BSOLO: BOUNDARY-AWARE ONE-STAGE INSTANCE SEGMENTATION SOLO

Yuxuan Zhang[1], *Wei Yang*[1*]

[1] University of Science and Technology of China, School of Computer Science and Technology, China

ABSTRACT

Current one-stage instance segmentation methods ignore the boundary information of masks, resulting in coarse masks that are far from the ground truth. In this paper, we propose a boundary-aware method to refine boundary information, called BSOLO. The core idea of BSOLO is to design a Hungarian-Algorithm-based boundary loss to calculate matching costs between boundaries. This loss effectively measures the difference between boundaries and suits for boundary regression, contributing to generating refined instance masks with high-quality boundaries. Besides, we propose a Feature Fusion Network (FFN) to capture long-range dependency. Through constructing the relationship between pixels, such a module is beneficial for predicting masks with large or uncontinuous region. Furthermore, we introduce a Prototype Attention Module (PAM) for mask assembling through channel attention, which enhances informative features and spotlights important prototypes. To evaluate the performance of BSOLO, we conduct extensive experiments. Experimental results show that BSOLO achieves 39.3 AP on MS COCO test-dev2017, outperforming SOLO and other methods by a large margin. We hope that BSOLO broadens the perspective for designing more valid boundary constraints.

Index Terms— Instance segmentation, boundary loss, BSOLO, Hungarian Algorithm

1. INTRODUCTION

Instance segmentation is a fundamental task in computer vision, which has wide applications in autonomous driving, augmented reality and video surveillance. Currently, one-stage instance segmentation [1, 2, 3] has become a mainstream method, since it leaves out the redundant process of detect-then-segment. SOLO [1], as a classic one-stage paradigm, is not only efficient but also more accurate in mask accuracy.

However, one-stage models have ample room for the improvement of boundary quality due to no boundary constraints, resulting in the detail of the predicted mask being far from the ground truth. To solve the challenging issue, PointRend [4] employs an iterative subdivision algorithm, thus detailed information of boundary is refined. [5] introduces an extra framework to refine boundary quality based on predicted masks. However, these methods require post-processing procedures, which complicate the whole pipeline.

In this paper, we propose BSOLO based on SOLO [1] through directly enhancing boundary constraints. The main idea of BSOLO is to design a specific loss function for boundary regression. To achieve this goal, we propose a boundary loss inspired by Hungarian Algorithm [6, 7]. Compared with BMask-RCNN [8], our boundary loss takes distance factor into account, which is suitable for measuring the difference between boundary sets, as shown in Fig.2(a) and (b). Moreover, we introduce two extra modules, i.e., Feature Fusion Network (FFN) and Prototype Attention Module (PAM). The former targets to refine Feature Pyramid Network (FPN) [9], which fuses multi-scale features effectively and captures long-range dependencies. The latter module focuses channel attention on crucial prototypes for assembling segmented masks. The whole framework is trained on the challenging MS COCO dataset [10] with various semantic categories and numerous objects. In addition, experiments show that our method BSOLO surpasses the vanilla SOLO by 1.5 points on MS COCO test-dev2017.

We summarize our contributions as follows:

- We propose a simple yet effective boundary-aware method, named BSOLO. Through directly constraining boundary information with Hungarian loss, the predicted masks become more accurate and refined.
- To capture long-range dependency, we propose FFN to extend FPN, notably improving large masks accuracy.
- We design PAM for assembling informative prototypes, which enables us to focus on crucial channel features.
- To evaluate the performance of BSOLO, extensive experiments have been conducted. The results demonstrate that our BSOLO outperforms state-of-the-art methods without boundary constraints on accuracy.

