ORIGINAL RESEARCH Open access

Adaptive Spatial–Temporal Context Aggregation for Lightweight UAV Siamese Tracking

Ahmed Youssef[1*], Khaled Hassan[1], Mahmoud Elamin[2]

Abstract

The Siamese tracker achieves outstanding tracking results by leveraging the powerful feature-learning capabilities of deep neural networks. Nonetheless, aerial tracking presents persistent obstacles stemming from intricate scene conditions—encompassing limited image resolution, target occlusion, confusable objects, diminutive targets, size fluctuations, varying aspect ratios, shape distortion, and restricted onboard processing power—that make efficient, reliable tracking difficult. Our work introduces a lightweight yet effective Siamese architecture for aerial tracking, built on adaptive temporal contextual aggregation, incorporating two key components—a parallel atrous module (PAM) and an adaptive temporal context aggregation model (ATCAM)—to address the aforementioned difficulties. The PAM operates by running several atrous convolutional branches at different dilation rates simultaneously, thereby capturing and merging multiscale representations infused with spatial context at a single feature level, thereby substantially bolstering robustness against appearance alterations induced by aspect ratio changes, occlusion, and scale variation. Complementing this, the ATCAM injects temporal context into the current frame adaptively via an encoder-decoder scheme, equipping the tracker to overcome distractors and maintain target identification when extracting fine-grained features is hampered, as occurs at low resolution or among similar-looking objects. Extensive evaluation on the UAV20L, UAV123@10fps, and DTB70 datasets confirms the strong performance of our proposed architecture, achieving a swift frame rate of 75.5 fps or higher on an NVIDIA 3060 Ti GPU.

Keywords Aerial tracking, Atrous convolution, Siamese tracker, Temporal context

*Correspondence:
Ahmed Youssef
ahmed.youssef@uofk.edu
[1] Department of AI for Sorghum and Millet Yield Prediction under Salinity Stress, Faculty of Agriculture, University of Khartoum, Khartoum, Sudan
[2] Department of Smart Drip Fertigation Management, Faculty of Agriculture, Sudan University of Science and Technology, Khartoum, Sudan

Introduction

Visual object tracking is a fundamental problem in computer vision. Given the target specification in the opening frame of a video, the tracker must subsequently pinpoint that target in all subsequent frames using computational methods. Aerial tracking, as a subfield, is significant in remote sensing analytics and serves numerous practical applications, including intelligent surveillance [1, 2], airborne detection [3], and moving object analysis [4, 5]. Accordingly, a substantial body of research is devoted to advancing aerial tracking techniques [6-10].

Two principal paradigms dominate the aerial tracking landscape: those grounded in correlation filters and those built upon Siamese architectures. Correlation filter methods gained widespread popularity in visual tracking after MOSSE (Minimum Output Sum of Squared Errors) [11], chiefly due to their computational speed. Their Achilles' heel, however, lies in limited representational capacity, which curtails their effectiveness when confronting the demanding scenarios typical of aerial

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

footage—such as poor resolution, minuscule targets, occlusions, scale changes, and interference from similar distractors [12, 13]. In parallel, the surge of deep learning—driven by the formidable representational power of convolutional neural networks (CNNs)—has permeated countless domains [14]. Siamese-based trackers [15] leverage this deep feature extraction, operating on a template-matching principle to markedly improve tracking quality through learned deep representations.

Even so, when measured against generic tracking tasks, aerial tracking continues to grapple with a distinctive set of unresolved challenges induced by complex observation environments and airborne platform constraints. Compared with standard tracking scenarios, aerial sequences introduce challenges due to the bird's-eye perspective [16], including low resolution, occlusion, changing aspect ratios, small target footprints, distractors with similar appearance, and scale fluctuations. In the first instance, handling perturbations such as occlusion, aspect-ratio shifts, and scale variance requires a tracker to be equipped with robust feature extraction. Regrettably, most current lightweight backbone architectures rely on CNNs with static convolutional kernel sizes, which are insufficient for the rigors of aerial tracking. Conversely, more heavyweight backbones exact a computational toll too steep to reconcile with the real-time speed imperative. Secondly, the combined assault of low resolution, tiny objects, and visually akin distractors renders the extraction of discriminative and robust features exceedingly tough. The stream of temporal context spanning successive frames encodes valuable motion patterns, which we contend can empower a tracker to withstand these challenges. Despite this, prevailing aerial trackers predominantly draw upon the template frame paired with the current search frame alone, overlooking the informative temporal continuity offered by neighboring frames. Lastly, because aerial tracking solutions must ultimately be fielded on platforms with modest computational budgets, the trade-off between model efficiency and tracking accuracy remains an enduring concern.

Extracting powerful representations without incurring significant computational expense is a key requirement for aerial tracking. We draw on the property that atrous convolution [17] enlarges the effective receptive field of a filter without introducing additional learnable parameters or floating-point operations, and, on this basis, we devise a parallel atrous module (PAM) that extends a standard backbone. Our PAM deploys multiple parallel branches, each employing a 3×3 atrous convolution with a distinct dilation factor, to simultaneously collect and blend multiscale features enriched by spatial context while preserving fast inference speeds. Spatial context, in this sense, refers to the relational information shared between proximate and more distant pixel locations.

Further, to equip the network with temporal awareness, we propose an adaptive temporal context aggregation module (ATCAM). The proposed ATCAM absorbs temporal context from a sequence of consecutive frames and uses an encoder-decoder architecture to adaptively modulate the response map for the current frame. Importantly, the ATCAM recycles computation outputs generated by the network during prior frame processing, introducing temporal context with only a trivial additional computational footprint. The central contributions of our work are enumerated below.

(1) We present a Siamese aerial tracking framework that jointly leverages multiscale feature maps and continuous temporal context, delivering enhanced aerial tracking accuracy whilst maintaining a careful balance between parameter count and computational workload, thereby realizing an efficient, lightweight solution.

(2) We design the PAM, which aggregates multiscale information and enhances spatial contextual richness with minimal extra computational burden, thereby elevating the expressive quality of the learned features.

(3) We design the ATCAM, which retains response maps from preceding frames and uses them to adaptively recalibrate the response map of the current frame, integrating temporal contextual cues with minimal additional computational overhead.

(4) Comprehensive experimental assessments across three established aerial tracking benchmarks—DTB70, UAV20L, and UAV123@10fps—verify that our method delivers leading aerial ground-tracking performance compared to contemporary state-of-the-art approaches and attains an operating speed of 75.5 fps on an NVIDIA 3060Ti GPU.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

The remainder of this manuscript unfolds as follows. Section 2 provides a concise review of the related literature on aerial tracking. Section 3 delineates our proposed aerial tracking network and provides in-depth details on the PAM and ATCAM designs. Section 4 reports and discusses the experimental validation. Section 5 summarizes the work with concluding remarks.

Related Work

Siamese trackers for aerial tracking

Siamese-based trackers have attracted substantial research interest in object tracking, leveraging expressive feature representations enabled by deep neural networks. SINT [18] is an early contribution that presents a deep Siamese architecture for tracking via template matching, yielding state-of-the-art results.

The anatomy of a Siamese network can be decomposed into three stages: a feature encoding backbone, a cross-correlation fusion operation, and a prediction module. The encoding backbone comprises two structurally identical, weight-tied branches: one processes the target template, and the other handles the search region. During the fusion stage, the template-derived features serve as a convolutional kernel, cross-correlated with the search region features to generate a similarity response map. The prediction module subsequently decodes this map to output the tracking result.

SiamFC [15] introduced an end-to-end trainable fully convolutional Siamese network that delivers real-time, accurate tracking. SiamRPN [19] augmented the Siamese architecture with a region proposal network and box regression branches to achieve tighter bounding-box predictions. SiamCAR [20] shifted to an anchor-free paradigm, eliminating the cumbersome hyper-parameter selection associated with anchor-based approaches while simultaneously lowering computational demands. The versatility of Siamese-based architectures has led to their widespread adoption in aerial tracking tasks. As notable examples, SiamAPN [21] developed an adaptive anchor generation and refinement mechanism within the Siamese pipeline, reducing computational burden while improving robustness under complex aerial conditions. SiamAPN++ [6] crafted an attention enhancement module that blends self-attention and cross-attention, yielding an attentional Siamese framework with improved tracking fidelity. SiamSA [22] embedded a paired-scale attention block into the Siamese design to combat the pronounced scale variation inherent in aerial tracking.

These prior works demonstrate that diverse architectural innovations layered on the Siamese template can substantially enhance tracking performance, affirming the framework's promise for aerial scenarios. Building on this foundation, our work presents a lightweight Siamese network that fuses spatial and temporal contextual cues, specifically designed for unmanned aerial vehicle tracking.

Atrous convolution

A convolution operation's receptive field specifies the spatial window of the input feature map that influences a given output unit; broader receptive fields permit the capture of more holistic, global-level patterns. With conventional convolutions, increasing the receptive field requires inflating the kernel dimensions, which inevitably multiplies both the number of learnable parameters and the number of floating-point operations. This trade-off motivated us to investigate atrous convolution. Unlike its standard counterpart, which is constrained to a fixed receptive extent, atrous convolution expands its field of view by interspersing zeros among kernel entries, accomplishing this enlargement without increasing parameters or computation. The step distance between adjacent effective kernel positions, ignoring inserted zeros, is termed the dilation rate. A visual explanation of atrous convolution is presented in **Figure 1**.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

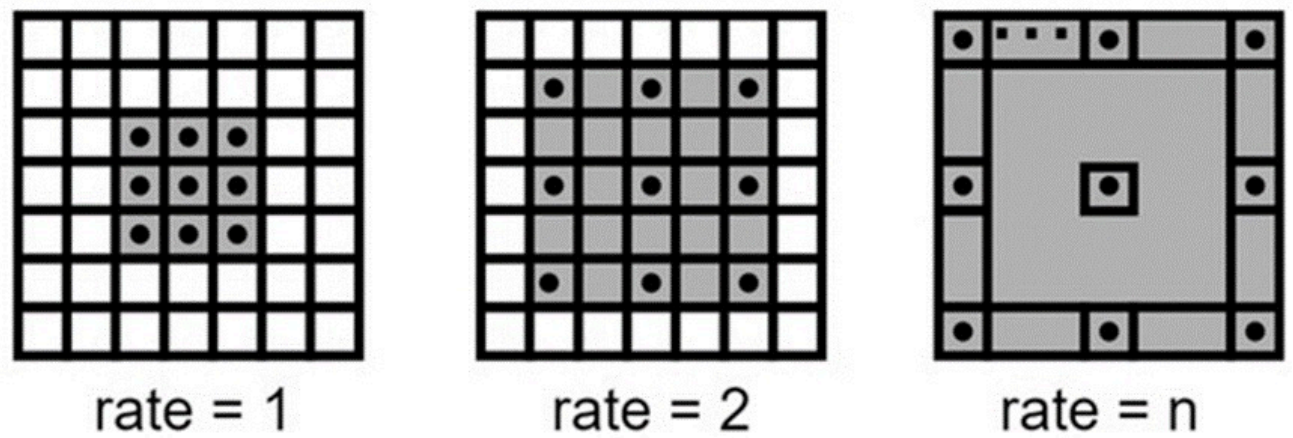


Figure 1. The schematic of the atrous convolution, 'rate' means the dilation rate.

