SM3Det: A Unified Model for Multi-Modal Remote Sensing Object Detection

Yuxuan Li, Xiang Li†, Yunheng Li, Yicheng Zhang, Yimian Dai, Qibin Hou, Ming-Ming Cheng, Jian Yang †

PCA Lab, VCIP, Computer Science, NKU
yuxuan.li.17@ucl.ac.uk, {yunhengli, zhangyc}@mail.nankai.edu.cn, {xiang.li.implus, yimian.dai, houqb, cmm, csjyang}@nankai.edu.cn

Abstract

With the rapid advancement of remote sensing technology, high-resolution multi-modal imagery is now more widely accessible. Conventional object detection models are trained on a single dataset, often restricted to a specific imaging modality and annotation format. However, such an approach overlooks the valuable shared knowledge across multi-modalities and limits the model's applicability in more versatile scenarios. This paper introduces a **new task** called Multi-Modal Datasets and Multi-Task Object Detection (M2Det) for remote sensing, designed to accurately detect horizontal or oriented objects from any sensor modality. This task poses challenges due to 1) the trade-offs involved in managing multi-modal modelling and 2) the complexities of multi-task optimization. To address these, we establish a benchmark dataset and propose a unified model, **SM3Det** (**S**ingle **M**odel for **M**ulti-Modal datasets and **M**ulti-Task object **Det**ection). SM3Det leverages a grid-level sparse MoE backbone to enable joint knowledge learning while preserving distinct feature representations for different modalities. Furthermore, we propose a novel consistency and synchronization optimization mechanism, allowing it to effectively handle varying levels of learning difficulty across modalities and tasks. Extensive experiments demonstrate SM3Det's effectiveness and generalizability, consistently outperforming the combination of specialized models on individual datasets.

Code — github.com/zcablii/SM3Det
Datasets — www.kaggle.com/datasets/greatbird/soi-det
Extended Version — https://arxiv.org/pdf/2412.20665

Introduction

Remote sensing object detection (Yuan et al. 2025; Ni et al. 2025; Li et al. 2025, 2024a; Dai et al. 2024) typically involves multiple sensors employing different imaging mechanisms, resulting in diverse data modalities. Traditionally, detection models are developed for specific datasets associated with a single modality and a predefined format detection task (Li et al. 2024b; Yang et al. 2021; Dai et al. 2021), as shown in Figure 1 (b). This conventional approach overlooks the valuable and inherent joint knowledge within a unified remote sensing context. Furthermore, airborne platforms such as UAVs and satellites often carry multiple sensors, making it critical to process images from various modalities simultaneously. Previous multi-source object detection methods (Liu, Chen, and Wang 2021; Zhang, Huang, and Kuruoglu 2024; Zhang et al. 2024) have heavily relied on scarce, impractical, and inflexible spatially well-aligned paired images and spatial alignment algorithms (Devaraj and Shah 2013; Ahamed et al. 2012). These methods are also limited to performing single-format detection tasks, as depicted in Figure 1 (a). Thus, it is essential to develop a unified model capable of handling all modalities without requiring spatially aligned image pairs and performing multiple format detection tasks (referred to as "multi-tasks" throughout the paper), which is not thoroughly studied.

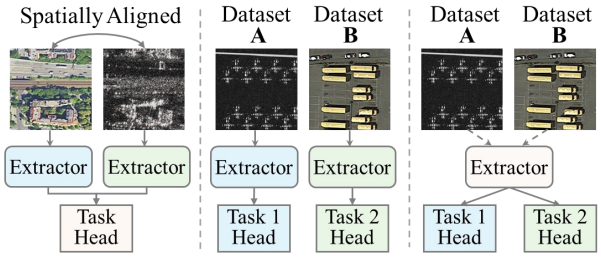

Figure 1: Comparison of tasks: (a) Spatially Aligned Multi-Modality, (b) Traditional Single Dataset, and (c) M2Det. M2Det aims to utilize a unified model for detecting objects in any modality, handling various detection tasks.

To fill this research gap, we propose a new task called Multi-Modal Datasets and Multi-Task Object Detection (M2Det). M2Det aims to detect objects in any given image, regardless of its modality, and across predefined detection tasks—whether horizontal bounding boxes or oriented bounding boxes—as illustrated in Figure 1 (c).

The M2Det task is closely related to two key research areas: multi-dataset object detection (Wang et al. 2019; Zhou, Koltun, and Krähenbühl 2022) and multi-task learning (Zhang and Yang 2021; Chen et al. 2018). However, the M2Det task presents unique challenges. In traditional multi-dataset object detection, even though images may have different attributes—such as natural images and paintings—they often share similar underlying concepts (optical

†Corresponding authors.

concepts). A simple joint training approach is effective, with a single model trained on the combined dataset typically outperforming models trained on individual datasets (Wang et al. 2019). In contrast, multi-modal datasets in remote sensing—such as RGB (Xia et al. 2018; Sun et al. 2022a), SAR (Li et al. 2024c; Zhang et al. 2021), IR (Sun et al. 2022b), and multi-spectral images (for Photogrammetry and ISPRS)—exhibit fundamentally different pattern concepts (as in Figure 6). While certain common knowledge may be shared across these modalities, the significant differences in data representation create a substantial modality gap, complicating the integration of information across modalities. Additionally, remote sensing datasets often include diverse annotation types, such as horizontal (Li et al. 2020, 2024c) and oriented (Xia et al. 2018; Sun et al. 2022b) bounding boxes, further adding complexity to model learning.

