Real-time synthetic-to-real human detection for robotics applications

Maria Tzelepi, Charalampos Symeonidis, Nikos Nikolaidis, and Anastasios Tefas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
email: {mtzelepi, charsyme, nnik, tefas}@csd.auth.gr.

***Abstract*—During the recent years, Deep Learning achieved exceptional performance in various computer vision tasks, paving auspicious research directions for its application in robotics. A key component for its exceptional performance is the availability of sufficient training data. However obtaining such amount of training data constitutes a challenging task, especially considering robotics applications. Thus, synthetic data have recently been regarded as a promising tool to overcoming the data availability problem. In this work we first build a synthetic human dataset, and then we train a lightweight model, capable of operating in real-time for high-resolution input on low-power GPUs, for discriminating between humans and non-humans. The target of this work is to assess the generalization of the model trained on synthetic data, to real data, and also to explore the effect of using (few) real images in the training phase. As it is shown through quantitative and qualitative results the use of only few real images can beneficially affect of the performance of the synthetic-to-real real-time model.**

***Index Terms*—Synthetic-to-real, human detection, real-time, heatmaps, robotics.**

I. Introduction

During the recent years, Deep Learning (DL) attained widespread popularity due to its exceptional performance on various computer vision tasks [1]–[4]. Its impressive performance on computer vision, paved auspicious research directions for its application in robotics [5]–[8]. A key component for the successful performance of DL algorithms is the availability of sufficient training data. State-of-the-art DL models require millions of training examples [9]. However, obtaining such amount of training data, especially considering robotics applications, constitutes a challenging task. Thus, synthetic data, i.e., data generated artificially rather than by actual events, have recently been regarded as a very promising tool to circumvent the data availability problem [10].

The use of synthetic data is accompanied, in general, by various benefits linked with their low-cost nature and ability to meet specific requirements imposed by the application, which may not be feasible in real data. Thus, synthetic data have been utilized in a wide range of robotics applications, e.g., [11]–[14]. Their application on robotics applications is associated with a series of specific advantages. A few of those follow below: 1) synthetic data provide detailed annotations, since these are automatically produced, without containing errors usually occurring in the manual annotation process; 2) they are usually large in scale, since they are procedurally generated; 3) they minimize the risk of DL methods deployed in simulation environments in robotics to exhibit unstable behaviours or complete failures, due to not having been adapted to the visual differences between the virtual and the real world data.

A key issue associated with the successful use of synthetic data in robotics is the gap between the generated data and their deployment considering real data (that is, synthetic-real gap). The need for bridging this gap has fueled a new research area [15]–[17].

In this work, we first build a synthetic dataset for discriminating between humans and non humans, and use it to train a lightweight fully convolutional model that is capable of operating in real-time (about 25 Frames Per Second - FPS) utilizing a low-power GPU for high resolution input [18]. The target is to use the model to provide semantic heatmaps of human presence on real data. That is, we train the real-time model on the synthetic data, and we test the model on unseen images that contain real humans, producing semantic heatmaps, as explained in [18]. A main objective of this work is to assess the generalization of the model to real data, and investigate the effect of using real images in the training phase. As it is demonstrated in the experimental evaluation the use of even few real training examples can considerably improve the performance of training merely with synthetic data, while this is also reflected in the qualitative evaluation through the produced heatmaps.

The remainder of the manuscript is organized as follows. Section II presents in detail the proposed synthetic-to-real real-time human detection model, including the real-time model and the constructed synthetic dataset. Subsequently, in Section III the experiments conducted to assess the performance the synthetic-to-real real-time model, both quantitatively and qualitatively, are provided. Finally the conclusions are drawn in Section IV.

II. Real-time Synthetic-to-Real Human Detection

In this work we propose a *synthetic-to-real real-time model* for discriminating between humans and non humans. The core objective of this work is to assess the generalization of the model trained on synthetic data, to real data, and also to explore the effect of using (few) real images in the training phase. In the following Sections we describe the real-time

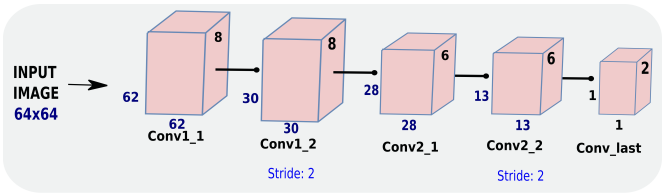

Fig. 1. Architecture of the real-time VGG-1080p model.

