Petroleum Science 23 (2026) 2686–2697

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

MSA-DETR: Multi-scale attention enhanced DETR for object detection in oilfield surveillance

Qian-Wen Cao [a,b], Jin-Rong Ma [a], Lai-Bin Zhang [a,b,*]

[a] China University of Petroleum (Beijing), Beijing, 102249, China
[b] Key Laboratory of Oil and Gas Production Equipment Quality Inspection and Health Diagnosis, State Administration for Market Regulation, Beijing, 102249, China

ARTICLE INFO

Article history:
Received 14 October 2025
Received in revised form
18 March 2026
Accepted 22 March 2026
Available online 25 March 2026

Edited by Jia-Jia Fei

Keywords:
Object detection
Petroleum drilling safety
Multi-scale perception
Drilling engineering
Artificial intelligence

ABSTRACT

In modern petroleum engineering, ensuring operational safety at drilling sites is of critical importance. Visual object detection plays a key role in intelligent safety monitoring systems by enabling real-time supervision of personnel and equipment. However, safety-critical targets in drilling scenes are often small, partially occluded, and embedded in cluttered environments, leading to decreased detection accuracy and potential safety risks. Existing convolutional neural networks (CNN)-based detectors, although effective in natural scenes, often exhibit limited robustness under such complex industrial conditions. To address these challenges, this paper proposes MSA-DETR, a Transformer-based detection framework designed to enhance multi-scale perception in drilling monitoring scenarios. By improving the ability to capture both global contextual information and fine-grained visual cues, the proposed approach enhances sensitivity to safety-relevant objects. Extensive experiments conducted on two real-world drilling monitoring datasets demonstrate that MSA-DETR consistently outperforms state-of-the-art detection methods, providing more reliable visual perception for petroleum safety management and accident prevention.

1. Introduction

Ensuring personnel and equipment safety during drilling and production operations is a fundamental requirement in modern petroleum engineering. Drilling sites are typically characterized by high operational intensity, complex workflows, and harsh working conditions, where even minor mistakes may lead to severe accidents, economic losses, or environmental hazards. To improve safety management efficiency, surveillance video systems have been widely deployed in drilling sites and industrial parks, providing continuous visual monitoring of personnel activities and equipment status (Neto et al., 2024). However, the massive amount of video data generated in practice makes manual inspection inefficient, subjective, and unreliable (Ahmed et al., 2023). As a result, automated visual analysis has become an indispensable component of intelligent petroleum safety management systems.

Object detection serves as a core technique in automated visual analysis and has achieved remarkable progress in many application domains (Zhu et al., 2020a). In drilling surveillance scenarios, object detection aims to identify personnel, equipment, and potentially hazardous behaviors from video streams to support risk assessment and accident prevention (Gong et al., 2021). As illustrated in Fig. 1, surveillance cameras are often installed at elevated or distant positions, leading to significant scale variation (Wang et al., 2020). In addition, complex backgrounds, illumination changes, and frequent occlusions further increase detection difficulty. These factors make drilling surveillance a representative and challenging scenario for multi-scale object detection, where targets of different sizes coexist and require unified modeling. From a visual understanding perspective, industrial object detection faces challenges at multiple semantic levels. At the macro level, models must capture global scene structure and suppress background clutter, while at the micro level, they must precisely localize fine-grained and scale-sensitive regions (Iqbal et al., 2024; Peng et al., 2024). The coexistence of large, medium, and small

* Corresponding author.
E-mail address: zhanglb@cup.edu.cn (L.-B. Zhang).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.03.045

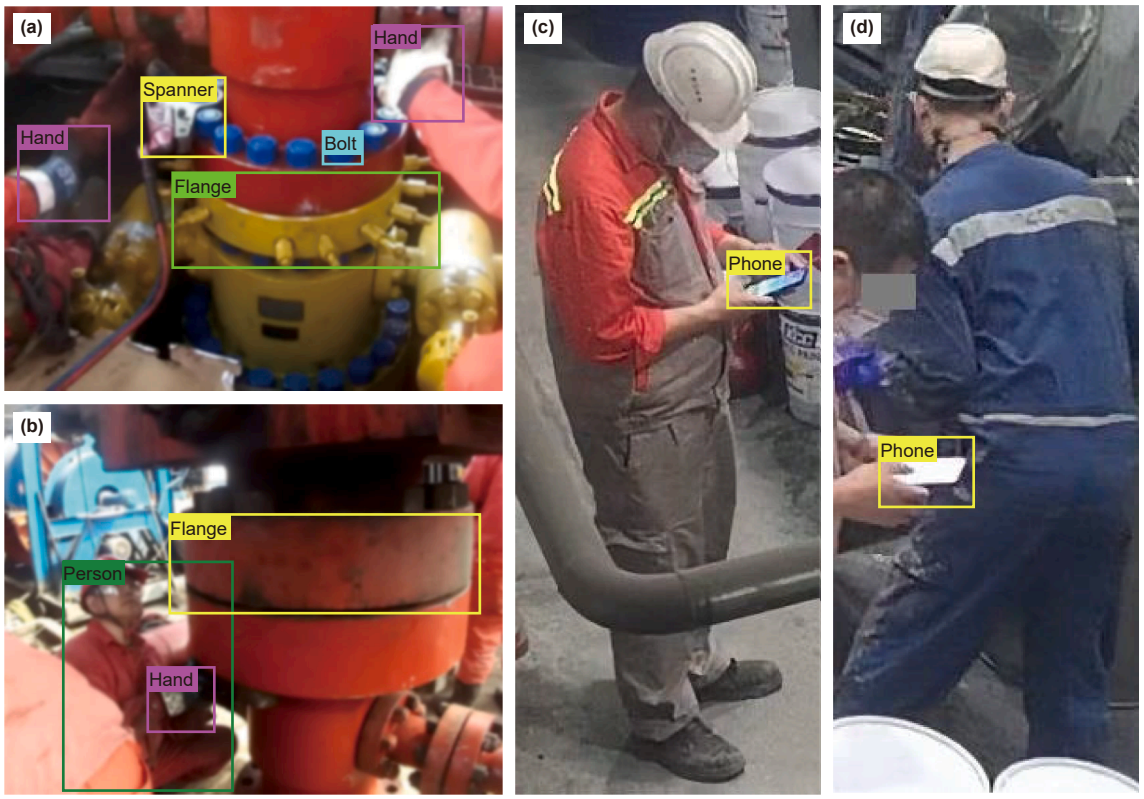

Fig. 1. Sample images extracted from oil and gas operation monitoring videos. Images (**a**) and (**b**) show workers operating objects of different scales in drilling sites, whereas images (**c**) and (**d**) capture unauthorized mobile phone use.

objects introduces severe scale imbalance, making effective multi-scale feature representation and cross-scale interaction essential for robust detection in drilling environments (Wang et al., 2025).

CNN-based detectors, such as Faster R-CNN (Ren et al., 2015), SSD (Liu et al., 2016), and the YOLO (Redmon et al., 2016) family, have been widely adopted due to their efficiency and strong performance in natural scenes. These methods typically rely on fixed receptive fields and feature pyramid structures to approximate multi-scale representations (Rijayanti et al., 2023). However, in complex drilling scenes, such strategies often struggle to achieve a balanced representation across heterogeneous object scales (Tao et al., 2024). Deep down-sampling operations tend to suppress fine-grained features, while shallow features lack sufficient semantic context for large objects (Miri Rekavandi et al., 2025). As a result, CNN-based detectors face inherent limitations in jointly modeling global semantics and local details under severe scale variation, which restricts their robustness in industrial surveillance applications.

