LEARNING TASK-ALIGNED MASK QUERY FOR INSTANCE SEGMENTATION

Bin Fu[1], *Hongliang He*[1,2], *Pengxu Wei*[3], *Jie Chen*[1,2*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Sun Yat-sen University, Guangzhou, China

ABSTRACT

Recently, query-based instance segmentation methods have achieved comparable performance to previous state-of-the-art methods. However, the query lacks the learning of the consistency between classification and segmentation tasks, which may lead to misalignment between classification score and mask quality (i.e., mask IoU) and can not result in a reliable ranking for predictions. In this work, we propose a novel instance segmentation method, termed AlignMask, which effectively learns task-aligned mask queries for instance end-to-end. Specifically, we propose Aligned Query Learning (AQL) to learn task-aligned features for pixel embedding and transformer decoder, which helps segmentation quality estimation of the mask query. We also use Aligned Label Assignment to explicitly align the optimization goals for classification score and mask quality of the query. Extensive experiments on MS-COCO show that our proposed AlignMask achieves competitive performance with state-of-the-art models.

Index Terms— Instance segmentation, Task-aligned query, Transformer on set prediction, Label assignment

1. INTRODUCTION

Instance segmentation is one of the classic and challenging computer vision tasks. Classical instance segmentation methods [1, 2, 3] follow the two-stage paradigm and generate masks from dense bounding boxes. Some recent works [4, 5, 6] use dynamic kernels to generate a dynamic number of masks. However, these methods need to remove duplicated predictions through non-maximum suppression (NMS). Until DETR [7] is proposed, which regards object detection as a set prediction problem and uses one-to-one label assignment. DETR uses learnable query embeddings to represent objects. Such a design effectively eliminates hand-designed anchors and NMS. Recent end-to-end instance segmentation methods [8, 9, 10, 11] are typically based on set prediction.

*Corresponding author. This work was supported in part by the National Key R&D Program of China (2022ZD0118201), Natural Science Foundation of China (Grant 61972217, 32071459, 62176249, 62006133, 62271465), and the Natural Science Foundation of Guangdong Province in China (grant 2019B1515120049).

Fig. 1. Illustration of segmentation results (mask IoU with Ground-truth and classification scores) predicted by Mask2Former [8] (top row) and the proposed AlignMask (bottom row). Ground-truth is indicated by red mask, prediction is green mask, and there overlap is yellow mask. Classification scores predicted by AlignMask better reflect the segmentation quality of mask.

Query-based instance segmentation methods typically generate binary masks associated with global classification score through query embedding and pixel embedding, such as [8, 11, 12]. However, pixel embedding and transformer decoder lack the learning of the consistency between the classification and segmentation tasks, the instability of bipartite matching[13] and strict division of positive and negative samples may give query inconsistent optimization goals for two tasks. As a result, the misalignment between classification score and mask IoU is illustrated in Fig. 1. Due to the unreliability of classification score, a mask with lower mask IoU is vulnerable to be ranked with high priority if it has a high classification score, and the final average precision is consequently degraded.

To address this problem, we propose a novel query generating and learning approach to learn task-aligned mask query, which generates additional query by ground-truth mask and pixel embedding, and through transformer decoder to learn corresponding ground-truth mask and mask quality estimation. We use mask quality to guide the classification optimization goals and adjust sample weights to explicitly align the classification score and mask quality of the query.

The main contributions of this work can be summarized as follows:

