remote sensing

MDPI

Article

ProtoWave-Mamba: A Semi-Supervised Approach for Ocean Internal Wave Segmentation with Wavelet Enhancement and Long-Range Sequence Modeling

Kaizhe Feng [1], Dongfang Zhang [1,2], Yaqiong Yu [1], Chengjun Li [1,*], Tingzi Chi [1], Junnan Guo [1] and Yi Yan [1]

[1] School of Computer Science, China University of Geosciences (Wuhan), 388 Lumo Road, Wuhan 430074, China; fengkaizhe2004@cug.edu.cn (K.F.); zhangdongfang@cug.edu.cn (D.Z.); yuyaqiong@cug.edu.cn (Y.Y.); chitingzi@cug.edu.cn (T.C.); guojunnan@cug.edu.cn (J.G.); yanyi@cug.edu.cn (Y.Y.)

[2] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences (Wuhan), Wuhan 430074, China

* Correspondence: chengjunli@cug.edu.cn

Highlights

What are the main findings?

- ProtoWave-Mamba achieves an IoU of 0.7655 and a Dice of 0.8669 on Dataset 1, and an IoU of 0.7154 and a Dice of 0.8333 on Dataset 2, using only 30 labeled images per dataset and outperforming state-of-the-art semi-supervised baselines such as Mean Teacher.
- Module-level ablations demonstrate that wavelet frequency enhancement, Mamba-based long-range modeling, and class prototype alignment each contribute to boundary recovery and structural continuity of slender internal wave strips.

What are the implications of the main findings?

- Prototype-guided pseudo-label denoising mitigates noise accumulation in semi-supervised learning, substantially reducing the demand for costly pixel-level annotations in marine remote sensing.
- The proposed framework provides an accurate and efficient segmentation solution for elongated, low-contrast oceanic targets within wave-containing scenes, providing a step towards large-scale automated internal wave observation.

Abstract

Ocean internal wave segmentation from remote sensing imagery is a challenging task due to the elongated morphology of targets, blurred boundaries, and the high cost of pixel-level annotations. To address these issues, this paper proposes ProtoWave-Mamba (PWM-Net), an end-to-end semi-supervised segmentation framework specifically tailored to the physical characteristics of ocean internal waves. Specifically, it synergistically integrates multi-scale wavelet transforms to enhance weak texture information in the frequency domain, a Mamba architecture with linear-complexity long-range modeling to capture global contextual dependencies of slender structures, and a class prototype alignment module to extract category-wise feature centers from a very limited set of labeled samples. By enforcing cross-sample alignment in the feature space, the proposed method guides the distribution of unlabeled data and performs prototype-guided pseudo-label denoising to suppress the noise accumulation inherent in pseudo-label-based semi-supervised learning. Consequently, ProtoWave-Mamba achieves efficient semi-supervised knowledge transfer and significantly improves boundary representation and feature discrimination with only 30 labeled images per dataset. Experimental results on two representative ocean internal

Check for updates

Academic Editors: Jiaqi Yao, Nan Xu and Shijin Li

Received: 13 August 2026
Revised: 23 September 2026
Accepted: 26 September 2026
Published: 1 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

wave datasets demonstrate that the proposed method consistently outperforms state-of-the-art semi-supervised methods. In particular, compared with the best-performing baseline Mean Teacher, it achieves substantial improvements in Intersection over Union (IoU), Dice coefficient, and Recall, especially under complex background conditions. Ablation studies further confirm the effectiveness of each core module in maintaining contour continuity and fine-grained segmentation quality.

Keywords: ocean internal waves; semi-supervised segmentation; wavelet transform; Mamba; prototype learning; pseudo-label denoising; remote sensing imagery

1. Introduction

As a key mesoscale phenomenon in ocean dynamics, ocean internal waves play a significant role in driving ocean energy dissipation and modulating upper-ocean mixing, making them an important subject for both scientific research and practical applications [1]. The process of internal wave breaking and the turbulence mixing it induces are considered critical elements in oceanic internal mixing and climate-scale circulation simulations [2]. These waves exhibit a strong generation–propagation–dissipation chain characteristic in typical ocean regions [3,4]. Beyond their climatic significance, internal waves pose direct operational hazards in the marine environment: the strong surface currents and enhanced turbulence associated with internal solitary waves can threaten offshore engineering structures, interfere with underwater navigation and drilling operations, and modulate acoustic propagation paths in ways that affect sonar systems and underwater communication. Consequently, accurate and timely monitoring of internal wave activity is of direct relevance to maritime safety, marine resource management, and disaster early warning. With the widespread use of multi-source high-resolution satellite remote sensing technologies, pixel-level segmentation and fine detection of internal wave strips using optical or Synthetic Aperture Radar (SAR) images have become an important technological pathway for automated monitoring of vast ocean areas [5–7].

More recently, the Surface Water and Ocean Topography (SWOT) mission has added a complementary observational dimension: its wide-swath Ka-band radar interferometer measures sea surface height in two dimensions with sufficient precision to resolve the amplitudes and energy fluxes of internal solitary waves, and it has provided the first global view of these waves together with other submesoscale features [8]. Internal solitary wave research is thereby shifting from regional analysis toward global-scale mapping: the global SWOT analysis locates internal solitary wave activity across the South China Sea, the Andaman Sea, the Indonesian archipelago, the Mascarene Plateau and the Amazon shelf, and measures a wave packet in the Andaman Sea with a 20 cm sea surface height amplitude and a wavelength of about 5 km, whose vertically integrated energy flux far exceeds that of the local coherent internal tide. Subsequent studies based on the same observations have characterised the spatiotemporal distribution and vertical dynamics of internal solitary waves in the South China Sea [9], quantified the refraction and diffraction of internal solitary waves induced by mesoscale eddies off the Amazon shelf [10], and established dedicated sea surface height anomaly processing and amplitude-inversion frameworks for internal solitary wave characterisation [11]. In parallel, artificial intelligence is being coupled with these new observational streams so as to generalise internal wave analysis beyond individual study sites; physics-informed neural networks trained on paired satellite observations, for example, can forecast internal solitary wave propagation across an entire marginal sea [12]. Deep learning has similarly been used to convert decades of satellite

archives into long-term regional products, for example a TransUNet-based framework that segmented internal solitary waves from multi-source imagery and produced a 22-year spatiotemporal map of the East China Sea [13]. Radar and optical imagery, which record only the surface expression of the waves, and SWOT, which records the associated sea surface height perturbation, are therefore best regarded as complementary sources for internal wave monitoring rather than as substitutes, and this combination of dense spatial coverage with quantitative amplitude information is directly aligned with the motivation of this work, namely to provide an automated segmentation framework that can be deployed over large ocean areas while keeping the annotation cost of such a deployment within reach. The imaging mechanisms of SAR for internal wave surface textures and strips have also been systematically established through early observations [14–16]. In recent years, deep learning frameworks for internal wave feature extraction, using multi-sensor (SAR and optical) and large-scale data, have gradually matured, making automated recognition and product generation more feasible [6,7,17,18]. This development has further facilitated the construction of long-term datasets and regional statistical analyses [19].

However, semantic segmentation of ocean internal waves is not a generic visual task, as the objects involved possess highly specific characteristics. In remote sensing imagery, internal waves usually appear as low-contrast and weak-textured structures, while their strips are extremely elongated with vast spatial spans, which imposes stringent requirements on feature representation for pixel-level fine segmentation [20]. In addition, the boundaries of internal waves often merge with the continuously evolving oceanic background clutter, leading to severe boundary blurring and increasing the difficulty of continuous modeling. Furthermore, in certain oceanic regions, multiple internal wave packets can interact and intersect, forming complex interference fringes and overlapping topologies. This physical phenomenon further exacerbates the difficulty of distinguishing individual, continuous wave strips. To alleviate these issues, integrated deep learning frameworks for strip extraction and direction estimation have been developed [21]. Nevertheless, the intrinsic attributes of "weak textures, long strips, and blurry boundaries" remain fundamental challenges for accurate feature extraction and continuous modeling [20].

Before the widespread application of deep learning technologies, automatic recognition and segmentation of ocean internal waves mainly relied on classical digital signal processing theories and handcrafted feature engineering. Traditional detection and recognition frameworks were dominated by manually designed features combined with classifiers [22,23]. One representative line of research exploited the periodicity and directional characteristics of internal wave textures by enhancing features through frequency-domain analysis or directional filtering techniques. For example, Gabor transforms were applied to extract internal wave strip features followed by clustering-based segmentation [24]. Other studies adopted fast Fourier transform (FFT), continuous wavelet transform, and related time–frequency analysis methods for internal wave detection and representation [16]. Another category of methods focused on geometric modeling, employing edge detection operators to identify internal wave boundaries, followed by morphological refinement, connected component analysis, and energy function-based active contour models to extract wave skeletons and envelopes under specific rule constraints. In SAR imagery, where boundaries are weak and the background noise is multiplicative, active contour and variational segmentation models were often used to fit blurred edges as closely as possible [15].

Although these traditional approaches demonstrated effectiveness under specific imaging conditions, such as high signal-to-noise ratios and relatively homogeneous sea states, they heavily relied on manually predefined mathematical models and discriminative criteria. This reliance leads to notable limitations when applied to complex real-world ocean environments. First, internal wave SAR images are often contaminated by speckle

noise and unclear texture details, and traditional enhancement or detection pipelines tend to amplify noise, making them highly sensitive to dynamic background interference and resulting in poor cross-scene generalization ability [25]. Second, localized window-based filtering and edge extraction methods struggle to establish global consistency and semantic-level understanding of long-range structural continuity. With the rapid development of deep neural networks in computer vision, research focus has gradually shifted toward end-to-end semantic segmentation networks with stronger global modeling capabilities, such as Transformer-based segmentation frameworks, which aim to automatically learn more robust feature representations from large-scale remote sensing data [17,26].

End-to-end segmentation frameworks based on deep learning, particularly encoder–decoder architectures represented by U-Net and its variants, have become dominant paradigms for pixel-level fine segmentation. On this basis, techniques such as residual connections, multi-scale feature fusion (e.g., feature pyramid networks), and attention mechanisms have been introduced to enhance joint modeling of multi-scale semantics and fine details [27,28]. Meanwhile, Vision Transformers (ViT) further strengthen long-range context modeling by leveraging self-attention mechanisms, alleviating the limited receptive field of traditional convolutional neural networks and enabling more flexible segmentation framework designs [26,29]. However, the unique properties of internal wave targets, including low contrast, weak textures, and extremely slender linear structures, lead to several specific failure modes when applying these mainstream frameworks. Deep downsampling and feature aggregation often cause excessive feature smoothing and loss of strip details, resulting in blurred boundaries. In densely distributed wave groups or in regions exhibiting complex interference fringes, adjacent strips may be incorrectly fused due to limited background separation. Moreover, insufficient modeling of continuity and topological consistency can cause structural breakage, fragmenting originally continuous internal wave strips. The challenge of maintaining connectivity in slender structures has been widely discussed in topology-preserving segmentation research [30], and similar issues have been observed in internal wave segmentation experiments [20].

Although Transformer-based models offer enhanced global perception, they often incur high computational and memory costs when processing large-scale high-resolution remote sensing imagery, and boundary delineation remains challenging. Consequently, efficient Transformer designs combined with boundary enhancement strategies have been proposed for remote sensing semantic segmentation [31]. More importantly, most fully supervised methods require large amounts of pixel-level annotations, which are expensive and time-consuming to obtain in dynamic marine environments. To alleviate this annotation bottleneck, approaches such as generative segmentation frameworks trained on small datasets for internal wave extraction have been explored [32].

Given the abundance of unlabeled remote sensing imagery and the scarcity of expert annotations, semi-supervised semantic segmentation (SSS) has emerged as a promising solution for internal wave recognition [33]. Existing SSS methods can be broadly categorized into pseudo-labeling and self-training strategies, such as Cross Pseudo Supervision, which iteratively generate and refine their own labels [33]; consistency regularization approaches, exemplified by the Mean Teacher framework and its robust teacher–student variants, which constrain prediction consistency under different perturbations [34,35]; and contrastive representation alignment methods that enhance class separability by optimizing feature-space distributions [36,37]. However, these general SSS methods face significant challenges when applied to the low-contrast and slender structures of internal wave targets. Erroneous pseudo-labels caused by structural defects in weak-texture regions tend to be amplified during iterative training, leading to error self-reinforcement and difficulty in correcting long-range continuity errors [35,38]; excessive reliance on prediction consistency in low-contrast

backgrounds can also result in over-smoothing that sacrifices critical boundary and texture details. Although explicit boundary and geometric constraints have been incorporated into semi-supervised methods to improve edge quality and shape stability [39], existing representation alignment strategies still lack stable class anchors and provide insufficient support for the long-range continuity of slender targets, highlighting the need for more structure-aware semi-supervised frameworks [30].

To address the challenges of stability, detail loss, and structural breakage in semi-supervised internal wave segmentation, this paper proposes a unified semi-supervised approach termed ProtoWave-Mamba (PWM-Net). First, instead of relying solely on noisy pseudo-labels, a class prototype alignment mechanism is introduced, where class centers extracted from labeled data in the feature space serve as global anchors to guide the distribution alignment of unlabeled samples, providing robust semantic consistency constraints during semi-supervised learning. Second, to alleviate weak texture representation and boundary blurring, an in-network, end-to-end trainable wavelet-based frequency-domain enhancement module is designed. Unlike traditional offline filtering that requires manual intervention, this module automatically strengthens multi-scale high-frequency edge information during feature extraction, enabling the model to capture subtle textures and distinguish closely overlapping edges within interference fringes. Finally, to ensure long-range structural continuity of extremely slender wave strips with balanced computational efficiency, the linear state-space model Mamba is incorporated to capture long-distance dependencies across spatial noise and maintain the directional trajectory of intersecting waves in interference areas. The synergistic integration of these components enables robust utilization of unlabeled data, enhanced boundary representation, and preservation of topological integrity in internal wave segmentation. PWM-Net is thus designed as a task-driven solution tailored to the physical topology of ocean internal waves, embedding these domain-specific characteristics directly into the architecture rather than relying on a generic visual backbone.

More importantly, these three components are not independent additions but form a collaborative pipeline in which each module conditions and reinforces the others. The wavelet module first sharpens weak high-frequency cues and suppresses sea clutter in the frequency domain, providing clean local evidence; the Mamba module then propagates this evidence over long spatial ranges to bridge broken segments and maintain the topological continuity of slender bands; and the prototype alignment module finally anchors the refined features to stable class-level semantic centers, preventing pseudo-label noise from accumulating during semi-supervised training while feeding stabilized guidance back to earlier stages. This joint design, rather than a simple concatenation of off-the-shelf components, underpins the robustness of PWM-Net under extremely limited annotation budgets.

The main contributions of this paper are summarized as follows:

- A unified semi-supervised segmentation approach (ProtoWave-Mamba) is proposed for ocean internal wave segmentation under complex marine conditions, integrating class prototype alignment, wavelet-based frequency enhancement, and Mamba-based long-range modeling to effectively address feature smoothing and structural breakage.
- A prototype-guided semi-supervised alignment strategy is designed to enhance conventional pseudo-label learning by introducing feature-space prototype anchors that denoise and stabilize the pseudo-label guidance, enhancing training stability while significantly reducing the reliance on costly pixel-level annotations.
- Extensive experiments on two challenging internal wave remote sensing datasets demonstrate that the proposed method outperforms mainstream semi-supervised models in both segmentation accuracy and topological consistency, with qualitative visualizations providing interpretative insights into boundary recovery in weak-texture regions.

2. Related Work

2.1. Internal Wave Segmentation in Remote Sensing

The automatic recognition and segmentation of oceanic internal waves has long been a focal topic in ocean remote sensing research [5–7]. Early studies can be broadly categorized into two classical approaches. The first category is based on filtering, directional texture analysis, and frequency-domain enhancement, where methods exploit the periodic texture and directional characteristics of internal waves in imagery. These approaches separate target energy peaks using techniques such as Gabor filters [40], wavelet transforms [41], or directional gradient operators [42]. The second category focuses on segmentation methods based on edge detection, energy functions, and spectral features, emphasizing geometric modeling. Representative methods employ edge detectors such as Canny [43] or active contour models [44] to track wavefront envelopes under predefined energy function constraints. Alongside these image-based approaches, altimetric observation of internal solitary waves has advanced rapidly since the launch of SWOT, whose two-dimensional sea surface height measurements allow the amplitude and energy flux of the waves to be quantified rather than only their surface signatures to be delineated [8]. Dedicated studies have since used SWOT sea surface height anomalies to map the spatiotemporal distribution of internal solitary waves in the South China Sea [9], to characterise their refraction and diffraction during interaction with mesoscale eddies off the Amazon shelf [10], and to invert wave amplitudes in the Maluku and Sulu Seas [11]. Because altimetry constrains the internal dynamics of the wave whereas SAR imagery constrains its horizontal footprint, the two observation types are complementary, and satellite datasets are increasingly being coupled with physics-informed learning models to forecast wave propagation [12]. Taken together, these developments mark the transition of internal wave remote sensing from regional case studies toward global-scale mapping [8–12].

Despite their effectiveness under certain conditions, these traditional approaches suffer from notable limitations. They are highly dependent on manual parameter tuning and offline pre-processing [22], exhibit poor generalization capability in complex marine environments such as sea clutter or plankton aggregation [23], and are difficult to integrate into end-to-end learning frameworks that jointly model long-range structural continuity and fine boundary details [45].

With the rapid development of deep learning, encoder–decoder architectures represented by U-Net [27] and vision transformers [29] have significantly advanced semantic feature extraction. However, when confronted with the unique characteristics of internal waves, namely low contrast, weak texture, and extremely elongated linear structures, existing deep supervision models still exhibit several common failure modes. First, excessive feature smoothing occurs, where deep downsampling leads to the loss of weak band features and blurred boundaries [22]. Second, spatial adhesion arises; even with global attention mechanisms [26], adjacent bands are often incorrectly merged due to the absence of explicit modeling of narrow background gaps in densely distributed multi-wave groups [45]. Finally, structural breakage results from insufficient continuity modeling for extremely elongated structures, causing originally connected internal wave bands to fragment and lose their physical topological integrity [22].

Given the high cost of pixel-level annotations in dynamic marine environments, semi-supervised semantic segmentation offers a promising alternative. A dedicated review of SSS methods and their domain-specific limitations for internal wave segmentation is provided in Section 2.5.

2.2. Prototype-Based Learning for Semantic Segmentation

Prototype-based learning has become a powerful paradigm in representation learning, where category prototypes act as semantic centers to facilitate feature alignment and clustering consistency. In semantic segmentation, prototypes are commonly used to represent global category distributions, particularly in domain adaptation and semi-supervised or small-sample scenarios, helping models establish stable semantic decision boundaries in latent embedding spaces. For example, ProDA, proposed by Zhang et al., utilizes category prototypes and prototype distances to denoise pseudo-labels online in the target domain, thereby promoting tighter target structures and improving the stability of domain-adaptive segmentation [46]. Similarly, SPCL, introduced by Xie et al., employs semantic prototypes as anchors for pixel-level contrastive learning, enhancing intra-class compactness and inter-class separability through pixel–prototype contrastive constraints [47].

Despite their advantages, prototype-based methods exhibit limitations in internal wave segmentation scenarios. While prototypes provide stable global semantic references, they are often insensitive to local boundary variations and weak texture details. For fine-grained targets such as internal waves, where the transition between target and background is extremely subtle, simple class-level alignment is insufficient. This necessitates complementary mechanisms, such as wavelet-based frequency-domain enhancement, to strengthen local texture information that may be overlooked by prototype alignment alone. Related works, such as SePiCo, which guide pixel representation learning through semantic concepts and prototype statistics, further highlight the importance of prototype anchors for stabilizing training and improving separability [48]. More recently, Zhang et al. [49] proposed a self-aware and cross-sample prototypical learning framework for semi-supervised medical image segmentation, which adaptively refines prototypes by modeling intra-class feature variations at both pixel and sample levels. This work shares conceptual similarities with our Category Prototype Alignment Module (CPAM) in using prototypes as semantic anchors, but differs in that our approach is specifically designed for the elongated, low-contrast maritime targets in SAR imagery and incorporates a momentum-based prototype update mechanism to handle pseudo-label noise in marine remote sensing.

2.3. Frequency-Domain Enhancement and Wavelet Analysis for Vision

Frequency-domain techniques, particularly the discrete wavelet transform (DWT), have been widely adopted for edge enhancement, texture analysis, and multi-scale information decomposition due to their ability to isolate high-frequency components from background noise. For instance, WaveSNet, proposed by Li and Shen, reformulates DWT and inverse DWT as trainable network layers, enabling end-to-end learning that preserves details during downsampling and recovers them during upsampling, thereby enhancing segmentation performance on edges and textures [50]. In visual perception tasks, wavelet-based modules effectively suppress low-frequency sea clutter while highlighting the steep gradients characteristic of internal wave ripples. Li et al. also provided a systematic review of wavelet-based feature enhancement for remote sensing semantic segmentation, demonstrating that multi-scale frequency-band decomposition can significantly improve high-frequency structural representation and boundary delineation [51]. Building upon this modern paradigm, our proposed framework embeds the wavelet transform as a fully automated, differentiable layer rather than a manual pre-processing step.

However, pure frequency-domain enhancement suffers from inherent semantic blindness. Without strong semantic constraints, frequency enhancement modules may amplify background noise or false edges, leading to false positives. Therefore, frequency-domain enhancement must be stabilized by global semantic anchors, such as class prototypes, and constrained by spatial modeling mechanisms capable of understanding global structural

layouts, such as Mamba-based long-range modeling. In the context of SAR signal processing, Zhang et al. [52] proposed a time–frequency-aware hierarchical feature optimization method for SAR jamming recognition, demonstrating the effectiveness of frequency-domain feature enhancement for SAR-specific signals under complex interference conditions. This insight aligns with our wavelet enhancement design, where multi-scale frequency decomposition is employed to reinforce weak texture information while suppressing background clutter in internal wave imagery. Spectral and directional descriptors are also increasingly used to characterise wave fields statistically over multi-decadal records: a directional wave entropy analysis of a 1979–2023 hindcast along the southwestern coast of Lake Michigan identified a strongly bi-directional wave system with pronounced seasonal modulation and direction-dependent extremes [53]. This reflects a broader trend in which large-domain wave records are processed automatically to extract statistics that single-site or single-scene analyses cannot provide—a trend that equally underlies the interest in automated, large-scale internal wave monitoring pursued in this work.

2.4. Long-Range Modeling with Selective State Space Models

Selective State Space Models (SSMs), represented by Mamba, have recently attracted significant attention for long-sequence modeling due to their linear computational complexity and strong long-range dependency modeling capability. Gu and Dao introduced input-dependent selective state-space modeling in Mamba, enabling selective propagation and forgetting of long-range information with efficient parallel implementation [54]. This property makes Mamba particularly suitable for patch-token sequences derived from high-resolution remote sensing imagery, where modeling the topological continuity of extremely slender internal wave bands over large spatial extents is essential.

In the vision domain, Zhu et al. proposed Vision Mamba (Vim), which incorporates bidirectional state-space modeling into visual representation learning for efficient global context aggregation [55]. Liu et al. further adapted Mamba as a visual backbone by introducing two-dimensional selective scanning to aggregate contextual information along multiple directions, supporting a wide range of visual perception tasks [56]. Nevertheless, the effectiveness of Mamba heavily depends on the quality of input token representations. In internal wave segmentation scenarios, weak boundaries or semantic drift in early features may cause long-range modeling to propagate errors across the entire sequence. Consequently, Mamba requires semantic stabilization from prototype alignment and precise local representations enabled by wavelet-based enhancement.

Beyond sequence modeling, semantic graph reasoning over multi-domain fusion has been applied to radar jamming recognition [57], demonstrating that joint reasoning across heterogeneous feature spaces can improve structural understanding in SAR-related tasks. Similarly, a multi-level feature distribution learning method has been proposed for automatic modulation open-set recognition [58], showing that jointly modeling local fine-grained details and global structural context is crucial for handling complex signal patterns. More recently, a scalable subspace ensemble framework with similarity-based prototype completion has been proposed for radar jamming class-incremental recognition [59], which leverages subspace-specific feature learning and prototype similarity to address the stability–plasticity dilemma in dynamic electromagnetic environments. These insights resonate with our framework's design philosophy, where the Mamba module captures global dependencies while the wavelet module preserves local texture details. Outside ocean remote sensing, the same combination of an efficient sequence operator with semi-supervision has been used to relieve annotation scarcity: a recent study on freshwater mussel detection in turbid water couples a Mamba-based super-resolution reconstruction with a semi-supervised YOLOv10 detector, so that a small set of labelled

acoustic images suffices for reliable detection [60]. This pattern—a Mamba backbone paired with semi-supervision—is closely related to the design rationale of the present work.

2.5. Semi-Supervised Semantic Segmentation for Remote Sensing

SSS aims to reduce annotation costs by leveraging both limited labeled data and abundant unlabeled data. Existing SSS methods can be broadly categorized into three paradigms. The first is pseudo-labeling and self-training, where models iteratively generate and refine their own predictions as training targets, as exemplified by Cross Pseudo Supervision (CPS) [33].

The second is consistency regularization, such as the Mean Teacher framework [34], which enforces prediction consistency under different perturbations to exploit unlabeled data distributions. Among consistency-based approaches, Zhang et al. [61] further proposed an evidential-enhanced tri-branch consistency learning method for medical image segmentation, which leverages uncertainty estimation to improve the reliability of consistency constraints. Although effective in medical imaging settings, such methods assume relatively homogeneous target morphologies, which differ substantially from the slender, low-contrast structures of internal wave targets. The third is contrastive learning and representation alignment, including U2PL [38] and UniMatch [62], which enhance feature discriminability through contrastive constraints or weak-to-strong consistency.

It is important to note, however, that these general-purpose SSS methods were originally developed and validated on medical imaging datasets (e.g., LA, ACDC) or natural image benchmarks (e.g., PASCAL VOC, Cityscapes). The task characteristics, imaging mechanisms, and noise patterns in these domains differ substantially from those in SAR-based ocean internal wave segmentation. Specifically, medical images typically feature compact, well-contrasted anatomical structures with relatively stable acquisition conditions, whereas SAR internal wave images exhibit extremely elongated slender targets, low contrast against background clutter, and multiplicative speckle noise [25].

