IEEE Signal Processing Letters 25.10 (2018): 1495 – 1499l

DOI: https://doi.org/10.1109/LSP.2018.2865833

Automatic Semantic Parsing of the Ground-Plane in Scenarios Recorded with Multiple Moving Cameras

Alejandro López-Cifuentes, Marcos Escudero-Viñolo, and Jesús Bescós

***Abstract*—Nowadays, video surveillance scenarios usually rely on manually annotated focus areas to constrain automatic video analysis tasks. Whereas manual annotation simplifies several stages of the analysis, its use hinders the scalability of the developed solutions and might induce operational problems in scenarios recorded with Multiple and Moving Cameras (MMC). To tackle these problems, an automatic method for the cooperative extraction of Areas of Interest (AoIs) is proposed. Each captured frame is segmented into regions with semantic roles using a state-of-the-art method. Semantic evidences from different junctures, cameras and points-of-view are then spatio-temporally aligned on a common ground plane. Experimental results on widely-used datasets recorded with multiple but static cameras suggest that this process provides broader and more accurate AoIs than those manually defined in the datasets. Moreover, the proposed method naturally determines the projection of obstacles and functional objects in the scene, paving the road towards systems focused on the automatic analysis of human behaviour. To our knowledge, this is the first study dealing with this problematic, as evidenced by the lack of publicly available MMC benchmarks. To also cope with this issue, we provide a new MMC dataset with associated semantic scene annotations.**

***Index Terms*—multiple moving cameras, semantic segmentation, area of interest, PTZ, video surveillance, scene parsing.**

I. Introduction

THE field of view (FOV) of Pan-Tilt-Zoom (PTZ) cameras can be dynamically modified, a functionality that may be beneficial for video surveillance scenarios as it generally maximizes the sensor coverage, reducing the number of sensors required to fully cover a scene [1]. To cope with occlusions in these scenes, solutions based on Cooperative Camera Networks (CCNs) [2], [3]—in which each camera captures the scene from a different point of view—prevail over mono-camera systems. However, these solutions generally include static cameras or PTZ ones that are not moved. This, in our opinion, is due to the complexity of calibrating PTZ cameras as continuously updating is required to cope with the camera motion and mechanical fluctuations [4].

Video surveillance applications usually start with a detection stage to locate the objects of interest—e.g. humans—in the scene. This stage feeds subsequent analysis stages as object tracking and event detection. In CCNs where the cameras' FOVs overlap, cooperation can be carried out by combining, refining and aggregating individual detections on a common reference plane which, for simplicity, is usually the ground plane [5], [6]. The performance of detection methods has improved enormously in recent years thanks to the advent of deep learning solutions based on convolutional neural networks [7]. This performance can be further improved, both in terms of effectiveness and efficiency, by using scene information [8]. A common strategy is to filter out detections if their projection falls outside an Area of Interest (AoI) or focus area in the scene. The extent and shape of the AoI depend on the application; but AoI is usually enclosed in the ground plane where cameras' FOVs overlap and where best calibration precision is achieved. AoI is generally defined manually and by unqualified installation personnel [9], [10]. In this context, we claim that semantic segmentation may be useful to bypass this manual stage and automatize the process.

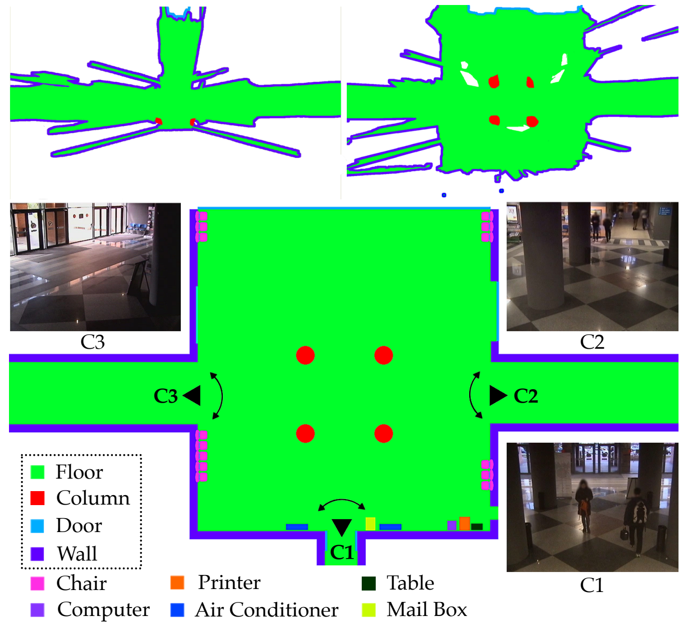

Fig. 1. Ground-plane semantic parsing in the MMC Dataset. Top row: Automatic partition with the proposed method for static (left) and moving (right) cameras. Bottom row: Example frames from the dataset, ground truth partition and semantic legend.

