

FDMA-Net: Four-directional Mamba alignment for tiny object detection in visible and thermal imagery

Sungju An

Gachon University

Siwoo Kim

Gachon University

Jhonghyun An

jhonghyun@gachon.ac.kr

Gachon University

Article

Keywords: Visible–thermal object detection, Tiny object detection, Mamba, Cross-modal alignment, Thermal detail enhancement, Distributional localization

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11225780/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

FDMA-Net: Four-directional Mamba alignment for tiny object detection in visible and thermal imagery

Sungju An¹, Siwoo Kim¹, Jhonghyun An^{1*}

^{1*}Department of Artificial Intelligence and Software, Gachon University, 1342 Seongnam-daero, Sujeong-gu, Seongnam-si, 13120, Gyeonggi-do, Republic of Korea.

*Corresponding author(s). E-mail(s): jhonghyun@gachon.ac.kr;
Contributing authors: samsj1209@gachon.ac.kr; ksw5830@gachon.ac.kr;

Abstract

Tiny object detection in paired visible and thermal aerial images is challenging because spatial misalignment weakens cross-modal correspondence and small targets provide limited spatial detail. To address these challenges, we propose the Four-Directional Mamba Alignment Network (FDMA-Net), which integrates cross-modal feature alignment, thermal detail enhancement, and distributional localization. The Four-Directional Mamba Alignment (FDMA) module uses bidirectional Mamba scans along rows and columns to estimate sampling displacements, with learned confidence weights aggregating opposite-direction estimates for feature alignment. Thermal Detail Enhancement (TDE) supplies shallow thermal features to the finest detection level and the center-prediction branch to retain fine spatial detail. The Center-Size Distribution Head (CSDH) models signed center offsets and object dimensions using discrete probability distributions to support tiny-object localization. Experiments on RGBTDronePerson and VTUAV-det achieve mean AP50 values of 49.08% and 75.7%, respectively, across three seeds. These results exceed the highest AP50 values reported for the compared methods by 3.63 and 3.9 percentage points, respectively.

Keywords: Visible–thermal object detection, Tiny object detection, Mamba, Cross-modal alignment, Thermal detail enhancement, Distributional localization

1 Introduction

Detecting tiny objects in aerial imagery requires distinguishing targets from complex backgrounds using only a small number of pixels [1, 2]. Visible images provide appearance and texture, whereas thermal imagery can support detection under low illumination [3]. Cross-modal fusion exploits this complementarity [4], but requires reliable association of information with the same target [5, 6].

Spatial correspondence is therefore central to RGB–thermal fusion [5, 7]. When a target occupies only a few feature locations, a displacement between modalities can cause target information in one image to be combined with background information in the other. Existing multispectral detectors address this issue through feature alignment and cross-modal calibration [5, 7, 8]. For aerial tiny objects, COXNet combines adaptive alignment with cross-layer fusion and scale integration [6]. These approaches highlight the need to estimate reliable sampling displacements when tiny targets provide weak or ambiguous cross-modal correspondence cues.

When local correspondence cues are weak or ambiguous, surrounding spatial context can provide additional evidence for estimating cross-modal displacements. Selective state-space models such as Mamba accumulate input-dependent information along sequences [9], while visual Mamba models extend this capability to image features through multidirectional scanning [10]. This motivates using bidirectional scans along image rows and columns to incorporate spatial context into horizontal and vertical displacement estimation.

We therefore propose the Four-Directional Mamba Alignment Network (FDMA-Net) for tiny object detection in paired visible and thermal aerial imagery. Its core module, Four-Directional Mamba Alignment (FDMA), estimates visible-feature sampling displacements from four-directional cross-modal context and combines opposing estimates using learned confidence weights. Directional consistency regularization encourages agreement between opposing predictions during training. The aggregated displacements guide visible-feature sampling relative to thermal references before cross-modal fusion. To complement feature alignment, Thermal Detail Enhancement (TDE) transfers shallow thermal detail to the finest detection level and the center-prediction branch. The Center-Size Distribution Head (CSDH) then models signed center offsets and object dimensions as discrete probability distributions for localization.

The main contributions are summarized as follows:

- We propose Four-Directional Mamba Alignment (FDMA), which uses bidirectional Mamba scans along rows and columns to estimate visible-feature sampling displacements. Learned confidence weights aggregate opposing directional estimates, and directional consistency regularization encourages their agreement during training. The resulting displacements guide visible-feature sampling relative to thermal references before cross-modal fusion.
- We introduce Thermal Detail Enhancement (TDE), which transfers shallow thermal C2 features to the finest P3 detection level to supplement aligned features with fine spatial detail. The thermal detail pathway also supplies subcell information to the center-prediction branch.

- We develop a Center-Size Distribution Head (CSDH) that separately models signed center offsets and full object dimensions using discrete probability distributions. This representation separates center displacement from object extent for bounding-box prediction.

2 Related Work

2.1 Visible–thermal alignment

Spatial correspondence is central to combining visible and thermal evidence. AR-CNN aligns region features [5], whereas C²Former uses inter-modality cross-attention to learn calibrated representations [7]. For aerial tiny objects, COXNet combines convolutional sampling-offset prediction with cross-layer fusion and scale refinement [6]. MSDF-Mamba derives deformable alignment offsets from a shared reference before selective-scan fusion [11]. These approaches establish explicit spatial correction as a component of multispectral detection. FDMA investigates the contextual evidence used for displacement estimation, combining opposing directional predictions before sampling visible features relative to a thermal reference.

2.2 State-space models for multimodal interaction

Mamba uses input-dependent state-space parameters to propagate information along sequences, with linear scaling in sequence length for fixed model dimensions [9]. VMamba extends selective scanning to visual features through multiple traversal routes [10]. Multimodal applications include patch-level fusion in RemoteDet-Mamba [12], wavelet-based fusion in WaveMamba [13], disparity-guided fusion in DMM [14], and interaction between thermal and generated pseudo-visible features in PFGF [15]. Our use of Mamba places directional context within sampling-displacement estimation, with confidence weighting and consistency coupling opposite estimates. This design converts directional multimodal context into an explicit sampling field for feature alignment.

