remote sensing

MDPI

Article

FSLA-RTDETR: An Improved Four-Scale RT-DETR Network for Dense Small-Object Seabed Pockmark Detection in Shaded-Relief Bathymetric Imagery

Chenyang Wang [1,2], Xiaofei Yan [1,3], Weibo Rao [2], Shaodong Li [1,3], Gang Chen [2], Rui Li [1,3] and Gang Li [1,3,*]

[1] Guangzhou Marine Geological Survey, China Geological Survey, Guangzhou 511458, China; wangchenyang_@cug.edu.cn (C.W.); yanxiaofei@mail.cgs.gov.cn (X.Y.); lishd23@alumni.sysu.edu.cn (S.L.); lrui@mail.cgs.gov.cn (R.L.)
[2] College of Marine Science and Technology, China University of Geosciences, Wuhan 430079, China; rwb@cug.edu.cn (W.R.); ddwhcg@cug.edu.cn (G.C.)
[3] Key Laboratory of Marine Mineral Resources, Ministry of Natural Resources, Guangzhou Marine Geological Survey, China Geological Survey, Guangzhou 511458, China
* Correspondence: lig@mail.cgs.gov.cn

Highlights

What are the main findings?

- A four-scale RT-DETR framework (FSLA-RTDETR) was developed for accurate detection of dense small-object seabed pockmarks in shaded-relief bathymetric imagery.
- The proposed framework combines the SAC-C2f backbone, LPEA Encoder, a DGCST-based feature fusion neck, and an additional high-resolution P_2 feature level to enhance small-object feature representation and localization accuracy.

What are the implications of the main findings?

- Under the present no-pretraining setting, FSLA-RTDETR outperforms the representative CNN- and Transformer-based object detectors evaluated in this study, demonstrating its effectiveness for dense small-object detection in complex seabed scenes.
- The proposed framework provides an efficient solution for automated seabed pockmark mapping, facilitating marine geological investigation and seabed hazard assessment.

Abstract

Seabed pockmarks are important geomorphological indicators of shallow fluid migration, shallow gas activity, and gas hydrate-related processes. However, automatic pockmark detection from shaded-relief bathymetric imagery remains challenging because many pockmarks have small image-space footprints, occur densely, have weak boundaries, and are embedded in complex seabed backgrounds. To address these issues, this study proposes FSLA-RTDETR, an improved four-scale RT-DETR network for dense small-object seabed pockmark detection. The model introduces SAC-C2f to enhance high-level feature representation, the LPEA Encoder to improve global contextual modeling, and a DGCST-based bidirectional feature fusion network to integrate shallow spatial details and deep semantic information. In addition, the high-resolution P_2 feature is incorporated into the decoder to improve small-object localization. A pockmark detection dataset was constructed from publicly available BOEM bathymetric data from the northern Gulf of Mexico, containing 880 images and 8013 annotated pockmark instances. Experimental results show that FSLA-RTDETR achieves mAP50, mAP75, and mAP50–95 values of 0.9218, 0.6221, and 0.5485, respectively. Compared with the baseline RT-DETR, mAP75 and

Check for updates

Academic Editor: Sergei Badulin

Received: 15 September 2026
Revised: 28 September 2026
Accepted: 30 September 2026
Published: 22 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.

mAP50–95 increase by 6.01 and 2.22 percentage points, respectively. Under the present no-pretraining setting, comparative experiments further demonstrate that FSLA-RTDETR achieves superior high-IoU localization performance compared with the representative object detection models evaluated in this study, providing an effective method for automatic seabed pockmark detection and geomorphological mapping.

Keywords: seabed pockmark; shaded-relief bathymetric imagery; small-object detection; RT-DETR; deep learning; seabed geomorphological mapping

1. Introduction

Submarine topography provides essential evidence for interpreting seafloor geological processes, evaluating marine resource potential, and assessing the environmental conditions relevant to offshore engineering [1]. Among various seafloor landforms, pockmarks are widely recognized as negative-relief features that commonly occur on continental shelves, continental slopes, and deep-sea basins. They are usually expressed as circular, subcircular, or elliptical depressions in bathymetric data [2–4]. The development of pockmarks is generally associated with fluid escape from shallow sediments, including methane seepage, pore-water expulsion, and shallow gas migration [5–8]. In many areas, pockmark fields are also related to sediment instability and submarine landslide processes, which may affect the safety of seabed pipelines, cables, and other engineering structures [9,10]. In addition, these depressions can create local habitat heterogeneity and may influence benthic communities and carbon cycling at the seafloor [8,11,12]. Reliable pockmark detection is therefore important for marine geological mapping, hydrocarbon exploration, gas hydrate assessment, and environmental studies of the seabed [13].

The increasing availability of multibeam echo-sounding and 3D seismic data has made it possible to map seabed geomorphology at much higher spatial resolution than before [14–17]. On this basis, a range of methods has been developed for identifying pockmarks from bathymetric data. Existing studies can be roughly grouped into manual interpretation, rule-based geomorphometric analysis, semi-automated GIS workflows, template matching, machine learning methods, and more recent deep learning approaches.

Manual interpretation was used in many early studies and remains valuable when detailed geomorphological judgment is required. For example, ref. [18] identified pockmarks through visual analysis of multibeam bathymetry in a GIS environment. Pockmark centers were determined according to slope-aspect variations, and their boundaries were delineated with the aid of multiple terrain-derived maps. This type of interpretation can provide detailed and reliable mapping results, but it is time-consuming and depends strongly on expert experience. When pockmarks occur in large numbers or need to be mapped over broad regions, manual delineation becomes difficult to apply efficiently [19].

To reduce the workload of manual mapping, rule-based geomorphometric methods have been introduced. Ref. [20] proposed a semi-automated framework that uses terrain attributes such as slope, curvature, and relative elevation, together with hydrological tools including depression filling and watershed segmentation, to identify pockmark-like depressions. Such methods improve processing efficiency and are relatively easy to implement. However, their performance is often sensitive to threshold selection and terrain assumptions. As a result, they may not generalize well to seabed areas with complex relief, variable illumination effects, or irregular pockmark morphology.

Semi-automated GIS workflows have further improved the efficiency of large-scale pockmark mapping. Refs. [21,22] developed GIS-based toolchains that combine terrain anal-

ysis and spatial segmentation for extracting and characterizing pockmarks from multibeam bathymetry. Ref. [23] also proposed a multi-step workflow for detecting large pockmark populations and analyzing their spatial distribution and hydrodynamic setting. These approaches are useful for regional mapping and statistical analysis, but they still require expert-defined parameters and several processing stages. This limits their adaptability when the seabed background, data resolution, or pockmark morphology changes.

Template matching provides another way to detect pockmarks by comparing bathymetric patterns with predefined geometric models. Ref. [24] proposed the POSIT method, which detects pockmarks by correlating spherical or ellipsoidal cap models with bathymetric data through a convolution–correlation framework. This strategy can achieve accurate localization without requiring training samples. Nevertheless, its effectiveness depends on the assumption that pockmarks can be approximated by simple geometric shapes. For irregular, partly eroded, or densely distributed pockmarks, exhaustive parameter searching also increases the computational cost.

Machine learning methods have recently been used to improve the discrimination between true pockmarks and other seafloor depressions. Ref. [25] combined Bathymetric Position Index-based candidate extraction, morphometric feature calculation, and Random Forest classification to refine pockmark detection. This approach helps reduce false positives compared with purely rule-based extraction. However, it still depends on handcrafted descriptors and pre-segmented candidate regions, so the detection process is not fully end-to-end.

Deep learning methods have also begun to be applied to seabed geomorphological interpretation. Ref. [26] used convolutional neural networks and generative adversarial networks to detect and segment pockmarks from bathymetric data [27–30]. Compared with handcrafted feature-based methods, deep networks can learn hierarchical representations directly from the input data and are better suited to complex seabed backgrounds. However, conventional CNN-based models may still have difficulty capturing long-range spatial relationships, and densely distributed pockmark targets with small image-space footprints can be weakened or lost after repeated downsampling.

Overall, existing pockmark detection methods have evolved from manual interpretation and rule-based geomorphometric analysis to data-driven learning-based approaches. Nevertheless, several challenges remain unresolved. First, many pockmark targets occupy only a limited portion of the image, making them vulnerable to information loss during repeated downsampling. Second, pockmarks are often densely distributed, which increases the difficulty of separating adjacent targets and accurately localizing their boundaries. Third, shaded-relief bathymetric images contain complex seabed textures, illumination-induced variations, and weak morphological contrasts, which can easily lead to false detections or missed detections. Therefore, an end-to-end detection framework that can jointly preserve shallow spatial details, enhance high-level semantic representation, and exploit global contextual information is needed for automatic pockmark detection from shaded-relief bathymetric imagery.

To address these challenges, this study proposes FSLA-RTDETR, a Four-Scale Synergistic Linear-Attention RT-DETR network for dense small-object seabed pockmark detection. The proposed model introduces SAC-C2f into the deeper stages of the backbone to enhance discriminative semantic representation for weak pockmark-related morphological patterns. The LPEA Encoder is employed to strengthen global contextual modeling of the deepest feature map while preserving positional information and local spatial structure. Inspired by [31], DGCST is adopted as a lightweight fusion block and embedded into the bidirectional feature fusion neck to improve multi-scale feature interaction. Finally, the high-resolution P_2 feature is incorporated into the RT-DETR [32] decoder, forming a

four-scale detection structure based on P_2, P_3, P_4, and P_5 features. The proposed framework is evaluated on a manually annotated pockmark dataset constructed from publicly available BOEM shaded-relief bathymetric data and is validated through ablation studies and comparisons with representative one-stage, two-stage, and DETR-like detectors.

The main contributions of this study are summarized as follows:

- A manually annotated seabed pockmark detection dataset was constructed from publicly available BOEM shaded-relief bathymetric data in the northern Gulf of Mexico, containing 880 images and 8013 annotated pockmark instances.
- A four-scale RT-DETR-based detection framework was developed for dense small-object pockmark detection, in which the high-resolution P_2 feature is incorporated to improve the localization of small-object targets with weak boundaries.
- Three task-oriented components are incorporated into FSLA-RTDETR, including the SAC-C2f-enhanced backbone, the LPEA Encoder, and a DGCST-based bidirectional feature fusion neck, to enhance semantic representation, global contextual modeling, and multi-scale feature interaction.
- Extensive ablation and comparative experiments demonstrate that the proposed FSLA-RTDETR improves high-IoU localization performance and outperforms representative one-stage, two-stage, and DETR-like detectors in terms of mAP75 and mAP50–95.

2. Study Area and Data

2.1. Study Area

The study area is located on the deepwater continental slope of the northern Gulf of Mexico, as shown in Figure 1. This area is recognized as one of the most active fluid seepage provinces worldwide, where submarine geomorphological features such as pockmarks, cold seeps, and mud volcanoes are widely developed [33].

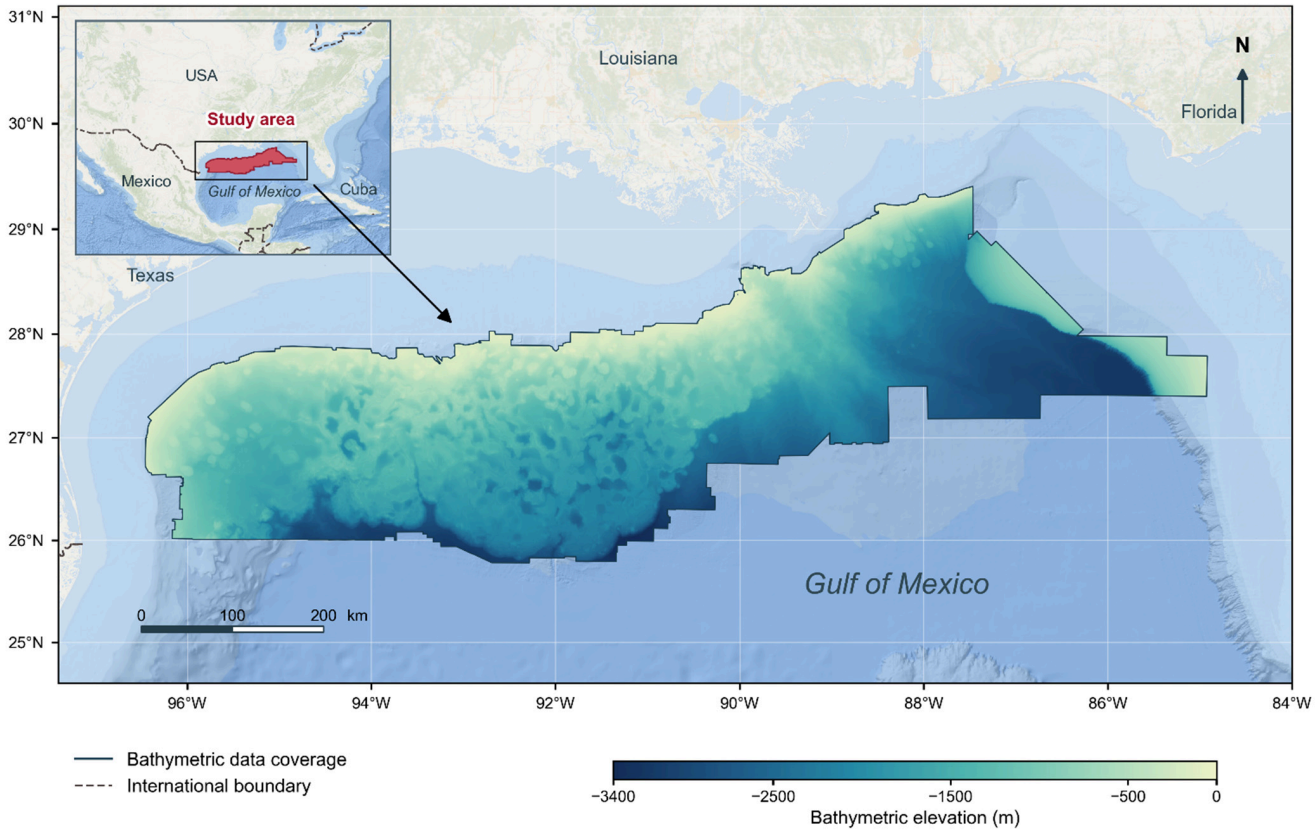


Figure 1. Location and bathymetric depth range of the study area.

The formation of pockmarks in this region is closely related to subsurface fluid migration, fault activity, and gas hydrate dissociation [34,35]. Due to high sedimentation rates and abundant fine-grained sediments, the northern Gulf of Mexico provides favorable conditions for the development and preservation of pockmarks [36].

In bathymetric imagery, pockmarks are expressed as local depressions with variable outlines and sizes, and they commonly occur in clustered patterns across different spatial scales [37]. Such morphology and spatial organization make the northern Gulf of Mexico a challenging case for testing whether automatic detection models can separate adjacent pockmarks and maintain stable performance across scale variations.

2.2. Data Description

The bathymetric data were taken from the BOEM open-access dataset Northern Gulf of Mexico Deepwater Bathymetry Grid and 3D Seismic, which provides regional deepwater bathymetric information for the northern Gulf of Mexico. The original digital elevation model (DEM) consists of approximately 1.4 billion grid cells, each with a spatial resolution of 40 × 40 feet, covering a depth range from −130 ft to −11,087 ft (approximately −40 m to −3379 m) [38]. The average vertical uncertainty of the DEM is estimated to be approximately 1.3% of water depth, which is acceptable for regional-scale geomorphological interpretation. The large spatial coverage and relatively fine grid resolution of this dataset make it suitable for examining pockmark morphology over the northern Gulf of Mexico.

