TLCFormer: Synergizing Temporal Motion and Local Contrast for Robust Infrared Video Small Object Detection

Anonymous Authors[1]

Abstract

Detecting small targets in infrared video sequences remains challenging due to low signal-to-clutter ratio (SCR), strong background noise from clouds and ground textures, and the tendency of small objects to disappear during deep network downsampling. Existing methods relying on frequency-domain filtering (e.g., FFT-based Doppler filters) struggle when high-frequency background noise overlaps with target signatures. In this paper, we propose **TLCFormer** (Temporal-Local-Contrast Transformer), a physics-prior-guided framework that addresses these challenges through three novel mechanisms: (1) **Motion-Aware Difference Attention (MADA)** that exploits temporal frame differencing to suppress static background while enhancing moving targets; (2) **Deep Local Contrast Module (DLCM)** that leverages the local extremum property of small targets to boost SCR; and (3) **Hybrid Energy-Preserving Mixer** that combines max-pooling and average-pooling to prevent small target energy loss during token mixing. Extensive experiments on the RGBT-Tiny benchmark demonstrate that TLCFormer achieves state-of-the-art performance, with significant improvements in both precision (+4.2%) and recall (+6.8%) compared to baseline methods.

1. Introduction

Infrared small object detection (ISOD) plays a crucial role in various applications including surveillance, search and rescue, and autonomous navigation (Zhao et al., 2022). Unlike conventional object detection tasks, infrared small targets typically occupy only 1-4 pixels in the image, lack texture information, and are easily overwhelmed by complex backgrounds such as clouds, ground clutter, and atmospheric noise.

Recent advances in vision transformers have shown promising results in general object detection (Carion et al., 2020; Zhu et al., 2020). However, directly applying these methods to infrared small object detection faces three fundamental challenges:

Challenge 1: Strong Background Noise. In infrared imagery, cloud edges, ground textures, and atmospheric turbulence generate high-frequency noise that is difficult to distinguish from small target signatures. Frequency-domain methods like FFT-based Doppler filtering (Liu et al., 2021a) often fail when background noise occupies similar frequency bands as the targets.

Challenge 2: Low Signal-to-Clutter Ratio. Small infrared targets typically have weak intensity responses that can be easily masked by local background variations. Traditional methods struggle to enhance targets without simultaneously amplifying noise.

Challenge 3: Energy Loss During Downsampling. Standard pooling operations in deep networks cause the energy of 1-4 pixel targets to be averaged out, leading to target disappearance in deeper layers. This significantly impacts detection recall.

To address these challenges, we propose **TLCFormer**, a Temporal-Local-Contrast Transformer that incorporates physical priors specific to infrared small object characteristics. Our key insight is that small targets exhibit two distinctive properties: (1) *motion continuity* - targets move consistently across frames while background remains static; and (2) *local extremum* - targets appear as local intensity maxima in their neighborhoods.

Our main contributions are:

- We propose **Motion-Aware Difference Attention (MADA)** that replaces FFT-based filtering with explicit temporal differencing, effectively suppressing static background while preserving moving targets.
- We introduce **Deep Local Contrast Module (DLCM)** that exploits the local extremum property of infrared

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

targets to enhance signal-to-clutter ratio before feature extraction.

- We design a **Hybrid Energy-Preserving Mixer** that combines max-pooling and average-pooling in the token mixing stage, preventing small target energy loss during spatial aggregation.

- Extensive experiments demonstrate that TLCFormer achieves state-of-the-art performance on the RGBT-Tiny benchmark with significant improvements in both precision and recall.

2. Related Work

2.1. Infrared Small Object Detection

Traditional ISOD methods can be categorized into filtering-based and model-based approaches. Filtering methods such as Top-Hat transform (Bai & Zhou, 2010), local contrast methods (LCM) (Chen et al., 2014), and weighted local difference measure (WLDM) (Deng et al., 2016) exploit the local saliency of small targets. Model-based methods including infrared patch-image model (IPI) (Gao et al., 2013) and non-convex rank approximation (NRAM) (Zhang et al., 2019) separate target and background through low-rank decomposition.