2.3 Spatial detail and distributional localization

Feature Pyramid Networks combine information across scales [16], while COXNet transfers information between modality levels through cross-layer fusion [6]. Such designs motivate access to shallow features when targets contain limited spatial evidence [17]. TDE transfers thermal detail to the finest detection level and supplies subcell features for center prediction. For localization, center-based detection represents objects through center points and dimensions [18]; distributional regression models box-boundary locations with discrete probabilities [19, 20]. CSDH combines center-based object representation with distributional regression by modeling signed center offsets and full object dimensions as discrete probability distributions. Within FDMA-Net, TDE enriches the finest detection features with shallow thermal detail, while CSDH predicts object centers and sizes for bounding-box localization.

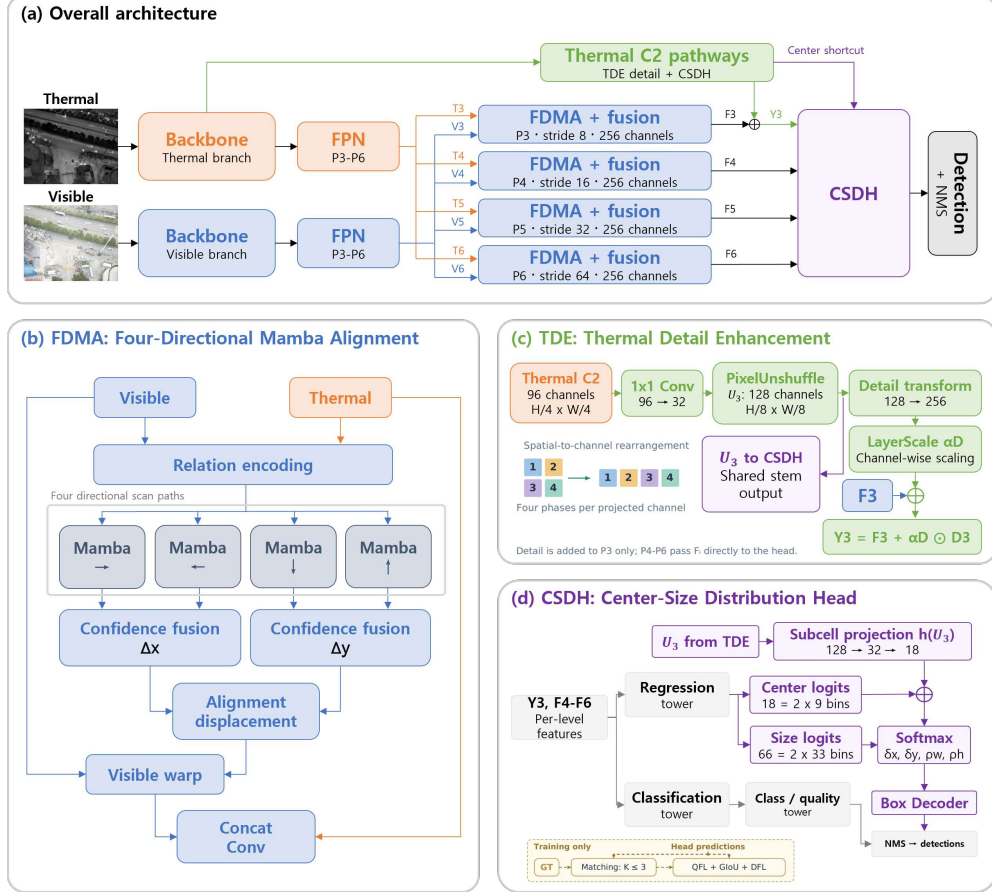


Fig. 1 Overall architecture of FDMA-Net. (a) Two modality-specific backbones and feature pyramids provide visible and thermal features at P3–P6. FDMA produces fused features F_3 – F_6 ; the thermal C2 pathway enhances F_3 into Y_3 and supplies a center-prediction shortcut. (b) Four-directional Mamba alignment combines horizontal and vertical displacement estimates before visible-feature warping and fusion. (c) TDE forms the subcell tensor U_3 by channel projection and PixelUnshuffle, then generates the scaled detail update $\alpha_D \odot D_3$. (d) CSDH receives Y_3, F_4 – F_6 and predicts class scores, 18 center logits, and 66 size logits. The projection $h(U_3)$ updates the P3 center logits before softmax and box decoding.

3 Methods

3.1 Architecture

FDMA-Net follows a GFL-based dense detection framework [19] with two independent YOLOv8m backbones [21] and feature pyramids [16] to produce $V_l, T_l \in \mathbb{R}^{B \times 256 \times H_l \times W_l}$ at levels $l = 3, \dots, 6$, with strides $s_l = 8, 16, 32, 64$. Thermal annotations define the detection coordinates. FDMA fuses the modalities at every level; TDE enhances P3; CSDH predicts classes and boxes. The fused outputs are $F_3, \dots, F_6$; after TDE, CSDH receives Y_3, F_4, F_5, F_6 , as labeled in Fig. 1.

