applied sciences

MDPI

Article

SwinMR: A Mutual Refinement Enhanced SwinTrack Framework

Shifeng Zhao, Chuanyuan Yang * and Yanfang Fu *

School of Computer Science and Engineering, Xi'an Technological University, Weiyang Campus, Xi'an 710021, China; zhaoshifeng@xatu.edu.cn

* Correspondence: yangchuanyuan2@163.com (C.Y.); fuyanfang@xatu.edu.cn (Y.F.)

Abstract

The task of tracking weak targets in low-altitude UAV scenarios requires high robustness and generalization ability of the model. Against this backdrop, this paper proposes a novel annotation and training mechanism based on SwinTrack. To improve the model's tracking ability for weak targets, this paper proposes a pseudo-label consistency screening and background enhancement annotation strategy. This strategy enlarges the target box proportionally before training to obtain more effective background information. Furthermore, pseudo-labels are screened using a hybrid gating system of geometric overlap and confidence consistency to reduce the negative transfer interference of noise generated in different domains on the model. Since the data feature distribution varies significantly in tracking tasks, this paper introduces a mutual-teaching pseudo-label iterative training method into the field of weak target tracking. This aims to continuously transfer the model from the source domain to the target domain during iteration, thereby improving the model's generalization ability. Experiments have shown that, when faced with a completely new dataset of weak target tracking, the proposed method improves upon recent strong baselines in single-target tracking by 0.05 in both P@20 and NP-AUC, and by 0.04 in SUS, demonstrating the enhanced tracking performance and generalization ability of the proposed method in the field of weak target tracking.

Keywords: weak target tracking; mutual refinement; background enhancement strategy

check for updates

Received: 4 November 2025
Revised: 2 December 2025
Accepted: 10 December 2025
Published: 11 December 2025

Citation: Zhao, S.; Yang, C.; Fu, Y. SwinMR: A Mutual Refinement Enhanced SwinTrack Framework. *Appl. Sci.* **2025**, *15*, 13070. https://doi.org/10.3390/app152413070

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The application of low-altitude drone photography in urban inspection, security patrol and emergency search and rescue is increasing, but the task of tracking weak targets still has major challenges to be solved: on the one hand, due to scale changes and motion jitter caused by aerial angle and imaging height, the target appearance and background interference frequently alternate, and the occlusion and illumination changes in complex environment lead to short-term prediction instability (as shown in Figure 1, low-altitude drone overhead shot of a moving pedestrian on the ground and Figure 2, drone long-distance imaging) [1]. On the other hand, there are often significant cross-domain differences between the source domain and the target domain, and directly transferring the model to a completely new target domain will lead to a significant performance drop [2].



Figure 1. Small pedestrian targets captured by a drone.



Figure 2. Schematic diagram of long-distance imaging of weak targets [3].

Since this paper adopts the unified COCO data format, and the COCO format clearly indicates that the size of a small target is 32 × 32 pixels, this paper uses this as the size judgment boundary for small targets. The most important issue in the problem of the ability to discriminate small targets is the ability to represent small targets. However, the existing methods have not achieved good results in addressing the limitation of effective pixels for small targets, and therefore cannot have good robust performance in the task of tracking small targets in complex and variable scenarios [4]. Moreover, the current methods have not solved the problem of model generalization ability in the task of tracking small targets. If the same target is tracked for a long time and multiple different background data are involved, the model will have difficulty achieving good performance in all data without manual annotation [5].

Some scholars have designed multi-scale feature extraction and combined it with a pyramid structure to enhance the representation ability of small targets in order to deal with the small size of the target [6]. However, these methods not only failed to solve the problem but also increased the computational cost. At the same time, some scholars have tried to introduce cross-modal fusion in order to highlight the information of the target in the context of background interference [7]. Besides, such methods require accurate label information and cannot cope with the scenario without human annotation. At the same time, although Transformer-based tracking frameworks (such as SwinTrack) have achieved good results in labeled domain conditions in recent years, there is still a lack of methods that can enable the model to have good generalization ability when facing different scenarios and data without human annotation, when facing small targets and cross-domain settings [8].

This paper proposes a novel annotation and training architecture. Its core idea is to improve the detection and tracking capabilities of weak targets in target tracking tasks by using pseudo-labels and background enhancement annotations for cross-domain training, forming a closed-loop training process [9]. In the initial data processing stage, this paper proposes a background enhancement annotation (BEA) method. This method does not change the backbone network and decoder structure, but only enhances the annotation information in the classification branch. By expanding the shared pseudo-label bounding box and setting positive sample regions, background noise is suppressed, and the discriminability between the target and the background is improved. Regarding the training framework, this paper first uses two initial models trained independently in the source

domain to predict the same frame in the target domain. Based on this, a joint gating mechanism is introduced, which only retains the frame for further training when the predictions of both models reach specific thresholds in terms of geometric position and confidence. This effectively suppresses the negative transfer effect caused by noisy supervision, thereby improving training stability. After consistency screening, the method employs a mutually optimized strategy, obtaining shared pseudo-labels through confidence-weighted soft fusion and uniqueness operations [10]. These pseudo-labels are learned symmetrically between the two models, allowing the models to gradually adapt to the feature distribution of the target domain, thereby continuously improving the model's generalization ability.

2. Related Works

2.1. Visual Target Tracking

Visual tracking methods based on correlation filtering are built upon a framework of shallow handcrafted features and frequency domain optimization. They achieve extremely high efficiency by transforming the entire convolution operation into frequency domain multiplication. MOSSE [11], within this paradigm, establishes a fast filter update mechanism using a frequency domain solution that minimizes output error, enabling online tracking to be completed at millisecond speeds. However, due to insufficient feature representation, the model is prone to crashing under drastic lighting changes, target deformation, and background interference. KCF [12] introduces kernel tricks into correlation filtering, utilizing a cyclic matrix structure to improve sample utilization and enhance feature discrimination, resulting in higher robustness in conventional scenes. However, because it still relies on handcrafted features, it still faces drift risks in complex dynamic backgrounds. STRUCK [13] directly optimizes the target boundary position through a structured SVM [14], enabling the model to handle partial occlusion, local deformation, and gradual changes in target appearance. However, its linear structure still lacks robustness to large-scale changes and extreme occlusion. Methods at this stage generally exhibit the characteristics of "extremely fast but limited expression," performing exceptionally well in simple scenes but struggling to adapt to the highly dynamic changes in real-world environments.

Deep Siamese networks overcome the limitations of traditional correlation filtering, which relies on weakly expressive features, by learning end-to-end similarity metrics and significantly improve cross-scene generalization capabilities within a unified template matching framework. SiamFC [15] uses a convolutional Siamese structure to construct the similarity calculation between the template and the search region, ensuring high speed while achieving stronger feature stability than shallow methods; however, its localization accuracy is limited by the simple fully convolutional structure. SiamRPN [16] and SiamRPN++ [17] improve feature expressiveness and bounding box prediction capabilities by fusing a region proposal network with a deeper backbone, and address the gradient instability problem that easily occurs during training of deep Siamese structures. SiamMask [18] adds an explicit segmentation branch, enabling the model to have stronger geometric consistency in occluded and non-rigid deformation scenarios. In another approach, ATOM [19] and DiMP [20] introduce an IoU prediction module and discriminative online updates, using explicit modeling of prediction quality to compensate for the shortcomings of pure template matching in fine-grained localization, allowing the model to maintain more stable tracking performance in long sequences and violently dynamic scenes. Overall, the technological advancements at this stage stemmed from the combination of deep features, large-scale data training, and online adaptation mechanisms, enabling tracking performance to exceed the capabilities of correlation filtering frameworks.