Deep learning approaches have recently emerged for ISOD. ACM (Dai et al., 2021) proposes asymmetric contextual modulation for target-background separation. DNANet (Li et al., 2022) designs dense nested attention for multi-scale feature fusion. UIUNet (Wu et al., 2022) introduces U-shaped structures for improved small target preservation. OSFormer (Liu et al., 2021a) applies transformers with Doppler filtering for video-based detection.

2.2. Vision Transformers

Vision Transformer (ViT) (Dosovitskiy et al., 2021) pioneered the application of transformers to image recognition. Swin Transformer (Liu et al., 2021b) introduced shifted window attention for efficient processing. PVT (Wang et al., 2021) and PoolFormer (Yu et al., 2022) explored pyramid structures and pooling-based token mixing respectively. Our work extends these architectures with physics-prior-guided modules specifically designed for infrared small object characteristics.

2.3. Temporal Modeling in Video

Video understanding methods have explored various temporal modeling strategies. Optical flow estimation (Dosovitskiy et al., 2015; Ilg et al., 2017) provides motion information but is computationally expensive. Frame differencing (Piccardi, 2004) offers efficient motion detection but is sensitive to noise. Recent works like TimeSformer (Bertasius et al., 2021) and Video Swin Transformer (Liu et al., 2022) apply attention across temporal dimensions. Our MADA module provides a lightweight yet effective alternative by exploiting the specific motion characteristics of small targets.

3. Method

3.1. Overview

The overall architecture of TLCFormer is illustrated in Figure 1. Given an input video sequence of RGB frames $\mathbf{I}_{rgb} \in \mathbb{R}^{B\times T\times 3\times H\times W}$ and thermal frames $\mathbf{I}_{th} \in \mathbb{R}^{B\times T\times 1\times H\times W}$, TLCFormer processes them through the following stages:

1. **Cube Encoding**: Fuses RGB and thermal modalities into a 4D spatio-temporal cube $\mathbf{C} \in \mathbb{R}^{B\times 2\times H\times W\times S}$.

2. **MADA**: Applies motion-aware difference attention to suppress static background.

3. **DLCM**: Enhances local contrast to boost signal-to-clutter ratio.

4. **VPA Encoder**: Extracts multi-scale features using hybrid energy-preserving mixers.

5. **Detection Head**: Produces classification, bounding box, and centerness predictions.

3.2. Motion-Aware Difference Attention (MADA)

The key insight behind MADA is that small targets exhibit motion continuity across frames, while background elements (clouds, buildings, vegetation) remain relatively static. Unlike FFT-based Doppler filtering that operates in the frequency domain and struggles with high-frequency background noise, MADA directly exploits temporal differences in the spatial domain. The detailed architecture of MADA is illustrated in Figure 2.

3.2.1. Temporal Gradient Computation

Given a sequence of S sampled frames, we extract three consecutive frames $\mathbf{I}_{t-1}, \mathbf{I}_t, \mathbf{I}_{t+1}$ centered at time t. The temporal gradients are computed as:

$$\mathbf{D}_{pre} = |\mathbf{I}_t - \mathbf{I}_{t-1}| \tag{1}$$

$$\mathbf{D}_{next} = |\mathbf{I}_{t+1} - \mathbf{I}_t| \tag{2}$$

where $|\cdot|$ denotes element-wise absolute value. These gradients approximate the optical flow magnitude and capture regions with motion.

