scientific reports

Check for updates

OPEN

Integrating simplified Swin-T with modified EFS-Net for attention-guided underwater pipelines segmentation in complex underwater environments

Niloufar Hosseini, Farahnaz Mohanna✉ & Mohammad Kazem Moghimi

Accurate segmentation of underwater pipelines is essential for marine infrastructure inspection. However, deep learning models often struggle with extreme underwater conditions such as low light, sea snow, and sea fog, leading to poor generalization on unseen data. Existing approaches typically focus on either accuracy or computational efficiency, leaving the challenge of achieving an optimal balance between the two unresolved. This paper introduces a novel hybrid architecture, the Swin Transformer-EFSNet fusion network, which delivers state-of-the-art accuracy with significantly reduced computational complexity and strong generalization capability. The model employs a dual-encoder design: a lightweight Swin Transformer branch to capture contextual relationships and a modified EFSNet branch optimized for efficient local feature extraction. Their outputs are dynamically integrated using a three-head cross-attention fusion module which prioritizes salient spatial and contextual information before decoding the final segmentation mask. We also present the HOMOMO dataset, a new benchmark containing images with challenging conditions such as low light, fog, sea snow, and complex occlusions (e.g., pipelines buried under sand or covered by vegetation). Extensive experiments on HOMOMO and two public datasets demonstrate that our method outperforms strong baselines, including UNet, SwinUNet, TransUNet, Mask2Former, YOLOv5, YOLOv11, and YOLOv12. On HOMOMO, our model achieves a *mIoU* of 98.44% and an F-boundary of 82.01%, surpassing the best-performing method by 8.43% and 5.34%, respectively. Crucially, the proposed model exhibits outstanding generalization to unseen data, demonstrating robustness against domain shifts. By effectively balancing global and local processing, our hybrid design achieves high accuracy without imposing heavy computational costs. These results establish a new paradigm for efficient and reliable visual perception in underwater environments, paving the way for practical autonomous inspection systems.

Keywords Underwater pipelines segmentation, Swin-Transformer, EFS-Net, Multi-head cross attention, Feature fusion

Underwater pipelines are prone to corrosion and deformation, which, if undetected, can cause serious economic and environmental damage. Traditionally, inspection relies on remotely operated vehicles (ROVs), with data manually analyzed by skilled operators[1]. However, this process is costly, time-consuming, and prone to human error[2], motivating the need for automated solutions. Despite advances in artificial intelligence (AI), the noisy and low-quality data collected in underwater environments limit the effectiveness of existing approaches[1]. ROVs typically provide operational, ultrasonic, and vision-based data.

In[3], pipelines were identified by adjusting the ROV's thrusters: the vertical thruster maintained depth, equal forces on the side thrusters enabled forward motion, and differential forces induced angular deviations. Experiments showed that the ROV followed the pipeline path with errors of 0.072 m and 0.037 m along the x- and y-axes, respectively. A different method for identifying and positioning buried pipelines was presented in[4], where a multi-sensor surveying system acquired acoustic profile images and both over- and under-water

Department of Communications Engineering, University of Sistan and Baluchestan, Zahedan, Iran. ✉email: f_mohanna@ece.usb.ac.ir

topography. A position deviation correction method improved identification accuracy, reducing the average error to 0.06 m and 0.07 m along the x- and y-axes. Deep learning techniques have also been explored. In[5], GoogleNet was applied to side-scan sonar images, achieving a 90% identification accuracy. The study further highlighted the importance of dataset selection in transfer learning, reporting that pre-training with ImageNet improved accuracy by about 10%. Similarly, field experiments in[6] used multi-beam and forward-looking sonar to construct a dataset capturing various structural damages. A segmentation network combining channel and spatial attention mechanisms demonstrated strong segmentation performance while maintaining relatively fast inference speed. More recently, a method was presented in[7] for identifying and positioning exposed underwater pipelines in 3D sonar images using an enhanced YOLOv5 model. The region-growing algorithm, initialized from YOLOv5 detected bounding box centers, refined the pipeline localization. The final positioning was determined using spatial relationships among the pipelines, ROV, and tracking ship. This method achieved a reported performance of 77%. While methods using magnetic[8], sonar[9], or ultrasonic data have been explored, each suffers from nonlinear distortions, low contrast, or multipath reflections. In contrast, optical sensors offer high-resolution, fast acquisition, and additional cues such as shadows, markings, and textures[10], making them well-suited for pipeline inspection. Yet optical images also face challenges such as blur, distortion, and scattering[11].

Classical methods for optical image analysis—edge detection, morphological operations, and machine learning models such as SVMs and random forests[12–14]—depend heavily on edge information and require complex preprocessing. Their accuracy degrades in harsh underwater environments[15]. Deep learning offers significant advantages by automatically extracting robust features, reducing preprocessing, and improving accuracy[16,17]. Recent studies have applied CNNs and transformers to pipeline identification, with methods ranging from YOLO-based detectors and segmentation networks such as UNet and DeepLabv3 + to transformer-enhanced models. While these approaches achieve strong results on specific datasets, performance often drops under different underwater conditions, highlighting poor generalization. For pipeline identification, a transfer learning approach[18] was applied to six neural networks: UNet, SegNet, DeepLabv3+(ResNet-18), DeepLabv3+(ResNet-34), DeepLabv3+(ResNet-50), and DeepLabv3+(ResNet-101). The best-performing network was DeepLabv3+(ResNet-101), which was initially trained on ImageNet and then fine-tuned on the PASCAL VOC2012 dataset. Using 11,463 underwater pipeline images, this network achieved a mean Intersection over Union (*mIoU*) of 99.12%. The performance of YOLOv5, YOLOv6, YOLOv7, and YOLOv8 was also evaluated for underwater pipeline identification[19]. Experiments were conducted on the same dataset of 3,021 images collected by an ROV. The results showed that YOLOv5 achieved the highest *mAP* of 97.10%, followed by YOLOv7 with 96.30%, YOLOv6 with 95.30%, and YOLOv8 with 95.10%. The effect of data augmentation and transfer learning techniques has been investigated in MobileNet, MobileNet-V2, Inception-V3, Xception, and Inception-ResNet-V2 networks for underwater cable image identification[20]. Following a comparative analysis of these models, MobileNet-V2 outperformed the others, achieving the highest accuracy of 93.50% while also requiring the lowest computational time.

Recently, transformers have been widely adopted for underwater image processing due to their ability to model long-range dependencies and capture complex features. These capabilities are particularly useful for underwater images, which often suffer from poor lighting and challenging environmental conditions such as sea fog or sand cover[21,22]. A method named TR–YOLOv5s was introduced in[23], comprising preprocessing, down-sampling, automatic identification, and localization steps. Identification was performed using a transformer module combined with the YOLOv5s model, enhanced by an attention mechanism. Experiments demonstrated that this method achieved a *mAP* of 85.6%, representing a 12.5% improvement over YOLOv5s, with a mean test time of approximately 0.068 s. An underwater object identification algorithm based on Faster R-CNN was presented in[24] to address challenges such as color offset, low contrast, and object blur. In this approach, the Swin-Transformer (Swin-T) served as the backbone, and deep and shallow feature maps were fused using a path aggregation network. Online hard example mining improved training efficiency, and replacing ROI pooling with ROI align enhanced identification accuracy, achieving a *mAP* of 80.54% on the URPC2018 dataset containing 5,543 images. The Swin-T[25] generates a hierarchical feature representation using a shifted windowing process, which limits self-attention computation to non-overlapping local windows while allowing cross-window connections. This mechanism enables efficient extraction of local features while preserving global information. Pre-trained weights on ImageNet further improve the robustness of feature representations. Swin-T has been applied at various scales and demonstrates linear computational complexity with respect to image size. It has achieved state-of-the-art performance on COCO object detection and ADE20K semantic segmentation, significantly surpassing previous methods.

For underwater object segmentation, a method was proposed in[26], using the Efficient Fish Segmentation Network (EFS-Net) and Multi-level Feature Accumulation-based Segmentation Network (MFAS-Net). EFS-Net employed convolutional layers in the early stages for optimal feature extraction, while MFAS-Net used feature refinement and transfer blocks to enhance low-level information and propagate it to deeper stages. Additionally, MFAS-Net applied multi-level feature accumulation to improve pixel-wise predictions for indistinct objects. The networks were evaluated on the DeepFish and SUIM datasets, achieving *mIoUs* of 76.42% and 92.0%, respectively. A simple scaling method was introduced in[27] to allow users to scale baseline models to target resource constraints while maintaining efficiency. This method was applied to MobileNets and ResNet, and neural architecture search was used to design a baseline network that could be scaled into a family of models, called EfficientNets. EfficientNets achieved accuracies of 91.70% on CIFAR-100 and 98.80% on Flowers, demonstrating that mobile-sized models can be effectively scaled while surpassing state-of-the-art performance with significantly fewer parameters.

To address these limitations, we propose a hybrid segmentation model combining the Simplified Swin-Transformer (Swin-T) and a modified Efficient Fish Segmentation Network (EFS-Net). Swin-T leverages

hierarchical self-attention with shifted windows, enabling simultaneous local feature extraction and global context preservation. The modified EFS-Net incorporates trainable layers (Strided-Conv, Tra-Conv) and EfficientNetB0[27] as the initial feature extractor, providing stable, multiscale representation even from imperfect data. This hybrid design enhances accuracy, robustness, and efficiency for underwater pipeline segmentation. An additional contribution of this work is the introduction of a large-scale and challenging underwater pipeline dataset, HOMOMO, consisting of 123,876 RGB images captured across 1.2 *km* of seabed pipelines under diverse conditions including sea fog, sea snow, low light, and complex occlusions. The main contributions of this paper are summarized as follows:

1. A novel hybrid segmentation framework that combines a Simplified Swin-Transformer and a Modified EFS-Net via a three-head cross-attention fusion module. This design leverages global contextual and local spatial features simultaneously while maintaining lightweight computational cost.
2. Introduction of the HOMOMO dataset, a large-scale and challenging underwater pipeline segmentation benchmark comprising over 120,000 expert-annotated images captured under diverse real-world conditions.
3. Extensive experimental validation across three datasets, demonstrating superior accuracy, robustness, and generalization compared to state-of-the-art CNN and Transformer-based segmentation models.
4. A practical contribution toward enabling efficient and reliable autonomous inspection of subsea pipelines, with potential applications in offshore oil and gas, renewable energy, and critical infrastructure monitoring.

The remainder of this paper is organized as follows: Some related works for object recognition based on deep learning are investigated in Sect. 2. Section 3 describes the proposed method in detail. Section 4 presents experimental results, dataset descriptions, and comparisons with state-of-the-art approaches. Section 5 concludes the paper.

Related work

In this section, some related works for object recognition based on deep learning are investigated, and the reasons for not using these architectures in the proposed model are explained. This helps to understand the proposed method more clearly.