As illustrated in **Figure 1**, the receptive field scale shifts as a function of the dilation rate, and their relationship is governed by the following formula:

$$RF = [r(n-1)+1] \times [r(n-1)+1] \quad (1)$$

Where r represents the dilation rate, n denotes the kernel size, and RF indicates the receptive field. Because inserting zero gaps introduces no additional parameters or computational cost, many research efforts have capitalized on atrous convolution for feature learning. Bolme *et al.* [11] leveraged atrous convolution to build feature pyramids for multiscale representation learning. CFPNet [23] adopted asymmetric atrous convolutions to construct a channel-centric feature pyramid tailored for segmentation. Cao *et al.* [24] assembled a multi-dilation component using dilated convolutions to harvest information across multiple scales. Wang *et al.* [25] refined a residual learning framework with dilated convolutions, enabling the propagation of high-frequency details while correcting chromatic inconsistencies.

Although atrous convolution can indeed enlarge the receptive field, relying on a single dilation rate still confines the network to a single fixed spatial extent. To overcome this, we devise a parallel atrous module (PAM) that simultaneously executes multiple 3×3 atrous convolutional branches with different dilation rates to merge features across scales and infuse greater global awareness.

Temporal context information for object tracking

By its very nature, object tracking operates on video streams composed of successive frames; the dense temporal signals embedded in these frames are critical. In aerial tracking, specifically, adverse conditions such as low resolution and diminutive targets frequently arise, making motion and continuity cues spanning adjacent frames indispensable for reliable tracking.

Prior work has explored ways to integrate temporal context into tracking pipelines. For instance, AutoTrack [26] formulated a temporal regularization scheme that automates hyper-parameter tuning while exploiting temporal relationships throughout the tracking process to lift performance. Within the Siamese tracking paradigm, transformer-based designs have become the predominant vehicle for fusing temporal information. TrDiMP [27] engineered a transformer encoder-decoder structure to pool temporal context from frame sequences. PTSEFormer [28] likewise used transformers, adding attention layers to unify the current frame representation with historical temporal cues. TCTrack harnesses a transformer backbone to amplify feature extraction and adjust response maps informed by temporal

© 2023 The Author(s). | Published by Agro Intelligence Research Press

dynamics. Yet transformers exact a high computational cost, rendering them poorly suited to aerial deployment scenarios where computing resources are tightly constrained.

In response, our work introduces ATCAM, a lightweight module for temporal context aggregation. The module caches response maps computed from prior frames. It uses them to adaptively recalibrate the response map of the current frame, thereby weaving in temporal contextual cues across consecutive frames with an almost negligible increase in computational load. This enhances tracking resilience within the constraints of aerial operation.

Materials and Methods

This section provides an in-depth walkthrough of the architecture we introduce. **Figure 2** presents an overview of the system, which is built around four core modules: a backbone for feature encoding, a cross-correlation-based fusion stage, the ATCAM component, and a bounding box prediction head. To close out this section, we outline the optimization and evaluation protocols used during training and testing.

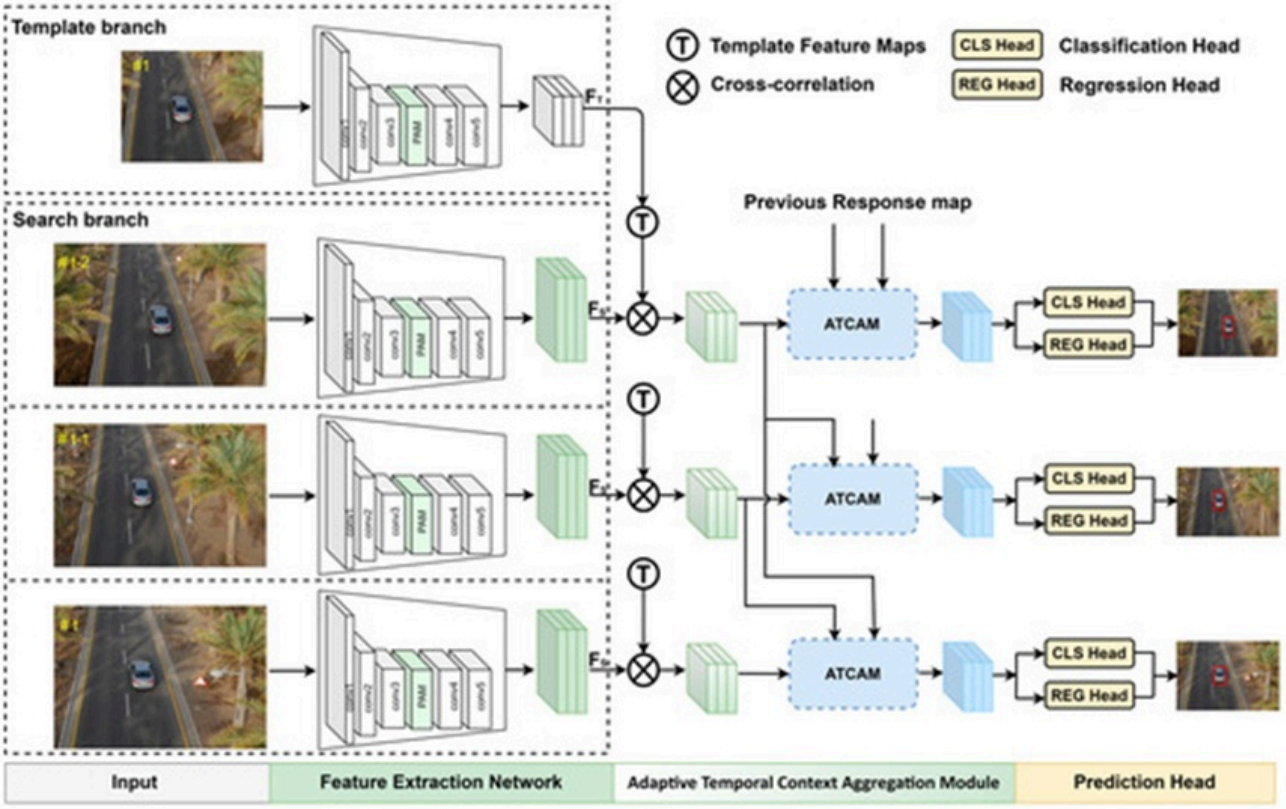


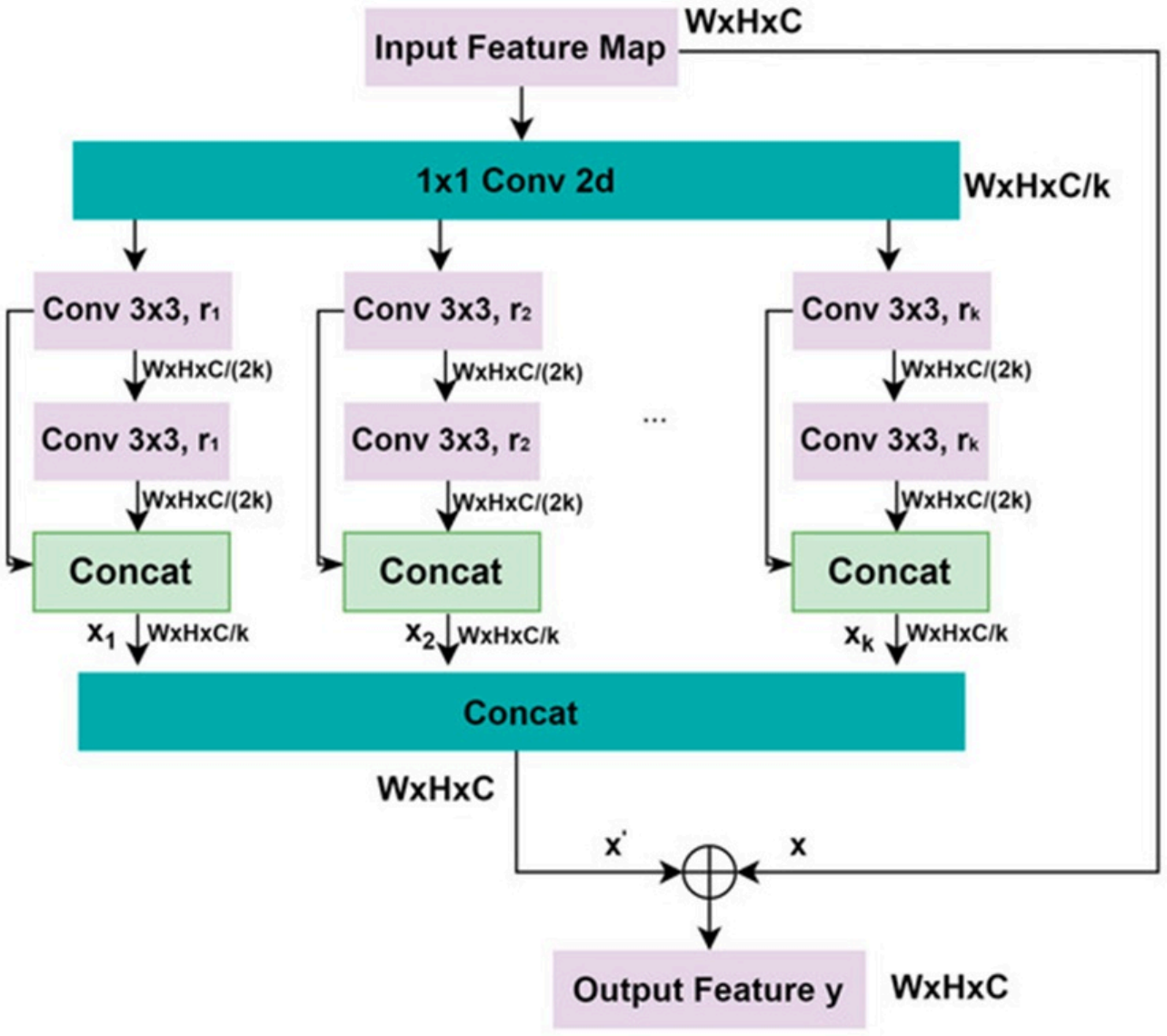
Figure 2. The schematic structure of the proposed tracking network.