Fig. 2. Sample images of Synthetic Human dataset.

model architecture and the generation of the *synthetic human* dataset.

A. Real-Time Model

In this work, we train a fully convolutional lightweight on synthetic data, that is able to operate in real-time for detecting humans in real-images, considering high-resolution input on a low-power GPU. That is, the VGG-1080p model [18] is used, consisting of five convolutional layers 11K parameters. The model's architecture is illustrated in Fig. 1. The model runs in real-time (i.e., 25.6 FPS) on a Jetson TX-2 for 1080p input. More specifically, the network is trained on synthetic images of size 64×64, and then in the test phase, real images of size 1920×1080 are propagated to the network, and for every window 64×64 the output of the network at the output layer is computed, in order to generate the heatmaps of human presence.

B. Synthetic Human Dataset

The *synthetic human* dataset consists of real background images populated with 3D human models in various poses. PIFu [19], a state-of-the-art deep learning method for generating realistic 3D human models from single-view images, is used to generate the human models. The dataset consists in 133 human models, generated using full-body images of people from the Clothing Co-Parsing [20] dataset as PIFu's input. The Cityscapes [21] dataset which is composed of video sequences depicting street scenes in various cities, was used to take background images. The 3D human models are placed on potential 2D image locations (e.g., pavements, roads), based on coarse annotations for semantic image segmentation provided by Cityscapes, so as to manage a higher level of realism.

Since, the target is to train models that can run in real-time on high-resolution input for producing heatmaps of human presence [18], the generated images are cropped, and a train set of 20,000 synthetic cropped images containing humans is constructed. The train set also contains 20,000 non human images, cropped from images of the Cityscapes dataset. The test set consists of 4,000 real images containing humans and 4,000 real images without humans, cropped from video frames that were gathered by querying YouTube video search engine with random keywords. The cropped images are of size 64×64. Since a main objective of this work is to evaluate the effect of real-human images on the train set, we also construct four additional versions of the train set where 100, 200, 500, and 1000 out of 20,000 images are real-human images, while the rest are synthetic. The real human images are derived from the CUHK Person Re-identification datasets [22], [23]. Sample images of the constructed dataset are provided in Fig. 2.

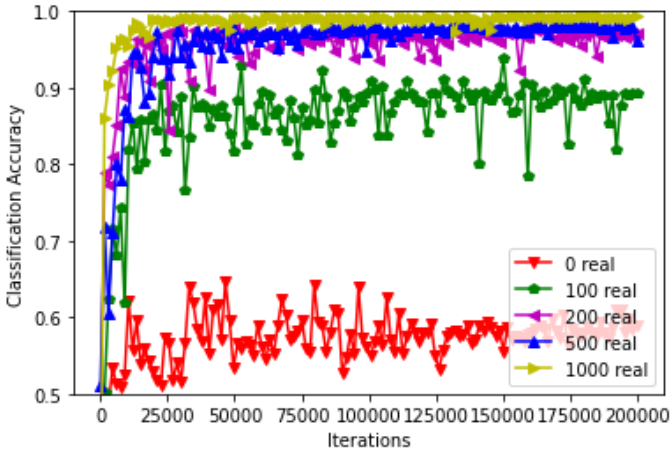

Fig. 3. Classification accuracy using the synthetic-to-real real-time model trained with 0, 100, 200, 500, and 1000 real images throughout the training iterations.

III. Experimental Evaluation

A. Evaluation Metrics and Implementation Details

Two sets of experiments were conducted. First, the performance of the synthetic-to-real real-time human detection model is evaluated using classification accuracy (test accuracy) as evaluation metric. Furthermore, the training curves of classification accuracy throughout the training iterations. Second, qualitative results are provided using the proposed synthetic-to-real real-time model. The model is used to produce heatmaps of human presence on real unseen high-resolution test images. The models are trained for 200,000 iterations (i.e., 320 epochs) using the mini-batch gradient descent with mini-batch of 64 samples, and we set the learning rate to 10^{-3}.