In this study, the shaded-relief bathymetric imagery provided directly by BOEM was used as the model input rather than being independently generated from the raw depth grid by the authors. The imagery was obtained from BOEM's deepwater bathymetry and hillshade product, "Northern GoA Deepwater Bathymetry Grid from 3D Seismic." According to the official BOEM product documentation, the hillshaded rasters were generated in ArcMap using illumination from due north at a 45° angle and a vertical exaggeration factor of 5. No additional image normalization, tonal adjustment, or hillshade transformation was applied to the BOEM shaded-relief imagery. Hillshade visualization emphasizes subtle relief changes by introducing an illumination effect, making shallow depressions and their margins easier to recognize in areas with weak terrain contrast. This representation is therefore better aligned with the visual characteristics used during manual pockmark interpretation, especially for subtle depressions embedded in complex seabed textures.

Because the original bathymetric grid covers a broad region, it was first divided into smaller image patches using a sliding-window cropping strategy. A window size of 512 × 512 pixels and a stride of 512 pixels were used in both the horizontal and vertical directions, resulting in an overlap ratio of 0% between adjacent patches. No image resizing was applied during patch generation. For areas near the boundary of the study region that could not be fully covered by the sliding window, zero-padding was applied so that all patches had the same input size. This processing step allowed the model to use a fixed input dimension while still retaining edge-region information.

Only patches containing visible pockmarks were retained for annotation and construction of the primary dataset used for model training, validation, and testing. Patches without identifiable pockmark targets were excluded from this primary dataset to reduce redundant background samples. To further assess false-positive behavior in completely target-free seabed areas, 88 pockmark-free patches were subsequently selected from the excluded patches to form a separate target-free evaluation set. Each of these patches was manually inspected to confirm the absence of visible pockmarks and was used only for post-training evaluation, without being involved in model training, validation, or parameter optimization. However, the pockmark-containing patches in the primary dataset still include large areas of non-pockmark seabed, such as irregular terrain textures, illumination-

related shading, and other local relief variations. As a result, the detector still needed to distinguish true pockmark depressions from visually similar background structures within each patch.

The selected image patches were manually annotated using Labelme. Each pockmark was labeled with an axis-aligned bounding box and assigned to one category, "pockmark". Pockmark identification was based primarily on the geomorphological characteristics visible in the shaded-relief bathymetric imagery, with the corresponding original DEM used as an auxiliary reference. Candidate features were identified as pockmarks when they exhibited a coherent negative-relief morphology, typically with an approximately circular, subcircular, or elliptical planform, and a discernible depression relative to the surrounding seabed. The original DEM was consulted to verify that the interpreted feature corresponded to an actual bathymetric depression rather than solely to an illumination-related tonal variation in the shaded-relief image. Elongated or highly irregular terrain features, isolated shading variations without a corresponding bathymetric depression, and ambiguous features that could not be confidently distinguished from the surrounding seabed were excluded from annotation. During annotation, particular attention was given to clustered pockmarks and depressions with weak boundaries, because these cases were more likely to cause missed labels or inconsistent box placement.

The initial annotations were completed by the first author and then checked by a second expert. When disagreements occurred, the two annotators reviewed the corresponding image patches together and reached a final decision through discussion. To further evaluate annotation consistency, 100 image patches were randomly selected and independently annotated by both annotators. Within each image, bounding boxes from the two independent annotation sets that referred to the same visually identifiable pockmark instance were manually matched on a one-to-one basis. The IoU was then calculated for each matched pair of bounding boxes, and the reported mean IoU was obtained by averaging the IoU values across all matched pockmark instances. The average bounding-box IoU between the two annotation sets was 0.91, indicating a high level of inter-annotator consistency in bounding-box placement.

The primary pockmark dataset contains 880 images and 8013 annotated pockmark instances, corresponding to an average of about 9.1 targets per image. The dataset was randomly split into training, validation, and test sets at a ratio of 8:1:1. As a result, 704 images were used for training, 88 images for validation, and 88 images for testing. The additional 88 target-free patches were kept separate from these three subsets and were used exclusively for false-positive evaluation. Because the sliding-window stride was equal to the patch size, no overlapping image regions or duplicated pixels occurred between adjacent patches. However, since the primary dataset was randomly split at the patch level, spatially adjacent but non-overlapping patches could potentially be assigned to different subsets. Therefore, some degree of spatial autocorrelation between neighboring patches cannot be completely excluded.

Figure 2 summarizes the statistical characteristics of the annotated pockmark dataset. All annotations belong to the single class "pockmark". As shown in Figure 2a, the bounding-box aspect ratios are mainly concentrated around 1.0, indicating that most pockmarks have similar horizontal and vertical extents, which is consistent with their generally circular or elliptical morphology. Figure 2b shows a clear positive relationship between normalized bounding-box width and height, further confirming the relatively regular shape characteristics of most annotated pockmarks. The mean bounding-box width and height were 12.31 and 13.32 pixels, respectively, with a mean bounding-box area of 193.68 pixels^2. Given the raster cell size of 40 ft (approximately 12.192 m) per pixel, these mean bounding-box dimensions correspond to approximate physical extents of 150.06 m $\times$ 162.35 m. These physical

dimensions represent the approximate spatial extent of the annotated bounding boxes rather than a strict measure of pockmark diameter or an independent geomorphological size classification.

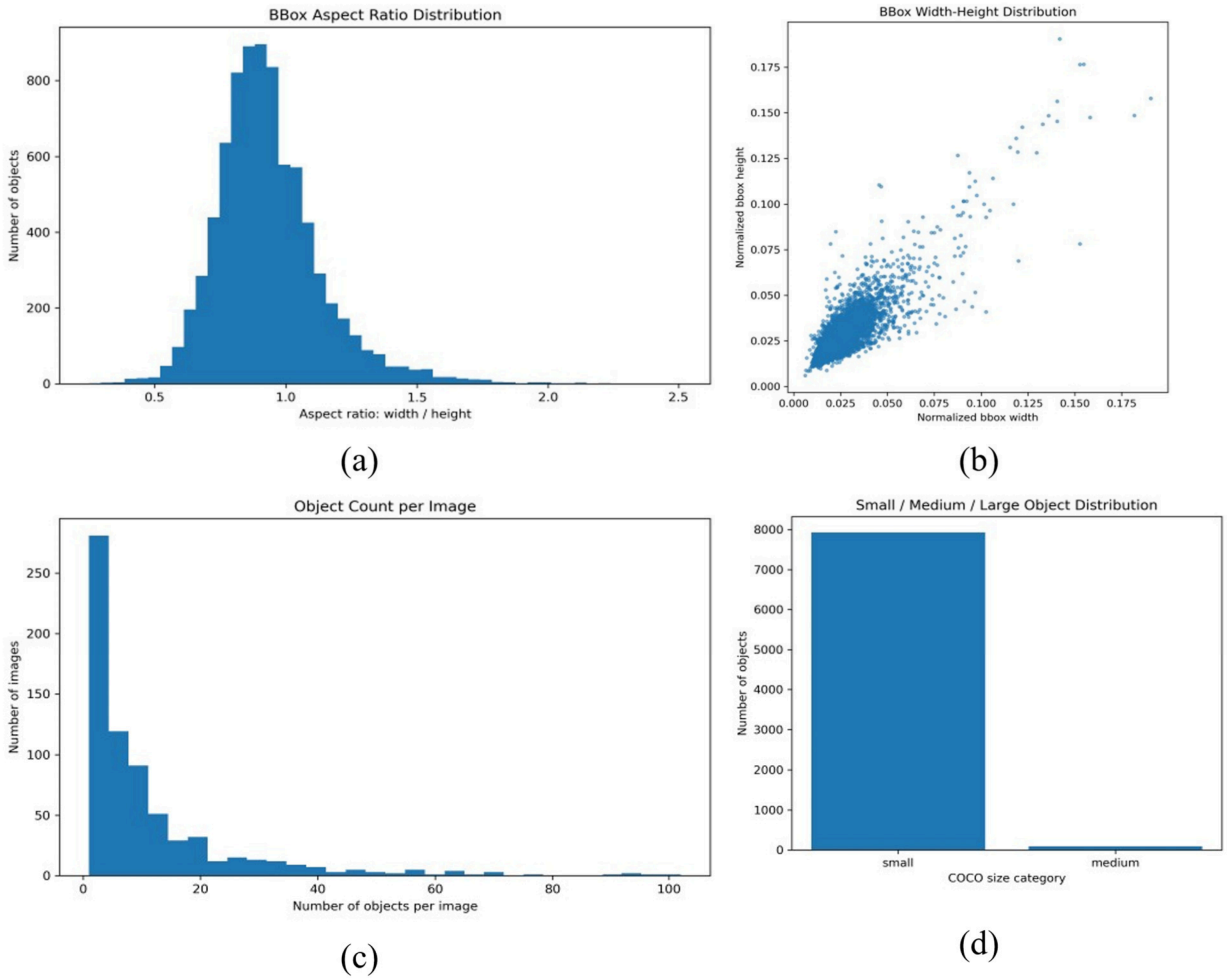


Figure 2. Statistical characteristics of the annotated pockmark dataset. (**a**) Bounding-box aspect-ratio distribution; (**b**) normalized bounding-box width–height distribution; (**c**) distribution of the number of pockmark instances per image; and (**d**) COCO-based object-size distribution, showing that the dataset is overwhelmingly dominated by small objects.

Figure 2c shows that the number of pockmarks per image varies considerably and exhibits a long-tailed distribution, indicating that some image patches contain densely clustered pockmarks whereas others contain only a small number of targets. Figure 2d further demonstrates the strong predominance of COCO-defined small objects in the dataset. According to the COCO object-size convention, 7925 of the 8013 annotated pockmarks (98.90%) were classified as small objects, while only 88 instances (1.10%) were classified as medium objects, and no large objects were present. Here, "small object" refers specifically to the image-space object-size category defined by COCO and should not be interpreted as indicating that the corresponding seabed pockmarks are necessarily small in absolute geomorphological dimensions. Overall, the dataset is therefore strongly dominated by small-object, densely distributed, and low-contrast seabed targets. These characteristics make pockmark detection challenging because target boundaries are often weak, neighboring depressions may overlap visually, and repeated downsampling in detection networks can progressively reduce fine-scale spatial information. The pronounced predominance of small objects also provides a clear motivation for introducing high-resolution feature repre-

sentations and enhanced multi-scale feature fusion in the proposed detector. In addition, the use of open-access BOEM bathymetric data improves the reproducibility of the dataset construction process and enables future studies to revisit the same source region.

3. Methods

3.1. Overall Architecture of FSLA-RTDETR

Dense pockmark fields in shaded-relief bathymetric images are difficult to detect because many targets occupy small image-space regions, exhibit weak boundaries, and are visually mixed with background terrain textures. To improve detection under these conditions, we developed FSLA-RTDETR, a Four-Scale Synergistic Linear-Attention RT-DETR network for automatic seabed pockmark detection. As illustrated in Figure 3, the network is organized around four parts: a SAC-C2f-enhanced backbone for feature extraction, an LPEA Encoder for high-level contextual modeling, a DGCST-based bidirectional neck for multi-scale feature fusion, and a four-scale RT-DETR decoder for final prediction.

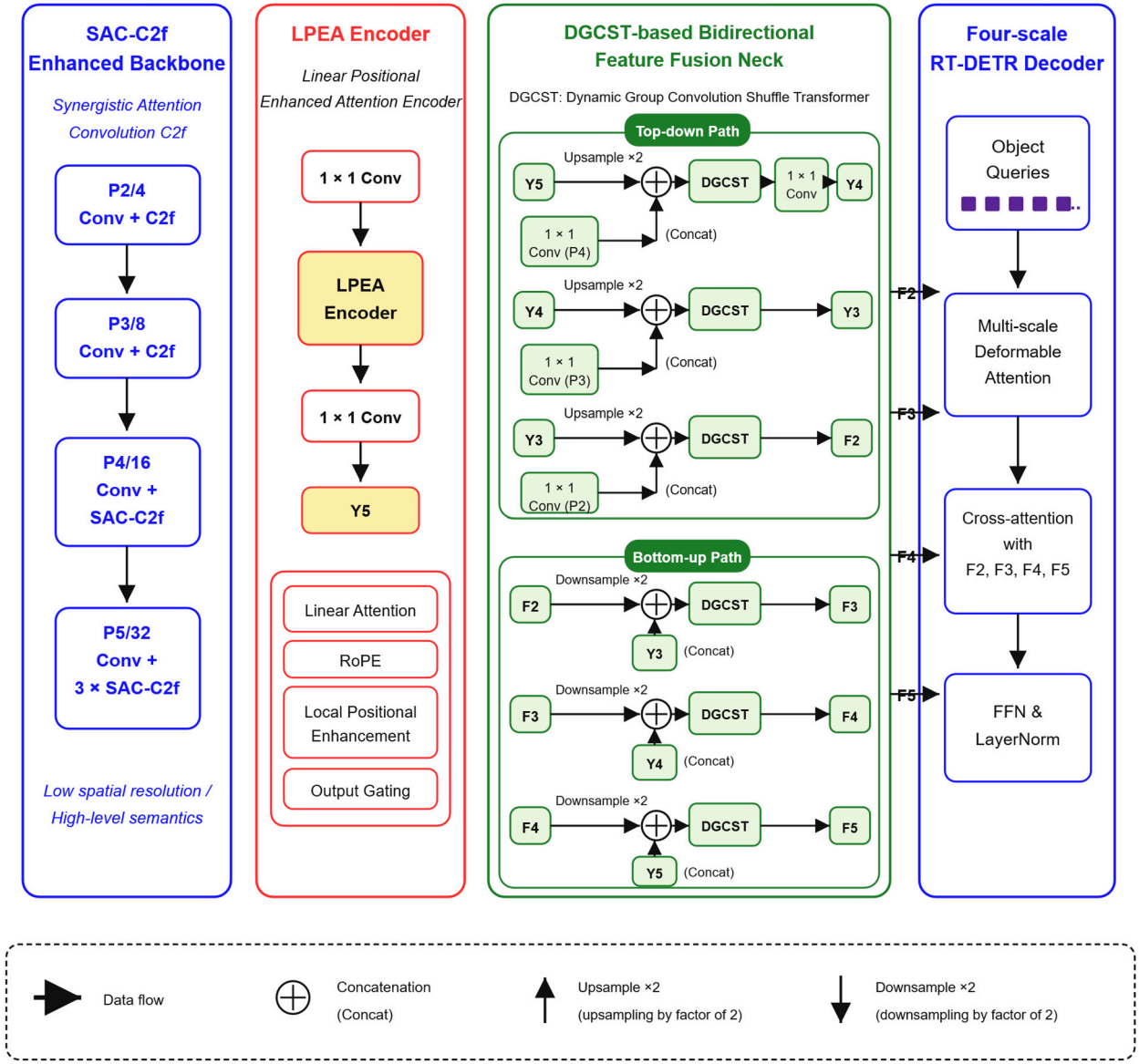


Figure 3. Overall architecture of FSLA-RTDETR.

Given an input shaded-relief bathymetric image $I \in R^{3\times H\times W}$, the backbone first extracts hierarchical features at four spatial scales:

$$\{P_2, P_3, P_4, P_5\} = \mathcal{B}(I), \tag{1}$$

where $\mathcal{B}(\cdot)$ denotes the backbone network. P_2, P_3, P_4, and P_5 represent feature maps with downsampling ratios of 4, 8, 16, and 32, respectively. Compared with the standard RT-DETR backbone, the proposed backbone introduces SAC-C2f modules into the deeper stages to enhance high-level semantic representation while maintaining efficient feature extraction.

The deepest semantic feature P_5 is further processed by the LPEA Encoder:

$$Y_5 = \mathcal{E}_{\text{LPEA}}\ (P_5), \tag{2}$$

where $\mathcal{E}(\cdot)$ denotes the LPEA Encoder and Y_5 represents the enhanced high-level semantic feature. By incorporating linear attention, positional enhancement, and local spatial refinement, the LPEA Encoder improves global contextual modeling for complex seabed backgrounds.