Feature extraction network

In virtually all computer vision pipelines, the quality of feature extraction is a foundational determinant of downstream performance. Within aerial tracking specifically, the ability to isolate salient and robust representations substantially elevates a tracker's immunity to background clutter, underscoring its critical role. The Siamese feature extraction backbone, visualized in **Figure 2**, splits into two weight-coupled pathways—one dedicated to encoding the exemplar template, the other to encoding the candidate search window. These twin pathways accept a template crop $Z \in \mathbb{R}^{127\times127\times3}$ and a search crop $S \in \mathbb{R}^{255\times255\times3}$, where the figures 127 and 255 specify the spatial extents of each respective input.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

To satisfy the latency requirements inherent to onboard aerial processing, we deliberately avoid parameter-heavy designs such as ResNet [29] and instead adopt the compact AlexNet [30] architecture, which contains only five convolutional layers. The trade-off, of course, is that a leaner backbone possesses inherently weaker representational capacity. To remedy this shortcoming and to more effectively counteract the characteristic difficulties of aerial tracking—occlusion, non-rigid deformation, scale fluctuation, among others—we contribute a parallel atrous module (PAM). The PAM runs a bank of atrous convolutions configured with heterogeneous dilation rates side by side, enabling simultaneous multiscale feature coverage and substantially widening the aggregate receptive field, which aids in distilling discriminative cues amidst the challenging aerial tracking milieu. **Figure 3** diagrams the internal layout of the PAM. Starting from an input tensor $x \in \mathbb{R}^{w \times h \times c}$, with w and h indicating spatial dimensions and c the channel depth, a 1×1 convolution first compresses the channel count before the data stream fans out into k parallel atrous pathways, each tuned to a unique dilation factor. Every such pathway contains a pair of 3×3 atrous convolutions operating at an identical dilation rate. Within each pathway, we implement a miniature feature pyramid motif [31] wherein the outputs of the two consecutive convolutions are skip-connected via concatenation, producing the per-branch representation $xi \in \mathbb{R}^{w \times h \times (c/k)}$, $i \in \{1,2,\ldots,k\}$. The dilation factor varies from branch to branch. Repeating this scheme across all branches harvests a collection of multiscale descriptors. The ensemble of branch outputs is subsequently merged into $x' \in \mathbb{R}^{w \times h \times c}$. A residual additive connection then combines x with x' to generate the module's final output $y \in \mathbb{R}^{w \times h \times}$.



© 2023 The Author(s). | Published by Agro Intelligence Research Press

Figure 3. The schematic structure of the proposed parallel atrous module.

Owing to the distinct dilation rate assigned to each parallel pathway, every branch effectively observes a differently scaled subregion of the identical input tensor. Branches configured with modest dilation rates naturally gravitate toward low-level textural detail. In contrast, those with aggressive dilation rates become receptive to high-level semantic patterns shaped by distant contextual relationships. Through this mechanism, the PAM consolidates multi-granularity features imbued with holistic scene awareness, boosting the model's representational power and helping the tracker weather perturbations such as occlusion, deformation, and scale shifts. In our setup, the dilation rates are 1, 2, 3, and 5, yielding receptive field spans of 3, 5, 7, and 11, respectively.

Because the PAM preserves the tensor shape, it can seamlessly replace any conventional layer in the feature extraction backbone. We gather the PAM outputs with those from the last two convolutional stages to form the feature extraction network's final multi-level output. The resulting feature bundles from the two pipelines are written as $FTi \in \mathbb{R}^{7\times7\times256}, \in 1,2,3$ and $FSi \in \mathbb{R}^{31\times31\times256}, i \in 1,2,3$, where the numbers 7 and 31 reflect the spatial resolution of the template and search feature grids, each possessing 256 channels.

Cross-correlation fusion network

Once feature extraction concludes, we hold the template-derived features FT and the search-region-derived features FS. The cross-correlation fusion stage is tasked with computing response maps that quantify the similarity between these two sets of representations by correlating and integrating FT with FS. Up until this stage, the template and search pathways remain entirely decoupled; the fusion network therefore establishes the first point of information interchange between them, marking a crucial junction in the tracking workflow.

A well-known property of convolutional hierarchies is that shallow layers preserve fine spatial detail at the expense of semantic abstraction. In contrast, deep layers sacrifice spatial granularity for high-level conceptual understanding. If one were to base the fusion exclusively on the terminal layer's feature map, the nuanced cues held in earlier layers would be entirely discarded. Accordingly, to simultaneously draw upon both shallow and deep representations and thus achieve a richer information merger, we source feature maps from multiple depth levels for response map computation. **Figure 4** illustrates this multi-level fusion scheme.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

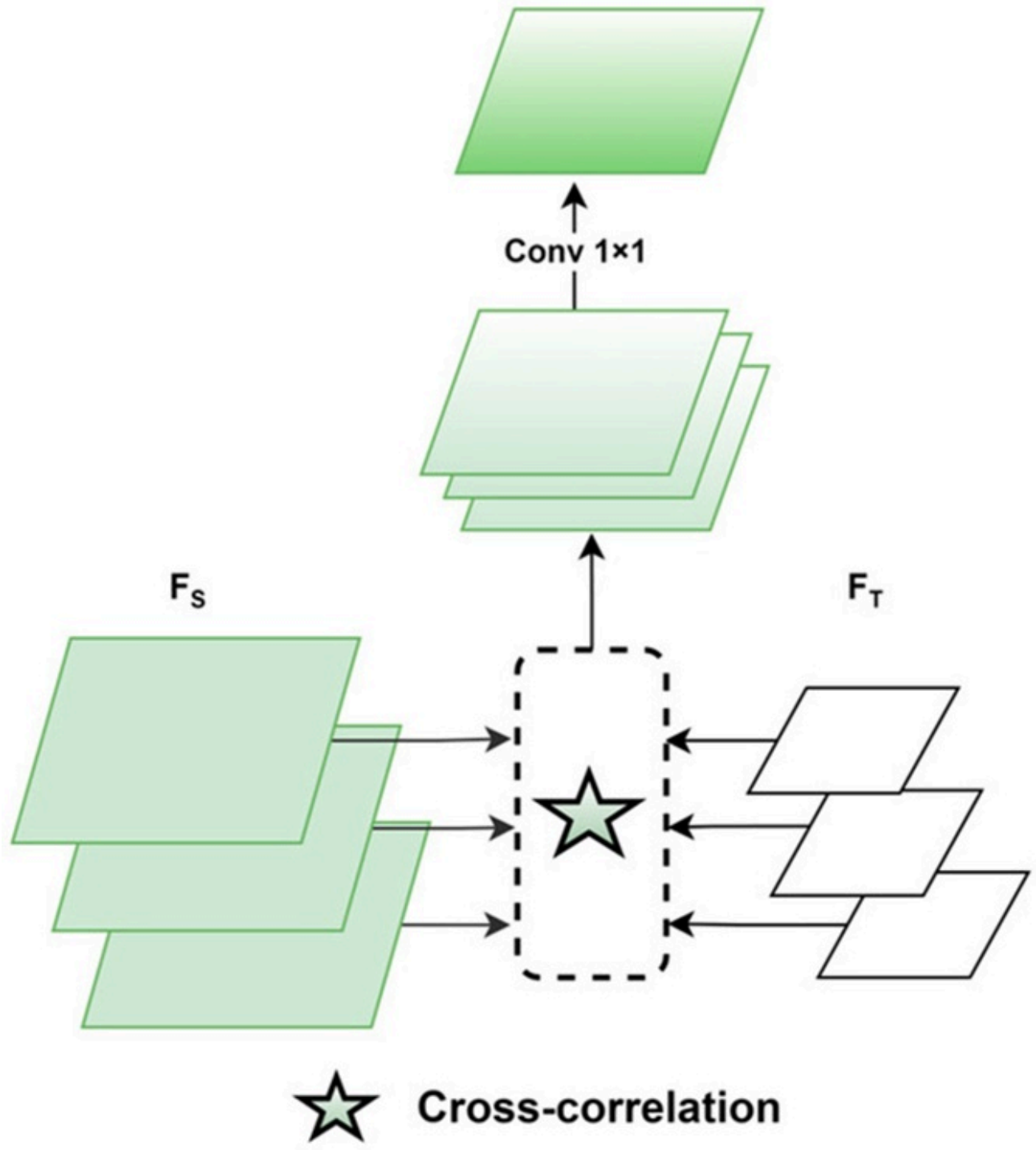


Figure 4. The schematic structure of the cross-correlation fusion network.

Operationally, the fusion network receives $FTi \in \mathbb{R}^{7\times7\times256}, \in 1,2,3$ and $FSi \in \mathbb{R}^{31\times31\times256}, i \in 1,2,3$ as inputs. The template feature maps FTi are repurposed as convolution kernels and slid across FSi in a cross-correlation manner to yield the response maps. The computation follows the expression below:

$$R_I = F_{TI} \bigotimes F_{SI}, i = 1,2,3 \qquad (2)$$

© 2023 The Author(s). | Published by Agro Intelligence Research Press

With $\otimes$ signifying the cross-correlation operator and Ri $Ri \in \mathbb{R}^{25\times25\times256}$, $i \in 1,2,3$ denoting the resultant response maps. A 1×1 convolution subsequently merges the trio of response maps Ri, a step summarized by:

$$R = \mathrm{Conv}\,1 \times 1(\mathrm{Concat}(R_t)) \quad (3)$$

Where $R \in \mathbb{R}^{25\times25\times256}$ stands for the consolidated response map.

