$region$	$mAP^{1.0}$
$circle$	55.46
$rect._{1:1}$	55.39
$rect._{2:1}$	54.77
$rect._{1:2}$	54.96

(a) $region$

u_0	$mAP_{1.0}^{all}$
12	55.25
8	55.46
6	55.36
4	55.42

(b) u_0

δ_1	$mAP_{1.0}^{all}$
0.10	55.46
0.15	55.98
0.20	56.04
0.25	55.78

(c) δ_1

δ_2	$mAP_{1.0}^{all}$
0.25	55.19
0.50	55.46
0.75	55.22
1.0	55.41

(d) δ_2

Table 5. Ablation study for hyper-parameters.

Appendix

A. More experiment.

Sample shape and sample density. In Table 5b, we ablate on sample density u_0 and validate that u_0 has little impact. Theoretically, circles are orientation equivalent, which makes it suitable to model multi-view object. Experimentally, the sampling results of circle and rectangle are very close due to the discrete sampling, leading to comparable performance (Table 5a, where different aspect ratio $rect._{w:h}$ are studied).

Threshold of refinement. δ_1 and δ_2 are thresholds, well defined in detection and localization tasks. δ_1, δ_2, u_0 have little impact to performance (Table 5).

B. SeaPerson

SeaPerson is building for tiny person localization, which can help maritime quick rescue, beach safety inspection and so on. The resolution of images are mainly 1920 $\times$ 1080 and the person size is extremely low (about 22.6 pixels). Therefore, there is no privacy sensitive information.

Dataset Collection. SeaPerson is collected as : i) Videos are recorded in various seaside scenes by a RGB camera on a Unmanned Aerial Vehicle. ii) We sample an image of every 50 frames from video and remove images with high homogeneity. iii) We annotated all persons in all sampled images with bounding boxes. iv) Following the rules of coarse point annotation in Sec 4.1, coarse point annotation is obtained on SeaPerson for POL task.

Dataset Splitting. We randomly split dataset into three subsets (training set, valid set and test set), while images from the same video sequence cannot be separated into different subsets. As shown in Table 6, the ratio of images' number in training set, valid set and test set is about 10:1:10.

Dataset Properties. Our proposed SeaPerson is similar with TinyPerson while the volumn of SeaPerson is about 7 times that of TinyPerson. The absolute size and relative size of objects are very small as shown in Table 7 and Fig. 7. In such scenario, we only care about the position of the object rather than the size of the object, which makes it very suitable for POL task. In addition, SeaPerson can also be used as a dataset for tiny object detection due to the bounding box annotation.

	TinyPerson V2		TinyPerson	
	#images	#annotations	#images	#annotations
train set	5711	262063	794	42197
valid set	568	42399	816	30454
test set	5753	315165	-	-
sum	12032	619627	1610	72651

Table 6. Statistic information of SeaPerson and TinyPerson.

dataset	absolute size	relative size	aspect ratio
COCO	99.5 $\pm$ 107.5	0.190 $\pm$ 0.203	1.213 $\pm$ 1.337
TinyPerson	17.0 $\pm$ 16.9	0.011 $\pm$ 0.010	0.690 $\pm$ 0.422
TinyPerson V2	22.6 $\pm$ 10.8	0.016 $\pm$ 0.007	0.723 $\pm$ 0.424

Table 7. The mean and standard deviation of absolute size, relative size and aspect ratio of objects in different datasets. Size is defined as the square root of the product of width and height and aspect ratio is the value of width divided by height. Absolute size is the size of object and relative size is the value of the object's size divided by the image's size. These settings are followed as [45].

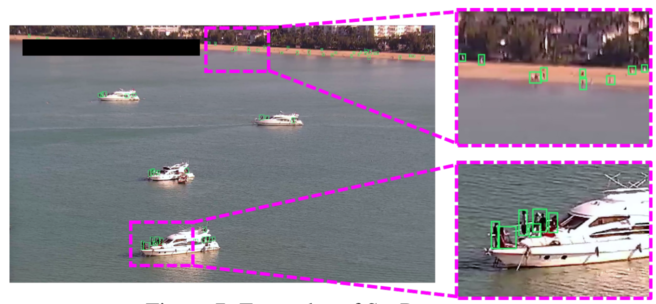

Figure 7. Examples of SeaPerson.

C. Implementation Details for CPRNet

ResNet-50 is used as the backbone network unless otherwise specified and FPN is adopted for feature fusion. P2 (stride is 4) is used for SeaPerson and P3 (stride is 8) is used for COCO and DOTA. The mini-batch is 64/4/4 images and all models are trained with 8/2/4 GPUs for COCO/DOTA/SeaPerson. The training epoch numbers are set as 12/12/6, the learning rate are set as 0.0025, 0.00025, 0.0125 and decays by 0.1 at the 8-th/8-th/4-th and 11-th/11-th/5-th epoch for COCO/DOTA/SeaPerson, respectively. In default settings, the backbone is initialized with the pretrained weights on ImageNet and other newly added layers are initialized with Xavier. The sampling radius R is set as 8/7/5 for COCO/DOTA/SeaPerson by defalut.

