MA-DyNN: Modal-Adaptive Dynamic Neural Network for Crowd-Counting on Consumer Drones

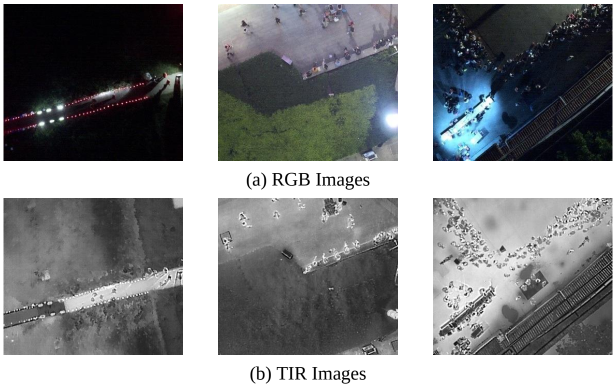
Hang Shen, *Member, IEEE,* Qi Liu, Yu Liu, Tianjing Wang, *Member, IEEE,* and Guangwei Bai

***Abstract*—Consumer drones are increasingly used for crowd-counting in complex environments; however, their deployment faces challenges from adverse external conditions such as low illumination and inclement weather, as well as inherent limitations like constrained onboard computational resources. To address these constraints, we present MA-DyNN (Modal-Adaptive Dynamic Neural Network), a lightweight and robust framework that dynamically adapts to varying modality conditions for accurate crowd counting. This framework employs an efficient single-stream architecture with specialized modal extractors to capture and integrate complementary information from both visible and thermal infrared (TIR) inputs. Based on the extracted modal features, we design a modality-adaptive gating mechanism to dynamically select the optimal modality based on environmental conditions, favoring visible imagery for inference efficiency in well-lit scenarios and leveraging TIR as auxiliary support under low-light or degraded conditions. To enhance robustness against sensor failure or missing modalities, we develop a density-aware modality converter that adds crowd density constraints to a cycle-consistent generative adversarial learning framework to generate high-fidelity TIR images. This enables consistent performance by aligning synthetic and real TIR-based counting outcomes through adversarial learning. Extensive experiments on DroneRGBT and RGBT datasets show that MA-DyNN achieves superior accuracy, generalization, and real-time performance compared to state-of-the-art multimodal baselines. Its inference acceleration performance approaches single-modality models without compromising the accuracy gains provided by multimodal learning**

***Index Terms*—Consumer drone, crowd counting, multimodal, dynamic neural networks.**

I. Introduction

The rapid advancements in consumer drone technology have significantly expanded their roles in real-time aerial analytics, particularly for latency-sensitive applications such as crowd counting, public safety monitoring, large-scale event management, disaster response, and urban population studies [1]–[3], where on-device inference is often preferred to ensure timely responsiveness. Compared to industrial-grade drones, consumer drones offer cost-effective and accessible solutions for data collection and analysis. Their high mobility, wide coverage, and flexible deployment facilitate rapid acquisition of aerial imagery, making them highly suitable for complex scenarios. Consequently, they have become indispensable tools for enabling efficient and scalable visual perception on edge platforms.


(a) RGB Images

(b) TIR Images

Fig. 1: Aerial images in low-light conditions.

Despite these advantages, aerial imagery captured by consumer drones inherently exhibits scale variations due to fluctuating flight altitudes and dynamic viewpoints. As shown in Fig. 1(a), the apparent size of individuals varies significantly between different regions of an image. To address this challenge, researchers have explored multiscale feature extraction strategies [4], [5], employing feature pyramid networks [6], [7] and expanded receptive fields [8] to enhance the ability to model targets across diverse scales. Beyond scale variations, real-world environmental conditions further complicate aerial crowd counting. Most existing methods often assume ideal conditions with clear visibility, allowing for capturing detailed appearance and texture features [9]. However, in practical deployment, adverse weather, overexposure, underexposure, and shadow occlusion can severely degrade RGB image quality [10], as illustrated in Fig. 1(a). Under such conditions, RGB-based counting methods struggle to accurately detect and identify valid targets. To mitigate these issues, TIR cameras have been integrated into consumer drones to capture surface thermal emissions imperceptible to the human eye. As depicted in Fig. 1(b), TIR images provide essential target-related information, such as location and spatial context, complementing missing details in RGB images [11]. However, TIR imaging may highlight non-human heat-emitting objects, such as lamp posts, leading to potential misidentifications. Consequently, multimodal crowd-counting combining RGB and TIR (RGB-T) offers improved robustness and reliability [12].

Manuscript received XX March 2025; revised XX XXX 2025; accepted XX XXX 2025. This work was supported in part by the National Natural Science Foundation of China under Grants 61502230 and 61501224, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20201357, and in part by the Six Talent Peaks Project in Jiangsu Province under Grant RJFW-020. (Corresponding author: Hang Shen)

Hang Shen, Qi Liu, Yu Liu, Tianjing Wang, and Guangwei Bai are with the College of Computer and Information Engineering (College of Artificial Intelligence), Nanjing Tech University, Nanjing, 211816, China. E-Mail: {hshen, qiliu, wangtianjing, bai}@njtech.edu.cn, liuyu_ligion@163.com

Leveraging the complementary strengths of RGB-T modalities requires effective fusion strategies, which have evolved from early-stage simple feature concatenation [13] to high-order feature interactions based on bilinear pooling [14]. More recently, attention mechanisms have been employed to adjust the significance of each modality based on contextual cues [15]. However, these fusion strategies may lead to computationally expensive multimodal models that are difficult to deploy on resource-constrained consumer drones. Furthermore, environmental factors such as lighting and weather variations influence the relative importance of each modality, making static fusion approaches less effective in varying conditions.

To overcome these limitations, dynamic neural networks (DyNNs) have emerged as a promising paradigm, improving computational efficiency and robustness by adaptively adjusting computational paths based on input data features. This dynamic inference is well-suited for consumer drones, where onboard processing power is limited, and real-time adaptation is critical. Representative approaches include early exiting [16], layer skipping [17], and selective activation in mixture-of-experts (MoE) [18]. By leveraging on-demand computation, DyNNs provide a flexible and efficient framework for multimodal fusion, enabling adaptive inference resource allocation while maintaining high performance on aerial platforms.

A. Challenging Issues and Related Works

Despite recent advances in RGB-T fusion and DyNN methods, several challenging issues remain in the practical deployment and online inference on consumer drones:

1) Lightweight multimodal feature alignment. The inherent spatial misalignment between RGB and TIR modalities makes directly applying traditional feature extraction impractical. A common strategy is to employ dual-stream architectures [19], where parallel branches extract features independently of each modality. For example, DSCDNet [20] introduces dual-stream feature extraction across spatial and frequency domains to enhance RGB-T object detection, while ADNet [21] uses an asymmetric dual-stream design combined with a feature interaction module to accommodate differences in information density between RGB and TIR data. Recent transformer-based methods, such as M3DETR [22], unify multiple point cloud representations with multi-scale features for 3D object detection, while vision-language models, such as Flamingo [23] and BLIP-2 [24], achieve sophisticated cross-modal reasoning through attention mechanisms. However, these approaches are computationally intensive, making them impractical for deployment on resource-constrained consumer drones where real-time inference is essential.

2) Modality adaptation in aerial reasoning. In drone-based perception under diverse environmental conditions, the relative importance of different modalities varies dynamically with spatiotemporal context. For instance, RGB data is more informative during daylight, while TIR is essential at night or in low-visibility weather. Most existing fusion methods treat all modalities equally throughout inference, lacking the ability to selectively activate or deactivate specific sensors based on current conditions. This leads to unnecessary computation and energy waste. While multimodal learning aims to exploit complementary information from multiple sources, prior work [25] has shown that fully processing all modalities can introduce redundancy and increase computational cost. This highlights the need for adaptive modality selection mechanisms that dynamically identify the most informative subset under resource constraints. Some methods, such as MFGNet [26] and CANNET [27], use attention mechanisms to weight modal contributions, yet they still rely on complete modal processing. In contrast, DyNN-based methods like D-gate [28] and GateNet [29] leverage gating strategies to route samples through different computation paths based on instance complexity, offering a viable solution for dynamic modality adaptation in aerial scenarios.

