frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 17 September 2025
DOI 10.3389/frsen.2025.1662024

OPEN ACCESS

EDITED BY
Rui Li,
University of Warwick, United Kingdom

REVIEWED BY
Fengxiang Wang,
National University of Defense Technology, China
Yi Liu,
Chongqing University, China
Di Wang,
Wuhan University, China

*CORRESPONDENCE
Yiping Song,
yiping928@163.com
Xun Huang,
hobbery@163.com

RECEIVED 08 July 2025
ACCEPTED 11 August 2025
PUBLISHED 17 September 2025

CITATION
Zhan H, Song Y, Huang X, Tan X and Zhang T (2025) CARP: cloud-adaptive robust prompting of vision-language models for ship classification under cloud occlusion.
Front. Remote Sens. 6:1662024.
doi: 10.3389/frsen.2025.1662024

CARP: cloud-adaptive robust prompting of vision-language models for ship classification under cloud occlusion

Haoke Zhan, Yiping Song*, Xun Huang*, Xiao Tan and Ting Zhang

Naval University of Engineering, Wuhan, China

Fine-grained few-shot ship classification under cloud occlusion is vital for maritime safety but remains challenging due to corrupted features and limited data utility. While the advent of large pre-trained vision-language models (VLMs) provides promising solutions, the lack of specialized benchmarks hinders their effective application. To address this, we introduce SeaCloud-Ship, the first benchmark dedicated to this task. It comprises 7,654 high-resolution, high-quality annotated images across 30 classes, featuring quantified cloud coverage (12.5%–75%) for standardized evaluation. We innovatively propose CARP, a cloud-aware prompting framework built upon CoOp, to combat feature corruption, semantic misalignment, and utility decay. Our core contributions include: (1) GCE Loss dynamically adjusting classification weights to suppress cloud interference based on feature degradation severity; (2) Adaptive Optimization Prompt Design (AOPD) utilizing distortion-aware vectors for effective multi-modal feature alignment and semantic deviation repair; (3) Dynamic Weight Adjustment Mechanism (DWAM) real-time balancing of multi-source feature fusion by evaluating inter-modal information gain. Extensive experiments on SeaCloud-Ship demonstrate CARP's superior robustness and state-of-the-art performance, establishing a strong baseline for cloud-occluded ship classification.

KEYWORDS

remote sensing ship classification, vision-language models, few-shot learning, cloud occlusion, prompt tuning

1 Introduction

Advancements in remote sensing have led to widespread Earth observation programs, elevating Remote Sensing Fine-Grained Ship Classification (RS-FGSC) as a critical task. However, RS-FGSC faces significant challenges, particularly severe data scarcity. Acquiring sufficient high-quality, labeled multi-category satellite ship imagery is extremely costly, often resulting in few-shot or zero-shot scenarios. Under these low-data conditions, traditional deep learning models struggle to generalize effectively.

Pretrained on large image-text pairs, Vision-Language Models (VLMs) like Contrastive Language-Image Pretrainin (CLIP) (Radford et al., 2021) offer a new paradigm for RS-FGSC, leveraging rich cross-modal semantics to lead in various remote sensing tasks (Bao et al., 2022; Chen et al., 2020; Jia et al., 2021; Lee et al., 2018; Li et al., 2021; Harold Li et al., 2019; Li X. et al., 2020; Lu et al., 2019; Singh et al., 2022; Su et al., 2019; Tan et al., 2019; Wang et al., 2025a; Wang et al., 2024a; Wang et al., 2025b; Wang et al., 2024b; Wang et al., 2021; Zhang et al., 2024). Standard adaptation fine-tunes these VLMs often updating only

the classification head or prompts—using annotated RS-FGSC data. Yet for cloud-occluded few-shot RS-FGSC, this strategy encounters key obstacles: (1) Absence of dedicated benchmarks for cloud robustness evaluation; (2) Vulnerability to Cloud occlusion interference distinct to satellite imagery.

To bridge the first gap, we establish SeaCloud-Ship–the first benchmark for cloud-occluded fine-grained ship classification. Comprising 7,654 high-resolution images across 30 categories with quantified cloud coverage (12.5%-75%), it enables rigorous evaluation under controlled noise conditions. We further develop CloudGEN, a physics-based cloud synthesis method, to extend benchmark versatility.

Regarding the second challenge, three critical sub-problems emerge: Feature Corruption: Clouds obscure ships' discriminative local features, compromising reliable feature extraction from limited samples. Semantic Misalignment: Cloud distortions warp the visual feature space, causing severe mismatch between degraded features and VLMs' textual prompts. Data Utility Degradation: Cloud contamination reduces effective training samples, intensifying few-shot learning difficulties. Conventional fine-tuning and prompt learning methods show significantly reduced robustness and generalization under these combined pressures.

To address these challenges, we propose cloud-adaptive robust prompt (CARP)—a novel prompt-learning framework built on CoOP. Its core innovations: Mitigating feature corruption by replacing CE loss with adaptive gradient-weighted generalized cross entropy (GCE); Rectifying semantic misalignment via an Adaptive Optimization Prompt Design with Distortion-aware compensation (AOPD), using learnable cloud-occlusion-aware vectors; Countering data utility degradation through a Dynamic Weight Adjustment Mechanism (DWAM) that adaptively balances visual/textual feature weights in cross-modal contrastive learning. This framework establishes a new paradigm for cloud-occluded few-shot fine-grained ship classification. In summary, our contributions are as follows:

1. We introduce SeaCloud-Ship, the first benchmark dataset specifically designed for fine-grained ship classification under cloud occlusion conditions;
2. We propose the cloud-adaptive robust prompt (CARP) framework to enhance cloud-occluded few-shot learning via generalized cross entropy for cloud-noise resistance, Adaptive Optimization Prompt Design for feature distortion rectification, and Dynamic Weight Adjustment Mechanism for cross-modal alignment optimization;
3. We experimentally demonstrate CARP's superior performance over existing models on SeaCloud-Ship, achieving state-of-the-art results across diverse cloud coverage ratios and few-shot settings.

2 Related work

2.1 Prompt learning in vision-language models

Prompt learning has emerged as a dominant paradigm for efficiently adapting large-scale vision-language models (VLMs) to downstream vision tasks without the computational burden of full fine-tuning. Pioneered by CLIP (Radford et al., 2021), which utilized hand-crafted textual prompts like "a photo of a [class]" to align image-text representations, this approach evolved significantly with Context Optimization (CoOp) (Zhou et al., 2022b). CoOp replaced manual prompts with learnable continuous vectors optimized through gradient descent, substantially improving few-shot generalization by dynamically adapting prompts to target datasets. Subsequent research expanded this foundation along several dimensions: CoCoOp (Zhou et al., 2022a) introduced instance-conditional prompts to enhance generalization beyond base categories; VPT (Jia et al., 2022) unified visual and textual prompting within a shared optimization framework; while methods like MaPLe (Uzair Khattak et al., 2023) and ProGrad (Zhu et al., 2023) enforced hierarchical multimodal alignment through constraint-based learning. Domain-specific adaptations also emerged, such as ship-targeted prompt tuning (Lan et al., 2024) that customizes maritime semantics, and training-free variants like Tip-Adapter (Zhang et al., 2021) leveraging cached embeddings for zero-shot transfer. Collectively, these methods demonstrate robust performance across diverse vision tasks including open-vocabulary classification and object detection, provided they operate on high-resolution, unobstructed imagery.