Transformer-based detectors have recently emerged as a promising alternative by introducing global attention mechanisms into object detection. Detection transformer (DETR) and its variants enable end-to-end detection and long-range dependency modeling without relying on handcrafted anchor designs (Carion et al., 2020). Nevertheless, standard DETR encodes only the final-stage backbone features, which limits its ability to preserve high-resolution spatial information critical for multi-scale perception (Huang and Li, 2024, Miri Rekavandi et al., 2025). Subsequent approaches, such as deformable DETR (Zhu et al., 2020b), conditional DETR (Meng et al., 2021), and lite-DETR (Li et al., 2023), improve convergence and efficiency through sparse attention and query refinement. Despite these advances, existing Transformer-based detectors still exhibit limitations in explicitly modeling hierarchical multi-scale interactions and adaptively balancing attention across different feature resolutions. In industrial scenarios with severe scale imbalance, such deficiencies may lead to diluted attention responses and unstable localization across object scales.

To address the above challenges, we propose MSA-DETR (multi-scale attention enhanced DETR), a Transformer-based detection framework tailored for complex drilling surveillance environments. The proposed approach enhances multi-scale representation capability by incorporating attention-based feature interaction across different resolutions, enabling more effective integration of global semantics and local details. By explicitly modeling cross-scale relationships in both the encoder and decoder, MSA-DETR improves detection robustness under extreme scale variation and complex background interference. The main contributions of this work are summarized as follows.

(1) We design a multi-scale object detection framework tailored for drilling scenarios, which aggregates and compresses attention information across multiple feature layers. This design makes our model capture objects of varying scales under complex visual conditions.
(2) We construct a multi-scale attention fusion model, in which features at different resolutions are extracted from the backbone and passed into the encoder for hierarchical attention interaction. The attention representations from high-level and low-level features are updated in an alternated manner, enabling more effective attention-based feature aggregation across scales.
(3) We adopt a downsampling-based feature fusion method to expand the receptive field for sensitive targets. This strategy is particularly effective in detecting small-scale objects and

contributes to the end-to-end modeling and optimization of the detection process under challenging drilling environments.

2. Related work

2.1. Object detection

In recent years, deep learning-based object detection methods (Cao et al., 2025) have achieved remarkable progress and can be broadly categorized into two classes: CNN-based architectures, such as Faster R-CNN (Ren et al., 2015) and YOLO (Redmon et al., 2016), and the more recent Transformer-based architectures. These two paradigms differ significantly in their mechanisms for feature extraction and spatial modeling, and each faces distinct challenges. Early object detection algorithms, represented by Faster R-CNN (Ren et al., 2015), adopt a two-stage framework that combines region proposal mechanisms with classification and regression networks. This approach offers high detection accuracy, particularly for prominent objects in high-resolution images. However, its relatively slow inference speed and complex training pipeline limit its applicability in real-time scenarios. To address this, single-stage detectors such as the YOLO series were introduced. From YOLOv1 (Redmon et al., 2016) to YOLOv13 (Ramos and Sappa, 2025), continuous iterations have transformed object localization and classification into a unified regression task, significantly improving detection efficiency. These models have demonstrated strong performance in real-world applications such as traffic surveillance and face recognition. Nevertheless, the YOLO series still faces challenges in detecting small or densely packed objects (Nikouei et al., 2025), especially under occlusion or in cluttered backgrounds, where false positives or missed detections are more likely to occur. In contrast to these approaches, our method achieves more efficient and robust detection of multi-scale objects and targets affected by environmental factors.

With the rapid success of Transformers in natural language processing (NLP), Vision Transformers have also been introduced into object detection tasks (Shehzadi et al., 2025). DETR (Carion et al., 2020) pioneered end-to-end Transformer-based detection by removing the region proposal network (RPN) (Ren et al., 2015), greatly simplifying the pipeline. Equipped with global self-attention, DETR models long-range dependencies effectively but suffers from slow convergence and suboptimal performance on small or fine-grained targets.

To mitigate these limitations, numerous DETR variants have emerged, many of which explicitly incorporate multi-scale information. Deformable DETR (Zhu et al., 2020b) adopts sparse deformable attention for efficient multi-scale sampling; Conditional DETR (Meng et al., 2021) accelerates convergence by refining the query update process; and lite DETR (Li et al., 2023) uses a lightweight design for real-time applications. Other representative extensions include SMCA-DETR (Gao et al., 2021), which guides queries toward specific scales via multi-scale attention maps; DAB-DETR (Liu et al., 2022), which parameterizes queries as dynamic anchor boxes to enhance localization; DN-DETR (Li et al., 2022), which uses denoising queries to reduce training cost; and anchor DETR (Wang et al., 2022), which initializes queries with anchor points for improved alignment.

Despite these advances, most prior efforts focus on sampling sparsity, query initialization, or training efficiency, while paying less attention to explicitly encoding hierarchical multi-scale cues in the encoder and propagating scale-aware semantics to the decoder. In contrast, we introduce a multi-scale attention-pooling encoder that fuses features across resolutions through structured attention pooling rather than relying on sparse sampling or decoder-driven scale weighting. This preserves fine-grained textures crucial for detecting small objects in drilling environments. Moreover, our scale-aware decoder incorporates explicit scale embeddings to guide cross-attention. Unlike conditional DETR (Meng et al., 2021) or DAB-DETR (Liu et al., 2022), which modify query formulations without modeling intrinsic scale semantics, our decoder leverages scale-aware representations to improve localization accuracy and robustness under complex drilling scenarios. In contrast, we introduce a multi-scale attention-pooling encoder that explicitly fuses features across different resolutions through structured hierarchical attention. This design enables effective interaction between high-level semantic features and low-level spatial details, forming scale-consistent representations prior to decoding. Unlike conditional DETR (Meng et al., 2021) or DAB-DETR (Liu et al., 2022), which focus on modifying query formulations, the proposed approach enhances localization robustness and stability across heterogeneous object scales in complex drilling scenarios.

2.2. Petroleum surveillance analysis

With the rapid development of Industry 4.0 and intelligent manufacturing, industrial monitoring systems are evolving from traditional passive recording to active perception, intelligent warning, and real-time decision-making (Serror et al., 2021). Against this backdrop, intelligent monitoring technologies based on AI, CV, and the internet of things (IoT) (Gubbi et al., 2013) are becoming key enablers for ensuring industrial safety. Conventional industrial video surveillance systems primarily rely on manual observation, which often results in delayed responses and high false alarm rates. To overcome these limitations, researchers have proposed intelligent video surveillance systems powered by deep learning and computer vision, enabling real-time analysis and early warning of human behaviors, equipment states, and environmental changes. For example, a research team (Zhang et al., 2023) proposed an approach that integrates an adaptive recursive path aggregation network (AR-PANet) with YOLOv4 to enable accurate and efficient detection of small tools and equipment in tunnel construction monitoring under low artificial lighting conditions. Lyu et al. proposed a feature fusion framework for single and multiple object detection in industrial automation (Lyu et al., 2024). However, despite such advances, intelligent surveillance systems still face several challenges in real-world industrial applications, including object recognition in complex backgrounds, occlusion handling, and multi-scale object detection.

In the petroleum industry, surveillance analysis has also become a crucial component for ensuring production safety, environmental protection, and operational efficiency. Recent studies have focused on developing vision-based monitoring systems tailored for oil and gas environments, where extreme lighting, sea spray, and distant camera angles make object detection particularly challenging. For instance, a research team developed a combinatorial reasoning-based abnormal sensor recognition method for subsea production control systems in offshore oil and gas platforms, which improves fault identification in sensor data under harsh operational conditions (Zhang et al., 2024). Similarly, Zhang et al. conducted modeling and field investigations of a catastrophic oil spill and vapor-cloud explosion caused by pipeline leakage in a confined space, highlighting the severity of risks associated with equipment failures and the importance of timely detection (Zhang et al., 2020). To further address small-target and remote-area monitoring, Wang et al. proposed an optimized Faster R-CNN model for oil well detection from high-resolution remote sensing images, significantly improving localization accuracy of

small wellheads and valves (Wang et al., 2023b). Beyond safety monitoring, petroleum surveillance also extends to environmental anomaly detection. Wang et al. designed a cyber-physical framework for oil spill detection and risk management that integrates optical and thermal data to enable early warning of leakage events (Wang et al., 2023a). Zhan et al. further combined hyperspectral imaging with CNN and DBSCAN clustering to achieve high-precision offshore oil-spill identification (Zhan et al., 2024). These advances collectively highlight that petroleum surveillance analysis requires robust multi-scale perception, domain-specific data augmentation, and real-time adaptability to handle diverse operational scenarios.

