

Supplementary Materials: LDA-AQU: Adaptive Query-guided Upsampling via Local Deformable Attention

Anonymous Authors

1 MORE EXPERIMENTAL DETAILS

We evaluate the effectiveness of LDA-AQU across four dense prediction tasks (*i.e.*, object detection, instance segmentation, panoptic segmentation, and semantic segmentation) by substituting the base upsampler (*i.e.*, Nearest Neighbor Upsampler or Bilinear Upsampler) with LDA-AQU. Following the same process, we further compare the performance of LDA-AQU with other state-of-the-art upsamplers, including CARAFE [14], IndexNet [9], FADE [10], SAPA [11], DySample [8], *etc.*

Object Detection. We utilize Faster R-CNN [12] with ResNet [4] as the baseline model to conduct the performance comparison between various upsamplers and LDA-AQU on object detection task using the MS COCO dataset [7]. Specifically, we compare the performance of various upsamplers by replacing the upsampler utilized in the Feature Pyramid Network (FPN) [6]. During both the training and testing stages, the short side of input images is resized to 800 pixels. We perform performance comparisons and ablation studies on 4 GPUs with 2 images per GPU. The SGD optimizer is utilized to train the network with a momentum of 0.9 and a weight decay of 0.0001. Following previous work [2], we set the initial learning rate to 0.01. We employ the $1\times$ (12 epochs) training schedule to train networks, with the learning rate decreasing by a factor of 0.1 at the 9-*th* and 11-*th* epochs.

Instance Segmentation. We utilize Mask R-CNN [3] with ResNet [4] as the baseline model to conduct the performance comparison on instance segmentation task using the MS COCO dataset [7]. We perform performance comparison by replacing the upsamplers of FPN and mask head in Mask R-CNN. All other settings remain the same as for object detection.

Panoptic Segmentation. For panoptic segmentation task, we use Panoptic FPN [5] with ResNet [4] as the baseline model to conduct the performance comparison using the MS COCO dataset [7]. We perform performance comparison by replacing the upsamplers of FPN in Panoptic FPN. All other settings remain the same as for object detection.

Semantic Segmentation. For semantic segmentation task, we utilize UperNet [15] with ResNet [4] as the baseline model to conduct the performance comparison using the ADE20K dataset [16]. We train the model on 4 GPUs with 2 images per GPU. We utilize the SGD optimizer with an initial learning rate of 0.005, a momentum of 0.9, and a weight decay of 0.0005. The models are trained for 160K iterations, utilizing the *Poly* learning rate policy with a *power* of 0.9 and a *min_lr* of 0.0001.

2 MORE ABLATION STUDIES

As in the main text, all ablation studies are conducted on models based on Faster R-CNN [12] and ResNet-50 [4] using the MS COCO dataset [7].

Kernel Sizes. We verify the impact of different kernel sizes of information encoding k_e and neighboring points k_u in LDA-AQU on

model performance. As illustrated in Table 1, the impact of k_e and k_u on model performance is not as significant as that of the local deformation ranges. Larger values of k_e and k_u typically result in greater performance improvements, but they also lead to a notable increase in computational complexity and parameters. Therefore, we prefer to set k_e and k_u to a kernel size of 3×3 to achieve a better balance between computational complexity and performance.

Table 1: Performance comparison of various kernel sizes in the FPN of Faster R-CNN.

k_e	k_u	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	3	38.9	60.5	42.3	22.8	42.4	50.1
1	5	39.1	60.4	42.8	23.3	43.3	49.9
3	3	39.2	60.7	42.7	22.9	43.0	50.1
3	5	39.2	60.5	42.6	23.2	42.9	50.2
5	3	39.3	60.8	42.7	23.3	43.3	50.3

Upsampling scheme. We also evaluate the impact of different upsampling methods for the queries in LDA-AQU on model performance. As shown in Table 2, the nearest neighbor interpolation upsampler performs the same as the bilinear upsampler, with an AP of 39.2. However, their detection performance on objects of different scales differs significantly. We posit that, due to the higher resolution of feature maps utilized for detecting small objects, employing nearest neighbor points as precise object features to guide the upsampling task is more appropriate. In contrast, for low-resolution feature maps with a notable semantic gap between points, bilinear interpolated features are preferred for guiding the upsampling task due to their greater resemblance to the features of the objects.