These domain differences lead to specific failure modes when general SSS methods are applied to internal wave segmentation. Pseudo-labeling methods (CPS, FixMatch) suffer from error amplification in weak-texture regions, where early prediction errors in broken wave strips are reinforced through iterative self-training [33,38]. Consistency regularization (Mean Teacher) tends to produce over-smoothing in low-contrast SAR backgrounds, sacrificing critical boundary details. Contrastive learning methods (U2PL) struggle to establish reliable contrastive pairs for slender structures where positive and negative samples are spatially ambiguous. These observations underscore the need for domain-specific SSS frameworks that explicitly account for the elongated topology, weak boundary signals, and complex background noise of maritime targets, which directly motivates the proposed ProtoWave-Mamba approach.

2.6. Summary and Positioning

Existing semi-supervised segmentation methods rarely address the simultaneous requirements of stable alignment without pseudo-label noise, fine boundary recovery, and long-range structural continuity. Most frameworks excel in one aspect but struggle to balance others. The proposed ProtoWave-Mamba approach aims to bridge this gap by unifying class prototype alignment, wavelet-based frequency enhancement, and Mamba-driven long-range modeling into a collaborative semi-supervised segmentation architecture.

It is also worth clarifying the relationship between our task and structurally adjacent recognition problems studied in recent literature, such as radar jamming recognition and automatic modulation classification. Although these studies share low-level operations with our framework—for example, time–frequency feature enhancement for SAR jamming

recognition [52], semantic graph reasoning over multi-domain fusion [57], multi-level feature distribution learning for modulation open-set recognition [58], and prototype-based incremental recognition for radar jamming [59]—they differ fundamentally from ocean internal wave segmentation in three respects. First, the targets in these tasks (jamming signals and modulated waveforms) generally exhibit compact or spectrally localized structures, whereas internal waves appear as extremely elongated, low-contrast band structures spanning large spatial extents with strong directional continuity. Second, the imaging mechanisms differ substantially: SAR sea-surface imagery is corrupted by multiplicative speckle noise and dynamically evolving background clutter, which is not characteristic of one-dimensional signal waveforms. Third, the annotation bottleneck in marine remote sensing is far more severe, because pixel-level labels of weak, thin wave bands are extremely expensive and ambiguous to produce, which directly motivates the semi-supervised design of PWM-Net. Given these differences, methods developed for these adjacent tasks are not directly transferable to internal wave segmentation; the task-specific adaptation of wavelet enhancement, long-range sequence modeling, and prototype alignment to the physical topology of internal waves constitutes the core contribution of this work.

3. Proposed Approach

To address the instability of semi-supervised segmentation models, the loss of weak texture boundaries, and the fragmentation of elongated internal wave structures in complex oceanic environments, we propose an end-to-end semi-supervised semantic segmentation approach named ProtoWave-Mamba. The proposed approach integrates category prototype alignment, wavelet-based frequency enhancement, and state space sequence modeling to enhance feature representation and preserve structural continuity under limited annotation conditions. Importantly, all these collaborative modules, including the wavelet frequency enhancement, are fully embedded into the network architecture. They operate automatically as end-to-end trainable components without requiring any manual offline preprocessing or human intervention.

Specifically, ProtoWave-Mamba operates within a semi-supervised learning interaction mechanism: The category prototype alignment module establishes global semantic anchors for each category in feature space, guiding both labeled and unlabeled data to align features, significantly alleviating the noise accumulation problem in traditional pseudo-label iteration processes. The wavelet frequency enhancement module uses multi-scale decomposition and reconstruction to strip away the chaotic background signals and actively enhance the edge features and fine details of internal wave strips, allowing the model to detect fuzzy boundaries even with limited annotations. The Mamba sequence modeling module leverages the long-range dependency capture advantage of selective state-space models, aggregating long-distance spatial information in the form of pixel-level token sequences, ensuring the structural continuity and topological integrity of extremely thin strips. These three core modules are deeply integrated in the unified semi-supervised framework through multi-scale skip connections and semantic fusion mechanisms. The decoder gradually restores spatial resolution and integrates multi-level features. ProtoWave-Mamba aims to enhance segmentation performance for weak texture and elongated targets in complex marine scenarios through prototype-guided semantic prediction, frequency-domain-driven detail completion, and sequence modeling-driven structural integration. The overall architecture is shown in Figure 1. In the decoder, the enhanced features are combined by channel-wise concatenation followed by convolutional fusion and upsampled to the input resolution. All feature-fusion operations in the framework take this concatenation-and-convolution form, and the residual paths inside the wavelet and sequence-modeling blocks are additive; consequently no fusion operation is an attention module, and the only

attention-like operators are the channel-wise frequency attention gates inside the wavelet module and the prototype similarity computation in CPAM. The entire network is trained in an end-to-end manner to ensure both feature enhancement and structural integrity in semi-supervised oceanic internal wave segmentation tasks.

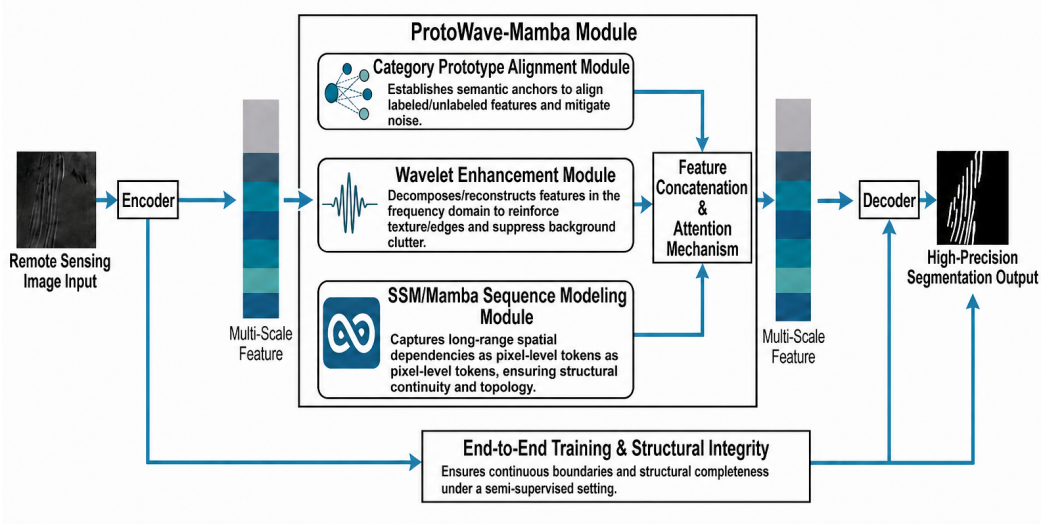


Figure 1. Overview of the ProtoWave-Mamba approach for semi-supervised ocean internal wave segmentation.

The overall semi-supervised training procedure of PWM-Net is summarized as follows. (i) The encoder processes both labeled and unlabeled images to produce multi-scale feature representations. (ii) For unlabeled samples, soft predictions are generated by the decoder, from which pseudo-labels are obtained by taking the class of maximum predicted probability (Equation (2)); pixels whose confidence falls below the threshold $\kappa = 0.85$ are discarded by the confidence-guided filtering strategy. (iii) The CPAM aggregates features of the filtered labeled and high-confidence unlabeled pixels through masked average pooling to estimate class prototypes, which are maintained with a momentum-based moving average (Equation (7)) to suppress batch-specific noise. (iv) A prototype alignment loss pulls pixel features toward their corresponding class prototypes, establishing stable semantic anchors that implicitly denoise the pseudo-label signal. (v) In parallel, the MWTM enhances weak high-frequency cues in the frequency domain, and the Mamba module propagates global context along the spatial sequence to maintain the topological continuity of slender wave bands; a consistency loss between weakly and strongly augmented views (Equation (32)) further regularizes the learning on unlabeled data. All modules are optimized jointly in an end-to-end manner under the composite objective of Equation (33). Through this pipeline, pseudo-labels are not used as differentiable training targets: after confidence filtering, a discrete pseudo-label enters only as a binary indicator mask inside the masked average pooling that estimates the class prototypes, which is the key mechanism that distinguishes PWM-Net from conventional self-training methods.

3.1. Category Prototype Alignment Module

3.1.1. Module Design Motivation

Existing semi-supervised segmentation methods predominantly rely on pixel-level pseudo-labels or prediction consistency. In the low-contrast, weak-texture regime of internal wave imagery, however, erroneous pseudo-labels caused by structural defects tend to be amplified during iterative training, and pure consistency constraints easily over-smooth fragile boundaries. More critically, without stable class-level anchors, the learned feature space drifts across training iterations, which becomes particularly severe when the labeled

set is extremely small (e.g., 10–30 images). To address this issue of insufficient class discriminability and pseudo-label noise accumulation under limited labeled data, this paper proposes the Category Prototype Alignment Module (CPAM). The core idea of this module is to use category prototypes in feature space as "semantic anchors," transforming the traditional pixel classification task into a distance measurement task within feature space. This approach significantly enhances the model's ability to recognize weak internal wave targets under limited labeled samples. The overall framework of CPAM is shown in Figure 2. The module is organised in four stages: it receives the multi-scale feature map $F \in \mathbb{R}^{C\times H\times W}$ together with the ground-truth mask of a labeled sample or the confidence-filtered pseudo-label of an unlabeled one; it estimates the class prototypes $\{P_{\text{iw}}, P_{\text{bg}}\}$ by masked average pooling and maintains them with a momentum-based moving average; it measures the cosine similarity between every pixel feature and every prototype; and it rectifies negative similarities to produce the prototype-aware representation F_{enhanced} that is passed to the decoder.

Although ocean internal waves exhibit substantial visual diversity across different sea regions, imaging conditions, and scales, they share a unified physical topology: elongated band structures with strong directional continuity and characteristic frequency signatures. In a binary segmentation setting (wave vs. background), the intra-class feature variance within the wave class is inherently constrained by these physical priors. The wavelet enhancement module explicitly suppresses scale- and orientation-dependent noise, while the Mamba sequence modeling captures long-range spatial dependencies that are invariant to local appearance variations. Consequently, despite the superficial visual diversity, the feature representations projected by our encoder converge to a relatively compact manifold in the learned feature space, making a single prototype per class a reasonable approximation.

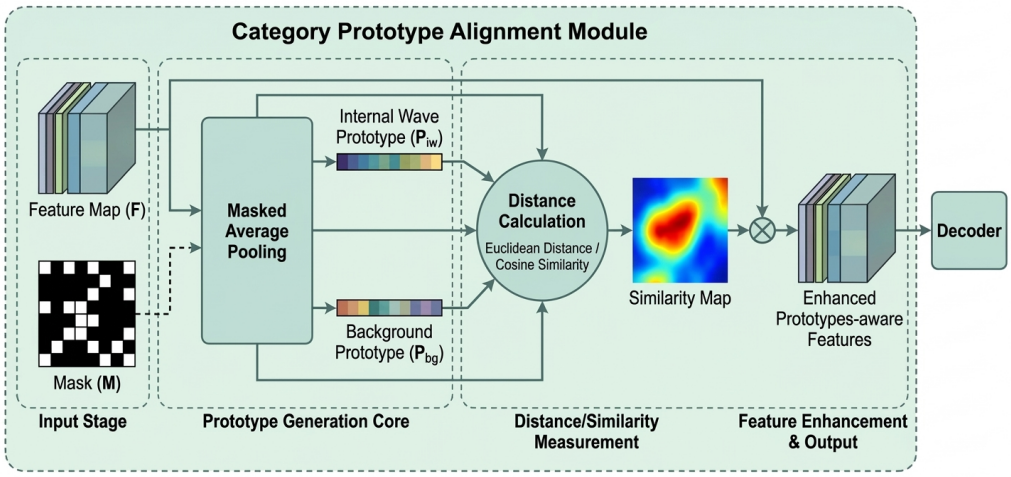


Figure 2. Detailed architecture of the Category Prototype Alignment Module (CPAM).

3.1.2. Module Structure and Implementation Steps

Input Stage

Given an input remote sensing image, the encoder first extracts multi-scale feature representations, which are denoted as

$$F \in \mathbb{R}^{C\times H\times W}, \tag{1}$$

where H and W represent the spatial height and width of the feature map, respectively, and C denotes the number of feature channels. In the semi-supervised learning setting, the module simultaneously receives both labeled and unlabeled samples. For labeled samples,

M corresponds to the ground-truth annotation, while for unlabeled samples, M denotes the pseudo-label generated from the model's prediction. Specifically, for each unlabeled pixel i, the pseudo-label is obtained by taking the class with the maximum predicted probability:

$$M_i = \arg \max_{c \in \{\text{iw,bg}\}} p(c \mid \mathbf{f}_i), \tag{2}$$

where $p(c \mid \mathbf{f}_i)$ is the softmax output of the decoder for class c at pixel i. The corresponding confidence score is defined as $\text{conf}_i = \max_c p(c \mid \mathbf{f}_i)$, which serves as the basis for the subsequent filtering mechanism.

To mitigate the impact of noisy pseudo-labels on prototype quality, a confidence-guided filtering strategy is adopted. Specifically, only pixels whose prediction confidence exceeds a threshold $\kappa = 0.85$ are retained for prototype generation. This value is not a free parameter of the method: it was fixed once by a grid search over $\{0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$ on the validation split (Section 4.1) and is then shared by the prototype filtering here and by the consistency mask of Equation (32), so that a single value governs every confidence-based decision in the framework. Low-confidence pixels are excluded from the masked average pooling operation, preventing erroneous features from contaminating the prototype centroids. The filtered pseudo-labels then guide the subsequent prototype generation process. It is important to clarify that our method does not eliminate pseudo-labels entirely; rather, it introduces prototype-based denoising to mitigate the error accumulation commonly associated with raw pseudo-label iteration, and it never promotes a pseudo-label to a differentiable training target.

Prototype Generation Core

To obtain representative feature prototypes for wave targets and background regions, a masked average pooling (MAP) operation is employed. By applying the mask M to the feature map, the module computes the feature centers corresponding to different semantic categories. The prototype of category c is defined as

$$P_c = \frac{\sum_{i=1}^{N} \mathbb{1}(M_i = c) \cdot \mathbf{f}_i}{\sum_{i=1}^{N} \mathbb{1}(M_i = c)}, \tag{3}$$

where P_c denotes the prototype of category c, with $c \in \{\text{iw}, \text{bg}\}$ representing the internal wave target and background, respectively, $\mathbf{f}_i$ is the feature vector of the i-th pixel, $\mathbb{1}(\cdot)$ is the indicator function that equals 1 when $M_i = c$ and 0 otherwise, and N denotes the total number of pixels in the image.

Through this process, the internal wave prototype P_{iw} and the background prototype P_{bg} are generated. These prototypes serve as semantic anchors for measuring the feature distance between wave targets and background regions, thereby facilitating subsequent distance-based feature alignment and enhancement.

Three implementation details of Equation (3) need to be stated explicitly, because they determine how the prototypes relate to the optimization. First, the prototypes are non-parametric: they are stored as non-trainable buffers of the module and are never registered in the optimizer's parameter list, so no gradient descent is applied to them at any point of the training. Second, the masked average pooling is detached from the backward graph; i.e., the prototype is computed under a stop-gradient operator; this ensures that the prototype alignment loss in Section 3.4 shapes the pixel features rather than the prototype itself. Third, if an image contains no valid pixel of a given class after confidence filtering, the corresponding prototype is simply left unchanged for that iteration instead of being updated with an empty average. These three conventions turn the prototype into a running

statistic of the feature distribution rather than a set of free parameters, which is the reason why prototype quality does not collapse when noisy pseudo-labels are involved.

Distance/Similarity Measurement

To obtain class-specific prototypes, the similarity between each pixel feature and the class prototype is calculated to generate a similarity map. The similarity map reflects the correlation between pixel features and class prototypes, thereby measuring distances in the feature space. In this work, cosine similarity is adopted to measure the similarity between image features and class prototypes, which is formulated as

$$S_{i,c} = \frac{\mathbf{f}_i \cdot P_c}{\|\mathbf{f}_i\| \|P_c\|}, \tag{4}$$

where $S_{i,c}$ represents the similarity between the i-th pixel feature and the prototype of class c, the symbol $\cdot$ denotes the dot product operation, and $\|\cdot\|$ represents the L2 norm. The similarity value is constrained within the range $[-1, 1]$, where values closer to 1 indicate higher semantic similarity between pixel features and class prototypes. Negative similarity values indicate that the pixel feature and the class prototype point in opposite directions in the feature space; i.e., the pixel is highly unlikely to belong to that class; since the subsequent feature enhancement (Equation (5)) multiplies the feature map by the similarity map, such negative weights would invert the feature direction and corrupt the representation. We therefore apply a rectification that clamps negative similarities to zero before the enhancement step, keeping the operation consistent with a soft-assignment interpretation in which each pixel is assigned to a class prototype with a continuous weight in $[0, 1]$. This soft assignment preserves the uncertainty information of ambiguous boundary pixels, in contrast to the hard assignment adopted in the multi-prototype extension (Section Multi-Prototype Extension), where each pixel is assigned to its nearest prototype before the (soft) prototype alignment loss is computed. In addition, Euclidean distance can also be used for distance measurement, and the optimal distance metric can be selected through experimental comparison.

Feature Enhancement and Output

Finally, based on the similarity measurement between pixel features and class prototypes, the features are enhanced and output. For each class $c \in \{\text{iw}, \text{bg}\}$, the enhanced feature representation is calculated as

$$F_{\text{enhanced},c} = F \odot S_c, \tag{5}$$

where the symbol $\odot$ denotes element-wise multiplication, and $S_c \in \mathbb{R}^{H \times W}$ is the cosine similarity map with respect to the prototype of class c, which is broadcast along the channel dimension before multiplication. The class-specific enhanced features are then concatenated along the channel dimension to form the final representation fed into the decoder. This enhancement process strengthens the model's focus on regions with higher semantic similarity and improves the pixel-level feature representation for the final decoder.

3.1.3. Semi-Supervised Learning Mechanism

We further clarify the specific role of pseudo-labels in our framework, because the term refers to two different quantities that enter the optimization in two different ways. The proposed method does not eliminate pseudo-labels. What distinguishes it from conventional self-training is where and in what form they are used. In standard pseudo-labeling methods (e.g., CPS, FixMatch), the noisy prediction of a pixel is promoted to a hard training target for the next iteration, so that early errors are progressively amplified in a self-reinforcing loop.

The first mechanism is hard and acts on the prototype branch. The pseudo-label produced by Equation (2) is a discrete class index, and it is used strictly as a binary indicator mask inside the masked average pooling of Equation (6). Pixels whose confidence does not exceed $\kappa = 0.85$ are discarded before they can contribute. Because this mask carries no gradient, an incorrect pseudo-label can at worst perturb a prototype estimate, and that perturbation is further attenuated by the momentum update of Equation (7), which mixes each batch-wise estimate with a long running average over previous iterations. In this sense the pseudo-label acts as a noisy sampling rule for estimating class statistics, not as a supervision signal.

The second mechanism is soft and acts on the consistency branch. The foreground probability of the teacher view is used directly as a soft regression target in Equation (32), so that the student is asked to reproduce the teacher's soft response rather than a single hard class assignment. No argmax operation is applied on this branch.

The distinction matters, because the hard argmax pseudo-labels used in the prototype loss and the soft probability targets used in the consistency loss are not two competing descriptions of the same operation but two parallel components of the framework. The prototypes are additionally stabilized by (i) confidence-based filtering, which removes unreliable pixels before they contribute to prototype generation, and (ii) the momentum-based moving average, which dilutes the impact of batch-specific noise. The resulting prototypes aggregate information from large numbers of valid pixels across multiple batches and are therefore substantially more robust than the raw pseudo-label of any single forward pass. Consequently, our method is best characterized as "pseudo-label denoising via prototype-guided alignment" rather than "pseudo-label avoidance".

Pseudo-Label Noise Suppression

The two mechanisms described above act at different stages of the pipeline and are complementary: confidence-guided filtering prevents an unreliable pseudo-label from entering the class statistics at all, whereas the momentum-based average attenuates the batch-specific noise that survives the filter. The quantity that ultimately guides the training process is therefore the running prototype estimated from large numbers of validated pixels, not the pseudo-label of any individual image.

Cross-Sample Feature Alignment

The alignment enforced by CPAM is deliberately *cross-sample* rather than image-internal. A prototype P_c is a single centroid shared by every image that contributes to it: for the labeled images it is estimated from the ground-truth masks, and for the unlabeled images it is estimated from the confidence-filtered pseudo-labels of Equation (6); the same centroid is then applied to the features of all images through Equation (31). The training signal therefore does not require a prediction to be consistent within a single image, as consistency regularization does. It requires features belonging to the same class in *different* samples to be brought together around a common anchor, which is how the semantic content of the small labeled set is transferred to the unlabeled data. This also explains why the mechanism tolerates inaccurate individual pseudo-labels: a wrong pseudo-label perturbs one term of an average taken over the whole batch, while the shared prototype continues to be driven by the remaining high-confidence pixels.

Prototype Update Strategy

To further suppress the cumulative effect of erroneous pseudo-labels across training iterations, the prototypes are dynamically maintained via a momentum-based moving average rather than being recomputed from scratch at each step. At each training iteration,

the current-batch prototype estimate is first obtained by masked average pooling restricted to the selected pixels:

$$\hat{P}_c^{(t)} = \frac{\sum_i \mathbb{1}(M_i = c)\,\mathbb{1}(\mathrm{conf}_i > \kappa)\,\mathbf{f}_i}{\sum_i \mathbb{1}(M_i = c)\,\mathbb{1}(\mathrm{conf}_i > \kappa)}, \tag{6}$$

and the running prototype is then updated as:

$$P_c^{(t)} = \alpha_m \cdot P_c^{(t-1)} + (1 - \alpha_m) \cdot \hat{P}_c^{(t)}, \tag{7}$$

where $P_c^{(t-1)}$ is the prototype from the previous iteration, $\hat{P}_c^{(t)}$ is the current-batch estimate obtained from Equation (6) using only the pixels whose confidence exceeds $\kappa = 0.85$, and $\alpha_m = 0.99$ is a momentum coefficient. Both Equation (6) and Equation (7) are evaluated under a stop-gradient operator: the gradient of the prototype alignment loss is propagated to the pixel features $\mathbf{f}_i$ only, and never to the prototypes, which are maintained as non-trainable buffers rather than as learnable parameters. This momentum update ensures smooth prototype evolution, preventing drastic shifts caused by batch-specific noise while allowing gradual adaptation to emerging feature distributions. For the first iteration, prototypes are initialized exclusively from labeled samples to establish reliable semantic anchors before incorporating unlabeled data, so that the warm-up estimate is unaffected by early pseudo-label noise. This strategy effectively decouples the prototype quality from the instantaneous quality of pseudo-labels in any single batch.

Loss Overview

The overall training objective of the proposed framework integrates three complementary loss terms: a supervised segmentation loss, a prototype alignment loss, and a consistency regularization loss. The detailed formulation and weighting strategy are presented in Section 3.4.

Multi-Prototype Extension

To explicitly address potential intra-class diversity, we further extend CPAM with a multi-prototype variant. Instead of maintaining a single centroid per class, we maintain K prototypes $\{P_c^1, P_c^2, \ldots, P_c^K\}$ for each category $c \in \{\mathrm{iw}, \mathrm{bg}\}$. These prototypes are non-parametric buffers exactly as in the single-prototype case and are updated by the same momentum rule; the only difference is that K instead of one running average is maintained per class. During training, each pixel feature $\mathbf{f}_i$ is assigned to the nearest prototype of its own class via:

$$k_i^* = \arg\min_k \|\mathbf{f}_i - P_{y_i}^k\|_2. \tag{8}$$

The intra-class prototype alignment loss is then computed as:

$$\mathcal{L}_{\mathrm{proto}}^{\mathrm{intra}} = -\frac{1}{|\Omega|} \sum_{i \in \Omega} \log \frac{\exp\left(\mathbf{f}_i \cdot \mathrm{sg}(P_{y_i}^{k_i^*})/\tau\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{f}_i \cdot \mathrm{sg}(P_{y_i}^k)/\tau\right)}, \tag{9}$$

where Ω denotes the set of pixels whose class assignment is trusted (ground-truth labels for labeled pixels, and hard argmax pseudo-labels for unlabeled pixels with $\mathrm{conf}_i > \kappa = 0.85$), and $\tau = 0.1$ is the same temperature used in Equation (31). Note that Equation (9) contrasts the pixel only against the K prototypes *of its own class* and is therefore a purely intra-class objective; the inter-class separation is supplied by the main prototype alignment loss of Equation (31). This variant is used only in the dedicated ablation of Section 4.6 and is not part of the default configuration.

This multi-prototype formulation allows the model to capture potential sub-clusters within each class, thereby explicitly addressing intra-class feature diversity; because the

additional prototypes follow the same momentum rule with the same coefficient $\alpha_m = 0.99$, they introduce neither trainable parameters nor additional hyper-parameters.

3.2. Multi-Scale Wavelet Transform Module (MWTM)

3.2.1. Module Design Motivation

Ocean internal waves in remote sensing images often manifest as faint and intricate texture features, with their signals frequently submerged in complex background noise (such as sea clutter). Traditional convolutional operators, limited by their local receptive fields, struggle to effectively separate noise from fine edges in the spatial domain. To address this, this paper introduces a multi-scale wavelet transform module.

Unlike traditional wavelet filtering techniques that require manual parameter tuning or offline image preprocessing, our proposed module is fully embedded within the PWM-Net. Implemented through differentiable tensor operations, it serves as an in-network, end-to-end trainable layer that automatically executes during the forward pass. This module leverages the decoupling capability of wavelet transforms in the time–frequency domain, extracting fine edges and topological structures of the target through multi-level decomposition. It also employs frequency-domain processing techniques to enhance the signal and suppress noise. The detailed architecture is shown in Figure 3.