These challenges may impede traditional model learning and optimization in the following ways: **1)** Representation Constraints: A dense model that shares the same parameters across multiple tasks and modalities may encounter limitations in representation capacity, as a single set of parameters may struggle to effectively fit the diverse distributions inherent in each dataset. **2)** Optimization Inconsistencies: The varying learning difficulties across different modalities and tasks can lead to unsynchronized optimization rates or optimization directions for various components of the model. This inconsistency can result in conflicting optimization outcomes, adversely affecting the model's ability to achieve different loss objectives.

To address these challenges, we first establish a comprehensive benchmark dataset by merging SARDet-100K (Li et al. 2024c), DOTA (Xia et al. 2018), and DroneVehicle (Sun et al. 2022b), which collectively span SAR, optical, and infrared modalities. Subsequently, we propose a unified model, SM3Det, tailored for the M2Det task in remote sensing, addressing the challenges from both model architecture and model optimization perspectives:

Model Architecture: We propose integrating a plug-and-play grid-level sparse Mixture of Experts (MoE) architecture into backbone networks, enabling the model to capture both shared knowledge and modality-specific representations. In contrast to prior multi-dataset object detection models that use hard-coded, image-level routing (Wang et al. 2019; Jain et al. 2024), our approach introduces grid-level experts with dynamic routing. These experts operate on spatial grid features, allowing the model to adaptively process information at a grid level, which is crucial for object detection tasks.

Model Optimization: We introduce a novel dynamic submodule optimization (DSO) mechanism for model optimization consistency and synchronization. It adaptively adjusts the learning rates of various network components based on tailored policies. DSO accommodates the varying learning complexities across different tasks and modalities by balancing the relative convergence rate and guaranteeing optimization direction consistency. Unlike traditional techniques that primarily modify loss weights or gradients—often lacking precise manipulation over specific network submodules or suffering from inefficiencies—our DSO provides fine-grained control while maintaining optimization efficiency.

Intensive experiments indicate that our unified single SM3Det model significantly outperforms individual models across all modality datasets. Our lightweight SM3Det variant not only demonstrates excellent performance but also features a substantially reduced number of parameters. Furthermore, the SM3Det model exhibits strong generalizability, enabling it to adapt to various backbones and detectors. Our contributions are summarized as follows:

- We introduce a new task: Multi-Modal Datasets and Multi-Task object detection in remote sensing using a unified detection model.
- We propose the SM3Det model, which addresses the challenges of the M2Det task by offering innovative solutions from both architecture and optimization perspectives.
- Extensive experiments and analyses on the established benchmark dataset demonstrate that our proposed single model is effective and outperforms individual models across all modalities.

Related Work

Multi-Dataset Object Detection

Multi-dataset object detection aims to leverage a diverse collection of datasets to learn general knowledge and achieve universal object detection. Leveraging multiple datasets in training has proven to be a highly effective strategy for enhancing the performance of deep learning models across various applications (Kapidis, Poppe, and Veltkamp 2021; Zhao et al. 2020; Zhang et al. 2025; Ye et al. 2025). This approach has also been widely explored in the domain of object detection. The DA network (Wang et al. 2019), for instance, employs specialized SE layers (Hu, Shen, and Sun 2018) that serve as domain-specific attention mechanisms for individual datasets. Universal-RCNN (Xu et al. 2020) introduces a partitioned detector trained across multiple datasets, integrating features through an inter-dataset graph-based attention module. Unidet (Zhou, Koltun, and Krähenbühl 2022) advances this concept by proposing a unified label space and underscoring the importance of batch sampling strategies.

Models trained on combined optical-concept datasets typically outperform those trained on individual datasets, as multi-dataset training can serve as a powerful form of data augmentation. However, the diverse imaging modalities in remote sensing present unique challenges for joint training. This area remains largely unexplored.

Multi-Task Learning

Multi-task learning involves utilizing a single model to learn multiple objectives, typically with multiple task heads and loss functions. In multi-task learning, various strategies (Chen et al. 2018; Sener and Koltun 2018; Guo et al. 2018; Kendall, Gal, and Cipolla 2018) have been developed to address task imbalances and optimize learning outcomes. GradNorm (Chen et al. 2018) focuses on correcting gradient imbalances during backpropagation by adjusting the gradient sizes for each task's loss function. Methods like Multi-Gradient Descent Algorithm (Sener and Koltun 2018) employ Pareto optimization for gradient backpropagation,

though they can be inefficient due to the additional gradient calculations required. Similar to GradNorm, DWA (Wang et al. 2019) also uses task losses to assess convergence rates, however, it dynamically adjusts the weight of each task's loss instead. Uncertainty (Kendall, Gal, and Cipolla 2018) loss takes a different approach by incorporating homoscedastic uncertainty into the weighted loss function.

Unlike loss reweighting or gradient manipulation, our method dynamically adjusts the learning rate for network submodules, enhancing multi-modal datasets and multi-task learning by maintaining optimization consistency.