3.2 Four-Directional Mamba Alignment

Let $\mathcal{S}(X, g)$ denote bilinear sampling with zero padding, where normalized coordinates $(-1, -1)$ and $(1, 1)$ correspond to the centers of the top-left and bottom-right feature pixels, respectively. At pyramid level l , the normalized reference coordinate at spatial location (i, j) is defined as

$$r_l(i, j) = \left(\frac{2(j + 0.5)}{W_l} - 1, \frac{2(i + 0.5)}{H_l} - 1 \right).$$

The reference-sampled visible and thermal features are $V_l^0 = \mathcal{S}(V_l, r_l)$ and $T_l^0 = \mathcal{S}(T_l, r_l)$, respectively. This reference sampling is not an identity operation; its exact coordinate mapping is given in Supplementary Methods. Separate convolution–normalization–GELU projections produce 32-channel embeddings v_l and t_l , which are combined into a cross-modal relation feature:

$$R_l = \psi_l([v_l, t_l, v_l - t_l, v_l \odot t_l]), \quad (1)$$

where ψ_l projects the concatenated relation feature to 32 channels.

As shown in Fig. 2, each row and column of R_l forms an independent sequence. For axis $a \in \{h, v\}$, the residual block $\mathcal{B}_{l,a}(X) = X + \text{Mamba}_{l,a}(\text{LN}(X))$ processes forward and reversed sequences; reversed outputs are restored to spatial order. Opposite directions share the block and prediction head, while axes and levels have separate parameters. Each level thus uses two parameterized blocks and four evaluations. The Mamba model and state dimensions are 32 and 16; convolution width is four and expansion is two.

A 3×3 convolution, GELU, and 1×1 convolution predict displacement and confidence logits a_l^d, b_l^d in each direction. For $D_x = \{\rightarrow, \leftarrow\}$ and $D_y = \{\downarrow, \uparrow\}$,

$$\begin{aligned} d_l^d &= r_l^{\max} \tanh(a_l^d), & c_l^d &= \sigma(b_l^d), \\ \bar{d}_l^k &= \frac{\sum_{d \in D_k} c_l^d d_l^d}{\sum_{d \in D_k} c_l^d + 10^{-6}}, \\ \Delta k_l &= \alpha_l \bar{d}_l^k, & \alpha_l &= 2\sigma(\theta_l), \quad k \in \{x, y\}. \end{aligned} \quad (2)$$

The parameter $r_l^{\max}$ bounds each directional displacement before learnable level scaling. We use a base displacement range of $R = 16$ input-image pixels per coordinate axis and set $r_l^{\max} = R/s_l$. For pyramid strides $s_l = (8, 16, 32, 64)$, this gives $(r_3^{\max}, r_4^{\max}, r_5^{\max}, r_6^{\max}) = (2, 1, 0.5, 0.25)$ in feature-map pixels. This setting maintains the same base displacement range across pyramid levels. The learnable scale $\alpha_l \in (0, 2)$ subsequently adjusts the fused displacement at each level. Initial offsets are zero, confidences are 0.5, and $\alpha_l = 1$. All directional predictions use the same coordinate signs. The fused estimates $\bar{d}_l^x, \bar{d}_l^y$ correspond to $\bar{d}_x, \bar{d}_y$ in Fig. 2,

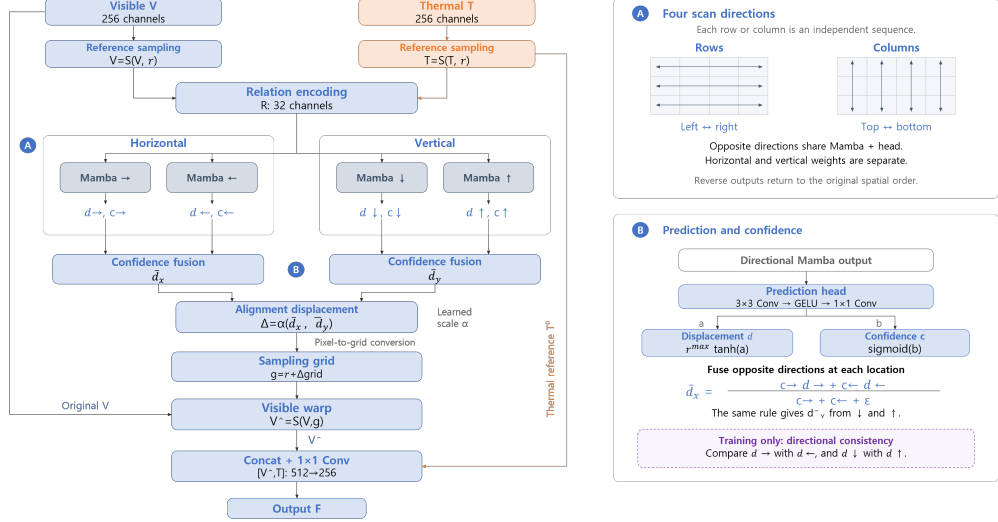


Fig. 2 Detailed FDMA architecture at one pyramid level. Reference sampling and relation encoding produce the 32-channel relation feature R . (A) Each row or column is processed as an independent sequence in opposite directions, with shared Mamba and prediction-head parameters within each axis. (B) Directional heads predict bounded displacements d and confidence values c . Confidence fusion gives $\bar{d}_x$ and $\bar{d}_y$, which are scaled by α and converted to the sampling grid g . The original visible feature is sampled at g and concatenated with the thermal reference, followed by a 1×1 projection from 512 to 256 channels. Directional consistency is applied during training.

where the pyramid-level index is omitted. The displaced grid and fused feature are

$$g_l = r_l + \left(\frac{2\Delta x_l}{\max(W_l - 1, 1)}, \frac{2\Delta y_l}{\max(H_l - 1, 1)} \right), \quad (3)$$

$$\hat{V}_l = \mathcal{S}(V_l, g_l), \quad F_l = \text{Conv}_{1 \times 1}([\hat{V}_l, T_l^0]).$$

The final sampler acts on the original visible feature. Before confidence weighting and level scaling, opposite estimates are regularized by

$$\mathcal{L}_{\text{dir}} = \frac{1}{4} \sum_{l=3}^6 \text{mean}_{b,i,j} \left(|d_l^{\rightarrow} - d_l^{\leftarrow}| + |d_l^{\downarrow} - d_l^{\uparrow}| \right). \quad (4)$$

The consistency term is averaged over all spatial locations and complements the detection objective during training.