In industrial environments, accurately identifying human behaviors is critical for preventing safety incidents. Reddy et al. proposed an intelligent monitoring system that demonstrated clear performance advantages over conventional methods in industrial safety scenarios (Reddy et al., 2024). The field of industrial monitoring and analysis is rapidly evolving toward greater intelligence, integration, and real-time responsiveness. Despite significant progress, challenges remain in multi-scale object detection, behavior recognition under complex environments, and real-time system responsiveness. To address these issues, this paper proposes MSA-DETR, a multi-scale attention-based DETR model designed to enhance object detection performance in complex drilling scenarios. By leveraging a multi-scale attention mechanism, the model provides more effective technical support for ensuring drilling safety.

3. Methodology

In this section, we begin by introducing the overall framework of our proposed method. Subsequently, we delve into the architectural details of MSA-DETR, emphasizing the design of the multi-scale attention pooling encoder and the multi-scale attention enhanced visual decoder, which are central to achieving robust detection in complex drilling environments.

3.1. Solution overview

Inspired by the recent success of DETR-based models in object detection tasks (Shehzadi et al., 2025), we propose the MSA-DETR framework for object detection in drilling surveillance scenarios. As illustrated in Fig. 2, the model first processes the extracted image features into a sequence of tokens. Specifically, drilling surveillance videos are first divided into frames, which are then fed into a backbone network to obtain feature maps. These feature maps are subsequently flattened and concatenated to form the input token sequence. The token sequence is then passed through two key components: the multi-scale attention pooling encoder and the multi-scale attention enhanced visual decoder. The encoder is designed to guide the model to focus on multi-scale and key targets. To this end, we introduce a hierarchical attention interaction mechanism and an enhanced attention pooling module to effectively retain informative features. In the decoder, we propose a scale-aware weighting mechanism to efficiently receive the context-rich feature sequences, enabling precise object localization and bounding box regression.

3.2. Multi-scale attention pooling encoder

3.2.1. Multi-scale feature extraction and token alignment

Effective feature extraction is fundamental to model performance. In the original DETR, features are extracted solely from the final layer of the backbone, which contains highly abstract information. However, this is insufficient for accurately capturing key targets, especially those that are small, occluded, or susceptible to background interference.

To address this issue, we adopt ResNet-50 (Pang et al., 2021) as the backbone to extract multi-scale feature maps. In order to uniformly process features at different scales, each feature map is first flattened into a sequence of tokens. We then append a hierarchical embedding to each group of tokens to indicate their corresponding feature level. Finally, the token groups are concatenated to form a unified multi-scale feature sequence $S = [S_1; S_2; S_3; S_4]$ which is fed into the Transformer encoder as input,

$$S_i = Flatten(Conv(F_i, c = d_{\text{model}})) \quad (1)$$

where F_i denotes the i-th layer feature map of the backbone network, $Conv$ represents the channel adjustment convolution, d_{model} is the unified feature dimension, and $Flatten$ is the flatten operation, by which multi-scale features form a dimension-consistent sequence.

3.2.2. Hierarchical alternated self-attention interaction

The multi-scale feature maps extracted from the backbone contain abundant information; however, lower-level feature maps have larger spatial resolutions, which tend to introduce excessive

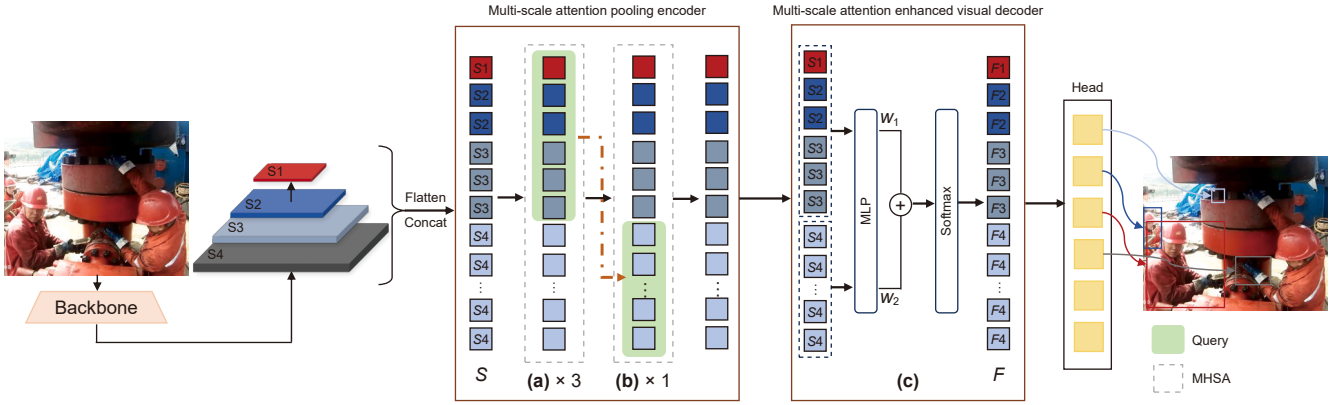

Fig. 2. The framework of MSA-DETR uses S_2–S_4 to denote feature layers from the backbone, which are taken from the 5th to 3rd residual stages in ResNet-50, representing progressively deeper and more semantic features. S_1 is obtained by downsampling S_2 by a factor of 2. We consider S_1–S_3 as high-level feature layers, and S_4 as a low-level one. In (**a**) and (**b**), the enhanced attention pooling module is used to iteratively update multi-scale features with high-level cues, followed by low-level refinement. (**c**) Shows the multi-scale attention enhanced visual decoder, which performs weighted fusion across all scales to generate the enhanced feature F.

background information and may overwhelm meaningful signals. To address the representational gap and redundancy across multi-scale features, we propose a hierarchical alternated attention update strategy. This strategy leverages the abstraction capability of high-level features while retaining fine-grained details from lower levels, which is crucial for preserving effective information.

In the encoder, the fused multi-scale token sequence S is decomposed into high-level features F_L and low-level features F_H. Specifically, we define $F_L = [S_1; S_2; S_3]$ as high-level feature tokens (more abstract and semantically abundant), $F_H = S_4$ as the lowest-level feature tokens (with minimal downsampling).

High-level features are intended to capture semantically sensitive regions, while low-level features retain local spatial detail often lost in deeper layers. Notably, low-level features have a larger number of tokens due to their higher resolution, but they also carry more background noise. In drilling monitoring scenarios, such redundant background information can significantly affect global self-attention, reducing the focus on critical regions.

To mitigate this issue, we propose a hierarchical alternated attention mechanism, where tokens from different scales interact sequentially. Specifically, we perform attention updates in an alternating manner: high-level features and low-level features are used as queries in turn, attending to the full multi-scale key/value set.

High-level queries are updated more frequently to enhance object discrimination, while low-level queries are updated more sparsely to suppress background noise. This mechanism is illustrated in Fig. 2. Formally, the update process can be described as:

$$Q = F_H, K = V = Concat(F\_H, F\_L)$$

$$F_H{}' = MHSA(Q, K, V) \tag{2}$$

where *Concat* denotes the concatenation of low-level and high-level features into a full-scale feature set. The query Q corresponds to the initial high-level features, while the keys and values K, V are derived from the initial features across all scales. F_H represents the high-level tokens, and F'_H denotes the updated high-level features after interaction. MHSA stands for multi-head self-attention.

$$Q = F_L, K = V = Concat(F'_H, F_L)$$

$$F'_L = MHSA(Q, K, V)$$

$$Output_{enc} = Concat(F'_L, F'_H) \tag{3}$$

where F'_H denotes the updated high-level features after the current encoder layer; the query is set as the initial low-level features. F_L represents the low-level tokens, and F'_L denotes their updated representations, $Output_{enc}$ means the training result of this round of encoder.

