Keypoints-aware Object Detection

Ayush Jaiswal, Yue Wu, Pradeep Natarajan, Premkumar Natarajan
Amazon Alexa Natural Understanding
Manhattan Beach, CA, USA
{ayujaisw, wuayue, natarap, premknat}@amazon.com

Abstract

We propose a new framework for object detection that guides the model to explicitly reason about translation and rotation invariant object keypoints to boost model robustness. The model first predicts keypoints for each object in the image and then derives bounding-box predictions from the keypoints. While object classification and box regression are supervised, keypoints are learned through self-supervision by comparing keypoints predicted for each image with those for its affine transformations. Thus, the framework does not require additional annotations and can be trained on standard object detection datasets. The proposed model is designed to be anchor-free, proposal-free, and single-stage in order to avoid associated computational overhead and hyperparameter tuning. Furthermore, the generated keypoints allow for inferring close-fit rotated bounding boxes and coarse segmentation for free. We propose to evaluate our model on the standard PASCAL VOC and MS COCO datasets and metrics along with new specialized experiments designed for assessing robustness to translation and rotation. Finally, the segmentation utility of generated keypoints would be evaluated on the MS COCO dataset.

1 Introduction

Object detection is formulated as the localization and classification of objects in an image, where the former is typically encoded as rectangular bounding boxes that contain object instances. Intuitively, this task is a core component of automated visual scene understanding, allowing for images to be parsed in terms of objects. As such, the field has amassed extensive research interest in the last couple of decades [1], with large improvements in performance.

Existing detectors can be categorized into two-stage and one-stage models, where the former first generates region proposals in an image followed by precise localization, while the latter directly predicts detections across the entire image. Both model families commonly comprise a backbone network, e.g., ResNet [2] pretrained on ImageNet [3], which generates convolutional features from the image, optionally followed by a variant of Feature Pyramid Networks [4] for combining information from different depths of the backbone. These features are then used to predict the object classes and bounding boxes through sibling head modules. A majority of detectors employ anchor or default boxes of various sizes and shapes with bounding box predictions made in terms of adjustments to these anchors. However, methods for anchor-free detection have been proposed recently [5, 6], which vastly reduce hyperparameters related to anchors. Recent works [1] in object detection have proposed improvements to all the aforementioned components, training schemes including better losses, sampling strategies, etc., and new approaches for corner-points-based anchor-free detection.

In this work, we propose a method for improving the robustness of detectors by training them to explicitly focus on object keypoints that are invariant to affine transformations of the image and objects contained in them. Figure 1 describes the high-level approach. Specifically, we propose a framework for training one-stage anchor-free and proposal-free detectors that treats each pixel as an object center and produces corresponding object keypoints and classes at each scale of the