3) Robustness optimization of multimodal models. During dynamic deployment, multimodal data captured by onboard sensors is susceptible to quality degradation or even modality loss due to hardware limitations or environmental interference. Cross-modal generation has emerged as a key technique for reconstructing missing modalities [30]. For instance, Khan et al. [31] recently applied Pix2Pix GANs to generate synthetic TIR images from RGB inputs for crowd counting, demonstrating effectiveness in data augmentation but employing static multimodal fusion without considering dynamic environmental adaptation. Li et al. [32] further proposed an invertible neural network with reversible operations for lossless modality transformation. Although diffusion models [33] can produce highly realistic results, their iterative denoising process limits inference speed. However, many existing approaches lack task-specific constraints, such as enforcing spatial consistency in target distributions, making it difficult to preserve structural and textural features during reconstruction.

B. Contributions and Organization

In this paper, we propose a Modal-Adaptive DyNN (MA-DyNN) framework that enhances adaptability, robustness, and computational efficiency for real-time aerial crowd-counting on consumer drones operating under challenging conditions, including low illumination, adverse weather, and nighttime. Our primary contributions are threefold:

- Based on a shared-backbone single-stream architecture, a lightweight feature fuser (F2) is designed and equipped with compact extractors tailored to each input modality. Inter-modal complementarity is reinforced through a CGAN-based learning process that promotes feature alignment and consistency.
- A density-aware modality converter (DMC) is developed for missing modality generation. A pre-trained crowd-counting model is incorporated to provide spatial supervision, constraining local object structures in the CycleGAN-generated TIR images and improving their fidelity to real TIR distributions.
- A modality-adaptive gating (MG) mechanism is presented to optimize computational efficiency. By analyzing density maps produced by dedicated decoders under both single- and dual-modal inputs, the system dynamically determines the necessity of TIR input, generating binary control signals for adaptive modality activation and resource-efficient inference.

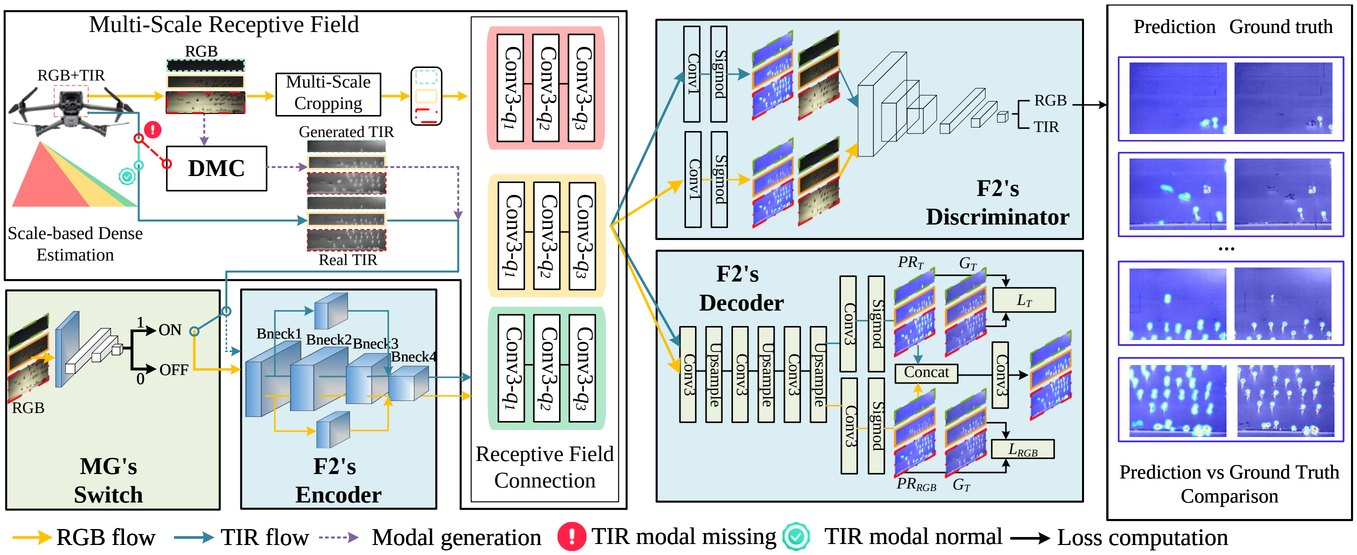


Fig. 2: MA-DyNN architecture (F2: Feature Fuser; DMC: Density-aware Modality Converter; MG: Modality-adaptive Gating).

Extensive experiments on two authoritative RGB-T crowd-counting benchmarks, DroneRGBT [34] and RGBT-CC [35], demonstrate the effectiveness and superiority of the proposed approach. Specifically, our experimental results provide clear answers to the following research questions:

- How robustly does MA-DyNN perform under unstable or missing modality conditions?
- To what extent can MA-DyNN simultaneously optimize counting accuracy and inference latency in real-time?
- Can MA-DyNN approach the computational efficiency of single-modality methods while preserving the accuracy advantages of multimodal integration?

The remainder of this paper is organized as follows. Section II details the MA-DyNN architecture, including the FFE, the MG, and the DMC modules. Section III presents ablation studies and performance comparisons against state-of-the-art methods. Section IV concludes this work with discussions on future research directions.

II. Proposed Solution

Airborne crowd-counting demands robust scale adaptation across varying flight altitudes. Previous approaches for drone imagery analysis, such as ARCNN [36], primarily enhance tiny-object detection through specialized mechanisms like progressive attention and density maps. However, our approach addresses the more challenging scenario of handling multi-scale targets and high-density crowd distributions. To address this issue, we adopt a multiscale receptive field strategy that integrates simplified altitude-aware scale estimation, adaptive image cropping, and optimized dilation rate combinations [37]. This strategy employs three adjustable dilated convolutions with dilation rates represented as triplets $Q = (q_1, q_2, q_3)$, where different Q configurations are designed to handle different crowd scales effectively. This design eliminates the need for explicit distance measurement while enabling consistent scale-invariant feature extraction across diverse flight heights. Built upon this scale-normalized foundation, the proposed MA-DyNN architecture incorporates dynamic computation and multimodal fusion through three interdependent modules, F2, DMC, and MG, as illustrated in Fig. 2:

1) F2. This module is constructed within a CGAN framework, where the generator comprises an encoder-decoder architecture. The encoder extracts complementary features from RGB (yellow flow) and TIR (blue flow) inputs, while the decoder generates the final density map. Meanwhile, the discriminator determines modality attribution, enforcing cross-modal alignment. This adversarial design facilitates a shared representation space that preserves modality-specific benefits while enabling effective fusion.

2) DMC. This function synthesizes high-quality pseudo-TIR images from RGB inputs in the absence of TIR data (indicated by reddish-brown flow). With density distribution consistency loss, the model enhances structural and contextual alignment between generated and real TIR images, ensuring reliable crowd-counting even under partial modality conditions. When TIR data is available (denoted by a green checkmark), the system directly employs the real thermal images.

3) MG. This controller can dynamically activate modality branches based on ambient lighting conditions. Under favorable illumination, only the RGB branch reduces unnecessary computation. The TIR branch is activated in low-light scenarios to enable efficient on-demand multimodal inference.

A. F2 Design

As illustrated in Fig. 2, the proposed multimodal feature fusion mechanism is built upon a CGAN framework, comprising a generator and a discriminator. The generator produces modality-aware features passed through scale-aware convolution operations before being evaluated by the discriminator. The discriminator determines the modality origin of the generated features, thereby enforcing cross-modal alignment through adversarial learning. This process allows the generator to learn modality-consistent representations while preserving modality-specific characteristics.

1) Generator. The generator serves as the core component for multimodal feature extraction and density map generation. Generator's workflow involves two sequential stages where input RGB and TIR images first undergo feature extraction and fusion, followed by density map generation to produce accurate crowd counting results. Specifically, due to the spatial distribution inconsistency between RGB and TIR images, we must reconstruct the feature extraction method. Unlike two-stream networks with large parameters, this subsection employs a universal feature extractor to learn shared parameters, in which MobileNet v3 [38] is used to construct the shared-backbone network.

Let (x_i^r, x_i^t) denote a multimodal sample pair, comprising an RGB sample x_i^r and its corresponding TIR sample x_i^t. Two lightweight modality-specific feature extractors are integrated into the backbone network to extract complementary modality-specific features from each sample, ensuring the preservation of essential modality information. After extracting universal (modality-consistent) features via the backbone, these features are element-wise combined with modality-specific features extracted from the modality extractors. Taking x_i^t as an example, the fused feature representation is computed as

$$f_i^t = F^u(x_i^t; \vartheta^u) + F^t(x_i^t; \vartheta^t) \tag{1}$$

where F^u and F^t denote the universal and TIR-specific feature extractors with corresponding parameter sets ϑ^u and ϑ^t. At the j-th convolutional layer, let W_j^u and W_j^t represent convolution parameters for universal and modality-specific extractors, respectively. The expression in (1) at layer j can be simplified as

$$\begin{aligned} f_{i,j}^t &= W_j^u \times x_i^t + W_j^t \times x_i^t \\ &= (W_j^u + W_j^t) \times x_i^t = W_j^{\mathrm{T}} \times x_i^t \end{aligned} \tag{2}$$

where $\times$ is the convolution operation, and W_j^{T} denotes the fused convolution parameters, capturing modality-specific representation.

