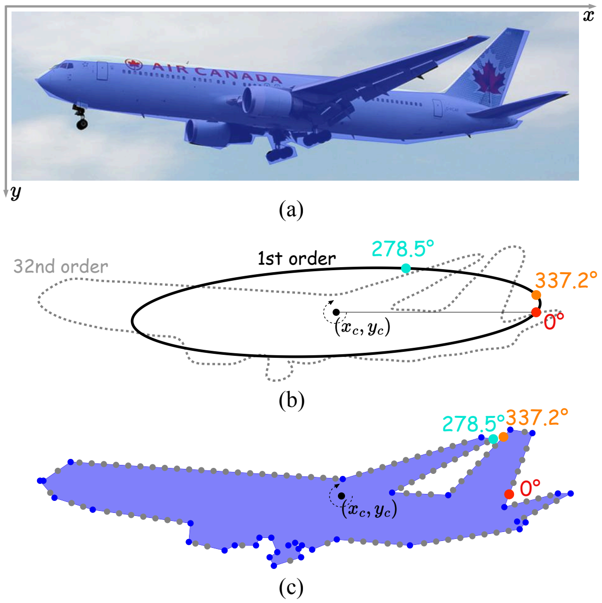
EFDTR: Learnable Elliptical Fourier Descriptor Transformer for Instance Segmentation

Jiawei Cao[1 2] Chaochen Gu[1 2] Hao Cheng[1 2] Xiaofeng Zhang[1 2] Kaijie Wu[1 2] Changsheng Lu[3 4]

Abstract

Polygon-based object representations efficiently model object boundaries but are limited by high optimization complexity, which hinders their adoption compared to more flexible pixel-based methods. In this paper, we introduce a novel vertex regression loss grounded in Fourier elliptic descriptors, which removes the need for rasterization or heuristic approximations and resolves ambiguities in boundary point assignment through frequency-domain matching. To advance polygon-based instance segmentation, we further propose EFDTR (**E**lliptical **F**ourier **D**escriptor **Tr**ansformer), an end-to-end learnable framework that leverages the expressiveness of Fourier-based representations. The model achieves precise contour predictions through a two-stage approach: the first stage predicts elliptical Fourier descriptors for global contour modeling, while the second stage refines contours for fine-grained accuracy. Experimental results on the COCO dataset show that EFDTR outperforms existing polygon-based methods, offering a promising alternative to pixel-based approaches. Code is available at `https://github.com/chrisclear3/EFDTR`.



(a)

(b)

(c)

Figure 1. (a) Pixel-wise instance mask from the COCO dataset. (b) First-order reconstruction (black line) compared with 32nd-order EFD reconstruction (gray dashed line). (c) Dense contour sampling with 128 points, initially equidistant, followed by target snapping, where sampling points are adjusted to the nearest target points.

1. Introduction

Instance segmentation is a fundamental task in computer vision, focusing on the detection and delineation of individual objects within an image by assigning unique masks to each detected instance. This capability is essential for applications in autonomous driving, medical image processing (Chen et al., 2018; 2023), and robotics (Ma et al., 2024; Liu et al., 2025; Zhang et al., 2024a), where precise object boundaries are critical to decision making and interaction. Most existing instance segmentation methods rely on pixel-wise masks to delineate objects (He et al., 2017; Kirillov et al., 2020; Wang et al., 2020; Cheng et al., 2022b;a; Li et al., 2023a). However, contour-based or polygon-based approaches have gained attention due to their ability to represent object shapes with fewer parameters while retaining key topological properties. Notably, these methods can directly output contour coordinates or keypoints (Lu et al., 2024; Lu & Koniusz, 2024; 2022; Lu, 2024; Lu et al., 2023) without post-processing, making them particularly well suited for tasks such as remote sensing (Zhang et al., 2024b) and BEV vector map generation (Liao et al., 2024) in autonomous

[1]Department of Automation, Shanghai Jiao Tong University, Shanghai, China [2]Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China [3]School of Computing, The Australian National University, Canberra, Australia [4]Australian Institute for Machine Learning, University of Adelaide. Correspondence to: Kaijie Wu & Changsheng Lu <kaijiewu@sjtu.edu.cn, changshengluu@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

driving. Since contours are typically represented as collections of polygons, these methods are collectively referred to as polygon-based approaches in this paper to highlight their focus on polygon regression.

In recent years, polygon-based methods have significantly advanced polygon regression accuracy and performance. Representative approaches include Curve GCN (Ling et al., 2019), PolyTransform (Liang et al., 2020), Deep Snake (Peng et al., 2020), PolarMask (Xie et al., 2020), DANCE (Liu et al., 2021), E2EC (Zhang et al., 2022b), and BoundaryFormer (Lazarow et al., 2022). These methods can be broadly categorized into single-stage and two-stage approaches.

Single-stage models directly predict contour coordinates in an end-to-end manner. While efficient, their precision is often limited, making two-stage methods preferable. Two-stage methods typically generate a coarse initial contour, followed by refined polygon regression through displacement or offset learning. However, both approaches depend on the regression of 2D vertex coordinates, which requires precise target assignment. Apart from differentiable rasterization methods like PolyTransform (Liang et al., 2020) and BoundaryFormer (Lazarow et al., 2022), which perform supervised learning via rendered pixel maps, existing target assignment strategies can be broadly categorized into three types:

- **Polar Target Assignment**: This strategy aligns targets and predictions using the same polar angle. For multi-polygon instances, it selects the largest polygon and defines its centroid as the origin for uniform ray sampling (see Figure 2b). While conceptually simple, this approach is limited to star-convex shapes and struggles with non-convex structures.

- **Cartesian Target Assignment**: This strategy minimizes the Euclidean distance between predicted and ground-truth vertices to determine assignments. However, it lacks global context, leading to ambiguities in matching (Figure 2c). Additionally, it disregards the sequential topology of contours, often causing structural inconsistencies in polygon regression.

