KP-Mask2Former: A fish occlusion contour recovery method based on KANsformer and PCBAM

Yunfeng Wang[1], Beibei Li[2], Jun Yue[1*], Yixun Zhao[1], Lei Yuan[1], Peiheng Li[1], Guangjie Kou[1]

(1. *College of Computer Science and Artificial Intelligence, Ludong University, Yantai 264025, Shandong, China*;
2. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*)

Abstract: Recovering complete contours of occluded fish is crucial for accurate fish size estimation and automated counting. Nevertheless, frequent and severe occlusions in high-density aquaculture environments make reliable contour recovery particularly challenging. To address this problem, this study proposes KP-Mask2Former, a fish contour recovery framework specifically designed for occlusion-intensive scenarios. First, a Parallel Convolutional Block Attention Module (PCBAM) is introduced into the patch-merging stages of the backbone to strengthen the representation of fish contour features. Second, Swin KANsformer is incorporated into the backbone to enhance the reconstruction of incomplete boundaries caused by occlusion. Third, a SEAM-Mask Head is developed to suppress interference from occluding instances and improve the recovery of complete fish contours. The proposed framework is evaluated on a self-constructed Occluded *Oplegnathus punctatus* Dataset (OOD). Experimental results demonstrate that KP-Mask2Former achieves an AP of 87.1%, representing a 4.3-percentage-point improvement over the baseline. It also outperforms representative state-of-the-art methods, including PolySnake, E2EC, PatchDCT, BCNet, Transfiner, MP-Former, and FastInst. Further analyses confirm the superiority of the proposed method in boundary delineation, intra-class occlusion handling, and multi-scale instance segmentation.

Keywords: fish, instance segmentation, occlusion recovery, attention mechanism, deep learning
DOI: 10.25165/j.ijabe.20261904.10285

Citation: Wang Y F, Li B B, Yue J, Zhao Y X, Yuan L, Li P H, et al. KP-Mask2Former: A fish occlusion contour recovery method based on KANsformer and PCBAM. Int J Agric & Biol Eng, 2026; 19(4): 235–242.

1 Introduction

Accurate recovery of fish contours is essential for reliable phenotypic measurement and automated counting[1]. Complete object boundaries provide fundamental morphological information, including body length and width, which supports biomass estimation and precision feeding. Inaccurate contour reconstruction can therefore introduce substantial errors into morphological measurements and further compromise downstream applications such as biomass estimation and feeding management. In high-density aquaculture environments, fish frequently overlap with conspecific individuals, resulting in intra-class occlusion and incomplete visible contours. Conventional fish measurement and counting methods typically rely on manual observation or annotation, making them labor-intensive, inefficient, and susceptible to subjective bias. In recent years, computer vision methods have demonstrated considerable potential for recovering the complete shapes of occluded objects because of their non-contact, objective and efficient characteristics. Nevertheless, existing occlusion-handling methods are generally designed for generic objects and rarely consider the complex appearance, deformable boundaries, and severe intra-class occlusion of fish.

Current De-occlusion methods can generally be divided into two categories: two-stage methods and one-stage methods. Two-stage methods typically first detect or localize the occluded regions and subsequently reconstruct the missing content using a separate recovery module. For example, Guo et al.[2] proposed a computer vision–based method that detects target regions, identifies and removes occluding devices by combining the Hough transform with color features, and reconstructs the missing contours using linear interpolation and ellipse fitting, achieving a recovery accuracy of over 80%. Huang et al.[3] introduced CClusnet-Inseg, a centroid clustering network for instance segmentation of mutually occluded piglets in farrowing pens, in which pixel-level centroid prediction and clustering are employed to distinguish individual instances. Hao et al.[4] developed the Hand De-occlusion and Removal (HDR) method for hand de-obscuration, which alleviates severe hand occlusion and appearance ambiguity in human–object interaction scenarios. Ju et al.[5] presented Swap-R&R, a self-supervised face recovery method that reconstructs unobstructed frontal facial images from occluded profile views. Yin et al.[6] proposed a face de-obscuration model that combines facial segmentation with 3D face reconstruction and recovers missing facial textures through the segmentation module and inpainting module. Similarly, Zhou et al.[7] developed a two-stage framework that first refines visible-instance annotations and then predicts complete-object annotations, thereby improving reconstruction accuracy under complex occlusion conditions. Despite their effectiveness, two-stage methods depend heavily on the accuracy of intermediate predictions. Errors produced during occlusion localization or visible-region segmentation may propagate to the subsequent reconstruction stage, thereby degrading the quality of the final recovered contours.

Received date: 2025-10-23 **Accepted date:** 2026-07-15
Biographies: Yunfeng Wang, MS, research interest: artificial intelligence, Email: wangyunfeng823@163.com; **Beibei Li**, PhD, Researcher, research interest: computer vision, Email: lbbflx@163.com; **Yixun Zhao**, MS candidate, research interest: artificial intelligence, Email: arthur121381@outlook.com; **Lei Yuan**, MS candidate, research interest: artificial intelligence, Email: 2011469323@qq.com; **Peiheng Li**, MS candidate, research interest: artificial intelligence, Email: lphworkmail@163.com; **Guangjie Kou**, PhD, Professor, research interest: artificial intelligence, Email: kouguangjie@126.com.
***Corresponding author: Jun Yue**, PhD, Professor, research interest: artificial intelligence technology. College of Computer Science and Artificial Intelligence, Ludong University, Yantai 264025, Shandong, China. Tel: +86-535-6681245, Email: yuejuncn@sohu.com.

In contrast, single-stage methods employ a unified network to simultaneously localize occluded regions and reconstruct the missing content. Li et al.[8] proposed Mask-FPAN for high-precision parsing of occluded faces in real-world scenes. By incorporating a semi-supervised Occ-Autoencoder de-occlusion module, the method achieved an mIoU of 90.1%. Zhan et al.[9] introduced PCNets, a single-stage self-supervised model for scene de-obscuration that separates and reconstructs foreground objects in complex scenes, achieving an IoU of 87.1% on the COCOA dataset. By jointly performing occlusion localization and restoration, single-stage methods eliminate intermediate operations such as bounding-box regression and feature alignment, thereby enabling faster inference. However, their capability to handle complex occlusion patterns remains limited. Therefore, generic contour recovery methods provide useful insights into recovering occluded fish contours, particularly for addressing missing regions and incomplete boundaries. Nevertheless, fish-specific characteristics introduce additional challenges. Non-rigid body deformation[10] and motion blur caused by rapid swimming further increase the difficulty of contour recovery. Moreover, fish of the same species in high-density aquaculture tanks often exhibit highly similar textures, colors, and body sizes. Under severe mutual occlusion, low-level boundary cues and high-level semantic features become spatially entangled. Existing generic instance segmentation networks often lack sufficient local discriminative capability to separate adjacent individuals under such visually similar and dynamically blurred conditions. Consequently, features from different fish may be incorrectly associated, resulting in merged boundaries, inaccurate contours, and loss of fine-grained details.

