2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

# Intelligent Fish Detection System with Similarity-Aware Transformer

Shengchen Li[1,†], Haobo Zuo[2,†], Changhong Fu[1,*], Zhiyong Wang[3], Zhiqiang Xu[3]

***Abstract*— Fish detection in water-land transfer has significantly contributed to the fishery. However, manual fish detection in crowd-collaboration performs inefficiently and expensively, involving insufficient accuracy. To further enhance the water-land transfer efficiency, improve detection accuracy, and reduce labor costs, this work designs a new type of lightweight and plug-and-play edge intelligent vision system to automatically conduct fast fish detection with high-speed camera. Moreover, a novel similarity-aware vision Transformer for fast fish detection (FishViT) is proposed to onboard identify every single fish in a dense and similar group. Specifically, a novel similarity-aware multi-level encoder is developed to enhance multi-scale features in parallel, thereby yielding discriminative representations for varying-size fish. Additionally, a new soft-threshold attention mechanism is introduced, which not only effectively eliminates background noise from images but also accurately recognizes both the edge details and overall features of different similar fish. 85 challenging video sequences with high framerate and high-resolution are collected to establish a benchmark from real fish water-land transfer scenarios. Exhaustive evaluation conducted with this challenging benchmark has proved the robustness and effectiveness of FishViT with over 80 FPS. Real work scenario tests validate the practicality of the proposed method. The code and demo video are available at https://github.com/vision4robotics/FishViT.**

## I. Introduction

Intelligent vision systems can effectively solve the problems of low efficiency, low accuracy, and high cost associated with traditional crowd-collaboration mode for fish detection in water-land transfer. Water-land transfer, *i.e.*, transferring fresh fish from surface fish culture vessels to vehicles for sale, while keeping fish alive and injury-free. Currently, as shown in Fig. 1, numerous terminals still rely on manual detection methods to detect fish during the process of water-land transfer. However, due to the rapid expansion of the fishery market [1], traditional manual fish detection has gradually revealed its shortcomings of low efficiency, high cost, and low precision. The specific reasons mainly include the following aspects: *1)* manual fish detection is prone to fatigue and errors due to the prolonged attention and concentration required; *2)* this kind of detection is subject to inconsistencies that arise from variations in individual perception, training, and attention levels, leading to mix-ups and misidentifications; *3)* this way relies heavily on human labor, which can be expensive, prone to fish injury,

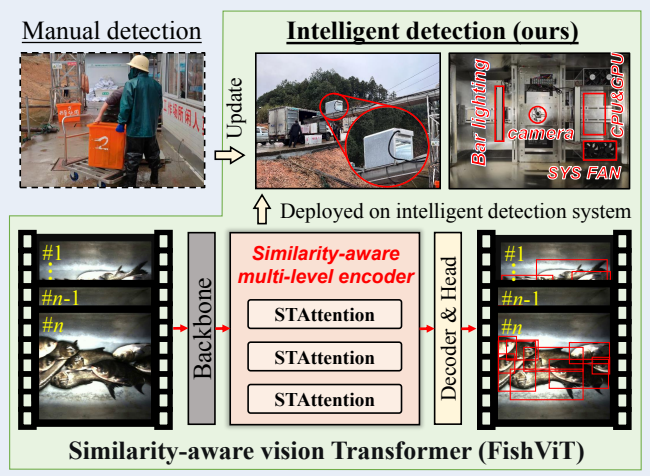


Fig. 1. Fish detection in water-land transfer and the proposed fish detection method (FishViT). The picture with the dotted box represents the traditional manual detection mode, which is inefficient and costly. The proposed intelligent detection system significantly improves production efficiency, enhances accuracy, and reduces costs. Fresh fish slide fast from the pipeline into vehicles for sale and the intelligent detection device stand detects above the pipeline. The proposed FishViT is embedded into the intelligent detection system to effectively realize high-speed fish detection. FishViT mainly consists of three components: Backbone, Similarity-aware multi-level encoder, and Decoder & Head. The Similarity-aware multi-level encoder is composed of three parallel soft-threshold attention (STAttention) modules.

and challenging to scale up or down based on demand. ***Therefore, how to design a high efficiency, high precision and fish-injury-free intelligent fish detection system is of great urgency.***

With the development of deep learning, learning-based fish detection has drawn considerable attention due to its prosperous applications in fishery, *e.g.*, fish economic benefit estimation [2], feedstuff feeding [3], and culture density adjustment [4]. It can bring enormous economic benefits to the fishery. Early learning-based detection methods are mostly based on convolutional neural networks (CNNs) [5], [6]. However, since CNNs lack global information integration ability, CNN-based detection methods struggle to identify and locate the fish precisely in the presence of abundant highly similar appearance interference.

Recently, Transformer-based object detection methods have gained increasing popularity in the academic community due to their ability for global modeling high robustness, and detection accuracy. [7]. Despite remarkable advancements, these techniques have difficulties to be applied for efficient and accurate fish detection. Firstly, the fish slide down at high speed in the pipeline for water-land transfer, which makes it hard to meet real-time application needs

[1]Shengchen Li and Changhong Fu are with the School of Mechanical Engineering, Tongji University, Shanghai 201804, China. changhongfu@tongji.edu.cn
[2]Haobo Zuo is with the Department of Computer Science, University of Hong Kong, Hong Kong 999077, China.
[3]Zhiyong Wang and Zhiqiang Xu are with the Fishery Machinery and Instrument Research Institute, Shanghai 200092, China.
†Equal contribution *Corresponding author

because of the high computational complexity of classical Transformer structures. Moreover, the presence of abundant highly similar fish in the water-land transfer pipeline and the frequent occurrence of challenging scenarios such as occlusion and cluttered backgrounds disturb the global information integration of Transformer. ***Thus, how to develop a lightweight, efficient, and robust Transformer-based detector to handle fish detection challenges, thereby satisfying the increasing demand for fish water-land transfer is an urgent problem.***

The essential component of the Transformer is the attention mechanism. The performance of attention mechanisms has been greatly improved and innovated by state-of-the-art (SOTA) works [8]. However, when facing fish detection in water-land transfer with high similarity and density, traditional attention shows poor robustness due to a lack of sufficient identification ability.