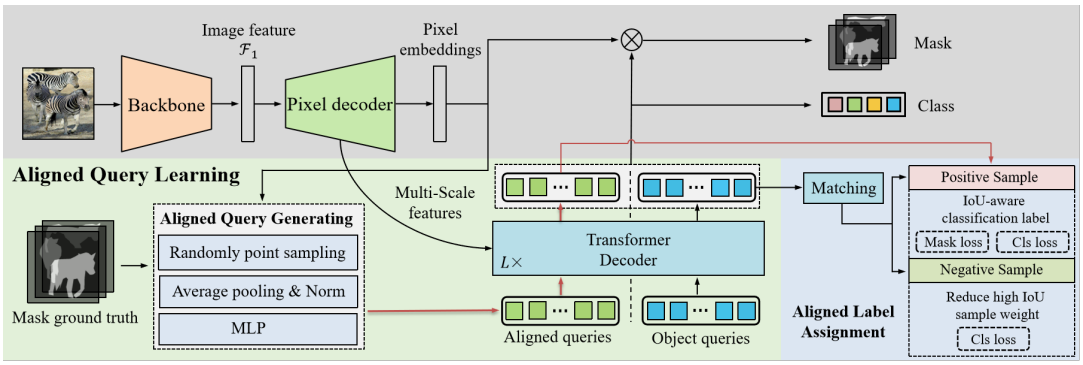

Fig. 2. The overview architecture of our proposed AlignMask effectively learn task-aligned mask queries.

- We propose a novel transformer-based framework AlignMask, which effectively learn task-aligned mask query for instance end-to-end.
- We propose Aligned Query Learning (AQL) to learn task-aligned features for pixel embedding and transformer decoder and use task-aligned learning strategy for set prediction label assignment.
- We conducted extensive experiments on MS COCO [14], showing that our AlignMask achieves competitive performance and validating the effectiveness of our task-alignment approaches.

2. METHOD

2.1. Overview architecture

We integrate the Aligned Query Learning(AQL) into a unified framework based on Mask2Former [8], termed as AlignMask shown in Fig. 2. The framework consists of a backbone, a pixel decoder, and a transformer decoder. Same as Mask2Former, AlignMask is compatible with most backbone architecture, uses the more advanced multi-scale deformable attention Transformer (MSDeformAttn) [15] as default pixel decoder, and uses transformer decoder with masked attention.

AQL is the core of AlignMask, which helps the pixel embedding and transformer decoder learn the consistency between classification and segmentation tasks, use Aligned Label Assignment explicitly align the optimization goals of two tasks for both aligned queries and object queries, and more details of AQL will be given in the next subsection.

2.2. Aligned Query Learning

Aligned Query Generating. Specifically, for one instance i, we generate the aligned query aq_i by the pixel embeddings PE and its ground-truth mask m_i, which can be defined as:

$$\begin{aligned} PPE_i &= \phi(PE) \\ Pm_i &= \phi(m_i) \\ MPf_i &= PPE_i \otimes Pm_i \\ aq_i &= MLP(LN(GAP(MPf_i))) \end{aligned} \tag{1}$$

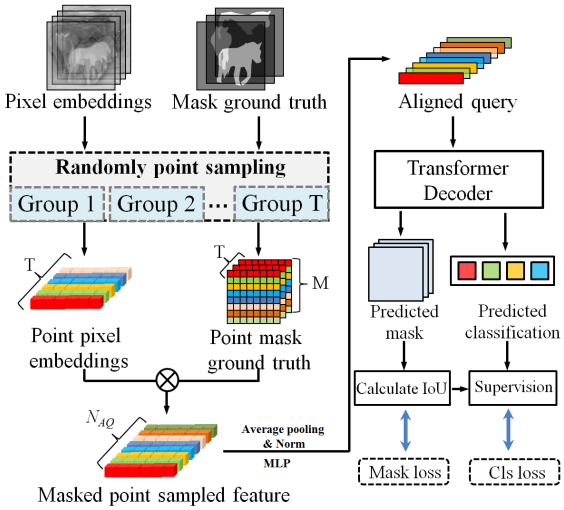

Fig. 3. Illustration of the details of our proposed Aligned Query Learning.

where $PE \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$ denotes the pixel embedding, $m_i \in \{0,1\}^{H\times W}$ denotes the binarized ground-truth mask, $\phi(\cdot)$ denotes randomly point sampling [17] from a uniform distribution, $PPE_i \in \mathbb{R}^{K\times C}$ and $Pm_i \in \mathbb{R}^{K\times 1}$,$MPf_i \in \mathbb{R}^{K\times C}$ denotes the point sampled feature of pixel embedding masked by ground-truth mask, $aq_i \in \mathbb{R}^{C}$ denotes the initial value of an aligned query, MLP is a Multi-Layer Perceptron with 2 hidden layers, LN is Layer Normalization, GAP is global average pooling, H, W are the height and width of the image, K is the number of sampled point. We set K = 12544, i.e., 112 × 112 points.

