Appendix: Supplementary Materials for "LWGANet: Addressing Spatial and Channel Redundancy in Remote Sensing Visual Tasks with Light-Weight Grouped Attention"

Wei Lu[1], **Xue Yang**[2], **Si-Bao Chen**[1]*

[1] School of Computer Science and Technology, Anhui University, Hefei 230601, China
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
luwei_ahu@qq.com, yangxue-2019-sjtu@sjtu.edu.cn, sbchen@ahu.edu.cn

This document provides supplementary materials to the main paper, "LWGANet: Addressing Spatial and Channel Redundancy in Remote Sensing Visual Tasks with Light-Weight Grouped Attention." The purpose of this appendix is to offer comprehensive details that ensure the full reproducibility of our work and provide deeper insights into our methodology and findings. We elaborate on architectural specifics, dataset characteristics, implementation protocols, and present additional experimental results that complement the analyses in the main paper.

The appendix is organized as follows:

- **Appendix A** presents a thorough breakdown of the LWGANet architectural configurations.
- **Appendix B** details the 12 public benchmark datasets used in our evaluation.
- **Appendix C** outlines the precise experimental protocols for all downstream tasks.
- **Appendix D** contains extended quantitative results and qualitative visualizations.
- **Appendix E** provides a comprehensive analysis of our ablation studies.
- **Appendix F** discuss a further rationale design.

LWGANet Architecture Details

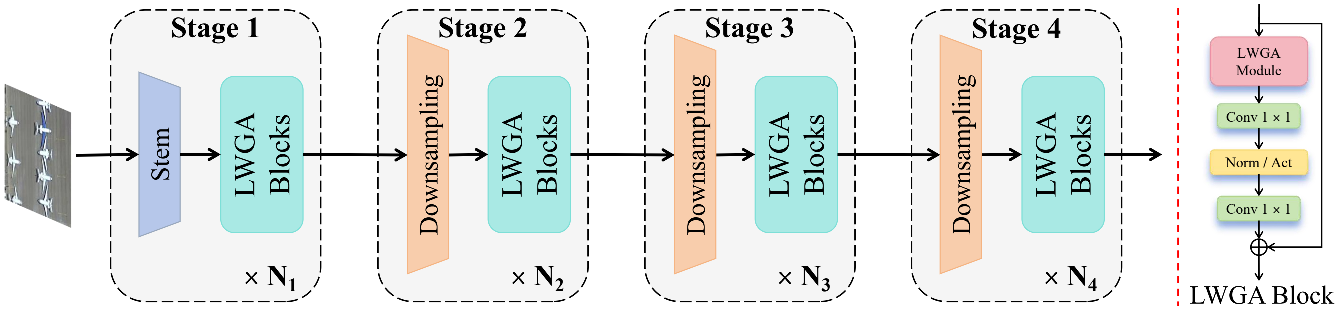

Figure 1: An overview of the LWGANet architecture. The network is organized into four hierarchical stages, producing feature maps with spatial dimensions of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$. The corresponding channel dimensions are C, $2C$, $4C$, and $8C$, where H, W, and C denote the input height, width, and base channel number, respectively.

The LWGANet architecture is designed as a lightweight, hierarchical backbone for efficient multi-scale feature extraction in remote sensing (RS) visual tasks. As illustrated in Figure 1, the architecture adopts a four-stage pyramidal structure. Each stage progressively reduces the spatial resolution of the feature maps while increasing the channel depth, enabling the network to learn a rich hierarchy of features from local details to global context. To accommodate varying computational budgets and performance requirements, we instantiate LWGANet in three distinct variants: LWGANet-L0, LWGANet-L1, and LWGANet-L2. The detailed configurations for these variants are summarized in Table 1.

*Corresponding author.

Stage	Feature Map Size	Layer	Channels (In/Out)		
			L0	L1	L2
1	$\frac{H}{4} \times \frac{W}{4}$	Stem Layer	$3/32$	$3/64$	$3/96$
		$[LWGABlock] \times N_1$	$32/32$	$64/64$	$96/96$
2	$\frac{H}{8} \times \frac{W}{8}$	DRFD Module	$32/64$	$64/128$	$96/192$
		$[LWGABlock] \times N_2$	$64/64$	$128/128$	$192/192$
3	$\frac{H}{16} \times \frac{W}{16}$	DRFD Module	$64/128$	$128/256$	$192/384$
		$[LWGABlock] \times N_3$	$128/128$	$256/256$	$384/384$
4	$\frac{H}{32} \times \frac{W}{32}$	DRFD Module	$128/256$	$256/512$	$384/768$
		$[LWGABlock] \times N_4$	$256/256$	$512/512$	$768/768$
SMA distance			$[11, 11, 11, 11]$	$[11, 11, 11, 11]$	$[11, 11, 11, 11]$
Number of Blocks			$[1, 2, 4, 2]$	$[1, 2, 4, 2]$	$[1, 4, 4, 2]$
Activation			GELU	GELU	ReLU
Dropout			0.0	0.1	0.1
Parameters (224×224 input)			1.72M	5.90M	13.0M
FLOPs (224×224 input)			0.186G	0.709G	1.87G

Table 1: Detailed architectural configurations of the LWGANet variants (L0, L1, L2).

Architectural Blueprint. LWGANet's design follows a consistent pattern across its four stages, which operate at spatial resolutions of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$ respectively, where H and W are the height and width.

- **Stage 1** begins with a `Stem Layer`, which performs an initial 4× downsampling of the input image and projects it into a higher-dimensional feature space. This is followed by a series of N_1 `LWGA Blocks` for initial feature refinement.
- **Stages 2, 3, and 4** each commence with a `DRFD Module` (Lu et al. 2023). The DRFD module is responsible for spatial downsampling (by a factor of 2) and channel expansion (typically doubling the channels), mitigating information loss during resolution reduction. Following the DRFD module, a stack of N_k ($k \in \{2, 3, 4\}$) `LWGA Blocks` is employed to enhance the feature representation at the new scale. The `LWGA Block`, the network's core computational unit, employs a lightweight grouped attention mechanism to capture multi-scale features.

Model Variants and Complexity. To provide a flexible trade-off between model capacity and computational cost, we define the L0, L1, and L2 variants by adjusting the network's width, depth, and other hyperparameters.

- **Channel Dimensions:** The base channel dimension C is set to 32, 64, and 96 for the L0, L1, and L2 variants, respectively. The channels at each stage scale accordingly, as detailed in Table 1.
- **Block Depth:** The number of `LWGA Blocks` per stage, denoted by $[N_1, N_2, N_3, N_4]$, is configured as $[1, 2, 4, 2]$ for the L0 and L1 models, and increased to $[1, 4, 4, 2]$ for the larger L2 model to enhance its representational power.
- **Hyperparameters:** The L0 and L1 variants utilize the GELU activation function, while the L2 variant employs ReLU. To manage overfitting in the larger models, a dropout rate of 0.1 is applied to L1 and L2, whereas L0 is trained without dropout.
- **Computational Profile:** These design choices result in a clear progression of model scale. The L0 model is the most compact with 1.72M parameters and 0.186G FLOPs. The L1 model offers a balanced profile with 5.90M parameters and 0.709G FLOPs. The L2 model provides the highest capacity, with 13.0M parameters and 1.87G FLOPs.

In summary, the hierarchical design of LWGANet, combined with its specialized DRFD and LWGA modules, facilitates effective and efficient multi-scale feature learning. The availability of three scalable variants makes LWGANet a versatile and powerful solution, adaptable to a wide spectrum of RS applications with diverse resource constraints.

Experimental Datasets

This section provides a detailed overview of the 12 benchmark datasets employed in our experimental evaluation, spanning four major RS visual tasks: scene classification, oriented object detection, semantic segmentation, and change detection.

Scene Classification Datasets

UCMerced Land Use (UCM). The UCM dataset (Yang and Newsam 2010) is a seminal benchmark for RS scene classification. It consists of 2,100 aerial images, each with a size of 256×256 pixels and a spatial resolution of 0.3 meters. The dataset is evenly distributed across 21 land-use categories (e.g., agricultural, residential, forest), with 100 images per category. Sourced from the United States Geological Survey (USGS) National Map, UCM is widely used for initial benchmarking due to its diversity and manageable size.