To address the aforementioned issues, the proposed similarity-aware vision Transformer for fast fish detection (FishViT) is deployed on the self-built intelligent vision device to effectively recognize every single fish in a dense and similar group, as indicated in Fig. 1. Specifically, the similarity-aware multi-level encoder is designed to enhance multi-scale features in parallel, generating discriminative representations for fish of various sizes. Importantly, the soft-threshold attention (STAttention) is introduced to suppress background noise thereby accurately recognizing the edge details of diverse, similar fish. The main contributions of this work are as follows:

- A lightweight and plug-and-play intelligent system is designed to automatically realize dense and fast fish detection with the onboard camera, heat sink, and processor. Compared with traditional manual detection, it can improve efficiency and accuracy at lower cost.
- A similarity-aware vision Transformer for fast fish detection is introduced for the efficient detection of individual fish within dense and visually similar groups, deployed on the self-built intelligent system. A similarity-aware multi-level encoder is proposed to enhance multi-scale feature representation capabilities in parallel.
- An innovative soft-threshold attention mechanism is presented to effectively eliminate background noise from images, thereby precisely discerning the edge information of diverse yet similar fish. This mechanism aims to clarify every boundary of similar fish, leading to improved performance in fish detection.
- Comprehensive evaluations on the 85 high-quality video sequences of real fish water-land transfer validate the promising performance of FishViT compared with other SOTA detectors. Work scenario tests have demonstrated the superior practicability of the proposed FishViT.

## II. Related Works

### A. Fish Detection

Compared with early manual fish detection, the rapid development of deep learning makes fish detection intelligent. Categorized by application scenarios, fish detection can be divided into two categories: underwater and overwater. Most modern fish detection applications and datasets focus on underwater scenarios and yield many impressive achievements [9]. Currently, there is little research has been done on detectors for fish water-land transfer. In terms of detection methods, fish detectors can be categorized as CNN-based and Transformer-based detectors. CNN-based fish detectors, such as Faster R-CNN [10] and YOLO [11], achieve promising results on various detection benchmarks. Compared with CNN-based detectors, the Transformer-based detectors have higher robustness and detection accuracy, such as DETRs [12], [13]. However, due to the high computational complexity of classical Transformer structures, the Transformer-based detectors are difficult to achieve a trade-off between high performance and real-time speed. Furthermore, due to the presence of abundant highly similar fish in the water-land transfer pipeline and the frequent occurrence of challenging scenarios such as occlusion and cluttered backgrounds, existing fish detectors are difficult to apply effectively in water-land transfer.

### B. Attention Mechanism

The attention mechanism plays a crucial role in deep learning with its outstanding capabilities for global feature modeling [14]. However, the high computational complexity of self-attention limits its application in visual tasks. There have been many research attempts to address this problem from multiple perspectives. One line of research is using linear attention to address high computation complexity [15], which replaces the Softmax function in self-attention with separate kernel functions. K. M. Choromanski *et al.* [16] propose Performers, approximating the Softmax operation with orthogonal random features. EfficientViT [17] uses depth-wise convolution to improve linear attention's local feature extraction capacity. W. Xu *et al.* [18] build on the above work to design a factorized attention mechanism, significantly enhancing computation efficiency. On the other hand, in order to further improve the model performance, M. Zhao *et al.* [19] add the soft-threshold [20] to deep residual networks, which effectively helps the model to better capture the key features and filter the noise. However, most of the above works are based on generalized scenario datasets, which are less robust when facing dense, occluded complex scenarios with highly similar fish in water-land transfer.

### C. Feature Fusion

Except for attention, another major challenge in object detection is effectively utilizing multi-scale features, which have been demonstrated to significantly improve performance, especially for varying-size objects. In modern CNN-based detectors [21], feature pyramid network (FPN) [22] has become the primary solutions to exploit multi-scale features. Analogously, many Transformer-based detectors also attempt to improve DETRs by feature fusion. Deformable-DETR [23] first introduces multi-scale features into DETR, exchanging information among multi-scale feature maps