3.2.3. Enhanced attention pooling module

On the other hand, drilling monitoring scenarios often involve numerous fine-grained yet critical targets that exhibit weak responses in complex backgrounds and are prone to being overlooked by self-attention mechanisms. To enhance the sensitivity to such targets, we introduce an improved attention pooling module into the self-attention interaction process. This design is motivated by the idea of expanding the representational capacity and sensitivity range for small-scale salient regions.

As illustrated in Fig. 3, we perform downsampling on the key and value features to enlarge the influence area of high-sensitivity regions, while keeping the query resolution unchanged. This

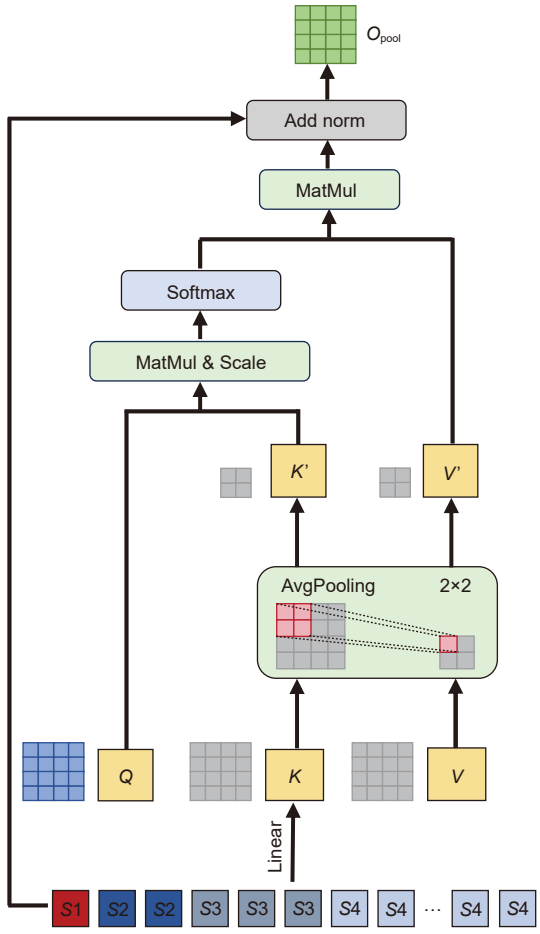

Fig. 3. The detailed architecture of proposed enhanced attention pooling module. Based on the standard self-attention mechanism, multi-scale features are input, and avgpooling is applied to the K (key) and V (value) to generate K' and V' to enhance the response capability of high-saliency regions. The original Q (query) is retained and participates in the attention computation together with the scale-compressed K/ V, which improves the perception robustness for small objects and fine-grained information, while effectively reducing background noise interference.

ensures that the expressive capacity of each pixel in the query is preserved as much as possible. To formally define this process, let $X \in R^{H\times W\times C}$ denote the input feature map from a specific scale, where H, W are the spatial dimensions and C is the channel dimension (equivalent to d_{model}). The enhanced attention pooling applies an average pooling operation $\mathcal{P}_{\dashv\sqsubset\}}$ with kernel size k and stride s. The downsampled feature map X' is calculated as:

$$X' = \mathcal{P}_{\dashv\sqsubset\}}(X), X' \in R^{H'\times W'\times C} \tag{4}$$

The spatial dimensions H' and W' are updated as follows:

$$H' = \left\lfloor \frac{H-k}{s} + 1 \right\rfloor, W' = \left\lfloor \frac{W-k}{s} + 1 \right\rfloor \tag{5}$$

Subsequently, the feature maps are flattened into token sequences. Let $L = H \times W$ be the sequence length of the original query, and $L' = H' \times W'$ be the reduced sequence length for keys and values. The flattening operation is expressed as:

$$Q = Flatten(X) \in R^{L\times C}, K', V' = Flatten(X') \in R^{L'\times C} \tag{6}$$

Here, K' encodes pooled contextual representations for scale-aware attention matching, while V' preserves the aggregated

feature content within each pooled region and serves as the information carrier for cross-scale contextual propagation.

These expanded receptive fields are crucial for detecting small-scale objects, while for large-scale targets, the pooled regions occupy only a minimal proportion and thus have negligible impact on detection accuracy. Based on the dimensionality reduction defined above, we introduce the scale-aware attention computation. Unlike standard self-attention where queries, keys, and values share the same length, our module computes the interaction between the high-resolution query Q and the context-aggregated key K'. The attention matrix $Attention_{\text{pool}}$ is derived as:

$$Attention_{\text{pool}} = softmax\left(\frac{Q(K'W_K)^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right) \tag{7}$$

where, $Q(K'W_K)^{\mathrm{T}}$ results in a matrix of shape $\mathbb{R}^{L\times L'}$, representing the correlation between each fine-grained pixel and the expanded receptive regions. The final output O_{pool} is obtained by aggregating the pooled values:

$$O_{\text{pool}} = Attention_{\text{pool}}(V'W_V) \tag{8}$$

This formulation explicitly reduces the computational complexity of the key/value projection from $\mathcal{O}\left(L^2\right)$ to $\mathcal{O}(L\cdot L')$, while expanding the receptive field for small object detection

3.3. Multi-scale attention enhanced visual decoder

In the decoder stage, MSA-DETR faces the challenge of accurately restoring and localizing target objects. To efficiently receive multi-scale attention information from the Transformer encoder, we introduce a scale-aware weighting mechanism that enables adaptive focus on object scales. By assigning different attention weights to regions of different scales, the model can attend to targets of various sizes more effectively, thereby significantly improving detection accuracy and robustness.

Specifically, we divide the fused feature maps into two subgroups based on spatial resolution: a large-scale group F_{large}, representing low-resolution and high-level semantic features, and a small-scale group F_{small}, corresponding to high-resolution, fine-grained features extracted from lower encoder layers.

To achieve dynamic scale-wise weighting, we design a MLP that generates the corresponding attention weights w_1 and w_2 for F_{large} and F_{small} respectively. The detailed design is described as follows:

$$w_1 = MLP\left(F_{\text{large}}\right), w_2 = MLP(F_{\text{small}})$$

$$[w'_1, w'_2] = softmax([w_1, w_2])$$

$$\text{s.t.} \quad w'_1 + w'_2 = 1 \tag{9}$$

where the MLP takes F_{large} and F_{small} as inputs and produces a pair of weights, w'_1 and w'_2, which are then normalized via a softmax function to ensure that their sum equals 1.

Each sub-feature map is then scaled by its corresponding weight and aggregated to form the fused feature F:

$$F = w'_1 \cdot F_{\text{large}} + w'_2 \cdot F_{\text{small}} \tag{10}$$

This dynamic weighting process enables the model to adaptively emphasize scale-relevant information while suppressing irrelevant background noise, based on the current scene context and decoder iteration status.

During self-attention interaction, the decoder employs the enhanced fused feature F as both the key and value, engaging in hierarchical interactions with the query:

$$Attention(Q_t, F) = softmax\left(\frac{Q_t W^{Q}\left(FW^{K}\right)^{\mathrm{T}}}{\sqrt{d}}\right)\left(FW^{V}\right) \tag{11}$$

where Q_t represents the query embeddings at the t-th decoding layer, and d is the feature dimension. $W^i(i = Q, K, V)$ are the weighted fusion weights for Q, K, and V.

At each decoding stage, hierarchical attention and attention pooling are performed over the fused multi-scale feature map F. The output is further refined by a feed-forward network (FFN), which updates the predicted bounding box coordinates. This iterative refinement allows the decoder to progressively converge toward accurate object locations:

$$\Delta b_t = FFN(Attention(Q\_t, F))$$

$$b_t = b_{t-1} + \Delta b_t \tag{12}$$

where b_t denotes the position of the predicted bounding box from the t-th layer Decoder, and Δb_t denotes its offset predicted by the same layer.

4. Experiment

In this section, we evaluate the performance of the proposed object detection model and method on two typical drilling surveillance scenarios, and compare it with several baseline learning algorithms.

4.1. Experiment setup

4.1.1. Data collection

We conduct evaluation experiments based on two real drilling surveillance scenario datasets. Fig. 4 shows the distribution characteristics of object categories and scales in the two datasets. The scale is divided according to the COCO standard format (Lin et al., 2014) into small (area $\leq 32^2$), medium ($32^2 \leq$ area $\leq 96^2$), and large (area $\geq 96^2$).