3.3 Thermal Detail Enhancement

TDE projects the native stride-4 thermal C2 feature from 96 to 32 channels and applies PixelUnshuffle with a downsampling factor of two to rearrange spatial information into channels, producing the 128-channel tensor U_3 (Fig. 1c). A depthwise convolution, group normalization, GELU, and pointwise projection form a 256-channel update D_3 . The enhanced feature is $Y_3 = F_3 + \alpha_D \odot D_3$, where the

learnable channel scale starts at 0.1; higher levels are unchanged. In FDMA-Net, a separate projection $h(U_3)$ with channel dimensions $128 \rightarrow 32 \rightarrow 18$, scaled by a fixed 0.1, updates P3 center logits after level scaling and before softmax. The per-level regression-logit scale is learnable and initialized to one. The subcell shortcut directly modifies neither size logits nor higher levels.

3.4 Center-Size Distribution Head

As shown in Fig. 1d, CSDH uses separate four-layer classification and regression towers shared across levels. It predicts C classification logits per spatial location, where $C = 3$ for RGBTDronePerson and $C = 1$ for VTUAV-det, together with 84 regression logits: nine bins per signed center axis over $[-2, 2]$ and 33 bins per full-size axis over $[0, 32]$, in stride units. Relative to the prior center $q_{l,i,j} = (js_l, is_l)$, targets are

$$t = (\delta_x, \delta_y, \rho_w, \rho_h) = \left(\frac{c_x - q_x}{s_l}, \frac{c_y - q_y}{s_l}, \frac{w}{s_l}, \frac{h}{s_l} \right). \quad (5)$$

Softmax is applied independently to the four distributions. Probability-weighted expectations yield the continuous values $\hat{\delta}_x, \hat{\delta}_y, \hat{\rho}_w, \hat{\rho}_h$ used by the Box Decoder in Fig. 1d; box corners are the decoded center plus or minus half the dimensions. Distribution focal supervision interpolates between adjacent support bins, with exact endpoint handling [19].

Training candidates must be valid, non-ignored, within 1.5 strides of the target center along both axes, and representable by the distribution supports. A two-stage cardinality-first minimum-cost assignment first maximizes the number of objects assigned distinct locations, minimizing cost at that cardinality, and then adds at most two unused locations per matched object, giving $K \leq 3$ positive locations as shown in Fig. 1d. Its no-gradient cost combines negative log class probability, twice the generalized-IoU loss [22], and 0.25 times the mean distribution loss across four axes.

3.5 Training objective

The objective combines quality focal classification [19], generalized-IoU [22] and distribution losses weighted by detached maximum class scores, the object-region feature regularizer of Peng et al. [6], and directional consistency:

$$\mathcal{L} = \mathcal{L}_{\text{QFL}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{dist}} + 0.1\mathcal{L}_{\text{feat}} + \lambda_{\text{dir}}(e)\mathcal{L}_{\text{dir}}. \quad (6)$$

Here, e denotes the training epoch. The box and distribution terms include weights 2 and 0.25, respectively. The feature term compares channel distributions of fused and unsampled thermal features within expanded object regions; both retain gradients. The directional weight is 0.01 in epochs 1–8, 0.005 in epochs 9–10, and zero in epochs 11–12.

4 Experiments

4.1 Experimental setup

We evaluate FDMA-Net on RGBTDronePerson [2] and VTUAV-det [2], a detection reannotation of the VTUAV tracking benchmark [23]. RGBTDronePerson contains 4,841 filtered training pairs and 1,225 validation pairs with person, rider, and crowd annotations. For VTUAV-det, person is the sole foreground class in both training and evaluation.

The models are implemented using the MMDetection framework [24] and trained on two RTX A5000 GPUs with a total batch size of 16 and a 12-epoch SGD schedule. RGBTDronePerson images are resized to fit 640×512 while preserving aspect ratio and padded to multiples of 32, with learning-rate drops at epochs 9 and 12. For VTUAV-det, 11,208 person-positive pairs are used for training and all 5,378 original test pairs for evaluation; images are resized with aspect ratio preserved and padded to 640×384 , with paired horizontal flipping ($p = 0.5$) and learning-rate drops at epochs 8 and 11.

We report size-specific detection performance on RGBTDronePerson [2]. Object scale is defined as $s = \sqrt{wh}$, where w and h are the annotated bounding-box width and height in pixels. The Tiny group covers $s \in [1, 20]$ and includes three finer groups: Tiny1 ($s \in [1, 8]$), Tiny2 ($s \in [8, 12]$), and Tiny3 ($s \in [12, 20]$). The Small group covers $s \in [20, 32]$. All size-specific results in Tables 1 and 4 report AP at an IoU threshold of 0.50. Tiny AP is computed jointly over the full Tiny range. The superscript “all” denotes the full evaluated size range, and AP25 denotes AP at an IoU threshold of 0.25.

FDMA-Net results in Table 1 and Table 2 and all ablation results in Table 3 and Table 4 are averaged over runs with seeds 0, 1, and 2. Each run uses its highest-validation-AP50 checkpoint, with checkpoint selection and reporting performed on the same validation split. AP follows the tiny-object evaluator and its ignore rules; AP50:95 averages ten IoU thresholds from 0.50 to 0.95 under this protocol. VTUAV-det results are evaluated on the test split for the person class. AP denotes person-class average precision averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05. AP50 and AP75 use IoU thresholds of 0.50 and 0.75, respectively. For VTUAV-det, AP_s , AP_m , and AP_l follow the standard COCO area-based definitions [25]: small objects have area $A < 32^2$, medium objects satisfy $32^2 \leq A < 96^2$, and large objects have $A \geq 96^2$ pixels². These groups are distinct from the Tiny/Small ranges defined above for RGBTDronePerson. RGBTDronePerson inference uses a score threshold of 0.05, class-wise NMS at IoU 0.3, and at most 100 detections per image. Our timing measurements use one RTX A5000, batch size one, and FP32 tensors, with 50 warm-up iterations followed by three repetitions of 300 images; decoding and NMS are included, while image loading is excluded. Additional implementation and evaluation settings are provided in Supplementary Table S3 and Supplementary Methods.