Methods

Task Definition

The proposed M2Det task is designed to utilize a unified model for detecting objects in images from any modality, handling various predefined detection tasks, such as horizontal and rotated bounding boxes. The significance of this task is evident in various real-world applications, including low-altitude economy (Jiang et al. 2023; Huang et al. 2024), aerial surveillance (Avola et al. 2021; Bozcan and Kayacan 2020), earth observation (Li et al. 2017; Anderson et al. 2017), and other research domains (Khan, Yanmaz, and Rinner 2014; Jensen 2016). For instance, platforms equipped with M2Det models can fully leverage available multi-modal data while benefiting from simplified version control and the seamless integration of multiple sensors without requiring model updates on the device. This significantly reduces model maintenance costs in industrial applications. Furthermore, processing images of different modalities in a single model within one mini-batch maximizes the parallel computing capabilities of GPUs, thereby enhancing computational and energy efficiency on edge devices.

Methodological Overview

The overall network architecture follows the classic design of multi-task learning models (Wang et al. 2019; Zhou, Koltun, and Krähenbühl 2022). It consists of a relatively heavy feature space shared component (backbone) and relatively lightweight feature space independent components (task heads). The backbone is responsible for joint representation learning, with most parameters being shared, thus ensuring parameter efficiency. The lightweight heads are separated to accommodate distinct features and task learning. However, as discussed in Section , modality and task gaps may degrade the performance of such classic multi-task models. To address this issue, we propose the SM3Det model, which consists of two parts:

Model architecture: A sparse MoE backbone where experts are activated on local image features of multi-modality dataset images at the grid level.

Model optimization: An efficient dynamic submodule optimization mechanism, to handle the varying learning difficulties and optimization inconsistency across multiple tasks and modalities.

Grid-level MoE

Previous approaches to multi-dataset object detection (Zhou, Koltun, and Krähenbühl 2022; Xu et al. 2020) utilize dense models that leverage shared concepts among datasets to enhance joint knowledge representation. In the case of multi-modal remote sensing images, this joint knowledge also exists (Li et al. 2024c), though it may be less explicit, with common weak cues such as shape and scale across modalities. However, due to inherent modality and task gaps, employing a dense model that utilizes the same parameters across multiple tasks and modalities can result in a congested feature/representation space, ultimately reducing the model's expressiveness. Therefore, it is essential to explore methods that leverage joint knowledge across modalities while enabling distinct representation learning for each modality to prevent feature space interference.

Drawing inspiration from the success of Sparse MoE networks (Shazeer et al. 2017), which are characterized by their sparsity and high capacity, we propose leveraging MoE for the M2Det task. For transformer-based backbones (Liu et al. 2021; Wang et al. 2022), we integrate MoE experts within the FFN components. For modern CNNs (Liu et al. 2022; Guo et al. 2022; Li et al. 2023), which often employ 1×1 convolutions (Lin 2013) for feature interaction or dimensionality reduction/expansion, we introduce sparse experts to enhance these layers. Unlike previous transformer-based detectors that route an entire image's features through a single expert (Jain et al. 2024), our design allows experts to operate on local grid features within the backbone. This approach ensures that experts process similar spatial patterns across modalities, facilitating shared representation learning. Simultaneously, multiple experts capture distinct patterns across modalities, enabling independent representation learning. Specifically, for the local spatial input feature x_{ij} at the i-th row and j-th column of a deep image feature, the output feature $f_{MoE}(x_{ij})$ after the MoE layer is:

$$f_{MoE}(x_{ij}) = \sum_{n=1}^{N} G_n(x_{ij}) \cdot Conv_n^{1\times1}(x_{ij}), \tag{1}$$

$$G(x_{ij}) = \text{TOP}_k\left(\text{Softmax}\left(\frac{E^T W x_{ij}}{\tau \|W x_{ij}\| \|E\|}\right)\right), \tag{2}$$

where N is the total number of experts, G is the gating function and $Conv_n^{1\times1}$ is the n-th 1×1 convolutional expert. Each expert has a representation embedding in the matrix E. The input feature x is first transformed by the matrix W. The product of Wx is then compared with each expert embedding in E to calculate the similarity. This comparison is then normalized by the product of the norms of Wx and E, ensuring scale-invariance. The similarity scores are passed through a $Softmax$ function, converting them into a probability distribution. This means the gating function assigns a probability to each expert, indicating its relevance to the input feature x. Finally, the TOP_k operator selects the top-k experts with the highest probabilities. It reweights each expert by assigning the $Softmax$ probability to the top-k experts, setting the rest to zero. This step sparsifies the model by focusing only on a small subset of experts, reduc-

