UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

D-Cog: Distilling Human Cognitive Representations into Visual Prompts for Few-Shot SAR Recognition

Permalink

https://escholarship.org/uc/item/6bs4x7t9

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Shang, Jingjie
Gao, Yuan
He, Zhongjiang
et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

D-Cog: Distilling Human Cognitive Representations into Visual Prompts for Few-Shot SAR Recognition

Jingjie Shang[1], Yuan Gao[2], Zhongjiang He[2], Ying Li[3], Boran Zhao[2], Liming Fan[3], Zi-Gang Huang[3], Pengju Ren[†4], Yuan Wang[†1] & Xiaochen Bo[†5]

[1]School of Integrated Circuits, Peking University
[2]School of Software Engineering, Xi'an Jiaotong University
[3]School of Life Science and Technology, Xi'an Jiaotong University
[4]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[5]Academy of Military Medical Sciences, Beijing

Abstract

Synthetic Aperture Radar (SAR) recognition is frequently impeded by severe speckle noise and data scarcity. In contrast, human vision excels in these challenging conditions due to robust top-down cognitive modulation. This work investigates whether human cognitive representations can be transferred to SAR recognition models. We introduce the first paired fMRI–SAR dataset designed to capture the cognitive neural responses of human viewing SAR images. Building upon this dataset, we propose a cognitive fusion–distillation framework. This framework integrates fMRI-derived cognitive representations into a teacher network and distills the cognitive capability into a lightweight student model via prompt tuning. Extensive experiments demonstrate consistent performance improvements in few-shot, cross-domain, and out-of-distribution conditions. Furthermore, Representational similarity analysis reveals that cognition-guided models learn representation patterns distinct from purely data-driven models. These results suggest that neuro-cognitive signals can serve as transferable inductive biases to enhance robustness and generalization in specialized vision tasks.

Keywords: Cognitive Representations; fMRI; SAR Recognition.

Introduction

Synthetic Aperture Radar (SAR) is indispensable for all-weather earth observation, playing a pivotal role in military reconnaissance and disaster monitoring (Lv et al., 2024). Unlike natural optical images, SAR data is characterized by non-intuitive scattering geometries, severe speckle noise, and ambiguous visual semantics. These factors make reliable recognition exceptionally challenging. Although Deep Neural Networks dominate standard computer vision (Liu et al., 2022), their effectiveness in the SAR domain remains limited, especially under few-shot constraints. When data is scarce or noise is pervasive, these models lack the semantic stability required to disentangle targets from background clutter. This limitation stems primarily from an over-dependence on statistical patterns. Consequently, given the high risks of recognition errors, human analysts remain essential in the interpretation loop.

Experienced human observers can reliably recognize SAR targets across diverse imaging conditions, exhibiting superior robustness and generalization even with limited data. Cognitive science studies attribute this proficiency not only to low-level visual details, but also to high-level cognitive representations guided by top-down modulation (Fernandes et al., 2021). When faced with degraded stimuli, the human visual system suppresses noise to extract semantically invariant features, thereby maintaining object constancy. As illustrated in Fig. 2, while speckle noise misleads conventional models, the human brain leverages higher cortical regions to successfully reconstruct target identity. This contrast motivates our core research question: Can this human visual cognition of SAR signals be effectively transferred to computer vision models?

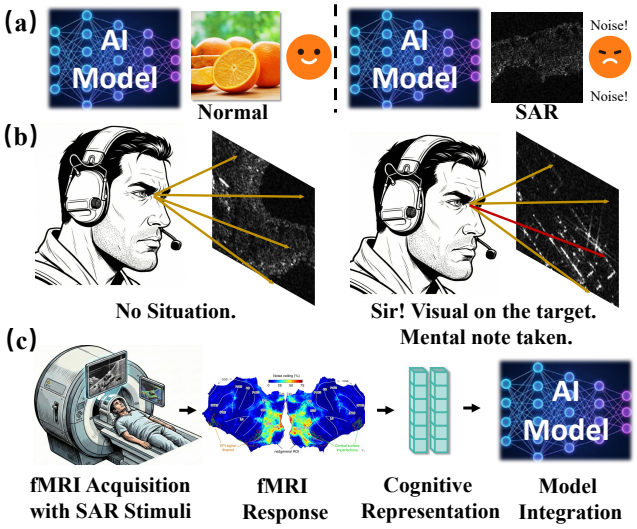

Figure 1: Motivation of this work.