The introduction of Transformers elevated visual tracking from local convolutional matching to a unified global spatiotemporal modeling paradigm. TransT [21] utilizes a mutual attention mechanism to directly fuse template features and search region features, overcoming the limitations of traditional twin structures in information interaction and significantly improving the ability to suppress interfering targets. STARK [22] extends Transformers to the temporal direction, enabling the model to model the dynamic evolution of targets in consecutive frames, thus achieving end-to-end position prediction without complex post-processing. MixFormer [23] adopts a single-stream Transformer structure to uniformly encode templates and search regions, eliminating the information fragmentation caused by these two in multi-branch networks, ensuring consistent representation of appearance changes, non-rigid deformations, and background perturbations. OS-Track [24] introduces continuous-temporal priors, giving Transformers more stable temporal coherence in long video sequences. SwinTrack [25] combines hierarchical window attention with a quality-aware prediction head, achieving a good balance between high-resolution modeling and anti-drift performance. Transformer-based methods are driving visual tracking toward unified feature representation, global spatiotemporal dependency modeling, and long-term robustness, becoming an important trend in current visual tracking research.

2.2. Tracking Methods for Small Targets

As shown in Figure 3, in typical low-altitude aerial photographs of backgrounds such as vast grasslands and high-contrast building complexes, the scene textures are complex, and the structures change frequently. The target pixel (green box in the figure) appears as only a small bright or dark area within the background, its area being extremely small and isolated compared to the surrounding large, highly textured background. In this situation, the appearance prior relied upon by traditional visual tracking models is difficult to establish sufficiently, limiting the model's representational ability and making it difficult to maintain stable and accurate tracking performance. This significant imbalance causes the tracking model to easily overlook the target itself during feature extraction and response focusing, even mistaking it for background texture, leading to misjudgments in the detection and tracking stages.

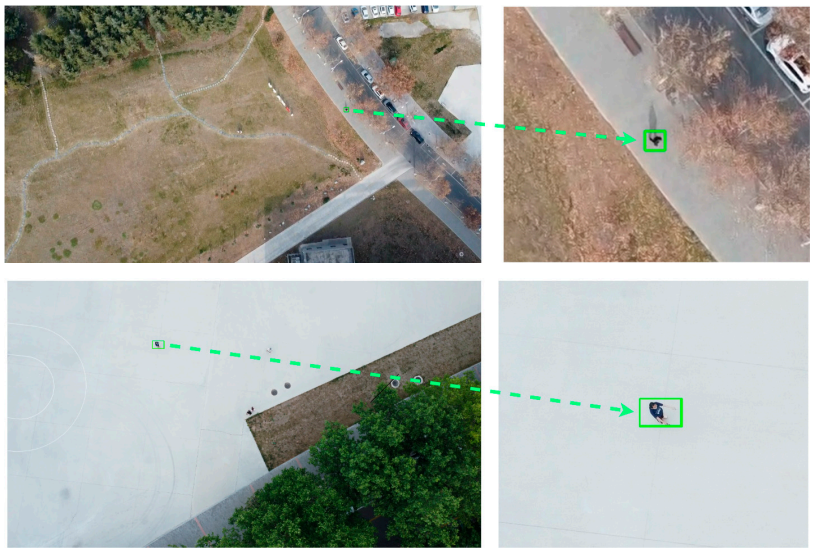

Figure 3. Example of tracking small targets.

Current research on small target tracking mainly focuses on enhancing discrimination ability, improving localization and scale robustness, and strengthening interference suppression and feature representation. To improve discrimination ability and suppress

interference, DaSiamRPN [26] introduces a distractor-aware mechanism in its Siamese network training, enabling the feature space to distinguish background interference similar to the target. Combined with a local-to-global search strategy, it effectively reduces the risk of target drift and loss. SRDCF [27] and ECO [28], through spatial regularization, background suppression, and multi-resolution feature fusion techniques, provide stable and reliable discrimination capabilities in weakly textured or complex background environments, offering a robust baseline for small target tracking.

Regarding anchor-box constraint removal and localization stability, SiamFC++ [29] adopts an anchor-free structure, effectively solving the instability problem of small targets in traditional anchor-box matching. It also introduces a quality assessment module to correct inconsistencies between classification and localization confidence. SiamBAN [30] designs an adaptive anchor-free regression head that adjusts boundary estimation based on local geometric characteristics, thereby mitigating the phenomenon of boundary oscillations for small targets. SiamCAR [31] incorporates a dual branch of centrality and quality into its anchorless structure, improving the consistency and stability of weak target prediction.

For feature enhancement and fine-grained boundary optimization of weakly textured and small targets, Ocean [32] unifies classification and regression through an object-aware anchor-free regression head, thus addressing the "high score but poor localization" problem for weakly textured targets. HiFT [33] utilizes a hierarchical Transformer to capture multi-scale long-range dependency information, significantly improving the semantic representation of weak regions in scenes with dense small targets or across regions. AlphaRefine [34] corrects coarse localization biases for small targets through a high-precision boundary refinement branch, significantly improving boundary accuracy and overall tracking performance in weak target tracking and occlusion recovery.

2.3. Unsupervised Domain Adaptive Cross-Domain Transfer Learning

Unsupervised domain adaptation (UDA) [35] and cross-domain transfer learning aim to train an initial model using only labeled data from the source domain and enable it to generalize effectively in the unlabeled target domain. Core methods for achieving this goal include pseudo-label generation, iterative training, feature alignment, and consistency constraints. Model performance is highly dependent on the accuracy of the pseudo-labels, while noise is suppressed through bidirectional mutual teaching, exponential moving average (EMA) smoothing, or consistency constraints, thereby enhancing the model's robustness in the target domain.

In the field of person re-identification, advancements in methods reflect continuous optimization of cross-domain feature discriminative ability and pseudo-label quality. The SSG [36] method generates pseudo-labels through clustering of global and local features, achieving consistent clustering and iterative optimization. ECN [37] and its improved version ECN++ [38] mitigate the sensitivity to clustering noise through neighbor consistency propagation, thus improving pseudo-label robustness. AD-Cluster [39] combines adaptive clustering and discriminative learning to enhance the discriminative power of cross-domain features. HCT [40] enhances pseudo-label quality through hard sample mining and robust clustering. MMT [41] employs a teacher-student bi-branch framework to generate pseudo-labels and suppresses noise through EMA teacher mutual distillation. SpCL [42] methods enhance intra-class similarity and inter-class discriminability, utilizing large-scale memory for contrastive learning to strengthen supervisory signals. Recent research trends focus on multi-branching, multi-granularity mutual refinement, and the introduction of soft-supervised losses to improve the separability and discriminative power of target domain features.

In dense prediction tasks such as object detection and semantic segmentation, cross-domain methods mainly include feature alignment and pseudo-label self-training. Feature alignment methods utilize adversarial learning domain discriminators, maximum mean difference (MMD), and kernel mean embedding (KME) to align the feature distributions of the source and target domains while maintaining discriminative power. Pseudo-label self-training methods, such as STAC [43], Unbiased Teacher [44], and Soft Teacher [45], generate high-confidence pseudo-labels through a teacher model, while the student model learns under a data augmentation view. Key techniques include selecting high-quality pseudo-labels to improve self-training stability and combining confidence thresholds and consistency constraints to enhance pseudo-label quality.

In object tracking tasks, UDA methods are relatively new, but their technical path is similar to that of person re-identification. Self-supervised temporal consistency methods utilize forward-backward consistency, trajectory periodicity constraints, and occlusion recovery consistency to generate supervision signals online [46]. Pseudo-label self-training methods generate target domain predictions as pseudo-labels using a source domain tracker, and then filter them using confidence thresholds, temporal smoothing, and cross-domain consistency, while simultaneously optimizing the model through mutual teaching iterations [47]. Other methods, such as style transfer and domain randomization, are used to reduce input distribution differences.

Based on the analysis of the above research status, it can be seen that unsupervised cross-domain research is gradually reaching a consensus: (1) It is necessary to ensure the quality of pseudo-labels through constraints such as consistency constraints; (2) The mutual teaching strategy of dual models or multi-branch can enable high-quality subsets to effectively supervise the adaptability of the model; (3) Teacher-student and soft supervision loss can reduce the noise impact in the case of no label. Therefore, based on the above consensus, this paper introduces a closed-loop framework on the SwinTrack framework, which is "the dual-source domain model infers two sets of pseudo-labels in the target domain, and then the pseudo-labels after consistency screening are exchanged to refine the initial model" [48]. The self-training and pseudo-label consistency ideas that have been experimentally verified in Re-ID [49] are transferred to the single target tracking scenario of weak targets, aiming to enable the model to achieve stable cross-domain tracking capability in the target domain without manual labeling.