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

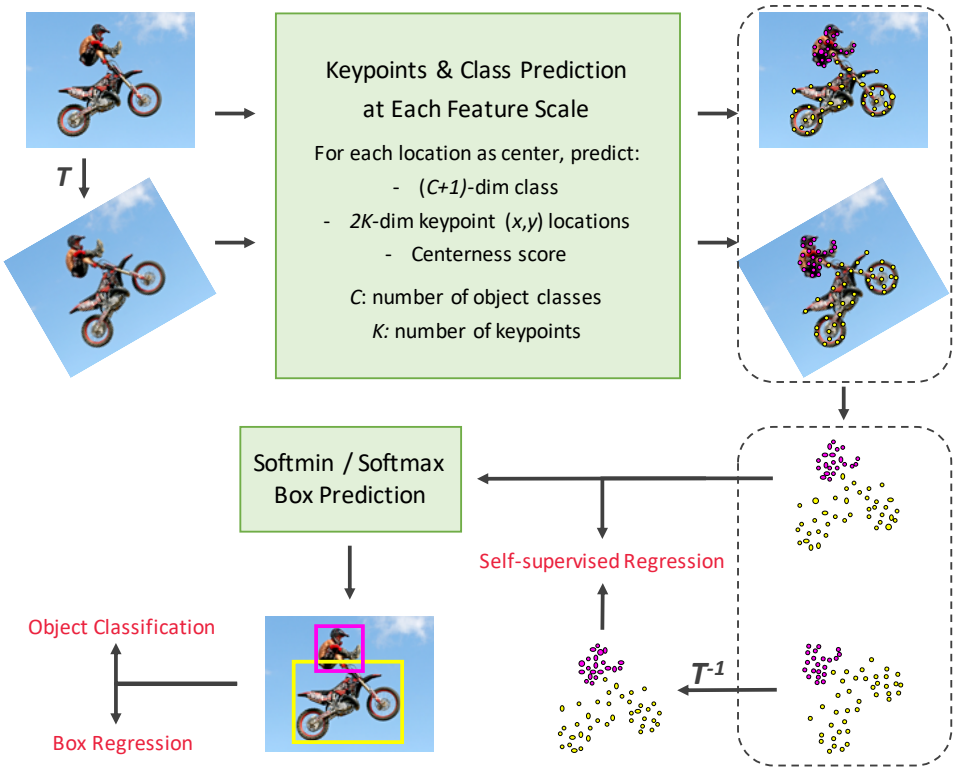

Figure 1: The proposed keypoints-aware object detection framework. The model treats each location on the image as an object center and predicts (1) object class, (2) keypoints, and (3) a centerness score. Bounding boxes are then computed through softmin and softmax operations on the keypoints. Classification, box regression, and centerness are learned in the standard supervised learning setup. Keypoints, on the other hand, are learned through self-supervision by comparing keypoints generated for images and their affine-transformed variants. Best viewed digitally and zoomed in.

backbone network. The conventional bounding boxes can then be derived from the keypoints through straightforward maxima and minima operations.

The object classification and bounding box regression tasks are trained in the standard supervised fashion, along with a centerness loss [5] to avoid predictions far from object centers. In contrast, to avoid dependence on ground-truth keypoint annotations, we learn their prediction in a self-supervised manner. We first predict keypoints for the original image and its affine transformation. We then transform keypoints of the transformed image to coordinates on the original image by applying the inverse transformation on them. The self-supervised loss then becomes the distance between the keypoints of the original image and those from the inverse transformation. Besides the regular bounding box localization, the proposed framework provides close-fit rotated bounding boxes and coarse segmentation masks for free, as illustrated in Figure 2. These can be achieved with simple computational geometery techniques and can benefit downstream applications.

We propose to evaluate the proposed framework on the standard PASCAL VOC [7] and MS COCO [8] benchmark datasets. Furthermore, we propose to create new evaluation-only versions of these datasets by applying various affine transformations to their images in order to compare the robustness of our model with the state-of-the-art methods in terms of invariance to the said transformations. Finally, we propose to additionally evaluate the quality of the coarse segmentations generated from the keypoints on the MS COCO dataset in order to quantify their downstream utility.

2 Related Work

A number of works [9–13] have been proposed recently that regress bounding box corner locations and object centers directly instead of relying on anchors. Corner proposal [14] has also been employed

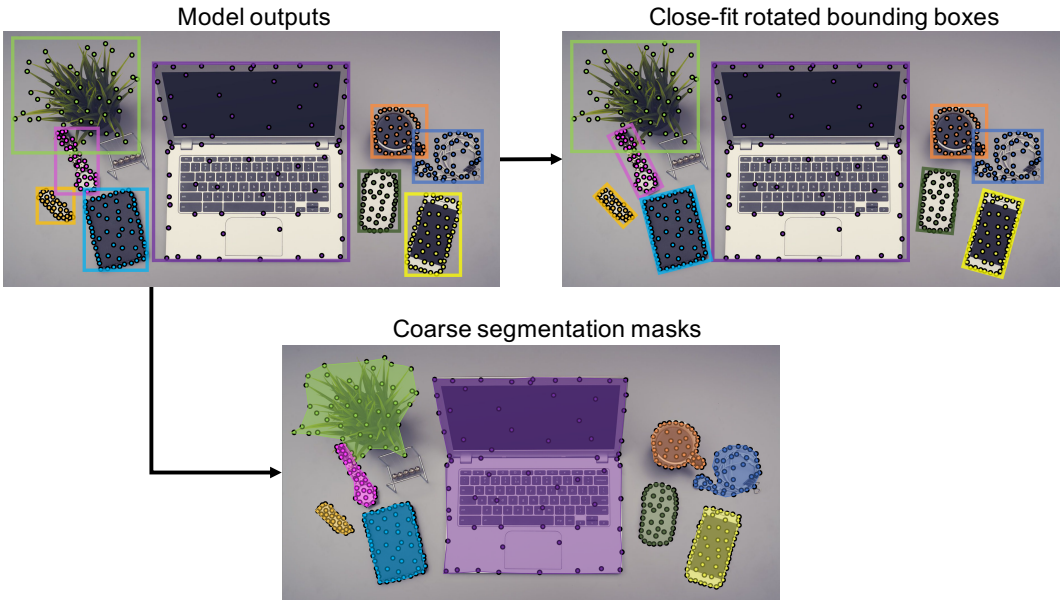

Figure 2: Auxiliary benefits of the proposed model. The model outputs keypoints and rectangular bounding boxes along with the class labels. The keypoints can then be used to calculate close-fit rotated bounding boxes and coarse segmentation masks using computational geometry algorithms.

as a replacement for region proposal in two-stage object detectors. These methods are termed as keypoint-based object detectors, where keypoints refer to the coordinates of box corners and centers. In contrast, our work generates object keypoints that lie on the spatial regions spanned by the objects instead of bounding box corners, which often lie outside object regions. Furthermore, keypoints in our method are learned in a self-supervised fashion without requiring additional annotations.