A soft-assignment color histogram strategy was introduced in[28] to develop a differentiable underwater color disparity for underwater images. Also, an underwater image enhancement framework was developed based on visual–textual fusion. In addition, the discrete wavelet transform was employed to decompose images into low and high-frequency components. Low-frequency color restoration was guided by differentiable underwater color disparity, and high-frequency detail refinement was guided by detail-intensity regularization. This guided fusion ensured that enhanced images exhibited both natural color appearance and finely reconstructed textures. However, a non-deep learning-based histogram-based color compensation method was also introduced in[29], which applied multiple attribute adjustment techniques, including max-min intensity stretching, luminance map-guided weighting, and high-frequency edge mask fusion. In addition, a multilayer information fusion and self-organized stitching method was introduced in[30] for improving the clarity of underwater scene. However, the approach's limitation was that it struggled in turbid waters with severe blue-green attenuation. An end-to-end architecture[31] based on AquaSketch-enhanced cross-scale information fusion was presented to address the underestimation of basic sketch features caused by underwater image distortion. The architecture used a top-down dual-branch pyramid for cross-scale information fusion to overcome the insufficient integration of multiscale representations of underwater objects. An adaptive attenuated channel compensation approach was developed in[32] based on optimal channel pre-correction and a salient absorption map-guided fusion method to eliminate color deviations in the RGB color space. Then, it used an algorithm to enhance the contrast of channel L and an adaptive color distribution specification method to improve contrast and match the color distribution in the Lab color space. Finally, an edge-enhanced mask fusion technique was applied for correcting blurry details. This non-deep learning approach improved underwater images to be as colorful as natural images. A plot classification network, named S2G-GCN, was presented in[33], integrating spectrum-to-graph modeling and graph convolutional network (GCN). First, the constant false alarm rate detection algorithm was applied to R-D spectra to capture potential target plots. For each plot, an echo energy diffusion region was built to include several resolution cells around its spectral peak. Then, these cells were modeled as a graph, where each node corresponded to a cell and edges were defined by spatial proximity and energy similarity between neighboring nodes. Finally, a (GCN)-based classifier was employed to learn discriminative features from the constructed graph and classified each detected plot into one of the true target, sea clutter, ground clutter, or noise classes. In the proposed method, no initial processing is performed on the input image; instead, the architecture itself identifies pipelines in the image, even if it is completely noisy and of low quality.

An underwater salient instance segmentation architecture was introduced in[34] based on the Segment Anything Model (SAM) for the underwater domain. The architecture used an underwater-adaptive ViT encoder to incorporate underwater-domain visual prompts into the segmentation network. An out-of-the-box underwater salient feature prompter generator (SFPG) was also designed to generate salient prompters instead of explicitly providing foreground points or boxes as prompts in SAM. A WaterMask was designed in[35] for underwater image instance segmentation. First, the differences-similarity graph attention module was devised to recover lost detailed information due to image quality degradation and down-sampling. Then, the multi-level feature refinement module was presented to predict foreground and boundary masks separately using features at different scales, and to guide the network via a boundary-mask strategy with a boundary-learning loss to yield finer predictions. Our model's segmentation performance is more efficient than that in[34 and 35].

A geometric mapping framework was presented in[36] to address the multiple matches in cross-modal retrieval. The rectangular matching of P2RM and R2RM were developed. The P2RM treated all retrieved candidates as

rectangles with zero volume and the query as a box. While the R2RM encoded all heterogeneous data into rectangles. Both strategies can be employed to improve the retrieval performance of baselines using off-the-self approaches. An evidence-based multi-feature fusion model was introduced in[37] to prevent DNNs from being deceived by the contaminated features only from a single block view. First, the model introduced an evidential deep learning approach to produce a reasonable uncertainty estimate for features from different blocks within an architecture. Then, it integrated multi-block features at the evidence level using Demster-Shapfer's theory for trusted prediction. A geometric representation was presented in[38] to estimate the semantics of heterogeneous data via sector embedding. The input data (image/text) was projected onto a sector, with its symmetric axis representing mean semantics and the aperture estimating uncertainty. A sector matching loss was also introduced to encourage candidates to be contained within the apertures of a query sector. An approach, named ACMR, was presented in[39] to learn both discriminative and modality-invariant representations for cross-modal retrieval. The ACMR employed two processes: a feature projector that generated modality-invariant and discriminative representations, and a modality classifier that detected the modality of an item given an unknown feature representation. A triplet constraints were also introduced to ensure that the cross-modal semantic data structure was well preserved when projected into a common subspace. Our model does not apply any retrieval techniques for underwater pipeline recognition.

A model was presented in[40] for collision-free intelligent vehicle navigation to avoid obstacles using deep reinforcement learning. The model utilized multimodal perception to achieve reliable online interaction with the surroundings. It used transfer learning to implement the virtual learning policies in the real-world environments. The model integrated camera, Lidar, and inertial measurement unit data, so, a series of cross-domain self-attention layers were applied. The ASHT-KD teacher-student approach was presented in[41] for all-day mobile visual place recognition. The framework learned the all-day place recognizer through knowledge transfer from several teachers to a limited number of students, depending on the environment's complexity. The teacher network was a two-level sampling ViT pipeline, while the Siamese student network was a lightweight pipeline consisting only of one-level down-sampling ViT. The model has drawbacks, including high computational complexity, strong dependence on the quality of the teacher model, complex hyper-parameter tuning, and the risk of information loss. An end-to-end trainable dark-enhanced Net was presented in[42] to alleviate the impact of poor illumination and environmental noise for mobile place recognition. First, a lightweight dark enhancement module was trained to improve image illumination quality. Then, a dual-level sampling pyramid transformer was constructed to extract discriminative features through aggregating descriptors. Moreover, a re-ranking method based on the cross-entropy loss was used for final place matching. The objectives and applications of methods in[40–42] and our method are distinct. Applying transfer learning through a teacher-student approach is our future research for underwater pipelines recognition.

A network was introduced in[43] for landslide extraction that leveraged the characteristics of context association. A two-branch multiscale context feature extraction module that captured the contextual relationships across different scales through an attention mechanism while concurrently extracted context information within the same scale. A supervised classifier was also presented to refine the prediction accuracy. The model's limitations are its accuracy in landslide delineation and its adaptability across diverse scenarios, both of which require further improvement. A multifaceted collaborative salient object detector was presented in[44] based on Transformer architecture for optical remote sensing images, incorporating aspects of localization, balance, and affinity. The network focused on locating targets, balancing detailed features, and modeling global image-level context. The global distribution affinity learning module concentrated on leveraging deep features to construct image-level global context association graphs through explicit affinity learning to recognize global patterns within images. This module also fostered a more cohesive expression of multilayer features by applying deep supervision within the decoding layer. A cross-view intelligent person search method was presented in[45] based on multi-feature constraints. First, the global-local context-aware module was stablished to extract differential personnel features. Second, the semantic complementarity and feature aggregation module was constructed to address personnel-scale feature constraints across different contexts. Third, the method was constrained to use only person spatial, person identity, and detection confidence features to improve person search accuracy. Our EFSNET-Swin-T segmentation model achieves high accuracy on underwater pipelines images. The objectives and applications of methods in[43–45] and our method are distinct.

The Mamba architecture was explored[46] for remote sensing change detection tasks. Three architectures, MambaBCD, MambaSCD, and MambaBDA were developed for binary change detection, semantic change detection, and building damage assessment tasks, respectively. The encoders across all three architectures used the cutting-edge Visual Mamba architecture, enabling full learning of global spatial contextual information from input images. For the change decoder, three spatial-temporal relationship modeling mechanisms were introduced that leveraged their attributes to capture the spatial-temporal interactions among multi-temporal features, thereby obtaining accurate change information. A Mamba-Convolution network was presented in[47] for underwater image enhancement, which inherited the global dependencies modeling of Mamba architecture and the local dependencies modeling of convolution architecture to improve enhancement performance. To capture global and local dependencies in image features, a Mamba-convolution hybrid block was introduced, integrating the global modeling capability of mamba blocks with the local modeling capability of the CNN-based feature attention module. Moreover, a cross fusion Mamba block was presented to fuse the image feature maps of encoder-decoder layers at different levels. A multi-label method was presented in[48] that used an ensemble learning approach to detect coral reef conditions and extract ecological information. The method's architecture combined Swin-Transformer-Small, Swin-Transformer-Base and EfficientNet-B7. The method classified the coral reef conditions as healthy, compromised, dead and rubble. It also identified corresponding stressors, including competition, disease, predation and physical issues. Our method modifies EfficientNet-B0 and combines it with the simplified Swin-T for underwater pipelines recognition, achieving a mAP of 98.98%.

Proposed method

We propose a hybrid framework for semantic segmentation of underwater pipelines that integrates two complementary architectures: the Simplified Swin-Transformer (Swin-T) and a Modified Efficient Fish Segmentation Network (EFS-Net). The overall flow of the method is illustrated in Fig. 1, and its major components are described below.

Input image

Each input is an underwater RGB image with dimensions of 1080×1920. To reduce the computational cost, these images are resized to 256×256 before processing. The 256×256 size was selected after conducting numerous experiments on underwater pipeline datasets to achieve the highest recognition accuracy. However, reducing the 256×256 size to lower sizes removed details of edges and texture in underwater pipelines images, especially those covered with sand and underwater plants, and reduced the performance of the proposed method. Therefore, a size of 256×256 is the best for our experiments.

Modified EFS-Net

The architecture of the Modified EFS-Net is shown in Fig. 2. Unlike the original EFS-Net, which uses five convolutional layers for feature extraction, we replace them with a simplified EfficientNetB0 to capture multiscale features while preserving spatial details. The original EfficientNetB0 consists of seven MBConv (Mobile Inverted Bottleneck Convolution) blocks, but only the first four are retained to emphasize low-level features such as edges and textures, which are crucial for pipeline segmentation. Extracted features by the Modified EFS-Net are then processed through Strided-Conv layers, and subsequently passed through down-sampling and up-sampling modules.

MBConv

Each MBConv block consists of depthwise separable convolutions, linear bottlenecks, and inverted residuals[49]. This design significantly reduces parameters and computational cost while maintaining feature extraction accuracy.

Down-sampling

The Down-Sampling block comprises Strided-Conv, ReLU-BN, and Conv layers across three stages. Unlike pooling, Strided-Conv preserves spatial information, enabling more accurate segmentation. At the final stage ("max-depth"), deeper semantic features are extracted, producing output feature maps of size $256 \times 4 \times 4$.

Up-sampling

The Up-Sampling block mirrors the down-sampling process, employing Transposed Convolutions (Tra-Conv), ReLU-BN, and Conv layers over three stages. This reconstruction restores spatial resolution while avoiding the spatial loss typically introduced by un-pooling.