Subsequently, the extracted multimodal features undergo three successive upsampling operations within the decoder module, restoring them to the original input resolution. Separate density maps M_i^r and M_i^t are generated for RGB and TIR modalities, respectively, and fused through a final convolution operation, yielding the integrated multimodal density map M_i^*. To achieve stable convergence, we employ a k-iteration optimization with root-based loss [37]. Given the ground truth density map $\hat{M}_i$, the training objective for the generator over a batch $\mathcal{I}$ containing I sample pairs is expressed as

$$\mathcal{L}_G = \frac{1}{I}\sum_{i=1}^{I}\sqrt[k]{\|M_i^* - \hat{M}_i\|_2^2}. \tag{3}$$

The optimal value of k is determined experimentally, as detailed in Section III.

2) Discriminator. Based on the adversarial learning principle commonly used in crowd counting, where the focus lies in distinguishing predicted density maps from ground truth distributions, we construct a modality-oriented adversarial framework that focuses on the feature representation of RGB and TIR modalities. This enables the two modalities to complement each other and achieve effective alignment in the feature space. The discriminator is designed to identify the modality origin of the features generated by the CGAN. It facilitates alignment between RGB and TIR modalities by learning to distinguish between them. The generator outputs modality-specific features, which are passed through a 1×1 convolutional layer followed by a Sigmoid activation to produce S_R and S_T. These are concatenated with the RGB image to form a 4-channel input to the discriminator. A SalGAN-like [39] architecture is used to predict modality labels and enforce feature consistency across modalities.

3) Adversarial training. The features of RGB and TIR modalities participating in adversarial learning are mapped to the same feature space and then passed through the convolutional layer and Sigmoid function to obtain the discrimination results of the discriminator. When training the discriminator, the generator is fixed. The generator parameters are updated during generator training, while the discriminator's loss is backpropagated. Note that the discriminator is only used in the training phase for adversarial learning between modality features. The discriminator loss function is expressed as

$$\mathcal{L}_D = \frac{1}{I}\sum_{i=1}^{I}[\mathcal{L}(D(x_i^r, S_R), 1) + \mathcal{L}(D(x_i^r, S_T), 0)] \tag{4}$$

where $\mathcal{L}(\cdot, \cdot)$ is the binary cross-entropy loss, and the targets 1 and 0 represent ground-truth modality labels for RGB and TIR features, respectively. The final training loss combines the generator loss and the discriminator loss, expressed as

$$\mathcal{L}_1 = \mathcal{L}_G + \mathcal{L}_D. \tag{5}$$

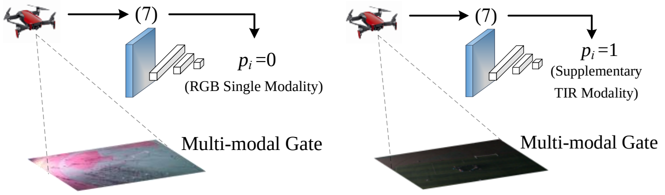
This joint optimization ensures the fused features are not only effective for crowd density regression but also aligned in modality representation space, thus enhancing generalization and robustness during inference.

B. MG Mechanism

Accurate and resource-efficient crowd counting requires adaptive selection of modality branches based on input scene characteristics. To this end, this subsection introduces an MG mechanism that enables selective activation of the TIR branch under challenging environmental conditions (e.g., low illumination, dust, or adverse weather), while defaulting to RGB-only inference when appropriate.

The MG mechanism is implemented as a lightweight binary classifier that determines whether the TIR modality is necessary, based solely on input RGB images. They are passed through a shallow convolutional encoder, optimized to extract global illumination and contextual cues. These features are fed into a fully connected layer to generate a modality selection confidence score, guiding the gating decision.

To supervise the training of the binary classifier, we introduce an illumination-adaptive thresholding mechanism that dynamically adjusts the gating criterion based on ambient lighting. Specifically, a perceptual luminance score ρ_i is com-



(a) Good illumination ($M_i^r = 97$, $M_i^* = 106$) (b) Poor illumination ($M_i^r = 4$, $M_i^* = 16$)

Fig. 3: Decision-making of multimodal gating.

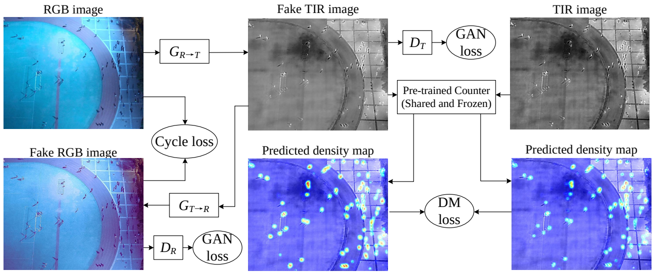


Fig. 4: Modal transformation framework.

puted for each RGB image using the ITU-R BT.601[1] standard. An instance-specific threshold ζ_i is then calculated as

$$\zeta_i = \exp(-\alpha \cdot \rho_i) \tag{6}$$

where α is a scaling factor empirically optimized on the validation set.

During training, each RGB–TIR pair (x_i^r, x_i^t) is fed into the network to generate three density maps: M_i^r, produced using only the RGB branch; M_i^t, obtained from the TIR branch alone; and M_i^*, resulting from the fusion of RGB and TIR inputs. This setup enables comprehensive evaluation of each modality's contribution to the counting performance and provides supervision signals for the subsequent gating strategy. The binary gating signal p_i indicating the necessity of the TIR branch is computed as

$$p_i = \begin{cases} 1, \ \text{if } (M_i^* - M_i^r)/M_i^* > \zeta_i \\ 0, \ \text{otherwise.} \end{cases} \tag{7}$$

$p_i = 1$ indicates that TIR significantly improves counting accuracy and should be activated, typically observed in scenarios with substantial shadowed regions or poor illumination in RGB imagery. Conversely, $p_i = 0$ implies that RGB alone provides adequate counting accuracy, allowing the TIR branch to remain inactive to conserve computational resources and enhance inference efficiency.

For accurately predicting optimal modality selections, we employ the following binary cross-entropy loss

$$\mathcal{L}_2 = -\sum_{i=1}^{I} [p_i \log(\hat{p}_i) + (1 - p_i) \log(1 - \hat{p}_i)] \tag{8}$$

where $\hat{p}_i \in (0, 1)$ denotes the predicted activation probability from the gating network. This loss encourages the network to correctly identify scenarios where integrating TIR modality significantly enhances counting accuracy.

As illustrated in Fig. 3, during inference, the gating network dynamically selects modality branches based on illumination conditions. Under favorable lighting (Fig. 3(a)), the RGB modality alone is sufficient to yield accurate crowd counting results. Consequently, the TIR branch remains inactive, and the model outputs the RGB-based density map (M_i^r). Conversely, in poorly illuminated environments (Fig. 3(b)), the gating mechanism automatically activates the TIR modality to supplement RGB information, and the model outputs the fused multimodal density map (M_i^*). This modality-adaptive inference behavior ensures accurate crowd counting performance under diverse illumination conditions while optimizing computational resources.

C. DMC Design

In complex scenarios, TIR imagery plays a pivotal role in crowd counting, especially under adverse illumination. However, drone-based platforms may occasionally fail to acquire TIR data due to sensor limitations, environmental obstructions, or cost constraints. To address such modality incompleteness, we develop a DMC module that synthesizes high-fidelity pseudo-TIR images from RGB inputs to support robust multimodal inference.

Traditional image-to-image translation methods, such as Pix2Pix GAN [31], typically rely on pixel-level losses (e.g., L1 loss and adversarial loss) and require strictly aligned image pairs. While effective for general translation tasks, these pixel-level losses treat all spatial regions equally and lack sensitivity to crowd-relevant structures. Furthermore, they operate in raw intensity space, making them sensitive to photometric shifts between RGB and TIR modalities. Although CycleGAN relaxes the paired-data requirement by introducing cycle consistency, its adversarial and reconstruction losses mainly promote visual plausibility and bidirectional domain mapping, without explicitly preserving spatial crowd distributions critical for accurate counting.

