DyFCLT: Dynamic Frequency-Decoupled Cross-Modal Learning Transformer for Multimodal Tiny Object Detection

Supplementary Material

6. Appendix

6.1. More Implementation Details

Design details of the baseline. Our baseline is a dual-branch architecture extended from RT-DETR. Specifically, the infrared and visible inputs are first fed into two independent RT-DETR backbones and encoders to obtain multi-scale feature maps for each modality. Then, the features at each scale are fused using a CSPBlock, producing the final multi-scale fused representations (which are identical to those used in our proposed method). Finally, the fused multi-scale features are flattened and passed into the RT-DETR decoder to generate the final predictions.

Detailed experimental settings. For all three datasets used in this study, all training configurations remain identical except for the number of training epochs. We adopt a base learning rate of 0.00025, together with a piecewise decay scheduler whose decay factor is fixed to 1.0. A linear warm-up strategy is applied at the beginning of training, where the warm-up factor starts from 0.001 and lasts for 2000 steps. For optimization, we follow the AdamW optimizer with a weight decay of 0.0001, and apply gradient clipping with a maximum L2-norm of 0.1. All experiments strictly follow this unified configuration unless otherwise specified.

6.2. Loss Function

The overall training objective of our model follows DETR-style detectors and is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{dn}} + \lambda\,\mathcal{L}_{\text{mask}},$$

where $\mathcal{L}_{\text{cls}}$ is the IoU-aware classification loss (as in RT-DETR [61]), $\mathcal{L}_{\text{box}}$ combines L_1 loss and generalized IoU loss for bounding box regression, and $\mathcal{L}_{\text{dn}}$ is the denoising loss for query denoising [17]. We set $\lambda = 4$ to balance the mask loss with the other terms.

The term $\mathcal{L}_{\text{mask}}$ is a *Focal Tversky Loss*, originally proposed by Abraham and Khan [1], which extends the Tversky index to emphasize difficult pixels while balancing false positives and false negatives. Its definition is based on the Tversky index T:

$$T = \frac{\text{TP}}{\text{TP} + \alpha\,\text{FP} + \beta\,\text{FN}},$$

and the focal Tversky loss is then

$$\mathcal{L}_{\text{mask}} = (1 - T)^{\gamma},$$

where α, β adjust the trade-off between false positives and false negatives, and γ focuses training on hard (misclassified) pixels. In the text, we use the default settings from the original paper. We apply this mask loss to supervise the foreground mask prediction in the *Irrelevant Background Smoothing* module. At training time, the network predicts a binary mask M, and $\mathcal{L}_{\text{mask}}$ encourages M to correctly discriminate foreground and background regions. By doing so, it strengthens the learning of true foreground (target) regions, reduces background interference, and yields more reliable mask guidance for background smoothing.

Overall, the Focal Tversky-based mask loss offers a robust and balanced supervision for the binary mask, which is crucial in settings with highly imbalanced foreground and background pixels.

6.3. Details of the frequency analysis in Fig. 1

To investigate the intrinsic frequency characteristics of objects across different scales, we perform a comprehensive radial frequency–band analysis on both modalities. Following the octave-based partitioning strategy [34], we divide the normalized radial frequency domain into three octave-style bands: a low-frequency band $[0, \frac{1}{8})$, a mid-frequency band $[\frac{1}{8}, \frac{1}{4})$, and a high-frequency band $[\frac{1}{4}, \frac{1}{2}]$, where $\frac{1}{2}$ corresponds to the Nyquist limit. This octave-based division allows us to systematically separate coarse, intermediate, and fine structural information in the frequency domain.

For each annotated instance in the dataset, we first crop the object region according to its bounding box and classify its scale into five categories (*tiny*, *extremely_small*, *small*, *medium*, *large*) following the official scale definitions provided in the RGBT-Tiny benchmark. To suppress boundary-induced spectral leakage, the cropped patch is enlarged into a square using reflection padding, which reduces artificial edge discontinuities, and further expanded to the nearest power-of-two resolution, improving frequency resolution and numerical stability of the FFT. Before applying the Fourier Transform, we remove the DC component through mean subtraction to eliminate global brightness bias and apply a 2D Hann window, which attenuates edge artifacts; the window power is explicitly compensated during the computation of the power spectral density (PSD) to preserve the correct relative energy distribution.

We then perform a 2D FFT and compute the radially distributed frequency magnitude using an FFT-shifted isotropic radial frequency grid, ensuring a consistent mapping from spatial to radial frequency. For each object patch,

the PSD is integrated within the three octave-based bands, and the resulting band energies are normalized to form a per-instance percentage distribution that sums to 1. Finally, all instances are aggregated within each RGBT-Tiny scale category to obtain the average low-, mid-, and high-frequency energy ratios.