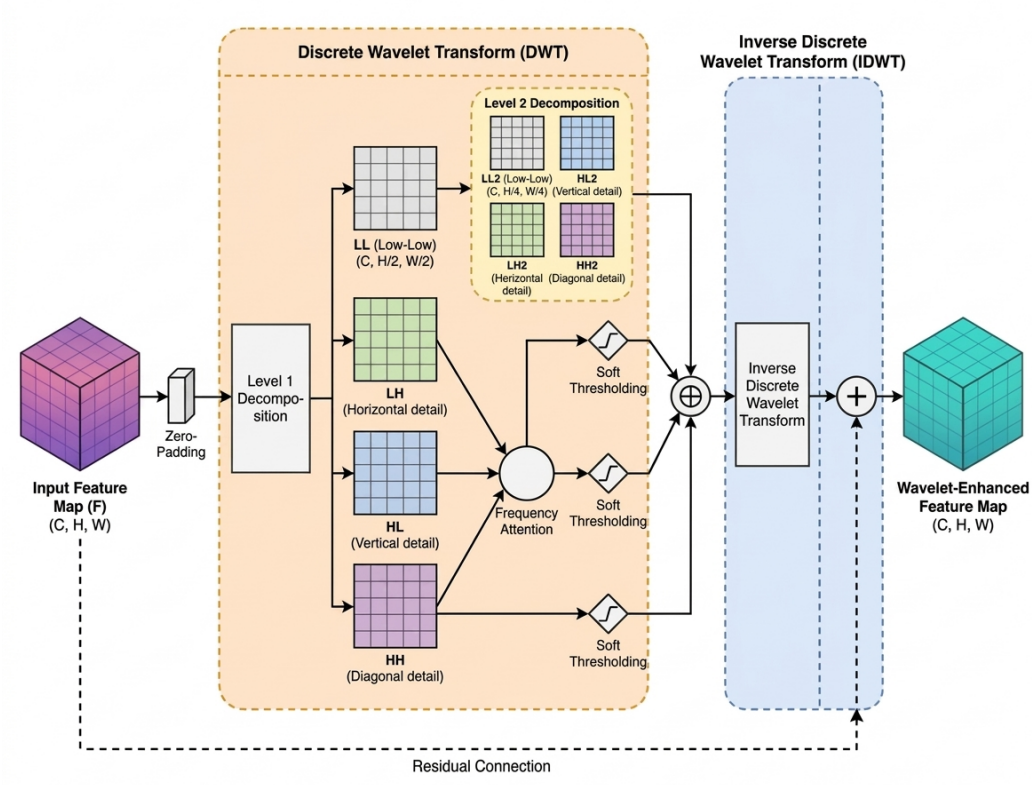


Figure 3. Multi-Scale Wavelet Transform Module architecture.

3.2.2. Module Implementation Flow

Feature Normalization and Discrete Wavelet Transform (DWT)

Given an input feature map $F \in \mathbb{R}^{C \times H \times W}$, its spatial dimensions are first padded so that both H and W become multiples of 4, which is the necessary condition for a two-level decomposition followed by an exact reconstruction. Symmetric (replicate) padding is adopted rather than zero padding, so that the padded rows and columns simply duplicate the nearest valid feature values. This choice avoids the spurious high-frequency responses

that zero padding would create along the feature-map border, which would otherwise be indistinguishable from genuine wavefront edges.

A Level 1 orthonormal Haar DWT is then applied to the padded feature map, decomposing it into four frequency subbands:

$$\{F_{LL1}, F_{LH1}, F_{HL1}, F_{HH1}\} = \mathrm{DWT}(F), \tag{10}$$

where F_{LL1} represents the low-frequency subband that preserves the coarse approximation information of the feature map, and $\{F_{LH1}, F_{HL1}, F_{HH1}\}$ denote the high-frequency subbands corresponding to horizontal, vertical, and diagonal directional details, respectively.

To capture deeper and more distinctive multi-scale features, a Level 2 decomposition is further applied to F_{LL1} using the same Haar filters:

$$\{F_{LL2}, F_{LH2}, F_{HL2}, F_{HH2}\} = \mathrm{DWT}(F_{LL1}), \tag{11}$$

where F_{LL2} is the finer low-frequency subband and $\{F_{LH2}, F_{HL2}, F_{HH2}\}$ are the corresponding level-2 directional detail subbands. After the two-level decomposition the module therefore holds seven subbands in total: the deepest approximation F_{LL2} plus six detail subbands, $\{F_{LH1}, F_{HL1}, F_{HH1}\}$ at level 1 and $\{F_{LH2}, F_{HL2}, F_{HH2}\}$ at level 2. These subbands provide complementary scales for describing both the sharp local gradients and the broader structural layout of internal wave strips.

Implementation-wise, the Haar analysis filters are realized as fixed (non-learnable) depth-wise convolution kernels of stride 2, so that the DWT is an exactly differentiable linear operator embedded in the network. Padding, decomposition, and the subsequent reconstruction are all executed inside the forward pass under full numerical precision, which keeps the round-trip error of the module negligible and requires no offline preprocessing or manual parameter tuning.

Frequency Attention

To recover detailed substructures of internal waves under noisy backgrounds, a collaborative frequency-domain optimization mechanism is introduced. It consists of two components applied sequentially at each decomposition level: a frequency attention that re-weights the directional detail subbands, and a learnable soft thresholding that shrinks the residual noise. Attention is applied first, so that the thresholding operates on the attention-weighted responses.

For a given decomposition level $l \in \{1, 2\}$, the three detail subbands of that level are stacked along the channel dimension to form a compact tensor

$$Z_l = \mathrm{Concat}\left(F_{LH_l}, F_{HL_l}, F_{HH_l}\right) \in \mathbb{R}^{3C \times H_l \times W_l}, \tag{12}$$

where H_l and W_l are the spatial dimensions at level l. The three detail subbands of one level are strongly correlated: they originate from the same separable filtering along the two spatial axes, so their directional energies are coupled, and clutter amplified in one orientation frequently leaks into the other two. Treating them jointly, rather than independently, therefore allows the module to suppress orientation-coupled clutter while keeping the directional signatures of internal wave fronts.

One frequency attention branch is instantiated per level, giving two branches in total. Each branch computes a channel-wise attention map by global average pooling followed by a two-layer bottleneck MLP with reduction ratio $r = 2$ and a sigmoid gate:

$$A_l = \sigma\Big(W_l^2\, \phi\big(W_l^1\, \mathrm{GAP}(Z_l)\big)\Big) \in \mathbb{R}^{3C}, \tag{13}$$

where GAP$(\cdot)$ denotes global average pooling, $\phi(\cdot)$ is the ReLU activation, $\sigma(\cdot)$ is the sigmoid function, and $W_l^1 \in \mathbb{R}^{3C/r \times 3C}$ and $W_l^2 \in \mathbb{R}^{3C \times 3C/r}$ are the weights of the bottleneck MLP. The attention weights are broadcast along the spatial dimensions and modulate the subband responses element-wise:

$$\tilde{Z}_l = A_l \odot Z_l. \tag{14}$$

The modulated tensor $\tilde{Z}_l$ is then split back into three subbands and passed to the learnable soft thresholding stage. The deepest approximation subband F_{LL2} is left untouched by both operations, because it carries the coarse structural information of the wave strips and contains little of the high-frequency clutter that these two components are designed to remove.

Learnable Soft Thresholding

Subsequently, a soft-thresholding operation is applied to the attention-weighted detail subbands to further suppress noise and strengthen key structural components. The soft-thresholding function is defined as

$$f_{thr}(x) = \mathrm{sgn}(x) \cdot \max(|x| - \rho, 0), \tag{15}$$

where x denotes an element of a detail subband and ρ is a learnable threshold. Here ρ is used instead of λ deliberately, because λ already denotes the Dice and cross-entropy weights in Equation (30) and the two quantities are unrelated.

Rather than sharing one global threshold, the module maintains an independent threshold for every detail subband. With three detail subbands at each of the two levels, this yields six learnable thresholds $\{\rho_{LH1}, \rho_{HL1}, \rho_{HH1}, \rho_{LH2}, \rho_{HL2}, \rho_{HH2}\}$. Independent thresholds are necessary here because the two levels characterize different spatial scales of the wave strips: the level-1 subbands concentrate on the fine ripple texture of the wavefronts, whose amplitude is intrinsically low and must be preserved, whereas the level-2 subbands respond to coarser sea-clutter structures of considerably larger amplitude that can tolerate stronger shrinkage.

To keep every threshold non-negative and to avoid the degenerate solution in which a threshold collapses to zero (thereby silently disabling the denoising), each threshold is re-parameterized through a softplus function of an unconstrained learnable scalar:

$$\rho_j = \mathrm{softplus}(\beta_j) = \ln\left(1 + e^{\beta_j}\right), \qquad j \in \{LH1, HL1, HH1, LH2, HL2, HH2\}, \tag{16}$$

where β_j is the unconstrained parameter optimized by gradient descent. All six offsets are initialized to the same value $\beta_j^{(0)} = \mathrm{softplus}^{-1}(0.01) = \ln(e^{0.01} - 1) \approx -4.60$, so that $\rho_j^{(0)} = 0.01$ and the module starts from an almost identity mapping; the shrinkage then increases only when this reduces the training objective. Because Equation (15) is subdifferentiable everywhere and has a well-defined gradient with respect to both x and ρ, namely $\partial f_{thr}/\partial \rho = -\mathrm{sgn}(x)\mathbb{1}(|x| > \rho)$, the thresholds are learned jointly with the rest of the network in a single end-to-end optimization.

IDWT and Residual Connection

Because the decomposition is hierarchical, the reconstruction is also hierarchical and requires two successive inverse discrete wavelet transforms (IDWT), matching the two forward decomposition levels. First, the enhanced level-2 detail subbands are recombined with the untouched approximation subband F_{LL2} to recover the level-1 low-frequency subband:

$$\hat{F}_{LL1} = \mathrm{IDWT}\left(F_{LL2}, \tilde{F}_{LH2}, \tilde{F}_{HL2}, \tilde{F}_{HH2}\right), \tag{17}$$

where $\tilde{F}_{\cdot 2}$ denotes the attention-modulated and thresholded level-2 detail subbands. Then $\hat{F}_{LL1}$ is recombined with the enhanced level-1 detail subbands to reconstruct the feature map in the original spatial domain:

$$\tilde{F}_{wavelet} = \mathrm{IDWT}\left(\hat{F}_{LL1}, \tilde{F}_{LH1}, \tilde{F}_{HL1}, \tilde{F}_{HH1}\right). \tag{18}$$

The synthesis filters of both IDWT stages are the transposed counterparts of the fixed Haar analysis kernels used in Equations (10) and (11), and are likewise implemented as differentiable operations. After the second reconstruction, the feature map is cropped back to the original size $H \times W$ in order to remove the padding introduced at the beginning of the module.

To prevent information loss and potential boundary artifacts caused by multi-level decomposition, a residual connection is introduced by directly combining the reconstructed feature map with the original input feature:

$$F_{output} = F + \tilde{F}_{wavelet}. \tag{19}$$

This design ensures that the original signal information is preserved while fine-grained features are enhanced through wavelet-based processing: the low-frequency content flows along the identity path, whereas the two-level detail branches contribute only the selectively amplified high-frequency residual.

3.2.3. Semantic Enhancement and Noise Suppression Balance

The multi-scale wavelet transform module utilizes a "divide and conquer" strategy in the frequency domain, enabling the model to effectively address the low-contrast challenges of internal wave images. Compared to single-domain convolution, this module ensures the structural continuity of internal wave bands through the LL component, improves the perception accuracy of elongated boundaries through the LH, HL, and HH components, and suppresses background noise via a soft-threshold mechanism. This multi-dimensional, multi-scale feature fusion ensures highly sensitive extraction of ocean internal waves in complex deep-sea environments.

3.3. Mamba Sequence Modeling Module

3.3.1. Module Design Motivation

Ocean internal waves exhibit strong directional and cross-regional topological continuity in their spatial distribution, often forming elongated band structures spanning tens of kilometers. Traditional Convolutional Neural Networks (CNNs), limited by their local receptive fields, struggle to capture these long-range dependencies, often leading to segmentation results that are fragmented. On the other hand, traditional Transformer architectures face quadratic computational costs when processing high-resolution remote sensing features. To address this, this paper introduces the Mamba sequence modeling module, based on state space models (SSM). This module, with its linear complexity in sequence scanning, efficiently establishes global contextual relationships for internal wave targets, enhancing the model's ability to perceive elongated textures. Its overall framework is shown in Figure 4.

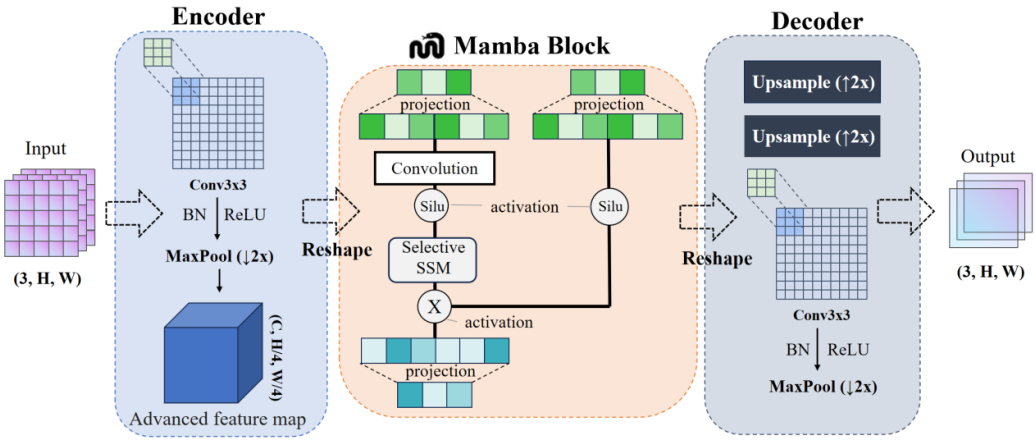


Figure 4. Architecture of the lightweight encoder–decoder segmentation network based on Mamba.

3.3.2. Module Architecture and Implementation Workflow

The lightweight segmentation network is composed of three core stages, including an encoder, a Mamba block, and a decoder.

Encoder Stage: Feature Extraction

Given an input image $I \in \mathbb{R}^{3\times H\times W}$, the encoder is employed to perform hierarchical semantic modeling. The three input channels are not the RGB channels of an optical image: they are the co-polarized backscatter, the cross-polarized backscatter, and their difference, all derived from the same Sentinel-1 SAR scene. Their construction and the accompanying normalization are described in detail in Section 4.1.

Each encoding stage consists of a 3×3 convolutional layer (Conv3×3), followed by a batch normalization (BN) layer and a nonlinear activation function (ReLU or SiLU). After convolution, spatial downsampling is conducted using max pooling (MaxPool, $\downarrow 2\times$). The channel width doubles at each stage following the schedule $32 \rightarrow 64 \rightarrow 128$, so that the feature map entering the sequence modeling stage is $C = 128$ channels wide.

After two successive downsampling operations, the input image is transformed into an advanced semantic feature map

$$F_{\text{adv}} \in \mathbb{R}^{C\times(H/4)\times(W/4)}, \tag{20}$$

This process aims to suppress small-scale noise in the original image while compressing the spatial resolution, thereby preparing the features for subsequent sequence modeling.

Mamba Block: Sequential Feature Enhancement

The Mamba block serves as the core modeling component of the network. First, the feature map $F_{\text{adv}} \in \mathbb{R}^{C\times(H/4)\times(W/4)}$ is flattened from a two-dimensional spatial tensor into a one-dimensional feature sequence.

Because a state space model is a causal operator, its output depends on the order in which the spatial positions are visited, and the scanning order is therefore a design parameter rather than an implementation detail. A single raster scan would impose an anisotropic bias: information would propagate freely from the top-left towards the bottom-right of the feature map but only indirectly in the opposite direction. This is inappropriate here, because the orientation of internal wave strips is not known in advance and can be arbitrary with respect to the image axes. The module therefore performs a four-direction scan over the same feature map. Let flatten$(\cdot)$ denote row-major raster flattening of a spatial tensor; the four input sequences are

$$\begin{aligned} X_{\rightarrow} &= \mathrm{flatten}(F_{\mathrm{adv}}), & X_{\leftarrow} &= \mathrm{flatten}\big(\mathrm{Rot}_{180}(F_{\mathrm{adv}})\big), \\ X_{\downarrow} &= \mathrm{flatten}\big(F_{\mathrm{adv}}^{\top}\big), & X_{\uparrow} &= \mathrm{flatten}\big(\mathrm{Rot}_{180}(F_{\mathrm{adv}}^{\top})\big), \end{aligned} \tag{21}$$

where $F_{\mathrm{adv}}^{\top}$ denotes transposition of the two spatial dimensions and $\mathrm{Rot}_{180}(\cdot)$ denotes a 180° in-plane rotation. The fourth sequence $X_{\uparrow}$ is thus the reverse of the transposed sequence, so that the four scans sweep the feature map in both directions along each of the two image axes, and no orientation of the wave strips is favoured a priori. Each sequence is processed by its own selective state space recursion, and the four outputs are inverted by the corresponding inverse transformation and merged by summation:

$$F_{\mathrm{mamba}} = \mathrm{unflat}\big(\hat{Y}_{\rightarrow}\big) + \mathrm{Rot}_{180}\Big(\mathrm{unflat}\big(\hat{Y}_{\leftarrow}\big)\Big) + \mathrm{unflat}\big(\hat{Y}_{\downarrow}\big)^{\top} + \Big(\mathrm{Rot}_{180}\big(\mathrm{unflat}(\hat{Y}_{\uparrow})\big)\Big)^{\top}, \tag{22}$$

where $\hat{Y}_{\bullet}$ denotes the state space output of the corresponding sequence and $\mathrm{unflat}(\cdot)$ restores the two-dimensional spatial layout. It should be noted explicitly that the four scan branches share the same parameters—the input projections, the input-dependent **B** and **C** generators, and the output projection—and differ only in their scanning orders. The summation in Equation (22) therefore merges four directional views of the same feature map without increasing the parameter count of the module.

The input sequence is then expanded through linear projection and processed in parallel by a dual-branch structure.

In the selective scan branch, a one-dimensional convolution is first applied to capture local sequential dependencies, followed by a SiLU activation function. The features are then fed into a selective state space model (Selective SSM), whose parameters are dynamically adjusted according to the input. The Selective SSM acts as a scanning filter to selectively suppress irrelevant background signals while extracting shared structural patterns of internal wave bands. Within this branch the sequence is processed by the four direction-specific recursions of Equation (21), and their outputs are merged by the summation of Equation (22) before the sequence is restored to its spatial layout. Because the four scans operate on the same sequence and share their parameters, this step changes the effective receptive field of the branch rather than its capacity.

For information gated fusion, the output of the Selective SSM is fused with the activated auxiliary branch through element-wise multiplication (denoted as $\otimes$), enabling selective enhancement of informative features. Finally, a projection layer compresses the feature dimension back to its original size, completing the modeling of long-range dependencies within the sequence.

Decoder Stage: Feature Reconstruction

The enhanced sequential features are reshaped back into the spatial domain to restore the original topological structure. The decoder adopts a symmetric two-stage upsampling architecture (Upsample, $\uparrow 2\times$), combined with 3×3 convolutional layers to progressively recover spatial resolution.

During feature reconstruction, shallow fine-grained features from the encoder are fused to compensate for spatial information loss caused by downsampling. The final output is a probability map of size $(2, H, W)$, whose two channels correspond to the internal wave class and the background class, $\{\mathrm{iw}, \mathrm{bg}\}$, matching the class set used by the pseudo-label definition in Equation (2) and by the prototype loss in Equation (31). In the released implementation this two-class output is parameterised by a single foreground logit, z, so that $p(\mathrm{iw} \mid \mathbf{f}) = \sigma(z)$ and $p(\mathrm{bg} \mid \mathbf{f}) = 1 - \sigma(z)$; the two forms are mathematically equivalent, and the single-logit parameterisation is the one used by the consistency loss of Equation (32). This output is used to achieve high-precision binary segmentation of ocean internal waves.

3.3.3. Mathematical Formulation and Modeling Principle

The core principle of the Mamba module is to map discrete feature sequences into a continuous state space. The continuous state evolution system is formulated as

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{23}$$

$$y(t) = \mathbf{C}h(t) + \mathbf{D}x(t), \tag{24}$$

where $x(t)$ denotes the input internal wave feature sequence, $h(t)$ represents the hidden state that simulates long-term memory, and $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ correspond to the state transition matrix, input matrix, and output matrix, respectively. The diagonal matrix $\mathbf{D}$ is a feed-through term that lets the current input reach the output without passing through the recurrence. Its presence matters for the present task, because it preserves a direct and unattenuated path for thin wavefront features: a weak but locally distinct edge is not required to survive the full state recursion before it can influence the prediction.

Following the selective state space formulation, the input-dependent parameters are generated from the input itself by linear projections:

$$\mathbf{B} = W_B x, \qquad \mathbf{C} = W_C x, \qquad \Delta = \mathrm{softplus}(W_\Delta x + b_\Delta), \tag{25}$$

where $W_\bullet$ and b_Δ are learnable weights and bias, and Δ is the step size. The softplus function ensures that Δ remains strictly positive, which is required for a stable discretization: a non-positive step size would make $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ non-contractive and the recursion unstable.

To enable efficient computation within neural networks with discrete weights, the step size parameter Δ is used to discretize the continuous system. The discretized parameters are given by

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}), \tag{26}$$

$$\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \tag{27}$$

Because $\mathbf{A}$ is diagonal and Δ is small, Equation (27) should not be read as an exact algebraic reduction of the continuous system; it is the standard first-order zero-order-hold approximation adopted in practice, and the two expressions coincide to first order in Δ whenever $\Delta|\mathbf{A}| \ll 1$. This is a deliberate design choice inherited from the original Mamba formulation, and it is the form actually implemented in our network.

Based on the discretized parameters, the discrete state update equations are defined as

$$h_k = \bar{\mathbf{A}}h_{k-1} + \bar{\mathbf{B}}x_k, \tag{28}$$

$$y_k = \mathbf{C}h_k + \mathbf{D}x_k. \tag{29}$$

Different from conventional state space models (SSMs), in which $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are constants fixed after training, the Mamba formulation generates $\mathbf{B}$, $\mathbf{C}$, and Δ dynamically from the input through Equation (25). This is what makes the recursion selective: the amount of information that is written into the state and read out of it becomes a function of the local content rather than a constant of the layer. In our setting this selectivity allows the model to prolong the memory of a coherent wave band and to shorten it when the sequence crosses a region of unstructured sea clutter. For the four-direction configuration of Equation (21), the direction-specific recursions are evaluated with a chunked parallel associative scan, so that the four directional passes are processed in parallel rather than strictly one after another. Consequently, the module attains a global receptive field with a computational complexity that grows as $O(L)$ in the sequence length L, in contrast to the quadratic dependence of self-attention.

3.3.4. Global Perception and Topology Preservation

The Mamba module provides a global receptive field beyond that of traditional convolution kernels. For cross-scale geophysical targets such as ocean internal waves, the selective scanning mechanism establishes texture continuity among wave bands, ensuring the integrity of macroscopic structures in segmentation predictions. Combined with the multi-channel feature fusion strategy in the decoder, the network can effectively distinguish internal wave boundaries from visually similar sea surface textures, significantly improving robustness in low signal-to-noise ratio environments.

3.4. Optimization Objective and Joint Training

The proposed PWM-Net is optimized end-to-end with a composite loss function that jointly supervises the three core modules:

3.4.1. Segmentation Loss

The primary output of the Mamba decoder is supervised by a combination of Dice loss and binary Cross-Entropy (CE) loss:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{Dice}}(P, G) + \lambda_{\text{ce}} \cdot \mathcal{L}_{\text{CE}}(P, G), \tag{30}$$

where P denotes the predicted probability map and G the ground-truth binary mask.

3.4.2. Prototype Alignment Loss

To enforce semantic consistency in the feature space, a prototype alignment loss is applied on the features produced by the wavelet-enhanced and Mamba-encoded branches. The loss is evaluated only on pixels that carry a trustworthy class assignment: the labeled pixels, which carry their ground-truth label, and the unlabeled pixels whose prediction confidence exceeds $\kappa = 0.85$, which carry the hard argmax pseudo-label of Equation (2). Let Ω denote this pixel set; the loss is defined as

$$\mathcal{L}_{\text{proto}} = -\frac{1}{|\Omega|} \sum_{i \in \Omega} \log \frac{\exp\left(\mathbf{f}_i \cdot \text{sg}(P_{y_i})/\tau\right)}{\sum_{c \in \{\text{iw,bg}\}} \exp\left(\mathbf{f}_i \cdot \text{sg}(P_c)/\tau\right)}, \tag{31}$$

where $\mathbf{f}_i$ is the pixel-level feature vector, P_c is the prototype of category c obtained by the masked average pooling of Equation (6), y_i is the class assignment of pixel i, $\tau = 0.1$ is a temperature hyper-parameter, and $\text{sg}(\cdot)$ denotes the stop-gradient operator. The prototypes are therefore treated as fixed reference points within a training step: $\text{sg}(\cdot)$ blocks any gradient from reaching them, so the loss shapes the pixel features only, while the prototypes are updated exclusively by the momentum rule of Equation (7). This loss pulls features toward their corresponding class prototype while pushing them away from the opposite class prototype, and it is the mechanism by which the prototypes act as semantic anchors for the unlabeled data.

3.4.3. Consistency Regularization Loss (For Semi-Supervised Setting)

Following the mainstream semi-supervised paradigm, a consistency loss is applied between the predictions from the weakly augmented and strongly augmented views of the same input:

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} (p_i^{\text{s}} - p_i^{\text{w}})^2, \qquad \Omega = \{\, i : \max(p_i^{\text{w}}, 1 - p_i^{\text{w}}) > \kappa \,\}, \tag{32}$$

where p_i^{w} and p_i^{s} denote the foreground probabilities of the teacher (weakly augmented) and student (strongly augmented) views at pixel i, respectively, and $\kappa = 0.85$ is the same confidence threshold used for prototype filtering in Equation (6). The regression target is the foreground

probability of the teacher view rather than its argmax, so that the student reproduces the soft response of the teacher instead of a hard class assignment; the set Ω is defined by a hard confidence mask and $|\Omega|$ is its cardinality, so pixels below κ are removed from the average rather than down-weighted. The weight of this term is held at $\lambda_u = 0.5$ throughout training.

Soft Targets Versus Hard Masks

The two semi-supervised mechanisms of this work deliberately use two different forms of the network's own predictions—a soft probability target on the consistency branch (Equation (32)) and a stop-gradient indicator mask on the prototype branch (Equation (6))—and they should not be confused with each other. This distinction has been stated in full in Section 3.1.3 and is not repeated here.