3. Method

3.1. Overall Methodology Overview

This section will elaborate on the overall network architecture process. All networks use SwinTrack as the baseline tracker, denoted as network F(·). The differences between the models are only reflected in the parameters, denoted as $F(\theta_1)$, $F(\theta_2)$ and $F(\theta_3)$, respectively. $F(\theta_1)$ and $F(\theta_2)$ are obtained from two different common source domains, Domain D_1 and Domain D_2, respectively, learning two significantly different feature distributions. $F(\theta_3)$ is used for label-driven teacher learning on the self-made target domain, Target Domain T_1, and is also the convergence center of the entire iterative training. Structurally, the three networks are all composed of a template and a search dual-branch architecture, SwinTransformer, a quality-aware classification head, and a boundary regression head. The functional division of each role is completed through iterative parameter updates.

Figure 4 shows a schematic diagram of the overall network architecture. The middle part of the figure shows two student models, $F(\theta_1^t)$ and $F(\theta_2^t)$, independently performing forward inference on T_1 in round t, generating two sets of candidate pseudo-labels: $PL_1^t = \sum_{i=0}^{N-1}\left(b_{1,i}^t, s_{1,i}^t\right)$ and $PL_2^t = \sum_{j=0}^{M-1}\left(b_{1,j}^t, s_{1,j}^t\right)$, where $b_{1,i}^t$ and $b_{1,j}^t$ represent the predicted bounding box and $s_{1,i}^t$ and $s_{1,j}^t$ represent the corresponding quality score. Then,

the two pseudo-labels undergo consistency screening under a unified coordinate system, retaining only those that meet the consistency screening threshold and the geometric consistency threshold as refined pseudo-labels. These pseudo-labels are then uniquely merged to obtain $PL_3^t = \sum_{k=0}^{k_t} \left(b_k^{sup}, s_k^{sup} \right)$. In the classification response map, this is represented by a unique positive sample grid centered at b_k^{sup}, with the quality score at s_k^{sup}, and the rest as background grids, thus constructing a clear and robust supervision signal. Furthermore, the orange arrows indicate the parameter updates for both the teacher and student models. The teacher model $F\left(\theta_3^t\right)$ uses the unique shared pseudo-label PL_3^t as a supervision signal in T_1 to update the teacher model's parameters, obtaining θ_3^{t+1}. Subsequently, the student model's parameters are updated using the exponential moving average (EMA) method:

$$\theta_1^{t+1} = \alpha\theta_1^t + (1-\alpha)\theta_3^{t+1}, \theta_2^{t+1} = \alpha\theta_2^t + (1-\alpha)\theta_3^{t+1} \quad (1)$$

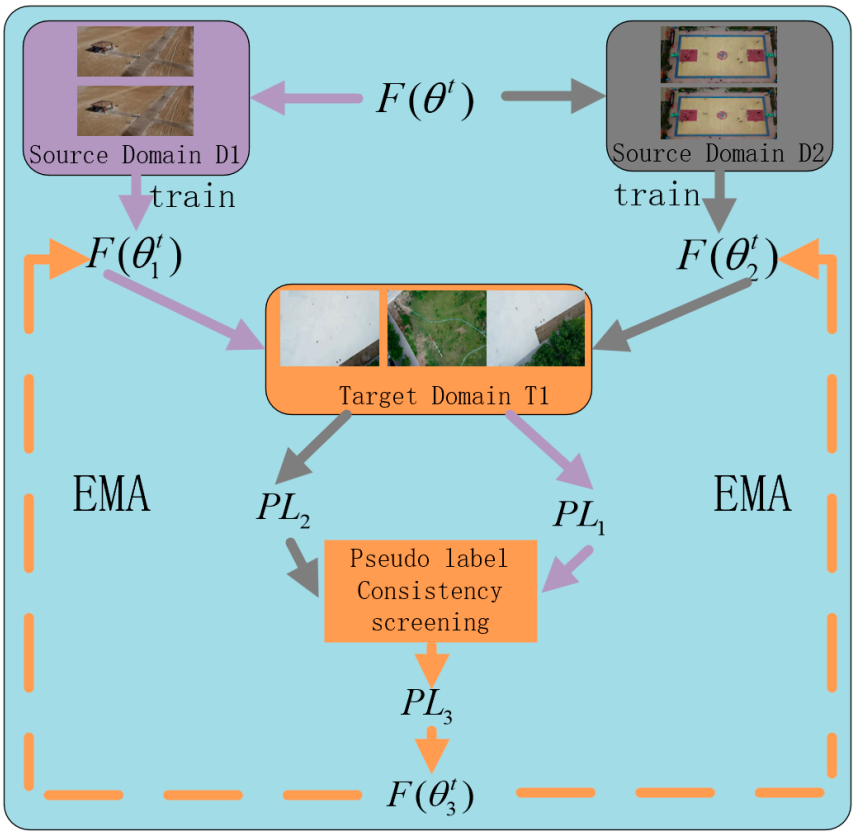

Figure 4. Overall Method Framework Diagram.

Then, the process moves to the next round, where $F\left(\theta_1^{t+1}\right)$ and $F\left(\theta_2^{t+1}\right)$ again generate the next round of pseudo-label candidate boxes PL_1^{t+1} and PL_2^{t+1} in T_1, thus forming the entire closed-loop operation. Throughout the process, while maintaining the SwinTrack structure, the model's adaptability is gradually transferred from the source domain to the target domain through iterative training, thereby improving the robustness and cross-domain capability of tracking single weak targets.

3.2. Background Enhancement Annotation

This section introduces a proposed annotation strategy for enhancing backgrounds to alleviate the limited information content of small targets. In the original coordinate system, the dimensions of the target bounding box are enlarged proportionally based on its center to serve as the template bounding box. A supervised construction strategy is employed:

classification focuses on the enhanced bounding box while regression points to the original bounding box. This involves quality-aware and supervised classification operations on the positive samples corresponding to the enhanced bounding boxes, while regression uses the original bounding box as the target, ensuring consistency in evaluation criteria and predicted geometry.

In implementation, the template and search bounding boxes are cropped at the center with the same magnification factor and resampled to a fixed size. If any bounding errors occur, edge padding is applied. The entire computation and memory overhead is equivalent to the baseline. Specifically, during training, the source domain bounding boxes and generated target domain pseudo-labels are magnified at the same ratio to ensure consistent cross-domain supervised training. During inference, the uniformly magnified enhanced bounding boxes are only used for cropping and feature extraction, and the final prediction result is written back to the original bounding box in the original coordinate system. The above strategy explicitly encodes the prior of the effective reference of the steady-state background in adjacent frames: it not only increases the proportion of effective tokens for weak targets in the entire input, but also provides a more robust geometric contrast in the case of occlusion or slight deformation; thus, on the baseline, it obtains effective gains for weak target tasks under a consistent decision index system, which facilitates subsequent ablation and comparison experiments.

As shown in Figure 5, the left side shows the default annotation, where the bounding box only indicates the target size, resulting in limited effective target information within the ROI. The right side shows the uniformly enlarged augmented bounding box, where the center remains unchanged while the bounding box is proportionally enlarged according to global parameters, thus including more background information within the ROI. During the training phase, the classification branch generates positive samples and quality supervision within the augmented bounding box shown on the right, thereby improving attention concentration and cross-domain robustness. The regression branch continues to use the original bounding box on the left to backpropagate, maintaining consistency in the output geometric parameters and evaluation criteria.



Figure 5. Example diagram of background enhancement strategy.

To verify the effectiveness of this method, this paper compares the response plot with the baseline and obtains the comparative data, which is intuitively displayed in the following comparison plot.

As shown in Figure 6 above, the response comparison diagram shows the baseline model on the left and the model with background enhancement on the right. It can be clearly seen that the response values in the baseline model's response map are relatively scattered, with higher response values in the background region, resulting in low distinction between the target and the background. In contrast, the response map with background enhancement on the right shows an increase in response values around the target, exhibiting

clearer response peaks, while the response values in the background region are significantly lower than the baseline. This demonstrates that the background enhancement annotation method improves model learning and thus enhances the model's performance in tracking small targets by incorporating information from the target's near-field background.

