Supplementary Materials of Layout-induced Video Representation for Recognizing Agent-in-Place Actions

Ruichi Yu[1,2*] Hongcheng Wang[2]
Ang Li[1] Jingxiao Zheng[1] Vlad I. Morariu[3†] Larry S. Davis[1]
[1]University of Maryland, College Park [2]Comcast Applied AI Research [3]Adobe Research
[1]{yrcbsg, angli, jxzheng, lsd}@umiacs.umd.edu,
[2]hongcheng_wang@comcast.com, [3]morariu@adobe.com

1. Mobile App for Annotating Places

We developed a mobile app for users to efficiently annotate their camera regions. The app is shown in Figure 2. Users can simply use points to define polygons and choose a pre-defined category for each place. The annotation process is efficient since usually the users will fix the cameras for a long time, and they only need to spend a few seconds to annotate the segmentation maps one time per camera.

2. Dataset Statistics

Detailed dataset statistics are shown in Table 1.

3. Network Architecture

The architecture of our network is shown in Table 2.

4. Decoupling Spatial-temporal Max Pooling

Traditional 3D ConvNets conduct max pooling along both spatial and temporal dimensions of feature maps to increase the size of receptive field. In home surveillance scenario, it is reported in [6] that decoupling the max pooling by first conducting spatial-only max pooling on some 3D-conv blocks, then adding more conv blocks with temporal-only max pooling leads to better performance. One possible reason is that conducting temporal-wise max pooling early will capture motion patterns of only local body of the moving objects. Since in a home surveillance video, the moving objects are usually large, and we need to apply several conv blocks with spatial-only max pooling layers to capture the motion of the entire object. We tried both methods and the per-category average precision results is shown in Fig.2. "ST Max Pool" denotes that we use 5 conv blocks with spatial-temporal max pooling to abstract both spatial and temporal information at the same time. "Decouple ST Max Pool" denotes our network structure that has 5 conv blocks with spatial-only max pooling, and followed by 4 more blocks with temporal-only max pooling. Since the second network has more conv blocks, to make the two network structures have similar depth, we add one more conv layer in each conv block of "ST Max Pool". For both methods, we use our full model with PD+DD+Topo-Agg. The hyper-parameters setting is: we decompose semantics on different places after the second conv blocks ($L = 2$); we conduct distance-based place discretization on $PL_{DT} = \{walkway, driveway, lawn\}$ and choose $k = 3$; for topological feature aggregation, we choose $h = 1$. We can observe that "Decouple ST Max Pool" leads to better performance.

5. Per-category Performance.

Fig. 7 in the main paper shows the average precision for each action on unseen scenes. LIVR outperforms the baseline methods by a large margin on almost all action categories. When comparing the orange and green bars in Fig. 7, we observe that the proposed topological feature aggregation (Topo-Agg) leads to consistently better generalization for almost all actions. The blue dashed box highlights the actions that include moving directions, and consistent improvements are brought by distance-based place discretization (DD). For some actions, especially the ones occurring on street and sidewalk, since they are relatively easy to recognize, adding DD or Topo-Agg upon the place-based feature descriptions (PD) does not help much. Overall, LIVR improves the generalization capability of the network, especially for actions that are more challenging, and are associated with moving directions.

*Work done at the Comcast Applied AI Research.
†Was affiliated with the University of Maryland during the work.