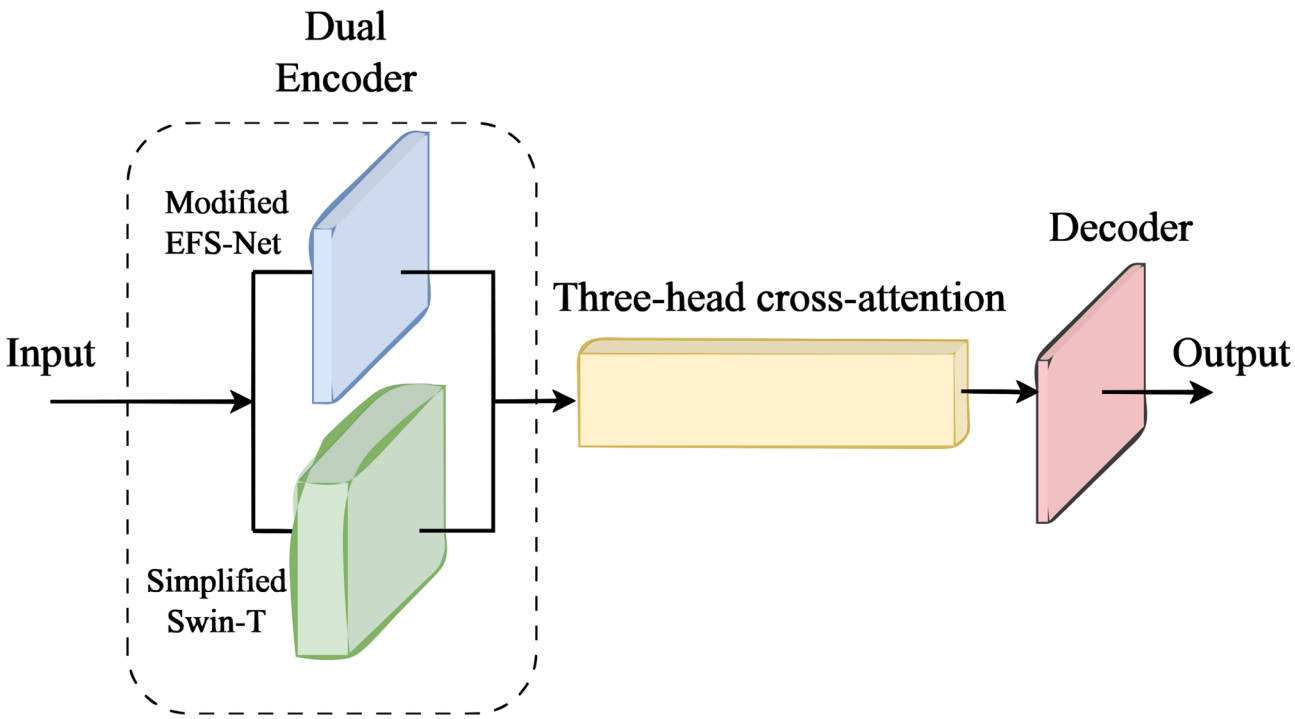

Fig. 1. The flowchart of the proposed method.

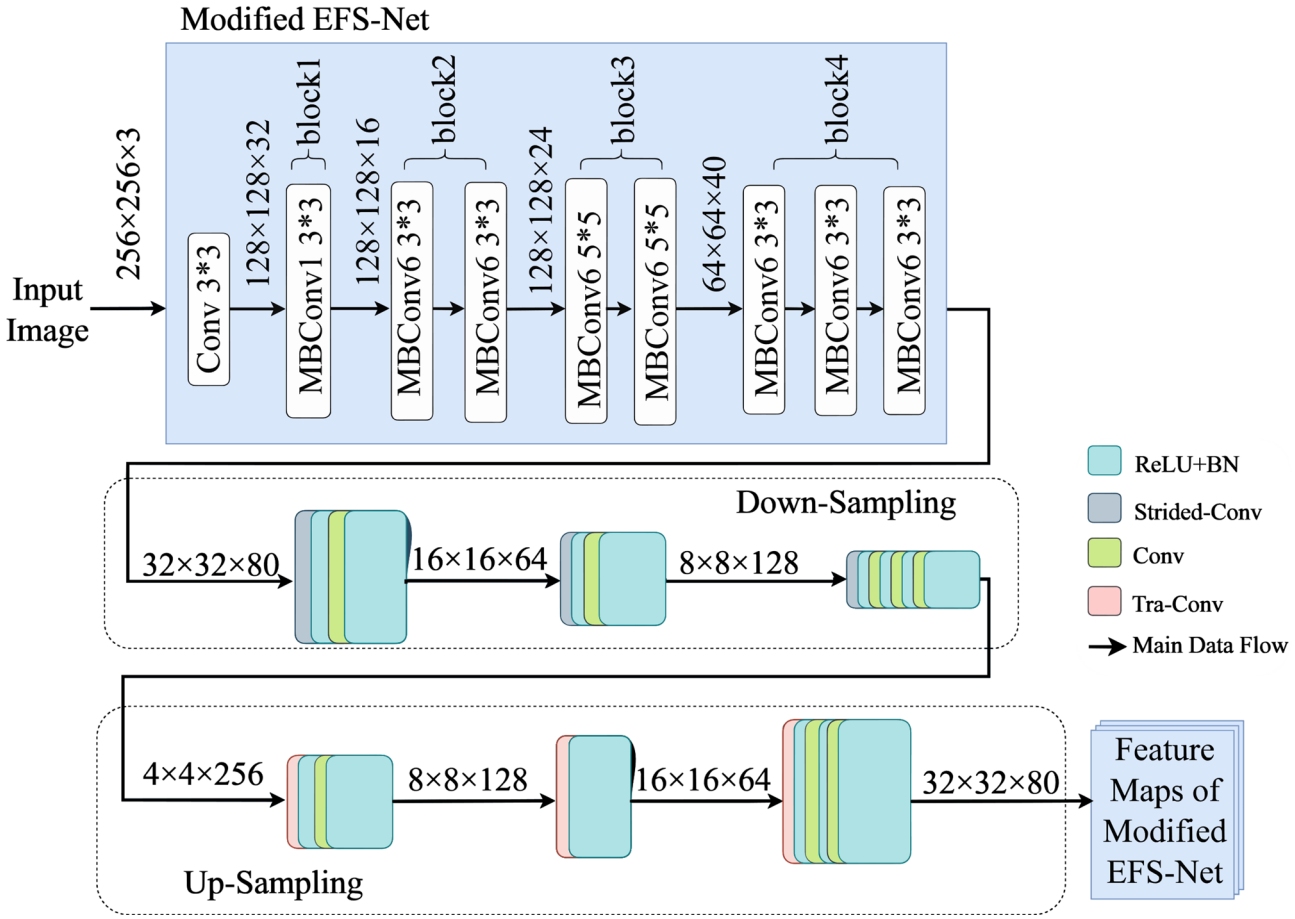

Fig. 2. The structure of modified EFS-Net.

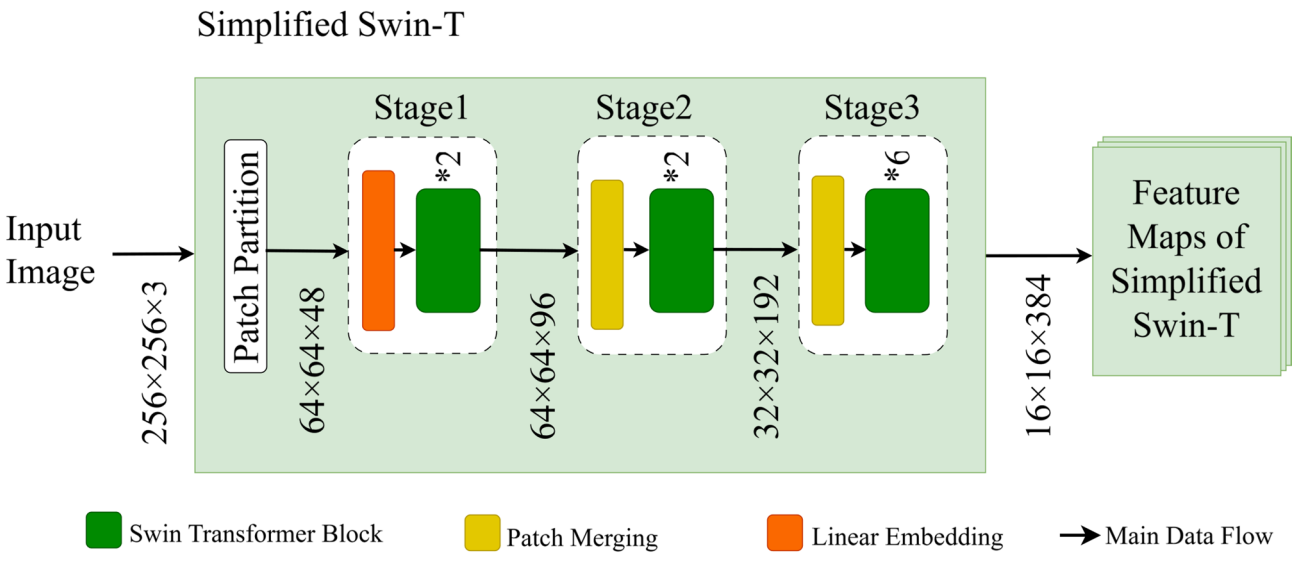

Fig. 3. The structure of simplified Swin-T.

Simplified Swin-T

The Simplified Swin-Transformer (Fig. 3) is derived from the original Swin-T (Swin-UperNet), but only the first three stages are retained for efficiency in underwater pipeline identification.

The components of Simplified Swin-T are:

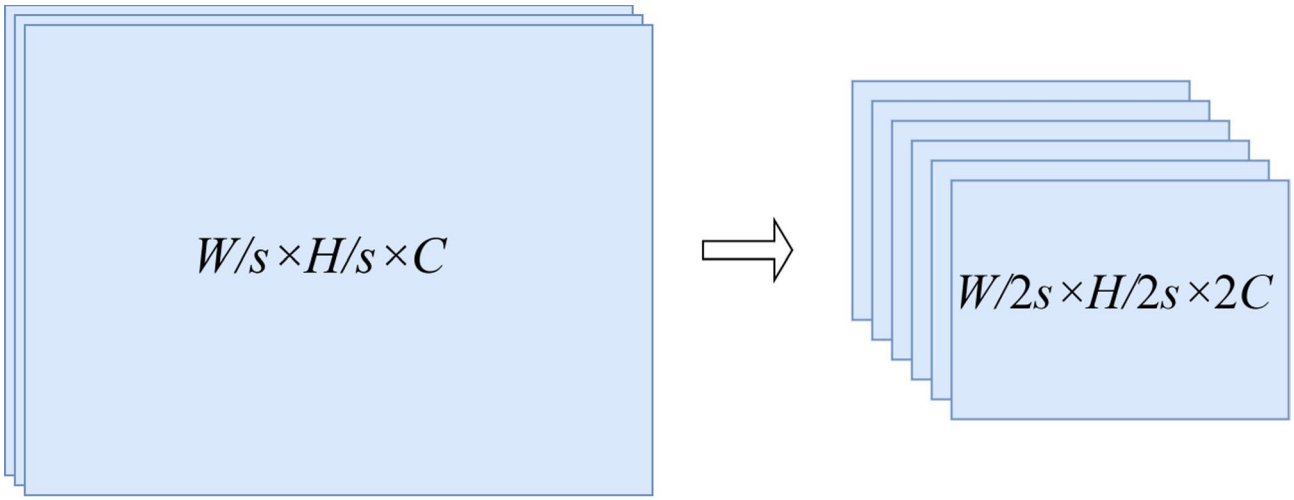

Fig. 4. The schematic of patch merging.