Semantic segmentation is the task of partitioning an image into a set of role-annotated segments with human-wise significance. Semantic segmentation is increasingly becoming an essential baseline for applications such as self-driving, robot sensing, and pedestrian detection. The performance of semantic segmentation has been boosted by the recent development of deep learning solutions [11]–[13] over vast and varied datasets [14], [15].

A. López-Cifuentes, M. Escudero-Viñolo, and J. Bescós are with the Video Processing & Understanding Lab, Universidad Autónoma de Madrid, Spain. E-mail Addresses: {alejandro.lopezc, marcos.escudero, j.bescos}@uam.es

This study has been partially supported by the Spanish Government through its TEC2014-53176-R HAVideo project.

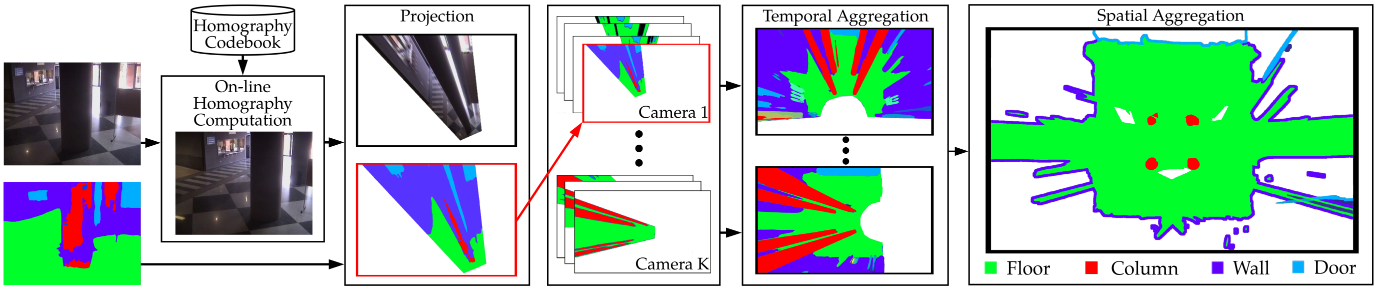

Fig. 2. Flowchart of the proposed method. From left to right: A frame is segmented into semantic regions. The closest view to this frame in a small codebook is obtained via local-feature matching. The frame and the semantic map are aligned to this view and projected onto the ground plane. The set of projected maps is temporally aggregated for each camera. Camera maps are finally spatially aggregated to yield an overall semantic partition of the ground plane.

In this paper, we pose a video surveillance scenario captured by three PTZ cameras moving in surveillance mode, i.e. each under a predefined but unsynchronized moving pattern (see Figure 1). We target to assign to every point in the ground plane a semantic label by projecting labels obtained for each camera view. These labels not only permit to define the AoI automatically but may also be used to identify the projection of functional (e.g. doors) and non-functional (e.g., columns or other obstacles) objects.

In order to project the labels obtained for each camera view, we need the respective calibration parameters or at least the homography that maps image pixels to the ground-plane. Our first contribution is an image alignment method which, based on the matching of local features, is able to estimate the ground-plane homography for every possible view. Our second contribution is a technique to efficiently combine the set of labels generated by the different non-synchronized moving cameras.

Due to the lack of datasets recorded with multiple and non-static PTZ-cameras a new Dataset for evaluation has been generated and made publicly available [16]. Results on this dataset suggest that the proposed method outperforms static and non-collaborative camera configurations. Additional qualitative results on widely-used datasets indicate that our method yields a broader AoI which is tighter to scene objects.

II. Proposed Method

A. Preliminaries

In a 3D scenario captured by K PTZ cameras which are moving, let $S = \{s_i\}$ be the set of semantic classes trained in a semantic segmentation algorithm, where classes represent roles as $floor$, $window$, etc. Let G be the ground plane of the scenario and let $\{\mathbf{P}_j,\ j = 1...J\}$ be the set of J points in this plane. Our aim is to assign semantic labels to these points. This is achieved by running the segmentation algorithm on every captured frame and then projecting and combining the semantic labels from each camera to the ground plane G.

