sensors

MDPI

Article

MDF-Det: Motion-Aware Decoupling and Scene Filtering for Wide Area Small Moving Target Detection

Kangqiushi Li †, Xiaoran Zhang †, Zheng Zhang *, Huaxin Xiao and Yu Liu

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China; likangqiushi20@nudt.edu.cn (K.L.); zhangxiaoran@nudt.edu.cn (X.Z.); xiaohuaxin@nudt.edu.cn (H.X.); jaspnyuliu@nudt.edu.cn (Y.L.)

* Correspondence: zhangzheng@nudt.edu.cn

† These authors contributed equally to this work.

Highlights

What are the main findings?

- MDF-Det establishes a unified coarse-to-fine framework that jointly addresses weak-target preservation, dense-target decoupling, and scene-induced false-alarm suppression in wide-area motion imagery.
- The spatial attention-guided decoupling and learned scene-prior filtering mechanisms separate merged target responses and suppress contextually implausible detections, enabling MDF-Det to achieve an average F_1 score of 0.878 across six WPAFB AOIs.

What are the implications of the main findings?

- The results demonstrate that reliable WAMI detection requires coordinated control of candidate recall, dense-target separation, and scene-level contextual filtering rather than relying solely on local motion cues.
- The proposed modular framework provides an interpretable and extensible solution for detecting tiny and densely distributed moving targets under complex wide-area aerial imaging conditions.

Abstract

Object detection in Wide Area Motion Imagery (WAMI) is crucial for large-scale intelligent surveillance and monitoring systems. However, detecting extremely small moving targets in low-frame-rate grayscale WAMI remains highly challenging. In particular, three key problems limit the performance of existing methods: weak motion responses caused by low target contrast can lead to missed detections; densely distributed targets often produce merged responses that are difficult to separate; and registration artifacts, parallax, and dynamic background clutter can generate numerous false alarms. To address these problems, we propose MDF-Det, a coarse-to-fine spatiotemporal framework for WAMI small moving target detection. The framework consists of three complementary components. First, a Motion and Appearance Feature Fusion (MAFF) strategy integrates dense optical-flow-derived motion saliency with grayscale frame differencing to enhance weak target responses and improve candidate preservation. Second, a Spatial Attention-Guided Target Decoupling (SA-TD) module employs fine-grained heatmap decoding and step-threshold degradation to separate merged responses in densely populated scenes. Finally, a Scene-Prior Guided Filtering (SPGF) mechanism learns complementary vehicle-accessibility and motion-activity priors from scene context to suppress contextually implausible false alarms caused by complex background interference. Extensive experiments on six evaluation AOIs

Check for updates

Academic Editor: Zafar Ali

Received: 4 August 2026
Revised: 8 September 2026
Accepted: 10 September 2026
Published: 14 September 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.

of the WPAFB 2009 dataset demonstrate that MDF-Det achieves an average F_1 score of 0.878, corresponding to a relative improvement of 5.5% over the strongest baseline.

Keywords: wide area motion imagery; small moving target detection; motion-appearance fusion; target decoupling; scene-prior filtering

1. Introduction

Wide-Area Motion Imagery (WAMI) is an advanced aerial imaging technology that enables persistent coverage of large geographical areas, often spanning entire cities or tens of square kilometers [1]. The detection and tracking of moving objects such as vehicles and pedestrians in WAMI constitute important tasks for a range of emerging applications, including urban monitoring, critical infrastructure protection, disaster response, border surveillance, and defense [2–4]. Recent studies have also explored neuromorphic event cameras for WAMI, demonstrating their potential for effective aerial sensing under low-frame-rate and challenging illumination conditions [5]. With continued advancements in sensor and platform technology, WAMI data collection has become increasingly accessible, driving a growing demand for object detection and tracking algorithms to support effective wide-area surveillance.

However, object detection in WAMI presents unique and formidable challenges when compared to the task of object detection in ordinary videos [1,6,7]. First, the individual frames have extremely high resolutions. For example, the WPAFB dataset [8] contains frames exceeding 315 megapixels, whereas moving vehicles typically occupy only about 50 pixels on average [9], as illustrated in Figure 1a. Second, WAMI data are often acquired at a low frame rate (typically about 1∼2 frames per second). This slow temporal sampling causes fast-moving vehicles to exhibit large interframe displacements that often exceed the target's own size, which severely undermines conventional motion-based methods that rely on small, smooth motion assumptions [6,10]. Third, these tiny targets are densely distributed across complex, heterogeneous aerial backgrounds, making efficient and accurate localization highly challenging.

To address these challenges, WAMI object detection has evolved from traditional motion-based techniques [4,10] to deep spatiotemporal frameworks [6,11,12]. Yet, despite these advances, the extreme imaging conditions of WAMI still expose several critical limitations in existing approaches, as illustrated in Figure 1. First, the extremely small target size, weak appearance contrast, and complex aerial backgrounds often obscure subtle motion responses, making faint targets difficult to preserve during candidate generation (Figure 1b) [6,10]. Second, in high-density traffic scenes, neighboring vehicles may produce highly overlapping responses, which severely hinders conventional heatmap regression from separating densely packed targets (Figure 1c) [3,4,12]. Third, without scene-level contextual constraints, existing detectors may misclassify registration artifacts, building parallax, swaying vegetation, and other deceptive background responses as genuine moving targets, resulting in numerous contextually implausible false positives (Figure 1d) [3,10].

To deal with these problems, we propose MDF-Det, a coarse-to-fine framework for vehicle detection in WAMI. By integrating temporal motion priors with grayscale appearance-change features, the proposed framework adaptively enhances low-contrast targets and subtle motion signals at the front end. It then exploits spatial attention mechanisms to separate densely clustered targets in the deep feature space, while further incorporating scene-level contextual priors to suppress background artifacts. Together, these components

form an effective cascaded detection pipeline that addresses the inherent challenges of wide-area small moving target detection.

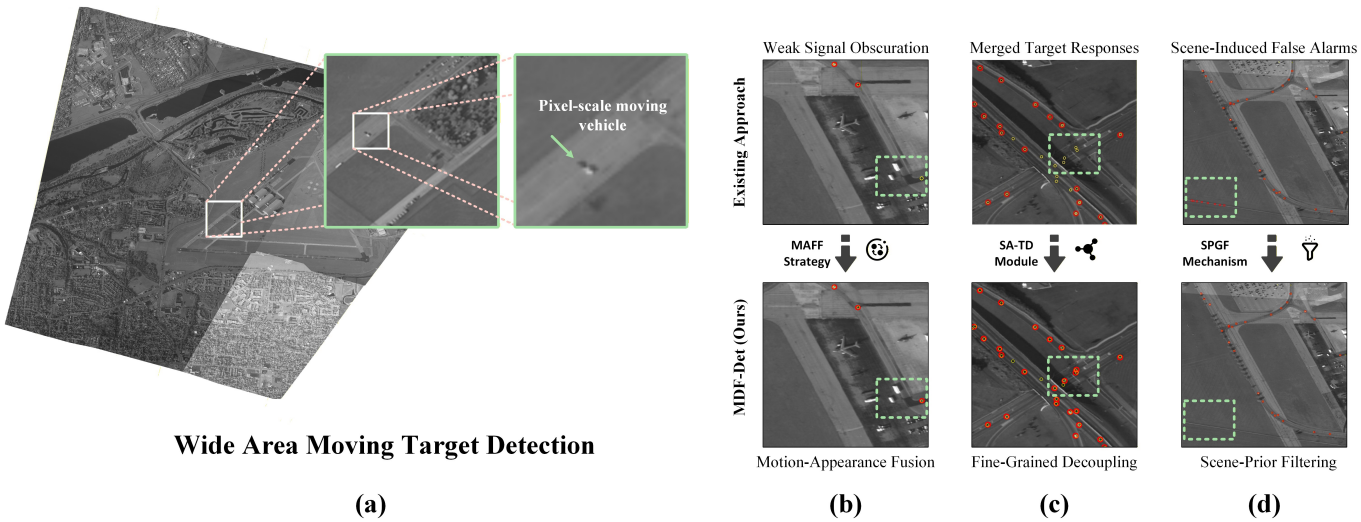


Figure 1. Conceptual illustration of MDF-Det for wide-area moving target detection. (**a**) Full-scale WAMI scene with zoomed-in pixel-scale vehicles. (**b**–**d**) Comparison with conventional methods under weak signal obscuration, merged target responses, and scene-induced false alarms. MDF-Det addresses these challenges through motion-appearance fusion, fine-grained decoupling, and scene-prior filtering, respectively. Green dashed boxes mark representative regions, red circles denote predictions, and yellow circles denote ground truths.

The main contributions of this work are summarized as follows:

- We propose MDF-Det, a unified coarse-to-fine framework for small moving target detection in WAMI. The proposed cascaded pipeline sequentially improves candidate quality through motion–appearance fusion, separates densely clustered targets, and suppresses false positives via scene-level filtering.
- We develop a Motion and Appearance Feature Fusion (MAFF) strategy. By integrating temporal motion saliency derived from optical flow with grayscale difference features, MAFF adaptively enhances low-contrast signals, thereby improving the detection capability for faint targets.
- We design a Spatial Attention-Guided Target Decoupling (SA-TD) module. By combining fine-grained heatmap decoding with a step-threshold degradation algorithm, the proposed module effectively separates overlapping targets in highly crowded scenarios.
- We introduce a Scene-Prior Guided Filtering (SPGF) mechanism. By learning complementary vehicle-accessibility and motion-activity priors from trajectory-supervised scene context, SPGF suppresses false positives caused by background clutter in contextually implausible regions.
- Extensive experiments on the WPAFB 2009 dataset demonstrate that MDF-Det achieves the highest F_1 scores across all six evaluation AOIs, with an average F_1 score of 0.878, corresponding to a relative improvement of 5.5% over the strongest baseline.

The remainder of this paper is organized as follows. Section 2 reviews the literature on moving object detection in WAMI. Section 3 presents the proposed MDF-Det framework and its core components. Section 4 describes the dataset, evaluation metrics, baseline methods, implementation details, and network training settings. Section 5 presents the experimental results. Finally, Section 6 discusses the main findings and limitations, and Section 7 concludes the paper.

2. Related Work

The primary objective of object detection in WAMI is to efficiently localize and track high-value dynamic targets, such as moving vehicles, across gigapixel-scale aerial backgrounds. Based on methodological evolution, existing detection approaches can be broadly categorized into three distinct classes: traditional techniques relying on temporal motion cues, two-stage spatiotemporal architectures, and end-to-end single-stage networks driven by heatmap regression.

2.1. Traditional Methods Utilizing Temporal Motion Cues

Early research in WAMI primarily relies on temporal motion cues to compensate for the scarcity of spatial appearance details. Background subtraction and multi-frame differencing constitute the foundational techniques for extracting motion pixels from quasi-static backgrounds. Although initial approaches frequently employ two- or three-frame differencing to capture dynamic responses, they suffer from the aperture problem for slow-moving targets and ghosting artifacts induced by the displacement of objects [12]. To adapt to gradual environmental variations, subsequent research introduces more statistical techniques for background modeling—such as adaptive Gaussian Mixture Models [13] and median or mean filtering algorithms [4,14]—to separate moving vehicles from static scenes.

Despite the computational efficiency of these methods, they remain highly susceptible to errors in image registration. Specifically, parallax-induced artifacts frequently generate false positives within these traditional frameworks. Furthermore, they inherently struggle to detect slow-moving targets and objects with low visual contrast against the background.

2.2. Two-Stage Architectures with Spatiotemporal Context

To balance the extensive background search space against computational constraints, two-stage coarse-to-fine detection architectures emerge as a pivotal strategy in this domain. Early hybrid frameworks typically employ traditional techniques, such as background subtraction, to generate candidate regions for initial coarse screening. Subsequently, they cascade lightweight Convolutional Neural Networks (CNNs) as backend classifiers to filter out false positives and separate tightly clustered targets [15]. Although these pipelines prioritize efficiency, the overall performance remains highly susceptible to the quality of the initial physical preprocessing.

To circumvent these limitations, subsequent research transitions the two-stage approach toward fully deep convolutional architectures. Pioneering this shift, ClusterNet [11] introduces an anchor-free, fully deep hierarchical framework. In this architecture, the first stage extracts large-scale motion clusters, whereas the second stage performs refined localization via heatmap regression. This design improves detection sensitivity and represents an early deep-learning approach in WAMI for detecting tiny vehicles.

