# A Sinkhorn Mixture-of-Experts Switch for Efficient RGB-Thermal Fusion

Aditya Viswakumar
Department of Artificial Intelligence
Indian Institute of Technology Hyderabad
ai23resch11001@iith.ac.in

Rajalakshmi Pachamuthu
Department of Electrical Engineering
Indian Institute of Technology Hyderabad
raji@ee.iith.ac.in

***Abstract*—In this work, we propose a computationally efficient RGB-Thermal (RGB-T) sensor fusion framework based on Sinkhorn Attention. The proposed method integrates a Sinkhorn-based alignment mechanism within a Mixture-of-Experts (MoE) architecture to enable effective and adaptive fusion of multi-modal features. The Sinkhorn module explicitly aligns features from RGB and thermal modalities, facilitating robust cross-modal interaction. The network adopts a multi-stage hierarchical fusion strategy, in which modality-specific and shared representations are progressively aligned and fused at multiple semantic levels. The proposed Sinkhorn fusion network performs competitively against existing methods while requiring reduced memory and computational resources. Owing to its lightweight design, the model is capable of real-time edge inference, making it suitable for resource-constrained environments.**

***Index Terms*—Sinkhorn Attention, Mixture-of-Experts, RGB-Thermal Fusion, Multi-Modal Learning**

## I. Introduction

RGB based digital perception relies on measuring the intensity of visible light reflected off object surfaces. This closely mimics the biological visual perception system that relies on capturing light from the visible region of the electromagnetic spectrum. RGB sensors are the most commonly used visual sensors because of their low cost and ease of integration with existing computing systems.

However, RGB perception is susceptible to illumination conditions and the optical property of the object under observation. As a result, they are sensitive to low-light, adverse weather and, surface material properties of the object.

To mitigate these limitations, recent research has increasingly explored the use of thermal modalities, which capture information emitted rather than reflected by objects [1]–[6]. Thermal imaging captures electromagnetic radiation in the infrared region and encodes an object's heat signature as two-dimensional intensity maps. Unlike RGB imagery, thermal images are largely invariant to external illumination and therefore provide robust perceptual cues under challenging conditions such as darkness, fog, and partial occlusion.

Thermal information is naturally produced by many entities in nature, making it particularly effective for detection and tracking tasks. Humans and other mammals maintain relatively stable body temperatures, resulting in distinctive thermal signatures [1], [2]. Thermal emissions from vegetation strongly correlate with a crop's physiological stage [3], [4]. Thermography is used in the field of medicine for diagnosing diseases such as arthritis and osteoarthritis in a non-invasive manner [5]. In industrial systems, thermal patterns reflect operational health and efficiency [6].

In this work, we investigate RGB–Thermal (RGB-T) fusion for efficient edge inference. We particularly focus on human spatial density estimation, a task that aims to estimate the number of humans in a scene by predicting a spatial density map [7], [8]. Spatial density estimation remains a challenging problem even with multi-modal inputs because of perspective distortion, scale variation, ambient lighting, clutter, and occlusion [9]–[11]. Most existing RGB-T density estimators rely on fully data-driven fusion strategies utilizing computationally heavy fusion networks [9]–[11]. Although effective, such approaches incur significant memory and computational overhead, limiting their applicability to real-time and edge-based deployment, where the availability of compute, memory, and power is limited.

We propose a computationally lightweight RGB-T fusion framework for density estimation. Our approach is based on a Mixture-of-Experts (MoE) formulation, where multiple experts selectively stream and contribute towards a fused representation [12], [13]. We incorporate recently introduced Sinkhorn attention mechanism to regulate cross-modal interactions under a doubly stochastic constraint [14]. This mechanism enforces structured one-to-one correspondences and controls how information is exchanged between modalities. Experiments on benchmarks demonstrate that the proposed approach achieves competitive performance while being significantly more compute efficient, making it well suited for real-time deployment.

## II. Related Work

RGB based density estimation methods have traditionally relied on regressing Gaussian density map [7], [8]. Density-based methods model spatial distributions in a differentiable manner and have emerged as the dominant regression paradigm owing to their superior accuracy and robustness under high-density conditions [15], [16].