Let I_n be an arbitrary frame n captured by the k-th camera oriented in a specific direction, and let $\mathbf{p}_j = (x, y, 1)^T$ be a pixel in I_n, expressed in homogeneous coordinates. We start from a semantic segmentation[1] of I_n, such that $\forall \mathbf{p}_j \in I_n,\ l_n(\mathbf{p}_j) = s_i,\ s_i \in S$.

[1]Extracted by [11] trained with the ADE20K Dataset [17]

Let $\mathcal{H}^k_{n,G}$ be a 3 by 3 homography matrix that projects pixels $\mathbf{p}_j$ onto G:

$$\mathbf{P}_j = \mathcal{H}^k_{n,G} \times \mathbf{p}_j \quad (1)$$

Each projected point $\mathbf{P}_j$ inherits from the I_n frame the semantic label assigned to $\mathbf{p}_j$:

$$l_n(\mathbf{P}_j) = l_n(\mathbf{p}_j) = s_i \in S. \quad (2)$$

B. Precomputed Homographies Codebook

For a static camera, a single homography matrix $\mathcal{H}_{n,G}$ is enough to project every captured frame onto G. However, if the PTZ camera moves capturing frame I'_n instead of frame I_n, then $H^k_{n,G} \neq H^k_{n',G}$. Obtaining all the ground-plane homographies for the k-th camera beforehand is infeasible, mainly due to the mechanical fluctuations. In this paper, we propose to pre-calculate a small subset or codebook of reference ground-plane homographies, via homogeneously sampling the orientations covered by a PTZ camera. In particular, we propose a default sampling in which a 40% of minimum overlap between consecutive reference views is imposed, which results in 9 view/homography pairs in each camera's codebook. The effect of this sampling is studied in Section III. The homography matrix for any other view/orientation is on-line computed via homography composition from the closest reference, thanks to the geometric properties of the pan-tilt motion. This requires that the codebook keeps reference homographies and the corresponding frame or view used to pre-compute each.

C. On-line Homography Computation

Let $\{I_r, H^k_{r,G}\}$ be the codebook of view/homography pairs for the k-th camera. For a new captured frame I_n, its closest view in the codebook $I_{\hat{r}}$ is obtained by a comparison strategy based on local feature descriptors. In particular, we use the AKAZE [18] method to extract a set of local features points $\Phi_n = \{\phi_{n,m}\}$ and the MLDB [18] method to describe them, leading to a set of descriptors $\Gamma_n = \{\gamma_{n,m}\}$.

Let Φ_r and Γ_r be the respective sets of features points and descriptors for a codebook view I_r. A match between two features $\phi_{n,m} \in \Phi_n$ and $\phi_{r,m} \in \Phi_r$ is defined by a distance [19] controlled by a resemblance weight $\alpha = 0.8$:

$$d(\gamma_{n,m}, \gamma_{r,m}) < \alpha\ d(\gamma_{n,m}, \gamma_{r,m'}), \quad \forall m \neq m'. \quad (3)$$

Let $\Phi_{n,r}$ be the set of matching feature points between I_n and I_r; this comparison process is repeated for every view in the codebook, and the closest view $I_{\hat{r}}$ is obtained by maximizing the number of matching points: $\hat{r} = \text{argmin}_r(|\Phi_n| - |\Phi_{n,r}|)$, where $|\Phi_n|$ is the cardinality of Φ_n.

We propose to obtain the homography to project pixels in I_n onto G by composing two matrices: $\mathcal{H}^k_{n,G} = \mathcal{H}^k_{\hat{r},G} \times \mathcal{H}^k_{n,\hat{r}}$. $\mathcal{H}_{n,\hat{r}}$ is the perspective rectification homography between I_n and $I_{\hat{r}}$. It is obtained by minimizing the back-projection error between corresponding points in $\Phi_{n,r}$. RANSAC [20] is used to fix the maximum allowed re-projection error to treat a point pair as an inlier. $\mathcal{H}_{\hat{r},G}$ is the precomputed homography in the codebook. The projection of every pixel $\mathbf{p}_j \in I_n$ (Eq. 1) and the assigning of their associated semantic labels (Eq. 2) results in the partial ground plane segmentation for the n-th frame of the k-th camera.