WSM-phone (well site monitoring-phone): The improper use of safety equipment can distract workers and increase operational safety risks. To address this issue, a dataset collected to identify unauthorized mobile phone usage by personnel within drilling parks, aiming to enhance production safety management. The data are collected from real-world surveillance videos, comprising a total of 19,000 images, with 16,000 for training and 3,000 for testing. The dataset includes only two object categories, with a roughly equal number of bounding boxes per class. As shown in Fig. 4, small- and medium-scale targets are predominant, while large-scale targets are relatively scarce. Mobile phones, being fine-grained and thin-shaped objects, are highly affected by viewpoint variations. In addition, the presence of low lighting and cluttered backgrounds in the environment increases the difficulty of detection.

WSM-safety (well site monitoring–device): The safety of personnel operating equipment also represents a critical potential risk in well sites. A dataset focuses on equipment installation monitoring in drilling environments, covering the identification of personnel and various drilling components. It contains 1,600 images with a total of 23,000 annotated bounding boxes, including 1,200 images for training and 400 for testing. The scene involves a

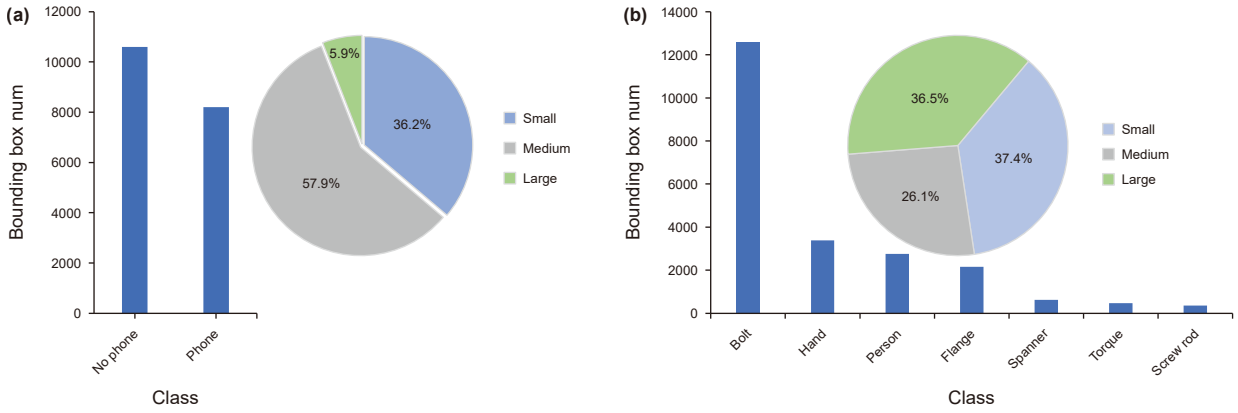

Fig. 4. Category and object size distribution characteristics of the two scenario datasets. (**a**) WSM-phone. (**b**) WSM-safety. The bar chart shows the number of objects of each category in the dataset, and the pie chart shows the proportion of objects at each scale.

wide variety of detection targets, with significant imbalance in category frequency and a large range of object sizes. The dataset requires the detection of both small components and large machinery, with a relatively balanced distribution across different scales. Due to the high target density and complex scenes, this dataset is especially suitable for research on multi-class object detection, large-scale variation modeling, and safety hazard identification.

4.1.2. Implement details

The proposed model is implemented in PyTorch 2.4.1 with Python 3.8.20. Experiments are conducted on an Ubuntu system equipped with an NVIDIA A800 GPU. Our MSA-DETR is developed based on the deformable DETR framework (Zhu et al., 2020b), with ResNet-50 (He et al., 2016) serving as the backbone, initialized using ImageNet-1K (Deng et al., 2009) pretrained weights. The number of attention heads is set to 8, and 300 object queries are employed. The model is trained using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Initial learning rates are set to 1×10^{-5} for the backbone and 1×10^{-4} for the Transformer, with a batch size of 2. Following the standard training strategy of DETR-based methods, the learning rate is decayed by a factor of 0.1 at epoch 30.

During training, we adopt standard data augmentation strategies following the settings of DETR (Carion et al., 2020) and deformable DETR (Zhu et al., 2020b), including random horizontal flipping, multi-scale image resizing, and random size cropping, followed by image normalization.

4.2. Comparison with SOTA

4.2.1. Qualitative analysis

We compare the proposed MSA-DETR with a broad range of representative object detection methods. These include CNN-based detectors such as Faster R-CNN (Ren et al., 2015) and YOLOv8 (Varghese and Sambath, 2024), as well as end-to-end Transformer-based detectors including DETR (Carion et al., 2020), deformable DETR (Zhu et al., 2020b), DN-DETR (Li et al., 2022), and lite-DETR (Li et al., 2023). The quantitative comparison on the WSM-phone and WSM-safety datasets is summarized in Table 1.

From an overall accuracy perspective, MSA-DETR achieves the highest mAP on both datasets. On WSM-phone, it reaches 68.6 on mAP after 50 training epochs, outperforming Faster R-CNN and YOLOv8 trained for 100 epochs and DETR trained for 300 epochs. On the more challenging WSM-safety dataset, MSA-DETR achieves 56.1 on mAP under the same training schedule, indicating faster convergence and more stable optimization compared with standard DETR-based baselines.

In terms of localization quality, MSA-DETR also shows clear advantages. It achieves 72.4 on AP_{75} on WSM-phone and 60.1 on AP_{75} on WSM-safety, demonstrating its ability to generate more precise and well-aligned bounding boxes. Compared with DETR and deformable DETR, the proposed model exhibits noticeably improved localization stability in complex industrial scenes with cluttered backgrounds and scale variation.

The performance gain is particularly significant for small objects. On WSM-phone, MSA-DETR achieves 52.4 on AP_S, exceeding deformable DETR by more than 12 points and YOLOv8 by over 6 points. On WSM-safety, the AP_S further increases to 51.4, outperforming all competing methods. These results indicate that the proposed multi-scale attention pooling encoder and scale-aware decoder effectively enhance sensitivity to fine-grained targets, which are critical in drilling surveillance scenarios.

Beyond detection accuracy, we further analyze model complexity and inference efficiency. MSA-DETR contains 48 M parameters and requires 138 GFLOPs per image, which is lower than most Transformer-based competitors with comparable accuracy. Despite incorporating explicit multi-scale attention mechanisms, the model achieves an inference speed of 25 FPS. Compared with DN-DETR and lite-DETR, which focus on accelerating training or reducing encoder complexity, MSA-DETR provides a more balanced trade-off between accuracy, robustness, and runtime efficiency. This balance makes it well suited for real-time or near real-time industrial monitoring.

4.2.2. Quantitative analysis

We further compared the visual performance of different methods. To simulate occasional adverse environmental conditions in drilling monitoring scenarios, we introduced gaussian noise to generate blurred environments. The comparison results are illustrated in Fig. 5, which presents qualitative detection outcomes from Faster R-CNN, YOLOv8, DETR, deformable DETR, and MSA-DETR on both the WSM-phone and WSM-safety datasets under clear and blurred settings.

On the WSM-phone dataset, all methods can accurately detect the main targets under normal conditions. However, when gaussian blur is applied, Faster R-CNN and DETR frequently miss small or partially occluded objects, while YOLOv8 produces additional false positives around background regions. Deformable DETR exhibits improved localization compared with DETR but still struggles with blurred edges and fine-grained targets. In contrast, MSA-DETR consistently maintains high-confidence predictions

Table 1
Comparison with SOTA on different datasets.