- **Hybrid Strategies**: This strategy segment the contour into regions based on polar angles and apply Euclidean distance-based assignment within each subregion. While this mitigates certain topological errors, it does not fundamentally resolve the limitations of distance-based regression.

Despite their merits, existing methods impose restrictive geometric assumptions, struggle with topological consistency, or lack a robust global contour representation. To address these issues, we introduce Elliptic Fourier Descriptors (EFDs) as a compact, reliable representation for polygonal contours. Any closed 2D contour can be expressed as a Fourier series with ellipses as basis functions (Kuhl & Giardina, 1982). As illustrated in Figure 1b, the inverse transformation of a first-order EFD (solid) captures a rough global shape represented by an ellipse (Lu et al., 2019; 2017), while a thirty-second-order EFD (dashed) restores fine-grained details. Furthermore, the phase information encoded in EFDs establishes a bijective mapping between contour points and their positions along the shape. This facilitates precise regression target assignment and resolves the inherent ambiguity found in Euclidean distance-based methods. For example, as shown in Figure 1c, points that are spatially close but topologically distinct are clearly disambiguated through their EFD phase values. To fully leverage the advantages of EFDs, we propose EFDTR (Elliptic Fourier Descriptor-based Transformer), a two-stage regression model that addresses the limitations of existing target assignment strategies through phase-domain sampling. In the first stage, EFD parameter regression captures the global coarse topology, while the second stage refines the regression through phase-based target assignment to achieve precise boundary localization. The framework effectively harnesses the Transformer's capability for sequence modeling, with deformable attention facilitating flexible feature aggregation.

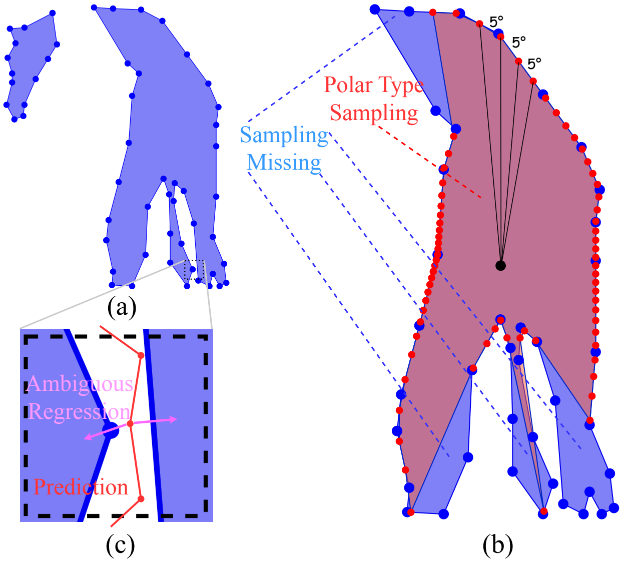


Figure 2. Limitations of regression matching between the Polar and Cartesian coordinate systems. (a) Original polygon segmentation. (b) Inadequate sampling of target points in the polar coordinate system. (c) Ambiguous regression in the Cartesian coordinate system due to minimum Euclidean distance assignment.

By integrating EFDTR into an end-to-end differentiable segmentation pipeline, we achieve superior polygon regression accuracy, surpassing prior polygon-based approaches. EFDTR emerges as a promising alternative to pixel-based approaches, offering a structured and efficient solution for shape representation in learning-based segmentation.

2. Related Work

2.1. Instance Segmentation with Polygon

Various methods have been proposed for contour prediction and instance segmentation using deep learning techniques due to their advantage of capturing object shapes (Shi et al., 2024). Curve-GCN (Ling et al., 2019) employs a graph convolutional network for prediction, fine-tuning the process with a cyclic matching loss and a differentiable rasterization loss. PolyTransform (Liang et al., 2020) extracts initial contours via a segmentation network and regresses points using a deforming network. DeepSnake (Peng et al., 2020) adapts the ExtremeNet framework (Zhou et al., 2019), initially detecting an octagonal contour, which is then iteratively refined via ring convolution. Point-Set Anchors (Wei et al., 2020) regresses sampled points on anchor points placed on bounding boxes, facilitating instance segmentation. Dance (Liu et al., 2021) extracts boundary features using a contour extraction network and applies attentive deformation to regress the boundaries. These methods typically rely on predicting target points from object boundaries, but the resulting point allocation tends to be uneven, limiting their ability to model complex contours effectively.

PolarMask (Xie et al., 2020) represents contours in polar coordinates through their center, radius, and angle, utilizing a FOCS-based (Tian et al., 2022) architecture and an IoU loss function for contour regression. FourierNet (Riaz et al., 2021) predicts contours by transforming contour parameters into Fourier space and directly regressing Fourier coefficients. These representation methods struggle to accurately model general object contours due to their structural constraints.

E2EC (Zhang et al., 2022b) aligns points within intervals using multi-directional alignment and supervises point regression through a dynamic matching loss, effectively combining the allocation strategies of DANCE and PolarMask. C-AOI (Zhu et al., 2023), an improvement of E2EC, introduces a Transformer into the model architecture and enhances the matching loss with additional constraints, proposing an adaptive matching loss to reduce allocation ambiguity.

PolyFormer (Liu et al., 2023) is a referring task model that can also be applied to instance segmentation, treating polygons as natural language statements, where each token corresponds to predicting a point on the contour. Finally, BoundaryFormer (Lazarow et al., 2022), a hybrid RCNN-Transformer model, utilizes iterative upsampling and deformable attention to predict point offsets, incorporating a differentiable rasterization loss into the objective function for effective supervision.

Recent studies in geometric representation provide additional insights. PolygonGNN (Yu et al., 2024) introduces a heterogeneous graph structure for polygon modeling, and its spanning tree sampling strategy inspires our MST-based contour merging. PolyhedronNet (Yu et al., 2025), though focused on 3D polyhedra, offers useful ideas for future extensions of surface-based contour representation.