Functional magnetic resonance imaging (fMRI) provides a high-resolution, whole-brain perspective on human visual cognition (Allen et al., 2022). Compared with non-invasive electrophysiological (EEG) recordings, fMRI is particularly suitable for capturing stable and high-level cognitive representation structures underlying visual understanding. Therefore, integrating artificial visual models with these cognitive representations to constrain feature learning presents a promising strategy for enhancing SAR recognition.

[†]Corresponding authors.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

In this paper, we explore a cognition-driven paradigm for SAR recognition by explicitly leveraging human visual representations measured via fMRI. We construct the first paired fMRI-SAR dataset to capture neural responses elicited by diverse SAR scene categories, providing a high-fidelity cognitive carrier. Building on this foundation, we design a cognitive deep fusion framework using the proposed Universal Cognitive Fusion module, that injects fMRI-derived cognitive representations into a hierarchical vision backbone. To overcome the dependence on fMRI signals during deployment, we further propose a cognitive prompt distillation strategy that transfers cognition-induced capabilities into a prompt-based student model operating solely on SAR images. Extensive experiments demonstrate that the proposed framework consistently outperforms state-of-the-art models, particularly under few-shot and cross-domain conditions. Additionally, representational analyses validate the effectiveness of cognition-guided learning for SAR recognition. Our main contributions are summarized as follows:

- We present **Cog-SAR**, a novel dataset comprising SAR images and synchronously recorded fMRI signals. To the best of our knowledge, this is the first dataset designed to align human neural activities with SAR visual interpretation.
- We propose a cognitive fusion and prompt-based distillation framework that integrates fMRI-derived cognitive representations into SAR recognition models and transfers human cognitive priors into a student network, enabling efficient and fMRI-free inference.
- Extensive experiments demonstrate significant gains in few-shot recognition, cross-dataset generalization, and out-of-distribution robustness, while representational similarity analysis reveals that cognition-guided models learn feature structures that differ significantly from data-driven models.

Cog-SAR Datasets

Preparation and Behavioristic Factor

Prior to data acquisition, participants completed a personal information and screening questionnaire to assess their eligibility. The questionnaire covered handedness, recent sleep quality, and current emotional state. Given that the participants lacked prior experience with SAR imagery, a ten-minute practice session was conducted preceding the formal scan. This session utilized a set of images distinct from the experimental stimuli to familiarize subjects with the visual characteristics of the different categories. During this phase, participants were encouraged to identify discriminative features and formulate their own cognitive strategies for classification.

Stimulus Paradigm

The visual stimuli consisted of 700 SAR images with a spatial resolution of 256×256 pixels, covering a diverse set of representative scene categories. All images were displayed centrally on a white background with a superimposed red fixation cross to reduce eye movements. We employed an event-related experimental design. Each trial lasted 4 seconds (2 TRs), beginning with a 1-second fixation period to facilitate cognitive reset, followed by a 3 seconds presentation of the SAR image. To improve the estimation of the hemodynamic response function (HRF), null trials were randomly inserted as temporal jitter, with each null trial including a 1 second fixation cross and a 3 seconds blank interval.

The recording process comprised two sessions, each session containing two runs. Each run comprised 175 image trials (700 seconds) and 35 null trials (140 seconds). These null trials, constituting approximately 20% of the total trials, were distributed pseudo-randomly throughout the run. Additionally, three null trials were fixed at both the beginning and the end of each run to serve as baseline periods. To maintain consistency across participants, the specific pseudo-random stimulus sequence was identical across all participants.

MRI Data Acquisition

The dataset included high-resolution T1-weighted structural scans, resting-state BOLD functional MRI, and task-based BOLD functional MRI. Participants were positioned supine, and their heads were stabilized with padding within the head coil to minimize motion. fMRI data was acquired using a gradient-echo echo-planar imaging (EPI) sequence. Identical acquisition parameters were applied to both resting-state and task-based scans: TR = 2000 ms, TE = 21 ms, and FA = 66°. Spatial parameters included an FOV of 220 mm × 220 mm, a 94 × 94 acquisition matrix, 33 slices with a thickness of 4 mm, and no inter-slice gap. The resting-state scans lasted for 150 TRs (5 minutes), while task-based fMRI lasted for 1700 TRs (60 minutes). To alleviate fatigue associated with the long scanning duration, data collection was split into two sessions separated by a 10-15 minute break.