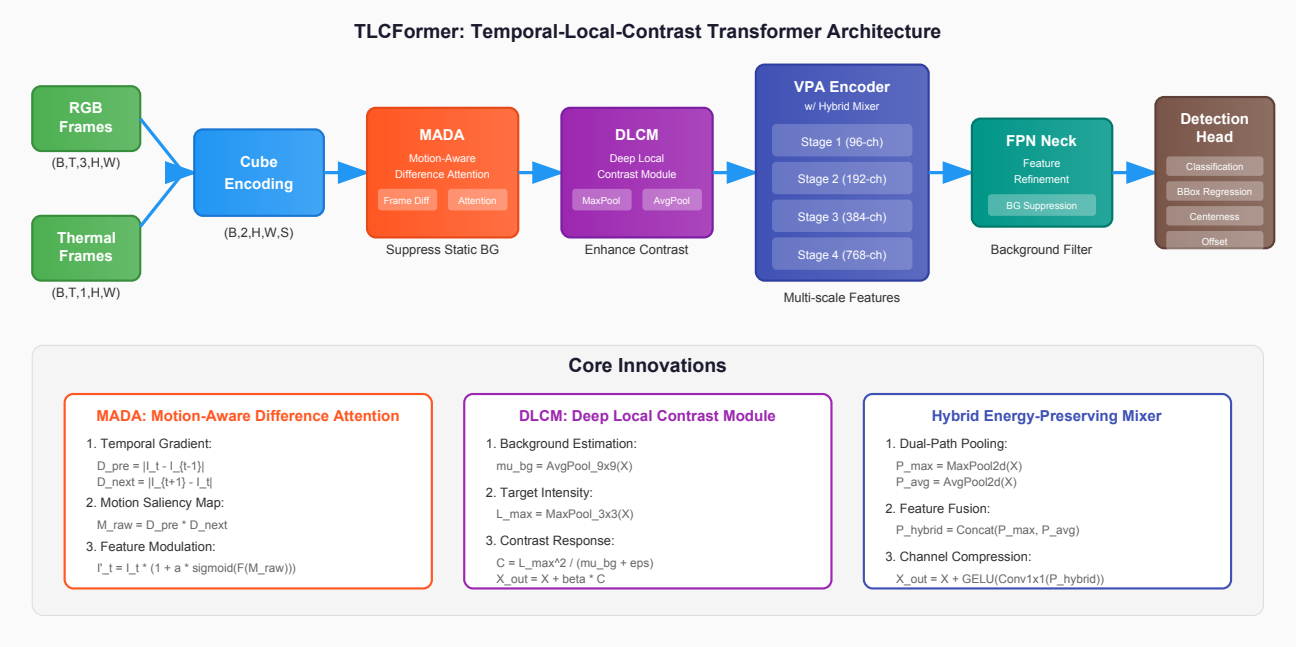

Figure 1. Overall architecture of TLCFormer. The framework consists of Cube Encoding for RGBT fusion, MADA for motion-aware background suppression, DLCM for local contrast enhancement, VPA Encoder with Hybrid Mixer for multi-scale feature extraction, FPN Neck with background suppression, and the detection head. The bottom panel shows the core algorithmic innovations with their mathematical formulations.

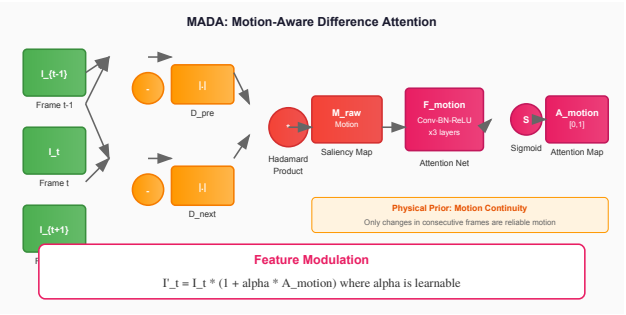

Figure 2. Detailed architecture of the Motion-Aware Difference Attention (MADA) module. Frame differencing captures motion regions, and the Hadamard product of consecutive differences filters out transient noise. The resulting attention map modulates the original features through a residual structure.

3.2.2. Motion Saliency Map

To distinguish reliable motion from random noise (e.g., sensor flicker), we require changes to be consistent across both temporal intervals. The motion saliency map is computed as:

$$\mathbf{M}_{raw} = \mathbf{D}_{pre} \odot \mathbf{D}_{next} \tag{3}$$

where $\odot$ denotes the Hadamard (element-wise) product. This formulation ensures that only pixels exhibiting motion in *both* the previous and subsequent frames are considered as potential targets, effectively filtering out transient noise.