Table 2: Performance comparison of different upsampling methods for query upsampling in LDA-AQU.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Nearest	39.2	60.6	42.3	23.8	42.7	50.3
Bilinear	39.2	60.7	42.7	22.9	43.0	50.1

3 OBJECT DETECTION ON PASCAL VOC

We further evaluate the effectiveness of LDA-AQU on Pascal VOC dataset [1]. Specifically, we utilize the VOC 2012 and VOC 2007 trainval splits for model training, and evaluate their performance across different models on the VOC 2007 test split. We resize both the training and test images to 640×640 , ensuring consistency with the training strategy and hyperparameters employed in the experiments conducted on MS COCO. Similarly, we employ Faster

Table 3: Performance comparison of Object Detection on Pascal VOC based on Faster R-CNN (F-RCNN). The best in each column is highlighted in bold, and the second best is underlined.

F-RCNN [12]	mAP	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv
Nearest	78.7	80.5	87.5	78.1	67.2	66.3	86.1	87.5	<u>89.2</u>	63.5	84.6	74.1	87.8	87.0	80.8	80.1	52.3	78.5	80.2	85.3	76.5
Deconv	77.7	80.7	80.7	79.3	66.8	64.9	<u>86.3</u>	87.1	88.5	61.8	84.4	74.6	87.0	85.8	79.7	79.8	51.8	77.1	79.7	86.4	71.1
PS [13]	78.5	81.0	81.0	78.4	66.4	65.9	<u>85.4</u>	87.4	89.1	62.2	86.3	74.5	87.5	87.8	79.9	80.0	51.6	<u>84.0</u>	79.5	85.2	77.5
CARAFE [14]	79.2	81.1	81.2	79.2	73.3	65.8	<u>86.3</u>	<u>88.1</u>	89.0	63.0	84.8	73.4	88.2	88.0	80.2	<u>80.2</u>	<u>53.2</u>	84.8	81.3	<u>86.1</u>	76.3
FADE [10]	79.3	81.5	81.1	79.2	68.1	<u>66.4</u>	<u>85.8</u>	<u>88.1</u>	89.1	<u>63.9</u>	<u>86.4</u>	75.4	87.5	88.4	87.0	<u>80.2</u>	53.9	79.7	<u>82.4</u>	<u>86.1</u>	76.4
SAPA-B [11]	<u>79.9</u>	<u>87.0</u>	<u>86.8</u>	<u>79.7</u>	<u>72.9</u>	65.9	85.9	87.8	88.7	63.4	86.6	75.1	88.9	89.0	86.2	80.0	51.9	78.9	80.0	85.5	<u>77.9</u>
DySample [8]	78.9	80.9	81.0	80.1	66.3	66.3	86.8	87.9	89.3	64.4	86.3	<u>75.9</u>	87.9	88.4	86.5	<u>80.2</u>	51.4	79.3	<u>82.0</u>	79.8	<u>77.4</u>
LDA-AQU (ours)	80.3	88.8	81.0	79.1	72.7	67.3	86.8	88.3	88.7	63.5	<u>86.4</u>	79.2	<u>88.7</u>	<u>88.6</u>	<u>86.8</u>	80.4	<u>53.2</u>	79.8	82.6	85.9	78.4

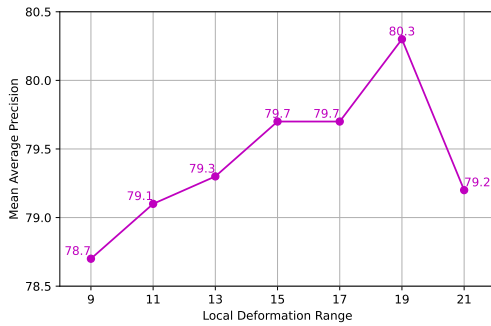


Figure 1: Performance comparison of different local deformation ranges in LDA-AQU on the Pascal VOC dataset.

R-CNN [12] with nearest neighbor interpolation as the baseline model and compare the effects of different upsampling methods by modifying the upsamplers implemented in FPN.

