remote sensing

MDPI

Article

CVAC-Net: Cross-View Alignment and Adaptive Collaboration Network for Aerial–Ground Visual Tracking

Mingli Xie [1,2,3,4], Congjian Li [1,2], Haoran Zhou [1,2], Shixiong Liu [1,2], Yunfeng Wan [1,2,3,4], Shuli Zhu [1,2] and Wensheng Wang [1,2,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; xiemingli24@mails.ucas.ac.cn (M.X.); licongjian@aircas.ac.cn (C.L.); zhouhr@aircas.ac.cn (H.Z.); liusx@aircas.ac.cn (S.L.); wanyunfeng24@mails.ucas.ac.cn (Y.W.); zhusl@aircas.ac.cn (S.Z.)
[2] Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: wangws@aircas.ac.cn

Highlights

What are the main findings?

- We propose CVAC-Net, a novel framework for aerial–ground visual tracking that aligns, filters, and fuses complementary cross-view features.
- CVAC-Net outperforms several state-of-the-art trackers on UGVT in terms of SR, NPR, and PR, with further validation on Griffin.

What are the implications of the main findings?

- Adaptive cross-view interaction enables the use of complementary aerial and ground information under changing tracking conditions without camera calibration.
- Our scene-adaptive aerial–ground collaboration enables persistent object tracking in multi-platform remote sensing systems, with demonstrated robustness in challenging scenarios.

Abstract

Aerial–ground visual tracking facilitates dynamic remote sensing and scene monitoring by integrating complementary observations from unmanned aerial vehicle (UAV) and ground cameras, allowing one view to compensate when the target is poorly observed or temporarily invisible in the other. However, substantial cross-view variations in target appearance, scale, and geometry make it difficult to establish reliable target correspondences and effectively fuse complementary cues. To address these challenges, we propose CVAC-Net, a Cross-View Alignment and Adaptive Collaboration Network for aerial–ground visual tracking. Specifically, CVAC-Net uses cross-view contrastive learning to reduce semantic differences and a learned spatial offset to adjust the relative positions of aerial- and ground-view target features within their normalized search crops. Next, it uses the similarity between the main-view template and the aligned auxiliary-view search features to generate an importance map that reweights redundant and target-irrelevant regions, thereby improving the reliability of feature aggregation. Finally, its view-specific branches select and weight the top-K encoder outputs to integrate complementary cross-view features. Extensive experiments on the UAV–ground visual tracking (UGVT) benchmark show that CVAC-Net improves the baseline by 3.2, 3.1, and 1.7 percentage points in success rate (SR), normalized precision rate (NPR), and precision rate (PR), respectively. We further evaluate CVAC-Net using a two-dimensional (2D) single-object tracking protocol derived from Griffin using its predefined training and validation splits, providing supplementary validation on a simulated aerial–ground distribution.

Check for updates

Academic Editor: Peter Hofmann

Received: 10 August 2026
Revised: 25 September 2026
Accepted: 7 October 2026
Published: 13 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: aerial–ground visual tracking; cross-view alignment; adaptive collaboration

1. Introduction

Visual object tracking uses target information in the first frame to estimate its states (e.g., position and scale) in subsequent frames [1]. With the widespread utilization of this technology, it has been employed for numerous Earth observation applications, including environmental monitoring [2–4], disaster management [5,6], and precision agriculture [7,8]. Most existing trackers operate on a single video stream, where target appearance and motion are modeled within a consistent observation space [9–15]. This setting allows the tracker to focus on temporal target localization within a single sequence. However, as illustrated in Figure 1b, when the target is occluded or disappears from the field of view, a single-view tracker lacks an independent observation from which to recover the missing target information [16].

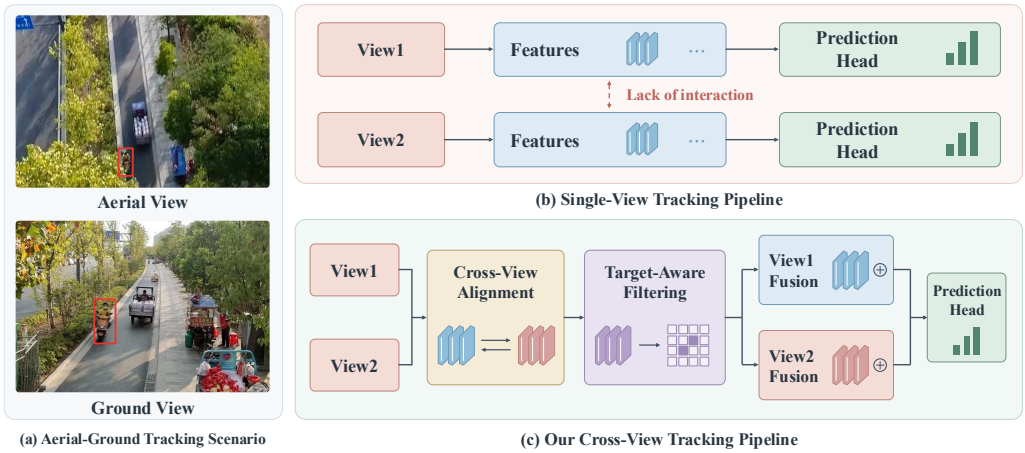


Figure 1. Motivation and information flow of aerial–ground visual tracking. (**a**) Paired aerial and ground observations of the same target, marked by red bounding boxes. (**b**) Independent single-view tracking without cross-view information exchange. (**c**) CVAC-Net connects the two tracking branches through cross-view alignment, target-aware filtering, and view-adaptive fusion, allowing complementary information to support target prediction in both views.

To mitigate this limitation, researchers have proposed aerial–ground visual tracking for the same target, which combines observations from aerial and ground cameras. The ground view provides detailed appearance information but lacks a comprehensive view of the target. In contrast, the aerial view captures global motion over a wide field of view but provides fewer appearance details. By combining the two views, the tracking system reduces the effects of occlusion, out-of-view motion, or image degradation that may affect only one view. Sun et al. [17] first formulate the UAV-Ground visual tracking task and release the UGVT benchmark. Based on this dataset, multi-view collaborative learning (MvCL) [17] uses dot-product cross-attention to locate regions with similar target appearance in the other view, establishing cross-view association and enabling complementary feature fusion. However, establishing cross-view correspondence alone does not ensure effective collaboration, since the transferred information may not always be reliable or suitable for the current tracking condition. This leaves the effective exploitation of cross-view complementarity under heterogeneous and changing observations unresolved.

This problem remains difficult to address for three main reasons. First, large differences in viewpoint, appearance, scale, and target position create a huge cross-view gap. For example, as shown in Figure 1a, the red-boxed rider is a small, blurred, and visually

compressed patch in the aerial image, whereas in the ground image it appears much larger and exhibits a more recognizable silhouette. This gap makes it difficult to identify corresponding target information and combine features from aerial and ground views. Second, complementarity does not mean that every feature from the other view is useful. The search region from the other view also contains background clutter and target-irrelevant content. Directly fusing the whole region may transfer this redundant information and cause cross-view interference. The interference may weaken the target response and increase localization errors. Third, different views require different complementary cues, whose usefulness also varies with the tracking condition. Fine-grained ground-view appearance is more useful for small or poorly resolved aerial targets, whereas wider aerial context better supports ground tracking under occlusion or limited coverage. A fixed fusion strategy may therefore struggle to accommodate these changing cross-view relations.

To address these challenges, we propose CVAC-Net, a Cross-View Alignment and Adaptive Collaboration Network for aerial–ground visual tracking, whose general tracking pipeline is illustrated in Figure 1c. First, inspired by Contrastive Language–Image Pre-training (CLIP)'s contrastive alignment of image and text representations [18], we apply cross-view contrastive learning to reduce semantic differences between aerial and ground search features and then use a learned spatial offset to adjust the relative positions of the target features from the two views. Second, we compute the feature similarity between the main-view template and the aligned auxiliary-view search region to generate an importance map and use this map to reweight redundant and target-irrelevant regions before cross-view fusion. Third, inspired by sparsely gated mixture-of-experts architectures [19], we build view-specific multi-encoder branches with an end-to-end routing network. The router selects and fuses the top-K encoder outputs, while a spatial gate adjusts their contribution to the main-view feature.

We conducted comprehensive experiments to evaluate the performance of our network on UGVT [17] and Griffin [20] datasets. On UGVT, CVAC-Net improves the TransT baseline by 3.2, 3.1, and 1.7 percentage points in SR, NPR, and PR, respectively. On Griffin sequences with low visibility or small targets, the corresponding improvements are 2.43, 1.92, and 1.12 percentage points. These results highlight the potential of CVAC-Net for future multi-platform remote sensing systems, in which aerial–ground collaboration can support persistent object tracking and fine-grained scene monitoring in complex and dynamic environments.

The main contributions of this work are summarized as follows:

1. We propose CVAC-Net, a cross-view collaboration framework for aerial–ground visual tracking. By sharing search-region features between the two tracking branches, CVAC-Net exploits both target and contextual information from the complementary view without requiring explicit camera calibration or coordinate transformation.
2. We develop a progressive cross-view interaction mechanism, including two-stage alignment and target-aware filtering, that improves the reliability of auxiliary information before fusion. Two-stage alignment uses cross-view contrastive learning to reduce semantic differences and a learned spatial offset to adjust the relative positions of aerial- and ground-view target features, while target-aware filtering reweights redundant and target-irrelevant regions using the main-view template.
3. We design a collaboration mechanism to account for differences between aerial and ground tracking. Separate branches learn view-specific feature transformations, while a routing network selects encoder outputs and a spatial gate adjusts their contribution during fusion.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the proposed CVAC-Net and details its cross-view collaboration

mechanism. Section 4 presents the experimental setup and evaluates the proposed method through comparisons, ablation studies, and qualitative analysis. Section 5 discusses the method and its limitations. Finally, Section 6 summarizes the main findings and discusses future research directions.