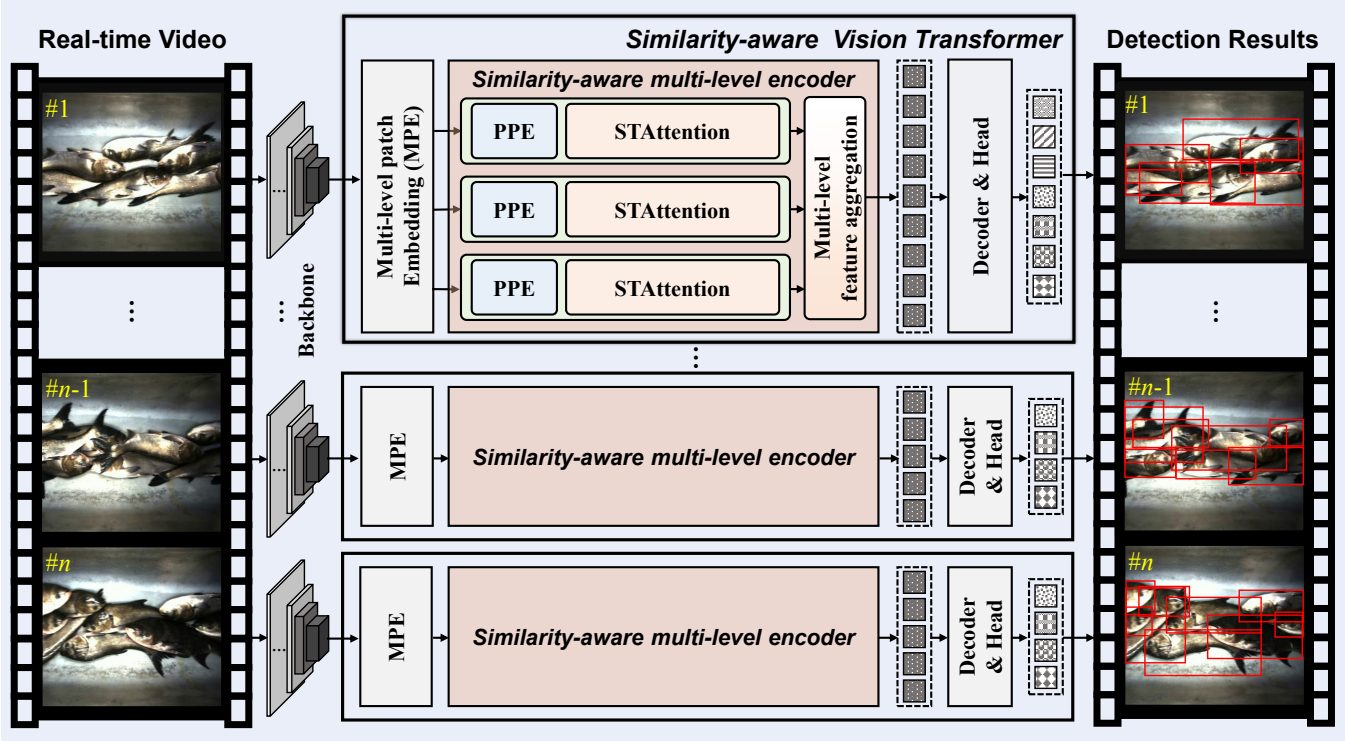


Fig. 2. Overview of the proposed FishViT. The components from the left to right are *Backbone*, *Similarity-aware multi-level encoder*, *Decoder & Head*. The last three feature maps extracted by backbone are fed to similarity-aware multi-level encoder with parallel structure, each branch of encoder contains pooling positional encoding (PPE) and soft-threshold attention (STAttention). Finally, the feature map of each branch after multi-level aggregation are fed into decoder & head for detection. Best viewed in color.

while improving the performance and convergence speed. SMCA-DETR [24] achieves efficient multi-scale feature coding in the encoder by introducing multi-scale self-attention and utilizes concatenate for feature fusion. CF-DETR [25] integrates the Transformer encoder within an FPN architecture to generate feature pyramids. Iterative multi-scale feature aggregation [26] designs a sparse sampling strategy for multi-scale features to boost the performance of Transformer-based object detectors significantly with only slight computational overhead. Although the above methods enable the use of multi-scale features in Transformer-based detectors, they introduce huge computational overhead, making it difficult to effectively realize end-to-end real-time object detection.

## III. Methodology

The workflow of FishViT is shown in Fig. 2. It can be divided into three modules: *Backbone*, *Similarity-aware multi-level encoder*, *Decoder & Head*. The encoder mainly consists of two novel parts, *i.e.*, similarity-aware multi-level parallel structure, and soft-threshold attention mechanism.

### A. Backbone

To meet the real-time applications onboard the intelligent vision system, FishViT uses a lightweight backbone known as ResNet18 [27], which is used to extract multi-scale features. Specifically, the last three output feature maps of the backbone are utilized as the input to the subsequent process.

***Remark 1:*** Given an image $\mathcal{I} \in \mathbb{R}^{W_0 \times H_0 \times 3}$, the backbone network generates its feature maps. For descriptive purposes, the last three output feature maps, which are further fed to the Transformer encoder, are uniformly represented by $\mathcal{F}_l \in \mathbb{R}^{W \times H \times C}$ in the following introduction ($C$, $W$, $H$ represent the channel, width, and height of the feature maps respectively, and $l \in \{3, 4, 5\}$).

### B. Similarity-aware Multi-level Encoder

The prime challenge to water-land transfer fish detection is the high similarity of fish, which tends to result in missed and false detections in high-density and occlusion scenarios. To cope with this issue, a lightweight similarity-aware multi-level encoder based on STAttention is designed. Specifically, the multi-level parallel structure is designed to enhance the multi-scale feature representation for addressing the fish appearance similarity issues. Through the multi-level patch embedding, feature maps of different levels are obtained and fed to each branch in parallel. The self-attention mechanism is utilized in every branch for global modeling and establishing context relationships.

***Remark 2:*** Through the design of the similarity-aware multi-level encoder, different levels of features are fully extracted. The proposed STAttention performs global modeling and establishes context relationships at different levels, which greatly improves the robustness of the model and increases the detecting accuracy.

*1) Pooling positional encoding:* Positional encoding is a critical component for the self-attention mechanism to be able to understand and process sequence data. Inspired by [18], the pooling positional encoding method (PPE) is proposed to speed up the model while guaranteeing its effect. As shown in Fig. 3, the process of PPE can be described by

the formula as:

$$\mathcal{M}_P = \text{AvgPooling}(\mathcal{M}) + \mathcal{M} \quad , \tag{1}$$

where $\mathcal{M}$ represents the feature vector that after Multi-level patch embedding (MPE), and $\mathcal{M}_P$ represents the $\mathcal{M}$ completed the positional encoding.

***Remark 3:*** The use of the proposed pooling position encoding effectively reduces redundant information and speeds up computation. $\mathcal{F} \in \mathbb{R}^{W\times H\times C}$ and $\mathcal{M} \in \mathbb{R}^{N\times C}$ represent the input and output of the MPE separately in the following introduction.