To address the inherent limitations of RGB-only approaches, recent research has explored multi-modal sensor fusion. Lian et al. [17] introduced an RGB-Depth regression framework to mitigate underestimation caused by small head sizes and heavy

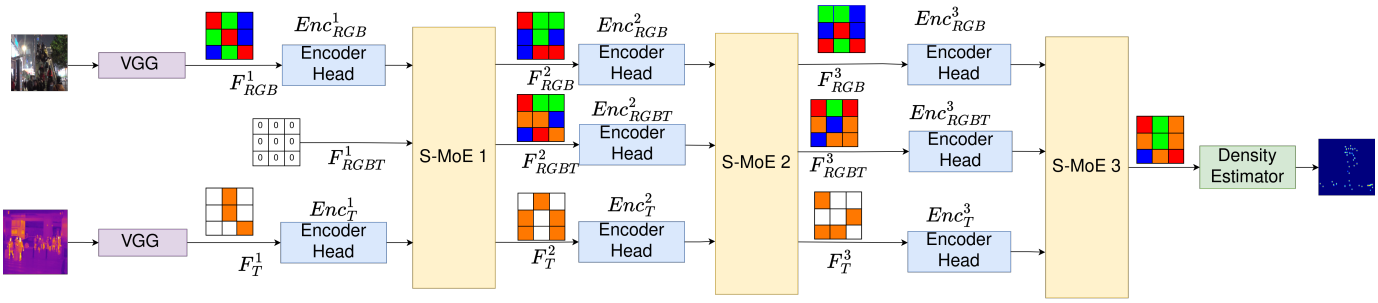


Fig. 1: RGB and thermal features obtained from the backbone network are fused using a sequence of S-MoE switches. Each S-MoE switch employs Differential Cross-Attention (DCA) and Sinkhorn normalization to extract and align modal features. A spatially adaptive gating module fuses the aligned features to generate a unified feature representation. The final fused representation is passed to a convolutional regressor to produce the density map.

occlusions. By incorporating depth cues, their method improves scale awareness and spatial consistency in dense scenes. Similarly, Liu et al. [1] proposed an RGB-Thermal density regression approach that leverages complementary thermal information to enhance robustness under challenging illumination and background conditions. Notably, this work was among the first to introduce an efficient bidirectional fusion framework for cross-modal density regression, enabling effective information exchange between RGB and thermal modalities. More recently, transformer-based architectures have gained significant traction in multi-modal density estimation due to their strong capability for modeling long-range dependencies and cross-modal interactions. Liu et al. [9] proposed a patch-based transformer framework employing count-guided tokens to facilitate multi-scale cross-modal feature interaction. Meng et al. [10] introduced a distillation based framework that extracts a third pseudo-modality from RGB-Thermal inputs using a diffusion model. Furthermore, Meng et al. [11] proposed a novel training strategy that enhances multi-modal density maps without introducing additional model parameters.

## III. Our Approach

The proposed architecture is illustrated in Fig. 1. It consists of a sequence of Sinkhorn Mixture-of-Experts (S-MoE) switches that progressively fuse RGB and thermal information. The network takes an RGB image and its corresponding thermal image as input and outputs a fused feature map. This feature map is then fed into density regressor to generate a spatial density map.

The overall network is organized into L stages, where each stage comprises of a single S-MoE switch. At each stage, modality-specific features are refined and selectively fused, enabling the network to incrementally build a robust multi-modal representation. We describe each component of the architecture in detail below.

### A. Fusion Network

**Feature Extractor.** Features are extracted using VGG backbone. Let $I_m$ denote the inputs to the network, where $m \in \{\mathrm{RGB}, \mathrm{T}\}$ corresponds to RGB and thermal modalities, respectively. The extracted feature maps are given by

$$\mathbf{F}_m = \mathrm{VGG}(I_m) \tag{1}$$

where $\mathbf{F}_m \in \mathbb{R}^{C\times H\times W}$. Here, $C$ denotes the feature dimensionality and $H \times W$ represents the spatial resolution.