Although existing generic segmentation networks have achieved substantial progress on standard public datasets, they remain limited in handling non-rigid deformation and occlusion-induced topological ambiguity in complex underwater aquaculture environments. These limitations often result in boundary misalignment and confusion between adjacent fish instances. To address these challenges, this study proposes KP-Mask2Former, an occluded fish contour reconstruction framework integrating PCBAM, Swin KANsformer, and a SEAM-enhanced mask head. The main contributions are summarized as follows:

(1) PCBAM is proposed to enhance fish contour feature extraction under occlusion conditions. Unlike the conventional Convolutional Block Attention Module (CBAM), which sequentially applies channel and spatial attention, PCBAM employs parallel attention branches and weighted feature fusion to alleviate feature attenuation caused by cascaded attention operations.

(2) Swin KANsformer is constructed by replacing the multilayer perceptron (MLP) layers in Swin Transformer with Kolmogorov–Arnold Network (KAN) layers. KAN employs learnable nonlinear functions on network edges, enabling more flexible modeling of irregular boundary variations caused by occlusion and non-rigid fish deformation.

(3) A SEAM-enhanced mask head is introduced to improve spatial alignment during mask prediction. By suppressing feature misalignment and local feature aliasing caused by intra-class occlusion, the proposed mask head improves the accuracy of occluded fish contour reconstruction.

2 Materials and methods

2.1 Data preparation

2.1.1 Data collection

The OOD was collected in South Workshop No. 8 at Shandong Laizhou Mingbo Aquatic Co., Ltd. Microalgae were added to the tank to simulate practical seawater aquaculture conditions, and thirty *Oplegnathus punctatus* individuals were placed in the D10 aquaculture tank. The tank measured 3.3 m in diameter and 0.64 m in height, with a water depth of 0.60 m. The water temperature was maintained between 21°C and 24°C. Image data were captured using a stereo underwater camera ZF-USB-02B10 (Figure 1).

Figure 1 Schematic of the data acquisition platform for the *Oplegnathus punctatus* occlusion dataset

2.1.2 Data annotation

The original side-by-side stereo video (3840×1080, 10.79 fps) was processed using PotPlayer to extract 10 000 frames, from which 500 unobstructed images were selected. Each stereo image was then split down the middle into left and right views, yielding 1000 single-view JPG images at 1920×1080 resolution. The semi-automatic annotation tool ISAT-SAM was then used to accurately delineate the contour of each *Oplegnathus punctatus* instance, providing the source images and annotations required for subsequent occluded-data synthesis. After annotation, ISAT-SAM automatically generated JSON files containing the file name, storage path, image dimensions, category labels, and contour coordinates.

2.1.3 Data augmentation

To simulate intra-class occlusion among *Oplegnathus punctatus* individuals and construct a synthetic occlusion dataset, a Copy-Paste procedure[11] was implemented using 1000 original unoccluded images. Specifically, source images (I_{src}) and main target images (I_{main}) were randomly selected and subjected to augmentations including flipping, cropping, scale jittering, and padding. Instances from I_{src} and their annotations ($mask_{src}$) were then blended onto I_{main}, $mask_{src} = I_{main} \times \alpha + I_{src} \times (1-\alpha)$, where α is a weight coefficient controlling image blending. Out of 6000 generated candidates, 5800 images exhibiting distinct occlusions were retained, formatted in COCO style, and split into training and test sets (8:2 ratio) to construct the OOD dataset. The generation pipeline and resulting annotations are illustrated in Figure 2.

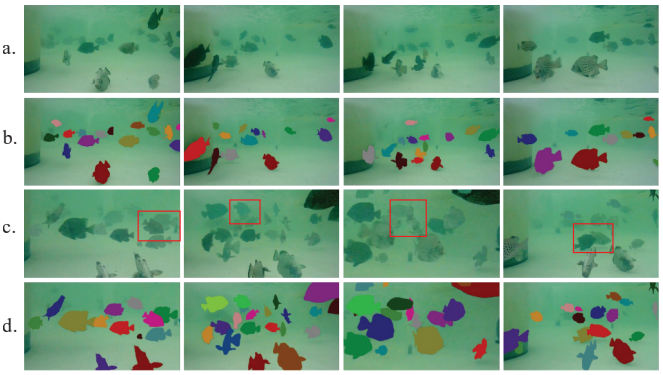


Figure 2 Visualization of the synthetic generation process for partially occluded *Oplegnathus punctatus* images

2.2 KP-Mask2Former

In this study, KP-Mask2Former is proposed as an extension of the Mask2Former instance segmentation framework for occluded fish contour reconstruction. Its overall architecture is illustrated in Figure 3. Mask2Former[12] is a general-purpose segmentation framework with a simple meta-architecture consisting of a backbone, a pixel decoder[13], and a transformer decoder[14]. The backbone can adopt either a CNN-based architecture, such as ResNet[15] or a Transformer-based architecture[16], while the pixel decoder is typically implemented using multi-scale deformable attention[13].

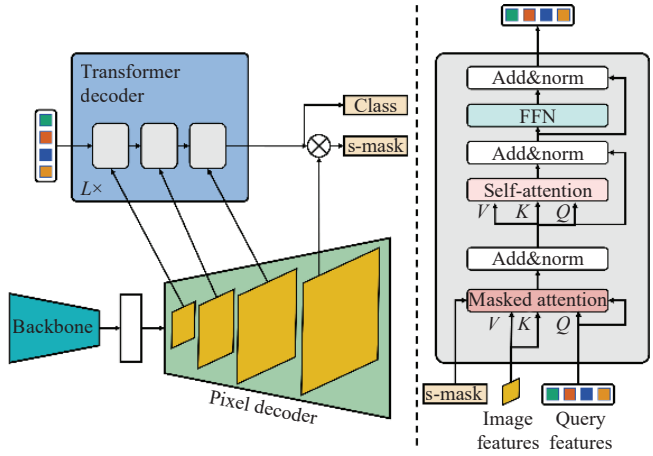


Figure 3 Overall architecture of KP-Mask2Former

In KP-Mask2Former, the input image is first processed by the backbone network to extract target features. The resulting low-resolution features are then passed to the pixel decoder, which upsamples them to generate high-resolution features. The transformer decoder subsequently employs these high-resolution features for object query processing, and the refined segmentation head performs pixel-wise decoding. The head predicts binary annotations to differentiate between instances.

