

CSRN: Cross-Sensor Robust Recognition Network for Multi-modal Aerial View Object Classification

Hongli Liu¹ Yu Wang^{1*} Shengjie Zhao^{1*}

¹School of Computer Science and Technology, Tongji University

2411501@tongji.edu.cn yuwangtj@yeah.net shengjiezhaot@tongji.edu.cn

Abstract

Target detection and classification in aerial imagery presents significant challenges due to the scarcity of target information. Electro-Optical (EO) images have limited resolution and perform poorly under adverse weather conditions. On the other hand, Synthetic Aperture Radar (SAR) images are capable of effective detection in diverse weather and low-light environments, but their performance is hindered by speckle noise, which impairs deep learning models' ability to extract meaningful features. Therefore, the use of a single sensor may not achieve the desired accuracy. To address this challenge, we propose a Cross-Sensor Robust Recognition Network (CSRN) that leverages the complementary advantages of EO and SAR imagery to overcome their individual limitations and improve the performance of Automatic Target Recognition (ATR) systems. Specifically, we design a cross-modal adaptation framework that learns a domain-invariant feature space, effectively mitigating domain discrepancies between different sensor data. This framework enhances the robustness and classification accuracy of the system by strengthening cross-modal feature learning. Experimental results demonstrate the superiority of the proposed approach on the PBVS MAVOC 2025 challenge dataset, achieving **1st place** in SAR classification with a total score of **0.43** and a Top-1 accuracy of **31.78%**. This framework provides a novel solution for improving multi-source information fusion and target recognition accuracy. The code for our CSRN framework is publicly available at: https://github.com/HongliLiu1/CSRN_PBVS2025.

1. Introduction

Automatic Target Recognition (ATR) in aerial imagery plays a pivotal role in remote sensing applications such as environmental monitoring, disaster response, and strategic surveillance [13]. With the advancement of deep learn-

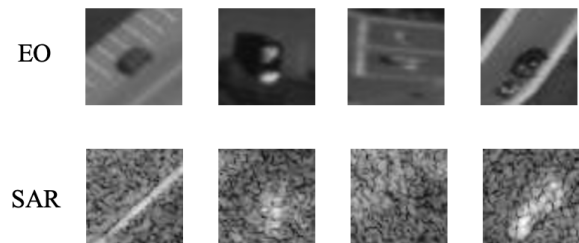


Table 1. A few EO samples along with their corresponding SAR samples from the PBVS MAVOC challenge dataset.

ing, Electro-Optical (EO) based object recognition has achieved remarkable success by leveraging large-scale annotated datasets and powerful convolutional backbones [6]. However, EO systems suffer from performance degradation under adverse conditions such as cloud cover, fog, and low illumination.

In contrast, Synthetic Aperture Radar (SAR) provides robust imaging capabilities independent of weather or lighting, making it an indispensable modality for all-weather, day-and-night operations [18]. Despite these advantages, SAR-based object recognition remains significantly less developed than its EO counterpart, due to two persistent challenges: (1) SAR images suffer from strong speckle noise, which introduces semantic ambiguity and disrupts discriminative feature extraction; and (2) aerial datasets often exhibit severe class imbalance, where minority classes are underrepresented and underperforming in standard training pipelines [2].

Recent advances in multi-modal learning have motivated the fusion of EO and SAR data to exploit their complementary characteristics. Nevertheless, existing cross-modal domain adaptation approaches—such as bidirectional teacher-student frameworks [24] and Sliced Wasserstein Distance (SWD)-based alignment methods [18]—remain suboptimal. These methods typically overlook two critical aspects: (i) the asymmetric noise and texture statistics between EO and SAR modalities, and (ii) the long-tailed nature of real-world object distributions, which is especially problematic

*Corresponding Author.

in low-shot SAR domains. As shown in Fig. 1, SAR images often exhibit degraded resolution and object indistinctness, further exacerbating these issues.

