Style-transfer GANs for bridging the domain gap in synthetic pose estimator training

Pavel Rojtberg  
Fraunhofer IGD, Darmstadt  
TU Darmstadt  
pavel.rojtberg@igd.fraunhofer.de

Thomas Pöllabauer  
Fraunhofer IGD, Darmstadt  
TU Darmstadt  
thomas.poellabauer@igd.fraunhofer.de

Arjan Kuijper  
Fraunhofer IGD, Darmstadt  
TU Darmstadt  
arjan.kuijper@igd.fraunhofer.de

Abstract—Given the dependency of current CNN architectures on a large training set, the possibility of using synthetic data is alluring as it allows generating a virtually infinite amount of labeled training data. However, producing such data is a non-trivial task as current CNN architectures are sensitive to the domain gap between real and synthetic data.

We propose to adopt general-purpose GAN models for pixel-level image translation, allowing to formulate the domain gap itself as a learning problem. Here, we focus on training the single-stage YOLO6D [20] object pose estimator on synthetic CAD geometry only, where not even approximate surface information is available.

Our evaluation shows a considerable improvement in model performance when compared to a model trained with the same degree of domain randomization, while requiring only very little additional effort.

I. INTRODUCTION

The ability to detect known objects and their 3D position relative to the viewer is crucial for many Augmented Reality applications and robotic tasks. Recent advances with deep convolutional models such as SSD-6D [14], PoseCNN [27] and YOLO6D [20] allow solving this problem using only RGB images in real-time. However, to achieve state-of-the-art performance these models require a large amount of labeled training data. The assembly of such a training-set is an expensive, error-prone and time-consuming process [9], making it cumbersome and often times inapplicable for use with custom applications.

When 3D geometry is available, one can resort to synthetically generate training data by rendering, which allows to create a virtually infinite training set in an automated fashion. However, it was shown that deep CNN models, even when applying cross-validation, tend to over-fit to the specific dataset [22] and show significantly degraded performance when presented with data from a different domain [4]. Particularly, there is a strong domain-gap between real and synthesized images, which typically prevents the use of synthetic images for training.

To overcome this limitation, existing approaches apply domain randomization (DR) to enforce domain invariance by overwhelming the model with variation [23, 21], requiring the network to learn deeper, more abstract, features that are invariant across domains. An alternative direction is to reduce the gap by employing photo-realistic rendering [25] and structurally correct context generation [17]. Notably, Tremblay et al. [24] apply both for the task of object pose estimation. However, designing a randomization method for a specific domain requires a domain expert to define which parts must stay invariant. Conversely, increasing photo-realism requires an artist to carefully model the specific environments in detail. This in turn increases the cost of generating the data thus negating the primary selling point of using synthetic images in the first place.

When some real images are available, transfer learning can be exploited, e.g. by fine-tuning a synthetically trained model with real images [16]. Alternatively, [8] propose to pre-train a network on real data and fine-tune on synthetic data. Furthermore, it is possible to use the real data to enforce domain invariance during training [4]. Recent work [29] extend this approach to guide domain-randomization, thus introducing some benefits of learning-based domain adaptation. However, it is still required to correctly select and design randomization modules.

Recent advances on generative adversarial networks (GANs) [6, 12, 2] have shown great improvements regarding image quality and plausibility, training stability and variation of output. Of particular interest in the task of closing the domain gap are conditional GANs [11]. These are networks that, unlike traditional GANs, take additional inputs to condition generated output. Here, the image-conditional GANs form a general-purpose framework for image-to-image translation problems, like semantic segmentation, colorization and other style transfer tasks. Existing solutions can be split into paired...
models [11, 26] and unpaired models [30]. The former are trained to adapt source to target images, paired in an supervised fashion, while the latter do not require supervision and instead directly learn to transfer the distribution of image features found in two unstructured data-sets.

This work focuses on employing such models to formulate the domain gap between real and synthetic images as a style-transfer problem. At this, we introduce training pipelines for both paired and unpaired image translation and evaluate the results on the task of object pose estimation. This is a particularly challenging scenario which requires a high fidelity of the object contours, which was, to best of our knowledge, not previously addressed with GAN based image translation. In the context of paired image translation, we propose the use of the intermediate edge domain to do away with the need of real images for supervised training. Here, we evaluate different mapping strategies for transferring CAD geometry with unknown surface properties into the edge domain.