2. Related Work

2.1. Visual Object Tracking

Visual object tracking aims to locate the position of a target in each frame of a video based on its initial annotation. In remote sensing videos, this task is more challenging because the targets are usually small and may be affected by camera movement, scale changes, occlusion, and complex backgrounds. Existing research on satellite and unmanned aerial vehicle tracking improves tracking performance by leveraging motion information, temporal signals, and stronger target features [21–29].

Most single-object trackers focus on enhancing the contrast between the target template and the current search area. Some methods use direct template matching [11,12,30], while others distinguish the target from the background by online updating the model [9,10,31,32]. Recently, trackers based on Transformers utilize the attention mechanism to enhance the interaction between the template and the search features [13,33–35]. Although these designs are different, they all rely on a single video stream. When the target is severely occluded or moves out of the camera's field of view, the tracker cannot obtain other observational information to improve tracking performance.

2.2. Multi-View Tracking

Multi-view tracking uses observations from several cameras so that one view can support another. Multi-drone trackers share templates or target features and model the relation between aerial views [16,36,37]. These methods are designed for cameras on the same platform type and do not directly address the differences between aerial and ground observations. Other multi-camera methods project features onto a common ground plane or into three-dimensional (3D) space [38–41]. Such a projection provides a clear geometric relation between views, but it requires camera calibration. This requirement is difficult to satisfy when the UAV pose changes during tracking.

Aerial–ground visual tracking observes the same target through both drones and ground cameras. UGVT introduces paired aerial and ground videos and uses the cross-attention method to achieve target association and feature fusion [17]. Recent work has introduced language-guided cross-view interaction for this task [42]. Griffin investigated three-dimensional detection and multi-object tracking through aerial–ground collaboration [20]. Although its tasks and annotation methods are different from two-dimensional single-object tracking, its synchronized vehicle and drone images provide beneficial supplementary conditions for cross-view evaluation.

The key feature of CVAC-Net is that it directly realizes the interaction between the features of the search region in the air and on the ground without the need for camera calibration or explicit coordinate transformation. This interaction process consists of three stages: feature alignment, template-guided reweighting of auxiliary regions, and adaptive cross-view fusion. These stages jointly establish the cross-view correspondence, weight auxiliary information according to its relevance, and dynamically adjust the weight contribution of each view according to the changes in tracking conditions.

3. Proposed Method

This section presents CVAC-Net for aerial–ground visual tracking. Given paired observations from UAV-borne and ground-based cameras, CVAC-Net enhances each main-view

search representation through cross-view alignment and adaptive collaboration before target localization. The method is designed around two central requirements: establishing reliable correspondence across heterogeneous views and integrating complementary cues without introducing target-irrelevant interference. We first formulate the dual-branch search-region sharing strategy, then detail the two-stage alignment, target-aware filtering, and adaptive collaboration modules. Finally, we summarize the inference process of CVAC-Net.

3.1. Dual-Branch Siamese-like Framework

In multi-view tracking, including multi-drone tracking [16,36,37] and aerial–ground visual tracking [17], a straightforward baseline is to run two independent single-view trackers. The i-th tracker computes a within-view response as

$$f_{ii}(z_i, x_i) = f(\phi(z_i), \phi(x_i)), \quad i = 1, 2, \tag{1}$$

where z_i and x_i are the template and search images, and $\phi(\cdot)$ denotes the shared feature extractor. Despite being simple and scalable, the single-view tracker lacks explicit cross-view interaction, thereby failing to leverage complementary information.

To incorporate complementary information from different views, recent methods [16,36,37] adopt template sharing to propagate target appearance across views, enabling cross-view matching. Specifically, ASNet [36] and TransMDOT [16] compute

$$\begin{cases} f_{ji}(z_j, x_i) = f\big(\phi(z_j), \phi(x_i)\big), \\ f_{ij}(z_i, x_j) = f\big(\phi(z_i), \phi(x_j)\big), \\ \qquad \text{s.t. } i, j \in \{1, 2\},\ i \neq j, \end{cases} \tag{2}$$

and CRM [37] further suppresses heterogeneous information between the shared template and the search region to improve the exploitation of cross-view information. However, the template contains little contextual background and is often fixed or weakly updated, failing to provide sufficient spatio-temporal information.

To fully leverage cross-view complementarity, we propose a search region sharing strategy that captures both the target and its local context. However, effective exploitation faces three primary challenges. (1) Cross-view misalignment: Significant viewpoint discrepancies cause severe spatial non-correspondence, hindering direct feature matching across views. (2) Noise interference: The expanded search regions inevitably introduce background clutter and redundancy, which distract the model from the actual target. (3) Dynamic view relationships: In diverse tracking scenarios, the relationships between views are complex and time-varying, making it difficult to effectively model the dynamic dependencies for information fusion. To address these issues, CVAC-Net aligns and integrates cross-view search-region features before the final tracking head:

$$\begin{cases} \hat{x}_i = \mathrm{CVAC}\big(\phi(x_i), \phi(x_j), \phi(z_i)\big), \\ f_i(z_i, x_i, x_j) = f\big(\phi(z_i), \hat{x}_i\big), \\ \qquad \text{s.t. } i, j \in \{1, 2\},\ i \neq j, \end{cases} \tag{3}$$

where $\mathrm{CVAC}(\cdot)$ denotes cross-view alignment and adaptive collaboration in CVAC-Net, and the template feature $\phi(z_i)$ is matched against a fused search representation computed from $\phi(x_i)$, $\phi(x_j)$, and $\phi(z_i)$.

3.2. CVAC-Net Architecture

CVAC-Net, as shown in Figure 2, contains three main components: two-stage alignment (TSA), target-aware filtering (TAF), and adaptive collaboration. The latter is implemented by view-adaptive fusion (VAF). Together, these components implement cross-view alignment and adaptive collaboration between two tracking branches that share search-region features. We first use two-stage alignment to reduce semantic differences and compensate for spatial offsets between aerial- and ground-view target features in their normalized search crops. Target-aware filtering and VAF then produce fused features for each tracking branch.

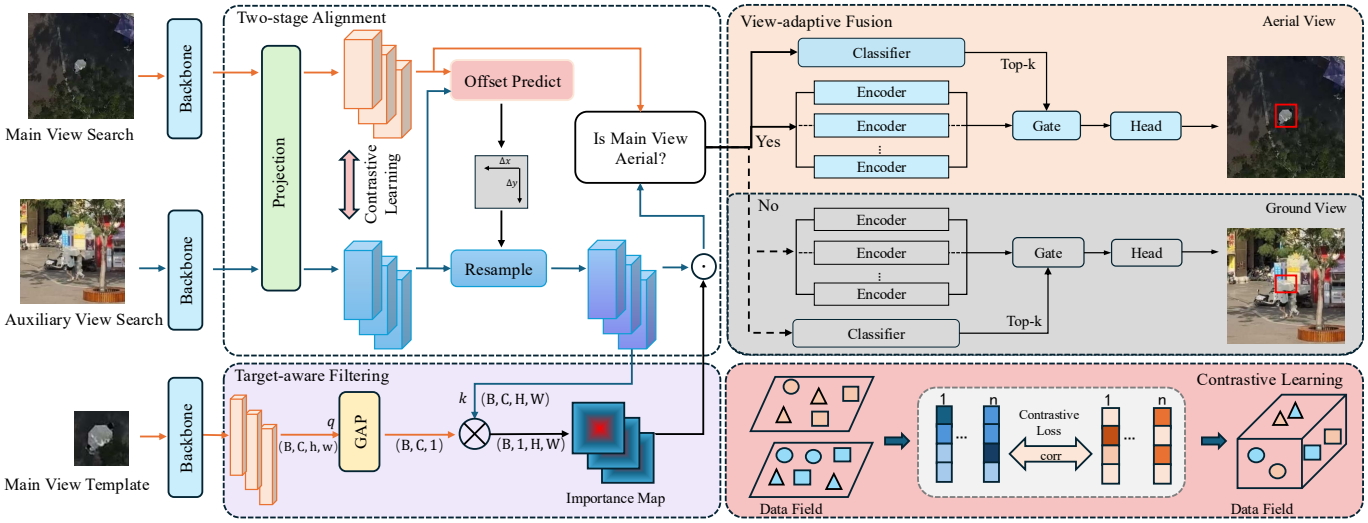


Figure 2. Overall architecture of CVAC-Net. The network extracts paired aerial–ground template and search features with a shared backbone, performs two-stage alignment, target-aware filtering, and view-adaptive fusion, and then feeds the fused representation into the final tracking head. Orange and blue feature maps and arrows denote the main and auxiliary views, respectively. Dashed outlines delimit modules; dashed arrows indicate the ground-view branch. Red boxes mark tracked targets. GAP denotes global average pooling.