In dataset pre-processing, for COCO dataset, the short side of the images is resized to 400, and the ratio of width and height is kept. In dota dataset, images are split into sub-images (1024 $\times$ 1024 pixels) with overlap (200 $\times$ 200 pixels). And in SeaPerson, images are split into sub-images (640 $\times$ 640 pixels) with overlap (100 $\times$ 100 pixels). For data augmentation, only random horizontal is utilized in our CPRNet training.

D. Details of Semantic Variance

The $Var(x')$ and $Var(y')$ in Eq. 13 are calculate as:

$$
\begin{aligned}
Mean(x') &= \frac{1}{M} \sum_{1 \le j \le M} x'_j; \\
Mean(y') &= \frac{1}{M} \sum_{1 \le j \le M} y'_j; \\
Var(x') &= \frac{1}{M} (x'_j - Mean(x'))^2; \\
Var(y') &= \frac{1}{M} (y'_j - Mean(y'))^2;
\end{aligned}
\tag{15}
$$

where (x'_j, y'_j) is relative position of j-th object, M is the number of objects in dataset. For the objects whose annotated points or refined points are out of the bounding box, they often are regarded as wild points during the learning procedure. The wild points account for a small proportion and will not be learned by the network. However, their RSV will be very large since the RSV is relative to the height and width. Therefore, to better reflect the semantic variance of the points that really affect network learning, only the object whose annotated point or refined point is inside bounding box is used for calculating RSV.

E. Visualization of CPR

The visulization of CPR on COCO, DOTA and SeaPerson are shown as Fig. 9, Fig. 10 and Fig. 8, respectively.

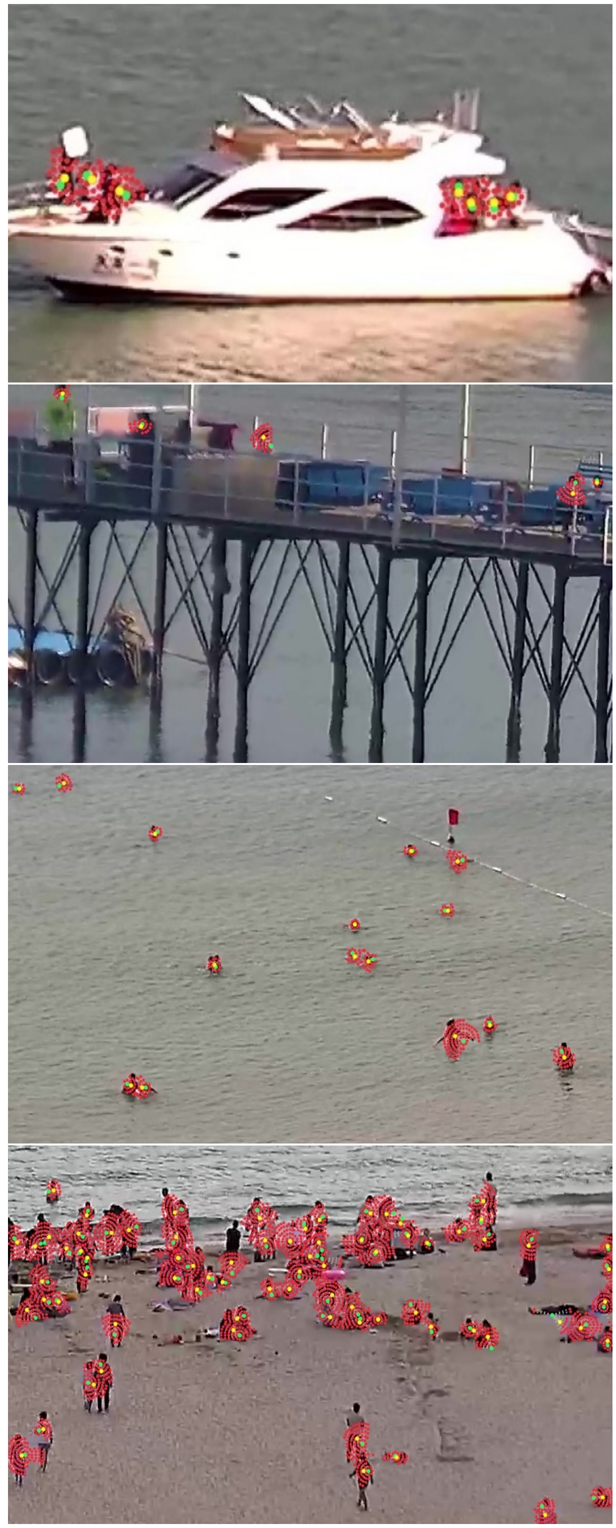

Figure 8. Visualization of CPR on SeaPerson. The images are cut from original images for better visualization. Semantic points (red) around the annotated point (green) are weighted averaged to obtain the semantic center (yellow) as final refined point (see Sec. 3.3).