Although two-stage architectures effectively balance search efficiency and detection accuracy, they remain heavily dependent on the quality of proposals generated in the first stage. If a missed detection occurs during this initial phase, the downstream deep networks lack the inherent capacity to recover the target, which ultimately leads to irrecoverable target loss.

2.3. Heatmap-Based Single-Stage Regression Networks

Inspired by anchor-free methods such as CenterNet [16], regression based on heatmaps gradually emerges as a leading approach in the field of WAMI. By representing targets as two-dimensional Gaussian keypoints and directly regressing the center positions, these methods effectively bypass the complex design and matching heuristics associated with traditional anchor boxes. Representative networks, such as HM-Net [6] and HMRN [17],

extend this concept by incorporating the image and the predicted heatmap of the previous frame as temporal feedback priors alongside the input of the current frame. This architecture enables the joint detection and prediction of target displacement in a single forward pass, thereby efficiently facilitating the tracking of multiple targets.

To capture complex long-range spatiotemporal dynamics, Negin et al. [12] introduce a self-attention mechanism that leverages VQ-VAE to extract unsupervised, discrete spatiotemporal embeddings. These latent embeddings are subsequently integrated into an encoder-decoder architecture based on the Transformer to generate refined heatmaps of keypoints. Furthermore, by incorporating the regularization of shape priors into the loss function, the framework encourages the network to produce responses that align with the physical morphology of small targets.

However, lacking decoupling mechanisms for complex dense scenarios, these methods frequently suffer from overlapping responses in heatmaps when targets are clustered, which leads to a significant degradation in the accuracy of localization.

Limitations and Motivation: In summary, although existing approaches have advanced vehicle detection in WAMI, they still struggle to jointly preserve weak motion responses and separate densely clustered targets during subsequent refinement. Motivated by these limitations, we reformulate the conventional coarse-to-fine detection pipeline and propose MDF-Det. By enhancing faint motion signals through motion priors, separating overlapping targets via fine-grained heatmap decoding, and suppressing contextually implausible false positives using scene-level contextual constraints, MDF-Det provides a unified framework for wide-area moving target detection.

3. Proposed Method

3.1. Overview

As illustrated in Figure 2, MDF-Det adopts a coarse-to-fine architecture for wide-area moving target detection. The pipeline begins with the Motion and Appearance Feature Fusion (MAFF) strategy, which incorporates temporal motion priors to enhance weak target responses and generate coarse candidates. These candidates are subsequently processed by the Spatial Attention-Guided Target Decoupling (SA-TD) module, where fine-grained heatmap decoding separates merged target responses in densely populated scenes. Finally, the Scene-Prior Guided Filtering (SPGF) mechanism exploits learned scene-level contextual priors to suppress false responses caused by parallax and dynamic background clutter, yielding the final detection results.

3.2. Motion and Appearance Feature Fusion Strategy

In WAMI, traditional background subtraction algorithms predominantly rely on pixel-level grayscale differences. Consequently, when dealing with low-contrast or slow-moving targets, rigid difference thresholds may lead to missed detections [18]. To improve candidate region extraction, we propose the MAFF strategy, which combines two complementary cues. Specifically, dense optical flow provides explicit displacement-based motion information by estimating inter-frame pixel motion, whereas grayscale differencing captures photometric appearance changes through local intensity variations between the current frame and the temporally aligned reference background. MAFF integrates these motion and appearance-change cues to enhance weak target responses while preserving the efficiency of the conventional grayscale difference framework.

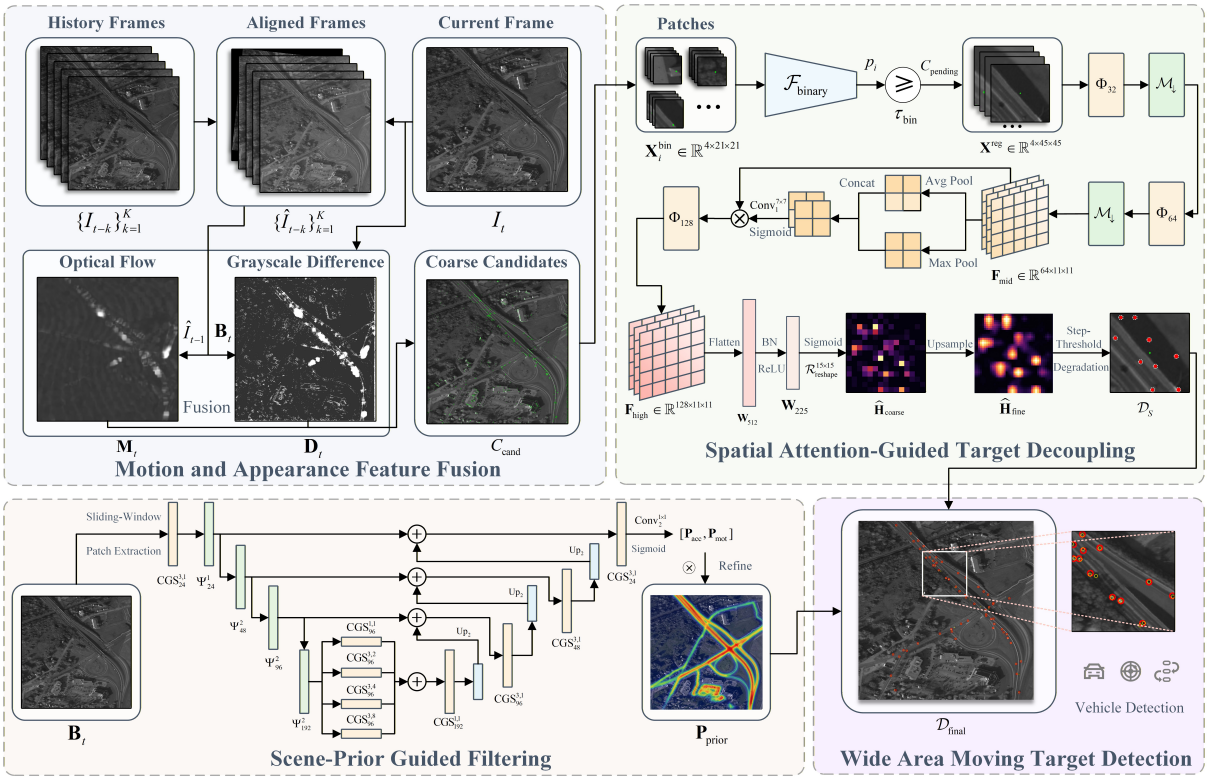


Figure 2. Overview of the proposed MDF-Det framework. The pipeline consists of three core components: MAFF: fuses dense optical-flow motion cues with grayscale differencing to enhance weak target responses and generate coarse candidates; SA-TD: combines spatial attention, fine-grained heatmap decoding, and step-threshold degradation to separate densely clustered targets; and SPGF: learns complementary vehicle-accessibility and motion-activity priors from scene context to suppress contextually implausible false alarms.

3.2.1. Spatial Registration and Photometric Compensation

Given a current frame I_t and a sequence of historical frames $\{I_{t-k}\}_{k=1}^{K}$, our initial objective is to eliminate the global background displacement induced by the ego-motion of the aerial platform. To achieve this, we first employ the ORB feature extractor [19] to compute image correspondences and estimate the global homography matrix $\mathbf{T}_k \in \mathbb{R}^{3\times3}$, which maps each historical frame to the current frame. Utilizing this matrix, we execute a perspective transformation [20] to project the historical frames into the unified spatial coordinate system of I_t. This spatial registration process yields the aligned template set, denoted as $\{\tilde{I}_{t-k}\}_{k=1}^{K}$:

$$\tilde{I}_{t-k}(\mathbf{x}) = I_{t-k}\left(\mathcal{W}(\mathbf{x}; \mathbf{T}_k^{-1})\right) \tag{1}$$

where $\mathbf{x} = [x, y]^T$ denotes the two-dimensional pixel coordinates in the current frame, and $\mathcal{W}(\cdot)$ represents the spatial warping and bilinear resampling operator parameterized by the inverse homography matrix. During the registration process, out-of-bounds regions and void areas lacking pixel values—induced by the perspective transformation—are systematically extracted and morphologically dilated to construct a global invalid mask, Ω_{invalid}, for subsequent processing.

Despite achieving precise spatial alignment, dynamic factors such as cloud occlusion and viewpoint variations frequently introduce significant photometric inconsistencies. Consequently, prior to constructing the reference background, we apply a large-scale Gaussian low-pass filter, denoted as $\mathcal{G}_{\text{illum}}$, to perform adaptive photometric compensation on the aligned templates [15]. This operation effectively isolates the low-frequency illumination drift components:

$$\hat{I}_{t-k} = \mathrm{Clip}\left(\tilde{I}_{t-k} + \mathcal{G}_{\mathrm{illum}}\left(I_t - \tilde{I}_{t-k}\right),\ 0, 255\right) \tag{2}$$

where $\mathrm{Clip}(\cdot, 0, 255)$ is a clipping function that bounds the pixel values within the standard grayscale interval of $[0, 255]$; and $\hat{I}_{t-k}$ represents the resulting photometrically aligned template.

Utilizing the photometrically aligned template set $\{\hat{I}_{t-k}\}_{k=1}^{K}$, we construct a local reference background, $\mathbf{B}_t$, via temporal median filtering $\mathrm{Median}(\cdot)$. Subsequently, we compute the initial absolute grayscale difference map $\mathbf{D}_t$:

$$\mathbf{B}_t(x, y) = \mathrm{Median}_{1 \le k \le K}\{\hat{I}_{t-k}(x, y)\}, \quad \mathbf{D}_t = \left|I_t - \mathbf{B}_t\right| \tag{3}$$

3.2.2. Enhanced Detection and Candidate Generation

To overcome the inherent limitations of traditional grayscale differencing—which frequently fails to detect low-contrast, slow-moving targets—we adopt a dense optical flow algorithm based on local polynomial expansion [21,22]. This approach explicitly captures subtle motion information, thereby enhancing weak motion responses during candidate generation. Specifically, we extract the pixel-level motion field $\mathbf{d} = (u, v)^T$ between the current frame I_t and the preceding frame $\hat{I}_{t-1}$ following spatial registration and photometric compensation. For a given spatial coordinate $\mathbf{x} = [x, y]^T$, the algorithm first approximates the image intensity signal within its local neighborhood as a quadratic polynomial surface:

$$I_t(\mathbf{x}) \approx \mathbf{x}^T \mathbf{A}_t \mathbf{x} + \mathbf{b}_t^T \mathbf{x} + c_t \tag{4}$$

where $\mathbf{A}_t$ denotes a symmetric matrix encoding the local spatial second-order derivatives; $\mathbf{b}_t$ is a column vector representing the first-order spatial gradients; and c_t is a scalar corresponding to the locally smoothed central grayscale intensity.

Assuming the target satisfies the brightness constancy constraint over an infinitesimal time interval—formally expressed as $I_t(\mathbf{x}) = \hat{I}_{t-1}(\mathbf{x} - \mathbf{d})$—equating the polynomial coefficients of adjacent frames yields the ideal theoretical constraint: $\mathbf{A}_t\mathbf{d} = -\frac{1}{2}(\mathbf{b}_t - \hat{\mathbf{b}}_{t-1})$. However, to ensure effective motion estimation in the presence of image noise, we formulate a weighted least squares optimization objective with respect to the displacement vector $\mathbf{d}$ over a local spatial neighborhood $\mathcal{Q}$:

$$\mathbf{d}^*(\mathbf{x}) = \arg\min_{\mathbf{d}} \sum_{\Delta\mathbf{x} \in \mathcal{Q}} w(\Delta\mathbf{x}) \left\| \mathbf{A}(\mathbf{x} + \Delta\mathbf{x})\mathbf{d} - \Delta\mathbf{b}(\mathbf{x} + \Delta\mathbf{x}) \right\|^2 \tag{5}$$

where $\mathbf{A}(\cdot) = \frac{1}{2}(\mathbf{A}_t(\cdot) + \hat{\mathbf{A}}_{t-1}(\cdot))$ denotes the arithmetic mean of the second-order coefficient matrices across the adjacent frames; $\Delta\mathbf{b} = -\frac{1}{2}(\mathbf{b}_t - \hat{\mathbf{b}}_{t-1})$ represents the scaled difference vector of the first-order gradients; $\Delta\mathbf{x}$ indicates the spatial coordinate offset within the local neighborhood $\mathcal{Q}$; and $w(\cdot)$ is a two-dimensional Gaussian weighting function that governs the neighborhood contributions.