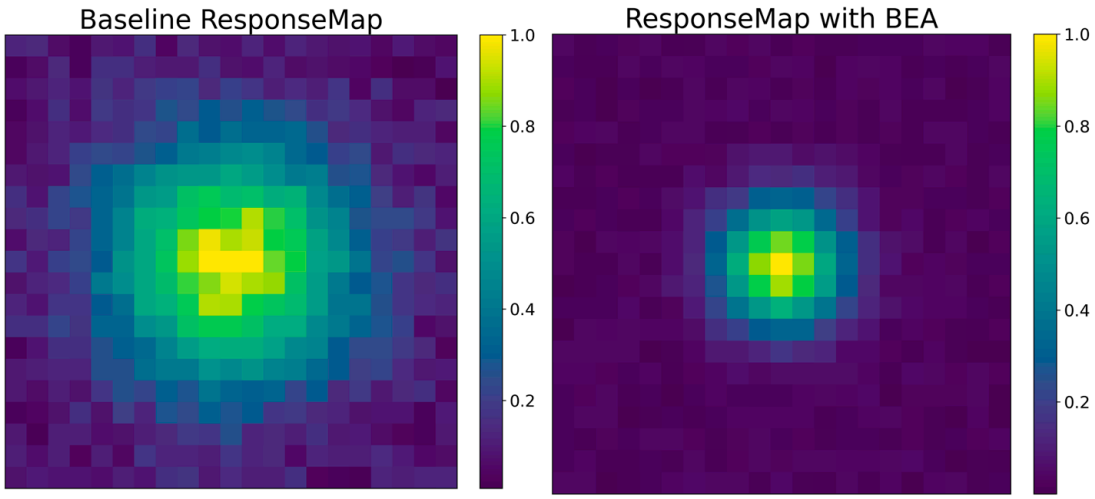

Figure 6. Baseline vs. Background Enhancement Response Map Comparison.

As shown in Figure 7 above, the score distribution is as follows: the left side represents the baseline score, and the right side represents the score distribution after introducing the background enhancement strategy. The results show that the baseline model's score distribution has a large overlap between the target and background scores, indicating a weak distinction between them. In contrast, the right side, after introducing background enhancement, shows that the target area scores significantly more than the background, thus better distinguishing the target from the background. This demonstrates that the model with background enhancement strengthens the distinction between the template and the background, providing a clearer criterion for subsequent pseudo-label selection and model iteration.

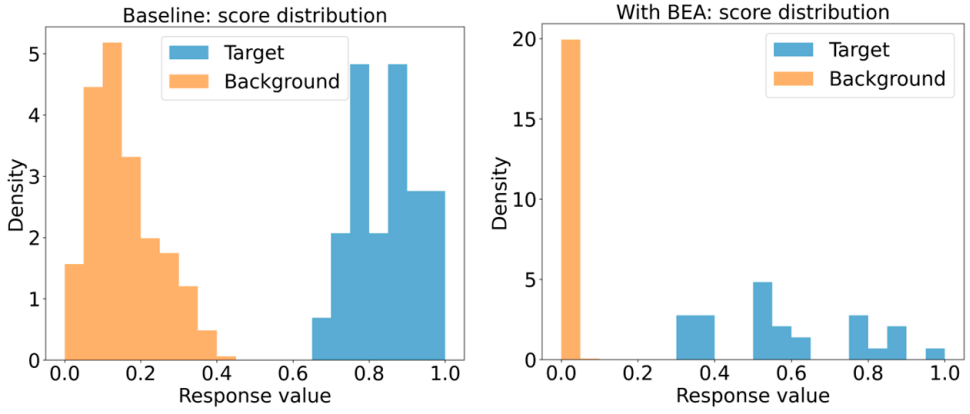

Figure 7. Comparison of baseline and background enhancement score distribution.

3.3. Consistency Screening Gate

To address the issue of reliable monitoring signals for unlabeled target domains, this section proposes a consistency screening gating mechanism combined with uniqueness for pseudo-label refinement. The pseudo-label consistency screening process is described below:

Suppose that in the t-th iteration, the source domain models (θ_1^t) and $F(\theta_2^t)$ are used to perform inference in the target domain T_1, resulting in two initial pseudo-labels PL_1^t and PL_2^t:

$$PL_1^t = \sum_{i=0}^{N-1}\left(b_{1,i}^t, s_{1,i}^t\right),\ PL_2^t = \sum_{j=0}^{M-1}\left(b_{1,j}^t, s_{1,j}^t\right) \tag{2}$$

Then, the two initial pseudo-labels are screened using consistency gating:

$$IoU\left(b_{1,i}^t, b_{2,j}^t\right) \geq \tau_{iou},\ \left|s_{1,i}^t - s_{2,j}^t\right| \leq \tau_{cls} \tag{3}$$

The screened pseudo-labels are retained and merged to make them unique:

$$PL_3^t = \left\{\left(b_{t,k}^{sup}, s_{t,k}^{sup}\right) \middle| \left(b_{1,i}^t, s_{1,i}^t\right), \left(b_{1,j}^t, s_{1,j}^t\right) satisfy(3)\right\}_{k=1}^{K_t} \tag{4}$$

where K_t is the number of pseudo-labels retained after consistency screening. Then, a teacher model is trained under the supervision of these pseudo-labels.

As shown in Figure 8, the horizontal axis represents the consistency score, and the vertical axis represents the sample count. Color coding indicates that blue areas represent rejected samples, and orange areas represent accepted samples. The dashed line in the figure represents the threshold, which is approximately 0.6. Observing the distribution plot, a clear bimodal structure can be observed, indicating that there are two significant concentrated areas in the consistency score. Therefore, setting a threshold can effectively remove low-quality mislabeled data, thus providing cleaner and more reliable input data for subsequent data fusion.

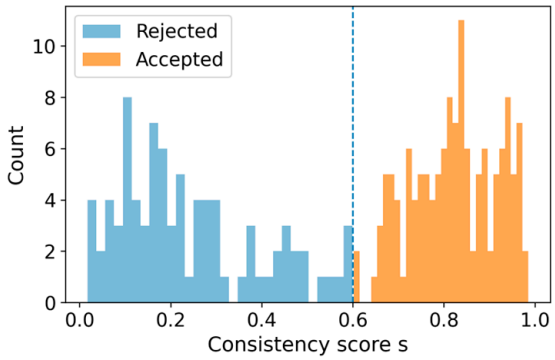

Figure 8. Consistency-gated distribution histogram.

As shown in Figure 9, the left figure illustrates the prediction results of two initial models for the same target frame. It can be seen that both models have low confidence levels and low geometric overlap, resulting in a multi-peaked response distribution and a shift in the target center. These factors lead to poor frame quality, therefore this frame was not included in the supervision process and no fused bounding box was generated. The right figure shows the filtered pseudo-label, whose performance is significantly better than the initial model prediction results in the left figure. Specifically, the prediction results of the two models highly overlap, have high confidence levels, and exhibit a single-peaked response, indicating good target alignment. Since the prediction results meet the fusion criteria, the pseudo-labels are merged into a single label and used as the supervision signal for subsequent training. The quality of the pseudo-label directly affects its effectiveness as a supervision signal. When the pseudo-label quality is low, it will have a significant negative impact on cross-domain training; therefore, quality filtering of the pseudo-labels is crucial to ensuring the stability and efficiency of the training process.