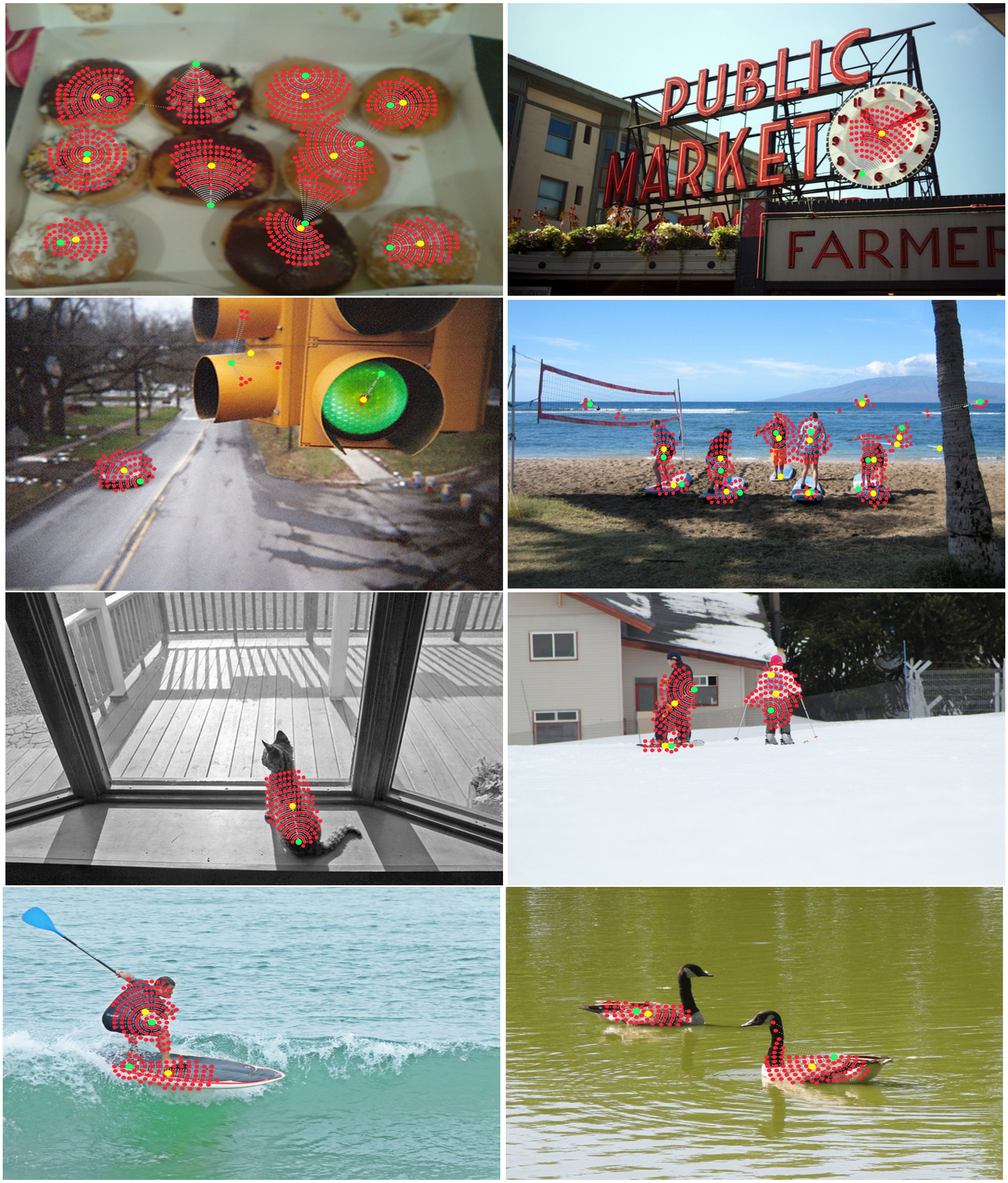

Figure 9. Visualization of CPR on COCO. The images are cut from original images for better visualization. Semantic points (red) around the annotated point (green) are weighted averaged to obtain the semantic center (yellow) as final refined point (see Sec. 3.3).

Figure 10. Visualization of CPR on DOTA. The images are cut from original images for better visualization. Semantic points (red) around the annotated point (green) are weighted averaged to obtain the semantic center (yellow) as final refined point (details in Sec. 3.3).