Adaptive temporal context aggregation module

Motivated by the desire to fully exploit temporal continuity in video streams, we propose an adaptive temporal context aggregation module (ATCAM), as shown in **Figure 5**. Architecturally, the ATCAM splits into an encoder that compresses temporal context and a decoder that reconstructs the temporally enriched representation. The encoder synthesizes a temporal prior Re by collectively processing the response map belonging to the current frame alongside response maps propagated from adjacent time steps. The decoder then integrates temporal context into the current frame's response map via concatenation-based fusion with the prior. Our decision to perform temporal reasoning at the response map level, rather than earlier feature stages, stems from two considerations: first, the response map already embodies fused content from both the template and search pipelines of nearby frames, providing richer temporal signals than either branch alone; second, operating on response maps is substantially lighter in terms of memory footprint and computational expense compared to manipulating higher-dimensional feature volumes.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

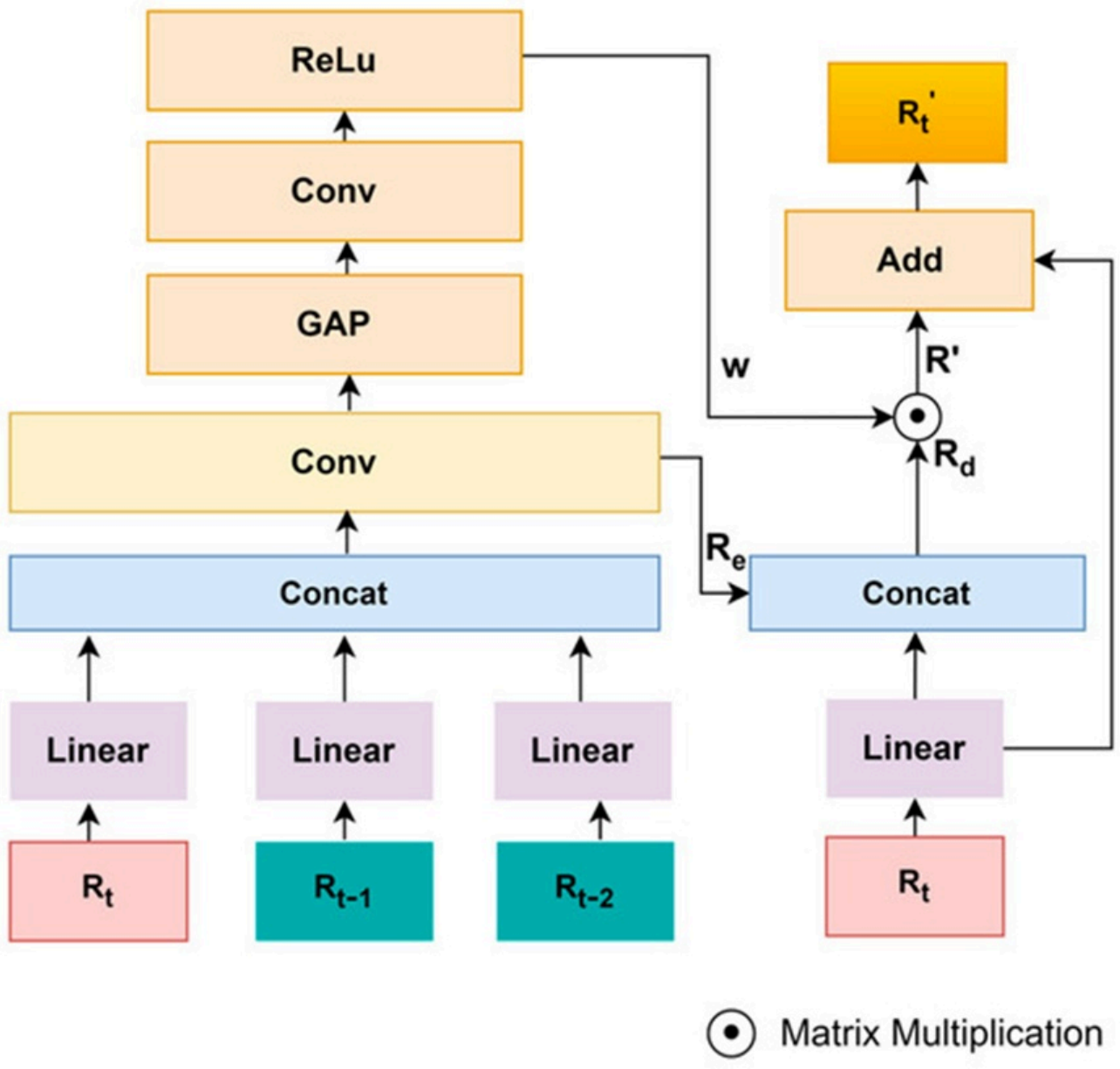


Figure 5. The schematic structure of the proposed adaptive temporal context aggregation module.

In detail, the encoder ingests the historical response maps $Rt-i \in \mathbb{R}^{25\times25\times256}, \in 1,\ldots,n$ together with the present response map $Rt \in \mathbb{R}^{25\times25\times256}$, and distills these through a linear-Concat-fusion arrangement to produce the temporal prior $Re \in \mathbb{R}^{25\times25\times256}$. This encoding operation is captured by:

$$Re = \text{Conv}(\text{Concat}(\text{Linear}(Rt, \ldots, Rt - n))) \quad (4)$$

Subsequently, the decoder concatenates this prior with the current response map and fuses them, outputting $Rd \in \mathbb{R}^{25\times25\times256}$, as formalized below:

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$Rd = \mathrm{Conv}(\mathrm{Concat}(Re, \mathrm{Linear}(Rt))) \quad (5)$$

It merits emphasis that frames neighboring the current time step are not uniformly beneficial; certain temporal contexts—instances of motion blur, heavy occlusion, and the like—may be counterproductive or even detrimental to localizing the target at the present moment. To endow the module with the capacity to sift through temporal signals adaptively, we incorporate an adaptive temporal context regulator. This regulatory unit performs global average pooling (GAP) across the encoded prior Re and routes the pooled descriptor through a feed-forward network to emit a modulation coefficient $w \in \mathbb{R}^{1\times1\times256}$. This coefficient is subsequently used in the decoder to gate feature aggregation in a channel-selective manner.

$$w = \mathrm{ReLu}(\mathrm{GAP}(\mathrm{Conv}(Re))) \quad (6)$$

An element-wise product between the modulation coefficient w and the aggregated feature Rd yields R', expressed as:

$$R' = w \cdot Rd \quad (7)$$

As a final safeguard, residual skip connections are woven into the decoder to ensure that the temporal enhancement introduced by adjacent frames supplements rather than overwhelms the current frame's own response. The concluding output of the ATCAM is denoted as $R't \in \mathbb{R}^{25\times25\times256}$.

$$Rt' = R' + \mathrm{Linear}(Rt) \quad (8)$$

Prediction head

Once the feature encoding backbone, cross-correlation fusion stage, and ATCAM have each performed their respective computations, we are left with a temporally conditioned response map denoted as $R't$. Sitting atop the network, the prediction head splits into two dedicated substreams: one for classifying each spatial bin and another for regressing box coordinates. The classifier's job is to scan every grid cell within the response map and assign a binary label—target foreground versus background—while the regressor estimates a bounding quadrilateral tied to that same cell. The mechanics are detailed below.

To begin, $R't$ is forwarded into the classification pathway, which yields a two-channel score tensor $Acls \in \mathbb{R}^{H\times W\times2}$. Here, H and W capture the vertical and horizontal granularity of the classification output, and the depth of 2 reflects that each spatial index $(i,)$ stores a duplet of real values indicating foreground and background likelihood, respectively. In tandem, the regression pathway emits a four-channel tensor $Areg \in \mathbb{R}^{H\times W\times4}$, with the quartet of numbers at any given location $(i,)$ encoding a displacement vector. Each such location corresponds to a particular subwindow inside the search crop; we use $(x,)$ to reference the subwindow's center. The four-dimensional descriptor, written compactly as $(i,)=(l,t,r,b)$, captures how far (x,y) lies from the left, top, right, and bottom extremities of the regressed box centered upon it.

Experiments

This section first lays out the configuration details of our framework. Subsequently, we benchmark against a panel of 19 cutting-edge trackers [6-8, 19-22, 26, 32-42] spanning three authoritative evaluation suites: UAV20L [43],

© 2023 The Author(s). | Published by Agro Intelligence Research Press

UAV123@10fps, and DTB70 [44]. An ablation campaign follows, designed to isolate the contributions of the PAM and ATCAM modules introduced herein. We further contrast the runtime throughput of our method with four aerial tracking baselines. A concluding qualitative segment presents visual side-by-side comparisons and targeted case discussions that bring the tracking outcomes to life.

Implementation detail

Model parameters were optimized using a composite of four training corpora: COCO [45], GOT10K [46], LaSOT [47], and VID [48]. All optimization runs were executed on an RTX 3080 GPU within a Python 3.7 environment with PyTorch v1.7.1. The chosen optimizer was Stochastic Gradient Descent (SGD) with a momentum term of 0.9, and training proceeded for 20 full passes over the data. Mini-batches contained 64 samples apiece. Because our formulation draws on temporal evidence from neighboring frames, a multi-frame sampling regimen was required during training. Within any given source video, one frame is designated as the exemplar template, another as the search view, and a random selection of temporally adjacent frames serves as contextual references for tuning the ATCAM.