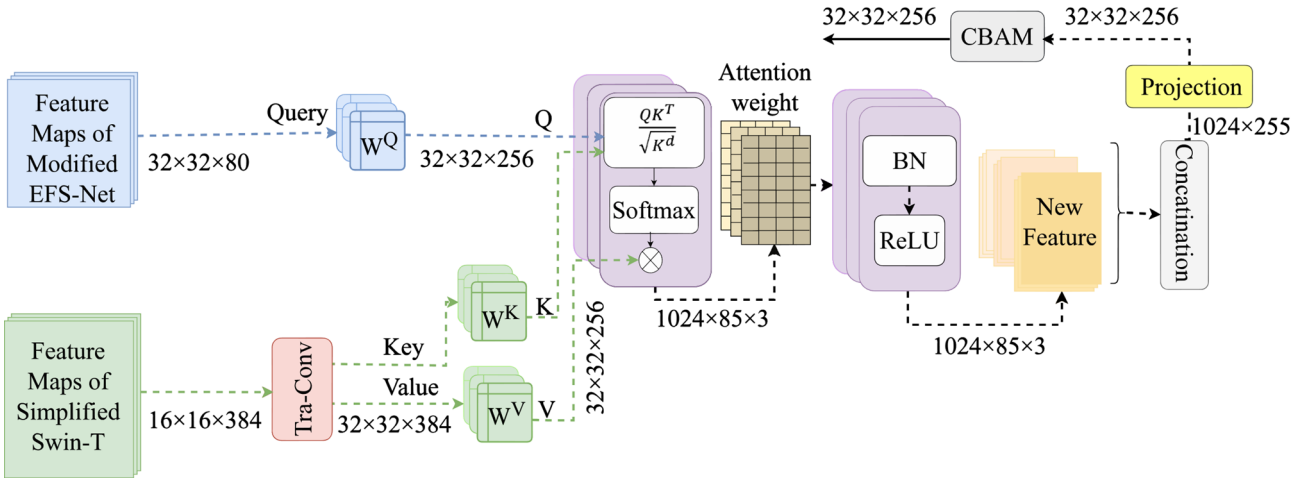

Fig. 5. The structure of three-head cross attention[50], and CBAM.

- *Patch Partitioning*: The input image $x \in R^{3*W*H}$ is divided into 4 × 4 non-overlapping patches.
- *Linear Embedding*: Each patch is embedded into a 96-channel feature representation.
- *Swin-Transformer Block*: Using shifted window multi-head self-attention (SW-MSA)[25], local windows (8 × 8) are processed efficiently, while window shifting allows inter-region information exchange.
- *Patch Merging*: Features are hierarchically merged across scales, reducing spatial size while increasing channel depth for multiscale representation.

The schematic of patch merging is shown in Fig. 4, where W is the input width, H is the input length, C is the number of channels, and s is the stage number.

Feature fusion

Features from the two encoders are fused using a three-head cross-attention mechanism (Fig. 5). Cross-attention adaptively weights feature maps from each encoder using the trainable weight matrices of W^Q, W^K, and W^V, thereby emphasizing salient spatial and contextual cues. Compared to concatenation, this approach reduces computational complexity and avoids redundant information.

The three-head cross-attention module takes two feature maps: one from the Modified EFS-Net (80 channels, 32 × 32) and the other from the Simplified Swin-T (384 channels, 16 × 16). After spatial alignment via Tra-Conv, both features are projected to 32 × 32 × 256. The three-head attention computes the attention weights by (1).

$$Attention = Softmax(\frac{QK^T}{\sqrt{K^d}}) \otimes V \quad (1)$$

Where Q, K, and V are linear projections of the input features, and symbol of $\otimes$denotes element-wise multiplication with broadcasting. Each head specializes in different aspects of pipeline recognition, including local edges, segment relationships, and global structure. The output of attention module ($1024 \times 85 \times 3$) after applying BN, and ReLU is concatenated with a feature of size (1024×255). Then it is projected to $32 \times 32 \times 256$. This output is refined by CBAM block. CBAM stands for Convolutional Block Attention Module, which consists of channel attention and spatial attention modules. The CBAM's channel attention module generates Channel-Wise Weights (CWW) according to (2), to emphasize pipeline's relevant features.

$$CWW = Sigmoid\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \quad (2)$$

Where F is the final output of three-head cross attention after projection. Subsequently, CBAM's spatial attention module generates Spatial-Wise Weights (SWW) according to (3) to produce a spatial mask that focuses computation on regions likely to contain pipelines.

$$\begin{aligned} F' &= CWW \otimes F \\ F'\_Concatinate &\left(AvgPool\left(F'\right); MaxPool\left(F'\right)\right) \\ SWW &= Sigmoid\left(Conv_{7\times7}\left(F'\_Concatinate\right)\right) \end{aligned} \quad (3)$$

The final output of CBAM is as $SWW \otimes CWW \otimes F$.

The inclusion of CBAM further enhances spatial and channel attention. It suppresses noise and stabilizes training. We observed experimentally that using three attention heads yielded the best trade-off between accuracy and complexity for underwater pipeline segmentation.

When comparing the cross-attention feature fusion method with other commonly used feature fusion methods, several advantages become apparent. First, cross-attention is performed adaptively and dynamically, based on the importance of the features extracted by each network. Consequently, more attention is given to the most important features and, therefore, to the most relevant parts of the image. Second, cross-attention can better model the relationships between features, allowing for more accurate identification of objects in complex images with multiple features and intricate relationships. Third, cross-attention has lower computational complexity than the concatenation method due to its use of the attention mechanism and dynamic weighting for feature fusion. However, calculating attention scores requires additional time during training, which can be problematic, especially for large datasets and complex networks. In contrast, the concatenation fusion method combines features without considering their relative importance, increasing the dimensionality of the feature vector, which can lead to higher computational complexity and the inclusion of redundant information. In transformer-based fusion methods[51], both cross-attention and self-attention are employed to fuse features of different types, such as images and text from different inputs. Since the proposed method uses only image inputs, there is no need for the additional complexity of self-attention alongside cross-attention.

Decoder

The decoder (Fig. 6) reconstructs the segmentation mask through four steps:

1. A 3×3 Conv layer (128 filters) with ReLU and BN processes fused features.
2. Up-sampling quadrupled the feature dimensions; a 3×3 Conv (64 filters) reduces channel count.
3. Another up-sampling restores input resolution with a 3×3 Conv (32 filters).
4. Finally, a 1×1 Conv with sigmoid activation generates the binary segmentation mask.

The decoder architecture employs a designed progressive up-sampling strategy that optimally balances computational efficiency with reconstruction accuracy. The gradual channel reduction (256→128→64→32) following each up-sampling operation ensures efficient memory usage while preserving essential features for pipeline segmentation. Sequential 3×3 convolutions with BN provide sufficient receptive field and training stability, while the final 1×1 convolution with sigmoid activation directly produces the binary segmentation mask. This design is particularly advantageous for detecting thin and partially occluded pipelines, as the staged

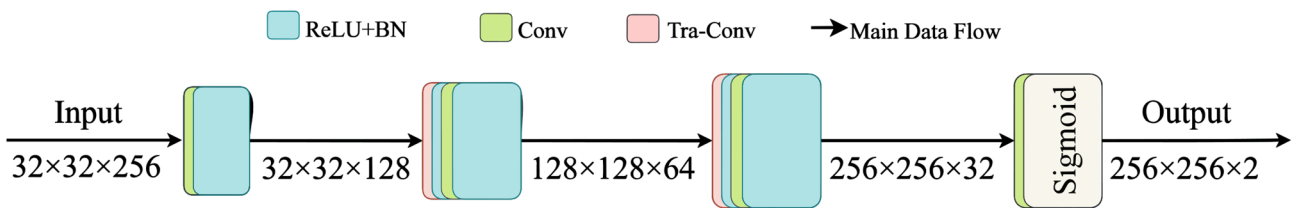

Fig. 6. The structure of decoder[26].

reconstruction process minimizes information loss and maintains boundary precision throughout the decoding stages.

Table 1 compares the main characteristics of the Simplified Swin-T and Modified EFS-Net feature extractors. By combining global contextual reasoning (Swin-T) with efficient local feature extraction (EFS-Net), the proposed hybrid design achieves accurate, robust, and computationally efficient segmentation of underwater pipelines.

Totally, our contribution is not merely an improved Swin-T, but rather a novel fusion paradigm for underwater pipeline recognition that consists of, three-head cross-attention fusion, and sequential feature refinement pipeline. In the three-head cross-attention fusion:

- Query is from Modified EFS-Net which leverages local spatial features for attention guidance.
- Key/Value from Simplified Swin-T which Provides global contextual information.
- Three-head cross attention which enables multi-scale feature interaction across different representation. Sub-spaces.

In addition, our sequential feature refinement pipeline represents a deliberate design choice where:

- Cross-attention enables dynamic feature selection.
- Concatenation preserves maximum information from both branches.
- CBAM provides adaptive spatial-channel refinement of combined features, which is particularly effective for challenging underwater conditions.
- Decoder reconstructs precise segmentation masks.

Therefore, our key conceptual contribution lies in the purpose-built design for underwater pipeline recognition, which presents unique challenges that generic segmentation models fail to address effectively. We intentionally leverage the complementary strengths of each component in the proposed.

- Modified EFS-Net's strength: Preserves spatial details and edge information critical for pipeline boundary detection.
- Simplified Swin-T's strength: Provides global contextual understanding for pipeline trajectory prediction.
- Our fusion innovation hybrid architecture: The cross-attention mechanism intelligently balances these aspects specifically for linear structure detection.

The Simplified Swin-T configuration is not mere parameter reduction. Our analysis revealed that for linear structure detection:

- Deep abstraction in Stage 4 of Swin-T actually harms linear feature preservation.
- Early-stage Swin features (Stages 1–3) provide optimal balance of context and spatial resolution.
- This configuration represents domain-aware architectural optimization, not simplification.

While individual components exist separately, our integration methodology introduces:

Network	Layer name	Layer type	Output	Number of parameters
Modified EFS-Net	Conv	Conv3 × 3 stride 2	128 × 128 × 32	864
	block1	MBConv 1, 3 × 3	128 × 128 × 16	1728
	block2	2×(MBConv 6, 3 × 3)	128 × 128 × 24	2 × 10,296
	block3	2×(MBConv 6, 5 × 5)	64 × 64 × 40	2 × 43,360
	block4	3×(MBConv 6, 3 × 3)	32 × 32 × 80	3 × 81,920
	Down-Sampling (step1)	Strided-Conv+ Conv3 × 3	16 × 16 × 64	864 + 9216
	Down-Sampling (step2)	Strided-Conv+ Conv3 × 3	8 × 8 × 128	18,432 + 36,864
	Down-Sampling (step3)	Strided-Conv + 3×Conv3 × 3	4 × 4 × 256	73,728 +(147,456)×3
	Up-Sampling (step1)	Tra-Conv+ Conv3 × 3	8 × 8 × 128	73,728 + 36,864
	Up-Sampling (step2)	Tra-Conv	16 × 16 × 64	18,432
	Up-Sampling (step3)	Tra- Conv + 2× (Conv3 × 3)	32 × 32 × 80	864+(9,216)×2
Simplified Swin-T	Linear Embedding	Conv4 × 4, stride4	64 × 64 × 96	4,656
	2×(Swin Transformer Block)	2 × (3-head Self-Attention, MLP hidden = 384)	64 × 64 × 96	2 × 27,648 2 × 73,728
	Patch Merging	Linear projection	32 × 32 × 192	147,456
	2×(Swin Transformer Block)	2 × (3-head Self-Attention, MLP hidden = 768)	32 × 32 × 192	2 × 110,784 2 × 294,912
	Patch Merging	Linear projection	16 × 16 × 384	294,912
	6×(Swin Transformer Block)	6 × (3-head Self-Attention, MLP hidden = 1536)	16 × 16 × 384	6 × 369,536 6 × 1,179,648

Table 1. Characteristics of two feature extractors of simplified Swin-T and modified EFS-Net.