As Fig. 3 shown, for each image have M instances, we generate T sampled point sets, and each point set generate M aligned queries. Due to the randomly point sampling set, aligned queries generated by same ground-truth mask in different groups are different. This enhances the robustness of mask quality learning through diverse aligned queries. We keep the first N_{AQ} queries generated from all $T \times M$ queries, if the number of queries is small than N_{AQ}, we pad zero vectors. Here we set $N_{AQ} = 100$.

Aligned queries are generated only during training. It obtains predicted masks and mask quality through the same

Table 1. Comparison with state-of-the-art instance segmentation methods on COCO val2017. Backbones pre-trained on ImageNet-22K are marked with †.

Method	Backbone	Query type	Epochs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	GFLOPS	Params
Mask R-CNN [1]	R50	dense anchors	400	42.5	-	-	23.8	45.0	60.0	358	46M
SOLOv2 [6]	R50	dense anchors	36	37.5	58.2	40	15.8	41.4	56.6	-	-
QueryInst [11]	R50	300 queries	36	40.6	63	44	23.4	42.5	52.8	-	-
Mask2Former [8]	R50	100 queries	50	43.7	66.0	46.9	23.4	47.2	64.8	226	44M
AlignMask(ours)	R50	100 queries	50	**45.4(+1.7)**	**66.3**	**50.1(+3.2)**	**25.8**	**48.3**	**65.7**	246	44M
Mask2Former [8]	Swin-S	100 queries	50	46.3	69.3	50.2	25.3	50.3	68.4	313	69M
AlignMask(ours)	Swin-S	100 queries	50	**48.6(+2.3)**	**70.4**	**53.6(+3.4)**	**28.8**	**52.1**	**69.8**	333	69M
QueryInst [11]	Swin-L†	300 queries	50	48.9	74.0	53.9	30.8	52.6	68.3	-	-
Swin-HTC++ [16]	Swin-L†	dense anchors	72	49.5	-	-	**31.0**	52.4	67.2	1470	284M
Mask2Former [8]	Swin-L†	200 queries	100	50.1	**74.1**	54.7	29.9	53.9	**72.1**	868	216M
Mask2Former [8]	Swin-B†	100 queries	50	48.1	72.1	52.1	27.8	52.0	71.1	466	107M
AlignMask(ours)	Swin-B†	100 queries	50	**50.5(+2.4)**	72.9	**55.8(+3.7)**	29.1	**54.1**	71.7	486	107M

transform decoder as object queries. The labels of aligned queries are the GT masks that generate it and IoU between predicted masks and GT masks.

Aligned Label Assignment. We further introduce a label assign method for set predictions to explicitly align the classification score and mask quality of the query. We use mask quality to guide the classification optimization goals for positive samples and adjust sample weights for negative samples.

For the positive samples selected by bipartite matching, we follow recent one-stage methods [18, 19] to adopt quality focal loss (QFL) that make IoU between predicted mask and matched GT as the target of classification. According to the evaluation metric of COCO, an IoU smaller than 0.5 is a sufficient condition for a false prediction. We only reduce negative sample weights for queries that IoU bigger than $\theta = 0.5$. Here IoU denotes the maximum IOU between predicted mask with all GT mask. We set the negative sample weights w_{neg} as a monotonically decreasing function defined within the interval $[0.5, 1]$, which can be defined as:

$$w_{neg} = \begin{cases} 1, & \text{if } IoU < 0.5, \\ -k \times IoU^{\gamma} + b, & \text{if } IoU >= 0.5, \end{cases} \tag{2}$$

where $\gamma = 2$ and w_{neg} passes through the points $(0.5, 1)$ and $(1, 0)$, so the k and b is determined constant.