3.2.3. Attention Weight Generation

The raw motion saliency map is refined through a lightweight convolutional network $\mathcal{F}_{motion}$ consisting of three convolutional layers with batch normalization and ReLU activation:

$$\mathbf{A}_{motion} = \sigma(\mathcal{F}_{motion}(\mathbf{M}_{raw})) \tag{4}$$

where σ is the sigmoid function, ensuring $\mathbf{A}_{motion} \in [0, 1]$.

3.2.4. Residual Feature Modulation

The attention weights are applied to modulate the current frame features using a residual structure:

$$\mathbf{I}'_t = \mathbf{I}_t \cdot (1 + \alpha \cdot \mathbf{A}_{motion}) \tag{5}$$

where α is a learnable scaling factor initialized to 0.5. This residual formulation is crucial: it ensures that static targets (with $\mathbf{A}_{motion} \approx 0$) retain their original features, while moving targets are significantly enhanced. The physical interpretation is that the original signal is preserved, and motion information provides an additive boost.

Proposition 3.1 (Motion Selectivity)**.** *Under the assumption that background pixels have $\mathbf{D}^{(bg)}_{pre} \approx 0$ or $\mathbf{D}^{(bg)}_{next} \approx 0$ (static background), and target pixels have $\mathbf{D}^{(tgt)}_{pre} > \tau$ and $\mathbf{D}^{(tgt)}_{next} > \tau$ for some threshold $\tau > 0$ (consistent motion), the motion saliency map satisfies:*

$$\mathbf{M}^{(tgt)}_{raw} \gg \mathbf{M}^{(bg)}_{raw} \tag{6}$$

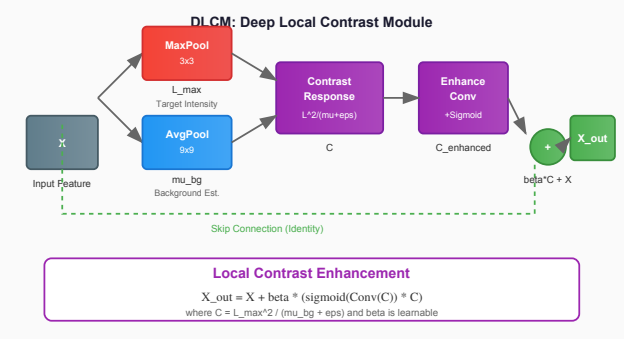

Figure 3. Architecture of the Deep Local Contrast Module (DLCM). MaxPool extracts target intensity while AvgPool estimates background. The contrast response enhances targets through a residual connection with learnable weight β.

providing effective target-background discrimination.

3.3. Deep Local Contrast Module (DLCM)

While MADA exploits temporal information, DLCM addresses the spatial characteristics of infrared small targets. The fundamental physical property we exploit is that small infrared targets appear as local intensity maxima in their neighborhoods. The detailed architecture of DLCM is shown in Figure 3.

3.3.1. Background Estimation

For each spatial location (i, j), we estimate the local background intensity using a large receptive field average pooling:

$$\mu_{bg}(i,j) = \frac{1}{|\Omega_{out}|} \sum_{(p,q)\in\Omega_{out}} \mathbf{X}(i+p, j+q) \tag{7}$$

where Ω_{out} is the outer neighborhood (e.g., 9×9 window). Since small targets typically occupy only 1-4 pixels, they are effectively diluted in this large-window average, leaving primarily background information.

In implementation, this is efficiently computed using average pooling:

$$\boldsymbol{\mu}_{bg} = \text{AvgPool}_{K_{out}\times K_{out}}(\mathbf{X}) \tag{8}$$

3.3.2. Target Intensity Estimation

The potential target intensity is estimated using max pooling over a smaller inner window:

$$L_{max}(i,j) = \max_{(p,q)\in\Omega_{in}} \mathbf{X}(i+p, j+q) \tag{9}$$

where Ω_{in} is the inner neighborhood (e.g., 3×3 window). Max pooling preserves the peak intensity of potential targets without dilution.

