ABSTRACT

We present MVMO (Multi-View, Multi-Object dataset): a synthetic dataset of 116,000 scenes containing randomly placed objects of 10 distinct classes and captured from 25 camera locations in the upper hemisphere. MVMO comprises photorealistic, path-traced image renders, together with semantic segmentation ground truth for every view. Unlike existing multi-view datasets, MVMO features wide baselines between cameras and high density of objects, which lead to large disparities, heavy occlusions and view-dependent object appearance. Single view semantic segmentation is hindered by self and inter-object occlusions that could benefit from additional viewpoints. Therefore, we expect that MVMO will propel research in multi-view semantic segmentation and cross-view semantic transfer. We also provide baselines that show that new research is needed in such fields to exploit the complementary information of multi-view setups.

Index Terms— multi-view, cross-view, semantic segmentation, synthetic dataset

1. INTRODUCTION

The task of semantic segmentation [1] aims at, given an input image, performing pixel-wise classification over a predefined set of categories. As in many other dense prediction problems, the end-to-end convolutional neural networks (CNN)-based fully supervised approach to this task has become the de facto standard to solve it, leading to robustly performing models [2] at the expense of a large amount of human annotations. Nevertheless, understanding scenes based on a single 2D input is challenging when applied on (i) scenes with significant inter-object and self-occlusions that hide class-distinctive features (ii) scenes covering a wide spatial range, where distant objects can show a small apparent size.

In this context, we hypothesize that posing data-driven models that exploit multi-view camera setups that provide complementary information over the imaged scenes could be of potential interest for improving the results obtained by single-camera baselines. However, so far multi-view semantic segmentation has primarily been approached for
close-baseline setups [5] i.e. those where the distance between cameras (and thus, the resulting disparities) are small, whereas solving the aforementioned obstacles requires wide baselines. Scenarios that could benefit from this approach are frequent in real life, in domains as diverse as industry (e.g. conveyor belts), surveillance, or traffic management.

In this paper, we introduce MVMO, the Multi-View Multi-Object dataset, which addresses the current lack of publicly available large-scale datasets of densely annotated wide-baseline multi-view scenes containing multiple objects. MVMO is a synthetic, path tracing-based set of 116,000 scenes with per-view semantic segmentation annotations of 10 object categories. Each scene is observed from 25 camera locations distributed uniformly in the upper hemisphere (see Fig. 1). Unlike most existing multi-view image datasets (which are designed to be camera-centric and exhibit very close baselines while sensing their surroundings [5]), MVMO features wide baselines between many camera pairs as a result of a scene-centric design, and a large amount of objects per scene. This leads to large disparities, notable occlusions and variable apparent object geometry, size and surface appearances across views. Therefore, MVMO sets a particularly challenging arrangement that aims at contributing to push research on the fields of multi-view semantic-segmentation and cross-view semantic knowledge transfer. The experiments presented show that simple baselines fail to be of much help in transferring learned models to novel views, hence suggesting the need for novel research in this direction.

Related work. Our work relates to a number of previous datasets from various research fields, some of which already leverage wide-baseline multi-view datasets in an attempt to improve upon their respective single-view performances: In multi-view object detection, [10] introduces a multi object detection dataset with bounding box annotations for pedestrians, cars and buses from 6 calibrated cameras. Advances on multi-view human pose estimation were possible by leveraging various wide baseline datasets over RGB [11, 12, 3] and depth [13] images of both groups [14] and individuals.

The field of multi-view semantic segmentation (see Table 1) has been addressed from diverse perspectives. Many early works prior to the irruption of deep learning techniques focused on the binary segmentation of a single static foreground object from a sequence of close-baseline views from a class-agnostic point of view [15, 16], often learning sequence-specific models and relying on diverse cues: object-background color distributions, central object fixation or stereo geometry constraints. More recently, [17] used deep self-supervised training to extend the single subject segmentation task to three dynamic scenes in wide-baseline setups.