*2) STAttention:* Given an input feature $\mathcal{F} \in \mathbb{R}^{W\times H\times C}$, the representation is first transformed into the embedding space, denoted as $\mathcal{M} \in \mathbb{R}^{N\times C}$, with a feature dimension of $C$. The generalized attention function can be computed as [28]:

$$\text{Att}(\mathcal{M}) = \sum \frac{\mathcal{S}(\mathbf{Q},\mathbf{K})}{\sum \mathcal{S}(\mathbf{Q},\mathbf{K})}\mathbf{V} \quad , \tag{2}$$

where $\mathcal{S}(\cdot)$ denotes the function for measuring the similarity between queries and keys, if $\mathcal{S}(\mathbf{Q},\mathbf{K}) = \exp(\mathbf{Q}\mathbf{K}^{\mathrm{T}})$, then Eq. 2 simplifies to the $scaled \quad dot-product$ attention mechanism with softmax normalization. However, this approach incurs a computational complexity of $O(N^2C)$, rendering it unsuitable for real-time applications. The use of decomposable similarity functions enables attentions to be computed in a linear manner. Specifically, two functions $\phi(\cdot)$ and $\psi(\cdot)$ are used and the second matrix multiplication is computed:

$$\mathcal{S}(\mathbf{Q},\mathbf{K})\mathbf{V} = (\phi(\mathbf{Q})\psi(\mathbf{K})^T)\mathbf{V} = \phi(\mathbf{Q})(\psi(\mathbf{K})^T\mathbf{V}) \quad , \tag{3}$$

the result leads to a $O(NC^2)$ computation complexity, which greatly reduces its complexity in comparison for $N$ is generally much larger than $C$ value for images. Factorized attention (FactorAtt) is obtained when $\phi$ is the identity function and $\psi$ is the softmax [18]:

$$\text{FactorAtt}(\mathbf{X}) = \frac{\mathbf{Q}}{\sqrt{C}}(\text{Softmax}(\mathbf{K})^T\mathbf{V}) \quad . \tag{4}$$

In the face of high speed and high similarity dense fish detecting scenarios, the means of denoising can convert useful information into active and effective features while turning noisy information into invalid or near-zero features, which enables the detector to better discriminate the edge information of the fish and improve the detection accuracy. The combination of soft-threshold (ST) and deep learning is a promising method for denoising [29], and soft-threshold calculation formula can be expressed as follows:

$$\text{ST}(x) = \begin{cases} x-\tau & , \quad x > \tau \\ 0 & , \quad -\tau < x < \tau \\ x+\tau & , \quad x < -\tau \end{cases} , \tag{5}$$

where $x$ is the input feature, $\tau$ indicates the threshold. Inspired by [19], a deep residual shrinkage network is introduced to automatically determine the threshold. Firstly, as shown in Fig. 3, the absolute operation and GAP layer are used for the result of $(\mathbf{K})^T\mathbf{V}$ to simplify the feature

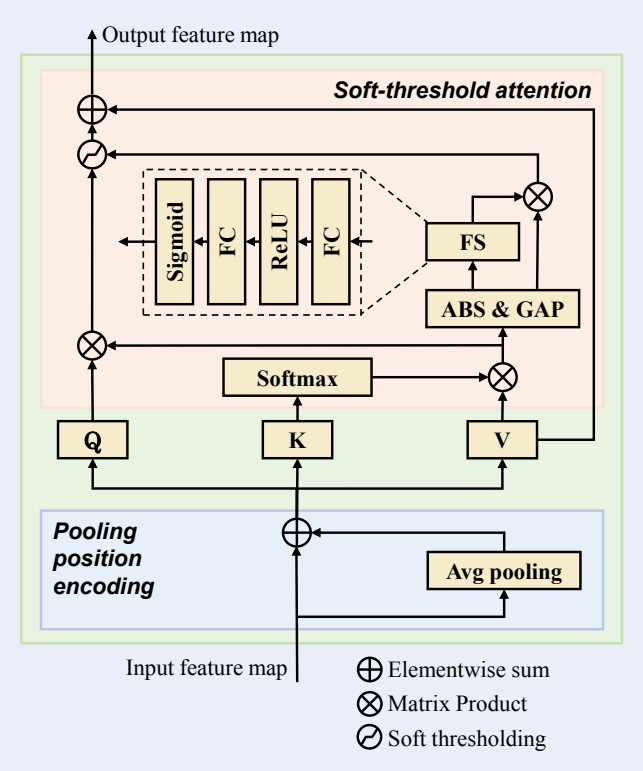


Fig. 3. Detailed workflow of the STAttention. The input feature maps are processed through STAttention after pooling position encoding. The result of $\text{Softmax}(\mathbf{K})^T\mathbf{V}$ is used to generate a soft-threshold, and the linear attention result is filtered by the soft-threshold. Specifically, the soft-threshold mechanism consists of two main modules: ABS & GAP and FS. ABS stands for Absolute Value, and GAP represents Global Average Pooling, while the FS module refers to the operations within the dashed box. Finally, $\mathbf{V}$ is added to the filtered result through the shortcut technique, and the added result is used as the final output feature map. This soft-threshold mechanism effectively suppresses background noise, enabling precise identification of the edges of each individual fish within within dense and visually similar groups. Best viewed in color.

mapping to 1-D vector, and then the feature dimension is reduced through FS module. Eventually, the output feature of FS is scaled to the range (0,1) by sigmoid layer, the formula is described as:

$$\alpha_c = \frac{1}{1+e^{-z_c}} \quad , \tag{6}$$

where $\alpha_c$ is the scaling parameter of feature map $cth$ channel, $z_c$ is the feature at the $cth$ neuron of 1-D vector. Ultimately, the thresholds are calculated by the formula as:

$$\tau_c = \alpha_c \cdot \text{average}\left|((\mathbf{K})^T\mathbf{V})_{i,j,c}\right| \quad , \tag{7}$$

where $\tau_c$ is the threshold for the $cth$ channel of the feature map, and $i$, $j$, and $c$ are the indexes of width, height, and channel of the feature map, respectively.

Meanwhile, the shortcut technique is introduced to solve the problem of gradient vanishing or gradient explosion in deep neural network training. In summary, the formula given in STAttention is as follows:

$$\text{STAttention}(\mathbf{X}) = \text{ST}(\frac{\mathbf{Q}}{\sqrt{C}}(\text{Softmax}(\mathbf{K})^T\mathbf{V})) + \mathbf{V} \quad . \tag{8}$$

TABLE I

MAIN DETECTION RESULTS. RED REPRESENTS THE BEST RESULT AND BLUE REPRESENTS THE SECOND BEST RESULT. ↑ INDICATES THAT A HIGHER VALUE FOR THIS METRIC IS BETTER, WHILE ↓ SIGNIFIES THAT A LOWER VALUE IS PREFERABLE FOR THIS METRIC. ALL REAL-TIME DETECTORS SHARE A COMMON INPUT SIZE OF 640 AND ALL END-TO-END OBJECT DETECTORS SHARE A COMMON INPUT SIZE OF 800. OUR FISHVIT ACHIEVED THE BEST PERFORMANCE ACROSS ALL METRICS.