These design choices—including reflection padding, Hann windowing, power compensation, and per-patch normalization—effectively mitigate artifacts introduced by cropping and windowing, allowing a fair and unbiased estimation of the inherent radial frequency composition of object regions across scales.

6.4. More Experiments

Ablation study in terms of model complexity. As shown in Table 7, incorporating DFCA and SSE introduces only a marginal increase in model complexity. Specifically, enabling both modules increases the parameter count by merely 0.5M and adds approximately 2.3G FLOPs compared with the baseline configuration, reflecting less than a 3% overhead. However, despite this minimal computational cost, our method achieves the substantial performance gains reported in the main paper. Therefore, the slight increase in complexity is well justified by the significant improvement in detection accuracy, demonstrating that the proposed components offer an excellent trade-off between efficiency and effectiveness.

DFCA	SSE	Params (M)	FLOPs (G)
✗	✗	85.0	148.0
✓	✗	85.3	149.3
✓	✓	85.5	150.3

Table 7. Ablation study in terms of model complexity. DFCA and SSE indicate whether each submodule is enabled (✓) or disabled (✗). Params: number of parameters in millions (M), FLOPs: floating-point operations in gigaflops (G).

Complexity and efficiency comparison with other methods. As summarized in Table 8, DyFCLT maintains a moderate parameter size (85.3M) and a relatively high computational cost (150.3 GFLOPs). However, when compared with several strong competitors such as C2Former (118.7M, 145.5 GFLOPs), QFDet (60.2M, 162.9 GFLOPs), and IM-CMDET (105.0M, 133.9 GFLOPs), our method achieves a noticeably better balance between complexity and efficiency. Specifically, although DyFCLT operates with similar or even higher FLOPs than these models, it achieves a faster inference speed of 24.5 FPS, outperforming C2Former (20.0 FPS), QFDet (22.7 FPS), and IM-CMDET (19.3 FPS). This indicates that the proposed frequency-decoupled and lightweight interaction designs introduce only a mild computational burden while preserving practical runtime efficiency. Moreover, despite operating in a comparable computational regime, DyFCLT consistently surpasses all competing methods in detection accuracy (as reported in the main results), demonstrating that the additional FLOPs are effectively utilized. Together, these observations show that DyFCLT provides a favorable efficiency–accuracy trade-off: its runtime remains competitive among high-performing models, while its detection performance exceeds all existing baselines, highlighting its suitability for real-world RGBT tiny object detection scenarios.

Method	Params (M)	GFLOPs (G)	FPS
Faster R-CNN*	97.2	111.8	30.6
Cascade R-CNN*	92.6	113.8	26.1
RetinaNet*	59.8	99.0	**33.0**
FCOS*	78.6	<u>94.4</u>	<u>32.0</u>
ATSS*	<u>59.5</u>	97.5	31.3
HRfuser	**48.8**	**52.9**	11.5
TINet	88.8	108.2	26.4
C2Former	118.69	145.5	20.0
QFDet	60.2	162.9	22.7
IM-CMDET	105.0	133.9	19.3
DyFCLT (Ours)	85.3	150.3	24.5

Table 8. Model complexity and efficiency comparison on RGBT-DronePerson. * indicates that the method is modified to the RGBT dual-stream baseline. Note that all experiments are conducted on an RTX 3090.

6.5. More Visualizations

Visual analysis of DFCA. To further verify the effectiveness of our Dynamic Frequency-Decoupled Cross-Modal Attention (DFCA), we visualize the normalized frequency spectra of the same layer before and after applying DFCA across different input instances, as shown in Figure 7. From the visualization, it can be observed that the frequency responses exhibit a clear stratified structure after DFCA processing, which validates the effectiveness of our radial frequency-band decomposition design. The red dashed boxes in the figure indicate the approximate boundaries of these frequency layers. Notably, different inputs yield distinct frequency distributions after DFCA: for Input 1, the lowest-frequency band (closest to the center) shows strengthened responses after DFCA, whereas Input 2 displays the opposite trend. Moreover, the resulting frequency-layer patterns of Input 1 and Input 2 differ significantly from each other. These observations demonstrate that DFCA dynamically adjusts its frequency-decoupling behavior according to the characteristics of each input, enabling adaptive enhancement or suppression of specific frequency components. This adaptive frequency modeling allows DFCA to better preserve informative structures while mitigating redundant or noisy frequency responses, ultimately contribut-