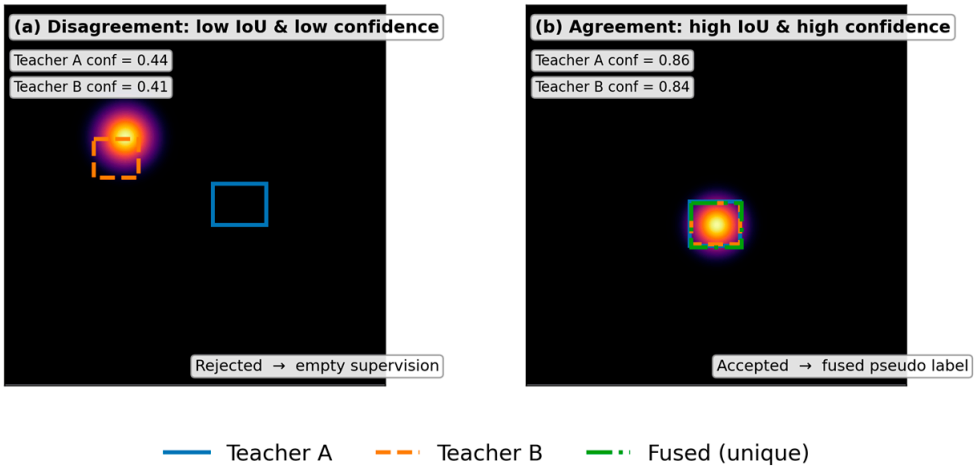

Figure 9. Example of target frame comparison.

As shown in Figure 10 above, the pass rate and loss variance statistics are as follows: the left axis and solid line represent the pass rate, while the right axis and dashed line represent the training loss variance of the teacher model. As the epochs are iterated, the pass rate gradually increases, indicating that the quality of the pseudo-labels is also improving. At the same time, the loss variance shows a rapid decrease and gradually tends to stabilize, which represents the gradual improvement of the stability of PL_3 and is also evidence of the model's convergence.

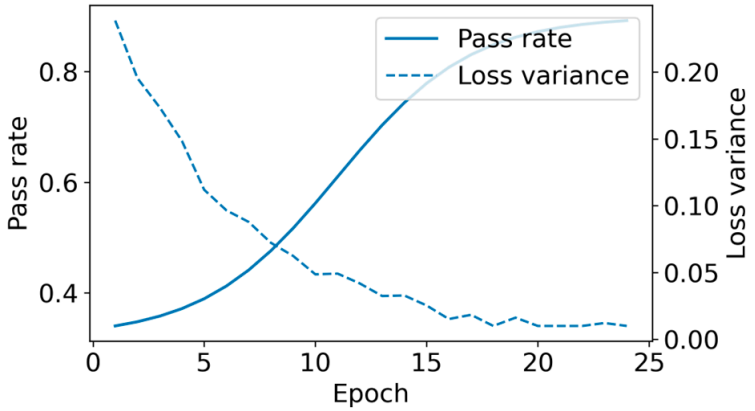

Figure 10. Gating pass rate and loss variance.

3.4. Overall Algorithm Design

This section focuses on the algorithm flow of SwinMR as a whole. The two student models in the algorithm flow box in the table below were trained under supervised training using the source domains D_1 and D_2 for the first time, and subsequently updated by the teacher EMA to ensure stability during the iteration process (Algorithm 1). Unlabeled data in the target domain is generated independently by the two students, while retaining their classification confidence and regression quality indices. Predictions with large biases are then eliminated through consistency constraints of IoU and response scores, preventing erroneous pseudo-labels from being introduced into the training loop. Among the selected candidate pseudo-labels, the system introduces a weighted soft fusion mechanism to integrate the predictions of the two students according to their stability and confidence, thereby obtaining higher-quality pseudo-labels in terms of geometric location and response accuracy.

Algorithm 1. SwinMR-Mutual Refinement Self-Training Framework

Input:
D_1, D_2: labeled source-domain datasets
T_1: unlabeled target-domain dataset
$F(\cdot|\theta)$: SwinTrack-based tracking model
θ_1, θ_2: student model parameters are obtained only initially from training on D_1 *and* D_2.
θ_1', θ_2': updated by the teacher model EMA
Hyper-parameters: T_{max}, t_{cons}, α, $K_{iter}(epoch)$
Output:
Adapted model parameters θ_1, θ_2 and fused teacher θ_3
/* when t = 0, supervised pretraining on source domains */
1 $\theta_1 \leftarrow$ train_supervised(D_1)
2 $\theta_2 \leftarrow$ train_supervised(D_2)

/* Iterative mutual-refinement loop */
For t = 1 to T_{max} do
  /* Stage1: teacher inference on target domain */
  $P_1^{teacher} \leftarrow$ infer(θ_1, D_3)/boxes b_1, scores s_1, *ResponseMap* R_1/
  $P_2^{teacher} \leftarrow$ infer(θ_2, D_3)/boxes b_2, scores s_2, *ResponseMap* R_2/

  /* Stage2: consistency filtering of predictions */
  for each frame f in T_1:
    for each prediction pair(p_1, p_2):
      c $\leftarrow$ consistency_score(p_1, p_2)
      if $c \geq t_{cons}$ then
        Add (p_1, p_2) to S_f

  /* Stage3: unique fusion and ResponseMap construction */
  $PL_3 \leftarrow \varnothing$
  for each frame f:
    $PL_f \leftarrow$ fuse_unique($P_1^{teacher}[f]$, $P_2^{teacher}[f]$, S_f)
    Add PL_f to PL_3

  /*Stage4:train teacher using fused pseudo-labels */
  $\theta_3 \leftarrow$ train_supervised(PL_3, K_{iter})

  /* Stage5:EMA undate of teachers form θ_3 */
  $\theta_1' \leftarrow \alpha\theta_1 + (1-\alpha)\theta_3$
  $\theta_2' \leftarrow \alpha\theta_2 + (1-\alpha)\theta_3$

  /* optional fine-tuning students on pseudo-labels */
  $\theta_1 \leftarrow$ train_supervised(θ_1, PL_3)
  $\theta_2 \leftarrow$ train_supervised(θ_2, PL_3)

end for

The final pseudo-labels adopt a ResponseMap structure compatible with SwinTrack, expressing the target location through a unique positive sample grid and CXCYWH normalization, allowing the teacher model to directly connect to existing tracking heads for

training. During fixed-round training, the teacher model relies entirely on pseudo-labels for self-supervised learning. The classification branch uses Varifocal loss to enhance sensitivity to low-response regions of small targets, while the regression branch uses GIoU loss to ensure geometric accuracy in range prediction. After each training round, the teacher model's weight updates are synchronized to both students via EMA, forming a stable closed-loop mutual teaching mechanism that jointly drives the continuous evolution of both teacher and students.

This multi-model collaborative mutual teaching design effectively avoids the uncontrollable accumulation of pseudo-label noise in single-teacher self-training, ensuring that pseudo-label quality is consistently improved in a controlled manner during iterations. The stability of the teacher model is also strengthened, allowing the student model to gradually learn more robust tracking capabilities across different scenarios. In experiments with real low-altitude overhead videos, this framework demonstrates significant advantages under challenging conditions such as extremely small scales, complex background interference, and target blur, with stable overall performance improvements, fully demonstrating its effectiveness and application value in cross-domain tracking tasks for small targets.

4. Experiments

4.1. Implementation Details

In the initial model training in the source domain, the model used two manually labeled public datasets, uav123 and VisDrone2019-SOT. The cropping sizes of the input template and the search region were fixed at 128 × 128 and 256 × 256, respectively. The patch size was 4, the window was divided into 7 × 7, and the backbone network used SwinTransformerV2Block. Each batch contained 32 pairs of samples. The optimizer used was AdamW, and the initial learning rate and weight decay were set to 1×10^{-4}. In addition, the value β was set to (0.9, 0.999). The learning rate was gradually reduced every 40 epochs, i.e., a cosine annealing strategy was adopted. To prevent overfitting of the model, data augmentation operations such as random flipping and color perturbation were added during training. Target domain 1 consists of pedestrian sequences captured by drones over the campus, used for pseudo-label iterative training to drive the model's generalization to the target domain. Target domain 2 also consists of pedestrian sequences captured by drones over the campus, but its overall feature distribution differs significantly from target domain 1, demonstrating the model's generalization ability between two domains with large feature distributions. Of the two target domains, one participates in iterative optimization, while the other serves as the data for the comparative experiment, testing al models.