To tackle these challenges, we propose a novel **Cross-Sensor Robust Recognition Network (CSRN)** that integrates multi-modal representation learning, class-balanced training, and domain-level alignment to enhance fine-grained SAR object recognition. Specifically, CSRN begins by addressing the severe class imbalance inherent in aerial datasets through a class-balanced data partitioning strategy. By employing stratified sampling, the dataset is split into balanced labeled and unlabeled subsets, ensuring proportional representation across head and tail classes and thus stabilizing the training process. Following this, CSRN performs dual-modal feature extraction, where an EfficientNet encoder captures spatial detail from EO imagery, and a modified ResNet-50 processes SAR inputs to extract noise-resilient representations. This design allows the model to leverage the complementary advantages of both modalities. Finally, to reduce the domain discrepancy between EO and SAR data, CSRN introduces a cross-modal distribution alignment module based on Sliced Wasserstein Distance (SWD), aligning features in a domain-invariant latent space and enhancing cross-modal generalization.

In summary, our contributions are outlined as follows:

- We propose a novel Cross-Sensor Robust Recognition Network (CSRN) that integrates multi-modal learning, class rebalancing, and domain adaptation to enhance SAR-based object recognition in aerial imagery.
- We design a class-balanced data partitioning strategy to alleviate long-tail distribution issues, and develop a dual-modal feature extraction module combining EfficientNet and a modified ResNet-50, along with a Sliced Wasserstein Distance (SWD)-based alignment mechanism to enable domain-invariant and complementary cross-modal representation learning.
- Evaluation results on the PBVS MAVOC 2025 challenge dataset demonstrate that our approach achieves **1st place** in the SAR classification task, with a Top-1 accuracy of **31.78%**.

2. Related Work

2.1. Multi-modal EO-SAR Fusion

The evolution of SAR image classification has transitioned from physics-driven approaches to data-centric deep learning. Early methods relied on polarimetric decomposition [4] and Wishart classifiers [10], heavily dependent on expert-designed scattering models. The emergence of deep neural networks, particularly through transfer learning from AlexNet [8] and ResNet [6], enabled automatic and scalable feature extraction [3], marking a fundamental shift in SAR image understanding. More recently, metric-based few-shot

learning [25] and transformer-based architectures [23] have been introduced, but remain sensitive to limited data and class imbalance.

In the context of multi-modal fusion, combining EO and SAR data has shown promise due to their complementary sensing characteristics. Prior efforts have explored early/late fusion [28], attention-guided integration [14], and semi-supervised transfer between modalities [18]. However, most methods suffer from oversimplified alignment schemes that fail to capture modality-specific noise asymmetry and intra-class variance. Cross-modal domain adaptation techniques, especially EO→SAR transfer, have leveraged adversarial discriminators [5] for feature alignment, but often struggle with mode collapse in high-dimensional spaces. Optimal transport methods, such as Sliced Wasserstein Distance (SWD) [16], offer a more stable alternative for distribution alignment, owing to their mathematical rigor and gradient stability [7]. Rostami *et al.* [17] applied SWD for global domain alignment, but their approach overlooked category-level structure. In contrast, our CSRN explicitly aligns EO and SAR features in a shared latent space using category-aware SWD-based alignment, improving class-consistent transfer and robustness to domain shift.

2.2. Long-tailed Recognition

Long-tailed class distributions are prevalent in aerial recognition scenarios, where dominant classes overwhelm rare categories in both labeled and unlabeled data [27]. Traditional approaches such as SMOTE [1] and focal loss [11] are known to degrade under extreme data scarcity. Hybrid methods [26] and instance re-weighting schemes [29] provide alternatives, yet introduce significant hyperparameter tuning complexity or overfitting risks. To address these challenges, our method departs from conventional oversampling and loss shaping by employing a dual-strategy. First, we adopt a stratified sampling approach to preserve the natural class distribution in labeled data, ensuring representative supervision across the label space. Second, we incorporate uncertainty-aware selection [20] to enhance the diversity of unlabeled samples, thus maximizing learning utility without relying on synthetic data generation. Compared to prior solutions, our approach is architecture-agnostic, tuning-free, and seamlessly integrates with multi-modal pipelines. It further improves performance on rare categories by enhancing both data representativeness and distributional balance.