2.2.1 PCBAM attention mechanism

CBAM[17] is a hybrid attention mechanism incorporating both channel attention[18] and spatial attention. Its detailed architecture is illustrated in Figure 4. The CBAM model first receives the feature map F1 weighted by channel attention; subsequently, it outputs the feature map F2 weighted by spatial attention, forming a "cascade connection" structure. The process is formulated as follows:

$$\begin{cases} F_1 = M_C(F) \otimes F \\ F_2 = M_S(F_1) \otimes F_1 \end{cases} \tag{1}$$

where, $M_C(F)$ denotes the output weighting of F obtained via channel attention, while $M_S(F_1)$ represents the output weighting of F derived through spatial attention; ⊗ Represents the feature map weighted multiplication operator CBAM attention uses a simple "cascade connection", where the weights at the back of the line are generated from the feature map at the front of the line. The back-ranked attentional feature inputs are partially shaped by the preceding attentional mechanisms, potentially leading to interference and instability within the model.

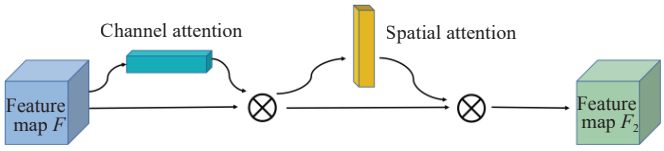


Figure 4 Architecture of the CBAM attention module

Although CBAM effectively captures global contextual information, occluded fish contour reconstruction requires greater attention to locally discriminative regions and incomplete boundary cues. To address this limitation, PCBAM is proposed to enhance the representation of critical object features. PCBAM replaces the sequential attention structure of CBAM with a parallel attention mechanism, thereby reducing feature attenuation and strengthening the representation of important contour regions.

PCBAM adopts a parallel weighting strategy to circumvent the issue of information attenuation inherent in sequential processing, as shown in Figure 5. Specifically, PCBAM performs global max pooling and average pooling along both the channel and spatial dimensions independently. These pooled outputs are mapped by their respective MLPs and subsequently aggregated via element-wise addition to yield the initial channel and spatial descriptors. Following this, the two sets of descriptors are concatenated to capture interactive relationships and fed into a shared MLP to jointly model the cross-correlation between channel and spatial features. Finally, the joint features are decoupled back into separate channel and spatial pathways, where they pass through Sigmoid activation functions to generate their corresponding attention weights. These weights are then element-wise multiplied with the original input feature map, ensuring that the final fused features simultaneously possess channel discriminativeness and spatial structural integrity.

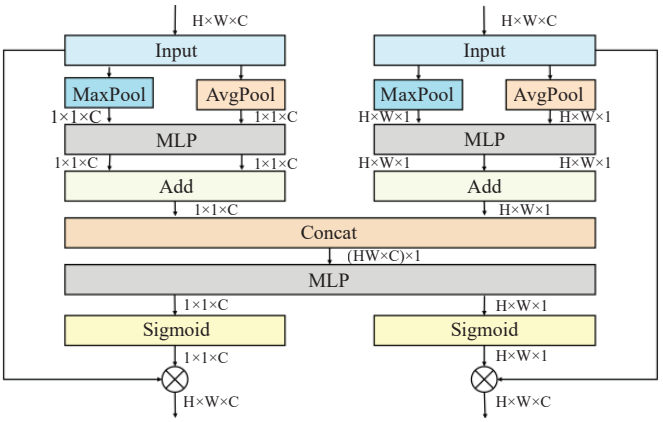


Figure 5 Architecture of the proposed PCBAM attention module

2.2.2 Swin Transformer

Mask2Former employs Swin Transformer[16] as its backbone. Swin Transformer is a hierarchical vision Transformer that extracts multi-scale features and captures long-range dependencies through a shifted-window mechanism. It restricts self-attention computation to non-overlapping local windows and periodically shifts the window partition to enable cross-window information exchange while reducing computational complexity.

The Swin Transformer module in the backbone comprises three core components: window-based multi-head self-attention (W-MSA), shifted window multi-head self-attention (SW-MSA), and an MLP. Layer normalization is applied before each attention and MLP module, followed by residual connections. At the end of each stage, a patch-merging module reduces the number of tokens and generates hierarchical feature representations.

The MLP layers, also known as fully connected feedforward neural networks, are commonly used in machine learning for approximating nonlinear functions. An MLP consisting of K layers can be described as a transformation matrix W and the activation function σ interactions. This can be expressed mathematically as:

$$\mathrm{MLP}(Z) = (W_{K-1} \circ \sigma \circ W_{K-2} \circ \sigma \circ \ldots \circ W_1 \circ \sigma \circ W_0)Z \quad (2)$$

MLP aim at approximating complex functional mappings through sequences of multi-layer nonlinear transformations. Despite their widespread use, dense MLP layers may introduce substantial parameter and computational costs when processing high-dimensional image features, limiting their ability to flexibly model irregular occluded fish boundaries. To improve nonlinear feature modeling, Liu et al.[19] proposed the KAN as an alternative to conventional MLPs. While MLPs are inspired by the universal approximation theorem, KAN is based on the Kolmogorov–Arnold representation theorem. KAN retains a fully connected architecture but replaces the fixed node-based activation functions and linear weights of MLPs with learnable univariate functions on network edges.

$$\mathrm{KAN}(Z) = (\Phi_{K-1} \circ \Phi_{K-2} \circ \ldots \circ \Phi_1 \circ \Phi_0)Z \quad (3)$$

where, Φ_i denotes the i-th layer of the entire KAN network. Each KAN layer has n_{in} inputs and n_{out} outputs, including $n_{in} \times n_{out}$ learnable activation functions ϕ, as shown in Equation (4):

$$\Phi = \{\phi_{q,p}\}, \quad p = 1,2,\ldots,n_{in}, \quad q = 1,2,\ldots,n_{out} \quad (4)$$

