UAV-CB: A Complex-Background RGB–T Dataset and Local Frequency Bridge Network for UAV Detection

Supplementary Material

A. Theoretical Characterization of the Frequency–Spatial Fusion Gap and the Cross-Modality Discrepancy Gap

Let $X^m : \Omega \to \mathbb{R}$ denote an image of modality $m \in \{r, t\}$ (RGB, thermal) defined on a discrete spatial domain $\Omega \subset \mathbb{Z}^2$. A spatial encoder produces a feature field

$$F_s^m = \psi^m(X^m) \in \mathbb{R}^{H\times W\times C}, \tag{20}$$

where $F_s^m(x, y)$ is indexed by pixel coordinates $(x, y) \in \Omega$. A frequency representation is obtained through a (global or local) linear transform $\mathcal{F}$, yielding

$$F_f^m = \mathcal{F}(X^m) \in \mathbb{C}^{U\times V}, \tag{21}$$

with coefficients indexed on the spectral domain $\Omega_f \subset \mathbb{Z}^2$.

A.1. Frequency–Spatial Fusion Gap

From a functional-analytic perspective, the spatial random field $F_s^m(x, y)$ and the spectral random field $F_f^m(u, v)$ inhabit two Hilbert spaces with incompatible coordinate systems and inductive biases. The spatial representation F_s^m encodes *local* structures through convolutional neighborhoods, whereas the frequency representation F_f^m encodes *global* oscillatory patterns while discarding explicit spatial locality due to the globally mixing nature of $\mathcal{F}$.

Ideally, one could posit the existence of a latent representation Z^m with mappings $\phi_s^m : Z^m \to F_s^m$ and $\phi_f^m : Z^m \to F_f^m$ such that (F_s^m, F_f^m) are conditionally consistent views:

$$P(F_s^m, F_f^m \mid Z^m) = P(F_s^m \mid Z^m)\, P(F_f^m \mid Z^m). \tag{22}$$

However, in cluttered UAV scenarios, discriminative spatial cues (weak edges, low-contrast contours) and spectral patterns (high/low-frequency energy distribution) lack any *localized* correspondence. Because the Fourier transform aggregates contributions from all spatial positions, there generally does not exist a *simple* spatially local operator T such that

$$T : F_f^m \mapsto \tilde{F}_f^m(x, y) \tag{23}$$

simultaneously recovers useful spectral information and preserves strict pixelwise correspondence with $F_s^m(x, y)$. Consequently, the joint distribution $P(F_s^m, F_f^m \mid Y)$ typically does not admit a factorization compatible with a spatially anchored latent representation $Z^m(x, y)$.

We refer to this structural incompatibility as the **Frequency–Spatial Fusion Gap**:

> *The gap denotes the absence of a shared, spatially anchored latent representation capable of capturing both global spectral structure and local spatial details, i.e., the difficulty of establishing a coherent localized correspondence between $F_s^m(x, y)$ and $F_f^m(u, v)$.*

A.2. Cross-Modality Discrepancy Gap

For multimodal inputs $m \in \{r, t\}$, spatial features $F_s^r(x, y)$ and $F_s^t(x, y)$ at the same pixel need not reside in a common semantic subspace because RGB and thermal signals arise from fundamentally different physical imaging processes. Let $Y(x, y) \in \{0, 1\}$ denote the class label. An idealized formulation assumes a modality-invariant latent representation $Z(x, y)$ satisfying

$$F_s^r(x, y) = \psi_r(Z(x, y)), \qquad F_s^t(x, y) = \psi_t(Z(x, y)). \tag{24}$$

In practice, the class-conditional distributions

$$\mathcal{P}_r^y = P(F_s^r(x, y) \mid Y = y), \qquad \mathcal{P}_t^y = P(F_s^t(x, y) \mid Y = y) \tag{25}$$

are often significantly misaligned due to differences in reflectance, illumination, emissivity, and atmospheric conditions. Their divergence

$$D(\mathcal{P}_r^y \,\|\, \mathcal{P}_t^y) \tag{26}$$

(e.g., Wasserstein, KL, f-divergence) can be large, making it difficult for any low-complexity mapping T to satisfy $T_{\#}\mathcal{P}_r^y \approx \mathcal{P}_t^y$. Thus RGB and thermal spatial features generally do not inhabit a shared modality-invariant manifold.