Comparison with Other State-of-the-Art Upsamplers. As illustrated in Table 3, our LDA-AQU achieves the best performance with an AP of 80.3, surpassing the baseline model by 1.6 AP (80.3 AP vs. 78.7 AP).

Ablation Study of Local Deformation Ranges. Then, we evaluate the impact of varying the local deformation ranges in LDA-AQU on the Pascal VOC dataset. As depicted in Figure 1, the model achieves optimal performance when θ is set to 19. This is attributed to the larger object scale typically in the Pascal VOC dataset compared to the MS COCO dataset, necessitating a higher value of θ to achieve better shape matching with the objects.

4 MORE VISUAL INSPECTION AND ANALYSIS

We augment our study with additional visualizations, including illustrations of deformed neighboring points, alongside qualitative experiments conducted across the aforementioned four tasks. All visualizations are based on the backbone network of ResNet-50 [4].

Deformed Neighboring Points. As shown in Figure 2, we visualize more upsampled points (*i.e.*, queries) and their corresponding deformed neighboring points. Many examples in Figure 2 demonstrate that our LDA-AQU can adaptively adjust the position of neighboring points according to the shape (*e.g.*, umbrella, pole, leg,

etc.) and context (*e.g.*, boundary of two different objects) of the objects, thereby focusing more on features related to the queries.

Qualitative Experiments on Object Detection. As depicted in Figure 3, we visualize more object detection results to compare the effects of Faster R-CNN with Bilinear Interpolation (BI) and Faster R-CNN with LDA-AQU.

Qualitative Experiments on Instance Segmentation. As illustrated in Figure 4, we present additional instance segmentation results to compare the performance of Mask R-CNN with BI and Mask R-CNN with LDA-AQU.

Qualitative Experiments on Panoptic Segmentation. In Figure 5, we present additional panoptic segmentation results, comparing the effectiveness of Panoptic FPN with BI and Panoptic FPN with LDA-AQU.

Qualitative Experiments on Semantic Segmentation. In Figure 6, we present more semantic segmentation results, comparing the effectiveness of UperNet with BI and UperNet with LDA-AQU.

REFERENCES

- [1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88 (2010), 303–338.
- [2] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).
- [3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [5] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6399–6408.
- [6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.
- [7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- [8] Wenzhe Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. 2023. Learning to Upsample by Learning to Sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6027–6037.
- [9] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. 2019. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3266–3275.

- [10] Hao Lu, Wenze Liu, Hongtao Fu, and Zhiguo Cao. 2022. FADE: Fusing the assets of decoder and encoder for task-agnostic upsampling. In *European Conference on Computer Vision*. Springer, 231–247.
- [11] Hao Lu, Wenze Liu, Zixuan Ye, Hongtao Fu, Yuliang Liu, and Zhiguo Cao. 2022. SAPA: Similarity-aware point affiliation for feature upsampling. *Advances in Neural Information Processing Systems* 35 (2022), 20889–20901.
- [12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [13] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1874–1883.
- [14] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. 2019. Carafe: Content-aware reassembly of features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3007–3016.
- [15] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*. 418–434.
- [16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348

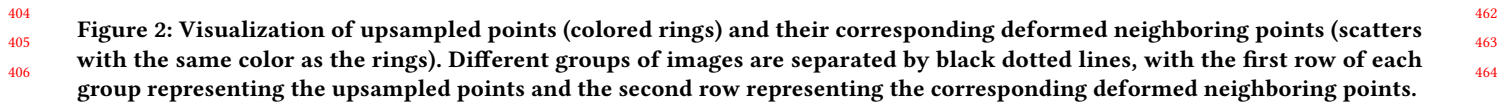




Figure 3: Visualization of prediction results based on Faster R-CNN on MS COCO. Different groups of images are separated by black dotted lines, with the first row of each group representing the results of Faster R-CNN w/ BI and the second row representing the results of Faster R-CNN w/ LDA-AQU.