2.2. DEtection with TRansformer (DETR)

The decoding mechanism of our model is inspired by key developments in the DETR series, which has significantly influenced the field of object detection (Wang et al., 2022a). Carion et al. (Carion et al., 2020) introduced DETR, an end-to-end object detection framework powered by Transformers, which surged waves due to its novel approach. Compared to traditional approaches, the standout characteristic of DETR is eliminating the need for hand-crafted anchor boxes and the step of Non-Maximum Suppression (NMS) commonly used in earlier detection paradigms. Since its inception, several variants of DETR have been proposed to tackle its limitations and further improve performance. For example, Deformable-DETR (Zhu et al., 2020) enhances training efficiency by incorporating multi-scale features and optimizing the attention mechanism. Conditional DETR (Meng et al., 2021) and Anchor DETR (Wang et al., 2022b) address the challenges of query design optimization. DAB-DETR (Liu et al., 2022) introduces the concept of 4D reference points and refines the model by optimizing prediction boxes in successive layers. DINO (Zhang et al., 2022a), building upon these methods, achieves state-of-the-art performance, establishing a solid foundation for further advancements in DETR-based detection.

Incorporating insights from DINO (Zhang et al., 2022a), RT-DETR (Zhao et al., 2024) combines YOLO-inspired techniques to enable real-time performance for DETR-based detection systems. BoundaryFormer takes it a step further by incorporating Deformable Attention and applying progressive upsampling in the decoder to improve the sequential regression of polygon points. These innovations lay the groundwork for the two-stage decoder architecture proposed in this paper, contributing to the evolution of DETR-style object detectors.

3. Method

3.1. Multiple Polygon Connection

Elliptical Fourier descriptors are designed for single closed polygons. For instances containing multiple polygonal contours, these must be converted into a single polygon by introducing auxiliary edges. Intuitively, the objective is to minimize the total length of the connecting edges, as shorter edges help reduce the error in Intersection over Union (IoU) when predicted auxiliary edges do not perfectly align with the ground truth.

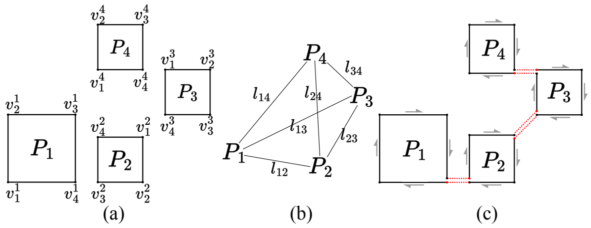


Figure 3. The illustration of multiple polygon connection.

Figure 3 illustrates an example of connecting four polygons. The distance l_{ij} between two polygons P_i and P_j, with vertex sets $V_i = \{v_1^i, v_2^i, \ldots, v_{n_1}^i\}$ and $V_j = \{v_1^j, v_2^j, \ldots, v_{n_2}^j\}$, respectively, is given by:

$$l_{ij} = \min_{v_p^i \in V_i, v_q^j \in V_j} \|v_p^i - v_q^j\|. \tag{1}$$

Each polygon is treated as a node in a fully connected graph, with edge weights l_{ij}. Finding the optimal auxiliary edges reduces to solving the minimum spanning tree (MST) of this graph. We adopt Prim's algorithm to compute the MST, efficiently merging multiple polygons into a single closed contour.

3.2. Elliptic Fourier Phase Assignment

The contour coordinates of a single closed polygon with m vertices $\{(v_x^1, v_y^1), (v_x^2, v_y^2), \ldots, (v_x^m, v_y^m)\}$ can be described by an elliptical Fourier series, as defined by the following equation:

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} x_c \\ y_c \end{bmatrix} + \sum_{n=1}^{\infty} \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} \cos\left(\frac{2\pi n t}{T}\right) \\ \sin\left(\frac{2\pi n t}{T}\right) \end{bmatrix}, \tag{2}$$

where (x_c, y_c) represents the centroid of polygon, and $(x(t), y(t))$ denotes the coordinates of a point along the polygon, parameterized by its cumulative arc-length t. The arc-length parameterization is defined as follows:

$$t_p = \sum_{i=1}^{p} \|\mathbf{v}_{i+1} - \mathbf{v}_i\|_2, \quad t_0 = 0, \quad t_m = T, \tag{3}$$

where T denotes the perimeter of the polygon and $\mathbf{v}_{m+1} = \mathbf{v}_1$. The Fourier coefficients A_n, B_n, C_n, D_n associated with the harmonic number n are computed as follows:

$$A_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{m} \frac{\Delta x_p}{\Delta t_p} [\cos(\frac{2\pi n t_p}{T}) - \cos(\frac{2\pi n t_{p-1}}{T})], \tag{4}$$

$$B_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{m} \frac{\Delta x_p}{\Delta t_p} [\sin(\frac{2\pi n t_p}{T}) - \sin(\frac{2\pi n t_{p-1}}{T})], \tag{5}$$

$$C_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{m} \frac{\Delta y_p}{\Delta t_p} [\cos(\frac{2\pi n t_p}{T}) - \cos(\frac{2\pi n t_{p-1}}{T})], \tag{6}$$

$$D_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{m} \frac{\Delta y_p}{\Delta t_p} [\sin(\frac{2\pi n t_p}{T}) - \sin(\frac{2\pi n t_{p-1}}{T})]. \tag{7}$$

Δx_p and Δy_p indicate the displacement along the x- and y-axes, respectively, between points p and $p+1$. Δt_p is the step length between points p and $p+1$.

As Fourier coefficients are sensitive to the choice of the initial point, reparameterization strategies are commonly adopted to maintain consistency across different applications. To achieve a standardized representation, we enforce a phase constraint at $t = 0$, which corresponds to the starting point of the contour. Specifically, we require that the first-order elliptical Fourier coefficient aligns with the positive x-axis relative to the centroid.