3.2.1. Two-Stage Alignment

Taking the aerial-view branch as an example, its inputs are the aerial template I_z^a and the paired aerial and ground search images (I_x^a, I_x^g). The shared backbone $\phi(\cdot)$ first extracts search features, after which the projection network $g(\cdot)$ maps both views into the same feature space:

$$X^a = \mathrm{norm}_2(g(\phi(I_x^a))), \qquad X^g = \mathrm{norm}_2\Big(g\Big(\phi(I_x^g)\Big)\Big). \tag{4}$$

The projection network is

$$g(X) = \rho\Big(\mathrm{BN}_2\Big(W_2^{3\times3} * \mathcal{D}\Big(\rho\Big(\mathrm{BN}_1(W_1^{1\times1} * X)\Big)\Big)\Big)\Big), \tag{5}$$

where $*$ denotes convolution, ρ is the rectified linear unit (ReLU), BN denotes batch normalization, $\mathcal{D}$ is dropout, and norm_2 normalizes the channel vector at each spatial location to unit length. To reduce the semantic gap between views, paired aerial and ground features from the same scene are treated as positives and features from different pairs in the mini-batch as negatives. For the k-th pair, $\bar{X}_k^v = \mathrm{GAP}(X_k^v)$ for $v \in \{a, g\}$, and

$$S_{km} = \frac{\mathrm{sim}(\bar{X}_k^a, \bar{X}_m^g)}{\tau}, \tag{6}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity and τ is the temperature. We optimize correspondence in both directions with the symmetric information noise-contrastive estimation (InfoNCE) loss [43]:

$$\mathcal{L}_{a\to g} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(S_{kk})}{\sum_{m=1}^{B}\exp(S_{km})}, \tag{7}$$

$$\mathcal{L}_{g\to a} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(S_{kk})}{\sum_{m=1}^{B}\exp(S_{mk})}, \tag{8}$$

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{2}\left(\mathcal{L}_{a\to g} + \mathcal{L}_{g\to a}\right). \tag{9}$$

Semantic alignment alone does not account for spatial offsets between aerial- and ground-view target features within their normalized search crops. At inference, each view's search crop is centered on its own target estimate from the previous frame. Errors in these estimates can shift the target within the normalized crops and change the relative positions of the two views' target features. Independent center jitter of the two search crops during training exposes the model to analogous offsets. We therefore use a two-dimensional translation as a low-degree-of-freedom correction of the auxiliary feature, rather than attempting to recover a general aerial–ground warp without camera calibration. For main-view feature X^m and auxiliary-view feature X^u, the lightweight predictor is

$$\delta = s\tanh\Big(W_2^{1\times1} * \rho\Big(W_1^{1\times1} * \mathrm{GAP}([X^m, X^u])\Big)\Big), \tag{10}$$

where $[\cdot,\cdot]$ denotes channel concatenation, s is a learned scale, and $\delta = (\Delta x, \Delta y) \in \mathbb{R}^{B\times2}$ contains the horizontal and vertical offsets Δx and Δy shown in Figure 2, respectively. The aligned auxiliary feature is then

$$\widehat{X}^u_{:,h,w} = \mathrm{bilinear}(X^u, q_{hw} + \delta), \tag{11}$$

where q_{hw} is the regular sampling grid. This differentiable sampling follows the spatial-transformer formulation [44]. Restricting the transform to translation preserves the relative layout of the feature map and uses the predicted offset as a correction within the normalized search crop.

3.2.2. Target-Aware Filtering

After alignment, the auxiliary search region still contains background and view-specific interference. TAF uses the first-frame template of the current main-view branch to weight target-correlated auxiliary responses. For the aerial branch, the aerial view is the main view and the ground view is auxiliary. Their roles are reversed for the ground branch. Let $Z^m \in \mathbb{R}^{B\times C\times H_z\times W_z}$ be the main-view template feature and $\widehat{X}^u \in \mathbb{R}^{B\times C\times H\times W}$ the aligned auxiliary search feature. We first form a template descriptor

$$t^m = \mathrm{GAP}(Z^m) \in \mathbb{R}^{B\times C}. \tag{12}$$

The correlation response at each auxiliary location is

$$R^m_{h,w} = \frac{(t^m)^{\top}\widehat{X}^u_{:,h,w}}{\sqrt{C}}, \qquad R^m \in \mathbb{R}^{B\times1\times H\times W}, \tag{13}$$

and the filtered feature is

$$\widetilde{X}^u = R^m \odot \widehat{X}^u, \tag{14}$$

where $\odot$ denotes spatial weighting broadcast across channels. The response R^m directly weights auxiliary locations according to their similarity to the current main-view template before cross-view fusion.

3.2.3. Adaptive Collaboration

To model view-dependent complementarity, adaptive collaboration is implemented with view-specific fusion branches that share the same architecture but use independent parameters. The active branch is determined by the main view. For the aerial branch, $m = a$ and $u = g$, while the ground branch exchanges these roles. Let X^m be the main-view search feature and $\widetilde{X}^u$ the filtered auxiliary feature from the preceding stage. Their concatenation $S = [X^m, \widetilde{X}^u]$ is the input to the active fusion branch Φ^m, which produces $F^m_{\text{fuse}} = \Phi^m(S)$. The following equations describe the active branch. Its router, encoders, and gate have separate parameters for the aerial and ground branches.

To account for the current pair of search regions, a lightweight router pools S and assigns a score to each of the N encoders:

$$q = \text{softmax}\left(\frac{W_r^{1\times1} * \text{GAP}(S)}{T_r}\right), \tag{15}$$

where $W_r^{1\times1}$ is the routing convolution, q_i is the weight assigned to encoder i before selection, and $T_r = 3$ is the routing temperature. Each encoder maps the paired features to a candidate representation through a 1×1–3×3–1×1 convolutional bottleneck:

$$E_i(S) = W_{i,3}^{1\times1} * \rho\left(W_{i,2}^{3\times3} * \rho\left(W_{i,1}^{1\times1} * S\right)\right). \tag{16}$$

Here, ρ is ReLU, and each $E_i(S)$ has the same channel and spatial dimensions as X^m. Using several encoders gives the branch multiple learnable ways to combine the main and auxiliary features. To form the output, we retain the K largest router weights, denoted by $\mathcal{T}_K = \text{TopK}(q, K)$, and renormalize them:

$$\pi_i = \frac{q_i}{\sum_{j\in\mathcal{T}_K} q_j + \epsilon} \quad (i \in \mathcal{T}_K), \qquad Y = \sum_{i\in\mathcal{T}_K} \pi_i E_i(S), \tag{17}$$

where ϵ prevents division by zero. Encoders outside $\mathcal{T}_K$ have zero mixture weight. The observed utilization and contributions of these encoders are examined in Section 4.4.6.

Finally, a spatial gate combines the candidate cross-view representation with the original main-view feature:

$$A = \sigma\left(W_g^{1\times1} * [X^m, Y]\right), \qquad F^m_{\text{fuse}} = A \odot Y + (1 - A) \odot X^m. \tag{18}$$

Here, σ is the sigmoid. The single-channel map $A \in \mathbb{R}^{B\times1\times H\times W}$ varies across spatial locations, allowing the branch to combine cross-view information with the main-view representation locally.

3.2.4. Training Objective

Following TransT [13], the tracking objective contains the cross-entropy classification loss $\mathcal{L}_{\text{cls}}$, the ℓ_1 box-regression loss $\mathcal{L}_1$, and the generalized intersection-over-union (GIoU) loss $\mathcal{L}_{\text{GIoU}}$. We add the cross-view alignment loss in Equation (9):

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_1 + \lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{align}}(e)\mathcal{L}_{\text{align}}. \tag{19}$$

The coefficients $\lambda_{\text{cls}} = 8.334$, $\lambda_1 = 5$, and $\lambda_{\text{GIoU}} = 2$ follow the original TransT training objective so that the inherited tracking losses retain their established balance. The added alignment coefficient is $\lambda_{\text{align}}(e) = 0.5 \times 0.8^e$, where e is the training epoch. Cross-view alignment is emphasized at the beginning, when the newly introduced projection and collaboration modules must establish correspondence. Its coefficient then decreases so that classification and localization increasingly determine the later optimization. The coefficient and schedule controls are reported in Section 4.4.5.

3.3. Tracking with CVAC-Net

After completing cross-view alignment and adaptive collaboration, CVAC-Net utilizes the fused search representation and its corresponding template features for target localization. For each branch, the main-view search features are enhanced by filtering and combining the auxiliary-view features, and the resulting representation is used to predict the target state. This tracking process keeps the matching stage simple while fully conducting cross-view collaboration.

The detailed inference procedure of CVAC-Net is summarized in Algorithm 1. Here, f_t^a and f_t^g denote the synchronized aerial and ground frames at time t, while s_t^a and s_t^g are the corresponding target states. The operator $\text{crop}(\cdot)$ extracts the template or search region, $\phi(\cdot)$ is the shared backbone, and $\tilde{z}$ and $\tilde{x}$ denote the extracted features. The two calls to $\text{CVAC}(\cdot)$ exchange the main and auxiliary roles and produce F_{fuse}^a and F_{fuse}^g. Finally, $\text{Head}(\cdot)$ denotes the TransT prediction head that estimates the updated target state in each view.

Algorithm 1: Tracking Algorithm of CVAC-Net.

