Learning to Select Views for Efficient Multi-View Understanding

Supplementary Material

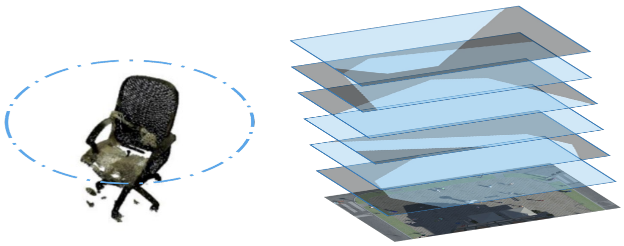

Figure 1. Example of multi-view camera setups. **Left**: multi-view classification jointly considers multiple camera views (blue dots) to identify the object. **Right**: multi-view detection estimates pedestrian occupancy from multiple cameras (blue FoV maps) over bird's-eye-view (bottom colored image). For both classification and detection tasks, due to hardware constraints, camera layouts are usually pre-defined.

1. MVSelect Architecture

As shown in Fig. 2, we design MVSelect architecture $d(\cdot)$ with two branches. The first branch expands the camera selection result $s_t^{\text{cam}} \in \mathbb{R}^N$ into D-dimensional learnable camera embeddings, and then sums over the selected embeddings to formulate a hidden vector. The second branch processes the observation $s_t^{\text{obs}} \in \mathbb{R}^D$, and converts that into another hidden vector. By combining the two hidden vectors, MVSelect outputs the action-value $Q(s, a)$, which measures the expected cumulative rewards for taking an action a in a given state s.

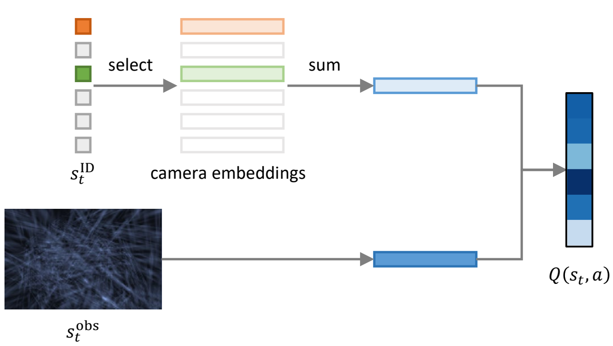

Figure 2. MVSelect architecture.

2. Additional Details on Experimental Setup

Datasets. We verify the performance of the proposed approach on multiview classification and detection tasks.

ModelNet40 is a subset of 3D CAD models in ModelNet [17]. It includes 40 categories of synthetic 3D objects with 9,843 training models and 2,468 test models. For multiview classification experiments, we use two different configurations: the *12-view* circular configuration from MVCNN [13] and the *20-view* dodecahedral configuration from RotationNet [9].

ScanObjectNN is a 3D dataset scanned from real-world objects. Introduced by Uy et al. [15], it contains 2902 3D objects across 15 categories. Traditionally used for point cloud classification, we re-purpose this dataset for multi-view classification by rendering textured meshes from the point clouds and use the same 12 views setup as ModelNet40 [13, 17].

Wildtrack [2] is a real-world multiview detection dataset with 7 camera views covering a 12×36 square meter area, which is represented as a 480×1440 grid from BEV. It contains 360 frames for training and 40 frames for testing.

MultiviewX [8] is a synthetic multiview detection dataset created using the Unity [14] engine. It has 6 cameras with higher pedestrian density than Wildtrack. It focuses on a 16×25 square meter area, which is discretized into 640×1000 BEV grid. Like Wildtrack, MultiviewX also contains 360 training frames and 40 testing frames.

Evaluation metrics. For multiview classification, we follow previous methods [6, 9, 11, 16, 19, 20] and report instance-averaged accuracy as the primary indicator.