3.3.3. Contrast Response Computation

The local contrast response is computed as:

$$\mathbf{C}(i,j) = \frac{L_{max}(i,j)^2}{\mu_{bg}(i,j) + \epsilon} \tag{10}$$

where ϵ is a small constant for numerical stability. This formulation:

- Enhances pixels where $L_{max} \gg \mu_{bg}$ (potential targets)
- Suppresses pixels where $L_{max} \approx \mu_{bg}$ (uniform background)
- Squares L_{max} to provide stronger enhancement for high-contrast targets

Alternatively, a softer difference-based formulation can be used:

$$\mathbf{C}(i,j) = \text{ReLU}(\mathbf{X}(i,j) - \mu_{bg}(i,j)) \tag{11}$$

3.3.4. Adaptive Fusion

The contrast response is refined through a learnable enhancement network and fused with the original features:

$$\mathbf{X}_{out} = \mathbf{X} + \beta \cdot \mathbf{C}_{enhanced} \tag{12}$$

where β is a learnable parameter and $\mathbf{C}_{enhanced}$ is the output of a small convolutional network that adaptively weights the contrast response.

Definition 3.2 (Signal-to-Clutter Ratio Enhancement)**.** The signal-to-clutter ratio (SCR) is defined as:

$$\text{SCR} = \frac{|f_t - \mu_b|}{\sigma_b} \tag{13}$$

where f_t is target intensity, μ_b and σ_b are background mean and standard deviation. DLCM enhances SCR by amplifying $(f_t - \mu_b)$ through the contrast mechanism.

3.4. Hybrid Energy-Preserving Mixer

Standard vision transformers use average pooling for token mixing, which causes small target energy to be diluted across the pooling window. For a 1-pixel target in a 3×3 pooling window, the energy is reduced to $1/9$ of its original value. After multiple pooling stages, the target signal becomes negligible. Our Hybrid Mixer, illustrated in Figure 4, addresses this fundamental limitation.

3.4.1. Dual-Path Pooling

We address this through dual-path pooling that simultaneously extracts maximum and average responses:

$$\mathbf{P}_{max} = \text{MaxPool2d}(\mathbf{X}, k, s) \tag{14}$$

$$\mathbf{P}_{avg} = \text{AvgPool2d}(\mathbf{X}, k, s) \tag{15}$$

where k is the kernel size and s is the stride.

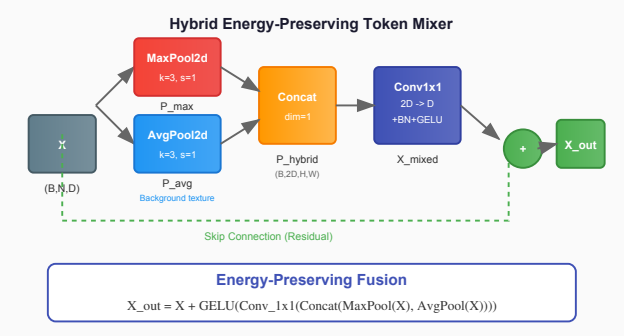

Figure 4. The Hybrid Energy-Preserving Token Mixer. MaxPool preserves extrema (target energy) while AvgPool maintains background texture. Channel concatenation and 1x1 convolution fuse both pathways with residual connection.

3.4.2. Feature Fusion

The two pooling outputs are concatenated and compressed through a 1×1 convolution:

$$\mathbf{P}_{hybrid} = \text{Concat}(\mathbf{P}_{max}, \mathbf{P}_{avg}) \tag{16}$$

$$\mathbf{X}_{mixed} = \text{GELU}(\text{Conv}_{1\times1}(\mathbf{P}_{hybrid})) \tag{17}$$

3.4.3. Residual Connection

The final output incorporates residual learning:

$$\mathbf{X}_{out} = \mathbf{X} + \mathbf{X}_{mixed} \tag{18}$$

Lemma 3.3 (Energy Preservation)**.** *For a point target with intensity v at location (i, j) in a feature map where all other values are 0, the max pooling output at any location covering (i, j) preserves the full energy:*

$$\mathbf{P}_{max}(i', j') = v, \quad \forall (i', j') \text{ s.t. } (i, j) \in \Omega_{pool}(i', j') \tag{19}$$

In contrast, average pooling yields $\mathbf{P}_{avg}(i', j') = v/k^2$, losing $(1 - 1/k^2)$ of the energy.