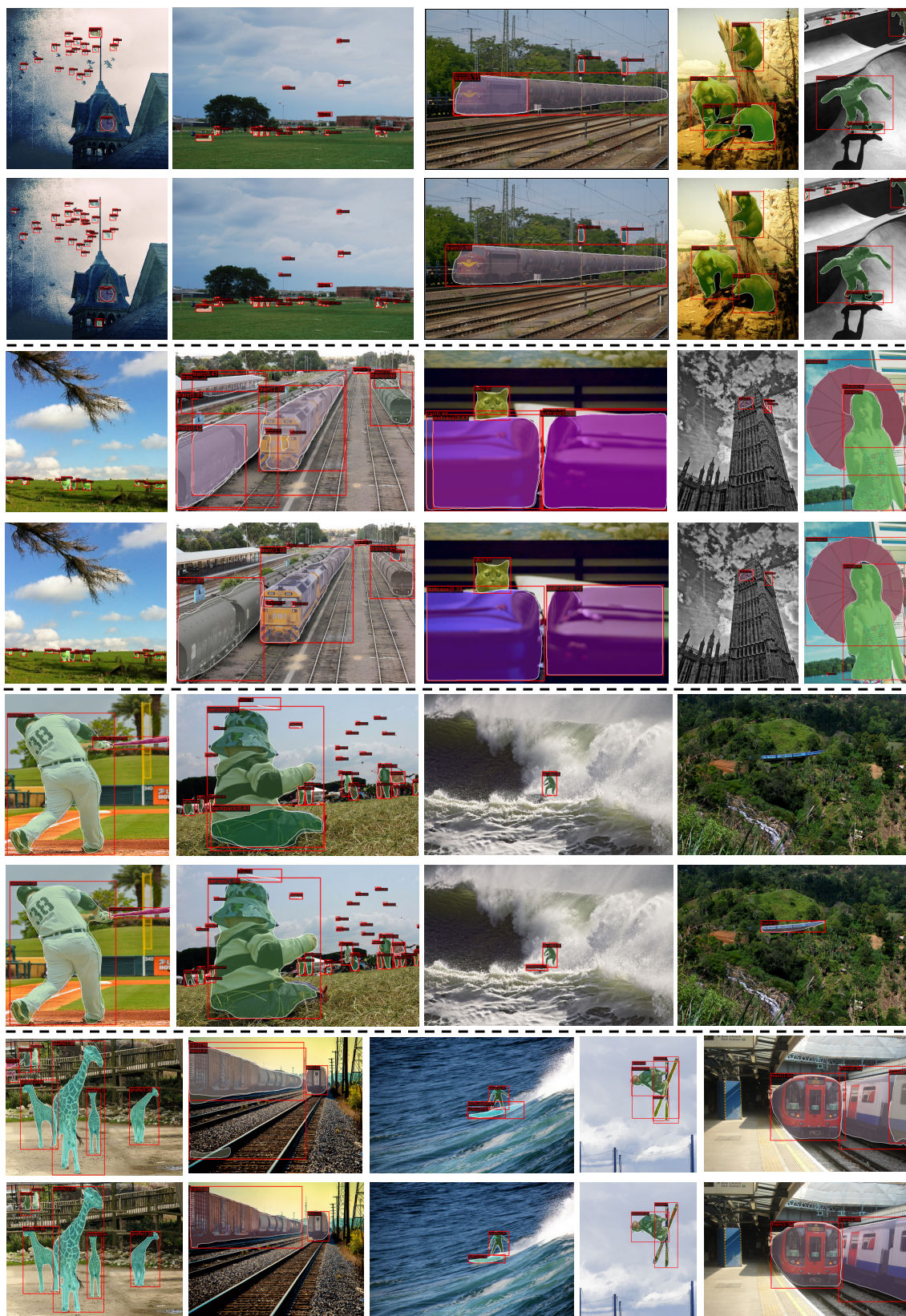


Figure 4: Visualization of prediction results based on Mask R-CNN on MS COCO. Different groups of images are separated by black dotted lines, with the first row of each group representing the results of Mask R-CNN w/ BI and the second row representing the results of Mask R-CNN w/ LDA-AQU.