Subsequently, the DGCST-based bidirectional feature fusion neck integrates shallow spatial details and deep semantic information. The enhanced four-scale features are generated as follows:

$$\{F_2, F_3, F_4, F_5\} = \mathcal{N}_{DGCST}(P_2, P_3, P_4, Y_5), \tag{3}$$

where $\mathcal{N}(\cdot)$ denotes the bidirectional feature fusion neck, and F_2, F_3, F_4, and F_5 are the fused multi-scale features. The high-resolution P_2 feature is first projected to the unified channel dimension through a 1×1 convolution. Meanwhile, the top-down P_3-level representation is upsampled to the P_2 spatial resolution. These two feature maps are then concatenated along the channel dimension and refined by the DGCST module to generate the enhanced high-resolution feature F_2. Starting from F_2, the subsequent bottom-up pathway progressively propagates the fused spatial and semantic information to produce F_3, F_4, and F_5. The high-resolution feature F_2 retains fine spatial details that are important for localizing small-object pockmarks, while the deeper features provide stronger semantic and contextual information.

Finally, the four-scale RT-DETR decoder takes the fused features as input and outputs a fixed set of object predictions:

$$\hat{Y} = \mathcal{D}_{\text{RT-DETR}}\ (F_2, F_3, F_4, F_5), \tag{4}$$

where $\mathcal{D}(\cdot)$ denotes the RT-DETR decoder and $\hat{Y}$ represents the predicted pockmark detection results, including class probabilities and bounding-box coordinates. Specifically, the four fused feature maps $\{F_2, F_3, F_4, F_5\}$ are supplied to the decoder as four independent multi-scale feature inputs. The P_2-level information is therefore not directly concatenated with the other scales at the decoder entrance; instead, it is first incorporated into the bidirectional neck to form F_2, which is then processed together with F_3, F_4, and F_5 by the existing RT-DETR decoding mechanism. No additional attention module is introduced specifically for the P_2 branch.

Compared with the original RT-DETR, FSLA-RTDETR introduces four major improvements. First, SAC-C2f modules are embedded into the deeper stages of the backbone to strengthen discriminative feature representation for weak pockmark-related morphological patterns. Second, the LPEA Encoder is introduced to enhance global contextual modeling of the deepest feature map while preserving positional and local spatial information. Third, a DGCST-based bidirectional feature fusion neck is constructed to improve cross-scale

feature interaction. Fourth, the high-resolution P_2-level information is fused within the neck to generate F_2, and the resulting four-scale feature set $\{F_2, F_3, F_4, F_5\}$ is supplied to the RT-DETR decoder, forming a four-scale detection structure that is better suited to dense small-object pockmark detection.

3.2. SAC-C2f Enhanced Backbone

The backbone network is responsible for extracting hierarchical visual features from shaded-relief bathymetric imagery. In the shallow stages, standard convolutional layers and C2f modules are used to preserve local texture information and fine spatial details. However, pockmark targets often occupy small image-space regions and exhibit weak boundaries and low-contrast morphological expression, which can be easily confused with complex seabed background textures. Therefore, stronger semantic representation is required in the deeper stages of the backbone.

To enhance high-level feature representation, inspired by [39,40], this study introduces the Synergistic Attention-Convolution C2f module, termed SAC-C2f. SAC-C2f is built upon the original C2f structure and replaces the internal bottleneck units with SAC blocks. In this way, SAC-C2f preserves the efficient split-concat feature aggregation mechanism of C2f while improving the ability of the deeper backbone stages to capture discriminative pockmark-related morphological patterns.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the C2f structure first applies a convolutional projection and splits the projected feature into two branches:

$$X_1, X_2 = Split(Conv(X)). \tag{5}$$

One branch is retained for feature reuse, while the other branch is sequentially processed by n SAC blocks:

$$Y_i = \mathcal{S}_i(Y_{i-1}), i = 1, 2, \ldots, n, \tag{6}$$

where $Y_0 = X_2$, $\mathcal{S}_i(\cdot)$ denotes the i-th SAC block, and n represents the number of repeated SAC blocks. The intermediate features generated by the SAC blocks are concatenated with the retained branch and fused by a final convolution:

$$Y = Conv(Concat(X_1, X_2, Y_1, \ldots, Y_n)). \tag{7}$$

Inside each SAC block, the input feature $Y_{i-1} \in \mathbb{R}^{C\times H\times W}$ is first flattened into a token sequence:

$$T = Flatten(Y_{i-1}), T \in \mathbb{R}^{N\times C}, \tag{8}$$

where $N = H \times W$. The query, key, and value are generated from the same token sequence:

$$Q = K = V = T. \tag{9}$$

Multi-head self-attention is then applied to model contextual dependencies among spatial positions:

$$T_m = MHA(Q, K, V). \tag{10}$$

The attention output is reshaped into a two-dimensional feature map:

$$A = Reshape(T_m), A \in \mathbb{R}^{C\times H\times W}. \tag{11}$$

To further refine the attention response, a synergistic modulation module is introduced. The modulation module consists of pixel attention, channel attention, and multi-scale spatial attention. Pixel attention is formulated as

$$A_p = A \odot \sigma(BN(Conv_{1\times1}(A))), \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ represents element-wise multiplication. Channel attention is calculated by global average pooling followed by a lightweight fully connected transformation:

$$A_c = A \odot \sigma(FC(GAP(A))). \tag{13}$$

Multi-scale spatial attention is implemented using parallel dilated convolutions with dilation rates of 1, 6, 12, and 18:

$$A_s = Conv_{1\times1}\Big(Concat\big(Conv^r_{3\times3}(A)\big)_{r\in1,6,12,18}\Big). \tag{14}$$

The pixel-channel attention interaction is used to modulate the multi-scale spatial feature:

$$A_m = BN\big(Conv_{1\times1}\big(Softmax\big(A_pA_c\big)A_s\big)\big) + b + A, \tag{15}$$

where b is a learnable bias term. This modulation process adaptively combines pixel-level saliency, channel-wise discrimination, and multi-scale spatial context, which is beneficial for enhancing weak pockmark-related structures in complex seabed backgrounds.

The modulated feature is then flattened back into a token sequence and added to the original token sequence through a residual connection:

$$T_r = Dropout(LN(Flatten(A_m) + T)). \tag{16}$$

The normalized token feature is reshaped back into a two-dimensional feature map:

$$Z = Reshape(T_r), Z \in \mathbb{R}^{C\times H\times W}. \tag{17}$$

After that, a ConvolutionalGLU [39] module is applied for local feature refinement and gated feature selection. Specifically, Z is first projected by a 1×1 convolution and split into a feature branch and a gating branch:

$$U, V_g = Split(Conv_{1\times1}(Z)). \tag{18}$$

The feature branch is enhanced by depthwise convolution, while the gating branch controls information flow through element-wise multiplication:

$$G = DWConv(U) \odot V_g. \tag{19}$$

The gated feature is then projected back to the original channel dimension and added to the input feature through a residual connection:

$$Z_g = Z + Conv_{1\times1}(G). \tag{20}$$

Finally, the SAC block output is obtained through two-dimensional layer normalization and dropout:

$$\mathcal{S}_i(Y_{i-1}) = Dropout\big(LN_{2d}\big(Z_g\big)\big). \tag{21}$$

Compared with the original C2f module, SAC-C2f maintains the efficient cross-stage feature aggregation strategy but replaces the internal bottleneck transformation with SAC blocks. Through multi-head self-attention, pixel-channel-spatial synergistic modulation, and convolutional gated feed-forward refinement, SAC-C2f enhances both contextual representation and local feature discrimination. This design is particularly suitable for shaded-relief bathymetric imagery, where pockmark targets often have small image-space footprints and weak boundaries and are embedded in complex seabed backgrounds.

The detailed structure of the proposed module is illustrated in Figure 4.

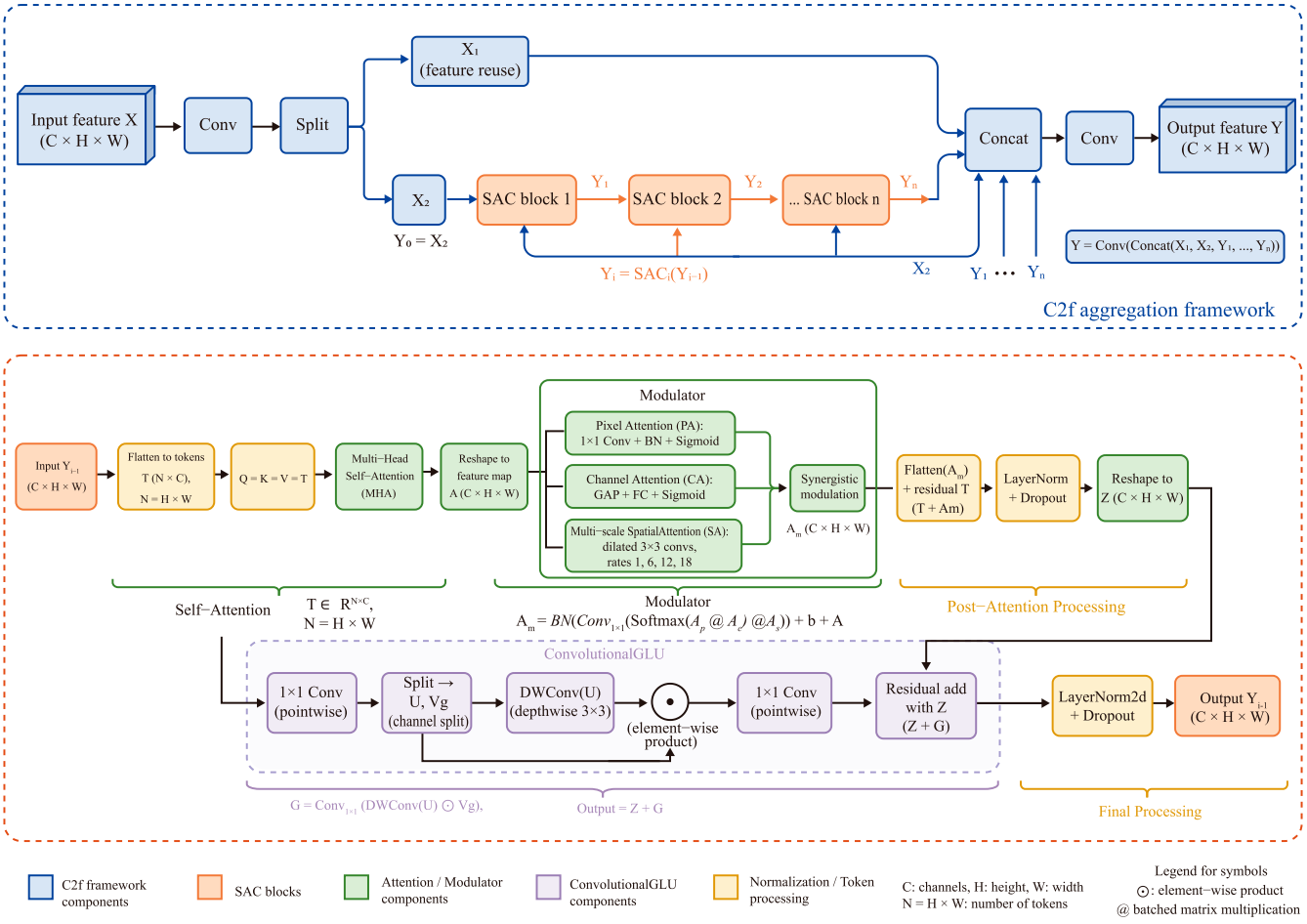


Figure 4. Architecture of the SAC-C2f module.

3.3. LPEA Encoder

To enhance the global contextual representation of high-level semantic features, the deepest feature map P_5 is fed into the Linear Positional Enhanced Attention Encoder, abbreviated as LPEA Encoder. In FSLA-RTDETR, the LPEA Encoder is applied only to the P_5 feature level. The encoder combines magnitude-aware linear attention [41], rotary positional embedding [42], local positional enhancement, output gating, and a convolutional feed-forward network to strengthen high-level semantic representation.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the attention module first generates query, key, value, and output-gating features through a 1×1 convolution:

$$[Q, K, V, O] = Conv_{1\times 1}(X), \tag{22}$$

where Q, K, V, and O denote the query, key, value, and output-gating features, respectively. The query, key, and value features are reshaped into multi-head token representations:

$$Q, K, V \in \mathbb{R}^{B \times h \times N \times d}, \tag{23}$$

where B is the batch size, h is the number of attention heads, $N = H \times W$ is the number of spatial tokens, and $d = C/h$ is the channel dimension of each head.

To enable linear attention computation, the query and key features are first mapped into a positive feature space using an ELU activation:

$$\begin{aligned} Q^{+} &= ELU(Q) + 1, \\ K^{+} &= ELU(K) + 1. \end{aligned} \tag{24}$$

The magnitude-aware normalization term is calculated before applying positional rotation:

$$z = Q^{+}\left(Mean\left(K^{+}\right)\right)^{T} \cdot d^{-\frac{1}{2}}, \tag{25}$$

where $d^{-1/2}$ is the scaling factor and the mean operation is performed along the token dimension.

To preserve spatial positional information, rotary positional embedding is then applied to the positive query and key features:

$$\hat{Q} = \Theta\left(Q^{+}\right), \hat{K} = \Theta\left(K^{+}\right), \tag{26}$$

where $\Theta(\cdot)$ denotes the RoPE operation. This operation introduces two-dimensional spatial positional information into the attention computation and helps the model preserve relative spatial relationships among seabed morphological structures.

The linear key-value aggregation is formulated as

$$S = \left(\frac{d^{-1/2}}{N}\right)\hat{K}^{T}V. \tag{27}$$

The magnitude-aware linear attention response is calculated as

$$R = \hat{Q}S \odot \left(1 + \frac{1}{z + \epsilon}\right) - z \odot Mean(V), \tag{28}$$

where $\boldsymbol{\epsilon}$ is a small constant for numerical stability, and $\odot$ denotes element-wise multiplication.

To enhance local spatial structure representation, a 5×5 depthwise convolution is applied to the value branch:

$$L = \mathrm{DWConv}_{5\times5}(V). \tag{29}$$

The local positional enhancement $\boldsymbol{L}$ is added to the linear attention response:

$$R' = R + L. \tag{30}$$

Finally, the enhanced response is modulated by the output-gating feature $\boldsymbol{O}$ and projected by a 1×1 convolution:

$$Y = \mathrm{Conv}_{1\times1}((R + L) \odot O). \tag{31}$$

The complete LPEA Encoder follows a post-normalization residual structure. The attention output is first added to the input feature and normalized:

$$X' = Norm(X + Dropout(Y_{MALA})). \tag{32}$$

Then, a convolutional feed-forward network consisting of two 1×1 convolutional layers and a GELU activation is applied:

$$FFN(X') = Conv_{1\times1}(Dropout(GELU(Conv_{1\times1}(X')))). \tag{33}$$

The final output of the LPEA Encoder is obtained through another residual connection and normalization:

$$Y_5 = Norm(X' + Dropout(FFN(X'))). \tag{34}$$

Through this design, the LPEA Encoder enhances long-range contextual modeling with linear attention while preserving positional and local spatial information through RoPE and depthwise convolution. This makes it suitable for detecting densely distributed pockmark targets with small image-space footprints in complex seabed terrain.

The detailed structure of the proposed module is illustrated in Figure 5.

