defined as $K = \lfloor \rho\, HW \rfloor$. Unlike prior random or Cox-process-based masking that rely on a fixed and handcrafted intensity field, COM formulates an adaptive clustered spatial process with controllable clustering and adaptive density. Specifically, an Adaptive Cluster Parameterization (ACP) mechanism is introduced, parameterized by $\mathcal{P} = (\lambda_p, \mu, \sigma)$, where λ_p determines the expected number of occlusion clusters, μ controls the average occlusion strength within each cluster, while σ regulates the spatial dispersion. These parameters jointly define the occlusion configuration over the spatial domain $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$, modeled as a mixture of Gaussian fields:

$$\pi(p) \propto \sum_{i=1}^{N_p} w_i \exp\left(-\frac{\|p - c_i\|^2}{2\sigma^2}\right), \tag{3}$$

where $N_p \sim \text{Poisson}(\lambda_p HW)$ denotes the number of clusters, c_i denote the sampled cluster centers, and $w_i \sim \text{Poisson}(\mu)$ specifies the relative occlusion strength. Based on the intensity field $\pi(p)$, occlusion points are generated through a clustered sampling process, where each cluster i yields w_i occluded points drawn from a Gaussian distribution centered at c_i:

$$p_{i,j} \sim \mathcal{N}(c_i, \sigma^2 I_2), \quad j = 1,\dots,w_i. \tag{4}$$

The union of all sampled points $\{p_{i,j}\}$ forms the final occlusion mask $\mathcal{M}$, and the mask ratio ρ is maintained by randomly subsampling points if necessary.

By tuning these parameters, COM adaptively transitions between sparse, balanced, and dense occlusion regimes, corresponding to variations in cluster count, per-cluster intensity, and spatial compactness. This parametric flexibility enables the generation of occlusion distributions with controllable granularity and structural coherence, closely reflecting the heterogeneous and spatially correlated nature of real-world UAV occlusions. To achieve invariance under structured visibility degradation, COM introduces an occlusion-robust representation objective. It aligns the latent features of the original and masked templates. Given a clustered-masked version $Z' = \mathcal{M}_{\text{com}}(Z)$, the model minimizes their representational discrepancy:

$$\mathcal{L} = \left\| t_Z^{(L)} - t_{Z'}^{(L)} \right\|_2^2, \tag{5}$$

where $t_Z^{(L)}$ and $t_{Z'}^{(L)}$ denote the extracted feature representations of Z and Z', respectively. This regularization preserves semantic consistency and suppresses appearance drift caused by clustered occlusion, leading to representations robust to spatially correlated visibility loss. Since it operates only during training, COM adds no computational overhead at inference and can be seamlessly integrated with any ViT-based tracker.

3.3. Cost-Aware Depth Bias

UAV tracking requires high-precision performance under strict computational and real-time constraints. While Vision Transformer (ViT)-based trackers possess strong representational capability, their computational cost grows linearly

