An Overview of YOLOv8 and Vision Transformer (ViT)

Jingyang Chen, Xuzhong Wang, Huizhi Zhao

October 17, 2025

Abstract

This document introduces two state-of-the-art computer vision architectures used in the DEEPSEA project: YOLOv8 for object detection and Vision Transformer (ViT) for image classification. We describe their core design principles, mathematical foundations, and their integration into marine species detection and classification.

1 Introduction

Deep learning has transformed computer vision tasks such as object detection and image classification. In the DEEPSEA framework, two models play central roles:

- **YOLOv8 (You Only Look Once, version 8):** A fast, real-time object detector that predicts bounding boxes and class probabilities in a single forward pass.
- **Vision Transformer (ViT):** A transformer-based architecture that applies self-attention to image patches, achieving state-of-the-art accuracy in image classification.

2 YOLOv8: Object Detection Model

2.1 Model Overview

YOLOv8 belongs to the YOLO (You Only Look Once) family of detectors, which reformulates object detection as a single regression problem rather than a two-stage pipeline. The model outputs bounding box coordinates and class probabilities directly from input images.

2.2 Mathematical Formulation

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, YOLO divides it into $S \times S$ grid cells. Each grid cell predicts B bounding boxes, where each bounding box is represented by (x, y, w, h, c):

$$x, y \in [0, 1], \quad w, h \in [0, 1], \quad c = P(\text{object})$$

and class probabilities $P(\text{class}_i|\text{object})$ for each class i.

The final detection confidence for each class i is:

$$P(\text{class}_i) = P(\text{object}) \cdot P(\text{class}_i|\text{object})$$

2.3 Loss Function

YOLOv8 minimizes a composite loss function combining localization, confidence, and classification terms:

$$\mathcal{L} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$$

where

$$\begin{aligned}
\mathcal{L}_{\text{box}} &= \sum_i \|b_i - \hat{b}_i\|_2^2, \\
\mathcal{L}_{\text{obj}} &= \sum_i (c_i - \hat{c}_i)^2, \\
\mathcal{L}_{\text{cls}} &= -\sum_i y_i \log(\hat{y}_i)
\end{aligned}$$

The model uses anchor-free detection with feature pyramids to efficiently detect multi-scale objects, improving both precision and inference speed.

3 Vision Transformer (ViT): Image Classification Model

3.1 Patch Embedding

Unlike CNNs that use convolutional filters, ViT treats an image as a sequence of fixed-size patches. Given an image $x \in \mathbb{R}^{H\times W\times C}$, it is divided into N patches $x_p^i \in \mathbb{R}^{P^2\times C}$, where P is the patch size.

Each patch is linearly projected into a D-dimensional embedding:

$$z_0^i = x_p^i E + E_{\text{pos}}^i, \quad i = 1, \ldots, N$$

where $E \in \mathbb{R}^{(P^2\cdot C)\times D}$ is the learned embedding matrix and E_{pos}^i adds positional information.

3.2 Transformer Encoder

The embeddings are fed into a stack of L Transformer encoder layers, each consisting of:

- Multi-Head Self-Attention (MHSA)
- Multi-Layer Perceptron (MLP)
- Layer Normalization (LN) and residual connections

The self-attention mechanism computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices.

The multi-head variant concatenates multiple attention outputs:

$$\text{MHA}(Q, K, V) = [\text{head}_1, \dots, \text{head}_h]W^O$$

3.3 Classification Head

A special `[CLS]` token is prepended to the patch embeddings. After processing through the encoder, the final representation of this token $z_L^{(0)}$ is passed through an MLP for classification:

$$\hat{y} = \text{softmax}(W_{\text{cls}} z_L^{(0)} + b)$$

The cross-entropy loss is then minimized:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

4 Integration in DEEPSEA

In DEEPSEA:

- **YOLOv8** performs localization of marine species, outputting bounding boxes and detection confidences in real time.
- **ViT** classifies cropped objects or full images to determine the specific benthic species category.

This hybrid pipeline achieves both spatial detection and semantic understanding, with ViT providing high accuracy (∼92%) and YOLOv8 offering fast detection (∼80% mAP).

5 Conclusion

The combination of YOLOv8 and ViT provides a powerful framework for underwater computer vision. YOLOv8's efficiency in spatial localization and ViT's superior classification performance together enable accurate, real-time marine species identification.