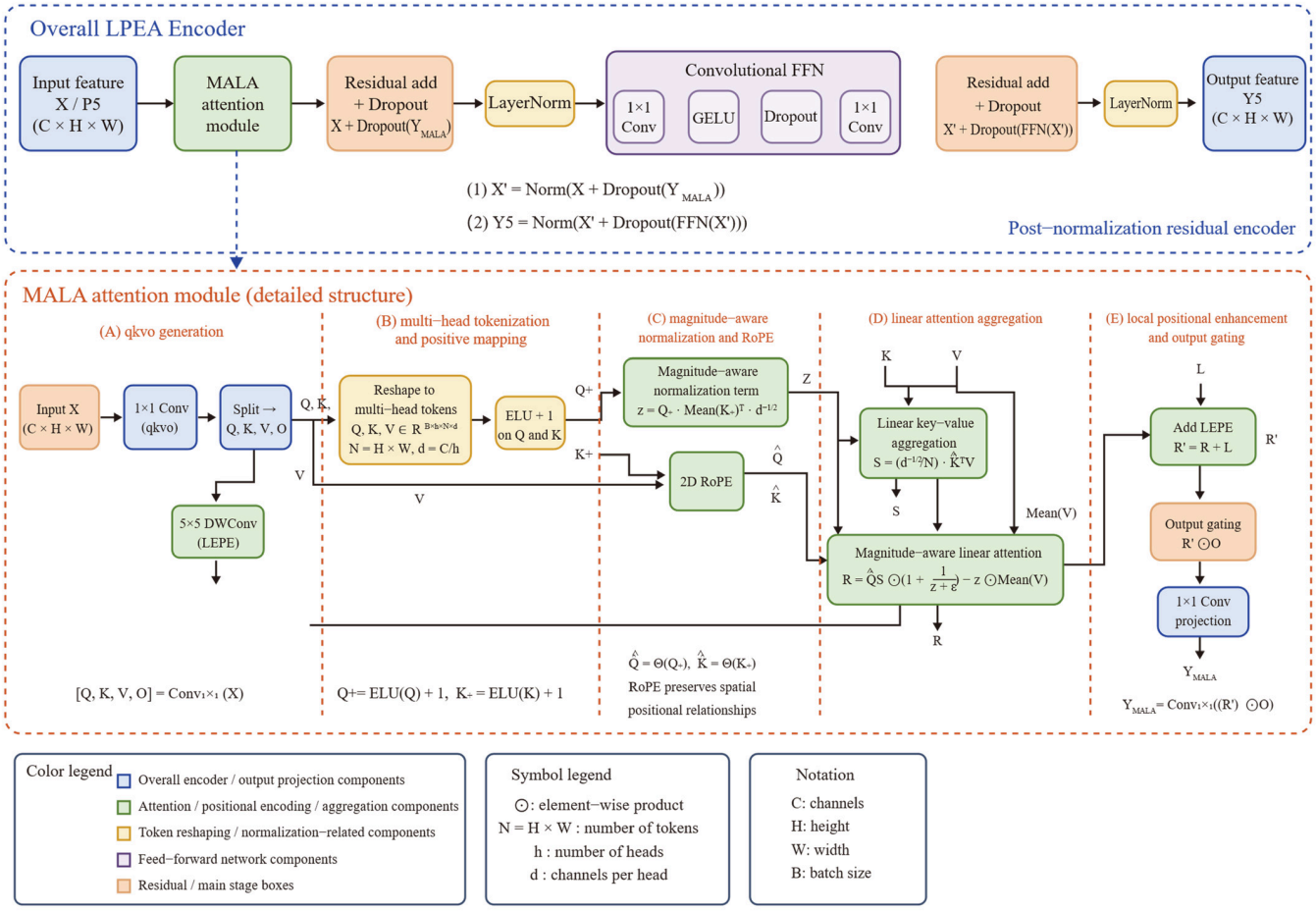


Figure 5. Architecture of the LPEA Encoder.

3.4. DGCST-Based Bidirectional Feature Fusion Neck

Although high-level features contain rich semantic information, small-object pockmarks are often more clearly represented in shallow feature maps with higher spatial resolution. Therefore, effective multi-scale feature fusion is essential for dense small-object pockmark detection. In FSLA-RTDETR, a DGCST-based bidirectional feature fusion neck is

constructed to integrate multi-scale features from P_2, P_3, P_4, and the encoded high-level feature Y_5.

The proposed feature fusion neck consists of a top-down pathway and a bottom-up pathway. In the top-down pathway, the encoded feature Y_5 is progressively upsampled and fused with lower-level backbone features after 1×1 channel projection. Specifically, Y_5 is upsampled and fused with the projected P_4 feature to generate Y_4. Then, Y_4 is upsampled and fused with the projected P_3 feature to generate Y_3. Finally, Y_3 is upsampled and fused with the projected P_2 feature to generate the high-resolution fused feature F_2. This top-down process transfers high-level semantic information to shallow high-resolution features.

The bottom-up pathway further enhances cross-scale information interaction. Starting from F_2, the feature is progressively downsampled and fused with Y_3, Y_4, and Y_5, producing the final fused features F_3, F_4, and F_5, respectively. Through this bidirectional design, shallow spatial details and deep semantic information can be effectively integrated.

The fusion process can be formulated as

$$\begin{aligned} Y_4 &= DGCST\left(Concat\left(Up(Y_5), P_4^{proj}\right)\right), \\ Y_3 &= DGCST\left(Concat\left(Up(Y_4), P_3^{proj}\right)\right), \\ F_2 &= DGCST\left(Concat\left(Up(Y_3), P_2^{proj}\right)\right), \\ F_3 &= DGCST(Concat(Down(F_2), Y_3)), \\ F_4 &= DGCST(Concat(Down(F_3), Y_4)), \\ F_5 &= DGCST(Concat(Down(F_4), Y_5)), \end{aligned} \tag{35}$$

where $\mathrm{Up}(\cdot)$ and $\mathrm{Down}(\cdot)$ denote upsampling and downsampling operations, respectively. P_i^{proj} denotes the feature obtained by projecting P_i with a 1×1 convolution. The final multi-scale features $\{F_2, F_3, F_4, F_5\}$ are used as the input to the four-scale RT-DETR decoder.

The Dynamic Group Convolution Shuffle Transformer, abbreviated as DGCST, is used as the basic fusion block in the neck. Given an input feature X, DGCST first applies a 1×1 convolution to unify the channel dimension:

$$X' = Conv_{1\times1}(X). \tag{36}$$

Then, the projected feature is split into two channel groups:

$$X_1, X_2 = Split(X'), \tag{37}$$

where $X_1 \in \mathbb{R}^{\frac{C}{4}\times H\times W}$ represents a lightweight spatial transformation branch, and $X_2 \in \mathbb{R}^{\frac{3C}{4}\times H\times W}$ is preserved as an information-retention branch. The first branch is processed by group convolution:

$$\widetilde{X}_1 = GConv(X_1), \tag{38}$$

where the number of groups is equal to the number of channels in X_1, making this operation equivalent to depthwise convolution on the partial channel branch. To enhance cross-channel information exchange, channel shuffle is then applied:

$$\hat{X}_1 = Shuffle\left(\widetilde{X}_1\right). \tag{39}$$

The shuffled branch is concatenated with the preserved branch:

$$X_f = Concat\left(\hat{X}_1, X_2\right). \tag{40}$$

Finally, two consecutive 1×1 convolutions are used to refine cross-channel interaction, and a residual connection is introduced to preserve the fused feature information:

$$Y_{DGCST} = X_f + Conv_{1\times1}\Big(Conv_{1\times1}\Big(X_f\Big)\Big). \tag{41}$$

DGCST differs from standard fusion blocks in the way it handles channel transformation. Instead of applying spatial convolution to the whole feature map, it processes only one channel subset and keeps the remaining channels as an information-preserving branch. This reduces unnecessary spatial computation while maintaining the original feature response to a large extent. After the partial group convolution, channel shuffle is used to mix the transformed features across channels, and the following 1×1 convolutions further refine the fused representation. In the proposed neck, this structure helps combine fine spatial cues from high-resolution features with semantic information from deeper layers. This is important for pockmark detection because shallow features are useful for localizing pockmark targets with small image-space footprints, but they also contain strong seabed texture variations that may interfere with target discrimination.

3.5. Four-Scale RT-DETR Decoder and Prediction Head

The enhanced multi-scale features generated by the DGCST-based bidirectional feature fusion neck are fed into the RT-DETR decoder for final object prediction. Unlike the conventional three-scale detection structure that mainly uses P_3, P_4, and P_5, FSLA-RTDETR incorporates the high-resolution F_2 feature into the decoder. Therefore, the decoder receives four feature levels:

$$\mathcal{F} = \{F_2, F_3, F_4, F_5\}, \tag{42}$$

where F_2, F_3, F_4, and F_5 correspond to feature maps with downsampling ratios of 4, 8, 16, and 32, respectively.

The inclusion of the F_2 feature is particularly important for seabed pockmark detection. Since many pockmarks occupy only a small fraction of the input image, repeated downsampling may weaken or even remove fine boundary cues. By introducing the high-resolution F_2 feature, the decoder can access richer shallow spatial information, which helps improve the localization of densely distributed pockmark targets with small image-space footprints.

Given a set of object queries Q_0, the RT-DETR decoder updates the query features by interacting with the four-scale memory features:

$$Q_l = \mathcal{D}_l(Q_{l-1}, \mathcal{F}), l = 1, 2, \ldots, L, \tag{43}$$

where Q_l denotes the object query features at the l-th decoder layer, L is the number of decoder layers, and $\mathcal{D}_l(\cdot)$ represents the l-th decoder layer. Through multi-scale feature interaction, the object queries progressively learn target-related semantic and spatial information from different feature levels.

After the final decoder layer, the prediction head generates classification scores and bounding-box coordinates for each object query:

$$\begin{aligned} \hat{p}_i &= Head_{cls}(q_i), \\ \hat{b}_i &= Head_{box}(q_i), \end{aligned} \tag{44}$$

where q_i denotes the i-th object query after decoding, $\hat{p}_i$ represents the predicted probability of the pockmark class, and $\hat{b}_i$ denotes the predicted bounding-box. Since this study focuses on single-class pockmark detection, the classification branch mainly distinguishes pockmark targets from background regions, while the regression branch estimates their bounding-box locations.

The final prediction set can be expressed as

$$\hat{Y} = \left\{ \left(\hat{p}_i, \hat{b}_i \right) \right\}_{i=1}^{N_q}, \tag{45}$$

where N_q denotes the number of object queries. Compared with the three-scale decoder, the four-scale decoder provides stronger small-object representation by combining high-resolution spatial details from F_2 with semantic features from deeper layers. This design is beneficial for detecting densely distributed, weakly bounded pockmark targets with small image-space footprints in shaded-relief bathymetric imagery.

To further investigate the feature-response behavior of FSLA-RTDETR, Grad-CAM visualization was conducted at two representative stages of the network, as shown in Figure 6. Two representative pockmark scenes were selected for qualitative comparison with the baseline RT-DETR.

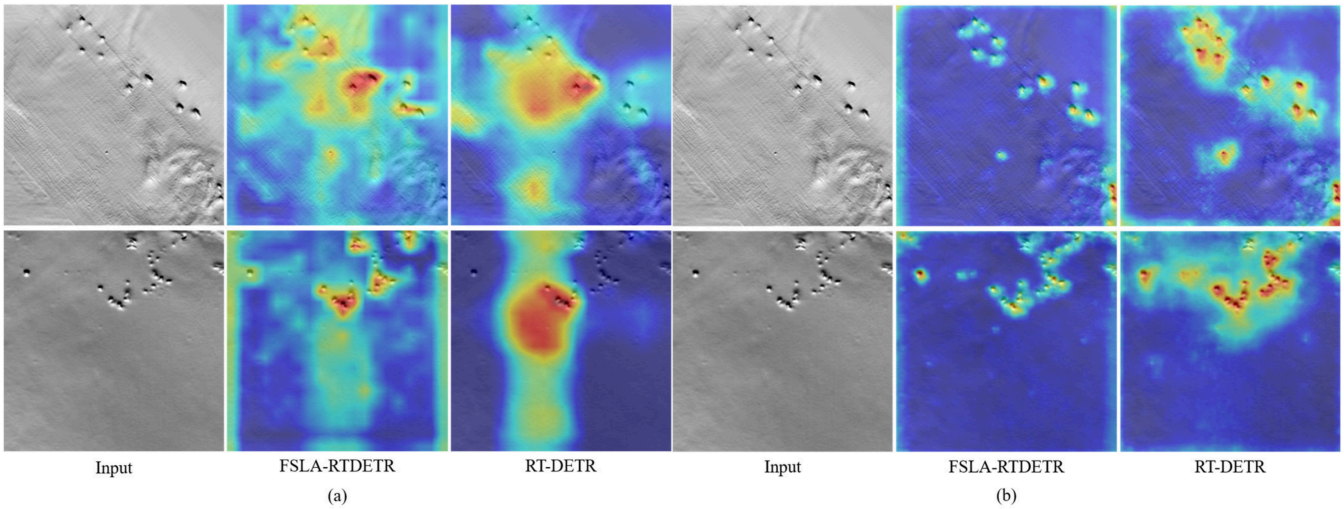


Figure 6. Grad-CAM visualization comparison between RT-DETR and FSLA-RTDETR for two representative pockmark scenes. (**a**) Feature responses after AIFI and LPEA, respectively; (**b**) multi-scale feature responses immediately before decoder input. Warmer colors indicate stronger activation responses.

In Figure 6a, the heatmaps are generated from the high-level encoded features after the AIFI module in RT-DETR and after the LPEA Encoder in FSLA-RTDETR. Compared with the baseline, FSLA-RTDETR generally produces more localized and target-oriented activation responses around pockmark regions, while the responses of RT-DETR are relatively broader and more diffuse over the surrounding seabed background. In densely distributed areas, the activation regions produced by FSLA-RTDETR also show better correspondence with individual pockmark targets. These qualitative observations suggest that the LPEA Encoder improves the discrimination of pockmark-related high-level features and reduces irrelevant responses to complex seabed textures.

Figure 6b further compares the multi-scale feature responses immediately before decoder input. For RT-DETR, the heatmaps are generated from its three decoder-input feature levels, whereas for FSLA-RTDETR, the visualization is derived from the four fused feature levels $\{F_2, F_3, F_4, F_5\}$. The responses from the selected multi-scale layers are aggregated to obtain the final Grad-CAM visualization. Compared with the baseline, FSLA-RTDETR exhibits more spatially concentrated responses around densely distributed small-object pockmarks, with less activation spreading into surrounding background regions. Neighboring targets also tend to retain more distinguishable response regions. This indicates that the additional high-resolution F_2 feature and DGCST-based bidirectional

feature fusion improve the preservation of fine spatial information and strengthen the representation of dense small-object pockmarks before decoding.

3.6. Training Objective

FSLA-RTDETR follows the set prediction training strategy of RT-DETR. Given a ground-truth set

$$Y = \left\{ (c_j, b_j) \right\}_{j=1}^{M}, \tag{46}$$

and a prediction set

$$\hat{Y} = \left\{ \left(\hat{p}_i, \hat{b}_i \right) \right\}_{i=1}^{N_q}, \tag{47}$$

where M is the number of ground-truth pockmarks and N_q is the number of object queries, Hungarian matching is used to establish a one-to-one assignment between predictions and ground-truth objects:

$$\hat{\sigma} = \underset{\sigma}{argmin} \sum_{j=1}^{M} \mathcal{C}_{\text{match}} \left(y_j, \hat{y}_{\sigma(j)} \right) \tag{48}$$

The matching cost consists of classification cost, $L1$ bounding-box cost, and GIoU cost:

$$\mathcal{C}_{match} = \lambda_{cls} \mathcal{C}_{cls} + \lambda_{L1} \left\| b_j - \hat{b}_{\sigma(j)} \right\|_1 + \lambda_{giou} \mathcal{C}_{giou} . \tag{49}$$

For Hungarian matching, the weighting coefficients for the classification, L1 bounding-box, and GIoU costs were set to 2, 5, and 2, respectively.