**Encoder Heads.** Prior to fusion by the S-MoE switch, encoder heads are used to extract modality-specific representations. Each encoder is implemented using patch-based self-attention, which captures modality-specific information while maintaining computational efficiency. As illustrated in Fig. 1, we employ three encoders to process the RGB, thermal, and fused feature streams, respectively.

At stage $l$, let the input feature be denoted by $\mathbf{F}_m^{(l)}$, where $m \in \{\mathrm{RGB}, \mathrm{T}, \mathrm{RGBT}\}$. RGBT denotes the fused representation. The corresponding encoder outputs are computed as

$$\mathbf{E}_m^{(l)} = \mathrm{Enc}_m^{(l)}\left(\mathbf{F}_m^{(l)}\right) \tag{2}$$

where $\mathrm{Enc}_m^{(l)}$ represents the encoder corresponding to modality $m$, and $\mathbf{E}_m^{(l)} \in \mathbb{R}^{C\times H\times W}$. For the first stage ($l = 1$), the encoded fused representation is manually initialized to a zero tensor

$$\mathbf{E}_{\mathrm{RGBT}}^{(1)} = \mathbf{0}. \tag{3}$$

**S-MoE Switch.** Each Sinkhorn MoE switch consists of two components (i) Differential Cross-Attention (DCA) and, (ii) spatially adaptive modality fusion.

**Differential Cross-Attention.** This module is illustrated in Fig. 2. DCA identifies complementary cross-modal correspondences while suppressing information already captured in the fused stream. Instead of directly attending between modalities, DCA operates in a differential feature space, explicitly isolating information absent from the current fused representation. At stage $l$, the differential features are defined as

$$\mathbf{D}_{\mathrm{RGB}}^{(l)} = \mathbf{E}_{\mathrm{RGB}}^{(l)} - \mathbf{E}_{\mathrm{RGBT}}^{(l)}, \quad \mathbf{D}_{\mathrm{T}}^{(l)} = \mathbf{E}_{\mathrm{T}}^{(l)} - \mathbf{E}_{\mathrm{RGBT}}^{(l)}. \tag{4}$$

At $l = 1$, since $\mathbf{E}_{\mathrm{RGBT}}^{(1)} = \mathbf{0}$, the differential features reduce to the encoder outputs, and the first S-MoE switch directly fuses the raw modality features. The differential feature maps are

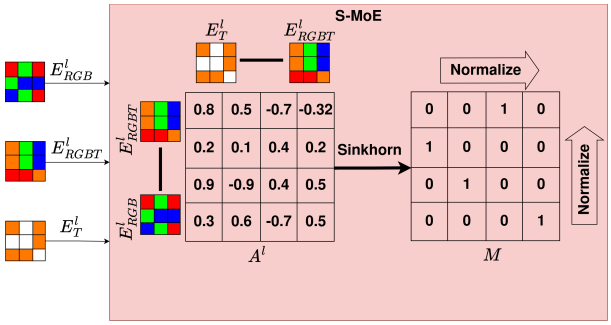


Fig. 2: Illustration of the proposed Differential Cross-Attention (DCA) module. Pairwise attention scores are computed on differential RGB and thermal features, followed by normalization using the Sinkhorn operator.

flattened into $N = H \times W$ spatial tokens, yielding matrices in $\mathbb{R}^{N\times C}$. The cross-modal affinity matrix is then computed as

$$\mathbf{A}^{(l)} = \frac{\mathbf{D}^{(l)}_{\mathrm{RGB}} \left(\mathbf{D}^{(l)}_{\mathrm{T}}\right)^{\top}}{\sqrt{C}} \in \mathbb{R}^{N\times N}, \tag{5}$$

where $\sqrt{C}$ is the standard attention scaling factor.