3. Methodology

3.1. Overview Architecture

The proposed Cross-Sensor Robust Recognition Network (CSRN) establishes a unified end-to-end pipeline for robust

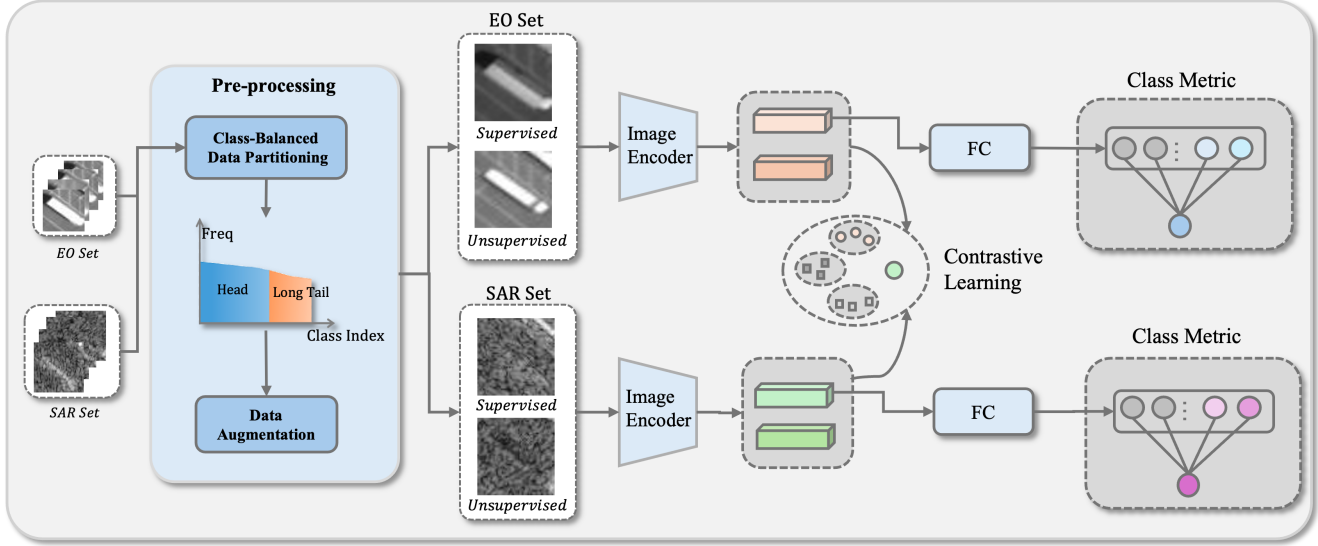


Figure 1. Overview of CSRN: a dual-stream architecture that integrates class-balanced sampling, EO–SAR feature extraction, cross-modal fusion, and SWD-based alignment. The model is jointly supervised by classification and alignment objectives for robust multi-modal recognition.

object recognition across Electro-Optical (EO) and Synthetic Aperture Radar (SAR) modalities, targeting two core challenges in multi-modal aerial recognition: severe class imbalance and cross-modal domain shift (Fig. 1).

To mitigate distribution imbalance, CSRN first applies a class-balanced sampling strategy (Sec. 3.2) based on stratified splitting and dynamic resampling, ensuring representative supervision across head and tail categories. A dual-stream backbone (Sec. 3.3) then extracts modality-specific features, with an EfficientNet encoder for EO data and a modified ResNet-50 for SAR, preserving spatial and geometric characteristics respectively. These features are fused via a cross-modal integration module that retains modality-specific information while promoting semantic alignment. To further address domain discrepancy, CSRN incorporates a Sliced Wasserstein Distance (SWD)-based alignment mechanism (Sec. 3.4), which aligns distributions in a shared latent space to facilitate domain-invariant learning.

Together, these components form a coherent pipeline that jointly optimizes representation learning, distribution rebalancing, and cross-modal adaptation, ultimately improving recognition performance in challenging aerial scenarios.