The results of the KAN network from k layer to $k+1$ layer can be represented in matrix form, as shown in Equation (5):

$$Z_{k+1} = \underbrace{\begin{pmatrix} \phi_{k,1,1}(\cdot) & \phi_{k,1,2}(\cdot) & \cdots & \phi_{k,1,n_k}(\cdot) \\ \phi_{k,2,1}(\cdot) & \phi_{k,2,2}(\cdot) & \cdots & \phi_{k,2,n_k}(\cdot) \\ \vdots & \vdots & & \vdots \\ \phi_{k,n_{k+1},1}(\cdot) & \phi_{k,n_{k+1},1}(\cdot) & \cdots & \phi_{k,n_{k+1},n_k}(\cdot) \end{pmatrix}}_{\Phi_k} Z_k \quad (5)$$

To better model non-rigid fish deformation and occlusion-induced topological ambiguity, the original MLP layers in Swin Transformer are replaced with KAN layers to construct Swin KANsformer. This modification is designed to improve nonlinear feature modeling for fragmented boundary reconstruction rather than serving as a simple component substitution. The architecture of Swin KANsformer is shown in Figure 6. It alternates between W-MSA and SW-MSA blocks, each incorporating layer normalization, a KAN module, and residual connections. Layer normalization is applied before each submodule, followed by residual feature aggregation. This design enhances the representation of irregular fish boundaries and improves occluded contour reconstruction.

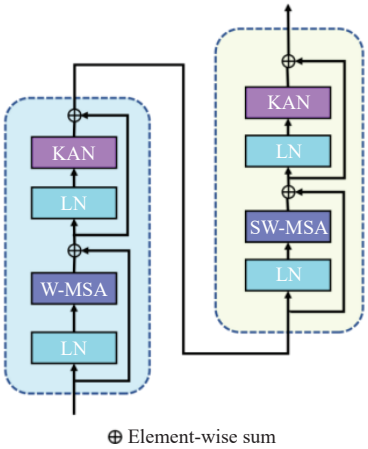


Figure 6 Architecture of the Swin KANsformer module

Building on Swin KANsformer and PCBAM, this study constructs the Swin-s_KP backbone. In this backbone, the original Swin Transformer blocks are replaced with Swin KANsformer blocks, while PCBAM is embedded into the Stage 2 patch-merging module to enhance feature representation. The architecture of Swin-s_KP is shown in Figure 7. In Stage 1, the input image is partitioned into non-overlapping patches and projected into feature embeddings through a linear embedding layer. The resulting features are then processed by Swin KANsformer blocks for initial feature extraction. In Stage 2, PCBAM is incorporated into the patch-merging module to adaptively refine the features while integrating local and global information, thereby improving the representation of occluded fish regions.

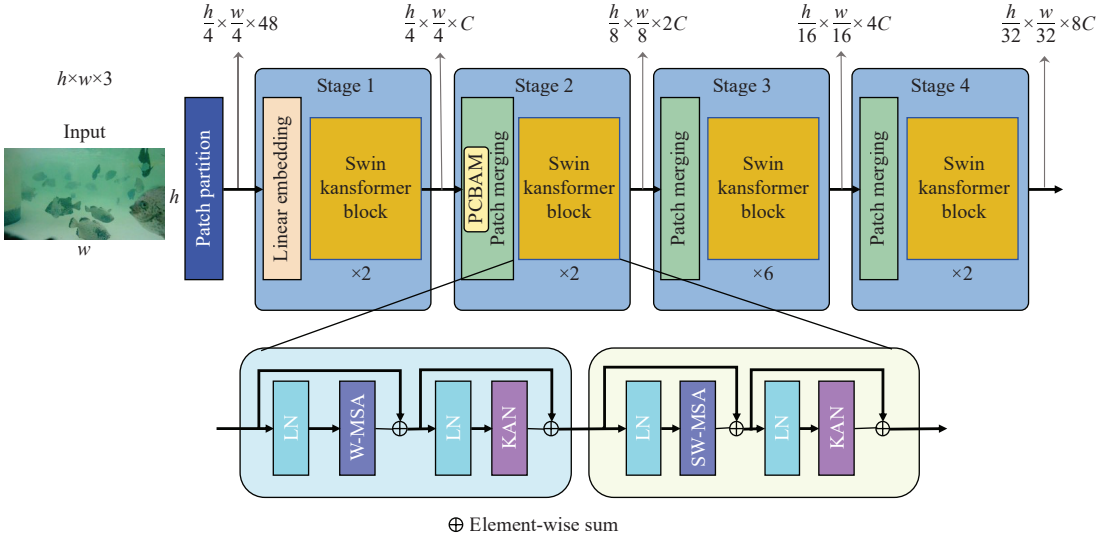


Figure 7 Architecture of the Swin-s_KP backbone

2.2.3 SEAM-Mask head

Spatially Enhanced Attention Module (SEAM) is an attention module designed to enhance feature representation under occlusion, and its architecture is shown in Figure 8. It consists of depthwise separable convolutions and residual connections[20]. First, pointwise convolutions fuse the outputs of different depthwise convolution branches. A two-layer fully connected network then integrates channel information to strengthen inter-channel dependencies. By exploiting the relationships between occluded and visible object features, SEAM compensates for feature loss caused by occlusion.

The normalized channel responses are subsequently mapped through an exponential function, expanding their range from [0, 1] to [1, e]. This monotonic mapping increases tolerance to localization errors. Finally, the generated attention weights are multiplied by the original features to enhance the representation and reconstruction of occluded fish contours.

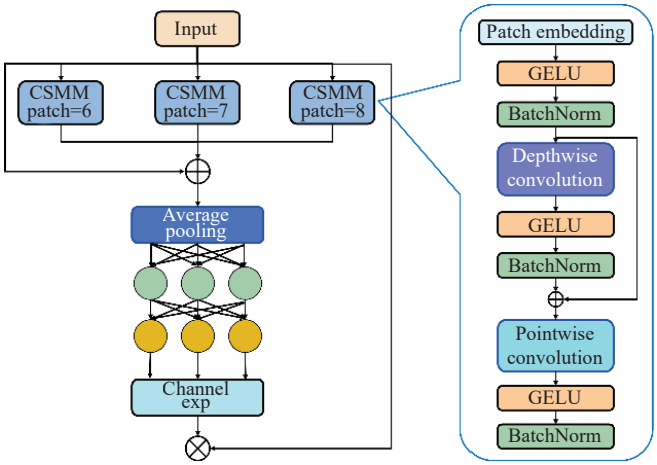


Figure 8 Architecture of the SEAM attention module

To alleviate the boundary ambiguity caused by intra-class occlusion, SEAM[21] is incorporated into the mask head of KP-Mask2Former to construct the SEAM-Mask Head. This module is designed to support multi-scale fish segmentation, enhance foreground feature representation, and suppress background interference.