Following the estimation of the optical flow field, we derive a normalized two-dimensional temporal motion saliency map, $\mathbf{M}_t \in [0, 1]^{W \times H}$, by computing the pixel-wise motion magnitude and subsequently applying a small-kernel Gaussian smoothing filter, $\mathcal{G}_{\mathrm{flow}}$:

$$\mathbf{M}_t = \mathcal{N}\left(\mathcal{G}_{\mathrm{flow}}\left(\sqrt{u^2 + v^2}\right)\right) \tag{6}$$

where W and H denote the spatial dimensions (width and height) of the input image, respectively; and $\mathcal{N}(\cdot)$ represents a linear normalization operator that scales the local motion magnitude into the interval $[0, 1]$.

Subsequently, we introduce an independent additive motion offset term to explicitly integrate the motion saliency as a physical prior, thereby yielding the enhanced difference map $\mathbf{D}'_t$:

$$\mathbf{D}'_t = \begin{cases} \mathbf{D}_t, & \text{if } (x, y) \in \Omega_{\text{invalid}} \\ \mathbf{D}_t + \alpha(\mathbf{M}_t - T_m), & \text{otherwise} \end{cases} \tag{7}$$

where $\mathbf{M}_t$ denotes the estimated motion magnitude; T_m represents a static noise threshold; and α is a scaling factor that modulates the motion offset. This additive formulation selectively amplifies the signal in dynamic regions ($\mathbf{M}_t > T_m$) while penalizing static background responses.

Subsequently, we apply a binarization threshold, τ_{diff}, to extract the potential target regions:

$$\mathbf{B}_{\text{cand}}(\mathbf{x}) = \mathbb{I}\big(\mathbf{D}'_t(\mathbf{x}) \geq \tau_{\text{diff}}\big) \tag{8}$$

where $\mathbf{B}_{\text{cand}}$ denotes the resulting binary candidate mask, wherein pixels with a value of 1 designate potential moving targets; $\mathbb{I}(\cdot)$ is the standard indicator function.

Following the application of morphological denoising to $\mathbf{B}_{\text{cand}}$, we partition the binary mask into 3×3 non-overlapping spatial cells to facilitate lightweight, grid-based non-maximum suppression and eliminate redundant, locally clustered responses [23]. As illustrated in Figure 3, by jointly executing local max pooling and center-anchored mapping, the extraction of the initial candidate center set, C_{cand}, is mathematically formulated as:

$$C_{\text{cand}} = \Big\{ (3x_d + 1, 3y_d + 1) \notin \Omega_{\text{invalid}} \;\Big|\; \max_{i,j \in \{0,1,2\}} \mathbf{B}_{\text{cand}}(3x_d + i, 3y_d + j) = 1 \Big\} \tag{9}$$

where (x_d, y_d) denotes the downsampled grid coordinates; the constraint $\notin \Omega_{\text{invalid}}$ ensures that the remapped candidate coordinates strictly avoid the predefined global invalid regions; and the constant offset of $+1$ serves to anchor these coordinates to the geometric center of each activated cell. Subsequently, this sparsified candidate set is forwarded for classification and target decoupling.

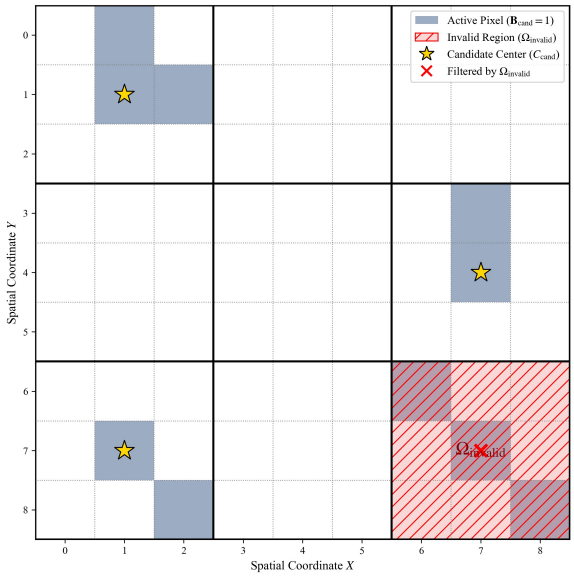


Figure 3. Illustration of the center-anchored candidate extraction. Locally aggregated active pixels (blue) in $\mathbf{B}_{\text{cand}}$ are downsampled into unified candidate centers (yellow stars) via grid-based max pooling, while redundant responses within the global invalid boundary Ω_{invalid} are explicitly suppressed.

The pseudocode for the proposed MAFF strategy is summarized in Algorithm 1.

Algorithm 1: Motion and Appearance Feature Fusion Strategy

Input : Current frame I_t and sequence of historical frames $\{I_{t-k}\}_{k=1}^{K}$; Hyperparameters $\alpha, T_m, \tau_{\text{diff}}$.

Output: Coarse candidate center set C_{cand}.

$\{\tilde{I}_{t-k}\}, \Omega_{\text{invalid}} \leftarrow \text{SpatialRegistration}(\{I_{t-k}\}_{k=1}^{K}, I_t)$

$\{\hat{I}_{t-k}\} \leftarrow \text{PhotometricCompensation}(\{\tilde{I}_{t-k}\}, I_t)$

$B_t \leftarrow \text{Median}_{1 \le k \le K}(\{\hat{I}_{t-k}\})$

$D_t \leftarrow |I_t - B_t|$

$(u, v) \leftarrow \text{DenseOpticalFlow}(I_t, \hat{I}_{t-1})$

$M_t \leftarrow \mathcal{N}(\mathcal{G}_{\text{flow}}(\sqrt{u^2 + v^2}))$

$D'_t \leftarrow \text{EnhanceDifference}(D_t, M_t, T_m, \alpha, \Omega_{\text{invalid}})$

$B_{\text{cand}} \leftarrow \mathbb{I}(D'_t \ge \tau_{\text{diff}})$

$C_{\text{cand}} \leftarrow \text{GridNMS}(B_{\text{cand}}, \Omega_{\text{invalid}})$

return C_{cand}

3.3. Spatial Attention-Guided Target Decoupling Module

Although the front-end perception stage captures moving pixels, the resulting candidate regions frequently exhibit severe spatial overlap among targets [24]. To resolve this challenge, we propose the SA-TD module. This module transitions the detection task from conventional coarse-grained bounding box regression to fine-grained heatmap decoding, thereby enabling the decoupling of densely clustered targets within the deep feature space.

3.3.1. Candidate Screening via Binary Classification

To eliminate false positives induced by parallax effects and dynamic background clutter, we initially employ a lightweight binary classifier to serve as a spatiotemporal filter [15]. Let $C_{\text{cand}} = \{\mathbf{c}_i\}_{i=1}^{N_c}$ denote the set of coarse candidate centers derived from the preceding background subtraction stage. For each candidate center $\mathbf{c}_i$, we crop localized patches from the current frame t and three consecutive historical templates (i.e., $t-1$, $t-2$, and $t-3$) to construct a spatiotemporal input tensor, $\mathbf{X}_i^{\text{bin}} \in \mathbb{R}^{4\times21\times21}$. Ultimately, the probability of genuine target presence, p_i, is predicted by the binary classification network $\mathcal{F}_{\text{binary}}$:

$$p_i = \mathcal{F}_{\text{binary}}(\mathbf{X}_i^{\text{bin}}; \theta_{\text{bin}}) \tag{10}$$

where θ_{bin} denotes the learnable parameters of the network. To suppress evident background clutter while preserving potential challenging targets, we establish a confidence threshold, τ_{bin}, to conduct a preliminary screening of the initial candidate set:

$$C_{\text{pending}} = \{\mathbf{c}_i \in C_{\text{cand}} \mid p_i \ge \tau_{\text{bin}}\} \tag{11}$$

Following this lightweight filtering stage, the retained candidate set, C_{pending}, is forwarded to the subsequent multi-target heatmap regression network.

3.3.2. Spatial Attention-Guided Heatmap Generation

To establish a unified mathematical formulation for the network architecture, we initially define a generalized dual-convolutional base operator, denoted as $\Phi_m(\cdot)$. For any given input feature tensor $\mathbf{F}$, its strictly cascaded expansion is mathematically defined as:

$$\Phi_m(\mathbf{F}) = \mathrm{ReLU}\Big(\mathcal{BN}\big(\mathrm{Conv}_m^{3\times3}\big(\mathrm{ReLU}(\mathcal{BN}(\mathrm{Conv}_m^{3\times3}(\mathbf{F})))\big)\big)\Big) \tag{12}$$

where m denotes the output channel dimension of the module; $\mathrm{Conv}_m^{3\times3}$ represents a 3×3 convolutional operator for feature mapping that yields m output channels; $\mathcal{BN}$ signifies the batch normalization operation; and ReLU corresponds to the Rectified Linear Unit non-linear activation function.

With the base operator established, for each candidate point $\mathbf{c}_i \in C_{\text{pending}}$, we extract a localized spatiotemporal patch of 45×45 pixels to formulate the regression input tensor, $\mathbf{X}^{\text{reg}} \in \mathbb{R}^{4\times45\times45}$. This tensor is subsequently processed through two cascaded feature extraction stages—each equipped with a 2×2 spatial max pooling operation ($\mathcal{M}_\downarrow$)—and is ultimately projected to yield the mid-level feature map $\mathbf{F}_{\text{mid}} \in \mathbb{R}^{64\times11\times11}$:

$$\mathbf{F}_{\text{mid}} = \mathcal{M}_\downarrow\Big(\Phi_{64}\big(\mathcal{M}_\downarrow(\Phi_{32}(\mathbf{X}^{\text{reg}}))\big)\Big) \tag{13}$$

To accentuate the subtle motion signatures of tiny targets amidst complex background clutter, we integrate a spatial attention module within the mid-level feature space [25]. Specifically, we initially derive two-dimensional spatial descriptors by applying average pooling ($\mathcal{P}_{\text{avg}}$) and max pooling ($\mathcal{P}_{\max}$) operations across the channel axis:

$$\mathbf{F}_{\text{avg}}^s = \mathcal{P}_{\text{avg}}(\mathbf{F}_{\text{mid}}), \quad \mathbf{F}_{\max}^s = \mathcal{P}_{\max}(\mathbf{F}_{\text{mid}}) \tag{14}$$

where $\mathbf{F}_{\text{avg}}^s \in \mathbb{R}^{1\times11\times11}$ denotes the average-pooled feature descriptor, which aggregates global statistical information; and $\mathbf{F}_{\max}^s \in \mathbb{R}^{1\times11\times11}$ represents the max-pooled feature descriptor, which captures the most salient local motion responses. By collapsing the channel dimension, these paired descriptors effectively steer the network's attention toward spatial regions exhibiting prominent motion signatures.

These descriptors are concatenated along the channel axis and subsequently processed through a 7×7 convolutional layer, $\mathrm{Conv}_1^{7\times7}$, coupled with a Sigmoid activation function to derive the spatial attention weight map $\mathbf{A}_s \in \mathbb{R}^{1\times11\times11}$:

$$\mathbf{A}_s = \mathrm{Sigmoid}\big(\mathrm{Conv}_1^{7\times7}([\mathbf{F}_{\text{avg}}^s; \mathbf{F}_{\max}^s])\big) \tag{15}$$

where $\mathrm{Conv}_1^{7\times7}$ denotes a spatial convolutional operator featuring a 7×7 kernel and yielding a single output channel; and $[\cdot;\cdot]$ represents the channel-wise feature concatenation operation.

Subsequently, the attention mask $\mathbf{A}_s$ is element-wise multiplied with the mid-level feature map $\mathbf{F}_{\text{mid}}$, and the resulting tensor is directly propagated to the high-order feature extraction stage. By employing the previously defined generalized operator $\Phi_m(\cdot)$, these saliency-enhanced features are further projected into high-order semantic representations, yielding $\mathbf{F}_{\text{high}} \in \mathbb{R}^{128\times11\times11}$:

$$\mathbf{F}_{\text{high}} = \Phi_{128}\big(\mathbf{F}_{\text{mid}} \otimes \mathbf{A}_s\big) \tag{16}$$

where $\otimes$ denotes the spatial element-wise multiplication operation.