3.2. Class-Balanced Sampling Strategy

To address the challenge of long-tailed distributions in multi-sensor aerial datasets, we propose a structured three-stage preprocessing framework that integrates stratified splitting, dynamic resampling, and modality-specific augmentation. This pipeline rebalances class distributions while enhancing cross-modal generalization under sensor-

specific perturbations.

We first perform stratified sampling to partition the dataset into labeled (\mathcal{D}_L) and unlabeled (\mathcal{D}_U) subsets, ensuring that the original class distribution is preserved across both splits. Maintaining inter-subset consistency in class frequency reduces bias introduced by random imbalance and provides a more stable foundation for model training.

To mitigate distribution skew from dominant and rare classes, we apply a dynamic resampling mechanism that adjusts the number of training samples per class based on their frequency. Specifically, we define the head class set as:

$$C_{\text{head}} = \{j \mid N_j > \tau_{\text{head}}\}, \quad \tau_{\text{head}} = 1000, \quad (1)$$

where N_j denotes the number of samples in class j . For head classes, we downsample using a fixed scaling factor:

$$N_j^{\text{train}} = \lfloor \rho_{\text{head}} N_j \rfloor, \quad \rho_{\text{head}} = 0.8. \quad (2)$$

In contrast, tail classes are defined as:

$$C_{\text{tail}} = \{j \mid N_j < \tau_{\text{tail}}\}, \quad \tau_{\text{tail}} = 800. \quad (3)$$

For these underrepresented classes, we apply upsampling with a dynamic ratio:

$$\rho_{\text{tail}} = \min\left(\frac{\tau_{\text{tail}}}{N_j}, 3\right), \quad (4)$$

$$N_j^{\text{train}} = \min(\tau_{\text{tail}}, \lceil \rho_{\text{tail}} N_j \rceil), \quad (5)$$

To handle extreme cases, an exponential decay function is used as an alternative downsampling scheme for head classes:

$$N_j^{\text{train}} = N_j \cdot e^{-\lambda(N_j - \tau_{\text{head}})}, \quad (6)$$

where λ controls the decay rate based on the class frequency deviation.

Finally, we introduce modality-specific data augmentation to improve robustness under sensor-induced variation. For EO images, we apply `ColorJitter` to adjust brightness, contrast, and saturation, along with `RandomAffine` for rotations of $\pm 15^\circ$ and scaling factors sampled from $[0.8, 1.2]$ [21]. For SAR images, we introduce speckle noise using `SpeckleNoise` ($\sigma = 0.2$) [9] and apply `Resample` ($r_{\text{decay}} = 0.7$) to simulate resolution degradation [19]. These transformations embed sensor-aware perturbations into the training data, enhancing the model’s generalization capacity across modalities.

3.3. Dual-Network Feature Extraction

To effectively exploit the complementary properties of EO and SAR modalities, CSRN employs a parallel dual-stream feature extraction backbone. Specifically, EO inputs are processed using an EfficientNet-B4 encoder [22], which captures fine-grained spatial and spectral patterns. In contrast, SAR inputs are handled by a modified ResNet-50 [6] encoder, designed to be robust against speckle noise and geometric distortions.

For EO inputs x^{EO} , we extract high-resolution feature maps using EfficientNet-B4 up to block 6a:

$$F_{\text{eff}} = \phi_{\text{eff}}(x^{\text{EO}}), \quad (7)$$

where ϕ_{eff} denotes the EO encoder, and $F_{\text{eff}} \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ with $H_1 = W_1 = 14$ and $C_1 = 1792$.

Simultaneously, SAR inputs are processed by a modified ResNet50, whose stem layer has an increased stride to better handle speckle noise:

$$F_{\text{res}} = \psi_{\text{res}}(x^{\text{SAR}}) \quad (8)$$

where $F_{\text{res}} \in \mathbb{R}^{28 \times 28 \times 512}$ and ψ_{res} represents the initial convolution and pooling layers of ResNet50’s stem:

$$\psi_{\text{res}}(x^{\text{SAR}}) = \text{MaxPool}^{\text{stride}=2} \left(\text{Conv}_{7 \times 7}^{\text{stride}=2}(x^{\text{SAR}}) \right). \quad (9)$$

where ψ_{res} consists of a 7×7 convolution with stride 2 followed by a 3×3 max pooling layer.