Input: Video frames $\{f_t^a\}_{t=1}^L$ of aerial view and $\{f_t^g\}_{t=1}^L$ of ground view, with initial target states s_1^a, s_1^g
Output: Target states $\{s_t^a, s_t^g\}_{t=1}^L$
for $t \leftarrow 1$ **to** L **do**
  if $t = 1$ **then**
    Obtain templates z^a, z^g in f_1^a, f_1^g based on s_1^a, s_1^g
    $z_{\text{crop}}^a, z_{\text{crop}}^g : 128 \times 128 \times 3 \leftarrow \text{crop}(z^a, z^g)$
    $\tilde{z}^a, \tilde{z}^g : 16 \times 16 \times 1024 \leftarrow \phi(z_{\text{crop}}^a, z_{\text{crop}}^g)$
  else
    Obtain search regions x_t^a, x_t^g in f_t^a, f_t^g based on s_{t-1}^a, s_{t-1}^g
    $x_{\text{crop}}^a, x_{\text{crop}}^g : 256 \times 256 \times 3 \leftarrow \text{crop}(x_t^a, x_t^g)$
    $\tilde{x}^a, \tilde{x}^g : 32 \times 32 \times 1024 \leftarrow \phi(x_{\text{crop}}^a, x_{\text{crop}}^g)$
    `// Cross-view alignment and fusion`
    $F_{\text{fuse}}^a \leftarrow \text{CVAC}(\tilde{x}^a, \tilde{x}^g, \tilde{z}^a)$
    $F_{\text{fuse}}^g \leftarrow \text{CVAC}(\tilde{x}^g, \tilde{x}^a, \tilde{z}^g)$
    `// CVAC-Net tracking head`
    $s_t^a \leftarrow \text{Head}(F_{\text{fuse}}^a, \tilde{z}^a)$
    $s_t^g \leftarrow \text{Head}(F_{\text{fuse}}^g, \tilde{z}^g)$

4. Experiments

This section evaluates CVAC-Net on paired aerial and ground imagery acquired under different platform and annotation settings. UGVT [17] provides the primary benchmark, and a two-dimensional single-object tracking version of Griffin [20] provides supplementary evaluation on the full validation set and selected challenging scenarios. We first describe the implementation details, datasets, and metrics. We then compare CVAC-Net with state-

of-the-art trackers and conduct ablation studies. Finally, visualizations demonstrate the effectiveness of CVAC-Net under several representative tracking challenges.

4.1. Implementation Details

The experiments were conducted using Python 3.8.20 and PyTorch 1.13.1 with CUDA runtime 11.6. CVAC-Net uses a ResNet-50 backbone [45]. The template and search crops are resized to 128×128 and 256×256, respectively. Training uses search-center and scale jitter factors of 3 and 0.25, paired grayscale conversion with probability 0.05, and per-image brightness jitter within $[0.8, 1.2]$ before normalization. Repeated full-model runs and additional controls use AdamW [46] for 20 epochs with a batch size of 24, learning rates of 5×10^{-6} for non-backbone parameters and 1×10^{-6} for the backbone, and weight decay 10^{-5}. The learning rates remain constant over these 20 epochs. We set $\tau = 0.07$, $N = 9$, and $K = 4$.

CVAC-Net uses pretrained TransT weights to initialize its backbone and prediction components. Its newly added cross-view modules are randomly initialized, and the complete network is then trained on the UGVT training split.

4.2. Datasets and Evaluation Metrics

Datasets. We use UGVT [17] as the main dataset for aerial–ground visual tracking. Because the official training/test division is not public, we randomly divide its 210 paired sequences into 165 training pairs and 45 test pairs using random seed 1. The same split is used throughout the experiments.

UGVT is a unified multi-view benchmark for training and evaluating aerial–ground visual tracking. It contains 210 paired sequences captured simultaneously by a ground camera (Sony sensor, 1–4 m) and a UAV camera (DJI Phantom 4, 10–50 m). The UAV view is mostly moving, while the ground view is usually static, with some sequences introducing intentional ground-camera motion. UGVT includes 21 target classes and 10 challenge attributes. More details are available in [17].

Griffin [20] is a CARLA–AirSim co-simulation dataset originally developed for cooperative aerial–ground 3D detection and tracking. We derive a paired 2D single-object tracking protocol from its synchronized vehicle and drone videos, shared identities, camera calibration, and 3D cuboid annotations. Each cuboid is projected into both image planes using the official view-specific camera parameters, and the projected extrema form an axis-aligned box clipped to the image boundary. Camera calibration is used only to derive these boxes and is not provided to the tracker at inference. Only continuous segments with valid boxes in both synchronized views are retained. The official scene-disjoint split yields 303 training pairs from 73 scenes (22,706 frames per view) and 67 validation pairs from 16 scenes (4863 frames per view).

We additionally define a difficult subset using two predeclared conditions. A sequence is assigned to the low-visibility category if target visibility is below 60% in at least 10% of its frames in either view, and to the small-target category if the mean box area is no greater than 400 pixels^2 in either view. Their union contains 26 paired validation sequences, referred to as Difficult26. All 67 validation pairs and Difficult26 use the same one-pass evaluation (OPE) protocol.

Evaluation metrics. We use the OPE protocol with three common metrics [47,48]: SR, PR, and NPR.

- SR is the area under the success plot obtained by varying the intersection-over-union threshold between predicted and ground-truth boxes.
- PR is the fraction of frames whose predicted center is within 20 pixels of the ground-truth center.

- NPR is the fraction of frames whose center error, normalized by the ground-truth width and height, is below 0.20.

4.3. Comparison with State-of-the-Art Trackers

In this section, we assess the performance of CVAC-Net compared with mainstream tracking methods, including SiamBAN [30], TransT [13], CTTrack [14], OSTrack [35], ATOM [9], DiMP50 [10], SuperDiMP [49], ToMP50 [32], MixFormer [34], KeepTrack [15] and the MvCL-based method [17].

Except for the MvCL results quoted from [17], the comparison rows are evaluated on the same 45-pair test set with the same OPE protocol using retained tracker outputs. CVAC-Net and the TransT baseline are fine-tuned on the corresponding 165-pair training split from pretrained TransT weights. The other single-view trackers are evaluated from available pretrained checkpoints without additional training on this split. The comparison rows share the 45-pair test protocol.

4.3.1. Overall Performance Evaluation

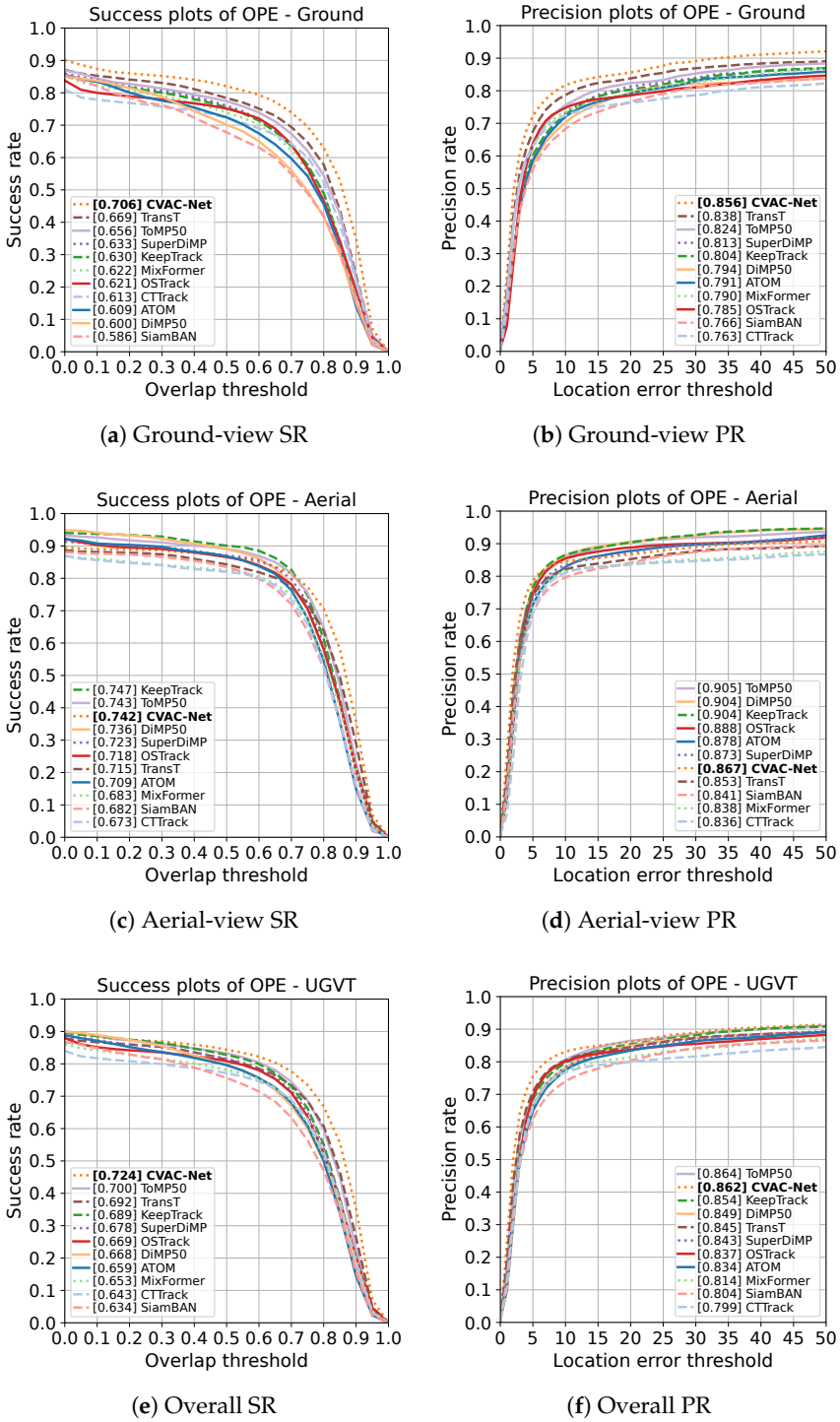
Table 1 reports the results for the ground view, aerial view, and the overall UGVT benchmark. The official implementation and trained models of the MvCL variants are not publicly available, and their exact test-sequence identities cannot be matched to ours. Their scores are therefore reported from [17] as contextual references rather than reproduced under our unified test pipeline, and they are excluded from the ranking. The CVAC-Net row reports the descriptive mean and sample standard deviation of three full-model results. On the ground view, its mean SR, NPR, and PR scores are 0.706, 0.802, and 0.856, exceeding TransT by 3.7, 3.3, and 1.8 percentage points, respectively. For the aerial view, the corresponding means are 0.742, 0.863, and 0.867, exceeding TransT by 2.7, 3.0, and 1.5 percentage points, respectively. CVAC-Net ranks first on all three ground-view metrics among the compared trackers, while its aerial-view mean SR is below KeepTrack and ToMP50.