3.4.4. Total Loss

The overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_s \cdot \mathcal{L}_{\text{seg}} + \lambda_p \cdot \mathcal{L}_{\text{proto}} + \lambda_u \cdot \mathcal{L}_{\text{cons}}, \tag{33}$$

where λ_s, λ_p, and λ_u are weighting coefficients balancing the contribution of each module. The symbol α_m is deliberately not reused here for a loss weight, because it already denotes the prototype momentum coefficient in Equation (7). In our experiments we set $\lambda_s = 1.0$, $\lambda_p = 0.1$, and $\lambda_u = 0.5$; these three weights were fixed before the test evaluation and their effect is examined in the loss-level ablation of Table 17.

4. Experiments and Analysis

4.1. Datasets

To evaluate the adaptability and robustness of the proposed ProtoWave-Mamba approach across different ocean regions and complex environments, two representative SAR datasets of ocean internal waves were selected. These datasets are complementary in terms of temporal coverage, spatial extent, and annotation schemes, providing strong support for performance evaluation in small-sample and label-scarce scenarios.

Throughout this paper, *scene* denotes one Sentinel-1 acquisition and patch denotes one spatial sample fed to the network; unless stated otherwise, a patch is obtained by cropping a scene to the network input size.

Dataset 1 (S1-IW-2023) [63]: This dataset contains 742 ocean internal wave patches acquired by Sentinel-1 A/B in Interferometric Wide swath (IW) mode between June 2014 and February 2023. The dataset was deliberately assembled by retrieving internal wave scenes from multiple sea areas worldwide rather than from a single region, and it is therefore distributed without any designated study area. The images are provided in GeoTIFF format, with high spatial resolution and clearly distinguishable internal wave textures, and the dataset is distributed together with internal wave annotations.

Dataset 2 (IWs Dataset v1.0) [64]: The IWs Dataset v1.0 consists of 390 Sentinel-1 SAR scenes acquired between April 2014 and December 2021 over four internal wave-prone seas in south-east Asia, namely the Andaman Sea, the South China Sea, the Sulu Sea and the Celebes Sea. The dataset exhibits a long temporal span and a wide geographic distribution. Each scene is accompanied by corresponding internal wave labels, making it suitable for evaluating the behaviour of models under diverse noise backgrounds and sampling compositions.

The two collections differ in construction rather than in location alone: Dataset 1 is a global, multi-region collection without a designated study area, whereas Dataset 2 is a regional collection spanning four named seas. The two are not fully disjoint in geography, since the South China Sea contributes to both. Throughout this paper, the term cross-dataset

therefore denotes a change in sampling composition, imaging scale and background clutter rather than a strictly cross-region transfer.

Dataset Composition and Screening

Our experiments build on two publicly available Sentinel-1 internal-wave datasets that differ both in coverage and in annotation origin. The first, S1-IW-2023 [63], is a global collection that aggregates SAR acquisitions from many sea areas without being restricted to any single region; it contains 742 patches of 1024 × 1024 pixels cropped from 457 Sentinel-1 scenes, together with 1039 internal-wave annotations. The second, the IWs Dataset v1.0 [64], concentrates on four well-known internal-wave seas—the Andaman Sea, the South China Sea, the Sulu Sea and the Celebes Sea—and contains 390 Sentinel-1 scenes with 775 annotations. Both collections were originally released with detection-level labels (bounding boxes in PASCAL VOC/COCO and YOLO format, respectively) rather than pixel-level masks; the conversion into the segmentation ground truth used here is described in the next paragraph.

Because the two collections differ considerably in acquisition quality and in annotation density, we apply a two-stage screening protocol. In the first stage we discard patches that cannot support a boundary-level segmentation task: (i) those in which no unambiguous internal-wave signature survives visual inspection; (ii) those whose bright–dark band pair is severely degraded by rain cells, high wind speed or strong speckle, so that the wave front cannot be delineated consistently; and (iii) those whose original detection label is inconsistent with the imaged wave packet, for example when the box covers ship wakes, slick-like features or land. This stage removes a much larger fraction of the first dataset (392 of 742 patches, 52.8%), which is expected for a global collection that inevitably includes many acquisitions of marginal quality, while it removes only 40 of the 390 scenes of the second dataset, which is already curated and confined to internal-wave-rich seas. In the second stage we keep 350 patches from each source, so that the two datasets contribute equally to training and evaluation and neither collection dominates the optimisation or the cross-dataset comparison. The screening statistics are summarised in Table 1 and the resulting composition in Table 2.

Table 1. Screening statistics of the two source datasets. "Removed (Stage 1)" reports the number of patches (Dataset 1) or scenes (Dataset 2) discarded by the quality criteria described in Section 4.1, together with the corresponding share of the original collection.

Source Dataset	Original Size	Removed (Stage 1)	Removed (%)	Retained (Stage 2)
S1-IW-2023 [63] (global, multi-sea)	742	392	52.8	350
IWs Dataset v1.0 [64] (four seas)	390	40	10.3	350
Total	1132	432	38.2	700

Table 2. Summary of the two collections used in this study. Dataset 1 was assembled from multiple sea areas worldwide and has no designated study region; Dataset 2 is regionally focused on four named seas. The two collections are not fully disjoint in geography, since the South China Sea contributes to both.

Property	Dataset 1 (S1-IW-2023)	Dataset 2 (IWs Dataset v1.0)
Reference	[63]	[64]
Coverage	Global, multiple sea areas (no designated region)	Andaman Sea, South China Sea, Sulu Sea, Celebes Sea
Sensor and mode	Sentinel-1 A/B, IW mode (Stripmap scenes excluded in this work)	Sentinel-1, IW mode (Stripmap scenes excluded in this work)
Acquisition period	June 2014–February 2023	April 2014–December 2021
Released scale	742 patches	390 scenes
Used in this work	30 labeled + 200 unlabeled + 20 validation + 100 test	30 labeled + 200 unlabeled + 20 validation + 100 test
Input used here	256 × 256 patches	256 × 256 patches

Pixel-Level Mask Generation and Quality Control

Both source datasets provide only detection-level annotations, so the pixel-level ground truth required by a segmentation model had to be constructed explicitly. For every retained patch the original bounding box was converted into a pixel-level mask by manual delineation inside the box: the box acts as a region of interest, and the left–right envelope of the alternating bright–dark band pair associated with the internal-wave packet is traced manually with polygon tools in CVAT. Annotators were instructed (i) to enclose the complete wave-packet envelope rather than the individual crests, (ii) to exclude ship wakes, oil slicks, oceanic fronts, land and speckle-only stripes, and (iii) to keep the delineated band at least three pixels wide so that the boundary remains well defined after resizing. All images were annotated independently by two team members with more than two years of experience in SAR internal-wave interpretation, and conflicts were adjudicated by a third, senior annotator. To quantify the reliability of the protocol, a randomly selected 15% subset was re-annotated: the two annotators reached a mean Dice similarity of 0.93 and a boundary F1 of 0.89 within a two-pixel tolerance. Each mask was additionally validated against its source box and accepted only if it lies inside the box and covers at least 80% of the box area, which flags both missed wave packets and the accidental delineation of unrelated structures; a final visual pass was carried out over the complete mask set. The resulting labels are binary (internal-wave/non-internal-wave); we note that the internal-wave "boundary" is by nature an envelope rather than a physical interface, and the protocol above fixes that envelope consistently for all images. The derived masks and the exact subset identifiers are available from the corresponding author upon reasonable request; the two source collections from which they were derived remain publicly available at their original repositories (Section 4.1).

Split Protocol and Scene-Level Disjointness

The partition is performed at the *scene* level rather than at the level of individual image patches. Before the four subsets are drawn, every patch is grouped by its Sentinel-1 acquisition identifier—product name, acquisition date and relative orbit—and, where applicable, by the internal-wave packet to which it belongs, and each group is assigned as a whole to exactly one subset. Consequently, no patch that shares a scene, an acquisition or a wave packet with a training sample can appear in the validation or the test subset, and the correlation that a patch-level random split would introduce is structurally excluded. The assignment is stratified by source collection and uses a fixed random seed; it yields, for each collection, 30 labeled and 200 unlabeled training patches, 20 validation patches and 100 test patches, so that 350 patches per collection are retained, as summarised in Table 3. The same scene-level grouping is applied when the labeled, unlabeled and test subsets are re-sampled for the stability analysis of Table 13 (seeds 42, 123 and 2024), so that every re-sampled partition is scene-disjoint as well. For the second collection, each retained scene is cropped to a single 1024×1024 patch centred on its annotated wave packet, so that the number of patches equals the number of retained scenes (350); for the first collection, which is already distributed as 742 patches cropped from 457 scenes, the retained unit is the distributed patch. Throughout Table 3 and all subsequent experiments, counts for both collections therefore refer to (256×256) input patches obtained from the retained units.

Table 3. Scene-level partition of the two collections. Patches sharing a Sentinel-1 acquisition, or covering the same internal-wave packet, are assigned to the same subset, so no acquisition is shared between the training, validation and test subsets.

Subset	Dataset 1 (S1-IW-2023)	Dataset 2 (IWs v1.0)	Total
Labeled training	30	30	60
Unlabeled training	200	200	400
Validation	20	20	40
Test	100	100	200
Total retained	350	350	700

To simulate the common practical scenario in ocean internal wave extraction where images are abundant but annotations are scarce, and to evaluate the effectiveness of the proposed approach in leveraging unlabeled data without ever using pseudo-labels as direct training targets, the original datasets were reorganized into four subsets—a labeled training set, an unlabeled training set, a validation set and an independent test set—as follows. The remaining data (392 patches for S1-IW-2023 and 40 scenes for IWs Dataset v1.0) were not adopted in the experiments.

Labeled Training Set: 30 patches together with their corresponding masks were drawn as the labeled subset of the scene-level partition defined above and summarised in Table 3. These samples are used to estimate the class prototypes, providing the supervised guidance of the feature space throughout training.

Unlabeled Training Set: Approximately 200 unlabeled patches were selected as the unlabeled training set. During training, these images were aligned through the proposed prototype learning module. The objective was to refine feature representations by learning the intrinsic structural distribution of large-scale background and wave patterns, thereby enhancing robustness against interference signals without promoting any pseudo-label to a differentiable training target.

Validation Set: To select hyper-parameters without touching the test set, a small validation set was constructed by drawing 20 patches from the remaining data of each collection under the same scene-level protocol and annotating them manually; these images were used only for hyper-parameter tuning and were never involved in training or test evaluation.

Testing Set: An independent testing set consisting of 100 labeled patches was retained to evaluate the segmentation accuracy of the model on previously unseen scenarios. It should be noted that this test set, like the two source collections from which it is drawn, contains only scenes that were selected because they exhibit internal wave signatures; it does not include wave-free patches, nor scenes whose elongated surface signatures are produced by other oceanic or atmospheric phenomena. All accuracies reported below are therefore conditional on the presence of an internal wave in the scene. The Precision values in particular quantify the false-positive burden within wave-containing images rather than the false-alarm rate of the model over arbitrary ocean scenes.

Input Construction and Preprocessing

All experiments operate on Sentinel-1 Level-1 ground range detected (GRD) products acquired in IW mode. The products have a nominal pixel spacing of 10 m, and every patch is resized to 256×256 pixels, which is the input size of all compared baselines as well as of the proposed network. Every patch used in this study is dual-polarimetric; i.e., it contains both the VV and the VH channel: no single-polarization scene is retained, and no channel is zero-filled or interpolated, so that the three-channel descriptor defined below is well defined for every sample. The two polarizations are co-registered before the

three-channel tensor is built. All experiments are restricted to IW-mode GRD products, so the few Stripmap (SM) scenes contained in the original archives were not used.

The GRD products are distributed as 16-bit digital numbers (DN), which are not directly comparable across scenes. We therefore convert the DN of every pixel into a calibrated backscatter coefficient σ^0 following the standard Sentinel-1 calibration procedure, in which the DN is divided by the square of the calibration vector supplied with the product, and the calibrated coefficient is expressed in decibels,

$$\sigma^0_{\mathrm{dB}} = 10 \log_{10}(\sigma^0). \tag{34}$$

The logarithmic scale is used because internal waves modulate the sea-surface roughness over a wide dynamic range, and the dB representation makes this modulation comparable across scenes acquired under different wind speeds and incidence angles.

The three input channels consumed by the network are constructed from the two calibrated polarizations as

$$I = \left[\sigma^0_{VV},\ \sigma^0_{VH},\ \sigma^0_{VV} - \sigma^0_{VH} \right], \tag{35}$$

where all three quantities are expressed in decibels. Rather than using the two polarizations independently, we include the difference channel $\sigma^0_{VV} - \sigma^0_{VH}$ because the modulation of the co-polarized return by an internal wave is systematically stronger than that of the cross-polarized return; being the difference of two dB values, this channel is equivalent to $10 \log_{10}(\sigma^0_{VV}/\sigma^0_{VH})$ and therefore enhances the surface-roughness contrast associated with the wavefronts while partially cancelling the scene-wide radiometric level, which is dominated by the local wind field and the incidence angle. Supplying this contrast explicitly avoids requiring the encoder to reconstruct it from two numerically correlated input channels.

Each channel is normalized independently: its stored 8-bit values are first mapped to $[0, 1]$ by dividing by the full-scale value, and then standardized with a fixed mean of 0.50 and a fixed standard deviation of 0.25. These two constants are identical for every image and for every subset and are not estimated from the data; because the mapping is fixed, the normalized input carries no information about the labeled, unlabeled, validation or test split.

Speckle is handled implicitly rather than by an explicit despeckling filter. A filter strong enough to remove multiplicative speckle would also smooth the narrow bright–dark bands that delimit a wavefront, and would thus attenuate the signal we intend to segment. Instead, the two-level wavelet decomposition of the MWTM (Section 3.2.2) separates the high-frequency detail subbands, in which speckle mainly resides, from the low-frequency approximation, and the learnable soft-thresholding stage suppresses the small coefficients dominated by speckle while retaining the coherent wavefront edges.

No color or photometric transformation designed for optical imagery, such as color jitter, is applied, because hue and saturation are defined in an RGB color space and have no counterpart in a calibrated SAR backscatter tensor. Geometric augmentation, consisting of random horizontal and vertical flips together with random 90° rotations, is applied to every sample. For the strongly augmented view required by the consistency term, the geometric augmentation is combined with two SAR-specific photometric perturbations only: a multiplicative speckle term with unit mean, $x \leftarrow x \cdot \eta$ with $\eta \sim \Gamma(4, 1/4)$, that is, a Gamma-distributed speckle with an equivalent number of looks of four, whose result is clipped to $[0, 1]$; and a global contrast rescaling $x \leftarrow x^{\gamma}$ with $\gamma \sim \mathcal{U}(0.8, 1.25)$, which reproduces the contrast change induced by wind speed and incidence angle. Both perturbations affect the input image only and never the mask.

4.2. Implementation Details

All experiments were conducted with the following configuration. Hardware: a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB memory) and an Intel Xeon Gold 6330 CPU, running Ubuntu 20.04. Software: Python 3.9, PyTorch 2.1.0, CUDA 11.8, and cuDNN 8.9. Training setup: the models are optimized with AdamW, i.e., Adam with decoupled weight decay, using the default moment coefficients ($\beta_1 = 0.9$, $\beta_2 = 0.999$), an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, and a cosine annealing schedule that decays the learning rate to 1×10^{-6} over the training budget; the batch size is 4 and the maximum number of epochs is 30 for both the labeled and unlabeled branches; gradients are clipped to an ℓ_2 norm of 1.0; the checkpoint with the lowest validation loss is retained, and training is stopped early if the validation loss does not improve for five consecutive epochs; all input images were resized to 256×256 pixels. Data augmentation: both branches use the augmentation policy defined in Section 4.1, that is, random horizontal and vertical flips together with random 90° rotations for every view, and additionally the multiplicative speckle term and the global contrast rescaling for the strong view of the unlabeled branch. Both photometric perturbations affect the input image only and never the mask; no transformation designed for optical RGB imagery was applied to either branch.

Hyper-parameters: Table 4 collects every hyper-parameter of the framework together with its value and its provenance. In the provenance column, "prior work" denotes a value taken from the cited publication, "validation" a value selected on the validation split before the test evaluation, "empirical" a constant fixed in the implementation without a tuning procedure, and "learned" a parameter optimized by gradient descent; the segmentation loss whose two components are listed in the table is $\mathcal{L}_{\text{seg}} = \lambda_{\text{dice}}\,\mathcal{L}_{\text{Dice}} + \lambda_{\text{ce}}\,\mathcal{L}_{\text{CE}}$ (Equation (30)), and the total objective is Equation (33). In brief, the confidence threshold $\kappa = 0.85$ was fixed once by a grid search over $\{0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$ on the validation split (Section 4.1) and is shared by the prototype filtering and the consistency mask, whereas the temperature $\tau = 0.1$ of the prototype alignment losses follows the range used in prototype-based pseudo-label denoising [46]; the two terms of the segmentation loss carry equal weight ($\lambda_{\text{dice}} = \lambda_{\text{ce}} = 1.0$), so that λ_s scales the segmentation objective as a whole rather than its two components independently; the momentum coefficient of the prototype running average $\alpha_m = 0.99$ and the default number of prototypes per class $K = 1$ are those of the single-prototype configuration, and the sensitivity of the framework to them is quantified in Table 16; and the soft-thresholding offsets β_j of the wavelet module are not tuned by hand but initialized so that $\rho_j^{(0)} = 0.01$ and then learned jointly with the network. All hyper-parameters were fixed before the test evaluation and were not revisited afterwards. All models were trained from scratch with a fixed random seed (42), and the reported metrics were computed on the independent test set.

Table 4. Hyper-parameters of the proposed framework and their provenance.

Symbol	Meaning	Value	Provenance
κ	Confidence threshold, used both for prototype filtering (Equation (6)) and as the consistency mask (Equation (32))	0.85	Grid search over $\{0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$ on the validation split
τ	Temperature of the prototype alignment losses (Equations (31) and (9))	0.1	Prior work [46]
α_m	Momentum coefficient of the prototype running average (Equation (7))	0.99	Empirical; sensitivity in Table 16
K	Number of prototypes per class, default configuration	1	Empirical; sensitivity in Table 16

Table 4. *Cont.*

Symbol	Meaning	Value	Provenance
λ_s	Weight of the segmentation loss in Equation (33)	1.0	Empirical
λ_p	Weight of the prototype alignment loss in Equation (33)	0.1	Empirical (fixed); the contribution of the term is isolated in Table 17
λ_u	Weight of the consistency loss in Equation (33)	0.5	Empirical (fixed); the contribution of the term is isolated in Table 17
λ_{dice}, λ_{ce}	Weights of the Dice and binary cross-entropy terms inside $\mathcal{L}_{\text{seg}}$	1.0, 1.0	Fixed, not tuned
β_j	Unconstrained offsets of the six wavelet soft thresholds (Equation (16))	$\rho_j^{(0)} = 0.01$	Initialized so that $\rho_j^{(0)} = 0.01$, i.e., $\beta_j^{(0)} = \ln(e^{0.01} - 1) \approx -4.60$; learned thereafter

Evaluation Protocol

All models are trained on the labeled and unlabeled training subsets and their checkpoints are selected on the validation subset alone; the test subset is evaluated once, for the final report. No test-time adaptation, no self-training on test images and no test-set hyper-parameter tuning is applied to any method, including the semi-supervised baselines, which are all trained under the identical scene-disjoint partition of Table 3, the identical optimisation schedule described above and the identical augmentation policy. Every metric is computed independently on each test image and then averaged over the test set; we deliberately avoid pooling all pixels of the test set into a single confusion matrix, because pooling would allow a few large wave packets to dominate the score and would hide per-image failures.

4.3. Performance Measure

To comprehensively evaluate the performance of the proposed model in ocean internal wave detection, multiple commonly used image segmentation metrics were adopted. Combined with color-coded visual analysis, quantitative and qualitative comparisons were conducted at the pixel level, region level, and boundary level. The evaluation scheme not only includes traditional segmentation accuracy metrics, but also introduces boundary-aware indices that better conform to the morphological characteristics of internal waves.

4.3.1. Pixel Accuracy (PA)

Pixel Accuracy (PA) measures the overall consistency between the predicted image and the ground-truth mask at the pixel level, which is defined as the proportion of correctly predicted pixels over the total number of pixels. However, when the target region is relatively small or the background occupies a large proportion (i.e., class imbalance), PA may mask certain prediction errors. The mathematical formulation of PA is defined as

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}}, \tag{36}$$

where p_{ii} denotes the number of true positive (TP) pixels, and p_{ij} represents the number of pixels whose ground-truth class is i but are predicted as class j.

4.3.2. Intersection over Union (IoU)

Intersection over Union (IoU) computes the ratio between the intersection and the union of the predicted region and the ground-truth region. It is one of the most represen-

tative core metrics in semantic segmentation, where a higher IoU value indicates more accurate spatial localization and more complete detection. IoU is defined as

$$IoU = \frac{TP}{TP + FP + FN}, \tag{37}$$

where TP, FP, and FN represent the numbers of true positive, false positive, and false negative pixels, respectively.

4.3.3. Dice Coefficient

The Dice coefficient is regarded as a spatial variant of the F1-score. It emphasizes the overlap between the predicted mask and the ground-truth mask, assigning double weight to the intersection region. Compared with IoU, Dice is more sensitive to small targets, discontinuous boundaries, and class imbalance. It is particularly suitable for evaluating the spatial fitting performance of slender targets such as ocean internal waves. The Dice coefficient is defined as

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{38}$$

4.3.4. Precision/Recall

Precision and Recall reflect the accuracy and coverage of positive-class predictions, respectively. A higher Precision indicates fewer false alarms, while a higher Recall implies fewer misses. The balance between Precision and Recall is critical for reliability in practical applications, as these two metrics are commonly used to analyze the trade-off between false positives and false negatives. The definitions of Precision and Recall are given as

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{39}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{40}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively.

4.3.5. Boundary F1-Score

The boundary of a binary mask M is extracted as its one-pixel inner contour,

$$B(M) = M \setminus \text{erode}(M, K_3), \tag{41}$$

where K_3 is a 3×3 structuring element and the same rule is applied to the prediction P and to the ground truth G. Predicted boundary points are matched to ground-truth boundary points within a tolerance of $\theta = 2$ pixels using a Euclidean distance transform, and the matching is performed in both directions:

$$P_{\text{boundary}} = \frac{\left|\{p \in B(P) : d(p, B(G)) \le \theta\}\right|}{|B(P)|}, \qquad R_{\text{boundary}} = \frac{\left|\{q \in B(G) : d(q, B(P)) \le \theta\}\right|}{|B(G)|}, \tag{42}$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance between a boundary point and the nearest point of the other boundary. The Boundary F1-score is then defined as

$$\text{BF1} = \frac{2 \times P_{\text{boundary}} \times R_{\text{boundary}}}{P_{\text{boundary}} + R_{\text{boundary}}}. \tag{43}$$

The tolerance is fixed at $\theta = 2$ pixels for every method and both datasets; the ranking reported below is unchanged for $\theta \in \{1, 2, 3\}$.

To facilitate intuitive performance analysis, a color-coded mask visualization strategy is adopted. Specifically, blue indicates correctly predicted regions (TP), red denotes missed regions (FN), and cyan represents false positive regions (FP). This visualization method helps to intuitively assess boundary details, target completeness, and background false alarms, thereby complementing the quantitative evaluation results. It should be kept in mind that every test image contains at least one internal wave, so the FP regions shown here and the Precision values in Tables 9 and 11 quantify over-segmentation within wave-containing scenes; the false-alarm behaviour of the model on wave-free ocean scenes lies outside the scope of the present evaluation and is discussed in Section 5.7.

4.3.6. Skeleton-Based clDice

The centreline Dice (clDice) [30] compares the skeletons of the predicted and ground-truth masks and is therefore sensitive to connectivity rather than to area overlap alone. Let S_P and S_G denote the skeletons of the prediction P and of the ground truth G. Topology precision and topology sensitivity are defined as

$$T_{\text{prec}} = \frac{|S_P \cap D(G)|}{|S_P|}, \qquad T_{\text{sens}} = \frac{|S_G \cap D(P)|}{|S_G|}, \tag{44}$$

where $D(\cdot)$ denotes a one-pixel morphological dilation of a mask, so that each skeleton is intersected with the dilated counterpart of the other mask. This dilation is a one-pixel matching tolerance: it prevents the single-pixel-wide skeleton from being penalised by a residual localisation offset between the mask pair, so that clDice measures connectivity rather than sub-pixel alignment. The clDice is the harmonic mean of the two terms,

$$\text{clDice} = \frac{2\, T_{\text{prec}}\, T_{\text{sens}}}{T_{\text{prec}} + T_{\text{sens}}}. \tag{45}$$

A wave band that is recovered only partially, or that is split into several disjoint pieces, loses skeleton overlap and is penalised even when its area overlap remains high.

4.3.7. Connectivity Accuracy (CA)

Region- and skeleton-based overlaps are both global measures and do not distinguish a prediction that recovers a structure with one connected component from a prediction that recovers it with several. Connectivity accuracy therefore asks whether every ground-truth structure is covered by a single predicted component. Let $S_j^{(t)}$ denote the one-pixel morphological skeleton of the j-th ground-truth component of the t-th evaluated image, and let $\tilde{C}_i^{(t)}$ denote the one-pixel dilation of the i-th connected component of the corresponding predicted mask (thresholded at 0.5). The j-th ground-truth structure is considered covered if

$$\frac{\left|S_j^{(t)} \cap \tilde{C}_i^{(t)}\right|}{\left|S_j^{(t)}\right|} \geq 0.9 \tag{46}$$

for a *single* index i, and CA is the fraction of ground-truth structures that are covered:

$$\text{CA} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_G^{(t)}} \sum_{j=1}^{N_G^{(t)}} \mathbb{1}\left[\exists\, i : \ \frac{\left|S_j^{(t)} \cap \tilde{C}_i^{(t)}\right|}{\left|S_j^{(t)}\right|} \geq 0.9\right], \tag{47}$$

where T is the number of evaluated images. The one-pixel dilation of the prediction acts as a matching tolerance and absorbs the localisation error of the segmentation, while no sub-pixel registration is applied to the predictions. We evaluate coverage on the skeleton rather than on the full ground-truth region because the masks used here are comparatively thick

bands: a set of many small blobs can jointly cover most of the band *area* while individually tracing none of the crest. Reducing each ground-truth component to its skeleton and requiring a single predicted component to account for it makes CA a statement about whether the crest is followed end-to-end by one connected prediction, which is the property that matters for internal-wave delineation.