Furthermore, we adhere to the right-hand rule, where counterclockwise motion is considered positive in the standard Cartesian coordinate system, while clockwise motion is positive in the image coordinate system. Consequently, the first-order Fourier coefficients must satisfy:

$$\begin{cases} A_1 > 0 \\ C_1 = 0 \\ D_1 > 0 \end{cases} \tag{8}$$

This guarantees that the zero-phase frequency component in the Fourier domain is properly aligned with the positive x-axis of the centroid, as depicted in Figure 1d.

The elliptical Fourier series computed for each polygon should then be multiplied by the following rotation matrix to satisfy the normalization requirements of the first-order Fourier series. Here, θ represents the phase shift required to align the standardized starting point with that of the input polygon:

$$\theta = \begin{cases} \arctan2(-C_1, D_1), & A_1 D_1 > B_1 C_1 \\ \arctan2(C_1, -D_1), & A_1 D_1 \leq B_1 C_1 \end{cases} \tag{9}$$

Then apply the rotation matrix to the original elliptical Fourier coefficients, yielding:

$$\begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix} = \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix} \tag{10}$$

The phase θ_p of point p is computed as:

$$\theta_p = \frac{2\pi t_p}{T}. \tag{11}$$

Here, θ_p represents a normalized arc-length parameter, mapping each point along the contour to a phase in the Fourier representation. Consequently, there exists a bijective relationship between the phase and the contour points of the closed polygon, ensuring a unique correspondence.

As illustrated in Figure 1c, the cyan and orange points are spatially close in Cartesian coordinates, making Euclidean distance an ambiguous matching criterion. However, their distinct phase values in the frequency domain enable precise differentiation. Similarly, the green and magenta points, which originate from overlapping vertices introduced by auxiliary edges, further highlight the importance of phase information in resolving under-matching and ambiguity in polygonal contour regression.

3.3. Model Structure with Leanable Elliptic Fourier

The model architecture proposed in this paper is shown in Figure 4. It follows the design principles of two-stage object detection models and incorporates key elements from the DETR series in the decoder. Specifically, in the feature extraction network, we primarily draw inspiration from RT-DETR (Zhao et al., 2024), which uses an improved Pyramid Attention Network (PAN) (Li et al., 2018) called Hybrid Feature Fusion module to obtain multi-scale feature maps P_2, P_3, P_4, P_5. In our case, since we aim to predict instance contours, we retain feature maps with up to 4× downsampling P_2. This module performs feature fusion using self-attention on the 32× downsampled feature map P_5, while applying convolutional layers for fusion in the other feature map paths. The final output consists of feature vectors with a size of $\frac{85WH}{1024}$, where W and H represent the width and height of the input image, respectively.

In the EFD decoder, the input comprises multi-scale features from P_3, P_4, and P_5, striking a balance between computational efficiency and prediction accuracy. Since the EFD decoder primarily predicts Fourier parameters to capture a coarse global topology, utilizingP_3, P_4, and P_5 is sufficient. Incorporating P_2would introduce a significant computational overhead without a proportional gain in performance.

For the first EFD decoder layer, a Score Encoder ranks the flattened feature representations and selects the top-k features based on their assigned scores. These selected features are then concatenated with the noised target, forming the primary input to the decoder. The added noise facilitates denoising, aiding in convergence acceleration and improving accuracy. Furthermore, the EFD Encoder extracts the Fourier parameters corresponding to the top-k features, which serve as reference information to refine the decoding process. Consequently, each decoder layer operates with two key inputs: the concatenated top-k features with the noised target, and the EFD reference parameters.

The input features are first encoded by the Self-Attention module, where positional embeddings are incorporated, and the resulting representations serve as queries to retrieve relevant flattened features. These queries are then processed by the Deformable Attention module, which adaptively attends to spatially significant regions to refine the feature representations. The output features are subsequently passed through two prediction heads: the Class Head, which predicts the instance class, and the EFD Head, which estimates the Fourier descriptor parameters. The predicted EFD parameters serve as reference information for the next layer, enabling iterative refinement. The total number of parameters at each stage follows the formulation $4n + 1$, as illustrated in Figure 4.

The Fourier inverse transform is applied to reconstruct the 2D coordinates corresponding to the elliptic Fourier descriptors predicted in the first stage. To facilitate precise feature extraction, we employ the `grid_sample` function for feature sampling. Given the need for accurate contour regression in the subsequent stage, we integrate multi-scale fusion information from feature maps P_2 to P_5, where four-scale normalized coordinate sampling produces four sets of feature vectors.

To reduce computational overhead, we adopt a group-based feature sampling strategy, where instead of sampling all 128 predicted points individually, only $\frac{128}{g}$ points are sampled, where g denotes the group size, set to 4 in our design. Additionally, explicit EFD parameters and class information are leveraged to predict the fusion weights for the four scales, enabling adaptive weighting across hierarchical features. The fused single-scale feature representation is then used as the query input for the Polygon Decoder, refining the polygonal structure prediction in an iterative manner.

3.4. Loss

We utilize the VFL loss (Zhang et al., 2021) in the classification loss function to correlate the target value of positive samples with the IoU prediction. For the elliptical Fourier descriptor (EFD) regression, we apply the L1 loss, while the Smooth-L1 loss is used for polygons under phase alignment.

Based on the information in Section 3.2, the phase set for the target polygon points is defined as:

$$\mathcal{P}_t = \left\{ \frac{2\pi t}{T} \,\middle|\, t \in \{t_0, t_1, \ldots, t_{m-1}\} \right\}, \tag{12}$$

where T represents the total perimeter of the contour, and m is the number of vertices of the target polygon.