However, these successes falter dramatically under cloud occlusion in remote sensing contexts. VLMs' reliance on discriminative visual features renders them acutely vulnerable to cloud-induced local information loss, which corrupts feature extraction. Static prompt embeddings cannot dynamically compensate for such distortions, leading to progressive misalignment between visual and textual representations as cloud density increases. This drift is exacerbated in few-shot settings, where limited data impedes robust calibration against noise. While recent multimodal alignment techniques address generic domain shifts, they remain fundamentally unequipped to handle structured atmospheric interference (Wang et al., 2025b; Wang et al., 2024b). Consequently, existing frameworks fail to resolve the core challenges of feature corruption, semantic drift, and data scarcity in cloud-occluded fine-grained ship classification, highlighting a critical research void.

2.2 Ship classification

Ship classification in remote sensing imagery is challenged by fine-grained inter-class variations, complex maritime backgrounds, and pervasive cloud occlusion—a uniquely disruptive factor that catastrophically degrades visual features. Recent deep learning approaches have made significant strides through two primary strategies: domain adaptation for sensor invariance and multimodal learning for feature enrichment. Zheng et al. (2023) pioneered a dual-teacher framework (SCSD) that decomposes optical/SAR supervision into interactive cross-domain and semi-supervised tasks, substantially improving pseudo-label reliability on unlabeled SAR imagery. This direction was extended by the Multi-Level Alignment Network (Xu et al., 2022), which integrates pixel-, instance-, and feature-level alignment to mitigate domain shifts in detection tasks. For fine-grained discrimination, Huang et al. (2022) combined CNN-Swin hybrid architectures with multi-branch

feature extraction, setting new benchmarks on FGSC-23 and military ship datasets. Concurrently, Lu et al. (2025) developed the Multi-Scale Context Aggregation Network (MSCAN), leveraging hierarchical convolution-attention fusion to suppress coastal clutter in SAR imagery and significantly enhance small-vessel localization. In multimodal learning, Li W. et al. (2020) achieved cross-sensor alignment via Cross-Modal Contrastive Learning (CMCL), constructing a shared semantic space that maximizes inter-modal consistency while preserving sensor-specific diversity.

Despite these innovations, cloud occlusion remains a critical unsolved vulnerability. Existing methods predominantly optimize for domain shifts (Xu et al., 2022; Zheng et al., 2023) or background interference (Lu et al., 2025), while cloud corruption—which intrinsically obliterates local discriminative features—demands fundamentally different mitigation mechanisms. Although unsupervised change detection (Zheng et al., 2021) offers partial relief by exploiting temporal sequences, it fails under persistent cloud cover common in maritime monitoring. Crucially, no current approach addresses the tripartite challenge of cloud-induced feature loss, semantic distortion, and data scarcity in few-shot settings. Our work bridges this gap by integrating physics-aware cloud modeling with prompt-based vision-language adaptation, establishing the first unified framework for occlusion-robust fine-grained classification.

3 Methods

3.1 Preliminary

CoOp (Zhou et al., 2022b) pioneers prompt learning for vision-language models by optimizing continuous context vectors in text prompts to adapt models like CLIP to downstream tasks. Unlike manual prompt engineering, CoOp automatically learns task-specific contextual representations through backpropagation. Given input image $\mathbf{x}$ and class label c, the text prompt is constructed as:

$$\mathbf{t}_c = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M, \mathbf{c}_c] \tag{1}$$

Here $\mathbf{v}_1, \ldots, \mathbf{v}_M$ are M learnable context vectors (each matching CLIP's text embedding dimension), and $\mathbf{c}_c$ is the class name embedding. During training, these context vectors are optimized to minimize prediction loss while keeping CLIP's original image and text encoders (f and g) frozen. Classification probability is computed as:

$$p(y = c|\mathbf{x}) = \frac{\exp\left(\cos\left(f(\mathbf{x}), g(\mathbf{t}_c)\right)/\tau\right)}{\sum_{c'=1}^{C} \exp\left(\cos\left(f(\mathbf{x}), g(\mathbf{t}_{c'})\right)/\tau\right)} \tag{2}$$

where $\cos(\cdot)$ represents cosine similarity, τ is the temperature parameter, and C denotes the number of classes. CoOp substantially boosts few-shot performance (e.g., increasing CLIP's accuracy by 10%–20% in 1–16 shot tasks) while demonstrating strong cross-dataset generalization and maintaining parameter efficiency.

Although CoOp shows strong capabilities in zero-shot recognition tasks for vision-language models, it faces significant challenges in fine-grained ship recognition under cloud occlusion in remote sensing imagery. Standard prompt learning struggles to overcome visual feature degradation due to heavy cloud obstruction, causing misalignment between image semantics and textual class prompts while hindering the capture of critical fine-grained visual features (e.g., distinguishing warships from cargo vessels). Furthermore, cloud corruption degrades visual features, increasing model sensitivity to label noise. Heavily occluded samples exhibit ambiguous representations, which amplify the impact of mislabeled data during training.

We therefore propose the cloud-adaptive robust prompt (CARP)—a novel prompt learning paradigm for cloud-occluded scenarios. Building upon the CoOp framework, CARP systematically addresses cloud interference through two strategies: 1) Introducing Adaptive Optimization Prompt Design (AOPD) and Dynamic Weight Adjustment Mechanism (DWAM) modules to resolve visual-semantic misalignment and fine-grained feature extraction; 2) Replacing traditional cross-entropy (CE) loss with generalized cross entropy (GCE) to mitigate label noise sensitivity. By incorporating a tunable parameter q, GCE adaptively balances attention to hard samples (potentially noisy) while maintaining high-confidence predictions for clear samples, thereby reducing negative impacts from noisy labels. CARP provides the first end-to-end robust prompt learning solution for fine-grained ship recognition under cloud occlusion in remote sensing (see structure in Figure 1).

3.2 Adaptive optimization prompt design (AOPD)

To enhance model robustness against cloud occlusion, we propose an Adaptive Optimization Prompt Design (structure shown in Figure 2). This integrates a compensation vector dynamically into the prompt learning process, forming a dual-stream prompt architecture: one stream employs standard learnable context vectors ($\mathbf{C}_{\text{ctx}}$) to capture task-related general semantics, while the other stream utilizes a specifically learned occlusion compensation vector to adaptively adjust and enhance semantic prompts in response to potential occlusion patterns in input images.

