Scene Parsing with Global Context Embedding Supplementary Materials

Wei-Chih Hung¹, Yi-Hsuan Tsai^{1,2}, Xiaohui Shen³, Zhe Lin³, Kalyan Sunkavalli³, Xin Lu³, Ming-Hsuan Yang¹ ¹University of California, Merced ²NEC Laboratories America ³Adobe Research

1. Overview

In this document, we present additional results and analysis. First, we visualize the trained global context features using the t-SNE [4] method. Second, we provide per-class performance comparison on the MIT ADE20k dataset [7]. Third, we show more qualitative comparisons with other state-of-the-art methods on both MIT ADE20k and PASCAL Context dataset [5].

2. t-SNE Visualization of Global Context Feature

In Figure 1, we show the visualization in the feature space of the trained global context features. We sample 1000 images from the MIT ADE20k dataset [7] and pass these images through our proposed global context network to extract 4096-dimensional context feature vectors. We use the t-SNE [4] algorithm to reduce the dimension of the extracted feature vectors from 4096 to 2 for better visualization. Since we do not have the scene category labels, we use the thumbnail images with scene parsing ground truth annotations to draw the nodes in Figure 1. The t-SNE visualization shows that our global context network can group similar scenes into one cluster (e.g., the bedroom in Figure 1) and thus provide global cues to the scene parsing task.

In Figure 2, we show the t-SNE visualization of the VGG-16 features pre-trained on the ImgaNet for comparison. We observe that the VGG features can differentiate the indoor and outdoor scenes well. However, there are no detailed scene clusters for VGG features when compared to our trained features as shown in Figure 1. In Figure 3, we show additional t-SNE visualization using hand-crafted global features (HOG [2], LBP [1], Dense SIFT [3], and GIST [6]) on the MIT ADE20k dataset. We observe that these features have worse semantic embedding than the CNN features.

3. Per-class Performance Comparison on the MIT ADE20k Dataset

In Figure 4, we present the per-class comparisons using the mean IU metric. The MIT ADE20k dataset consists of 150 classes. It is worth noting that the indexes of classes are sorted by the pixel appearance frequency in the descending order, i.e., *wall* (index 1) is the most frequent class in the dataset while *flag* (index 150) is the least frequent one. As a result, classes in Figure 4 (b) and (c) represent the rare categories within the dataset. We observe that our method achieves higher mean IU than the baseline model particularly on those rare classes, since the proposed method can utilize the global context information to discover rare objects instead of predicting those objects as dominating "stuff" classes.

4. Additional Qualitative Comparisons on the MIT ADE20k Dataset

In Figure 6-15, we show additional qualitative comparisons of our method on the MIT ADE20K dataset. The 150 color coded scene categories are shown in Figure 5. The results show that our method can utilize the global context to eliminate false positives of the parsing results and discover small and rare objects in various scene-type images.

5. Additional Qualitative Comparisons on the PASCAL Context Dataset

In Figure 17-20, we show additional qualitative comparisons of our method on the PASCAL Context dataset. The color coded scane categories of the dataset are shown in Figure 16.



Figure 1. t-SNE visualization of trained global context features on the MIT ADE20k dataset (better viewed in color). We annotate a few observed scene clusters in the visualization (e.g., bedroom, bathroom, mountain).



Figure 2. t-SNE visualization of VGG-16 features pre-trained on ImageNet. When compared to Figure 1, the VGG-16 features can also differentiate the indoor and outdoor scenes well but not the detailed scene clusters.



Figure 3. t-SNE visualization of hand-crafted features on the MIT ADE20k dataset. Compare to Figure 1 and 2, these hand-crafted features perform worse on the task of differentiating different scene categories.





wall	building	sky	floor	tree	ceiling
road	bed	windowpane	grass	cabinet	sidewalk
person	earth	door	table	mountain	plant
curtain	chair	car	water	painting	sofa
shelf	house	sea	mirror	rug	field
armchair	seat	fence	desk	rock	wardrobe
lamp	bathtub	railing	cushion	base	box
column	signboard	drawers	counter	sand	sink
skyscraper	fireplace	refrigerator	grandstand	path	stairs
runway	case	pool table	pillow	screen door	stairway
river	bridge	bookcase	blind	coffee table	toilet
flower	book	hill	bench	countertop	stove
palm	kitchen island	computer	swivel chair	boat	bar
arcade machin	hovel	bus	towel	light	truck
tower	chandelier	awning	streetlight	booth	television
airplane	dirt track	apparel	pole	land	bannister
escalator	ottoman	bottle	buffet	poster	stage
van	ship	fountain	conveyer belt	canopy	washer
plaything	swimming poc	stool	barrel	basket	waterfall
tent	bag	minibike	cradle	oven	ball
food	step	tank	trade name	microwave	pot
animal	bicycle	lake	dishwasher	screen	blanket
sculpture	hood	sconce	vase	traffic light	tray
ashcan	fan	pier	crt screen	plate	monitor
bulletin board	shower	radiator	glass	clock	flag

Figure 5. Color coded of scene categories of the MIT ADE20K dataset.



Figure 6. Scene parsing results from the MIT ADE20k dataset.



Figure 7. Scene parsing results from the MIT ADE20k dataset.



Figure 8. Scene parsing results from the MIT ADE20k dataset.



Figure 9. Scene parsing results from the MIT ADE20k dataset.



Figure 10. Scene parsing results from the MIT ADE20k dataset.



Figure 11. Scene parsing results from the MIT ADE20k dataset.



Figure 12. Scene parsing results from the MIT ADE20k dataset.



Figure 13. Scene parsing results from the MIT ADE20k dataset.



Figure 14. Scene parsing results from the MIT ADE20k dataset.



Figure 15. Scene parsing results from the MIT ADE20k dataset.

background	aeroplane	bicycle	bird	boat	bottle
bus	car	cat	chair	cow	table
dog	horse	motorbike	person	pottedplant	sheep
sofa	train	tvmonitor	bag	bed	bench
book	building	cabinet	ceiling	cloth	computer
cup	door	fence	floor	flower	food
grass	ground	keyboard	light	mountain	mouse
curtain	platform	sign	plate	road	rock
shelves	sidewalk	sky	snow	bedclothes	track
tree	truck	wall	water	window	wood

Figure 16. Color coded scene categories of the PASCAL Context dataset.



Figure 17. Scene parsing results from the PASCAL Context dataset.



Figure 18. Scene parsing results from the PASCAL Context dataset.



Figure 19. Scene parsing results from the PASCAL Context dataset.



Figure 20. Scene parsing results from the PASCAL Context dataset.

References

- [1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 2006. 1, 3
- [2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005. 1, 3
- [3] D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004. 1, 3
- [4] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. JMLR, 2008. 1
- [5] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *CVPR*, 2014. 1
- [6] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1, 3
- [7] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016. 1