Regarding multiview detection, we report the following metrics: multi-object detection accuracy (MODA), multi-object detection precision (MODP), precision, and recall [10]. During evaluation, we first compute false positives (FP), false negatives (FN), and true positives (TP), and then use them to calculate the metrics. Specifically, MODA is calculated as $1 - \frac{\text{FP}+\text{FN}}{\text{GT}}$, where GT is the number of ground truth pedestrians. MODP is calculated as $\frac{\sum 1-\text{dist}[\text{dist}<\text{thres}]/\text{thres}}{\text{TP}}$, where dist is the distance from the estimated pedestrian location to its ground truth and thres is the threshold of 0.5 meters. MODP indicates the BEV localization accuracy. Precision and recall are calculated as $\frac{\text{TP}}{\text{TP}+\text{FP}}$ and $\frac{\text{TP}}{\text{GT}}$, respectively.

All metrics are reported in percentages.

3. Evaluation against State-of-the-Arts

In Table 1, we compare our implementations of MVCNN [13] and MVDet [8] with their original implementations and state-of-the-art methods. On 3 datasets and 4 settings, our implementations outperform the original implementations and achieve competitive results. Although our focus is not on improving these classic architectures, the results indicate that they can still serve as strong baselines.

Table 1. Performance comparison with state-of-the-art multiview classification and multiview detection methods. Results are averaged from 5 runs. * indicates that the camera poses are dynamically chosen and do not follow a pre-defined layout. We also report the MVSelect and task network joint training results in the last line.

	ModelNet40 [17]	
	12 views	20 views
MVCNN [13]	90.1	92.0
GVCNN [4]	92.6	-
MHBN [20]	93.4	-
RotationNet [9]	-	94.7
RelationNet [19]	94.3	97.3
ViewGCN [16]	-	**97.6**
MVTN* [6]	93.8	93.5
MVCNN (our implementation)	**94.5**	96.5
MVCNN + MVSelect (2 views)	94.3	94.4

	Wildtrack [2]				MultiviewX [8]			
	MODA	MODP	prec.	recall	MODA	MODP	prec.	recall
RCNN & cluster [18]	11.3	18.4	68	43	18.7	46.4	63.5	43.9
POM-CNN [5]	23.2	30.5	75	55	-	-	-	-
DeepMCD [3]	67.8	64.2	85	82	70	73	85.7	83.3
Deep-Occlusion [1]	74.1	53.8	95	80	75.2	54.7	97.8	80.2
MVDet [8]	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
SHOT [12]	90.2	76.5	96.1	94.0	88.3	82.0	96.6	91.5
MVDeTr [7]	**91.5**	**82.1**	**97.4**	94.0	**93.7**	**91.3**	**99.5**	94.2
MVDet (our implementation)	90.0	80.9	95.4	**94.5**	93.0	90.3	98.7	**94.4**
MVDet + MVSelect (3 views)	88.6	79.9	93.3	94.2	88.1	89.8	98.2	89.7

Compared to state-of-the-arts that use full N cameras, joint training the tasks network along with MVSelect gives competitive results while only using $T = 2$ or $T = 3$ cameras for multiview classification and multiview detection.

References

[1] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 271–279, 2017. 2

[2] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. 1, 2

[3] Tatjana Chavdarova et al. Deep multi-camera people detection. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 848–853. IEEE, 2017. 2

[4] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 264–272, 2018. 2

[5] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2007. 2

[6] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2021. 1, 2

[7] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1673–1682, 2021. 2

[8] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2020. 1, 2

[9] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5010–5019, 2018. 1, 2

[10] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, 2008. 1

[11] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 1

[12] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6049–6057, 2021. 2

[13] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 1, 2

[14] Unity Technologies. Unity. https://unity.com/. 1

[15] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[16] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1850–1859, 2020. 1, 2

[17] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1, 2

[18] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu.

Multi-view people tracking via hierarchical trajectory composition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4256–4265, 2016. 2

[19] Ze Yang and Liwei Wang. Learning relationships for multi-view 3d object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7505–7514, 2019. 1, 2

[20] Tan Yu, Jingjing Meng, and Junsong Yuan. Multi-view harmonized bilinear network for 3d object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 186–194, 2018. 1, 2