Task	Dataset Name	# Images/Pairs	# Classes	# Instances	Image Size (px)	Spatial Res. (m)	Reference
Scene Classification	UCMerced (UCM)	2,100	21	N/A	256×256	0.3	(Yang and Newsam 2010)
	Aerial Image (AID)	10,000	30	N/A	600×600	0.5 – 8	(Xia et al. 2017)
	NWPU-RESISC45	31,500	45	N/A	256×256	0.2 – 30	(Cheng, Han, and Lu 2017)
Oriented Object Detection	DOTA 1.0	2,806	15	188,282	800^2 – $20{,}000^2$	Varies	(Xia et al. 2018)
	DOTA 1.5	2,806	16	402,089	800^2 – $20{,}000^2$	Varies	(Xia et al. 2018)
	DIOR-R	23,463	20	192,472	800×800	0.5 – 30	(Cheng et al. 2022)
Semantic Segmentation	UAVid	300	8	N/A	up to 4096×2160	Varies	(Lyu et al. 2020)
	LoveDA	5,987	7	N/A	1024×1024	0.3	(Wang et al. 2021)
Change Detection	LEVIR-CD	637 pairs (10,240 patches)	2	N/A	1024×1024 (cropped)	0.5	(Chen and Shi 2020)
	WHU-CD	1 pair (7,432 patches)	2	N/A	Large (cropped)	0.075 – 0.5	(Ji, Wei, and Lu 2018)
	CDD-CD	11 pairs (16,000 patches)	2	N/A	Varies (cropped)	0.03 – 1	(Lebedev et al. 2018)
	SYSU-CD	20,000 pairs	2	N/A	256×256	0.5	(Shi et al. 2021)

Table 2: Summary of the 12 benchmark datasets. The datasets cover four major RS tasks.

Aerial Image Dataset (AID). AID (Xia et al. 2017) is a larger-scale dataset for aerial scene classification, comprising 10,000 images across 30 distinct scene types (e.g., airport, forest, viaduct). Each image measures 600×600 pixels, with spatial resolutions varying from 0.5 to 8 meters. Sourced from Google Earth, AID's scale and resolution variability make it a robust benchmark for evaluating the generalization capabilities of deep learning models.

NWPU-RESISC45. The NWPU-RESISC45 dataset (Cheng, Han, and Lu 2017) is a comprehensive and challenging benchmark for scene classification. It contains 31,500 images distributed among 45 scene classes, with 700 images per class. The images have a resolution of 256×256 pixels and spatial resolutions ranging from 0.2 to 30 meters. Its large scale, high number of classes, and significant intra-class diversity make it an ideal testbed for modern deep learning architectures.

Oriented Object Detection Datasets

DOTA 1.0. The DOTA 1.0 dataset (Xia et al. 2018) is a large-scale benchmark for object detection in aerial images. It contains 2,806 high-resolution images (ranging from 800×800 to 20,000×20,000 pixels) and 188,282 object instances across 15 categories. Objects are annotated with oriented bounding boxes (OBB) using arbitrary quadrilaterals, making it a standard for evaluating oriented object detection algorithms. The 15 object categories include: Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Small vehicle (SV), Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HA), Swimming pool (SP), and Helicopter (HC).

DOTA 1.5. DOTA 1.5 (Xia et al. 2018) is an updated version of DOTA 1.0, using the same images but with revised annotations. It addresses challenges such as small object detection by adding numerous new instances, bringing the total to 402,089. It also introduces a new category, ”container crane (CC),” increasing the total to 16.

DIOR-R. The DIOR-R dataset (Cheng et al. 2022) is an extension of the DIOR dataset, specifically tailored for oriented object detection. It comprises 23,463 images and 192,472 instances across 20 object categories, with spatial resolutions from 0.5 to 30 meters. All objects are annotated with rotated bounding boxes, providing a rich resource for developing and testing robust oriented detectors. The 20 common object categories include: Airplane (APL), Airport (APO), Baseball field (BF), Basketball court (BC), Bridge (BR), Chimney (CH), Expressway service area (ESA), Expressway toll station (ETS), Dam (DAM), Golf field (GF), Ground track field (GTF), Harbor (HA), Overpass (OP), Ship (SH), Stadium (STA), Storage tank (STO), Tennis court (TC), Train station (TS), Vehicle (VE) and Windmill (WM).

Semantic Segmentation Datasets

UAVid. The UAVid dataset (Lyu et al. 2020) is designed for high-resolution semantic segmentation from Unmanned Aerial Vehicle (UAV) platforms. It includes 30 video sequences, from which 300 images are densely annotated with pixel-level labels. The images feature resolutions up to 4,096×2,160 pixels and cover eight semantic classes (e.g., buildings, roads, moving cars), presenting challenges related to large scale variations and oblique viewing angles.

LoveDA. The LoveDA dataset (Wang et al. 2021) targets land-cover mapping and is notable for its focus on domain adaptation between urban and rural scenes. It contains 5,987 images of 1,024×1,024 pixels with a 0.3-meter spatial resolution. The dataset is annotated with seven land-cover classes and is split into urban and rural subsets, making it ideal for evaluating cross-domain segmentation performance.

Change Detection Datasets

LEVIR-CD. The LEVIR-CD dataset (Chen and Shi 2020) is a large-scale benchmark for building change detection. It consists of 637 pairs of high-resolution (1,024×1,024 pixels, 0.5-meter resolution) bi-temporal images from Google Earth. For standardized evaluation, the images are cropped into 256×256 non-overlapping patches, yielding 7,120 training, 1,024 validation, and 2,048 test pairs.

WHU-CD. The WHU-CD dataset (Ji, Wei, and Lu 2018) focuses on building change detection in very high-resolution aerial imagery (0.075–0.5 meters). It contains a single large image pair covering a region that underwent significant urban development, which is cropped into 7,432 non-overlapping patch pairs for analysis.

CDD-CD. The CDD-CD dataset (Lebedev et al. 2018) provides 11 pairs of season-varying satellite images with resolutions from 0.03 to 1 meter per pixel. It captures changes in both man-made objects and natural landscapes. The dataset is preprocessed into 256×256 patches, resulting in 10,000 training, 3,000 validation, and 3,000 test pairs.

SYSU-CD. The SYSU-CD dataset (Shi et al. 2021) contains 20,000 pairs of 256×256 pixel images with a 0.5-meter resolution, focusing on urban changes in Hong Kong. It is divided into 12,000 training, 4,000 validation, and 4,000 test pairs, providing a large-scale resource for training data-hungry change detection models.

Experimental Setup

This section details the experimental protocols, including training configurations, data processing, and evaluation environments for each downstream task.

Classification Experimental Setup

For scene classification, all models were trained from scratch for 300 epochs to assess their intrinsic learning capabilities without external pre-training. The datasets were partitioned into training and validation sets using an 80/20 split. Key training parameters are listed below:

- **Input Resolution:** All images were resized to 224×224 pixels.
- **Optimizer:** We used the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of 1×10^{-4} and a weight decay of 5×10^{-2}.
- **Learning Rate Schedule:** A cosine decay schedule (Loshchilov and Hutter 2017) was employed, preceded by a 2-epoch linear warmup phase with an initial warmup factor of 1×10^{-3}.
- **Batch Size:** A batch size of 64 was used for all experiments.
- **Data Augmentation:** Standard augmentation techniques were applied, including RandomResizedCrop, RandomHorizontalFlip, and RandAugment (Cubuk et al. 2020).
- **Loss Function:** The standard cross-entropy loss was used for optimization.
- **Environment:** Experiments were conducted on an NVIDIA RTX 3090 GPU using the PyTorch framework with Ubuntu20.04. Inference speeds were benchmarked on three platforms: NVIDIA RTX 3090 (GPU), Intel i9-11900K (CPU), and NVIDIA AGX-XAVIER (ARM).

Detection Experimental Setup

For oriented object detection, we adopted the Oriented R-CNN (Xie et al. 2021b) framework implemented in MMRotate (Zhou et al. 2022). All backbone models were pre-trained on ImageNet-1k (Deng et al. 2009) for 300 epochs.

- **Data Preprocessing:** For DOTA 1.0 and DOTA 1.5, original images were cropped into 1,024×1,024 patches with a 200-pixel overlap. For DIOR-R, the input size was 800×800.
- **Training Schedule:** Models were trained for 36 epochs using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 0.05. The batch size was set to 8.
- **Data Augmentation:** Random resizing and flipping were applied during training to enhance model robustness.
- **Evaluation:** Following standard evaluation protocols, we trained models on the trainval splits and reported performance on the test splits.

Segmentation Experimental Setup

For semantic segmentation, we employed UnetFormer (Wang et al. 2022a) as the segmentation head. The backbones were pre-trained on ImageNet-1k.

- **UAVid Dataset:** Models were trained for 30 epochs with a batch size of 8. Input images were resized to 1,024×1,024. Augmentations included random vertical/horizontal flips and random brightness adjustments. Test-time augmentation (TTA) with flips was used for evaluation.

- **LoveDA Dataset:** Models were trained for 70 epochs with a batch size of 16. Images were randomly cropped to 1,024×1,024. Augmentations included random scaling, flips, and rotations. Multi-scale and flip augmentations were used during testing.
- **Optimizer:** The AdamW optimizer was used with a base learning rate of 6×10^{-4}, adjusted by a cosine annealing schedule.

Change Detection Experimental Setup

For change detection, we integrated our backbone into two different decoders, A2Net (Li et al. 2023c) and CLAFA (Wang et al. 2023), following their respective training protocols. Backbones were pre-trained on ImageNet-1k.