The composite training objective is the sum of a classification penalty and a regression penalty. Regarding the former, each grid cell $(i,)$ inside the classification score tensor $Acls\in\mathbb{R}^{H\times W\times 2}$ houses foreground and background activation scores δpos and δneg, and the discrepancy is measured via the standard cross-entropy formulation [49]:

$$L_{cls} = 0.5 \times (\mathrm{L}_{\mathrm{BCE}}(\delta_{pos}, I) + \mathrm{L}_{\mathrm{BCE}}(\delta_{neg}, I)) \quad (9)$$

Where I provides the ground-truth class indicator at cell $(i,)$, and $LBCE$ denotes the binary cross-entropy operator. Now, let $(x0,0)$ and $(x1,y1)$ stand for the upper-left and lower-right vertices of the ground-truth bounding rectangle, and let (x,y) be the search-region point that aligns with regression map index (i,j).

$$\begin{aligned} l' &= x - x_0, t' = y - y_0 \\ r' &= x_1 - x, b' = y_1 - y \end{aligned} \quad (10)$$

In **Eq. 10**, the symbols $l',',r',b'$ measure the separation between (x,y) and the left, top, right, and bottom borders of the ground-truth box, respectively. To counteract the heavy skew between foreground and background samples, we do not regress over the entire spatial grid; instead, we restrict regression supervision to those indices (i,j) whose associated search-space coordinates (x,y) reside strictly within the confines of the ground-truth box—equivalently, those satisfying $l',t',r',b'>0$. This selection rule is expressed as:

$$\mathbb{I}(i,j) = \begin{cases} 1 & l', t', r', b' > 0 \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Where (i,j) acts as a gating function, evaluating to one when position (i,j) should contribute to the regression loss and zero otherwise. With $(x,)$ representing the annotated box and $(i,)$ the predicted counterpart, the regression penalty follows:

$$L_{reg} = \frac{1}{\sum_{(i,j)} \mathbb{I}_{(i,j)}} \sum_{(i,j)} \mathbb{I}_{(i,j)} \times L_{\mathrm{IoU}}\left[T_{(x,y)}, t_{(i,j)}\right] \quad (12)$$

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$L_{\mathrm{IoU}}\left[T_{(x,y)}, t_{(i,j)}\right] = -\ln\left(\frac{T_{(x,y)} \cap t_{(i,j)}}{T_{(x,y)} \cup t_{(i,j)}}\right) \quad (13)$$

Where $Lreg$ is the regression error, and $LIoU$ refers to the Intersection over Union (IoU) loss [50] evaluated between the ground-truth annotation (x,y) and the estimate $t(i,j)$. The regression term aims to reduce the spatial mismatch between the hypothesized box and the actual annotation; we use an IoU-based loss to quantify this mismatch.

At test time, the manually annotated bounding box from the inaugural frame of each sequence serves as the target template, and all subsequent frames serve as search images. Both the template and the search undergo the forward pass described in **Figure 2**. A subtlety arises in the operation of the ATCAM: proceeding in temporal order, the very first search frame has no accumulated historical context to rely upon. The ATCAM is therefore skipped for this frame, though the resulting response map is cached to inform processing of subsequent frames. Throughout our experiments, we retain the two most recent frames preceding the current time step as the temporal reference set.

Comparison with the state-of-the-arts

Tracking fidelity is gauged on the UAV20L, UAV123@10fps, and DTB70 platforms using two canonical metrics: success rate and precision rate. The success rate is derived from the IoU score between the tracker's output box and the reference annotation. The success plot traces the percentage of frames whose IoU exceeds a variable decision threshold. The precision metric quantifies the Euclidean distance between the predicted and ground-truth box centers; the precision plot records the proportion of frames for which this distance is below a given cutoff. Unless stated otherwise, the precision threshold adopted herein equals 20 pixels, and the success threshold is taken as 0.5.

DTB70 benchmark

The DTB70 collection comprises 70 aerial-captured sequences, deliberately curated to address numerous challenges endemic to airborne tracking. Its diverse difficulty profile makes DTB70 an apt testbed for probing tracker robustness. As shown in **Figure 6**, our approach achieves both precision (0.815) and success rate (0.635) higher than those of the SOTA panel.

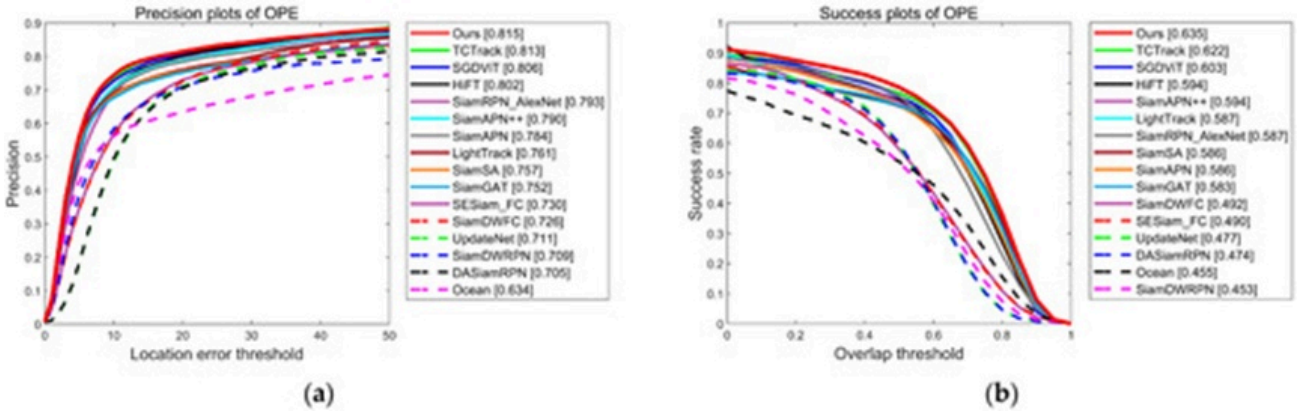


Figure 6. Precision plots (a) and success plots (b) on the DTB70 benchmark.

Digging deeper, the per-attribute success breakdown plotted in **Figure 7** confirms that the proposed tracker achieves standout results across multiple distinct challenge categories within DTB70. This body of evidence corroborates that the PAM and ATCAM introduced in this study, when used together, collectively strengthen feature extraction and inject adaptive contextual signals, resulting in tangible gains in aerial tracking quality. Of particular note, our method edges out TCTrack by margins of 0.2% on precision and 1.3% on success rate. TCTrack relies on transformer modules to weave temporal information into the tracking loop.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

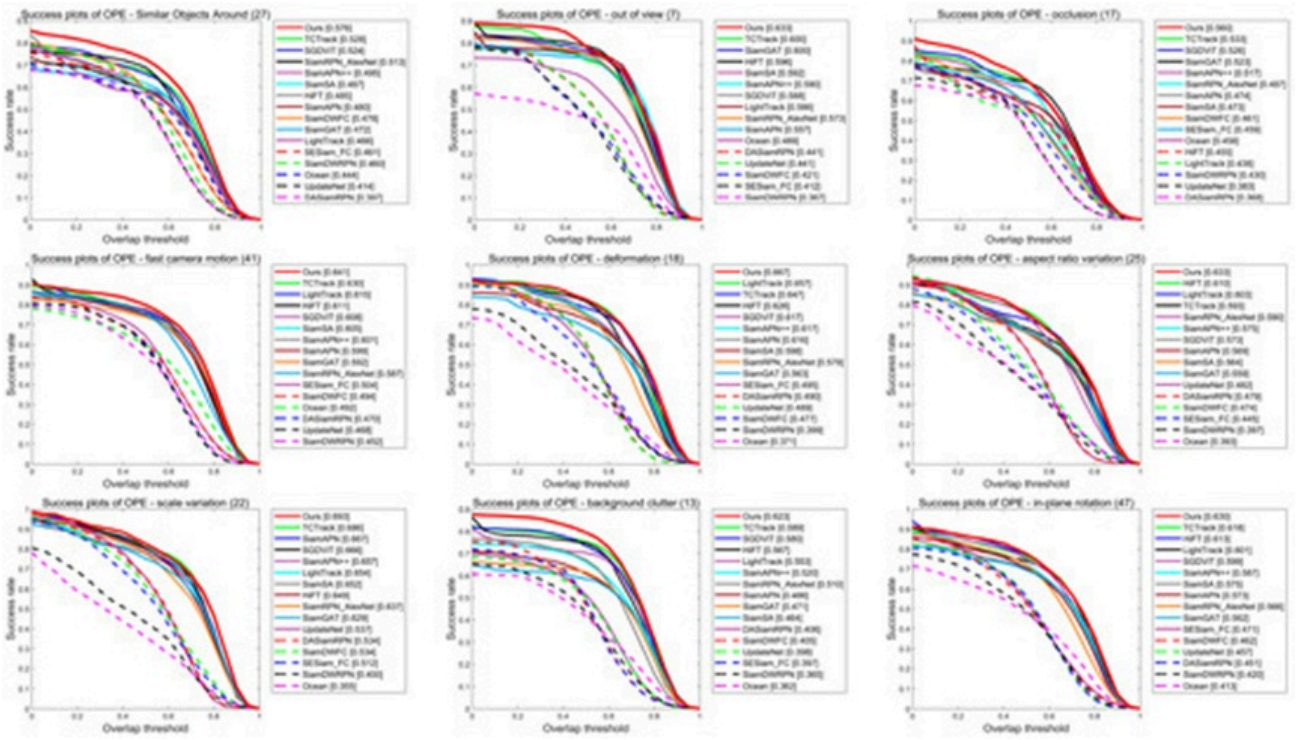


Figure 7. Success rate comparison across 15 SOTA trackers for 9 attributes on the DTB70 benchmark.

Inspecting the attribute-stratified results in **Figure 7**, one observes that our tracker and TCTrack jointly lead the field under challenges such as similar objects, out-of-view episodes, occlusion, and rapid camera motion. This pattern is intuitive: extracting faithful and discriminative features becomes exceedingly difficult when confusable distractors, target disappearance, heavy occlusion, or sudden motion are present. We believe the strong performance of both our method and TCTrack stems from their ability to leverage temporal context, which enables the trackers to maintain target continuity by referencing trustworthy signals from neighboring frames. Where the comparison diverges is in challenges such as deformation, aspect-ratio shifts, and scale changes—here, TCTrack lags behind the other evaluated algorithms. We interpret this gap as evidence that TCTrack's feature extraction capacity leaves room for improvement. Deformation and scale excursions necessitate a tracker that can pool information across multiple spatial granularities. Our tracker addresses exactly this need by embedding the PAM within the feature extraction backbone, thereby harvesting multiscale representations endowed with global contextual reach, and consequently pulls ahead of TCTrack.

To sum up, the combination of improved feature extraction and temporal context aggregation allows our tracker to surpass the competing SOTA methods on the DTB70 benchmark.

UAV20L benchmark

The UAV20L benchmark furnishes 20 extended-duration tracking videos filmed from unmanned aerial platforms. Prolonged tracking sequences increase the likelihood of cumulative disturbances—such as aspect-ratio fluctuations, camera egomotion, target occlusion, scale drift, and the like. As a result, UAV20L is exceptionally well-suited for stress-testing the longevity and robustness of trackers. For these reasons, we select UAV20L for performance validation here, again relying on success rate and precision rate as the yardsticks.