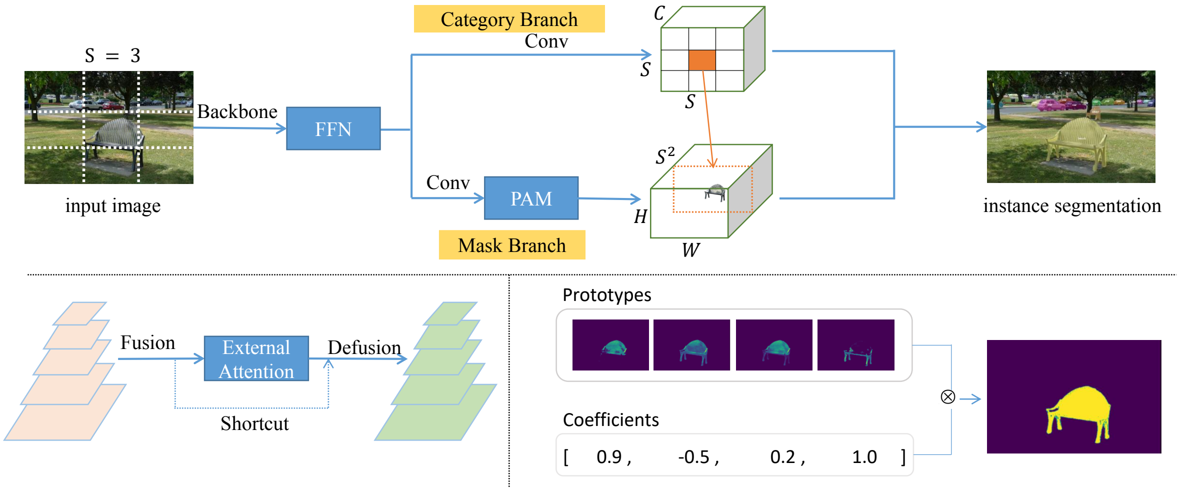
2. BSOLO

In this section, we dive into the detail of our proposed BSOLO. A Hungarian-Algorithm-based boundary loss is introduced. Two extra modules, i.e., FFN and PAM, are proposed for feature fusion and mask assembling, respectively. The overall architecture is shown in Fig.1.

* Corresponding authors:qubit@ustc.edu.cn

This work was supported by the National Natural Science Foundation of China (No. 62172385) and the Anhui Initiative in Quantum Information Technologies (No. AHY150300).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE



(a) Feature Fusion Network(FFN)

(b) Prototype Attention Module(PAM)

Fig. 1. Overall architecture of BSOLO. Based on SOLO [1], we introduce two extra modules to the framework. Here we set $S = 3$ for an example. The orange grid is responsible for predicting the mask of the chair. Two proposed modules, i.e., FFN and PAM, are shown in (a) and (b) respectively. In (a), initial features ranging from different resolutions are shown as red squares, and the fused features are demonstrated as green squares. In (b), we set the prototype number $T = 4$. The predicted mask is assembled after the linear combination of prototypes with coefficients.

2.1. SOLO

SOLO obtains $S \times S \times C$ category map in category branch and $H \times W \times S^2$ segmented map in mask branch, as shown in Fig.1, where H and W represent height and width of the input image. C and S denote total categories and grids. Each grid is responsible for predicting a mask in segmented maps.

2.2. Boundary Loss

Boundary set: Following the traditional edge detection method, we first obtain the ground truth boundary set $\{y_b^i\}_{i=1}^N$ by Laplacian operation on the basis of the ground truth mask $\{y_m^i\}_{i=1}^N$, where $y_b^i \in R^{N_1 \times 2}$ denotes the *boundary set* of the i-th instance's coordinates among total N instances for the input image, and N_1 represents the total number of boundary pixels of the i-th instance mask. y_m^i represents the i-th instance mask with map size $H \times W$. Then, we calculate the predicted boundary set $\{p_b^i\}_{i=1}^N$ by means of the dilation and the erosion residue operators. $p_b^i \in R^{N_2 \times 2}$ represents the set of predicted boundary pixels coordinates, where N_2 denotes the number of boundary pixels coordinates in predicted masks.

Issue of existing boundary loss: Most existing methods [8, 11, 12] take advantage of cross-entropy loss or dice loss for boundary regression. However, these measurements ignore the distance between p_b^i and y_b^i. That is, such losses are not suitable for measuring the difference between boundaries. Fig.2(a) and (b) demonstrate this issue.

Hungarian-based boundary loss: To deal with the problem above, we propose a novel boundary loss based on Hungarian Algorithm [6]. The core idea is to calculate the minimum cost of matching p_b^i with y_b^i. Once the matching cost decreases, the distance between p_b^i and y_b^i will get closer. In this fashion, the distance factor is taken into consideration for reasonable boundary regression.