4.3.8. Fragmentation Rate (FR)

Skeleton-based measures are insensitive to over-segmentation: splitting a structure into several pieces leaves both the skeleton and its overlap with the prediction essentially unchanged. We therefore report the fragmentation rate, which compares the number of connected components of the prediction with that of the ground truth.

Let $\mathcal{C}_A(\cdot)$ denote the number of 8-connected components of a binary mask whose area is at least $A_{\min} = 10$ px, computed after thresholding the prediction at 0.5, and write $N_G^{(t)} = \mathcal{C}_A(Y^{(t)})$ and $N_P^{(t)} = \mathcal{C}_A(\hat{Y}^{(t)})$. The signed component-count deviation of image t is $r_t = N_P^{(t)} - N_G^{(t)}$ (r for the residual component count; the symbol Δ is reserved for the state-space step size of Equation (25)): $r_t > 0$ indicates over-segmentation (the prediction is broken into more pieces than the ground truth), whereas $r_t < 0$ indicates under-segmentation (adjacent ground-truth structures are merged into a single predicted component). The fragmentation rate is

$$\mathrm{FR} = \frac{1}{T} \sum_{t=1}^{T} \frac{\left| N_P^{(t)} - N_G^{(t)} \right|}{N_G^{(t)}}, \tag{48}$$

where T is the number of evaluated images. Because the denominator contains the ground-truth count only, FR is unbounded above: values larger than one simply mean that the prediction is more than twice as fragmented as the ground truth, and the same numerical deviation weighs very differently depending on $N_G^{(t)}$ (e.g., $r_t = 2$ gives FR $= 0.33$ for $N_G^{(t)} = 6$ but FR $= 0.13$ for $N_G^{(t)} = 15$). To obtain a scale that is directly comparable across images and bounded in $[0, 1]$, we additionally report the normalised fragmentation rate

$$\mathrm{FR}' = \frac{1}{T} \sum_{t=1}^{T} \frac{\left| N_P^{(t)} - N_G^{(t)} \right|}{\max\left\{ N_P^{(t)},\, N_G^{(t)} \right\}}, \tag{49}$$

which equals 0 when the predicted and ground-truth component counts coincide and approaches 1 when one of the two sets is empty. Per image and for over-segmented predictions ($N_P^{(t)} \geq N_G^{(t)}$) it satisfies $\mathrm{FR}'_t = \mathrm{FR}_t / (1 + \mathrm{FR}_t)$, so that fragmentation rates of 1, 4 and 95 correspond to FR′ values of 0.5, 0.8 and 0.99, respectively, while for $N_P^{(t)} \leq N_G^{(t)}$ the two measures coincide.

FR preserves the absolute scale of the fragmentation error, whereas FR′ provides a bounded and image-independent summary; both are reported jointly in Table 10. Note finally that FR is complementary to CA. CA is computed on the skeleton of the ground truth and therefore penalises a prediction that fails to trace a crest with a single component, while FR penalises the opposite failure mode, namely the emergence of spurious or duplicated components. A method may therefore obtain a high CA together with a poor FR, or vice versa, and the two measures should be read together.

4.4. Comparison Methods

To comprehensively evaluate the effectiveness of the proposed PWM-Net, comparative experiments were conducted from two complementary perspectives. For segmentation

accuracy, PWM-Net was compared against five state-of-the-art semi-supervised semantic segmentation baselines under identical partial-label settings. For computational efficiency, four additional fully supervised baselines (Wavelet-UNet [65], WaveSNet [50], PMM [66], and VM-UNet [67]) were included to cover the frequency-domain, prototype-based, and Mamba-based architectural paradigms. Each method is referred to by the name under which it was published and is reproduced from its published reference design, i.e., the network and the training strategy described in the original publication, re-implemented in our common training and evaluation pipeline rather than executed from the authors' official code repository; the method-specific settings follow the corresponding publication, with the single exception of the UniMatch confidence threshold described below, and are listed in Table 6. To avoid the ambiguity that arises when published methods are re-implemented under a new name, each method is referred to by its published name throughout this paper; Table 6 lists the network and implementation used for every compared method. These baselines were strategically selected to ensure a rigorous assessment of the specific technical components utilized in our framework. Table 5 summarizes the core characteristics and applicability of all compared methods.

The comparison methods comprise two categories:

1. Fully supervised baselines (reference designs, trained with full annotations): Wavelet-UNet [65] and WaveSNet [50] represent frequency-domain feature enhancement techniques; PMM [66] validates the effectiveness of prototype-based feature learning; VM-UNet [67] serves as a state-of-the-art Mamba-based visual backbone for long-range dependency modeling.
2. Semi-supervised baselines (trained under identical partial-label settings): SS-Net [68] serves as a medical-oriented semi-supervised baseline with smoothness constraints; Mean Teacher [34] represents the classical consistency regularization paradigm using EMA-based teacher–student learning; CPS [33] employs cross pseudo supervision between dual decoders for co-training; U2PL [38] adopts contrastive learning with unreliable pseudo-labels to fully utilize unlabeled data; UniMatch [62] represents the current state-of-the-art in general semi-supervised semantic segmentation through dual-stream weak-to-strong consistency.

The accuracy comparison is restricted to the five semi-supervised methods and to PWM-Net, because all six consume the same labeled and unlabeled data under the same partial-label protocol; placing the four fully supervised frameworks in the same table would compare paradigms that use different amounts of supervision. A labelled-only reference on our own architecture is nevertheless provided by the module ablation of Section 4.6: configuration A1 (plain U-Net) and configuration A8 (the full architecture trained on the same 30 labeled images with the same data pipeline, optimisation schedule and augmentation policy as the semi-supervised methods, but with the unlabeled branch removed) are trained under the same labeled-data budget and the same training pipeline as the semi-supervised methods, although without unlabeled data; the gap between A8 (IoU 0.5373) and the full PWM-Net (IoU 0.7655) therefore isolates the contribution of the semi-supervised mechanism itself. The four fully supervised frameworks are therefore reported as architectural and efficiency references in Table 7.

Table 5. Core characteristics and applicability of different comparison methods.

Method	Publication (Year)	Core Innovation	Rationale for Selection
Fully supervised:			
Wavelet-UNet [65]	ICANN 2020	Wavelet U-Net with DWT/IDWT layers for detail-preserving segmentation.	Represents frequency-domain modeling.
WaveSNet [50]	PRCV 2022	DWT/IDWT as trainable network layers for detail-preserving segmentation.	Represents end-to-end frequency-domain feature enhancement.
PMM [66]	ECCV 2020	Mixture-of-prototypes representation estimated by EM.	Validates the effectiveness of prototype representations in discriminative feature learning.
VM-UNet [67]	arXiv 2024	Pure vision Mamba architecture.	Highlights the capability of state-space models in capturing long-range dependencies.
Semi-supervised:			
SS-Net [68]	MICCAI 2022	Pixel-level smoothness and inter-class separation constraints.	Medical-oriented semi-supervised baseline representative of smoothness-constrained segmentation under limited-label settings.
Mean Teacher [34]	NeurIPS 2017	Consistency regularization via EMA teacher–student framework.	Represents the classical consistency-based semi-supervised paradigm.
CPS [33]	CVPR 2021	Cross Pseudo Supervision between dual decoders.	Represents the pseudo-labeling co-training paradigm.
U2PL [38]	CVPR 2022	Semi-supervised contrastive learning using unreliable pseudo-labels.	Represents contrastive learning-based semi-supervised segmentation.
UniMatch [62]	CVPR 2023	Weak-to-strong dual-stream consistency for semi-supervised segmentation.	Serves as the current state-of-the-art in general semi-supervised semantic segmentation.
PWM-Net (Ours)	–	A collaborative segmentation framework integrating prototype learning, wavelet transform, and Mamba.	Proposed framework addressing weak textures and slender topologies.

To keep the comparison interpretable, all semi-supervised methods were trained and evaluated in the same codebase, on the same scene-level partition (Section 4.1), with the same augmentation pipeline, the same optimisation family, the same epoch budget and the same evaluation code. Method-specific hyper-parameters—the EMA decay of Mean Teacher, the confidence thresholds of CPS, the contrastive weights of U2PL, and the smoothness and class-separation weights of SS-Net—were taken from the corresponding original publications and verified on the same validation split, with one exception: the confidence threshold of UniMatch is set to $\kappa = 0.85$, i.e., to the same value used by the proposed method, instead of the 0.95 recommended in its original implementation, so that every method in the accuracy comparison is evaluated under the same confidence criterion. All method-specific settings are listed together with the network and the implementation source of each method in Table 6. No method, including the proposed one, was tuned on the test set.

Table 6. Implementation of the compared methods. All methods share the same scene-level data partition, the same augmentation pipeline, the same optimisation family, the same epoch budget and the same evaluation code. The third column gives the reference network design of each method and the fourth column the settings that are specific to it and are not shared with the other methods.

Method	Venue	Network (Reference Design)	Method-Specific Setting
Fully supervised (efficiency references only):			
Wavelet-UNet [65]	ICANN 2020	Wavelet U-Net	—
WaveSNet [50]	PRCV 2022	Wavelet-integrated encoder–decoder	—
PMM [66]	ECCV 2020	Prototype-based segmentation network	—
VM-UNet [67]	arXiv 2024	Pure vision state-space (Mamba) backbone	—
Semi-supervised (accuracy comparison):			
SS-Net [68]	MICCAI 2022	Encoder–decoder with dual decoders and an EMA teacher	Smoothness and class-separation weights as in [68]
Mean Teacher [34]	NeurIPS 2017	Encoder–decoder with an EMA teacher	EMA decay 0.999
CPS [33]	CVPR 2021	Shared encoder with two decoders	Cross-supervision weight 1.0
U2PL [38]	CVPR 2022	Encoder–decoder with a contrastive head	Contrastive weight and temperature as in [38]
UniMatch [62]	CVPR 2023	Shared encoder with dual weak–strong branches	Confidence threshold set to $\kappa = 0.85$ (our unified value; the original implementation recommends 0.95)
PWM-Net (Ours)	—	Compact CNN encoder (32–64–128), Mamba block, U-Net decoder	$\kappa = 0.85$, $\alpha_m = 0.99$, $\tau = 0.1$

4.5. Experimental Results

4.5.1. Computational Efficiency Analysis

To comprehensively evaluate the computational efficiency of different segmentation models, we assess model complexity using five complementary metrics: the number of parameters (Params), floating-point operations (FLOPs) at multiple input resolutions, peak GPU memory consumption, model storage size, and average training time per epoch. For semi-supervised methods (Mean Teacher, CPS, U2PL, UniMatch, SS-Net), we report the complexity of the student network during inference, as the teacher or auxiliary branches are only required during training. The four fully supervised frameworks are additionally reported here as efficiency references; as discussed in Section 4.4, they were not evaluated for segmentation accuracy under the semi-supervised protocol.

Table 7 summarizes the computational efficiency of all compared models across five metrics. As shown, the proposed PWM-Net achieves a highly favorable balance between model complexity and segmentation performance. With only 8.98 M parameters, 18.4 G FLOPs (at 256×256), 923 MB GPU memory, and a model size of 33 MB, PWM-Net is among the most lightweight of the compared frameworks: it has fewer parameters than Mean Teacher (9.14 M) and CPS (15.23 M), and is of the same order of magnitude as UniMatch (7.20 M) and U2PL (8.24 M), while outperforming all of them in segmentation accuracy (Tables 9 and 11). In contrast, the Mamba-based VM-UNet requires 21.52 M parameters and 44.2 G FLOPs at the same resolution, while SS-Net exhibits the heaviest footprint across all metrics (39.54 M parameters, 98.6 G FLOPs, 3438 MB memory). This indicates that PWM-Net's multi-module design does not incur excessive overhead, thanks to the parallel architecture and the linear complexity of its core components.

The cost of the proposed model grows in proportion to the number of input pixels. When the input resolution increases from 128 × 128 to 512 × 512, that is, by a factor of 16 in pixel count, PWM-Net's FLOPs grow from 5.1 G to 70.2 G, that is, roughly in proportion to the input pixel count, in line with the linear per-token cost of the selective scan. The Mamba-based VM-UNet follows a similarly linear trend (11.4 G to 174.3 G) but from a considerably larger constant: at 512 × 512 it requires 174.3 G FLOPs against 70.2 G for PWM-Net, so that PWM-Net uses 59.7% fewer FLOPs at the same resolution. The two state-space backbones therefore differ in scale rather than in asymptotic behaviour, and the comparison against a self-attention operator of comparable width under our own protocol is reported separately in Section 4.5.2.

In terms of training efficiency, PWM-Net requires only 142 s per epoch, faster than all the other compared baselines except the two wavelet-based reference models, Wavelet-UNet (73 s/epoch) and WaveSNet (116 s/epoch), which are trained under the fully supervised baseline configuration without the additional unlabeled branch. Notably, it is faster than Mean Teacher (169 s/epoch) despite similar parameter counts, as the teacher network in Mean Teacher introduces an extra forward pass and EMA update. CPS (223 s/epoch) and UniMatch (215 s/epoch) incur further overhead from dual-branch architectures and augmented data streams. Taken together, these results demonstrate that PWM-Net delivers superior segmentation accuracy while maintaining competitive computational efficiency—a crucial property for practical deployment in large-scale ocean remote sensing applications.

Table 7. Computational efficiency comparison of different segmentation models.

Method	Params (M)	FLOPs (G)			Model Size (MB)	Training Time (s)	GPU Memory (MB)
		128 × 128	256 × 256	512 × 512			
Wavelet-UNet	5.59	–	13.3	–	22	73	551
WaveSNet	13.17	–	29.8	–	55	116	1032
PMM	14.08	–	34.1	–	55	147	1186
VM-UNet	21.52	11.4	44.2	174.3	82	205	1763
SS-Net	39.54	–	98.6	–	151	397	3438
Mean Teacher	9.14	–	16.4	–	28	169	750
CPS	15.23	–	31.6	–	55	223	1462
U2PL	8.24	–	19.4	–	31	184	883
UniMatch	7.20	–	17.1	–	30	215	832
PWM-Net (Ours)	8.98	5.1	18.4	70.2	33	142	923

4.5.2. Mamba Versus Self-Attention Under an Identical Protocol

To separate the contribution of the state-space sequence module from the choice of a long-range operator, we trained a control variant of the proposed architecture in which the four directional Mamba blocks are replaced by window-based multi-head self-attention blocks of comparable width and depth (window size 8, eight heads). Every other component—the encoder–decoder backbone, the wavelet enhancement module, the prototype alignment module, the loss weights, the augmentation policy and the data pipeline—is kept identical, and both variants are trained under the same labeled-only protocol used for the module ablation of Table 14 (30 labeled images, the training schedule of Section 4.2, seed 42). The efficiency figures of the proposed model are reproduced from Table 7, and Time denotes the training time per epoch. Table 8 summarises the results.

Table 8. Controlled comparison between the proposed Mamba-based sequence module and a window-based self-attention block of comparable width and depth.

Variant	Acc.	IoU	Dice	Prec.	Recall	BF1	Params (M)	FLOPs (G)	Time (s)
Window self-attention control	0.9448	0.5112	0.6719	0.7216	0.6028	0.1417	10.15	25.9	190
PWM-Net (Mamba, ours)	0.9467	0.5373	0.6990	0.7428	0.6314	0.1524	8.98	18.4	142

Under this controlled setting, the attention variant does not improve accuracy: it attains an IoU of 0.5112 and a BF1 of 0.1417 against 0.5373 and 0.1524 for the Mamba variant, while requiring 1.13× the parameters, 1.41× the FLOPs and 1.34× the training time per epoch. The margins are modest (0.0261 in IoU) and derive from single runs with identical random seeds, so they should not be read as evidence that state-space models are intrinsically superior to self-attention for this task. They do, however, support the weaker but sufficient statement that, at matched width and depth, the linear-complexity operator does not cost accuracy while requiring clearly less computation and memory.

4.5.3. Test Results on Dataset 1

To clearly present the segmentation performance of the proposed method and the compared approaches on the oceanic internal wave remote sensing dataset, Table 9 summarizes the key evaluation metrics of the methods on the Dataset 1 test set.

Table 9. Segmentation performance comparison of different methods on the Dataset 1 test set.

Method	Pix Accuracy	IoU	Dice	Precision	Recall	BF1
SS-Net	0.9453	0.5589	0.7085	0.7390	0.6900	0.1331
Mean Teacher	0.9533	0.5755	0.7274	0.8668	0.6516	0.1545
CPS	0.8515	0.3025	0.4613	0.3676	0.6336	0.0592
U2PL	0.9350	0.4244	0.5899	0.7644	0.5023	0.0905
UniMatch	0.9354	0.4771	0.6444	0.6949	0.6101	0.1001
PWM-Net (Ours)	0.9724	0.7655	0.8669	0.8286	0.9100	0.2579

As detailed in Table 9, PWM-Net attains the best score on five of the six metrics: PA of 0.9724, IoU of 0.7655, Dice of 0.8669, recall of 0.9100 and BF1 of 0.2579. The only exception is precision, where Mean Teacher is marginally higher (0.8668 versus 0.8286). The margins over the strongest baseline in each column are consistent rather than metric-specific: 0.0191 in PA, 0.1900 in IoU, 0.1395 in Dice and 0.1034 in BF1 over Mean Teacher, and 0.2200 in recall over SS-Net. Mean Teacher is the strongest of the compared semi-supervised methods, and its behaviour is complementary to ours: it reaches a precision comparable to that of PWM-Net while operating at a markedly lower recall (0.6516); i.e., it recovers the bright interior of the wave packet but misses the faint leading and trailing crests, where the backscatter modulation is weakest. CPS is the weakest configuration (IoU of 0.3025, Dice of 0.4613), which is consistent with the degradation of its cross-supervision signal when the two branches disagree on thin, low-contrast crests. Across the labelled-data partitions evaluated in Table 13, the IoU of PWM-Net varies with a standard deviation of only 0.0045, indicating that its advantage is not attributable to a particular data partition.

Table 10 evaluates structural fidelity with four topology-aware criteria that are much stricter than region overlap. The analysis uses 30 test images drawn at random from the 100-image test set of Dataset 1, and the same 30 images are used for every method, so that any difference in the table originates from the model rather than from the evaluated images. In Table 10, the arrows indicate whether larger (↑) or smaller (↓) values are better, and the best value in each column is shown in bold. clDice penalises a prediction that fails to trace a ridge continuously, CA measures whether every ground-truth structure is recovered by a single connected prediction, and FR (together with its bounded form FR′) quantifies

how strongly the prediction is fragmented into spurious components. Averaged over these 30 test images, PWM-Net (Ours) attains the highest clDice (0.9765), the highest CA (0.7778) and the lowest FR (0.1556) and FR′ (0.14), compared with the best baseline values of clDice 0.8568 (Mean Teacher), CA 0.6444 (UniMatch), FR 0.8111 (SS-Net) and FR′ 0.44 (SS-Net).

Table 10. Topology-aware evaluation of the compared methods on 30 test images drawn at random from the 100-image test set of Dataset 1; the same 30 images are used for every method.

Method	clDice ↑	CA ↑	FR ↓	FR′ ↓
SS-Net	0.8303	0.2667	0.8111	0.44
Mean Teacher	0.8568	0.4778	4.4444	0.79
CPS	0.3958	0.4333	95.7222	0.99
U2PL	0.7711	0.3667	34.7889	0.96
UniMatch	0.7408	0.6444	63.7111	0.98
PWM-Net (Ours)	0.9765	0.7778	0.1556	0.14

The absolute scale of FR is only meaningful when it is read together with the underlying component counts. For these 30 test images, the ground truth contains on average $N_G = 9.0$ components, whereas the number of predicted components spans two orders of magnitude across methods: 15.7 (SS-Net), 46.7 (Mean Teacher), 250.3 (U2PL), 469.7 (UniMatch) and 716.3 (CPS), against 7.7 for PWM-Net (Ours). The mean absolute component-count deviation follows the same ordering (6.7, 37.7, 241.3, 460.7, 707.3 and 1.3). FR therefore disperses over 0.81–95.72 for the baselines, and its magnitude is dominated by the extreme over-fragmentation of the pseudo-label-based methods rather than by subtle differences in boundary quality: on a ground truth of nine components, CPS produces on average more than seven hundred fragments. The bounded counterpart FR′ compresses this range to 0.44–0.99 while preserving the ranking, and shows that the proposed method keeps the predicted component count close to the ground truth from both sides (7.7 versus 9.0; i.e., it occasionally merges rather than splits adjacent fragments, mean absolute component-count deviation r_t of 1.3, FR′ = 0.14).

To provide a spatial-level assessment of the segmentation performance, Figure 5 visualizes the model predictions alongside corresponding error maps, where true positive (TP), false positive (FP), and false negative (FN) regions are highlighted in blue, cyan, and red, respectively. Each row of the figure corresponds to one randomly selected test image and shows, from left to right, the original image, the model prediction, the ground truth and the error map. The error maps illustrate the behaviour quantified in Table 10: the baselines tend to recover the bright interior of the packet while either omitting the faint crests (false negatives, red) or breaking them into isolated segments.

In the visualizations presented in Figure 5, the predicted segmentation of PWM-Net closely matches the ground truth. The error maps indicate that TP (blue) regions densely and continuously cover the main wave patterns. False negatives (red) occur mainly at the narrow wave termini and where wave signatures intersect other structures. FPs (cyan) are sparse, indicating that, within these wave-containing scenes, the model produces few responses on background texture even in Sample 2, which contains high morphological variation and local discontinuities. This visual impression is consistent with the Precision values in Table 9; because all test images contain an internal wave, it should not be read as evidence that the model would remain equally free of false alarms on wave-free ocean scenes.

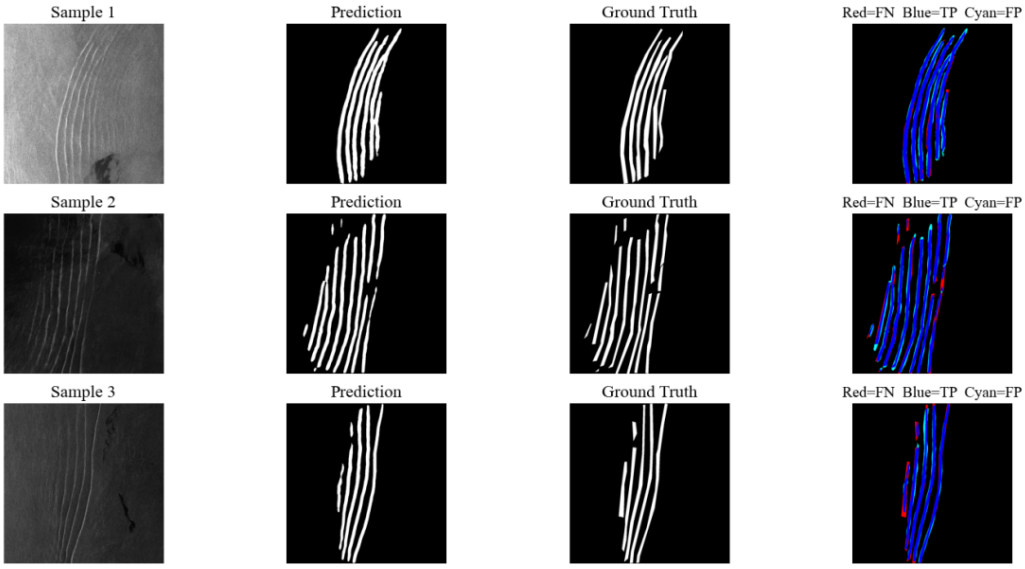


Figure 5. Visualization of segmentation results and error distribution on randomly selected test samples from Dataset 1. In the rightmost column, red, blue and cyan indicate false negatives (FN), true positives (TP) and false positives (FP), respectively.

Figure 6 presents a comparative visualization among the evaluated models across three distinct samples. Each group of the figure contains the original image, the ground truth and the predictions of the six compared methods, and the proposed approach is the last one (highlighted in red). In Sample 1, SS-Net (Method 1) roughly delineates the wave contours but generates numerous false positive clusters in the background. Mean Teacher (Method 2) captures the wave regions more completely yet still suffers from notable fragmentation. CPS (Method 3) and U2PL (Method 4) exhibit visible discontinuities and missing fine details along the slender wave regions, with CPS producing particularly sparse predictions. UniMatch (Method 5) shows scattered activations with limited structural integrity. In contrast, PWM-Net (Method 6, Ours) produces continuous wave bands with sharp boundaries, closely matching the ground truth.

In Sample 2 and Sample 3, where background interference is more severe and wave patterns are highly curved, SS-Net (Method 1) and CPS (Method 3) suffer from severe structural loss and excessive false positives. Mean Teacher (Method 2) and U2PL (Method 4) yield relatively cleaner backgrounds, yet their segmentations remain fragmented, with broken streaks and incomplete wave trajectories. UniMatch (Method 5) partially captures the wave morphology but lacks precision in the curved regions. By contrast, PWM-Net (Method 6, Ours) maintains the continuity of the elongated morphology and curved trajectories, yielding results structurally consistent with the ground truth across all challenging samples.