The Occlusion Compensation Vector $\mathbf{V}_{\text{occ}}$—a learnable parameter with the same dimensionality as $\mathbf{C}_{\text{ctx}}$—strengthens model resilience to visual occlusions. This mechanism optimizes prompt embedding generation by dynamically compensating for representation deviations in occluded regions. Its core operational principle involves:

$$\mathbf{C}_{\text{Cloud}} = \mathbf{C}_{\text{ctx}} + \mathbf{V}_{\text{occ}} \tag{3}$$

$$\mathbf{V}_{\text{occ}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \quad \sigma = 0.02 \tag{4}$$

Here, $\mathbf{C}_{\text{ctx}} \in \mathbb{R}^{n_{\text{ctx}} \times d}$ denotes the original context vector, and $\mathbf{V}_{\text{occ}}$—a compensation parameter optimized via backpropagation and initialized from a zero-mean Gaussian distribution—is jointly optimized with $\mathbf{C}_{\text{ctx}}$ during training. This produces the final prompt embedding:

$$\mathbf{P} = \Gamma(\mathbf{C}_{\text{Cloud}}) \tag{5}$$

which dynamically handles visual occlusion scenarios, significantly enhancing representation capability with missing visual information. This design allows prompts to perceive and compensate for occlusion interference, improving text-visual feature correlation under occlusions.

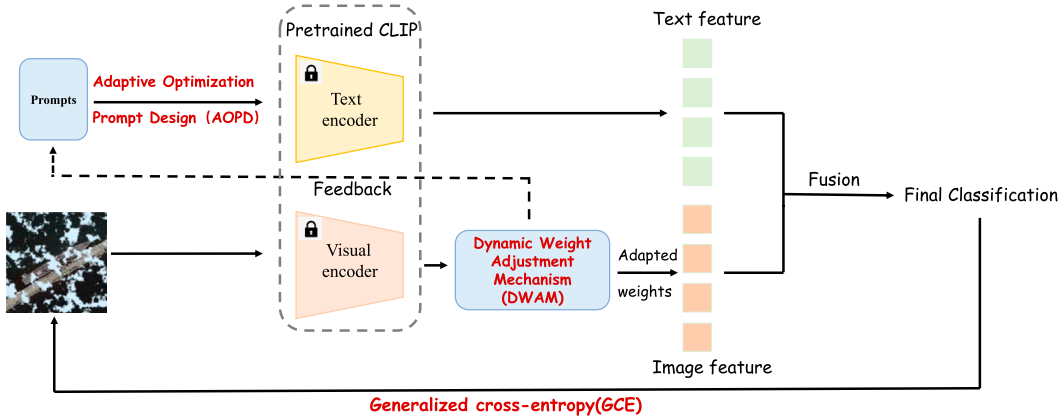

FIGURE 1
Our cloud-adaptive robust prompt (CARP) architecture processes input images through a visual encoder to extract features, which feed into the Dynamic Weight Adjustment Mechanism (DWAM) to automatically balance text and visual feature weights. Text features pass through the Adaptive Optimization Prompt Design (AOPD) to learn occlusion vectors before entering the text encoder. The final output undergoes gradient backpropagation via the generalized cross entropy (GCE).

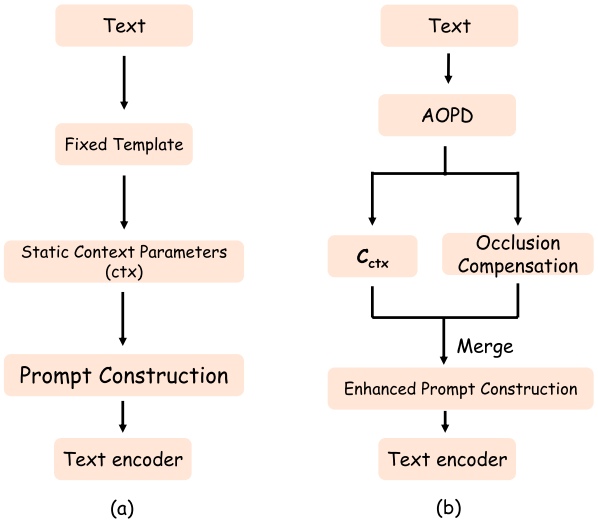

FIGURE 2
(a) shows the state before improvement, and **(b)** shows the state after improvement. AOPD Dual-Stream Architecture: Introducing occlusion vectors to enhance model robustness.

3.3 Dynamic weight adjustment mechanism (DWAM)

When input images suffer severe cloud occlusion, degrading visual feature reliability and impeding text-visual alignment, we propose the Dynamic Weight Adjustment Mechanism (DWAM). This strategy adaptively reduces visual modality weighting while enhancing reliance on language inputs (e.g., reliable categorical text descriptions), maintaining robust semantic alignment under heavy occlusion. Crucially in few-shot scenarios, this adaptive balancing efficiently utilizes limited information and prevents overfitting to noisy data.

DWAM captures real-time multimodal gradients via backpropagation, generating dynamic weight coefficient α from gradient L_2-norms:

$$\alpha = \sigma\left(\frac{\|\nabla_{\mathrm{V}}\|_2}{\|\nabla_{\mathrm{V}}\|_2 + \|\nabla_{\mathrm{T}}\|_2}\right) \tag{6}$$

where $\|\nabla_{\mathrm{V}}\|_2$ and $\|\nabla_{\mathrm{T}}\|_2$ denote image/text encoder gradient norms, and $\sigma(\cdot)$ is the Sigmoid function. The coefficient $\alpha \in [0, 1]$ re-weights features:

$$\mathbf{V}_{\mathrm{new}} = \alpha\mathbf{V}, \quad \mathbf{T}_{\mathrm{new}} = (1-\alpha)\mathbf{T} \tag{7}$$

Normalized features compute cosine similarity:

$$\mathrm{logits} = \exp(s) \cdot \left(\frac{\mathbf{V}_{\mathrm{new}}}{\|\mathbf{V}_{\mathrm{new}}\|_2}\right)\left(\frac{\mathbf{T}_{\mathrm{new}}}{\|\mathbf{T}_{\mathrm{new}}\|_2}\right)^{\top} \tag{8}$$

This dynamically allocates representation weights based on per-modality optimization difficulty, where s is a learnable log scale, and $\mathbf{V}$, $\mathbf{T}$ are original image/text features.