After matching, the final training loss is defined as

$$\mathcal{L} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{L1} \mathcal{L}_{L1} + \lambda_{giou} \mathcal{L}_{giou} , \tag{50}$$

where $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{L1}}$, and $\mathcal{L}_{\text{giou}}$ denote the classification loss, L1 bounding-box regression loss, and generalized IoU loss, respectively. For the final training loss, the corresponding weighting coefficients were set to 1, 5, and 2, respectively. The classification term employed Varifocal Loss, with a focusing parameter of $\gamma = 1.5$ and a balancing factor of $\alpha = 0.25$. Auxiliary losses from the intermediate decoder layers were also retained during training. This objective enables the model to distinguish pockmark targets from complex seabed backgrounds while improving bounding-box localization accuracy.

4. Experiments and Analysis

4.1. Ablation Study

4.1.1. Experimental Design

To evaluate the contribution of each component in the proposed FSLA-RTDETR, ablation experiments were conducted on the seabed pockmark detection dataset. The original RT-DETR was used as the baseline model. Based on this baseline, SAC-C2f, LPEA, the high-resolution P_2 feature level, and the DGCST-based bidirectional feature fusion neck were introduced under different configurations.

All models were trained and evaluated using the same dataset split, input image size, training strategy, and evaluation protocol. Precision, Recall, F1-score, mAP50, mAP75, and mAP50–95 were adopted as evaluation metrics. Since many pockmark targets have small image-space footprints, occur densely, and often exhibit weak boundaries in shaded-relief bathymetric imagery, mAP75 and mAP50–95 were considered important indicators for evaluating high-quality localization performance.

To further isolate the individual contributions of the P_2 feature level and DGCST, two additional variants were evaluated. In the P_2-only variant, the decoder feature dimension was kept at 256, identical to the baseline, so that the effect of introducing the additional

high-resolution feature level could be assessed independently. Similarly, a DGCST-only variant was constructed by replacing the baseline feature-fusion module with DGCST while keeping the other components unchanged. In the progressively integrated four-scale variants, the decoder dimension was reduced to 128 to balance detection performance and model complexity. The detailed component configurations are summarized in Table 1, while the corresponding quantitative performance and computational complexity are reported in Table 2.

Table 1. Component configurations of the ablation variants.

Variant	SAC-C2f	LPEA	P_2	DGCST	Decoder Dim.
V0 Baseline	–	–	–	–	256
V1	✓	–	–	–	256
V2	–	✓	–	–	256
V3	–	–	✓	–	256
V4	–	–	–	✓	256
V5	✓	✓	–	–	256
V6	✓	✓	✓	–	128
V7 FSLA-RTDETR	✓	✓	✓	✓	128

Note: "–" indicates that the corresponding modification is not added, and "✓" indicates that it is added.

Table 2. Quantitative performance and model complexity of the ablation variants.

Variant	Params (M)	GFLOPs	FPS	Precision	Recall	F1-Score	mAP50	mAP75	mAP50–95
V0 Baseline	19.87	56.9	275.70	0.8668	0.8329	0.8495	0.9173	0.5620	0.5263
V1	20.92	57.5	158.52	0.8687	0.8227	0.8451	0.9132	0.5677	0.5263
V2	19.95	57.2	270.27	0.8765	0.8177	0.8461	0.9173	0.5658	0.5255
V3	21.08	94.0	190.05	0.8708	0.8213	0.8453	0.9215	0.5804	0.5316
V4	18.50	**50.1**	**280.08**	0.8707	0.8290	0.8494	0.9167	0.5824	0.5308
V5	20.99	57.8	220.68	**0.8791**	0.8363	**0.8572**	**0.9283**	0.5967	0.5395
V6	19.73	79.8	126.54	0.8614	0.8128	0.8364	0.9157	0.6101	0.5434
V7 FSLA-RTDETR	**18.40**	77.1	194.94	0.8480	**0.8408**	0.8444	0.9218	**0.6221**	**0.5485**

Note: Bold indicates the best value for each metric, and underline indicates the second-best value. Lower is better for Params and GFLOPs, while higher is better for FPS, Precision, Recall, F1-Score, mAP50, mAP75, and mAP50–95.

4.1.2. Quantitative Results

Compared with the baseline RT-DETR, the complete FSLA-RTDETR improved Recall from 0.8329 to 0.8408, mAP50 from 0.9173 to 0.9218, mAP75 from 0.5620 to 0.6221, and mAP50–95 from 0.5263 to 0.5485. Among these metrics, mAP75 showed the most notable improvement, increasing by 6.01 percentage points. This indicates that the proposed model improves high-quality bounding-box localization for dense small-object pockmarks under stricter IoU conditions.

However, Precision decreased from 0.8668 to 0.8480, and F1-score slightly decreased from 0.8495 to 0.8444. This suggests that FSLA-RTDETR tends to detect more pockmark candidates, which improves Recall and high-IoU localization performance but may also introduce additional false positives. For dense seabed pockmark mapping, reducing missed detections and improving localization quality are particularly important, because missed pockmarks may affect subsequent spatial distribution analysis and geomorphological interpretation.

To further examine the robustness of the reported performance improvements, the baseline RT-DETR and the proposed FSLA-RTDETR were independently trained using

three random seeds (42, 123, and 3407), while all other training and evaluation settings were kept unchanged. As shown in Table 3, the baseline RT-DETR achieved mean mAP50, mAP75, and mAP50–95 values of 0.9142 ± 0.0029, 0.5585 ± 0.0084, and 0.5225 ± 0.0042, respectively. In comparison, FSLA-RTDETR achieved 0.9226 ± 0.0010, 0.6171 ± 0.0107, and 0.5478 ± 0.0020, corresponding to average improvements of 0.84, 5.86, and 2.53 percentage points, respectively. In particular, the improvements in mAP75 and mAP50–95 were consistently observed across all three random seeds, indicating that the gains in high-IoU localization performance were not limited to a single training run. However, Precision and Recall showed greater run-to-run variation for FSLA-RTDETR, suggesting that these threshold-dependent metrics remain more sensitive to stochastic training variation.

Table 3. Repeated-run performance of the baseline RT-DETR and FSLA-RTDETR over three random seeds. Values are reported as mean ± standard deviation.

Model	Precision	Recall	F1-Score	mAP50	mAP75	mAP50–95
Baseline RT-DETR	0.8655 ± 0.0015	0.8281 ± 0.0042	0.8463 ± 0.0028	0.9142 ± 0.0029	0.5585 ± 0.0084	0.5225 ± 0.0042
FSLA-RTDETR	0.8535 ± 0.0105	0.8339 ± 0.0187	0.8435 ± 0.0047	0.9226 ± 0.0010	0.6171 ± 0.0107	0.5478 ± 0.0020

Note: The results were obtained from three independent training runs using random seeds 42, 123, and 3407 under otherwise identical training and evaluation settings. Standard deviations were calculated across the three runs.

4.1.3. Effect of SAC-C2f and LPEA

When SAC-C2f was introduced alone, Precision slightly increased from 0.8668 to 0.8687, and mAP75 increased from 0.5620 to 0.5677. However, Recall decreased from 0.8329 to 0.8227, and F1-score decreased from 0.8495 to 0.8451. This indicates that SAC-C2f can slightly improve feature discriminability and high-IoU localization quality, but using it alone may make the detector more conservative for weak or ambiguous pockmark targets.

Similarly, when only the original encoder was replaced by LPEA, Precision increased from 0.8668 to 0.8765, and mAP75 increased from 0.5620 to 0.5658. However, Recall decreased from 0.8329 to 0.8177, and mAP50–95 slightly decreased from 0.5263 to 0.5255. This suggests that LPEA improves the contextual discrimination of high-level semantic features, but it does not independently improve the detection of all weak pockmark targets.

When SAC-C2f and LPEA were jointly used, the model achieved consistent improvements over the baseline. Precision increased from 0.8668 to 0.8791, Recall increased from 0.8329 to 0.8363, F1-score increased from 0.8495 to 0.8572, mAP50 increased from 0.9173 to 0.9283, mAP75 increased from 0.5620 to 0.5967, and mAP50–95 increased from 0.5263 to 0.5395. These results demonstrate that SAC-C2f and LPEA are complementary. SAC-C2f enhances high-level feature representation through synergistic attention-convolution modeling, while LPEA improves contextual encoding using linear positional attention.

4.1.4. Effect of the High-Resolution P_2 Feature Level

To isolate the contribution of the high-resolution P_2 feature level, an additional P_2-only variant (V3) was evaluated while keeping the RT-DETR decoder dimension fixed at 256, identical to the baseline configuration. Compared with the baseline (V0), introducing P_2 increased Precision from 0.8668 to 0.8708, mAP50 from 0.9173 to 0.9215, mAP75 from 0.5620 to 0.5804, and mAP50–95 from 0.5263 to 0.5316. The improvement was particularly evident for mAP75, which increased by 0.0184. These results indicate that the high-resolution P_2 feature provides additional fine-grained spatial information that is beneficial for precise localization of small-object pockmarks under stricter IoU criteria.

However, Recall decreased from 0.8329 to 0.8213 and F1-score decreased slightly from 0.8495 to 0.8453. This suggests that introducing high-resolution shallow features alone does

not uniformly improve all detection metrics and may also introduce additional background responses from irregular terrain textures and illumination-related shading.

The isolated P_2 experiment also clarifies the change in model complexity. With the decoder dimension fixed at 256, adding P_2 increased the parameter count from 19.87 M to 21.08 M and the computational cost from 56.9 to 94.0 GFLOPs. Therefore, the reduction in parameter count observed in the progressive V6 configuration should not be attributed to P_2 itself. In V6, the decoder dimension was simultaneously reduced from 256 to 128, which substantially decreased the number of learnable parameters in decoder-related projection, attention, and feed-forward operations. This parameter reduction exceeded the additional parameters introduced by the P_2 branch, resulting in a lower overall parameter count (19.73 M). Nevertheless, because P_2 operates on a substantially higher-resolution feature map, the computational cost still increased to 79.8 GFLOPs.

4.1.5. Effect of DGCST-Based Feature Fusion

To independently evaluate the contribution of DGCST, a DGCST-only variant (V4) was first compared with the baseline model. Introducing DGCST increased Precision from 0.8668 to 0.8707, mAP75 from 0.5620 to 0.5824, and mAP50–95 from 0.5263 to 0.5308. In particular, mAP75 improved by 0.0204, indicating that DGCST contributes to more accurate localization under stricter IoU criteria. Meanwhile, the parameter count decreased from 19.87 M to 18.50 M and GFLOPs decreased from 56.9 to 50.1, while inference FPS increased from 275.70 to 280.08. These results show that DGCST improves high-IoU localization performance while reducing model complexity and computational cost.

The effect of DGCST was further examined in the complete four-scale configuration. Compared with V6, which contains SAC-C2f, LPEA, and P_2 but does not use DGCST, the complete FSLA-RTDETR (V7) improved Recall from 0.8128 to 0.8408, F1-score from 0.8364 to 0.8444, mAP50 from 0.9157 to 0.9218, mAP75 from 0.6101 to 0.6221, and mAP50–95 from 0.5434 to 0.5485. At the same time, the parameter count decreased from 19.73 M to 18.40 M and GFLOPs decreased from 79.8 to 77.1.

These results indicate that DGCST provides complementary benefits in the four-scale architecture. By combining partial depthwise convolution, channel shuffle, and residual pointwise transformation, it facilitates interaction between shallow spatial details and deeper semantic information while limiting computational redundancy. This helps mitigate the additional background interference introduced by high-resolution features and improves both localization quality and model efficiency.

4.2. Comparative Experiments

4.2.1. Experimental Settings

To comprehensively evaluate the effectiveness of the proposed FSLA-RTDETR for seabed pockmark detection, several representative object detection models were selected for comparative experiments, including one-stage detectors, two-stage detectors, and DETR-like Transformer-based detectors. Specifically, YOLO11m [43] and YOLO26m [44] were adopted as representative one-stage detectors; Faster R-CNN-R50-FPN [28] and Cascade R-CNN-R50-FPN [45] were selected as classical two-stage detection baselines; and Deformable DETR-Refine-R50 [46] and DINO-4scale-R50 [47] were included as representative DETR-like Transformer detectors.

To establish a controlled comparison under the same data-availability condition, all models were trained from scratch without using pretrained weights and evaluated using the same training, validation, and test splits. This no-pretraining setting was adopted to exclude the influence of external pretraining datasets and to ensure that all models learned directly from the constructed pockmark dataset. The common training hyperparameters were

kept consistent across the compared models, including an input size of 512 × 512 pixels, a batch size of 4, 300 training epochs, the AdamW optimizer, an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, approximately 5 warm-up epochs, and a random seed of 42. Meanwhile, architecture-specific components and training mechanisms, including assignment or matching strategies, detection heads, loss formulations, proposal-generation procedures, and decoder structures, were retained according to the corresponding model implementations rather than being forcibly unified. The detailed implementation settings, including hardware environment, software versions, training hyperparameters, and FPS measurement protocol, are summarized in Table 4.

Table 4. Implementation details for model training and evaluation.

Item	Setting
GPU	NVIDIA GeForce RTX 5090
Python version	3.10.20
PyTorch version	2.7.0+cu128
Ultralytics version	8.4.26
MMDetection version	3.3.0
Input image size	512 × 512 pixels
Batch size	4
Training epochs	300
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Warm-up	5 epochs
Random seed	42
Pretrained weights	None
Maximum detections per image	300
Hungarian matching weights (classification/L1/GIoU)	2/5/2
Training loss weights (classification/L1/GIoU)	1/5/2
Varifocal Loss α	0.25
Varifocal Loss γ	1.5
Auxiliary decoder losses	Enabled for intermediate decoder layers
FPS measurement	Pre-processing + model inference + post-processing

The proposed FSLA-RTDETR differs from the compared baseline models in that it adopts a self-designed backbone network within an RT-DETR-style detection framework. This backbone is specifically designed to enhance feature representation for densely distributed, low-contrast pockmark targets with small image-space footprints in shaded-relief bathymetric imagery.

All models were evaluated on the test set. Precision, Recall, and F1-score were computed under a unified fixed-threshold setting, where the confidence threshold was set to 0.5, the IoU threshold for matching predicted boxes with ground-truth boxes was set to 0.5, and the maximum number of detections per image was set to 300. The mAP metrics, including mAP50, mAP75, and mAP50–95, were calculated using the standard COCO-style evaluation protocol. Because different detector architectures may benefit differently from large-scale pretraining, particularly for a relatively small, single-class dataset, the comparative results reported in this study should be interpreted specifically under the present no-pretraining setting rather than as an absolute ranking of the detector families under pretrained conditions.

4.2.2. Quantitative Results

Table 5 presents the quantitative comparison of different detectors on the pockmark test set. Overall, FSLA-RTDETR achieves the best performance in terms of F1-score, mAP50, mAP75, and mAP50–95 among all compared models. The proposed model obtains a Precision of 0.8480, Recall of 0.8408, F1-score of 0.8444, mAP50 of 0.9218, mAP75 of 0.6221, and mAP50–95 of 0.5485.

Table 5. Quantitative comparison of different detectors on the pockmark test set under the no-pretraining setting.