In contrast to conventional networks that regress a single coordinate offset, the SA-TD module decodes the high-order semantic representations, $\mathbf{F}_{\text{high}}$, into a spatial probability distribution to effectively decouple multiple densely clustered targets within the localized patch. Specifically, the flattened high-order feature tensor is mapped to a probability vector

via a Multi-Layer Perceptron and subsequently reshaped to form a coarse local heatmap, $\hat{\mathbf{H}}_{\text{coarse}} \in [0,1]^{15\times 15}$:

$$\hat{\mathbf{H}}_{\text{coarse}} = \mathcal{R}^{15\times 15}_{\text{reshape}}\Big(\text{Sigmoid}\Big(\mathbf{W}_{225} \text{ ReLU}\big(\mathcal{BN}(\mathbf{W}_{512}\,\text{Flatten}(\mathbf{F}_{\text{high}}))\big)\Big)\Big) \tag{17}$$

where $\text{Flatten}(\cdot)$ denotes the operation that flattens a three-dimensional feature tensor into a one-dimensional vector; $\mathbf{W}_{512}$ and $\mathbf{W}_{225}$ represent the weight matrices of the fully connected layers, projecting the features into 512 and 225 dimensions, respectively; and $\mathcal{R}^{15\times 15}_{\text{reshape}}(\cdot)$ signifies the spatial reshaping operator that rearranges the resulting 225-dimensional probability vector into a 15×15 two-dimensional grid.

3.3.3. Optimization Objective and Step-Degradation Inference

During the training phase, given a localized patch encompassing N_p genuine moving targets, let $\{\boldsymbol{\mu}_j\}_{j=1}^{N_p}$ denote their corresponding relative center coordinates within the local patch coordinate frame. The ground-truth heatmap, $\mathbf{H}$, for this specific patch is subsequently synthesized by rendering two-dimensional Gaussian kernels:

$$\mathbf{H}(\mathbf{x}) = \max_{1\le j\le N_p} \exp\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}_j\|_2^2}{2\sigma_p^2}\right) \tag{18}$$

where $\mathbf{x} = [x, y]^T$ denotes the two-dimensional spatial coordinate vector within the localized patch; $\boldsymbol{\mu}_j = [\mu_{x_j}, \mu_{y_j}]^T$ represents the relative center coordinates of the j-th target; and σ_p signifies the standard deviation parameter governing the spatial extent of the Gaussian kernel.

To model the spatial distribution of targets, the network is optimized via the Mean Squared Error loss function:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N_r}\sum_{\mathbf{x}} \left\|\hat{\mathbf{H}}_{\text{coarse}}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\right\|_2^2 \tag{19}$$

where N_r denotes the total number of spatial pixels comprising the coarse heatmap.

During the inference phase, the predicted coarse heatmap is bilinearly upsampled to the original localized patch resolution of 45×45, yielding the refined heatmap $\hat{\mathbf{H}}_{\text{fine}}$. To decouple and extract tiny targets exhibiting severe spatial overlap within densely clustered scenes—and to effectively mitigate the missed detection issue induced by the mutual suppression of adjacent genuine targets—we introduce a step-threshold degradation algorithm founded on topological connected components.

Specifically, we define a monotonically decreasing discrete threshold sequence, $\mathcal{T} = \{\tau_0, \tau_1, \ldots, \tau_S\}$, where S represents the total number of degradation steps. The initial threshold is initialized as $\tau_0 = \max(\hat{\mathbf{H}}_{\text{fine}})$ and is iteratively decayed by a constant step size of $\Delta\tau$ until reaching a predefined lower bound, $\tau_{\min}$.

During the s-th iteration step ($s \in \{0, 1, \ldots, S\}$) associated with the specific threshold τ_s, we extract the set of independent connected components, denoted as $\mathcal{O}^{(s)} = \{o_j^{(s)}\}$, wherein the refined heatmap satisfies the condition $\hat{\mathbf{H}}_{\text{fine}} \ge \tau_s$. The spatial coordinates of the target candidates are subsequently determined by identifying the local maxima within each respective connected component. This yields the candidate set for the current iteration, $\mathcal{V}_s$, formulated as:

$$\mathcal{V}_s = \left\{ \arg\max_{\mathbf{x} \in o_j^{(s)}} \hat{\mathbf{H}}_{\text{fine}}(\mathbf{x}) \;\middle|\; o_j^{(s)} \in \mathcal{O}^{(s)} \right\} \tag{20}$$

To preclude the redundant extraction of identical salient targets across varying threshold levels, we introduce an inter-step exclusion mechanism designed to dynamically maintain a global set of valid coordinates, $\mathcal{D}$. Initializing with $\mathcal{D}_0 = \mathcal{V}_0$, the subsequent updates to this global set strictly adhere to a set difference operational rule:

$$\mathcal{D}_s = \mathcal{D}_{s-1} \cup \left(\mathcal{V}_s \setminus \mathcal{D}_{s-1} \right) \tag{21}$$

This exclusion operation $(\setminus)$ prevents previously detected coordinates from being repeatedly added across different threshold levels. Upon reaching the terminal threshold, τ_S, the resulting set, $\mathcal{D}_S$, provides the spatial coordinates of the separated densely clustered targets within the localized patch.

To illustrate this procedure, Figure 4 shows the progressive target decoupling process. Starting from the refined heatmap $\hat{\mathbf{H}}_{\text{fine}}$ (Figure 4a), a high initial threshold ($\tau_1 = 0.97$, Figure 4b) extracts the most salient targets as independent connected components. As τ gradually decreases ($\tau_7 = 0.90$, Figure 4c), additional weaker target responses progressively emerge. At the terminal threshold ($\tau_{\text{min}} = 0.17$, Figure 4d), further target responses are separated, while the inter-step exclusion mechanism reduces redundant detections across different threshold levels.

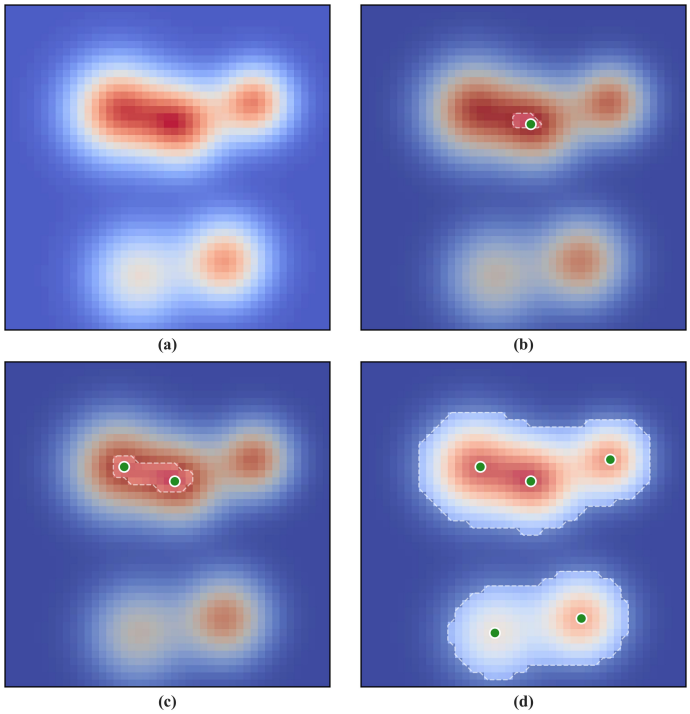


Figure 4. Step-threshold target decoupling process. (**a**) Original heatmap ($\hat{\mathbf{H}}_{\text{fine}}$). (**b**–**d**) Progressive extraction at decreasing thresholds τ_s (dashed outlines: connected domains; green dots: target centers). (**b**) $\tau_1 = 0.97$: most salient target extraction. (**c**) $\tau_7 = 0.90$: domain expansion for additional targets. (**d**) $\tau_{\text{min}} = 0.17$: final target separation via cross-layer exclusion.

The pseudocode for the proposed SA-TD module is summarized in Algorithm 2.

Algorithm 2: Spatial Attention-Guided Target Decoupling Module

Input : Coarse candidate center set C_{cand}; Current frame I_t and historical templates; Hyperparameters $\tau_{\text{bin}}, \tau_{\min}, \Delta\tau$.

Output: Decoupled target coordinate set $\mathcal{D}_S$.

$C_{\text{pending}} \leftarrow \{c_i \in C_{\text{cand}} \mid \mathcal{F}_{\text{binary}}(X_i^{\text{bin}}) \geq \tau_{\text{bin}}\}$
$\mathcal{D}_S \leftarrow \emptyset$
for ***each*** $c_i \in C_{pending}$ **do**
  $X^{\text{reg}} \leftarrow \text{ExtractPatch}(I_t, \{I_{t-k}\}, c_i)$
  $F_{\text{mid}} \leftarrow \mathcal{M}_{\downarrow}(\Phi_{64}(\mathcal{M}_{\downarrow}(\Phi_{32}(X^{\text{reg}}))))$
  $A_s \leftarrow \text{SpatialAttention}(F_{\text{mid}})$
  $F_{\text{high}} \leftarrow \Phi_{128}(F_{\text{mid}} \otimes A_s)$
  $\hat{H}_{\text{fine}} \leftarrow \text{DecodeAndUpsample}(F_{\text{high}})$
  $\tau_s \leftarrow \max(\hat{H}_{\text{fine}}); \quad \mathcal{D}_{\text{patch}} \leftarrow \emptyset$
  while $\tau_s \geq \tau_{min}$ **do**
    $\mathcal{O}^{(s)} \leftarrow \text{ConnectedComponents}(\hat{H}_{\text{fine}} \geq \tau_s)$
    $\mathcal{V}_s \leftarrow \{\arg\max_{x \in o_j^{(s)}} \hat{H}_{\text{fine}}(x) \mid o_j^{(s)} \in \mathcal{O}^{(s)}\}$
    $\mathcal{D}_{\text{patch}} \leftarrow \mathcal{D}_{\text{patch}} \cup (\mathcal{V}_s \setminus \mathcal{D}_{\text{patch}})$
    $\tau_s \leftarrow \tau_s - \Delta\tau$
  end
  $\mathcal{D}_S \leftarrow \mathcal{D}_S \cup \mathcal{D}_{\text{patch}}$
end
return $\mathcal{D}_S$

3.4. Scene-Prior Guided Filtering Mechanism

Although SA-TD effectively separates densely clustered target responses, candidate-level features alone remain insufficient to suppress false alarms caused by building parallax, registration residuals, vegetation motion, and other structured background interference [26]. We therefore introduce the SPGF mechanism to provide a soft scene-level constraint for candidate verification. SPGF estimates vehicle-compatible scene responses directly from visual context and generates a continuous probability prior for suppressing contextually implausible detections. Unlike rigid binary masks constructed from external geographic data such as OpenStreetMap (OSM) [27], the learned scene prior does not require precise map-to-image registration and provides soft contextual constraints that reduce the risk of discarding valid targets near uncertain or transitional regions.

3.4.1. Formulation of the Network Topology

SPGF reuses the temporally aggregated reference background $\mathbf{B}_t$ constructed in MAFF as its scene-context input. Since $\mathbf{B}_t$ is obtained from spatially and photometrically aligned historical frames through temporal median filtering, transient foreground responses are attenuated while relatively stable scene structures are retained. The network adopts a fully convolutional residual encoder–decoder architecture to extract contextual features relevant to vehicle accessibility and motion activity. The encoder progressively captures multi-scale representations, while an atrous spatial pyramid enlarges the effective receptive field at the bottleneck [28]. The decoder then progressively restores spatial resolution through bilinear upsampling and skip connections.

For compact notation, let $\text{CGS}_n^{k,r}(\cdot)$ denote a $k \times k$ convolution with n output channels and dilation rate r, followed by Group Normalization (GN) [29] and Swish activation [30]. Moreover, $\Psi_n^{\delta}(\cdot)$ denotes a residual block with n output channels and stride δ, where a 1×1

projection is applied to the shortcut when the feature dimensions change. The number of output channels at the l-th encoder stage is defined as $n_l = 24 \times 2^{l-1}$, $l = 1, \ldots, 4$, with $\delta_1 = 1$ and $\delta_l = 2$ for $l = 2, \ldots, 4$.