3.4 Generalized cross entropy (GCE)

To address label noise sensitivity from cloud occlusion, we replace conventional cross-entropy (CE) with generalized cross entropy (GCE). This module dynamically balances noise robustness and training efficiency via tunable parameter q: degenerating to standard CE when $q \to 0$ and equaling MAE loss at $q = 1$. It is defined as:

$$\mathcal{L}_{\mathrm{GCE}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1 - p_{y_i}^{q}}{q}, \quad q \in (0, 1] \tag{9}$$

where p_{y_i} is sample i's predicted probability for its true class, and N is batch size. Experiments confirm $q = 0.3$ optimally adapts to remote sensing cloud occlusion: retaining CE's fast convergence while suppressing cloud-induced label noise overfitting through

moderate loss curvature adjustment. For enhanced robustness, gradient backpropagation halts when prediction confidence $p_{y_i} < \kappa$ ($\kappa = 0.5$):

$$\nabla_\theta^{(i)} = \begin{cases} 0 & p_{y_i} < \kappa \\ \nabla_\theta \left(\frac{1 - p_{y_i}^q}{q} \right) & \text{otherwise} \end{cases} \tag{10}$$

Combined with AOPD's compensation vector, this dual robust optimization significantly boosts generalization under degraded visual features. When q approaches 0, GCE degenerates into the standard cross-entropy (CE), which is sensitive to noisy samples. When $q = 1.0$, it is equivalent to the mean absolute error (MAE), which enhances noise tolerance but impairs the efficiency of gradient updates. Experiments show that $q = 0.3$ achieves the optimal balance in the cloud occlusion noise environment—it not only suppresses the interference of label noise by adjusting the loss curvature but also retains the discriminative ability for key ship features. Algorithm 1 provides complete CARP pseudocode.

Require:
T: Training epochs
$\mathcal{X}$: Input dataset
C: Number of classes
τ: Temperature parameter
σ: $\mathbf{V}_{\text{occ}}$ init std dev
(q, κ): GCE parameters ($q \in (0, 1], \kappa \in [0, 1]$)
$\mathbf{C}_{\text{ctx}}$: Learned context vectors
Ensure: $\mathbf{V}_{\text{occ}}$: Occlusion compensation vectors
s: Scaling factor

Initialize $\mathbf{C}_{\text{ctx}}$, $\mathbf{V}_{\text{occ}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, s
for $t \leftarrow 1$ to T do
  Sample $(\mathbf{x}, c) \in \mathcal{X}$
  Extract visual feature $\mathbf{V} = f(\mathbf{x})$
  Generate prompt: $\mathbf{t}_c = [\Gamma(\mathbf{C}_{\text{ctx}} + \mathbf{V}_{\text{occ}}), \mathbf{c}_c]$ (Equations 3, 5)
  Extract text feature $\mathbf{T}_c = g(\mathbf{t}_c)$
  Compute weight: $\alpha = \sigma(\|\nabla_{\mathbf{V}}\|_2 / (\|\nabla_{\mathbf{V}}\|_2 + \|\nabla_{\mathbf{T}_c}\|_2))$ (Equation 6)
  Adjust features: $\mathbf{V}_{\text{new}} = \alpha \mathbf{V}$, $\mathbf{T}_{\text{new}} = (1 - \alpha)\mathbf{T}_c$ (Equation 7)
  Compute similarity: $\text{logits}_c = \exp(s) \cdot \cos(\mathbf{V}_{\text{new}}, \mathbf{T}_{\text{new}})$ (Equation 8)
  Calculate probability: $p_c = \frac{\exp(\text{logits}_c / \tau)}{\sum_{c'} \exp(\text{logits}_{c'} / \tau)}$ (Equation 2)
  if $p_c \geq \kappa$ then
    Compute loss: $\mathcal{L}_{\text{GCE}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1 - p_{y_i}^q}{q}$, $q \in (0, 1]$ (Equation 9)
    Obtain gradients: $\nabla = \nabla_{\mathbf{C}_{\text{ctx}}, \mathbf{V}_{\text{occ}}, s} \mathcal{L}$
  else
    Set gradients: $\nabla = \mathbf{0}$ (Equation 10)
  end if
  Update parameters with ∇
end for

Algorithm 1. Cloud-adaptive robust prompt (CARP) Training.

4 Experiments

In this section, we first introduce SeaCloud-Ship, a dataset specifically constructed for few-shot ship fine-grained classification under cloud occlusion conditions. Based on this dataset, we conduct comprehensive experiments to evaluate the performance of the proposed method under varying levels of cloud occlusion and validate the effectiveness of the cloud-adaptive robust prompt (CARP). The experiments cover dataset construction details, baseline comparisons, implementation parameters, and multi-dimensional result analysis.

4.1 Datasets

With the growing demand for cloud detection technology, an increasing number of open-source datasets (Aybar et al., 2022; Foga et al., 2017; Mohajerani et al., 2019; Shendryk et al., 2019) have become available for training purposes. However, the absence of fine-grained ship recognition datasets under cloud occlusion conditions hinders validation of model robustness in such scenarios. To address this, we developed SeaCloud-Ship—a publicly accessible benchmark dataset specifically designed for fine-grained ship classification under cloud interference, with Figure 3 illustrating the dataset construction workflow.

4.1.1 Source screening

Due to the lack of open-source fine-grained ship datasets under cloud occlusion, the challenges in collecting such data, the inability to cover all common categories, and the difficulty in controlling cloud coverage ratios to accurately validate model performance, our data primarily derives from three public fine-grained ship classification datasets: FGSC-23 covers 23 categories (22 ship types and 1 non-ship category) with 4,080 samples. Partially sourced from the HRSC2016 dataset, its main imagery comes from Google Earth and GF-1 satellite, covering vessels under various lighting conditions, land/sea backgrounds, and arbitrary orientations. FGSCR-42 comprises 9,320 images across 42 common ship categories (including naval and civilian vessels), with approximately 200 images per class. Image resolutions range from 50× 50 to 1,500× 1,500 pixels, reflecting multi-scale ship features. Sources include Google Earth and mainstream remote sensing datasets like DOTA, HRSC 2016, and NWPU VHR-10. FGSCM-52 expands FGSCR-42 by adding 10 categories, forming 52 fine-grained ship classes (e.g., warships, civilian vessels). It incorporates multi-scale optical remote sensing imagery (50× 50 to 1,600× 1700 pixels, 0.4–2 m resolution from GF-2/Sentinel-2), integrating multi-source data from Google Earth, DOTA, and HRSC 2016. This dataset enhances annotations for rare hull types (e.g., auxiliary vessel subtypes) and adds complex scenarios (ports, offshore areas), emphasizing feature variations under diverse lighting, angles, and backgrounds.