- Asymmetric feature alignment: EFS-Net features (preserving pipe edges) guide Swin-T feature selection.
- Progressive context integration: From local pipe textures to global pipeline networks.
- Robustness to underwater degradation: Specifically designed to handle turbidity and low visibility.

The proposed architecture addresses pipeline-specific challenges:

- Continuous structure maintenance across long distances.
- Occlusion resilience from marine fouling.
- Scale invariance for pipes of varying diameters.
- Orientation awareness for pipeline route analysis.

As a result, the proper design of these components in the proposed model and proper adjustment of their inputs, and outputs stablishes the proposed method with high performance for special task of underwater pipeline recognition in the comprehensive and challenging dataset of HOMOMO. The complete architecture diagram of the proposed model, including (Input → Dual Encoder → Three-head cross-attention → Decoder), is shown in Fig. 7.

Results, evaluation, and comparison

The proposed method was implemented on an NVIDIA A100 GPU (40 *GB* HBM2, 1555 *Hz* bandwidth). Experiments were conducted on three datasets.

Image datasets

HOMOMO: A custom dataset of 123,876 RGB images (1920 × 1080, resized to 256 × 256) captured from 1.2 *km* of seabed pipelines under challenging conditions (sea fog, sea snow, sand, vegetation) by professional diver to have real underwater pipelines videos. Data were split into training (60%), validation (20%), and testing (20%). Training data were augmented (rotation, scaling, color inversion, Gaussian noise), producing 187,000 images resized to 256 × 256. Images were labeled into two classes: *pipeline* and *non-pipeline*.

Roboflow: 5,980 simulated RGB images (640 × 640, resized to 256 × 256) containing synthetic underwater pipelines with varying lighting and reflection conditions. Available at: https://universe.roboflow.com/underwaterpipes/underwater_pipes_orginal_pictures.

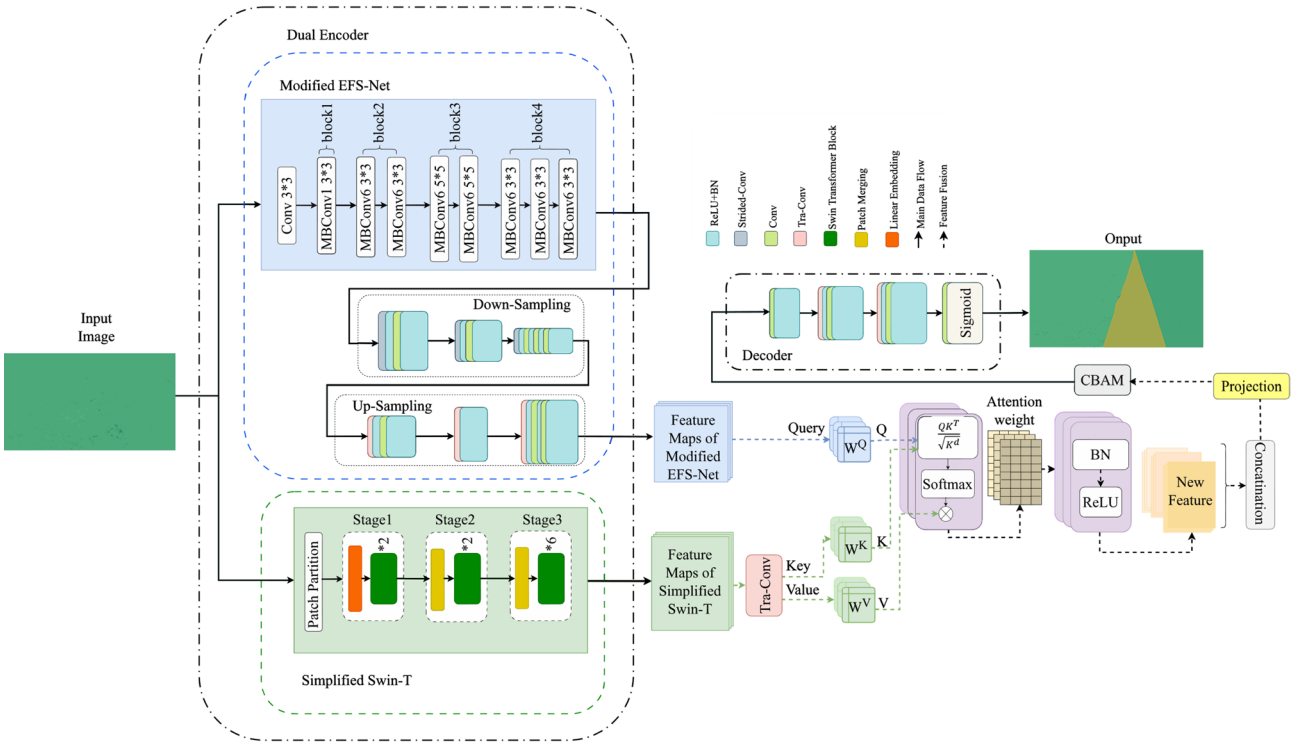

Fig. 7. A complete architecture diagram including (Input → Dual Encoder → Three-Head Cross-Attention → Decoder).

YouTube: 28,622 real RGB images (720 × 1280, resized to 256 × 256) with fog, buried pipelines, and vegetation, gathered from YouTube online sources.

Experimental setup

The proposed model was trained with a fused loss function including Cross Entropy and Dice Loss[52], optimized using Adam (l_r = 0.0001, batch = 16, 100 epochs). Five-fold cross-validation was applied. Transfer learning with ImageNet pre-trained weights was used, followed by fine-tuning on HOMOMO. To evaluate generalizability, the fine-tuned model was tested directly on Roboflow and YouTube without retraining. The Hyper-parameters of the proposed model are shown in Table 2.

Evaluation metrics

Performance was assessed using standard metrics: mean Intersection over Union (*mIoU*), Accuracy, Precision, Recall, F-score, and F-boundary, defined in (4) to (10)[3,26]. The F-boundary metric was computed with a threshold of 2 pixels, determined experimentally, which shows the maximum acceptable distance between an identified boundary pixel, and that pixel on its corresponding ground truth to be considered as a true positive.

$$IoU = \frac{T_p}{T_p + F_N + F_p} \tag{4}$$

$$Accuracy = \frac{T_p + T_N}{T_P + T_N + F_P + F_N} \tag{5}$$

$$\mathrm{Pr}\,ecision = \frac{T_p}{T_p + F_p} \tag{6}$$

$$\mathrm{Re}\,call = \frac{T_p}{T_p + F_N} \tag{7}$$

$$Specificity = \frac{T_N}{T_N + F_p} \tag{8}$$

$$F - score = \frac{2 \times \mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{9}$$

$$F - boundary = \frac{2 \times \mathrm{Pr}\,ecision_b \times \mathrm{Re}\,call_b}{\mathrm{Pr}\,ecision_b + \mathrm{Re}\,call_b} \tag{10}$$

Where, T_p is the number of correctly identified pixels for pipelines, T_N is the number of correctly identified pixels for non-pipelines, F_N is the number of pixels of pipelines that are not identified, and F_P is the number of pixels incorrectly identified as the pipelines. For computation of $\mathrm{Re}\,call_b$ and $\mathrm{Pr}\,ecision_b$, T_p is the number of correctly identified pixels for edges of pipelines, T_N is the number of correctly identified pixels for edges of non-pipelines, F_N is the number of edges pixels of pipelines that are not identified, and F_P is the number of pixels that are incorrectly identified as edges of pipelines.

Ablation, and simulation results

The ablation experiments results of different components of the proposed model are indicated in Table 3. These results clearly demonstrate the additive effect of each component of the proposed architecture in improving the performance of underwater pipeline segmentation. At the baseline level, the single models Swin-T with 1-head (87.5% mIoU) and EFS-Net with 1-head (85.2% mIoU) each demonstrated their complementary capabilities in understanding the overall context and extracting spatial details. The integration of these two architectures in

Hyper parameter	Description
Epochs no.	100
Train dataset	60% of dataset
Test dataset	20% of dataset
Validation dataset	20% of dataset
Learning algorithm	Adam
Learning rate	0.0001
Activation function	Rectified Linear Unit (RLU)
Batch normalization	16
Validation	k-fold (k = 5)
Loss function	(Cross Entropy)+(Dice Loss)

Table 2. Hyper-parameters of the proposed method.

Components	Criteria						
	Mean IoU	Mean Accuracy	Mean Precision	Mean Recall	Mean Specificity	Mean F1-score	Mean F-boundary
(Simplified Swin-T)+ (1-head Cross Attention)	87.5%	89.74%	88.1%	87.9%	89.2%	88.6%	76.8%
(Modified EFS-Net))+ (1-head Cross Attention)	85.2%	87.42%	85.8%	85.5%	86.48%	85.6%	78.5%
(Simplified Swin-T)+ (Modified EFS-Net))+ (1-head Cross Attention)	92.8%	94.83%	93.5%	93.1%	93.97%	93.3%	79.8%
(Simplified Swin-T)+ (Modified EFS-Net))+ (3-head Cross Attention)	96.3%	97.31%	96.9%	96.6%	97.27%	96.7%	80.9%
(Simplified Swin-T)+ (Modified EFS-Net))+ (1-head Cross Attention)+CBAM = proposed method	98.44%	99.5%	98.98%	98.52%	99.1%	98.74%	82.01%

Table 3. Ablation experiments results of different components of the proposed model.

Criteria	YouTube	Roboflow	HOMOMO
Mean IoU	81.32% ± 6.17	85.4% ± 4.21	98.44% ± 0.91
Mean accuracy	93.0% ± 4.21	94.6% ± 2.50	99.5% ± 0.62
Mean precision	85.73% ± 5.31	86.98% ± 3.81	98.98% ± 0.88
Mean recall	84.76% ± 5.78	86.76% ± 3.92	98.52% ± 0.78
Mean specificity	85.41% ± 3.98	87.36% ± 2.1	99.1% ± 0.69
Mean F1-score	85.24% ± 5.43	86.86% ± 3.86	98.74% ± 0.84
Mean F-boundary	70.01% ± 8.35	75.99% ± 7.81	82.01% ± 1.3

Table 4. Results of underwater pipelines recognition using the proposed method in HOMOMO, Roboflow, and YouTube including their standard deviations.