Model	Params	FPS	Precision	Recall	F1-Score	mAP50	mAP75	mAP50–95
YOLO11m	20.03M	151.52	0.8441	0.8227	0.8333	0.9010	0.4536	0.4818
YOLO26m	20.35M	158.73	0.8384	0.7825	0.8095	0.8773	0.4819	0.4860
Faster R-CNN-R50-FPN	43.35M	65.87	0.7054	**0.8752**	0.7812	0.8735	0.3627	0.4298
Cascade R-CNN-R50-FPN	69.15M	43.62	0.7025	0.8639	0.7748	0.8581	0.3544	0.4262
Deformable DETR-Refine-R50	40.76M	47.99	0.8225	0.7399	0.7790	0.8057	0.3816	0.4129
DINO-4scale-R50	47.56M	36.98	**0.9136**	0.7128	0.8008	0.8513	0.4062	0.4410
FSLA-RTDETR	**18.40M**	**194.94**	0.8480	0.8408	**0.8444**	**0.9218**	**0.6221**	**0.5485**

Note: Bold indicates the best result for each metric, and underline indicates the second-best result. For Params, lower is better; for FPS, higher is better; for all other metrics, higher is better.

Although DINO-4scale-R50 achieves the highest Precision of 0.9136, its Recall is relatively low, indicating that it tends to produce more conservative detections. Faster R-CNN-R50-FPN obtains the highest Recall among the compared baselines, reaching 0.8752, but its Precision and localization-related mAP metrics are relatively limited. In comparison, FSLA-RTDETR achieves a better balance between Precision and Recall, resulting in the highest F1-score of 0.8444. More importantly, FSLA-RTDETR shows a clear advantage in mAP75 and mAP50–95, demonstrating its stronger localization capability for seabed pockmarks.

Figure 7 further compares the precision–recall (PR) curves of the seven detectors at IoU thresholds of 0.50 and 0.75. As shown in Figure 7a, all models maintain relatively high precision over a broad recall range at IoU = 0.50, indicating that most detectors are capable of identifying pockmark targets under a relatively moderate localization criterion. Nevertheless, FSLA-RTDETR maintains a favorable precision–recall trade-off, particularly in the high-recall region. When the IoU threshold is increased to 0.75, as shown in Figure 7b, the differences among the detectors become substantially more pronounced. The PR curve of FSLA-RTDETR remains clearly above those of the other models over most of the recall range and extends to a higher recall level while retaining comparatively high precision. This result is consistent with its superior mAP75 performance and indicates that FSLA-RTDETR provides more accurate high-IoU bounding-box localization for dense small-object pockmarks. The larger performance advantage at IoU = 0.75 also suggests that the proposed high-resolution feature representation and multi-scale feature-fusion strategy are particularly beneficial when stricter spatial agreement between predicted and ground-truth bounding boxes is required.

4.2.3. Comparison with One-Stage Detectors

YOLO11m and YOLO26m were selected as representative one-stage detectors. Both models show competitive detection performance under the no-pretraining setting. YOLO11m achieves a Precision of 0.8441, Recall of 0.8227, F1-score of 0.8333, mAP50 of 0.9010, mAP75 of 0.4536, and mAP50–95 of 0.4818. YOLO26m obtains a Precision of 0.8384, Recall of 0.7825, F1-score of 0.8095, mAP50 of 0.8773, mAP75 of 0.4819, and mAP50–95 of 0.4860.

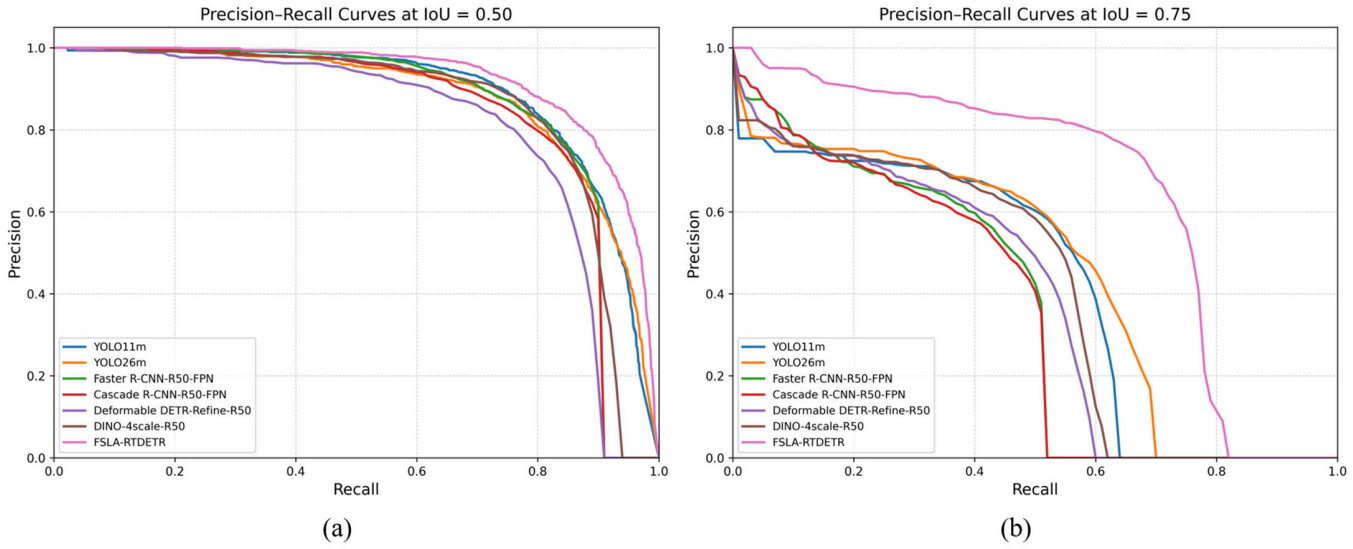


Figure 7. Precision–recall curves of different detectors on the pockmark test set at (**a**) IoU = 0.50 and (**b**) IoU = 0.75.

Compared with YOLO11m, FSLA-RTDETR improves mAP50 from 0.9010 to 0.9218, mAP75 from 0.4536 to 0.6221, and mAP50–95 from 0.4818 to 0.5485. Compared with YOLO26m, FSLA-RTDETR improves mAP50 by 0.0445, mAP75 by 0.1402, and mAP50–95 by 0.0625. These results indicate that although YOLO-based models have strong one-stage detection capability, FSLA-RTDETR provides more accurate pockmark localization, especially under stricter IoU thresholds.

The improvement in mAP75 is particularly important for this task. Since many pockmark targets have small image-space footprints, occur densely, and exhibit blurred boundaries in shaded-relief bathymetric imagery, a small deviation in the predicted bounding box may lead to a substantial decrease in IoU. Therefore, the higher mAP75 achieved by FSLA-RTDETR indicates that the proposed model provides more accurate high-IoU bounding-box localization and better spatial alignment between predicted and ground-truth boxes.

4.2.4. Comparison with Two-Stage Detectors

Faster R-CNN-R50-FPN and Cascade R-CNN-R50-FPN were used as two-stage baselines. Faster R-CNN obtains an mAP50 of 0.8735, mAP75 of 0.3627, and mAP50–95 of 0.4298. Cascade R-CNN obtains an mAP50 of 0.8581, mAP75 of 0.3544, and mAP50–95 of 0.4262. These results show that two-stage detectors can effectively detect pockmark targets at a relatively loose IoU threshold, but their high-IoU localization performance remains limited under the no-pretraining setting.

Compared with Faster R-CNN, FSLA-RTDETR improves mAP50 from 0.8735 to 0.9218, mAP75 from 0.3627 to 0.6221, and mAP50–95 from 0.4298 to 0.5485. Compared with Cascade R-CNN, FSLA-RTDETR improves mAP50 by 0.0637, mAP75 by 0.2667, and mAP50–95 by 0.1223. Although Cascade R-CNN is designed to improve localization accuracy through multi-stage bounding-box refinement, its performance under the present no-pretraining setting remains lower than that of FSLA-RTDETR, particularly at stricter IoU thresholds. This may be related to the limited training data scale, the lack of pretrained weights, and the ambiguous boundaries of seabed depressions.

Under the present no-pretraining setting, FSLA-RTDETR benefits from its task-oriented backbone design and RT-DETR-style detection framework, which enables more effective feature extraction and object localization for pockmarks in complex seabed geomorphological backgrounds. These results should therefore be interpreted as a comparison

under the current experimental conditions rather than as a general ranking of two-stage and RT-DETR-based detector families under pretrained settings.

4.2.5. Comparison with DETR-like Transformer Detectors

Deformable DETR-Refine-R50 and DINO-4scale-R50 were adopted as representative DETR-like Transformer detectors. Deformable DETR-Refine-R50 achieves an mAP50 of 0.8057, mAP75 of 0.3816, and mAP50–95 of 0.4129. DINO-4scale-R50 achieves an mAP50 of 0.8513, mAP75 of 0.4062, and mAP50–95 of 0.4410. DINO performs better than Deformable DETR-Refine, which may be related to its stronger query optimization and denoising training strategy.

However, both DETR-like baselines achieve lower detection accuracy than FSLA-RTDETR under the current no-pretraining setting. Compared with DINO, FSLA-RTDETR improves mAP50 from 0.8513 to 0.9218, mAP75 from 0.4062 to 0.6221, and mAP50–95 from 0.4410 to 0.5485. Compared with Deformable DETR-Refine, FSLA-RTDETR improves mAP50 by 0.1161, mAP75 by 0.2405, and mAP50–95 by 0.1356.

The comparatively lower performance of the R50-based DETR-like models under the present experimental setting suggests that these architectures may be more difficult to optimize from scratch on this relatively small, single-class dataset with densely distributed small-object targets. This performance advantage may reflect the task-oriented design of FSLA-RTDETR, which is tailored to weak-contrast, densely distributed pockmark targets with small image-space footprints. The increase in mAP75 also shows that the proposed model improves bounding-box localization at stricter IoU thresholds, rather than only increasing low-threshold detection accuracy. Nevertheless, these conclusions are restricted to the current no-pretraining setting. Because different detector architectures may benefit differently from large-scale pretraining, their relative performance may change when pretrained weights or alternative model-specific training strategies are used.

To facilitate a more intuitive comparison of the detection performance across several challenging scenarios, representative results are presented in Figures 8–10.

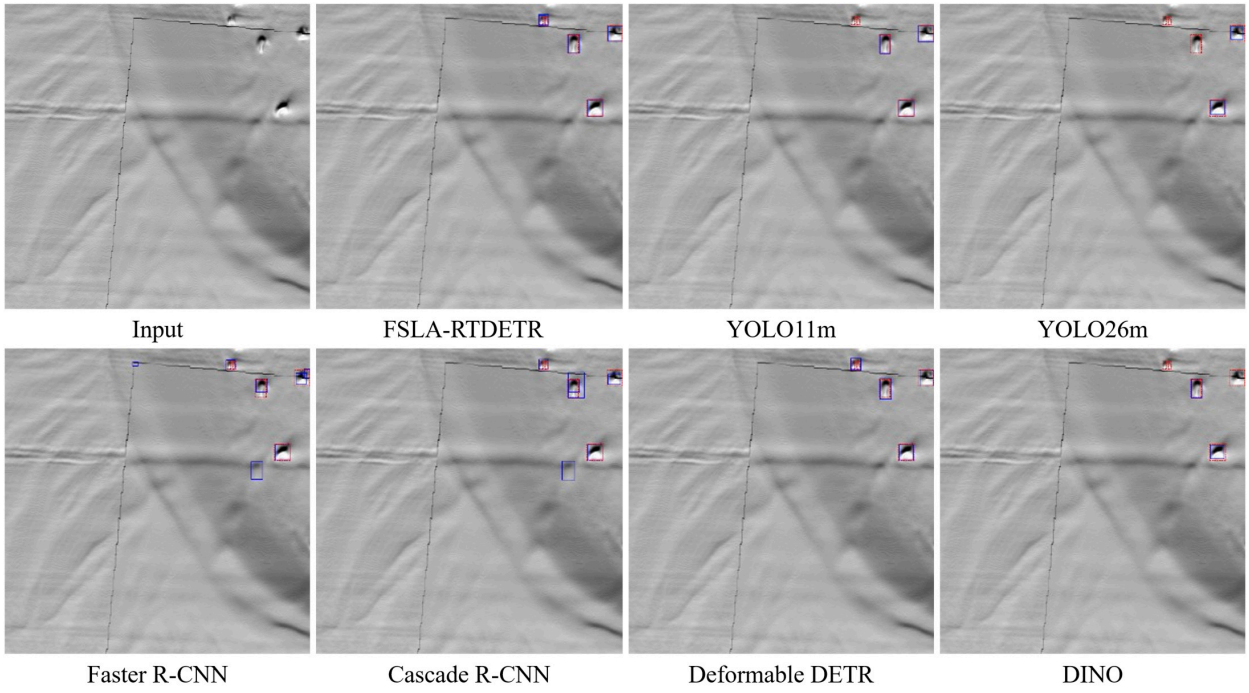


Figure 8. Qualitative comparison of different detectors for small-object pockmarks located near image boundaries. Ground-truth annotations are shown as red dashed boxes, and predicted bounding boxes are shown as blue solid boxes.

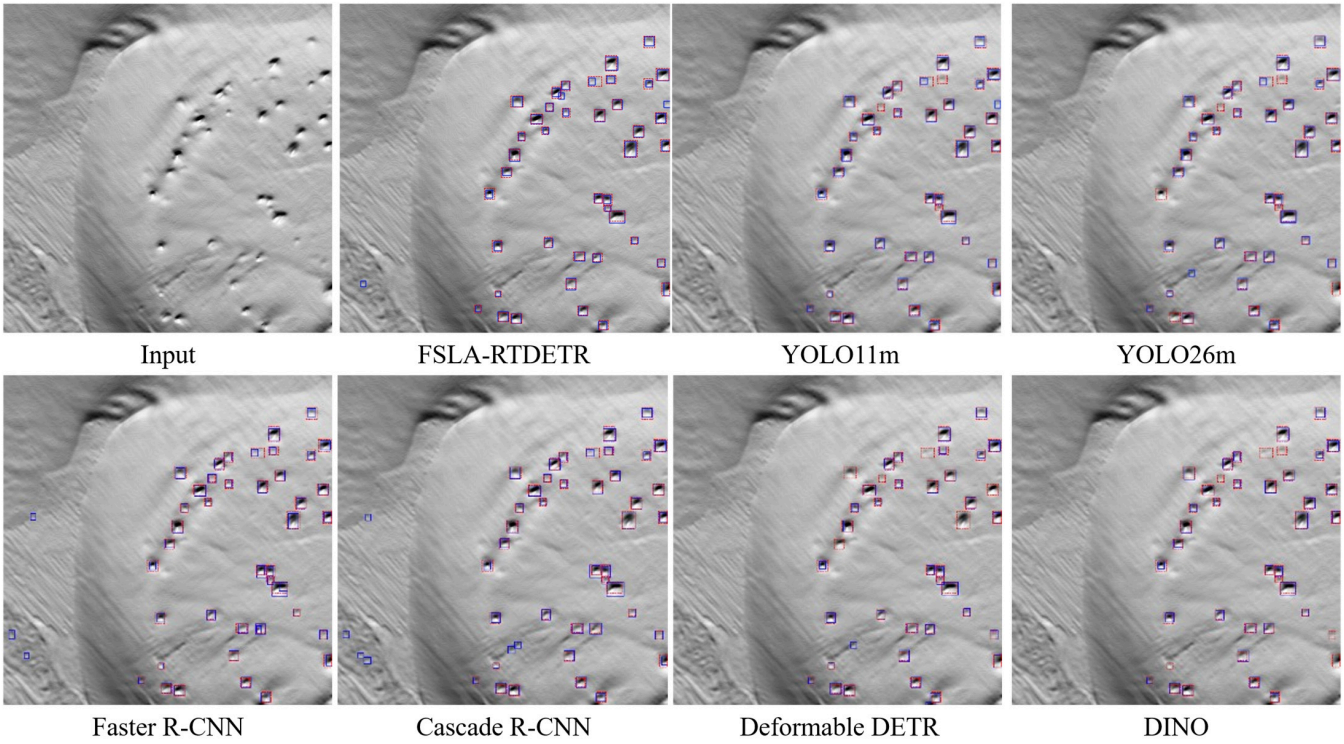


Figure 9. Qualitative comparison of different detectors for densely distributed small-object pockmarks. Ground-truth annotations are shown as red dashed boxes, and predicted bounding boxes are shown as blue solid boxes.