4.1.2 Selection rule

To construct a cloud-interference-aware fine-grained ship recognition benchmark with scenario generalization capabilities, we selected source dataset samples using a three-tiered criterion:

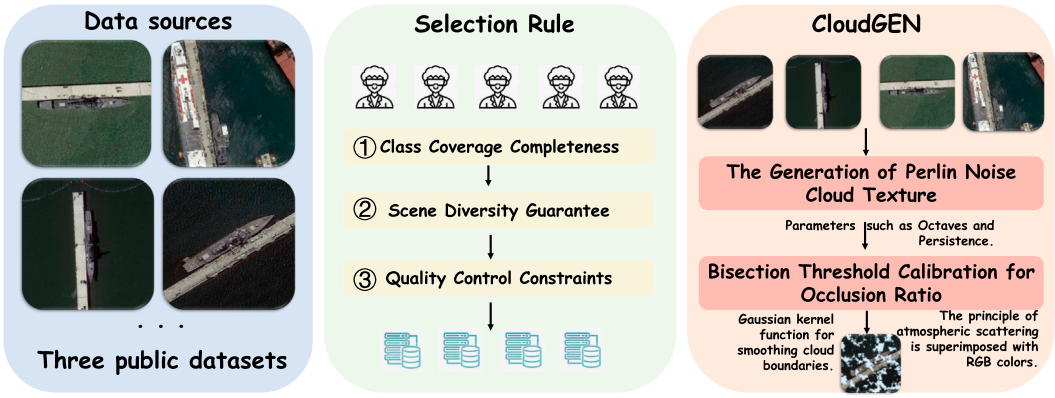

FIGURE 3
Pipeline of SeaCloud-Ship. Our dataset construction pipeline includes data collection, three levels of manual quality inspection, and final generation via the CloudGEN.

First, First, ensure the completeness of category coverage, and prioritize the selection of multi-view and multi-scale samples to retain fine-grained discriminative features; Second, guaranteeing scene diversity by mandating coverage across four typical scenarios: open waters, coastal zones, complex channels, and extreme illumination conditions; Finally, applying quality constraints to retain relatively high-resolution images inherently free from cloud occlusion, thereby preventing pre-existing interference. Through this process, 7,654 high-quality images were selected to form the foundational dataset. Controllable occlusion was applied via the automated cloud generation system (CloudGEN), while strictly preserving the original directory structure and class balance. Figure 4 demonstrates visual comparisons under varying occlusion ratios, validating the effectiveness of data simulation.

4.1.3 Cloud generation method

Based on the rigorously curated high-quality ship image dataset, we developed the automated cloud generation system CloudGEN, designed to controllably add simulated cloud layers with varying occlusion ratios. While existing cloud generation methods approach realistic cloud quality, genuine cloud occlusion data presents challenges in controlling obstruction ratios and collecting sufficient category coverage. Consequently, we employ an automated approach for generating cloud occlusion ratios. To authentically replicate real cloud interference characteristics in remote sensing imagery, we utilize a Perlin noise-based physical simulation method. By modeling optical properties under diverse cloud thicknesses, density distributions, and lighting conditions, we dynamically generate realistic cloud noise matching actual satellite observations directly onto raw ship images. First, we dynamically create naturalistic cloud textures using a fractal noise algorithm, adjusting octaves and persistence parameters to control cloud complexity and distribution. Second, a dichotomy-based threshold calibration technique precisely governs cloud coverage, ensuring accurate occlusion ratios per image. Finally, Gaussian blurring achieves optical transitions at cloud edges, while overlaying spectrally specific cloud colors simulates authentic atmospheric scattering. This process batch-generates cloud-occluded samples conforming to remote sensing imaging principles while preserving the original dataset's directory structure. Figure 4 illustrates comparative results under varying cloud occlusion ratios.

4.1.4 Statistic

The SeaCloud-Ship dataset comprises 30 fine-grained ship categories (detailed in Table 1), spanning warships (e.g., destroyers, aircraft carriers, frigates), civilian vessels (e.g., yachts, cargo ships), and specialized ships (e.g., hospital ships, tugboats). It contains 7,654 high-quality optical remote sensing images with a mean resolution of 512× 512 pixels, ranging from low-scale (92× 92) to high-scale (1024× 1024). The dataset integrates multi-source heterogeneous data, including Google Earth imagery, China's Gaofen satellites (GF-1/GF-2), the European Sentinel-2 satellite, and aerial platforms like HRSC2016 and DOTA, primarily using high-resolution optical sensors (0.4–2 m). Samples originate from global maritime zones—Pacific, Atlantic, and Indian Oceans—covering diverse scenarios such as open seas, coastal waters, complex channels, and extreme lighting. The largest category, Towing_vessel, includes 778 images, averaging 255 per class with balanced military-civilian ratios. High-resolution classes focus on critical military targets: aircraft carriers (Nimitz-class averaging 812× 812 pixels) and destroyers (Arleigh Burke-class averaging 540× 540 pixels). Civilian classes like container ships (mean 442× 442 pixels) and sand carriers (Sand_carrier averaging 230× 230 pixels) emphasize low-resolution robustness testing. This multi-scale, multi-source, multi-scenario framework establishes SeaCloud-Ship as the first benchmark dataset supporting fine-grained vessel recognition under cloud-occluded conditions.

4.2 Datasets and baselines

Datasets Building upon (Di et al., 2021; Lan et al., 2024; Zhang et al., 2020), we construct the first cloud occlusion benchmark dataset SeaCloud-Ship. Using CloudGEN4.1.3, we synthesize progressive cloud occlusion (coverage levels: 12.5%, 25%, 37.5%, 50%, 62.5%, 75%), resulting in 7,654 images.