MRI Data Preprocessing

Following data acquisition, all neuroimaging data underwent visual quality inspection to assess obvious artifacts, consistency between functional and structural images, head motion (thresholds of >3 mm translation or >3° rotation), temporal alignment, and consistency in slice numbers and acquisition parameters across participants. Based on these criteria, eight healthy participants were retained in the final dataset. We performed preprocessing utilizing fMRIPrep (Esteban et al., 2019) to mitigate noise and artifacts. The pipeline included intensity non-uniformity, head motion correction, and skull stripping for both structural images and the mean functional image. Spatial normalization was performed using nonlinear registration to the ICBM 152 Nonlinear Asymmetrical template version 2009c (MNI152NLin2009cAsym), with voxel resampling to $2.973 \times 2.973 \times 3.885\text{mm}^3$. Subsequently, fMRIPrep calculated confounding time-series to be used for linear detrending and covariate regression. Resting-state fMRI data were corrected using the CompCor (Behzadi et al., 2007) method to retain low-frequency neural fluctuations, whereas task-based fMRI data were not filtered.

Methods

In this section, we introduce our cognitive framework, which consists of two stages: (1) The construction of a Cognitive Deep Fusion Network (teacher network) that integrates exogenous cognitive signals into the SAR image recognition process, and (2) a Cognitive-Prompt Distillation framework that transfers this cognitive capability to a student network with prompt tuning.

Cognitive Deep Fusion Network

We utilize fMRI as a cognitive signal to provide high-level priors for SAR image recognition. Given a SAR image I, the HiViT backbone (Zhang et al., 2023) initially generates visual embeddings via a patch embedding layer. Concurrently, the paired fMRI signal is encoded by a pretrained fMRI encoder (D. Li et al., 2025) into a compact cognitive representation $\mathbf{F}_{cog} \in \mathbb{R}^{N\times C}$. The core component of our method is the Universal Cognitive Fusion (UCF) module, designed to seamlessly inject cognitive representations into visual streams. Let $\mathbf{X}_{vis} \in \mathbb{R}^{N\times C}$ denote the visual representations derived from the SAR image. The module first aligns the cognitive representation dimension with the visual embedding dimension via a linear projection and normalization:

$$\mathbf{F}_0 = \text{LayerNorm}(\mathbf{W}_p\mathbf{F}_{cog} + \mathbf{b}_p) \tag{1}$$

To capture internal correlations within the cognitive signal before fusion, we employ a self-refinement step using Multi-Head Self-Attention (MHSA). Subsequently, the refined cognitive features $\mathbf{F}_{ref}$ serve as the Key (K) and Value (V), while the normalized visual tokens function as the Query (Q) in a Multi-Head Cross-Attention (MHCA) mechanism. This configuration allows the visual features to dynamically attend to relevant cognitive contexts:

$$\mathbf{Z}_{attn} = \text{MHCA}(Q = \text{LayerNorm}(\mathbf{X}_{vis}), K = \mathbf{F}_{ref}, V = \mathbf{F}_{ref}) \tag{2}$$

To ensure training stability and preserve original visual information, we incorporate a learnable gating mechanism with a zero-initialized parameter γ, forming a gated residual connection:

$$\mathbf{X}_{out} = \mathbf{X}_{vis} + \tanh(\gamma) \odot \mathbf{Z}_{attn} \tag{3}$$

where $\odot$ denotes the Hadamard product. This design enables the network to adaptively regulate the magnitude of cognitive injection.

Instead of a single fusion point, we adopt a cognitive deep fusion strategy that matches the architecture in HiViT. Specifically, the UCF module is injected (i) at the input level after patch embedding, and (ii) densely after each hierarchical transformer block (BlockWithRPE) throughout both the early (low-level) and main (high-level) stages. Patch merging layers (PatchMerge) do not perform fusion and are implemented as identity placeholders to keep one-to-one alignment between visual blocks and fusion modules. This dense injection ensures that cognitive information modulates both local structure cues and high-level semantics, yielding a teacher network that is strongly cognition-aware.

Cognitive-Prompt Distillation Framework

Although the teacher network benefits from cognition-guided deep fusion, fMRI acquisition is costly and unavailable during deployment. Consequently, we propose a cognitive prompt distillation framework that trains a student model using only SAR images during inference while retaining the cognition-induced capability learned by the teacher.