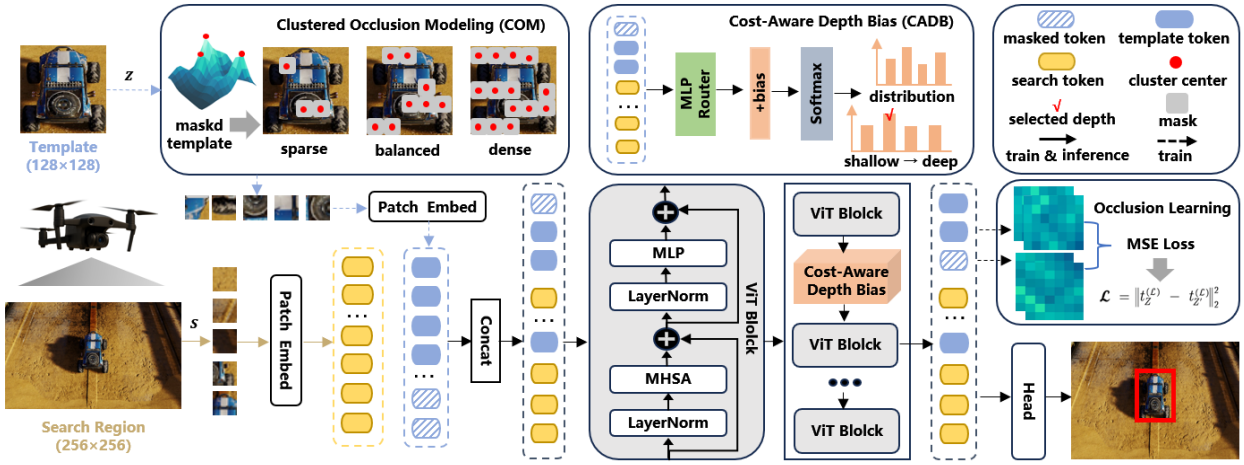

Figure 3. Overview of the proposed OCTrack framework. It consists of a single-stream transformer backbone and a prediction head. Clustered Occlusion Modeling generates realistic occlusion patterns to enhance feature robustness, while Cost-Aware Depth Bias adjusts the routing tendency of the network by introducing a computation-sensitive prior, guiding inference toward more economical depths while preserving robustness.

with network depth, and redundant features tend to emerge in later encoder layers [54]. Existing dynamic inference strategies [50, 54, 56] improve efficiency by dynamically activating, skipping, or selecting layers guided by task relevance or inter-layer similarity, yet they lack explicit modeling of computational preference, limiting adaptability to varying scene complexity. To address this, we propose the Cost-Aware Depth Bias (CADB), which formulates adaptive layer selection as a bias-regulated routing process with a computation-sensitive prior. The depth-dependent bias promotes shallower inference, while alignment-based supervision naturally guides the router toward deeper layers when larger feature discrepancies are observed.

Specifically, we employ a MLP router to estimate the importance of subsequent encoder layers. The router consists of two fully connected layers with a ReLU activation in between. It takes the feature of the first token from the fused template-search sequence at the reference layer i_{ref} as input, serving as a compact global descriptor summarizing the joint context. The router outputs a layer-importance score vector $y = f(x) \in \mathbb{R}^L$, where L denotes the number of candidate layers following i_{ref}. For each encoder layer following the reference layer i_{ref}, a depth-dependent bias encodes computational preference:

$$b_i = \kappa\,[1 - 2(i - i_{\text{ref}})/(L-1)], \tag{6}$$

where $\kappa > 0$ controls the strength of shallow-layer preference. The final routing probability is computed as $p = \text{softmax}((y+b)/\tau)$, with temperature $\tau > 0$ adjusting the sharpness of the selection distribution.

During training, pseudo-labels are derived from the inter-layer alignment signal between the reference layer $F_{i_{\text{ref}}}$ and subsequent features F_i. The alignment score is defined as $s_i = (F_{i_{\text{ref}}} \cdot F_i)/(\|F_{i_{\text{ref}}}\|_2 \|F_i\|_2)$, where $F_{i_{\text{ref}}}$ serves as a fixed anchor representing the baseline feature representation. The layer with the highest alignment score, $i^\star = \arg\max_i s_i$, is regarded as the representative layer, and its index is used for supervision. We denote the corresponding one-hot label as t_i, where $t_{i^\star} = 1$ and zeros elsewhere. The router is optimized by minimizing the cross-entropy loss over the bias-corrected logits:

$$\mathcal{L}_{\text{route}} = -\sum_{i=1}^{L} t_i \log \frac{\exp(y_i + b_i)}{\sum_{j=1}^{L} \exp(y_j + b_j)}. \tag{7}$$

At inference, CADB determines the execution depth from bias-corrected routing scores, activating only the representative layer and bypassing the others to reduce effective computation, while requiring no additional runtime mechanisms.