The final loss function of classification is as follows:

$$\begin{aligned} L_{cls} = & \sum_{i=1}^{N_{pos}} |t_i - s_i|^{\gamma} BCE(s_i, t_i) \\ & + \sum_{j=1}^{N_{neg}} w_{neg}^{j} \times s_j^{\gamma} BCE(s_j, 0) \end{aligned} \tag{3}$$

where i denotes i-th query that matched in bipartite matching, j denotes j-th query that not matched, t is the mask IoU between predicted mask and matched GT, s is the classification score, γ is a hyperparameter and we set $\gamma = 2$, BCE is $Binary\ Cross\ Entropy$ loss.

3. EXPERIMENTS

3.1. Datasets and evaluation metrics

Our experiments are performed on MS COCO 2017 dataset [14]. Following the common practice, all models are trained on the train2017 split and evaluated on the val2017 split. We report the standard mean average precision (AP) result on the COCO validation dataset under different IoU thresholds and object scales.

3.2. Implementation details

We follow the basic settings of Mask2Former on the COCO dataset, except for the form of classification loss. If not stated otherwise, we use 4 V100 GPU to train our models for 50 epochs with a batch size of 16 and 8 V100 GPU for model with Swin-Base backbone. For ablation studies, we train our models for 12 epochs with ResNet-50 backbone. For Aligned Query Learning(AQL), we set $T = 5$ and $N_{AQ} = 100$.

3.3. Main results

Comparison with state-of-the-arts. We compare AlignMask with state-of-the-art models on the COCO dataset in Table 1. AlignMask outperforms a strong Mask R-CNN [1] baseline using large scale jittering (LSJ) augmentation [20, 21] and longer training iterations. And AlignMask achieve 45.4 AP with ResNet-50 backbone[22], outperforming the state-of-the-art Mask2Former [8] 1.7AP. AlignMask have more improvement with Swin-Transformer backbone, 2.3 AP and 2.4 AP for Swin-S and Swin-B backbone[16] respectively. Due to the limitation of computing resources, we do not conduct experiment for Swin-L backbone. AlignMask with Swin-B backbone outperforms the state-of-the-art QueryInst [11], HTC++ [3], and Mask2Former with Swin-L backbone. Note that for a fair comparison, we only consider single-scale inference and models trained with only COCO train2017 set data.

Table 2. Quantitative analysis of the effectiveness of Alignmask on task-alignment with a ResNet-50 backbone

Method	AP	top-2% (10000 predictions)			top-10% (50000 predictions)		
		PCC	Mean Score	Mean IoU	PCC	Mean Score	Mean IoU
Mask2Former [8]	43.7	0.221	0.974	0.848	0.609	0.830	0.480
AlignMask(ours)	**45.4**	**0.291**	0.786	**0.865**	**0.610**	0.470	**0.511**

Table 3. Ablation results for different shared parameters of transformer decoder between aligned query and object query.

Shared Parameters		AP	AP_{50}	AP_{75}
self-attention	masked attention[8]			
No Aligned Query		39.7	58.7	43.3
		40.6	59.7	44.5
✓		40.8	60.2	44.7
	✓	40.8	60.1	44.6
✓	✓	**41.4**	61.0	45.2

Table 4. Ablation results for different number of Aligned Query N_{AQ} and point sampling times T of GT Mask in AQL.