Table 1. Tracking performance comparison on UGVT under ground view, aerial view, and overall. Metrics include SR, NPR, and PR under the OPE protocol. The best three scores are highlighted in red, green, and blue, respectively.

Tracker	Ground View			Aerial View			UGVT		
	SR	NPR	PR	SR	NPR	PR	SR	NPR	PR
TransT	0.669	0.769	0.838	0.715	0.833	0.853	0.692	0.801	0.845
SiamBAN	0.586	0.659	0.766	0.682	0.822	0.841	0.634	0.740	0.804
DiMP50	0.600	0.688	0.794	0.736	0.877	0.904	0.668	0.782	0.849
ToMP50	0.656	0.752	0.824	0.743	0.881	0.905	0.700	0.817	0.864
CTTrack	0.613	0.709	0.763	0.673	0.806	0.836	0.643	0.757	0.799
KeepTrack	0.630	0.734	0.804	0.747	0.893	0.904	0.689	0.813	0.854
SuperDiMP	0.633	0.739	0.813	0.723	0.862	0.873	0.678	0.800	0.843
OSTrack	0.621	0.735	0.785	0.718	0.856	0.888	0.669	0.796	0.837
MixFormer	0.622	0.717	0.790	0.683	0.814	0.838	0.653	0.766	0.814
ATOM	0.609	0.705	0.791	0.709	0.860	0.878	0.659	0.783	0.834
MvCL-TransT †	0.637	0.750	0.844	0.623	0.740	0.786	0.630	0.745	0.815
MvCL-DiMP50 †	0.632	0.745	0.866	0.658	0.783	0.839	0.645	0.764	0.853
CVAC-Net (Ours)	0.706 ±0.008	0.802 ±0.010	0.856 ±0.010	0.742 ±0.010	0.863 ±0.015	0.867 ±0.014	0.724 ±0.005	0.833 ±0.007	0.862 ±0.006

Note: The gains over TransT are 3.2, 3.1, and 1.7 percentage points for overall SR, NPR, and PR. Results marked with † are reported from [17] and provided for contextual reference only. They are excluded from ranking and color highlighting.

Overall, CVAC-Net obtains UGVT SR, NPR, and PR of 0.724, 0.833, and 0.862, respectively. These means exceed those of TransT by 3.2, 3.1, and 1.7 percentage points. CVAC-Net has the highest overall SR and NPR among the compared trackers, while ToMP50 has the highest overall PR. Figure 3 shows the corresponding mean success and precision curves.



(**a**) Ground-view SR (**b**) Ground-view PR

(**c**) Aerial-view SR (**d**) Aerial-view PR

(**e**) Overall SR (**f**) Overall PR

Figure 3. Success and precision curves on UGVT. The CVAC-Net curves are the equal mean of the three complete-model results used in Table 1. Ground, aerial, and overall curve scores agree with that table.

4.3.2. Inference Efficiency

Table 2 compares TransT and CVAC-Net over all 45 UGVT test sequences on an NVIDIA GeForce RTX 4090 graphics processing unit (GPU; NVIDIA Corporation, Santa

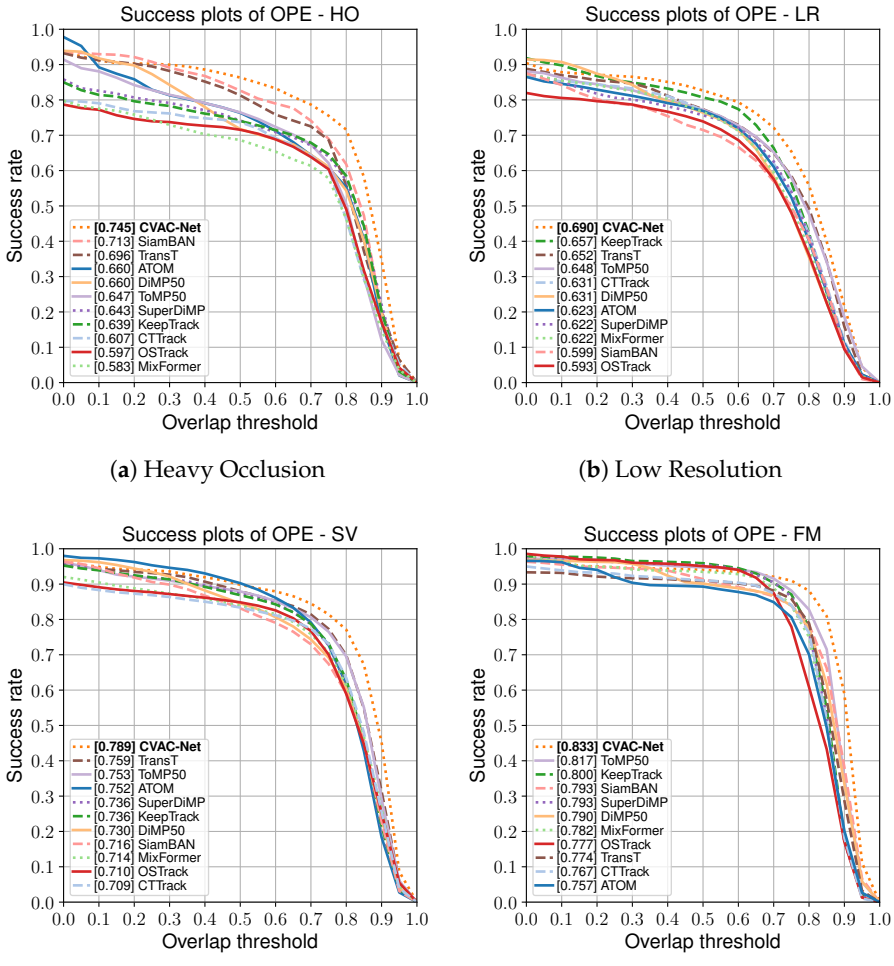
Clara, CA, USA). CVAC-Net achieves 27.37 synchronized-pair updates per second, equivalent to 54.74 tracked-view outputs per second. Its alignment and adaptive collaboration modules require more computation and memory than TransT.

Table 2. Two-view inference efficiency on UGVT. Profiler-covered GFLOPs (billions of floating-point operations) include both view updates. M denotes millions of parameters, and MiB denotes mebibytes. Memory is the maximum sampled GPU memory occupied by each inference process across all 45 test sequences, measured at 0.5-s intervals after warm-up on NVIDIA GeForce RTX 4090 GPUs. Pairs/s counts synchronized frame pairs; FPS (frames per second) counts the two tracked-view outputs per pair. Speeds are mean per-sequence rates excluding first-frame initialization.

Method	Parameters (M)	GFLOPs	Memory (MiB)	Pairs/s	FPS
TransT	23.016	92.043	1982	42.21	84.43
CVAC-Net	104.410	322.933	2340	27.37	54.74

4.3.3. Challenge-Based Evaluation and Analysis

Figure 4 reports the SR results of CVAC-Net under heavy occlusion, low resolution, scale variation, and fast motion. Compared with TransT, CVAC-Net improves SR by 4.9, 3.8, 2.9, and 5.9 percentage points under these four attributes, respectively. It achieves the highest SR in all four settings. The complete-test-set scores are reported in Table 1.



(**a**) Heavy Occlusion (**b**) Low Resolution

(**c**) Scale Variation (**d**) Fast Motion

Figure 4. Attribute-based success plots of CVAC-Net on UGVT. Each selected sequence-view contributes equally to its attribute plot, and the plotted scores include all frames of those sequences.

4.3.4. Supplementary Validation on Griffin

Table 3 reports validation on the derived Griffin 2D protocol. Alongside TransT, we compare CVAC-Net with ToMP50 [32], KeepTrack [15], CTTrack [14], and DiMP50 [10]. On all 67 eligible pairs, CVAC-Net obtains mean SR, NPR, and PR scores of 0.7574, 0.8871, and 0.8984, respectively. It ranks first among the selected trackers and exceeds TransT by 2.18, 0.80, and 1.38 percentage points in the three metrics.

On Difficult26, CVAC-Net obtains mean SR, NPR, and PR scores of 0.6959, 0.8331, and 0.8673, respectively, exceeding TransT by 2.43, 1.92, and 1.12 percentage points. These results provide supplementary evidence on a simulated aerial–ground distribution.

Table 3. Retained-trajectory evaluation on the derived Griffin 2D single-object tracking (SOT) protocol. Difficult26 is the pre-declared low-visibility-or-small-target subset of the same full-set trajectories. Each score first averages sequences equally within a view and then averages the vehicle and drone views equally. Gains in the text are computed from unrounded scores.

Tracker	Full Set			Challenging Scenarios		
	SR	NPR	PR	SR	NPR	PR
TransT	0.7355	0.8791	0.8846	0.6715	0.8139	0.8560
ToMP50	0.7180	0.8628	0.8732	0.6568	0.7911	0.8315
CTTrack	0.6836	0.8415	0.8504	0.6260	0.7832	0.8336
KeepTrack	0.6536	0.7727	0.8070	0.5952	0.7121	0.7728
DiMP50	0.6531	0.7790	0.8174	0.5875	0.6996	0.7821
CVAC-Net (Ours)	0.7574	0.8871	0.8984	0.6959	0.8331	0.8673

4.4. Ablation Study

4.4.1. Comparison with Basic Collaboration Strategies

To examine whether generic feature fusion is sufficient for aerial–ground collaboration, Table 4 compares three basic strategies using retained results on the same UGVT test split. Direct addition sums the synchronized feature maps element by element. Concatenation stacks the two feature maps along the channel dimension and restores the original channel number through a 1×1 projection. Cross-attention uses the main-view features as queries and the auxiliary-view features as keys and values. The reported scores are averaged over the ground and aerial views.