3.4. OCTrack for UAV Tracking

Overall Pipeline. The proposed OCTrack integrates Clustered Occlusion Modeling (COM) and Cost-Aware Depth Bias (CADB) within a unified single-stream transformer architecture for UAV tracking, as shown in Fig. 3. Given a target template $Z \in \mathbb{R}^{3\times H_z \times W_z}$, its clustered-masked counterpart $Z' = \mathcal{M}_{\text{com}}(Z)$ generated by the proposed COM, and a search region $S \in \mathbb{R}^{3\times H_s \times W_s}$, COM simulates realistic occlusion patterns with controllable spatial densities to enhance representation robustness. The masked and unmasked templates are concatenated with the search region and jointly embedded into patch tokens, which are processed by a ViT-based backbone to learn occlusion-

	Method	Source	DTB70		UAVDT		VisDrone2018		UAV123		Avg.		Avg.FPS		FLOPs	Param
			P(%)	AUC(%)	P(%)	AUC(%)	P(%)	AUC(%)	P(%)	AUC(%)	P(%)	AUC(%)	GPU	CPU	(GMac)	(M)
DCF-based	KCF [21]	TPAMI'15	46.8	28.0	57.1	29.0	68.5	41.3	52.3	33.1	56.2	32.9	–	624.3	–	–
	fDSST [12]	TPAMI'17	53.4	35.7	66.6	38.3	69.8	51.0	58.3	40.5	62.0	41.4	–	193.4	–	–
	ECO_HC [13]	CVPR'17	63.5	44.8	69.4	41.6	80.8	58.1	71.0	49.6	71.2	48.5	–	83.5	–	–
	AutoTrack [32]	CVPR'20	71.6	47.8	71.8	45.0	78.8	57.3	68.9	47.2	72.8	49.3	–	57.8	–	–
	RACF [29]	PR'22	72.6	50.5	77.3	49.4	83.4	60.0	70.2	47.7	75.9	51.8	–	35.6	–	–
CNN-based	HiFT [3]	ICCV'21	80.2	59.4	65.2	47.5	71.9	52.6	78.7	59.0	74.0	54.6	160.3	–	7.2	9.9
	TCTrack [4]	CVPR'22	81.2	62.2	72.5	53.0	79.9	59.4	80.0	60.5	78.4	58.8	149.6	–	8.8	9.7
	SGDViT [57]	ICRA'23	78.5	60.4	65.7	48.0	72.1	52.1	75.4	57.5	72.9	54.5	110.5	–	11.3	23.3
	DRCI [61]	ICME'23	81.4	61.8	84.0	59.0	83.4	60.0	76.7	59.7	81.4	60.1	281.3	62.7	3.6	8.8
	PRL-Track [18]	IROS'24	79.5	60.6	73.1	53.5	72.6	53.8	79.1	59.3	76.1	56.8	132.3	–	7.4	12.0
ViT-based	AbaViTrack [30]	ICCV'23	85.9	66.4	83.4	59.9	86.1	65.3	86.4	66.4	85.5	64.5	167.5	48.6	2.4	8.0
	SMAT [19]	WACV'24	81.9	63.8	80.8	58.7	82.5	63.4	81.8	64.6	81.8	62.6	102.5	–	3.2	8.6
	AVTrack [50]	ICML'24	84.3	65.0	82.1	58.7	86.0	65.3	84.8	66.8	84.2	63.8	210.5	68.7	1.9	7.9
	ORTrack [51]	CVPR'25	86.2	66.4	83.4	60.1	88.6	66.8	84.3	66.4	85.6	65.0	182.6	65.3	2.4	7.9
	SGLATrack [54]	CVPR'25	84.4	65.1	81.9	59.9	–	–	84.9	66.9	–	–	243.9	75.6	1.7	5.8
	OCTrack-b	–	85.7	66.0	85.0	63.0	88.5	67.0	84.9	66.9	86.0	65.7	265.2	82.5	1.7-2.4	5.8-8
	OCTrack-d	–	83.6	64.6	81.8	59.9	87.0	66.3	85.2	67.7	84.4	64.6	/	/	/	/
	OCTrack-s	–	84.8	65.7	84.2	61.8	84.9	65.2	82.8	64.7	84.2	64.4	/	/	/	/