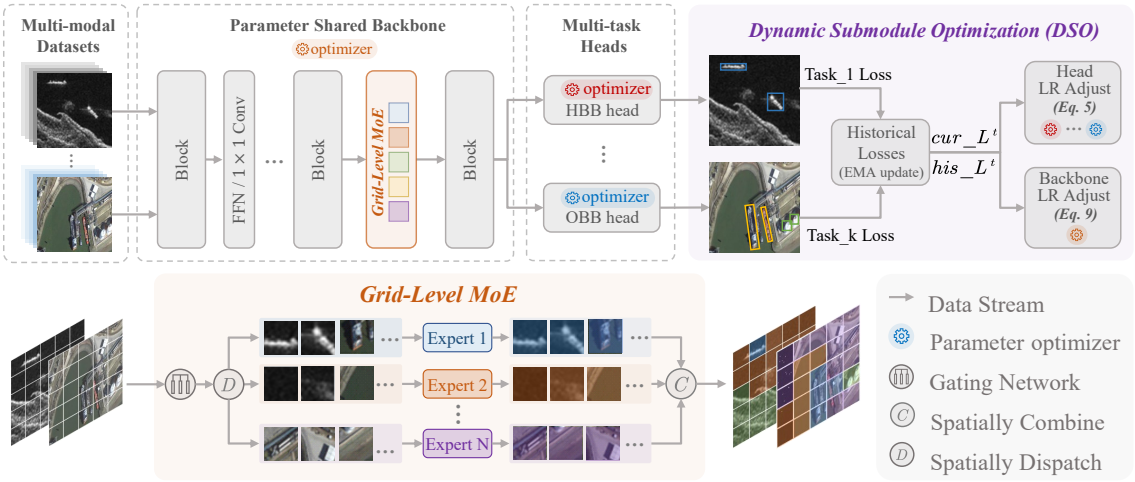

Figure 2: A conceptual illustration of SM3Det model. "HBB": horizontal bounding box, "OBB": oriented bounding box.

ing computational complexity and enhancing the model's expressiveness to handle diverse tasks and modalities.

In summary, $f_{MoE}(x_{ij})$ is a weighted sum of the outputs from top-k experts. The weights are determined by the gating function G, which dynamically selects the most relevant expert(s) for each local feature. The MoE creates a sparser feature space in the backbone model. By focusing on local patterns, the model can learn independently to model multiple modalities and local object patterns. Our design effectively addresses the challenges of crowded feature spaces and enhances the expressiveness of the model.

In practical implementation, to fully utilize the pretrained backbone weights, we initialize the weights of added experts by duplicating the corresponding pretrained 1×1 convolutional layers' weights before downstream model fine-tuning, ensuring all experts can be evenly chosen at the beginning of fine-tuning. For the task heads, we maintain simplicity and adhere to the existing design of task heads as in (Wang et al. 2019; Zhou, Koltun, and Krähenbühl 2022).

Dynamic Submodule Optimization (DSO)

In multi-modal, multi-dataset, and multi-task object detection tasks, one primary challenge is the varying learning difficulties (Kendall, Gal, and Cipolla 2018; Chen et al. 2018) across modalities and tasks. To address this problem, we propose a novel Dynamic Submodule Optimization (DSO) mechanism to manage the differing learning difficulties across tasks and modalities.

DSO takes each task head's loss as indicator to determine the current convergence rate of each task and the overall optimization direction of the network, adjusting the learning rate (LR) accordingly. Specifically, one policy is for the LR of each task head submodule (non-shared weights) to balance each task's relative convergence rate, and another policy is for the backbone submodule (shared weights) to ensure optimization direction consistency.

We denote the training loss from the iteration i of task t as $cur\_L_i^t$. Each task's loss maintains an exponential moving average (EMA) value as the smoothed historical statistic, denoted as $his\_L_i^t$, i.e.,

$$his\_L_i^t = \alpha \cdot cur\_L_i^t + (1-\alpha) \cdot his\_L_{i-1}^t \,. \tag{3}$$

For the head submodule's LR adjustment, we use the ratio of his_L to cur_L as the inverse of the convergence rate for iteration i of task t as:

$$w_i^t = \frac{his\_L_i^t}{cur\_L_i^t} \,. \tag{4}$$

The $Softmax$ with temperature θ is then used to reweight the LR of the corresponding network task head, aiming to balance the convergence speed of each task. The reweighting factor λ_i^t for task t at training iteration i is denoted as:

$$\lambda_i^t = \frac{T \cdot e^{w_i^t/\theta}}{\sum_k^T e^{w_i^k/\theta}} \,, \tag{5}$$

where T is the total number of tasks. As a result, a relatively large value of $cur\_L_i^t$ indicates faster convergence for task t, leading to a smaller w_i^t and, consequently, a lower reweighting factor λ_i^t to prevent overly rapid convergence. Conversely, a smaller value of $cur\_L_i^t$ results in a larger λ_i^t. This strategy ensures that the convergence rate of each task remains balanced throughout training.

For the backbone submodule's LR adjustment, the reweighting is based on the historical consistency of each loss. To measure the training convergence consistency, we define a consistency score C based on cur_L and his_L. Specifically, cur_L and his_L are first converted into probability distributions using the function P, which employs a simple $Softmax$ function:

$$P(L) = Softmax(L). \tag{6}$$

Next, the Kullback-Leibler divergence, D_{KL}, is calculated to evaluate whether the current losses from each task remain stable and consistent with their historical values:

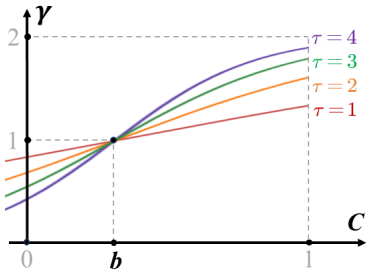

Figure 3: Reweighting curves for various temperature (τ).

$$C = 1 - D_{KL}\left(\, P(cur\_L) \parallel P(his\_L) \,\right) \tag{7}$$

$$= 1 - \sum_{t}^{T} P(cur\_L^t) \cdot \log \frac{P(cur\_L^t)}{P(his\_L^t)} \,, \tag{8}$$

therefore C is in the range of $(-\infty, 1]$. A larger C indicates that the relative values of the current iteration losses are similar to their historical values, suggesting that the current batch of samples stabilizes the network updates. In this case, the LR has to be increased to make the network converge faster. Conversely, a lower C indicates instability, suggesting that the current samples make some tasks more difficult and others easier to learn compared to the previous average state. If the network updates the shared weights too aggressively in such cases, the network will be optimized in the direction of the harder task of the current iteration, which might harm the easier tasks. Therefore, the network should update cautiously to reduce the LR.