Inspired by task-guided generative modeling approaches (e.g., NGGAN [40]), we augment the CycleGAN architecture with a density map-constrained loss to enforce crowd-specific spatial alignment, as illustrated in Fig. 4. Specifically, a pre-trained crowd-counting network is employed as a spatial mapping function $\psi(\cdot)$, transforming both real and generated TIR images into corresponding density maps. This enables the generator to receive task-driven supervision in the density space rather than raw image space. Let $G_{R\to T}$ denote the generator that translates an RGB image x_i^r into a synthetic TIR image, i.e., $\hat{x}_i^t = G_{R\to T}(x_i^r)$. The density map-constrained loss is defined as the root mean squared error (RMSE) between the density maps derived from the synthetic and real TIR images, given by

$$\mathcal{L}_{\text{DM}}(G_{R\to T}) = \mathbb{E}_{(x_i^r, x_i^t)\sim P_{\text{data}}(x^r, x^t)} \left[M\left(\psi(\hat{x}_i^t), \psi(x_i^t)\right) \right] \tag{9}$$

where $\psi(\cdot)$ denotes the spatial mapping via a pre-trained counting model and $M(\cdot, \cdot)$ represents the RMSE between two density maps. Unlike traditional losses operating in raw

[1] https://www.itu.int/rec/R-REC-BT.601/

pixel space, the density map constraint enforces supervision in task-oriented feature space, explicitly modeling crowd spatial distributions. For two images x_i^r and $x_{i'}^r$ with similar crowd distributions but different illumination, $\mathcal{L}_{DMC}$ approaches zero in density space, while pixel-level L1 loss $\|x_i^r - x_{i'}^r\|_1$ remains large due to photometric variations. This indicates that the density map constraint provides photometrically robust structural supervision. By enforcing consistency between generated and real samples in density space, this constraint prevents the generator from converging to visually plausible but counting-inaccurate local optima. The density map constraint provides task-specific gradient signals from the crowd counting model, ensuring the generator learns representations directly relevant to downstream counting tasks, thereby enhancing task adaptability and training stability.

The adversarial loss used in CycleGAN for the TIR domain is expressed as

$$\begin{aligned}&\mathcal{L}_{\text{GAN}}(G_{R\to T}, D_T)\\&= \mathbb{E}_{x_i^t\sim P_{\text{data}}(x^t)}\left[\log D_T(x_i^t)\right]\\&+ \mathbb{E}_{x_i^r\sim P_{\text{data}}(x^r)}\left[\log\left(1 - D_T(G_{R\to T}(x_i^r))\right)\right]\end{aligned} \tag{10}$$

where D_T is the discriminator for domain T, trained to distinguish real TIR images from generated ones. The cycle-consistency loss encourages invertible mappings between domains R and T, expressed as

$$\begin{aligned}&\mathcal{L}_{\text{cycle}}(G_{R\to T}, G_{T\to R})\\&= \mathbb{E}_{x_i^r\sim P_{\text{data}}(x^r)}\left[\|G_{T\to R}(G_{R\to T}(x_i^r)) - x_i^r\|_1\right]\\&+ \mathbb{E}_{x_i^t\sim P_{\text{data}}(x^t)}\left[\left\|G_{R\to T}(G_{T\to R}(x_i^t)) - x_i^t\right\|_1\right].\end{aligned} \tag{11}$$

Combining the adversarial loss (10), cycle-consistency loss (11), and density map-constrained loss (9), the total objective for training the DMC is to minimize

$$\begin{aligned}\mathcal{L}_3 &= \mathcal{L}_{GAN}(G_{R\to T}, D_T) + \mathcal{L}_{GAN}(G_{T\to R}, D_R)\\&+ \lambda\mathcal{L}_{cycle}(G_{R\to T}, G_{T\to R}) + \mu\mathcal{L}_{DM}(G_{R\to T})\end{aligned} \tag{12}$$

where λ and μ are hyperparameters controlling the relative weight of cycle consistency and density map constraints. This enhanced training strategy ensures that the generated TIR images not only appear visually realistic but also accurately represent spatial crowd distributions aligned with downstream counting tasks.

III. Experimental Design and Results Analysis

Three authoritative datasets containing diverse and complex scenarios were utilized to enhance the validity and applicability of experimental results. The DroneRGBT dataset [34] comprises 3,600 RGB-TIR image pairs captured from varying altitudes under different lighting conditions (dark, dusk, bright), with crowd densities ranging from 1 to 401 individuals across diverse scenes including parks, streets, and shopping malls. The RGBT-CC dataset [35] contains 2,030 image pairs (640×480 pixels) from urban scenarios, with 1,013 pairs under bright illumination and 1,017 in dark environments, providing realistic crowd density distributions. The CARPK [41] dataset includes 1,448 drone-captured vehicle images from four parking lots, split into 989 training and 459 testing images with 90,000 annotations, enabling evaluation beyond crowd counting scenarios.

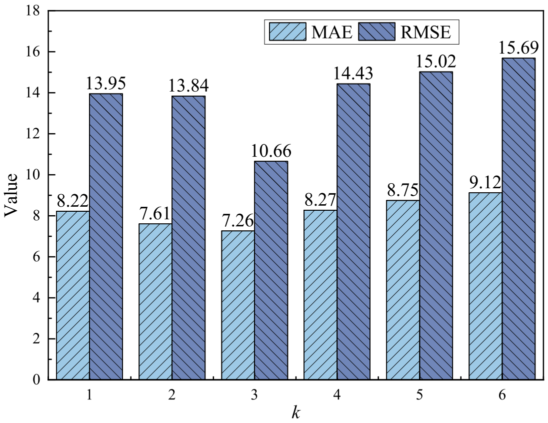


Fig. 5: MAE and RMSE for different values of k.

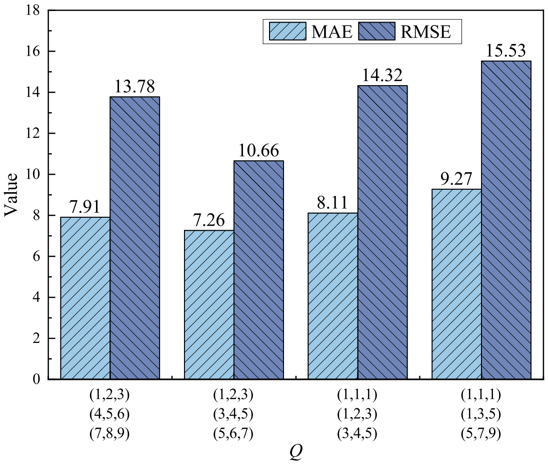


Fig. 6: MAE and RMSE for Q at (large, medium, small) scales.

Model training utilized a workstation featuring an Intel i9-13900K CPU and 128GB DDR5-3600 memory. Graphics processing was handled by dual RTX 4090 (24GB), while storage comprised a 1TB NVMe SSD and 6TB HDD. During training, the Adam optimizer was employed with a learning rate of 1×10^{-5} for density map generation. Adversarial loss was adopted to optimize the alignment between RGB and TIR features, utilizing alternating discriminator models. The Adam optimizer learning rate during adversarial training was set to 1×10^{-5}. These parameter selections ensured model stability and convergence, preserving accuracy and robustness in crowd counting. λ and μ in (12) were set to 0.1. For visualization purposes, density maps are displayed using a jet colormap with the colorbar range from 0 to 20, where warmer colors (red/yellow) indicate higher crowd density and cooler colors (blue) represent lower density areas.

To determine the optimal k in (6) and the combinations of dilation rate Q, we conducted a series of experiments on the DroneRGBT dataset. Fig. 5 shows the best performance at $k = 3$, which was used in all subsequent evaluations. Moreover, as shown in Fig. 6, the best accuracy was achieved at different combinations for different crowd densities: $Q = (1, 2, 3)$ for

TABLE I: Discriminator composition

Layer	Kernel	Activation	Out-channels
Conv1	3×3	ReLU	32
Conv2	3×3	ReLU	32
Max-pooling	2×2	N/A	32
Conv3	3×3	ReLU	64
Conv4	3×3	ReLU	64
Max-pooling	2×2	N/A	64
Conv5	3×3	ReLU	64
Conv6	3×3	ReLU	64
Max-pooling	2×2	N/A	64
Fc7	N/A	Tanh	100
Fc8	N/A	Tanh	2
Fc9	N/A	Sigmoid	1

small-scale, $Q = (3, 4, 5)$ for medium, and $Q = (5, 6, 7)$ for large-scale, respectively.

Table I details the discriminator architecture, comprising six 3×3 convolutional layers (Conv1–Conv6) with ReLU activations, interleaved with max-pooling every two layers. Three fully connected layers (Fc7–Fc9), followed by Tanh and Sigmoid activations, output the final domain classification. This design enabled the discriminator to effectively distinguish real and synthetic TIR samples.