The complete network is formulated recursively as

$$\begin{cases} \mathbf{E}_0 = \mathrm{CGS}^{3,1}_{n_1}(\mathbf{B}_t), \\ \mathbf{E}_l = \Psi^{\delta_l}_{n_l}(\mathbf{E}_{l-1}), \qquad l = 1, \ldots, 4, \\ \mathbf{Z} = \mathrm{CGS}^{1,1}_{n_4}\Big([\mathrm{CGS}^{1,1}_{n_3}(\mathbf{E}_4), \mathrm{CGS}^{3,2}_{n_3}(\mathbf{E}_4), \mathrm{CGS}^{3,4}_{n_3}(\mathbf{E}_4), \mathrm{CGS}^{3,8}_{n_3}(\mathbf{E}_4)]\Big), \\ \mathbf{U}_4 = \mathbf{Z}, \\ \mathbf{U}_l = \mathrm{CGS}^{3,1}_{n_l}\Big([\mathrm{Up}_2(\mathbf{U}_{l+1}), \mathbf{E}_l]\Big), \qquad l = 3, 2, 1, \\ [\mathbf{P}_{\mathrm{acc}}, \mathbf{P}_{\mathrm{mot}}] = \mathrm{Sigmoid}\Big(\mathrm{Conv}^{1\times 1}_2(\mathbf{U}_1)\Big). \end{cases} \tag{22}$$

where $\mathbf{E}_l$ and $\mathbf{U}_l$ denote the encoder and decoder features at the l-th scale, respectively, $\mathbf{Z}$ denotes the bottleneck feature obtained through multi-scale atrous aggregation, and $\mathrm{Up}_2(\cdot)$ denotes bilinear upsampling by a factor of two. At the bottleneck, parallel convolutions with dilation rates $\{2, 4, 8\}$ aggregate broader scene context while preserving local structural information. The decoder progressively fuses the bottleneck representation with encoder features and produces two complementary probability maps: $\mathbf{P}_{\mathrm{acc}}$ characterizes vehicle-accessibility compatibility, whereas $\mathbf{P}_{\mathrm{mot}}$ characterizes vehicle-motion compatibility.

3.4.2. Network Optimization Objective

The supervision maps are constructed from temporally aggregated vehicle trajectories rather than from a coarse binary scene mask. For each training anchor, trajectory observations from the preceding 20 frames are aligned to the current reference frame. The accessibility branch uses all valid trajectory points, whereas the motion branch uses only trajectory points associated with displacements of at least 0.8 m. Each trajectory point is represented by a localized Gaussian response with a radius of 10 pixels and a standard deviation of 5 pixels at the native image resolution, yielding two soft supervision maps for vehicle accessibility and motion activity, respectively.

The two prediction branches are jointly optimized using spatially weighted variants of focal loss [31] and Dice loss [32]. Specifically, $\mathcal{L}^{\mathrm{acc}}_{\mathrm{focal}}$ and $\mathcal{L}^{\mathrm{mot}}_{\mathrm{focal}}$ emphasize hard-to-classify pixels and reduce the dominance of easy background samples in the vehicle-accessibility and vehicle-motion branches, respectively. Meanwhile, $\mathcal{L}^{\mathrm{acc}}_{\mathrm{dice}}$ and $\mathcal{L}^{\mathrm{mot}}_{\mathrm{dice}}$ encourage spatial consistency between the predicted and ground-truth prior maps. The overall objective is formulated as:

$$\mathcal{L}_{\mathrm{SPGF}} = (\mathcal{L}^{\mathrm{acc}}_{\mathrm{focal}} + 0.5\mathcal{L}^{\mathrm{acc}}_{\mathrm{dice}}) + 0.7(\mathcal{L}^{\mathrm{mot}}_{\mathrm{focal}} + 0.5\mathcal{L}^{\mathrm{mot}}_{\mathrm{dice}}). \tag{23}$$

The coefficient 0.5 balances the region-level Dice constraint with the pixel-wise focal objective, while the coefficient 0.7 balances the contribution of motion-activity supervision relative to vehicle-accessibility supervision.

3.4.3. Prior Refinement and Filtering

During inference, the two complementary predictions are first fused through their geometric mean:

$$\mathbf{P}_{\mathrm{scene}} = \sqrt{\mathbf{P}_{\mathrm{acc}} \odot \mathbf{P}_{\mathrm{mot}}}, \tag{24}$$

where $\odot$ denotes element-wise multiplication. This fusion assigns high prior confidence to regions jointly supported by vehicle-accessibility and vehicle-motion cues.

Although $\mathbf{P}_{\text{scene}}$ provides a continuous scene-level constraint, isolated responses and short discontinuities may remain in the predicted map. We therefore apply a lightweight topology-preserving refinement $\mathcal{R}_{\text{topo}}(\cdot)$, consisting of Gaussian smoothing, weak- and strong-confidence support extraction, multi-directional morphological closing, connected-component filtering according to the presence of strong-confidence support, area, and elongation, and soft attenuation of unreliable responses. The overall refinement process is compactly formulated as:

$$\mathbf{P}_{\text{prior}} = \mathcal{R}_{\text{topo}}(\mathbf{P}_{\text{scene}}). \tag{25}$$

This refinement bridges short discontinuities and suppresses isolated or topologically inconsistent responses. Unlike binary scene masks, the refinement retains continuous confidence values, allowing uncertain contextual evidence near structural boundaries to be preserved rather than converted into hard binary decisions.

The resulting prior $\mathbf{P}_{\text{prior}}$ is finally applied to the target set $\mathcal{D}_S$ produced by SA-TD:

$$\mathcal{D}_{\text{final}} = \left\{ \mathbf{p} \in \mathcal{D}_S \;\middle|\; \mathbf{P}_{\text{prior}}(\mathbf{p}) \geq \tau_{\text{prior}} \right\}. \tag{26}$$

where τ_{prior} denotes the scene-prior confidence threshold.

Thus, SPGF acts as a soft contextual filter rather than an independent target detector, suppressing detections inconsistent with the inferred scene context while preserving the fine-grained localization produced by SA-TD.

The pseudocode of the proposed SPGF mechanism is summarized in Algorithm 3.

Algorithm 3: Scene-Prior Guided Filtering Mechanism

Input : Target set $\mathcal{D}_S$ produced by SA-TD; temporally aggregated reference background $\mathbf{B}_t$ constructed in MAFF; scene-prior confidence threshold τ_{prior}.

Output: Final filtered target set $\mathcal{D}_{\text{final}}$.

Obtain $\mathbf{P}_{\text{acc}}$ and $\mathbf{P}_{\text{mot}}$ from $\mathbf{B}_t$ according to Equation (22)
$\mathbf{P}_{\text{scene}} \leftarrow \sqrt{\mathbf{P}_{\text{acc}} \odot \mathbf{P}_{\text{mot}}}$
$\mathbf{P}_{\text{prior}} \leftarrow \mathcal{R}_{\text{topo}}(\mathbf{P}_{\text{scene}})$
$\mathcal{D}_{\text{final}} \leftarrow \emptyset$
for ***each*** $\mathbf{p} \in \mathcal{D}_S$ **do**
    if $\mathbf{P}_{\text{prior}}(\mathbf{p}) \geq \tau_{\text{prior}}$ **then**
        $\mathcal{D}_{\text{final}} \leftarrow \mathcal{D}_{\text{final}} \cup \{\mathbf{p}\}$
    end
end
return $\mathcal{D}_{\text{final}}$

4. Experimental Setup

4.1. Datasets and Evaluation Metrics

We evaluate MDF-Det on the WPAFB 2009 dataset [8], a standard WAMI benchmark acquired at 1.25 Hz using a camera system composed of six optical sensors. The individual sensor images are spatially stitched into gigapixel-scale mosaics covering an area of approximately 35 km^2. The dataset contains 1025 frames and is divided into a 512-frame training sequence and a 513-frame test sequence. The binary classification and heatmap regression networks in SA-TD, together with the scene-prior network in SPGF, are trained exclusively using samples constructed from the training sequence, while the test sequence is reserved solely for final performance evaluation.

The dataset presents substantial challenges for moving-target detection, including frame resolutions exceeding 315 megapixels, dense traffic with approximately 2000 vehicles per frame, and extremely small targets with an average size of approximately 9×18 pixels. Following the commonly adopted evaluation protocol [11], quantitative evaluation is conducted on six representative AOIs, namely AOIs 01, 02, 03, 34, 40, and 41.

For qualitative cross-scene evaluation of SPGF, representative scenes from the Greene 2007 dataset [33] are additionally used. Greene 2007 is used exclusively for zero-shot inference and is not involved in training, validation, model selection, or hyperparameter tuning.

Detection performance is quantitatively assessed using Recall, Precision, and the F_1 score [11]. A predicted target location is considered a true positive if it is within 20 pixels of the corresponding ground-truth location. Computational efficiency is further evaluated in terms of the number of trainable parameters, model size, FLOPs per frame, peak GPU memory, and inference time in seconds per frame. Runtime contribution is additionally reported to characterize the relative computational cost of each module.

4.2. Baseline Methods and Implementation Details

To evaluate the proposed approach, MDF-Det is compared with five representative baseline methods, including ClusterNet [11], the CNN-based method of Zhou et al. [15], the heatmap-regression-based HMRN [17] and HM-Net [6], and CATLoss [34]. Since official implementations directly applicable to our experimental setting are unavailable for these methods, all baselines are independently reimplemented according to the architectures, training procedures, and parameter settings described in their original publications. The principal implementation and training settings of the reproduced baselines are summarized in Table 1. ClusterNet, HMRN, HM-Net, and CATLoss are trained from random initialization, whereas CNNs are initialized from the publicly released weights of the reference implementation and subsequently retrained on the adopted training partition. To ensure a fair comparison, all reproduced baselines and MDF-Det use the same training, validation, and test data partitions, as well as the same evaluation protocol. Model selection and hyperparameter tuning are performed exclusively on the validation set, after which the selected configurations are fixed and evaluated on the test set without further adjustment. Therefore, all baseline results reported in this study are obtained from our reproduced implementations rather than directly adopted from the original publications.

The MDF-Det pipeline is implemented using PyTorch and TensorFlow/Keras. All experiments are conducted on a workstation equipped with an Intel Core i9-14900K CPU and an NVIDIA RTX 3090 GPU. All methods are evaluated on the same six AOIs under an identical detection matching criterion and evaluation procedure.

The hyperparameters of MDF-Det are set as follows. For MAFF, we use $K = 5$, $T_m = 0.1$, $\alpha = 15$, and $\tau_{\text{diff}} = 1$. For SA-TD, we set $\tau_{\text{bin}} = 0.7$, $\Delta\tau = 0.01$, and $\tau_{\text{min}} = 0.17$. For SPGF, the scene-prior confidence threshold is set to $\tau_{\text{prior}} = 0.2$. These hyperparameters are selected using the temporally separated validation set and are subsequently fixed for all test-set evaluations unless otherwise specified.

Table 1. Principal implementation and training settings used for the reproduced baseline methods.

Method	Input Configuration	Training Configuration				Training Objective
		Optimizer	Learning Rate	Batch Size	Training	
ClusterNet [11]	Five consecutive frames; $128 \times 128 \times 5$ patches	AdamW	3×10^{-4}	256	12 epochs	Focal binary cross-entropy; Dice loss
CNNs [15]	$21 \times 21 \times 4$ classification patches; $45 \times 45 \times 4$ regression patches	SGD	1×10^{-2} (cls.); 5×10^{-3} (reg.)	100	80 epochs (cls.); 100 epochs (reg.)	Cross-entropy for classification; MSE for heatmap regression
HMRN [17]	Current frame, previous frame; previous center heatmap	Adam	1.25×10^{-4}	16	75 epochs	CenterNet-style focal loss; center-masked displacement L1 loss
HM-Net [6]	Current frame, previous frame; previous heatmap	Adam	3×10^{-4}	6	40 epochs	Center focal loss; center-masked motion and precision L1 losses
CATLoss [34]	$21 \times 21 \times 4$ objectness patches; $45 \times 45 \times 4$ localization patches	AdamW	3×10^{-4}	128 (obj.); 64 (loc.)	20 epochs	Binary focal loss for objectness; CATLoss for localization

4.3. Network Training Details

The binary classification network used for candidate screening in SA-TD follows the network architecture of Zhou et al. [15], but all model parameters are retrained in this study rather than using the pretrained weights from the original method. The binary classification, heatmap regression, and SPGF networks are all trained and validated using temporally separated and non-overlapping intervals constructed exclusively from the 512-frame training sequence: frames 100–379 are used for training, whereas frames 400–499 are reserved for validation. No random frame-level splitting is performed, thereby reducing temporal correlation between adjacent WAMI frames across the training and validation subsets. The 513-frame test sequence is reserved exclusively for final evaluation and is not involved in model training, validation, model selection, or hyperparameter tuning.