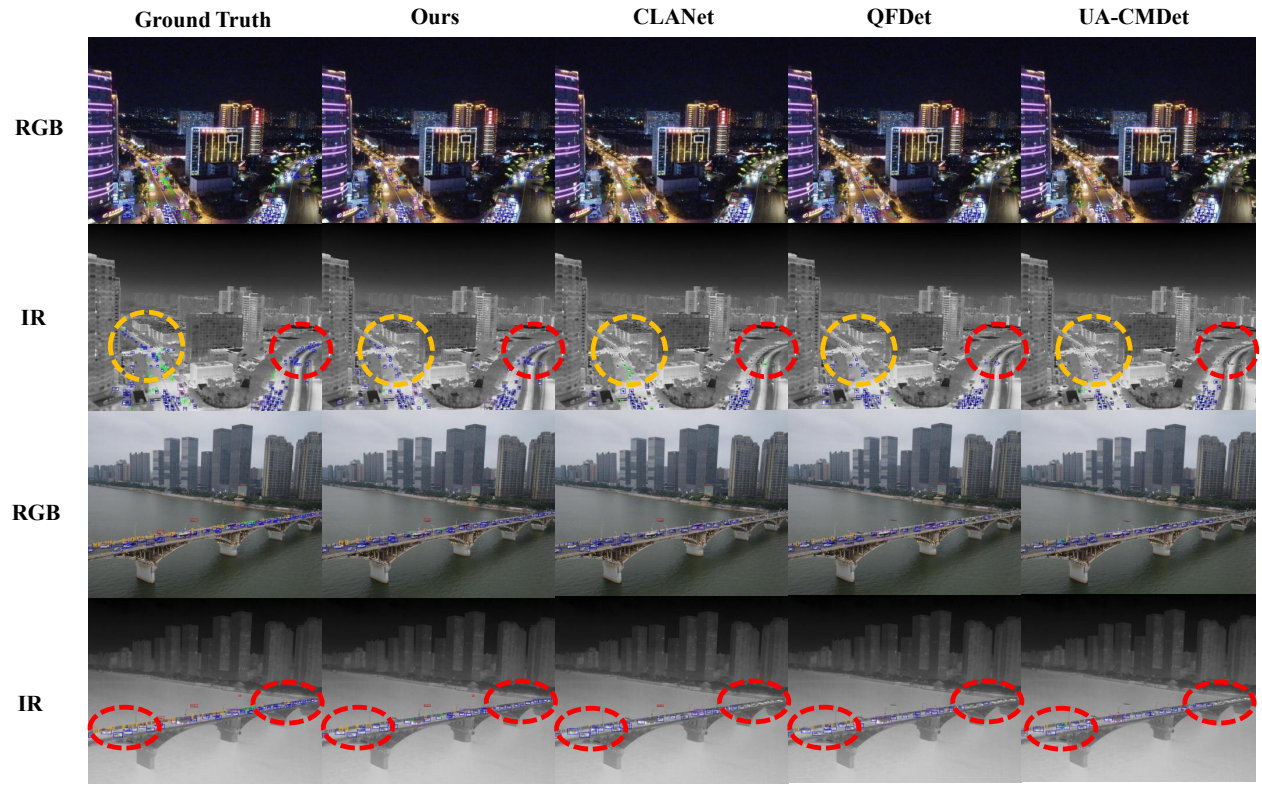

Figure 6. Comparison results with other RGB-T methods on RGBT-Tiny. The categories within the blue, purple, green, and yellow boxes in the figure correspond to *car*, *bus*, *cyclist*, and *pedestrian*, respectively.

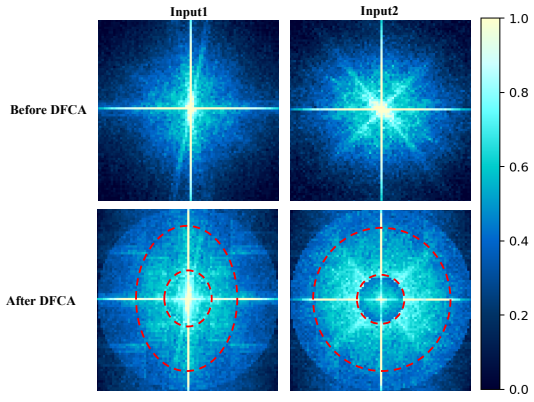

Figure 7. Visualization of the normalized frequency spectra before and after applying our DFCA. The red dashed boxes highlight the approximate boundaries of the stratified frequency layers formed by DFCA's radial frequency-band decomposition.

ing to more discriminative multimodal feature representations.

More qualitative comparisons. Figure 6 presents additional qualitative comparisons, where each column from left to right corresponds to: *Ground Truth*, *Ours*, *CLANet*, *QFDet*, and *UA-CMDet*. The red dashed boxes highlight regions with severe missed detections, typically containing extremely tiny targets or targets heavily degraded by low illumination, occlusion, or cluttered backgrounds. Meanwhile, the yellow dashed boxes indicate areas where both missed detections and false positives frequently occur. As illustrated in the figure, the compared methods struggle to accurately localize tiny or heavily occluded objects, often overlooking small pedestrians, distant vehicles, or producing spurious responses on background structures. In contrast, our method consistently produces clearer and more complete detections, with significantly fewer false alarms in cluttered regions. These visual results further confirm the robustness and superior discriminative capability of our approach for challenging tiny object detection scenarios.