| Methods | Backbone | GFLOPs ↓ | Params ↓ | $FPS_{bs=16}$ ↑ | $AP_{50}^{val}$ ↑ | $E_{cls}$ ↓ | $E_{loc}$ ↓ | $E_{miss}$ ↓ |
|---|---|---|---|---|---|---|---|---|
| *Real-time Object Detectors* | | | | | | | | |
| YOLOv8-M | | 79 | 26 M | 62.6 | 93.2 | 0.48 | 3.29 | 1.91 |
| YOLOv8-L | | 165 | 79 M | 30.3 | 93.6 | 0.27 | 2.85 | 1.47 |
| YOLOv8-X | | 258 | 165 M | 17.7 | 94.4 | 0.24 | 2.6 | 1.23 |
| *End-to-end Object Detectors* | | | | | | | | |
| DETR-DC5 | R50 | 187 | 41 M | 5.1 | 84.1 | 13.38 | 10.92 | 2.36 |
| DETR-DC5 | R101 | 253 | 60 M | 3.7 | 84.5 | 16.31 | 16.74 | 3.96 |
| Anchor-DETR-DC5 | R50 | 172 | 39 M | 5.4 | 87.6 | 2.46 | 1.53 | 0.38 |
| Anchor-DETR-DC5 | R101 | 240 | 58 M | 1.5 | 89.6 | 2.71 | 1.49 | 0.36 |
| Conditional-DETR-DC5 | R50 | 195 | 44 M | 4.3 | 91.1 | 2.59 | 2.14 | 0.5 |
| Conditional-DETR-DC5 | R101 | 262 | 63 M | 2.6 | 92.7 | 4.37 | 6.78 | 1.23 |
| *Real-time End-to-end Object Detectors* | | | | | | | | |
| RT-DETR | R18 | 56.9 | 20 M | 64.7 | 93.6 | 0.29 | 1.85 | 0.61 |
| **FishViT (ours)** | R18 | **45.6** | **18 M** | **82.3** | **94.7** | **0.11** | **1.65** | **0.28** |

***Remark 4:*** STAttention accelerates the model speed by linearization of self-attention. Meanwhile, through the setting of soft-threshold, the influence of background noise can be better eliminated, thus precisely discerning the edge information of diverse yet similar fish, improving the detection accuracy. The specific process of Eq. 7 is:

$$\begin{aligned} \text{STAttention}(\mathbf{X}) &= \text{ST}(\frac{\mathbf{Q}}{\sqrt{C}}(\text{Softmax}(\mathbf{K})^T\mathbf{V})) + \mathbf{V} \\ &= \text{Sign}(\text{FactorAtt}(\mathbf{X}))\tau_c + \mathbf{V} \quad , \end{aligned} \tag{9}$$

where $\text{Sign}(*)$ represents the sign function.

### C. Decoder & Head

As shown in Fig. 2, the similarity-aware multi-level encoder transforms multi-scale features into a sequence of image features. In order to provide more encoder features with accurate classification and precise location for object queries, the output sequence of the encoder needs to go through the IoU-aware query selection [30] to obtain a fixed number of image features as initial object queries for the decoder. Then, the selected object queries are optimized by the decoder and mapped to classification scores and bounding boxes by the prediction head. Finally, the proposed FishViT uses a Transformer decoder with auxiliary prediction heads to get fish detection results.

***Remark 5:*** As shown in Fig. 4, in challenging scenarios such as tumbling, flowing, and dense, FishViT is able to clearly discriminate the edges of each fish in a dense fish group by benefiting from the parallel multi-scale representation as well as the denoising effect of STAttention.

## IV. EXPERIMENTS

### A. Implementation Details and New Benchmark

This work develops a new type of lightweight plug-and-play intelligent vision system to implement high-speed dense fish detection autonomously. The whole system mainly includes the following two components: a pipeline with a total length of 16 meters and a retractable tail to match different types of land sales vehicles. An intelligent detection device with a high framerate camera (to capture high-speed images of fish), a high-performance processor (Intel *i*9-12900KF+NVIDIA RTX 3060), two bar lightings (to avoid outdoor light interference), four SYS FAN (for heat dissipation), a metal housing (for light resistant and waterproof), and an indoor console. Utilizing the proposed intelligent vision system, 85 challenging video sequences with a high frame rate of 90 frames/s are collected in real fish water-land transfer scenarios to comprehensively assess FishViT's effectiveness in fish detection. The video sequences contain two fish species, *i.e.*: 'Black carp' and 'Silver carp', of which 50 sequences are used for training, 5 sequences are used for validation and 30 sequences for testing. The sequences share six challenging attributes, including 'spume', 'overexposure', 'tumbling', 'dense', 'flowing', and 'underexposure'. ResNet18 [27] serves as the backbone for FishViT. FishViT is trained by AdamW optimizer with the values of $weight\_decay$ and $base\_learning\_rate$ as $10^{-4}$ for 150 epochs.

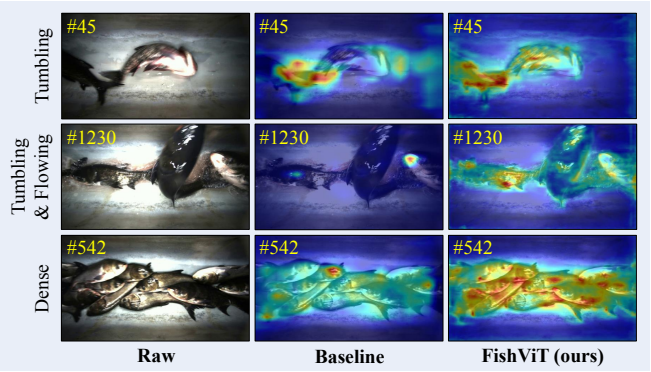


Fig. 4. Visualization of the confidence maps of the Baseline and the proposed FishViT. FishViT can effectively reduce background interference and focus on the detailed representation of the fish to cope with extremely complex situation such as high-speed tumbling, flowing water, and high density. Best viewed in color.