Yang *et al.* [15] propose a detection method that aggregates information from keypoints in images through deformable convolutions with learned offsets. In contrast, our method predicts keypoints as model outputs for each object in the image. Kulkarni *et al.* [16] also predict object keypoints in an unsupervised way but their method is designed for tracking objects in videos, requiring pairs of frames as inputs. Jakab *et al.* [17] predict keypoints by reconstructing a target image from a source image and keypoints extracted from the target through a bottleneck procedure. Our method, instead, explicitly predicts keypoints for each object in the image.

Anchor-points detectors such as [18] predict anchor points on the image and then bounding boxes as vertical and horizontal offsets from the anchor locations. Wei *et al.* [19] design object-specific anchors as sets of points and make predictions relative to these point-set anchors. Our work, on the other hand, is similar in spirit to [5] and treats each location as a potential object center for making detection predictions. The self-supervised keypoint prediction in our work also falls under the umbrella of consistency-based learning through data augmentation [20].

3 Keypoints-aware Model for Object Detection

In the following sections, we describe (1) the complete model architecture, (2) self-supervised training for learning to predict object keypoints, and (3) computation of close-fit rotated bounding boxes and coarse segmentation masks as auxiliary *post hoc* outputs. In the rest of the text, we denote the set of predicted keypoints as P and its cardinality as K.

3.1 Model Architecture

The proposed method is compatible with any existing detection framework, but we design it like [5] as a one-stage proposal-free and anchor-free model for validating its effectiveness. These attributes of the model significantly reduce [5] computational overhead and sensitive hyperparameters. The model

starts with a backbone convolutional feature extractor, for which we experiment with two commonly used alternatives [2] – ResNet-50 and ResNet-101, pretrained on ImageNet [3]. The features are then passed through a standard Feature Pyramid Network (FPN) [4] to combine features from various depths of the backbone such that both low-level and high-level image features are utilized at each scale of prediction. Additional convolutional and upsampling operations are applied to ensure that all features have the same spatial dimensions as the original image and a fixed number of channels.

Objects of different sizes are detected at different feature levels, as standard in one-stage detectors. In this work, we follow the approach of [5] to treat each location on a feature map as an object center and make predictions relative to center locations. Hence, for each feature level, the FPN outputs are then fed to three sibling modules that make predictions for each location – (1) object-class prediction, (2) keypoints prediction as $2 \times K$ channels representing $(\Delta x, \Delta y)$ distances to keypoints at each location, and (3) centerness prediction, which prevents predicted detections from being far from the center location. Object bounding boxes are then derived from the keypoints as top-right and bottom-left coordinates through softmax and softmin operations, which facilitates gradient flow through all keypoints during backpropagation. The training losses for classification, box regression, and centerness are Focal loss [21] ($\mathcal{L}_{\text{Foc}}$), IoU loss [22] ($\mathcal{L}_{\text{IoU}}$), and Centerness loss [5] ($\mathcal{L}_{\text{Cen}}$), respectively. We generate prediction targets at different scales, limiting output ranges at each scale, and resolving overlapping ground-truth boxes by picking the one with the minimal area, following [5].

3.2 Self-supervised Training for Keypoint Predictions

In order to avoid dependence on ground-truth annotations for object-keypoints, we train the model to predict rotation and translation invariant keypoints in a self-supervised manner. Specifically, given an image I, we first generate its variant I_T by applying an affine transformation $I_T = TI$. Next, we generate keypoints (at each location and each feature level) P and P_T for I and I_T, respectively, using the method described in Section 3.1. The keypoints P_T are then transformed back to the space of I by applying the inverse transformation $P'_T = T^{-1}P_T$. Intuitively, P and P'_T should match for the keypoints to be transformation invariant, and the smooth L_1 loss ($\mathcal{L}_{\text{SL}_1}$) between them can be used as the self-supervised loss. However, P and P'_T represent sets of keypoints with no inherent ordering. We employ Hungarian matching [23] (HM) to find the best alignment between the two sets and minimize the distances between the matched points. Furthermore, T could throw parts of the image out of frame and introduce empty space in the image frame. We mitigate corrupt losses from such locations by computing a mask m and backpropagating gradients only from valid object-center locations. Thus, the self-supervised keypoint-prediction loss $\mathcal{L}_{\text{Key}}$ for each object-center location o and at each scale s can be written as shown in Equation (1). The complete training loss is presented in Equation (2), with β denoting the Lagrange multiplier for the new semi-supervised loss $\mathcal{L}_{\text{key}}$.