Dataset	Methods	Epochs	mAP, %	AP_{50}, %	AP_{75},%	AP_S, %	AP_M,%	AP_L, %	Params, M	GFLOPs	FPS
WSM-phone	*CNN-based object detectors*										
	Faster R-CNN	100	46.5	76.1	49.7	25.8	34.3	57.7	42	180	25
	YOLOv8	100	67.2	84.6	66.0	46.3	56.6	61.4	68	257	45
	Transformer based object detectors										
	DETR	300	58.0	80.8	51.6	14.0	37.7	61.9	41	187	11
	Deformable DETR	50	66.7	85.7	63.4	40.1	48.9	63.0	40	173	14
	DN-DETR	50	67.5	86.4	68.2	45.3	54.1	65.5	48	195	9
	Lite-DETR	50	68.1	87.1	70.5	49.6	57.8	67.2	47	203	21
	Ours										
	MSA-DETR	50	68.6	87.6	72.4	52.4	59.7	68.4	48	138	25
WSM-safety	*CNN-based object detectors*										
	Faster R-CNN	100	36.5	73.8	32.5	13.6	34.9	47.1	42	180	25
	YOLOv8	100	53.7	86.2	53.7	43.6	45.7	54.8	68	257	45
	Transformer based object detectors										
	DETR	300	49.7	84.8	50.2	28.2	39.1	50.8	41	187	11
	Deformable DETR	50	51.0	85.9	52.6	34.5	43.1	52.9	40	173	14
	DN-DETR	50	53.2	87.2	55.8	41.2	44.8	53.8	48	195	9
	Lite-DETR	50	195	54.9	88.1	58.4	47.6	45.9	54.7	47	21
	Ours										
	MSA-DETR	50	56.1	88.6	60.1	51.4	46.8	55.6	48	138	25

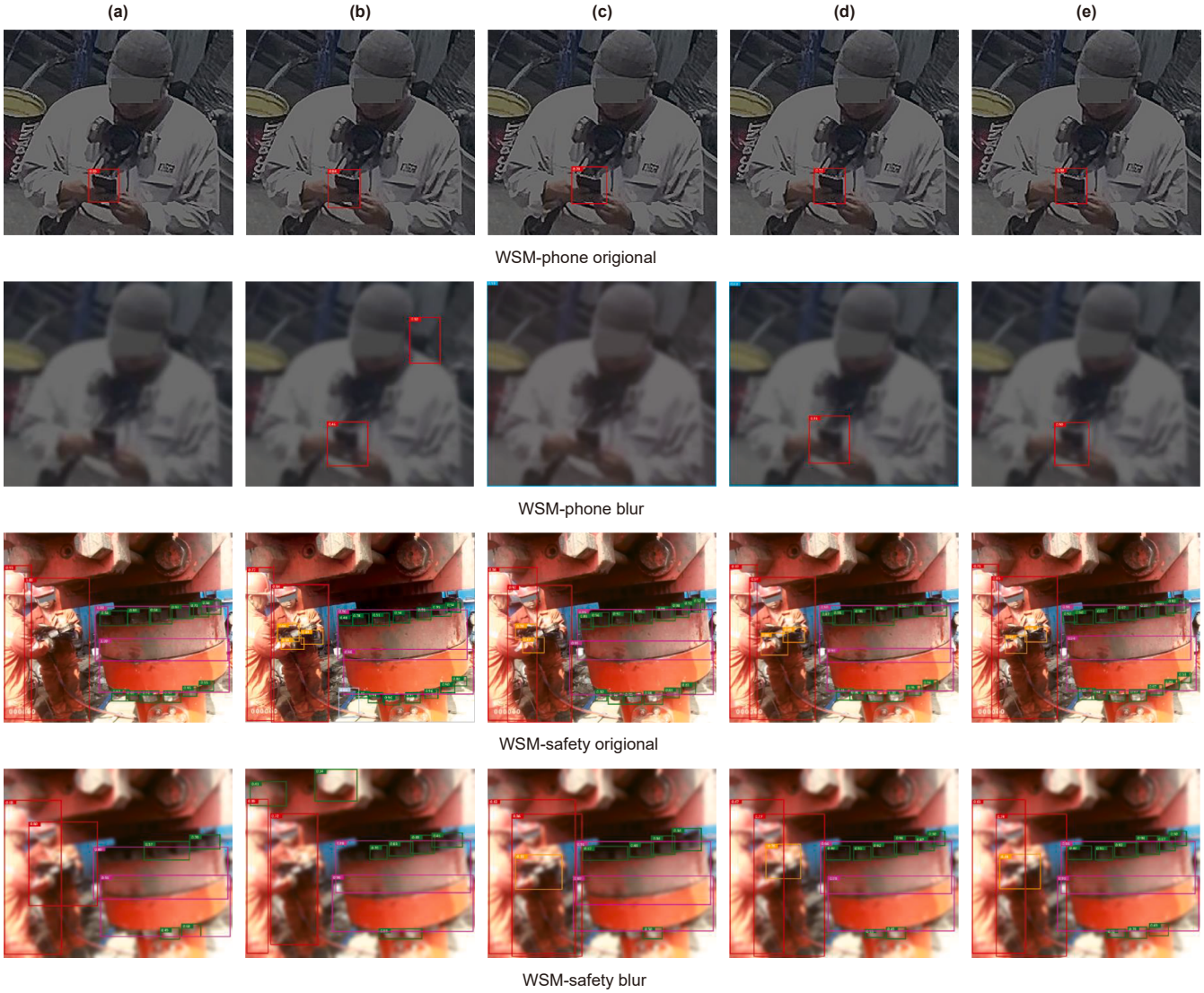

Fig. 5. Comparisons on object detection among different models and datasets. (**a**) Faster R-CNN. (**b**) YOLOv8. (**c**) DETR. (**d**) Deformable DETR. (**e**) MSA-DETR.

and precise bounding box alignment, even under degraded visual quality, demonstrating superior robustness.

For the WSM-safety dataset, all detectors perform well on clear images. Under blurred conditions, Faster R-CNN and DETR experience a notable decline in detection accuracy, and YOLOv8 shows unstable bounding boxes with reduced precision. Deformable DETR delivers more stable detections than these methods but occasionally misses small-scale components. MSA-DETR, benefiting from its multi-scale attention fusion, achieves the most complete and reliable detection coverage across multiple object categories, highlighting its strong adaptability to complex and noisy drilling environments.

Overall, these visual comparisons confirm that MSA-DETR provides the most robust and consistent performance across different environmental conditions, effectively balancing precision and stability in real-world drilling monitoring scenarios. Fig. 6 presents the attention distribution maps from the decoder of MSA-DETR. It can be observed that in lower-level feature layers, the attention is more dispersed, which facilitates better global context modeling. In contrast, higher-level feature layers exhibit more concentrated attention focused on highly sensitive regions. This hierarchical attention mechanism enables the model to maintain robustness in complex environments with multiple targets. The visualization results provide direct evidence of the effectiveness of modeling based on multi-scale feature layers.

4.3. Ablation study

4.3.1. Effectiveness of each proposed component

Setup. To validate the effectiveness of the proposed components, we conducted ablation studies. A DETR model extracting features from stages S_2 to S_4 of the backbone is used as the baseline, and evaluations are performed on the WSM-safety, which involves abundant multi-scale equipments. The results demonstrate that each proposed module contributes to performance improvements. Specifically, the hierarchical alternated self-attention interaction and the enhanced attention pooling module in the multi-scale attention pooling encoder enhance the capability to capture small-scale objects. Meanwhile, the scale-weighted mechanism in the multi-scale attention enhanced visual decoder improves the utilization of multi-scale features from the encoder, leading to gains across different object sizes.

Quantitative results. With deformable DETR as the baseline model, Table 2 presents the performance of MSA-DETR with each proposed component incrementally added. First, introducing the hierarchical alternated self-attention significantly strengthens the ability to model contextual structures, resulting in a 0.4 improvement on mAP overall, with a notable 4.3 gain on AP_S. Next, the enhanced attention pooling module further improves receptive field aggregation, pushing AP_S to 45.6, verifying its effectiveness in perceiving fine-grained targets in complex scenes.

Fig. 6. Attention heatmap visualization. (**a**)–(**d**) are Layer 1–4 respectively, corresponding to attention interactions from feature layers $S_1 - S_4$. **(e)** Integrates the attention from each layer. High-level layers (S_1–S_3) focus on salient targets with concentrated attention; the low-level layer (S_4) focuses on global attention modeling with dispersed attention.