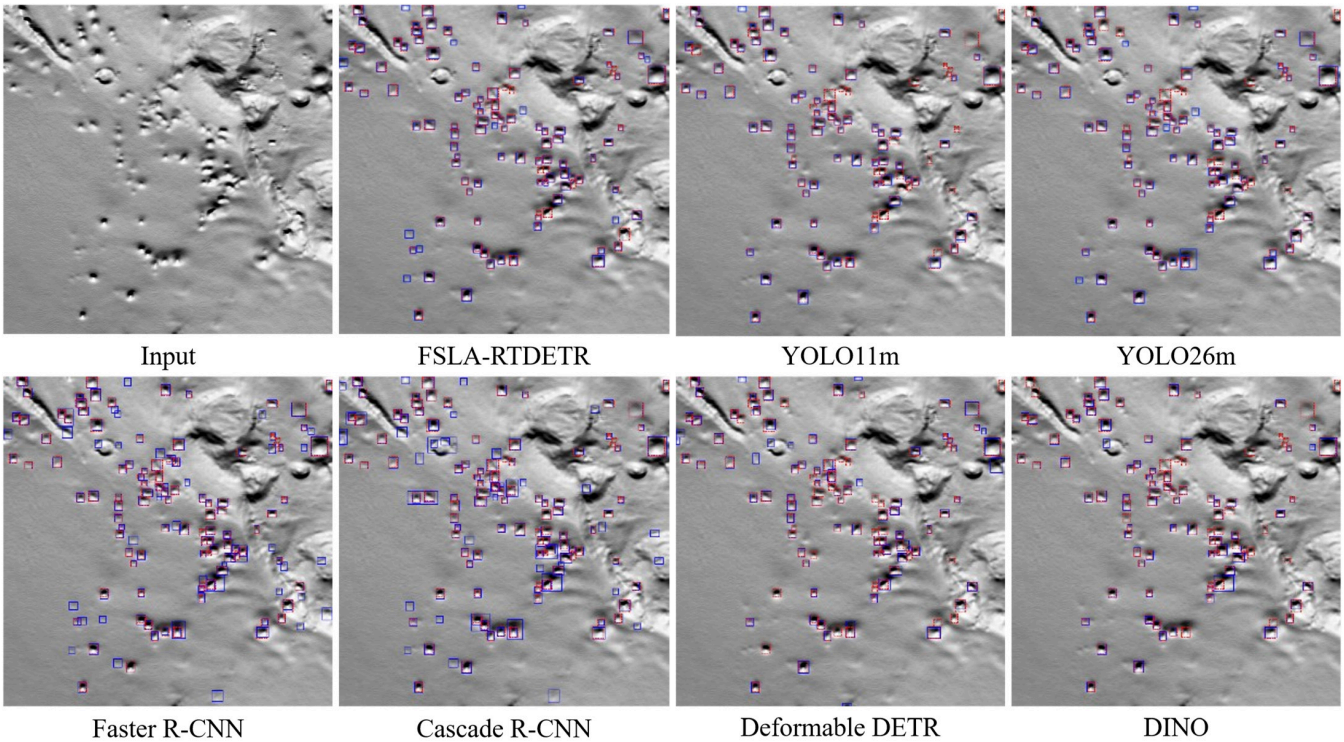


Figure 10. Qualitative comparison of different detectors for dense small-object pockmarks in complex seabed environments. Ground-truth annotations are shown as red dashed boxes, and predicted bounding boxes are shown as blue solid boxes.

4.3. Target-Free Background Evaluation

To further evaluate the false-positive behavior of FSLA-RTDETR in pockmark-free seabed areas, the trained model was tested on the additional target-free evaluation set described in Section 2.2. This set contains 88 manually verified 512 × 512-pixel patches with no visible pockmarks and was not involved in model training, validation, or parameter optimization. Since no ground-truth pockmarks are present in these images, all predicted pockmark instances are regarded as false-positive detections.

To characterize the false-positive behavior under different confidence criteria, the model was evaluated at confidence thresholds of 0.5, 0.6, 0.7, and 0.8. In addition to the total number of false-positive detections, the average number of false positives per image (FP/image) and the false-positive image rate (FPIR) were reported. FPIR represents the percentage of target-free images containing at least one false-positive detection. To further quantify the practical trade-off between false-positive suppression and true-pockmark retention, the original pockmark-containing test set was evaluated at the same confidence thresholds. At each threshold, predicted boxes were matched to the ground-truth annotations using an IoU threshold of 0.5 with one-to-one matching, and Precision, Recall, and F1-score were calculated.

As shown in Table 6, at a confidence threshold of 0.5, the model produced 57 false-positive detections across the 88 target-free patches, corresponding to 0.6477 false positives per image. Among these patches, 34 contained at least one false-positive detection, resulting in an FPIR of 38.64%. On the original pockmark-containing test set, the corresponding Precision, Recall, and F1-score were 0.8439, 0.8458, and 0.8448, respectively. When the confidence threshold was increased to 0.6, the total number of false-positive detections decreased to 34, with an FP/image of 0.3864 and an FPIR of 26.14%. At the same time, Precision increased to 0.9070, whereas Recall decreased to 0.7716 and the F1-score decreased slightly to 0.8338. At a threshold of 0.7, only 12 false-positive detections were observed in 11 patches, corresponding to an FP/image of 0.1364 and an FPIR of 12.50%. Precision further increased to 0.9633, while Recall and F1-score decreased to 0.6766 and 0.7949, respectively. When the threshold was further increased to 0.8, the model produced only four false-positive detections in four patches, resulting in an FP/image of 0.0455 and an FPIR of 4.55%. Under this stricter threshold, Precision reached 0.9891, but Recall decreased markedly to 0.4527, with an F1-score of 0.6212.

Table 6. Trade-off between true-pockmark detection and false-positive suppression at different confidence thresholds.

Confidence Threshold	Precision	Recall	F1-Score	Total FP	FP/Image	Images with ≥1 FP	FPIR (%)
0.5	0.8439	0.8458	0.8448	57	0.6477	34	38.64
0.6	0.9070	0.7716	0.8338	34	0.3864	23	26.14
0.7	0.9633	0.6766	0.7949	12	0.1364	11	12.50
0.8	0.9891	0.4527	0.6212	4	0.0455	4	4.55

Note: Precision, Recall, and F1-score were calculated on the original pockmark-containing test set using an IoU threshold of 0.5 for one-to-one prediction-to-ground-truth matching. Total FP, FP/image, Images with ≥1 FP, and FPIR were calculated on the 88 target-free patches. The Precision, Recall, and F1-score values in Table 6 were recomputed directly from saved predictions using explicit one-to-one matching for the confidence-threshold analysis and may therefore differ slightly from the framework-reported summary metrics in Table 5.

A consistent reduction in false-positive detections was observed as the confidence threshold increased. Increasing the threshold from 0.5 to 0.8 reduced the total number of false-positive detections from 57 to 4, corresponding to a reduction of approximately

93.0%. Meanwhile, the FPIR decreased from 38.64% to 4.55%. However, the improvement in false-positive suppression was accompanied by a progressive reduction in true-positive retention. For example, increasing the threshold from 0.5 to 0.6 reduced FP/image from 0.6477 to 0.3864 and FPIR from 38.64% to 26.14%, while Recall decreased from 0.8458 to 0.7716. At a threshold of 0.8, false-positive responses were strongly suppressed, but Recall decreased to 0.4527. These results demonstrate a clear threshold-dependent trade-off between false-positive suppression and true-pockmark retention.

Representative examples are presented in Figure 11. At relatively low confidence thresholds, some local depressions, irregular seabed textures, and illumination-induced shading patterns may be incorrectly identified as pockmarks because of their visual similarity to true pockmark features. As the confidence threshold increases, these false-positive detections are progressively reduced or eliminated. Nevertheless, a higher confidence threshold also removes an increasing proportion of lower-confidence true pockmark detections. Therefore, rather than defining a single universally optimal confidence threshold, the operating threshold should be selected according to the requirements of the specific application. Lower thresholds favor the retention of potential pockmark candidates, whereas higher thresholds provide stronger false-alarm suppression and may reduce the subsequent manual verification workload.

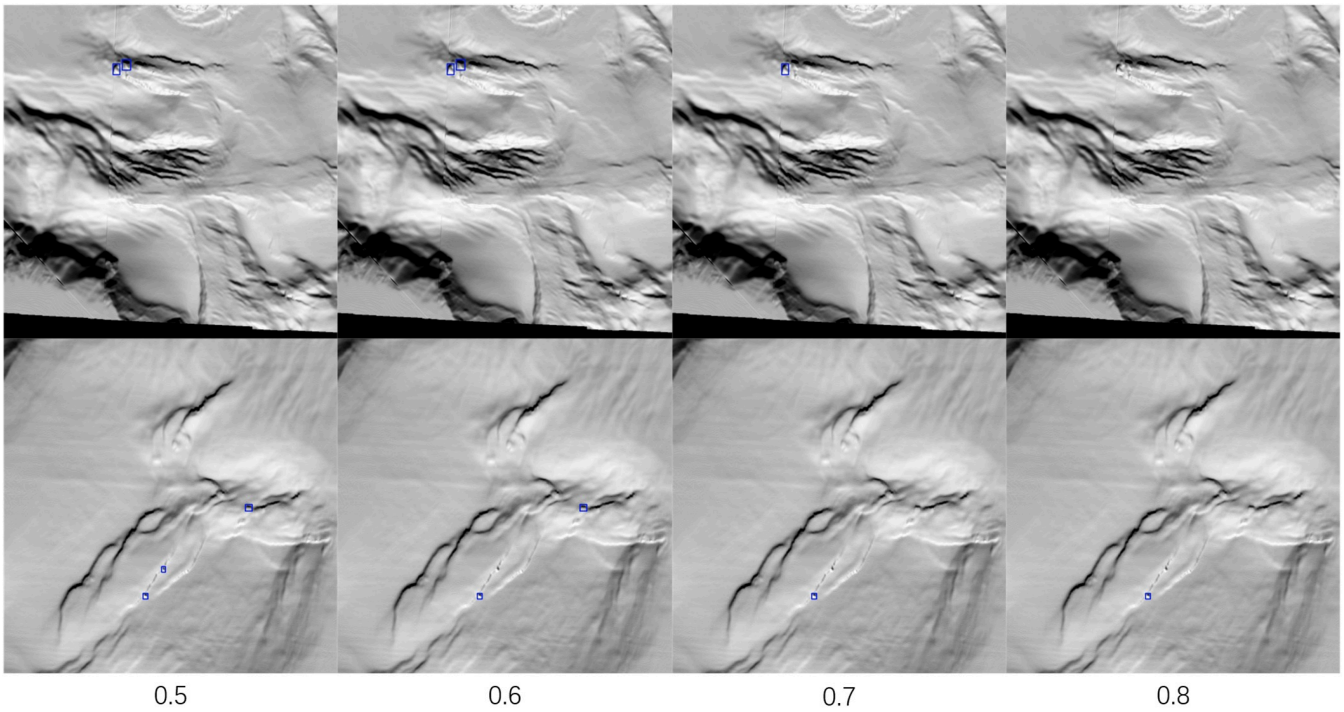


Figure 11. Representative false-positive detections on target-free bathymetric patches under different confidence thresholds. Bounding boxes indicate false-positive pockmark detections. As the confidence threshold increases from 0.5 to 0.8, most low-confidence false detections are progressively suppressed.

5. Discussion

FSLA-RTDETR was developed for the detection of dense small-object seabed pockmarks in shaded-relief bathymetric imagery. Compared with general object detection tasks, this problem is more difficult because many pockmarks occupy only a small part of the image, have weak or incomplete boundaries, and are often mixed with complex seabed textures. These characteristics make both target recognition and bounding-box localization challenging, especially when adjacent pockmarks occur in clusters.

On the constructed pockmark detection dataset, FSLA-RTDETR achieved mAP50, mAP75, and mAP50–95 values of 0.9218, 0.6221, and 0.5485, respectively. Compared with the baseline RT-DETR, mAP75 increased by 6.01 percentage points and mAP50–95 increased by 2.22 percentage points. The larger gain in mAP75 is particularly meaningful for this task. In pockmark mapping, a loose detection box may still contribute to mAP50, but it is less useful for describing the actual position and extent of an individual depression. The improvement under stricter IoU evaluation suggests that FSLA-RTDETR does not simply detect more candidate targets but also improves the spatial agreement between predicted boxes and annotated pockmarks.

The ablation experiments also indicate that the different components of FSLA-RTDETR contribute in different ways. SAC-C2f improves the representation of weak pockmark-related morphological patterns in deeper backbone stages. The LPEA Encoder helps the model use broader contextual information from high-level features, which is useful when individual pockmarks are not clearly separated from the surrounding terrain. Introducing the high-resolution P_2 feature gives the decoder access to finer spatial details and helps preserve small-target information that may be weakened after repeated downsampling. However, shallow features also contain strong seabed texture variations and illumination-related patterns. The DGCST-based bidirectional feature fusion neck helps balance these two effects by combining shallow localization cues with deeper semantic information, which contributes to the improvements in Recall, mAP75, and mAP50–95.

From the perspective of seabed geomorphological mapping, the proposed detector can be used as a candidate extraction tool for large bathymetric regions. The detection results can support pockmark distribution mapping, local density estimation, and preliminary analysis of spatial patterns. Since pockmarks are commonly associated with fluid escape, shallow gas activity, and gas hydrate-related processes, automatic detection can reduce the workload of manual interpretation during the early stage of regional seabed analysis. The method is therefore more suitable for large-area screening and object-level mapping than for replacing detailed expert geomorphological interpretation.

Several limitations remain. First, densely clustered pockmarks are still difficult to separate in some cases. When neighboring pockmarks are closely spaced, their shaded-relief expressions may merge into a connected depression cluster. The original shape of each pockmark then becomes less distinguishable, and the model may miss individual targets or produce incomplete separation. This problem is likely to become more obvious when such clustered samples are insufficiently represented in the training set.

Second, weak boundaries and low local contrast still affect detection quality. Some pockmarks gradually transition into the surrounding seabed terrain, and their margins are not clearly expressed in the hillshade image. In these cases, the predicted box may shift away from the annotated depression boundary, or weak pockmarks may be confused with ordinary seabed undulations. Hillshade imagery may also enhance scarps, sediment waves, illumination-related textures, and other non-pockmark depressions, which can produce false positives. Pockmarks located close to image boundaries are another difficult case because part of their spatial context may be missing after patch cropping.

Third, the current framework represents pockmarks using bounding boxes. This representation is efficient for object-level detection and regional candidate mapping, but it does not directly delineate geomorphological boundaries and cannot describe detailed geomorphological attributes such as boundary irregularity, elongation, internal relief variation, or depression depth. Therefore, FSLA-RTDETR is better suited for object-level spatial localization of potential pockmark targets than for full geomorphometric characterization. Accordingly, improvements in high-IoU detection metrics such as mAP75 should be inter-

preted as better agreement between predicted and ground-truth bounding boxes rather than as more accurate delineation of the actual geomorphological boundaries of pockmarks.

Fourth, the current random patch-level split does not ensure complete spatial independence between the training, validation, and test subsets. Although the 512 × 512-pixel patches were generated with a stride of 512 pixels and therefore do not spatially overlap, spatially adjacent patches may still be assigned to different subsets. Neighboring patches may share similar seabed morphology, local relief characteristics, and image appearance, which could make the present evaluation less challenging than a geographically blocked validation. Consequently, residual spatial dependence may lead to an optimistic estimate of model performance when transferred to spatially independent areas. Therefore, the reported results should be interpreted primarily as patch-level generalization within the same regional bathymetric domain rather than as fully spatially independent geographic generalization.