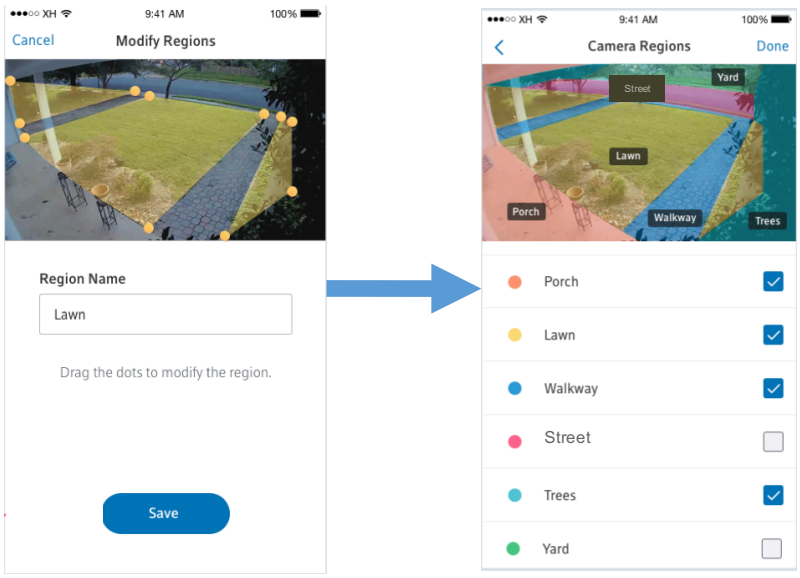


Figure 1. Mobile app for users to annotate camera regions. Note that this version of App contains not only the six places in the main paper but also some more fine-grained places.

Table 1. Dataset Statistics

	Observed Scene	Unseen Scene	Total
<vehicle, move along, street>	1397	957	2354
<person,move along,sidewalk>	539	299	838
<pet,move along,sidewalk>	149	169	318
<person,stay,lawn>	54	129	183
<person,move away (home),driveway>	218	173	391
<person,move toward (home),driveway>	222	201	423
<person,move toward (home),walkway>	153	91	244
<person,move away (home),walkway>	118	52	170
<vehicle,move away (home),driveway>	71	70	141
<vehicle,move toward (home),driveway>	52	66	118
<person,interact with vehicle,driveway>	171	81	252
<person,move across,lawn>	225	433	658
<person,stay,porch>	105	112	217
<person,move toward (home),porch>	310	139	449
<person,move away (home),porch>	260	136	396
Total	4044	3108	7152

Table 2. Network Structure of LIVR. We apply spatial-only max pooling after block Conv1-Conv5, and temporal-only max pooling after block Conv6-Conv9. From Conv3 to Conv9, each conv blocks consists of two identical 3D-conv layers with ReLU in between. The "on/off" status of each connection for the final gated FC layer is determined by Topo-Agg.

Block	Input Size	Kernel Size	Stride	# Filters	Block	Input Size	Kernel Size	Stride	# Filters
Conv1	15×90×160×3	3×3×3	1×1×1	64	Conv6	15×3×5×64	3×3×3	1×1×1	64
Pool1	15×90×160×3	1×2×2	1×2×2	-	Pool6	15×3×5×64	2×1×1	2×1×1	-
Conv2	15×45×80×64	3×3×3	1×1×1	64	Conv7	8×3×5×64	3×3×3	1×1×1	64
Pool2	15×45×80×64	1×2×2	1×2×2	-	Pool7	8×3×5×64	2×1×1	2×1×1	-
Conv3	15×23×40×64	3×3×3	1×1×1	64	Conv8	4×3×5×64	3×3×3	1×1×1	64
Pool3	15×23×40×64	1×2×2	1×2×2	-	Pool8	4×3×5×64	2×1×1	2×1×1	-
Conv4	15×12×20×64	3×3×3	1×1×1	64	Conv9	2×3×5×64	3×3×3	1×1×1	64
Pool4	15×12×20×64	1×2×2	1×2×2	-	Pool9	2×3×5×64	2×1×1	2×1×1	-
Conv5	15×6×10×64	3×3×3	1×1×1	64	SGMP	1×3×5×64	1×3×5	1×1×1	-
Pool5	15×6×10×64	1×2×2	1×2×2	-	Gated FC	1×1×1×384	-	-	-

6. Automatically Generating Segmentation Maps

In home surveillance scenario, we believe that it is reasonable to involve users to provide us with perfect segmentation maps of their own houses. However, to evaluate our proposed method's effectiveness with imperfect, automatically generated maps, we developed an algorithm using automatic semantic segmentation and historical statistics of the videos to generate place segmentation maps.