B. Experimental Results

In Table I we provide the classification accuracy of the synthetic-to-real real-time model trained merely with synthetic data, and with 100, 200, 500, and 1000 real images. As it is demonstrated, the model trained only with synthetic humans achieves sufficient performance, while as we include real human images, we can accomplish progressively increased performance. We can notice that even by adding only 100 real images the performance is remarkably improved. Furthermore,

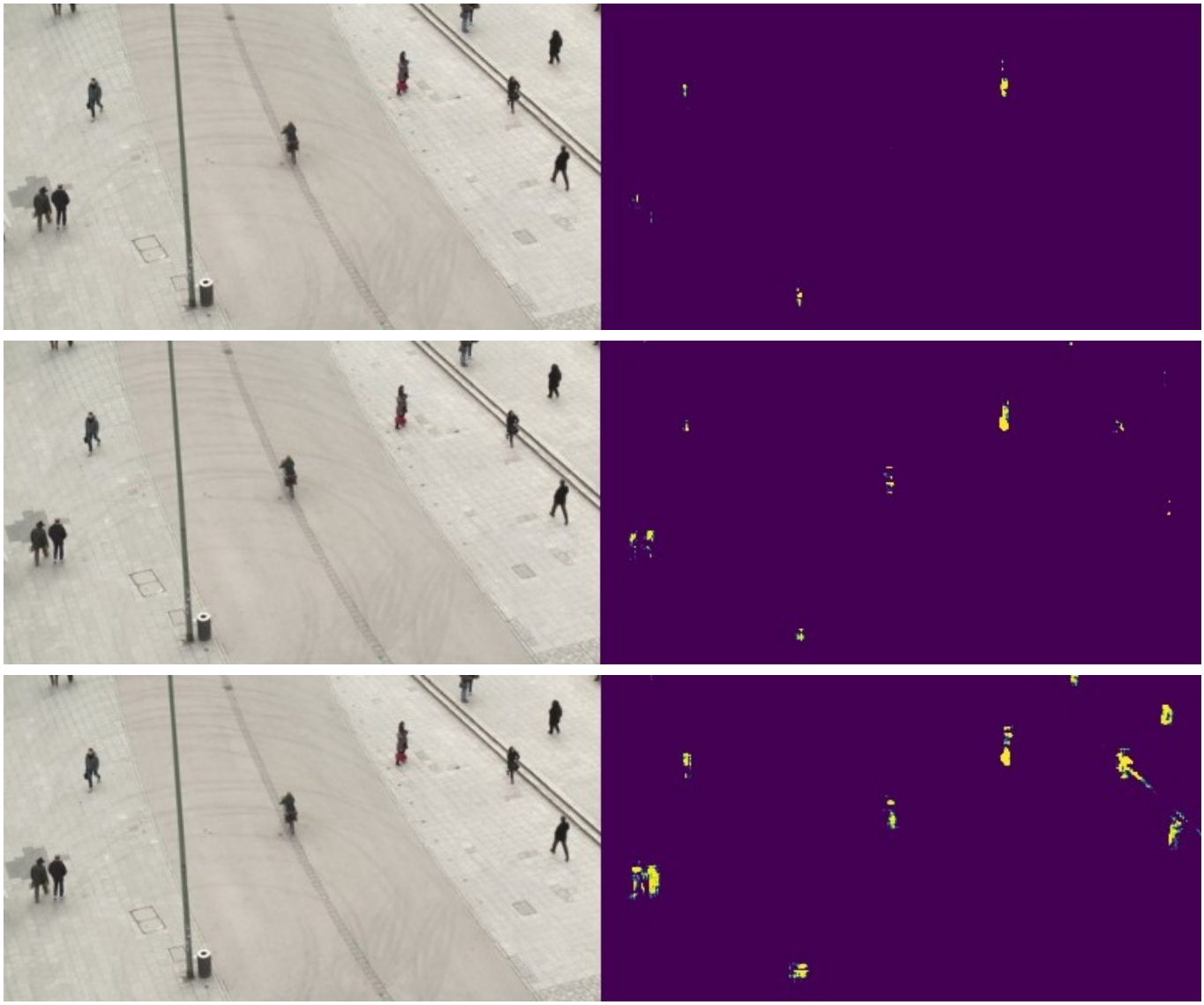

Fig. 4. Heatmaps on real-image containing humans using the synthetic-to-real real-time model trained with 0, 500, and 1000 real images respectively.

the same remarks are drawn in Fig. 3, where the training curves of the synthetic-to-real real-time model trained with 0, 100, 200, 500, and 1000 real images throughout the training iterations, are illustrated. Furthermore, another important remark is that the more real images we include in the training procedure, the more stable the performance is. That is, we notice that when training only with synthetic data, apart from the poorer performance in terms of classification accuracy, the model also exhibits unstable performance. This is also occurs in the case of training with only 100 real images, while when training with 200, and especially with 500 and 1000 real images a more stable performance is managed.