	No AQ	$N_{AQ} = 50$	$N_{AQ} = 100$
No AQ	39.7	-	-
T = 1	-	40.8	41.0
T = 5	-	41.4	41.4

3.4. Ablation studies and Analysis

Quantitative Analysis for Task-alignment. We quantitatively analyze the effect of the proposed methods on the alignment of two tasks. For Mask2Former [8] and our AlignMask, we collect their top-2% confident predictions(10000 predictions) and top-10% confident predictions(50000 predictions) on COCO val2017. As shown in Table 2, we calculate their average classification score and average IoU with ground-truth mask. Alignmask has 0.865 and 0.511 average IoU respectively, which is more than Mask2Former, and the average classification score also closer to the average IoU. To further verify the ability for task-alignment of our method, we calculate the Pearson correlation coefficient on these predictions. Alignmask has 0.291 PCC on top-2% confident predictions, which is more than 0.221 of Mask2Former. This indicates that Alignmask helps bridge the gap between classification confidence and segmentation quality, especially for high confidence predictions. It can also be qualitatively observed in the Fig. 4, our Alignmask reduces predictions that have high classification score but low Mask IoU.

Influence of Components in AQL. As Shown in Table 3, if use different Parameter in both self-attention and masked attention of transformer decoder, AQL still improves the performance. This result shows that AQL can help the pixel embedding learn task-aligned feature. Shared same transformer decoder for aligned query and object query is better than use different Parameter for self-attention or masked attention, it shows that AQL helps the transformer decoder learn the consistency of two tasks.

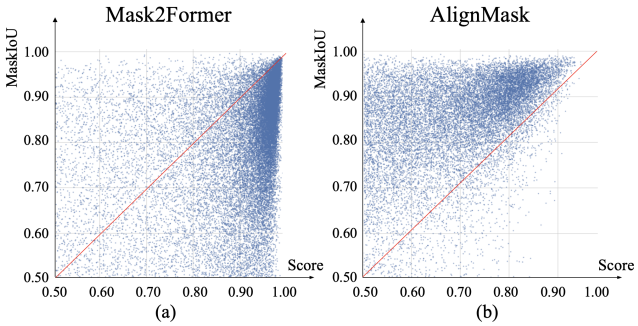

Fig. 4. Comparisons of Mask2Former and our proposed AlignMask. (a) shows the relationship between classification Score and MaskIoU of Mask2Former predictions, and the mask score has less relationship with MaskIoU. (b) shows the results of AlignMask;

Hyper-parameters in AQL. We explore the influence of different Aligned Query number N_{AQ} and point sampling times T of GT Mask in Table 4,. The results show that AQL significantly improve the performance with one point sample time of GT Mask, it shows the effectiveness and robustness of AQL. The performance improves and converges gradually by increasing the number of point samples times and Aligned Query number.

4. CONCLUSION

In this work, we illustrate the misalignment between classification and segmentation tasks in the existing query-based set prediction instance segmentation methods, and propose AlignMask to align the two tasks. In particular, we design Aligned Query Learning(AQL) to learn segmentation quality estimation of the mask query and use a task-aligned learning strategy for end-to-end set prediction methods, which give the query a task-aligned optimization goal and helps the pixel embedding and transformer decoder learn the consistency between the two tasks. With these improvements, AlignMask achieved a 50.5 AP on MS-COCO with Swin-Base backbone, surpassing the state-of-the-art instance segmentation methods with same backbone by a large margin.

5. REFERENCES

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[2] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6409–6418.

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al., "Hybrid task cascade for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4974–4983.

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee, "Yolact++: Better real-time instance segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[5] Zhi Tian, Chunhua Shen, and Hao Chen, "Conditional convolutions for instance segmentation," in *European conference on computer vision*. Springer, 2020, pp. 282–298.

[6] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17721–17732, 2020.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[9] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy, "K-net: Towards unified image segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10326–10338, 2021.

[10] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei, "Solq: Segmenting objects by learning queries," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21898–21909, 2021.

[11] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu, "Instances as queries," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6910–6919.

[12] Bowen Cheng, Alex Schwing, and Alexander Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17864–17875, 2021.

[13] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13619–13627.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[15] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[17] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick, "Pointrend: Image segmentation as rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9799–9808.

[18] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 3490–3499.

[19] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21002–21012, 2020.

[20] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin, "Simple training strategies and model scaling for object detection," *arXiv preprint arXiv:2107.00057*, 2021.

[21] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2918–2928.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.