To balance this, we propose dynamically reweighting the shared weight backbone with the following policy:

$$\gamma_i = \; 2 \cdot Sigmoid((C - b) \cdot \tau) \tag{9}$$

$$= \; \frac{2}{1 + e^{-(C-b)\cdot\tau}} \,. \tag{10}$$

The scalar factor of 2 ensures the reweighted value after the sigmoid function is in the range of (0, 2). b is the hyperparameter, bias, which can be interpreted as the reweithing threshold, i.e. when the C is b, the reweight is 1. τ is the temperature for value sensitivity adjustment. The reweighting curves for various temperatures and the relation between b and C are demonstrated in Figure 3.

Experiments and Analysis

To train and evaluate models for the M2Det task, we establish a new benchmark dataset by merging three detection datasets: SARDet-100K (Li et al. 2024c), DOTA-v1.0 (Xia et al. 2018), and DroneVehicle (Sun et al. 2022b), which correspond to SAR, optical, and infrared modalities, respectively. We refer to this combined dataset as the SOI-Det dataset.

Main Results

We evaluate the performance of our proposed SM3Det model against individual dataset training, simple joint training, and three SOTA methods that can be adapted for this task: UniDet (Zhou, Koltun, and Krähenbühl 2022) with a partitioned head, the DA network (Wang et al. 2019) implemented within the ConvNext-T backbone, and uncertainty loss (Kendall, Gal, and Cipolla 2018) implemented upon UniDet. The main results are presented in Table 1.

It can be observed that simple joint training of the three multi-modality datasets—i.e., merely merging the datasets and using a model with a shared backbone and separate task heads, along with a random data sampling strategy—results in a significant performance drop. This phenomenon highlights the increased challenge of this task compared to multi-dataset training for general object detection, where simple joint training typically enhances the performance of individual datasets (Wang et al. 2019; Zhou, Koltun, and Krähenbühl 2022). The previous SOTA methods, UniDet (Zhou, Koltun, and Krähenbühl 2022), DA (Wang et al. 2019) and uncertainty loss (Kendall, Gal, and Cipolla 2018), barely exceed the baseline by a small margin. In contrast, our proposed SM3Det model significantly improves overall mAP performance from 48.23 to 50.20, an increase of 1.97 mAP. To be noticed, our lightweight version of SM3Det which only incorporates DSO but without MoE structures, also easily outperforms other SOTA methods.

To assess the generalization capability of SM3Det, we evaluate its performance across different backbones and detectors. As illustrated in Figure 4, SM3Det significantly outperforms individual models across various modern convolutional backbones, including ConvNext (Liu et al. 2022), VAN (Guo et al. 2022), LSKNet (Li et al. 2023) and PVT-v2 (Wang et al. 2022). The model also exhibits reasonable scalability as the model size increases. We also evaluate SM3Det with different detectors. Since both the optical dataset (DOTA) and the infrared dataset (DroneVehicle) involve OBB regression tasks, we use the same head network structure in our model. In contrast, for the SAR dataset (SARDet-100K), which involves an HBB regression task, we implement a standard horizontal object detection head. Figure 5 shows our evaluation of SM3Det on one-stage (RetinaNet (Lin et al. 2017), GFL (Li et al. 2022) and S^2ANet (Han et al. 2020)) and two-stage (F-RCNN (Ren et al. 2015), Cascade F-RCNN (Cai and Vasconcelos 2018), O-RCNN (Xie et al. 2024) and RoI-Transformer (Ding et al. 2019)) detector combinations. The results consistently demonstrate that SM3Det significantly outperforms individual models across all detector combinations.

Ablation Study and Analysis

Expert Number and top-k Number. In sparse MoE architecture, the number of experts to add (N) and the top-k value play crucial roles in determining the model's performance and efficiency. Increasing N generally enhances the model's representation capacity, while a higher top-k value allows for more specialized knowledge to be applied to each input. However, these enhancements come at the cost of a larger model size, increased computational complexity, and potentially requiring more training data to ensure that each expert is adequately trained. Therefore, selecting the appropriate number of experts and top-k value is critical for achieving