In the target domain mutual teaching phase, two initial models infer predicted bounding boxes. These boxes are then filtered for consistency based on geometric distribution and confidence threshold. Qualified frames and pseudo-labels proceed to the next round of iterative training. The filtered pseudo-label predictions, after unique fusion, serve as supervisory signals for further training of both models. Frames that fail the fusion are skipped by default. The training process in this phase is consistent with the source domain, but the learning rate is further reduced to 5×10^{-5}, and the batch size is adjusted to 16 to prevent gradient oscillations that could lead to convergence instability. Furthermore, gradient clipping is enabled in each iteration, with the clip norm set to 0.5, accompanied by momentum smoothing to suppress the negative transfer effect of low-quality pseudo-label noise on the model. For loss, the classification branch uses a weighted combination of QFL and GIoU from the regression branch. The quality objective of the classification branch is the calibration score after unique fusion of pseudo-labels, while the regression branch regresses the unique positive sample grid to the geometric parameters of the pseudo-bounding boxes.

During cross-domain training, the model parameters in each round are updated with the refined results of the pseudo-labels from the previous round, forming an iterative generalization process in which the model gradually adapts to the target domain. During training, the backbone weights of several earlier layers are frozen, and a multi-stage warm-up strategy is used to mitigate training instability caused by the distribution differences between the source and target domains. The experiment was conducted on a 2080Ti × 4 platform, taking a total of 36 h. The experiment verifies that this configuration, while ensuring convergence speed is not affected, reduces the impact of pseudo-label noise, thereby improving the model's generalization ability in the weak target tracking task.

4.2. Loss Function

Let the positive sample grid of frame t be g_t^+, the supervision box be $\sum_{i=1}^{N_t} b_t^{sup,i}$, and the quality score target be $\sum_{i=1}^{N_t} s_t^{sup,i}$ (where the source domain is set to 1.0 and the target domain is the uniquely fused score). Furthermore, let the prediction of the classification branch be $\hat{y}_t(g) \in [0,1]$, and the geometric prediction of the regression branch be $b_t^{sup,i}$. Then, the QFL loss of the classification branch can be expressed as:

$$\mathcal{L}_{cls}(t) = \sum_{g \neq g_t{}^+} w(g)\hat{y}_t(g)^r[-\log(1-\hat{y}_t(g))] + (\sum_{i=1}^{N_t} s_t^{sup,i})(1-\hat{y}_t(g_t{}^+))^r[-log\hat{y}_t(g_t{}^+)] \quad (5)$$

where r is the consistent modulation factor in the baseline, and $w(g)$ is the background weight: to better utilize the background enhancement effect, the ordinary background is set to 1, while the suppression near the enhancement box boundary is $w_{nb} \in \left[1, w_{fb}\right]$, and the strong negative samples at a distance are set to $w_{fb} > 1$; the quality target $s_t^{sup,i}$ is only taken at the unique positive sample grid.

The GIoU used for the regression loss can be expressed as:

$$\mathcal{L}_{reg}(t) = \sum_{i=1}^{N_t} \left[1 - GioU\left(b_t^{(i)}, b_t^{sup,i}\right)\right] \quad (6)$$

The supervision loss is computed only at the unique positive sample grid g_t^+, and the source domain is aligned with the original ground truth box while the target domain uses a shared pseudo-label box. Therefore, the supervision loss for each frame can be expressed as:

$$\mathcal{L}_{frame}(t) = \lambda_{cls}\mathcal{L}_{cls}(t) + \lambda_{reg}\mathcal{L}_{reg}(t) \quad (7)$$

Since λ_{cls} and λ_{reg} are consistent with the baseline, the total loss can be expressed as:

$$L = \sum_{t \in S_{src}} \mathcal{L}_{frame}(t) + \beta(r) \sum_{t \in S_{tgt}} \mathcal{L}_{frame}(t) \quad (8)$$

where S_{src} and S_{tgt} represent the sample sets of the source domain and the target domain, respectively, and $\beta(r)$ represents the target domain weight that increases with r.

4.3. Evaluation Indicators

In terms of evaluation metrics, this paper adopts the OPE (One-Pass Evaluation) protocol: $\{I_t\}_{t=1}^T$ is defined as a given time-ordered set of consecutive frames, and $B_1 = (x_1, y_1, w_1, h_1)$ is the ground truth bounding box in the first frame. The tracker is initialized using (I_1, B_1), and the predicted bounding boxes in subsequent frames are $\hat{B}_t = \left(\hat{x}_t, \hat{y}_t, \hat{w}_t, \hat{h}_t\right)$. If a target is missing in a frame or fails to produce a valid output, IoU_t is considered 0. To avoid threshold drift caused by preprocessing scaling, all evaluation values are performed at the original image resolution. Results are given for all frames and

for the effective subset of weak targets, achieving comparison in both overall performance and weak target performance.

The success rate curve and the area under the curve (AUC) are used as evaluation metrics for localization stability.

Define

$$IoU_t = \left| \frac{B_t \cap \hat{B}_t}{B_t U \hat{B}_t} \right| \tag{9}$$

the Intersection over Union (IoU) ratio for each frame,

$$S(u) = \frac{1}{T} \Sigma_{t=1}^{T} \mathbf{1}\{IoU_t \geq u\} \tag{10}$$

as the success rate, and set by a threshold u ∈ [0, 1], while

$$AUC = \int_0^1 S(u) d_u \tag{11}$$

The overall stability of the method is evaluated comprehensively from the perspectives of scale estimation and position alignment, set as the overall metric and independent of the threshold.

$$e_t = \left\| \left(\hat{x}_t + \frac{\hat{w}_t}{2}, \hat{y}_t + \frac{\hat{h}_t}{2} \right) - \left(x_t + \frac{w_t}{2}, y_t + \frac{h_t}{2} \right) \right\|_2 \tag{12}$$

Based on pixel center error,

$$P(\tau) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{e_t \leq \tau\} \tag{13}$$

The point value P@20 at τ = 20 pixels is used as the reporting accuracy curve to characterize the intuitive pixel-level center alignment capability and thus define the center accuracy. To address the issue of intuitive comparability across sequences in scenes with weak targets and mixed resolutions, this paper requires scale-normalized center error: setting the diagonal length of the ground truth bounding box to.

Calculate the area under the curve with normalized precision in the range τ ∈ [0, 0.5].

$$\mathrm{NP} - \mathrm{AUC} = \int_0^{0.5} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left\{ \widetilde{e}_t \leq \tau \right\} \tau \tag{14}$$

Finally, AUC, P@20, and NP-AUC are complemented in three aspects: IoU stability, pixel-scale intuitive accuracy, and cross-scale comparability, while remaining independent of network structure and training strategy to facilitate experimental reproducibility.

Ultimately, the success rate curves (AUC), central accuracy (P@20), and normalized accuracy (NP-AUC) described above are used to evaluate the overall method across all test frames. A threshold $a_t < 1024$ is used to explain the data composition and annotation caliber. Ablation experiments and comparative experiments are conducted to quantify the overall performance of this method in terms of weak target feature representation and tracking. The small-scale caliber defined by the area threshold designed above is naturally compatible with the COCO annotation format of consecutive frame datasets and has a greater advantage for slender targets. It also retains the statistical rules of the evaluation itself and is independent of the resolution and window scheduling during training and inference stages, thus quantifying the improvement of this method's ability to preserve weak target features.

4.4. Ablation Studies

This section uses SwinTrack as a baseline to conduct ablation experiments on the pseudo-label consistency mutual refinement method (hereinafter referred to as SSL) and background enhancement annotation (hereinafter referred to as BEA). Except for whether SSL or BEA is used, all other experimental configurations are strictly consistent: the

supervised training phase of the source domain uses UAV123 and VisDrone2019-SOT weak target union and is uniformly converted to COCO format; the self-training method of pseudo-label mutual refinement is only performed on the unlabeled target domain T1; and the final test results are all completed on the same unlabeled test target domain T2.

The overall results are shown in Figure 11 and Table 1. Compared to the baseline SwinTrack, using SSL alone provides a stable improvement in all three metrics. These gains further demonstrate that in cross-domain weak target tracking tasks, the consistent pseudo-label selection strategy combined with mutually refined iterative training can suppress the impact of early noise on pseudo-label quality, thus making the cross-domain capability more robust. Meanwhile, when using BEA alone, the improvements in P@20 and NP-AUC are more significant than the slight improvement in SUC: due to the strategy of regressing to the original box based on the classification reference augmentation box, the classification heatmap is more likely to form repeatable and sharp peaks in a more relevant and stable context, thereby improving normalization accuracy while reducing center point quantization error. Since the geometric caliber of the regression strategy remains unchanged, the overall IoU area, i.e., SUC, shows a moderate upward trend.