Table 2
Ablation study on the effectiveness of different components. HA, EP, and SW represent hierarchical alternated self-attention interaction, enhanced attention pooling module, and scale-aware weighting mechanism, respectively. Values in parentheses indicate improvements over the baseline.

HA	EP	SW	mAP, %	**AP_S**, %	**AP_M**, %	**AP_L**, %
			51.0	34.5	43.1	52.9
√			51.4 (+0.4)	39.8 (+4.3)	44.1 (+1.0)	53.1 (+0.2)
	√		52.2 (+1.2)	45.6 (+11.1)	44.8 (+1.7)	53.7 (+0.8)
		√	51.7 (+0.7)	40.7 (+6.2)	45.9 (+2.8)	54.2 (+1.3)
√	√		54.4 (+3.4)	48.8 (+14.3)	46.0 (+2.9)	55.0 (+2.1)
√		√	53.8 (+2.8)	47.9 (+13.4)	46.6 (+3.5)	54.4 (+1.5)
√	√	√	**56.1(+5.1)**	**51.4(+16.9)**	**46.8(+3.7)**	**55.6(+2.7)**

Furthermore, the scale-weighted mechanism dynamically assigns weights to features at different scales, effectively enhancing multi-scale feature fusion and decoding in the encoder. This adjustment results in consistent gains, with AP_M improved by as much as 2.8 and AP_L improved by 1.3. When all three modules are integrated, the model achieves optimal performance, reaching an mAP of 56.1, which is 5.1 higher than the baseline. Notably, AP_S increases by 16.9, AP_M by 3.7, and AP_L by 2.7, validating the complementarity and collective value of the proposed modules in multi-scale object detection.

In summary, the proposed multi-scale attention modules significantly enhance small object detection while maintaining strong performance on medium and large targets, offering a more effective design for multi-scale object detection in complex drilling environments.

4.3.2. *Effectiveness of downsampling methods*

Setup. To verify the effectiveness of the proposed downsampling feature fusion strategy, we conducted a set of ablation studies under the same experimental setting. By training model variants based on different downsampling methods, we investigate how various operations affect the feature capturing capability. Specifically.

- Avgpooling expands the receptive region by averaging values, enhancing global feature consistency.
- Maxpooling focuses on local extrema, strengthening the representation of salient points by highlighting peak features.
- Conv extracts spatial local correlations via convolution kernels, emphasizing semantic abstraction.

Quantitative results. As shown in Table 3, avgpooling with kernel size two and stride two achieves the best overall performance, reaching 56.1 on mAP on WSM-safety and the highest AP_S of 51.4, which confirms its effectiveness for small-object detection. At the same time, avgpooling maintains low model complexity, with 48 M parameters, 138 GFLOPs, and an inference speed of 25 FPS. In contrast, convolution-based downsampling increases both parameter count and computational cost, leading to lower inference speed and less stable performance. Maxpooling remains computationally efficient but consistently underperforms avgpooling in detection accuracy, particularly for small and medium-sized objects. Overall, avgpooling provides the most favorable balance between accuracy and efficiency for attention pooling in drilling surveillance scenarios.

Overall, avgpooling provides the most favorable balance between detection accuracy, computational complexity, and inference speed. These results confirm that avgpooling is a more robust and deployment-friendly downsampling strategy for attention-based multi-scale feature fusion in real-world drilling surveillance scenarios.

Quantitative analysis. Fig. 7 presents attention visualizations for the three downsampling methods, using the best-performing parameter settings from Table 3. As shown, avgpooling yields more dispersed attention, covering a wider range of targets,

Table 3
Ablation study on the effectiveness of different downsampling methods.

Downsumpling func	Kernel/Stride	mAP, %	**AP_S**, %	**AP_M**, %	**AP_L**, %	Params, M	GFLOPs	FPS
Conv	$K = 2, S = 2$	0.54	0.512	0.463	0.515	52	165	18
	$K = 2, S = 1$	0.415	0.385	0.335	0.391	52	225	15
	$K = 3, S = 2$	0.460	0.402	0.399	0.441	56	184	17
	$K = 3, S = 1$	0.447	0.387	0.377	0.426	56	309	14
Maxpooling	$K = 2, S = 2$	0.556	0.509	0.462	0.551	48	138	25
	$K = 3, S = 3$	0.464	0.357	0.419	0.452	48	136	29
Avgpooling	$\boldsymbol{K = 2, S = 2}$	**0.561**	**0.514**	**0.468**	**0.556**	48	138	25
	$K = 3, S = 3$	0.478	0.375	0.431	0.456	48	136	29

Fig. 7. Comparison of attention visualizations for three different downsampling methods. (**a**) Avgpooling. (**b**) Maxpooling. (**c**) Conv. Among them, avgpooling shows more dispersed attention, effectively attending to various targets; Maxpooling exhibits concentrated and prominent attention; Conv has larger attention errors and poorer performance.

maxpooling produces highly concentrated and sharp attention maps, which, due to their excessive sharpness, lead to the loss of certain fine-grained attention, while conv exhibits greater attention deviation and poorer results.

In conclusion, avgpooling demonstrates superior performance in balancing global feature aggregation and multi-scale object adaptability, validating its effectiveness as a downsampling-based feature fusion strategy.

5. Conclusion

To optimize multi-scale object detection in oilfield surveillance scenarios and enhance safety assurance in operational environments, we propose a novel detection framework named MSA-DETR. This framework aggregates and compresses attention from multiple feature layers to identify critical objects related to safety monitoring. Specifically, we construct a multi-scale feature attention fusion model, where the encoder takes multi-scale features extracted from the backbone and performs hierarchical alternated attention updates. Furthermore, we introduce a downsampling-based feature fusion method to expand the receptive field for sensitive targets, effectively improving the detection of small-scale objects. In the decoder, a scale-aware weighting mechanism is designed to enable accurate bounding box regression and localization. Experiments and evaluations conducted on two representative drilling surveillance scenarios demonstrate that, compared with four baseline models, our proposed MSA-DETR is more suitable for complex environments characterized by cluttered backgrounds, varying object scales, and diverse viewpoints. In future work, we plan to extend MSA-DETR to more diverse oilfield surveillance datasets to further evaluate its generalization across different operational conditions. We will also explore model optimization and inference acceleration techniques to facilitate integration with real-time industrial monitoring systems. In addition, future research will focus on improving robustness under extreme weather, fog, haze, and low-light conditions by adopting more robust training strategies or incorporating additional contextual information.

CRediT authorship contribution statement

Qian-Wen Cao: Writing – review & editing, Visualization, Supervision, Methodology, Conceptualization. **Jin-Rong Ma:** Writing – original draft, Software, Project administration, Investigation, Data curation. **Lai-Bin Zhang:** Supervision, Investigation, Funding acquisition.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the Oil&Gas Major Project (Grant No. 2025ZD1403701), National Natural Science Foundation of China (No. 62402526), Beijing Natural Science Foundation (No. 4244086), and Science Foundation of China University of Petroleum, Beijing (Nos. 2462025PTJS003, 2462023YJRC029).

References

Ahmed, M.I.B., Saraireh, L., Rahman, A., Al-Qarawi, S., Mhran, A., Al-Jalaoud, J., Al-Mudaifer, D., Al-Haidar, F., Al Khulaif, D., Youldash, M., Gollapalli, M., 2023. Personal protective equipment detection: a deep-learning-based sustainable approach. Sustainability 15. https://doi.org/10.3390/su151813990.

Cao, J., Peng, B., Gao, M., Hao, H., Li, X., Mou, H., 2025. Object detection based on CNN and vision-transformer: a survey. IET Comput. Vis. 19. https://doi.org/10.1049/cvi2.70028.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Gao, P., Zheng, M., Wang, X., Dai, J., Li, H., 2021. Fast convergence of DETR with spatially modulated co-attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3621–3630. https://doi.org/10.1109/iccv48922.2021.00360.