The student network maintains the same HiViT-based recognition pipeline but excludes the fMRI input branch. To compensate for the missing cognitive tokens, we introduce prompt tuning with a small set of trainable parameters (learnable prompts) that are optimized during distillation. These prompts act as lightweight, task-adaptive cognitive surrogates and are integrated into the student's feature learning.

During training, the student network is optimized using a combination of label supervision, logit-level knowledge distillation, and cognition-aware feature distillation. Specifically, we apply a standard cross-entropy loss $\text{CE}(,)$ on the student predictions with respect to the ground-truth labels:

$$\mathcal{L}_{\text{CE}} = \text{CE}\big(z_s, y\big), \tag{4}$$

where z_s denotes the logits produced by the student network and y is the ground-truth class label.

To align the predictive distributions of the student and the teacher, we adopt temperature-scaled knowledge distillation based on the Kullback-Leibler divergence $\text{KL}(,)$. For the implementation, the distillation temperature is set to $T = 4$, and the loss is scaled by T^2 to preserve the gradient magnitude:

$$\mathcal{L}_{\text{KL}} = T^2\,\text{KL}\Big(\sigma\big(z_s/T\big) \,\Big\|\, \sigma\big(z_t/T\big)\Big), \tag{5}$$

where z_t denotes the teacher logits and $\sigma(\cdot)$ is the softmax function.

Beyond logit-level alignment, we introduce a cognition-aware feature distillation loss to transfer the representation capability induced by cognitive fusion. Dense multi-layer feature distillation is applied to intermediate features extracted from corresponding layers. To stabilize training across layers with different activation scales, each layer-wise loss is normalized by the squared standard deviation of the corresponding teacher feature:

$$\mathcal{L}_{\text{feat}}^{\text{total}} = \sum_{l} \frac{\big\|f_s^{(l)} - f_t^{(l)}\big\|_2^2}{\Big(\text{Std}\Big(f_t^{(l)}\Big) + \epsilon\Big)^2}, \qquad \epsilon = 10^{-6}. \tag{6}$$

where $l \in \{1, 3, 7, 13, 20, 27\}$. We define $l \in \{1\}$, $l \in \{3, 7\}$, and $l \in \{13, 20, 27\}$ as the Input, Middle, and End stages, respectively. Finally, the total training objective for the student network is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + 0.1\,\mathcal{L}_{\text{KL}} + 0.1\,\mathcal{L}_{\text{feat}}^{\text{total}}. \tag{7}$$

Experiment

In this section, we evaluate the performance and cognitive interpretability of our framework through a series of quantitative experiments. We begin by comparing few-shot classification

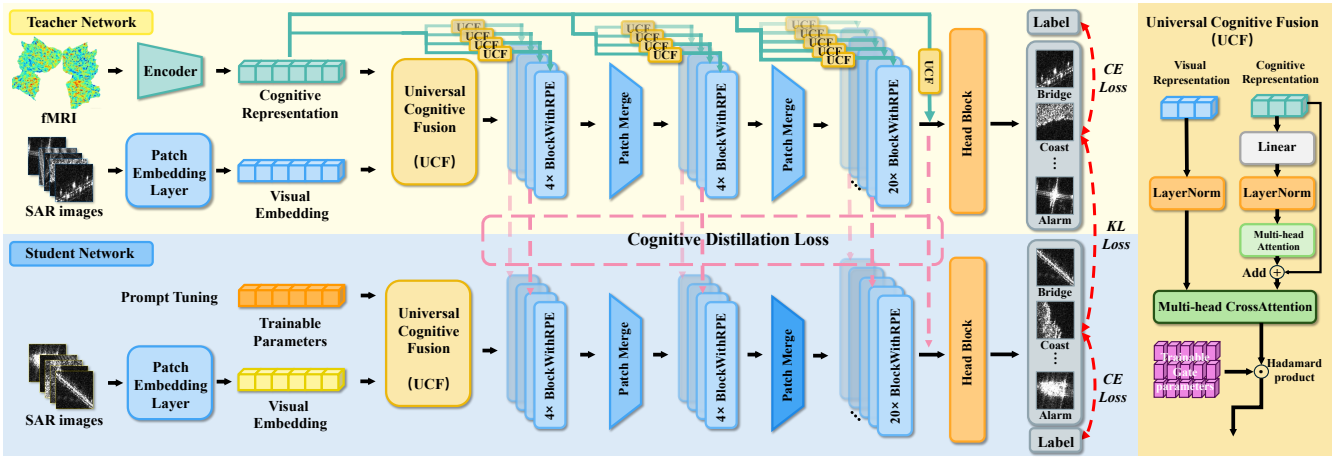

Figure 2: Overview of the proposed Cognitive-Prompt Distillation framework.