Table 1 Performance comparison on RGBTDronePerson [2]. AP values are percentages. GFLOPs denotes computational complexity, and Params(M) denotes the number of parameters in millions. FDMA-Net results are averaged over three independent runs with seeds 0, 1, and 2; only the mean values are shown in this table.

Method	AP ₂₅ ^{all}	AP ₅₀ ^{tiny}	AP ₅₀ ^{tiny1}	AP ₅₀ ^{tiny2}	AP ₅₀ ^{tiny3}	AP ₅₀ ^{small}	AP ₅₀ ^{all}	GFLOPs	Params(M)
Cascade R-CNN [†] [26]	42.47	31.99	0.00	29.43	37.77	33.61	31.55	76.23	92.35
RetinaNet [†] [27]	38.92	23.34	4.69	13.57	32.66	15.77	22.87	49.48	59.56
FCOS [†] [28]	45.21	30.71	9.40	22.73	34.87	26.00	29.89	47.21	78.32
ATSS [†] [29]	53.14	37.47	16.92	24.16	43.59	24.62	36.24	48.73	55.31
GFL [†] [19]	56.91	41.67	12.47	30.23	47.68	26.72	39.74	49.22	55.45
FCOS [†] + RFLA [1, 28]	51.41	39.45	25.87	30.93	44.31	25.32	38.20	110.62	78.32
QueryDet [†] [17]	55.16	37.75	17.89	24.90	44.05	25.52	37.07	121.67	62.58
TINet [30]	40.34	28.60	0.00	24.12	34.99	34.97	28.30	92.80	100.86
CFT [31]	37.32	22.83	16.72	18.14	27.52	8.14	22.69	112.02	206.26
HRFuser [32]	33.24	22.50	0.00	26.73	26.26	23.85	22.23	54.17	55.47
QFDet [2]	57.34	44.04	20.27	30.09	50.36	26.78	42.08	81.43	60.19
COXNet [6]	58.61	47.12	18.68	35.03	51.96	28.98	45.45	51.27	71.11
FDMA-Net (ours)	62.07	50.58	22.77	38.65	54.42	31.37	49.08	76.25	36.74

[†]Detector adapted to paired RGB-thermal inputs.

4.2 Comparison with published detectors

Table 1 shows that FDMA-Net achieves the highest overall AP25 (62.07%) and AP50 (49.08%) among the listed methods, improving AP50 over COXNet by 3.63 percentage points. Although FCOS with RFLA achieves higher Tiny1 AP50 and Cascade R-CNN and TINet achieve higher Small AP50, FDMA-Net surpasses all three in both overall and aggregate Tiny AP50. FDMA-Net achieves these performance gains with only 36.74 million parameters, the lowest reported parameter count among the listed detectors, highlighting its parameter efficiency.

Table 2 presents the VTUAV-det results. FDMA-Net achieves AP of 33.1% and AP50 of 75.7%, exceeding COXNet by 1.6 and 3.9 percentage points, respectively. AP_s increases from 15.3% to 19.8%, and AP_m increases from 30.6% to 32.7%. The improvements in AP_s and AP_m further support the effectiveness of FDMA-Net for detecting small and medium-sized objects in aerial visible-thermal imagery. Together, the findings on both datasets support the effectiveness of FDMA-Net for object detection in aerial visible-thermal imagery.

4.3 Component performance and inference cost

We evaluate the progressive integration of FDMA, TDE, and CSDH using the same data splits and optimization schedule. The paired-fusion baseline uses dual backbones with feature concatenation and projection. FDMA introduces directional alignment, while TDE enhances P3 with shallow thermal features. The first three configurations in Table 3 retain the baseline 17-bin boundary-distribution head and its associated assignment strategy. The final configuration retains FDMA and the TDE feature-enhancement pathway while replacing the baseline localization setup with

Table 2 Performance comparison on VTUAV-det [2]. AP values are percentages; AP averages IoU thresholds 0.50–0.95. Subscripts s, m, and l denote small, medium, and large objects. FDMA-Net results are averaged over three independent runs with seeds 0, 1, and 2; only the mean values are reported.

Method	AP	AP50	AP75	AP _s	AP _m	AP _l
ATSS [†] [29]	21.4	52.7	13.9	5.9	20.8	45.2
GFL [†] [19]	29.8	67.8	22.2	10.3	29.7	55.7
QueryDet [†] [17]	29.5	68.9	20.2	7.8	29.9	53.5
CFT [31]	8.7	29.3	2.4	3.9	8.5	23.4
HRFuser [32]	25.9	55.9	20.1	2.7	27.9	51.9
TINet [30]	26.8	59.4	20.1	1.2	29.0	53.7
QFDet [2]	31.1	70.4	22.9	12.5	20.4	56.8
COXNet [6]	31.5	71.8	23.1	15.3	30.6	56.0
FDMA-Net (ours)	33.1	75.7	23.3	19.8	32.7	52.8

[†]Adapted RGBT detector.

Table 3 Component ablation on RGBTDronePerson. Check marks indicate enabled modules. The first and last rows correspond to the paired-fusion baseline and the full FDMA-Net, respectively. AP values are percentages, reported as mean $\pm$ sample standard deviation over three seeds. Params(M) and FPS denote parameter count and inference throughput.