Table 1: Datasets for multi/cross-view semantic segmentation. The table shows the lack of datasets with wide baseline and high object density addressed by MVMO. **Object Density:** #objects/scene. Does not apply to close baseline scenarios. **Representation:** 2D→3DS: 3D Surface reconstructed from 2D. 3DVE: 3D Virtual Environment. 3DM→2D: 3D Model rendered to 2D images. **Photorealism:** S: Subject. B:Background. IBR: Image-Based Rendering. PCR: Point Cloud Rendering (view synthesis from Point Cloud). RT: Ray-Tracing. PT: Path-Tracing. UOM: Uniform Object Materials. ⋆: Needs to be placed/configured/generated by user; images are not readily available.

| Dataset                  | Wide Baseline | Object Density | Representation | Photorealism | # Scenes    | # Views | # Classes |
|--------------------------|---------------|----------------|----------------|--------------|-------------|---------|-----------|
| Human3.6M [3]            | Yes           | Low (1)        | 2D images      | Real         | 900,000 in 165 sequences | 4       | 24        |
| 3Dpeople [4]             | Yes           | Low (1)        | 3DM→2D         | S: High B: Low | 616,000 in 5,600 sequences | 4       | 8(clothes)/14(body) |
| SYNHIA [5]               | No            | N/A            | 3DM→2D         | Low          | 51,000 in 51 sequences | 8       | 13        |
| ScanNet [6]              | ⋆             | Low            | 2D→3DS         | High         | 1.5k        | ⋆       | 40        |
| House3D [7]              | ⋆             | Low            | 3DVE           | Low          | 45.6k       | ⋆       | 80        |
| Gibson [8]               | ⋆             | Low            | 3DVE           | High (IBR/PCR)| 1.4k        | ⋆       | 40        |
| CARLA [9]                | ⋆             | ⋆              | 3DVE           | Mid-High (RT)| ⋆           | ⋆       | 12        |
| MVMO (ours)              | Yes           | High (15-20)   | 3DM→2D         | High (PT, UOM)| 116k (uncorrelated) | 25      | 11        |
various large scale 3D virtual or reconstructed environments have been released. Their relevance comes from the fact that, through significant user intervention, parts of the 3D model and associated labels could be projected back to 2D to synthesise semantically annotated multi-view image sets from arbitrary camera locations with different degrees of realism. The House3D [7], Gibson [8] and CARLA [9] environments are some relevant examples, although only CARLA, being fully virtual, could yield high object densities via its API. This was shown in [25] for close baseline setups, proposing a multi-view semantic fusion scheme from up to 8 input views onto a new virtual zenithal view. Concurrent to our work, [26] presents Kubric, a new flexible dataset generation framework with multiple dense annotation features.

In conclusion, MVMO covers the lack of a standardised large scale photo-realistic multi-view dataset with wide-baselines (and hence, large disparities and relevant occlusions) across cameras and comprising semantic segmentation annotations for multiple objects of distinct classes.

2. MVMO DATASET CONSTRUCTION

We use Blender’s Python API for procedural 3D scene construction and image rendering, using the ModelNet10 3D object dataset [27] as repository of well-categorized 3D shapes of 10 common object classes. We build a basic scene with a grey plane at $z = 0$ and a single zenithal rectangular key light, and define a $2.8 \times 2.8m$ rectangular area for object placement. All cameras are projective cameras with a focal length of $f = 35mm$, oriented to the origin. The camera locations are determined by sampling the surface of a hemisphere of $r = 3m$ regularly so that they are equidistributed [28]. For our set of 25 samples, this yields locations at 4 levels (Fig. 1): 1 view at L0 (top, at $z = 3.0m$), 3 views at L1 ($z = 2.90m$), 9 views at L2 ($z = 2.12m$) and 12 views at L3 ($z = 0.78m$).

Then, for each scene: (i) we randomly select one of the 10 categories of ModelNet10 and (ii) sample one shape from the selected class, (iii) we normalize its scale so that its largest dimension is 1.0m, then applying a random scale in the $[0.3 – 0.8]$ range, (iv) we select a random base-color from a set of 9,284 predefined ones and apply a random combination of the specularity, roughness and metallic material modifier properties that -together with other fixed property values- define the Bidirectional Scattering Distribution Function (BSDF) of the materials applied to the whole shape. (v) we place it on the $z = 0$ plane of our base scene, in a random location (within the designated limit area) and angle, checking that the mesh does not intersect with any previously placed object. (vi) Once $15 – 20$ objects are placed, the scene and fine-detailed ground truth images are rendered with the Cycles engine for each of the 25 views at $256 \times 256$ pixels, producing photo-realistic, unbiased and physically consistent shading, reflectance and material effects, including specularities, and interreflections.