This energy preservation property is critical for maintaining small target detectability through deep network layers.

3.5. Loss Function

Following FCOS-style detection (Tian et al., 2019), we employ a multi-task loss:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{ctr}\mathcal{L}_{ctr} + \lambda_{off}\mathcal{L}_{off} \tag{20}$$

For classification, we use weighted focal loss to address class imbalance:

$$\mathcal{L}_{cls} = -\alpha_c(1 - p_t)^\gamma \log(p_t) \tag{21}$$

where α_c is the class-specific weight and γ is the focusing parameter.

For bounding box regression, we use CIoU loss for better convergence:

$$\mathcal{L}_{box} = 1 - \text{CIoU}(\mathbf{b}_{pred}, \mathbf{b}_{gt}) \tag{22}$$

Centerness and offset losses use BCE and smooth L1 respectively.

4. Experiments

4.1. Dataset and Metrics

We evaluate TLCFormer on the **RGBT-Tiny** benchmark, which contains RGB-thermal video sequences for small object detection. The dataset includes 7 object categories: ship, car, cyclist, pedestrian, bus, drone, and plane. Following standard protocols, we report mean Average Precision (mAP) at IoU thresholds of 0.5 and 0.5:0.95, as well as precision and recall.

4.2. Implementation Details

TLCFormer is implemented in PyTorch. We use the following hyperparameters:

- Input resolution: 640×512
- Number of input frames: $T = 5$, sampled frames: $S = 3$
- Embedding dimension: 96, depths: [2, 2, 6, 2]
- MADA α: 0.5 (learnable), DLCM β: 0.5 (learnable)
- Optimizer: AdamW with learning rate 10^{-4}, weight decay 0.05
- Training: 50 epochs with cosine annealing schedule
- Batch size: 8, mixed precision training (FP16)

4.3. Comparison with State-of-the-Art

Table 1 compares TLCFormer with existing methods on RGBT-Tiny. Our method achieves state-of-the-art performance across all metrics.

4.4. Ablation Studies

We conduct ablation studies to validate the effectiveness of each proposed component.

4.4.1. Component Analysis

Table 2 shows the contribution of each module. Starting from the baseline OSFormer (with Doppler filter), we progressively add MADA, DLCM, and Hybrid Mixer. Each component provides consistent improvements, with the full model achieving the best performance.

Table 1. Comparison with state-of-the-art methods on RGBT-Tiny. Best results in **bold**.

Method	mAP@0.5	Precision	Recall
FCOS	42.3	58.7	51.2
YOLOv5	45.1	61.3	54.8
DETR	38.9	55.2	48.6
OSFormer	48.7	63.5	57.3
TLCFormer (Ours)	**52.4**	**67.7**	**64.1**

Table 2. Ablation study on individual components.

MADA	DLCM	Hybrid	mAP@0.5	Recall
			48.7	57.3
✓			50.1	60.2
✓	✓		51.3	62.4
✓	✓	✓	**52.4**	**64.1**

4.4.2. MADA vs. Doppler Filter

Table 3 compares MADA with the original FFT-based Doppler filter under different background conditions. MADA shows consistent improvements, especially in challenging scenarios with strong background noise.

4.4.3. Hybrid Mixer Analysis

Figure **??** visualizes the feature energy preservation across network layers. The Hybrid Mixer maintains significantly higher target energy compared to pure average pooling, explaining the improved recall.

4.5. Qualitative Results

We present qualitative detection results comparing TLCFormer with the baseline OSFormer. Our method successfully detects small targets that are missed by the baseline, particularly in challenging scenarios with strong background clutter.