Gong, F., Ji, X., Gong, W., Yuan, X., Gong, C., 2021. Deep learning based protective equipment detection on offshore drilling platform. Symmetry 13, 954. https://doi.org/10.3390/sym13060954.

Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of things (IoT): A vision, architectural elements, and future directions. Future Gener. Comput. Syst. 29, 1645–1660. https://doi.org/10.1016/j.future.2013.01.010.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.90.

Huang, J., Li, T., 2024. Small object detection by DETR via information augmentation and adaptive feature fusion. In: Proceedings of 2024 ACM ICMR Workshop on Multimodal Video Retrieval, pp. 39–44. https://doi.org/10.1145/3664524.3675362.

Iqbal, E., Khan, S.U., Javed, S., Moyo, B., Zweiri, Y., Abdulrahman, Y., 2024. Multi-scale feature reconstruction network for industrial anomaly detection. Knowl. Base Syst. 305, 112650. https://doi.org/10.1016/j.knosys.2024.112650.

Li, F., Zeng, A., Liu, S., Zhang, H., Li, H., Zhang, L., Ni, L.M., 2023. Lite DETR: an interleaved multi-scale encoder for efficient DETR. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18558–18567. https://doi.org/10.1109/cvpr52729.2023.01780.

Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L., 2022. DN-DETR: accelerate DETR training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13619–13627. https://doi.org/10.1109/cvpr52688.2022.01325.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L., 2022. DAB-DETR: Dynamic anchor boxes are better queries for DETR arXiv preprint arXiv: 2201.12329.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. SSD: Single shot multibox detector. In: European Conference on Computer Vision. Springer, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

Lyu, P., Liu, J., Zhang, Y., Ye, B., Lan, T., Bai, L.P., Cai, Z., Jiang, Z.H., 2024. A novel feature fusion framework for industrial automation single-multiple object detection. IEEE Trans. Ind. Inf. 20, 7686–7697. https://doi.org/10.1109/TII.2024.3353814.

Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J., 2021. Conditional DETR for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3651–3660. https://doi.org/10.1109/iccv48922.2021.00363.

Miri Rekavandi, A., Rashidi, S., Boussaid, F., Hoefs, S., Akbas, E., Bennamoun, M., 2025. Transformers in small object detection: A benchmark and survey of state-of-the-art. ACM Comput. Surv. 58, 1l–33. https://doi.org/10.1145/3758090.

Neto, W.A.B.L., Goncalves, J.H.M., Dupslan, M., 2024. Video-powered AI solutions for active monitoring and safety improvement in offshore environments: the bram offshore and altave case. In: SPE Annual Technical Conference and Exhibition. SPE-221026-MS. https://doi.org/10.2118/221026-ms.

Nikouei, M., Baroutian, B., Nabavi, S., Taraghi, F., Aghaei, A., Sajedi, A., Moghaddam, M.E., 2025. Small object detection: a comprehensive survey on challenges, techniques and real-world applications. Intell. Syst. Appl. 27, 20056l. https://doi.org/10.1016/j.iswa.2025.200561.

Pang, G., Shen, C., Cao, L., Hengel, A.V.D., 2021. Deep learning for anomaly detection: A review. ACM Comput. Surv. 54, 1–38. https://doi.org/10.1145/3439950.

Peng, J., Shao, H., Xiao, Y., Cai, B., Liu, B., 2024. Industrial surface defect detection and localization using multi-scale information focusing and enhancement GANomaly. Expert Syst. Appl. 238, 122361. https://doi.org/10.1016/j.eswa.2023.122361.

Ramos, L.T., Sappa, A.D., 2025. A decade of you only look once (YOLO) for object detection: A review. IEEE Access 13, 192747–192794. https://doi.org/10.1109/access.2025.3630988.

Reddy, S.N., Kurrey, V., Nagar, M., Gupta, G.R., 2024. Action recognition based industrial safety violation detection. arXiv preprint, arXiv:2412.05531. https://doi.org/10.48550/arXiv.2412.05531.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, pp. 779–788. https://doi.org/10.1109/cvpr.2016.91.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28. https://doi.org/10.1109/tpami.2016.2577031.
Rijayanti, R., Hwang, M., Jin, K., 2023. Detection of anomalous behavior of manufacturing workers using deep learning-based recognition of human-object interaction. Appl. Sci. 13, 8584. https://doi.org/10.3390/app13158584.
Serror, M., Hack, S., Henze, M., Schuba, M., Wehrle, K., 2021. Challenges and opportunities in securing the industrial internet of things. IEEE Trans. Ind. Inf. 17, 2985–2996. https://doi.org/10.1109/TII.2020.3023507.
Shehzadi, T., Hashmi, K.A., Liwicki, M., Stricker, D., Afzal, M.Z., 2025. Object detection with transformers: A review. Sensors 25, 6025. https://doi.org/10.3390/s25196025.
Tao, H., Zheng, Y., Wang, Y., Qiu, J., Stojanovic, V., 2024. Enhanced feature extraction YOLO industrial small object detection algorithm based on receptive-field attention and multi-scale features. Meas. Sci. Technol. 35, 105023. https://doi.org/10.1088/1361-6501/ad633d.
Varghese, R., Sambath, M., 2024. YOLOv8: a novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE, pp. 1–6. https://doi.org/10.1109/adics58448.2024.10533619.
Wang, A., Sun, Y., Kortylewski, A., Yuille, A.L., 2020. Robust object detection under occlusion with context-aware CompositionalNets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12645–12654. https://doi.org/10.1109/cvpr42600.2020.01266.
Wang, X., Xie, Z., Yan, F., Wang, J., Fan, J., Zeng, Z., Lu, J., Zhang, H., Zeng, N., 2025. Towards more accurate industrial anomaly detection: a component-level feature-enhancement approach. Electronics 14, 1613. https://doi.org/10.3390/electronics14081613.
Wang, Y., Chen, X., Wang, L., 2023a. Cyber-physical oil spill monitoring and detection for offshore petroleum risk management service. Sci. Rep. 13, 4586. https://doi.org/10.1038/s41598-023-30311-w.
Wang, Y., Zhang, X., Yang, T., Sun, J., 2022. Anchor DETR: query design for transformer-based detector. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2567–2575. https://doi.org/10.1609/aaai.v36i3.20158.
Wang, Z., Bai, L., Song, G., Zhang, Y., Zhu, M., Zhao, M., Chen, L., Wang, M., 2023b. Optimized faster R-CNN for oil wells detection from high-resolution remote sensing images. Int. J. Rem. Sens. 44, 6897–6928. https://doi.org/10.1080/01431161.2023.2275322.
Zhan, C., Bai, K., Tu, B., Zhang, W., 2024. Offshore oil spill detection based on CNN, DBSCAN, and hyperspectral imaging. Sensors 24, 411. https://doi.org/10.3390/s24020411.
Zhang, J., Zhang, H., Liu, B., Qu, G., Wang, F., Zhang, H., Shi, X., 2023. Small object intelligent detection method based on adaptive recursive feature pyramid. Heliyon 9, e17730. https://doi.org/10.1016/j.heliyon.2023.e17730.
Zhang, R., Cai, B.P., Yang, C., Zhou, Y.M., Liu, Y.H., Qi, X.Y., 2024. Combinatorial reasoning-based abnormal sensor recognition method for subsea production control system. Pet. Sci. 21, 2758–2768. https://doi.org/10.1016/j.petsci.2024.02.015.
Zhang, S., Wang, X., Cheng, Y.F., Shuai, J., 2020. Modeling and analysis of a catastrophic oil spill and vapor cloud explosion in a confined space upon oil pipeline leaking. Pet. Sci. 17, 556–566. https://doi.org/10.1007/s12182-019-00403-2.
Zhu, H., Wei, H., Li, B., Yuan, X., Kehtarnavaz, N., 2020a. A review of video object detection: datasets, metrics and methods. Appl. Sci. 10, 7834. https://doi.org/10.3390/app10217834.
Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020b. Deformable DETR: Deformable Transformers for end-to-end object detection. arXiv preprint, arXiv:2010.04159. https://doi.org/10.48550/arXiv.2010.04159.