Finally, in the second set of experiments, we use the proposed trained model on the synthetic human dataset to generate heatmaps on unseen high-resolution images that contain real humans. That is, as previously mentioned, unseen images of size 1920×1080 are fed to the network, and for every window 64×64 we compute the output of the network at the output layer. First, in Fig. 4 we provide the heatmaps on an unseen high-resolution image, with the model trained with none, 500, and 1000 real images respectively. As it is shown, the beneficial effect of including a few real images in the training , demonstrated in the first set of experiments, is also reflected in the qualitative results. That is, as it is shown

TABLE I
CLASSIFICATION ACCURACY USING THE SYNTHETIC-TO-REAL REAL-TIME MODEL TRAINED WITH 0, 100, 200, 500, AND 1000 REAL IMAGES.

N. of real images	Classification accuracy
0	0.7725
100	0.9546
200	0.9848
500	0.9871
1000	0.9958

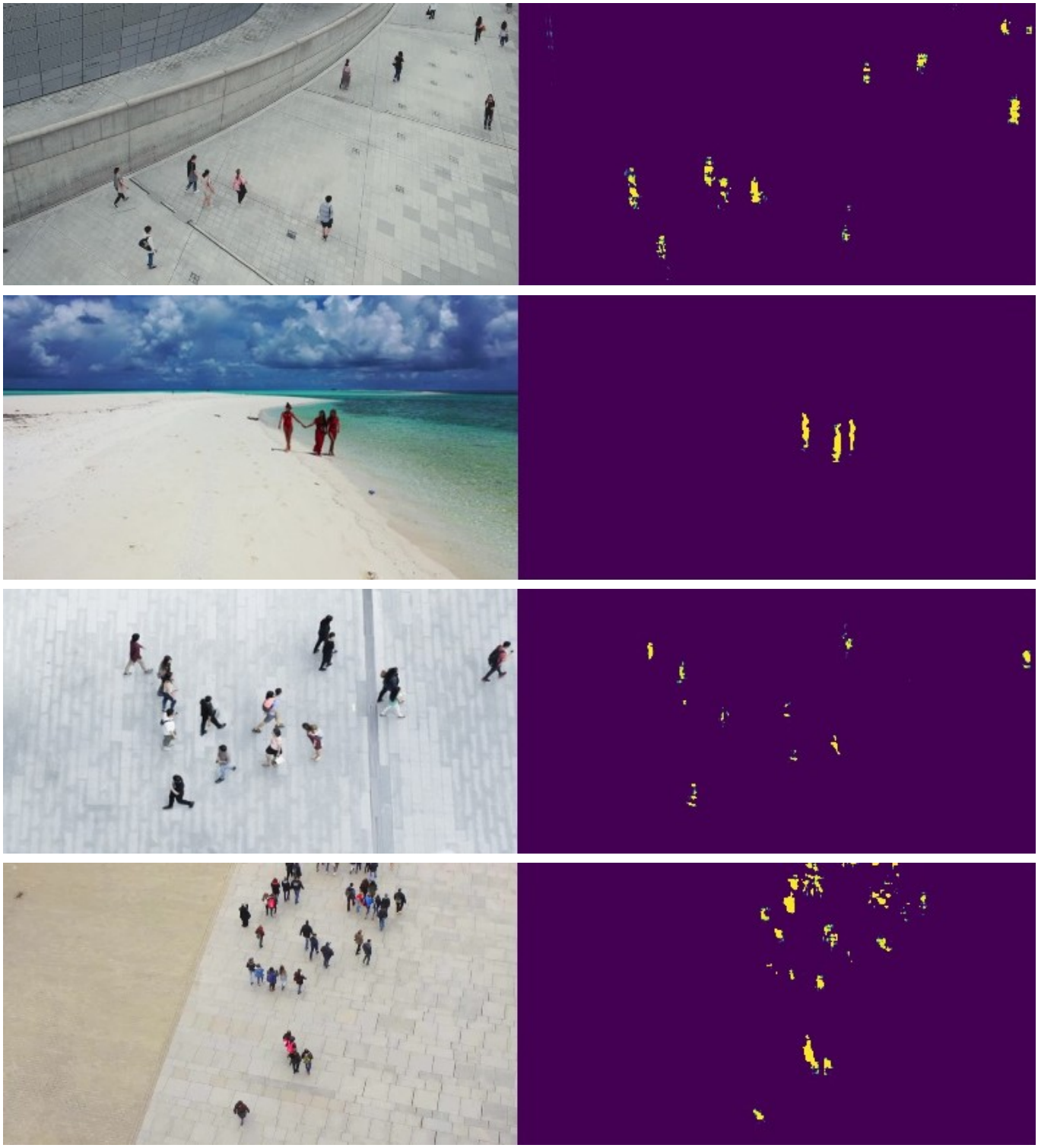

Fig. 5. Heatmaps on real-images containing humans using the synthetic-to-real real-time model trained with 1000 real images.

in the produced heatmaps, while using only synthetic data, only a few humans can be detected, when using 1000 real images in the training, all of them can be detected. Finally, in Fig. 5, we provide some heatmaps using the synthetic-to-real real-time model, trained with only 1000 real images. As it is demonstrated, the model achieves remarkable performance on detecting real humans.

IV. Conclusions

In this paper, we dealt with synthetic data considering robotics applications. More specifically, we first built a synthetic human dataset, and then we trained a lightweight model, capable of running in real-time for high-resolution input, for discriminating between humans and non-humans. The objective of this work is to assess the generalization of the model trained on synthetic data, to real data, and also to investigate the effect of using (few) real images in the training phase. As it is demonstrated in the experimental evaluation, the use of only few real images can beneficially affect of the performance of the synthetic-to-real real-time model.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871449 (OpenDR). This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[2] M. Tzelepi and A. Tefas, "Deep convolutional learning for content based image retrieval," *Neurocomputing*, vol. 275, pp. 2467 – 2478, 2018.

[3] C. Nasioutzikis, M. Tzelepi, and A. Tefas, "Deep hashing regularization towards hamming space retrieval," in *11th Hellenic Conference on Artificial Intelligence*, 2020, pp. 74–77.