- **Optimizer:** The Adam optimizer was used with a momentum of 0.9 and a weight decay of 1×10^{-4}.
- **Learning Rate Schedule:** The learning rate was initialized to 1×10^{-3} and decayed to zero over 20,000 iterations using a polynomial schedule.
- **Batch Size:** A batch size of 64 was used.
- **Model Selection:** The model weights that achieved the highest F1 score on the validation set were selected for final evaluation on the test set.

Method	Params. (M) ↓	FLOPs (G) ↓	Top-1 Accuracy (%) ↑			Speed (FPS) ↑		
			NWPU	AID	UCM	GPU (Memory)	CPU	ARM
MobileNet V2 1.0×	2.28	0.319	95.06	93.65	97.14	11301 (1443.71 MB)	49.11	785.4
FasterNet T0	2.68	0.338	93.30	92.85	94.75	18276 (606.11 MB)	106.4	839.7
StarNet S1	2.68	0.431	94.30	91.05	93.10	6045 (1019.07 MB)	71.70	459.8
EdgeViT XXS	3.79	0.546	94.75	93.10	95.24	4153 (1053.79 MB)	46.36	259.9
EfficientformerV2 S0	3.36	0.396	94.52	93.80	97.14	1299 (-)	54.00	272.0
EdgeNeXt XXS	1.17	0.197	92.35	88.10	88.10	8521 (605.24 MB)	100.8	-
⋆MobileViT XXS	1.03	0.333	94.37	93.30	95.00	4811 (2510.08 MB)	31.87	453.6
GhostNet V2 0.6×	2.16	0.077	94.65	92.70	94.29	9802 (660.67 MB)	69.96	867.1
LWGANet L0	1.72	0.186	95.49	94.60	98.57	13234 (561.04 MB)	80.00	687.8
MobileNet V2 2.0×	8.81	1.17	95.35	93.85	97.86	5567 (2747.40 MB)	17.03	372.3
FasterNet T1	6.37	0.855	93.73	93.20	94.52	11876 (788.16 MB)	51.42	650.1
StarNet S3	5.5	0.767	93.32	91.40	93.33	4438 (1256.22 MB)	47.86	336.8
EdgeViT XS	6.40	1.12	94.89	93.75	94.52	3310 (1360.09 MB)	29.95	205.8
PVT V2 B0	3.42	0.533	94.35	93.10	96.43	3843 (1248.48 MB)	40.26	243.4
EfficientformerV2 S1	5.87	0.661	94.97	93.95	96.90	1211 (-)	36.96	204.5
EdgeNeXt XS	2.15	0.408	92.79	90.45	88.10	5455 (753.13 MB)	56.42	-
⋆MobileViT XS	2.02	0.900	94.90	95.20	96.43	3300 (2594.06 MB)	12.99	306.3
GhostNet V2 1.0×	4.93	0.181	95.08	93.80	94.76	6596 (966.06 MB)	42.02	591.7
LWGANet L1	5.90	0.709	95.70	94.85	98.81	6418 (978.98 MB)	34.08	375.8
MobileNet V2 2.5×	13.7	1.80	95.48	94.45	97.86	3796 (3405.37 MB)	12.90	282.4
FasterNet T2	13.8	1.91	95.11	93.60	94.29	6852 (1141.44 MB)	26.40	669.8
StarNet S4	7.23	1.07	93.08	89.75	90.71	3093 (1262.88 MB)	34.20	235.1
EdgeViT S	12.7	1.90	95.05	93.35	95.95	2318 (1384.29 MB)	19.31	141.3
PVT V2 B1	13.5	2.04	94.62	93.45	95.71	2369 (2366.87 MB)	15.96	145.3
EfficientformerV2 S2	12.3	1.26	95.14	94.20	97.38	642 (-)	24.74	123.8
EdgeNeXt S	5.30	0.960	93.54	91.90	92.62	3844 (1061.58 MB)	30.00	-
⋆MobileViT S	5.03	1.75	95.19	95.25	97.14	2681 (2734.39 MB)	10.20	152.7
GhostNet V2 2.0×	16.7	0.632	95.44	94.30	95.95	3476 (1793.17 MB)	21.32	303.2
LWGANet L2	13.0	1.87	96.17	95.45	98.57	3308 (1429.27 MB)	16.18	274.3

Table 3: Comprehensive comparison on the NWPU, AID, and UCM classification datasets. '⋆' indicates a training image size of 256×256. FPS data are from RTX 3090 (GPU), Intel i9-11900K (CPU), and NVIDIA AGX-XAVIER (ARM) platforms.

Quantitative Experimental Results

This section presents a detailed quantitative analysis of LWGANet's performance across all evaluated tasks, benchmarked against state-of-the-art methods.

Classification Experimental Results

As demonstrated in Figure 2 and detailed in Table 3, our proposed LWGANet redefines the state-of-the-art trade-off between classification accuracy, inference speed, and model efficiency across the NWPU, AID, and UCM datasets. We conduct a comprehensive comparison against a suite of prominent lightweight models, including CNN-based architectures like MobileNetV2

(Sandler et al. 2018) and FasterNet (Chen et al. 2023), Transformer-based models such as PVT V2 (Wang et al. 2022b), EdgeViT (Pan et al. 2022), and EfficientformerV2 (Li et al. 2023b), as well as hybrid models like MobileViT (Mehta and Rastegari 2022). Across all benchmarks, LWGANet consistently delivers a superior performance profile.

Accuracy. LWGANet consistently outperforms its peers in Top-1 accuracy. For instance, the most compact variant, LWGANet-L0 (1.72M parameters), achieves impressive accuracies of 95.49%, 94.60%, and 98.57% on the NWPU, AID, and UCM datasets, respectively. This performance surpasses not only smaller models like EdgeViT-XXS but also larger competitors such as MobileNetV2-1.0×. This performance advantage is consistent across model scales; the LWGANet-L2 model reaches a remarkable 96.17% on NWPU, outclassing models with similar or greater parameter counts, including PVTV2-B1, while maintaining a competitive model size.