FDMA	TDE	CSDH	AP25	AP50	AP75	AP50:95	Params(M)	FPS
–	–	–	59.13 $\pm$ 1.30	43.44 $\pm$ 0.97	4.10 $\pm$ 0.27	13.92 $\pm$ 0.63	36.424	58.9
✓	–	–	59.33 $\pm$ 0.40	46.30 $\pm$ 0.12	5.65 $\pm$ 0.30	15.96 $\pm$ 0.17	36.661	34.2
✓	✓	–	59.96 $\pm$ 0.23	46.86 $\pm$ 0.52	6.01 $\pm$ 0.25	16.41 $\pm$ 0.29	36.698	33.5
✓	✓	✓	62.07 $\pm$ 0.44	49.08 $\pm$ 0.89	5.47 $\pm$ 0.41	16.71 $\pm$ 0.52	36.738	32.9

the complete CSDH configuration, which includes the center-size distribution head, its corresponding assignment strategy, and the thermal subcell-to-center shortcut.

Table 3 reports three-seed means. Adding FDMA increases mean AP50 from 43.44% to 46.30%. TDE further improves mean AP50 by 0.56 percentage points to 46.86%. The complete CSDH configuration provides an additional gain of 2.22 percentage points, bringing the full FDMA-Net to 49.08% mean AP50.

FDMA adds approximately 0.24 million parameters to the paired-fusion baseline, while TDE and CSDH together add a further 0.077 million. The complete FDMA-Net contains 36.738 million parameters and achieves an inference throughput of 32.9 FPS.

To further examine the directional design of FDMA, we compare single-direction, two-direction, and four-direction configurations with different aggregation schemes in Supplementary Table S1. The final FDMA configuration achieves 49.08% mean AP50, compared with 47.65% for unweighted four-direction averaging, supporting the benefit of reliability-aware aggregation.

Table 4 Ablation of directional consistency regularization in FDMA-Net on RGBTDronePerson. With and without denote training with the scheduled $\mathcal{L}_{\text{dir}}$ term and with its weight fixed to zero, respectively. Tiny, Tiny1–Tiny3, and Small report AP50 for the corresponding object-size groups. All values are percentages; the better result in each column is shown in bold. Results are reported as mean $\pm$ sample standard deviation over three independent training runs with seeds 0, 1, and 2.

Consistency loss	AP25	AP50	AP75	Tiny	Tiny1	Tiny2	Tiny3	Small
With	62.07 ± 0.44	49.08 ± 0.89	5.47 ± 0.41	50.58 ± 1.03	22.77 ± 7.73	38.65 ± 0.80	54.42 ± 0.95	31.37 ± 0.68
Without	61.14 ± 0.44	48.59 ± 0.20	5.56 ± 0.27	50.25 ± 0.26	15.04 ± 8.14	37.03 ± 0.75	56.58 ± 0.80	28.03 ± 0.73

4.4 Effect of directional consistency

We evaluate the contribution of the directional consistency term $\mathcal{L}_{\text{dir}}$ in Eq. 4 by comparing FDMA-Net trained with and without this regularizer. The configuration with consistency uses the scheduled $\lambda_{\text{dir}}(e)$ defined in the training objective, whereas the configuration without consistency sets $\lambda_{\text{dir}}(e) = 0$ throughout training. Both configurations retain FDMA, TDE, and CSDH, along with the remaining loss terms.

As shown in Table 4, directional consistency increases the three-seed mean AP25 from 61.14% to 62.07% and mean AP50 from 48.59% to 49.08%, corresponding to gains of 0.93 and 0.49 percentage points, respectively. The largest increase in mean AP50 occurs for Tiny1, from 15.04% to 22.77%, while Small AP50 increases from 28.03% to 31.37%.

4.5 Comparison before and after alignment

Figure 3 visualizes the visible sampling locations associated with a fixed thermal reference point p . For each object example, S_0 marks the initial visible location, and S_1 marks the location after the FDMA displacement is applied. The thermal point remains fixed while the visible sampling position is updated. The two examples illustrate how FDMA adjusts the visible information retrieved for a thermal reference. Supplementary Table S2 provides an object-center diagnostic of the learned sampling behavior using automatically matched cross-modal object pairs from a training subset.

4.6 Qualitative detection comparisons

Figure 4 compares COXNet, the paired-fusion baseline, and progressively augmented configurations of FDMA-Net. The examples include clustered targets, objects near buildings and vegetation, and sparsely distributed pedestrians. The comparisons illustrate how directional alignment, thermal detail enhancement, and center-size prediction jointly contribute to detecting targets in these scenes.

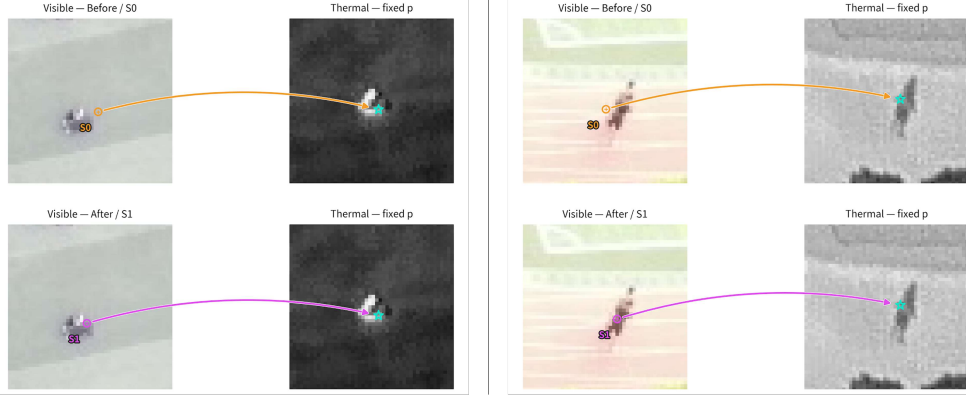


Fig. 3 Visible sampling locations before and after FDMA for two object examples. Each example pairs a visible crop with a thermal reference. The upper row marks the initial visible location S_0 in orange; the lower row marks the adjusted location S_1 in magenta. The cyan star indicates the fixed thermal location p . Arrows associate each visible sampling location with the same thermal reference.