Model	FLOPs	#P	Test on	**mAP**	@50	@75
3 models	403G	126M	Overall	48.23	79.39	51.26
GFL	131G	36M	SARDet-100K	57.31	87.44	61.99
O-RCNN	136G	45M	DOTA	45.31	77.70	46.45
O-RCNN	136G	45M	DroneVehicle	46.09	74.78	52.79
Simple Joint Training	403G	66M	Overall	47.05	77.56	50.11
			SARDet-100K	53.46	84.11	57.29
			DOTA	45.18	76.37	46.78
			DroneVehicle	44.99	73.28	51.50
DA +ConvNext-T	403G	66M	Overall	48.37	79.76	51.66
			SARDet-100K	53.86	84.93	58.09
			DOTA	46.23	78.47	47.58
			DroneVehicle	48.21	77.43	56.16
UniDet (Partitioned)	403G	66M	Overall	48.47	79.55	52.01
			SARDet-100K	53.81	84.70	57.43
			DOTA	46.49	78.28	48.59
			DroneVehicle	47.99	77.17	55.74
Uncertainty loss	403G	66M	Overall	48.79	79.99	52.50
			SARDet-100K	53.43	84.81	57.41
			DOTA	46.94	78.73	49.08
			DroneVehicle	48.78	77.96	56.88
SM3Det (DSO only)	403G	66M	Overall	49.40	80.19	52.93
			SARDet-100K	58.54	88.59	62.67
			DOTA	46.18	77.86	47.95
			DroneVehicle	48.09	77.09	56.20
SM3Det	487G	178M	Overall	50.20	80.68	53.79
			SARDet-100K	60.64	89.94	65.06
			DOTA	46.47	77.88	48.24
			DroneVehicle	48.87	77.99	56.90

Table 1: Model performance comparison on the SOI-Det dataset (SARDet-100K + DOTA + DroneVehicle). The proposed SM3Det model outperforms individual models and other SOTA models.

MoE (N, k)	w/o MoE	2, 2	4, 2	6, 2	8, 2	10, 2	8, 1	8, 2	8, 3	8, 2 Image-Level	8, 2 Grid-Level
FLOPs (G)	403	469	469	469	469	469	403	469	531	487	487
#P (M)	66	82	113	142	174	205	174	174	174	178	178
mAP	48.51	48.94	49.11	49.13	**49.31**	49.24	49.05	**49.31**	49.13	48.25	**50.20**
@50	79.70	80.25	80.10	79.74	**80.26**	80.18	79.72	**80.26**	79.98	79.10	**80.68**
@75	51.78	52.01	52.13	52.76	52.84	52.79	52.30	52.84	**52.77**	51.31	**53.79**

Table 2: Experiments on the MoE backbone with varying numbers of experts and top-k selection configurations. Experts are applied only to the even-indexed layers of the last two stages for validation efficiency, except for the last 2 columns. N: number of experts to add. k: number of experts to activate. The optimal configuration balancing performance and computational efficiency is identified as 8 experts with a top-k value of 2.

an optimal balance between model performance and computational efficiency. The results in Table 2 underscore the importance of tuning the number of experts and the top-k value in a sparse MoE architecture. It reveals that the optimal configuration for this sparse MoE architecture in terms of balancing performance and computational efficiency is 8 experts with a top-2 experts. This configuration maximizes the model's ability to learn from diverse inputs without introducing unnecessary complexity or overfitting.

Image-level v.s. Grid-level MoE. In Table 2, the grid-level MoE outperforms the image-level counterpart, indicating that grid-level experts more effectively capture spatial variations across different objects in multi-modal images. By processing features at a finer spatial granularity, experts are more attuned to object localization, making grid-level MoE particularly well-suited for object detection tasks.

Grid-level Experts Activation Behaviour Analysis. We visualize the selection results for each grid area across the last three stages of a well-tuned ConvNext-T backbone. In this visualization, each square grid represents the corresponding receptive field of that stage, with local deep features processed by different experts indicated by distinct colours. The

τ, b	3, 0.3	3, 0.4	3, 0.5	3, 0.6	2, 0.4	3, 0.4	4, 0.4	w/o DSO	w/o Head policy	w/o Backbone policy
mAP	50.14	**50.20**	50.07	50.03	49.92	**50.20**	50.03	49.47	49.86	50.11
@50	80.61	**80.68**	80.66	80.61	80.55	**80.68**	80.44	80.33	80.53	80.66
@75	**53.81**	53.79	54.00	53.98	53.56	53.79	53.79	52.98	53.44	53.70

Table 3: Experiments on the DSO method with varying temperature (τ) and bias (b). DSO is not sensitive to bias b.

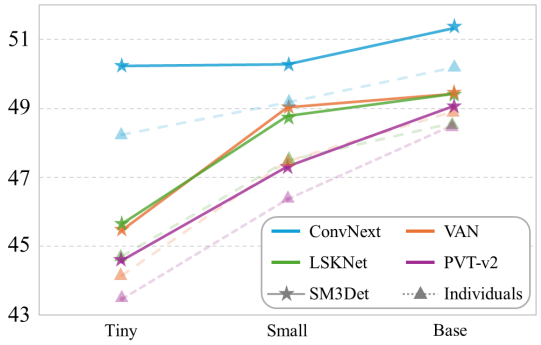

Figure 4: SM3Det on different backbones.

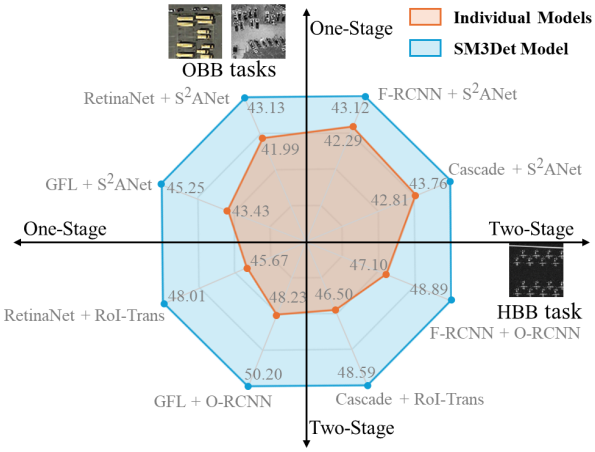

Figure 5: SM3Det on different detector heads.

top-1 selected experts are illustrated in Figure 6. For both RGB and IR images, a consistent pattern emerges: expert 1 predominantly processes salient objects, while Expert 3 focuses on background patches across all three stages. In contrast, the situation is more complex for SAR images. Particularly at stage 4, three experts (Expert 1, Expert 4, and Expert 6) are responsible for processing background areas, with Expert 1 also handling ship objects.