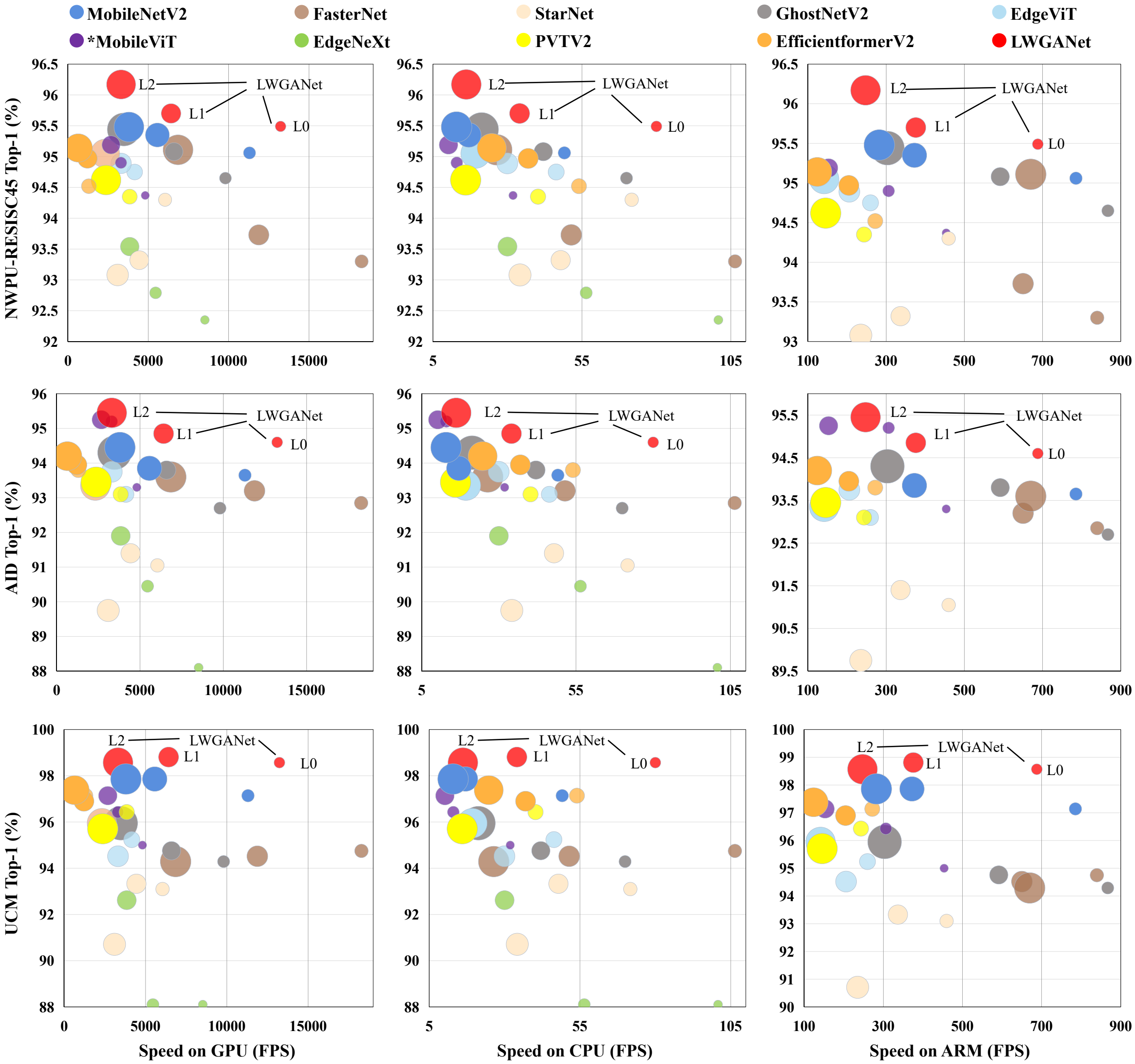

Figure 2: Comparison of Top-1 accuracy, inference speed (FPS), and model parameters on the NWPU, AID, and UCM datasets. The area of each circle is proportional to the model's parameter count. FPS was measured on NVIDIA RTX 3090 (GPU), Intel i9-11900K (CPU), and NVIDIA AGX-XAVIER (ARM) platforms with batch sizes of 256, 16, and 32, respectively. LWGANet (red circles) consistently achieves the best trade-off.

Inference Speed and Efficiency. LWGANet demonstrates exceptional throughput across diverse hardware platforms (GPU, CPU, and ARM). The LWGANet-L0 variant achieves 13,234 FPS on an RTX 3090 GPU and 80.00 FPS on an Intel i9-11900K CPU—speeds that are highly competitive and often superior to other models in its class. Even when scaled to the 13.0M-parameter LWGANet-L2, the model maintains excellent efficiency. Its GPU speed of 3,308 FPS is comparable to that of other models in its computational class but is achieved with significantly higher accuracy, highlighting the efficiency of our architecture.

Performance-Efficiency Trade-off. The bubble plot in Figure 2 visually summarizes LWGANet's dominance. The LWGANet series (red circles) consistently occupies the Pareto frontier, achieving a superior balance of accuracy, latency, and model complexity. It avoids the common compromise of sacrificing accuracy for speed or vice-versa, providing a more optimal solution than models that are either faster but less accurate (e.g., FasterNet) or more accurate but computationally heavier (e.g., some Transformer-based models).

Object Detection Experimental Results

DOTA 1.0 Results. Table 4 presents a comprehensive comparison of LWGANet-L2 against state-of-the-art oriented object detection methods on the DOTA 1.0 benchmark. We evaluate against a diverse set of approaches, including established detectors like SCRDet (Yang et al. 2019), RoI Transformer (Ding et al. 2019), Gliding Vertex (Xu et al. 2020), and ReDet (Han et al. 2021). We also benchmark against various backbones integrated within the Oriented R-CNN (Xie et al. 2021b) framework, such as the classic ResNet-50 (He et al. 2016) and more recent lightweight architectures like EfficientFormerV2-S2 (Li et al. 2023b), ARC-R50 (Pu et al. 2023), LSKNet-S (Li et al. 2023a), DecoupleNet-D2 (Lu et al. 2024), and PKINet-S (Cai et al. 2024). As shown in the table, our model, integrated into the Oriented R-CNN framework, achieves a new state-of-the-art mean Average Precision (mAP) of 79.02%. This result surpasses all competing methods.

Methods	Backbone	Params. (M) ↓	FLOPs (G) ↓	mAP (%) ↑	AP per category (%) ↑														
					PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
SCRDet	ResNet-50	41.9	-	72.61	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21
RoI Transformer	ResNet-50	55.1	225.3	74.05	89.01	77.48	51.64	72.07	74.43	77.55	87.76	90.81	79.71	85.27	58.36	64.11	76.50	71.99	54.06
Gliding Vertex	ResNet-50	41.1	211.3	75.02	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32
ReDet	ReResNet	31.6	-	76.25	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59
Oriented R-CNN	ResNet-50	41.1	211.4	75.87	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28
	*E-FormerV2-S2	29.2	145.1	76.70	89.55	84.12	53.39	74.40	80.70	84.84	87.92	90.89	87.44	84.47	60.88	67.43	77.63	67.62	59.16
	ARC-R50	74.4	212.0	77.35	89.40	82.48	55.33	73.88	79.37	84.05	88.06	90.90	86.44	84.83	63.63	70.32	74.29	71.91	65.43
	LSKNet-S	31.0	161.0	77.49	89.66	85.52	57.72	75.70	74.95	78.69	88.24	90.88	86.79	86.38	66.92	63.77	77.77	74.47	64.82
	DecoupleNet-D2	23.3	142.4	78.04	89.37	83.25	54.29	75.51	79.83	84.82	88.49	90.89	87.19	86.23	66.07	65.53	77.23	72.34	69.62
	PKINet-S	30.8	184.6	78.39	89.72	84.20	55.81	77.63	80.25	84.45	88.12	90.88	87.57	86.07	66.86	70.23	77.47	73.62	62.94
	LWGANet-L2	29.2	159.1	79.02	89.49	85.48	54.93	77.12	81.59	85.64	88.43	90.85	87.23	86.78	67.47	65.06	78.23	73.33	73.66

Table 4: Comparison with state-of-the-art oriented object detectors on the **DOTA 1.0 test set**. All methods are evaluated under a single-scale training and testing protocol Red and blue denote the top-two performance in each column. *E-FormerV2-S2 refers to EfficientFormerV2-S2.

Notably, LWGANet-L2 achieves this superior accuracy while maintaining high computational efficiency. With 29.2M parameters and 159.1 GFLOPs, it is more efficient than PKINet-S and significantly more accurate than models with lower complexity, such as DecoupleNet-D2 and EfficientFormerV2-S2. This demonstrates LWGANet-L2's exceptional ability to balance accuracy and resource consumption. In per-category performance, LWGANet-L2 excels on challenging classes like small vehicle (SV), large vehicle (LV), and helicopter (HC), underscoring the robustness of its feature representation.

DOTA 1.5 Results. On the more challenging DOTA 1.5 test set, as detailed in Table 5, we compare LWGANet-L2 against several state-of-the-art backbones within the Oriented R-CNN framework. These include ResNet-50 (He et al. 2016), ARC-R50 (Pu et al. 2023), LSKNet-S (Li et al. 2023a), DecoupleNet-D2 (Lu et al. 2024), PKINet-S (Cai et al. 2024), and EfficientFormerV2-S2 (Li et al. 2023b). LWGANet-L2 continues to demonstrate its superiority, achieving a new state-of-the-art mAP of 72.91%. This performance surpasses all other evaluated backbones. The consistent high performance across all 16 categories, without significant drops on any particular class, underscores the robustness of our model. This balanced accuracy profile underscores the effectiveness of our multi-scale attention mechanism, which effectively captures both fine-grained details and global context, a critical capability for handling the diverse object scales and dense scenes present in DOTA 1.5.

Semantic Segmentation Experimental Results

UAVid Results. On the UAVid dataset, as detailed in Table 6, we compare LWGANet-L2 against a variety of state-of-the-art lightweight semantic segmentation methods. These include CoaT (Xu et al. 2021), SegFormer (Xie et al. 2021a), and several backbones integrated with the UnetFormer head (Wang et al. 2022a), such as ResNet18 (He et al. 2016), EfficientFormerV2-S2 (Li et al. 2023b), FasterNet-T2 (Chen et al. 2023), and DecoupleNet-D2 (Lu et al. 2024). Our LWGANet-L2 backbone

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP ↑
ResNet-50	79.95	81.00	53.90	70.59	52.48	76.21	86.98	90.88	78.33	68.26	58.94	72.60	72.75	65.32	58.18	3.72	66.88
ARC-R50	80.27	82.40	54.57	73.03	52.37	80.28	87.93	90.88	83.33	69.18	57.37	72.35	71.97	65.40	68.35	3.24	68.31
LSKNet-S	72.05	84.94	55.41	74.93	52.42	77.45	81.17	90.85	79.44	69.00	62.10	73.72	77.49	75.29	55.81	42.19	70.26
DecoupleNet-D2	80.35	82.36	54.00	73.00	52.30	81.41	88.31	90.89	80.22	69.27	58.50	68.41	75.99	70.93	72.92	39.54	71.15
PKINet-S	80.31	85.00	55.61	74.38	52.41	76.85	88.38	90.87	79.04	68.78	67.47	72.45	76.24	74.53	64.07	37.13	71.47
*E-FormerV2-S2	80.22	82.44	51.78	72.85	52.26	77.82	88.43	90.87	79.57	68.19	62.46	70.15	77.21	71.92	76.37	52.43	72.19
LWGANet-L2 (ours)	80.00	84.40	55.31	74.10	52.46	82.26	88.84	90.86	79.34	69.03	65.47	72.06	76.98	74.74	75.24	45.40	72.91

Table 5: Experimental results on the DOTA 1.5 test set. All backbones are evaluated within the Oriented R-CNN detector.

achieves a state-of-the-art mean Intersection-over-Union (mIoU) of 69.1%. This result surpasses all lightweight competitors. While maintaining a competitive model size (12.6M parameters) and inference speed (67.3 FPS), LWGANet-L2 delivers a significant accuracy improvement of +1.3% mIoU over the strong ResNet18 baseline. It also outperforms other modern lightweight backbones like DecoupleNet-D2 and FasterNet-T2 by a substantial margin. The superior performance is consistent across most categories, with LWGANet-L2 achieving the highest IoU for five of the eight classes, including challenging ones like *Moving Car* and *Human*, demonstrating its powerful feature extraction capabilities.