[4] M. Tzelepi and A. Tefas, "Quadratic mutual information regularization in real-time deep cnn models," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020, pp. 1–6.

[5] H. A. Pierson and M. S. Gashler, "Deep learning in robotics: a review of recent research," *Advanced Robotics*, vol. 31, no. 16, pp. 821–835, 2017.

[6] A. I. Károly, P. Galambos, J. Kuti, and I. J. Rudas, "Deep learning in robotics: Survey on model structures and training strategies," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 266–279, 2020.

[7] N. Passalis, S. Pedrazzi, R. Babuska, W. Burgard, D. Dias, F. Ferro, M. Gabbouj, O. Green, A. Iosifidis, E. Kayacan *et al.*, "Opendr: An open toolkit for enabling high performance, low footprint deep learning for robotics," *arXiv preprint arXiv:2203.00403*, 2022.

[8] M. Tzelepi and A. Tefas, "Semantic scene segmentation for robotics applications," in *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE, 2021, pp. 1–4.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[10] S. I. Nikolenko, "Synthetic data for deep learning," *arXiv preprint arXiv:1909.11512*, 2019.

[11] D. Ward, P. Moghadam, and N. Hudson, "Deep leaf segmentation using synthetic data," *arXiv preprint arXiv:1807.10931*, 2018.

[12] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 211–220.

[13] Y. Lin, C. Tang, F.-J. Chu, and P. A. Vela, "Using synthetic data and deep networks to recognize primitive shapes for object grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 494–10 501.

[14] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 737–744.

[15] M. Yan, Q. Sun, I. Frosio, S. Tyree, and J. Kautz, "How to close sim-real gap? transfer with segmentation!" *arXiv preprint arXiv:2005.07695*, 2020.

[16] W. Chen, Z. Yu, Z. Wang, and A. Anandkumar, "Automated synthetic-to-real generalization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1746–1756.

[17] W. Zhao, J. P. Queralta, L. Qingqing, and T. Westerlund, "Towards closing the sim-to-real gap in collaborative multi-robot deep reinforcement learning," in *2020 5th International Conference on Robotics and Automation Engineering (ICRAE)*. IEEE, 2020, pp. 7–12.

[18] M. Tzelepi and A. Tefas, "Improving the performance of lightweight cnns for binary classification using quadratic mutual information regularization," *Pattern Recognition*, vol. 106, p. 107407, 2020.

[19] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[20] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[21] 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 *PProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012.

[23] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.