As illustrated in **Figure 8**, we draw comparisons against a roster of SOTA aerial trackers. The data clearly indicate that the tracker proposed in this paper achieves the most favorable results in the field. In quantitative terms, across the aggregate precision and success metrics, our tracker ranks first, outperforming both SiamRPN and SiamAPN++ by 2% (success rate) and 2.3% (precision rate).

© 2023 The Author(s). | Published by Agro Intelligence Research Press

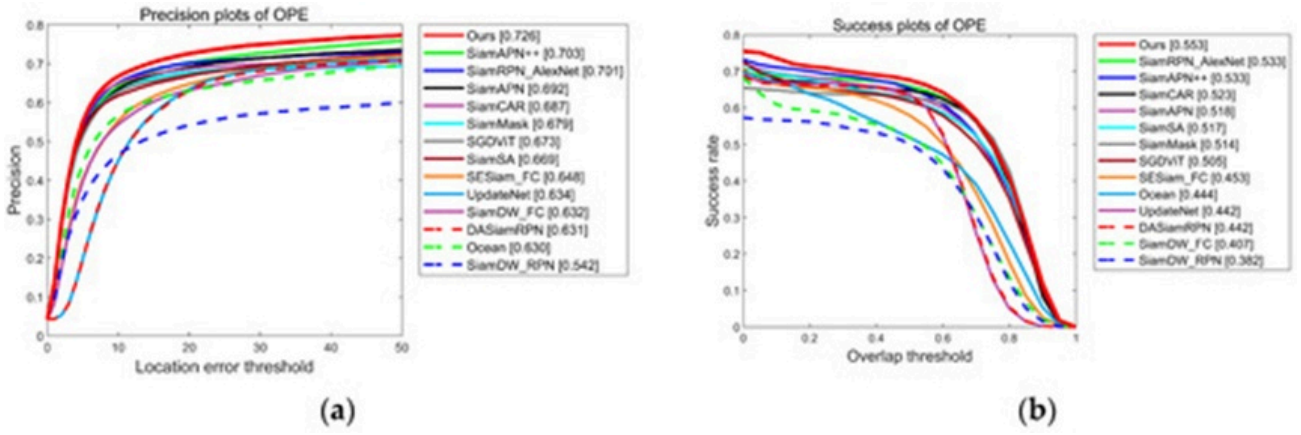


Figure 8. Precision plots (a) and success plots (b) on the UAV20L benchmark.

We further disaggregated performance by challenge attribute. As the results in **Figure 9** attest, the proposed tracker surpasses all competing methods under the conditions most commonly encountered during aerial operations—namely, similar objects, partial occlusion, low resolution, camera motion, aspect ratio changes, and scale variation. This demonstrates that our tracker is well-equipped to confront the multifaceted challenges that arise in extended aerial tracking scenarios and to deliver consistent performance improvements.

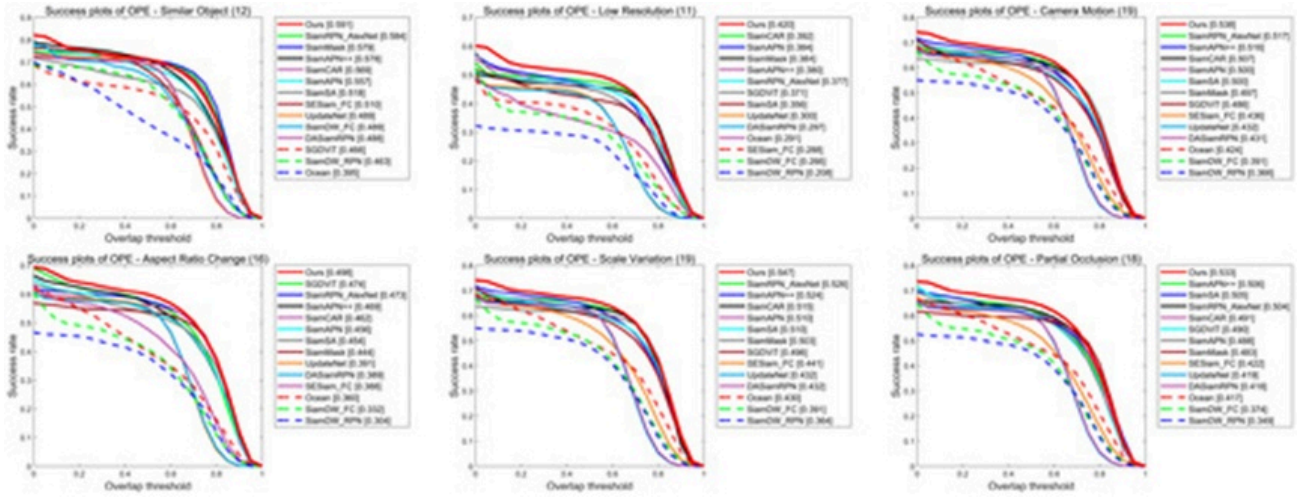


Figure 9. Success rate comparisons with 14 SOTA trackers across 6 attributes on the UAV20L benchmark.

UAV123@10fps benchmark

Where UAV20L emphasizes extended-duration tracking, UAV123@10fps pivots toward shorter sequences, assembling 123 videos that capture a wider cross-section of everyday tracking situations. The benchmark deliberately throttles the playback to 10 frames per second, which magnifies inter-frame displacement and creates a perception of swifter, more erratic target motion. This property renders UAV123@10fps particularly well-calibrated for probing how a tracker behaves when the target undergoes rapid, pronounced transformations. Success rate and precision continue to serve as the performance yardsticks.

Examining **Figure 10**, one sees that, when juxtaposed with the leading methods, our approach continues to rank in the top tier on both precision (0.777) and success rate (0.601). A telling comparison is SiamCAR, which our method surpasses by 1.1% on both metrics. The significance of this gap becomes clearer when considering that SiamCAR rests upon a ResNet50 backbone—an architecture whose computational heft and parameter volume dwarf the AlexNet

© 2023 The Author(s). | Published by Agro Intelligence Research Press

foundation we employ. This contrast highlights that our design strikes a practical compromise, delivering competitive accuracy without sacrificing runtime frugality. Attribute-conditioned results on UAV123@10fps, plotted in **Figure 11**, lend further weight to the tracker's resilience.

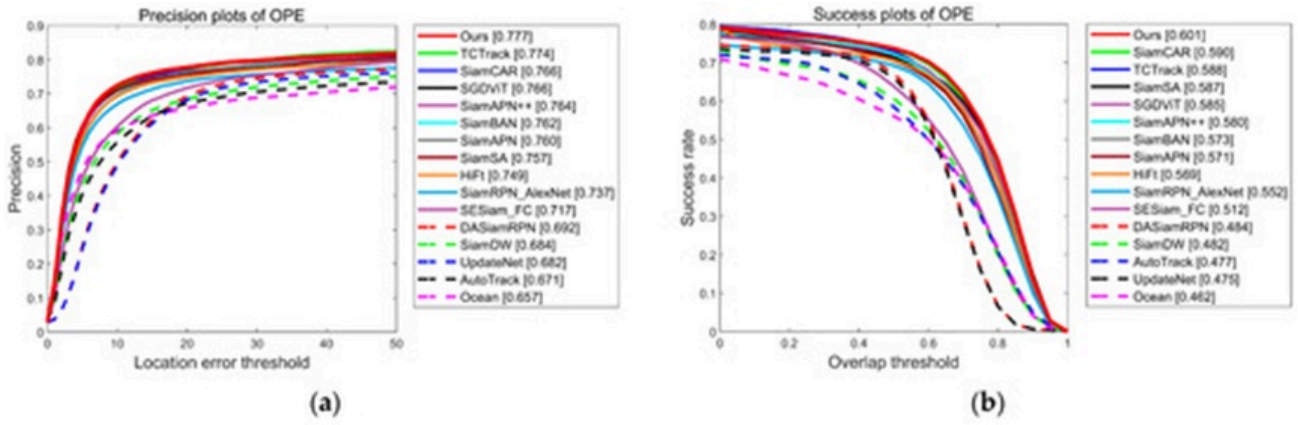


Figure 10. Precision plots (a) and success plots (b) on UAV123@10fps benchmark.

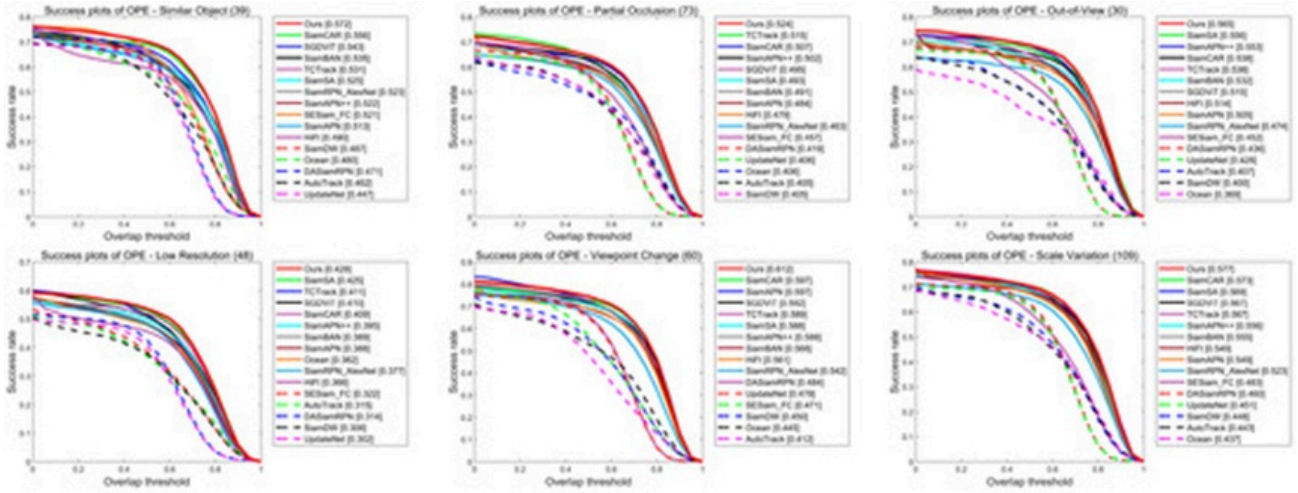


Figure 11. Success rate comparison with 15 SOTA trackers across 6 attributes on the UAV123@10fps benchmark.