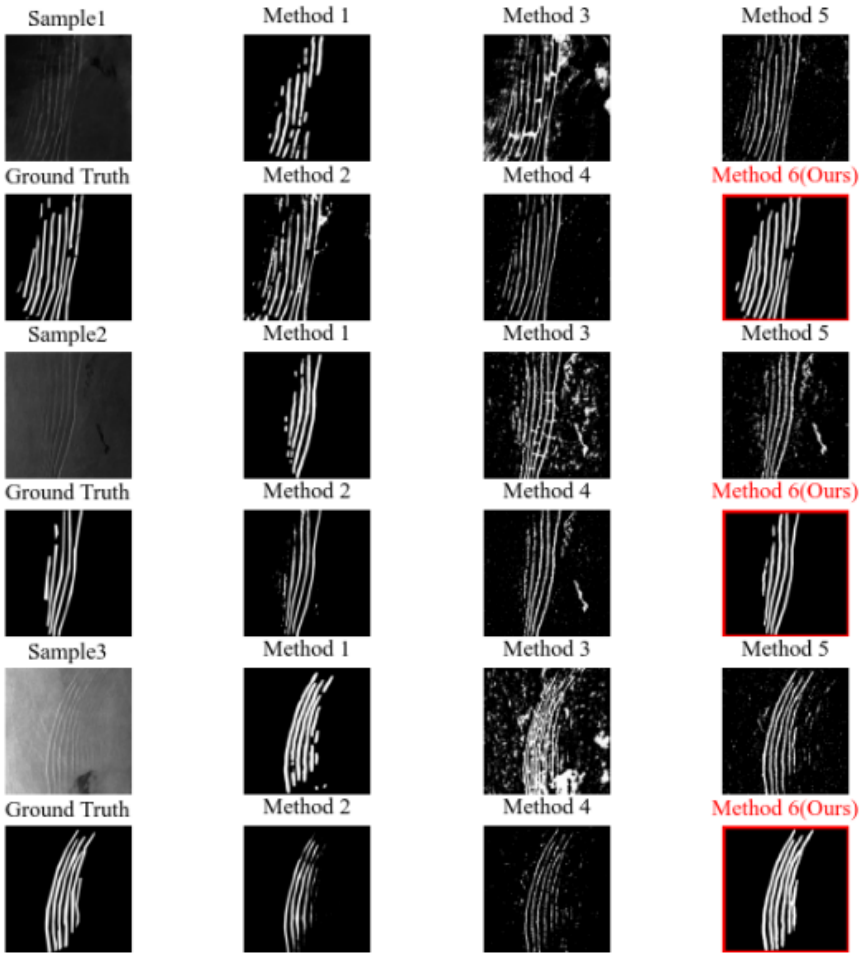


Figure 6. Comparison of segmentation results of different methods on Dataset 1 samples.

4.5.4. Test Results on Dataset 2

To assess the behaviour of the proposed method on a second collection with a different sampling composition, imaging scale and background clutter, Table 11 reports the key evaluation metrics of all compared methods on the Dataset 2 test set.

Table 11. Segmentation performance comparison of different methods on the Dataset 2 test set.

Method	Pix Accuracy	IoU	Dice	Precision	Recall	BF1
SS-Net	0.9345	0.5430	0.7009	0.6966	0.7094	0.1024
Mean Teacher	0.9428	0.5909	0.7422	0.7546	0.7448	0.1456
CPS	0.8194	0.3148	0.4738	0.3514	0.7369	0.0619
U2PL	0.9066	0.3568	0.5187	0.6138	0.4684	0.0631
UniMatch	0.9176	0.3693	0.5313	0.7046	0.4311	0.0582
PWM-Net (Ours)	0.9614	0.7154	0.8333	0.7891	0.8872	0.1895

As shown in Table 11, PWM-Net achieves the highest values across all metrics. Specifically, it attains a Pixel Accuracy of 0.9614, an IoU of 0.7154, a Dice of 0.8333, a Precision of 0.7891, a Recall of 0.8872, and a BF1 score of 0.1895. Compared with the best-performing baseline (Mean Teacher), PWM-Net improves IoU by 21.1% (from 0.5909 to 0.7154), Dice by 12.3% (from 0.7422 to 0.8333), Precision by 4.6% (from 0.7546 to 0.7891), and Recall by 19.1% (from 0.7448 to 0.8872). Among the baseline methods, Mean Teacher achieves the highest Recall (0.7448) and Pixel Accuracy (0.9428) but a comparatively lower IoU (0.5909) and

Dice (0.7422); its Precision is 0.7546, which the proposed method raises to 0.7891. SS-Net, a representative medical-oriented semi-supervised baseline, yields a competitive Pixel Accuracy of 0.9345; however, its BF1 (0.1024) remains substantially lower than that of the proposed method. CPS exhibits the lowest performance across most metrics, recording a Pixel Accuracy of 0.8194, an IoU of 0.3148, and a Dice of 0.4738. Consistent with Dataset 1, PWM-Net again shows the smallest performance fluctuation across random splits (IoU std. of 0.0049 in Table 13), indicating that its advantage on this collection is not tied to a particular partition of the data.

To intuitively understand the spatial performance under cross-dataset conditions, Figure 7 visualizes the prediction results and corresponding error maps for Dataset 2.

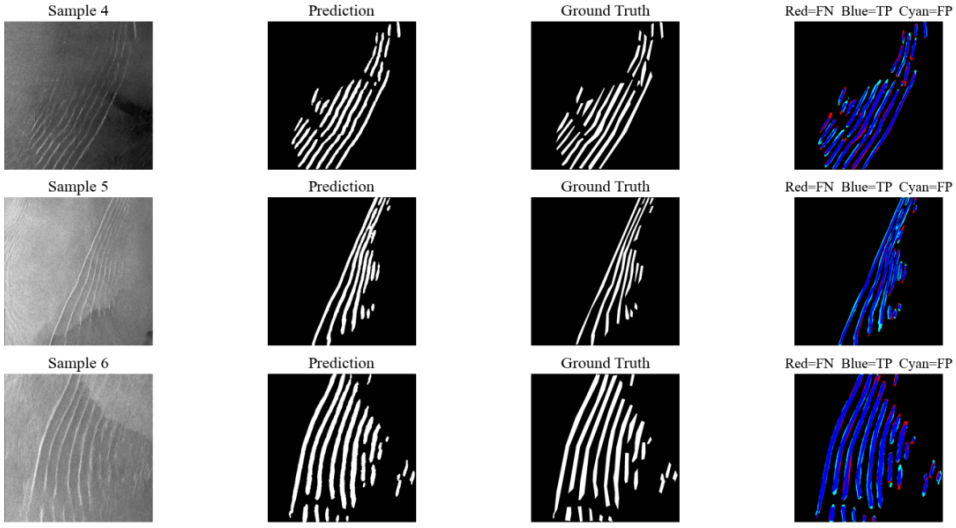


Figure 7. Visualization of segmentation results and error distribution on randomly selected test samples from Dataset 2. Three randomly selected test images from Dataset 2 are used to illustrate the comparison between predictions and ground-truth labels. Each row presents the predicted segmentation, the corresponding ground truth, and the error distribution map.

As observed in Figure 7, the true-positive regions (blue) continuously cover the main wave patterns. False negatives (red) are sparse and predominantly restricted to extremely narrow or weak-texture regions at local boundaries. False positives (cyan) appear sparsely in transitional texture areas and have limited influence on the overlap-based metrics reported in Table 11. As on Dataset 1, these observations describe behaviour on wave-containing scenes only. Notably, in Sample 6, where wave patterns are closely spaced, overlapping, and sharply curved, the predicted structures remain coherent with well-defined edges.

Figure 8 presents a comparative visualization of different models on Dataset 2 across three challenging samples. Each group of the figure contains the original image, the ground truth and the predictions of the six compared methods, with the proposed approach shown last (highlighted in red). In Sample 4, SS-Net (Method 1) is severely affected by background interference, producing scattered noisy pixels. Mean Teacher (Method 2) manages to extract the main structure but still suffers from notable discontinuity. CPS (Method 3) and U2PL (Method 4) exhibit visible missed detections on the thinner wave branches, with CPS producing particularly sparse and fragmented predictions. UniMatch (Method 5) achieves relatively good completeness but still shows minor discontinuities along the curved segments. In contrast, PWM-Net (Method 6, Ours) accurately recovers subtle wave details without fragmentation, closely matching the ground truth.

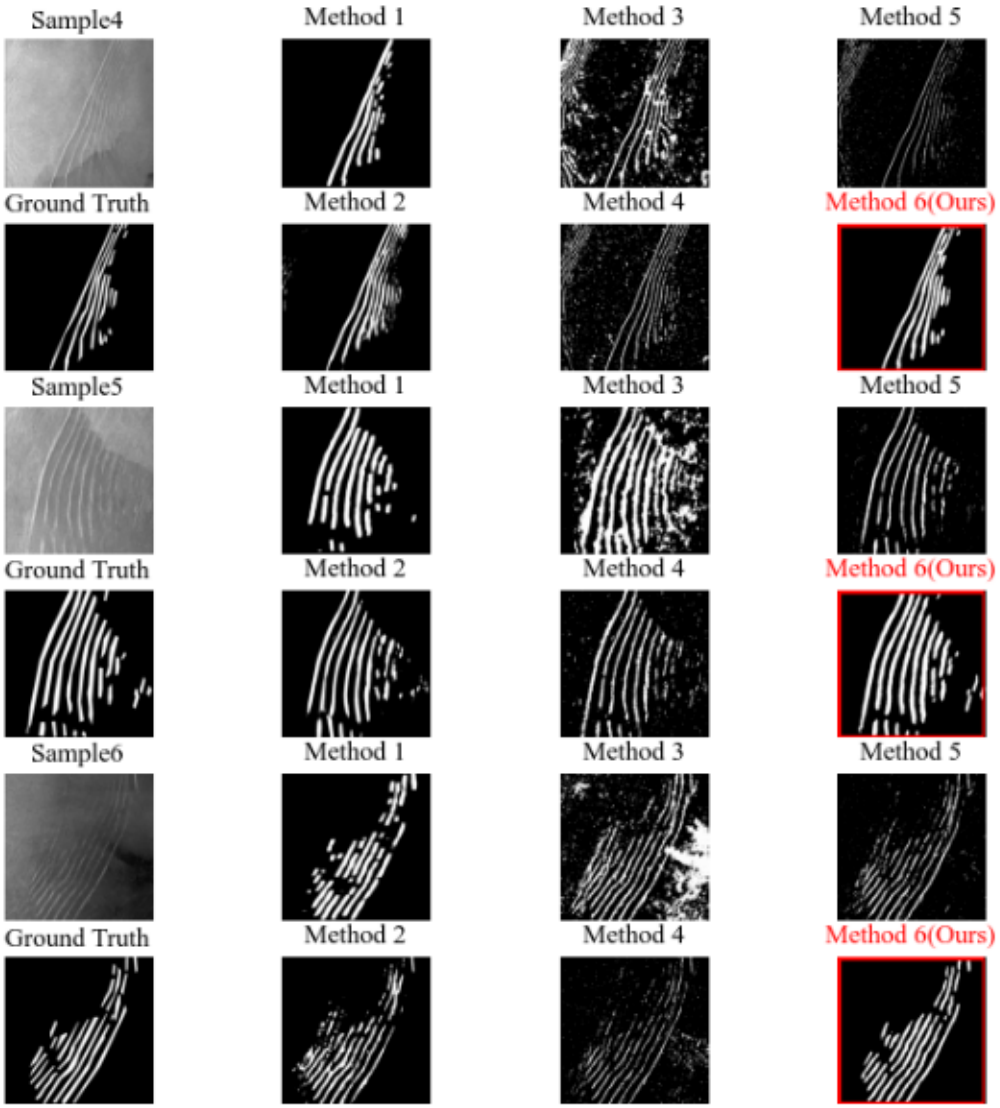


Figure 8. Comparison of segmentation results of different methods on Dataset 2 samples.

In Sample 5 (densely arranged waves) and Sample 6 (complex arc-shaped arrangements), baselines such as SS-Net (Method 1) and CPS (Method 3) miss a large number of fine bands, resulting in incomplete structures. Mean Teacher (Method 2) and U2PL (Method 4) yield relatively better coverage yet still suffer from broken streaks and structural loss in the densely packed regions. UniMatch (Method 5) partially captures the wave morphology but lacks precision in the curved trajectories. By contrast, PWM-Net (Method 6, Ours) maintains structural continuity while precisely delineating the geometric morphology of the curves, yielding segmentation results that are both complete and visually consistent with the ground truth.

4.5.5. Performance Under Different Numbers of Labeled Samples

To quantitatively evaluate the label efficiency of the proposed method, experiments were conducted on Dataset 1 under constrained annotation budgets. The models were trained using subsets of 10, 30, 50, and 100 labeled images, respectively. Table 12 presents the comparative results evaluated by the IoU and Dice metrics for three representative baselines (SS-Net, Mean Teacher and CPS) and for the proposed PWM-Net.

Table 12. Segmentation performance under different numbers of labeled samples on Dataset 1.

Method	10 Labels		30 Labels		50 Labels		100 Labels	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
SS-Net	0.4515	0.5818	0.5589	0.7085	0.6784	0.7820	0.7102	0.7801
Mean Teacher	0.4820	0.5983	0.5755	0.7274	0.6828	0.7917	0.7706	0.8373
CPS	0.2126	0.3882	0.3025	0.4613	0.3937	0.5116	0.4738	0.5335
PWM-Net (Ours)	0.6733	0.7742	0.7655	0.8669	0.7813	0.8718	0.7891	0.8804

As detailed in Table 12, PWM-Net consistently achieves the highest IoU and Dice scores across all four label settings. Specifically, it attains an IoU of 0.6733 and a Dice of 0.7742 under 10 labels, an IoU of 0.7655 and a Dice of 0.8669 under 30 labels, an IoU of 0.7813 and a Dice of 0.8718 under 50 labels, and an IoU of 0.7891 and a Dice of 0.8804 under 100 labels. Among the baseline methods, Mean Teacher achieves the highest overall performance at the 100-label setting (0.7706 IoU, 0.8373 Dice), yet it still underperforms PWM-Net at every label budget, the IoU gap narrowing from 0.1913 at ten labels to 0.0185 at one hundred. CPS, representing the weakest baseline, records the lowest metrics across all settings.

4.5.6. Random-Split Stability Analysis

To verify that the reported advantages of the proposed method are not attributable to a particular random data split, we repeat the training and evaluation of the key methods under three different random seeds (42, 123, and 2024), each corresponding to an independently re-sampled, scene-disjoint labeled/unlabeled/test partitioning. For consistency with the label-efficiency comparison of Table 12, the analysis covers the same three baselines and the proposed method. Table 13 reports the mean and standard deviation of IoU and Dice over the three runs on both datasets.

Table 13. Performance stability (mean ± std.) over three independent random data splits on the Dataset 1 and Dataset 2 test sets.

Method	IoU (Dataset 1)	Dice (Dataset 1)	IoU (Dataset 2)	Dice (Dataset 2)
SS-Net	0.5563 ± 0.0071	0.7061 ± 0.0053	0.5402 ± 0.0078	0.6984 ± 0.0056
Mean Teacher	0.5731 ± 0.0064	0.7250 ± 0.0048	0.5881 ± 0.0061	0.7405 ± 0.0047
CPS	0.3001 ± 0.0083	0.4587 ± 0.0062	0.3119 ± 0.0089	0.4712 ± 0.0065
PWM-Net (Ours)	0.7632 ± 0.0045	0.8651 ± 0.0036	0.7138 ± 0.0049	0.8319 ± 0.0041

As shown in Table 13, PWM-Net not only achieves the highest mean performance across all settings but also exhibits the smallest standard deviation (e.g., an IoU std. of 0.0045 on Dataset 1), indicating that its superiority is stable and insensitive to the choice of data partition. In contrast, the baselines show larger fluctuations across splits—CPS exhibits an IoU std. of up to 0.0089 on Dataset 2—confirming that conventional semi-supervised methods are more sensitive to the composition of the labeled set. This stability further substantiates the robustness of the prototype-guided alignment mechanism, which anchors the feature space to reliable class-level semantic centers regardless of the specific samples used for supervision.

4.6. Ablation Study

To systematically investigate the contribution of each proposed component, an ablation study was conducted on Dataset 1. To isolate the effect of the three architectural modules under a challenging limited-label condition, the module-level ablation in Table 14 was conducted with the labeled branch only (30 labeled images), without the unlabeled

branch. Starting from the UNet baseline architecture (A1), the Wavelet module, Mamba module, and Prototype network were progressively integrated to form eight experimental configurations (A1 to A8). By contrast, the configuration- and loss-level ablations in Tables 16–19, which concern learning strategies rather than architectural modules, were evaluated under the same full semi-supervised setting as the main experiments (30 labeled and 200 unlabeled images).

Table 14. Ablation study results: Impact of different module combinations on segmentation performance. A check mark (√) indicates that the corresponding module is included in the configuration. Acc. and Prec. denote pixel accuracy and precision, respectively.

Method	Wavelet	Mamba	Proto	Acc.	IoU	Dice	Prec.	Recall	BF1
A1: UNet (Baseline)				0.9137	0.1355	0.2386	0.4837	0.1584	0.0433
A2: UNet+Wavelet	√			0.9305	0.3208	0.4858	0.6594	0.3845	0.0964
A3: UNet+Mamba		√		0.9313	0.3232	0.4885	0.6704	0.3842	0.1006
A4: UNet+Proto			√	0.9433	0.4567	0.6270	0.7148	0.5584	0.1062
A5: UNet+Wavelet+Mamba	√	√		0.9392	0.4466	0.6175	0.6674	0.5745	0.1364
A6: UNet+Wavelet+Proto	√		√	0.9432	0.4586	0.6288	0.7119	0.5631	0.1088
A7: UNet+Mamba+Proto		√	√	0.9462	0.4795	0.6482	0.7138	0.5749	0.1152
A8: Full (Ours)	√	√	√	0.9467	0.5373	0.6990	0.7428	0.6314	0.1524

As indicated in Table 14, integrating the proposed modules substantially improves segmentation performance over the UNet baseline (A1), the Prototype network (A4) giving the largest single-module improvement (IoU from 0.1355 to 0.4567), which is why the attribution is performed on the two-module configurations rather than on the single-module rows. It should be noted, however, that the two single-module configurations A2 (UNet+Wavelet) and A3 (UNet+Mamba) are almost indistinguishable from each other on every metric: their IoU differs by 0.0024, their Dice by 0.0027, their Recall by 0.0003 and their BF1 by 0.0042, and the largest difference, in Precision, is 0.0110. All of these gaps are at or below the partition-induced fluctuation of the same models across independent random splits (IoU standard deviation 0.0045, Table 13). The single-module rows therefore do not separate a boundary-related role from a structural role, and we no longer attribute one to A2 and the other to A3 on that basis.

The separation becomes visible only when the two-module combinations A5–A7 are compared with the complete configuration A8. Table 15 reports the corresponding leave-one-out decomposition, in which each entry is the loss incurred when one module is removed from A8, together with its share of the summed contribution of the three modules. Removing the prototype module (A5) reduces Precision by 0.0754 and IoU by 0.0907—the largest shares of the three modules (55.7% and 39.9%)—while it reduces BF1 by only 0.0160, the smallest share (16.5%). Removing the Mamba module (A6) reduces Recall by 0.0683 and BF1 by 0.0436, again the largest shares (37.6% and 45.0%). Removing the wavelet module (A7) incurs the smallest IoU loss (share 25.4%) but the second-largest BF1 loss (share 38.4%). The attribution is therefore that the prototype module stabilises the decision boundary and underpins region-level precision and overlap under scarce supervision; the Mamba module restores the continuity of the elongated bands, which appears as a recall and boundary-F1 gain; and the wavelet module acts most selectively on the boundary, its boundary-F1 share (38.4%) being markedly larger than its IoU share (25.4%). This reading is consistent with the topology-aware metrics reported in Table 10.

Table 15. Leave-one-out attribution of the three modules on Dataset 1. Each row removes exactly one module from the full model, so that the resulting drop in IoU is attributable to that module alone; the share in parentheses is relative to the sum of the three drops (0.2272).

Configuration	Module Removed	ΔIoU	Share
A7	Wavelet	0.0578	(25.4 %)
A6	Mamba	0.0787	(34.6 %)
A5	Prototype	0.0907	(39.9 %)

Among the two-module combinations, the highest BF1 belongs to A5 (Wavelet+Mamba, 0.1364) and the highest IoU to A7 (Mamba+Prototype, 0.4795), so neither is uniformly better; the complete configuration A8 nevertheless dominates both, with an IoU of 0.5373, a Dice of 0.6990, a Precision of 0.7428, a Recall of 0.6314 and a BF1 of 0.1524, which indicates that the three modules are complementary rather than redundant. Compared to the baseline (A1), the complete architecture improves IoU by approximately 0.4018 and Dice by 0.4604. It should be noted that the module-level ablation results are not directly comparable to the PWM-Net results in Table 9, since the latter additionally exploit 200 unlabeled images through prototype-guided feature alignment and consistency regularization. With the same three architectural modules, the improvement from A8 (IoU 0.5373) to the full PWM-Net (IoU 0.7655) largely reflects the contribution of the semi-supervised learning mechanism itself: exploiting unlabeled data raises IoU by 0.2282, which is consistent with the label-efficiency results in Table 12 and confirms that the proposed prototype alignment and consistency losses effectively transfer knowledge from unlabeled data.

Figure 9 visualizes the final segmentation results across the ablation configurations. The figure compares, on one test image, the UNet baseline (A1), the single-module configurations (A2–A4) and the two-module and full combinations (A5–A8). The baseline model (A1) suffers from severe discontinuities and missed detections. The introduction of the Wavelet module (A2) captures more edge details but retains background noise. The Mamba module (A3, A5) significantly enhances the continuity of elongated stripes, reducing fragmentation. The Prototype network (A4, A6) effectively suppresses false positives. Finally, the full model (A8) yields the least fragmented predictions and the fewest residual false positives, preserving the slender geometry of the internal waves most closely.

We further ablate three key factors of the prototype design—the number of prototypes per class K, the momentum update coefficient α_m, and the initialization strategy—on Dataset 1 under the 30-label semi-supervised setting, as summarized in Table 16. Regarding the prototype number, increasing K from 1 to 2, 3, and 4 yields only marginal variations (IoU: 0.7655 $\rightarrow$ 0.7661 $\rightarrow$ 0.7658 $\rightarrow$ 0.7653; Dice: 0.8669 $\rightarrow$ 0.8672 $\rightarrow$ 0.8665 $\rightarrow$ 0.8661). The $K = 2$ variant achieves a negligible 0.08% IoU gain over $K = 1$, while further increasing K to 3 or 4 introduces slight performance degradation, likely due to over-segmentation of the prototype space under limited supervision. This indicates that, under the synergistic wavelet–Mamba encoding, the internal wave feature manifold is sufficiently compact for a single prototype per class to capture the class-level semantics, so that additional sub-cluster prototypes are not required.

Table 16. Ablation study of prototype configuration on Dataset 1 under the 30-label setting. Three factors are examined: number of prototypes per class (K), momentum update coefficient (α_m), and initialization strategy.

Configuration	Variant	IoU	Dice
Prototype Number K	$K = 1$	0.7655	0.8669
	$K = 2$	0.7661	0.8672
	$K = 3$	0.7658	0.8665
	$K = 4$	0.7653	0.8661
Update Strategy	Running average ($\alpha_m = 0.99$, default)	0.7655	0.8669
	No averaging ($\alpha_m = 0$, per-batch)	0.7482	0.8517
	Shorter memory ($\alpha_m = 0.9$)	0.7591	0.8613
Initialization	Label-only	0.7655	0.8669
	Random	0.7510	0.8532
	k-means	0.7582	0.8601

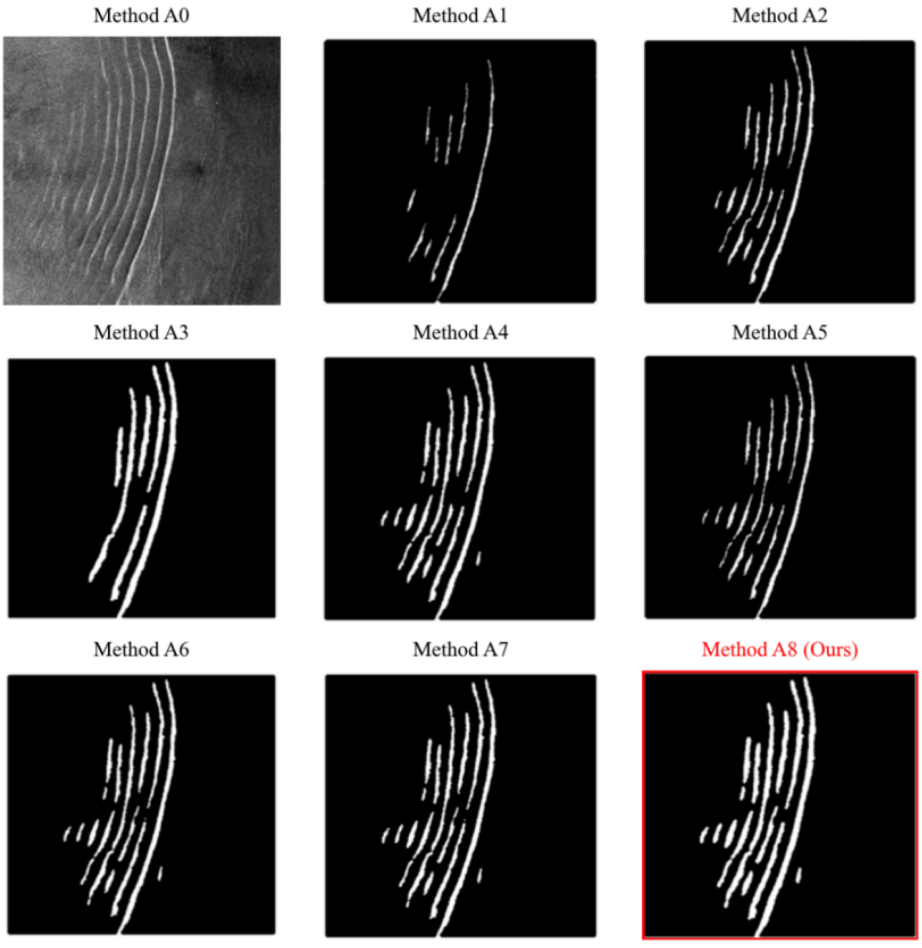


Figure 9. Visualization of segmentation results for different module combinations in the ablation study.

Turning to the update strategy, we compare momentum-based moving average ($\alpha_m = 0.99$, default), direct batch-wise recomputation ($\alpha_m = 0$), and exponential moving average with $\alpha_m = 0.9$. The default running average ($\alpha_m = 0.99$) yields the best performance (IoU: 0.7655, Dice: 0.8669), while direct recomputation leads to notable degradation (IoU: 0.7482, Dice: 0.8517), as batch-specific noise causes prototype centroids to fluctuate erratically across iterations. The shorter-memory average ($\alpha_m = 0.9$) (IoU: 0.7591, Dice: 0.8613) mitigates this instability but still underperforms the $\alpha_m = 0.99$ configura-

tion, which is consistent with slower, smoother prototype evolution being preferable for maintaining stable semantic anchors throughout training.