In contrast, the initial phase for the predicted polygon sampling points corresponds to uniform sampling in the interval $[0, 2\pi)$:

$$\tilde{\mathcal{P}}_p = \left\{ \frac{2\pi t}{n} \,\middle|\, t \in \{0, 1, 2, \ldots, n-1\} \right\}, \tag{13}$$

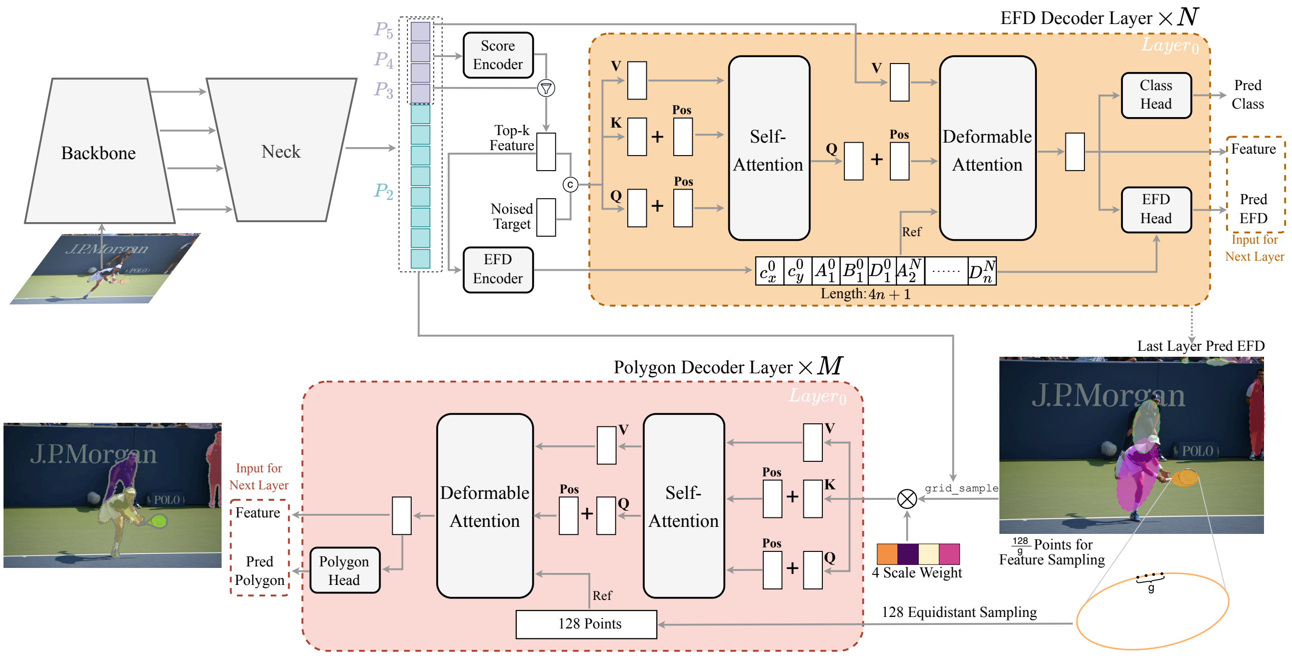


Figure 4. **Overview of EFDTR:** It consists of a feature extraction, learnable elliptical fourier descriptors (EFDs) decoder and polygon deocder three parts.

where n is the number of vertices of the predicted polygon and $n \geq m$. To better align with the target points, we apply a snapping operation $\mathcal{S}$, which is defined as follows:

$$\phi^* = \arg\min_{\phi \in \mathcal{P}_t} |\theta_p - \phi| \,, \tag{14}$$

$$\mathcal{S}(\theta, \mathcal{P}_t) = \begin{cases} \phi^* & \text{if } |\theta - \phi^*| < \frac{\pi}{n}, \\ & \text{and } |\theta - \phi^*| = \min_{\theta \in \tilde{\mathcal{P}}} |\theta - \phi^*| \,, \\ \theta & \text{otherwise.} \end{cases} \tag{15}$$

After applying the snapping operation, the phase set of the predicted points is updated as:

$$\mathcal{P}_p = \left\{ \mathcal{S}(\frac{2\pi t}{n}, \mathcal{P}_t) \middle| \; t \in \{0, 1, 2, \ldots, n-1\} \right\}. \tag{16}$$

Let $\mathcal{F}$ represent the inverse elliptical Fourier transform. The polygon vertex regression loss under phase alignment is defined as follows:

$$\mathcal{L}_1(pred, gt) = \frac{1}{n} \sum_{i=1}^{n} \text{smooth } l_1(\mathcal{F}_p(\theta), \mathcal{F}_t(\theta)), \theta \in \mathcal{P}_p \tag{17}$$

$$\mathcal{L}_2(pred, gt) = \frac{1}{m} \sum_{i=1}^{m} \text{smooth } l_1(\mathcal{F}_p(\theta), \mathcal{F}_t(\theta)), \theta \in \mathcal{P}_t \tag{18}$$

$$\mathcal{L}_{\text{polygon}} = \frac{\mathcal{L}_1(pred, gt) + \mathcal{L}_2(pred, gt)}{2} \tag{19}$$

The overall loss is as follows:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{efd}} + \beta \mathcal{L}_{\text{polygon}}, \tag{20}$$

where α=6 and β=10.

4. Experiments

4.1. Dataset and Pre-processing

The COCO dataset (Lin et al., 2014) is a widely used benchmark in computer vision, supporting tasks like object detection, segmentation, and captioning. It contains over 330,000 images across 80 categories with detailed annotations reflecting complex real-world object interactions.

COCO offers pixel-level annotations in Run-Length Encoding (RLE) and Polygon formats. This work focuses on Polygon annotations, for which we adjust the original labels during preprocessing to adhere to the right-hand rule. This ensures consistency, as our elliptical Fourier series is sensitive to orientation conventions.

4.2. Implementation Detail

The query number in the EFD decoder is set to 300, with adjacent 4 points grouped together. The EFDTR model is trained using the AdamW optimizer, with different learning rates for each model component and a multi-step learning rate scheduler. Additionally, Exponential Moving Average (EMA) is employed during training to stabilize the process. Data augmentation includes RandomFlip, RandomIoUCrop, and multi-scale training. During inference, the input image scale is fixed at 800 × 800.

Table 1. Quantitative Results on MS COCO. We compare our EFDTR with state-of-the-art models on `val2017`.