D. Temporal and Spatial Aggregation

In order to enlarge the covered area, and to globally reduce the impact of moving objects and segmentation errors, we propose to temporally aggregate the ground-plane semantic segmentations of several frames. Furthermore, to cope with scene occlusions and fuse semantic evidences from different cameras, we also propose a spatial aggregation process.

Temporal Aggregation: Given the semantic segmentation of a set of T frames $\{I_{n-T+1}, ..., I_n\}$ and the correspondent set of homography matrices, a given point in the ground plane $\mathbf{P}_j$ is assigned a set of $T_j \leq T$ semantic labels: $\{l_{n-T_j+1}(\mathbf{P}_j), ..., l_n(\mathbf{P}_j)\}$. A single temporally-smoothed label $\bar{l}_n(\mathbf{P}_j)$ is obtained as the mode value of this set. An example of this process is depicted in the fifth column of Figure 2.

Spatial Aggregation: Not every semantic class in S is projected correctly, points classified into a class not fully contained in G may produce distorted labelled regions (e.g. see regions labelled as *column* or *wall* in Figure 2 before spatial aggregation). These distortions can be corrected if evidences from multiple cameras are aggregated.

For a given scenario, we propose to divide the set S into three subsets: S_G, S_E and S_I.

- S_G includes the semantic classes or labels which represent ground classes, e.g. *floor, pavement, road*.
- S_E groups the semantic classes which represent scene enclosing concepts, e.g. *wall, door, window*.
- S_I contains the rest of the classes, which are associated to objects that are in the scene, e.g. *people, chair, column*.

Images of scene objects in the S_G or S_E subsets are usually views of the same object's face; the other object's faces are usually hidden. Differently, scene objects in the S_I subset may be captured from different points-of-view, which can be used to define their spatial extension. With this in mind, we define a hierarchy for the spatial aggregation.

First, points labelled in a S_G class by at least one of the $k = 1, ..., K$ cameras, are aggregated by union to shape the set of ground points $\mathcal{P}_G$:

$$\mathcal{P}_G = \bigcup_k^K \mathbf{P}_j, \quad \text{s.t } \bar{l}_{n,k}(\mathbf{P}_j) \in S_G \text{ for some } k. \qquad (4)$$

Second, points labelled in a S_E class that have not been previously assigned to $\mathcal{P}_G$, are also aggregated by union to create the set of enclosing points $\mathcal{P}_E$:

$$\mathcal{P}_E = \bigcup_k^K \mathbf{P}_j, \text{ s.t } \bar{l}_{n,k}(\mathbf{P}_j) \in S_E \text{ for some } k \text{ and } \mathbf{P}_j \notin \mathcal{P}_G. \quad (5)$$

This set needs to be refined to reduce the distortion produced due to the lack of additional views (see how *wall* areas extend to image boundaries in Figure 2 before spatial aggregation). To this aim, a subset $\mathcal{P}^*_E$ of $\mathcal{P}_E$ is created by removing all the points which are not neighbours of any point in $\mathcal{P}_G$.

The rest of the points, i.e. those labelled in S_I classes by all the cameras, are aggregated by intersection into $\mathcal{P}_I$ to reduce the distorted areas.

$$\mathcal{P}_I = \bigcap_k^K \mathbf{P}_j, \ s.t \ \bar{l}_{n,k}(\mathbf{P}_j) \in S_I \ \forall k \text{ and } \mathbf{P}_j \notin \{\mathcal{P}_G \cup \mathcal{P}_E\}. \quad (6)$$

The partition of the ground-plane is obtained by the union of the three sets $\mathcal{P} = \{\mathcal{P}_G \cup \mathcal{P}^*_E \cup \mathcal{P}_I\}$, keeping the labels assigned to each point. Qualitative examples of this process are depicted in Figures 1 (top row) and 2 (last column).

III. Experimental results

A. Experiments description

To assess the potential benefits of the proposed method, three experiments are carried out. **Ex.1** compares different semantic partitions $\mathcal{P}$ with respect to a manually generated ground-truth of a dataset which, in the absence of public datasets, has been generated to this aim [16]. **Ex.2** analyses, using the same dataset, the results' sensibility to the two parameters of the proposed algorithm (T and number of reference homographies). Finally, **Ex.3** evaluates the potential use of the method to automatically define AoIs in well-known datasets: Terrace [9], PETS2009 [21] and APIDIS [22].

