CMPT828 CVPR Workshop Selection

Grant Tingstad
NSID: gdt741
University of Saskatchewan
gdt741@usask.ca

1. Introduction

I have chosen the 7th Workshop on Benchmarking Multi-Target Tracking: How Far Can Synthetic Data Take Us? [1], as presented at the CVPR2022 conference. While the subject focuses on semantic segmentation challenges using synthetic data, it sets a precedent for using computer generated synthetic data for many computer vision research applications such as photogrammetry and vision-and-language navigation. Please note that I searched for projects included in the announced 2023 CVPR workshops but could not find any posted challenges that aligned with my career interests.

2. Background and Challenge Discussion

Deep learning requires large amounts of data to achieve favorable performance, and privacy concerns arise often when data is collected from public spaces. Moreover, annotations for large data sets can be very difficult and expensive to generate, the cost of which increases significantly as annotation variables increase; for example, a deep learning model for human pose estimation would require annotations of joint position and orientation for each image in the set. Considering these factors, the motivation for using photorealistic simulations to generate deep learning data sets should be clear – this method would enable unlimited data generation and automatic annotation with almost zero privacy concerns. When considering the use of synthetic data for deep learning applications, however, care must be exercised to ensure that avoidable bias is prevented. For example, sufficient randomization must be programmed into human models, environments, and lighting to ensure that the range of synthetic data is representative of randomization within the natural world.

The challenge associated with the selected workshop was the MOTSynth2MOT17 track competition, which required participants to build a multiple-object-tracking pedestrian segmentation network to generate pedestrian bounding boxes on real pedestrian footage when trained on the artificially generated MOTSynth data set. The MOTSynth data set contains a series of annotated pedestrian tracking videos generated within Grand Theft Auto V. In total, MOTSynth contains data for over 45,000 annotated pedestrians in a variety of simulated environments. The test set, MOT17, contains real-world footage of pedestrians in both indoor and outdoor environments, and includes annotated bounding boxes for over 750 pedestrians.

The competition was hosted on a website set up by the organizers [2] and had eight submitting participants. Projects were scored with a variety of methods; however, overall score was determined with Higher Order Tracking Accuracy [HOTA], which computes a mean of detection accuracy and association accuracy [6]. While the implementation of participants was not shared, some results varied significantly. For example, the first-place implementation, SIA_Track, achieved a HOTA score of 63.2 and a run speed of 24.2Hz, while the second-place implementation, SDTracker, achieved a similar HOTA score of 61.4 but a run speed of 1.2Hz. The competition concluded in May of 2022, but training and test data for this competition is still available on the organizer's web page and will be used for my implementation.

3. Literature Review

Multi Object Tracking implementations with synthetic training data have been on the rise as photorealistic simulations have become more available to the public. In 2018, Nvidia researchers and collaborators demonstrated that vehicle detection models trained on synthetic data, when fine tuned on real-world data outperform models trained on real-world data alone [11]. The researchers designed a simulation environment to load 3D models of cars and perform domain randomization by randomizing variables such as background, lighting, and camera orientation. Other 3D objects, like blocks and spheres, were also randomly introduced to the scenes to enhance variability of the output images. Ground truths (bounding boxes) were generated for vehicles in the scenes and over 1 million annotated images were generated. Object detection models were pre-

trained on ImageNet [10] and trained on a variety of CNN architectures, but particular focus was given to Faster R-CNN [9]. After fine-tuning with real-world data, the models were tested on the VKITTI [4] self-driving car data set, which includes annotated images of real vehicles. The work demonstrated that models trained on synthetic data and fine-tuned with real data achieved higher performance than models trained on real data alone using average precision at 50% IoU as a metric. The authors also found that fine tuning with real data significantly increased the performance of models trained on synthetic data regardless of the size of the training data set; although, it should be noted that the simulation environment used in these experiments did not produce photorealistic images to the extent that more modern software is capable.

The organizers of the MOTSynth2MOT17 track workshop created the synthetic data set MOTSynth from simulations within the popular video game Grand Theft Auto V. In a 2021 paper, Matteo Fabbri and his collaborators discuss some challenges in machine learning research that may be overcome with synthetic data [3]. A variety of annotations are included in MOTSynth, including pose information, depth maps, and instance segmentation with bounding boxes, making the data set an excellent candidate for pedestrian detection and more advanced fields of computer vision, like robot reinforcement learning. The authors trained pedestrian-detectors on both MOTSynth and a popular object segmentation data set, COCO [7], with a variety of object detection CNN models, including YOLOv3 [14], CenterNet [13], Faster R-CNN, and Mask R-CNN [5]. All networks were pre-trained with ImageNet weights. After testing each model on real-world annotated footage of pedestrians, the authors showed that all models trained on synthetic data outperformed models trained on COCO's pedestrian data alone. This research further supports the case for using synthetic data for real-world detection problems since synthetic data collection is efficient, inexpensive, and free of privacy concerns. Most importantly, it has been demonstrated that images generated from simulated environments can effectively train real-world object detectors with very good results.

2D human pose estimation is another area of computer vision that has been gaining attention in recent years. Successful pose estimation could help computers discern more information from scenes with humans, such as scene context and human behavior recognition (e.g., eating, jumping, sitting). Until recently, the field has been held back by a lack of annotated data; manual labelling of 3D poses for deep learning data sets is prohibitively expensive. In a 2017 paper, Varol et. al. discuss their work on a 2D pose estimation data set trained on synthetic data consisting of millions of generated RGB images, depth maps, and corresponding body part & pose annotations [12]. A series of models were trained using a limited set of real pose data from the FSitting data set [8], and synthetic pose images generated by the authors. The models were then tested on a variety of ground truth data sets. In one experiment, a model trained on real data achieved a mean body part segmentation IoU of 28%; the same experiment using a synthetic-trained model achieved a mean IoU of 40%. By fine tuning the synthetic model with data from real images, a mean IoU of 60% was achieved. This work shows that large sets of synthetic data can be very helpful in training models for challenging segmentation tasks, particularly when fine-tuning with significantly smaller data sets consisting of real images.

References

[1] How far can synthetic data take us? 7th workshop on benchmarking multi-target tracking. https://motchallenge.net/workshops/bmtt2022/. Accessed: 01.19.2023. 1

[2] Motsynth-mot-cvpr22. https://motchallenge.net/data/MOTSynth-MOT-CVPR22/. Accessed: 01.19.2023. 1

[3] Fabbri et al. Motsynth: How can synthetic data help pedestrian detection and tracking? In *ICCV*, 2021. 2

[4] Gaidon et al. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 2

[5] He et al. Mask r-cnn. *arXiv preprint*, 2017. 2

[6] Luiten et al. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*. 1

[7] Lin et al. Microsoft coco: Common objects in context. In *CVPR*, 2014. 2

[8] Oliveira et al. Deep learning for human part discovery in images. 2016. 2

[9] Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, pages 91–99, 2015. 2

[10] Russakovsky et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 2015. 2

[11] Tremblay et al. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *CVPR*, 2018. 1

[12] Varol et al. Learning from synthetic humans. In *CVPR*, 2017. 2

[13] Zhou et al. Objects as points. *arXiv preprint*, 2019. 2

[14] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint*, 2018. 2