3 Experimental results and analysis

3.1 Experimental environment and evaluation metrics

To evaluate the proposed method, a series of qualitative and quantitative comparative experiments were conducted. The hardware platform consisted of an Intel(R) Xeon(R) Gold 5118 CPU and three NVIDIA TITAN V GPUs. The models were implemented in PyTorch with CUDA 11.2 on CentOS 7.

During training, the initial learning rate and weight decay were set to 0.0001 and 0.0005, respectively. The momentum was set to 0.937, the batch size to 3, and the maximum number of iterations to 10000. To satisfy the 32× downsampling divisibility constraint of the Swin Transformer backbone and preserve fine features of occluded small fish for improved small-instance segmentation, raw images (1920×1080) were resized to 2048×1536 through proportional scaling and zero-padding. Experiments were conducted on the self-constructed OOD dataset containing 5800 images and the additional Fish Occlusion Dataset (FOD) containing 14 376 images.

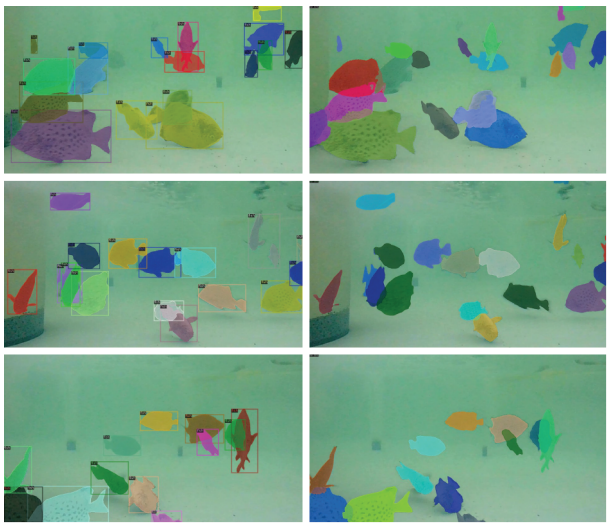
COCO evaluation metrics were adopted to assess fish instance segmentation performance, include Average Precision (AP), AP^{50}, AP^{75}, AP^{S}, AP^{M}, and AP^{L}, where IoU denotes the ratio of the intersection area to the union area between the predicted annotation and the ground truth annotation, as summarized in Table 1.

3.2 Qualitative segmentation

Figure 9 presents representative segmentation results of KP-Mask2Former. The first column shows the ground-truth annotations, while the second column presents the corresponding model predictions on the test set. The results show that KP-Mask2Former accurately reconstructs the contours of occluded fish across different object scales and preserves clear boundaries in occluded regions, demonstrating the effectiveness of the proposed method.

Table 1 Performance indicators for fish instance segmentation

Indicator	Description
AP	AP is the average precision value at 10 IoU thresholds from IoU = 0.5 to 0.95 (in steps of 0.05)
AP^{50}	Average precision at an IoU threshold of 0.5
AP^{75}	Average precision at an IoU threshold of 0.75
AP^{S}	Average precision for small targets (area ≤ 32^2 pixels) in the IoU = 0.5:0.05:0.95 range
AP^{M}	Average precision for medium targets (32^2 < area < 96^2 pixels) in the IoU = 0.5:0.05:0.95 range
AP^{L}	Average precision for large targets (area ≥ 96^2 pixels) in the IoU = 0.5:0.05:0.95 range



Ground truth Prediction

Figure 9 Segmentation results of the proposed model

3.3 Ablation experiment

In this section, ablation experiments were conducted to evaluate the contributions of the three proposed components to KP-Mask2Former. All experiments were performed on the OOD dataset using identical hyperparameter settings, and the results are reported in Table 2. Each component improves the model performance across all six evaluation metrics. Compared with the Mask2Former baseline, integrating PCBAM, further introducing Swin KANsformer, and finally adding SEAM produce cumulative AP improvements of 3.6, 3.7, and 4.3 percentage points, respectively.

Table 2 Ablation study of the proposed modules on instance segmentation performance

Base	P	S-K	S	AP↑	AP^{50}↑	AP^{75}↑	AP^{S}↑	AP^{M}↑	AP^{L}↑	Params (M)↓	FLOPs (G)↓
√				82.8	94.0	87.7	22.4	72.6	86.4	**62.2**	**78.3**
√	√			86.4	96.6	90.9	34.3	77.9	88.7	62.6	78.7
√	√	√		86.5	96.4	91.1	34.2	78.0	89.4	62.8	78.8
√	√	√	√	**87.1**	**96.9**	**91.4**	**36.7**	**78.2**	**89.5**	63.0	79.0

Note: Base, P, S-K, and S denote the Mask2Former baseline, PCBAM, Swin KANsformer, and SEAM modules, respectively.

The best performance is achieved when all three enhancements are integrated. Overall, the ablation results demonstrate that the proposed model effectively reconstructs occluded fish contours.

3.4 Comparison experiment

To validate the effectiveness of the proposed model, extensive experiments were conducted on the OOD dataset. KP-Mask2Former was compared with representative methods, including PolySnake[22], E2EC[23], PatchDCT[24], BCNet[25], Transfiner[26], MP-Former[27], and

FastInst[28]. The effects of different backbone networks on segmentation performance were also investigated. As reported in Table 3, KP-Mask2Former achieves an AP of 87.1% on the OOD dataset and outperforms all compared methods.