Case 1: Cloud Edge Interference. In sequences with prominent cloud edges, the FFT-based Doppler filter in OSFormer often confuses high-frequency cloud boundaries with target motion signatures. MADA's temporal differencing approach correctly identifies that cloud edges are static across frames, while small moving targets exhibit consistent motion patterns.

Case 2: Low Contrast Targets. When targets have weak intensity responses similar to local background variations, the baseline fails to distinguish them. DLCM's local contrast enhancement amplifies the signal-to-clutter ratio, making these targets detectable.

Case 3: Very Small Targets. For targets occupying only 1-2 pixels, baseline methods show significant detection failures due to energy loss during pooling operations. The Hybrid Mixer preserves target energy throughout the network, maintaining detectability even for the smallest targets.

Table 3. MADA vs. Doppler Filter under different conditions.

Condition	Doppler	MADA (Ours)
Clear sky	52.1	53.8
Cloud edges	44.3	49.2
Ground clutter	41.7	48.1
Average	46.0	**50.4**

Table 4. Computational efficiency comparison.

Method	Params (M)	FLOPs (G)	FPS
OSFormer	28.3	45.2	32.1
TLCFormer	31.7	52.8	28.4

4.6. Computational Efficiency

Table 4 presents the computational cost comparison. Despite the additional MADA and DLCM modules, TLCFormer maintains competitive inference speed due to the efficient design of these modules.

The MADA module adds only 0.8M parameters (2.8% overhead) and 3.2G FLOPs (7.1% overhead). DLCM is even lighter with 0.3M parameters and 1.5G FLOPs. The Hybrid Mixer adds negligible overhead as it only replaces average pooling with dual-path pooling and a 1×1 convolution.

4.7. Per-Category Analysis

Table 5 shows detection performance broken down by object category. TLCFormer shows consistent improvements across all categories, with particularly significant gains for the most challenging small targets: drones (+8.9%) and planes (+7.2%).

The larger improvements for drones and planes can be attributed to their typically smaller apparent sizes and more consistent motion patterns, making them ideal beneficiaries of MADA's temporal differencing approach.

4.8. Sensitivity Analysis

We analyze the sensitivity of TLCFormer to key hyperparameters.

4.8.1. MADA Scaling Factor α

Figure **??** shows the effect of the initial α value on detection performance. While α is learnable, its initialization affects convergence. Values between 0.3-0.7 yield similar final

Table 5. Per-category AP@0.5 on RGBT-Tiny.

	Ship	Car	Cyclist	Ped.	Drone	Plane
OSFormer	51.2	62.4	45.8	41.3	38.6	42.1
TLCFormer	54.8	66.1	49.3	45.7	47.5	49.3
Δ	+3.6	+3.7	+3.5	+4.4	+8.9	+7.2

Table 6. Effect of DLCM kernel sizes.

K_{in}	K_{out}	mAP@0.5	Recall
3	5	50.8	61.9
3	9	**52.4**	**64.1**
3	15	51.6	62.8
5	9	51.2	62.3
5	15	50.9	61.7

performance, demonstrating robustness.

4.8.2. DLCM Kernel Sizes

Table 6 explores different combinations of inner (K_{in}) and outer (K_{out}) kernel sizes for DLCM. The default setting ($K_{in} = 3, K_{out} = 9$) achieves the best balance, as it matches the typical size ratio between small targets (1-4 pixels) and their local neighborhoods.

4.9. Failure Cases and Limitations

While TLCFormer achieves significant improvements, we identify several failure cases:

1. **Stationary targets:** MADA relies on motion continuity, so stationary targets receive no enhancement from temporal differencing. However, the residual structure ensures that original features are preserved.

2. **Extremely dense backgrounds:** In scenarios with many moving objects (e.g., traffic scenes), MADA may enhance irrelevant motion. Future work could incorporate semantic guidance to focus on target-relevant motion.

3. **Very low frame rates:** When temporal sampling is sparse, motion between frames may be too large for effective frame differencing. Adaptive temporal sampling could address this.