Methods	Backbone	Params. (M) ↓	FLOPs (G) ↓	Speed (FPS) ↑	mIoU (%) ↑	IoU per category (%) ↑							
						Clutter	Building	Road	Tree	Vegetation	Moving Car	Static Car	Human
CoaT	CoaT-Mini	11.1	104.8	10.6	65.8	69.0	88.5	80.0	79.3	62.0	70.0	59.1	18.9
SegFormer	MiT-B1	13.7	63.3	31.3	66.0	66.6	86.3	80.1	79.6	62.3	72.5	52.5	28.5
UnetFormer	EfficientFormerV2-S2	12.7	35.8	54.1	65.2	63.8	82.5	78.9	78.0	63.0	73.5	51.7	29.9
	FasterNet-T2	13.3	49.1	198.8	65.7	65.3	86.2	79.6	78.8	62.1	70.9	54.8	28.1
	DecoupleNet-D2	6.8	32.1	-	65.8	65.1	85.4	80.6	78.8	62.1	74.1	49.7	30.8
	ResNet18	11.7	46.9	115.6	67.8	68.4	87.4	81.5	80.2	63.5	73.6	56.4	31.0
	LWGANet-L2	12.6	50.3	67.3	69.1	69.0	87.9	81.9	80.5	64.6	76.7	59.7	32.7

Table 6: Segmentation results on the UAVid test set. Speeds were evaluated on 1,024×1,024 inputs using an RTX 3090 GPU.

LoveDA Results. On the LoveDA benchmark, as shown in Table 7, we compare UnetFormer with an LWGANet-L2 backbone against several state-of-the-art models, including FactSeg (Ma et al. 2022), FarSeg (Zheng et al. 2020), UnetFormer (Wang et al. 2022a) (with a ResNet50 (He et al. 2016) backbone), RSSFormer (Xu et al. 2023), and LoveNAS (Wang et al. 2024). Our configuration again demonstrates its superiority by achieving the highest mIoU of 53.6%. Our model is not only the most accurate but also highly parameter-efficient. With only 12.58M parameters, it outperforms much larger models like FactSeg and FarSeg by over 3.5 percentage points in mIoU. Compared to the lightweight UnetFormer-ResNet18 baseline, our model provides a significant +1.2% mIoU gain for a marginal increase in model size (+0.85M parameters), indicating superior parameter utilization. LWGANet-L2's strong performance on diverse classes like *Barren* and *Forest* further validates its effectiveness for complex land-cover mapping tasks.

Decoder	Backbone	Params. (M) ↓	mIoU (%) ↑	IoU per category (%) ↑						
				Background	Building	Road	Water	Barren	Forest	Agriculture
FactSeg	ResNet50	33.44	50.0	42.51	54.62	55.88	77.96	16.51	44.72	57.81
FarSeg	ResNet50	31.37	50.1	43.15	55.41	55.91	78.88	16.51	43.94	56.96
UnetFormer	ResNet18	11.73	52.4	44.7	58.8	54.9	79.6	20.1	46.0	62.5
RSSFormer	RSS-B	30.82	52.4	52.38	60.71	55.21	76.29	18.73	45.39	58.33
LoveNAS	ResNet50	30.49	52.3	45.39	58.86	59.45	79.39	13.76	43.94	65.64
UnetFormer	**LWGANet-L2**	12.58	53.6	46.76	59.56	56.73	79.57	23.58	46.28	62.42

Table 7: Semantic segmentation results on the LoveDA test set, comparing with state-of-the-art models.

Qualitative Experimental Results

Object Detection Visualization

Figure 3 provides qualitative results for oriented object detection on the DOTA 1.0 test set. We compare LWGANet-L2 against other prominent backbones, including ARC-R50 (Pu et al. 2023), PKINet (Cai et al. 2024), FasterNet (Chen et al. 2023), and EfficientFormerV2 (Li et al. 2023b). The visualizations clearly demonstrate LWGANet's superior capability in handling objects with significant scale variations—a common challenge in aerial imagery. Our model accurately detects and delineates objects of vastly different sizes within the same scene, such as large harbors, medium-sized vessels, and small vehicles. In contrast, other lightweight backbones like FasterNet and EfficientFormerV2 exhibit robustness issues, such as missed detections for small vehicles and coarse, inaccurate bounding boxes for large vessels. In contrast, LWGANet's effective multi-scale feature fusion allows it to maintain high sensitivity to small targets while preserving precise localization for large objects, as evidenced by the tight and accurate bounding boxes. This visual evidence corroborates our quantitative findings and highlights the effectiveness of LWGANet's multi-scale feature representation.

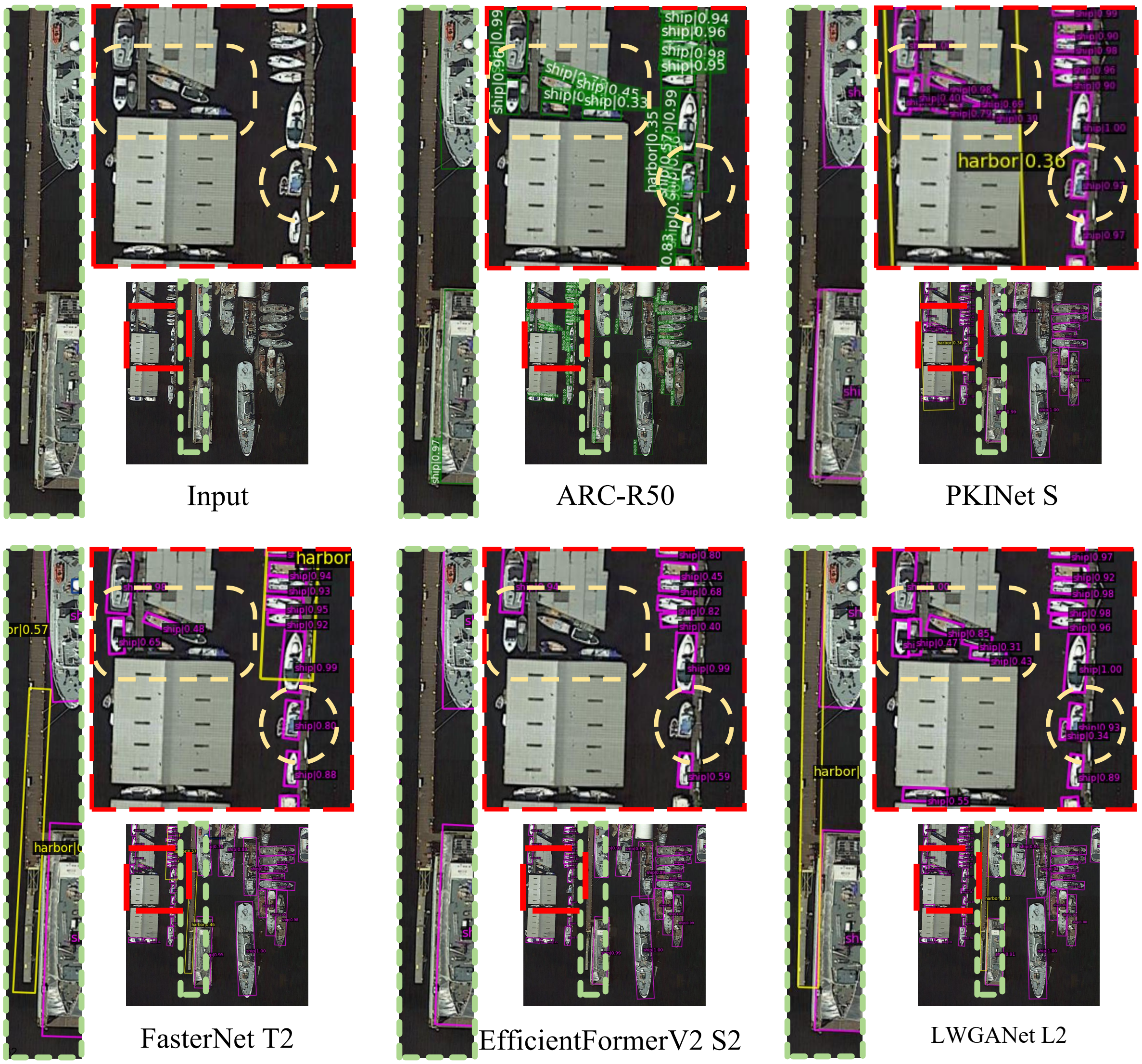

Figure 3: Qualitative comparison on the DOTA 1.0 test set. All backbones are integrated with the Oriented R-CNN detector. LWGANet demonstrates superior performance in detecting multi-scale objects, correctly identifying both large harbors and small ships, whereas other methods exhibit missed detections or inaccurate localizations.