Table 3 Performance comparison of different methods on the OOD dataset

Model	Backbone	AP↑	AP^{50}↑	AP^{75}↑	AP^{S}↑	AP^{M}↑	AP^{L}↑
PolySnake	DLA-34	21.3	27.0	23.1	15.0	19.2	22.2
E2EC	DLA-34	45.3	63.6	47.2	0.1	49.4	45.3
BCNet	ResNet-50	68.8	89.4	75.9	0.384	37.0	71.5
Transfiner	ResNet-50	76.5	93.1	83.9	14.8	66.2	79.0
PatchDCT	ResNet-50	78.6	93.1	84.9	12.8	67.2	81.0
MP-Former	ResNet-50	79.9	93.8	85.5	17.1	68.1	82.9
FastInst	ResNet-50	80.5	95.3	87.6	21.7	68.8	83.5
Mask2Former	ResNet-50	81.8	94.3	86.7	25.2	71.8	84.4
Mask2Former	Swin-s	82.8	94.0	87.7	22.4	72.6	86.4
KP-Mask2Former	Swin-t	82.6	95.8	88.0	29.0	72.1	85.3
KP-Mask2Former	Swin-s	**87.1**	**96.9**	**91.4**	**36.7**	**78.2**	**89.5**

Compared with the Mask2Former baseline, KP-Mask2Former improves AP by 4.3 percentage points, mainly owing to the enhanced feature representation of the backbone and mask head under occlusion. PCBAM independently models channel and spatial attention and then performs weighted feature fusion, thereby preserving complementary information and improving contour representation. KP-Mask2Former also exceeds FastInst and MP-Former by 6.6 and 7.2 percentage points, respectively. These results demonstrate the effectiveness of the Swin-s-based backbone in extracting discriminative features for occluded instances. Notably, E2EC and BCNet struggle on tiny instances, yielding extremely low AP^{S} scores (0.10 and 0.38, respectively) due to the lack of multi-scale feature enhancement under severe occlusion. In contrast, KP-Mask2Former leverages multi-scale attention and occlusion-aware modules to retain fine boundary details, achieving a superior AP^{S} of 36.7%. Overall, KP-Mask2Former provides an effective solution for segmenting mutually occluded fish with highly similar color and texture characteristics. Table 4 lists the segmentation performance of Mask2Former with different backbone networks.

In general, deeper backbones achieve better performance than shallower ones. Swin-s_KP and ResNet-101 obtain AP values of 87.1% and 85.1%, respectively. ResNet-101 outperforms ResNet-50 by 3.3 percentage points, while Swin-s exceeds Swin-t by 0.2 percentage points. These results also show that the Transformer-based Swin backbones achieve better segmentation performance than the corresponding residual networks. Specifically, Swin-t and Swin-s outperform ResNet-50 by 0.8 and 1.0 percentage points, respectively.

Table 4 Performance comparison of Mask2Former with different backbone networks

Backbone	AP↑	AP^{50}↑	AP^{75}↑	AP^{S}↑	AP^{M}↑	AP^{L}↑
ResNet-50	81.8	94.3	86.7	25.2	71.8	84.4
ResNet-101	85.1	96.2	90.1	23.5	74.9	87.7
Swin-t	82.6	95.8	88.0	29.0	72.1	85.3
Swin-s	82.8	94.0	87.7	22.4	72.6	86.4
Swin-s_KP	**87.1**	**96.9**	**91.4**	**36.7**	**78.2**	**89.5**

Figure 10 compares representative segmentation results for occluded *Oplegnathus punctatus* instances. Figure 10a shows the corresponding ground-truth annotations. In Figure 10b, FastInst fails to accurately recover some occluded regions, particularly in densely overlapping areas. Figure 10c shows that MP-Former incorrectly identifies some visible regions as occluded. In Figure 10d, PatchDCT produces blurred boundaries in the reconstructed contours. Figure 10e shows that Transfiner yields inaccurate boundary predictions. By contrast, KP-Mask2Former produces clear contours and segmentation results that closely agree with the ground truth, as shown in Figure 10f. Overall, KP-Mask2Former reconstructs occluded contours more accurately than the competing methods and preserves clearer object boundaries, further demonstrating the effectiveness of the proposed method.

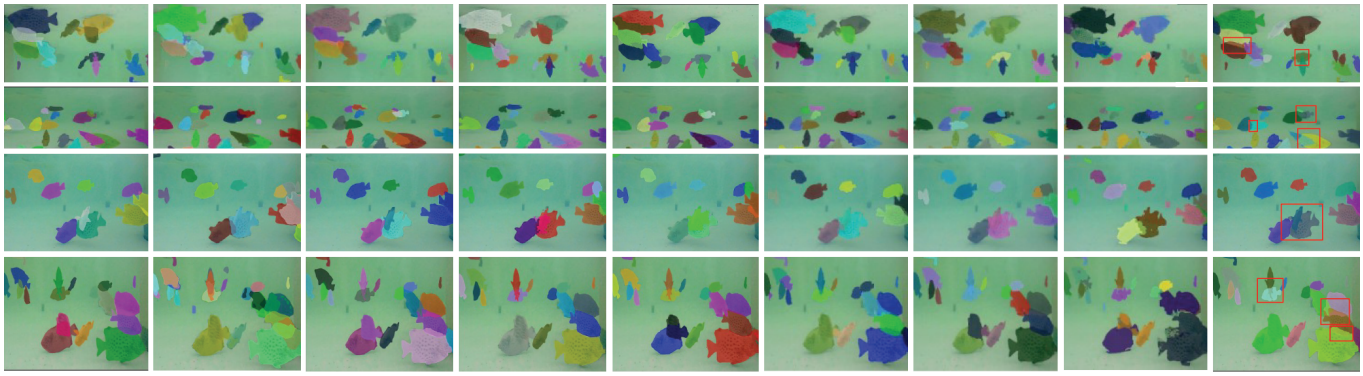


Figure 10 Segmentation results for occluded fish instances obtained by different methods

To evaluate the generalization capability of the proposed method, comparative experiments are conducted on the additional FOD dataset[29], with the results reported in Table 5. The FOD is a large-scale image dataset designed for training and evaluating instance segmentation models under challenging underwater occlusion conditions. It contains 14376 underwater images and 144894 pixel-level fish-instance annotations. Unlike the OOD dataset, FOD uses crucian carp as the target species and divides the samples into three categories according to occlusion severity: complete, partially occluded, and fragmented. In addition, the dataset incorporates realistic aquaculture disturbances, such as net-cage interference, enabling a more comprehensive evaluation of the generalization performance of different segmentation methods.