To comprehensively evaluate the impact of modality selection and fusion strategies on overall performance and system adaptability, the proposed approach was categorized into six variants under diverse inference configurations:

- Proposed-1: Dynamic inference integrating RGB and TIR modalities.
- Proposed-2: Inference exclusively using RGB modality.
- Proposed-3: Inference exclusively utilizing TIR modality.
- Proposed-4: Simultaneous inference with both RGB and TIR modalities.
- Proposed-5: Dynamic inference employing RGB and synthetic TIR data.
- Proposed-6: Dual-stream dynamic inference leveraging RGB and TIR modalities.

To rigorously evaluate the adaptability and counting accuracy of MA-DyNN, several state-of-the-art monomodal and multimodal baselines were selected for comparative analysis:

1) Single-modal Baselines

- Baseline-1 (MCNN [42]): Employs multiple convolutional branches with varying receptive fields to capture multi-scale crowd features.
- Baseline-2 (SANET [43]): Integrates scale aggregation modules to enhance multi-scale representation learning and counting robustness.
- Baseline-3 (CSRNET [44]): Utilizes dilated convolutions to enlarge receptive fields, effectively addressing dense crowd-counting scenarios.
- Baseline-4 (CANNET [27]): Implements context-aware attention mechanisms to incorporate contextual information and improve counting accuracy.
- Baseline-5 (BL [45]): Adopts Bayesian loss functions for precise point-level supervision in crowd estimation tasks.
- Baseline-6 (AAVCNET [37]): Optimizes aerial-view counting via altitude-aware spatial feature learning and scale normalization.

TABLE II: Monomodality comparison on DroneRGBT (TIR and RGB) and CARPK (RGB only) datasets

Method	**DroneRGBT**				**CARPK**	
	TIR		**RGB**		**RGB**	
	MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Baseline-1	13.64	19.77	20.45	27.3	19.10	43.30
Baseline-2	12.13	17.52	14.91	21.66	9.8	–
Baseline-3	8.91	13.80	13.06	19.06	11.48	13.32
Baseline-4	7.78	12.31	10.87	17.58	–	–
Baseline-5	7.41	11.56	10.90	16.80	9.58	11.38
Baseline-6	8.10	13.09	11.62	17.99	6.6	9.8
Proposed-3	8.31	12.12	N/A	N/A	10.24	12.05
Proposed-2	N/A	N/A	11.73	17.55	N/A	N/A
	DroneRGBT (TIR & RGB)					
	MAE		**RMSE**		N/A	
Proposed-1	7.26		10.66		–	

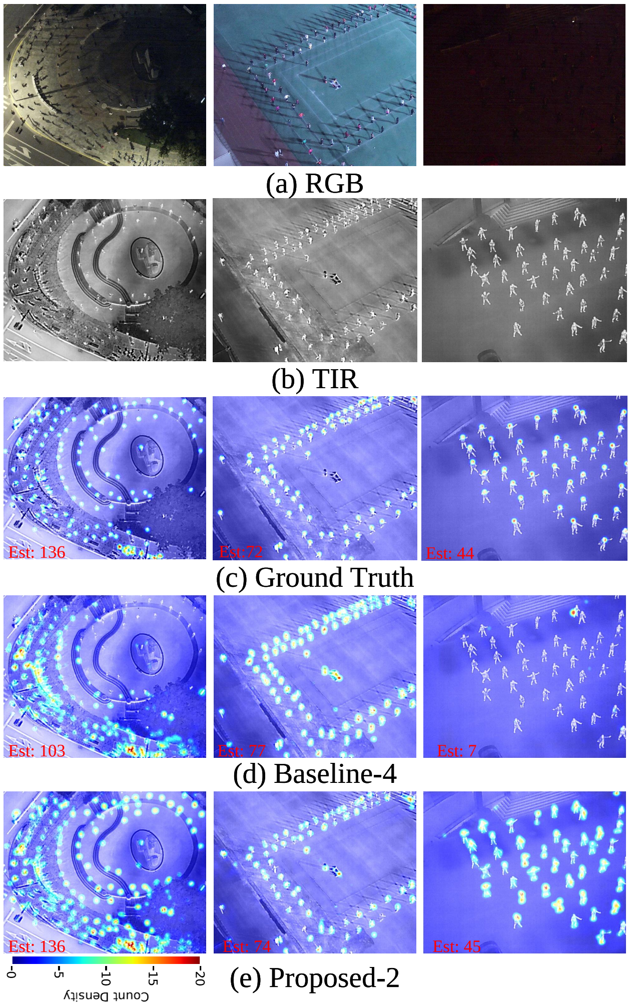
- Baseline-10 (LCDnet [46]): Provides a lightweight model tailored for real-time crowd-density estimation on resource-constrained edge devices.

2) Multimodal Baselines

- Baseline-7 (RMMCC [35]): Integrates transformer-based counting with multimodal RGB-T feature fusion to enhance crowd estimation accuracy.
- Baseline-8 (CMCRL [47]): Facilitates multimodal feature alignment and representation learning by enforcing distribution consistency between RGB and TIR modalities.
- Baseline-9 (MC^{3}Net [48]): Processes RGB and TIR streams through interactive fusion and cross-modality compensation, achieving state-of-the-art results.

A. Comparison with Monomodality Models

To demonstrate the superiority of the proposed multimodal approach over existing monomodality methods, extensive comparisons were conducted using monomodality data on both DroneRGBT and CARPK datasets. These baseline models were trained and evaluated individually on RGB or TIR data, with performance metrics summarized in Table II. Specifically, when utilizing RGB or TIR data alone on DroneRGBT, Proposed-2 and Proposed-3 exhibited substantial improvements over Baseline-1 to 3, although their performance slightly lagged behind Baseline-4 and other advanced methods. Similar trends were observed on CARPK. The comparative analysis against Baseline-6 and Baseline-5 indicated that our monomodality approach maintained competitive feature extraction capability. When incorporating TIR modality assistance into RGB-based counting, Proposed-1 consistently outperformed all monomodality baselines, confirming the multimodal integration's effectiveness. Fig. 7 illustrates density map visualizations, highlighting that Baseline-4, while proficient in RGB conditions, struggled in low-light detection. Conversely, Proposed-1 achieved accurate crowd counting under varying illumination scenarios, demonstrating its robustness under unstable or missing modality conditions.



(a) RGB

(b) TIR

(c) Ground Truth

(d) Baseline-4

(e) Proposed-2

Fig. 7: Density visualization from DroneRGBT (multimodality vs monomodality).

TABLE III: Multimodality comparison

Method	DroneRGBT		RGBT-CC		Params (M)	Delay (ms)
	MAE	RMSE	MAE	RMSE		
Baseline-7	6.98	10.25	13.72	18.79	15.8	164
Baseline-8	9.02	15.74	17.94	30.91	13.4	139
Baseline-9	6.98	12.19	11.47	20.59	113.0	235
Proposed-6	7.18	11.43	15.48	24.66	2.8	43
Proposed-1	7.26	10.66	16.39	27.01	2.0	34

B. Comparison with Multimodality Models

This subsection presents a comparative analysis between MA-DyNN and representative multimodal crowd-counting models. Proposed-6, designed as a dual-stream network, extracts modality-specific features using the initial four blocks of MobileNet V3 [38] for each modality branch. Subsequently, extracted feature maps are concatenated along the channel dimension, reduced via a 1×1 convolution, and processed through a regression module to generate the final density map.

Table III details the comparative performance on the DroneRGBT and RGBT-CC datasets, where Proposed-1 achieves an MAE of 7.26, corresponding to 96.1% of the state-of-the-art Baseline-9's accuracy on DroneRGBT. Although Proposed-1 achieves slightly higher MAE compared to Baseline-9 on RGBT-CC, Proposed-1 achieves satisfactory counting accuracy with substantially fewer parameters. Notably, Baseline-9's substantial computational requirements may not be suitable for deployment on consumer-grade drone platforms. Furthermore, Proposed-6 utilizes a parameter-intensive dual-stream design, achieving improved accuracy at the cost of increased inference latency. Due to multimodal inference often exhibiting modality redundancy, the proposed method efficiently and dynamically extracts critical information from multiple modalities. Specifically, the decoupled modality processing design in Proposed-1 enables adaptive switching to monomodality (RGB-only) inference under favorable lighting conditions, significantly reducing computational overhead while maintaining high accuracy. Thus, the modality-adaptive strategy of MA-DyNN effectively balances crowd-counting accuracy and inference latency, demonstrating practical value for real-world drone deployment scenarios.