To facilitate feature fusion, we first apply bilinear upsampling to F_{eff} for spatial alignment. The upsampled EO features and SAR features are then concatenated along the channel dimension and compressed via a 1×1 convolution:

$$F_{\text{fused}} = \text{Conv}_{1 \times 1}(\text{Concat}(\text{Up}_2(F_{\text{eff}}), F_{\text{res}})), \quad (10)$$

where $F_{\text{fused}} \in \mathbb{R}^{28 \times 28 \times 512}$ and Up_2 represents bilinear upsampling by a factor of 2 to align spatial dimensions before fusion. This dual-stream approach explicitly models EO’s spectral richness and SAR’s geometric invariance while maintaining parameter efficiency.

This dual-branch design allows each encoder to specialize in modality-specific characteristics—leveraging EO’s rich spatial semantics and SAR’s structural invariance—while maintaining a compact and efficient representation for subsequent cross-modal integration.

3.4. Cross-Modal Distribution Alignment

To ensure domain adaptation between EO and SAR modalities, we integrate cross-modal distribution alignment using a semi-supervised approach. The training objective consists of three key components: classification loss, cross-modal alignment loss, and regularization.

For labeled samples in $\mathcal{D}_L = \{(x_i^{\text{EO}}, y_i)\} \cup \{(x_j^{\text{SAR}}, y_j)\}$, the classification loss is computed separately for EO and SAR:

$$\mathcal{L}_{\text{EO}} = \frac{1}{N_{\text{EO}}} \sum_{i=1}^{N_{\text{EO}}} \mathcal{L}_{\text{CE}}(f_{\text{eff}}(x_i^{\text{EO}}), y_i) \quad (11)$$

$$\mathcal{L}_{\text{SAR}} = \frac{1}{N_{\text{SAR}}} \sum_{j=1}^{N_{\text{SAR}}} \mathcal{L}_{\text{CE}}(f_{\text{res}}(x_j^{\text{SAR}}), y_j). \quad (12)$$

where $f_{\text{eff}}(x_i^{\text{EO}})$ represents the classification output from the EfficientNet-B4 model for EO samples. $f_{\text{res}}(x_j^{\text{SAR}})$ represents the classification output from the ResNet50 model for SAR samples. The total supervised loss is computed as:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{EO}} + \mathcal{L}_{\text{SAR}} \quad (13)$$

To minimize feature space discrepancies between EO and SAR representations, we introduce an unsupervised alignment loss $\mathcal{L}_{\text{align}}$ applied exclusively to unlabeled data in \mathcal{D}_U . We employ the Sliced Wasserstein Distance (SWD) [16] to align their feature distributions:

$$\text{SWD}(A, B) = \mathbb{E}_{\theta \sim S^{d-1}} W_2(\mathcal{R}_\theta A, \mathcal{R}_\theta B), \quad (14)$$

where $W_2(\cdot, \cdot)$ denotes the Wasserstein-2 distance [15], and $\mathcal{R}_\theta A$ and $\mathcal{R}_\theta B$ represent the one-dimensional projections of feature distributions A and B along a random direction $\theta \sim S^{d-1}$. To ensure alignment across modalities and fused representations, the total alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = \text{SWD}(F_{\text{eff}}, F_{\text{res}}) + \text{SWD}(F_{\text{fused}}, F_{\text{eff}}) + \text{SWD}(F_{\text{fused}}, F_{\text{res}}). \quad (15)$$

The final training objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \quad (16)$$

where $\lambda_{\text{cls}} = 0.8$ and $\lambda_{\text{align}} = 0.2$ are hyperparameters that control the balance between classification and cross-modal alignment losses.