Table 1. Comparison of precision (P), AUC, and speed (FPS) between OCTrack-DeiT and lightweight trackers on DTB70 [28], UAVDT [15], VisDrone2018 [62], and UAV123 [1]. Red, blue, and green highlight the best, second, and third performance. Note that the slash mark (/) indicates results identical to those reported above, while OCTrack-b, OCTrack-d, and OCTrack-s denote the balanced, dense, and sparse occlusion modeling settings, respectively.

invariant representations. During inference, CADB adaptively adjusts the network depth by activating only the most informative transformer layers, maintaining a balance between accuracy and efficiency. Finally, the fused representations are decoded by a lightweight tracking head to produce precise target localization under UAV constraints.

Head and Loss. We employ a compact center-based prediction head following the general design in [10, 59]. It is constructed with several Conv-BN-ReLU layers that transform the search representation into spatial predictions. Given the reshaped search feature map, the head estimates three quantities simultaneously: a classification confidence map $p \in [0,1]^{H_s/P \times W_s/P}$, a local position offset $o \in [0,1]^2 \times H_s/P \times W_s/P$, and a normalized bounding box size $s \in [0,1]^2 \times H_s/P \times W_s/P$. For the tracking objective, we adopt a weighted focal loss [34] for classification and a combination of L1 loss and GIoU loss [42] for bounding box regression. The overall training loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \gamma\mathcal{L}_{\text{route}}, \tag{8}$$

where $\lambda_{\text{iou}} = 2$, $\lambda_{\text{L1}} = 5$, and $\gamma = 0.1$ are the weighting coefficients for the regularization terms.

4. Experiment

4.1. Implementation Details

Our tracking framework is implemented in Python 3.8 using PyTorch 1.10. For a fair runtime comparison, FPS of all ViT-based trackers was re-evaluated on our hardware (Intel i7-12700K, NVIDIA RTX 4060 Ti). FPS of non-ViT-based baselines follows the benchmark [51], as their runtime settings are not fully reproducible. We use three lightweight ViT variants, including DeiT-Tiny [45], ViT-Tiny [14], and Eva-Tiny [17], as backbones to construct corresponding tracker variants, namely OCTrack-DeiT, OCTrack-ViT, and OCTrack-Eva. The sizes of the template and search images are set to 128×128 and 256×256, respectively. We define three occlusion styles: *dense* (30 clusters, small σ), *balanced* (10 clusters, medium σ), and *sparse* (3 clusters, large σ), where σ scales with $\min(H, W)$. These styles correspond to small, medium, and large occlusions. We train OCTrack on a combination of GOT-10k [22], LaSOT [16], COCO [33], and TrackingNet [37]. The batch size is set to 32, and the AdamW [35] optimizer is used with a weight decay of 1×10^{-4} and an initial learning rate of 4×10^{-5}.

4.2. State-of-the-art Comparison

Comparison with Lightweight Trackers. Table 1 presents a comprehensive comparison between the proposed OCTrack and several representative lightweight trackers across four widely used UAV benchmarks. OCTrack achieves an excellent balance among accuracy, robustness, and efficiency. Compared with traditional DCF-based and CNN-based trackers, its advantage remains evident: RACF [29] records 75.9% precision and 51.8% AUC, while DRCI [61], representing the CNN category, attains 81.4% and 60.1%,

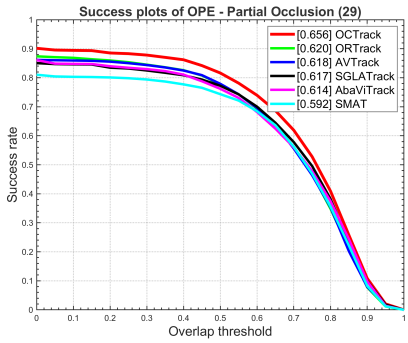

Figure 4. Success plot of OPE (One-Pass Evaluation) on UAVDT under the Partial Occlusion attribute.