Instead of conventional softmax normalization, we apply the Sinkhorn operator to obtain a doubly stochastic attention matrix

$$\mathbf{M}^{(l)} = \mathrm{Sinkhorn}\left(\mathbf{A}^{(l)}\right). \tag{6}$$

As explained in Section IV, the Sinkhorn normalization enforces one-to-one alignment by enforcing both row-wise and column-wise constraints. The cross-attended differential features are computed as

$$\tilde{\mathbf{D}}^{(l)}_{\mathrm{RGB}} = \mathbf{M}^{(l)}\mathbf{D}^{(l)}_{\mathrm{T}}, \quad \tilde{\mathbf{D}}^{(l)}_{\mathrm{T}} = \left(\mathbf{M}^{(l)}\right)^{\top}\mathbf{D}^{(l)}_{\mathrm{RGB}}. \tag{7}$$

Here, $\tilde{\mathbf{D}}^{(l)}_{\mathrm{RGB}}$ captures complementary information from the thermal modality relevant to RGB features, while $\tilde{\mathbf{D}}^{(l)}_{\mathrm{T}}$ captures complementary RGB information relevant to thermal features.

**Modality Fusion.** We perform spatially adaptive modality fusion, where the contribution of each modality is independently determined at every spatial location. The cross-attended differential features are first concatenated along the channel dimension

$$\tilde{\mathbf{D}}^{(l)} = \tilde{\mathbf{D}}^{(l)}_{\mathrm{RGB}} \oplus \tilde{\mathbf{D}}^{(l)}_{\mathrm{T}} \in \mathbb{R}^{2C\times H\times W}, \tag{8}$$

where $\oplus$ denotes channel-wise concatenation. Then a lightweight gating network, implemented using a $1 \times 1$ convolution followed by a softmax operation predicts the spatial weights

$$\boldsymbol{\alpha}^{(l)} = \mathrm{Softmax}\left(\mathrm{Conv}_{1\times 1}\left(\tilde{\mathbf{D}}^{(l)}\right)\right), \tag{9}$$

where

$$\boldsymbol{\alpha}^{(l)} = \left[\alpha^{(l)}_{\mathrm{RGB}}, \alpha^{(l)}_{\mathrm{T}}\right] \in \mathbb{R}^{2\times H\times W}. \tag{10}$$

**Algorithm 1:** Sinkhorn Normalization

**Input :** Attention matrix $A$, number of iterations T
**Output:** Doubly stochastic matrix $M$
$M \leftarrow \exp(A)$;
**for** $t \leftarrow 1$ **to** $T$ **do**
  // Row normalization
  $M_{ij} \leftarrow \frac{M_{ij}}{\sum_j M_{ij}}$;
  // Column normalization
  $M_{ij} \leftarrow \frac{M_{ij}}{\sum_i M_{ij}}$;
**return** $M$;

The fused representation for the next stage is computed as

$$\mathbf{F}^{(l+1)}_{\mathrm{RGBT}} = \alpha^{(l)}_{\mathrm{RGB}} \odot \tilde{\mathbf{D}}^{(l)}_{\mathrm{RGB}} + \alpha^{(l)}_{\mathrm{T}} \odot \tilde{\mathbf{D}}^{(l)}_{\mathrm{T}} + \mathbf{F}^{(l)}_{\mathrm{RGBT}}, \tag{11}$$

where $\odot$ denotes element-wise multiplication.

### B. Density Regression

The fused representation produced by the final S-MoE switch is passed to a lightweight regressor to generate the output spatial density map $\mathbf{DM}$

$$\mathbf{DM} = \mathrm{Conv}\left(\mathbf{F}^{(L)}_{\mathrm{RGBT}}\right), \tag{12}$$

where $\mathbf{DM} \in \mathbb{R}^{1\times H\times W}$.

### C. Loss Function

We adopt the Bayesian Loss [10], [18] to align the regressed density map with the discrete ground-truth point annotations. The loss is defined as

$$\mathcal{L}_c = \sum_{i=1}^{M} \left| 1 - \left\langle \frac{\mathcal{N}\left(z_i, \sigma^2\mathbf{I}_{2\times 2}\right)}{\sum_{n=1}^{M}\mathcal{N}\left(z_n, \sigma^2\mathbf{I}_{2\times 2}\right)}, \mathbf{DM} \right\rangle \right|, \tag{13}$$

where $M$ is the number of annotated points, $z_i$ denotes the $i$-th annotation, and $\langle\cdot,\cdot\rangle$ denotes the inner product between Gaussian kernels and predicted density map $\mathbf{DM}$.