Ablation experiments

To isolate and quantify the contributions of each proposed component—PAM and ATCAM—to aerial tracking, we conduct a set of ablation tests on UAV123@10fps. The starting point is a bare Siamese tracker comprising an AlexNet encoder, a cross-correlation matching layer, and a prediction module. By measuring how attribute-stratified success rates and aggregate accuracy shift after grafting PAM and ATCAM onto this minimal configuration, we can assess their respective contributions.

Table 1 catalogs the familiar challenges in aerial tracking: partial occlusion, scale fluctuations, aspect-ratio distortion, low resolution, confusable objects, and rapid motion. Versions augmented with either module register improvements across the board, albeit to differing degrees. When the target's visual signature morphs due to partial occlusion, scale variation, or aspect ratio change, the PAM-enhanced variant lifts success rates beyond the baseline by 1.9%, 2.1%, and 2.2%, respectively. These gains suggest that PAM's multiscale feature pooling, enriched by global contextual awareness, genuinely fortifies the model against appearance-degrading perturbations endemic to aerial footage.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Table 1. Ablation study of the proposed PAM and ATCAM. ↑ means the improvement in the success rate.

Model	Baseline + PAM + ATCAM	Baseline + ATCAM	Baseline + PAM	Baseline
Partial Occlusion	52.4% 4%↑	50.8% 2.4%↑	50.3% 1.9%↑	48.4%
Scale Variation	57.7% 3.4%↑	56.3% 2.0%↑	56.4% 2.1%↑	54.3%
Aspect Ratio Change	55.4% 2.6%↑	54.2% 1.4%↑	55.0% 2.2%↑	52.8%
Low Resolution	42.8% 2.2%↑	42.7% 2.1%↑	42.0% 1.4%↑	40.6%
Similar Object	57.2% 6.6%↑	55.0% 4.4%↑	53.8% 3.2%↑	50.6%
Fast Motion	51.3% 5.4%↑	49.8% 3.9%↑	48.0% 2.1%↑	45.9%
Precision	77.7% 2.9%↑	76.7% 1.9%↑	76.9% 2.1%↑	74.8%
Success	60.1% 3.2%↑	58.1% 1.2%↑	58.7% 1.8%↑	56.9%

On a separate front, low-resolution imagery and visually similar distractors pose a different category of difficulty: the tracker cannot latch onto features with sufficient discriminative salience. Here, the ATCAM proves its value by mining temporal regularities across neighboring frames. With ATCAM integrated, success rates for these two specific challenges increase by 2.1% and 4.4% relative to the baseline, demonstrating that temporal context can compensate for degraded spatial feature quality. The module also yields pronounced benefits under fast motion—a regime where motion blur routinely corrupts spatial features. We reason that inter-frame temporal continuity supplies implicit positional priors about where the target ought to appear, thus steering the tracker even when instantaneous appearance cues become unreliable.

Bringing both innovations together yields a further leap. **Table 1** shows that concurrent embedding of PAM and ATCAM boosts overall precision and success rate by 2.9% and 3.2%, respectively. Relative to adding either module singly, the dual configuration adds 0.8% and 1% in precision, and 1.4% and 2% in success rate. Across all challenge categories examined, the paired-module tracker outperforms both single-module variants, suggesting that PAM and ATCAM contribute synergistically rather than redundantly.

Tracking speed comparison

Because aerial trackers must eventually be deployed aboard resource-frugal platforms like small UAVs, chasing accuracy gains in isolation is an incomplete strategy; runtime throughput matters equally. We therefore dedicate this subsection to a head-to-head speed benchmark. Using the DTB70 dataset as the testbed, we timed four aerial tracking baselines—SiamSA, HiFT, TCTrack, and SGDViT—on identical RTX 3060 Ti hardware. The measurements, collated in **Table 2**, indicate that the algorithm proposed in this paper outperforms all comparators on both tracking fidelity and frames processed per second.

Table 2. Tracking speed comparison on the DTB70 benchmark.

	DTB70		
Model	**Fps**	**Success (%)**	**Precision (%)**
SiamSA	64.3	58.6%	75.7%
HiFT	65.8	59.4%	80.2%
TCTrack	72.1	62.2%	81.3%

© 2023 The Author(s). | Published by Agro Intelligence Research Press

SGDViT	72.6	60.3%	80.6%
Ours	75.5	63.5%	81.5%

Visual comparisons and case discussion

To convey a more tangible sense of tracking behavior, we have curated a set of visualized outputs in **Figure 12**. Consider first the Sheep2 clip: the footage is marred by exceedingly low resolution, the target occupies only a handful of pixels, and similar-looking objects populate the scene as viewpoints shift. As the sequence unfolds, SiamAPN, SGDViT, and SiamAPN++ each drift off target at different moments. Only our method and TCTrack maintain correct identification throughout frames 101–139, a pattern that aligns with our central hypothesis. When the target is tiny, poorly resolved, and surrounded by confusable entities, merely upgrading feature extraction alone is insufficient to protect against drift. The shared advantage of our tracker and TCTrack in this regime stems from their exploitation of temporal context. Yet as both the camera and the target continue moving, TCTrack eventually relinquishes the target, while our tracker alone persists—a testament to its superior robustness.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

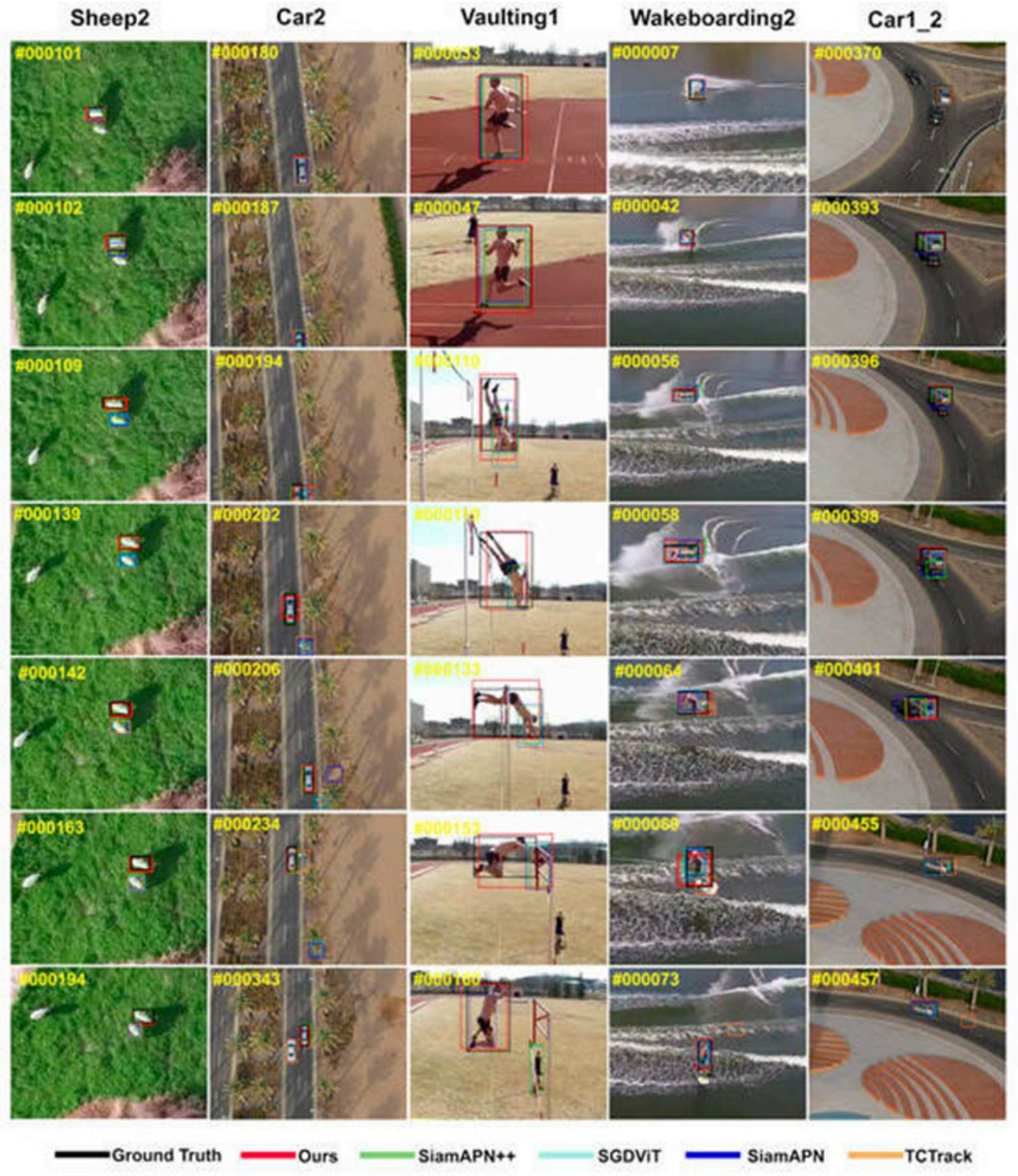


Figure 12. Visual Comparisons of our network and four SOTA trackers on five tracking videos.

Turning to the Car2 sequence, the target faces a gauntlet of aerial-tracking nuisances: temporary disappearances, cluttered surroundings, and similar-looking distractors. During frames 194–206, an occlusion event drastically reshapes the target's appearance; SGDViT, SiamAPN, and TCTrack are thrown off by street-lamp interference. Later, at frame 343, another similar-object episode occurs, and only our tracker alongside SiamAPN++ stays locked on. We credit

© 2023 The Author(s). | Published by Agro Intelligence Research Press

SiamAPN++'s resilience to its attention-based module that harvests global descriptors. By the same token, our tracker's unwavering performance likely reflects the PAM's capacity to synthesize multiscale, globally contextualized features.

The Vaulting1 sequence dramatizes extreme non-rigid deformation as the athlete executes a pole vault, compounded by occlusion and confusable elements. From frames 110 to 153, SGDViT, SiamAPN, SiamAPN++, and TCTrack collectively lose the ability to reliably pinpoint the target. At frame 160, SiamAPN++ and TCTrack are further derailed by a similar-looking distractor, whereas our method never wavers.

In Wakeboarding2, the subject again struggles with coarse resolution, background noise, and pronounced scale shifts. The visual outputs reveal that SiamAPN++, SGDViT, SiamAPN, and TCTrack all absorb various degrees of interference; our tracker, by contrast, achieves the cleanest performance among the tested set and delineates the target with greater fidelity. We believe the PAM-derived multiscale representations are what enable the tracker to exhibit the plasticity needed to cope with vigorous scale oscillations while still outputting a tight bounding box.