Table 5 shows that KP-Mask2Former outperforms Transfiner, PatchDCT, MP-Former, and FastInst across all major evaluation metrics. Specifically, it achieves an overall AP of 80.2%, exceeding

Transfiner and MP-Former by 11.7 and 17.4 percentage points, respectively. It also obtains the highest AP^{75} of 83.0%, demonstrating improved accuracy in segmenting fine boundary details and occluded object contours. In addition, KP-Mask2Former achieves the best AP^{M} and AP^{L} values among the compared methods, indicating consistent performance across objects of medium and large scale.

Table 5 Performance comparison of different models on the FOD dataset

Model	AP↑	AP^{50}↑	AP^{75}↑	AP^{M}↑	AP^{L}↑
PolySnake	22.6	27.5	20.4	15.0	25.0
E2EC	43.4	51.6	33.4	18.3	42.5
BCNet	61.3	72.2	59.2	32.8	60.1
Transfiner	68.5	79.5	71.5	36.0	68.6
PatchDCT	58.5	66.5	61.8	21.3	58.6
MP-Former	62.8	70.2	65.4	22.0	63.2
FastInst	33.6	48.2	36.4	7.16	33.7
KP-Mask2Former	**80.2**	**85.0**	**83.0**	**44.0**	**80.4**

Figure 11 presents segmentation results on the FOD dataset, comparing the performance of different methods on occluded fish instances. Figure 11a shows the ground-truth annotation of the occluded crucian carp. In Figure 11b, FastInst fails to accurately recover the partially occluded regions, particularly the internal areas. As shown in Figure 11c, MP-Former incorrectly separates two disconnected regions of the same instance into different objects. Figure 11d shows that PatchDCT produces blurred boundaries in the reconstructed contours. In Figure 11e, Transfiner incorrectly predicts an additional non-existent fish instance. By contrast, KP-Mask2Former produces clear boundaries and segmentation results that closely agree with the ground truth, as shown in Figure 11f. Overall, KP-Mask2Former recovers occluded contours more accurately than the competing methods and preserves clearer object boundaries, further demonstrating the effectiveness of the proposed method.

4 Discussion

The robust performance of KP-Mask2Former on the OOD and FOD datasets validates its targeted structural design for dense aquaculture instance segmentation. Traditional frameworks like Mask2Former frequently suffer from boundary blurring under severe intra-class occlusion due to feature attenuation in cascaded attention layers. By contrast, embedding PCBAM within the patch-merging phase allows concurrent extraction of channel and spatial attention maps, effectively preserving faint contour cues. Furthermore, replacing standard MLPs with Swin KANsformer blocks provides learnable edge-based univariate functions. This mathematical flexibility allows the network to effectively reconstruct continuous boundaries despite non-rigid fish deformations and motion blur. Enhanced by the SEAM-Mask Head, the framework successfully mitigates spatial misalignment and feature aliasing triggered by high-density overlaps. Comparative benchmarks reveal distinct limitations in generic contour- or query-based segmentation methodologies within aquaculture environments. Contour-based architectures (e.g., E2EC, PolySnake) yield low AP^{S} scores because continuous coordinate regression is easily disrupted by texture homophily among overlapping conspecific fish, causing boundary leakage. While bilayer extraction networks like BCNet attempt foreground-background separation, they fail when overlapping instances share identical chromaticity and patterning. Modern query-based models like FastInst and MP-Former also suffer from localized feature aliasing, occasionally omitting overlapping segments. Conversely, KP-Mask2Former achieves a superior AP^{75} of 91.4% on the OOD dataset. The integration of Swin KANsformer and SEAM ensures rigid boundary definition even when low-level semantic and boundary cues are spatially entangled.

Nevertheless, several limitations remain and should be addressed to improve its applicability in more complex practical environments. First, although the current experiments validate the effectiveness of the proposed model on the target fish species, further evaluations on a wider range of aquaculture species are required to assess its cross-species generalization. Second, the current dataset was primarily collected under controlled indoor tank conditions. Additional data from diverse outdoor aquaculture environments[30], such as ponds, cages, and open-water farms, are therefore needed to evaluate and improve the model's robustness to variations in illumination, water turbidity, water-flow disturbances, and complex backgrounds. Finally, this study primarily focuses on offline algorithmic evaluation. Future work will investigate lightweight optimization techniques, including model pruning, quantization, and inference acceleration, to support deployment on resource-constrained edge devices.

5 Conclusions

To address the challenge of occluded fish contour reconstruction, this study proposes KP-Mask2Former and constructs the OOD dataset, which contains 5800 images covering diverse occlusion patterns. First, PCBAM is embedded into the stage-two patch-merging module of the backbone to balance local and global information, thereby enhancing feature representation in occluded regions. Second, the MLP layers in Swin Transformer are replaced with KAN layers to improve the modeling of fragmented

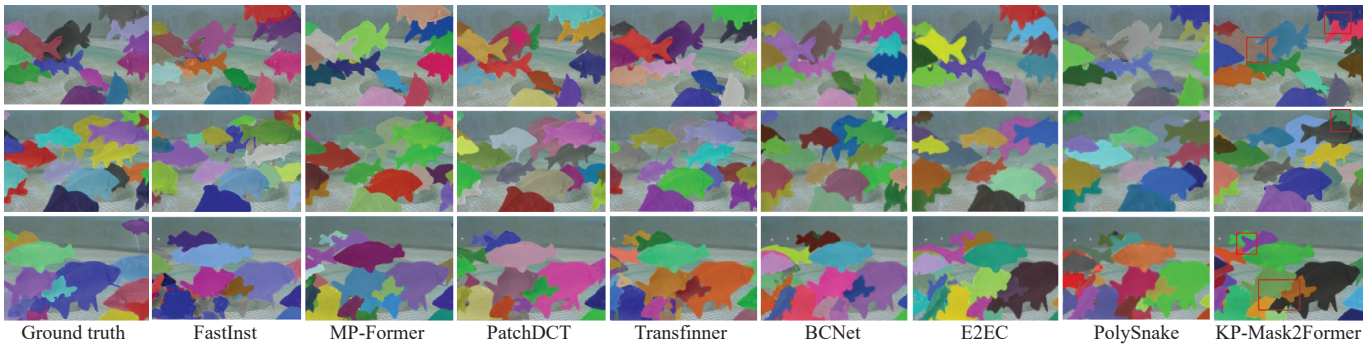


Figure 11 Segmentation results obtained by different methods on the FOD dataset

boundaries and enhance the adaptability of the network. Third, a SEAM-enhanced mask head is introduced to suppress interference from occluding instances and improve the reconstruction of occluded fish boundaries. Experiments conducted on both the self-constructed OOD dataset and the additional FOD dataset demonstrate that KP-Mask2Former achieves superior segmentation performance and substantially improves occluded contour reconstruction.