Criteria	Different components of the proposed architecture				
	Original EFS-Net	Modified EFS-Net	Original Swin-T (Swin-UperNet)	Simplified Swin-T	Proposed architecture
Mean IoU	77.12%	79.92%	86.71%	81.81%	98.44%
Mean accuracy	79.4%	83.12%	89.5%	87.02%	99.5%
Mean precision	78.83%	80.10%	88.43%	82.11%	98.98%
Mean recall	78.01%	81.01%	87.61%	82.71%	98.52%
Mean specificity	78.68%	81.98%	88.19%	83.96%	99.1%
Mean F-score	78.41%	80.05%	88.01%	82.40%	98.74%
Mean F-boundary	71.23%	79.99%	78.2%	69.15%	82.01%
Mean testing time (s)	0.010	0.009	0.021	0.0105	0.011
GFLOPs	2.2	2.9	4.5	3.1	6.2
Parameters no. (M)	3.27	4.68	31	21.5	25
FPS	100	111	48	95	91
Memory-usage (GB)	4.1	4.8	32	21.5	20

Table 5. Comparison of the proposed hybrid architecture with original EFS-Net, modified EFS-Net, original Swin-T (Swin-UperNet), and simplified Swin-T in complexity, speed, and recognition.

the hybrid model led to a significant jump to 92.8% mIoU, indicating the synergistic effect of combining these two approaches. Then, by introducing the three-head attention mechanism in the next model, the performance was improved to 96.3% mIoU, confirming the importance of simultaneous processing of different scales in pipeline detection. Finally, the addition of the CBAM module in the full model led to a final accuracy of 98.44% mIoU and 82.01% F-boundary, indicating the vital role of this module in boundary refinement. This continuous improvement trend in all metrics – precision from 85.8% to 98.98%, recall from 85.5% to 98.52%, and F1-score from 85.6% to 98.74% - clearly demonstrates the justification of the hierarchical design of the proposed architecture and the effect of each added component to it in achieving near-perfect accuracy.

The proposed method is trained and tested on HOMOMO dataset, but only tested on Roboflow and YouTube datasets. These experimental results including their standard deviations are shown in Table 4. From these results, the generalizability of the proposed method is confirmed. In fact, the results show the ability of the proposed architecture to recognize underwater pipelines on unseen data as well.

Method	Datasets			
	Criteria	HOMOMO	Roboflow	YouTube
Proposed	mIoU	**98.44%**	85.4%	81.32%
	Mean accuracy	**99.5%**	94.6%	93.0%
	Mean precision	**98.98%**	86.98%	85.73%
	Mean recall	**98.52%**	86.76%	84.76%
	Mean specificity	**99.1%**	87.36%	85.41%
	Mean F-score	**98.74%**	86.86%	85.24%
	Mean F-boundary	**82.01%**	75.99%	70.01%
	Mean test time (s)	**0.011**	0.018	0.012
DeeplabV3(ResNet101)[18]	mIoU	84.23%	68.5%	61.9%
	Mean accuracy	88.21%	87.2%	86.15%
	Mean precision	86.98%	73.32%	70.1%
	Mean recall	86.22%	72.98%	69%
	Mean specificity	86.83%	73.60%	70.08%
	Mean F-score	86.59%	73.14%	69.54%
	Mean F-boundary	71.23%	53.54%	51.02%
	Mean test time (s)	0.025	0.026	0.024
U-Net[18]	mIoU	82.10%	50.3%	48.45%
	Mean accuracy	86.90%	83.4%	82.71%
	Mean precision	86.20%	67.13%	59.98%
	Mean recall	84.99%	65.34%	58.18%
	Mean specificity	85.60%	66.43%	59.12%
	Mean F-score	85.59%	66.22%	59.06%
	Mean F-boundary	69.10%	41.93%	41.23%
	Mean test time (s)	0.043	0.039	0.040
Mask2Former[53]	mIoU	87.11%	69.14%	70.43%
	Mean accuracy	89.78%	87.18%	86.41%
	Mean precision	88.14%	77.25%	74.01%
	Mean recall	87.59%	75.12%	73.16%
	Mean specificity	88.15%	76.71%	74.31%
	Mean F-score	87.84%	76.17%	73.58%
	Mean F-boundary	70.2%	56.18%	57.31%
	Mean test time (s)	0.030	0.031	0.029
SwinUNet[54]	mIoU	90.01%	72.19%	71.31%
	Mean accuracy	91.13%	88.34%	87.12%
	Mean precision	90.39%	74.65%	73.03%
	Mean recall	90.23%	72.49%	72.12%
	Mean specificity	90.96%	73.11%	73.07%
	Mean F-score	90.31%	73.55%	72.57%
	Mean F-boundary	74.41%	58.98%	60.71%
	Mean test time (s)	0.033	0.032	0.31
TransUNet[55]	mIoU	88.78%	71.43%	72.1%
	Mean accuracy	88.99%	88.98%	89.11%
	Mean precision	88.91%	73.67%	73.26%
	Mean recall	88.86%	72.98%	73.19%
	Mean specificity	88.93%	73.14%	74.38%
	Mean F-score	88.88%	73.32%	73.22%
	Mean F-boundary	76.67%	60.34%	59.98%
	Mean test time (s)	0.036	0.0321	0.032

Continued

Method	Datasets			
	Criteria	HOMOMO	Roboflow	YouTube
YOLOv5[19]	mIoU	80.02%	61.27%	60.94%
	Mean accuracy	87.02%	80.63%	83.45%
	Mean precision	86.14%	70.34%	69.91%
	Mean recall	84.31%	69.76%	69.32%
	Mean specificity	84.90%	70.58%	70.1%
	Mean F-score	85.21%	70.04%	69.61%
	Mean F-boundary	63.12%	54.11%	58.43%
	Mean test time (s)	0.004	0.006	0.006
YOLOv11[56]	mIoU	82.96%	62.78%	63.01%
	Mean accuracy	89.62%	83.03%	85.14%
	Mean precision	89.14%	72.78%	72.44%
	Mean recall	87.21%	71.44%	70.99%
	Mean specificity	87.85%	72.23%	71.97%
	Mean F-score	88.16%	72.10%	71.70%
	Mean F-boundary	68.13%	55.91%	57.87%
	Mean test time (s)	0.0029	0.0025	0.0029
YOLOv12[56]	mIoU	81.12%	63.01%	62.11%
	Mean accuracy	89.43%	83.13%	85.41%
	Mean precision	88.14%	72.65%	73.06%
	Mean recall	79.87%	68.03%	68.31%
	Mean specificity	80.12%	69.17%	69.48%
	Mean F-score	83.80%	70.46%	70.60%
	Mean F-boundary	66.43%	55.44%	59.98%
	Mean test time (s)	0.0032	00.29	00.30

Table 6. Comparison of the proposed method with state-of-the-art methods in HOMOMO, Roboflow, and YouTube datasets.

Table 5 compares the proposed hybrid architecture with original EFS-Net, Modified EFS-Net, original Swin-T, and Simplified Swin-T in complexity, speed, and recognition to show the proposed method performance totally is higher, due to the combination of proper components; Modified EFS-Net and Simplified Swin-T.

As it is seen, the proposed architecture achieved the best performance across all metrics, e.g., mIoU improved by 11.73% over Swin-T, 21.32% over EFS-Net, 16.63% over Simplified Swin-T, and 18.52% over Modified EFS-Net. Testing time was slightly higher than original EFS-Net (+1–2 ms) but lower than original Swin-T (–10 ms) and Simplified Swin-T (–6.5 ms), represented an excellent accuracy–efficiency balance. In addition, the complexity of the proposed architecture according to the parameters number, GFLOPs, memory usage, and number of frames per second (FPS) confirmed that the proposed model provides an excellent trade-off between computational efficiency and segmentation accuracy, making the proposed method both theoretically sound and practically viable.

Generalization results

In the top rows of Table 6, the proposed method performance is reported on HOMOMO, Roboflow, and YouTube. Despite being trained only on HOMOMO, the method generalized well to unseen datasets, maintaining high segmentation accuracy under varying conditions (e.g., pipelines occluded by sand or vegetation, low light, fog).

Visual examples in Fig. 8 illustrate accurate detection even with challenges such as hidden pipelines by sand in seabed, in the presence of sea fog, sea snow, and limited light. The maximum observed error was 18.71%, calculated as the ratio of misclassified pixels to ground-truth pipeline pixels.

Comparison with state-of-the-arts

For fair benchmarking, U-Net, DeepLabV3 (ResNet101), SwinUNet, Mask2Former, YOLOv5, TransUNet, YOLOv11, and YOLOv12 were re-implemented under the same setup. Results (Table 5) show the proposed method consistently outperformed all baselines across *mIoU*, Accuracy, Precision, Recall, F-score and F-boundary. While YOLOv5 achieved lower inference time, its accuracy lagged behind; the small speed gap is negligible compared to the proposed model's substantial accuracy gains. Notably, the proposed method surpassed SwinUNet in both segmentation and boundary detection (F-boundary). Performance improvements over TransUNet highlight strong adaptability to diverse datasets. Figures 9 and 10 further illustrate performance