$$\mathcal{L}_{\text{Key}}^{(o,s)} = \sum_{\left(P_i^{(o,s)}, (P'_T)_j^{(o,s)}\right) \in \text{HM}(P^{(o,s)}, (P'_T)^{(o,s)})} m^{(i)} \mathcal{L}_{\text{SL}_1}\left(P_i^{(o,s)}, (P'_T)_j^{(o,s)}\right) \tag{1}$$

$$\mathcal{L} = \mathcal{L}_{\text{Foc}} + \mathcal{L}_{\text{IoU}} + \mathcal{L}_{\text{Cen}} + \beta \mathcal{L}_{\text{Key}} \tag{2}$$

We propose to further evaluate the model's sensitivity to K through ablation studies. We plan to determine good values of K in a data-driven manner by analyzing the number of vertices in segmentation masks in benchmark datasets and devising heuristic functions based on the results.

3.3 Auxiliary Derived Outputs

The proposed framework provides close-fit rotated bounding boxes and coarse segmentation masks without explicitly training for them with supervision. In order to generate close-fit rotated bounding boxes, we employ the Rotating Calipers method [24, 25] for computing the oriented minimum bounding box of a point-set, and apply it to the predicted object-keypoints. Generating segmentation masks, on the other hand, involves a simple calculation of the boundary of the predicted keypoints.

4 Experimental Evaluation

We evaluate the performance gains achieved from the proposed self-supervised keypoints-aware training and prediction strategy quantitatively on (1) object detection efficacy and (2) invariance to

translations and rotations. Experiment (2) would additionally quantify the improvements in invariance to affine transformations of objects in images due to our keypoints-based approach. Performance is compared with the standard FCOS object detector as the baseline. Furthermore, we provide qualitative results of the auxiliary close-fit rotated bounding boxes and coarse segmentation masks. These predictions are "free of cost" and are not expected to outperform models trained with rotated box and pixel-level annotations, respectively. Finally, we perform ablation studies for evaluating gains due to the proposed Hungarian matching, and the model's sensitivity to the number of keypoints.

Object detection efficacy is benchmarked on the standard PASCAL VOC [7] (training on VOC 07+12 training set; testing on VOC 07 validation set) and MS COCO [8] (training on COCO 2017 training set; testing on COCO 2017 validation set) datasets using the standard mean average precision (mAP) metrics defined for these datasets. In order to evaluate translation and rotation invariance, we generate two new datasets from both VOC and COCO test sets (total four) – one for translation and another for rotation. In case of the former, we randomly shift the image along X and Y axes, while in the latter case, we rotate each image by a random angle in the range $[-30°, 30°]$. The resulting images are cropped to remove empty-space artifacts from the transformations. mAP metrics are reported separately for each case. The goal here is to achieve high mAP scores for transformed images as indicators of robustness and invariance.

5 Conclusion

We have proposed a new keypoints-aware model for robust object detection invariant to affine transformations. The model predicts keypoints for each object, which are then used to compute the bounding boxes using simple minima and maxima operations. We have described the training process wherein box regression, classification and centerness are trained in a supervised manner while keypoints are learned through self-supervision by comparing keypoints generated for images and their affine-transformed variants. The proposed model also provides close-fit rotated bounding boxes and coarse segmentation masks for free. We have further proposed the experiment setup involving not only the standard object detection benchmarking but also quantification of invariance to translations and rotations, and evaluation of the utility of generated segmentation masks.

References

[1] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.

[2] 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*, pages 770–778, 2016.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[4] 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*, pages 2117–2125, 2017.

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

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[8] 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*, pages 740–755. Springer, 2014.

[9] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[10] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019.

[12] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10519–10528, 2020.

[13] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019.

[14] Kaiwen Duan, Lingxi Xie, Honggang Qi, Song Bai, Qingming Huang, and Qi Tian. Corner proposal network for anchor-free, two-stage object detection. *arXiv preprint arXiv:2007.13816*, 2020.

[15] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9657–9666, 2019.

[16] Tejas D Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. In *Advances in neural information processing systems*, pages 10724–10734, 2019.

[17] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in neural information processing systems*, pages 4016–4027, 2018.

[18] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. Soft anchor-point object detection. *arXiv preprint arXiv:1911.12448*, 2019.

[19] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. *arXiv preprint arXiv:2007.02846*, 2020.

[20] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] 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*, pages 2980–2988, 2017.

[22] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016.

[23] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955.

[24] Franco P Preparata and Michael I Shamos. *Computational geometry: an introduction*. Springer Science & Business Media, 2012.

[25] Godfried T Toussaint. Solving geometric problems with the rotating calipers. In *Proc. IEEE Melecon*, volume 83, page A10, 1983.