The 116,000 created scenes (each with 25 views) were then partitioned in a train set (100,000), two validation and two test sets (4,000 each). The latter are created based on whether the used ModelNet10 shapes were already used for the train set (SO: Same Objects) or come from a held-out set of shapes (OO: Other Objects) from the same categories, which poses a harder problem. Fig. 2 shows the resulting distributions of objects per category and scene for the train set.

This proposed wide-baseline multi-object dataset contains many occlusions, making semantic segmentation from a single view difficult. We think MVMO can facilitate research in multiple directions. We highlight two of them: (i) Multi-view semantic segmentation: existing close-baseline datasets have only few occlusions. Therefore, the proposed dataset makes for a more interesting setup for multi-view semantic segmentation. (ii) Cross-view semantic transfer: this is an especially exciting research direction which can be performed on MVMO. In real-life applications the dense labelling of all views is infeasible. Hence we believe that methods need to be designed that can learn to perform multi-view semantic segmentation based on labels from only a single view.

3. EXPERIMENTAL BASELINES

We run two baseline experiments for the cross-view semantic transfer problem. These experiments are included to show that there is no simple solution to this task and it is indeed an open research problem. To conduct them we select 5 representative views from three distinct levels: L0.cam0 (zenithal), L2.cam8, L2.cam12, L3.cam.13 and L3.cam.22 (see Fig. 1). In both cases we use a U-Net [2] as our semantic segmentation model, with an Imagenet-pretrained ResNet50 backbone.

**Experiment 1. Cross-view semantic transfer via direct testing** We train an independent model with each of the considered views and directly test them against every other camera’s test sets, without any specific adaptation. Table 2 shows the results in terms of Intersection over Union (IoU): The diagonals correspond to standard fully supervised single-view setups. We see that these improve as we adopt a higher perspective of the scene. As one might expect, direct semantic transfer between cameras placed within the same level (e.g. L2.cam8/L2.cam12) yields a minimal performance drop, on account of the quasi-invariance of the learned representations.
Table 2: IoU results for direct cross-view semantic transfer. Five models trained on 100% of the train set (100k scenes).

| Subset   | test\train | cam0 | cam8 | cam12 | cam13 | cam22 |
|----------|------------|------|------|-------|-------|-------|
| L0.cam0  | 71.12      | 29.09| 29.61| 14.28 | 14.88 |
| L2.cam8  | 24.63      | 70.21| 70.16| 28.14 | 28.54 |
| L2.cam12 | 25.14      | 69.09| 70.05| 27.73 | 28.29 |
| L3.cam13 | 12.18      | 31.26| 31.46| 59.18 | 58.72 |
| L3.cam22 | 12.11      | 30.10| 30.59| 58.39 | 59.41 |

Table 3: IoU results for planar homography-based transfer.

| Subset   | L0.cam0→L2.cam8 | L2.cam8→L0.cam0 |
|----------|-----------------|-----------------|
| Other obj.| 28.72           | 24.35           |
| Same obj. | 31.29           | 24.84           |

Fig. 3: Failure cases from monocular models in Table 2. a) self-occlusion (golden object) b) inter-object occlusion (sofa under the yellow desk) c) small objects (light pink and dark green objects) d) ambiguity from specular inter-reflection (light blue object with reflections of the cyan one). Last row shows a second view that could help solve the ambiguity.

\[ H_{z=0.2 \rightarrow 1} \] (ii) feed this to \( f_{v_2 \rightarrow ss_1} \) so as to obtain a semantic map referenced to \( v_1 \) (iii) transform this back to be referenced to \( v_2 \) with the inverse homography \( H_{z=0.1 \rightarrow 2} = H_{z=0.2 \rightarrow 1}^{-1} \). We test this on two cameras at distinct levels: L0.cam0 and L2.cam8. The lack of a significant performance gain in the results (see Table 3) over the direct transfer baseline from Table 2 shows that, as expected, the planar homography fails to help for the general, wide-baseline case, in which a good estimate of pixel-wise depth information from every secondary view is needed for unambiguous matching.