Most closely related to our work is [1] which employ GANs for data augmentation. In contrast, specifically address the domain gap and use synthetic data generation instead of data augmentation. [18] introduce the “pencil filter” as a domain with reduced expressiveness to tackle the domain gap and train a pose estimation network. However, the pose estimation network takes a strong performance hit as the “pencil domain” does not retain enough relevant features. We are avoiding this hit by learning a reversed mapping from the reduced domain to real images to reconstruct appropriate features. In the medical domain, [15] employ a GAN architecture for domain adaptation. However, they only consider the use of an unsupervised GAN model for reverse domain adaptation by making real images more synthetic, while we consider both paired and unpaired architectures and also consider the forward domain adaptation in the unsupervised case.

Based on the above, our key contributions are:
1) formulating the domain gap as a learning problem using off-the-shelf image-conditional GANs,
2) introduction of the intermediate edge domain for training paired translation networks purely from synthetic data and
3) evaluation of paired and unpaired models regarding pose estimation performance.

This paper is structured as follows: in Section II the general approach is introduced and the choice of suitable GAN models is discussed. In Section III the method is evaluated in terms of pose estimation performance of the YOLO6D [20] model on the LINEMOD [7] data-set.

We conclude with Section IV giving a summary of our results and discussing the limitations and future work.

II. APPROACH

The core idea of our approach is to formulate the domain gap as a learning problem that is addressed with generative CNN models. Here, we use the generative adversarial framework to train a conditional generator that is able to augment images such that the pose estimation network becomes invariant to the source domain. For this, the statistical distribution of image features found in both the real world and the synthetic domain must be matched, allowing the alignment of one domain to the other.

In this section we first discuss applicable GAN models and show qualitative results on the LINEMOD data-set to motivate the choice of specific image-conditional GAN models. Then, we present our pipeline for fully synthetic training based on supervised image translation, leveraging PIX2PIXHD [26]. Next, we turn to unsupervised image translation and introduce an alternative pipeline, that replaces the GAN model with CYCLEGAN [30], which simplifies data acquisition by lifting the requirement of pairing images from both domains.

A. Baseline methods

We define two baseline methods for synthetic training. Both follow the data augmentation scheme of [20] by using random backgrounds from a set of real images, apply random image scaling and randomly adjust exposure and saturation adjustment. However, instead of using crops from real image, we render the object on top of the background (See Figure 2) with the following methods:

a) Realistic texturing, by applying the true texture extracted from the data-set [19] and
b) randomly texturing the object during rendering.

The first option depends on having some real data available, but allows generating an arbitrary amount of realistic data by rendering. The second scheme applies blind DR by randomizing both the object and the background appearance. Note that this scheme neither requires plausible object placement nor plausible object appearance and thus is far easier to set up, compared to other DR solutions.

B. Suitable GAN models

The main requirement on the generator is that the resulting images have a high correlation with a given pose, such that the pose estimation model can be trained in a supervised fashion. In theory any GAN model can be used for this if images can be mapped into the latent space of the generator. Such a
mapping allows to condition the generator to create images, that resemble the input inside the latent space.

If the latent space is constructed in way that allows for interpolation by e.g. employing Kullback-Leibler loss [3], this approach would also allow synthesizing novel samples not seen during training. In the case of pose estimation, this is required to generate images for views not seen by the adaptation network during training. However, the resulting image must retain enough fidelity for the pose estimation network to predict the correct pose. Not all GAN architectures are practical for this use case. For a preliminary experiment, we use the STYLEGAN [13] model, which offers state-of-the-art generation performance and allows interpolation in the latent space of the generator.

We train STYLEGAN at a reduced resolution of $256 \times 256$ px for three days using approximately 110,000 renderings composed on top of real backgrounds. The model is initialized using the weights for the LSUN Cat data-set [28] of the original publication. This allows operating at the reduced training period, compared to training the model from scratch which required 13 days according to the original publication.

We then map a real image into the latent space and use it as the mean for the random vector to sample new images, which would ideally retain most of the original image content; particularly the pose of the target object. Figure 3 shows an exemplary result; while the images seem plausible (considering the reduced training time), it is obvious that the model is not able to sufficiently capture the locality of the object. This results in a significantly different pose in the sampled image, compared to the input image. This precludes this approach in being used for pose estimation. Therefore, we focus on conditional GAN models in the following.