For the binary classifier, 3,600,000 training samples and 36,000 validation samples are constructed as four-frame spatiotemporal patches of size $4 \times 21 \times 21$ from candidate locations in the corresponding training and validation intervals. For the heatmap regression network in SA-TD, 2,824,620 training samples and 696,090 validation samples are constructed as four-frame spatiotemporal patches of size $4 \times 45 \times 45$, with each target label represented by a 15×15 heatmap. A mean spatiotemporal patch, estimated from at most 50,000 training samples, is subtracted from the inputs for normalization. The heatmap regression network is trained for 100 epochs using Adam with an initial learning rate of 1×10^{-4}, a batch size of 256, and the mean squared error loss.

For SPGF, 5133 training samples and 1766 validation samples are independently constructed from the same temporally separated intervals using a sliding-window strategy on the original-resolution temporal reference background $\mathbf{B}_t$. Specifically, 384×384 patches and their corresponding trajectory-supervised probability maps are extracted without resizing the full image to a fixed low resolution. For each anchor frame, up to 96 patches are sampled using a 40:30:30 ratio of trajectory-centered positive samples, high-texture hard negatives, and random negatives, respectively. Negative samples are maintained at least 48 pixels from annotated trajectory points, while weak negative regions are assigned a reduced loss weight of 0.03 to account for the incompleteness of trajectory-based supervision.

To prevent spatial information leakage, all trajectory annotations and image patches overlapping the six evaluation AOIs are excluded together with an additional 64-pixel safety margin. The resulting spatial partition of the training, validation, and evaluation regions is illustrated in Figure 5. During inference, the same 384×384 sliding-window strategy is applied to the full-resolution reference background, and predictions from overlapping windows are merged to reconstruct the scene-prior probability maps. For each fixed AOI, the full-resolution scene-prior map is constructed once before online frame-by-frame detection and is subsequently reused for candidate filtering. The SPGF network is trained for a maximum of 80 epochs using Adam with an initial learning rate of 2×10^{-4} and a batch size of 4, using the combined focal and Dice losses described above.

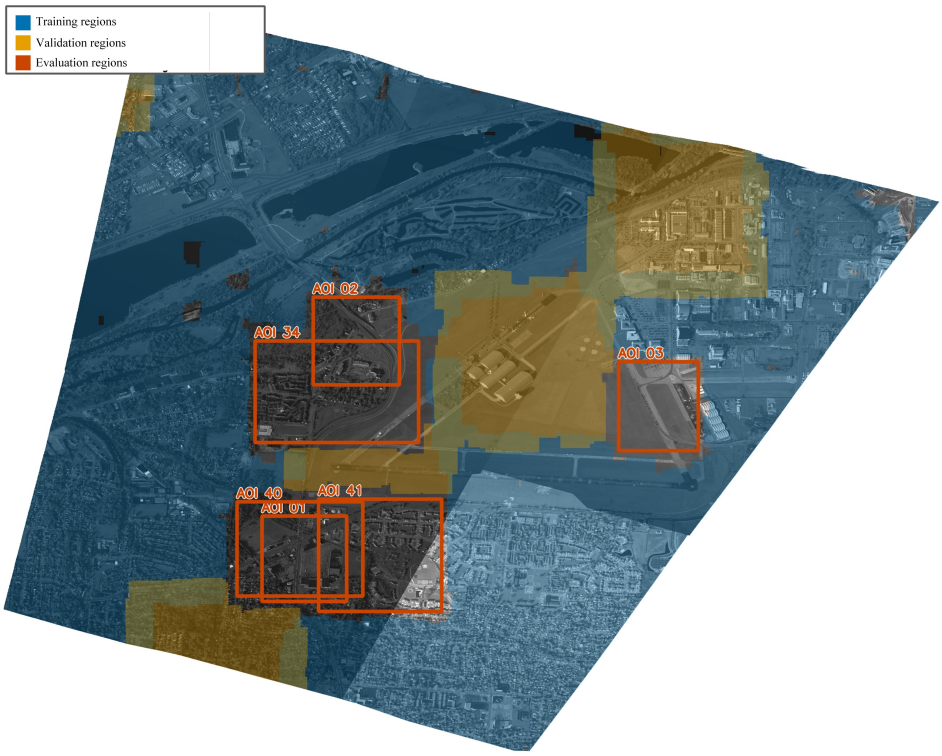


Figure 5. Spatial distribution of SPGF training, validation, and evaluation regions.

5. Results

5.1. Illustrative Operation

To qualitatively illustrate the operation of MDF-Det, Figure 6 presents representative intermediate results and final detections across four WAMI scenarios. Figure 6a shows the overall detection result on a full-scale scene, demonstrating that MDF-Det can maintain reliable target responses under an extremely low target-to-background ratio. The subsequent examples further illustrate the roles of the three cascaded modules: (1) **MAFF** enhances weak target responses by jointly exploiting motion and appearance cues, particularly for targets affected by shadows or weak motion contrast (Figure 6b); (2) **SA-TD** separates merged responses in densely distributed target regions, improving the discrimination of closely spaced targets (Figure 6c); and (3) **SPGF** suppresses false responses caused by registration residuals and structured background interference using the inferred scene-prior probability map (Figure 6d). The close correspondence between the predicted target locations (red circles) and the ground-truth annotations (yellow dots) further illustrates the ability of the proposed framework to localize moving targets while suppressing background-induced false alarms.

To evaluate the cross-scene generalization of SPGF, Figure 7 presents scene-prior probability maps on the six WPAFB 2009 evaluation AOIs and representative scenes from the Greene 2007 dataset. On WPAFB 2009, SPGF produces higher responses on vehicle-compatible structures such as roads, intersections, and parking areas, while assigning lower probabilities to most non-traffic regions. The model trained on WPAFB 2009 is further transferred directly to Greene 2007 without retraining, fine-tuning, or hyperparameter adjustment, where similar scene-aware response patterns are preserved across diverse environments. These results provide qualitative evidence that SPGF captures transferable scene-level compatibility rather than relying solely on scene-specific target distributions. Moreover, the inferred priors remain spatially selective: across the six WPAFB AOIs, the fraction of pixels with prior probability above 0.15 ranges from 23.40% to 48.41%, while only 12.02–24.90% of pixels exceed a probability of 0.50.

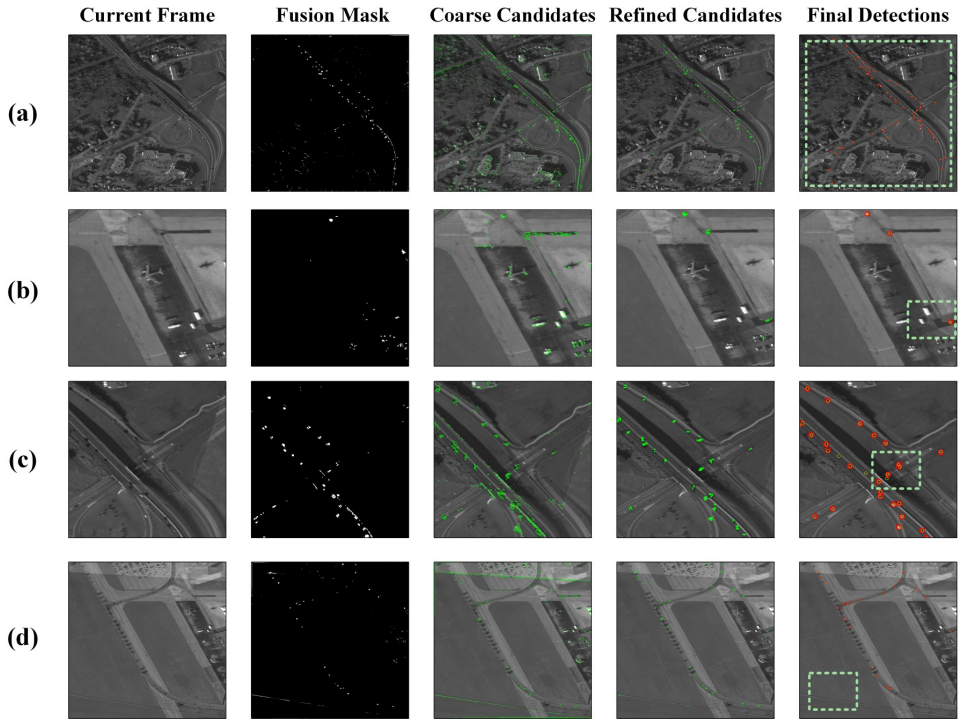


Figure 6. Qualitative visualization of the MDF-Det pipeline in four WAMI scenes: (**a**) curved intersection; (**b**) parking area; (**c**) multi-lane road; (**d**) sparse road. The columns show the current frame, fusion mask, coarse candidates, refined candidates, and final detections, respectively. Ground-truth locations and predictions are indicated by yellow dots and red circles, while green dashed boxes highlight challenging regions.

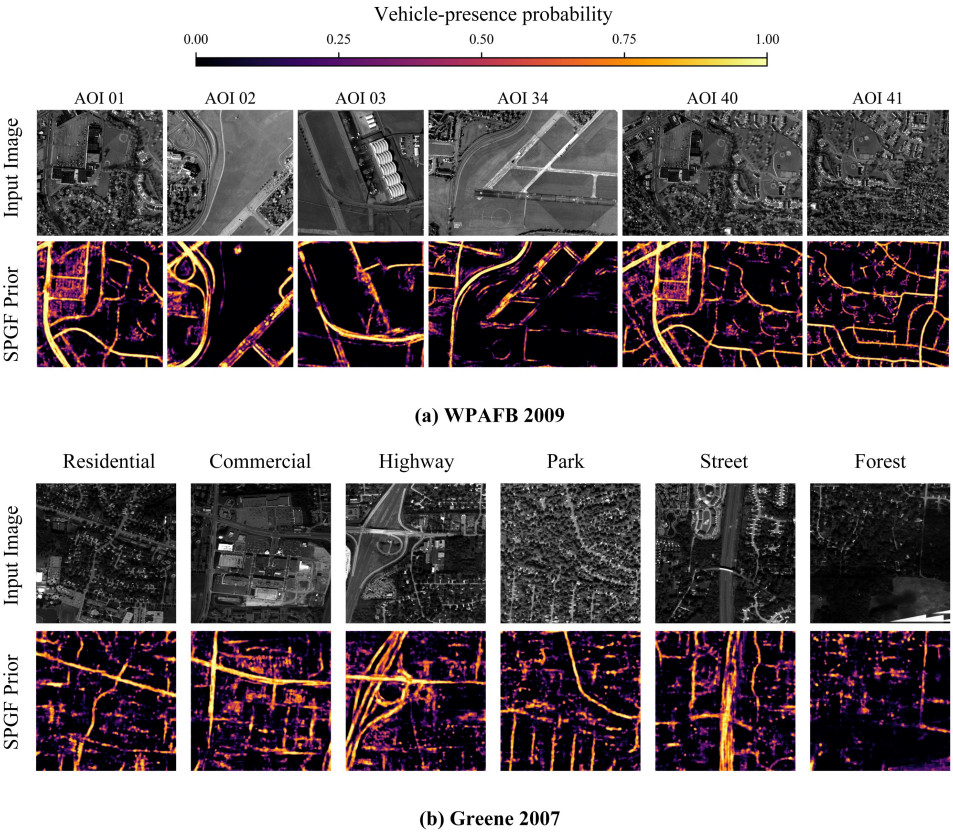


Figure 7. SPGF scene-prior probability maps on (**a**) WPAFB 2009 and (**b**) Greene 2007 under zero-shot transfer.