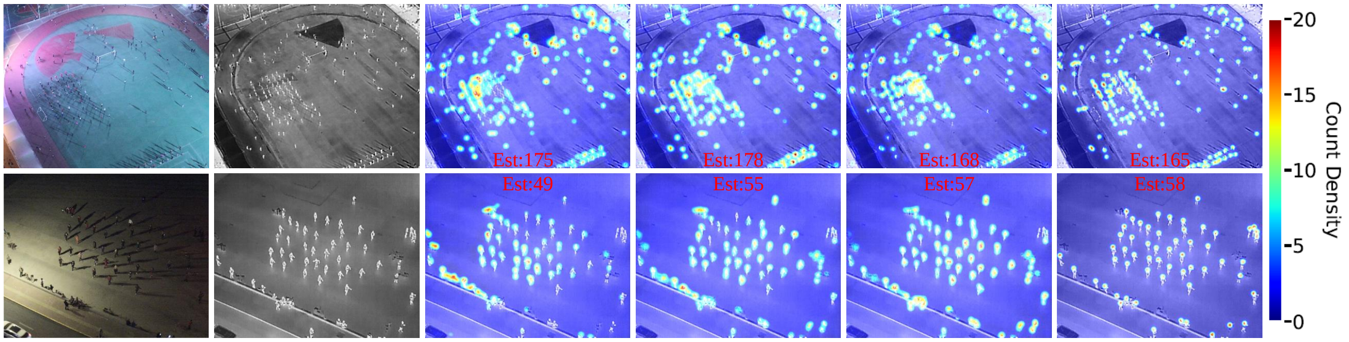
Inference latency, strongly correlated with environmental lighting conditions, is summarized in Table III. The average latency is non-linearly related to modality count, highlighting the efficiency of the proposed multimodal gating strategy, which dynamically activates TIR processing only in low-light scenarios. Visual results from DroneRGBT and RGBT-CC datasets, presented in Figs. 8 and 10, reveal substantial false positives and negatives in low-light RGB-only scenarios (Proposed-2). The baseline multimodal method effectively captured most targets under similar conditions; however, Proposed-1 outperformed by accurately detecting targets across diverse lighting situations with fewer errors.

C. Cross-Dataset Performance Comparison

To evaluate the generalization capability of MA-DyNN, comprehensive cross-dataset experiments were conducted using DroneRGBT and RGBT-CC datasets. Results summarized in Table IV reveal that MA-DyNN significantly outperforms Baseline-1 across all conditions. Although MA-DyNN achieves a slightly higher MAE than the parameter-intensive Baseline-9 within individual datasets, it exhibits superior robustness in cross-dataset evaluations, with markedly less performance degradation. Specifically, when evaluated from RGBT-CC to DroneRGBT, MA-DyNN demonstrates exceptional cross-domain stability with MAE increasing by only 1.78, compared to Baseline-9's increase of 2.05 and Baseline-1's substantial degradation of 3.88. Furthermore, MA-DyNN nearly matches Baseline-9's performance (9.04 vs. 9.03) while consuming considerably fewer computational resources. These findings strongly support MA-DyNN's suitability and practical applicability on resource-constrained consumer drones.

D. Inference Performance in Resource-limited Environments

To validate the deployment feasibility of MA-DyNN on consumer-grade drones, we conducted comprehensive inference performance evaluations on resource-constrained embedded platforms. Using Docker-based simulation environments to emulate the computational limitations of the NVIDIA



(a) RGB (b) TIR (c) Proposed-2 (d) Baseline-8 (e) Proposed-1 (f) Ground Truth

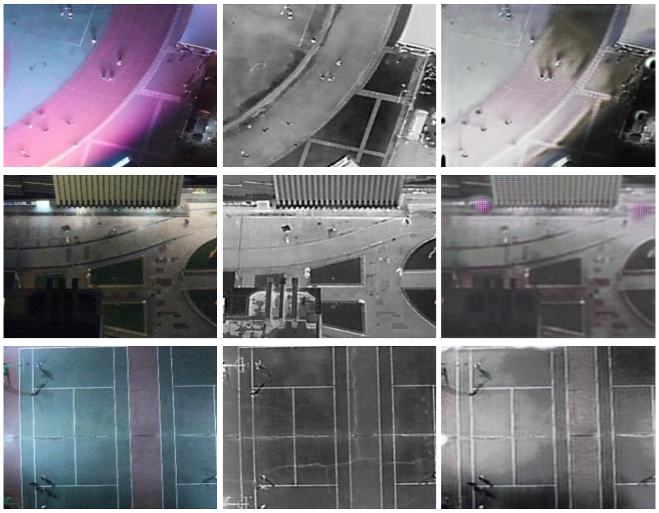
Fig. 8: Multimodal density visualization from DroneRGBT (compared with multi-modality models).

TABLE IV: Cross-dataset comparison

Transfer	Metric	Baseline-1	Baseline-9	Proposed-1
DroneRGBT ↓ RGBT-CC	MAE RMSE	25.89 40.11	14.29 24.52	17.93 28.34
RGBT-CC ↓ DroneRGBT	MAE RMSE	24.33 30.22	9.03 15.99	9.04 12.39

TABLE V: Reasoning latency on resource-limited platform

Method	Param (M)	SSIM ↑	PSNR ↑	Latency (ms)
Baseline-1	0.13	0.54	18.2	210
Baseline-3	16.26	0.72	21.70	1880
Baseline-2	0.25	0.59	19.4	230
Baseline-10	0.05	0.60	21.39	100
Proposed-2	2.0	0.68	20.2	220
Proposed-1	2.0	0.75	21.97	470



(a) RGB (b) Real TIR (c) Generated TIR

Fig. 9: Modality conversion rendering.

Jetson Nano, we compared MA-DyNN against several state-of-the-art lightweight models specifically designed for real-time inference. As shown in Table V, MA-DyNN achieves notably higher counting accuracy (SSIM: 0.75, PSNR: 21.97), significantly outperforming lightweight baselines in both structural fidelity and perceptual quality.

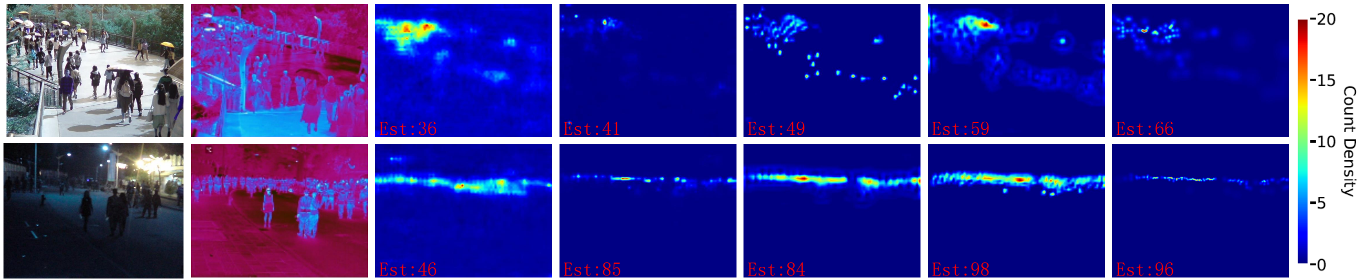
While MA-DyNN's multimodal inference incurs a longer runtime (0.47s) compared to unimodal models such as LCD-Net (0.10s) and MCNN (0.21s), the added latency is compensated by substantial gains in accuracy through effective modality fusion. Importantly, MA-DyNN leverages its MG module to dynamically bypass the TIR branch under favorable lighting, enabling fast RGB-only inference (0.22s) without sacrificing robustness. This adaptive behavior ensures a favorable trade-off between efficiency and accuracy depending on real-time conditions. Furthermore, as reported in Table III, MA-DyNN delivers faster inference than existing multimodal baselines, reinforcing its practicality for deployment in latency-sensitive aerial scenarios.

To further assess its energy efficiency, we estimated power consumption during inference using specifications from the NVIDIA Jetson developer documentation[2]. The Jetson Nano supports two official power modes: 5W and 10W. Assuming deployment under the default 10W mode for peak performance, MA-DyNN consumes approximately 4.7J per inference for full RGB–TIR processing (0.47s), and 2.2J for RGB-only inference (0.22s). Given a standard drone battery capacity of 60Wh, MA-DyNN supports around 2.6 hours of continuous adaptive inference. These results highlight the energy-aware design of MA-DyNN and confirm its suitability for resource-constrained aerial scenarios.