Fifth, the dataset used in this study was constructed from the northern Gulf of Mexico. Although this region contains abundant and diverse pockmark features, it still represents a specific seabed environment. Therefore, the external generalizability of FSLA-RTDETR has not yet been fully validated using independent datasets from other marine regions. Pockmark morphology, seabed sedimentary conditions, background relief patterns, and the visual appearance of shaded-relief bathymetric imagery may vary substantially among different geological settings. Variations in bathymetric spatial resolution may alter the apparent size, boundary sharpness, and local relief expression of pockmarks, while different sonar acquisition systems and processing workflows may introduce differences in spatial sampling, noise characteristics, artifacts, and image contrast. In addition, changes in hillshade illumination parameters may modify the apparent shading patterns and edge characteristics of the same seabed morphology. These domain shifts may potentially lead to reduced detection recall, decreased bounding-box localization accuracy, or increased false-positive responses to unfamiliar seabed structures. However, because independent cross-domain test data from other sea areas were not available in the present study, the magnitude of such performance degradation cannot yet be quantitatively determined. Future work will therefore focus on cross-region validation using bathymetric datasets from different geological environments, spatial resolutions, and acquisition systems, as well as transfer-learning, domain-adaptation, and domain-generalization strategies to improve the robustness of FSLA-RTDETR under domain shifts.

Future work should address these limitations from both the data and task-design perspectives. More bathymetric datasets from different regions should be introduced to evaluate the transferability of the model under different seabed conditions. Geographically blocked data partitioning and spatially independent test regions should also be adopted to provide a more rigorous assessment of spatial generalization and reduce the potential influence of spatial autocorrelation between training and evaluation samples. Additional training samples should also be collected for difficult cases, including densely clustered pockmarks, merged depressions, weak-boundary targets, boundary-located targets, and target-free background patches. These samples would help the model learn confusing patterns more effectively and reduce false detections. In addition, segmentation-based or multi-task frameworks could be explored to combine pockmark detection, boundary delineation, and geomorphometric attribute extraction. This would make the results more useful for geological interpretation beyond simple target localization.

6. Conclusions

This study proposed FSLA-RTDETR, an improved four-scale RT-DETR network for dense small-object seabed pockmark detection from shaded-relief bathymetric imagery.

The model introduces a SAC-C2f-enhanced backbone, an LPEA Encoder, a DGCST-based bidirectional feature fusion neck, and a four-scale decoder with the high-resolution P_2 feature. These components were designed to improve semantic representation, contextual modeling, multi-scale feature fusion, and small-object localization.

A pockmark detection dataset was constructed from BOEM bathymetric data in the northern Gulf of Mexico, containing 880 images and 8013 annotated pockmark instances. On this dataset, FSLA-RTDETR achieved mAP50, mAP75, and mAP50–95 values of 0.9218, 0.6221, and 0.5485, respectively. Compared with the baseline RT-DETR, the model improved mAP75 by 6.01 percentage points and mAP50–95 by 2.22 percentage points, showing better localization performance under stricter IoU thresholds. It should be noted that all comparative models in this study were trained from scratch without ImageNet-pretrained weights. Therefore, the reported comparative performance is specific to the present no-pretraining setting. Different detector architectures may benefit differently from pretrained initialization, particularly on a relatively small, single-class, geomorphology-oriented dataset, and the relative performance ranking may change when pretrained backbones are introduced.

The results indicate that FSLA-RTDETR is effective for detecting densely distributed and weakly bounded pockmark targets with small image-space footprints in complex seabed backgrounds. The method can provide candidate-level pockmark mapping results for large bathymetric datasets and may support subsequent analysis of pockmark distribution and seabed geomorphological patterns. However, further validation across different regions and data resolutions is still needed, and future work should extend the framework toward segmentation and geomorphometric attribute extraction.

Author Contributions: Conceptualization, C.W. and G.L.; methodology, C.W.; validation, S.L., W.R., and C.W.; formal analysis, C.W.; investigation, C.W.; resources, G.L.; writing—original draft preparation, C.W.; writing—review and editing, G.C. and C.W.; visualization, C.W.; supervision, R.L.; project administration, G.L. and X.Y.; funding acquisition, G.L. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Geological Survey Projects of the China Geological Survey (grant nos. DD202603102702, DD20230300603, and DD20221714).

Data Availability Statement: The source bathymetric data used in this study are publicly available from the U.S. Bureau of Ocean Energy Management (BOEM) through the "Northern GoA Deepwater Bathymetry Grid from 3D Seismic" product. The FSLA-RTDETR model configuration, core implementation code, software requirements, representative shaded-relief bathymetric image patches with corresponding annotations, and trained model weights are publicly available at https://github.com/Stride7/FSLA-RTDETR (accessed on 29 September 2026). The complete derived pockmark annotation dataset, precise patch-location information, and the complete training, validation, and test datasets form part of ongoing follow-up research and are not publicly available in full at the present stage.

Acknowledgments: We would like to acknowledge BOEM (Bureau of Ocean Energy Management) for providing the Northern Gulf of Mexico Deepwater Bathymetry Grid. This grid, derived from 3D seismic data, was used in this study.

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

References

1. Lecours, V.; Dolan, M.F.J.; Micallef, A.; Lucieer, V.L. A Review of Marine Geomorphometry, the Quantitative Study of the Seafloor. *Hydrol. Earth Syst. Sci.* **2016**, *20*, 3207–3244. [CrossRef]
2. Dimitrov, L.; Woodside, J. Deep Sea Pockmark Environments in the Eastern Mediterranean. *Mar. Geol.* **2003**, *195*, 263–276. [CrossRef]

3. Micallef, A.; Averes, T.; Hoffmann, J.; Crutchley, G.; Mountjoy, J.J.; Person, M.; Cohen, D.; Woelz, S.; Bury, S.J.; Ahaneku, C.V.; et al. Multiple Drivers and Controls of Pockmark Formation across the Canterbury Margin, New Zealand. *Basin Res.* **2022**, *34*, 1374–1399. [CrossRef]
4. Omeru, T.; Ayerume, D.; Bankole, S.; Ogbe, O.B.; Byami, J.A.; Obafemi, S.; Ifarajimi, W.; Pat-Nebe, O.C. Geomorphometric Analysis of Seabed Pockmarks, Offshore Western Niger Delta: A Case Study of the Freeman Field. *Interpretation* **2021**, *9*, T569–T584. [CrossRef]
5. Cathles, L.M.; Su, Z.; Chen, D. The Physics of Gas Chimney and Pockmark Formation, with Implications for Assessment of Seafloor Hazards and Gas Sequestration. *Mar. Pet. Geol.* **2010**, *27*, 82–91. [CrossRef]
6. Marcon, Y.; Ondréas, H.; Sahling, H.; Bohrmann, G.; Olu, K. Fluid Flow Regimes and Growth of a Giant Pockmark. *Geology* **2014**, *42*, 63–66. [CrossRef]
7. Mazzini, A.; Svensen, H.H.; Forsberg, C.F.; Linge, H.; Lauritzen, S.-E.; Haflidason, H.; Hammer, Ø.; Planke, S.; Tjelta, T.I. A Climatic Trigger for the Giant Troll Pockmark Field in the Northern North Sea. *Earth Planet. Sci. Lett.* **2017**, *464*, 24–34. [CrossRef]
8. Panieri, G.; Bünz, S.; Fornari, D.J.; Escartin, J.; Serov, P.; Jansson, P.; Torres, M.E.; Johnson, J.E.; Hong, W.; Sauer, S.; et al. An Integrated View of the Methane System in the Pockmarks at Vestnesa Ridge, 79°N. *Mar. Geol.* **2017**, *390*, 282–300. [CrossRef]
9. Lastras, G.; Canals, M.; Urgeles, R.; Hughes-Clarke, J.E.; Acosta, J. Shallow Slides and Pockmark Swarms in the Eivissa Channel, Western Mediterranean Sea. *Sedimentology* **2004**, *51*, 837–850. [CrossRef]
10. Pilcher, R.; Argent, J. Mega-Pockmarks and Linear Pockmark Trains on the West African Continental Margin. *Mar. Geol.* **2007**, *244*, 15–32. [CrossRef]
11. Giovannelli, D.; d'Errico, G.; Fiorentino, F.; Fattorini, D.; Regoli, F.; Angeletti, L.; Bakran-Petricioli, T.; Vetriani, C.; Yücel, M.; Taviani, M.; et al. Diversity and Distribution of Prokaryotes within a Shallow-Water Pockmark Field. *Front. Microbiol.* **2016**, *7*, 941. [CrossRef] [PubMed]
12. Lawal, M.A.; Cook, A.E. Gas Migration Phenomena and Pockmark Evolution in a Gulf of Mexico Minibasin. *Mar. Pet. Geol.* **2025**, *173*, 107270. [CrossRef]
13. Spatola, D.; Dahal, A.; Lombardo, L.; Casalbore, D.; Chiocci, F.L. First Pockmark Susceptibility Map of the Italian Continental Margins. *Mar. Pet. Geol.* **2025**, *176*, 107337. [CrossRef]
14. Cartwright, J.; Huuse, M. 3D Seismic Technology: The Geological 'Hubble'. *Basin Res.* **2005**, *17*, 1–20. [CrossRef]
15. Gaida, T.C.; Mohammadloo, T.H.; Snellen, M.; Simons, D.G. Mapping the Seabed and Shallow Subsurface with Multi-Frequency Multibeam Echosounders. *Remote Sens.* **2020**, *12*, 52. [CrossRef]
16. Posamentier, H.W.; Paumard, V.; Lang, S.C. Principles of Seismic Stratigraphy and Seismic Geomorphology I: Extracting Geologic Insights from Seismic Data. *Earth-Sci. Rev.* **2022**, *228*, 103963. [CrossRef]
17. Tasianas, A.; Bünz, S.; Bellwald, B.; Hammer, Ø.; Planke, S.; Lebedeva-Ivanova, N.; Krassakis, P. High-Resolution 3D Seismic Study of Pockmarks and Shallow Fluid Flow Systems at the Snøhvit Hydrocarbon Field in the SW Barents Sea. *Mar. Geol.* **2018**, *403*, 247–261. [CrossRef]
18. Cartwright, A.; Moss, J.; Cartwright, J. New Statistical Methods for Investigating Submarine Pockmarks. *Comput. Geosci.* **2011**, *37*, 1595–1601. [CrossRef]
19. Krämer, K.; Holler, P.; Herbst, G.; Bratek, A.; Ahmerkamp, S.; Neumann, A.; Bartholomä, A.; van Beusekom, J.E.E.; Holtappels, M.; Winter, C. Abrupt Emergence of a Large Pockmark Field in the German Bight, Southeastern North Sea. *Sci. Rep.* **2017**, *7*, 5150. [CrossRef] [PubMed]
20. Andrews, B.D.; Brothers, L.L.; Barnhardt, W.A. Automated Feature Extraction and Spatial Organization of Seafloor Pockmarks, Belfast Bay, Maine, USA. *Geomorphology* **2010**, *124*, 55–64. [CrossRef]
21. Gafeira, J.; Long, D.; Diaz-Doce, D. Semi-automated Characterisation of Seabed Pockmarks in the Central North Sea. *Near Surf. Geophys.* **2012**, *10*, 301–312. [CrossRef]
22. Gafeira, J.; Dolan, M.F.J.; Monteys, X. Geomorphometric Characterization of Pockmarks by Using a GIS-Based Semi-Automated Toolbox. *Geosciences* **2018**, *8*, 154. [CrossRef]
23. Picard, K.; Radke, L.C.; Williams, D.K.; Nicholas, W.A.; Siwabessy, P.J.; Howard, F.J.F.; Gafeira, J.; Przeslawski, R.; Huang, Z.; Nichol, S. Origin of High Density Seabed Pockmark Fields and Their Use in Inferring Bottom Currents. *Geosciences* **2018**, *8*, 195. [CrossRef]
24. Alonso del Rosario, J.J.; Canari, A.; Blázquez Gómez, E.; Martínez-Loriente, S. POSIT: An Automated Tool for Detecting and Characterizing Diverse Morphological Features in Raster Data—Application to Pockmarks, Mounds, and Craters. *Appl. Comput. Geosci.* **2024**, *23*, 100190. [CrossRef]
25. Giannakopoulos, V.; Feldens, P.; Fakiris, E. Semi-Automated Mapping of Pockmarks from MBES Data Using Geomorphometry and Machine Learning-Driven Optimization. *Remote Sens.* **2025**, *17*, 2917. [CrossRef]
26. Lundine, M.A.; Brothers, L.L.; Trembanis, A.C. Deep Learning for Pockmark Detection: Implications for Quantitative Seafloor Characterization. *Geomorphology* **2023**, *421*, 108524. [CrossRef]

27. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
29. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
30. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 5967–5976.
31. Gong, W. Lightweight Object Detection: A Study Based on YOLOv7 Integrated with ShuffleNetv2 and Vision Transformer. *arXiv* **2024**, arXiv:2403.01736.
32. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 16965–16974.
33. Lawal, M.A.; Cook, A.E.; Portnov, A.; Kumar, A. Pockmarks, Mud Volcanoes and Hydrocarbon Seeps in the Northern Gulf of Mexico: Trends and Controls on Widespread Fluid and Gas Venting. *Basin Res.* **2026**, *38*, e70095. [CrossRef]
34. Kumar, A.; Cook, A.E.; Lawal, M.A.; Portnov, A.; Lecours, V. Pockmark Occurrence in the Northern Gulf of Mexico Influenced by Glacial Cycles and Hydrate Stability. *Geochem. Geophys. Geosyst.* **2025**, *26*, e2024GC011781. [CrossRef]
35. Roelofse, C.; Alves, T.M.; Gafeira, J. Structural Controls on Shallow Fluid Flow and Associated Pockmark Fields in the East Breaks Area, Northern Gulf of Mexico. *Mar. Pet. Geol.* **2020**, *112*, 104074. [CrossRef]
36. Kumar, A.; Cook, A.E.; Portnov, A.; Skopec, S.; Frye, M.; Palmes, S. Bottom Simulating Reflections across the Northern Gulf of Mexico Slope. *Mar. Pet. Geol.* **2024**, *165*, 106870. [CrossRef]
37. Moss, J.L.; Cartwright, J.; Cartwright, A.; Moore, R. The Spatial Pattern and Drainage Cell Characteristics of a Pockmark Field, Nile Deep Sea Fan. *Mar. Pet. Geol.* **2012**, *35*, 321–336. [CrossRef]
38. Kramer, K.V.; Shedd, W.W. A 1.4-Billion-Pixel Map of the Gulf of Mexico Seafloor. *Eos*. 2017. Available online: https://eos.org/science-updates/a-1-4-billion-pixel-map-of-the-gulf-of-mexico-seafloor (accessed on 29 September 2026). [CrossRef]
39. Shi, D. TransNeXt: Robust Foveal Visual Perception for Vision Transformers 2024. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024.
40. Zheng, F.; Chen, X.; Liu, W.; Li, H.; Lei, Y.; He, J.; Pun, C.-M.; Zhou, S. SMAFormer: Synergistic Multi-Attention Transformer for Medical Image Segmentation. In *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024*; IEEE: New York, NY, USA, 2024; pp. 4048–4053.
41. Fan, Q.; Huang, H.; Ai, Y.; He, R. Rectifying Magnitude Neglect in Linear Attention. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–25 October 2025*; IEEE: New York, NY, USA, 2025; pp. 21505–21514.
42. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
43. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
44. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection. *arXiv* **2025**, arXiv:2509.25164.
45. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498. [CrossRef] [PubMed]
46. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
47. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.

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.