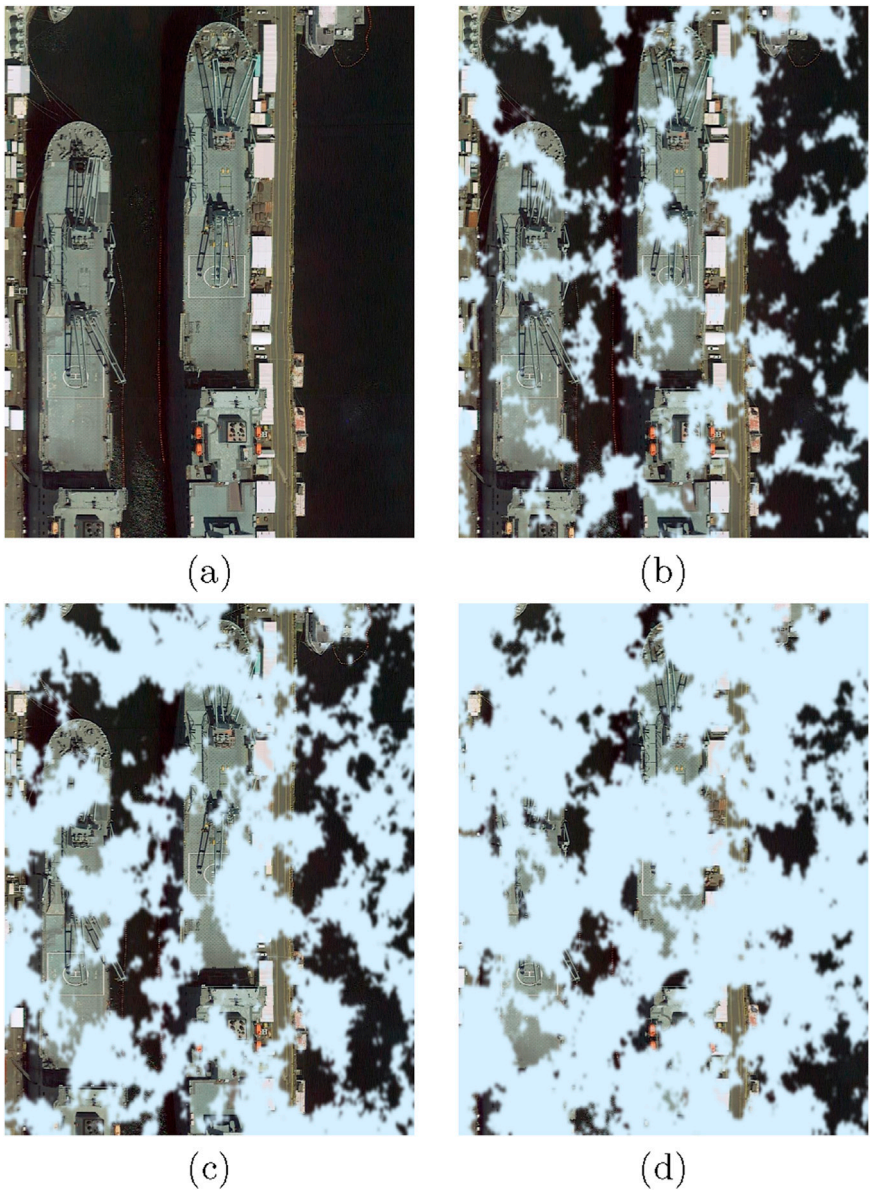

FIGURE 4
Figures **(a–d)** are comparative diagrams under cloud occlusion ratios of 0%, 25%, 50%, and 75%, respectively.

Baselines Four state-of-the-art prompt learning methods are selected for comparison: (1) CoOp (Zhou et al., 2022b): classical context optimization; (2) CoOp + MAE: incorporates masked autoencoder pretraining on CoOp for enhanced robustness; (3) CoCoOp (Zhou et al., 2022a): conditionally generated instance-level prompts; (4) NLPrompt (Pan et al., 2024): language-vision alignment method specifically designed for noisy scenarios, serving as the strongest relevant baseline.

4.3 Implementation details

Experiments utilize the CLIP-ViT/B-16 architecture with uniform training settings: SGD optimizer (initial learning rate 2×10^{-3}, momentum 0.9, weight decay 5×10^{-4}) coupled with cosine learning rate decay, batch size of 32 (16 samples per visual/text modality) across 200 epochs. Generalized cross entropy (GCE) employs noise tolerance parameter $q = 0.3$. The Dynamic Weight Adjustment Mechanism (DWAM) adopts a GRU structure (hidden dimension 64), maintains a gradient history buffer of length 10, and dynamically weights features by constraining $\alpha \in [0.3, 0.7]$.

Context learning configuration includes 16-dimensional context vectors ($n_{\text{ctx}} = 16$) with class tokens at text suffix positions; compensation vectors follow $\mathcal{N}(0, 0.02)$ initialization. Full training employs FP16 mixed precision. Hardware: NVIDIA GeForce RTX 3080 GPU and Intel Core i9-10920X CPU. Software: PyTorch 2.4.1/CUDA 12.1/cuDNN 8.0.5. Data augmentation includes random cropping (scale = [0.08, 1.0]), horizontal flipping, and ImageNet normalization. Reported results average three independent runs.

TABLE 1 Classification statistics of the ship dataset.

Category	Subclass	Quantity	Resolution (px)
Military Ships			
	Kidd Class Destroyer	68	615 × 605
	Crane Ship	142	1024 × 1024
	Independence Class Combat Ship	210	606 × 597
	Kuznetsov Class Aircraft Carrier	68	686 × 676
	Arleigh Burke Class Destroyer	580	548 × 541
	Akizuki Class Destroyer	18	715 × 705
	Murasame Class Destroyer	63	595 × 588
	Asagiri Class Destroyer	70	652 × 643
	Abukuma Class Frigate	78	683 × 672
	Freedom Class Combat Ship	177	522 × 514
	Whitby Island Class Dock Landing Ship	278	616 × 606
	Osumi Class Landing Ship	116	770 × 758
	Kitty Hawk Class Aircraft Carrier	68	875 × 858
	Sacramento Class Support Ship	50	915 × 903
	Izumo Class Helicopter Destroyer	63	951 × 937
	Type 45 Destroyer	159	710 × 699
	Midway Class Aircraft Carrier	208	545 × 538
	Hyuga Class Helicopter Destroyer	24	925 × 909
	Wasp Class Assault Ship	453	659 × 643
	Ticonderoga Class Cruiser	607	515 × 506
	Medical Ship	322	732 × 721
	Nimitz Class Aircraft Carrier	553	813 × 800
	San Antonio Class Transport Dock	319	721 × 710
Civilian Ships			
	Civil Yacht	777	115 × 114
	Megayacht	186	562 × 554
	Container Ship	455	443 × 436
	Tank Ship	160	743 × 732
	Towing Vessel	778	92 × 92
	Sand Carrier	226	230 × 226
	Cargo Ship	378	609 × 600

4.4 Main results

As shown in Table 2, our proposed method demonstrates significant advantages in remote sensing image recognition under cloud occlusion. It achieves optimal performance across all 30 comparative experiments, particularly excelling under high occlusion (75.00%) and few-shot (1-shot) scenarios. Compared to the best-performing baseline (CoOp + MAE), our method improves average accuracy by 3.6%, with the maximum gain in the 16-shot/12.50% scenario (61.03% vs. 56.56%). These results validate the effectiveness of our cloud-feature adaptive learning mechanism.

Our method exhibits exceptional robustness across varying occlusion levels. When occlusion increases from 12.50% to 75.00%, the best baseline (CoOp + MAE) shows 46.16% performance degradation under 1-shot conditions, while our method maintains degradation within 51.97%. Notably at the critical 50.00% occlusion threshold, our 1-shot accuracy reaches 23.63%, significantly exceeding CoOp + MAE (18.22%) and NLPrompt (16.55%). Even under

TABLE 2 Performance comparison under different cloud occlusion ratios.