TABLE VI: Ablation comparison on DroneRGBT

Method	MAE	RMSE	Latency (ms)
w/o CGAN	13.46	21.14	72
Proposed-4	7.11	9.83	87
Proposed-5	8.14	13.24	36
Proposed-2	8.31	12.12	24
Proposed-3	11.73	17.55	27
Proposed-1	7.26	10.66	34

[2] https://docs.nvidia.com/jetson/archives/l4t-archived/l4t-3275/index.html



(a) RGB (b) TIR (c) M^r (d) M^t (e) Early Fusion (f) $M*$ (g) Ground Truth

Fig. 10: Visualization of different ablation stages on RGBT-CC.

E. Ablation Experiments

To validate the contributions and effectiveness of individual components within MA-DyNN, comprehensive ablation experiments were conducted on DroneRGBT.

We first assessed the F2's necessity. As presented in Table VI, removing the CGAN-based fusion (forcing modalities to share a common backbone without explicit alignment) significantly increased MAE compared to Proposed-1. This performance degradation highlights interference between modality-specific features in the absence of effective fusion. Thus, the CGAN-based approach demonstrates a clear advantage by extracting coherent, modality-aligned representations, enhancing crowd-counting accuracy.

The contribution of MG in balancing counting accuracy and computational efficiency was next examined. Results in Table VI reveal that activating multimodal gating (Proposed-1) outperforms monomodality methods (Proposed-2 and Proposed-3). Conversely, disabling multimodal gating (Proposed-4), forcing simultaneous dual-modal processing, yields only a minor improvement in MAE (approximately 2.1%) but incurs significantly higher computational costs. This indicates that MG effectively prevents redundant computation, maintaining robustness without sacrificing inference efficiency.

To provide deeper insight into feature alignment effectiveness, we visualized intermediate density maps generated by individual RGB (M^r) and TIR (M^t) modalities before fusion, as shown in Fig. 10. For comparative analysis, an early fusion baseline was implemented using concatenated RGB-T inputs with CSRNet. The final fusion method (M^*) surpasses the early fusion baseline, validating the effectiveness and accuracy advantages of our feature fusion architecture.

We further evaluated the DMC's effectiveness in synthesizing TIR imagery. Fig. 11 provides visual comparisons between real TIR images and those generated by the DMC, demonstrating that synthetic images preserve object contours and luminance features. Quantitatively, substituting authentic TIR images with synthetic counterparts (Proposed-5) results in only a modest increase in MAE of 9.4% (Table VI). Furthermore, to quantitatively assess visual and semantic fidelity, we randomly selected 200 image pairs from the DroneRGBT test dataset and computed the SSIM and PSNR metrics between generated and real TIR images. As summarized in Table VII, DMC-generated TIR images exhibit significantly improved structural similarity and reconstruction quality compared to baseline generative methods, ensuring robust counting even when one modality is unavailable.

TABLE VII: Comparison of generated TIR image quality

Method	**SSIM↑**	**PSNR↑**
w/o density map	0.3905	24.23
Proposed DMC	0.4788	27.87

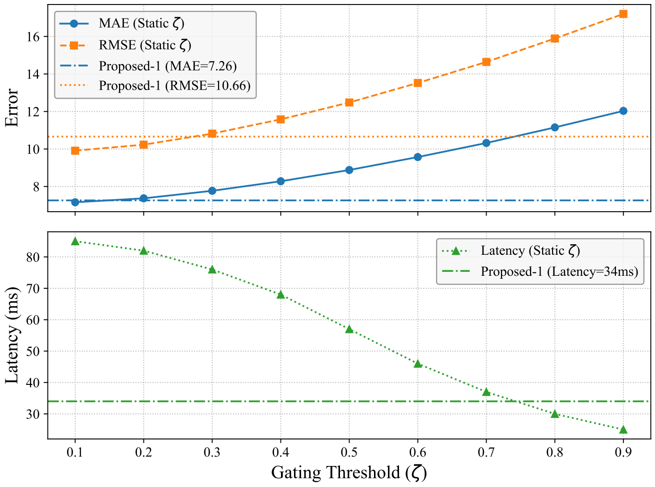


Fig. 11: Sensitivity analysis of ζ.

To investigate the impact of ζ, the threshold in (6), on model performance, we conducted a sensitivity analysis on DroneRGBT. In this experiment, the dynamic ζ computation module was disabled, and the model was evaluated across a range of fixed values. As illustrated in Fig. 10, increasing ζ from 0.1 to 0.9 caused the model to increasingly favor the faster RGB-only inference path, which led to a monotonic rise in prediction errors (MAE/RMSE) while significantly reducing inference latency. Crucially, our method exhibited a clear decoupling between accuracy and latency trade-offs: the performance equivalence points for these two metrics did not coincide under any single static threshold. In other words, no fixed value simultaneously achieved optimal accuracy and efficiency, whereas our dynamic thresholding mechanism adaptively selected the optimal modality based on real-time illumination. This behavior validated the effectiveness of our adaptive design in achieving both high accuracy and low latency, surpassing fixed-threshold baselines.

IV. Conclusion

In this study, we proposed MA-DyNN, a Modal-Adaptive Dynamic Neural Network framework that effectively integrates complementary RGB and TIR modalities for robust and efficient crowd counting on consumer drones. By leveraging the designed F2, DMC, and MG modules, MA-DyNN achieves a favorable balance between counting accuracy and computational efficiency under varying lighting conditions. Extensive experiments on the DroneRGBT and RGBT-CC benchmarks demonstrate that MA-DyNN consistently outperforms state-of-the-art methods in both accuracy and inference latency. Its adaptive inference strategy attains computational efficiency comparable to monomodal models while preserving the accuracy advantages offered by multimodal integration. Ablation studies validate the effectiveness of the F2 and MG components in enabling adaptive, resource-efficient inference. Moreover, MA-DyNN exhibits strong resilience to modality instability, showing minimal accuracy degradation when employing synthetic TIR data. These strengths underscore MA-DyNN's practical value in real-world consumer drone applications, including public safety surveillance, urban traffic monitoring, and crowd management.

Future research directions include incorporating additional sensing modalities (e.g., LiDAR or radar) to enhance robustness under severe conditions, as well as investigating domain adaptation methods to improve generalization across diverse geographical locations and crowd distributions.

References

[1] T. K. Behera, S. Bakshi, M. A. Khan, and H. M. Albarakati, "A lightweight multiscale-multiobject deep segmentation architecture for UAV-based consumer applications," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3740–3753, 2024.

[2] H. Shen, Z. Tong, T. Wang, and G. Bai, "UAV-relay-assisted live layered video multicast for cell-edge users in NOMA networks," *IEEE Trans. Broadcast.*, vol. 70, no. 1, pp. 135–147, 2024.

[3] H. Shen, Y. Tian, T. Wang, and G. Bai, "Slicing-based task offloading in space-air-ground integrated vehicular networks," *IEEE Trans. Mob. Comput.*, vol. 23, no. 5, pp. 4009–4024, 2024.

[4] P. Zhu, T. Peng, D. Du, H. Yu, L. Zhang, and Q. Hu, "Graph regularized flow attention network for video animal counting from drones," *IEEE Trans. Image Process.*, vol. 30, pp. 5339–5351, 2021.

[5] O. Elharrouss, N. Almaadeed, K. Abualsaud, A. Al-Ali, A. Mohamed, T. Khattab, and S. Al-Maadeed, "Drone-SCNet: Scaled cascade network for crowd counting on drone images," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 6, pp. 3988–4001, 2021.

[6] X. Yu, Y. Liang, X. Lin, J. Wan, T. Wang, and H.-N. Dai, "Frequency feature pyramid network with global-local consistency loss for crowd-and-vehicle counting in congested scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9654–9664, 2022.

[7] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in vhr optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[8] M. Ma, C. Xia, C. Xie, X. Chen, and J. Li, "Boosting broader receptive fields for salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1026–1038, 2023.

[9] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8198–8207.

[10] T. Peng, Q. Li, and P. Zhu, "RGB-T crowd counting from drone: A benchmark and MMCCN network," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 497—513.

[11] X. Li, H. Chen, Y. Li, and Y. Peng, "MAFusion: Multiscale attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–16, 2022.

[12] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13 955–13 965.

[13] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2019.

[14] Q. Xu, Y. Mei, J. Liu, and C. Li, "Multimodal cross-layer bilinear pooling for RGBT tracking," *IEEE Trans. Multimedia.*, vol. 24, pp. 567–580, 2021.

[15] H. Xing, W. Wei, L. Zhang, and Y. Zhang, "Multi-scale feature extraction and fusion with attention interaction for RGB-T tracking," *Pattern Recognit.*, vol. 157, p. 110917, 2025.