Among the basic strategies, concatenation obtains the best results, with SR, NPR, and PR scores of 69.4%, 79.2%, and 83.0%, respectively. Cross-attention obtains similar SR and NPR but a lower PR, while direct addition substantially reduces all three metrics. CVAC-Net exceeds concatenation by 3.0, 4.1, and 3.2 percentage points in SR, NPR, and PR, respectively. In these comparisons, directly mixing the two feature maps gives lower scores than CVAC-Net.

Table 4. Comparison with basic cross-view collaboration strategies on UGVT. SR, NPR, and PR are averaged over the ground and aerial views. The CVAC-Net row reports the mean of three full-model results.

Tracker	SR (%)	NPR (%)	PR (%)
TransT	69.2	80.1	84.5
Direct addition	52.7	61.8	64.8
Concatenation	69.4	79.2	83.0
Cross-attention	69.0	79.1	82.1
CVAC-Net (Ours)	72.4	83.3	86.2

4.4.2. Analysis of the Proposed Components

Table 5 combines the original progressive configurations with independent removal controls. Adding VAF and then TAF to TransT increases all three metrics in the progressive comparison. TSA-1 denotes projection and contrastive semantic alignment, and TSA-2 denotes learned spatial translation. Compared with the complete-model mean, removing TSA-1 reduces SR, NPR, and PR by 0.76, 1.82, and 0.85 percentage points, respectively. Removing TSA-2 reduces the same metrics by 0.74, 0.76, and 0.88 points, while removing TAF reduces them by 0.66, 0.87, and 0.74 points.

Replacing VAF with residual 1×1 fusion yields the largest decreases, 3.34, 4.16, and 3.86 percentage points in SR, NPR, and PR, respectively. The replacement removes both the multi-encoder transformations and spatial gating. Its lower performance therefore reflects their combined contribution. The weak direct-addition result in Table 4 also indicates that direct feature transfer is less effective than adaptive fusion.

Table 5. Component ablation on UGVT. A check mark indicates that the component is present. The VAF-only and VAF + TAF rows are retained from the original progressive comparison. The four removal rows are additional single-run controls. The complete-model row reports the mean of three full-model results. The row with TAF and TSA but without VAF uses residual 1×1 fusion.

VAF	TAF	TSA-1	TSA-2	UGVT		
				SR (%)	NPR (%)	PR (%)
–	–	–	–	69.2	80.1	84.5
✓	–	–	–	70.8	81.1	84.7
✓	✓	–	–	71.8	82.3	85.9
✓	✓	–	✓	71.63	81.44	85.32
✓	✓	✓	–	71.65	82.50	85.29
✓	–	✓	✓	71.73	82.39	85.43
–	✓	✓	✓	69.05	79.10	82.31
✓	✓	✓	✓	72.39	83.26	86.17

4.4.3. Analysis of the View-Specific Design

Table 6 compares shared and view-specific fusion branches. The view-specific design improves the reported ground-view SR, NPR, and PR from 0.694, 0.784, and 0.846 to 0.711, 0.811, and 0.863, respectively. The corresponding aerial-view scores increase from 0.740, 0.858, and 0.869 to 0.747, 0.867, and 0.871. These original comparisons suggest that the two views benefit differently from cross-view collaboration.

Table 6. Comparison of shared and view-specific fusion branches on UGVT. The CVAC-specific row is an individual full-model result; the main comparison table reports the three-result mean.

Tracker	Ground View			Aerial View		
	SR	NPR	PR	SR	NPR	PR
TransT	0.669	0.769	0.838	0.715	0.833	0.853
CVAC-shared	0.694	0.784	0.846	0.740	0.858	0.869
CVAC-specific	0.711	0.811	0.863	0.747	0.867	0.871

4.4.4. Sensitivity to the Number of Encoders and Selected Experts

We vary one routing parameter at a time and compare the alternative settings with the main-table mean for $N = 9, K = 4$ (Figure 5). This setting obtains 72.39% SR, 83.26% NPR, and 86.17% PR. With $K = 4$, $N = 4$ obtains 69.44%, 79.70%, and 82.59%, while $N = 12$ obtains 69.47%, 79.73%, and 82.41%, respectively. With $N = 9$, $K = 1$ obtains 70.99%, 81.67%, and 84.63%, while $K = 9$ obtains 71.01%, 81.68%, and 84.39%. The selected setting

has the highest observed scores on all three metrics among these configurations. The lower scores at $K = 1$ and $K = 9$ are consistent with a trade-off between limited feature coverage and the inclusion of less useful auxiliary information.

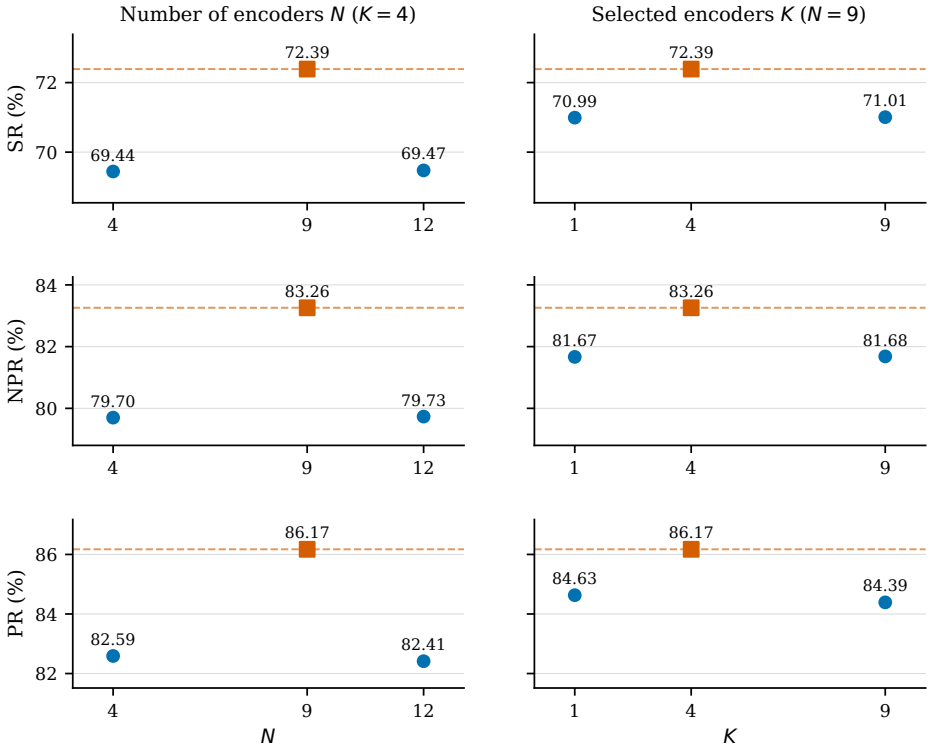


Figure 5. Sensitivity of SR, NPR, and PR to the number of encoders N at $K = 4$ (**left column**) and the number of selected encoders K at $N = 9$ (**right column**) on UGVT. Orange squares and dashed lines mark the main-table mean for $N = 9, K = 4$, while blue circles show the alternative settings from individual sensitivity runs.

4.4.5. Alignment-Loss Weight

To evaluate the role of cross-view alignment supervision, Table 7 compares its removal and two fixed weights with the decaying schedule. The complete model obtains 72.39% SR, 83.26% NPR, and 86.17% PR. Removing the alignment loss gives 71.23%, 82.00%, and 84.85%, respectively. A fixed weight of 0.5 gives 70.00%, 81.02%, and 83.97%, while a fixed weight of 1.0 gives 70.06%, 80.58%, and 83.66%. All three controls have lower scores on each metric. Tracking is the primary objective, so maintaining a fixed alignment weight may compete with later classification and localization optimization. The decaying schedule places more emphasis on correspondence early in training and more on tracking later, which is consistent with these results.

Table 7. Alignment-loss ablation on UGVT. The complete-model row reports the mean of three full-model results. The exponent e denotes the training epoch.

Alignment-Loss Weight	SR	NPR	PR
0	0.7123	0.8200	0.8485
0.5 (fixed)	0.7000	0.8102	0.8397
1.0 (fixed)	0.7006	0.8058	0.8366
0.5×0.8^e (ours)	0.7239	0.8326	0.8617

4.4.6. Feature and Encoder-Routing Analysis

To examine the shared projection, we compare features from 45 paired UGVT identities before and after projection. For each feature, cross-view recall at rank 1 (R@1) is one if its

true pair has the highest cosine similarity among the 45 features from the other view, and zero otherwise. The matched-minus-mismatched cosine margin is the mean similarity of true pairs minus the mean similarity of all nonmatching pairs. The projection increases R@1 from 26.67% to 33.33% for aerial-to-ground retrieval and from 24.44% to 37.78% in the reverse direction (Figure 6a). The cosine margin rises from 0.0745 to 0.1959. These results show better discrimination of paired features in this retrieval test.

To examine encoder utilization, we count how often each encoder belongs to the selected Top-4 over 20,409 routed frames per view-specific bank (Figure 6b). All nine encoders are selected. The adjacent-frame change rate is the fraction of successive frames within a sequence whose selected Top-4 sets differ. It is 13.37% for ground updates and 15.19% for aerial updates. We also compare the encoders' spatial residual responses on 360 sampled frames per bank. Spatial centered kernel alignment (CKA) measures similarity between the channel-wise root-mean-square residual maps of two encoders across these frames. The mean pairwise value is 0.363 in both banks. In this analysis, the Top-4 sets vary across frames, and the encoders produce different spatial response patterns in both view-specific branches.