DSO hyperparamers. We conduct an ablation study on each component of the proposed DSO method, as well as the sensitivity of its two key hyperparameters. The results are summarized in Table 3. Omitting the learning rate adjustment for either the head or backbone leads to significant performance degradation. The bias parameter b and temperature τ dynamically adjust learning rates to account for varying task and modality difficulties. Specifically, b serves as a reweighting balance point, meaning when the calculated

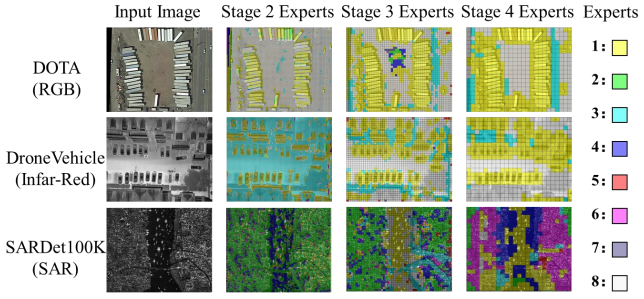

Figure 6: Visualization of grid expert activation across the last three stages of a well-tuned backbone on SAR, RGB, and IF images. Each square grid represents the receptive field at a given stage, with different colors indicating the local grid areas processed by distinct experts. The top-1 selected experts for each grid are shown. Each expert specializes in processing unique local patterns and semantics.

consistency score equals b, the reweighting factor is 1. A bias value of $b = 0.4$ proved optimal when the temperature was fixed at 3, striking a good balance in learning rate adjustments. Notably, variations in b did not significantly impact performance, indicating that the method is robust to changes in bias. Regarding temperature,τ, influences the reweighting curve in both the network head and the backbone's learning rate adjustment mechanism. Larger values result in sharper, more sensitive adjustments. A temperature of $\tau = 3$ provided the best balance between stability and responsiveness. In summary, the $\tau = 3$ and $b = 0.4$ yielded the best performance, effectively managing learning rate adjustments across diverse tasks and datasets.

Conclusion

In conclusion, this paper introduces a new and challenging task of Multi-Modal Datasets and Multi-Task Object Detection in remote sensing. To tackle this, we developed the SM3Det model, integrating a novel grid-level MoE approach and a dynamic submodule optimization mechanism. Intensive experiments and thorough analysis demonstrate SM3Det's strong performance and generalizability.

Acknowledgements

This work is supported by the National Science Fund of China (62361166670, U24A20330, 62576177, 62206134, 62225604), Shenzhen Science and Technology Program (JCYJ20250604184027034) and the Fundamental Research Funds for the Central Universities (070-63253222, 070-63253217).

References

Ahamed, T.; Tian, L.; Jiang, Y.; Zhao, B.; Liu, H.; and Ting, K. C. 2012. Tower remote-sensing system for monitoring energy crops; image acquisition and geometric corrections. *Biosystems engineering*, 112(2): 93–107.

Anderson, K.; Ryan, B.; Sonntag, W.; Kavvada, A.; and Friedl, L. 2017. Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-spatial Information Science*.

Avola, D.; Cinque, L.; Di Mambro, A.; Diko, A.; Fagioli, A.; Foresti, G. L.; Marini, M. R.; Mecca, A.; and Pannone, D. 2021. Low-altitude aerial video surveillance via one-class SVM anomaly detection from textural features in UAV images. *Information*.

Bozcan, I.; and Kayacan, E. 2020. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation*.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *CVPR*.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021. Attentional local contrast networks for infrared small target detection. *TGRS*.

Dai, Y.; Zou, M.; Li, Y.; Li, X.; Ni, K.; and Yang, J. 2024. DenoDet: Attention as Deformable Multi-Subspace Feature Denoising for Target Detection in SAR Images. *arXiv*.

Devaraj, C.; and Shah, C. A. 2013. Automated geometric correction of Landsat MSS L1G imagery. *IEEE Geoscience and Remote Sensing Letters*, 11(1): 347–351.

Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *CVPR*.

for Photogrammetry, T. I. S.; and (ISPRS), R. S. 2022. 2D Semantic Labeling Contest - Potsdam. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx.

Guo, M.; Haque, A.; Huang, D.-A.; Yeung, S.; and Fei-Fei, L. 2018. Dynamic task prioritization for multitask learning. In *ECCV*.

Guo, M.-H.; Lu, C.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S. 2022. Visual Attention Network. *Computational Visual Media*.

Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2020. Align Deep Features for Oriented Object Detection. *TGRS*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*.

Huang, C.; Fang, S.; Wu, H.; Wang, Y.; and Yang, Y. 2024. Low-Altitude Intelligent Transportation: system architecture, infrastructure, and key technologies. *Journal of Industrial Information Integration*, 100694.

