Appendix of Dense RepPoints: Representing Visual Objects with Dense Point Sets

Ze Yang[1,2†⋆], Yinghao Xu[3,4†*], Han Xue[5†*], Zheng Zhang[7]
Raquel Urtasun[6], Liwei Wang[1], Stephen Lin[7], and Han Hu[7]

[1] Peking University
[2] Zhejiang Lab
yangze@pku.edu.cn, wanglw@cis.pku.edu.cn
[3] Zhejiang University
[4] The Chinese University of Hong Kong
justimyhxu@gmail.com
[5] Shanghai Jiao Tong University
xiaoxiaoxh@sjtu.edu.cn
[6] University of Toronto
urtasun@cs.toronto.edu
[7] Microsoft Research Asia
{zhez,stevelin,hanhu}@microsoft.com

1 Network architecture

In this section, we describe the network architecture of Dense RepPoints in details. Similar to [2], we use a center point based initial object representation and utilize Dense RepPoints as the intermediate feature sampling locations. The overall architecture is illustrated in Figure 2 of main paper, using an FPN backbone like in [1, 2], where feature pyramid levels from 3 (downsampling ratio of 8) to 7 (downsampling ratio of 128) are employed. The head architecture is illustrated in Figure 1.

In addition to the class head and localization head, we introduce an optional attribute head to predict the score of each point. The localization subnet first computes offsets $\mathbf{o}_1$ for the Dense RepPoints, then the refinement and attribute predictions are obtained by bilinear sampling on the predicted refine fields $\mathbf{o}_2$ and attribute maps $\mathbf{a}_1$ with sampling locations based on $\mathbf{o}_1$. For the classification branch, we use group pooling to sample the features of Dense RepPoints with sampling locations based on $\mathbf{o}_1$, then fully-connected layers are used to predict the classification results. We use the same label assignment approach as in [2]. For the additional attribute branch, we use per-point binary cross entropy for the foreground / background attribute prediction.

⋆ Equal contribution. †This work was done when Ze Yang, Yinghao Xu and Han Xue were interns at Microsoft Research Asia.

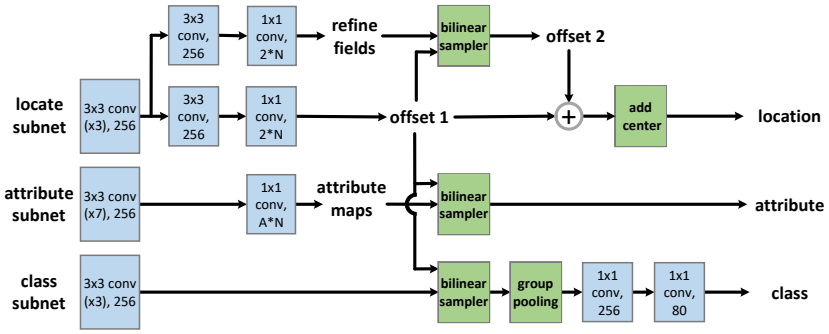

Fig. 1. Illustration of the head design. The attribute head is optional.

2 Details of triangulation post-processing

Dense RepPoints is defined as $\mathcal{R} = \{(x_k, y_k, \mathbf{a}_k)\}_{k=1}^{n}$, where $\mathbf{a}_k$ is the foreground score associated with the k-th point. We use Delaunay triangulation to triangulate the image space. Then, for any image pixel (x, y) inside of a triangle with vertices $(x_i, y_i, \mathbf{a}_i)$, $(x_j, y_j, \mathbf{a}_j)$ and $(x_k, y_k, \mathbf{a}_k)$, its barycentric coordinate $(\lambda_i, \lambda_j, \lambda_k)$ satisfies:

$$\begin{aligned} \lambda_i x_i + \lambda_j x_j + \lambda_k x_k &= x \\ \lambda_i y_i + \lambda_j y_j + \lambda_k y_k &= y \\ \lambda_i + \lambda_j + \lambda_k &= 1. \end{aligned}$$

By solving the above equation, we can obtain the barycentric coordinates $(\lambda_i, \lambda_j, \lambda_k)$ of (x, y) as

$$\begin{aligned} \lambda_i &= \frac{(y_j - y_k)(x - x_k) + (x_k - x_j)(y - y_k)}{(y_j - y_k)(x_i - x_k) + (x_k - x_j)(y_i - y_k)} \\ \lambda_j &= \frac{(y_k - y_i)(x - x_k) + (x_i - x_k)(y - y_k)}{(y_j - y_k)(x_i - x_k) + (x_k - x_j)(y_i - y_k)} \\ \lambda_k &= 1 - \lambda_i - \lambda_j. \end{aligned}$$

The foreground score of pixel (x, y) is computed by a linear interpolation using its Barycentric coordinates, as $\mathbf{a} = \lambda_i \mathbf{a}_i + \lambda_j \mathbf{a}_j + \lambda_k \mathbf{a}_k$.

2.1 Upper Bound Analysis

We design two oracle experiments to reveal the full potential of our method. *Upper bound of attribute scores.* The first experiment shows how much gain can be obtained when all the learned attribute scores are accurate and the learned point locations remain the same. In this experiment, we first calculate the IoU

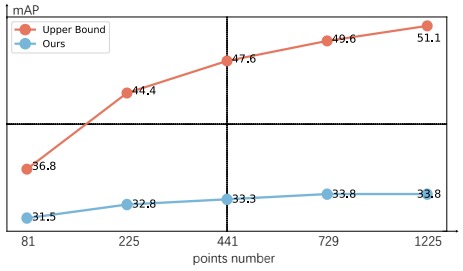

Fig. 2. Illustration for upper bound of *Dense RepPoints*.

between predicted bboxes and ground-truth bboxes to select positive samples (IoU threshold=0.5). Then, we change the predicted attribute scores of these positive samples to ground-truth scores. The attribute scores of negative samples remain the same. Finally, we utilize these new attribute scores to generate binary masks. This experiment on a ResNet-50 backbone yields about 39.4 detection mAP (the fluctuation of detection performance under different numbers of points is negligible). Results are shown in Figure 2. We observe large performance gain when the attribute scores are absolutely accurate, which suggests that our method still has great potential if the learned attribute scores are improved. When the number of points increases to 1225, the upper bound performance can improve nearly 60% over the original segmentation performance. Clearly, a better detection result (better point locations) will also boost the upper bound of our mask representation.

Upper bound of DTS and triangulation The second experiment examines the upper bound when all the attribute scores and point locations are equal to the ground truth. First, we use DTS to generate points for each ground-truth mask. Then we assign ground-truth attribute scores to these points. Finally, we use triangulation interpolation to predict masks. Table 1 shows the average IoU of our predicted masks and ground-truth masks. It can be seen that the IoU is nearly perfect (above 95%) when the points number increases, which indicates that our DTS and triangulation post-processing method can precisely depict the mask.

n	9	25	49	81	225	441	729
IoU	53.9	70.2	78.5	84.3	91.2	94.3	**95.6**

Table 1. The upper bound of IoU between predicted masks and ground-truth using DTS and triangulation post-processing under different point numbers.

References

1. Chen, X., Girshick, R.B., He, K., Dollár, P.: Tensormask: A foundation for dense object segmentation. In: ICCV (2019)
2. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: Point set representation for object detection. In: CVPR (2019)