Acknowledgements

This work was supported by Shandong Province Key Youth Research Projects in the Humanities and Social Sciences and the Shandong Province Key R&D Program Project (Grant No. 2022TZXD005).

[References]

[1] Costa C, Loy A, Cataudella S, Davis D, Scardi M. Extracting fish size using dual underwater cameras. Aquacultural Engineering, 2006; 35(3): 218–227.

[2] Guo Y, Aggrey S E, Oladeinde A, Johnson J, Chai L. A machine vision-based method optimized for restoring broiler chicken images occluded by feeding and drinking equipment. Animals, 2021; 11(1): 123.

[3] Huang E, Mao A, Hou J, Wu Y, Xu W, Ceballos M-C, et al. Occlusion-resistant instance segmentation of piglets in farrowing pens using center clustering network. Computers and Electronics in Agriculture, 2023; 210: 107950.

[4] Meng H, Jin S, Liu W T, Qian C, Lin M X, Ouyang W L, et al. 3D interacting hand pose estimation by hand de-occlusion and removal. Proceedings of the European Conference on Computer Vision, 2022; pp.380–397. DOI: 10.1007/978-3-031-20068-7_22

[5] Ju Y-J, Lee G-H, Hong J-H, Lee S-W. Complete face recovery GAN: Unsupervised joint face rotation and de-occlusion from a single-view image. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022; pp.3711–3721. doi: 10.1109/WACV51458.2022.00124

[6] Yin X, Huang D, Fu Z, Wang Y, Chen L. Segmentation-reconstruction-guided facial image de-occlusion. Proceedings of the 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), 2023; pp.1–8. doi: 10.1109/FG57933.2023.10042570

[7] Zhou Q, Wang S Y, Wang Y T, Huang Z L, Wang X G. Human de-occlusion: Invisible perception and recovery for humans. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp.3691–3701. doi: 10.1109/CVPR46437.2021.00369

[8] Li L, Zhang T F, Kang Z F, Jiang X K. Mask-FPAN: semi-supervised face parsing in the wild with de-occlusion and UV GAN. Computers & Graphics, 2023; 116: 185–193.

[9] Zhan X, Pan X, Dai B, Liu Z, Lin D, Loy C-C. Self-supervised scene de-occlusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp.3784–3792. doi: 10.1109/CVPR42600.2020.00384

[10] Huang J Z, Yu X, An D, Ning X, Liu J C, Tiwari P. Uniformity and deformation: A benchmark for multi-fish real-time tracking in the farming. Expert Systems with Applications, 2025; 264: 125653.

[11] Ghiasi G, Cui Y, Srinivas A, Qian R, Lin T-Y, Cubuk E-D, et al. Simple copy-paste is a strong data augmentation method for instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp.2918–2928. doi: 10.1109/CVPR46437.2021.00294

[12] Cheng B, Misra I, Schwing A G, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1290–1299. doi: 10.1109/cvpr52688.2022.00135

[13] Zhu X Z, Su W J, Lu L W, Li B, Wang X G, Dai J F. Deformable detr: Deformable transformers for end-to-end object detection. arXiv 2020. DOI: 10.48550/arXiv.2010.04159.

[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017; 30.

[15] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; pp.770–778. doi: 10.1109/CVPR.2016.90

[16] Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp.10012–10022. doi: 10.1109/ICCV48922.2021.00986.

[17] Woo S, Park J, Lee J-Y, Kweon I S. Cbam: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), 2018; pp.3–19. doi: 10.1007/978-3-030-01234-2_1

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; pp.7132–7141. doi: 10.1109/TPAMI.2019.2913372

[19] Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, et al. KAN: Kolmogorov-arnold networks. International Conference on Learning Representations, 2025; pp.70367–70413.

[20] Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; pp.1251–1258. DOI: 10.1109/CVPR.2017.195

[21] Yu Z, Huang H, Chen W, Su Y, Liu Y, Wang X. YOLO-Facev2: A scale and occlusion aware face detector. Pattern Recognition, 2024; 155: 10.

[22] Feng H, Zhou K Y, Zhou W G, Yin Y F, Deng J J, Sun Q, et al. Recurrent generic contour-based instance segmentation with progressive learning. IEEE Transactions on Circuits and Systems for Video Technology, 2024. https://arxiv.org/html/2301.08898v2

[23] Zhang T, Wei S Q, Ji S P. E2EC: An end-to-end contour-based method for high-quality high-speed instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022; pp.4443–4452. DOI: 10.48550/arXiv.2203.04074

[24] Wen Q R, Yang J R, Yang X, Liang K W. Patchdct: Patch refinement for high quality instance segmentation. Proceedings of the Eleventh International Conference on Learning Representations, 2022. DOI: 10.48550/arXiv.2302.02693

[25] Ke L, Tai Y-W, Tang C-K. Deep occlusion-aware instance segmentation with overlapping bilayers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp.4019–4028. doi: 10.1109/CVPR46437.2021.00401

[26] Ke L, Danelljan M, Li X, Tai Y-W, Tang C-K, Yu F. Mask transfiner for high-quality instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022; pp.4412–4421. doi: 10.1109/CVPR52688.2022.00437

[27] Zhang H, Li F, Xu H Z, Huang S J, Liu S L, Ni L M, et al. Mp-Former: Mask-piloted transformer for image segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023; pp.18074–18083. doi: 10.1109/CVPR52729.2023.01733

[28] He J J, Li P Y, Geng Y F, Xie X S. Fastinst: A simple query-based model for real-time instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023; pp.23663–23672. doi: 10.1109/CVPR52729.2023.02266

[29] Wang X Y, Yu H H, Zhang C, Liu Z N, Luo Z, Chen Y Y, et al. An underwater image dataset for occlusion-aware fish instance segmentation. Scientific Data, 2026; 13(1). doi: 10.1038/s41597-026-06898-w

[30] Li Y, Yang H H, Xing H R, Fu L Y, Gu N R, Liu S E. A Transformer-Mamba framework for cross-regional long-term forecasting of key water quality in aquaculture. Water Research, 2026; 302: 126135.