The failure of both experimental baselines, along with the fragility of photometric cues in wide baseline scenarios [17], suggests that exploiting the complementary information given by additional views of the scene in a data-driven multi-view learning setup or transferring the knowledge from trained models across views in unsupervised scenarios will require the development of new theoretical approaches.

4. CONCLUSION

We presented MVMO, a wide baseline multi-view synthetic dataset with semantic segmentation annotations that features a high object density and large amount of occlusions. We expect MVMO will propel research in multi-view semantic segmentation and cross-view semantic transfer and, likely through domain adaptation, address the current limitations of monocular setups in heavily-occluded real world scenes.
5. REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” in CVPR, 2015.

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in MICCAI, 2015.

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, July 2014.

[4] A. Pumarola, J. Sanchez, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, “3DPeople: Modeling the Geometry of Dressed Humans,” in ICCV, 2019.

[5] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez, “The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes,” in CVPR, 2016.

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner, “ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes,” in CVPR, 2017.

[7] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian, “Building Generalizable Agents with a Realistic and Rich 3D Environment,” arXiv:1801.02209 [cs], Jan. 2018.

[8] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, “Gibson Env: Real-World Perception for Embodied Agents,” in CVPR, 2018.

[9] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, “CARLA: An open urban driving simulator,” in CoRL, 2017.

[10] G. Roig, X. Boix, H. B. Shitrit, and P. Fua, “Conditional Random Fields for multi-camera object detection,” in ICCV, 2011.

[11] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, “Multi-view body part recognition with random forests,” in BMVC, 2013.

[12] Y. Yao, Y. Jafarian, and H. S. Park, “MONET: Multi-view Semi-Supervised Keypoint Detection via Epipolar Divergence,” in ICCV, 2019.

[13] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, “Towards viewpoint invariant 3D human pose estimation,” in ECCV, 2016.

[14] H. Joo et al., “Panoptic Studio: A Massively Multiview System for Social Motion Capture,” in ICCV, 2015.

[15] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, “Automatic 3D object segmentation in multiple views using volumetric graph-cuts,” Image Vis. Comput., vol. 28, no. 1, pp. 14–25, Jan. 2010.

[16] W. Lee, W. Woo, and E. Boyer, “Silhouette Segmentation in Multiple Views,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 7, pp. 1429–1441, July 2011.

[17] Y. Yao and H. S. Park, “Multiview Co-segmentation for Wide Baseline Images using Cross-view Supervision,” in WCACV, 2020.

[18] L. Ladicky et al., “Joint Optimization for Object Class Segmentation and Dense Stereo Reconstruction,” Int. J. Comput. Vis., vol. 2, no. 100, pp. 122–133, 2012.

[19] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, “Predicting Ground-Level Scene Layout From Aerial Imagery,” in CVPR, 2017.

[20] B. Coors, A. P. Condurache, and A. Geiger, “NoVA: Learning to see in novel viewpoints and domains,” in 3DV, 2019.

[21] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool, “Learning Where to Classify in Multi-view Semantic Segmentation,” in ECCV, 2014.

[22] L. Ma, J. Stückler, C. Kerl, and D. Cremers, “Multi-view deep learning for consistent semantic mapping with RGB-D cameras,” in IROS, 2017.

[23] A. Dai and M. Niessner, “3DMV: Joint 3D-Multi-View Prediction for 3D Semantic Scene Segmentation,” in ECCV, 2018.

[24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor Segmentation and Support Inference from RGBD Images,” in ECCV, 2012.

[25] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, “Cross-View Semantic Segmentation for Sensing Surroundings,” IEEE Robot. Autom. Lett., vol. 5, no. 3, pp. 4867–4873, July 2020.

[26] K. Greff et al., “Kubric: A scalable dataset generator,” in CVPR, 2022.

[27] Z. Wu et al., “3D ShapeNets: A Deep Representation for Volumetric Shapes,” in CVPR, 2015.

[28] M. Deserno, “How to generate equidistributed points on the surface of a sphere,” Tech. Rep., 2004.

[29] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, second edition, Apr. 2004.