We term this mismatch the **Cross-Modality Discrepancy Gap**:

> *The gap reflects the distributional and geometric misalignment between modality-dependent spatial feature distributions, which hinders the existence of a semantically consistent pixel-level isomorphism between RGB and thermal feature spaces.*

A.3. Why Localized Frequency Representations Can Bridge Both Gaps

Let $(\Omega, \mathcal{B}, \mu)$ denote the spatial domain with counting measure. The spatial encoder produces a random field

$$F_s^m : \Omega \to \mathbb{R}^C, \tag{27}$$

which resides in the Hilbert space $\mathcal{H}_s = L^2(\Omega, \mu; \mathbb{R}^C)$. The global Fourier transform $\mathcal{F} : \mathcal{H}_s \to \mathcal{H}_f$ is an isometry, but due to

$$\mathcal{F}(X^m)(u,v) = \sum_{(x,y)\in\Omega} X^m(x,y) e^{-j2\pi(ux+vy)/N}, \quad (28)$$

it destroys spatial locality. As a result, no spatially local operator $T : \mathcal{H}_f \to \mathcal{H}_s$ can, in general, recover spatially localized components from a globally mixed spectrum.

Now consider a localized Fourier operator on a patch cover $\{B_q\}_{q\in Q}$:

$$\mathcal{F}_{\text{loc}}(X^m) := \{\mathcal{F}(X^m \cdot \mathbf{1}_{B_q})\}_{q\in Q}, \quad (29)$$

yielding local spectra $F^m_{f,q}$. Each $F^m_{f,q}$ is indexed by the patch location q, giving rise to the product space

$$\mathcal{H}_{\text{loc-f}} = \prod_{q\in Q} \mathcal{H}_{f,q}, \quad (30)$$

which preserves spatial anchoring.

(i) Partial Spatial Locality. Because $F^m_{f,q}$ depends only on $X^m|_{B_q}$, there exists a spatially local operator T_q such that

$$T_q : \mathcal{H}_{f,q} \to \mathcal{H}_s, \quad \text{supp}(T_q(F^m_{f,q})) \subseteq B_q. \quad (31)$$

Thus, $(F^m_s|_{B_q}, F^m_{f,q})$ become conditionally consistent views of a patch-level latent structure $Z^m(q)$, reducing the Frequency–Spatial Fusion Gap.

(ii) Modality-Invariant Structural Alignment. When X^r and X^t share similar underlying geometry within B_q (e.g., edges or level sets), their local phase patterns tend to become increasingly consistent as the patch size decreases:

$$\Phi^r_q(u,v) \approx \Phi^t_q(u,v). \quad (32)$$

Since phase primarily encodes geometric structure while being less sensitive to modality-dependent amplitude variations, the distributional distance satisfies

$$W_2\big(P(F^r_{f,q}),\, P(F^t_{f,q})\big) \ll W_2\big(P(F^r_s),\, P(F^t_s)\big), \quad (33)$$

indicating that localized spectra provide a more modality-aligned representation than spatial features alone.

Taken together, localized frequency representations (i) preserve spatial locality at the patch level, and (ii) emphasize geometry-driven, modality-invariant structure in the frequency domain. Thus, they provide an effective intermediate space for jointly mitigating both the Frequency–Spatial Fusion Gap and the Cross-Modality Discrepancy Gap.

B. Differences Between Anti-UAV and UAV-CB in Camouflage Scenarios

Since both Anti-UAV and UAV-CB are RGB–T datasets designed for UAV perception, a direct comparison between them is essential for understanding their task orientation and difficulty characteristics. The Anti-UAV dataset is a general-purpose RGB–T UAV tracking benchmark collected mainly in urban environments. Although it includes diverse scenes for generic tracking, the proportion of camouflage-like cases within the dataset is very small. As shown in Fig. 6, the representative Anti-UAV samples exhibiting strong camouflage are rare exceptions, and nearly all of these challenging cases originate exclusively from building-dominated backgrounds. This indicates that Anti-UAV does not explicitly target camouflage or complex low-contrast conditions; instead, its difficult cases occur only incidentally during general urban data collection.