Semantic Segmentation Visualization

The qualitative results for semantic segmentation on the LoveDA test set are shown in Figure 4. The visualizations highlight the distinct advantage conferred by the LWGA module. LWGANet produces segmentation maps with remarkably sharp and accurate boundaries, especially for fine-grained structures like roads and buildings. Compared to other methods, our model generates more coherent and detailed predictions, minimizing noise and correctly delineating complex geometries. This visual superiority underscores the efficacy of our multi-scale attention mechanism in capturing the intricate spatial details crucial for high-fidelity semantic segmentation in RS imagery.

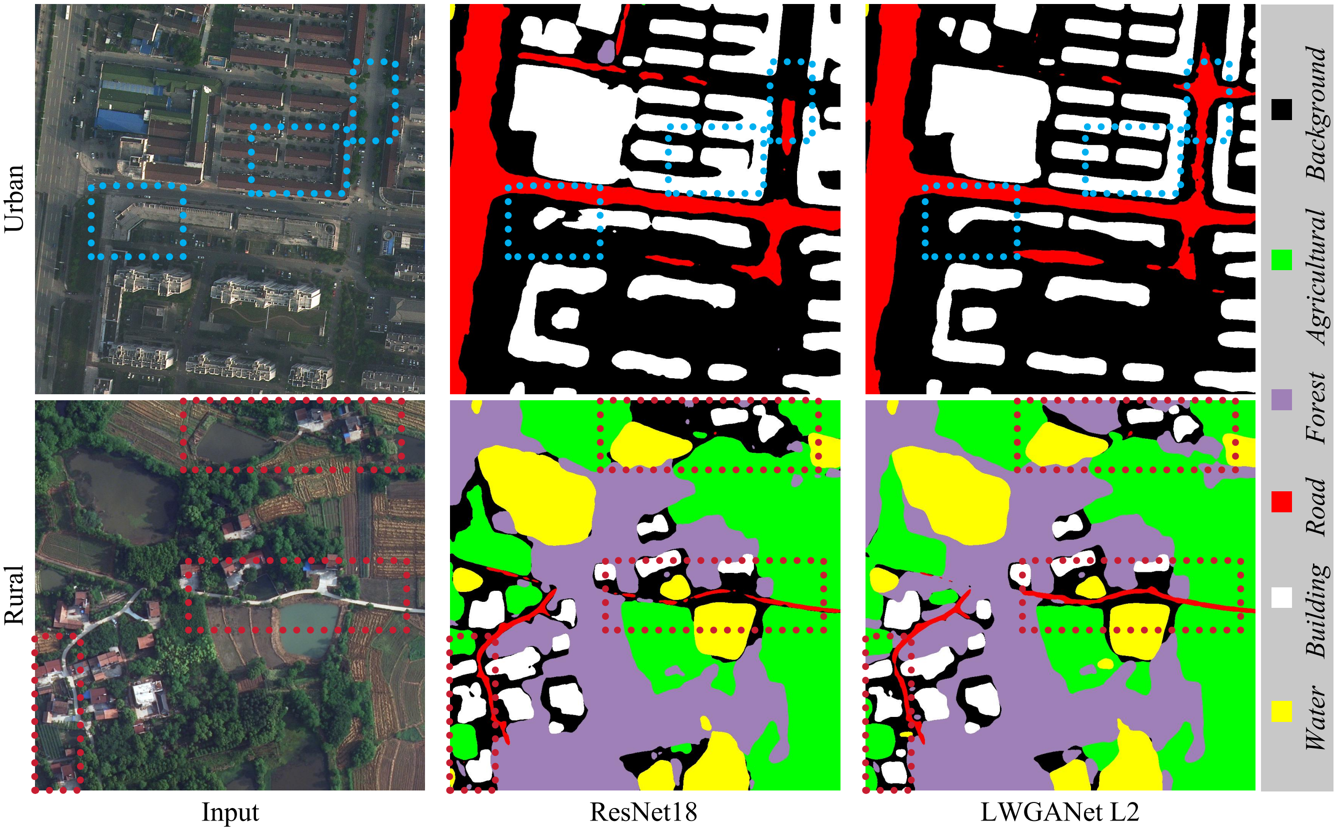

Figure 4: Qualitative results on the LoveDA test set, using UnetFormer as the segmentation head. LWGANet produces cleaner segmentation maps with more precise boundaries for roads and buildings, demonstrating its superior capability in capturing fine-grained details compared to other backbones.

Ablation Study Results

To validate the contribution of each component within the (`LWGA module`), we conducted a comprehensive ablation study, the results of which are presented in Table 8. The study evaluates the impact of the GPA, RLA, SMA, and SGA modules across four downstream tasks: classification, detection, segmentation, and change detection.

Our baseline model, shown in the first row, utilizes only the RLA module for feature extraction. Subsequent rows show the performance as we systematically remove one attention component at a time from the full LWGANet-L2 architecture (final row). To isolate the architectural contributions and ensure experimental efficiency, all models in this study were trained from scratch without ImageNet pre-training.

The results consistently demonstrate the positive contribution of each module. For instance, removing any single component generally leads to a performance drop across most tasks. The full LWGANet architecture, which combines all four attention mechanisms, achieves the highest performance in classification (96.17% accuracy), detection (71.59% mAP), and change detection (83.95% IoU).

An interesting observation from our ablation study is the seemingly modest impact of removing the Sparse Global Attention (SGA) module on the UAVid semantic segmentation task. We hypothesize this is attributable to the specific characteristics of the UAVid dataset. UAVid imagery often features lower-altitude, oblique views with a high density of small- to medium-sized objects. In such scenarios, local and medium-range context, effectively captured by the RLA and SMA modules, may be sufficient for robust pixel-level classification, diminishing the marginal benefit of long-range dependencies modeled by SGA.

However, this observation appears to be dataset-specific, as the value of global context is more pronounced in other tasks. For instance, in our main experiments, LWGANet achieved state-of-the-art results on the LoveDA dataset, where distinguishing between broad, similar-textured land-cover types benefits from a wider contextual view. Similarly, in change detection, a holistic understanding of bi-temporal scenes is critical for identifying large-scale structural changes and minimizing false positives, a role well-served by the SGA module. Therefore, while its contribution may vary with task-specific contextual requirements, the SGA module remains a vital component of LWGANet's architecture, ensuring its versatility and robustness across a wide spectrum of RS applications that demand comprehensive global scene understanding.

GPA	RLA	SMA	SGA	Classification (NWPU val)				Detection (DOTA 1.0 val)		Segmentation (UAVid test)		Change Detection (LEVIR-CD test)	
				Params.	FLOPs	Acc. (%) ↑	FPS ↑	Params. ↓	mAP (%) ↑	Params. ↓	mIoU (%) ↑	Params. ↓	IoU (%) ↑
	✓			27.69	4.35	95.35	3819	43.9	69.64	27.2	62.03	28.96	83.66
✓	✓	✓		12.54	1.84	95.83	3983	28.6	70.56	12.1	62.16	13.80	83.23
✓	✓		✓	12.90	1.87	95.51	3942	29.1	71.10	12.4	61.82	14.17	83.73
✓		✓	✓	11.95	1.70	95.86	3490	28.2	70.09	11.5	61.71	13.21	83.50
	✓	✓	✓	12.06	1.71	95.98	3638	28.3	71.16	11.6	61.90	13.33	83.74
✓	✓	✓	✓	13.01	1.87	96.17	3308	29.2	71.59	12.6	62.05	14.30	83.95

Table 8: Ablation study of attention components in the LWGA Block. The first row represents a baseline using only the RLA module. The final row is the full LWGANet-L2 model. Decoders used are Oriented R-CNN (Detection), UnetFormer (Segmentation), and A2Net (Change Detection).