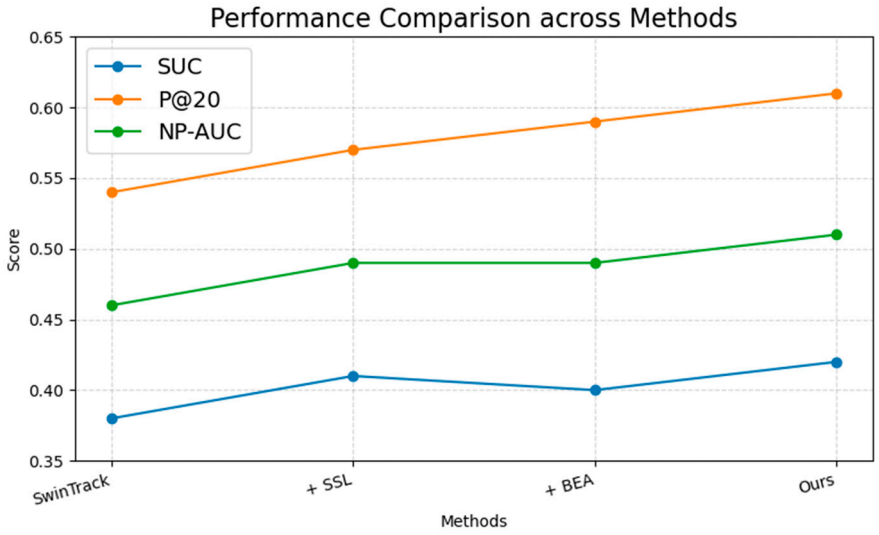

Figure 11. Ablation experiment comparison chart.

Table 1. Ablation experiment results.

Method	SUC	P@20	NP-AUC
SwinTrack	0.38	0.54	0.46
+SSL	0.41	0.57	0.49
+BEA	0.40	0.59	0.49
Ours	0.42	0.61	0.51

When the aforementioned SSL and BEA work together on the baseline (i.e., Ours), the three metrics are further improved: SSL provides high-quality pseudo-label signals for cross-domain mutual refinement training, thereby reducing the drift phenomenon caused by occlusion or shape deformation, while BEA provides more robust background information for each match as local upper and lower anchor points, thereby alleviating the limitation of effective information of weak targets themselves.

4.5. Comparison with Other Methods

This section presents comparative experiments on the proposed method's ability to generalize to weak targets and across domains. The target test domain is uniformly

set to the self-made dataset T2. The target domain used for generating pseudo-labels in the initial model inference for mutual supervision is set to the self-made dataset T1. Representative strong baselines from the same period are selected as comparison objects. Without additional post-processing at the inference end, both the proposed method and the baseline are tested on T2, and the following comparative experimental results are obtained.

As shown in Table 2 and Figure 12, the morphological characteristics of the three curves demonstrate that the proposed method significantly improves tracking performance in scenes with small targets. Specifically, the pseudo-label refining method effectively suppresses pseudo-label noise during training and utilizes relatively high-quality pseudo-labels for mutual supervision, thereby enhancing the overall generalization ability of the model. Furthermore, the background enhancement box method reduces pixel-level center error, thus improving P@20. When the background textures of the source and target domains differ significantly, SSL ensures a stable improvement in the model's generalization ability through improved pseudo-label quality. BEA, by providing stable background information, reduces model drift even in the event of slight occlusion or morphological deformation, thus ensuring overall performance improvement.

Table 2. Comparative experimental results.

Tracker	SUC	P@20	NP-AUC
Ocean	0.340	0.500	0.420
DiMP50	0.330	0.490	0.410
SiamR-CNN	0.360	0.520	0.440
TransT	0.350	0.510	0.430
STARK-ST50	0.360	0.520	0.440
KeepTrack	0.370	0.530	0.450
SwinTrack (T-224)	0.370	0.530	0.450
SwinTrack (plain)	0.380	0.540	0.460
Ours	0.420	0.610	0.510

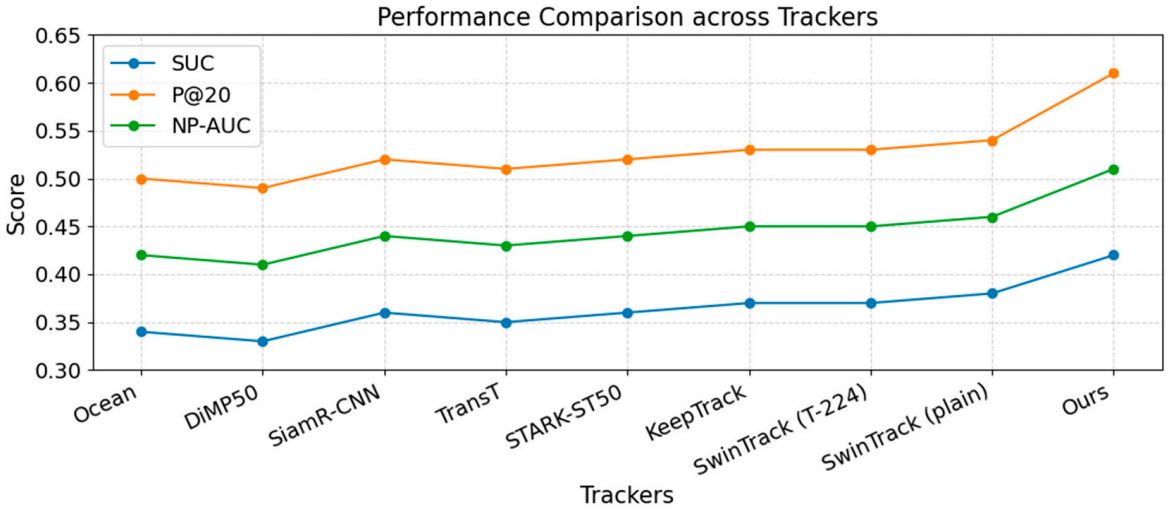

Figure 12. Comparison experiment comparison chart.

5. Conclusions

This paper addresses the limitations of effective information from small targets and the generalization ability of models in the field of small target tracking. Focusing on the research topic of small target tracking technology for low-altitude UAVs, it proposes a cross-domain self-training architecture for small target tracking, based on the SwinTrack as the baseline backbone network, which enhances the model's generalization ability

through mutual refinement. A complete technical system is comprehensively constructed, from theoretical framework modeling, algorithm design, experimental verification, to final engineering deployment. The baseline comparison video is publicly available at the following URL: https://github.com/yangchuanyuan/SwinMR (accessed on 3 November 2025). Specifically, it can be summarized as follows:

1. This paper first clarifies the problems of limited effective information and weak model generalization ability caused by the size of small targets. To address these problems, it first establishes the powerful feature extraction and self-attention mechanism of the Swin Transformer as the backbone network baseline architecture, laying the foundation for subsequent improvements.

2. To address the problem that each training set requires manual annotation to obtain satisfactory results due to the weak model generalization ability, this paper designs a mutual refinement iterative training mode based on the SwinTrack as the baseline architecture, inspired by MMT. This architecture introduces a dual-branch network into the field of small target tracking. Through mutual learning using pseudo-labels generated by the teacher model and filtered for consistency and then fused with unique identifiers, the model gradually adapts to the distribution of the target domain during continuous iteration, thereby improving its generalization ability.

3. To address the limitation of effective information caused by the scale of small targets, this paper proposes a Background Enhancement (BEA) strategy. This strategy utilizes the high background repetition rate between neighboring frames to expand the target bounding box by a uniform ratio, thus using the neighborhood background as effective information to improve the localization accuracy and tracking performance of small targets. The advantages of the BEA strategy and its robustness in feature discrimination are demonstrated in feature response visualization.