In contrast, UAV-CB is deliberately constructed for camouflage-oriented UAV detection. Instead of relying on naturally occurring difficult frames, UAV-CB systematically selects and curates samples containing severe background interference across a wide variety of challenging environments, including dense vegetation, powerline grids, and ground clutter in addition to urban structures. These diverse backgrounds introduce fundamentally different interference patterns—high-frequency texture aliasing, repetitive line structures, low-contrast silhouettes, and multimodal ambiguity—that are absent or extremely underrepresented in Anti-UAV. Furthermore, the interference intensity in UAV-CB is significantly greater: UAV targets are frequently blended into the background to the point of near invisibility in both RGB and thermal modalities.

Therefore, compared with Anti-UAV, UAV-CB provides both a broader spectrum of interference types and substantially stronger camouflage difficulty, making it a dedicated and more challenging benchmark for evaluating robust multimodal UAV detection under complex low-altitude conditions.

C. Qualitative Visualization Comparison

To further assess the effectiveness of our proposed LFBNet, we provide a method-wise qualitative comparison against two strong baselines: the best-performing single-modal detector RT-DETR and the best-performing multimodal fusion baseline C^{2}Former. Figure 7 presents representative samples covering challenging real-world cases such as low contrast, cluttered structures, and camouflage-like background textures.

RT-DETR (Single-Modality Baseline), despite being the strongest single-modality model in our quantitative evaluation, exhibits notable limitations when target contrast is weak or when background structures resemble UAV edges.