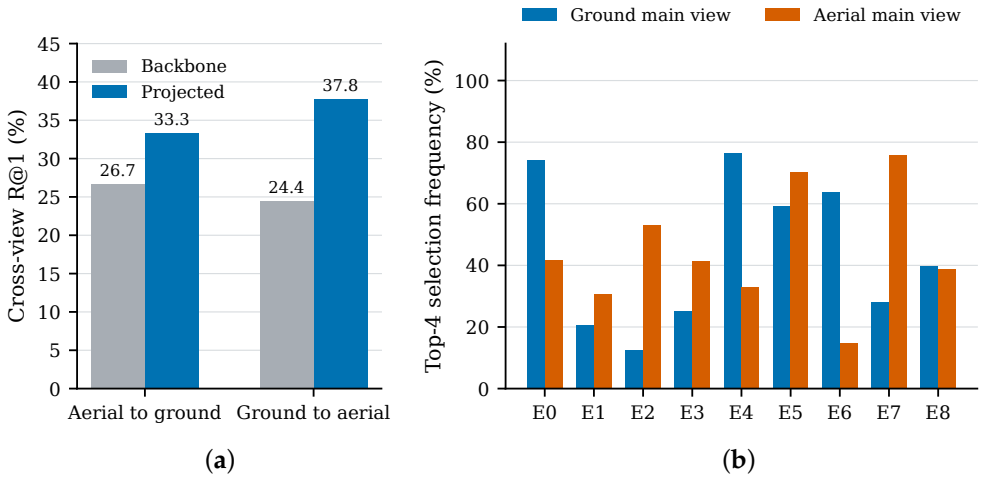


Figure 6. Feature and routing diagnostics on UGVT. (**a**) Cross-view identity retrieval before and after projection, measured by R@1. (**b**) Top-4 selection frequency for every encoder of CVAC-Net over 20,409 routed frames per bank. Encoder indices are local to each view-specific bank.

4.5. Visualization

4.5.1. Visualization of TAF Response

To visualize TAF, we bilinearly upsample R^m from Equation (13) to the search-image size and denote the result by R_{up}. We then normalize each map as

$$R_{\text{vis}} = \frac{R_{\text{up}} - \min(R_{\text{up}})}{\max(R_{\text{up}}) - \min(R_{\text{up}}) + 10^{-6}}. \tag{20}$$

We overlay values above 0.6 on the auxiliary-view image. This normalization is used only for display. The model uses R^m directly. Figure 7 shows the resulting response visualizations for representative paired observations.

4.5.2. Visualization of Tracking Results

Figure 8 compares the tracking results on electromobile-20, bus-1, car-43, and etricycle-4. These sequences contain low resolution, severe occlusion, fast motion, and scale variation, respectively. In the displayed frames, several comparison trackers drift from the target or fail after the corresponding challenge occurs. The bounding boxes pro-

duced by CVAC-Net remain closer to the ground truth in these examples. The qualitative observations agree with the attribute-based results in Figure 4.

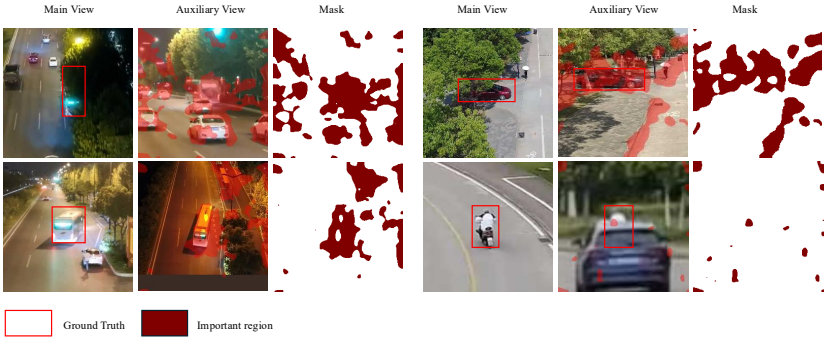


Figure 7. Visualization of TAF. From left to right: main-view image, auxiliary-view image, and regions with display-normalized response above 0.6. The displayed response illustrates the spatial weighting used before fusion.

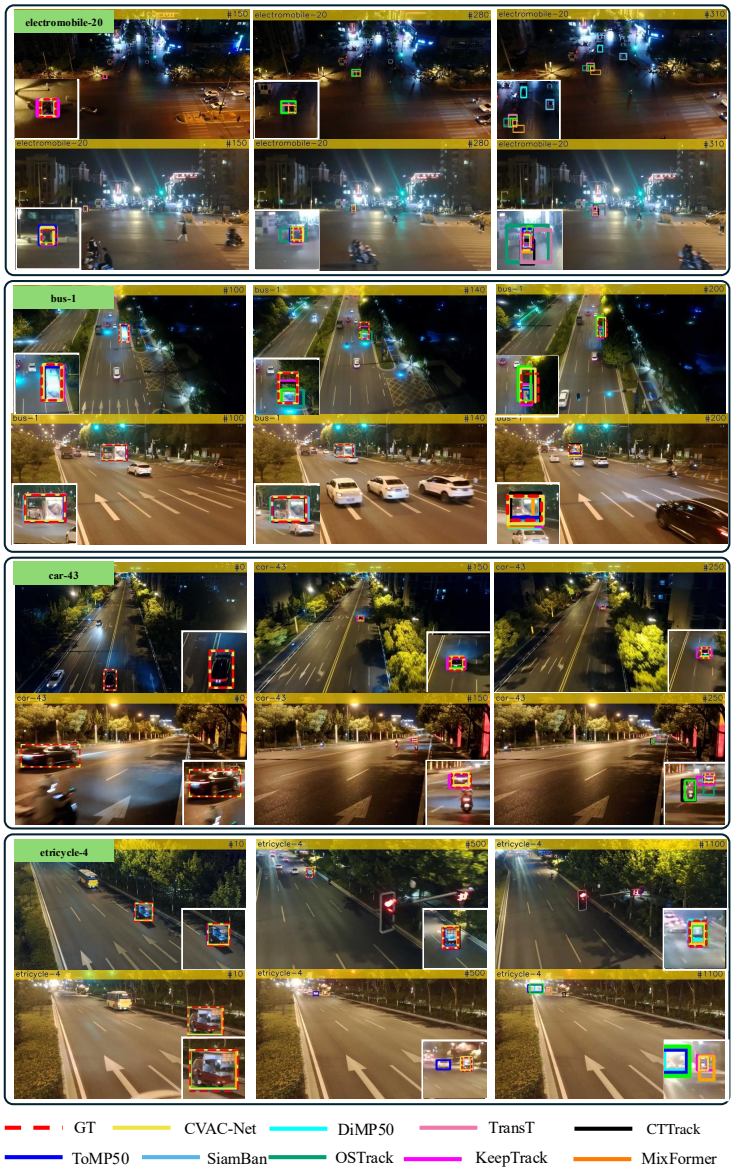


Figure 8. Visualization of tracking predictions on four UGVT sequences covering low resolution, occlusion, fast motion, and scale variation. CVAC-Net better preserves target continuity and reduces drift compared with single-view baselines. GT denotes ground truth.

5. Discussion

CVAC-Net is designed around a central difficulty of aerial–ground tracking: the auxiliary view may provide useful target information but also contain background and unrelated objects. To make this information usable, two-stage alignment addresses differences between the search features; target-aware filtering weights auxiliary responses according to the main-view template; and view-adaptive fusion adjusts their contribution to each tracking branch. Together, these components allow the two views to support each other without explicit camera calibration. Their separate roles may also help researchers identify whether cross-view correspondence, auxiliary feature weighting, or fusion needs attention in other paired-camera settings.

Despite these advantages, CVAC-Net still has several limitations. First, it assumes synchronized aerial and ground frames and target initialization in both views, which restricts its use when the observations are asynchronous or an initial target state is unavailable in one view. Second, the learned two-dimensional offset only adjusts relative displacement within normalized search crops and does not model scale changes, perspective effects, or local deformation. Finally, the current two-branch design supports one aerial and one ground view and cannot directly accommodate additional viewpoints. Its view-specific fusion branches also increase computation, which may limit deployment when processing resources are constrained.

6. Conclusions

This paper proposes CVAC-Net for aerial–ground visual tracking, which incorporates two-stage alignment, target-aware filtering, and view-specific collaboration to exchange target information between heterogeneous views. On the UGVT dataset, CVAC-Net outperforms TransT by 3.2, 3.1, and 1.7 percentage points in SR, NPR, and PR, respectively. The 2D evaluation derived from Griffin provides supplementary validation on simulated aerial–ground videos.

This study evaluates CVAC-Net using synchronized visible-light video from aerial and ground cameras. When the target is invisible for a long time from both perspectives, or when communication between the two platforms is interrupted, its effectiveness may be limited. Future work will focus on asynchronous observations and more efficient cross-view interactions, and will also incorporate more observation perspectives and modalities.

Author Contributions: M.X.: Conceptualization, Methodology, Software, Validation, Formal analysis, Visualization, Writing—original draft. C.L.: Methodology, Validation, Formal analysis, Writing—review and editing. H.Z.: Data curation, Validation, Visualization. S.L.: Investigation, Resources, Writing—review and editing. Y.W.: Data curation, Validation. S.Z.: Supervision, Funding acquisition, Writing—review and editing. W.W.: Supervision, Project administration, Funding acquisition, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Youth Innovation Promotion Association CAS under Grant No. 2023137 and in part by the Science and Disruptive Technology Program under Grant No. 2025-AIRCAS-SDTP-01.