Method	Epoch	Output	Supervision	AP	AP_{50}	AP_{75}	AP_{S}	AP_{M}	AP_{L}
ResNet-50 backbone									
Mask R-CNN(ICCV17)	400	masks	masks	42.5	–	–	23.8	45.0	60.0
DynaMask(CVPR23)	24	masks	edges+masks	38.2	58.1	41.5	20.5	40.8	52.7
Mask DINO(CVPR23)	50	masks	masks	46.0	69.0	50.7	26.1	49.3	66.1
MixingMask(PR24)	36	masks	polygons+masks	42.2	65.2	45.9	19.8	46.5	65.7
DeepSnake(CVPR20)	160	polygons	polygons	30.5	–	–	–	–	–
PolarMask(CVPR20)	24	polygons	polygons	32.1	53.7	33.1	14.7	33.8	45.3
DANCE(WACV21)	12	polygons	edges+polygons	34.5	55.3	36.5	17.2	37.5	48.0
SharpContour(CVPR22)	12	polygons	points+masks	37.8	–	–	24.3	49.4	59.1
BoundaryFormer(CVPR22)	12	polygons	masks	36.1	56.7	–	–	–	–
EFDTR(Ours)	36	polygons	polygons	43.6	64.8	47.2	23.4	46.5	64.1
ResNet-101 backbone									
Mask2Former(CVPR22)	50	masks	masks	44.2	–	–	23.8	47.7	66.7
DynaMask (CVPR23)	24	masks	edges+masks	39.0	59.1	42.2	20.9	42.1	53.3
BEIS(ECAI24)	36	masks	points+masks	42.1	–	–	25.0	45.4	55.4
SharpContour(CVPR22)	36	polygons	points+masks	40.8	–	–	–	–	–
EFDTR(Ours)	36	polygons	polygons	45.1	66.6	49.3	24.1	48.2	66.4

4.3. Comparison with State-of-the-Arts

Table 1 compares our EFDTR model with mask-based methods (e.g., Mask R-CNN (He et al., 2017), Mask DINO (Li et al., 2023a)) and polygon-based approaches (e.g., DeepSnake (Peng et al., 2020), PolarMask (Xie et al., 2020)). Our model significantly outperforms prior polygon-based methods, with an AP of 43.6, while remaining competitive with mask-based models like Mask R-CNN (AP 42.5) and Mask DINO (AP 46.0). EFDTR excels in both small and large object categories, and, when using a ResNet-101 backbone, further improves performance to 45.1 AP, surpassing Mask2Former and DynaMask (Cheng et al., 2022a; Li et al., 2023b). This highlights EFDTR as a promising alternative to pixel-based methods, advancing the state of polygon-based segmentation.

4.4. Ablation Study

In this section, we conduct ablation studies to evaluate the key components of our proposed EFDTR method and their impact on performance, validated on the COCO `val2017` dataset. For fairness and efficiency, all experiments are trained for 12 epochs.

Number of Decoder Layers. We investigate the effect of varying the number of EFD decoders (N_E) and polygon decoders (N_P) on instance segmentation performance. The experimental results, as shown in Table 2, indicate that the best performance is achieved with $N_E = 6$ and $N_P = 3$, yielding an AP of 40.5.

Order of EFD Prediction. To assess the impact of different orders of elliptical Fourier descriptors (EFDs) on segmentation, we compare the performance of 1st, 2nd, and 4th order harmonics. The first-order EFDs, which focus on predicting the ellipse shape, yield the best performance. In contrast, higher-order terms introduce noise due to their complexity, impairing coarse localization and hindering polygon convergence. This suggests that simpler, first-order EFDs are more effective for our task.

Number of Vertices in Group. We conducted experiments with different vertex groupings: 2, 4, 8, and 16 vertices per group. The results, shown in Table 4, indicate that using 4 vertices per group strikes the best balance between performance and computational cost. Although using 2

Table 2. Effect of the number of EFD decoders (N_E) and polygon decoders (N_P) on instance segmentaion.

(N_E, N_P)	**(4, 3)**	**(4, 4)**	**(5, 3)**	**(5, 4)**	**(6, 3)**	**(6, 4)**
AP	39.4	39.4	40.2	40.0	40.6	40.4
AP_{50}	58.7	58.6	60.2	59.8	60.8	60.5
AP_{75}	42.7	42.7	43.5	43.3	43.7	43.7
Params	49.6M	50.8M	50.8M	51.9M	51.9M	53.0M

Table 3. Effect of the order of EFD on instance segmentaion.

order	AP	AP_{50}	AP_{75}	AP_{S}	AP_{M}	AP_{L}
1	40.6	60.8	43.7	20.5	43.5	61.4
2	29.4	46.2	31.0	11.1	33.4	48.7
4	31.2	50.3	32.4	10.7	32.7	53.8