Shots	Method	Cloud occlusion ratio					
		12.50%	25.00%	37.50%	50.00%	62.50%	75.00%
1-shot	CoOp (Zhou et al., 2022b)	25.43	23.12	20.56	18.79	15.21	12.34
	CoCoOp (Zhou et al., 2022a)	22.89	21.05	19.47	17.33	14.88	11.76
	CoOp + MAE	24.17	22.68	20.91	18.22	15.93	13.02
	NLPrompt (Pan et al., 2024)	21.35	19.87	18.14	16.55	13.98	10.65
	Ours	**29.92**	**26.81**	**21.12**	**23.63**	**18.35**	**14.37**
2-shot	CoOp (Zhou et al., 2022b)	30.21	28.45	25.78	22.34	19.05	17.22
	CoCoOp (Zhou et al., 2022a)	27.98	26.11	23.44	21.56	18.77	16.89
	CoOp + MAE	29.56	27.79	24.92	22.11	19.33	17.55
	NLPrompt (Pan et al., 2024)	26.34	24.57	22.89	20.02	17.66	15.90
	Ours	**34.91**	**31.20**	**26.03**	**22.83**	**19.36**	**18.37**
4-shot	CoOp (Zhou et al., 2022b)	40.23	35.67	32.11	29.88	26.54	23.01
	CoCoOp (Zhou et al., 2022a)	38.76	34.55	31.09	28.77	25.33	22.12
	CoOp + MAE	41.56	36.89	33.22	30.99	27.44	24.35
	NLPrompt (Pan et al., 2024)	37.45	33.21	29.78	27.56	24.11	21.03
	Ours	**44.52**	**38.83**	**33.37**	**31.92**	**28.33**	**24.37**
8-shot	CoOp (Zhou et al., 2022b)	50.12	45.34	42.78	38.56	33.21	29.87
	CoCoOp (Zhou et al., 2022a)	48.76	44.55	41.09	37.88	32.54	28.90
	CoOp + MAE	51.56	46.89	43.22	39.99	34.44	30.35
	NLPrompt (Pan et al., 2024)	47.45	43.21	39.78	36.56	31.11	27.03
	Ours	**54.28**	**49.41**	**46.92**	**40.87**	**34.66**	**31.21**
16-shot	CoOp (Zhou et al., 2022b)	55.23	50.67	47.11	44.88	39.54	33.01
	CoCoOp (Zhou et al., 2022a)	53.76	49.55	46.09	43.77	38.33	32.12
	CoOp + MAE	56.56	51.89	48.22	45.99	40.44	35.35
	NLPrompt (Pan et al., 2024)	52.45	48.21	44.78	42.56	37.11	31.03
	Ours	**61.03**	**57.31**	**53.14**	**50.29**	**40.70**	**36.66**

The bolded content represents the values with the best results.

extreme 75.00% occlusion, our 16-shot result (36.66%) outperforms the best baseline by 1.31 percentage points.

Training sample size analysis reveals pronounced few-shot advantages. Under 1-shot conditions, our method improves average accuracy by 4.27% across occlusion levels; this advantage expands to 4.82% at 16-shot. Specifically in 4-shot scenarios at 62.50% occlusion, our method achieves 28.33% accuracy—a 3.00 percentage point improvement (11.84% relative gain) over CoCoOp (25.33%). These findings indicate that increased training samples enhance our method's ability to learn occlusion-invariant features.

TABLE 3 Module ablation study results based on the CoOp baseline.

GCE	AOPD	DWAM	25.0%	50.0%	75.0%
-	-	-	50.67	44.88	33.01
✓	-	-	54.50	45.50	34.30
-	✓	-	56.30	46.60	33.20
-	-	✓	55.70	46.30	34.70
✓	✓	✓	**57.31**	**50.29**	**36.66**

The bolded content represents the values with the best results.

4.5 Ablation experiments

As shown in Table 3, we systematically evaluate the synergistic effects of generalized cross entropy (GCE), Adaptive Optimization Prompt Design (AOPD), and Dynamic Weight Adjustment Mechanism (DWAM). Experiments demonstrate that individually introducing any module improves baseline performance, with DWAM yielding a 3.1% gain under 75% occlusion, highlighting

TABLE 4 Ablation study results of loss functions (classification accuracy%).

Loss	25.0%	50.0%	75.0%
MAE	53.90	47.30	32.70
CE	55.20	48.30	35.50
GCE ($q = 0.5$)	55.10	47.30	34.50
GCE ($q = 0.3$)	**57.30**	**50.30**	**36.60**

The bolded content represents the values with the best results.

its adaptability to extreme occlusion. Further analysis reveals functional complementarity: GCE enhances noise robustness, AOPD optimizes occlusion feature reconstruction, and DWAM achieves cross-modal dynamic balancing. When integrating all modules, the model reaches 50.30% accuracy under 50% occlusion—a 0.6 percentage point improvement over baseline, validating the efficacy of collaborative module design.

Loss function analysis (Table 4) uncovers key mechanisms: GCE with $q = 0.3$ significantly outperforms standard CE and MAE losses, showing a 2.10 percentage point improvement over CE at 25% occlusion. This stems from GCE's noise tolerance, dynamically weighting hard samples via q to suppress false signals in cloud-occluded regions. Notably, q exhibits sensitivity thresholds: at $q = 0.5$, diminished gradient smoothing causes noticeably lower performance gains under 75% occlusion compared to the $q = 0.3$ configuration.

5 Conclusion

This paper addresses the critical challenge of few-shot fine-grained ship classification in cloud-occluded remote sensing by proposing CARP (cloud-adaptive robust prompt)—an innovative framework that systematically tackles feature corruption, semantic misalignment, and data utility degradation. By designing a generalized cross entropy (GCE) loss for noise immunity, developing an Adaptive Optimization Prompt Design (AOPD) to repair semantic mismatches, and introducing a Dynamic Weight Adjustment Mechanism (DWAM) for cross-modal alignment, our method achieves breakthrough performance on the SeaCloud-Ship dataset.

As shown in Table 2, CARP outperforms all existing methods across 30 experimental settings, notably achieving 23.6% accuracy under 1-shot/50% occlusion (5.38% higher than CoOp + MAE) and setting a new state-of-the-art of 36.66% at 16-shot/75% occlusion. Experiments confirm CARP's superior robustness: when occlusion increases from 12.5% to 75.0%, its performance degradation (52.2%) is significantly lower than baselines (46.2%). The newly constructed SeaCloud-Ship dataset addresses a critical domain gap, providing reliable solutions for satellite imaging systems. Future work will explore multi-modal cloud occlusion modeling and zero-shot generalization enhancement.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HZ: Writing – review and editing. YS: Methodology, Visualization, Writing – original draft, Writing – review and editing. XH: Resources, Supervision, Writing – review and editing. XT: Supervision, Writing – review and editing. TZ: Supervision, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Aybar, C., Ysuhuaylas, L., Loja, J., Gonzales, K., Herrera, F., Bautista, L., et al. (2022). CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. *Sci. Data* 9, 782. doi:10.1038/s41597-022-01878-2

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., et al. (2022). Vlmo: unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inf. Process. Syst.* 35, 32897–32912.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., et al. (2020). "Uniter: universal image-text representation learning," in *European conference on computer vision* (Springer), 104–120.

