α	AP	AP$_{50}$	AP$_{75}$
0	57.3	82.0	63.8
0.3	57.5	82.1	64.1
0.5	**58.0**	**82.5**	**64.7**
0.7	57.6	82.3	64.3

(a) **Ratio α in Eq. (5)**

β	AP	AP$_{50}$	AP$_{75}$
0.3	56.4	81.5	62.4
0.5	57.1	82.1	63.8
0.7	**58.0**	**82.5**	**64.7**
0.9	56.4	81.5	62.5

(b) **Ratio β in Eq. (6)**

τ_s	τ_t	AP	AP$_{50}$	AP$_{75}$
0.1	0.04	57.7	82.3	64.4
0.1	0.07	**58.0**	**82.5**	**64.7**
0.2	0.15	57.3	82.1	64.2

(c) **Distillation *temp.*.** student (τ_s), teacher (τ_t)

P	AP	AP$_{50}$	AP$_{75}$
8	57.1	81.8	63.9
16	58.0	82.5	64.7
32	**58.2**	**82.7**	**65.1**

(d) **Number of sampled points P**

n	AP	AP$_{50}$	AP$_{75}$
2	57.1	82.1	63.7
4	**58.0**	**82.5**	**64.7**
8	57.6	82.2	64.3

(e) **Size of grid n**

R	AP	AP$_{50}$	AP$_{75}$
14×14	57.3	81.9	63.7
56×56	**58.0**	**82.5**	**64.7**
224×224	57.2	82.2	63.6

(f) **Feature map resolution R**

Table 4. **Ablation studies.** For all of them, we pre-train our representation on ImageNet-1K for 100 epochs, and report the transfer results on VOC object detection. Our default settings are shown in gray.

Acknowledgements. This work was partially supported by ONR N00014-21-1-2812. We thank the anonymous reviewers for their effort and valuable feedback to improve our work.

A. More Ablation Analysis

Beyond ablation analysis provided in the main paper (Sec. 4.5), we provide three more groups of analysis in this appendix. They are: 1) balance between different losses during pre-training; 2) student and teacher temperatures in point affinity distillation; and 3) hyper-parameters for point sampling. Unless otherwise specified, we use ImageNet-1K and pre-train for 100 epochs. The results are reported on VOC object detection, all summarized in Tab. 4.

A.1. Balance Between Losses

Contrastive & affinity distillation. The hyper-parameter α in Eq. (5) serves as the weight to balance the two point based loss terms. By default we set α as 0.5 and we report the results of different α values in Tab. 4a.

Image-level & point-level. On top of point-level computation, we further leverage image-level loss. The hyper-parameter β in Eq. (6) serves as the weight to balance the two loss terms. We report the results of different β values in Tab. 4b. We find when the image-level loss is small, the overall performance will be influenced, since the point-level task is harder to converge at the beginning. Adding image-level contrastive loss further enhances our method to balance localization and recognition capabilities.

A.2. Temperatures in Point Affinity Distillation

We now study the hyper-parameters for the student and teacher temperatures τ_s and τ_t. Intuitively, we hope the output from the teacher is closer to a 'one-hot' label [38], which means the teacher temperature is relatively smaller than the student one. We explored a few setups following this intuition, and summarize our observations in Tab. 4c.

From the default temperatures, our method is quite robust to the changes. For example, decreasing τ_t from 0.07 to 0.04 only slightly degenerates the performance. Varying both temperatures also do not affect much in the third row.

A.3. Point Sampling

Number of points P. For final loss which includes the point affinity, we also ablate the number of points. From the results in Tab. 4d we can observe the performance improves as the point number increases. We use point number P=16 as the default setting, where the performance starts to saturate. We report the results of different number of points.

Number of grids n. In the default setting, the adopted grid number is 4×4. We report the results of different number of grids in Tab. 4e. From the table, we observe the number of grid does not influence the results much.

Feature map resolution R. In the default setting, we upsample the feature map to 56×56. From Tab. 4f, we can observe that feature map resolution would also influence the results, with 56 x 56 further directing towards better AP.

B. More Visualizations

We provide more visualizations with our method in Fig. 8. We followed the same protocol which first pick a point (denoted by red circle), and then compute the affinity map with the pre-trained features. Brighter colors denote more similar points. Our method consistently generates masks with crisp boundaries.