**C. Paired intermediate domain translation**

In this section we introduce a training pipeline based on the Pix2PixHD [26] paired image translation model. This is a supervised approach, which requires aligned image pairs to learn the domain translation. Pairing synthetic images with real images requires an according training set of real images, which defies the goal of synthetic training. Therefore, we introduce a deterministic transform into an intermediate domain with reduce expressiveness, making real-world and rendered images less distinguishable. Here, we use the Laplace filter, which approximates the second order image derivative that is continuous and directly translates to edge strength without requiring a thinning step. The Pix2PixHD model is then employed to reconstruct real images from their Laplace filtered variants.

The pipeline now consists of two trainable models: the Pix2PixHD model for domain adaptation followed by the pose-estimation task network (see Figure 4). As the performance of the second model depends on the first model, the pipeline has to be trained in two stages. First, the domain adaptation network is trained until convergences on rendered images with random backgrounds and their Laplace filtered variants. Here, the network must simultaneously reconstruct the rendering as well as the real background which forces the network towards realistic reconstructions. Next, the pose estimation model is trained on the reconstructed images. Assuming the adaptation network was able to generalize from the training data, we now can create a virtually infinite amount of realistic views.

The remaining question is how to model the object surface before converting it to the Laplace domain, given that we do not want to impose any restrictions on possible object appearances. Here, one needs to balance the learning problem between the domain adaptation and the pose estimation network — e.g. enforcing discriminative features makes the problem for the adaptation network more challenging, yet it reduces the difficulty for the pose estimation network. Specifically, we opted for the following methods covering different work distributions of the involved networks:

1) Use the real world texture (see Figure 4). This should result in the best model performance as it is the easiest task. However, to obtain the texture real images are needed, which dissents with our goal of fully synthetic training. This option was therefore mainly included to assess the performance loss of mapping into the Laplace domain as well as the loss induced by the following options. It corresponds to baseline variant a) with domain
**Figure 5:** Examples of the reconstruction tasks for PiX2PiXHD with rendering methods. *Top row:* source image in the Laplace domain *Bottom row:* target images to be reconstructed.

2) Use a random texture (see Figure 5a). This prevents the networks from learning any surface related information of the object to be detected. While this should not affect the domain adaptation network which is already faced with the task of reconstructing arbitrary background images, it makes the task of pose estimation significantly more challenging. The important shading and contour cues can be arbitrarily degraded by the used texture. This corresponds to baseline variant b) with domain adaptation.  

3) Use a uniform color for the object (see Figure 5b). Instead of using a random texture, we assume the object to be of one uniform, yet arbitrary color. Given the Laplace intermediate domain, the adaptation network cannot learn a correct object colorization. Therefore, we set the target color to gray, which is the most likely guess the adaptation network can take, given this task. Keeping the surface properties fixed allows to apply a consistent shading to the object, which in turn can be exploited by the pose estimation network. However, the reconstruction of a plausible surface shading in turn makes the task of the domain adaptation network more challenging, but is a deliberate choice for balancing the work. To prevent the adaptation network to over-fit to a specific lighting position, we place several point lights randomly around the object.  

4) Use a fixed checkerboard pattern for the object surface (see Figure 6). Here, the Laplace images are generated from a uniformly colored object as above, while the reconstruction target is rendered with a fixed checkerboard texture. This makes the pose estimation task easier as the network can rely on stable cues on the object surface. However, this is particularly challenging for the adaptation network as it must encode the object geometry to reconstruct a correct appearance, while being confronted with merely an edge image generated from an uniformly colored object. 

Note that the translation of the object appearance to a different representation in the last two methods, requires the translation network to be executed at inference time as well. 

**D. Direct image domain translation** 

In this section we introduce a CYCLEGAN [30] based training pipeline for unsupervised domain adaptation. As this model does not require matching image pairs, there is no need for an explicit intermediate representation, and the model can directly learn the mapping between the domains of synthetic and real images. For the real domain we use the same generation scheme like in baseline b) of rendering randomly textured objects on random backgrounds. However, for the synthetic domain we cannot use real backgrounds to generate samples, as the different statistics would give away the synthetic object and bias the GAN towards object segmentation. Instead, we collect a separate data-set of synthetic background images from 3D-game footage on Youtube. Here, we select 50,000 random crops of randomly selected frames from a total of about 5 hours of video footage. 

Furthermore, the specific architecture is capable of translating images in both directions. This allows reversing the pipeline; instead of adapting synthetic images to the real domain at training time, it is also possible to adapt real images to the synthetic domain at run-time. While this eases training, it comes with the cost of having to execute the translation network for inference. 