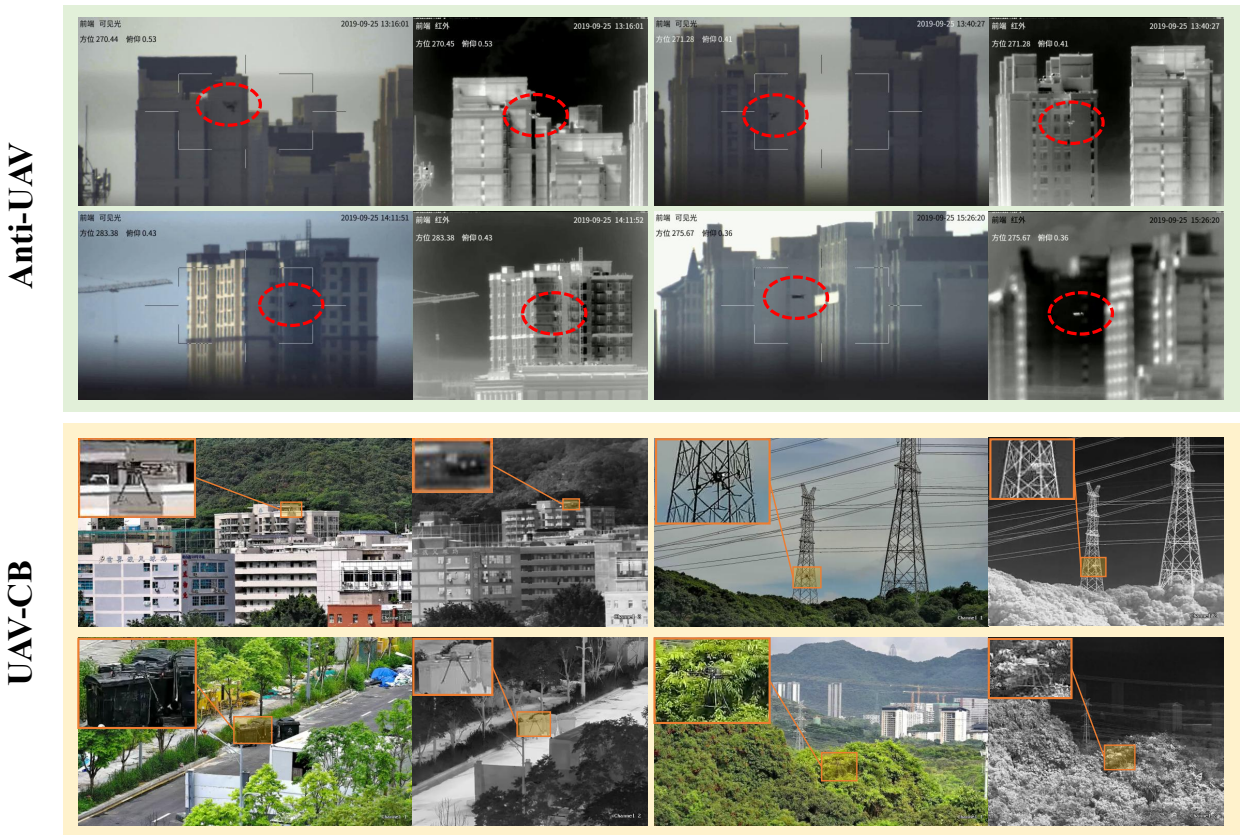

Figure 6. **Comparison between Anti-UAV (top two rows) and UAV-CB (bottom two rows)**. The Anti-UAV samples represent rare camouflage-like cases dominated by building backgrounds, while UAV-CB contains diverse and more challenging interference patterns, including building, vegetation, powerline, cloud, and ground.

Without thermal cues, RT-DETR fails to detect complete UAV boundaries, particularly thin structural parts such as rotor arms and support brackets. As shown in Fig. 7, these components are often partially or entirely missed, leading to under-localized bounding boxes.

C^2Former (Multimodal Fusion Baseline) integrates RGB and thermal features in the spatial domain, resulting in significantly improved boundary completeness compared to RT-DETR. However, due to its sensitivity to high-frequency clutter and modality misalignment, it frequently generates false positives in complex backgrounds. In Figure 7, the model mistakenly highlights strong edges from power lines, tree branches, and building textures as UAV-like structures.

LFBNet consistently produces the most reliable detections across all conditions. By jointly modeling localized frequency alignment (LFCA) and frequency-guided spatial fusion (FGSA), LFBNet achieves:

- accurate and complete UAV boundary reconstruction, including thin structural elements;
- robust suppression of clutter-induced false positives;
- stable detection performance under camouflage, low contrast, and complex background textures.

As seen in Figure 7, LFBNet is able to correctly detect UAVs even when strong structural interference (e.g., dense power lines, similar building edges, or heavy foliage textures) is present.

D. Evaluation results on Anti-UAV Detection

To further assess the generalization ability of our model under different data distributions, we construct an RGB–T detection subset from the Anti-UAV dataset. Since Anti-UAV is originally designed for tracking and contains many sequences with clean, textureless sky backgrounds that do not reflect complex-background challenges, we first remove these non-informative sequences. From each remaining RGB–T video, we then extract one frame every 25 frames, applying the same sampling strategy to both modalities to ensure temporal alignment. This preprocessing yields an RGB–T UAV detection dataset with meaningful background interference, consisting of 2,710 training pairs, 634 validation pairs, and 802 testing pairs.