Di, Y., Jiang, Z., and Zhang, H. (2021). A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sens.* 13 (4), 747. doi:10.3390/rs13040747

Foga, S., Scaramuzza, P. L., Guo, S., Zhu, Z., Dilley, R. D., Jr, Beckmann, T., et al. (2017). Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* 194, 379–390. doi:10.1016/j.rse.2017.03.026

Harold Li, L., Yatskar, M., Yin, D., Hsieh, C. J., and Chang, K. W. (2019). *Visualbert: a simple and performant baseline for vision and language*. arXiv preprint arXiv:1908.03557.

Huang, L., Wang, F., Zhang, Y., and Xu, Q. (2022). Fine-grained ship classification by combining CNN and swin transformer. *Remote Sens.* 14, 3087. doi:10.3390/rs14133087

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., et al. (2021). "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning* (PMLR), 4904–4916.

Jia, M., Tang, L., Chen, B.-C., Cardie, M., Belongie, S., Hariharan, B., et al. (2022). "Visual prompt tuning," in *European conference on computer vision* (Springer), 709–727.

Lan, L., Wang, F., Zheng, X., Wang, Z., and Liu, X. (2024). Efficient prompt tuning of large vision-language model for fine-grained ship classification. *IEEE Trans. Geoscience Remote Sens.* 63, 1–10. doi:10.1109/tgrs.2024.3509721

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. (2018). "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 201–216.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., et al. (2020a). *Unimo: towards unified-modal understanding and generation via cross-modal contrastive learning*. arXiv preprint arXiv:2012.15409.

Li, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., et al. (2020b). "Oscar: objectsemantics aligned pre-training for vision-language tasks," in Computer Vision–ECCV 2020: 16th European Conference Proceedings, Glasgow, United Kingdom, August 23–28, 2020 (Springer), 121–137.

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. (2021). Align before fuse: vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* 34, 9694–9705.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). "ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in neural information processing systems*.

Lu, X., Yue, T., Cai, J., Chen, Y., Lv, C., and Chu, S. (2025). *MSCA-net: multi-scale context aggregation Network for infrared small target detection*. arXiv preprint arXiv:2503.17193.

Mohajerani, S., and Saeedi, P. (2019). "Cloud-Net: an end-to-end cloud detection algorithm for Landsat 8 imagery," in IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium, Yokohama, Japan, 28 July 2019 - 02 August 2019 (IEEE), 1029–1032.

Pan, B., Li, Q., Tang, X., Huang, W., Fang, Z., Liu, F., et al. (2024). *NLPrompt: noise-label prompt learning for Vision-Language Models*. arXiv preprint arXiv:2412.01256.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *International conference on machine learning* (PMLR), 8748–8763.

Shendryk, Y., Rist, Y., Ticehurst, C., and Thorburn, P. (2019). Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. *ISPRS J. Photogrammetry Remote Sens.* 157, 124–136. doi:10.1016/j.isprsjprs.2019.08.018

Singh, A., Hu, R., Gowsami, V., Guillaume, G., Galuba, W., Rohrbach, M., et al. (2022). "Flava: a foundational language and vision alignment model," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 18-24 June 2022 (IEEE), 15638–15650.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., et al. (2019). *Vl-bert: pre-training of generic visual-linguistic representations*. arXiv preprint arXiv:1908.08530.

Tan, H., and Bansal, M. (2019). *Lxmert: learning cross-modality encoder representations from transformers*. arXiv preprint arXiv:1908.07490.

Uzair Khattak, M., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. (2023). "MaPLe: multi-modal prompt learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Vancouver, BC, Canada, 17-24 June 2023 (IEEE), 19113–19122.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. (2021). *Simvlm: simple visual language model pretraining with weak supervision*. arXiv preprint arXiv:2108.10904.

Wang, F., Huang, W., Yang, S., Fan, Q., and Lan, L. (2024a). *Learning to learn better visual prompts*, 5354–5363.

Wang, F., Wang, H., Wang, D., Guo, Z., and Zhong, Z. (2024b). *Scaling efficient masked autoencoder learning on large remote sensing dataset*. arXiv preprint arXiv:2406.11933.

Wang, F., Chen, M., Li, Y., Wang, D., Wang, H., Guo, Z., et al. (2025a). *GeoLLaVA-8K: scaling remote-sensing multimodal large Language Models to 8K resolution*. arXiv preprint arXiv:2505.21375.

Wang, F., Wang, H., Wang, Y., Wang, D., Chen, M., Zhao, H., et al. (2025b). *RoMA: scaling up mamba-based foundation models for remote sensing*. arXiv preprint arXiv:2503.10392.

Xu, C., Zheng, X., and Lu, X. (2022). Multi-level alignment network for cross-domain ship detection. *Remote Sens.* 14, 2389. doi:10.3390/rs14102389

Zhang, X., Lv, Y., Yao, L., Xiong, W., and Fu, C. (2020). A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 13, 1271–1285. doi:10.1109/jstars.2020.2981686

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., et al. (2021). *Tip-adapter: training-free clip-adapter for better vision-language modeling*. arXiv preprint arXiv:2111.03930.

Zhang, J., Huang, J., Jin, S., and Lu, S. (2024). Vision-language models for vision tasks: a survey. *IEEE Trans. Pattern Analysis Mach. Intell.* 46, 5625–5644. doi:10.1109/tpami.2024.3369699

Zheng, X., Chen, X., Lu, X., and Sun, B. (2021). Unsupervised change detection by cross-resolution difference learning. *IEEE Trans. Geoscience Remote Sens.* 60, 1–16. doi:10.1109/tgrs.2021.3079907

Zheng, X., Cui, H., Xu, C., and Lu, X. (2023). Dual teacher: a semisupervised cotraining framework for cross-domain ship detection. *IEEE Trans. Geoscience Remote Sens.* 61, 1–12. doi:10.1109/tgrs.2023.3287863

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022a). "Conditional prompt learning for vision-language models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 18-24 June 2022 (IEEE), 16816–16825.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022b). Learning to prompt for vision-language models. *Int. J. Comput. Vis.* 130 (9), 2337–2348. doi:10.1007/s11263-022-01653-1

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. (2023). "Prompt-aligned gradient for prompt tuning," in Proceedings of the IEEE/CVF international conference on computer vision, Paris, France, 01-06 October 2023 (IEEE), 15659–15669.