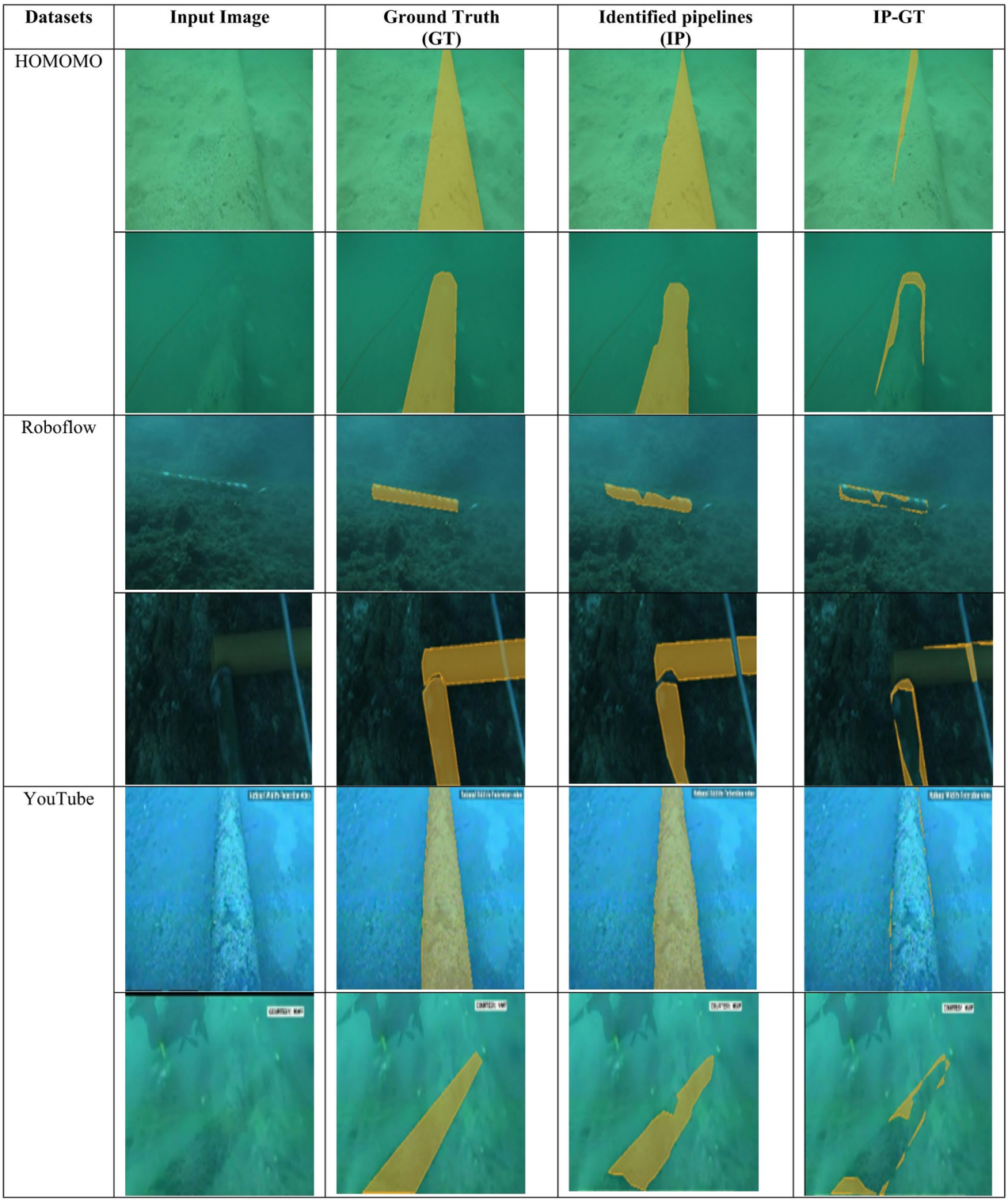

Fig. 8. Visual examples of the proposed method on HOMOMO, Roboflow, and YouTube.

gains over state-of-the-art models, confirming that the proposed hybrid approach provides both robust segmentation accuracy and generalization across unseen data.

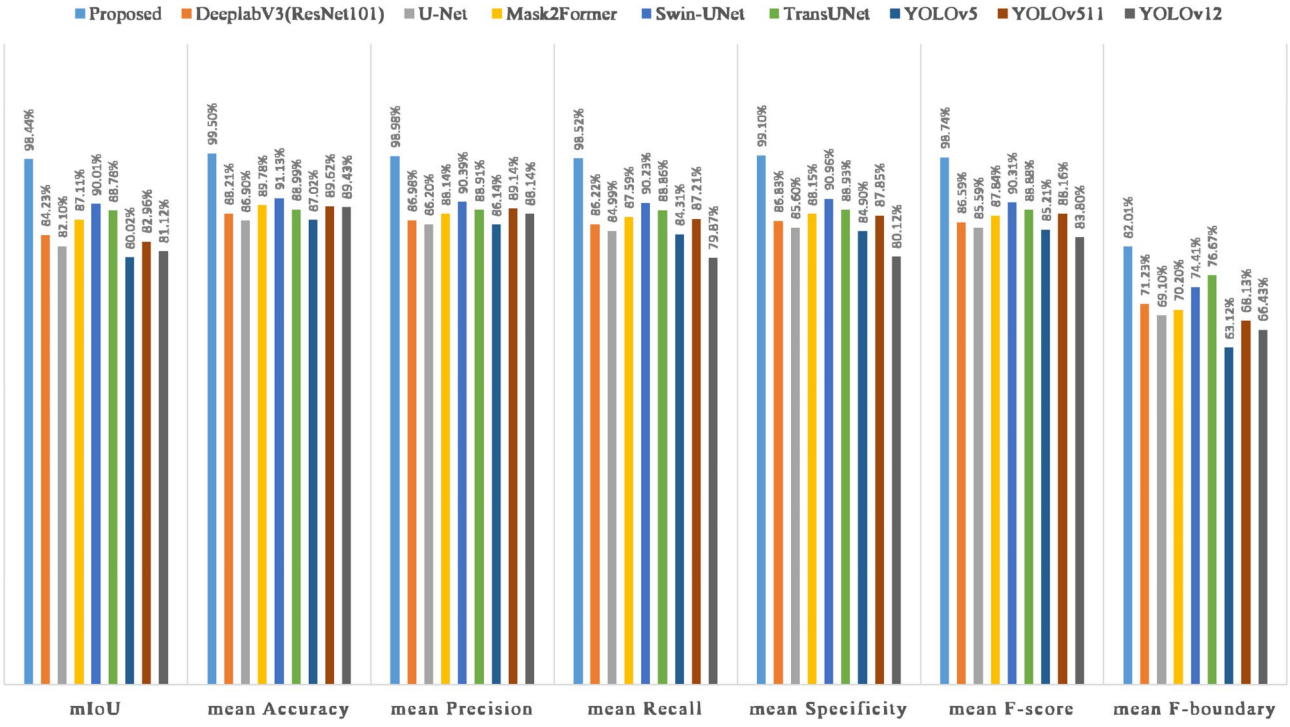

Fig. 9. Comparison of the proposed method performance over state-of-the-art methods on HOMOMO.

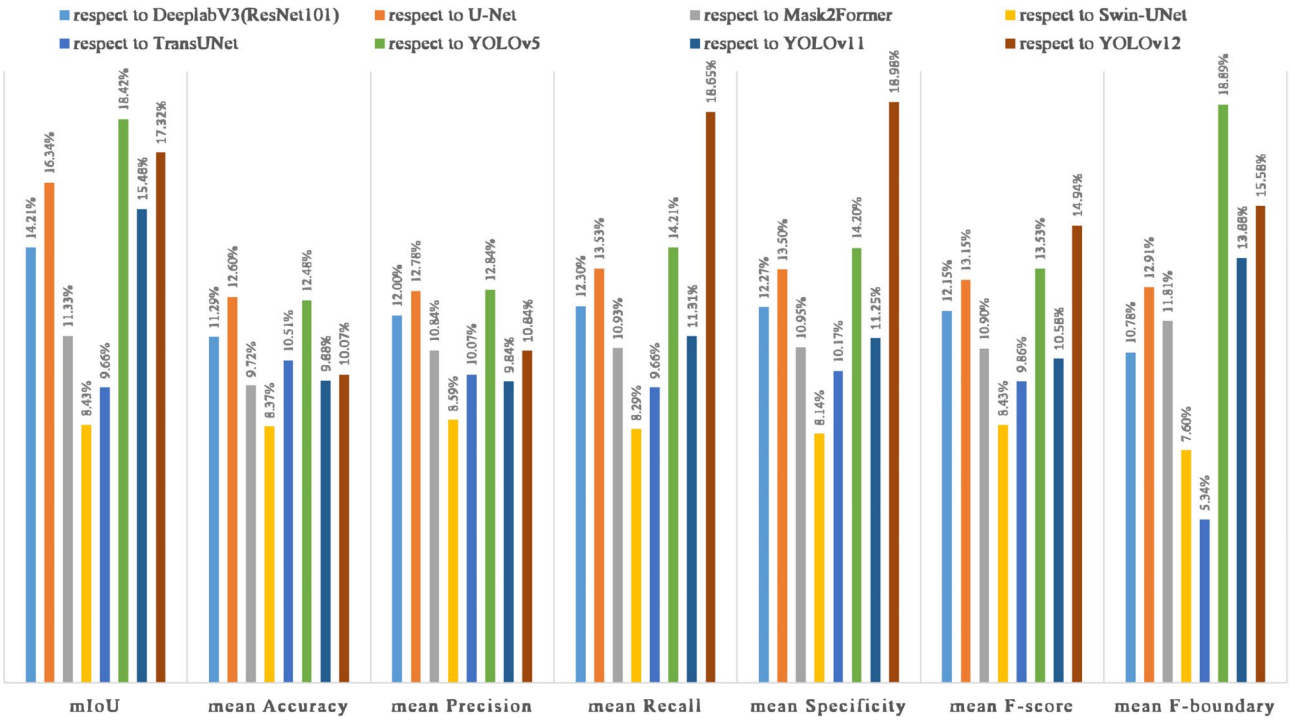

Fig. 10. Performance improvements of the proposed method respect to state-of-the-art methods on HOMOMO database.

Conclusion

Accurate identification of underwater pipelines is critical for the safety and reliability of marine infrastructure, yet existing models often fail under varying environmental conditions. This paper proposed a hybrid segmentation framework that integrates a Simplified Swin-Transformer with a Modified EFS-Net, fused through a cross-

attention module. The design leverages the Swin-Transformer's ability to capture long-range dependencies and EfficientNetB0's strength in extracting local features, while maintaining a lightweight structure for computational efficiency. Extensive experiments demonstrated the superiority of the proposed method. On the challenging HOMOMO dataset, the model achieved a *mIoU* of 98.44%, mean F-boundary of 82.01%, and consistently outperformed state-of-the-art approaches including U-Net, DeepLabV3 + ResNet101, Swin-UNet, TransUNet, Mask2Former, YOLOv5, YOLOv11, and YOLOv12. Importantly, the method generalized effectively to unseen datasets (Roboflow and YouTube), where it maintained strong accuracy despite variations such as sea fog, sand occlusion, and vegetation. Compared to baseline models, the proposed network achieved substantial improvements in segmentation metrics, while retaining competitive inference speed. These results establish the proposed framework as a practical and robust solution for real-world underwater inspection. By effectively balancing accuracy, generalization, and efficiency, it offers a new paradigm for underwater visual perception.

While the proposed method demonstrates state-of-the-art performance in underwater pipeline segmentation, it may experience reduced sensitivity for pipelines narrower than 2 pixels due to information loss during feature down-sampling, though this represents a fundamental trade-off between computational efficiency and spatial resolution common across deep learning approaches. Similarly, completely buried pipelines or those heavily obscured by marine growth pose significant challenges, as our vision-based method relies on visual cues - a limitation shared by all optical imaging techniques in turbid environments. Furthermore, our evaluation focused exclusively on computer vision-based methods to maintain a controlled comparison within the scope of this research, acknowledging that multi-sensor approaches incorporating sonar or Lidar could provide complementary advantages in scenarios where optical visibility is severely compromised. These limitations, however, highlight valuable directions for future work rather than diminish the substantive advances achieved by the proposed method in optical pipeline recognition. Future work will focus on extending the framework to multi-class segmentation tasks, real-time deployment on embedded hardware, and integration into autonomous inspection systems.

Data availability

The HOMOMO dataset generated and analysed during the current study is not publicly available due to its large size and ongoing research derived from it, but is available from the corresponding author (Farahnaz Mohanna, F_mohanna@ece.usb.ac.ir) upon reasonable request.