Due to the gap between appearance based segmenta-

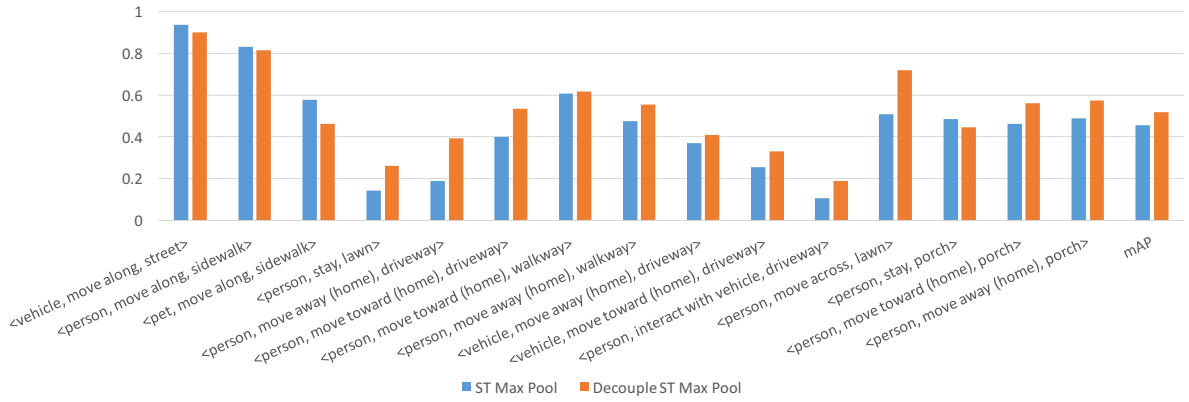
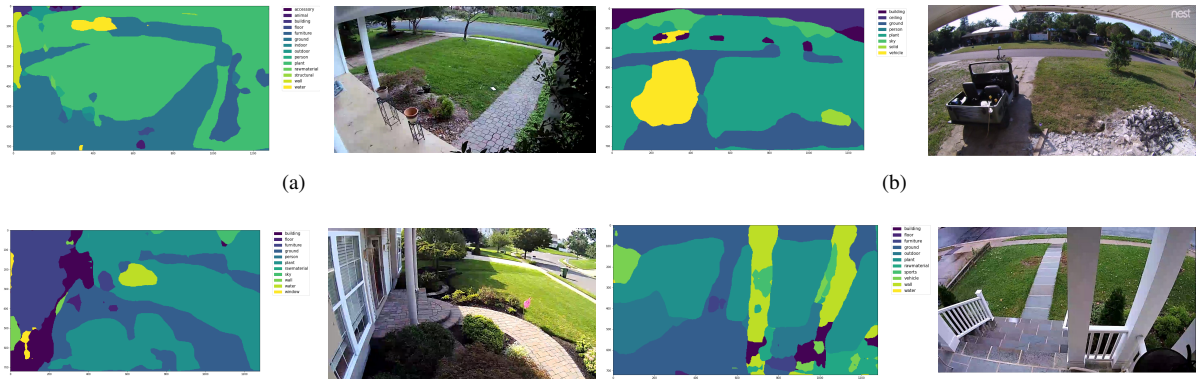


Figure 2. "ST Max Pool" denotes a network structure with 5 conv blocks with spatial-temporal max pooling. "Decouple ST Max Pool" denotes our network structure that has 5 conv blocks with spatial-only max pooling, and followed by 4 more blocks with temporal-only max pooling. We observe performance improvements on almost all action categories.