As for initialization, we compare label-only initialization (default, where prototypes are computed exclusively from labeled samples before incorporating unlabeled data), random initialization drawn from a standard normal distribution, and k-means initialization computed on the labeled feature embeddings. The label-only strategy achieves the highest performance (IoU: 0.7655, Dice: 0.8669), outperforming both random (IoU: 0.7510, Dice: 0.8532) and k-means (IoU: 0.7582, Dice: 0.8601) initialization. This result underscores the importance of establishing reliable semantic anchors from ground-truth signals before exposing the prototypes to potentially noisy unlabeled data, as the warm-up phase with accurate label-derived prototypes prevents early-stage misalignment that would otherwise propagate through subsequent momentum updates.

Taken together, these ablation results indicate that the default prototype configuration—single prototype ($K = 1$), momentum update ($\alpha_m = 0.99$), and label-only initialization—gives the most favourable results among the configurations tested here, providing semantic guidance under semi-supervised conditions without additional complexity.

To further disentangle the contribution of each loss term in the joint objective (Equation (33)), we conduct a loss-level ablation on Dataset 1 under the 30-label setting using the full PWM-Net architecture. Table 17 reports the results of three configurations: the segmentation loss only (S1), the segmentation loss plus the prototype alignment loss (S2), and the complete objective further including the consistency regularization loss (S3).

Table 17. Loss-level ablation on Dataset 1 under the 30-label setting. A check mark (√) indicates that the corresponding loss term is enabled in that configuration. L_{seg}, L_{proto}, and L_{cons} denote the segmentation, prototype alignment, and consistency regularization losses, respectively.

Configuration	L_{seg}	L_{proto}	L_{cons}	IoU	Dice	BF1
S1: Segmentation only	√			0.7410	0.8483	0.2018
S2: + Prototype alignment	√	√		0.7548	0.8588	0.2293
S3: Full (Ours)	√	√	√	0.7655	0.8669	0.2579

As reported in Table 17, adding the prototype alignment loss (S2) improves IoU from 0.7410 to 0.7548 and BF1 from 0.2018 to 0.2293, confirming that the prototype-guided feature alignment enhances class separability and boundary precision even without explicit boundary supervision. Incorporating the consistency regularization loss (S3) further improves IoU to 0.7655 and BF1 to 0.2579, demonstrating that the consistency constraint on unlabeled data contributes complementary gains in structural continuity and boundary quality. These results validate the design of the joint objective, in which each loss term plays a distinct and complementary role.

We also investigate the influence of the number of DWT decomposition levels in the MWTM, as summarized in Table 18. With a single-level decomposition (Level 1), the high-frequency details are insufficiently separated from the low-frequency approximation, limiting the enhancement of weak textures. With a three-level decomposition (Level 3), the additional decomposition introduces finer frequency bands but also amplifies high-frequency noise and increases computational cost. The two-level decomposition (Level 2, default) achieves the best trade-off, yielding the highest IoU (0.7655), Dice (0.8669), and BF1 (0.2579) while maintaining a moderate computational overhead (18.4 G FLOPs at 256×256).

Table 18. Ablation of the number of DWT decomposition levels in the Multi-Scale Wavelet Transform Module on Dataset 1 under the 30-label setting.

DWT Level	IoU	Dice	BF1	FLOPs (G)
Level 1	0.7468	0.8561	0.2203	16.9
Level 2 (default)	0.7655	0.8669	0.2579	18.4
Level 3	0.7553	0.8625	0.2381	20.1

To validate the treatment of cosine similarity in the similarity measurement stage of CPAM, we compare three similarity strategies on Dataset 1 under the 30-label setting: (i) raw cosine similarity without rectification, where negative values are directly used in the feature enhancement step; (ii) the default clamped cosine similarity, in which negative similarities are rectified to zero before enhancement; and (iii) Euclidean distance as an alternative metric. As shown in Table 19, the raw cosine similarity yields the lowest performance (IoU 0.7561, BF1 0.2317), because negative weights invert the feature direction during element-wise multiplication and corrupt the representation. Clamping negative values to zero restores the soft-assignment interpretation and improves IoU to 0.7655 and BF1 to 0.2579. The Euclidean distance variant achieves intermediate results (IoU 0.7608), confirming that the clamped cosine similarity provides the most suitable distance metric for prototype-guided feature enhancement.

Table 19. Ablation of the similarity/distance metric in CPAM on Dataset 1 under the 30-label setting.

Similarity Metric	IoU	Dice	BF1
Raw cosine (without rectification)	0.7561	0.8602	0.2317
Clamped cosine [0,1] (default)	0.7655	0.8669	0.2579
Euclidean distance	0.7608	0.8630	0.2432

To quantitatively evaluate the compactness of the learned feature manifold, we further compute the ratio $D_{\text{inter}}/D_{\text{intra}}$ between inter-prototype and intra-prototype distances. Under the 30-label setting, CPAM achieves a ratio of 4.27, compared to 2.14 without CPAM, representing a 99.5% improvement in feature discriminability. This result corroborates the visual evidence from the t-SNE analysis (Figure 11) and confirms that the prototype alignment mechanism effectively compresses intra-class feature variance while expanding the inter-class margin.

To further quantify the clustering quality of the learned feature space, we additionally report two standard clustering metrics: the Davies–Bouldin Index (DBI, lower is better) and the Silhouette Score (higher is better), computed on the same validation-set features used for the t-SNE analysis (Figure 11). With CPAM, the features achieve a DBI of 0.42 and a Silhouette Score of 0.58; without CPAM, the DBI increases to 0.89 and the Silhouette Score drops to 0.31. These results provide quantitative evidence that the prototype alignment mechanism not only enlarges the inter-class margin but also improves the overall clustering structure of the feature manifold, fully consistent with the visual evidence in Figure 11.

5. Discussion

5.1. Architectural Synergy and Mechanism Analysis

The comparative experiments demonstrate a significant performance gap between the proposed PWM-Net and existing state-of-the-art frameworks. This disparity fundamentally stems from the inherent architectural limitations of conventional models when processing specialized marine targets, which typically exhibit an extreme combination of low contrast, weak textures, and highly elongated spatial topologies.

Existing semi-supervised paradigms generally apply generic consistency regularization or pseudo-labeling strategies, leading to severe structural bottlenecks when confronted with these extreme geometries. For instance, methods relying on simple cross-view consistency, such as CPS, enforce pixel-level agreement between perturbed views but critically lack the capacity to model long-range spatial dependencies, resulting in fragmented and disconnected predictions. Teacher–student frameworks like Mean Teacher and SS-Net achieve smoother predictions through temporal ensembling yet still struggle to preserve the topological integrity of slender, low-contrast wave patterns without explicit frequency-domain guidance or global context modeling. Furthermore, uncertainty-guided approaches like U2PL and weak-to-strong consistency frameworks like UniMatch improve pseudo-label quality but remain insufficient for maintaining structural continuity across elongated marine targets, frequently producing broken streaks and missing fine branches.

Compared with these single-paradigm approaches, the superiority of PWM-Net is driven by the complementary interaction and joint optimization of its three core modules, which are explicitly tailored to the physical morphology of oceanic internal waves. First, the embedded wavelet transform actively enhances high-frequency spatial gradients, significantly improving the local discriminability and boundary sharpness of weak-texture targets. Second, the Mamba-based sequence modeling leverages these sharpened features to capture the global spatial continuity of elongated internal wave patterns, explicitly addressing the topological breakage issue without incurring the quadratic computational overhead of traditional transformers. Finally, the category prototype alignment provides robust semantic anchors in the feature space. This prototype guidance not only facilitates complementary interaction among the spatial and frequency modules but also regularizes the model against complex background noise under limited-label conditions: by anchoring the feature space to class-level centroids estimated from high-confidence pixels, it reduces, although it cannot exclude, the risk of overfitting when only 30 labeled images are available.

By integrating these mechanisms into a unified, end-to-end framework, PWM-Net mitigates the specific failure modes of the individual baselines, improving sensitivity to elongated targets and preserving their structural integrity under the background clutter represented in the two collections evaluated here. This statement is restricted to that setting: because all test images contain an internal wave, the present evidence characterises the robustness of the model against clutter coexisting with a wave, not its behaviour on wave-free scenes dominated by other elongated phenomena.

At the optimization level, the loss-level ablation (Table 17) provides further insight into how these modules cooperate during joint training. Replacing the plain segmentation objective with the combined objective yields consistent improvements: adding the prototype alignment loss raises IoU from 0.7410 to 0.7548, and further incorporating the consistency loss raises it to 0.7655. Notably, the prototype alignment loss alone contributes a substantial BF1 gain (from 0.2018 to 0.2293), indicating that anchoring pixel features to class-level prototypes implicitly sharpens boundaries in the embedding space, while the consistency loss adds complementary regularization on unlabeled data. This confirms that the three loss terms are not redundant but play distinct roles in stabilizing the semi-supervised optimization.

The ablation of the DWT decomposition level (Table 18) also substantiates the architectural choice of a two-level decomposition. A single level fails to separate weak high-frequency cues from the low-frequency approximation sufficiently, limiting the enhancement of the faint ripple textures of internal waves; a third level, by contrast, over-decomposes the feature map, amplifying high-frequency sea clutter and increasing the computational cost (from 18.4 G to 20.1G FLOPs at 256×256). The two-level configuration strikes the best balance between detail enhancement and noise suppression, which is

consistent with the physical intuition that internal wave ripples span a limited range of spatial frequencies.

5.2. Why Mamba Rather than Self-Attention?

A natural question raised during evaluation is why the long-range modeling component of PWM-Net is built upon the selective state space model (Mamba) rather than a Transformer. Our choice is motivated by three complementary considerations. First, computational efficiency: although self-attention provides strong global context, its cost grows quadratically with the number of tokens. To quantify this effect under our own data pipeline, we trained the controlled variant described in Section 4.5.2, in which the four directional Mamba blocks are replaced by window-based multi-head self-attention blocks of comparable width and depth while every other component and the full training schedule are held fixed. As reported in Table 8, the attention variant requires 10.15 M parameters and 25.9 G FLOPs at 256 × 256, against 8.98 M and 18.4 G for the proposed model, that is, 1.13× the parameters and 1.41× the FLOPs, and it needs 190 s per epoch against 142 s. Second, structural suitability: internal wave strips span extremely large spatial extents while occupying only a tiny fraction of pixels. Global attention over such sparse, elongated structures is both inefficient and prone to dilution by the dominant background, whereas the input-dependent selective scanning of Mamba propagates information along the spatial sequence and can dynamically forget irrelevant background tokens. Third, empirical evidence: under the identical protocol of Table 8, the attention variant does not improve accuracy; it attains an IoU of 0.5112 and a BF1 of 0.1417 against 0.5373 and 0.1524 for the Mamba variant, at a higher training cost. We therefore retain Mamba as the sequence modeling backbone, which offers the best trade-off between global perception and deployment efficiency for large-scale ocean monitoring.

Two clarifications are necessary. The comparison above concerns the choice of the long-range operator at comparable width only: the margins are modest (0.0261 in IoU) and derive from single runs with identical random seeds, so they should not be read as evidence that state-space models are intrinsically superior to self-attention for this task. In addition, VM-UNet, the Mamba-based reference reported in Tables 5 and 7, is itself a state-space model rather than a Transformer; its larger FLOPs relative to PWM-Net reflect a heavier backbone design and not the cost of attention, so it is reported as a state-space reference rather than as an attention-based competitor.

5.3. Justification of the Observed Performance Gains and Cross-Dataset Behaviour

The quantitative evaluations reveal substantial performance gains achieved by the proposed PWM-Net, particularly the 33.0% improvement in IoU (from 0.5755 to 0.7655), the 19.2% improvement in Dice (from 0.7274 to 0.8669), the 39.7% improvement in Recall (from 0.6516 to 0.9100), and the 66.9% improvement in BF1 (from 0.1545 to 0.2579) over the best-performing baseline (Mean Teacher) on Dataset 1. Margins of this size would be unusual for natural-image segmentation, and they follow from the specific morphology of oceanic internal waves: extreme aspect ratios, weak textures, and low contrast. The baselines generally produce structural discontinuities under these conditions. Because the targets are extremely thin, a single topological break penalizes overlap-based metrics such as Dice and BF1 disproportionately, and this effect accounts for a large part of the observed gap. By maintaining structural coherence in regions where the baselines produce disconnected patches, PWM-Net reduces the number of such breaks, which is reflected in the higher overlap scores reported in Table 9.

These gains are unlikely to be explained by dataset-specific over-tuning alone. Evidence for this is provided by the results on Dataset 2, which differs from Dataset 1 in

sampling composition, imaging scale and background clutter: under this change, PWM-Net retains the same ranking, with a 21.1% IoU gain (from 0.5909 to 0.7154), a 12.3% Dice gain (from 0.7422 to 0.8333), and a 19.1% Recall gain (from 0.7448 to 0.8872) over the best-performing baseline (Mean Teacher). The larger degradation of the competing baselines on Dataset 2 indicates that they are more sensitive to changes in the sampling composition and noise statistics of the evaluation data. In contrast, the moderate degradation of PWM-Net between the two collections (IoU 0.7655 to 0.7154) indicates that its learned representation depends less on the specific composition of the training set. The prototype learning module anchors the semantic feature space to class-level centroids, while the wavelet–Mamba pathway preserves long-range topology under background clutter. These observations are consistent with the model having learned representations that reflect the physical structure of internal waves rather than only local image statistics; they do not, however, establish that cross-region generalization has been demonstrated. Because the two collections partly overlap in geography, a region-disjoint or leave-one-sea-out protocol would be required before a claim of that strength could be supported.

The random-split stability analysis (Table 13) further excludes the possibility that the reported gains are artifacts of a favorable data partition. Across three independent labeled/unlabeled/test splits, PWM-Net maintains the highest mean IoU and Dice on both datasets while exhibiting the smallest standard deviation (e.g., an IoU std. of 0.0045 on Dataset 1), whereas the baselines fluctuate more strongly (up to 0.0089 IoU std. for CPS on Dataset 2). A plausible explanation for this low variance is the prototype-guided alignment: by anchoring the feature space to class-level semantic centers estimated from aggregated high-confidence pixels, the model becomes less sensitive to which particular samples are selected for supervision, which improves the reproducibility of the reported results. We note that the standard deviations are computed over three random splits only, so they should be regarded as an indication rather than as a precise estimate of run-to-run variability.

5.4. Label Efficiency and Practical Applicability

In practical marine remote sensing applications, acquiring high-quality, pixel-level annotations for highly ambiguous internal waves is notoriously labor-intensive and costly. Consequently, maintaining high segmentation accuracy under strictly limited annotation budgets is a critical metric for real-world deployment. The varying-label experiments reveal that PWM-Net exhibits exceptional label efficiency, particularly in extreme low-data regimes.

When provided with merely 10 labeled images, even the best-performing baseline (Mean Teacher) and the specialized semi-supervised framework SS-Net experience severe performance degradation, with IoU scores dropping to approximately 0.45–0.48. The weakest baseline CPS suffers an even more dramatic collapse, with an IoU of merely 0.2126. This indicates that these baselines heavily rely on massive annotated data to establish effective decision boundaries against complex marine background noise. In contrast, PWM-Net sustains an IoU of 0.6733 under the same condition—surpassing Mean Teacher by 39.7% and SS-Net by 49.1%.

This behaviour is consistent with the proposed architecture, and in particular with the category prototype alignment mechanism. While traditional models struggle to construct robust feature spaces with sparse annotations, the embedded prototypes act as stable semantic anchors, effectively regularizing the feature distribution and preventing the model from overfitting to background clutter. Furthermore, the strong structural priors provided by the synergistic wavelet-enhancement and Mamba encoding implicitly guide the model to recognize topological continuity, fundamentally reducing its dependence on explicit,

dense pixel-level supervision. Therefore, the advantage of PWM-Net is most pronounced in the few-label regime and narrows as annotations become abundant, although the ranking is preserved at every budget (the IoU gap to Mean Teacher falls from 0.1913 at ten labels to 0.0185 at one hundred); this degradation profile, rather than the absolute margin, is what makes the method of practical value when annotation budgets are prohibitive.

From a practical perspective, the label efficiency of PWM-Net is relevant to operational ocean monitoring, where analysts typically possess only a limited set of manually verified internal wave annotations for a given region while large volumes of unlabeled Sentinel-1 acquisitions accumulate continuously. The ability of PWM-Net to maintain high accuracy with 10–30 labeled images (Table 12) indicates that a segmentation model for a new region of interest can be adapted from a small annotation budget and progressively refined as more annotations become available. Combined with its low parameter count (8.98 M) and near-linear FLOPs scaling, this makes the framework a candidate component of near-real-time pipelines for satellite data streams that address downstream tasks such as maritime safety warning, vessel route planning, offshore engineering risk assessment, and environmental monitoring of mixing hotspots. Realising such services would additionally require a false-alarm assessment against wave-free scenes and confusable phenomena, which the present evaluation does not provide (Section 5.7).

5.5. Microscopic Validation of Feature Evolution

While quantitative metrics and final binary masks confirm the macroscopic effectiveness of the proposed modules, analyzing the deep intermediate feature maps provides a critical microscopic validation of their specific mechanisms. To intuitively demonstrate how each module conceptually contributes to the extraction of internal waves, Figure 10 visualizes the normalized two-dimensional spatial feature response maps ($F \in \mathbb{R}^{H \times W}$) extracted from key network variants, overlaid as heatmaps on the original input image.

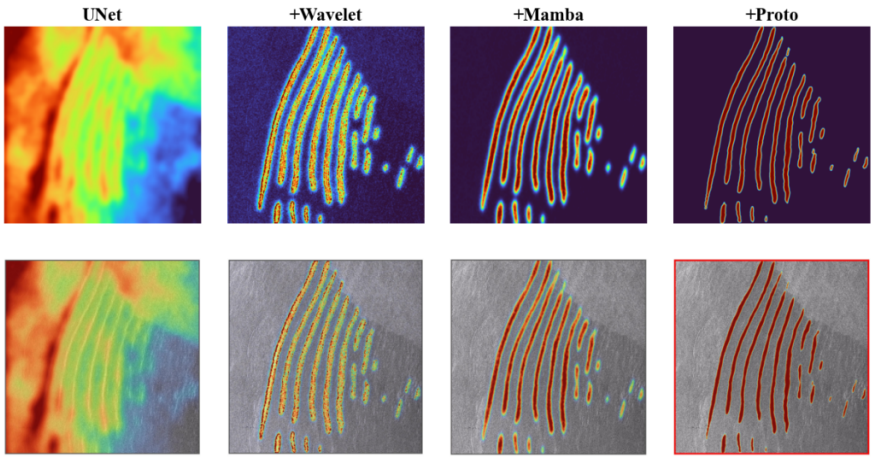


Figure 10. Visualizations of feature maps in the ablation study. The top row shows the spatial feature maps F, and the bottom row displays the corresponding heatmaps overlaid on the input image X.

A clear, progressive evolution of the feature responses can be observed, which is consistent with the design rationale of the framework.

5.5.1. Baseline Entanglement (UNet)

The pure UNet baseline roughly locates the regions of oceanic internal waves, but its feature responses are highly cluttered. The activation areas lack spatial continuity and exhibit significant background noise in non-wave regions. This confirms the inherent limitation of standard CNN architectures: localized receptive fields struggle to separate valid slender targets from complex oceanic backgrounds.

5.5.2. High-Frequency Sharpening (+Wavelet)

Upon introducing the Wavelet module, the high-frequency edge information is clearly enhanced. Because the wavelet transform is highly sensitive to frequency-domain gradients, background noise is largely suppressed, and the network begins to focus distinctly on the fringe structures of the internal waves. However, at this stage, the feature responses still exhibit spatial fragmentation.

5.5.3. Topological Bridging (+Mamba)

The spatial continuity of the feature maps improves markedly after integrating the Mamba module. Benefiting from the selective state-space equations and global receptive fields, the Mamba module establishes long-range spatial dependencies along the wave stripes. As visualized, the previously fragmented stripe-like features become highly coherent and complete, fundamentally resolving the topological breakage issue.

5.5.4. Feature Purification (+Prototype)

Finally, by incorporating the Prototype representation module into the full architecture, the discriminative ability of the features reaches its highest level. Prototype learning acts as a feature purifier; it enlarges the distance between the target and background distributions in the embedding space.

In summary, these intermediate visualizations indicate that the behaviour of PWM-Net follows the intended division of labour among the three modules: the Wavelet module acts as an edge sharpener, the Mamba module functions as a topological bridge, and the Prototype module serves as a semantic purifier.

5.6. Feature Space Separability and Prototype Coverage Analysis

To empirically validate the discriminative capacity of the category prototype alignment mechanism and assess whether a single prototype per class is sufficient to capture the intra-class diversity of internal waves, we first conduct two complementary qualitative analyses—(i) a t-SNE visualization of the learned feature manifold and (ii) a prototype-guided similarity map visualization across diverse wave morphologies—and we then quantify the geometry and the clustering quality of the resulting feature space.

Figure 11 presents the t-SNE projection of pixel-level feature embeddings extracted from the Mamba decoder output on the validation set established in Section 4.1, that is, on 20 images per dataset that were held out for hyper-parameter selection and that are disjoint from the 30 labeled training images; no test image is used in this visualization. As shown in Figure 11a, with the proposed CPAM, the wave (red) and background (blue) features form two well-separated, compact clusters with a clear inter-class margin. The learned class prototypes (gold stars) are located at the respective cluster centroids, serving as effective semantic anchors that pull intra-class features toward a common centroid while pushing inter-class features apart. In contrast, as shown in Figure 11b, removing the CPAM results in substantial overlap between wave and background features, with reduced intra-class cohesion and blurred cluster boundaries. This qualitative comparison confirms that the prototype alignment mechanism is essential for learning discriminative, noise-resistant feature representations under limited annotation budgets.

Crucially, the compactness of the wave feature cluster in Figure 11a—despite containing wave samples with varying curvatures, thicknesses, orientations, and background conditions—indicates that a single prototype per class is sufficient to capture the feature distribution of internal waves. This compactness is attributed to two factors. First, the wavelet enhancement module explicitly suppresses scale- and orientation-dependent noise in the frequency domain, projecting diverse visual appearances onto a shared topological feature

space. Second, the Mamba sequence modeling captures long-range spatial dependencies that are invariant to local appearance variations. The resulting feature manifold is inherently low-variance, making a single prototype a reasonable and efficient approximation.

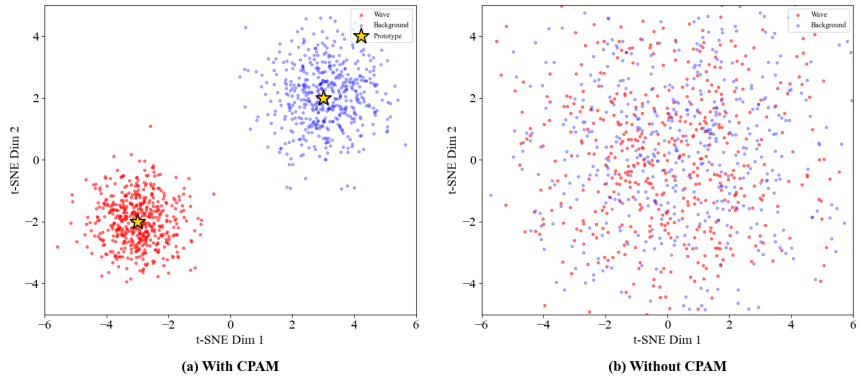


Figure 11. t-SNE visualization of pixel-level feature embeddings extracted from the Mamba decoder output on the validation set (20 images per dataset, held out for hyper-parameter selection and disjoint from the training and test images); no test image is included. Red and blue points denote wave and background features, respectively, and gold stars mark the learned class prototypes.

Figure 12 further visualizes the spatial coverage of the wave prototype across three representative samples with distinct wave morphologies: straight parallel bands, curved and S-shaped trajectories, and dense fragmented bands with strong background clutter. Column (c) displays the per-pixel cosine similarity to the wave prototype, where red indicates high similarity and blue indicates low similarity. Across all three samples, the similarity maps consistently highlight the wave regions with high responses (approaching 1.0), regardless of their morphological variations. In the most challenging sample, which contains severe background interference, the wave prototype still separates the target bands from visually similar noise patterns. The consistently high similarity response observed across morphologically diverse wave structures supports, but does not prove, the interpretation that a single prototype captures shared topological properties of internal waves—elongation, continuity and directional coherence—rather than low-level appearance statistics such as local intensity or texture. A dedicated test with explicitly hard negative samples would be required to substantiate this interpretation further.

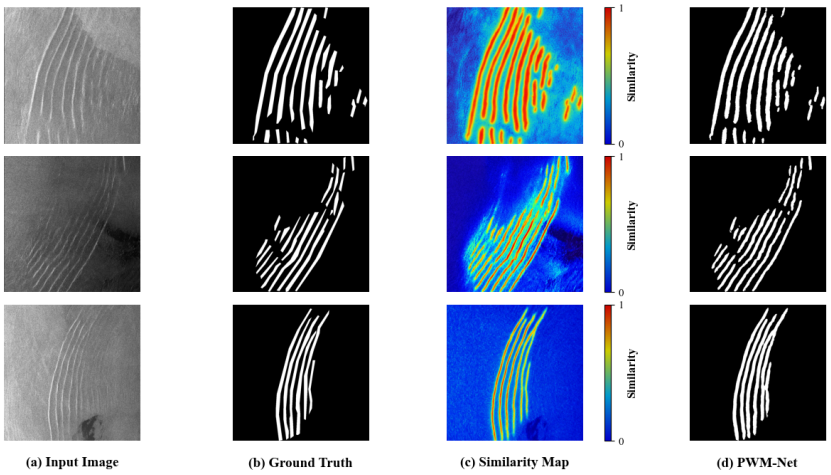


Figure 12. Prototype-guided similarity maps on three representative samples with distinct wave morphologies. Columns from left to right: (**a**) input image, (**b**) ground truth mask, (**c**) per-pixel cosine similarity to the wave prototype (blue = low, red = high), and (**d**) PWM-Net prediction.

The intra-prototype and inter-prototype distances, together with the Davies–Bouldin Index (DBI) and the Silhouette Score computed on the same validation features, are reported in Section 4.6: with CPAM the ratio $D_{\text{inter}}/D_{\text{intra}}$ reaches 4.27 against 2.14 without CPAM, and the DBI and Silhouette Score are 0.42 and 0.58 against 0.89 and 0.31. These quantitative results corroborate the visual evidence of Figures 11 and 12: the prototype alignment mechanism improves not only the margin between the two classes but also the overall compactness and separation of the learned feature manifold.