Table 1: Comparative evaluation of few-shot classification accuracy on the Cog-SAR dataset across varying support set sizes (3, 5, 10, and 20 shots). Results are reported as mean ± standard deviation.

Method Type	Method Name	Venue	3-shot	5-shot	10-shot	20-shot
Math-driven	ProtoNet	2017 NeurIPS	73.14 ± 2.83	76.99 ± 4.84	77.30 ± 5.26	78.21 ± 2.98
	TIM	2020 NeurIPS	50.42 ± 5.98	51.89 ± 4.25	52.79 ± 5.57	75.41 ± 2.67
	FRN	2021 CVPR	76.77 ± 5.72	77.04 ± 3.43	80.85 ± 5.31	85.75 ± 4.82
Task-specific	AConvNet	2016 TGRS	58.07 ± 4.49	61.36 ± 4.36	65.02 ± 3.09	78.46 ± 3.80
	ATL-Net	2020 IJCAI	58.49 ± 4.17	65.56 ± 3.57	68.75 ± 3.29	77.35 ± 3.06
	PGIL	2022 ISPRS	54.74 ± 2.64	57.57 ± 4.82	58.63 ± 5.62	69.32 ± 1.85
	BIDFC	2022 TGRS	55.38 ± 6.21	60.17 ± 3.73	66.96 ± 3.35	71.88 ± 2.18
Foundation	DINO-v2	2023 TMLR	67.81 ± 5.42	77.78 ± 2.86	85.67 ± 1.44	90.22 ± 1.28
	SARATR-X	2025 TIP	83.78 ± 5.31	88.80 ±3.85	95.05 ± 1.67	97.64 ± 0.85
Ours			**84.82 ± 2.28**	**93.34 ± 2.21**	**97.01 ± 0.99**	**98.24 ± 0.54**

performance with mainstream methods, followed by distillation and generalization analyses to assess knowledge transfer efficiency. Notably, Representational Similarity Analysis (RSA) is employed to compare the divergence between human neural representations and machine computational representations across our model and baselines. Finally, cross-subject evaluations and ablation studies quantify the robustness and individual contributions of each functional component.

Table 2: Cross-dataset evaluation results on FUSAR after training on the full CogSAR dataset.

Method	Venue	Acc(%)	F1(%)
AconvNet	2016 TGRS	73.2	72.1
PGIL	2022 ISPRS	80.51	78.13
SARATR-X	2025 TIP	88.8	86.2
Ours		**90.0**	**87.6**

Experimental Setting

All experiments were implemented using the PyTorch framework on a NVIDIA RTX 6000 Ada GPU. The models were optimized using the AdamW optimizer ($LR = 1 \times 10^{-3}$, $WD = 1 \times 10^{-4}$) for 30 epochs with a cosine decay schedule. The batch size was set to 14 for training and 8 for inference. The comparison methods include math-driven methods (ProtoNet (Snell et al., 2017),TIM (Boudiaf et al., 2020), FRN (Wertheimer et al., 2021)), task-specific networks (AConvNet (Chen et al., 2016),ATL-Net (Dong et al., 2020),PGIL (Huang et al., 2022),BIDFC (Zhai et al., 2022)), and foundation models (DINO-v2 (Oquab et al., 2024), SARATR-X (W. Li et al., 2025)). we employ accuracy (ACC) and F1-Score (F1) as our metrics.

Evaluation of Few-Shot Classification Performance

We conducted experiments on the Cog-SAR dataset, which includes seven categories. Following standard few-shot evaluation protocols, we assessed model performance under K-