Jain, Y.; Behl, H.; Kira, Z.; and Vineet, V. 2024. DAMEX: Dataset-aware Mixture-of-Experts for visual understanding of mixture-of-datasets. *NeurIPS*.

Jensen, O. B. 2016. Drone city–power, design and aerial mobility in the age of "smart cities". *Geographica Helvetica*.

Jiang, Y.; Li, X.; Zhu, G.; Li, H.; Deng, J.; and Shi, Q. 2023. 6G Non-Terrestrial networks enabled low-altitude economy: Opportunities and challenges. *arXiv preprint arXiv:2311.09047*.

Kapidis, G.; Poppe, R.; and Veltkamp, R. C. 2021. Multi-dataset, multitask learning of egocentric vision tasks. *TPAMI*.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*.

Khan, A.; Yanmaz, E.; and Rinner, B. 2014. Information merging in multi-UAV cooperative search. In *2014 IEEE international conference on robotics and automation*.

Li, D.; Wang, M.; Dong, Z.; Shen, X.; and Shi, L. 2017. Earth observation brain (EOB): An intelligent earth observation system. *Geo-spatial information science*.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS*.

Li, W.; Yang, W.; Liu, T.; Hou, Y.; Li, Y.; Liu, Z.; Liu, Y.; and Liu, L. 2024a. Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture. *ISPRS Journal o*.

Li, X.; Lv, C.; Wang, W.; Li, G.; Yang, L.; and Yang, J. 2022. Generalized focal loss: Towards efficient representation learning for dense object detection. *TPAMI*.

Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; and Li, X. 2023. Large Selective Kernel Network for Remote Sensing Object Detection. In *ICCV*.

Li, Y.; Li, X.; Dai, Y.; Hou, Q.; Liu, L.; Liu, Y.; Cheng, M.-M.; and Yang, J. 2024b. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *arXiv preprint arXiv:2403.11735*.

Li, Y.; Li, X.; Li, W.; Hou, Q.; Liu, L.; Cheng, M.-M.; and Yang, J. 2024c. SARDet-100K: Towards Open-Source Benchmark and ToolKit for Large-Scale SAR Object Detection. In *NeurIPS*.

Li, Y.; Zhang, Y.; Tang, W.; Dai, Y.; Cheng, M.-M.; Li, X.; and Yang, J. 2025. Visual Instruction Pretraining for Domain-Specific Foundation Models. *arXiv*.

Lin, M. 2013. Network in network. *arXiv*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *ICCV*.

Liu, J.; Chen, H.; and Wang, Y. 2021. Multi-source remote sensing image fusion for ship target detection and recognition. *Remote Sensing*, 13(23): 4852.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *CVPR*.

Ni, K.; Zou, M.; Li, Y.; Li, X.; Guo, K.; Cheng, M.-M.; and Dai, Y. 2025. DenoDet V2: Phase-Amplitude Cross Denoising for SAR Object Detection. *AAAI*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; Weinmann, M.; Hinz, S.; Wang, C.; and Fu, K. 2022a. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS*.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022b. Drone-based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*.

Wang, X.; Cai, Z.; Gao, D.; and Vasconcelos, N. 2019. Towards universal object detection by domain attention. In *CVPR*.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *CVPR*.

Xie, X.; Cheng, G.; Wang, J.; Li, K.; Yao, X.; and Han, J. 2024. Oriented r-cnn and beyond. *IJCV*.

Xu, H.; Fang, L.; Liang, X.; Kang, W.; and Li, Z. 2020. Universal-rcnn: Universal object detector via transferable graph r-cnn. In *AAAI*.

Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; and Tian, Q. 2021. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In *ICML*.

Ye, Y.; Teng, X.; Yang, H.; Chen, S.; Sun, Y.; Bian, Y.; Tan, T.; Li, Z.; and Yu, Q. 2025. 3MOS: a multi-source, multi-resolution, and multi-scene optical-SAR dataset with insights for multi-modal image matching. *Visual Intelligence*, 3(1): 1–27.

Yuan, X.; Zheng, Z.; Li, Y.; Liu, X.; Liu, L.; Li, X.; Hou, Q.; and Cheng, M.-M. 2025. Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection. *arXiv preprint arXiv:2501.03775*.

Zhang, H.; Huang, S.-L.; and Kuruoglu, E. E. 2024. HGR Correlation Pooling Fusion Framework for Recognition and Classification in Multimodal Remote Sensing Data. *Remote Sensing*, 16(10): 1708.

Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. 2021. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis. *Remote Sensing*.

Zhang, X.; Li, D.; Dong, X.; Wu, T.; Yu, H.; Wang, J.; Li, Q.; and Li, X. 2025. UniChange: Unifying Change Detection with Multimodal Large Language Model. *arXiv preprint arXiv:2511.02607*.

Zhang, Y.; and Yang, Q. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*.

Zhang, Z.; Zhang, L.; Wu, J.; and Guo, W. 2024. Optical and Synthetic Aperture Radar Image Fusion for Ship Detection and Recognition: Current state, challenges, and future prospects. *IEEE Geoscience and Remote Sensing Magazine*.

Zhao, X.; Schulter, S.; Sharma, G.; Tsai, Y.-H.; Chandraker, M.; and Wu, Y. 2020. Object detection with a unified label space from multiple datasets. In *2020*.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2022. Simple multi-dataset detection. In *CVPR*.