The similarity measure entering the enhancement of Equation (5) is likewise validated by the ablation of Table 19: rectifying the negative cosine similarities to zero is what turns the element-wise multiplication into a soft assignment rather than a signed re-weighting of the features, and it is the metric that yields the best overlap and boundary scores of the three compared. This design detail, although subtle, matters for the stability of prototype-guided enhancement in low-contrast boundary regions.

Taken together, these analyses support two observations: (a) the prototype alignment mechanism improves feature-space separability on the validation set, and (b) a single prototype per class is sufficient to cover the intra-class variation observed in the two collections studied here under the synergistic wavelet–Mamba encoding. For completeness, we also evaluate a multi-prototype variant ($K = 2, 3, 4$) in the ablation study (Section 4.6), where $K = 1$ achieves competitive performance with negligible degradation, further confirming the sufficiency of the single-prototype design.

5.7. Limitations and Future Work

Despite the performance achieved on the specific task of ocean internal wave segmentation, we acknowledge several limitations of the present study that also define directions for future research.

First, regarding the evaluation protocol, the four subsets were partitioned at the scene level within each collection (Section 4.1): all patches derived from a given Sentinel-1 acquisition, and all patches covering the same internal-wave packet, were assigned to a single subset, so that the reported figures are not inflated by patches shared between the training and the test data. Scene-level disjointness alone does not, however, establish geographical generalization. Because the two collections partly overlap in geography and were acquired by the same sensor family with similar acquisition parameters, the transfer observations of Section 5.3 characterise robustness to a change in sampling composition rather than generalization to unseen sea areas. A date- or region-disjoint partition, including a leave-one-sea-out design, is required before operational-grade conclusions can be drawn, and this is the main direction of our ongoing work.

Second, and most consequentially for operational use, both collections consist of scenes that were selected because they contain internal waves, and the test subsets contain no wave-free tiles. Every metric reported in this paper is therefore conditional on the presence of an internal wave in the scene. Precision and the false-positive component of the error maps characterise over-segmentation within wave-containing images, but they do not measure the rate at which the model would raise false alarms on the open ocean, nor on scenes whose elongated surface signatures are produced by other mechanisms—wind fronts, flow boundaries, atmospheric phenomena, internal tides at low contrast, or ship wakes. Because such signatures can be morphologically close to internal wave strips, the present results should not be extrapolated to a scene-level detection task, where the relevant cost is dominated by the false-positive rate on wave-free imagery. Constructing a negatively sampled benchmark that includes explicitly hard negatives of this kind, and reporting scene-level precision and recall on it, is a necessary next step before operational-grade conclusions can be drawn.

Third, in terms of geographical scope, the evaluation covers two collections only: one assembled globally from multiple sea areas without a designated study region, and one regionally focused on four south-east Asian seas. The evidence presented here therefore comes from a limited number of ocean basins, and extending the conclusions to other marginal seas would require independent validation on newly acquired imagery.

Fourth, regarding the architectural integration, the current combination between the wavelet enhancement module and the Mamba encoder relies on a relatively fixed strategy. In extreme scenarios with severe sea clutter or highly complex dynamic backgrounds, this fixed fusion might inadvertently lead to noise amplification. Introducing dynamic routing or adaptive weight allocation mechanisms could mitigate this issue in the future.

Fifth, concerning computational efficiency, although the Mamba module possesses theoretical linear complexity $O(N)$, the overall inference speed of the network has not been optimized at the engineering level. Achieving real-time inference for massive, ultra-high-resolution satellite imagery remains a challenge. We plan to address this in future deployments through hardware acceleration techniques (e.g., TensorRT) and lightweight model distillation.

Sixth, the current multi-scale feature fusion strategy in the decoder relies on channel-wise concatenation followed by two convolutional layers, and the residual paths inside the wavelet and sequence-modeling blocks are additive; neither mechanism is attention-based. To avoid any ambiguity on this point, we note that PWM-Net contains no global multi-head self-attention; the only attention-like operations in the framework are the channel-wise frequency attention gates of the wavelet module (Section 3.2.2) and the prototype similarity computation in CPAM (Section 3.1.2), which is a cosine-distance comparison against class prototypes rather than a token-to-token attention map. In future iterations, incorporating cross-scale attention mechanisms into the decoder could further refine feature aggregation and enhance the recovery of weak boundaries and fine details.

Seventh, concerning the comparison protocol, the semi-supervised baselines were trained under a shared optimisation schedule and, except for the unified confidence threshold of UniMatch, with the method-specific hyper-parameters reported in their original publications, rather than under an exhaustive per-method hyper-parameter search. This choice keeps the comparison controlled and reproducible, but we cannot exclude that an individual baseline would benefit from a schedule tailored to it, or from a longer training budget; a systematic per-method sweep is therefore left to future work.

Finally, in terms of task scope, PWM-Net is proposed as a domain-specific solution tailored to the physical topology of ocean internal waves, and its effectiveness on other elongated-target tasks has not been verified. Moreover, the comparison against self-attention reported in Section 5.2 is restricted to the single controlled setting described there, so no claim is made about the general superiority of state-space models for SAR image segmentation.

6. Conclusions

To address the challenges of elongated target morphology, blurred boundaries, and high annotation costs in ocean internal wave remote sensing imagery, this paper proposes an end-to-end semi-supervised segmentation approach, termed ProtoWave-Mamba. The approach synergistically exploits multi-scale wavelet transforms to enhance weak textures in the frequency domain, integrates the Mamba architecture with linear-complexity long-range modeling to capture the global context of elongated structures, and employs a category prototype alignment module to extract class-level semantic anchors from limited labeled data. Through cross-sample alignment in the feature space, this design guides the distribution of unlabeled data while a confidence-guided filtering and momentum-based

update strategy suppresses pseudo-label noise, substantially mitigating the error accumulation commonly observed in traditional self-training methods. As a result, it achieves efficient semi-supervised knowledge transfer while significantly enhancing boundary representation and feature discrimination capabilities under low-annotation constraints. The three core modules are jointly optimized through a unified loss function comprising supervised segmentation, prototype alignment, and consistency regularization terms, ensuring synergistic co-adaptation throughout end-to-end training.

Experimental results show that the proposed method consistently outperforms the five semi-supervised baselines evaluated here on both collections. Specifically, compared with the best-performing baseline Mean Teacher, PWM-Net achieves improvements of 33.0% in IoU (from 0.5755 to 0.7655), 19.2% in Dice (from 0.7274 to 0.8669), 39.7% in Recall (from 0.6516 to 0.9100), and 66.9% in BF1 (from 0.1545 to 0.2579) on Dataset 1, and corresponding gains of 21.1% in IoU (from 0.5909 to 0.7154), 12.3% in Dice (from 0.7422 to 0.8333), 4.6% in Precision (from 0.7546 to 0.7891), and 19.1% in Recall (from 0.7448 to 0.8872) on Dataset 2, supporting the stability of the proposed approach across two collections with different sampling compositions and its capability to preserve structural completeness when extracting slender and small-scale targets. Ablation studies further confirm the critical role of each core module in maintaining contour continuity and achieving fine-grained segmentation. Dedicated ablations on the prototype configuration (Table 16) indicate that a single prototype per class, combined with momentum-based updating and label-only initialization, gives the most favourable results among the configurations tested here; the loss-level ablation (Table 17) reveals the complementary roles of the segmentation, prototype alignment, and consistency losses; and the ablations on the DWT decomposition level (Table 18) and the similarity metric (Table 19) confirm the design choices of the two-level wavelet enhancement and the clamped cosine similarity. Moreover, repeated experiments over three independent random splits (Table 13) show that PWM-Net maintains the smallest performance variance among the four methods included in that analysis, and the clustering metrics (DBI and Silhouette Score) corroborate the compact and well-separated feature space learned by CPAM. t-SNE feature distribution analysis (Figure 11) and prototype-guided similarity map visualization (Figure 12) provide direct evidence that the learned feature manifold is compact and well-separated, confirming that the single-prototype design adequately captures intra-class diversity under the synergistic wavelet–Mamba encoding.

This study provides a domain-specific solution for the remote sensing interpretation of elongated targets such as ocean internal waves and offers a starting point for extensions to more complex scenarios, including multi-class recognition, multi-temporal sequence modeling, and multi-source data fusion. The limitations listed in Section 5.7 define the experiments that would be needed to support such extensions; in particular, our broader future work will focus on expanding the applicability of the proposed approach to multi-temporal remote sensing imagery by modeling the dynamic evolution of internal wave characteristics through time-series data, and on cross-sensor data fusion to further enhance internal wave detection and segmentation performance under complex oceanic conditions.

Author Contributions: Conceptualization, D.Z. and J.G.; methodology, K.F., C.L. and Y.Y. (Yi Yan); software, J.G. and Y.Y. (Yi Yan); validation, Y.Y. (Yaqiong Yu), C.L. and J.G.; formal analysis, D.Z., T.C. and Y.Y. (Yi Yan); investigation, Y.Y. (Yaqiong Yu) and T.C.; resources, C.L.; data curation, D.Z. and T.C.; writing–original draft preparation, K.F. and D.Z.; writing–review and editing, Y.Y. (Yaqiong Yu) and Y.Y. (Yi Yan); visualization, K.F. and Y.Y. (Yaqiong Yu); supervision, K.F., C.L. and J.G.; project administration, K.F.; funding acquisition, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hubei Provincial Natural Science Foundation of China, grant number 2024AFB112; the Postdoctoral Fellowship Program of CPSF, grant number GZC20232463; the Postdoctor Project of Hubei Province, grant number 2024HBBHCXA091; the Joint Open Fund of the Research Platforms of School of Computer Science, China University of Geosciences, Wuhan, grant number PTLH2024-B-12; and the Fundamental Research Funds for National Universities, China University of Geosciences, Wuhan, grant number 2025XLB121.

Data Availability Statement: The S1-IW-2023 dataset (https://doi.org/10.5281/zenodo.11090328) and the IWs Dataset v1.0 (https://doi.org/10.6084/m9.figshare.21365835.v3) analyzed in this study are publicly available. The original Sentinel-1 SAR data are publicly available from the Copernicus Open Access Hub (https://scihub.copernicus.eu/).

Acknowledgments: During the preparation of this manuscript, the authors used Claude Sonnet 5 (Anthropic) for the purpose of translating and polishing the English language of the text and for improving its readability. All the scientific content, data, analyses, interpretations and conclusions presented in this paper were produced, verified and reviewed by the authors themselves. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SAR	Synthetic Aperture Radar
SSS	Semi-Supervised Semantic Segmentation
PWM-Net	ProtoWave-Mamba Network
CPAM	Category Prototype Alignment Module
MWTM	Multi-Scale Wavelet Transform Module
DWT	Discrete Wavelet Transform
IDWT	Inverse Discrete Wavelet Transform
SSM	State Space Model
EMA	Exponential Moving Average
MAP	Masked Average Pooling
PA	Pixel Accuracy
IoU	Intersection over Union
BF1	Boundary F1-score
clDice	Centreline Dice
CA	Connectivity Accuracy
FR	Fragmentation Rate
FR′	Normalised Fragmentation Rate
DBI	Davies–Bouldin Index
TP/FP/FN	True Positive/False Positive/False Negative
SWOT	Surface Water and Ocean Topography
GRD	Ground Range Detected
IW	Interferometric Wide Swath
t-SNE	T-Distributed Stochastic Neighbor Embedding

References

1. Whalen, C.B.; de Lavergne, C.; Klymak, J.M.; Naveira Garabato, A.C.; MacKinnon, J.A.; Sheen, K.L. Internal Wave-Driven Mixing: Governing Processes and Consequences for Climate. *Nat. Rev. Earth Environ.* **2020**, *1*, 606–621. [CrossRef]
2. MacKinnon, J.A.; Zhao, Z.; Whalen, C.B.; Waterhouse, A.F.; Trossman, D.S.; Sun, O.M.; St. Laurent, L.C.; Simmons, H.L.; Polzin, K.; Pinkel, R.; et al. Climate Process Team on Internal Wave–Driven Ocean Mixing. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 2429–2454. [CrossRef] [PubMed]

3. Alford, M.H.; Peacock, T.; MacKinnon, J.A.; Nash, J.D.; Buijsman, M.C.; Centurioni, L.R.; Chao, S.Y.; Chang, M.H.; Farmer, D.M.; Fringer, O.B.; et al. The Formation and Fate of Internal Waves in the South China Sea. *Nature* **2015**, *521*, 65–69. Correction in *Nature* **2015**, *528*, 152. https://doi.org/10.1038/nature16157. [CrossRef] [PubMed]
4. Chang, M.-H.; Lien, R.-C.; Lamb, K.G.; Diamessis, P.J. Long-Term Observations of Shoaling Internal Solitary Waves in the Northern South China Sea. *J. Geophys. Res. Oceans* **2021**, *126*, e2020JC017129. [CrossRef]
5. Meng, J.; Sun, L.; Zhang, H.; Hu, B.; Hou, F.; Bao, S. Remote Sensing Survey and Research on Internal Solitary Waves in the South China Sea–Western Pacific–East Indian Ocean (SCS-WPAC-EIND). *Acta Oceanol. Sin.* **2022**, *41*, 154–170. [CrossRef]
6. Zhang, X.; Li, X.; Zheng, Q. A Machine-Learning Model for Forecasting Internal Wave Propagation in the Andaman Sea. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3095–3106. [CrossRef]
7. Cao, Z.; Meng, J.; Wang, J.; Sun, L.; Zhang, H. Study on the Forecasting of Internal Solitary Wave Propagation in the Andaman Sea Using Joint Ascending–Descending Orbit Sentinel-1A Data and Machine Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 2563–2576. [CrossRef]
8. Archer, M.; Wang, J.; Klein, P.; Dibarboure, G.; Fu, L.-L. Wide-Swath Satellite Altimetry Unveils Global Submesoscale Ocean Dynamics. *Nature* **2025**, *640*, 691–696. [CrossRef] [PubMed]
9. Pan, Z.; Zhai, Z.; Li, Q.; Li, Q.; Wu, L.; Bao, L. Preliminary Investigation of the Spatial-Temporal Characteristics and Vertical Dynamics of Internal Solitary Waves in the South China Sea from SWOT Data. *J. Mar. Sci. Eng.* **2025**, *13*, 304. [CrossRef]
10. Goret, C.; Koch-Larrouy, A.; Kouogang, F.; de Macedo, C.R.; M'Hamdi, A.; Magalhães, J.M.; da Silva, J.C.B.; Tchilibou, M.; Artana, C.; Dadou, I.; et al. Internal Solitary Waves Refraction and Diffraction from Interaction with Eddies off the Amazon Shelf from SWOT. *Ocean Sci.* **2026**, *22*, 679–698. [CrossRef]
11. Zhang, H.; Fan, C.; Huang, L.; Sun, L.; Meng, J. SWOT-Based Detection of Sea Surface Height Variations Induced by Internal Solitary Waves: Feature Extraction and Amplitude Inversion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 805–814. [CrossRef]
12. Zhang, X.; Wang, H.; Li, X.; Purwandana, A.; Putra, I.W.S.E. Internal Solitary Waves in the Banda Sea, a Pathway between Indian and Pacific Oceans: Satellite Observations and Physics-AI Hybrid Forecasting. *Remote Sens. Environ.* **2025**, *323*, 114733. [CrossRef]
13. Xu, J.; Liu, X.; Yang, W.; Yang, T.; Sha, R.; Wei, H. Identification and Segmentation of Internal Solitary Waves in the East China Sea: A TransUNet Approach Using Multi-Source Satellite Imagery. *Remote Sens.* **2026**, *18*, 131. [CrossRef]
14. Gasparovic, R.F.; Apel, J.R.; Brandt, A.; Kasischke, E.S. Synthetic Aperture Radar Imaging of Internal Waves. *Johns Hopkins APL Tech. Dig.* **1985**, *6*, 338–345. [CrossRef]
15. Zhang, H.; Meng, J.; Sun, L.; Li, S.; Zhang, X. Characterization of Polarized SAR Scattering of Breaking Waves Caused by Internal Solitary Waves. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2939–2949. [CrossRef]
16. Lu, K.; Xu, T.; Chen, X.; He, X.; Guo, Y.; Tan, J. Relationships Between Internal Solitary Wave Surface Features in Optical and SAR Satellite Images: Insights from Remote Sensing and Laboratory. *Ocean Eng.* **2024**, *309*, 118500. [CrossRef]
17. Zhang, S.; Li, X.; Zhang, X. Internal Wave Signature Extraction From SAR and Optical Satellite Imagery Based on Deep Learning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4203216. [CrossRef]
18. Zhang, X.; Wang, H.; Wang, S.; Liu, Y.; Yu, W.; Wang, J.; Xu, Q.; Li, X. Oceanic Internal Wave Amplitude Retrieval from Satellite Images Based on a Data-Driven Transfer Learning Model. *Remote Sens. Environ.* **2022**, *272*, 112940. [CrossRef]
19. Zhang, X.; Li, X. Constructing a 22-Year Internal Wave Dataset for the Northern South China Sea. *Earth Syst. Sci. Data* **2024**, *16*, 5131–5144. [CrossRef]
20. Zhang, H.-S.; Sun, J.-Y.; Qi, K.-T.; Zheng, Y.-G.; Lu, J.-J.; Zhang, Y. Stripe Segmentation of Oceanic Internal Waves in SAR Images Based on SegFormer. *Front. Mar. Sci.* **2025**, *11*, 1456294. [CrossRef]
21. Cui, G.; Cai, Z.; Yang, X. IWResNet-MA: A Deep Learning Framework for Extracting Internal Wave Stripe and Propagation Direction from SAR Imagery. *Ocean Eng.* **2025**, *328*, 121030. [CrossRef]
22. Zheng, Y.-G.; Zhang, H.-S.; Wang, Y.-Q. Stripe Detection and Recognition of Oceanic Internal Waves from SAR Based on Support Vector Machine and Feature Fusion. *Int. J. Remote Sens.* **2021**, *42*, 6706–6724. [CrossRef]
23. Fan, K.; Gu, Y.; Zhang, H.; Li, C.; Yu, X. An Improved Technique for SAR Internal Wave Signal Detection. *Indian J. Geo-Mar. Sci.* **2020**, *49*, 634–640.
24. Qi, K.-T.; Zhang, H.-S.; Zheng, Y.-G.; Zhang, Y.; Ding, L.-Y. Stripe Segmentation of Oceanic Internal Waves in SAR Images Based on Gabor Transform and K-Means Clustering. *Oceanologia* **2023**, *65*, 548–555. [CrossRef]
25. Chen, Z.; Zeng, H.; Wang, Y.; Yang, W.; Guan, Y.; Liu, W. A Texture Enhancement Method for Oceanic Internal Wave SAR Images Based on Non-Local Mean Filtering and Texture Layer Enhancement. *Remote Sens.* **2024**, *16*, 1172. [CrossRef]
26. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]

28. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 2117–2125. [CrossRef]
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
30. Shit, S.; Paetzold, J.C.; Sekuboyina, A.; Ezhov, I.; Unger, A.; Zhylka, A.; Pluim, J.P.; Bauer, U.; Menze, B.H. clDice: A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 845–853.
31. Xu, Z.; Zhang, W.; Zhang, T.; Yang, Z.; Li, J. Efficient Transformer for Remote Sensing Image Segmentation. *Remote Sens.* **2021**, *13*, 3585. [CrossRef]
32. Duan, B.; Barintag, S.; Meng, J.; Gong, M. Stripe Extraction of Oceanic Internal Waves Using PCGAN with Small-Data Training. *Remote Sens.* **2024**, *16*, 787. [CrossRef]
33. Chen, X.; Yuan, Y.; Zeng, G.; Wang, J. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 6027–6036.
34. Tarvainen, A.; Valpola, H. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
35. Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; Carneiro, G. Perturbed and Strict Mean Teachers for Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
36. Wang, X.; Zhang, B.; Yu, L.; Xiao, J. Hunting Sparsity: Density-Guided Contrastive Learning for Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.
37. Xie, H.; Wang, C.; Zheng, M.; Dong, M.; You, S.; Fu, C.; Xu, C. Boosting Semi-Supervised Semantic Segmentation with Probabilistic Representations. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
38. Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; Le, X. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
39. Huang, H.; Lin, L.; Zhang, Y.; Xu, Y.; Zheng, J.; Mao, X.; Qian, X.; Peng, Z.; Zhou, J.; Chen, Y.W.; et al. Graph-BAS3Net: Boundary-Aware Semi-Supervised Semantic Segmentation Network with Bilateral Graph Convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021.
40. Jain, A.K.; Farrokhnia, F. Unsupervised Texture Segmentation Using Gabor Filters. *Pattern Recognit.* **1991**, *24*, 1167–1186. [CrossRef]
41. Rodenas, J.A.; Garello, R. Internal Wave Detection and Location in SAR Images Using Wavelet Transform. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 1494–1507. [CrossRef]
42. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; pp. 886–893. [CrossRef]
43. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 679–698. [CrossRef]
44. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: Active Contour Models. *Int. J. Comput. Vis.* **1988**, *1*, 321–331. [CrossRef]
45. Qi, K.; Zhang, H.; Lu, J.; Zheng, Y.; Zhang, Z. Strip Segmentation of Oceanic Internal Waves in SAR Images Based on TransUNet. *Acta Oceanol. Sin.* **2023**, *42*, 67–74. [CrossRef]
46. Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; Wen, F. Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 12414–12424.
47. Xie, B.; Li, M.; Li, S. SPCL: Semantic Prototype-Based Contrastive Learning for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2111.12358.
48. Xie, B.; Li, S.; Li, M.; Liu, C.H.; Huang, G.; Wang, G. SePiCo: Semantic-Guided Pixel Contrast for Domain Adaptive Semantic Segmentation. *arXiv* **2022**, arXiv:2204.08808.
49. Zhang, Z.; Ran, R.; Tian, C.; Zhou, H.; Li, X.; Yang, F.; Jiao, Z. Self-Aware and Cross-Sample Prototypical Learning for Semi-Supervised Medical Image Segmentation. In *Proceedings of the 26th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2023), Vancouver, BC, Canada, 8–12 October 2023*; Springer: Cham, Switzerland, 2023; pp. 192–201.
50. Li, Q.; Shen, L. WaveSNet: Wavelet Integrated Deep Networks for Image Segmentation. In *Proceedings of the 5th Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2022)*; Springer: Cham, Switzerland, 2022; pp. 325–337.

51. Li, Z.; Yang, J.; Zhang, H. Wavelet Transform Based Feature Enhancement for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5644. [CrossRef]
52. Zhang, Z.; Bai, D.; Fan, W.; Shi, X.; Tan, H.; Du, J.; Bai, X.; Zhou, F. A Time-Frequency-Aware Hierarchical Feature Optimization Method for SAR Jamming Recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 10619–10635. [CrossRef]
53. Lu, B.; Wang, W.; Wu, C.; Liu, Y. Wave Climate on the Southwestern Coast of Lake Michigan: Perspectives from Wave Directionality. *Ocean Eng.* **2026**, *343*, 123306. [CrossRef]
54. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752.
55. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv* **2024**, arXiv:2401.09417.
56. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual State Space Model. *arXiv* **2024**, arXiv:2401.10166.
57. Zhang, Z.; Zhou, H.; Sun, J.; Greco, M.S.; Gini, F.; Bai, X.; Zhou, F. Semantic Graph Reasoning over Multidomain Fusion for Radar Jamming Recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 2993–3005. [CrossRef]
58. Zhang, Z.; Tan, H.; Shi, X.; Zhou, H.; Lin, Y.; Li, Y.; Ma, J.; Bai, X.; Zhou, F. A Multi-Level Feature Distribution Learning Method for Automatic Modulation Open-Set Recognition. *IEEE Trans. Cogn. Commun. Netw.* **2026**, *12*, 1508–1524. [CrossRef]
59. Li, P.; Zhang, Z.; Zhang, L.; Zhou, H.; Shi, X.; Zhou, F. S3C: Scalable Subspace with Similarity Completion for Radar Jamming Class Incremental Recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, early access. [CrossRef]
60. Zhao, F.; Xu, D.; Ren, Z.; Shao, X.; Wu, Q.; Liu, Y.; Wang, J.; Song, J.; Chen, Y.; Zhang, G.; et al. Mamba-Based Super-Resolution and Semi-Supervised YOLOv10 for Freshwater Mussel Detection Using Acoustic Video Camera: A Case Study at Lake Izunuma, Japan. *Ecol. Inform.* **2025**, *90*, 103324. [CrossRef]
61. Zhang, Z.; Zhou, H.; Shi, X.; Ran, R.; Tian, C.; Zhou, F. An Evidential-Enhanced Tri-Branch Consistency Learning Method for Semi-Supervised Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2535713. [CrossRef]
62. Yang, L.; Qi, L.; Feng, L.; Zhang, W.; Shi, Y. Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.
63. Ma, Z.; Yang, J. S1-IW-2023 Dataset (Ocean Internal Wave Dataset); Zenodo. 2024. Available online: https://zenodo.org/records/11090328 (accessed on 12 August 2026). [CrossRef]
64. Tao, M. *An Internal Wave Dataset from Sentinel-1 SAR Imagery (IWs Dataset v1.0)*; Figshare: London, UK, 2022. Available online: https://figshare.com/articles/dataset/IWs_Dataset_v1_0/21365835/3 (accessed on 12 August 2026). [CrossRef].
65. Li, Y.; Wang, Y.; Leng, T.; Wen, Z. Wavelet U-Net for Medical Image Segmentation. In *Proceedings of the 29th International Conference on Artificial Neural Networks (ICANN 2020)*; Springer: Cham, Switzerland, 2020; pp. 800–810.
66. Yang, B.; Liu, C.; Li, B.; Jiao, J.; Ye, Q. Prototype Mixture Models for Few-Shot Semantic Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV 2020), Virtual, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 763–778.
67. Ruan, J.; Li, J.; Xiang, S. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *arXiv* **2024**, arXiv:2402.02491.
68. Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; Cai, J. Exploring Smoothness and Class-Separation for Semi-Supervised Medical Image Segmentation. In *Proceedings of the 25th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2022), Singapore, 18–22 September 2022*; Springer: Cham, Switzerland, 2022; pp. 25–35.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.