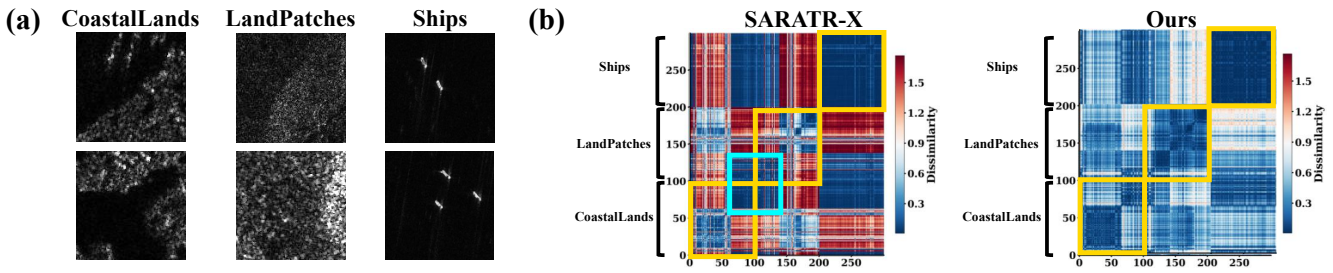

Figure 3: Representational Similarity Analysis. (a) Representative SAR images from CoastalLands, LandPatches, and Ships categories. (b) Comparison of Representational Dissimilarity Matrices (RDMs).

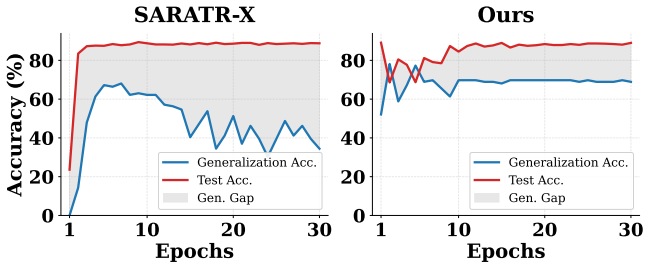

Figure 4: Generalization dynamics analysis of in-domain test accuracy and cross-domain generalization accuracy across epochs.

shot settings with $K \in \{3, 5, 10, 20\}$. The training set was constructed by randomly sampling K instances from each category (i.e., $7 \times K$ samples in total), and all remaining samples were used as the test set. To account for randomness, all results are represent the mean Top-1 accuracy and standard deviation over 10 independent runs with different random seeds.

As shown in Table 1, metric-based methods leverage distance metrics for generalization but struggle to capture the complex scattering properties inherent to SAR data. Conversely, while task-specific models demonstrate stronger performance under data-abundant conditions, they generalize poorly in few-shot regimes. Foundation models represent the current state-of-the-art by benefiting from extensive pretraining. However, our proposed framework significantly outperforms all competing methods across all shot settings, validating the efficacy of integrating cognitive representations into the learning loop. In the critical 5-shot scenario, which serves as a practical benchmark for SAR applications, our method achieves an accuracy of 93.34%, surpassing the strongest foundation model competitor (SARATR-X, 88.80%) by a substantial margin of 4.54%. This sustained superiority suggests that cognitive-guided features provide a more robust inductive bias, enabling rapid adaptation to SAR tasks even when training samples are scarce.

Evaluation of Cognitive Distillation and Generalization

To verify whether the student model successfully retains cognitive intelligence after distillation, we established a data-efficient learning protocol. We trained the model on the limited Cog-SAR dataset (700 samples) while evaluating it on the comprehensive FUSAR benchmark (2940 samples) (Hou et al., 2020). This configuration introduces a severe imbalance between training and testing data. As presented in Table 2, our method achieves an accuracy of 90.0%, surpassing both task-specific architectures and the foundation model SARATR-X (88.8%). This result confirms that the distilled student effectively inherits the representational power of the teacher even under data-constrained conditions.

Furthermore, to probe true out-of-distribution (OOD) generalization, we introduced an external validation set comprising 120 images collected from a distinct satellite (Xiang et al., 2025) source with different geographical characteristics. To prevent chance-level guessing, this dataset contains only four categories that intersect with the seven-class training label space. Fig. 4 illustrates the generalization dynamics. Data-driven foundation models gradually stabilize on the in-domain test set as training proceeds, yet their performance on the cross-domain generalization set deteriorates markedly. In contrast, our method maintains a substantially higher level of generalization accuracy without compromising in-domain performance, outperforming the baseline by approximately 20-30% in the OOD setting. We note that the initial training phase of the student model exhibits mild performance fluctuations due to partial initialization from the teacher network, which stabilizes as distillation progresses. Ultimately, the minimal generalization gap (grey area) demonstrates that cognitive distillation preserves the intrinsic robustness of human visual perception, preventing the model from collapsing into dataset-specific correlations.