respectively. In contrast, all OCTrack variants consistently exceed 84% in precision and 64% in AUC on average, demonstrating stronger occlusion robustness and more stable feature representation under complex aerial conditions. Within the ViT-based family, OCTrack ranks consistently among the top three in most evaluation metrics while maintaining competitive efficiency in both inference speed and model complexity. Aba-ViTrack [30] achieves comparable accuracy (85.5% / 64.5%) but runs slower (167 FPS), while SGLATrack [54] offers moderate speed (243.9 FPS) with slightly weaker robustness and generalization. Across multiple UAV benchmarks, OCTrack remains within the top three performers, achieving a well-balanced compromise between accuracy and efficiency, recording 86.0% average precision and 265.2 FPS on GPU, highlighting its suitability for real-time UAV tracking. Moreover, Table 1 also reports the results of three OCTrack variants, which correspond to the balanced, dense, and sparse occlusion modeling settings, respectively. Among them, OCTrack-b delivers the most balanced overall performance, achieving the best results on UAVDT (85.0% / 63.0%) and strong results on VisDrone2018 (best AUC 67.0% with near-best precision 88.5%). In particular, UAVDT achieves the best scores on both metrics, indicating that OCTrack-b maintains stable feature discrimination under moderate occlusion and dynamic viewpoint variations. OCTrack-d further validates the effectiveness of the proposed occlusion modeling strategy. Both UAV123 and VisDrone2018 contain extensive sequences with continuous occlusions and frequent object interactions. UAV123 emphasizes long-term visibility preservation under dynamic motion, while VisDrone2018 features higher object density and complex structural backgrounds. In these representative scenarios, OCTrack-d achieves the highest AUC of 67.7% on UAV123 and 87.0% / 66.3% on VisDrone2018, demonstrating that dense occlusion modeling effectively enhances the tracker's robustness in visibility recovery, feature aggregation, and cross-frame consistency, and further confirming the generalization capability of the proposed method in complex aerial environments.

Attribute-Based Evaluation. To further evaluate the robustness of the proposed method under occlusion, we conduct an attribute-based analysis on the UAVDT dataset, focusing on the subset with partial occlusion. Other attribute results are included in the supplementary material. For a fair and meaningful comparison, we evaluate OCTrack-DeiT against representative ViT-based trackers [19, 30, 50, 51, 54], as they share a similar paradigm. We adopt the configuration OCTrack-b in this analysis As shown in Fig. 4, OCTrack-DeiT achieves the highest AUC of 0.656 on the partial occlusion subset, surpassing the second-best method by approximately three percentage points. These results demonstrate that our method captures occlusion dynamics more adaptively and maintains stable representation and tracking continuity in complex aerial scenes.

Table 2. Effect of COM and CADB on the baseline trackers.

Tracker	COM	CADB	UAVDT		FPS
			P(%)	AUC(%)	
OCTrack-ViT			79.7	58.8	186.3
	✓		$83.7^{\uparrow 4.0}$	$62.0^{\uparrow 3.2}$	–
	✓	✓	$84.2^{\uparrow 4.5}$	$61.6^{\uparrow 2.8}$	**248.8**$^{\uparrow 34\%}$
OCTrack-Eva			77.6	57.3	224.5
	✓		$81.2^{\uparrow 3.6}$	$60.1^{\uparrow 2.8}$	–
	✓	✓	$80.6^{\uparrow 3.0}$	$59.9^{\uparrow 2.6}$	**280.6**$^{\uparrow 25\%}$
OCTrack-DeiT			80.4	59.6	197.6
	✓		$85.2^{\uparrow 4.8}$	$62.9^{\uparrow 3.3}$	–
	✓	✓	$85.0^{\uparrow 4.6}$	$63.0^{\uparrow 3.4}$	**265.2**$^{\uparrow 34\%}$