However, Hungarian loss is based on a crucial condition, namely p_b^i and y_b^i ought to be similar. Refer to Fig.2(c) and (d), assuming that p_b^i and y_b^i are totally different in terms of shape and location, such a matching mechanism will be unsuitable for boundary regression. Therefore, we utilize an extra constraint on boundary loss, i.e., intersection over union(IoU) of ground truth and predicted mask. We employ IoU as the measurement to judge the similarity between predicted mask and ground truth mask. Thus the boundary loss $\mathcal{L}_{bound}$ is calculated by:

$$\mathcal{L}_{bound} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{IoU(p_m^i, y_m^i)>t\}} \mathcal{L}_{match}(p_b^i, y_b^i)$$

where $\mathcal{L}_{match}$ denotes the minimum cost of pair-wise bipartite matching. p_m^i and y_m^i represent the predicted mask and ground truth mask of the i-th instance respectively. t represents a threshold that filters out boundary constraints on those predicted masks with low IoU with the corresponding ground truth. Boundary loss is ignored if mask IoU is not larger than t. In other words, boundary loss targets to refine boundary information for those coarse masks that are similar with the ground truth in terms of shape and location.

Multi-task total loss: We define a multi-task loss for training by incorporating classification loss, mask loss and boundary loss:

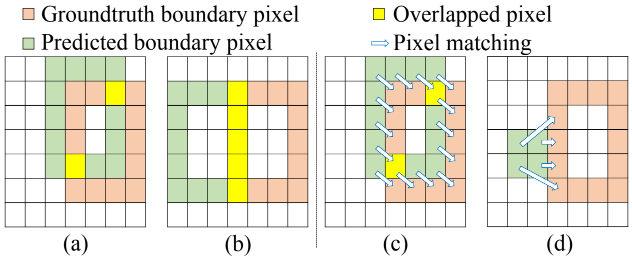


Fig. 2. Distance and shape consideration in boundary loss. Prediction in (a) is obviously better than (b), but the overlapped boundary pixels in (a) are less than (b). Thus boundary loss should contain not only overlapped area but also distance between boundary sets. The shape of boundary in (c) is similar to the ground truth, while (d) is totally different. Therefore, (d) is not suitable for direct Hungarian matching, which means the shape factor is also supposed to be considered in boundary loss.

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{mask} + \gamma\mathcal{L}_{bound}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{mask}$ denote classification loss and mask loss which are described in SOLO [1] respectively. Classification loss $\mathcal{L}_{cls}$ is a pixel-wise category binary cross-entropy loss. Mask loss $\mathcal{L}_{cls}$ is calculated with a pixel-wise dice loss. λ and γ represent two hyperparameters for adjusting the balance of total loss. We set $\lambda = 3$ and $\gamma = 1$ during training period.

2.3. Feature Fusion Network

Feature Pyramid Network (FPN) is widely applied in FCOS [17] and RetinaNet [18], which achieves great success in object detection. Such a structure takes advantage of extracting multi-scales features. To be specific, low-level features contain global spatial information while high-level features include detailed semantic meaning. In this work, we propose a Feature Fusion Network (FFN) to simplify the fusion of FPN and capture short-range and long-range dependency for each pixel.

As shown in Fig.1(a), features resolutions from shallow to deep range from $1/8$ to $1/128$ compared with the input image. We first interpolate features with different levels to $1/8$ resolution, and add them up for feature fusion. Then, an external attention module (EA) [19] is utilized to capture both short-range and long-range pixel dependencies. After that, we employ a reverse operation called "defusion" to rescale features to the initial sizes. Moreover, the enhanced features are added by the initial features through a shortcut connection.

FFN provides an effective way for the fusion of low-level and high-level features. Meanwhile, it takes dependencies among pixels into consideration. Such a self-attention based method contributes to broadening the perceptive field and generating high-quality masks with large size or uncontinuous region.

2.4. Prototype Attention Module

We introduce a prototype attention module (PAM) to the mask branch of SOLO for mask assembling. After passing through several convolution layers in the mask branch, the feature map $F \in R^{H\times W\times M}$ is divided into two subbranches: One is to employ a series of 3×3 conv layers to generate prototype matrix $\{P^i\}_{i=1}^{S^2}$ for each grid in category branch, where $P^i \in R^{H\times W\times T}$ represents the prototypes of the i-th grid and T represents the number of prototypes; The other branch employs a 3×3 conv layer and then a global-average-pooling layer to obtain coefficient vector $\{V^i\}_{i=1}^{S^2}$, where $V^i \in R^{1\times T}$ denotes the coefficients of i-th grid.