Although the proposed small target tracking framework has achieved good performance and is feasible for engineering deployment, several problems still need to be solved, and further research is needed to find solutions. Future work should focus on further improving model lightweighting, generalization ability, adaptability, and system synergy to promote the comprehensive development of the field of small target tracking.

While the model's structural optimization and lightweighting have been successfully ported to airborne computers, the required computing power is still insufficient for performing flight processing tasks in the air and requires further optimization. Future research could combine strategies such as dynamic pruning, knowledge distillation, or hybrid precision quantization to improve this problem. At the same time, it is also possible to consider combining edge computing or heterogeneous acceleration with a multi-level distributed inference architecture, that is, dividing the task level to the sensing layer, thereby improving the overall energy efficiency of the UAV architecture.

Author Contributions: Conceptualization, C.Y., Y.F. and S.Z.; Methodology, S.Z. and C.Y.; Writing—original draft, S.Z. and C.Y.; Writing—review & editing, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

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

Abbreviations

The following abbreviations are used in this manuscript:

SSL	Pseudo-label consistency semi-supervised learning
BEA	Background enhancement annotation
EMA	Exponential moving average
UAV	Unmanned Aerial Vehicle

References

1. Tian, X.; Jia, Y.; Luo, X.; Yin, J. Small target recognition and tracking based on UAV platform. *Sensors* **2022**, *22*, 6579. [CrossRef]
2. Zhao, M.; Yue, Q.; Sun, D.; Zhong, Y. Improved SwinTrack single target tracking algorithm based on spatio-temporal feature fusion. *IET Image Processing* **2023**, *17*, 2410–2421. [CrossRef]
3. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
4. Zeng, Z.; Li, X.; Fan, C.; Zou, L.; Chi, R. SwinEFT: A robust and powerful Swin transformer based event frame tracker. *Appl. Intell.* **2023**, *53*, 23564–23581. [CrossRef]
5. Mu, Q.; He, Z.; Wang, X.; Li, Z. SSTrack: An Object Tracking Algorithm Based on Spatial Scale Attention. *Appl. Sci.* **2024**, *14*, 2476. [CrossRef]
6. Wu, Q.E.; An, Z.; Chen, H.; Qian, X.; Sun, L. Small target recognition method on weak features. *Multimed. Tools Appl.* **2021**, *80*, 4183–4201. [CrossRef]
7. Kou, R.; Wang, C.; Yu, Y.; Peng, Z.; Huang, F.; Fu, Q. Infrared small target tracking algorithm via segmentation network and multistrategy fusion. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. [CrossRef]
8. Mirzaei, B.; Nezamabadi-Pour, H.; Raoof, A.; Derakhshani, R. Small object detection and tracking: A comprehensive review. *Sensors* **2023**, *23*, 6887. [CrossRef] [PubMed]
9. Zhao, M.; Li, W.; Li, L.; Hu, J.; Ma, P.; Tao, R. Single-frame infrared small-target detection: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 87–119. [CrossRef]
10. Sun, H.; Bai, J.; Yang, F.; Bai, X. Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
11. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
12. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef] [PubMed]
13. Hare, S.; Saffari, A.; Torr, P.; Struck, S. Structured output tracking with kernels. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 263–270.
14. Hao, P.Y.; Chiang, J.H.; Chen, Y.D. Possibilistic classification by support vector networks. *Neural Netw.* **2022**, *149*, 40–56. [CrossRef]
15. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Cham, Switzerland, 2016; pp. 850–865.
16. Cao, D.; Dai, R.; Zhao, T.; Alqahtani, F.; Tolba, A.; Sherratt, R.S.; Zhu, M. SiamIRPN: Siamese visual tracking with improved region proposal networks. In Proceedings of the 2023 International Conference on Frontiers of Robotics and Software Engineering (FRSE), Nanjing, China, 23–25 June 2023; pp. 165–173.
17. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
18. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H.S. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338.
19. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4660–4669.
20. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6182–6191.
21. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135.

22. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning spatio-temporal transformer for visual tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10448–10457.
23. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. Mixformer: End-to-end tracking with iterative mixed attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13608–13618.
24. Wang, Z.; Zhou, W.; Xia, Z.; Fan, B. Easy One-Stream Transformer Tracker. In Proceedings of the 2023 China Automation Congress (CAC), Chongqing, China, 17–19 November 2023; pp. 3183–3188.
25. Lin, L.; Fan, H.; Zhang, Z.; Xu, Y.; Ling, H. Swintrack: A simple and strong baseline for transformer tracking. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16743–16754.
26. 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, 8–14 September 2018; pp. 101–117.
27. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Learning spatially regularized correlation filters for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
28. Danelljan, M.; Bhat, G.; Khan, F.S.; FelsBerg, M. Efficient convolution operators for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6638–6646.
29. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G.S.F.C. Towards robust and accurate visual tracking with target estimation guidelines. *arXiv* **2020**, arXiv:1911.06188. [CrossRef]
30. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese box adaptive network for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6668–6677.
31. Jin, X.; Zhang, D.; Wu, Q.; Xiao, X.; Zhao, P.; Zheng, Z. Improved SiamCAR with ranking-based pruning and optimization for efficient UAV tracking. *Image Vis. Comput.* **2024**, *141*, 104886. [CrossRef]
32. Zhang, Z.; Peng, H.; Fu, J.; Li, B.; Hu, W. Ocean: Object-aware anchor-free tracking. In Proceedings of the 2020 European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 771–787.
33. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y.H.F.T. Hierarchical feature transformer for aerial tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15437–15446.
34. Yan, B.; Zhang, X.; Wang, D.; Lu, H.; Yang, X. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5289–5298.
35. Liu, X.; Yoo, C.; Xing, F.; Oh, H.; Fakhri, G.E.I.; Kang, J.W.; Woo, J. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Trans. Signal Inf. Process.* **2022**, *11*, e25. [CrossRef]
36. Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; Huang, T.S. Self-similarity grou: A simple unsupervised cross domain adaptation approach for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6112–6121.
37. Lu, Y.; Shen, M.; Ma, A.J.; Xie, X.; Lai, J.H. Mlnet: Mutual learning network with neighborhood invariance for universal domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 3900–3908.
38. Dai, Y.; Liu, J.; Bai, Y.; Tong, Z.; Duan, L.Y. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *IEEE Trans. Image Process.* **2021**, *30*, 7815–7829. [CrossRef]
39. Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; Tian, Y. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9021–9030.
40. Liu, Y.; Yang, X.; Zhou, S.; Liu, X.; Wang, Z.; Liang, K.; Tu, W.; Li, L.; Duan, J.; Chen, C. Hard sample aware network for contrastive deep graph clustering. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 8914–8922.
41. Ge, Y.; Chen, D.; Li, H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv* **2020**, arXiv:2001.01526. [CrossRef]
42. Ge, Y.; Zhu, F.; Chen, D.; Zhao, R. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11309–11321.
43. Oza, P.; Sindagi, V.A.; Patel, V.M. Unsupervised domain adaptation of object detectors: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 4018–4040. [CrossRef] [PubMed]
44. Liu, Y.C.; Ma, C.Y.; He, Z.; Kuo, C.W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; Vajda, P. Unbiased teacher for semi-supervised object detection. *arXiv* **2021**, arXiv:2102.09480. [CrossRef]
45. Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wei, F.; Bai, X.; Liu, Z. End-to-end semi-supervised object detection with soft teacher. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3060–3069.

46. Lu, Z.; Shuai, B.; Chen, Y.; Xu, Z.; Modolo, D. Self-supervised multi-object tracking with path consistency. *arXiv* **2024**, arXiv:2404.05136.
47. Panagiotakopoulos, T.; Dovesi, P.L.; Härenstam-Nielsen, L.; Poggi, M. Online domain adaptation for semantic segmentation in ever-changing conditions. In Proceedings of the 2020 European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 128–146.
48. Cai, B.; Ma, L.; Sun, Y. Dual consistent pseudo label generation for multi-source domain adaptation without source data for medical image segmentation. *Front. Neurosci.* **2023**, *17*, 1209132. [CrossRef] [PubMed]
49. He, T.; Shen, L.; Guo, Y.; Ding, G.; Guo, Z. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 879–887.

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.