(a) (b) (c) (d)

Figure 3. Segmentation maps generated by Deeplab [3] V2 pre-trained on COCO-Stuff [2] dataset. Without enough training data, deep learning based method cannot precisely differentiate places with different functionalities (*e.g.,* walkway, driveway and street).

tion (for most of the current semantic segmentation methods) and the functionality based segmentation (for our home surveillance scenario), directly applying the state of the art segmentation methods fails in our scenario. We applied Deeplab [3] V2 pre-trained on COCO-Stuff [2] dataset directly on our camera images and obtain poor results as shown in Fig.3. From the figures we can see that the appearance-based segmentation methods assigns same labels to pixels with similar appearance but different functionalities. Without a large enough training dataset containing different scene layouts with functionality labels, it is difficult to apply deep learning based segmentation methods to generate the place segmentation maps. Thus instead, we propose an approach using low-level cues and historical statistics for automatic segmentation. We cluster pixels into super-pixels based on their appearance and spatial relations. Then, on these segments, we utilize normalized cut (NCut) [5], which is an optimization based segmentation method, to further segment the images to multiple segments based on their appearance (Fig.4 (b)). Then, we apply a heuristic method that utilizes videos of the scenes to obtain heatmaps of some specific places (Fig.4 (c)), based on the fact that the object types and motion patterns on different functional places are different. For example, we can usually observe people/vehicles with different scales and walking in different ways at different functional places, from the view of a surveillance camera. Based on the observations, we first apply object detection algorithm [4] and tracking algorithm [1] to detect and track moving person and vehicles in the videos, and then generate the heatmaps of porch, walkway, street, sidewalk and driveway in each scene based on the above heuristics.

Given the results of NCut and the heatmaps, we label each segment with the majority place category of the heatmap. We label segments as lawn if their averaged color is close to one of the reference colors, *e.g.,* green, dark green, *etc.*The resulting annotation maps are shown in Fig.4(d). When compared to the manually labeled

ground truth, our automatically generated maps are reasonably good, especially for place categories such as walkway, porch, lawn and street. However, sometime our method may mistakenly label sidewalk or driveway as street. The NCut method cannot precisely separate sidewalk from street since sidewalk is usually a very narrow region in the camera view. Also, the appearance of sidewalk is very similar to street. An interesting future direction of this work is to integrate the estimation of the semantic maps into the network architecture in an end-to-end trainable framework, which would require collecting more scenes for training.

References

[1] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *CoRR*, abs/1602.00763, 2016. 3

[2] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *CoRR*, abs/1612.03716, 2016. 3

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016. 3

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 3

[5] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, Aug. 2000. 3

[6] Ruichi Yu, Hongcheng Wang, and Larry Davis. Remotenet: Efficient relevant motion event detection for large-scale home surveillance videos. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2018. 1

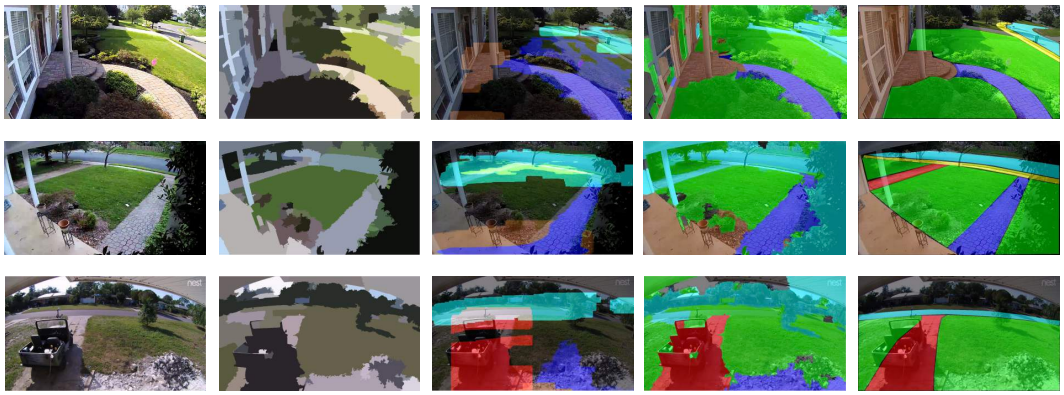


Figure 4. (a) shows the camera images. (b) shows the results of NCut. Each super pixel is represented using its color mean. (c) shows the heatmaps obtained from historical videos using our heuristic method. (d) shows the automatically generated maps. (e) shows the annotated maps.