[16] S. Tang, Y. Wang, Z. Kong, T. Zhang, Y. Li, C. Ding, Y. Wang, Y. Liang, and D. Xu, "You need multiple exiting: Dynamic early exiting for accelerating unified vision language model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10 781–10 791.

[17] H. Shen, Q. Liu, Y. Wang, T. Wang, and G. Bai, "MT-DyNN: Multi-teacher distilled dynamic neural network for instance-adaptive detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 5, pp. 6116–6129, 2025.

[18] F. Szatkowski, B. Wójcik, M. Piórczyński, and S. Scardapane, "Exploiting activation sparsity with dense to dynamic-k mixture-of-experts conversion," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 43 245–43 273, 2024.

[19] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[20] X. Yu, X. Cheng, Y. Liu, and Z. Zheng, "A dual-stream cross-domain integration network for RGB-T salient object detection," *IEEE Trans. Consum. Electron.*, vol. 71, no. 1, pp. 883–894, 2025.

[21] Y. Fang, R. Hou, J. Bei, T. Ren, and G. Wu, "ADNet: An asymmetric dual-stream network for RGB-T salient object detection," in *Proc. ACM Int. Conf. Multimed. Asia*, 2023, pp. 1–7.

[22] T. Guan, J. Wang, S. Lan, R. Chandra, Z. Wu, L. Davis, and D. Manocha, "M3DETR: Multi-representation, multi-scale, mutual-relation 3D object detection with transformers," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2293–2303.

[23] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23 716–23 736.

[24] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19 730–19 742.

[25] Y. He, R. Cheng, G. Balasubramaniam, Y.-H. H. Tsai, and H. Zhao, "Efficient modality selection in multimodal learning," *J. Mach. Learn. Res.*, vol. 25, no. 47, pp. 1–39, 2024.

[26] X. Wang, X. Shu, S. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu, "MFGNet: Dynamic modality-aware filter generation for RGB-T tracking," *IEEE Trans. Multimed.*, vol. 25, pp. 4335–4348, 2022.

[27] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5099–5108.

[28] M. Saeed Shafiee, M. Javad Shafiee, and A. Wong, "Dynamic representations toward efficient inference on deep neural networks by decision gates," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 667–675.

[29] Z. Chen, Y. Li, S. Bengio, and S. Si, "You look twice: Gaternet for dynamic filter selection in cnns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9172–9180.

[30] A. Lu, C. Li, J. Zhao, J. Tang, and B. Luo, "Modality-missing RGBT tracking: Invertible prompt learning and high-quality benchmarks," *Int. J. Comput. Vis.*, vol. 133, no. 5, pp. 2599–2619, 2025.

[31] M. A. Khan, H. Menouar, and R. Hamila, "Crowd counting in harsh weather using image denoising with pix2pix gans," in *Proc. Int. Conf. Image and Comput. Vis. New Zealand*, 2023, pp. 1–6.

[32] F. Li, Y. Zha, L. Zhang, P. Zhang, and L. Chen, "Information lossless multi-modal image generation for RGB-T tracking," in *Proc. Chinese Conf. Pattern Recognit. Comput. Vis.*, 2022, pp. 671–683.

[33] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-Fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Trans. Image Process.*, vol. 32, pp. 5705–5720, 2023.

[34] T. Peng, Q. Li, and P. Zhu, "RGB-T crowd counting from drone: A benchmark and MMCCN network," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 497–513.

[35] Z. Liu, W. Wu, Y. Tan, and G. Zhang, "RGB-T multi-modal crowd counting based on transformer," *arXiv preprint arXiv:2301.03033*, 2023.

[36] Y. Yu, K. Zhang, X. Wang, N. Wang, and X. Gao, "An adaptive region proposal network with progressive attention propagation for tiny person detection from UAV images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4392–4406, 2024.

[37] Y. Liu, H. Shen, T. Wang, and G. Bai, "Vehicle counting in drone images: An adaptive method with spatial attention and multiscale receptive fields," *ETRI Journal*, vol. 47, no. 1, pp. 7–19, 2025.

[38] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1314–1324.

[39] J. Pan, E. Sayrol, X. G.-i. Nieto, C. C. Ferrer, J. Torres, K. McGuinness, and N. E. OConnor, "SalGAN: Visual saliency prediction with adversarial networks," in *CVPR Scene Understanding Workshop*, 2017.

[40] Y.-R. Chien, P.-H. Chou, Y.-J. Peng, C.-Y. Huang, H.-W. Tsao, and Y. Tsao, "NGGAN: Noise generation gan based on the practical measurement dataset for narrowband powerline communications," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–15, 2025.

[41] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Int. Conf. Comput. Vis*, 2017, pp. 4145–4153.

[42] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 589–597.

[43] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. European Conf. Comput. Vis.*, 2018, pp. 734–750.

[44] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1091–1100.

[45] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6142–6151.

[46] M. A. Khan, H. Menouar, and R. Hamila, "LCDnet: a lightweight crowd density estimation model for real-time video surveillance," *J. Real-Time Image Process.*, vol. 20, no. 2, 2023.

[47] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4823–4833.

[48] W. Zhou, X. Yang, J. Lei, W. Yan, and L. Yu, "MC3Net: Multimodality cross-guided compensation coordination network for RGB-T crowd counting," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 4156–4165, 2024.

Hang Shen (Member, IEEE) received the Ph.D. degree with honors in Computer Science from the Nanjing University of Science and Technology, Nanjing, China, in 2015. He worked as a Full-Time Postdoctoral Fellow with the Broadband Communications Research (BBCR) Lab, Electrical and Computer Engineering Department, University of Waterloo, Waterloo, ON, Canada, from 2018 to 2019. He is an Associate Professor with the Department of Computer Science and Technology at Nanjing Tech University, Nanjing, China. His research interests involve space-air-ground integrated networks, drone/vehicle vision-language navigation, and cybersecurity. He serves as an Associate Editor for *Journal of Information Processing Systems*, *Frontiers in Blockchain*, and IEEE ACCESS, and was a Guest Editor for *Peer-to-Peer Networking and Applications*. He was a Program Committee Member of the IEEE International Conference on High Performance Computing and Communications (HPCC), the Annual International Conference on Privacy, Security and Trust (PST), the International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), and the International Conference on Artificial Intelligence Computing and Systems (AICompS). He is an IEEE member, a CCF Senior Member, and an Executive Committee Member of the ACM Nanjing Chapter.

Qi Liu received the B.Eng. degree in Computer Science from Shenyang Ligong University, Shenyang, China, in 2023. He is now pursuing his M.Eng. degree in Computer Science at Nanjing Tech University, Nanjing, China. His research interests include dynamic neural networks for autonomous driving object detection and unsupervised and semi-supervised learning for cybersecurity.

Yu Liu received the B.Eng. and M.Eng. degrees in Computer Science from Nanjing Tech University, Nanjing, China, in 2021 and 2024. His research interests include dynamic neural networks and multimodal deep learning for real-time object detection in drone images.

Tianjing Wang (Member, IEEE) received the B.Sc. degree in Mathematics from Nanjing Normal University in 2000, the M.Sc. degree in Mathematics from Nanjing University in 2002, and the Ph.D. degree in Signal and Information System from the Nanjing University of Posts and Telecommunications (NUPT) in 2009, all in Nanjing, China. From 2011 to 2013, she was a Full-Time Postdoctoral Fellow with the School of Electronic Science and Engineering, NUPT. From 2013 to 2014, she served as a Visiting Scholar with the Electrical and Computer Engineering Department at the State University of New York at Stony Brook, NY, USA. She is currently an Associate Professor in the Communication Engineering Department at Nanjing Tech University, Nanjing, China. Her research interests include embodied intelligence for the Internet of Vehicles/Drones. She is a member of IEEE and CCF.

Guangwei Bai received the B.Eng. and M.Eng. degrees in computer engineering from Xi'an Jiaotong University, Xi'an, China, in 1983 and 1986, respectively, and the Ph.D. degree in Computer Science from the University of Hamburg, Hamburg, Germany, in 1999. From 1999 to 2001, he worked as a Research Scientist at the German National Research Center for Information Technology, Germany. In 2001, he joined the University of Calgary, Calgary, AB, Canada, as a Research Associate. Since 2005, he has been working at Nanjing Tech University, Nanjing, China, as a Professor in Computer Science. From October to December 2010, he served as a Visiting Professor in the Electrical and Computer Engineering Department at the University of Waterloo, Waterloo, ON, Canada. His research interests include wireless networking, multimedia systems, and cybersecurity. He has authored and co-authored more than 80 peer-reviewed papers in international journals and conferences. He is an ACM member and a CCF Distinguished Member.