As a limitation, the original CYCLEGAN model is tuned to produce images at a resolution of $256 \times 256$ px, which is only a fraction of the YOLO6D receptive field of $416 \times 416$ px. Scaling the GAN up would require fine-tuning its hyperparameters like the size of the hidden layers, which is out of the scope of this work. Therefore, we perform our experiments with the limited resolution, which allows to judge the feasibility of the method. However, one should keep in mind that pose precision can be improved by scaling the network output to match YOLO6D. 

**III. Evaluation** 

In this section we quantitatively and qualitatively evaluate, whether our pipeline allows the training of the demanding pose estimation task network from synthetic, randomized and non-photorealistic renderings only. For comparability with related work, we use the LINEMOD data set for evaluation. Here, we focus on the "Driller" object as it exhibits a characteristic, non-uniform surface and is the most challenging object for the pose estimation model. 

In the following, we first present the results of the baseline approaches as introduced in section II. We then turn to the paired image translation via an intermediate, reduced domain and finally present the results for direct image domain translation. 

**A. Implementation details** 

For training the pipeline, we follow the procedure outlined by [20] in initially dropping the confidence loss, when training YOLO6D on a domain different from real images. This proved
Table I: Pose estimation error of the baseline methods a) and b) respectively.

| Texturing method / Domain translation | Mean error |
|--------------------------------------|------------|
|                                      | angle / translation / re-projection |
| real / none                          | 12.45° / 9 cm / 15 px |
| random / none                        | 63.81° / 45 cm / 111 px |

Table II: Pose estimation error of paired domain translation as introduced in Section II-C. The rows correspond to rendering methods 1-4) respectively.

| Texturing method / Domain translation | Mean error |
|--------------------------------------|------------|
|                                      | angle / translation / re-projection |
| real / laplace                        | 8.57° / 8 cm / 11 px |
| random / laplace                      | 54.45° / 36 cm / 88 px |
| uniform / laplace                     | 66.41° / 49 cm / 81 px |
| pattern / laplace                     | 82° / 68 cm / 133 px |

Table III: Pose estimation error of direct domain translation as introduced in Section II-D, as well as translation reversal (second row).

| Texturing method / Domain translation | Mean error |
|--------------------------------------|------------|
|                                      | angle / translation / re-projection |
| random / direct                      | 55.26° / 40 cm / 83 px |
| random / reverse                      | 19.77° / 10 cm / 25 px |

Figure 6: Checkerboard reconstruction results as required by surface rendering method 4). Top row: target image. Bottom row: reconstructions by Ptx2PtxHD.

B. Quantitative results

We measure the domain translation performance of the presented methods in terms of pose estimation error of the task network [20]. Here, we employ two different metrics; the 3D translation and 3D rotation as well as the 2D corner re-projection error. The latter measures the error in screen-space and therefore is well suited for augmented-reality, while the former is more meaningful for robotic applications.

Comparing the baseline methods as introduced in Section II-A (see Table I), we see that the performance of baseline method b) is significantly reduced, although we ensured model convergence in both cases. This shows, that our task network for pose-estimation, YOLO6D, is not sufficiently conditioned to overcome the domain gap on its own, even when presented with a virtually unlimited amount of images from the synthetic domain.

Looking at the domain translation results using the paired model as introduced in Section II-C in Table II, we see an improvement in pose estimation by about 15% in both cases. This indicates that the edge based intermediate representation leads to a more robust representation with YOLO6D. Likely, because we are able to reduce the texture bias [5] of the model. Notably, rendering the object using a uniform color as in 3) does not improve results compared to using a random texture. Probably the randomized lighting causes enough variation such that the pose estimator cannot benefit from the consistent shading. Training the pose estimation network was not possible when applying rendering method 4).

While the results improve over baseline, the margin is only moderate. Likely, the reason for this is that the Laplace filtering does not sufficiently reduce the expressiveness of the image and the domain gap is still present in the intermediate representation.

Turning to the direct domain translation using CycleGAN as introduced in Section II-D in Table II, the results of synthetic to real translation are consistent with paired translation, even though the network produces images only at quarter the resolution. This leads us to believe that unsupervised models are generally sufficient for domain adaptation.

When reversing the translation from real to synthetic, we see results in reach of our baseline method a). This is remarkable as the baseline uses realistic texturing of the object that is captured from real images. The domain translation method on the other hand only relies on the CAD geometry with both the object background and object surface being randomized.