## IV. Sinkhorn Alignment

The Sinkhorn operator is a differentiable, iterative normalization procedure that transforms an attention matrix $\mathbf{A}$ into a doubly stochastic matrix $\mathbf{M}$ such that the rows and columns of the $\mathbf{M}$ sum to one.

$$\mathbf{M1} = \mathbf{1}, \qquad \mathbf{M}^{\top}\mathbf{1} = \mathbf{1}, \tag{14}$$

where $\mathbf{1}$ denotes a vector of ones of appropriate dimension.

Algorithm 1 present the Sinkhorn normalization procedure. The algorithm takes as input the attention matrix and a predefined number of iterations. The number of iterations determines the extent of normalization. The input attention matrix is first exponentiated to ensure all entries are strictly positive and then the resulting matrix is iteratively normalized. Each iteration is a two step process comprising of row normalization followed by column normalization.

The Sinkhorn attention offers two major advantages over softmax-based attention normalization [14], [23], [24] (i) Each

TABLE I: Performance on RGBT-CC

| Method | RMSE | GAME(3) | GAME(2) | GAME(1) | GAME(0) |
|---|---|---|---|---|---|
| BL+IADM [1] | 28.18 | 32.89 | 24.69 | 19.95 | 15.61 |
| MSDTrans [9] | 18.79 | 26.14 | 19.02 | 14.81 | 10.90 |
| DEFNet [19] | 21.09 | 27.27 | 20.19 | 16.08 | 11.90 |
| MC3Net [20] | 20.59 | 27.95 | 19.40 | 15.06 | 11.47 |
| CGINet [21] | 20.54 | 27.73 | 20.06 | 15.98 | 12.07 |
| Broker Modality [10] | 17.32 | 23.64 | 17.65 | 13.61 | 10.19 |
| Free Lunch Counting [11] | 17.05 | 21.85 | 16.27 | 12.62 | 9.57 |
| RGB + Thermal + MoE + Softmax | 17.89 | 21.22 | 16.75 | 12.76 | 10.02 |
| **RGB + Thermal + MoE + Sinkhorn** | **16.53** | **20.98** | **15.75** | **11.98** | **9.12** |

TABLE II: Performance of DroneRGBT

| Method | RMSE | GAME(0) |
|---|---|---|
| BL+IADM [1] | 17.54 | 11.41 |
| CSRNet [22] | 13.80 | 8.91 |
| CGINet [21] | 13.45 | 8.37 |
| MC3Net [20] | 11.17 | 7.33 |
| Broker Modality [10] | 10.04 | 6.20 |
| Free Lunch Count [11] | 7.81 | 4.94 |
| RGB + Thermal + MoE + Softmax | 8.84 | 6.90 |
| **RGB + Thermal + MoE + Sinkhorn** | **7.21** | **4.51** |

attention score in $\mathbf{A}$ is normalized not just by its rows elements but also by the column elements. Over multiple iterations, this results in a global normalization scheme that aligns the two modalities. (ii) The doubly stochastic constraint encourages one-to-one correspondences between the feature pixels, reducing many-to-one or one-to-many alignments that often arise with row-wise softmax normalization. Together, these properties make the Sinkhorn attention well suited for cross-modal alignment, where redundant or ambiguous correspondences can significantly degrade fusion quality.

## V. Experiments

We evaluate our network on two datasets, RGBT-CC and DroneRGBT [1], [2]. Consistent with prior work, we use GAME and RMSE as evaluation metrics [10], [25]. The network is trained using the AdamW optimizer, with a learning rate of $1 \times 10^{-4}$ for the encoder heads and $1 \times 10^{-5}$ for the backbone [26]. The training batch size is set to 16. We use an end-to-end encoder embedding dimension of 256. The number of Sinkhorn iterations is set to 5. During training, we extract random patches of size $256 \times 256$ from the input images and apply horizontal flipping with a probability of 0.5.

### A. Results

We present the performance of our model in tables I and II. Visual results are illustrated in Fig. 3. Compared with transformer-based approaches such as MSDTrans [9], CGINet [21], and MC$^3$Net [20], our approach achieves superior performance without relying on complex attention mechanisms. Even though our approach utilizes attention mechanism, existing transformer-based methods depend on multiple layers of high-dimensional query-key-value projections and token-level attention, resulting in substantial computational and memory overhead.