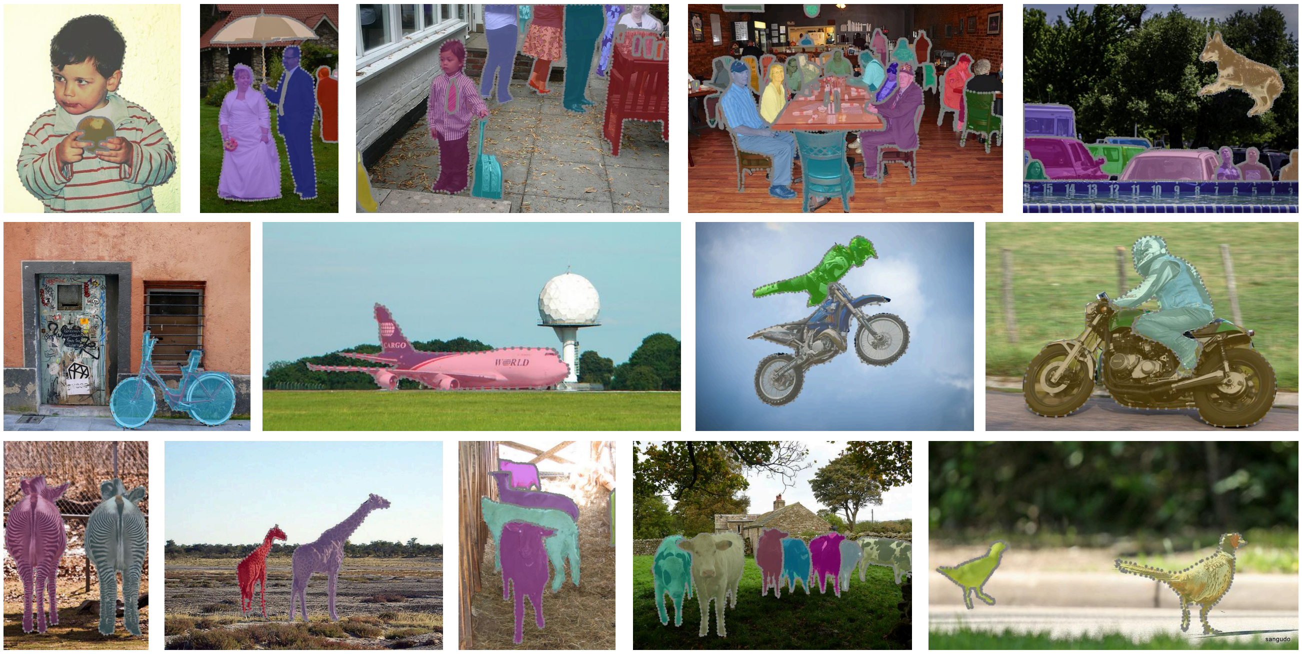


Figure 5. EFDTR visualization results on COCO `val2017`. The grey points on contour are predicted vertices of instances.

Table 4. Effect of the number of vertex group size on segmentation performance.

number	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
2	40.8	60.7	44.2	19.8	44.0	61.8
4	40.6	60.8	43.7	20.5	43.5	61.4
8	40.1	60.6	43.3	20.0	43.1	60.8
16	39.1	60.3	42.0	18.8	42.3	59.3

vertices yields the highest accuracy, the improvement in performance is not justified by the increased computational cost.

Multi-scale Feature Fusion Module Table 5 shows that the weighted fusion method outperforms both mean and ”only P_2” fusion, achieving the highest AP and improved performance across all metrics, highlighting its effectiveness in enhancing multi-scale feature integration.

IoU Type in varifocal loss. In varifocal loss (), the target value for positive samples depends on IoU. Table 6 shows that using axis-aligned IoU (AP 41.1) outperforms rotated IoU (AP 40.6). The simpler axis-aligned method provides higher IoU values, offering better supervisory signals and improving the overall performance.

Table 5. Effect of Multi-scale Fusion Methods on Segmentation Performance.

method	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
mean	40.3	60.1	43.6	18.9	43.5	61.2
only P_2	40.1	60.0	43.2	19.1	43.1	61.0
weighted fusion	40.6	60.8	43.7	20.5	43.5	61.4

Table 6. Impact of the IoU type in varifocal loss function.

IoU type	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
axis-aligned box	41.1	62.5	44.1	20.5	44.1	61.4
rotated box	40.6	60.8	43.7	20.5	43.5	61.4

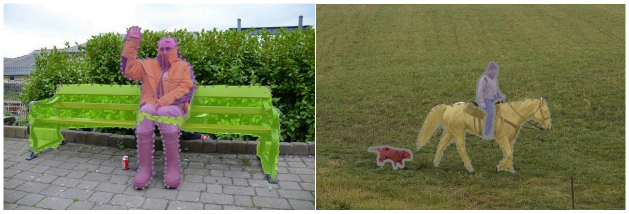


Figure 6. Segmentation of multiple polygon instance based on our EFDTR.

4.5. Qualitative Analysis

As shown in Figure 5, our model effectively segments fine-grained boundaries, such as umbrella handles and suitcase straps, and also performs well in segmenting internal boundaries of objects like bicycles and motorcycles. For relatively dense objects, EFDTR still demonstrates strong segmentation performance.

Additionally, Figure 6 illustrates the segmentation results for instances containing multiple polygons. The model can predict and connect contours, merging multiple polygons into a single output. However, there are still issues with the auxiliary boundaries not being brought close enough to each other, which remains a limitation of our current model.

5. Conclusion

In this paper, we propose EFDTR, a novel polygon-based framework for instance segmentation that leverages learnable Elliptic Fourier Descriptors (EFDs) to model object contours. Our method introduces several key innovations, including a multi-polygon connection strategy based on minimum spanning trees, a phase-aligned Fourier representation for more accurate contour parameterization, and a two-stage decoding architecture for polygon refinement. These contributions enable both precise boundary localization and scalable feature encoding. Experimental results on the COCO dataset demonstrate that EFDTR outperforms existing polygon-based approaches while remaining competitive with mask-based methods. We hope that the phase assignment of the Elliptic Fourier Descriptors will inspire further advancements in polygon contour learning and applications.

Impact Statement

This work introduces EFDTR, an instance segmentation framework using learnable Elliptical Fourier Descriptors for accurate and compact contour representation. By resolving regression ambiguity in the frequency domain, it improves both precision and efficiency. EFDTR shows potential in applications like autonomous driving and medical imaging. As it relies on public datasets and avoids sensitive data, societal risks are minimal, though caution is advised in safety-critical deployments.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62227811, 62473255, 62273235). Additional support was provided by the Deep Blue Program Fund Project of the Second Institute of Oceanography, Ministry of Natural Resources. The authors are also affiliated with the Key Laboratory for System Control and Information Processing, Ministry of Education of China and Shanghai Key Laboratory for Perception and Control in Industrial Network Systems.

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Chen, M., Zheng, H., Lu, C., Tu, E., Yang, J., and Kasabov, N. A spatio-temporal fully convolutional network for breast lesion segmentation in dce-mri. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part VII 25*, pp. 358–368. Springer, 2018.