C. Qualitative results

As shown in Figure 6, Ptx2PtxHD can not predict a consistent checkerboard texture. While the results are plausible, they do not exhibit enough detail to help the pose estimation model. Still, this is a remarkable result as the translation network is confronted with a significantly more demanding challenge; to correctly apply the texture pattern it not only has to infer the object pose, but also the object geometry.

Figure 1 shows some predictions of the pipeline trained with CycleGAN for domain adaptation. While, the model is able
To reliably detect the object, the accuracy of the object pose is lacking. This is likely due to the limited resolution of the adaptation network and can be further improved by scaling it up to the pose estimation input size.

Figure 7 shows the effect on image quality after applying CYCLEGAN on the objects “ape” and “benchvise”. The model is able to improve color reproduction as well as applying details like specular highlights.

IV. CONCLUSION

We have shown that employing a paired translation GAN for domain adaptation during training generally improves model robustness and hence the performance of the target network. Turning to unpaired translation GANs, we have shown that training solely on CAD geometry with neither knowing the surface properties nor the environment is possible and reaches the performance of training on realistic data, when using reverse translation. These results indicate that image-conditional GANs are indeed an effective measure to close the domain gap between real and synthetic images.

Furthermore, the introduced training is much simpler compared to existing solutions relying on domain randomization. The latter require a faithful setup of randomization modules — even when employing guided domain randomization. On the other hand, the presented style transfer pipelines only require collecting unstructured images from the target domain, which are fed into the adaptation network in an unsupervised fashion. This allows focusing on training the task network. At this, we have shown that the method is precise enough to train a pose-estimation model with satisfactory precision.

To further improve the performance of our approach, the obvious measure is to take out the resolution loss by scaling up the CYCLEGAN output to match the YOLO6D input. Furthermore, the method currently relies on blind randomization during rendering to consider the multitude of possible surface materials. Instead, it seems beneficial to employ a generator that is aware of the multi-modal data and therefore is capable to produce different surface materials based on a noise vector.

To this end, one could replace the image-conditional CYCLEGAN by a more recent variant like MUNIT [10]. Alternatively, further regularization on the unconditional STYLEGAN [13], could enforce its samples closely resemble the input image. In this case one could reconsider the architecture as it is also capable to generate multiple styles, by internally disentangling content and style.

REFERENCES

[1] A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340, 2017.
[2] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In International Conference on Learning Representations, 2019.
[3] P. K. Diederik, M. Welling, et al. Auto-encoding variational bayes. In Proceedings of the International Conference on Learning Representations (ICLR), volume 1, 2014.
[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1):2096–2030, 2016.
[5] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In International Conference on Learning Representations, 2019.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.
[7] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In Asian conference on computer vision, pages 548–562. Springer, 2012.
[8] S. Hinterstoisser, V. Lepetit, P. Wohlhart, and K. Konolige. On pre-trained image features and synthetic images for deep learning. In Proceedings of the European Conference on Computer Vision (ECCV), pages 0–0, 2018.
[9] S. Hinterstoisser, O. Pauly, H. Heibel, M. Martina, and M. Bokeloh. An annotation saved is an annotation earned: Using fully synthetic training for object detection. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 0–0, 2019.
[10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 172–189, 2018.
[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[13] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[14] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the International Conference on Computer Vision (ICCV 2017), Venice, Italy*, pages 22–29, 2017.

[15] F. Mahmood, R. Chen, and N. J. Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE transactions on medical imaging*, 37(12):2572–2581, 2018.

[16] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[17] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapiro, and S. Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7249–7255. IEEE, 2019.

[18] J. Rambach, C. Deng, A. Pagani, and D. Stricker. Learning 6dof object poses from synthetic single channel images. In 2018 *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 164–169. IEEE, 2018.

[19] P. Rojiberg and A. Kuiper. Real-time texturing for 6d object instance detection from rgb images. In 2019 *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 295–300. IEEE, 2019.

[20] B. Tekin, S. N. Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018.

[21] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[22] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[23] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 969–977, 2018.

[24] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning*, pages 306–316, 2018.

[25] A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger. Procedural modeling and physically based rendering for synthetic data generation in automotive applications. *arXiv preprint arXiv:1710.06270*, 2017.

[26] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[27] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. doi: 10.15607/RSS.2018.XIV.019.

[28] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[29] S. Zakharov, W. Kehl, and S. Ilic. Deceptionnet: Network-driven domain randomization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 532–541, 2019.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.