We also observe consistent performance improvements over two recent methods, Broker Modality and Free Lunch Enhancement (FLE) [10], [11]. both of which employ computationally intensive architectures. The Broker Modality framework introduces a third pseudo-modality in addition to RGB and thermal inputs to align sensor signals. However, in this design modality fusion is performed at the very end of the network which can limit effective cross-modal interaction. The FLE approach builds upon the Broker Modality framework by incorporating additional pre-training and post-training strategies to improve convergence and performance. While these enhancements benefit optimization, FLE retains the same underlying architecture as Broker Modality and therefore inherits its computational complexity and architectural limitations.

RGB Thermal Density Map

Fig. 3: Visual illustration of RGB and thermal inputs along with the spatial density map predicted by S-MoE.

TABLE III: Comparison of model size, computational cost.

| Method | Backbone (Millions) | Reg. Encoder (Millions) | Total Params (Millions) | GFLOPs | GPU-Mem (MB) |
|---|---|---|---|---|---|
| MC3Net [20] | 60 | 53 | 113 | 609 | 1000 |
| DEFNet [19] | 20 | 25 | 45 | 470 | 800 |
| Broker Modality [10] | 19 | 21 | 40 | 451 | 714 |
| **Ours (S-MoE)** | **14** | **10** | **24** | **245** | **301** |

In contrast, our approach adopts a simple yet effective Mixture-of-Experts strategy. At the first switch, sensor information is fused directly, while subsequent switches incrementally accumulate complementary information learned at deeper stages. This hierarchical fusion strategy ensures that each stage contributes novel information beyond the existing fused representation.

In Table III we compare the model size and computational cost of our approach with existing methods. All GFLOPs are measured for a single $512 \times 512$ image. Our model has approximately 40% fewer parameters and requires nearly 50% fewer GFLOPs. The reduction in computational requirements can be attributed to two key design choices. First, unlike existing methods, we do not employ pseudo-modalities to align RGB and thermal features, instead we utilize Sinkhorn normalization for cross-modal alignment. Second, we adopt a deep but lightweight regression head comprising 256-dimensional feature maps.

### B. Edge Device Experiments

Our edge inference setups are illustrated in Fig. 4. The acquisition system comprises co-located RGB and thermal cameras, which can be mounted on either a smart pole or a drone. The sensors capture images at 30 FPS. All processing is performed on-device using a Jetson Nano.

**FPS measurement.** From Fig. 5. We observe that S-MoE achieves a substantially higher processing speed, with performance approaching 30 FPS, closely matching the acquisition

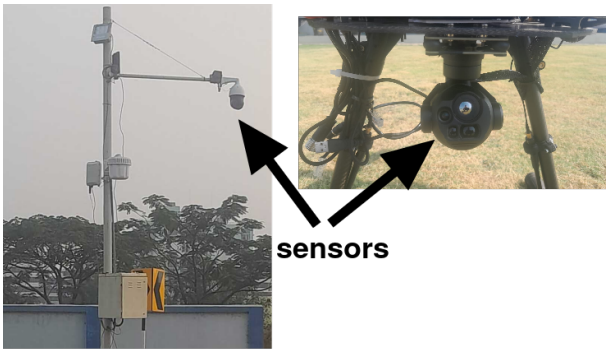


Fig. 4: A smart pole (left) and a drone (right) equipped with multi-modal sensors. Data processing is performed directly on-device. For the smart pole, inference runs on a Jetson compute box located near the base, while for the drone, it is executed on an onboard Jetson module.

rate of the sensors. In contrast, the Broker model exhibits significantly slower inference, achieving an effective throughput of approximately 8–12 FPS, which makes it unsuitable for practical real-time applications.