To validate our hypothesis that a synergistic, multi-pathway design is superior to any single-paradigm, we conducted an extensive ablation study on the DOTA-v1.0 dataset. We constructed several variants of our network, each exclusively using one type of attention module (GPA, RLA, SMA, or SGA) throughout all blocks. These experiments directly compare the feature extraction capabilities of the standalone modules against the full LWGANet, ensuring a fair comparison by utilizing comparable parameters and FLOPs (stem dims set to 64, 64, 96, 96, and 96, respectively). For this study, all backbones were pre-trained on ImageNet-1K for 100 epochs, and the Oriented R-CNN detector was subsequently trained on the DOTA-v1.0 train set for 30 epochs. Performance is reported on the DOTA-v1.0 validation set.

As summarized in Table 9, while the standalone modules exhibit predictable specializations, they are ultimately limited in scope. The convolution-based RLA excelled on objects with regular textures (e.g., Basketball Court), while the attention-based SMA and SGA performed better on large, amorphous structures (e.g., Harbor, Roundabout). However, each struggled outside its optimal domain; RLA failed on irregular objects (Helicopter), and SGA was less effective for small ones (Small Vehicle). This confirms that a single operator type cannot cover the full feature spectrum required by RS data. In stark contrast, the full LWGANet-L2, which integrates all four pathways, achieved a significantly higher mAP of 74.1%. It demonstrated a balanced and robust performance across nearly all categories. This substantial improvement is not merely additive; it highlights a synergistic effect where the combined, decoupled features create a more comprehensive and powerful representation than the sum of their parts. This result provides compelling evidence for our feature decoupling.

Method	#P ↓	FLOPs ↓	mAP ↑	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
GPA	28.1	157.0	68.4	89.1	73.9	46.9	70.7	62.1	82.2	88.8	90.6	71.8	60.8	63.3	61.8	67.1	54.8	42.4
RLA	29.0	159.9	70.3	89.8	75.7	49.2	75.9	65.2	84.3	89.0	90.7	76.5	62.3	64.8	65.6	75.2	55.1	34.4
SMA	27.1	152.0	70.5	89.7	77.1	47.5	78.2	64.5	85.2	89.1	90.7	68.1	62.2	64.3	65.1	76.4	50.1	49.1
SGA	34.3	161.7	71.0	89.6	77.0	48.8	73.4	59.0	85.2	89.0	90.6	72.3	61.6	66.1	66.7	74.8	57.7	52.8
LWGANet-L2	29.2	159.1	74.1	90.0	79.7	57.8	79.8	69.4	85.4	89.6	90.7	74.1	62.1	70.1	67.3	76.2	60.1	59.2

Table 9: Ablation experimental results on DOTA 1.0 val set with single-scale training and testing. We compare the LWGANet-L2 against variants constructed using only a single attention type, demonstrating the necessity of the multi-pathway design.

Further Design Rationale and Discussion

This appendix provides additional details and discussions to complement the main paper. We aim to offer deeper insights into the design choices of LWGANet, elaborate on the relationship between our work and existing methods, and discuss the broader applicability of our proposed principles.

Relationship to Prior Works

LWGA: From Homogeneous Grouping to Heterogeneous Multi-Scale Representation. Channel grouping is a well-established strategy for building efficient neural networks, as exemplified by the grouped convolutions in ShuffleNet (Zhang et al. 2018), depthwise separable convolutions in MobileNets (Sandler et al. 2018), and multi-head self-attention in Vision Transformers (Dosovitskiy et al. 2020). A common characteristic of these methods is their reliance on **homogeneous operations**, where each channel group is processed by identical, replicated computational units. While effective for general-purpose vision tasks, this paradigm proves suboptimal for RS imagery, which is characterized by extreme variations in object scale and morphology. In such contexts, a uniform feature extraction strategy leads to a specific form of representational inefficiency: channels become overly specialized to certain scales, resulting in redundancy and a diminished capacity to model the full spectrum of visual concepts present in RS data.

Our LWGA module introduces a departure from this homogeneous paradigm. It implements a **heterogeneous, multi-scale architecture** within a single block. Instead of merely parallelizing identical operations, LWGA partitions the feature channels and routes each subset through a distinct, computationally specialized pathway, each engineered to capture features at a specific scale—from point-wise details to global context. This design moves beyond simple ensembling, representing a detailed approach to **decouple the multi-scale representation learning problem into specialized, non-competing sub-tasks**. By assigning specialized operators to different channel subspaces, LWGA mitigates inter-scale interference and promotes a more comprehensive feature representation. The state-of-the-art results across diverse RS tasks validate this design's efficacy. More broadly, our work advocates for heterogeneous designs as a powerful tool for building efficient models for domains with high intra-class and inter-scale variance, offering a compelling alternative to conventional, homogeneous architectures.

TGFI: A Lightweight Mechanism for Sparse Global Interaction. The quadratic complexity of self-attention has spurred extensive research into sparse attention mechanisms. Our TGFI module aligns with this research direction but is differentiated by its design principles of simplicity and non-parametric efficiency. While many existing token sparsification methods employ learnable routing modules or computationally intensive clustering algorithms—introducing their own parametric or computational overhead—TGFI adopts a **simple, non-parametric sampling strategy based on feature activation magnitudes, making it computationally lean**. This approach is particularly well-suited for RS data, where salient foreground objects are often sparsely distributed across vast, low-information backgrounds. Critically, TGFI is not designed as a standalone attention replacement but as an integral component of the LWGA block. It functions as an efficient mechanism to enable a full-fledged global attention pathway on a reduced set of high-value tokens, ensuring that LWGANet can model long-range dependencies without prohibitive computational cost.

Detailed Implementation and Analysis of the TGFI Module

TGFI Implementation Details. The core concept of TGFI is to perform computationally expensive interactions on a sparse, salient subset of features to mitigate spatial redundancy. In our architecture, this principle is realized through a simple yet effective non-parametric mechanism using standard PyTorch layers, primarily `nn.MaxPool2d` with `return_indices=True` and its counterpart, `nn.MaxUnpool2d`. This approach directly implements "Top-1" selection from a defined spatial region. The process can be broken down into three steps:

1. **Sparse Feature Sampling:** We apply a max-pooling layer to the input feature map. The 'return_indices=True' argument is critical, as it not only returns the downsampled tensor (containing the maximum, or "most salient," value from each region) but also the spatial indices of these maxima within the original tensor. This step effectively creates a compact summary of the most informative features while preserving their original locations.
2. **Subspace Interaction:** The subsequent computationally-intensive operations (e.g., the convolutional proxy in stages 1-2, or self-attention in stages 3-4) are performed exclusively on this smaller, downsampled feature map. This drastically reduces the number of tokens and, consequently, the computational cost.
3. **Feature Restoration:** After the interaction, the enhanced feature map is restored to its original spatial resolution. For the SGA pathways (modules `GA12` and `D_GA`), we use `nn.MaxUnpool2d`, which takes the processed sparse tensor and the saved indices from the sampling step to place the enhanced features back into their exact original locations. This preserves spatial fidelity, which is crucial for dense prediction tasks. For the SMA pathway (`MRA` module), which produces an attention map, the resulting map is upsampled to the original resolution using nearest-neighbor interpolation (`F.interpolate`) before being applied to the original input feature map.

Hyperparameter Rationale (Region Size). The primary hyperparameter for TGFI is the "region size," which corresponds to the kernel size and stride of the max-pooling operation. Our design uses fixed, stage-aware region sizes based on empirical validation during model development:

- **For SMA (`MRA` module):** We employ a 3×3 downsampling region (implemented via an antialiased blur pool with a stride of 3). This moderately reduces the token count, allowing the subsequent convolutions to capture medium-range dependencies efficiently.
- **For SGA (modules `GA12`, `D_GA`):** We use a non-overlapping 2×2 region (`kernel_size=2, stride=2`). This choice represents a robust trade-off: it reduces the number of tokens by 75%, leading to a significant reduction in the computational complexity of global self-attention, while retaining sufficient spatial granularity to model long-range dependencies effectively.

While we did not perform an exhaustive sweep, these settings provided a consistent and strong balance between performance and efficiency across all tested tasks. Exploring adaptive or larger region sizes remains a valid direction for future work.

Comparison with Other Token Reduction Methods. TGFI's design philosophy distinguishes it from other contemporary token reduction techniques, positioning it as a lightweight and pragmatic solution for RS imagery.

- **vs. Token Merging (e.g., ToMe):** Token merging methods progressively combine similar tokens based on a learned similarity metric. While adaptive, this process introduces computational overhead for calculating token affinity. In contrast, TGFI is a **non-parametric and deterministic** method that relies on a simple saliency proxy (max activation), making it computationally lighter and simpler to implement.
- **vs. Token Pruning (e.g., Token Squeezing):** Token pruning often involves a learnable scoring module that decides which tokens to discard entirely. This adds parameters and an extra forward pass for the scoring network. TGFI, being parameter-free, avoids this overhead. Furthermore, by using a restoration mechanism (`MaxUnpool2d`), TGFI ensures that the spatial structure of the feature map is preserved, whereas pruning permanently discards spatial locations, which can be detrimental for dense prediction tasks like segmentation.