5.2. Comparison with Baseline Methods

Table 2 compares MDF-Det with five representative baseline methods across the six evaluation AOIs. MDF-Det achieves the highest F_1 score on all six AOIs, with an average performance of 0.878 ± 0.043, compared with 0.832 ± 0.071 for the strongest baseline, HM-Net. The corresponding mean F_1 improvement is 0.046 in absolute terms and 5.5% in relative terms. When compared with the best competing method on each individual AOI, MDF-Det yields absolute F_1 gains ranging from 0.013 to 0.048. The largest improvement is observed on AOI 03, where the F_1 score increases from 0.840 to 0.888. Improvements are also observed on AOIs 01 and 40, where the best baseline scores of 0.805 and 0.792 increase to 0.848 and 0.835, respectively. Overall, the results indicate that the performance improvement of MDF-Det is maintained across all six evaluated regions, while its cross-AOI standard deviation remains relatively low.

Table 2. F_1 score comparison of different methods across six AOIs. The Average column reports the mean $\pm$ standard deviation over all six AOIs. The best results are highlighted in **bold**.

Method	AOIs						Average
	01	02	03	34	40	41	
ClusterNet [11]	0.805	0.901	0.779	0.869	0.792	0.775	0.820 ± 0.052
CNNs [15]	0.691	0.799	0.558	0.774	0.681	0.726	0.705 ± 0.085
HMRN [17]	0.769	0.821	0.800	0.878	0.746	0.751	0.794 ± 0.050
HM-Net [6]	0.759	0.893	0.840	0.927	0.750	0.825	0.832 ± 0.071
CATLoss [34]	0.743	0.660	0.748	0.765	0.729	0.743	0.731 ± 0.037
MDF-Det	**0.848**	**0.914**	**0.888**	**0.941**	**0.835**	**0.843**	$\mathbf{0.878 \pm 0.043}$

The reported mean and standard deviation characterize cross-region variability but do not directly quantify the uncertainty of the paired improvement over the strongest baseline. To further assess this uncertainty while accounting for temporal dependence between adjacent WAMI frames, we additionally perform a paired temporal-block bootstrap analysis between MDF-Det and HM-Net on the test sequence [35]. Under this additional temporal-block analysis, 10,000 paired bootstrap replicates are generated using non-overlapping 20-frame blocks, yielding an estimated paired F_1 improvement of 0.0828 with a 95% percentile confidence interval of [0.0766, 0.0892]. Sensitivity analyses using 10- and 30-frame blocks yield confidence intervals of [0.0775, 0.0874] and [0.0763, 0.0887], respectively, demonstrating that the estimated paired improvement remains consistent across reasonable temporal block lengths.

5.3. Ablation Study

Table 3 evaluates the individual and complementary contributions of MAFF, SA-TD, and SPGF on AOI03. Among the individual modules, SA-TD provides the largest improvement, primarily by substantially increasing Recall, indicating its effectiveness in recovering and separating densely distributed target responses. In contrast, MAFF shifts the detection balance toward higher Precision, while SPGF improves Precision without reducing Recall, consistent with its role as a contextual false-alarm filter. More importantly, the modules exhibit clear complementarity: combining MAFF with SA-TD increases the F_1 score from 0.801 to 0.883, and further introducing SPGF improves Precision from 0.838 to 0.847 while preserving Recall at 0.933, yielding the best overall F_1 score of 0.888. These results indicate that MAFF improves candidate quality, SA-TD enhances target separability, and SPGF provides an additional scene-level constraint for false-alarm suppression.

Table 3. Ablation study of the proposed MDF-Det framework on AOI03. The best detection results are highlighted in **bold**.

Modules			Metrics			
MAFF	**SA-TD**	**SPGF**	**Recall**	**Precision**	F_1	**Time (s/Frame)**
–	–	–	0.861	0.748	0.801	5.81
✓			0.807	0.800	0.803	3.54
	✓		**0.950**	0.753	0.840	5.66
		✓	0.861	0.771	0.813	5.58
✓	✓		0.933	0.838	0.883	**3.26**
	✓	✓	**0.950**	0.774	0.853	5.71
✓		✓	0.807	0.810	0.808	3.30
✓	✓	✓	0.933	**0.847**	**0.888**	3.32

Table 4 further examines the internal contributions of spatial attention and threshold degradation (TD) in SA-TD. When TD is disabled, a simple greedy peak-selection strategy is used instead, which iteratively selects the maximum response from the heatmap, removes the selected response, and retains the top 20 candidate locations. The results show that spatial attention provides the dominant improvement by markedly increasing Precision while maintaining high Recall, whereas TD further suppresses ambiguous heatmap responses and improves the Recall–Precision balance. Their combination achieves the highest F_1 score of 0.888, confirming the complementary roles of feature enhancement and adaptive target decoupling within SA-TD.

Table 4. Internal ablation study of the SA-TD module on AOI03. The best detection results are highlighted in **bold**.

SA-TD Components		Metrics			
Spatial Attention	**Threshold Degradation**	**Recall**	**Precision**	F_1	**Time (s/Frame)**
–	–	**0.975**	0.604	0.746	3.15
–	✓	0.807	0.810	0.808	3.07
✓	–	0.958	0.817	0.882	3.16
✓	✓	0.933	**0.847**	**0.888**	3.32

5.4. Hyperparameter Analysis

The main hyperparameters of MAFF, SA-TD, and SPGF are analyzed on the temporally separated validation set to characterize their effects on the Recall–Precision trade-off and overall detection behavior. The selected settings are then fixed for all subsequent test-set inference without further adjustment.

(1) MAFF benefits from moderate motion enhancement, while excessive enhancement introduces additional background responses. As shown in Figure 8a, increasing α initially improves Recall, whereas larger values reduce Precision and F_1. Increasing τ_{diff} further shifts the operating point toward higher Precision and lower Recall. The F_1 score peaks at 0.917 when $\alpha = 15$ and $\tau_{\text{diff}} = 1$, and these values are therefore selected for subsequent test-set inference.

(2) SA-TD becomes increasingly precision-oriented as the screening thresholds increase, with intermediate settings remaining the most stable. Figure 8b shows that increasing τ_{bin} or τ_{min} generally reduces Recall while improving Precision, producing a unimodal F_1 trend. A stable high-performance region appears around $\tau_{\text{bin}} = 0.7$ and $\tau_{\text{min}} = 0.15$–0.17, with the maximum F_1 of 0.918 obtained at $\tau_{\text{bin}} = 0.7$ and $\tau_{\text{min}} = 0.17$; these values are therefore selected.

(3) SPGF progressively favors Precision as scene-prior filtering becomes stricter, while excessive filtering reduces overall performance. As shown in Table 5, increasing τ_{prior} consistently lowers Recall and improves Precision, causing F_1 to first increase and then decline. The highest F_1 of 0.922 is obtained at $\tau_{\text{prior}} = 0.2$ for both $\tau_{\text{bin}} = 0.7$ and 0.9. Since $\tau_{\text{bin}} = 0.7$ retains higher Recall, $\tau_{\text{prior}} = 0.2$ and $\tau_{\text{bin}} = 0.7$ are selected for subsequent test-set inference.

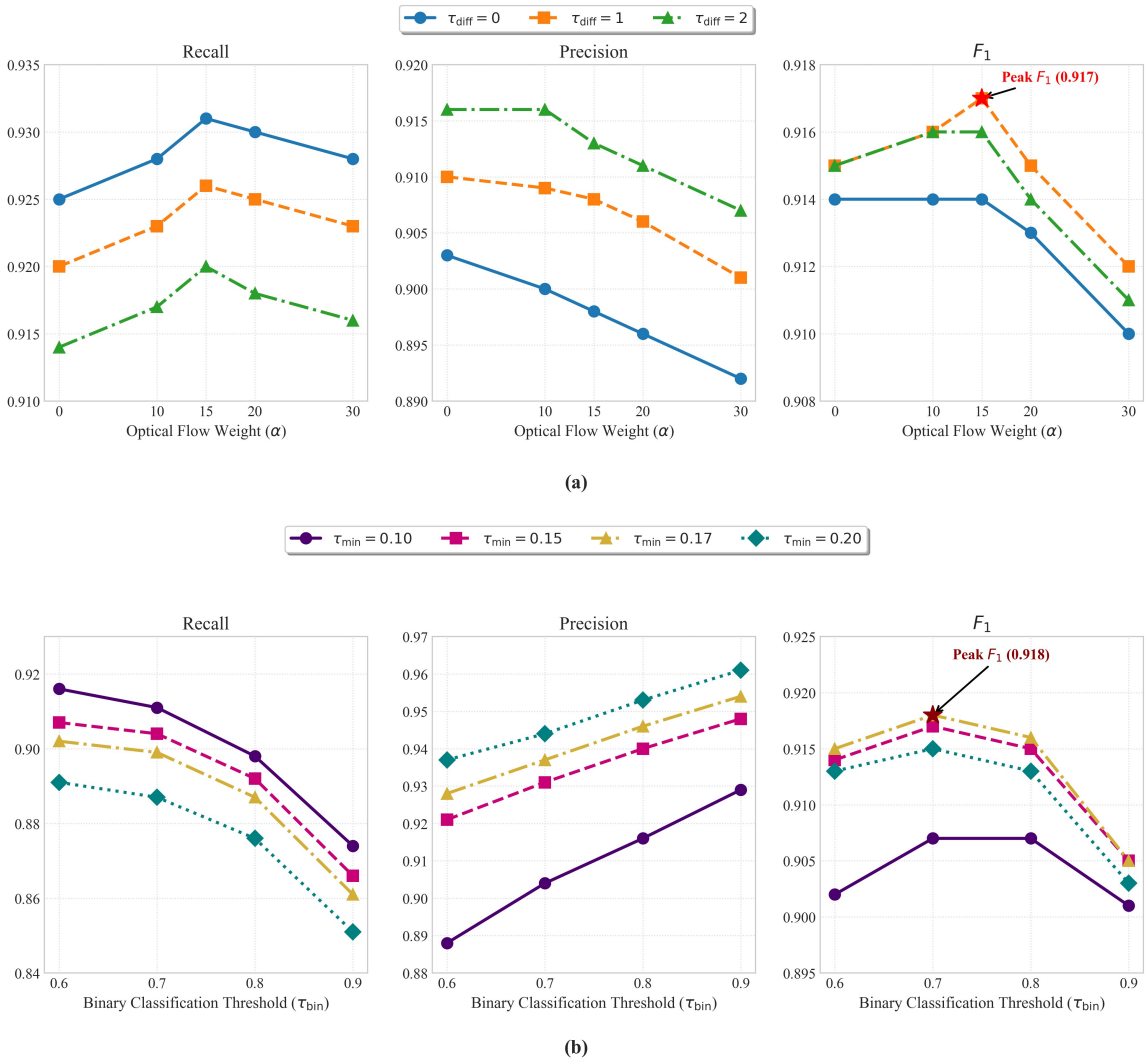


Figure 8. Validation-set hyperparameter sensitivity of (**a**) MAFF to α and τ_{diff}, and (**b**) SA-TD to τ_{bin} and τ_{min}.

Table 5. Validation-set performance under different combinations of the binary classification threshold τ_{bin} and scene-prior confidence threshold τ_{prior}.

τ_{prior}	$\tau_{\text{bin}} = 0.5$			$\tau_{\text{bin}} = 0.7$			$\tau_{\text{bin}} = 0.9$		
	Recall	**Precision**	F_1	**Recall**	**Precision**	F_1	**Recall**	**Precision**	F_1
0.0	0.942	0.841	0.889	0.937	0.866	0.900	0.919	0.893	0.906
0.1	0.938	0.872	0.904	0.933	0.898	0.915	0.911	0.923	0.917
0.2	0.928	0.898	0.913	0.925	0.918	0.922	0.903	0.942	0.922
0.3	0.913	0.910	0.912	0.906	0.931	0.918	0.886	0.948	0.916
0.4	0.894	0.924	0.909	0.886	0.941	0.913	0.860	0.955	0.905

5.5. Sensitivity to Localization Radius

To examine the influence of the localization criterion, we further evaluate MDF-Det using matching radii ranging from 5 to 25 pixels. All results are obtained from the same fixed detection outputs, with only the localization radius used for one-to-one matching being changed. Each ground-truth target is matched to at most one detection; unmatched detections and ground-truth targets are counted as false positives and false negatives, respectively.