Table 2. Class Distribution of Training Data

Class #	Class Name	# Train Images
0	sedan	364,291
1	suv	43,401
2	pickup truck	24,158
3	van	16,890
4	box truck	2,896
5	motorcycle	1,441
6	flatbed truck	898
7	bus	612
8	pickup truck with trailer	695
9	semi truck with trailer	353

4. Experiments

The PBVS 2025 MAVOC challenge employs an enhanced version of the NTIRE 2021 multi-view aerial dataset [12], which exhibits significant long-tail distribution characteristics. As shown in Table 2, head classes (e.g., "Sedan") contain over 360k samples, whereas tail classes (e.g., "Semi Truck with Trailer") have only 353 training samples. The dataset is captured using multi-sensor airborne systems, comprising the following sensor modalities:

- **Electro-Optical (EO) sensors:** 31×31 pixel visible-light images.
- **Synthetic Aperture Radar (SAR):** 55×55 pixel microwave images.

Evaluation consists of two primary metrics:

- **Top-1 classification accuracy.**
- **Out-of-distribution (OOD) detection capability**, measured using AUROC and TNR at 95% TPR.

The CSRN framework employs a joint training strategy to train two identical ResNet-50 [6] networks alongside an EfficientNet-B4 [22] model. The input images are resized to 224×224 pixels before being fed into the respective models. The resized EO and SAR images correspond to the same object across two different domains.

A two-stage training approach is adopted. First, a ResNet-50 [6] network is trained on the original imbalanced EO images from the challenge dataset for 15 epochs using Focal Loss [11], which helps mitigate class imbalance. The SAR ResNet-50 is initialized with pretrained weights. In parallel, an EfficientNet-B4 network is trained on EO images to extract high-level spatial features. The CSRN framework is then trained on a manually curated dataset using a weighted random sampler for 100 epochs with a batch size of 64. A learning rate scheduler is employed to enhance convergence, with an initial learning rate of $1e^{-3}$, and the AdamW optimizer is used.

In the testing phase, the final prediction is obtained by computing the weighted average of the ResNet-50-based SAR model, the EfficientNet-B4-based EO model, and an

Table 3. Test results for the PBVS 2025 MAVOC challenge in SAR Classification.

#Place	Team	Top-1% Accuracy	AUROC
1	CSRN	31.78%	0.77
2	Team 3	29.61%	0.76
3	Team 2	27.56%	0.87
4	Team 4	22.61%	0.88

auxiliary EfficientNet-B0-based SAR model. The experiments are conducted with PyTorch on two NVIDIA RTX 3090 GPUs.

4.1. Main results

Table 3 presents the test results for the PBVS 2025 MAVOC challenge on SAR classification, where our Cross-Sensor Robust Recognition Network (CSRN) achieves the highest Top-1% Accuracy of 31.78%, securing first place. This highlights the effectiveness of our cross-sensor learning paradigm, integrating Electro-Optical (EO) and Synthetic Aperture Radar (SAR) data for enhanced feature representation. Compared to the second-ranked team (29.61%), CSRN achieves a 2.17% improvement, demonstrating the advantages of dual-network feature extraction and class-balanced sampling. Leveraging EfficientNet-B4 for EO and ResNet50 for SAR, our model captures EO’s spectral richness and SAR’s structural robustness, while semi-supervised training ensures better generalization. These results validate the benefits of sensor-specific processing and cross-modal feature alignment for aerial object classification.

Despite CSRN ranking first in Top-1% Accuracy, its AUROC score (0.77) is lower than Team 2 (0.87) and Team 4 (0.88), indicating room for improvement in probability calibration and ranking across thresholds. Future work could explore uncertainty-aware learning and refined cross-modal fusion, such as attention mechanisms or adaptive weighting, to enhance both accuracy and AUROC. Nonetheless, CSRN’s top performance underscores the effectiveness of multi-sensor fusion and semi-supervised learning in aerial object classification.