In summary, TGFI was intentionally designed as a simple, efficient, and parameter-free mechanism that leverages standard, hardware-friendly operations to effectively mitigate spatial redundancy, aligning perfectly with the overall goal of creating a truly lightweight and practical backbone for remote sensing.

Discussion on Potential Broader Impacts

The core challenge addressed in this work—efficiently modeling data with high spatial redundancy and channel redundancy—is not unique to RS. We note that similar data characteristics are prevalent in other important domains. For instance, digital pathology with its gigapixel-scale images and high-resolution document analysis present analogous challenges of vast, uninformative backgrounds interspersed with critical, multi-scale features.

The design of LWGANet, particularly the idea of decoupling feature representation into specialized pathways, might offer a valuable perspective for researchers in these fields. This approach of decomposing a network into heterogeneous branches to address distinct task requirements is an emerging trend that not only enhances model efficiency and performance but also offers a more modular design philosophy. For instance, in all-in-one image restoration, researchers leverage multi-expert architectures to collaboratively tackle mixed degradations, as demonstrated in tasks for both general-purpose and specific underwater scenarios (Zhang et al. 2025a,b). Similarly, in other computer vision tasks, constructing functionally heterogeneous branches to separately capture low-quality features (Lu et al. 2025) or to disentangle image representations (Lu et al. 2026) has also proven to be an effective strategy.

While a direct application would require domain-specific adaptations and validation, we believe our work could serve as a conceptual starting point. We hope that our findings encourage the exploration of similar function-specialized, heterogeneous architectures for efficient visual analysis beyond the scope of RS.

Note on Experimental Protocols and Reproducibility

On Statistical Significance and Multiple Runs. We acknowledge the importance of reporting measures of variation (e.g., standard deviation over multiple runs) and conducting statistical tests to validate results. Due to the extensive scale of our evaluation—spanning 12 datasets across 4 distinct tasks—performing multiple training runs for every experiment was computationally prohibitive. Instead, we adopted an alternative strategy to ensure the reliability of our claims. We focused on demonstrating the **consistency and generalizability** of LWGANet's performance advantage. Our results show that LWGANet consistently resides on or near the Pareto frontier across this wide spectrum of benchmarks. The significant performance margins observed across diverse data distributions and task requirements provide strong evidence that the improvements are robust and not an artifact of random initialization. For all experiments, we followed deterministic training protocols with fixed random seeds to ensure direct reproducibility.

References

Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; and Yao, Y. 2024. Poly Kernel Inception Network for Remote Sensing Detection. In *CVPR*.

Chen, H.; and Shi, Z. 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *RS*, 12(10): 1662.

Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; and Chan, S.-H. G. 2023. Run, Don't walk: Chasing higher FLOPS for faster neural networks. In *CVPR*, 12021–12031.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE*, 105(10): 1865–1883.

Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; and Han, J. 2022. Anchor-Free Oriented Proposal Generator for Object Detection. *IEEE TGRS*, 60: 1–11.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 702–703.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

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*, 2849–2858.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Han, J.; Ding, J.; Xue, N.; and Xia, G.-S. 2021. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In *CVPR*, 2786–2795.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Ji, S.; Wei, S.; and Lu, M. 2018. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE TGRS*, 57(1): 574–586.

Lebedev, M.; Vizilter, Y. V.; Vygolov, O.; Knyaz, V. A.; and Rubis, A. Y. 2018. Change detection in remote sensing images using conditional adversarial networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42: 565–571.

Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; and Li, X. 2023a. Large Selective Kernel Network for Remote Sensing Object Detection. In *ICCV*, 16794–16805.

Li, Y.; Hu, J.; Wen, Y.; Evangelidis, G.; Salahi, K.; Wang, Y.; Tulyakov, S.; and Ren, J. 2023b. Rethinking Vision Transformers for MobileNet Size and Speed. In *ICCV*, 16889–16900.

Li, Z.; Tang, C.; Liu, X.; Zhang, W.; Dou, J.; Wang, L.; and Zomaya, A. Y. 2023c. Lightweight Remote Sensing Change Detection With Progressive Feature Aggregation and Supervised Attention. *IEEE TGRS*, 61: 1–12.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *Int. Conf. Learn. Represent. (ICLR)*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, W.; Chen, S.-B.; Li, H.-D.; Shu, Q.-L.; Ding, C. H. Q.; Tang, J.; and Luo, B. 2025. LEGNet: A Lightweight Edge-Gaussian Network for Low-Quality Remote Sensing Image Object Detection. In *ICCVW*, 2844–2853.

Lu, W.; Chen, S.-B.; Shu, Q.-L.; Tang, J.; and Luo, B. 2024. DecoupleNet: A Lightweight Backbone Network With Efficient Feature Decoupling for Remote Sensing Visual Tasks. *IEEE TGRS*, 62: 1–13.

Lu, W.; Chen, S.-B.; Tang, J.; Ding, C. H.; and Luo, B. 2023. A Robust Feature Downsampling Module for Remote Sensing Visual Tasks. *IEEE TGRS*, 61: 1–12.

Lu, W.; Li, H.-D.; Wang, C.; Chen, S.-B.; Ding, C. H.; Tang, J.; and Luo, B. 2026. UnravelNet: A Backbone for Enhanced Multi-Scale and Low-Quality Feature Extraction in Remote Sensing Object Detection. *ISPRS*, 231: 431–442.

Lyu, Y.; Vosselman, G.; Xia, G.-S.; Yilmaz, A.; and Yang, M. Y. 2020. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS*, 165: 108–119.

Ma, A.; Wang, J.; Zhong, Y.; and Zheng, Z. 2022. FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery. *IEEE TGRS*, 60: 1–16.

Mehta, S.; and Rastegari, M. 2022. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. In *ICLR*.

Pan, J.; Bulat, A.; Tan, F.; Zhu, X.; Dudziak, L.; Li, H.; Tzimiropoulos, G.; and Martinez, B. 2022. EdgeViTs: Competing Light-Weight CNNs on Mobile Devices with Vision Transformers. In *ECCV*, 294–311.

Pu, Y.; Wang, Y.; Xia, Z.; Han, Y.; Wang, Y.; Gan, W.; Wang, Z.; Song, S.; and Huang, G. 2023. Adaptive Rotated Convolution for Rotated Object Detection. In *ICCV*, 6589–6600.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 4510–4520.

Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; and Zhang, L. 2021. A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection. *IEEE TGRS*, 60: 1–16.

Wang, G.; Cheng, G.; Zhou, P.; and Han, J. 2023. Cross-Level Attentive Feature Aggregation for Change Detection. *IEEE TCSVT*.

Wang, J.; Zheng, Z.; Lu, X.; and Zhong, Y. 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In *Conf. on Neur. Inf. Proc. Syst. Datasets and Benchmarks Track*.

Wang, J.; Zhong, Y.; Ma, A.; Zheng, Z.; Wan, Y.; and Zhang, L. 2024. LoveNAS: Towards multi-scene land-cover mapping via hierarchical searching adaptive network. *ISPRS*, 209: 265–278.

Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; and Atkinson, P. M. 2022a. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS*, 190: 196–214.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022b. PVT v2: Improved baselines with Pyramid Vision Transformer. *CVM*, 8(3): 415–424.

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*, 3974–3983.

Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; and Lu, X. 2017. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE TGRS*, 55(7): 3965–3981.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021a. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS*, volume 34, 12077–12090.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021b. Oriented R-CNN for Object Detection. In *ICCV*, 3520–3529.

Xu, R.; Wang, C.; Zhang, J.; Xu, S.; Meng, W.; and Zhang, X. 2023. RSSFormer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation. *TIP*, 32: 1052–1064.

Xu, W.; Xu, Y.; Chang, T.; and Tu, Z. 2021. Co-Scale Conv-Attentional Image Transformers. In *ICCV*, 9981–9990.

Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; and Bai, X. 2020. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE TPAMI*, 43(4): 1452–1459.

Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In *ICCV*, 8232–8241.

Yang, Y.; and Newsam, S. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proc. 18th SIGSPATIAL Int. Symp. on Adv. in Geogr. Inf. Syst.*, 270–279.

Zhang, X.; Ma, J.; Wang, G.; Zhang, Q.; Zhang, H.; and Zhang, L. 2025a. Perceive-IR: Learning to Perceive Degradation Better for All-in-One Image Restoration. *TIP*, 1–1.

Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L.; and Du, B. 2025b. UniUIR: Considering Underwater Image Restoration as an All-in-One Learner. *TIP*, 34: 6963–6977.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *CVPR*, 6848–6856.

Zheng, Z.; Zhong, Y.; Wang, J.; and Ma, A. 2020. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In *CVPR*, 4096–4105.

Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; Zhang, W.; and Chen, K. 2022. MMRotate - A Rotated Object Detection Benchmark using PyTorch. In *ACM MM*, 7331–7334.