Supplementary Material: "IDNet: Information Decomposition Network for Fast Panoptic Segmentation"

Guangchen Lin, Songyuan Li, Yifeng Chen, and Xi Li

I. More technical details

As promised in Sec. III.B, we add the technical details in this section to make our pipeline easier to follow. You can ignore this section if you do not want to know the engineering details.

A. Training targets

There are two parts in our paper, which requires more technical details. One is the location head and the other is Information Composition. The rest parts have enough details to be implemented in engineering.

After the feature extraction, we get the location feature $\mathbf{F}^L$ and then generate center confidence scores sc, bounding boxes box and location embeddings θ. The training details of the location head are a bit similar to FCOS. There are K_{th} ground truth *thing* instances. For each instance, only the samples falling in the center region are considered positive. A location (x, y) on the feature map can be mapped onto the input image as $(\lfloor \frac{s}{2} \rfloor + xs, \lfloor \frac{s}{2} \rfloor + ys)$, where s is the stride of the location feature. The center region of an instance with the centroid (c_x, c_y) is defined as the box $(c_x - rs, c_y - rs, c_x + rs, c_y - rs)$, where r is set to 1.5. Besides, for the box, the range of regression targets is limited among different FPN layers as described in FCOS, which is another rule to determine positive samples. Only positive samples of instances are supervised in box, and the targets are set to 1/0 for positive/negative samples in sc. Although we no longer predict the "center-ness" in FCOS, it is still utilized for the weighted training of box. Therefore, there are several positive samples of each instance in training. To ensure an adequate amount of training data, they are all fed into Information Composition to predict corresponding location-aware masks. But only the location-aware mask with the top score of each instance is selected to match the category masks and join the panoramic decision.

B. Inference details.

In inference, the predicted objects in the location head are selected by bounding box NMS, where the score threshold is 0.1 and the IoU threshold is 0.6. Then the selected pieces of information are used to generate instance-distinct masks. After obtaining the instance masks, we utilize extra Matrix NMS to rule out duplicates among masks, where the IoU threshold is set to 0.6 and the score threshold is set to 0.35. At last, it is worth mentioning that we downsample the masks by four times in our Regional Matching (Eq. 9), which reduces the computation and memory overhead with a negligible performance drop.

II. Visualization

We show visual examples obtained by our method and compare it with UPSNet on COCO *val* dataset in Fig. S-1. Both methods use the ResNet50-FPN as the backbone. UPSNet is an efficient end-to-end two-stage method. From the figure, we observe that our method outperforms UPSNet with the same backbone in both recognition quality and segmentation quality, especially in object edge details.

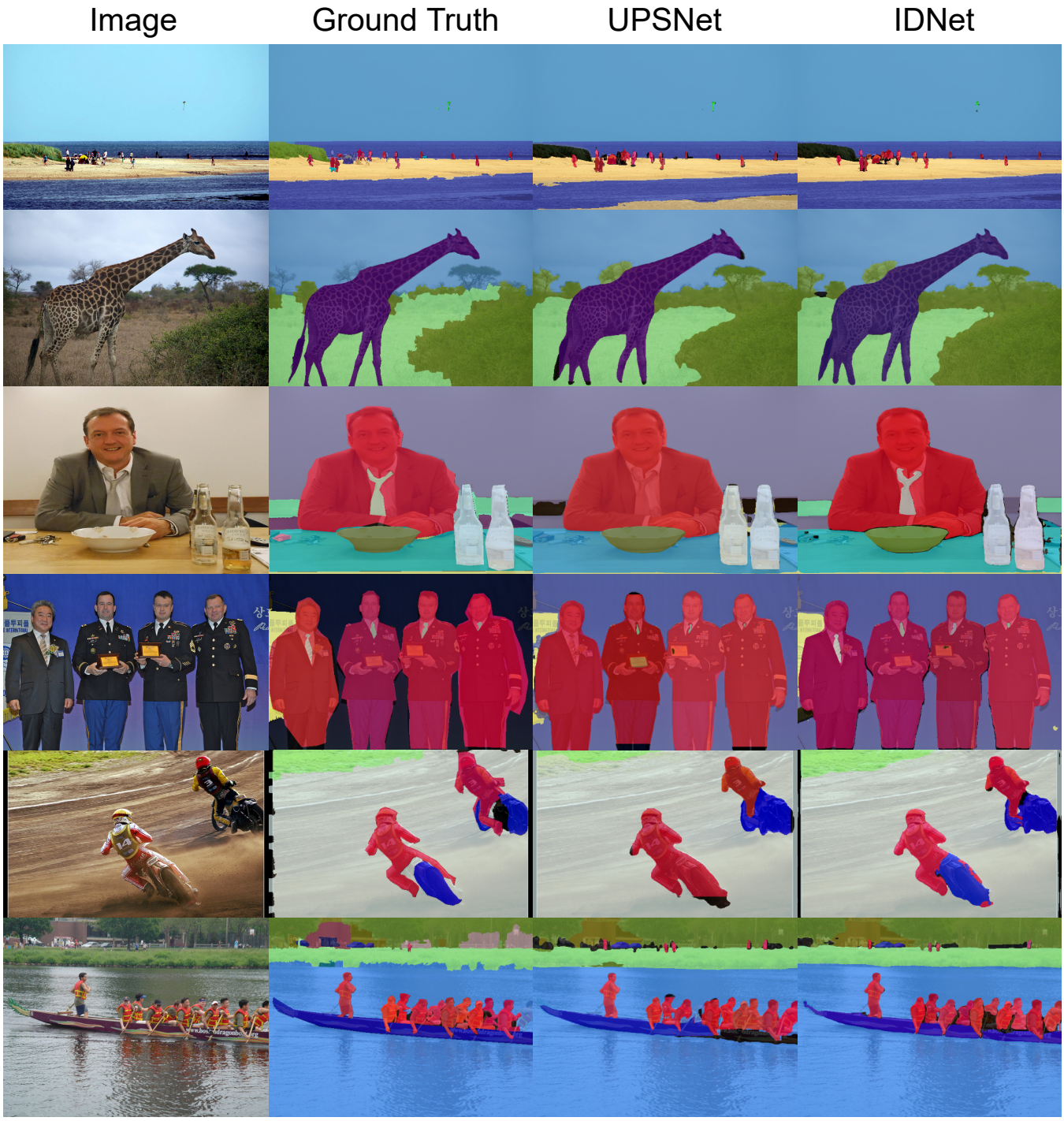

Fig. S-1: **The visualization on COCO *val* dataset**. The figure shows the improvement over UPSNet.