4.2. Ablation studies

Effectiveness of Each Component. As visualized in Table 4, the ablation study demonstrates the critical impact of each component in the Cross-Sensor Robust Recognition Network (CSRN). Starting with the baseline model, where none of the proposed components—Class-Balanced Data Partitioning (CBDP), Dual-Modal Feature Extraction (DMFE), and Cross-Modal Distribution Alignment (CMDA)—are applied, the model achieves a top-1% accuracy of 19.83%. This result highlights the challenge of class imbalance and domain discrepancies without any enhance-

Table 4. Ablation study showing the impact of different components in our Cross-Sensor Robust Recognition Network (CSRN).

CMDA	CBDP	DMFE	Top-1% Accuracy
-	-	-	19.83%
✓	-	-	26.61%
✓	✓	-	28.14%
✓	-	✓	29.79%
✓	✓	✓	31.78%

ments. When CBDP is introduced, the accuracy improves to 26.61%. CBDP addresses class imbalance by using stratified sampling, which ensures proportional representation across all categories, allowing the model to better generalize across both head and tail classes. The improvement observed here demonstrates the importance of balancing the dataset to mitigate biases toward dominant classes.

Further improvement is seen when DMFE is added, which increases the accuracy to 28.14%. DMFE combines EfficientNet for EO’s spatial detail preservation with modified ResNet-50 for SAR’s noise-resilient encoding. It effectively captures complementary features from both domains. This step enables the model to extract meaningful information from both EO and SAR data, which is crucial for aerial view object classification.

The introduction of CMDA brings the accuracy to 29.79%. CMDA uses SWD[16] to minimize feature space discrepancies between EO and SAR domains, ensuring better alignment between the two modalities. Finally, when all components — CBDP, DMFE, and CMDA — are combined, the model achieves its best performance with a top-1% accuracy of 31.78%. This result highlights the synergistic effect of the three components, demonstrating that each uniquely contributes to improving model accuracy.

Effect of Loss Balancing. We conduct an ablation study to investigate how different weightings between classification loss (λ_{cls}) and cross-modal alignment loss (λ_{align}) affect model performance, while ensuring $\lambda_{\text{cls}} + \lambda_{\text{align}} = 1$. As shown in Table 5, the best Top-1 accuracy of **31.8%** is achieved when $\lambda_{\text{cls}} = 0.8$ and $\lambda_{\text{align}} = 0.2$, suggesting that a moderate alignment objective effectively reduces domain shift without hindering label supervision. In contrast, relying solely on classification loss ($\lambda_{\text{align}} = 0$) or alignment loss ($\lambda_{\text{cls}} = 0$) results in notable performance drops, demonstrating the necessity of joint optimization for robust cross-modal representation learning.

5. Conclusion

In this paper, we present the Cross-Sensor Robust Recognition Network, Class-Balanced Data Partitioning, Dual-Modal Feature Extraction, and Cross-Modal Distribution Alignment. Our approach leverages both labeled and un-

Table 5. Ablation study on loss weighting. Best performance occurs at $\lambda_{\text{cls}} = 0.8$ and $\lambda_{\text{align}} = 0.2$.

λ_{cls}	λ_{align}	Top-1 Accuracy (%)
1.0	0.0	29.6
0.9	0.1	30.8
0.8	0.2	31.8
0.6	0.4	31.2
0.4	0.6	29.9
0.2	0.8	27.4
0.0	1.0	24.9

labeled data using a semi-supervised learning framework, enhancing the model’s generalization and accuracy across complex multi-modal domains. Through extensive ablation studies, we demonstrate that each component of the CSRN contributes uniquely to improving classification performance. Our method achieves significant improvements in top-1 accuracy, outperforming baseline models and placing us among the top-ranked solutions in the PBVS 2025 MAVOC Challenge. These results empirically validate the effectiveness of our multi-source domain fusion approach in aerial view object classification, reinforcing its potential for real-world applications.

6. Acknowledgements

This work was supported in part by the National Key Research and Development Project under Grant 2023YFC3806000, in part by the National Natural Science Foundation of China under Grant 62406226 and 61936014, in part sponsored by Shanghai Sailing Program under Grant 24YF2748700, in part sponsored by Tongji University Independent Original Cultivation Project under Grant 22120240326, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, in part by the Shanghai Science and Technology Innovation Action Plan under Grant 22511105300, and in part by the Fundamental Research Funds for the Central Universities under Grant 2022-5-YB-01.