Chen, M., Zheng, H., Lu, C., Tu, E., Yang, J., and Kasabov, N. Accurate breast lesion segmentation by exploiting spatio-temporal information with deep recurrent and convolutional network. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–9, 2023.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022a.

Cheng, T., Wang, X., Chen, S., Zhang, W., Zhang, Q., Huang, C., Zhang, Z., and Liu, W. Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4433–4442, 2022b.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kirillov, A., Wu, Y., He, K., and Girshick, R. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9799–9808, 2020.

Kuhl, F. P. and Giardina, C. R. Elliptic fourier features of a closed contour. *Computer graphics and image processing*, 18:236–258, 1982.

Lazarow, J., Xu, W., and Tu, Z. Instance segmentation with mask-supervised polygonal boundary transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4382–4391, 2022.

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L. M., and Shum, H.-Y. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023a.

Li, H., Xiong, P., An, J., and Wang, L. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

Li, R., He, C., Li, S., Zhang, Y., and Zhang, L. Dynamask: dynamic mask selection for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11279–11288, 2023b.

Liang, J., Homayounfar, N., Ma, W.-C., Xiong, Y., Hu, R., and Urtasun, R. Polytransform: Deep polygon transformer for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9131–9140, 2020.

Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., Huang, C., and Wang, X. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, pp. 1–23, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Ling, H., Gao, J., Kar, A., Chen, W., and Fidler, S. Fast interactive object annotation with curve-gcn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5257–5266, 2019.

Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R. K., Mahadevan, V., and Manmatha, R. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023.

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

Liu, Z., Liew, J. H., Chen, X., and Feng, J. Dance: A deep attentive contour model for efficient instance segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 345–354, 2021.

Liu, Z., Zhang, X., Liu, G., Zhao, J., and Xu, N. Leveraging enhanced queries of point sets for vectorized map construction. In *European Conference on Computer Vision*, pp. 461–477. Springer, 2025.

Lu, C. *General Keypoint Detection: Few-Shot and Zero-Shot*. PhD thesis, The Australian National University (Australia), 2024.

Lu, C. and Koniusz, P. Few-shot keypoint detection with uncertainty learning for unseen species. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19416–19426, 2022.

Lu, C. and Koniusz, P. Detect any keypoints: An efficient light-weight few-shot keypoint detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3882–3890, 2024.

Lu, C., Xia, S., Huang, W., Shao, M., and Fu, Y. Circle detection by arc-support line segments. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 76–80. IEEE, 2017.

Lu, C., Xia, S., Shao, M., and Fu, Y. Arc-support line segments revisited: An efficient high-quality ellipse detection. *IEEE Transactions on Image Processing*, 29: 768–781, 2019.

Lu, C., Zhu, H., and Koniusz, P. From saliency to dino: Saliency-guided vision transformer for few-shot keypoint detection. *arXiv preprint arXiv:2304.03140*, 2023.

Lu, C., Liu, Z., and Koniusz, P. Openkd: Opening prompt diversity for zero-and few-shot keypoint detection. In *European Conference on Computer Vision*, pp. 148–165. Springer, 2024.

Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., et al. The multimodality cell segmentation challenge: toward universal solutions. *Nature methods*, pp. 1–11, 2024.

Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., and Wang, J. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3651–3660, 2021.

Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., and Zhou, X. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8533–8542, 2020.

Riaz, H. U. M., Benbarka, N., and Zell, A. Fouriernet: Compact mask representation for instance segmentation using differentiable shape decoders. In *2020 25th international conference on pattern recognition (ICPR)*, pp. 7833–7840. IEEE, 2021.

Shi, W., Lu, C., Shao, M., Zhang, Y., Xia, S., and Koniusz, P. Few-shot shape recognition by learning deep shape-aware features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1848–1859, 2024.

Tian, Z., Chu, X., Wang, X., Wei, X., and Shen, C. Fully convolutional one-stage 3d object detection on lidar range images. *Advances in Neural Information Processing Systems*, 35:34899–34911, 2022.

Wang, T., Lu, C., Shao, M., Yuan, X., and Xia, S. Eldet: An anchor-free general ellipse object detector. In *Proceedings of the Asian Conference on Computer Vision*, pp. 2580–2595, 2022a.

Wang, X., Zhang, R., Kong, T., Li, L., and Shen, C. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020.

Wang, Y., Zhang, X., Yang, T., and Sun, J. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2567–2575, 2022b.

Wei, F., Sun, X., Li, H., Wang, J., and Lin, S. Point-set anchors for object detection, instance segmentation and pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 527–544. Springer, 2020.

Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., and Luo, P. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12193–12202, 2020.

Yu, D., Hu, Y., Li, Y., and Zhao, L. Polygongnn: Representation learning for polygonal geometries with heterogeneous visibility graph. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4012–4022, 2024.

Yu, D., Zhang, G., and Zhao, L. Polyhedronnet: Representation learning for polyhedra with surface-attributed graph. *arXiv preprint arXiv:2502.01814*, 2025.

Zhang, H., Wang, Y., Dayoub, F., and Sunderhauf, N. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8514–8523, 2021.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022a.

Zhang, H., Su, Y., Xu, X., and Jia, K. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23385–23395, 2024a.

Zhang, T., Wei, S., and Ji, S. E2ec: An end-to-end contour-based method for high-quality high-speed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4443–4452, 2022b.

Zhang, T., Wei, S., Zhou, Y., Luo, M., Yu, W., and Ji, S. P2pformer: A primitive-to-polygon method for regular building contour extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024.

Zhou, X., Zhuo, J., and Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 850–859, 2019.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Zhu, Y., Chen, L., Xiong, D., Chen, S., Du, F., Hao, J., He, R., and Sun, Z. C-aoi: Contour-based instance segmentation for high-quality areas-of-interest in online food delivery platform. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5750–5759, 2023.