Received: 23 September 2025; Accepted: 28 January 2026
Published online: 02 February 2026

References

1. Dang, T., Nguyen, T. T., Liew, A. W. C. & Eyad, E. Event classification on subsea pipeline inspection data using an ensemble of deep learning classifiers. *Cogn. Comput.* **17**(1), 10 (2025).
2. Xia, P., You, H. & Du, J. Visual-haptic feedback for ROV subsea navigation control. *Autom. Constr.* **154**, 104987 (2023).
3. Kartal, S. K. & Cantekin, R. F. Autonomous underwater pipe damage detection positioning and pipe line tracking experiment with unmanned underwater vehicle. *J. Mar. Sci. Eng.* **12**(11), 2002 (2024).
4. Guan, M. et al. An effective method for submarine buried pipeline detection via multi-sensor data fusion. *IEEE Access.* **7**, 125300–125309 (2019).
5. Du, X. et al. Revealing the potential of deep learning for detecting submarine pipelines in side-scan sonar images: an investigation of pre-training datasets. *Remote Sens.* **15**(19), 4873 (2023).
6. Tan, H., Zheng, L., Ma, C., Xu, Y. & Sun, Y. Deep learning-assisted high-resolution sonar detection of local damage in underwater structures. *Autom. Constr.* **164**, 105479 (2024).
7. Xiong, C., Lian, S. & Chen, W. An ensemble method for automatic real-time detection, evaluation and position of exposed subsea pipelines based on 3D real-time sonar system. *J. Civil Struct. Health Monit.* **13**(2), 485–504 (2023).
8. Bharti, V., Lane, D. & Wang, S. A semi-heuristic approach for tracking buried subsea pipelines using fluxgate magnetometers. In *Proceedings of 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 469–475 https://doi.org/10.1109/CASE48305.2020.9216755 (2020).
9. Su, J. et al. A high-accuracy underwater object detection algorithm for synthetic aperture sonar images. *Remote Sens.* **17**(13), 2112 (2025).
10. Jian, M., Yang, N., Tao, C., Zhi, H. & Luo, H. Underwater object detection and datasets: a survey. *Intell. Mar. Technol. Syst.* **2**(1), 9 (2024).
11. Saoud, L. S. et al. Seeing through the haze: A comprehensive review of underwater image enhancement techniques. *IEEE Access.* **12**, 145206–145233 (2024).
12. Sravya, N., Balakrishnan, A. A. & Supriya, M. H. An efficient underwater pipeline detection system using machine learning approach. In *Proceedings of 2019 IEEE International Symposium on Ocean Technology (SYMPOL)*, 181–190 (2019).
13. Rekik, F., Ayedi, W. & Jallouli, M. A trainable system for underwater pipe detection. *Pattern Recognit. Image Anal.* **28**(3), 525–536 (2018).
14. Sheng, M. et al. A new algorithm for AUV pipeline recognition and location. In *Proceedings of OCEANS 2018 MTS/IEEE*, https://doi.org/10.1109/OCEANS.2018.8604557 (2018).
15. Gasparovic, B., Lerga, J., Mausa, G. & Ivasic-Kos, M. Deep learning approach for objects detection in underwater pipeline images. *Appl. Artif. Intell.* **36**(1), 2146853 (2022).
16. Ma, Y., Cheng, Y. & Zhang, D. Comparative analysis of traditional and deep learning approaches for underwater remote sensing image enhancement: A quantitative study. *J. Mar. Sci. Eng.* **13**(5), 899 (2025).
17. Stamoulakatos, A. et al. Automatic annotation of subsea pipelines using deep learning. *Sensors* **20**(3), 674 (2020).
18. Medina, E., Campos, R., Gomes, J. G. R. C., Petraglia, M. R. & Petraglia, A. Convolutional neural networks for underwater pipeline segmentation using imperfect datasets. In *Proceedings of 2020 28th European Signal Processing Conference (EUSIPCO)*, https://doi.org/10.23919/Eusipco47968.2020.9287605 (2021).
19. Gasparovic, B., Mausa, G., Rukavina, J. & Lerga, J. Evaluating Yolov5, Yolov6, Yolov7, and Yolov8 in underwater environment: Is there real improvement? In *Proceedings of 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech)*, https://doi.org/10.23919/SpliTech58164.2023.10193505 (2023).
20. Thum, G. W., Tang, S. H., Ahmad, S. A. & Alrifaey, M. Toward a highly accurate classification of underwater cable images via deep convolutional neural network. *J. Mar. Sci. Eng.* **8**(11), 924 (2020).

21. Fathy, M. E., Mohamed, S. A., Awad, M. I. & Munim, H. E. A. E. A vision transformer based CNN for underwater image enhancement vitclaritynet. *Sci. Rep.* **15**(1), 16768 (2025).
22. Deng, R., Zhao, L., Li, H. & Liu, H. Cformer: An underwater image enhancement hybrid network combining Convolution and transformer. *IET Image Proc.* **17**(13), 3841–3855 (2023).
23. Yu, Y. et al. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. *Remote Sens.* **13**(18), 3555 (2021).
24. Liu, J., Liu, S., Xu, S. & Zhou, C. Two-stage underwater object detection network using Swin transformer. *IEEE Access.* **10**, 117235–117247 (2022).
25. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV48922.2021.00986 (2021).
26. Haider, A., Arsalan, M., Choi, J., Sultan, H. & Park, K. R. Robust segmentation of underwater fish based on multi-level feature accumulation. *Front. Mar. Sci.* **9**, 1010565 (2022).
27. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of International Conference on Machine Learning (ICML 2019)* https://doi.org/10.48550/arXiv.1905.11946 (2019).
28. Wang, H., Zhang, W., Xu, Y., Li, H. & Ren, P. WaterCycleDiffusion: Visual–textual fusion empowered underwater image enhancement. *Inf. Fusion*. **127**, 103693 (2026).
29. Wang, H., Frery, A. C., Li, M. & Ren, P. Underwater image enhancement via histogram similarity-oriented color compensation complemented by multiple attribute adjustment. *Intell. Mar. Technol. Syst.* **1**(12), 1–15 (2023).
30. Zhang, W., Wang, H., Ren, P. & Zhang, W. Underwater scene clarity reconstruction via multilayer information fusion and self-organized stitching. *IEEE Trans. Circuits Syst. Video Technol.* https://doi.org/10.1109/TCSVT.2025.3608828 (2025).
31. Li, H., Li, L., Wang, H., Zhang, W. & Ren, P. Underwater image captioning with AquaSketch-enhanced cross-scale information fusion. *IEEE Trans. Geosci. Remote Sens.* **63**, (2025).
32. Wang, H., Sun, S. & Ren, P. Underwater color disparities: cues for enhancing underwater images toward natural color consistencies. *IEEE Trans. Circuits Syst. Video Technol.* **34**(2), 738–753 (2024).
33. Li, X., Sun, W., Ji, Y. & Huang, W. S2G-GCN: A plot classification network integrating spectrum-to-graph modeling and graph convolutional network for compact HFSWR. *IEEE Geosci. Remote Sens. Lett.* **22**, 3506805 (2025).
34. Lian, S. et al. Diving into underwater: Segment Anything Model guided underwater salient instance segmentation and a large-scale dataset. In *Proceedings of the 41st International Conference on Machine Learning (PMLR 235)*, (2024).
35. Lian, S. et al. Instance segmentation for underwater imagery. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1305–1315 (2023).
36. Wang, Z. et al. Geometric matching for cross-modal retrieval. *IEEE Trans. Neural Networks Learn. Syst.* **36**(3), 5509–5521 (2025).
37. Wang, Z. et al. Evidence-based multi-feature fusion for adversarial robustness. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**(10), 8923–8937 (2025).
38. Whang, Z., Gao, Z., Han, M., Yang, Y. & Shen, H. T. Estimating the semantics via sector embedding for image-text retrieval. *IEEE Trans. Multimedia*. **26**, 10342–10353 (2024).
39. Wang, B., Yang, Y., Xu, X., Hanjalic, A. & Shen, H. T. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia* (*MM'17*), 154–162 (2017).
40. Li, Z., Shang, T. & Xu, P. Multi-modal attention perception for intelligent vehicle navigation using deep reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **26**(6), 8657–8669 (2025).
41. Li, Z., Xu, P., Dong, Z., Zhang, R. & Deng, Z. Feature-level knowledge distillation for place recognition based on soft-hard labels teaching paradigm. *IEEE Trans. Intell. Transp. Syst.* **26**(2), 2091–2101 (2025).
42. Li, Z., Shang, T., Xu, P., Deng, Z. & Zhang, R. Toward robust visual place recognition for mobile robots with an end-to-end dark-enhanced net. *IEEE Trans. Industr. Inf.* **21**(2), 1359–1368 (2025).
43. Xie, Y. et al. Landslide extraction from aerial imagery considering context association characteristics. *Int. J. Appl. Earth Obs. Geoinf.* **131**, 103950 (2024).
44. Xie, Y. et al. Localization, balance, and affinity: A stronger multifaceted collaborative salient object detector in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **63**, 4700117 (2024).
45. Zhu, J. et al. A cross-view intelligent person search method based on multi-feature constraints. *Int. J. Digit. Earth*, **17**(1), (2024).
46. Chen, H., Song, J., Han, C., Xia, J. & Yokoya, N. ChangeMamba: remote sensing change detection with spatio-temporal state space model. *IEEE Trans. Geosci. Remote Sens.* (2024).
47. Chen, H. et al. Mamba-convolution hybrid network for underwater image enhancement. *Sci. Rep.* **15**, 31975 (2025).
48. Shao, X. et al. Deep learning for multilabel classification of coral reef conditions in the indo-Pacific using underwater photo transect method. *Aquat. Conservation: Mar. Freshw. Ecosyst.* **34**, e4241 (2024).
49. Chen, X., Cai, Y., Wu, Y., Xiong, B. & Park, T. Multi-scale semantic segmentation with modified MBConv blocks. In *Proceedings of 2024 IEEE/CVF Conference on Computer Vision (ICCV)*, (2024).
50. Zheng, J., Liu, H., Feng, Y., Xu, J. & Zhao, L. CASF-Net: Cross-attention and cross-scale fusion network for medical image segmentation. *Comput. Methods Programs Biomed.* **229**, 107307 (2023).
51. Zhang, M. et al. Attention fusion of transformer-based and scale-based method for hyperspectral and lidar joint classification. *Remote Sens.* **15**(3), 650 (2023).
52. Nordstrom, M., Maki, A. & Hult, H. The impact label noise and choice of threshold has on cross-entropy and soft-dice in image segmentation. In *Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, https://doi.org/10.1109/CVPR52734.2025.01939 (2025).
53. Wang, L. et al. Cable segmentation based on Mask2Former in open-pit mining area. In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, https://doi.org/10.1109/INDIN58382.2024.10774442 (2024).
54. Shen, Z., Liu, W. & Xu, S. DS-SwinUNet: redesigning skip connection with double scale attention for land cover semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **18**, 4382–4395 (2025).
55. Upadhyay, A. K. & Bhandari, A. K. MaS-TransUNet: A multiattention Swin Transformer U-Net for medical image segmentation. *IEEE Trans. Radiation Plasma Med. Sci.* **9**(5), 613–626 (2025).
56. EL-Geneedy, M., Moustafa, E. L. D., Khater, H., Abd-Elsamee, H., Gamel, S. A. & S. & Advanced real-time detection of acute ischemic stroke using YOLOv12, YOLOv11, and YOLONAS: a comparative study for multi-class classification. *Sci. Rep.* **15**, 32546 (2025).

Author contributions

This article is derived from the Ph.D thesis of N.H, supervised by F.M and advised by M.K.M. All authors conceived and wrote the manuscript , provided the methodology and reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.M.