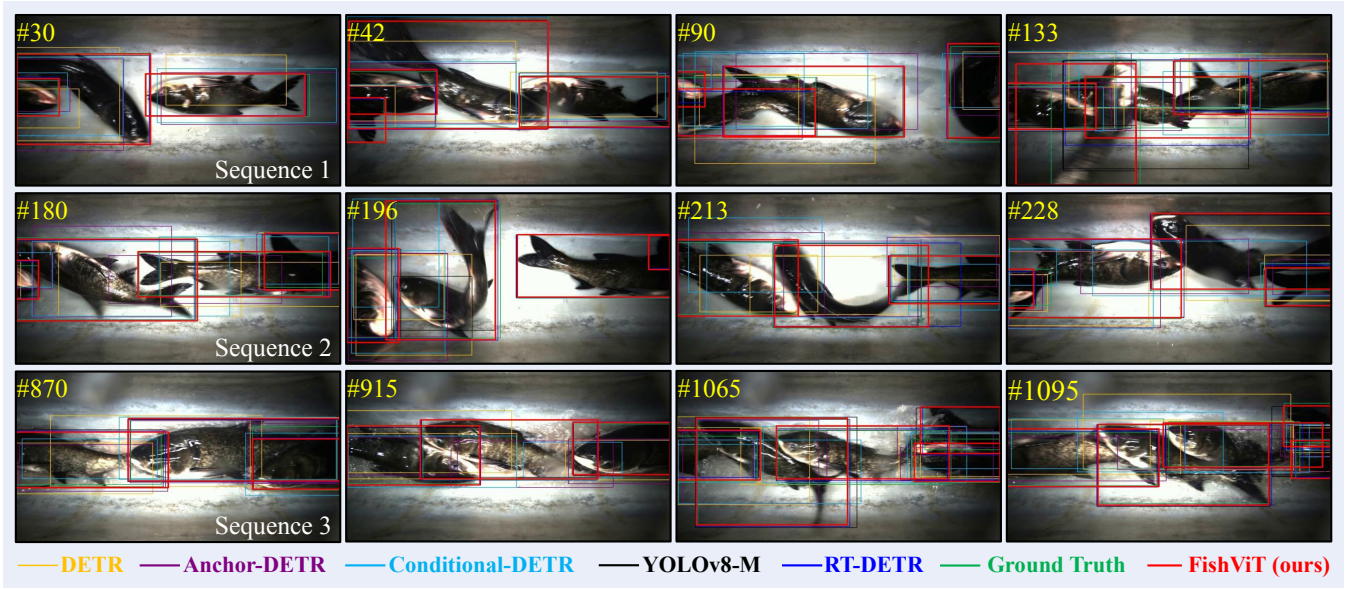


Fig. 5. Comparison of FishViT results with other SOTA detectors. YOLOv8-M has duplicate detection boxes due to post-processing, while other Transformer-based detectors exhibit missed detections. Our FishViT achieved the best results. Best viewed in color.

### *B. Evaluation on Detection Benchmarks*

To fully assess FishViT's fish detection capabilities, the crucial detection accuracy must be evaluated. The evaluation metric used is the standard COCO AP metric with a single scale image as input [31]. In object detection tasks, Average Precision (AP) is usually used as an indicator to measure the performance of a model. However, in different application scenarios, it is not appropriate to just look at the value of AP [32]. In the water-land fish transfer scenario, classification error ($\mathbf{E}_{cls}$) is intolerable because it directly relates to the selling price. In addition, detection tasks are often performed to serve downstream tasks, such as fish counting and fish size recording. Localization error ($\mathbf{E}_{loc}$), and missed GT error ($\mathbf{E}_{miss}$) can easily degrade the performance of downstream tasks. Therefore, the above three error types need to be comprehensively evaluated.

Real-time detectors have received widespread attention and applications due to their excellent performance as well as superior efficiency. YOLOs are the cutting-edge CNN-based real-time detectors, thereby YOLOv8 [33] with SOTA performance is selected as a representative for comparative experiments. DETRs are Transformer-based cutting-edge end-to-end detection method, and representative DETR [12], Anchor-DETR [34], Conditional-DETR [35], and RT-DETR [30] are selected for comparison experiments. For fairness, all real-time detectors adopted the same training strategy, and end-to-end object detectors are only fine-tuned according to official recommendations. TABLE I shows the overall detection performance. FishViT achieves a maximum FPS of **82.3**, which is fully meets the real-time requirements for fish detection in water-land transfer. Attributed to the excellent multi-scale feature expression ability of the similarity-aware multi-level encoder and the high similarity-aware capability of STAttention, FishViT yields the best $\mathbf{AP}_{50}$ (**94.7**). Meanwhile, FishViT reaches a minimum of **0.11** on the most concerned $\mathbf{E}_{cls}$, and a minimum of **1.65** and **0.28** on the $\mathbf{E}_{loc}$ and $\mathbf{E}_{miss}$, respectively. Figure 5 visualizes the results of FishViT and of other SOTA detectors, with FishViT demonstrating superior performance.

### *C. Evaluation on Attributes*

To exhaustively evaluate the performance of FishViT in fish detection, attribute-based detection accuracy evaluation experiments are conducted based on 30 sequences used for test. $\mathbf{AP}_{50}$ is used as the evaluation metric of the selected videos for each challenging attribute. As shown in Fig. 6, FishViT can consistently achieve satisfactory performance with optimal detection accuracy in all six challenging scenarios. Compared to other attributes, FishViT's benefits are most evident in the challenging scenario of dense. This is the primary factor considered in FishViT design, *i.e.*, to discriminate every single fish in the dense group with highly similar appearances. STAttention mechanism can effectively eliminate background noise from images, while precisely discerning the edge information of diverse yet similar fish. This mechanism aims to clarify every boundary of similar fish and, as a result, improves detection accuracy in challenging scenarios. Furthermore, multi-scale feature representation capability is enhanced by introducing similarity-aware multi-level encoder with parallel structure, which further improves the detection accuracy of similar fish.