Proposed dataset: Includes four sequences (5400 frames each at 30 fps) recorded on the same scenario but varying the cameras dynamics—*static*, *moving*—and the scenario occupancy—*crowded*, *empty*—.

Configurations: We pose two different configurations, one in which semantic maps for each camera are independently projected onto the ground plane (*non-cooperative*) and another in which the semantic maps are spatially aggregated by the process described in section II-D (*cooperative*). Temporal aggregation applies for both configurations.

Evaluation measures: Quantitative performance statistics are computed by measuring the spatial overlap between the manually annotated semantic ground-truth and the semantic segmentation $\mathcal{P}$. A point is a True Positive if it is assigned the same semantic role in both segmentations. Otherwise, it adds a False Positive for the class obtained and a False Negative for the class annotated. Precision (P), Recall (R) and F-Score (FS) measures are obtained from these statistics. For the non-cooperative configuration, per-camera measures are averaged.

TABLE I
Ex.1 Method's performance (%) in the proposed dataset

Configuration	Dataset		P	R	FS
	Occupancy	Dynamics			
Non-cooperative	*Empty*	*Static*	51.8	18.9	27.7
		Moving	53.4	50.0	51.6
	Crowded	*Static*	46.8	19.2	27.2
		Moving	51.9	48.9	50.4
Cooperative	*Empty*	*Static*	67.9	44.9	54.0
		Moving	**78.7**	**88.2**	**83.7**
	Crowded	*Static*	66.7	46.4	54.7
		Moving	**75.0**	**87.3**	**80.6**

TABLE II
Ex.1 Role-disaggregated performance (%) for moving cooperative cameras.

Semantic Role	*Empty*			*Crowded*		
	P	R	FS	P	R	FS
Floor	**88**	**96.5**	**92.0**	**85.7**	**95.7**	**90.4**
Column	56.8	91.4	70.0	59.2	86.8	70.3
Door	8.24	37.1	13.4	15.2	26.1	19.2
Wall	22.4	32.3	26.45	16.9	30.7	21.79

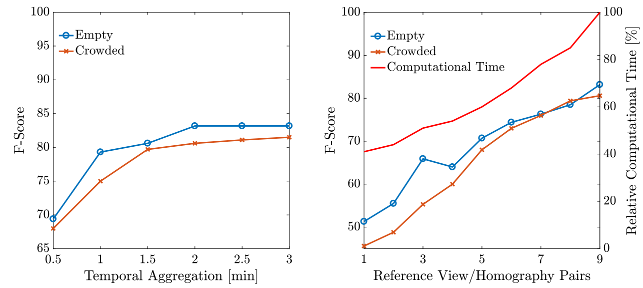

Fig. 3. **Ex.2** Sensitivity analysis: Overall F-Score with respect to the temporal aggregation parameter T (left) and the number of reference views with relative computational time (right) for moving cooperative cameras configuration.

Fig. 4. **Ex.3** Automatic AoI obtained by the proposed $\mathcal{P}_G$ partition (superimposed in green) compared to AoI manually annotated by the authors of Terrace [9], APIDIS [22] and PETS2009 [21] Datasets (red box).

System set-up: In **Ex.1** results are obtained using the default homography sampling (see Section II-B), and using all the frames for temporal aggregation, i.e. $T = 5400$. In **Ex.2** the impact of both T and codebook size is measured by sweeping on these values. For both **Ex.1** and **Ex.2** the three semantic sets S_G, S_E and S_I are used. Finally in **Ex.3**, we temporally aggregate results for the whole sequence and focus only on the S_G set; a single view/homography pair per camera is enough for these static-camera recorded scenarios.

B. Results and Discussion

Results for **Ex.1** presented in Table I suggest that overall performance is substantially increased if both temporal and spatial aggregation processes are used. In an empty scenario the use of static non-cooperative cameras yields a camera-averaged FS of 27%. It is relatively improved a 86% when using moving non-cooperative cameras, a 94% if results of static cameras are aggregated and a 202% if moving cooperative cameras are used. The presence of people in the scene produces a slightly decrease in the overall performance for every configuration but relative improvements are similar.