The results in Table 5 demonstrate that thermal-

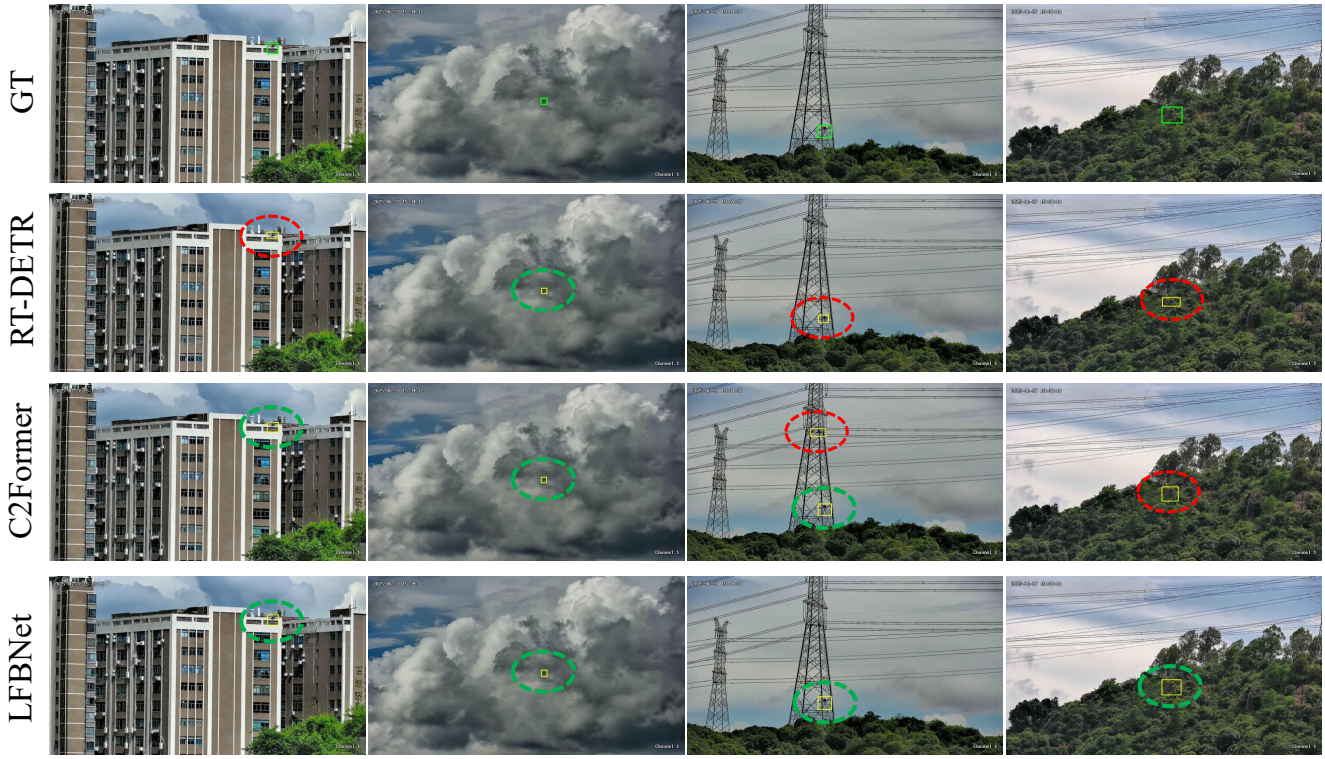

Figure 7. **Qualitative comparison on UAV-CB dataset.** Each method's detection output is visualized with green boxes indicating correct detections and red boxes highlighting missed or false detections.

Table 5. Performance comparison of different methods on the Anti-UAV dataset.

Methods	**Modality**	**AP$_{50}$(%)**	**AP$_{(0.5:0.95)}$(%)**
Faster R-CNN	Visible	63.4	32.4
YOLOv5s		69.6	41.5
RT-DETR		78.2	50.2
YOLOv13s		74.5	48.9
Faster R-CNN	Thermal	78.5	41.4
YOLOv5s		81.3	45.3
RT-DETR		88.1	56.7
YOLOv13s		83.8	51.4
YOLOv5s+Add	RGB+T	72.6	39.2
YOLOv5s+CMX		78.3	46.8
UA-CMDet		86.3	57.5
C^{2}Former		90.4	59.2
SFDFusion		91.2	60.3
LFBNet (Ours)		**92.4**	**62.1**

only detectors consistently outperform visible-only models on Anti-UAV, indicating that infrared modality provides stronger target–background separability under the dataset's illumination conditions. Multimodal fusion brings a clear performance boost, as methods such as UA-CMDet, C^{2}Former, and SFDFusion leverage complementary cross-modal cues to improve robustness against cluttered backgrounds. Among all compared approaches, LFBNet achieves the highest AP50 and AP$(0.5:0.95)$, surpassing the strongest baseline SFDFusion by 1.2% and 1.8%, respectively. This notable gain validates the effectiveness of our localized frequency modeling and frequency-guided alignment strategies, enabling more reliable cross-modal feature integration. Moreover, the superior performance on Anti-UAV—despite its distinct sensor configurations and scene distributions—demonstrates that LFBNet generalizes well beyond the proposed UAV-CB dataset and remains robust across diverse real-world UAV detection scenarios.