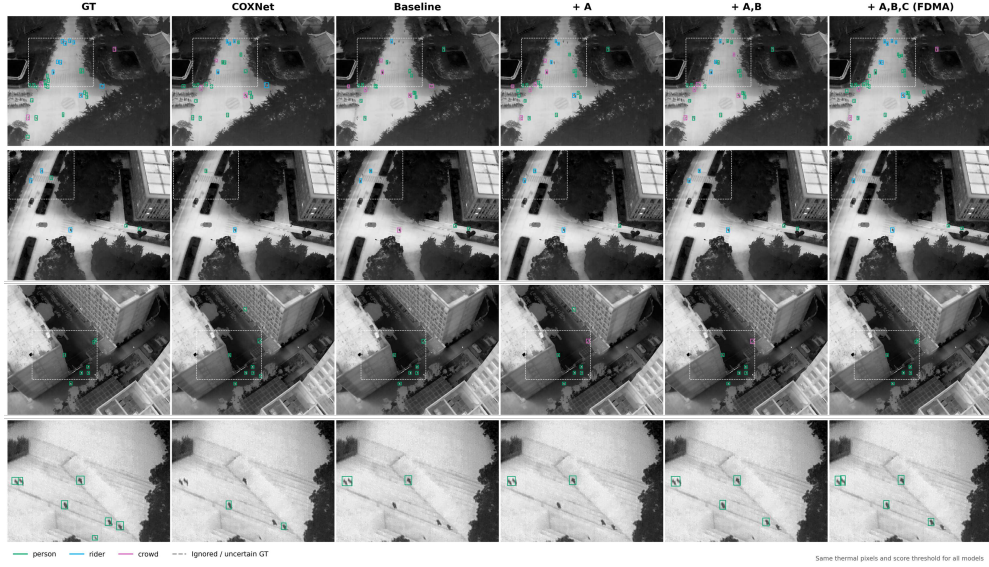


Fig. 4 Qualitative detection comparisons on RGBTDronePerson. GT denotes ground truth. A, B, and C represent FDMA, TDE, and CSDH, respectively; the final column corresponds to FDMA-Net with all three modules. Each row shows predictions on the same thermal image, using a common score threshold across models. Green, cyan, and magenta boxes indicate person, rider, and crowd, respectively. Gray dashed outlines denote ignored or uncertain ground-truth regions.

5 Discussion

The results support the integration of directional contextual modeling, thermal detail enhancement, and center-size localization for aerial visible–thermal tiny-object detection. FDMA-Net improves both overall and aggregate Tiny AP50 on RGBTDronePerson, while also improving small- and medium-object AP on VTUAV-det. In the RGBTDronePerson comparison, these gains are achieved with the lowest reported parameter count among the listed detectors, demonstrating the parameter efficiency of the integrated architecture.

Within the evaluated sequence of component configurations, adding FDMA provides the largest incremental AP50 gain. This result supports the use of surrounding row and column context to guide cross-modal sampling when tiny targets provide limited correspondence cues. The directional study further shows that reliability-aware aggregation outperforms unweighted four-direction averaging, highlighting the importance of confidence weighting when combining directional estimates. Directional consistency complements this design by encouraging agreement between opposing estimates during training and increasing overall mean AP50 in the reported ablation.

TDE and CSDH address feature detail and localization, respectively. TDE enriches the finest detection features with shallow thermal information, while the thermal subcell-to-center shortcut makes this detail directly available to center prediction. CSDH separately models signed center offsets and full object dimensions, with expectation-based decoding producing continuous localization estimates. The additional AP50 gains obtained with TDE and the complete CSDH configuration support combining high-resolution thermal cues with an explicit center-size representation for tiny-object detection.

The results also identify opportunities to refine strict-IoU localization, large-object detection, and inference throughput. Potential extensions include finer distribution supports, thermal detail enhancement at additional pyramid levels, and more efficient directional scanning. Adaptive displacement ranges or iterative refinement could improve the handling of larger residual offsets. Additional training runs and evaluation with an independent checkpoint-selection split would strengthen the assessment of robustness and generalization.

6 Conclusion

We presented FDMA-Net for tiny object detection in visible and thermal aerial imagery. FDMA estimates sampling displacements from four-directional Mamba context, TDE supplies shallow thermal detail, and CSDH predicts signed center offsets and full box dimensions. FDMA-Net achieves AP50 of 49.08% on RGBTDronePerson and 75.7% on VTUAV-det, improving on the compared COXNet configurations by 3.63 and 3.9 percentage points. Three-seed component experiments on RGBTDronePerson show that FDMA raises AP50 by 2.86 percentage points and that the complete model improves AP50 by 5.64 percentage points over paired feature fusion. The consistency ablation further shows a modest increase in overall mean AP50, while its effect varies across object-size groups. The results support

combining directional alignment, thermal detail transfer, and distributional localization for aerial RGB-thermal detection.

Data availability

RGBTDronePerson and the VTUAV-det detection annotations are accessible through the project page at <https://nnnnerd.github.io/RGBTDronePerson/> [2]. VTUAV-det is derived from the VTUAV tracking benchmark [23]. Aggregate results are provided in Tables 1–4.

Code availability

The source code and configuration files for FDMA-Net will be made publicly available on GitHub upon publication. During peer review, the implementation can be provided to editors and reviewers upon reasonable request.