As expected, the method performs best for ground classes, which are respected by the ground-plane homographies (Table II). The spatial aggregation is able to reduce the distortion of inner objects as indicated by the performance on the *column* class. Results for the classes in the enclosing set (*door*, *wall*) are less effective; in our opinion, this may be due to annotation biases—how thick is the wall?—and inaccuracies of the semantic segmentation. Objects that lie above the ground-plane (*chairs*), are not part of the model (*printers*), or move (*persons*), are out of the scope of this letter.

Fig. 3 depicts **Ex.2** results for the moving/cooperative configuration. The left graph suggests that performance benefits from the use of a higher recording time in the temporal aggregation stage. However, as expected, performance converges when the complete scene has been captured at least once by the cameras—which in the analysed sequence occurs at around 1' ($T = 1800$)—. After that, improvement is mainly driven by the correction of semantic inconsistencies. The right graph confirms that the better we estimate homographies (i.e., the more view/homography pairs we use), the better are the results, but at the expense of higher computational resources. Hence, a trade-off number of views should be adequately selected for each scenario.

Fig. 4 summarizes **Ex.3** results on automatic AoI extraction over publicly available datasets. The extracted AoIs are more tightly adjusted to the scene ground objects and cover a broader area than the projections of the manually annotated areas provided by the authors. Our method effectively handles multi-class ground partitions, as in the AoI for the PETS2009 dataset, where $\mathcal{P}_G$ encompasses *road*, *grass*, *pavement* and *side-walks* points.

IV. Conclusions

This paper describes a novel approach to automatically define Areas of Interest (AoIs) in scenarios recorded with multiple PTZ cameras. The ground-plane homography for every captured frame is estimated via feature-based matching respect to a small set of pre-calculated reference view/homography pairs homogeneously sampling the orientations covered by each moving camera. Temporal and spatial aggregation strategies are used to obtain an accurate semantic partition of the reference plane. Results on a new dataset confirm the advantages of the proposed collaborative method. Besides, a qualitative evaluation carried out on widely-used datasets yields of broader, more precise and role-annotated AoIs that may be used by forthcoming methods to improve and enhance the analysis.

REFERENCES

[1] C. Micheloni, B. Rinner, and G. L. Foresti, "Video analysis in pan-tilt-zoom camera networks," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 78–90, 2010.

[2] A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa, "Object detection, tracking and recognition for multiple smart cameras," *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1606–1624, 2008.

[3] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *European Conference on Computer Vision*. Springer, 2006, pp. 98–109.

[4] Z. Wu and R. J. Radke, "Keeping a pan-tilt-zoom camera calibrated," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 8, pp. 1994–2007, 2013.

[5] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst, "Sparsity driven people localization with a heterogeneous network of cameras," *Journal of Mathematical Imaging and Vision*, vol. 41, no. 1-2, pp. 39–58, 2011.

[6] F. Fleuret, R. Lengagne, and P. Fua, "Fixed point probability field for complex occlusion handling," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 1. IEEE, 2005, pp. 694–700.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[8] A. García-Martín and J. C. SanMiguel, "Context-aware part-based people detection for video monitoring," *Electronics Letters*, vol. 51, no. 23, pp. 1865–1867, 2015.

[9] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 2, pp. 267–282, 2008.

[10] P. Peng, Y. Tian, Y. Wang, J. Li, and T. Huang, "Robust multiple cameras pedestrian detection with multi-view bayesian network," *Pattern Recognition*, vol. 48, no. 5, pp. 1760–1772, 2015.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016. [Online]. Available: http://arxiv.org/abs/1612.01105

[12] S. R. Bulò, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of dnns," *arXiv preprint arXiv:1712.02616*, 2017.

[13] Z. Wu, C. Shen, and A. v. d. Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *arXiv preprint arXiv:1611.10080*, 2016.

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *International Conference on Computer Vision (ICCV)*, 2017.

[16] "MMC Dataset," http://www-vpu.eps.uam.es/DS/MMC/.

[17] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] P. F. Alcantarilla and T. Solutions, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, no. 7, pp. 1281–1298, 2011.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 1999, pp. 1150–1157 vol.2.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[21] "PETS 2009 Dataset," http://www.cvg.reading.ac.uk/PETS2009/a.html.

[22] "APIDIS Dataset," https://sites.uclouvain.be/ispgroup/Softwares/APIDIS.