4.3. Ablation Study and Analysis

Component-wise ablations. Table 2 presents the results on UAVDT as the proposed Clustered Occlusion Modeling (COM) and Cost-Aware Depth Bias (CADB) are progressively integrated into the baselines. All ablation experiments are performed with the balanced variant, OCTrack-b. Since COM operates only during training, its variants share the same inference speed as the baseline; thus, FPS is reported only after CADB is added. COM consistently improves performance across all backbones, boosting precision by **+4.0%**, **+3.6%**, and **+4.8%**, and AUC by **+3.2%**, **+2.8%**, and **+3.3%** on OCTrack-ViT, OCTrack-Eva, and OCTrack-DeiT, respectively. These results indicate that spatially clustered masking effectively regularizes feature learning under occlusion, leading to more robust and discriminative representations in complex aerial scenes. After integrating CADB, GPU throughput increases by **+34%**, **+25%**, and **+34%** for ViT, Eva, and DeiT, while accuracy remains nearly unchanged. CADB introduces a depth-aware bias that steers the layer-selection behavior toward shallower inference on easy frames while still enabling deeper reasoning when necessary. This bias-guided adjustment reduces the average inference depth while maintaining competitive accuracy. Together, COM and CADB balance tracking robustness and inference efficiency, making OC-

Table 3. Masking strategies integrated into OCTrack-DeiT and their performance on VisDrone2018.

Tracker	Masking Strategy	AUC (%)
OCTrack-DeiT	MAE [20]	62.9
	SAM [25]	64.8
	Cox Process [51]	65.9
	AdAutoMix [39]	63.2
	CutMix [60]	63.6
	COM	**67.0**

Table 4. Effect of bias strength κ on routing behavior and tracking performance, measured by AUC on DTB70 and UAV123 datasets.

κ	ASLI ↓	DTB70	UAV123	FPS ↑
0.0	10	65.9	66.9	197.6
0.1	9	64.8	66.2	230.4
0.3	8	66.0	66.9	265.2
0.5	7	65.1	66.0	278.4

Track well suited for real-world UAV tracking scenarios.

Effect of Masking Strategies. To evaluate the impact of different masking mechanisms, we test OCTrack-DeiT with common masking operators (Table 3). Random masking [20] yields the lowest AUC (62.9%), indicating that independent patch dropout cannot model structured visibility loss in UAV scenes. Spatial Cox process masking [51] improves AUC to 65.9% by introducing spatial correlation, but its fixed intensity field limits adaptability. Semantic and region-level augmentations such as SAM [25], AdAutoMix [39], and CutMix [60] bring moderate gains (63–65%), yet lack explicit spatial clustering. The proposed **COM** achieves the highest AUC of 67.0%, demonstrating that clustered masking improves feature robustness and preserves stable representations under complex aerial scenes.

Effect of bias strength. Table 4 summarizes how different bias strengths κ affect routing behavior and tracking accuracy. Since the first six layers serve as feature encoding, CADB influences routing only in the subsequent layers. As κ increases, the routing position consistently shifts toward shallower depths, indicating that the proposed bias provides controllable adjustment of inference depth. Because hard routing executes exactly one depth for each κ, the average selected layer index (ASLI) degenerates to the routed depth under that setting. AUC stays stable while FPS improves, suggesting that shallower inference is often adequate.

Qualitative Comparison and Visualization. For qualitative comparison, Fig. 5 presents results under challenging UAV scenarios, including occlusion, background clutter, pose and scale variations, and illumination changes. OCTrack maintains accurate and consistent localization across frames, delivering smoother temporal behavior when other trackers become unstable. These results demonstrate the robustness of our method in complex UAV environments. Fig. 6 visualizes the attention and feature maps of OCTrack-DeiT with and without COM. With COM, the model focuses more clearly on target regions and produces more spatially consistent feature maps across masking ratios.

Figure 5. Qualitative comparison between the proposed tracker and state-of-the-art methods across three sequences.