Data Availability Statement: The UGVT and Griffin benchmarks are available from their original sources [17,20] and remain subject to the corresponding access conditions. The CVAC-Net repository (https://github.com/Air-Eureka/CVAC-Net (accessed on 25 September 2026)) provides the exact UGVT training/test sequence lists and a split check. It also provides the Griffin-derived sequence lists, conversion and validation scripts, protocol specification, and source mapping. Additional time is needed to organize the CVAC-Net implementation and trained weights for public release. Third-party datasets and pretrained weights are not redistributed.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abdelaziz, O.; Shehata, M.; Mohamed, M. Beyond traditional visual object tracking: A survey. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 1435–1460. [CrossRef]
2. Tmušić, G.; Manfreda, S.; Aasen, H.; James, M.R.; Gonçalves, G.; Ben-Dor, E.; Brook, A.; Polinova, M.; Arranz, J.J.; Mészáros, J.; et al. Current practices in UAS-based environmental monitoring. *Remote Sens.* **2020**, *12*, 1001. [CrossRef]
3. Li, J.; Pei, Y.; Zhao, S.; Xiao, R.; Sang, X.; Zhang, C. A review of remote sensing for environmental monitoring in China. *Remote Sens.* **2020**, *12*, 1130. [CrossRef]
4. Manfreda, S.; McCabe, M.F.; Miller, P.E.; Lucas, R.; Pajuelo Madrigal, V.; Mallinis, G.; Ben Dor, E.; Helman, D.; Estes, L.; Ciraolo, G.; et al. On the use of unmanned aerial systems for environmental monitoring. *Remote Sens.* **2018**, *10*, 641. [CrossRef]
5. Zhang, N.; Nex, F.; Vosselman, G.; Kerle, N. Training a disaster victim detection network for UAV search and rescue using harmonious composite images. *Remote Sens.* **2022**, *14*, 2977. [CrossRef]
6. Alhaqbani, A.; Kurdi, H.; Youcef-Toumi, K. Fish-inspired task allocation algorithm for multiple unmanned aerial vehicles in search and rescue missions. *Remote Sens.* **2021**, *13*, 27. [CrossRef]
7. Caroppo, A.; Diraco, G.; Leone, A. A systematic review of available multispectral UAV image datasets for precision agriculture applications. *Remote Sens.* **2026**, *18*, 659. [CrossRef]
8. Vinci, A.; Brigante, R.; Traini, C.; Farinelli, D. Geometrical characterization of hazelnut trees in an intensive orchard by an unmanned aerial vehicle (UAV) for precision agriculture applications. *Remote Sens.* **2023**, *15*, 541. [CrossRef]
9. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ATOM: Accurate Tracking by Overlap Maximization. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 4660–4669. [CrossRef]
10. Bhat, G.; Danelljan, M.; Van Gool, L.; Timofte, R. Learning Discriminative Model Prediction for Tracking. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 6181–6190. [CrossRef]
11. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In *Proceedings of the Computer Vision—ECCV 2016 Workshops*; Springer: Cham, Switzerland, 2016; Volume 9914, pp. 850–865. [CrossRef]
12. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 4282–4291. [CrossRef]
13. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer Tracking. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 8126–8135. [CrossRef]
14. Song, Z.; Luo, R.; Yu, J.; Chen, Y.P.P.; Yang, W. Compact Transformer Tracker with Correlative Masked Modeling. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2023; Volume 37, pp. 2321–2329. [CrossRef]
15. Mayer, C.; Danelljan, M.; Paudel, D.P.; Van Gool, L. Learning Target Candidate Association to Keep Track of What Not to Track. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 13424–13434. [CrossRef]
16. Chen, G.; Zhu, P.; Cao, B.; Wang, X.; Hu, Q. Cross-Drone Transformer Network for Robust Single Object Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4552–4563. [CrossRef]
17. Sun, D.; Cheng, L.; Chen, S.; Li, C.; Xiao, Y.; Luo, B. UAV-Ground Visual Tracking: A Unified Dataset and Collaborative Learning Approach. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 3619–3632. [CrossRef]
18. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event, 18–24 July 2021*; Meila, M., Zhang, T., Eds.; Proceedings of Machine Learning Research; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 8748–8763.
19. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.V.; Hinton, G.E.; Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017*; Conference Track Proceedings; OpenReview.net: Newton Highlands, MA, USA, 2017.
20. Wang, J.; Cao, X.; Zhong, J.; Zhang, Y.; Han, Z.; Yu, H.; Zhang, C.; He, L.; Xu, S.; Wang, J. Griffin: Aerial-Ground Cooperative Detection and Tracking Dataset and Benchmark. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2026; Volume 40, pp. 9867–9875. [CrossRef]
21. Guo, Y.; Yang, D.; Chen, Z. Object Tracking on Satellite Videos: A Correlation Filter-Based Tracking Method with Trajectory Correction by Kalman Filter. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3538–3551. [CrossRef]

22. Xuan, S.; Li, S.; Han, M.; Wan, X.; Xia, G.S. Object Tracking in Satellite Videos by Improved Correlation Filters with Motion Estimations. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 1074–1086. [CrossRef]
23. Chen, S.; Wang, T.; Wang, H.; Wang, Y.; Hong, J.; Dong, T.; Li, Z. Vehicle Tracking on Satellite Video Based on Historical Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 7784–7796. [CrossRef]
24. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5612518 . [CrossRef]
25. Xue, Y.; Jin, G.; Shen, T.; Tan, L.; Wang, N.; Gao, J.; Wang, L. SmallTrack: Wavelet Pooling and Graph Enhanced Classification for UAV Small Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5618815. [CrossRef]
26. Li, S.; Yang, X.; Wang, X.; Zeng, D.; Ye, H.; Zhao, Q. Learning Target-Aware Vision Transformers for Real-Time UAV Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4705718. [CrossRef]
27. Chen, J.; Huang, B.; Li, J.; Wang, Y.; Ren, M.; Xu, T. Learning Spatio-Temporal Attention Based Siamese Network for Tracking UAVs in the Wild. *Remote Sens.* **2022**, *14*, 1797. [CrossRef]
28. Wu, D.; Zhu, H.; Lan, Y. A Method for Designated Target Anti-Interference Tracking Combining YOLOv5 and SiamRPN for UAV Tracking and Landing Control. *Remote Sens.* **2022**, *14*, 2825. [CrossRef]
29. Wang, Z.; Hu, Y.; Yang, J.; Zhou, G.; Liu, F.; Liu, Y. A Contrastive-Augmented Memory Network for Anti-UAV Tracking in TIR Videos. *Remote Sens.* **2024**, *16*, 4775. [CrossRef]
30. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese Box Adaptive Network for Visual Tracking. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 6667–6676. [CrossRef]
31. Danelljan, M.; Van Gool, L.; Timofte, R. Probabilistic Regression for Visual Tracking. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 7181–7190. [CrossRef]
32. Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D.P.; Yu, F.; Van Gool, L. Transforming Model Prediction for Tracking. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 8721–8730. [CrossRef]
33. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning Spatio-Temporal Transformer for Visual Tracking. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 10428–10437. [CrossRef]
34. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. MixFormer: End-to-End Tracking with Iterative Mixed Attention. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 13598–13608. [CrossRef]
35. Ye, B.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; Volume 13682, pp. 341–357. [CrossRef]
36. Zhu, P.; Zheng, J.; Du, D.; Wen, L.; Sun, Y.; Hu, Q. Multi-Drone-Based Single Object Tracking with Agent Sharing Network. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4058–4070. [CrossRef]
37. Xue, Y.; Jin, G.; Shen, T.; Tan, L.; Wang, N.; Gao, J.; Wang, L. Consistent Representation Mining for Multi-Drone Single Object Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10845–10859. [CrossRef]
38. Hou, Y.; Zheng, L.; Gould, S. Multiview Detection with Feature Perspective Transformation. In *Proceedings of the Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; Volume 12352, pp. 1–18. [CrossRef]
39. Hou, Y.; Zheng, L. Multiview Detection with Shadow Transformer (and View-Coherent Data Augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2021; pp. 1673–1682. [CrossRef]
40. Teepe, T.; Wolters, P.; Gilg, J.; Herzog, F.; Rigoll, G. EarlyBird: Early-Fusion for Multi-View Tracking in the Bird's Eye View. In *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*; IEEE: New York, NY, USA, 2024; pp. 102–111. [CrossRef]
41. Xu, M.; Zhu, Y.; Jiang, H.; Li, J.; Shen, Z.; Wang, S.; Huang, H.; Wang, X.; Zhang, H.; Yang, Q.; et al. MITracker: Multi-View Integration for Visual Object Tracking. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 27176–27185. [CrossRef]
42. Xu, B.; Hou, R.; Ren, T.; Wu, G. VL-UniTrack: A Unified Framework with Visual-Language Prompts for UAV-Ground Visual Tracking. *arXiv* **2026**, arXiv:2605.04574. [CrossRef]
43. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
44. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28, pp. 2017–2025.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 770–778. [CrossRef]
46. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*; ICLR: Amherst, MA, USA, 2019.
47. Wu, Y.; Lim, J.; Yang, M.H. Online Object Tracking: A Benchmark. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2013; pp. 2411–2418. [CrossRef]

48. Müller, M.; Bibi, A.; Giancola, S.; Al-Subaihi, S.; Ghanem, B. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *Proceedings of the Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; Volume 11205, pp. 310–327. [CrossRef]
49. Danelljan, M.; Bhat, G.; Mayer, C.; Paul, M. *PyTracking: Visual Tracking Library Based on PyTorch*; GitHub Repository: San Francisco, CA, USA, 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.