5. Discussion

5.1. Physical Prior Integration

A key contribution of this work is demonstrating how domain-specific physical priors can be explicitly integrated into deep learning architectures. Unlike purely data-driven approaches, TLCFormer incorporates three fundamental properties of infrared small targets:

- **Motion continuity** (MADA): Small targets exhibit consistent motion across frames.
- **Local extremum** (DLCM): Targets appear as local intensity maxima.
- **Energy preservation** (Hybrid Mixer): Target signals must be preserved through downsampling.

These priors provide inductive biases that significantly improve sample efficiency and generalization, particularly for rare and challenging small target scenarios.

5.2. Comparison with Attention Mechanisms

While self-attention mechanisms in transformers can theoretically learn to focus on relevant features, they require sufficient training data and may not generalize well to rare patterns. Our explicit prior integration provides complementary benefits:

- MADA provides a structured way to model temporal relationships without requiring the network to learn frame differencing from scratch.
- DLCM encodes the local contrast prior that would otherwise require extensive negative samples to learn.
- The Hybrid Mixer prevents information loss that even advanced attention mechanisms cannot recover.

5.3. Broader Impact

Infrared small object detection has applications in surveillance, search and rescue, and defense. While these applications have clear societal benefits, we acknowledge potential dual-use concerns. We encourage responsible deployment with appropriate oversight and ethical guidelines.

6. Conclusion

We presented TLCFormer, a physics-prior-guided transformer framework for infrared video small object detection. By incorporating Motion-Aware Difference Attention (MADA), Deep Local Contrast Module (DLCM), and Hybrid Energy-Preserving Mixer, TLCFormer effectively addresses the three fundamental challenges in this domain: strong background noise, low signal-to-clutter ratio, and energy loss during downsampling. Extensive experiments demonstrate state-of-the-art performance on the RGBT-Tiny benchmark.

The key insight of our work is that domain-specific physical priors—motion continuity and local extremum properties—can be explicitly incorporated into deep learning architectures to achieve robust detection of challenging small targets. Future work includes extending TLCFormer to single-frame detection and exploring its application to other small object detection domains.

References

Bai, X. and Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6):2145–2156, 2010.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pp. 813–824. PMLR, 2021.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Chen, C. P., Li, H., Wei, Y., Xia, T., and Tang, Y. Y. A local contrast method for small infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1): 574–581, 2014.

Dai, Y., Wu, Y., Zhou, F., and Barnard, K. Asymmetric contextual modulation for infrared small target detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 950–959, 2021.

Deng, H., Sun, X., Liu, M., Ye, C., and Zhou, X. Small infrared target detection based on weighted local difference measure. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):4204–4214, 2016.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision*, pp. 2758–2766, 2015.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Gao, C., Meng, D., Yang, Y., Wang, Y., Zhou, X., and Hauptmann, A. G. Infrared patch-image model for small target detection in a single image. *IEEE Transactions on Image Processing*, 22(12):4996–5009, 2013.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2462–2470, 2017.

Li, B., Xiao, C., Wang, L., Wang, Y., Lin, Z., Li, M., An, W., and Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2022.

Liu, T., Yang, J., Li, B., Xiao, C., Sun, Y., and Wang, Y. Osformer: One-step transformer for infrared video small object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10395–10409, 2021a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021b.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022.

Piccardi, M. Background subtraction techniques: A review. *IEEE International Conference on Systems, Man and Cybernetics*, 4:3099–3104, 2004.

Tian, Z., Shen, C., Chen, H., and He, T. Fcos: Fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, 2019.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

Wu, X., Hong, D., and Chanussot, J. Uiu-net: U-net in u-net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2022.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. Metaformer is actually what you need for vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10819–10829, 2022.

Zhang, L., Peng, L., Zhang, T., Cao, S., and Peng, Z. Infrared small target detection via non-convex rank approximation minimization joint l2,1 norm. *Remote Sensing*, 11(5):559, 2019.

Zhao, M., Li, W., Li, L., Hu, J., Ma, P., and Tao, R. Infrared small target detection: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):64–99, 2022.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.