Then, the output of mask branch M_{out} is calculated by:

$$M_{out} = \bigcup_{i=1}^{S^2} \sigma(conv(P^i(V^i)^T))$$

where σ is a normalization operator for mask predicting. $\bigcup$ denotes concatenating operation to build the output $M_{out} \in R^{H\times W\times S^2}$. PAM is followed by a 3×3 conv layer and the pixels are finally restricted to 0-1. At last, we filter out those masks with classification confidence lower than the threshold 0.5. The survival masks form p_m for loss calculation.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and metrics

To validate our proposed method, we train BSOLO on the commonly-used MS COCO [10] instance segmentation dataset train2017. We report COCO-style mask AP, AP_{50}/AP_{75} and $AP_S/AP_M/AP_L$ of test-dev2017 as our main results. Furthermore, we set several ablation studies based on val2017 using the same metrics. Training details follow the pattern of [1].

3.2. Main Result

We compare our proposed method with several state-of-the-art instance segmentation models on the dataset MS COCO test-dev2017, as shown in Table 1. BSOLO achieves 39.3 AP using the backbone ResNet-101 [20] with our proposed FFN, which surpasses all two-stage and one-stage methods without any boundary constraints. Meanwhile, our method outperforms all the previous boundary-aware segmentation methods including BMask-RCNN [8].

3.3. Ablation study

FFN: We evaluate the performance of FFN compared with the widely-applied FPN, as shown in Table 2. Several modules that can catch long-range dependency, i.e., traditional self-attention module (SA) [21], non-local module (NL) [22] and external attention module (EA) [19], are examined in FFN. We find that FFN surpasses FPN by a large margin on AP_L.

Table 1. Instance segmentation mask AP on COCO test-dev2017 for different methods. All entries represent single-model results. DCN means deformable convolutions network applied.

method	backbone	epoch	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCIS [13]	ResNet-101-C5-dilated	3x	29.2	49.5	-	7.1	31.3	50.0
Mask-RCNN [14]	ResNet-101-FPN	3x	35.7	58.0	37.8	15.5	38.1	52.4
PointRend [4]	ResNet-101-FPN	3x	36.3	-	-	-	-	-
CondInst [15]	ResNet-101-FPN	3x	39.1	60.9	42.0	**21.5**	41.7	50.9
BlendMask [16]	ResNet-101-FPN	3x	37.8	58.8	40.3	18.8	40.9	53.6
PolarMask [2]	ResNeXt-101-FPN-DCN	3x	36.2	59.4	37.7	17.8	37.7	51.5
SOLO [1]	ResNet-101-FPN	3x	37.8	59.5	40.4	16.4	40.6	54.2
BMask-RCNN [8]	ResNet-101-FPN	3x	37.7	59.3	40.6	16.8	39.9	54.6
BSOLO	ResNet-101-FFN	3x	**39.3**	**61.2**	**42.4**	19.5	**43.2**	**55.6**

Besides, FFN with EA performs better compared with other self-attention-based FFN on val2017. To be specific, FFN with EA achieves 36.6 AP on val2017, which is 0.8 AP higher than commonly-used FPN. As expected, the accuracy result has been improved with the usage of FFN and a proper dependency capturing module.

Table 2. FPN vs FFN with different self-attention modules.

method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FPN	35.8	57.1	37.8	15.0	**38.7**	53.6
FFN+NL [22]	36.4	57.4	**38.4**	15.3	38.5	54.6
FFN+SA [21]	35.4	55.9	37.0	14.4	38.2	54.3
FFN+EA [19]	**36.6**	**57.7**	38.3	**15.6**	**38.7**	**54.8**

Prototypes: Another significant component of our proposed method is the number of prototypes. We draw that the result can be further improved through selecting a proper prototype number. We set $T = 1, 2, 4, 8$ respectively in this ablation experiment and the results are shown in Table 3. As reported in Table 3, $T = 4$ achieves 36.4 AP on val2017, which outperforms others. Note that, a larger T is meaningless since it brought too much burden for network training.