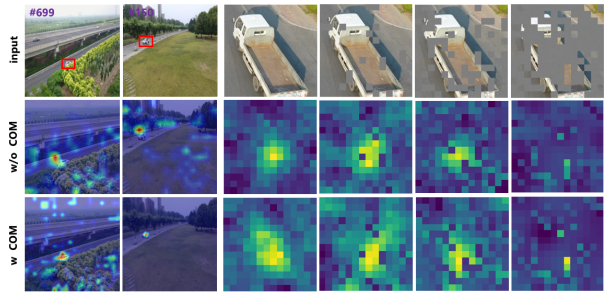

Figure 6. Comparison of attention maps (left two columns) and feature maps (right four columns) generated by OCTrack-DeiT with and without the COM. The first row displays the search images and their masked counterparts at different masking ratios (0%, 10%, 30%, and 70%), while the second and third rows illustrate the corresponding attention and feature distributions.

5. Conclusion

In this work, we rethink the approach to occlusion modeling in UAV tracking by treating occlusion as a spatially correlated phenomenon. We propose Clustered Occlusion Modeling (COM), which generates density-adaptive, spatially clustered occlusion patterns, allowing the model to learn robust feature representations under partial visibility. We also design Cost-Aware Depth Bias (CADB), an effective mechanism that introduces a depth-dependent prior to regulate inference depth, reducing unnecessary computation while maintaining competitive accuracy. Extensive experiments on multiple UAV tracking benchmarks demonstrate the effectiveness of our approach, providing a more robust solution for occlusion modeling in aerial scenarios.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants 62371323, U2333209, and U2433217; the Sichuan Science and Technology Program under Grants 2024YFG0010 and 2024ZDZX0046; and the Institutional Research Fund from Sichuan University under Grant 2024SCUQJTX030.

References

[1] U Benchmark. A benchmark and simulator for uav tracking. In *European conference on computer vision*, 2016. 6

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 2

[3] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Hift: Hierarchical feature transformer for aerial tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15457–15466, 2021. 2, 6

[4] Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. Tctrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14798–14808, 2022. 1, 6

[5] Satyaki Chakraborty and Martial Hebert. Learning to track object position through occlusion. *arXiv preprint arXiv:2106.10766*, 2021. 2

[6] T-H Chang and Shaogang Gong. Tracking multiple people with a multi-camera system. In *Proceedings 2001 ieee workshop on multi-object tracking*, pages 19–26. IEEE, 2001. 2

[7] Arnav Chavan, Zhiqiang Shen, Zhuang Liu, Zechun Liu, Kwang-Ting Cheng, and Eric P Xing. Vision transformer slimming: Multi-dimension searching in continuous optimization space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4931–4941, 2022. 3

[8] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *European conference on computer vision*, pages 375–392. Springer, 2022. 1

[9] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021. 1, 2

[10] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022. 1, 6

[11] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 4310–4318, 2015. 2

[12] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1561–1575, 2016. 2, 6

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6638–6646, 2017. 2, 6

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 6

[15] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 6

[16] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 6

[17] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023. 3, 6

[18] Changhong Fu, Xiang Lei, Haobo Zuo, Liangliang Yao, Guangze Zheng, and Jia Pan. Progressive representation learning for real-time uav tracking. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5072–5079. IEEE, 2024. 6

[19] Goutam Yelluru Gopal and Maria A Amer. Separable self and mixed attention transformers for efficient object tracking. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6708–6717, 2024. 6, 7

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 8

[21] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014. 6

[22] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 6

[23] Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons, 2008. 3

[24] Michal Irani and Shmuel Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of visual communication and image representation*, 4(4):324–335, 1993. 2

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 8

[26] Woojin Lee, Babar Shahzaad, Balsam Alkouz, and Athman Bouguettaya. Reactive composition of uav delivery services

in urban environments. *IEEE Transactions on Intelligent Transportation Systems*, 25(10):13453–13466, 2024. 1

[27] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 2

[28] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 6

[29] Shuiwang Li, Yuting Liu, Qijun Zhao, and Ziliang Feng. Learning residue-aware correlation filters and refining scale for real-time uav tracking. *Pattern Recognition*, 127:108614, 2022. 6