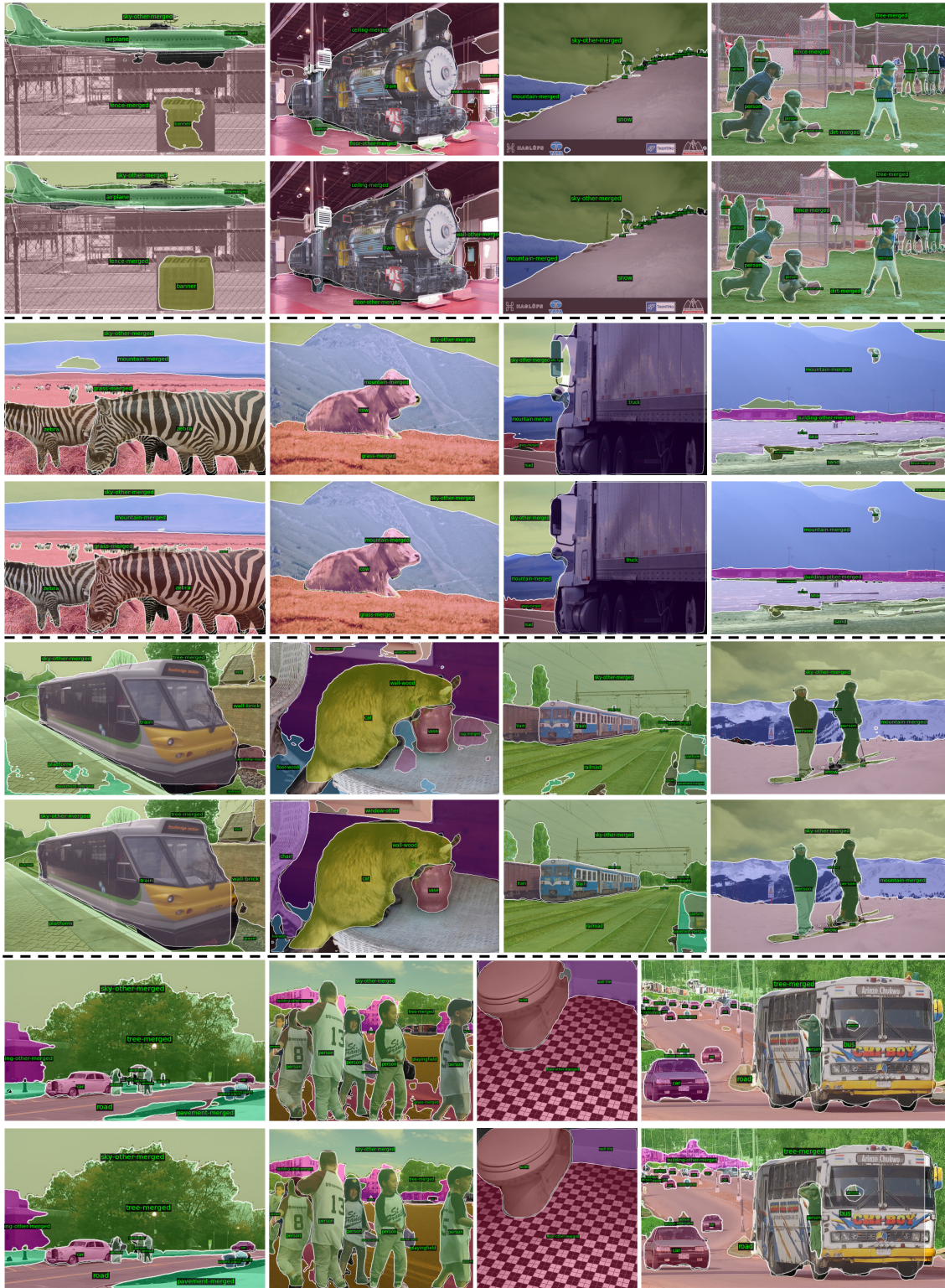


Figure 5: Visualization of prediction results based on Panoptic FPN on MS COCO. Different groups of images are separated by black dotted lines, with the first row of each group representing the results of Panoptic FPN w/ BI and the second row representing the results of Panoptic FPN w/ LDA-AQU.



Figure 6: Visualization of prediction results based on UperNet on ADE20K. Different groups of images are separated by black dotted lines, with the first row of each group representing the results of UperNet w/ BI and the second row representing the results of UperNet w/ LDA-AQU.