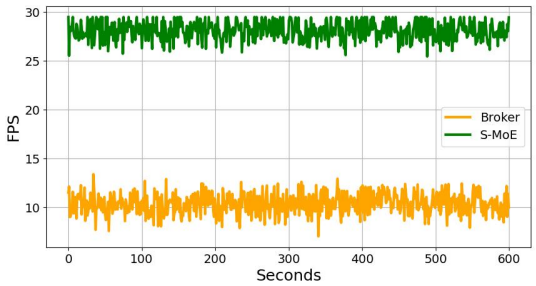


Fig. 5: FPS comparison of density estimators

**Drone flight time.** We evaluate the flight time of each method on a low-powered drone. The results are presented in Table IV. We observe that S-MoE provides additional flight time compared to the broker-based model. This improvement stems from reduced computational overhead, as S-MoE requires fewer parameters and FLOPs, leading to more efficient on-device computation.

TABLE IV: Flight time comparison across fusion methods.

| Method | Flight Time (min) |
|---|---|
| Broker Model | $12.9 \pm 0.5$ |
| S-MoE | $15.1 \pm 0.3$ |

## VI. Ablation Studies

**Impact of Sinkhorn.** From Tables I and II, we observe that replacing Sigmoid attention with Sinkhorn in the MoE switch leads to better model performance. Furthermore, in Fig. 6, we visualize the attention maps for captured RGB-T images. We observe that Softmax-based activation tends to concentrate attention within a single dominant region. In contrast,

RGB Thermal Softmax Sinkhorn

Fig. 6: Comparison of attention maps obtained using Softmax and Sinkhorn normalization.

Sinkhorn-based activation produces more distinct and well-structured attention maps, even in blurred and overlapping scenes, leading to more accurate detection and counting. We attribute this to the doubly stochastic constraint enforced by Sinkhorn normalization.

TABLE V: Impact of Multi-Modal Inputs

| | RGBT-CC | | DroneRGBT | |
|---|---|---|---|---|
| Method | GAME(0)↓ | RMSE↓ | GAME(0)↓ | RMSE↓ |
| RGB only | 14.23 | 25.25 | 10.25 | 15.45 |
| Thermal only | 18.23 | 30.52 | 12.56 | 20.12 |
| RGB + Thermal | 11.54 | 19.50 | 7.95 | 11.53 |

**Impact of Multi-Modal Fusion.** We analyze the impact of multi-modal information on density estimation in Table V. The results show that incorporating multi-modal inputs leads to improved performance.

TABLE VI: Impact of the number of S-MoE switches.

| Stages | GAME(0) | GAME(1) | GAME(2) | GAME(3) | RMSE |
|---|---|---|---|---|---|
| 3 | 9.60 | 12.74 | 16.45 | 21.98 | 17.21 |
| 4 | 9.12 | 11.98 | 15.75 | 20.98 | 16.53 |
| 5 | 9.91 | 12.91 | 16.66 | 22.54 | 17.89 |
| 6 | 9.65 | 13.01 | 16.42 | 21.67 | 17.76 |

**Impact of number of stages.** From Table VI, it is clear that increasing the number of stages beyond a certain point does not yield further performance gains and instead only increases the computational cost.

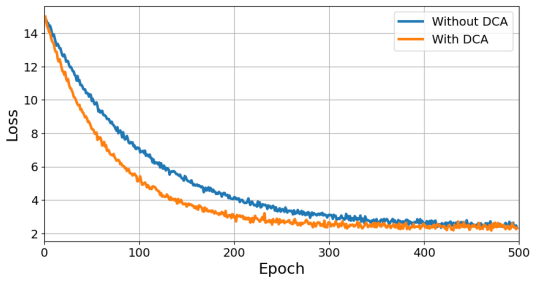


Fig. 7: Loss curve with and without DCA.

**Impact of DCA.** Instead of using Differential Cross-Attention (DCA), we directly compute attention between input modality representations. The results are presented in Fig. 7. The results indicate that both DCA and non-DCA networks converge to similar values. However, DCA achieves faster convergence. This improvement arises because DCA explicitly focuses on incorporating novel and complementary information from both

modalities into the fused representation, rather than redundantly reinforcing already shared features.

## VII. Conclusion

In this work, we introduce a lightweight and efficient RGB-Thermal fusion framework for spatial density estimation. Our approach integrates a Sinkhorn-based Mixture-of-Experts mechanism to enable adaptive modality selection and alignment while maintaining low computational and memory overhead. Experiments on public datasets demonstrate that our method achieves competitive performance while being significantly more memory and compute efficient. These results highlight the effectiveness of our design choices and make the proposed framework well suited for real-time deployment in resource-constrained environments.