References

- [1] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research (JAIR)*, 16:321–357, 2002. 2
- [2] Chen Chen, Mei Liu, and Boxiao Zhang. Small object detection in aerial images. *Remote Sensing*, 12(5):827, 2020. 1
- [3] Siwei Chen and Haipeng Wang. Deep learning for sar target classification: A data-dominated paradigm. *IEEE Access*, 7: 178366–178378, 2019. 2
- [4] Shane R. Cloude and Eric Pottier. Target decomposition the-

- orems in radar scattering. *Electronics Letters*, 32(13):1178–1179, 1996. 2
- [5] Yaroslav Ganin and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17(59):1–35, 2016. 2
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 2, 4, 5
- [7] Eric Heitz and Baptiste Guillard. Sliced wasserstein loss for neural texture synthesis. *ACM Transactions on Graphics*, 40(4):1–13, 2021. 2
- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. 2
- [9] Jong-Sen Lee. Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(2):165–168, 1980. 4
- [10] Jong-Sen Lee, Mitchell R Grunes, and Ronald Kwok. Classification of multi-look polarimetric sar data based on complex wishart distribution. *International Journal of Remote Sensing*, 15(11):2299–2311, 1994. 2
- [11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2, 5
- [12] Jerrick Liu, Nathan Inkawhich, Oliver Nina, and Radu Timofte. Ntire 2021 multi-modal aerial view object classification challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 588–595, 2021. 5
- [13] Wei Liu, Jie Yang, and Xiaolong Li. Atr in remote sensing: Applications and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):24–45, 2021. 1
- [14] Yichen Luo, Jinxing Liu, Gong Chen, Yuting Wang, and Wei Li. Cross-modal attention for eo and sar remote sensing image fusion in object detection. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3705–3708, 2022. 2
- [15] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. 4
- [16] Julien Rabin and Gabriel Peyré. Wasserstein regularization for imaging. *SIAM Journal on Imaging Sciences*, 4(4):1236–1263, 2011. 2, 4, 6
- [17] Mohammad Rostami, Soheil Kolouri, and Kyungnam Kim. Semi-supervised domain adaptation for sar-atr. *IEEE Transactions on Aerospace and Electronic Systems*, 57(4):2345–2358, 2021. 2
- [18] Mohammad Rostami, Soheil Kolouri, and Kyungnam Kim. Semi-supervised domain adaptation for SAR-ATR. *IEEE Transactions on Aerospace and Electronic Systems*, 57(4):2345–2358, 2021. 1, 2
- [19] Michael Schmitt and Xiaoxiang Zhu. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016. 4
- [20] Burr Settles. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences Technical Report*, 1648, 2009. 2
- [21] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. 4
- [22] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019. 4, 5
- [23] Qiang Wang, Lei Zhang, and Xiaodong Yang. Transformer meets sar: A novel framework for polarimetric sar image classification. *IEEE TGRS*, 61:1–13, 2023. 2
- [24] Lei Yang, Qiang Wang, and Hao Zhang. Bidirectional knowledge transfer for eo-sar fusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10231–10240, 2022. 1
- [25] Yuxiang Zhang and Hao Li. Few-shot learning for sar target recognition with metric embedding. *IEEE TGRS*, 60:1–12, 2021. 2
- [26] Yuxiang Zhang, Hao Li, and Lei Wang. Hybrid reweighting for class-imbalanced remote sensing. *IEEE TGRS*, 60:1–14, 2022. 2
- [27] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023. 2
- [28] Liangpei Zhao, Lei Zhang, Pengxiang Li, Bo Huang, and Yuming Xu. Fusion of sar and optical images for urban object detection using deep learning. *Remote Sensing*, 11(1):30, 2019. 2
- [29] Bohan Zhou, Quan Chen, Jiaying Wang, Chen Change Loy, Bo Dai, and Dahua Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9719–9728, 2020. 2