References

- [1] Xu, C. *et al.* RFLA: Gaussian receptive field based label assignment for tiny object detection. In *Computer Vision–ECCV 2022*, 526–543 (2022). https://doi.org/10.1007/978-3-031-20077-9_31.
- [2] Zhang, Y. *et al.* Drone-based RGBT tiny person detection. *ISPRS Journal of Photogrammetry and Remote Sensing* **204**, 61–76 (2023). <https://doi.org/10.1016/j.isprsjprs.2023.08.016>.
- [3] Hwang, S., Park, J., Kim, N., Choi, Y. & Kweon, I. S. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1037–1045 (2015). <https://doi.org/10.1109/CVPR.2015.7298706>.
- [4] Shen, J. *et al.* ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition* **145**, 109913 (2024). <https://doi.org/10.1016/j.patcog.2023.109913>.
- [5] Zhang, L. *et al.* Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019). <https://arxiv.org/abs/1901.02645>.
- [6] Peng, P. *et al.* COXNet: Cross-layer fusion with adaptive alignment and scale integration for RGBT tiny object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **36**, 596–608 (2026). <https://doi.org/10.1109/TCSVT.2025.3595147>.
- [7] Yuan, M. & Wei, X. C²Former: Calibrated and complementary transformer for RGB-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 1–12 (2024). <https://doi.org/10.1109/TGRS.2024.3376819>.

- [8] Chen, C. *et al.* Weakly misalignment-free adaptive feature alignment for UAVs-based multimodal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26826–26835 (2024). <https://doi.org/10.1109/CVPR52733.2024.02534>.
- [9] Gu, A. & Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling* (2024). <https://openreview.net/forum?id=tEYskw1VY2>.
- [10] Liu, Y. *et al.* VMamba: Visual state space model. In *Advances in Neural Information Processing Systems* **37** (2024). <https://doi.org/10.52202/079017-3273>.
- [11] Shen, J. *et al.* MSDF-Mamba: Mutual-spectrum perception deformable fusion Mamba for drone-based visible–infrared cross-modality vehicle detection. *Remote Sensing* **17**, 4037 (2025). <https://doi.org/10.3390/rs17244037>.
- [12] Ren, K., Wu, X., Xu, L. & Wang, L. RemoteDet-Mamba: A hybrid Mamba-CNN network for multi-modal object detection in remote sensing images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 3511–3515 (2026). <https://doi.org/10.1109/ICASSP55912.2026.11463789>.
- [13] Zhu, H. *et al.* WaveMamba: Wavelet-driven Mamba fusion for RGB-infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11219–11229 (2025). <https://doi.org/10.1109/ICCV51701.2025.01044>.
- [14] Zhou, M. *et al.* DMM: Disparity-guided multispectral Mamba for oriented object detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* **63**, 1–13 (2025). <https://doi.org/10.1109/TGRS.2025.3578309>.
- [15] Li, T. *et al.* Pseudo visible feature fine-grained fusion for thermal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6710–6719 (2025).
- [16] Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017). <https://arxiv.org/abs/1612.03144>.
- [17] Yang, C., Huang, Z. & Wang, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13668–13677 (2022). <https://arxiv.org/abs/2103.09136>.
- [18] Zhou, X., Wang, D. & Krähenbühl, P. Objects as points. Preprint at <https://arxiv.org/abs/1904.07850> (2019).

- [19] Li, X. *et al.* Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems* **33** (2020). <https://arxiv.org/abs/2006.04388>.
- [20] Li, X. *et al.* Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021). <https://arxiv.org/abs/2011.12885>.
- [21] Jocher, G., Chaurasia, A. & Qiu, J. Ultralytics YOLOv8. Software (2023). <https://github.com/ultralytics/ultralytics>.
- [22] Rezatofighi, H. *et al.* Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019). <https://arxiv.org/abs/1902.09630>.
- [23] Zhang, P., Zhao, J., Wang, D., Lu, H. & Ruan, X. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8886–8895 (2022). <https://doi.org/10.1109/CVPR52688.2022.00868>.
- [24] Chen, K. *et al.* MMDetection: Open MMLab detection toolbox and benchmark. Preprint at <https://arxiv.org/abs/1906.07155> (2019).
- [25] Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48.
- [26] Cai, Z. & Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6154–6162 (2018). <https://doi.org/10.1109/CVPR.2018.00644>.
- [27] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (2017). <https://doi.org/10.1109/ICCV.2017.324>.
- [28] Tian, Z., Shen, C., Chen, H. & He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 1922–1933 (2022). <https://doi.org/10.1109/TPAMI.2020.3032166>.
- [29] Zhang, S., Chi, C., Yao, Y., Lei, Z. & Li, S. Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020). <https://doi.org/10.1109/CVPR42600.2020.00978>.

- [30] Zhang, Y., Yu, H., He, Y., Wang, X. & Yang, W. Illumination-guided RGBT object detection with inter- and intra-modality fusion. *IEEE Transactions on Instrumentation and Measurement* **72**, 1–13 (2023). <https://doi.org/10.1109/TIM.2023.3251414>.
- [31] Fang, Q., Han, D. & Wang, Z. Cross-modality fusion transformer for multispectral object detection. Preprint at <https://arxiv.org/abs/2111.00273> (2021).
- [32] Brödermann, T., Sakaridis, C., Dai, D. & Van Gool, L. HRFuser: A multi-resolution sensor fusion architecture for 2D object detection. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 4159–4166 (2023). <https://arxiv.org/abs/2206.15157>.

Acknowledgements

This work was supported by the Gachon University research fund of 2025 (GCU-202500710001). This research was also financially supported by the Institute of Civil Military Technology Cooperation, funded by the Defense Acquisition Program Administration and the Ministry of Trade, Industry and Energy of the Korean government, under grant No. 25-PD-EL-01.

Author contributions

Sungju An: Methodology, Software, Investigation, Formal analysis, Writing. Siwoo Kim: Investigation, Formal analysis, Validation. Jhonghyun An: Supervision, Writing, Funding acquisition. All authors reviewed and approved the final manuscript.

Additional information

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FDMASupplementary.pdf](#)