## Acknowledgment

This work is supported by DST National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS) and Technology Innovation Hub on Autonomous Navigation (TIHAN) Foundation at Indian Institute of Technology Hyderabad.

## References

[1] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4823–4833, 2021.

[2] T. Peng, Q. Li, and P. Zhu, "Rgb-t crowd counting from drone: A benchmark and mmccn network," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[3] O. Shalash, A. Emad, F. Fathy, A. Alzogby, M. Sallam, E. Naser, M. El-Sayed, and E. Khatab, "Fusion of robotics, ai, and thermal imaging technologies for intelligent precision agriculture systems," *Sensors (Basel, Switzerland)*, vol. 25, no. 22, p. 6844, 2025.

[4] R. Ishimwe, K. Abutaleb, F. Ahmed, *et al.*, "Applications of thermal imaging in agriculture—a review," *Adv. Remote Sens*, vol. 3, no. 03, pp. 128–140, 2014.

[5] E. Ring and K. Ammer, "Infrared thermal imaging in medicine," *Physiological measurement*, vol. 33, no. 3, pp. R33–R46, 2012.

[6] J. A. Ramirez-Nunez, L. A. Morales-Hernandez, R. A. Osornio-Rios, J. A. Antonino-Daviu, and R. J. Romero-Troncoso, "Self-adjustment methodology of a thermal camera for detecting faults in industrial machinery," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*, pp. 7119–7124, IEEE, 2016.

[7] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.

[8] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Advances in neural information processing systems*, vol. 23, 2010.

[9] Z. Liu, W. Wu, Y. Tan, and G. Zhang, "RGB-T Multi-Modal Crowd Counting Based on Transformer," in *Procedings of British Machine Vision Conference*, pp. 1–14, 2022.

[10] H. Meng, X. Hong, C. Wang, M. Shang, and W. Zuo, "Multi-modal crowd counting via a broker modality," in *European Conference on Computer Vision*, pp. 231–250, Springer, 2024.

[11] H. Meng, X. Hong, Z. Lai, and M. Shang, "Free lunch enhancements for multi-modal crowd counting," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14013–14023, 2025.

[12] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017.

[13] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[14] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré, "Sinkformers: Transformers with doubly stochastic attention," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics* (G. Camps-Valls, F. J. R. Ruiz, and I. Valera, eds.), vol. 151 of *Proceedings of Machine Learning Research*, pp. 3515–3530, PMLR, 28–30 Mar 2022.

[15] Y. Ranasinghe, N. G. Nair, W. G. C. Bandara, and V. M. Patel, "Crowddiff: Multi-hypothesis crowd density estimation using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12809–12819, 2024.

[16] T. Han, L. Bai, L. Liu, and W. Ouyang, "Steerer: Resolving scale variations for counting and localization via selective inheritance learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21848–21859, 2023.

[17] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1821–1830, 2019.

[18] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6142–6151, 2019.

[19] W. Zhou, Y. Pan, J. Lei, L. Ye, and L. Yu, "Defnet: Dual-branch enhanced feature fusion network for rgb-t crowd counting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24540–24549, 2022.

[20] W. Zhou, X. Yang, J. Lei, W. Yan, and L. Yu, "Mc 3 net: Multimodality cross-guided compensation coordination network for rgb-t crowd counting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 5, pp. 4156–4165, 2023.

[21] Y. Pan, W. Zhou, X. Qian, S. Mao, R. Yang, and L. Yu, "Cginet: Cross-modality grade interaction network for rgb-t crowd counting," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106885, 2023.

[22] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, 2018.

[23] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *The annals of mathematical statistics*, vol. 35, no. 2, pp. 876–879, 1964.

[24] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, vol. 21, no. 2, pp. 343–348, 1967.

[25] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Onoro-Rubio, "Extremely overlapping vehicle counting," in *Iberian conference on pattern recognition and image analysis*, pp. 423–431, Springer, 2015.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.