C. License of Assets

Dataset	Licence
ImageNet	https://image-net.org/download.php
COCO	Creative Commons Attribution 4.0 License
Pascal VOC	http://host.robots.ox.ac.uk/pascal/VOC/
Cityscapes	https://www.cityscapes-dataset.com/license/

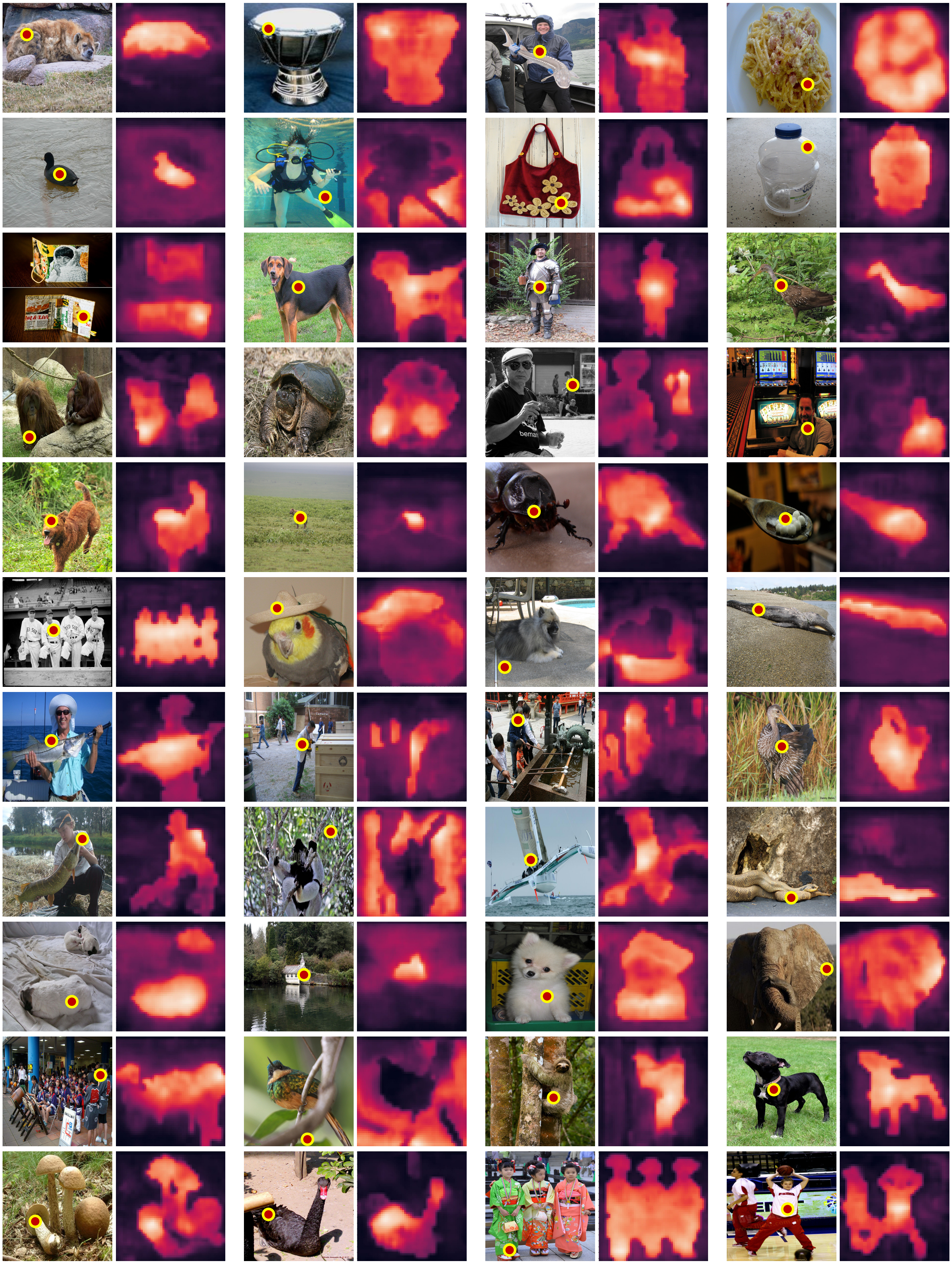

Figure 8. **More visualizations** on ImageNet-1K with our point-level region contrast.