[30] Shuiwang Li, Yangxiang Yang, Dan Zeng, and Xucheng Wang. Adaptive and background-aware vision transformer for real-time uav tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13989–14000, 2023. 1, 2, 3, 6, 7

[31] Shuiwang Li, Xiangyang Yang, Xucheng Wang, Dan Zeng, Hengzhou Ye, and Qijun Zhao. Learning target-aware vision transformers for real-time uav tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024. 2

[32] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11923–11932, 2020. 2, 6

[33] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[36] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999. 2

[37] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 6

[38] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 2

[39] Huafeng Qin, Xin Jin, Yun Jiang, Mounim A El-Yacoubi, and Xinbo Gao. Adversarial automixup. *arXiv preprint arXiv:2312.11954*, 2023. 8

[40] Delin Qu, Yizhen Lao, Zhigang Wang, Dong Wang, Bin Zhao, and Xuelong Li. Towards nonlinear-motion-aware and occlusion-robust rolling shutter correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10680–10688, 2023. 2

[41] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 3

[42] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 2, 6

[43] David C Schedl, Indrajit Kurmi, and Oliver Bimber. An autonomous drone for search and rescue in forests using airborne optical sectioning. *Science Robotics*, 6(55):eabg1188, 2021. 1

[44] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022. 3

[45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2, 3, 6

[46] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 3

[47] Qingmao Wei, Bi Zeng, Jianqi Liu, Li He, and Guotian Zeng. Litetrack: Layer pruning with asynchronous feature extraction for lightweight and efficient visual tracking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4968–4975. IEEE, 2024. 2

[48] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 3

[49] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14561–14571, 2023. 1

[50] You Wu, Yongxin Li, Mengyuan Liu, Xucheng Wang, Xiangyang Yang, Hengzhou Ye, Dan Zeng, Qijun Zhao, and Shuiwang Li. Learning an adaptive and view-invariant vision transformer for real-time uav tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1, 2, 5, 6, 7

[51] You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, and Shuiwang Li. Learning occlusion-robust vision transformers for real-time uav tracking. In *Proceedings of the Computer Vision and Pattern*

Recognition Conference, pages 17103–17113, 2025. 1, 2, 3, 6, 7, 8

[52] Jiaxu Xing, Giovanni Cioffi, Javier Hidalgo-Carrió, and Davide Scaramuzza. Autonomous power line inspection with drones via perception-aware mpc. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1086–1093. IEEE, 2023. 1

[53] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9981–9990, 2021. 3

[54] Chaocan Xue, Bineng Zhong, Qihua Liang, Yaozong Zheng, Ning Li, Yuanliang Xue, and Shuxiang Song. Similarity-guided layer-adaptive vision transformer for uav tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6730–6740, 2025. 1, 2, 3, 5, 6, 7

[55] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021. 1

[56] Xiangyang Yang, Dan Zeng, Xucheng Wang, You Wu, Hengzhou Ye, Qijun Zhao, and Shuiwang Li. Adaptively bypassing vision transformer blocks for efficient visual tracking. *Pattern Recognition*, 161:111278, 2025. 3, 5

[57] Liangliang Yao, Changhong Fu, Sihang Li, Guangze Zheng, and Junjie Ye. Sgdvit: Saliency-guided dynamic vision transformer for uav tracking. *arXiv preprint arXiv:2303.04378*, 2023. 6

[58] Ting Yao, Yehao Li, Yingwei Pan, and Tao Mei. Hiri-vit: Scaling vision transformer with high resolution inputs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6431–6442, 2024. 3

[59] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pages 341–357. Springer, 2022. 1, 2, 6

[60] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 8

[61] Dan Zeng, Mingliang Zou, Xucheng Wang, and Shuiwang Li. Towards discriminative representations with contrastive instances for real-time uav tracking. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1349–1354. IEEE, 2023. 6

[62] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Haibin Ling, Qinghua Hu, Haotian Wu, Qinqin Nie, Hao Cheng, Chenfeng Liu, et al. Visdrone-vdt2018: The vision meets drone video detection and tracking challenge results. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 6

[63] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. 2