Table 3. Different numbers of prototypes.

T	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	35.5	57.2	37.6	15.1	38.4	53.7
2	36.0	57.4	38.0	15.2	38.9	54.1
4	**36.4**	**57.9**	**38.3**	**15.7**	**39.2**	54.3
8	36.1	57.8	38.1	15.5	38.7	**54.4**

Boundary Loss: Table 4 compares several different loss functions for seeking the appropriate one as boundary loss. No boundary loss (-), binary cross-entropy (BCE), dice loss (DL), Hungarian loss without the IoU constraint (vanilla HL) and our proposed Hungarian Loss (HL) are taken into consideration. As reported in Table 4, HL that considers the IoU factor surpasses others by a large margin, which lives up to our expectation that such a distance-based loss is suitable for boundary preserving.

IoU threshold: Another important component is the threshold t, as described in Section 2.2. If the threshold is too small, Hungarian matching will be not appropriate, as shown in Fig.2(a) and (b). Once t is closer to 1, only few predicted masks are restricted by the boundary loss. To examine the best choice, we set $t = 0.3, 0.4, 0.5, 0.6, 0.7$, respectively, and the results are reported in Table 5. $t = 0.5$ is recognized as the ideal choice for our method, which achieves 36.8 AP, compared with other thresholds. That is, the boundary loss is taken into account for refinement if the IoU between predicted mask and ground truth is higher than threshold $t = 0.5$.

Table 4. Different boundary loss functions.

boundary loss	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
–	35.7	57.0	37.6	14.9	38.5	53.6
BCE	33.6	55.2	34.9	13.4	35.5	51.0
DL	35.9	56.8	37.8	15.2	38.5	54.0
vanilla HL	31.8	51.7	33.1	12.3	33.5	49.4
HL	**36.8**	**57.8**	**38.5**	**15.9**	**39.4**	**55.2**

Table 5. Different thresholds in Hungarian boundary loss.

t	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.3	33.7	54.8	36.4	14.4	37.5	50.4
0.4	36.0	56.6	37.7	15.4	38.7	53.6
0.5	**36.8**	**57.8**	**38.5**	15.5	**39.2**	**55.2**
0.6	36.6	57.4	38.3	**15.7**	38.9	54.7
0.7	36.5	57.2	38.3	15.5	39.0	54.5

4. CONCLUSIONS

In this paper, we proposed BSOLO, a novel boundary-aware method for instance segmentation. By designing a Hungarian-Algorithm-based boundary loss, the distance between boundary sets can be measured for reasonable boundary regression. Our proposed boundary loss outperforms other boundary losses on accuracy. Moreover, we introduced two plug-in modules, i.e., FFN and PAM. By utilizing EA, FFN captures long-range dependency. PAM pays attention to the crucial channels for mask assembling. Experiments show that BSOLO surpasses SOLO and other mainstream instance segmentation models. In addition, we hope that the proposed boundary loss can broaden the perspective for designing more effective boundary constraints.

References

[1] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li, "Solo: Segmenting objects by locations," in *European Conference on Computer Vision*. Springer, 2020, pp. 649–665.

[2] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, and Xuebo Liu, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12193–12202.

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9157–9166.

[4] 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.

[5] Chufeng Tang, Hang Chen, Xiao Li, and Jianmin Li, "Look closer to segment better: Boundary patch refinement for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13926–13935.

[6] Harold W Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, and Nicolas Usunier, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[8] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu, "Boundary-preserving mask r-cnn," in *European conference on computer vision*. Springer, 2020, pp. 660–676.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[10] 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.

[11] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji, "Salient object detection with pyramid attention and salient edges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1448–1457.

[12] Mengyang Feng, Huchuan Lu, and Errui Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1623–1632.

[13] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei, "Fully convolutional instance-aware semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2359–2367.

[14] 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.

[15] Zhi Tian, Chunhua Shen, and Hao Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 282–298.

[16] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8573–8581.

[17] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[19] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *arXiv preprint arXiv:2105.02358*, 2021.

[20] 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.

[21] Ashish Vaswani, Noam Shazeer, and Parmars, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.