TABLE II

ABLATION STUDY OF VARIOUS PARTS OF THE PROPOSED FISHVIT. $\Delta$ SYMBOLIZES THE IMPROVEMENT OVER THE BASELINE METHOD.

| Detecting Methods | $\mathbf{AP}_{50}$ | $\Delta_{AP_{50}}$ | $\mathbf{E}_{cls}$ | $\Delta_{E_{cls}}$ |
|---|---|---|---|---|
| Baseline | 89.1 | - | 1.42 | - |
| Baseline+ST | 92.6 | +3.5 | 0.30 | -1.12 |
| Baseline+ML | 93.2 | +4.1 | 0.23 | -1.19 |
| **Baseline+ST+ML (FishViT)** | **94.7** | **+5.6** | **0.11** | **-1.31** |

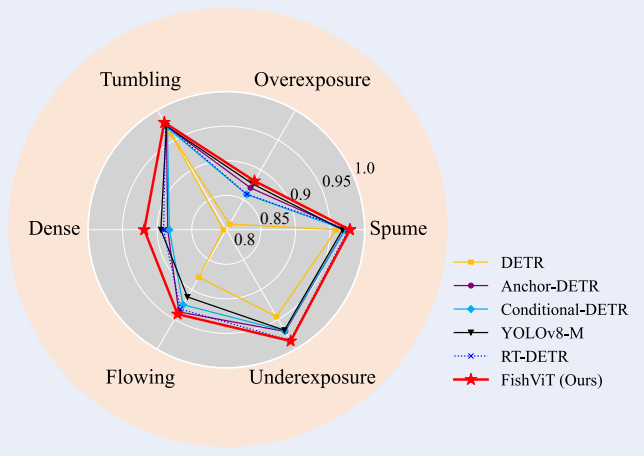


Fig. 6. The detection accuracy of FishViT and other SOTA detectors under six challenging attributes. Our FishViT achieves the best performance across all six challenging attributes. Best viewed in color.

### D. Ablation Study

To verify the effectiveness of each module in the proposed method, FishViT with different modules enabled is studied. This work considers Baseline as the model with only factorized attention encoder [18]. ST represents the STAttention and ML represents the parallel structure in similarity-aware multi-level encoder. The most important error in this fish detection task is the classification error, as it is directly linked to the revenue. Therefore, the ablation experiments used only $\mathbf{AP}_{50}$ and classification error as evaluation metrics.

**Discussion on STAttention:** As shown in TABLE II, with the addition of the soft-threshold (Baseline+ST), $\mathbf{AP}_{50}$ directly increased by **3.5**, indicating that STAttention can effectively improve the accuracy of the model. In addition, the classification error directly decreased by **1.12**, indicating that STAttention is able to remove the noise interference to a great extent and improve the classification accuracy under the condition of highly similar fish.

**Discussion on Similarity-aware Multi-level Encoder:** As shown in TABLE II, adding two extra Baseline branches (Baseline+ML) increases $\mathbf{AP}_{50}$ by **4.1**, demonstrating the multi-branch structure's positive impact on detection accuracy. Additionally, the classification error decreases by **1.19**, suggesting that the parallel structures help the model better capture edge information and improve classification accuracy by integrating multi-scale semantic data.

By combining ST and ML, FishViT (Baseline+ST+ML) efficiently learns the edge information while further enhancing the fish edge discrimination through the noise reduction effect of STAttention, which makes the $\mathbf{AP}_{50}$ rises by **5.6** and reducing classification error by **1.31**, fully demonstrating the strong robustness and high accuracy of FishViT.

### E. Real Work Scenario Test

The practicability of FishViT is further validated in real water-land transfer work scenario. FishViT achieves an average speed of over 80 FPS during the test, meeting real-time requirements. The unique software interface for fish detection and two real-world test sequences are shown in Fig. 7. Real-time images will be displayed in the interface's main window. A series of functions such as detection and recording can be achieved by clicking the button on the right, which is simple but efficient. The two tests contain several typical fish detection challenges, including dense and flowing. Test 1 demonstrates that high accuracy detection is maintained even as fish flipping and water velocity affect detection performance. In Test 2, FishViT encounters challenging, complex scenes. Nevertheless, FishViT manages to get remarkably elevated detection accuracy, which can be ascribed to STAttention's outstanding ability to identify similar objects. In conclusion, FishViT can accurately identify every fish in complex and challenging scenarios, which is incredibly convenient for real-world fish water-land transfer.

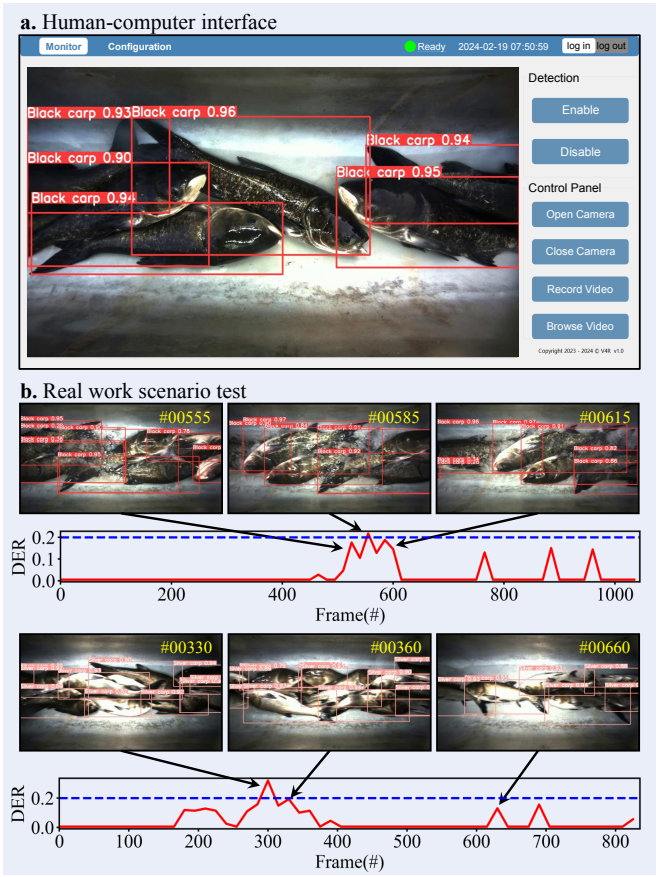


Fig. 7. Visualization of the real-world scenario test. **a.** The specially designed fish detection software interface. **b.** The detection results are marked with red boxes. The detection error rate (DER) is defined as $1 - \mathrm{AP}_{50}$ per frame. A DER score below the blue dotted line is considered accurate and reliable detection in the real-world test. Best viewed in color.