Representational Similarity Analysis

We performed representational similarity analysis (RSA) to investigate the divergence in learned representations between cognition-guided models and purely data-driven models. Our analysis focused on model behavior under two contrasting conditions: categories with highly similar visual distributions and

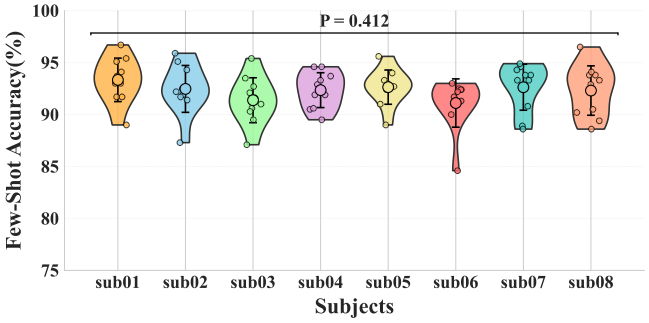

Figure 5: Evaluation of the cross-subject robustness.

Table 3: Ablation studies on distillation contributions (upper panel) and fusion positions (lower panel)

Input-stage	Middle-stage	End-stage	KL Loss	Acc(%)	F1(%)
✗	✓	✓	✓	89.7	87.1
✓	✗	✓	✓	88.0	85.2
✓	✓	✗	✓	88.6	86.0
✓	✓	✓	✗	89.7	87.3
Ours				**90.0**	**87.6**

Position	5-shot		10-shot	
	Acc(%)	F1(%)	Acc(%)	F1(%)
w.o. (i) Input level	89.28 ± 4.35	89.28 ± 4.30	95.70 ± 3.35	95.76 ± 3.21
w.o. (ii) low-level and high-level	89.94 ± 3.78	90.06 ± 3.46	95.6 ± 0.66	95.62 ± 0.62
w.o. (i) and (ii)	89.04 ± 3.54	89.16 ± 3.22	94.6 ± 0.54	94.58 ± 0.58
Ours	**93.34 ± 2.21**	**93.05 ± 2.35**	**97.01 ± 0.99**	**96.94 ± 0.92**

those that are visually dissimilar. Accordingly, three representative categories (CoastalLands, LandPatches, and Ships) from the Cog-SAR dataset were selected for RSA. As shown in Fig. 3(a), CoastalLands and LandPatches share highly similar visual characteristics, while Ships are visually distinct from both. Despite these visual ambiguities, humans can rely on semantic understanding and accumulated experience to abstract the intrinsic relationships and distinctions among the three categories. As shown in Fig. 3 (b), our model exhibits a well-defined diagonal structure, particularly in the Ship category (indices 200-300) and within the CoastalLands category (indices 0-100). In contrast, the SARATR-X RDM displays a significant structural anomaly in the 50-150 index range, where a dark blue intersection indicates a misidentification of coastal and inland samples as an overlapping cluster. These results demonstrate that cognition-guided method, by leveraging fMRI signals, captures key aspects of human cognitive representations, thereby enabling models to learn representational structures that are not solely constrained by the underlying data distribution.

Cross-Subject Analysis

We evaluate the cross-subject robustness of our proposed framework as shown in Fig. 5. The statistical analysis yields a p-value of 0.412, indicating that the model performance is statistically isotropic across different subjects without significant divergence. Despite this general consistency, nuanced fluctuations in individual contributions are observable. Specifically, accuracy distributionsexhibit slight variations in variance and median values (e.g., comparing sub01 and sub06). However, even with these discrepancies, our method maintains robust overall performance with high mean accuracy across the cohort, confirming its resilience to inter-subject variability in few-shot classification tasks.

Ablation Study

The ablation results are presented in Table 3. The upper panel details the impact of removing specific distillation loss components. Notably, the removal of the Input-stage results in the most significant performance decline, with accuracy dropping to a minimum of 88.0%. The lower panel highlights the importance of the Universal Cognitive Fusion (UCF) module's injected position. The results demonstrate that removing the module from optimal positions leads to noticeable performance degradation, with the 5-shot average accuracy consistently falling below the 90% benchmark achieved by our proposed framework.

Discussion and Conclusion

Our experimental results demonstrate that fMRI signals serve as an effective carrier of high-level visual cognition, bridging the semantic gap even in specialized tasks like SAR interpretation. Significant gains in few-shot settings confirm the value of these informative cognitive features. Crucially, RSA reveals that our model learns feature structures distinct from data-driven baselines, aligning closely with human visual abstraction. This validates the premise that human cognitive signals possess the representational power that extends beyond natural images to specialized domains.