Our tracker is the strongest performer in the comparative field, consistently delivering higher precision. In the Car1_2 footage, the target experiences occlusion, similar-object distraction, and deformation induced by turning maneuvers. The results show that our tracker effectively navigates these obstacles, maintaining accurate and uninterrupted target pursuit. Synthesizing these visual comparisons and case analyses, we arrive at the following picture: the tracker introduced in this paper capably navigates the spectrum of aerial tracking challenges—low resolution, occlusion, similar distractors, deformation, and beyond. This body of evidence supports the conclusion that the PAM successfully captures multiscale representations and sharpens feature distinctiveness for robust target discrimination. At the same time, the ATCAM proves decisive in scenarios like Sheep2, where degraded resolution collides with similar-object interference, enabling the tracker to preserve accurate target lock.

Conclusion

This paper presents a Siamese tracker, purpose-built for aerial tracking, that pursues accuracy gains through two mutually reinforcing strategies, each motivated by the characteristic difficulties of the airborne tracking environment—low resolution, aspect-ratio shifts, visually similar distractors, rapid motion, occlusion, and scale variation. The first strategy centers on the parallel atrous module (PAM), which we engineer to boost the model's representational power under aerial conditions. The PAM runs multiple atrous convolutional streams at distinct dilation rates side by side, enabling each stream to sample a differently scaled receptive patch from the same feature layer; the aggregated output fuses multiscale information laced with global context, all without undercutting model efficiency. The second strategy addresses a different pain point: the breakdown of spatial feature extraction at low resolution, with occlusion, or when targets are small. To this end, we introduce the adaptive temporal context aggregation module (ATCAM), which adaptively infuses temporal cues from adjacent frames into the current frame's processing, thereby assisting the tracker in resisting distractors and maintaining precise target recognition when high-fidelity spatial features are unavailable. Ablation analyses, comprehensive benchmarking against state-of-the-art trackers on DTB70, UAV20L, and UAV123@10fps, and runtime profiling collectively attest to the proposed network's accurate, computationally efficient aerial tracking.

Acknowledgements

The authors are grateful to the anonymous reviewers for their critical comments and constructive suggestions.

Conflict of interest

None

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Financial support

This work was supported by the Natural Science Foundation of Jilin Province (YDZJ202101ZYTS048).

Ethics statement

None

Received: 22 Aug 2022 Revised: 27 Oct 2022 Accepted: 11 Nov 2022
Published online: 25 November 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Panahi FH, Panahi FH, Ohtsuki T. A reinforcement learning-based fire warning and suppression system using unmanned aerial vehicles. IEEE Trans Instrum Meas. 2022;72:1-16.

2. Bai Y, Song Y, Zhao Y, Zhou Y, Wu X, He Y, et al. Occlusion and deformation handling visual tracking for uav via attention-based mask generative network. Remote Sens. 2022;14:4756.

3. Li J, Jiang S, Song L, Peng P, Mu F, Li H, et al. Automated optical inspection of FAST's reflector surface using drones and computer vision. Light Adv Manuf. 2023;4:1-11.

4. Shao J, Du B, Wu C, Zhang L. Tracking objects from satellite videos: A velocity feature based correlation filter. IEEE Trans Geosci Remote Sens. 2019;57:7860-71.

5. Su Y, Liu J, Xu F, Zhang X, Zuo Y. A novel anti-drift visual object tracking algorithm based on sparse response and adaptive spatial-temporal context-aware. Remote Sens. 2021;13:4672.

6. Cao Z, Fu C, Ye J, Li B, Li Y. SiamAPN++: Siamese attentional aggregation network for real-time UAV tracking. In: Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Prague, Czech Republic; 2021. p. 3086-92.

7. Cao Z, Huang Z, Pan L, Zhang S, Liu Z, Fu C. TCTrack: Temporal contexts for aerial tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); New Orleans, LA, USA; 2022. p. 14798-808.

8. Yao L, Fu C, Li S, Zheng G, Ye J. SGDViT: Saliency-Guided Dynamic Vision Transformer for UAV Tracking. arXiv. 2023;arXiv:2303.04378.

9. Javed S, Danelljan M, Khan FS, Khan MH, Felsberg M, Matas J. Visual object tracking with discriminative filters and siamese networks: A survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2022;45:6552-74.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

10. Fu C, Cao Z, Li Y, Ye J, Feng C. Onboard real-time aerial tracking with efficient Siamese anchor proposal network. IEEE Trans Geosci Remote Sens. 2021;60:1-13.

11. Bolme DS, Beveridge JR, Draper BA, Lui YM. Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; San Francisco, CA, USA; 2010. p. 2544-50.

12. Kiani Galoogahi H, Fagg A, Lucey S. Learning background-aware correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision; Venice, Italy; 2017. p. 1135-43.

13. Henriques JF, Caseiro R, Martins P, Batista J. High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell. 2014;37:583-96.

14. Zuo C, Qian J, Feng S, Yin W, Li Y, Fan P, et al. Deep learning in optical metrology: A review. Light Sci Appl. 2022;11:39.

15. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH. Fully-convolutional siamese networks for object tracking. In: Proceedings of the Computer Vision–ECCV 2016 Workshops; Amsterdam, The Netherlands. Part II 14; 2016. p. 850-65.

16. Hua X, Wang X, Rui T, Shao F, Wang D. Light-weight UAV object tracking network based on strategy gradient and attention mechanism. Knowl Based Syst. 2021;224:107071.

17. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv. 2017;arXiv:1706.05587.

18. Tao R, Gavves E, Smeulders AW. Siamese instance search for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. p. 1420-9.

19. Li B, Yan J, Wu W, Zhu Z, Hu X. High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, USA; 2018. p. 8971-80.

20. Guo D, Wang J, Cui Y, Wang Z, Chen S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Seattle, WA, USA; 2020. p. 6269-77.

21. Fu C, Cao Z, Li Y, Ye J, Feng C. Siamese anchor proposal network for high-speed aerial tracking. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA); Xi'an, China; 2021. p. 510-6.

22. Zheng G, Fu C, Ye J, Li B, Lu G, Pan J. Siamese Object Tracking for Vision-Based UAM Approaching with Pairwise Scale-Channel Attention. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Kyoto, Japan; 2022. p. 10486-92.

23. Lou A, Loew M. CFPNet: Channel-wise feature pyramid for real-time semantic segmentation. In: Proceedings of the IEEE International Conference on Image Processing (ICIP); Anchorage, AK, USA; 2021. p. 1894-8.

24. Cao J, Song C, Song S, Xiao F, Zhang X, Liu Z, et al. Robust Object Tracking Algorithm for Autonomous Vehicles in Complex Scenes. Remote Sens. 2021;13:3234.

25. Wang Q, Fan H, Sun G, Cong Y, Tang Y. Laplacian pyramid adversarial network for face completion. Pattern Recognit. 2019;88:493-505.

26. Li Y, Fu C, Ding F, Huang Z, Lu G. 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 (CVPR); Seattle, WA, USA; 2020. p. 11923-32.

27. Wang N, Zhou W, Wang J, Li H. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Nashville, TN, USA; 2021. p.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

1571-80.

28. Wang H, Tang J, Liu X, Guan S, Xie R, Song L. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In: Proceedings of the European Conference on Computer Vision (ECCV); Tel Aviv, Israel; 2022. p. 732-47.

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. p. 770-8.

30. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:84-90.

31. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honolulu, HI, USA; 2017. p. 2117-25.

32. Li B, Wu W, Wang Q, Zhang F, Xing J, Yan J. SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 4282-91.

33. Zhu Z, Wang Q, Li B, Wu W, Yan J, Hu W. Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV); Munich, Germany; 2018. p. 101-17.

34. Wang Q, Zhang L, Bertinetto L, Hu W, Torr PH. Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 1328-38.

35. Zhang L, Gonzalez-Garcia A, Weijer Jvd, Danelljan M, Khan FS. Learning the model update for siamese trackers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); Seoul, Republic of Korea; 2019. p. 4010-9.

36. Zhang Z, Peng H. Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 4591-600.

37. Zhang Z, Peng H, Fu J, Li B, Hu W. Ocean: Object-aware anchor-free tracking. In: Proceedings of the European Conference on Computer Vision (ECCV); Glasgow, UK; 2020. p. 771-87.

38. Cao Z, Fu C, Ye J, Li B, Li Y. Hift: Hierarchical feature transformer for aerial tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); Montreal, QC, Canada; 2021. p. 15457-66.

39. Guo D, Shao Y, Cui Y, Wang Z, Zhang L, Shen C. Graph attention tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Nashville, TN, USA; 2021. p. 9543-52.

40. Sosnovik I, Moskalev A, Smeulders AW. Scale equivariance improves siamese tracking. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); Waikoloa, HI, USA; 2021. p. 2765-74.

41. an B, Peng H, Wu K, Wang D, Fu J, Lu H. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Nashville, TN, USA; 2021. p. 15180-9.

42. Chen Z, Zhong B, Li G, Zhang S, Ji R, Tang Z, et al. SiamBAN: Target-aware tracking with siamese box adaptive network. IEEE Trans Pattern Anal Mach Intell. 2022;45:5158-73.

43. Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking. In: Proceedings of the Computer Vision–ECCV 2016: 14th European Conference; Amsterdam, The Netherlands. Proceedings, Part I 14; 2016. p. 445-61.

44. Li S, Yeung D-Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; San Francisco, CA, USA; 2017; 31.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

45. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Proceedings of the Computer Vision–ECCV 2014: 13th European Conference; Zurich, Switzerland. Proceedings, Part V 13; 2014. p. 740-55.

46. Huang L, Zhao X, Huang K. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Trans Pattern Anal Mach Intell. 2019;43:1562-77.

47. Fan H, Lin L, Yang F, Chu P, Deng G, Yu S, et al. LaSOT: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 5374-83.

48. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211-52.

49. Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization. Ann Statist. 2004;32:56-85.

50. Yu J, Jiang Y, Wang Z, Cao Z, Huang T. Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM International Conference on Multimedia; Amsterdam, The Netherlands; 2016. p. 516-20.

© 2023 The Author(s). | Published by Agro Intelligence Research Press