## V. Conclusion

A new type of lightweight, low-power, environmentally independent, and high-speed intelligent detection system deployed with the proposed FishViT is designed to automatically conduct high-speed fish detection in this work. The objective is to solve the problem of low efficiency and high cost of fish detection in traditional crowd-collaborative water-land transfer mode. To cope with the high similarity, high

speed, and high density problems of fish water-land transfer, a novel real-time end-to-end detector for fish detection is proposed. Additionally, STAttention with soft-threshold and similarity-aware multi-level encoder with parallel structure are presented. Extensive experiments prove that FishViT is capable of detecting fish at high speeds while delivering outstanding performance. FishViT will significantly enhance subsequent critical tasks in the field of water-land transfer, including the tracking, segmentation, sizing, and counting of fish, providing robust academic support for these processes. In conclusion, we firmly believe that the intelligent vision system with FishViT can aid in the advancement of fish detection in water-land transfer.

## ACKNOWLEDGMENTS

This work is supported by Hangzhou Qiandao Lake Development Group Co. LTD and Fishery Machinery and Instrument Research Institute, Chinese Academy of Fishery Sciences.

## REFERENCES

[1] S. Zhong, M. Crang, and G. Zeng, "Constructing Freshness: the Vitality of Wet Markets in Urban China," *Agriculture and Human Values*, vol. 37, no. 1, pp. 175–185, 2020.

[2] K. Thakur, M. Shetty, S. Singh, and J. Khanapuri, "Enhancing Fish Disease Detection: A Comprehensive Review for Sustainable Aquaculture," in *2023 6th International Conference on Advances in Science and Technology (ICAST)*. IEEE, 2023, pp. 191–196.

[3] D. An, J. Hao, Y. Wei, Y. Wang, and X. Yu, "Application of Computer Vision in Fish Intelligent Feeding System—A Review," *Aquaculture Research*, vol. 52, no. 2, pp. 423–437, 2021.

[4] C. Wang, Z. Li, T. Wang, X. Xu, X. Zhang, and D. Li, "Intelligent Fish Farm—the Future of Aquaculture," *Aquaculture International*, pp. 1–31, 2021.

[5] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[6] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6154–6162.

[7] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR Training by Introducing Query DeNoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 619–13 627.

[8] A. B. Amjoud and M. Amrouch, "Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review," *IEEE Access*, vol. 11, pp. 35 479–35 516, 2023.

[9] M. Park, W. Yang, Z. Cao, B. Kang, D. Connor, and M.-A. Lea, "Marine Vertebrate Predator Detection and Recognition in Underwater Videos by Region Convolutional Neural Network," in *Proceedings of Knowledge Management and Acquisition for Intelligent Systems (PKAW)*, 2019, pp. 66–80.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 28, 2015, pp. 1–9.

[11] K. M. Knausgård, A. Wiklund, T. K. Sørdalen, K. T. Halvorsen, A. R. Kleiven, L. Jiao, and M. Goodwin, "Temperate Fish Detection and Classification: a Deep Learning based Approach," *Applied Intelligence*, vol. 52, no. 6, pp. 6988–7001, 2022.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[13] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection," in *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2022, pp. 1–19.

[14] K. Li, D. Wang, X. Wang, G. Liu, Z. Wu, and Q. Wang, "Mixing Self-Attention and Convolution: A Unified Framework for Multi-source Remote Sensing Data Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[15] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are Rnns: Fast Autoregressive Transformers with Linear Attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 5156–5165.

[16] K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser *et al.*, "Rethinking Attention with Performers," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020, pp. 1–38.

[17] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 256–17 267.

[18] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-Scale Conv-Attentional Image Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9981–9990.

[19] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep Residual Shrinkage Networks for Fault Diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4681–4690, 2019.

[20] D. L. Donoho, "De-Noising by Soft-Thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[21] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 781–10 790.

[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.

[23] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020, pp. 1–16.

[24] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast Convergence of DETR With Spatially Modulated Co-Attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3621–3630.

[25] X. Cao, P. Yuan, B. Feng, and K. Niu, "CF-DETR: Coarse-to-Fine Transformers for End-to-End Object Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022, pp. 185–193.

[26] G. Zhang, Z. Luo, Z. Tian, J. Zhang, X. Zhang, and S. Lu, "Towards Efficient Use of Multi-Scale Features in Transformer-Based Object Detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6206–6216.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] Z. Qin, W. Sun, H. Deng, D. Li, Y. Wei, B. Lv, J. Yan, L. Kong, and Y. Zhong, "Cosformer: Rethinking Softmax in Attention," *arXiv preprint arXiv:2202.08791*, pp. 1–15, 2022.

[29] K. Isogawa, T. Ida, T. Shiodera, and T. Takeguchi, "Deep Shrinkage Convolutional Neural Network for Adaptive Noise Reduction," *IEEE Signal Processing Letters (SPL)*, vol. 25, no. 2, pp. 224–228, 2017.

[30] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "DETRs Beat YOLOs on Real-time Object Detection," *arXiv preprint arXiv:2304.08069*, pp. 1–11, 2023.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[32] D. Bolya, S. Foley, J. Hays, and J. Hoffman, "Tide: A General Toolbox for Identifying Object Detection Errors," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020, pp. 558–573.

[33] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[34] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor Detr: Query Design for Transformer-Based Detector," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022, pp. 2567–2575.

[35] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional Detr for Fast Training Convergence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3651–3660.