As shown in Table 6, detection performance improves gradually as the localization radius increases, while the improvement becomes marginal at larger radii. Increasing the radius from 10 to 20 pixels improves the F_1 score by only 0.005, whereas a further increase to 25 pixels yields an additional gain of only 0.003. This relatively small variation indicates that MDF-Det is not highly sensitive to the localization radius and that the reported performance is not primarily attributed to the adopted 20-pixel matching criterion.

Table 6. Sensitivity of MDF-Det to different localization radii. The **bold** value denotes the radius adopted in the main experiments.

Radius (px)	Precision	Recall	F_1
5	0.941	0.862	0.900
10	0.948	0.874	0.909
15	0.950	0.879	0.913
20	**0.952**	**0.879**	**0.914**
25	0.954	0.883	0.917

5.6. Computational Complexity and Runtime Analysis

Table 7 summarizes the computational complexity and runtime distribution of MDF-Det. The complete framework contains 10.794 M parameters, while its per-frame online neural-network inference requires 490.069 G FLOPs, with an overall runtime of 3.161 s/frame. FLOPs are reported for per-frame online neural-network inference: MAFF mainly consists of non-neural image-processing operations, whereas SPGF prior construction is performed once for each fixed AOI and is therefore excluded from the per-frame FLOP count. The binary classifier is the main computational bottleneck, accounting for 51.28% of the total runtime, mainly because it processes a large number of motion-generated candidate patches. In contrast, although SA-TD contains more parameters, it is applied only to a small subset of retained candidates, resulting in substantially lower FLOPs and runtime.

Table 7. Computational complexity and runtime profile of MDF-Det on the validation set.

Module	Parameters (M)	Model Size (MiB, FP32)	FLOPs (G/Frame)	Peak GPU Memory (MiB)	Time (s/Frame)	Runtime Contribution (%)
MAFF	0	–	–	–	0.864	27.32
Binary classifier	0.849	3.240	487.571	1176.0	1.621	51.28
SA-TD	8.337	31.803	2.498	84.2	0.676	21.40
SPGF	1.608	6.134	–	441.5	<0.001	0.01
Complete MDF-Det	10.794	41.177	490.069	1576.9	3.161	100.00

MAFF introduces no trainable parameters, with its computational cost mainly arising from multi-frame registration, motion compensation, optical-flow fusion, and background subtraction. SPGF incurs negligible online overhead because its scene prior is constructed only once for each fixed AOI, while online processing involves only prior lookup and candidate filtering. With a runtime of 3.161 s/frame, the current implementation does not yet meet the real-time processing requirement of WAMI systems operating at approximately

1–2 fps. Dense optical flow is currently executed on the CPU because the adopted OpenCV build does not provide CUDA optical-flow support, while the binary classifier remains the dominant computational bottleneck. Further acceleration could therefore be achieved through GPU-based optical-flow computation and more efficient candidate screening.

6. Discussion

The experimental results demonstrate that MDF-Det effectively addresses several key challenges in WAMI target detection. By integrating dense optical flow with grayscale differencing, MAFF enhances weak motion responses and improves the quality of foreground candidates. SA-TD further separates merged responses in densely populated regions through fine-grained heatmap decoding and step-threshold degradation, while SPGF introduces soft scene-level constraints to suppress false alarms caused by registration artifacts, vegetation motion, and other contextually implausible background responses. The ablation results confirm that the three components perform distinct yet complementary functions: MAFF shifts the detection balance toward higher Precision, SA-TD provides the largest gain in Recall and improves target separability, and SPGF further improves Precision through scene-level filtering.

Despite these advantages, MAFF remains dependent on the quality of inter-frame registration and optical-flow estimation. Strong parallax, abrupt illumination changes, or rapid camera motion may introduce residual misalignment or unreliable motion responses, thereby degrading candidate generation. MDF-Det also inherits a common limitation of candidate-driven cascaded pipelines. Since SA-TD and SPGF operate only on candidates retained by MAFF, weak targets completely missed during the initial candidate-generation stage cannot be recovered in subsequent refinement. Although the zero-shot transfer results indicate cross-scene transferability, SPGF is learned from trajectory-derived supervision within the training sequence and may therefore still exhibit some dependence on scene-specific spatial statistics.

A promising future direction is therefore to introduce a semantic feedback branch that complements the motion-based front end. By integrating motion confidence, foundation-model semantics, and soft road-network priors through uncertainty-aware gating, such a branch could reassess low-confidence regions and generate supplementary candidates in areas with weak motion responses but strong contextual evidence. This would provide a practical way to improve weak-target recovery while reducing the dependence of SPGF on scene-specific spatial statistics.

7. Conclusions

In this paper, we propose MDF-Det, a coarse-to-fine framework for tiny moving target detection in WAMI. By sequentially integrating motion–appearance fusion, spatial attention-guided target decoupling, and scene-prior filtering, MDF-Det effectively mitigates missed detections caused by weak motion responses, merged detections in dense scenes, and scene-induced false alarms. Experimental results on six evaluation AOIs of the WPAFB 2009 dataset demonstrate that MDF-Det achieves an average F_1 score of 0.878, corresponding to a relative improvement of 5.5% over the strongest baseline. These results demonstrate the effectiveness of MDF-Det across the six evaluated regions of the WPAFB benchmark.

Future work will focus on improving the cross-scene generalization of scene-prior modeling and reducing the dependence of the cascaded framework on the initial candidate-generation stage. We also plan to investigate semantic feedback mechanisms that integrate motion confidence, foundation-model semantics, and soft road-network priors, and to further validate MDF-Det on additional WAMI datasets and more diverse aerial imaging scenarios.

Author Contributions: Conceptualization, Z.Z.; funding acquisition, Z.Z. and Y.L.; supervision, Z.Z. and H.X.; project administration, H.X. and Y.L.; resources, H.X.; investigation, K.L.; methodology, K.L.; writing—original draft preparation, K.L. and X.Z.; writing—review and editing, X.Z. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant numbers 61402491 and 62071478. The APC was funded by the National Natural Science Foundation of China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The WPAFB 2009 and Greene 2007 datasets used in this study are publicly available from the U.S. Air Force Research Laboratory. The implementation of MDF-Det is available at https://github.com/likangqiushi20/MDF-Det (accessed on 8 September 2026).

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

References

1. Xiao, J.; Cheng, H.; Sawhney, H.; Han, F. Vehicle Detection and Tracking in Wide Field-of-View Aerial Video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2010; pp. 679–684.
2. Porter, R.; Fraser, A.M.; Hush, D. Wide-Area Motion Imagery. *IEEE Signal Process. Mag.* **2010**, *27*, 56–65. [CrossRef]
3. Cobb, M.; Reisman, M.; Killam, P.; Fiore, G.; Siddiq, R.; Giap, D.; Chern, G. Wide-Area Motion Imagery Vehicle Detection in Adverse Conditions. In *Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
4. Reilly, V.; Idrees, H.; Shah, M. Detection and Tracking of Large Number of Targets in Wide-Area Surveillance. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 186–199.
5. Nkrumah, I.; Moshrefizadeh, M.; Tahri, O.; Blasch, E.; Palaniappan, K.; AliAkbarpour, H. EC-WAMI: Event Camera-Based Pose Optimization in Remote Sensing and Wide-Area Motion Imagery. *Sensors* **2024**, *24*, 7493. [CrossRef] [PubMed]
6. Motorcu, H.; Ates, H.F.; Ugurdag, H.F.; Gunturk, B.K. HM-Net: A Regression Network for Object Center Detection and Tracking on Wide-Area Motion Imagery. *IEEE Access* **2021**, *10*, 1346–1359. [CrossRef]
7. Palaniappan, K.; Rao, R.M.; Seetharaman, G. Wide-Area Persistent Airborne Video: Architecture and Challenges. In *Distributed Video Sensor Networks*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 349–371.
8. Air Force Research Laboratory. *Wright-Patterson Air Force Base (WPAFB) 2009 Dataset*; Publicly available dataset; Air Force Research Laboratory (AFRL): Dayton, OH, USA, 2009.
9. Saleemi, I.; Shah, M. Multiframe Many–Many Point Correspondence for Vehicle Tracking in High-Density Wide-Area Aerial Videos. *Int. J. Comput. Vis.* **2013**, *104*, 198–219. [CrossRef]
10. Teutsch, M.; Grinberg, M. Robust Detection of Moving Vehicles in Wide-Area Motion Imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 27–35.
11. LaLonde, R.; Zhang, D.; Shah, M. ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatiotemporal Information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4003–4012.
12. Negin, F.; Tabejamaat, M.; Fraisse, R.; Bremond, F. Transforming Temporal Embeddings to Keypoint Heatmaps for Detection of Tiny Vehicles in Wide-Area Motion Imagery (WAMI) Sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1432–1441.
13. Zivkovic, Z. Improved Adaptive Gaussian Mixture Model for Background Subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2004; Volume 2, pp. 28–31.
14. Keck, M.; Galup, L.; Stauffer, C. Real-Time Tracking of Low-Resolution Vehicles for Wide-Area Persistent Surveillance. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2013; pp. 441–448.
15. Zhou, Y.; Maskell, S. Detecting and Tracking Small Moving Objects in Wide-Area Motion Imagery (WAMI) Using Convolutional Neural Networks (CNNs). In *Proceedings of the 22nd International Conference on Information Fusion*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
16. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
17. Ates, H.F.; Siddique, A.; Gunturk, B. HMRN: Heat Map Regression Network to Detect and Track Small Objects in Wide-Area Motion Imagery. *Signal Image Video Process.* **2023**, *17*, 39–45. [CrossRef]

18. Li, S.; Guo, A.; Wang, Y.; Li, L.; Wang, G.; Wu, Z. Combining Visual Motion and Luminance Features to Enhance Detection of Small Moving Objects in a Bioinspired Model. *PLoS Comput. Biol.* **2026**, *22*, e1014036. [CrossRef] [PubMed]
19. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An Efficient Alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2011; pp. 2564–2571.
20. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Nature: Berlin/Heidelberg, Germany, 2022.
21. Fortun, D.; Bouthemy, P.; Kervrann, C. Optical Flow Modeling and Computation: A Survey. *Comput. Vis. Image Underst.* **2015**, *134*, 1–21. [CrossRef]
22. Wang, T.; Zhu, L.; Huang, H. Enhancing Real-Time Object Detection with Optical Flow-Guided Streaming Perception. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 4816–4830. [CrossRef]
23. Zhang, T.; Chen, C.; Liu, Y.; Geng, X.; Aly, M.M.S.; Lin, J. PSRR-MaxpoolNMS++: Fast Non-Maximum Suppression with Discretization and Pooling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 978–993. [CrossRef] [PubMed]
24. Shi, Y.; Li, M.; Chen, N.; Luo, Y.; He, S.; An, W. Sparse-Gated RGB-Event Fusion for Small Object Detection in Wild. *Remote Sens.* **2025**, *17*, 3112. [CrossRef]
25. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
26. Zhou, Y.; Jiang, Y.; Yang, Z.; Yang, Y.; Wang, Y.; Li, X.; Sun, W.; Han, Z. A Small Moving Object Detection Method Based on Attention Mechanism and Location Analysis from Infrared Images. In *Proceedings of the International Conference on Image, Signal Processing, and Pattern Recognition*; SPIE: Bellingham, WA, USA, 2024; Volume 13180, p. 1318002.
27. Haklay, M.; Weber, P. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
28. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
29. Wu, Y.; He, K. Group Normalization. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
30. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for Activation Functions. *arXiv* **2017**, arXiv:1710.05941.
31. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
32. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 4th International Conference on 3D Vision, Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
33. Air Force Research Laboratory. *Greene 07 Dataset*; Publicly available wide-area electro-optical imagery dataset; Air Force Research Laboratory (AFRL): Dayton, OH, USA, 2007.
34. Hatipoglu, P.U.; Iyigun, C.; Kalkan, S. Crowd-Aware Thresholded Loss for Object Detection in Wide-Area Motion Imagery. *PFG Photogramm. Remote Sens. Geoinf. Sci.* **2023**, *91*, 339–364. [CrossRef]
35. Calhoun, G. Block Bootstrap Consistency under Weak Assumptions. *Econom. Theory* **2018**, *34*, 1383–1406. [CrossRef]

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.