The proposed framework demonstrates that cognitive distillation effectively captures human-like generalization capabilities. By significantly outperforming data-driven baselines in cross-domain tests, our results highlight the limitations of traditional optimization in acquiring robust, brain-like inductive biases. Crucially, this strategy offers a practical pathway to "compile" costly cognitive priors into models without requiring fMRI during deployment. Consequently, the student model retains superior few-shot adaptability and generalization efficiency.

Our analysis of cross-subject experiments highlights a critical insight: while the extracted cognitive structures exhibit inter-subject variability, they universally provide complementary information to the computational model. At the same time, the shared improvements across subjects indicate the presence of common representational structure that can be harnessed for subject-agnostic cognitive alignment. The future work may explicitly model them as a resource for personalization, robust aggregation, or learning population-level cognitive priors.

Conclusion. In summary, this work contributes a paired fMRI-SAR dataset and a cognitive fusion-distillation framework that improves SAR recognition, particularly under few-

shot and cross-domain conditions. Beyond performance, our analyses provide empirical support for a claim: neurocognitive measurements can function as a transferable source of representational constraints, shaping models toward more human-like robustness and generalization. This perspective suggests a promising direction for cognitive computational modeling in applied vision, where brain-derived representations not only to explain human behavior, but also to inform and regularize machine learning systems in domains where data are scarce and distribution shifts are unavoidable.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFB4405501); the Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project (Grant 2021ZD0201300).

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (compcor) for bold and perfusion based fmri. *NeuroImage*, *37*(1), 90–101.

Boudiaf, M., Ziko, I., Rony, J., Dolz, J., Piantanida, P., & Ben Ayed, I. (2020). Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, *33*, 2445–2457.

Chen, S., Wang, H., Xu, F., & Jin, Y.-Q. (2016). Target classification using the deep convolutional networks for sar images. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(8), 4806–4817.

Dong, C., Li, W., Huo, J., Gu, Z., & Gao, Y. (2020). Learning task-aware local representations for few-shot learning. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 716–722.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). Fmriprep: A robust preprocessing pipeline for functional mri. *Nature Methods*, *16*(1), 111–116.

Fernandes, E. G., Phillips, L. H., Slessor, G., & Tatler, B. W. (2021). The interplay between gaze and consistency in scene viewing: Evidence from visual search by young and older adults. *Attention, Perception, & Psychophysics*, *83*(5), 1954–1970.

Hou, X., Ao, W., Song, Q., Lai, J., Wang, H., & Xu, F. (2020). Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. *Science China Information Sciences*, *63*(4), 140303.

Huang, Z., Yao, X., Liu, Y., Dumitru, C. O., Datcu, M., & Han, J. (2022). Physically explainable CNN for SAR image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 25–37.

Li, D., Qin, H., Wu, M., Wei, C., & Liu, Q. (2025). Brainflora: Uncovering brain concept representation via multimodal neural embeddings. *Proceedings of the 33rd ACM International Conference on Multimedia*, 5577–5586.

Li, W., Yang, W., Hou, Y., Liu, L., Liu, Y., & Li, X. (2025). Saratr-x: Toward building a foundation model for sar target recognition. *IEEE Transactions on Image Processing*, *34*, 869–884.

Liu, J., Shang, J., Liu, R., & Fan, X. (2022). Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(8), 5026–5040.

Lv, J., Zhu, D., Geng, Z., Han, S., Wang, Y., Ye, Z., Zhou, T., Chen, H., & Huang, J. (2024). Recognition for sar deformation military target from a new minisar dataset using multi-view joint transformer approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, *210*, 180–197.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., . . . Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, *30*.

Wertheimer, D., Tang, L., & Hariharan, B. (2021). Few-shot classification with feature map reconstruction networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8012–8021.

Xiang, Y., Chen, J., & Hong, e. a., Zhonghua. (2025). Osdataset2.0: Sar-optical image matching dataset and evaluation benchmark. *Journal of Radars*, *14*, 1–13.

Zhai, Y., Zhou, W., Sun, B., Li, J., Ke, Q., Ying, Z., Gan, J., Mai, C., Labati, R. D., Piuri, V., & Scotti, F. (2022). Weakly contrastive learning via batch instance discrimination and feature clustering for small sample sar atr. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–17.

Zhang, X., Tian, Y., Xie, L., Huang, W., Dai, Q., Ye, Q., & Tian, Q. (2023). Hivit: A simpler and more efficient design of hierarchical vision transformer. *International Conference on Learning Representations*.