

# Closing the Reality Gap with Unsupervised Sim-to-Real Image Translation

Jan Blumenkamp<sup>1</sup>, Andreas Baude<sup>2</sup> and Tim Laue<sup>2</sup>

<sup>1</sup> University of Cambridge, Department of Computer Science and Technology,  
Cambridge, United Kingdom

`jb2270@cam.ac.uk`

<sup>2</sup> Universität Bremen, Fachbereich 3 – Mathematik und Informatik,  
Postfach 330 440, 28334 Bremen, Germany  
`{an.ba,tlaue}@uni-bremen.de`

**Abstract.** Deep learning approaches have become the standard solution to many problems in computer vision and robotics, but obtaining sufficient training data in high enough quality is challenging, as human labor is error prone, time consuming, and expensive. Solutions based on simulation have become more popular in recent years, but the gap between simulation and reality is still a major issue. In this paper, we introduce a novel method for augmenting synthetic image data through unsupervised image-to-image translation by applying the style of real world images to simulated images with open source frameworks. The generated dataset is combined with conventional augmentation methods and is then applied to a neural network model running in real-time on autonomous soccer robots. Our evaluation shows a significant improvement compared to models trained on images generated entirely in simulation.

## 1 Introduction

In recent years, deep learning approaches became the standard solution to many problems in computer vision, such as classification [5], object detection [16], or semantic segmentation [18]. Efforts were made to reduce the computational complexity in order to deploy them to mobile devices [9]. These approaches usually require a vast amount of training data which can either be generated through human labor or be generated synthetically. Generating training data through human labor can result in datasets of high quality, but it is a cumbersome and expensive task. The CamVid dataset [4] contains detailed semantic labels and uses preprocessing to assist human labelers, but annotating a single frame takes 20 to 25 minutes. Multiple volunteers were tasked to perform the labeling, but only about 15 % of the volunteers delivered acceptable results. Similar problems exist in other datasets such as the PASCAL VOC challenge [6] or in the COCO dataset [14].

Recently, a trend can be seen to approaches that rely on simulated and synthetic data. The SYNTHIA dataset, for instance, consists of 213400 images and pixel-accurate semantic annotations as well as depth maps generated with the

Unity framework [19]. Computer games can also be used to generate images that can then be labeled manually [17]. Unfortunately, data generated in a simulated environment often does not directly transfer to reality. This issue is referred to as the *reality gap* [12].

Hess *et al.* [8] introduced an environment to create annotated training data in a RoboCup Standard Platform League (SPL) setting and demonstrated the feasibility of performing a semantic segmentation on that data. A major challenge in the RoboCup SPL is the perception of the field in a diverse set of lighting conditions. Low quality cameras result in images with low contrasts and limited processing power usually requires using fast conventional computer vision approaches. Frameworks such as TensorFlow Lite [1] or CompiledNN [25] made utilizing neural networks in mobile and low-end devices more feasible.

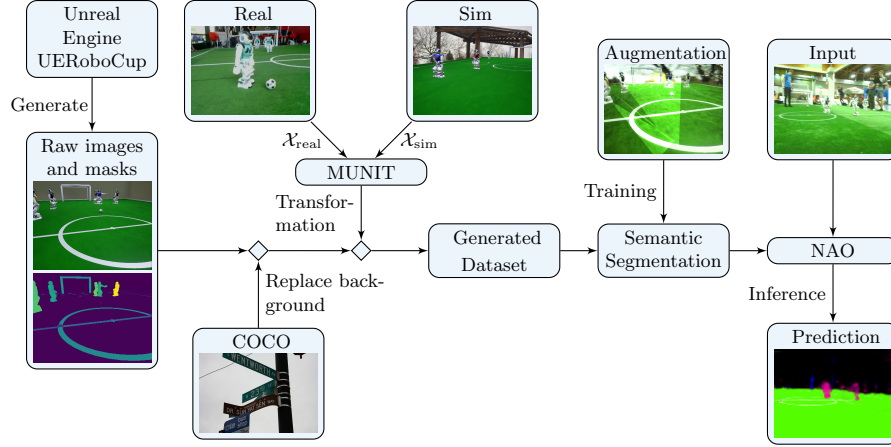
In our work, we synthetically generate images with the tools provided by [8] and transform them with unsupervised image-to-image translation [10] and domain randomization [26] so that they can be used as training data for any kind of deep learning task. We use this dataset to train a semantic segmentation that is able to run in real-time on a NAO v6 robot. Our approach can generally be applied to any other domain where computing power is sparse and flexibility and reliability plays an important role. All required software dependencies are open source. Our evaluation shows that models trained with our method perform noticeably better than models trained with data directly generated from simulation as well as with generated data that is expanded with conventional augmentation techniques. In summary, our main contributions are:

- We developed a method using publicly available state of the art image-to-image transformation frameworks and demonstrate that it can be used to generate high quality datasets that allow training highly performing models.
- We introduce a multi-class semantic segmentation model architecture that is capable of running in real-time on a NAO v6 robot.

The remainder of this paper is organized as follows: After summarizing the related work in Sec. 2, we describe our data generation approach in Sec. 3. In Sec. 4, we present our semantic segmentation model architecture. Lastly, multiple models generated with different data augmentations are evaluated and discussed in Sec. 5 and Sec. 6 respectively. An overview of the proposed method is depicted in Fig. 1.

## 2 Related Work

There are two remarkable works in the area of *closing the reality gap with image-to-image translation*: Bousmalis *et al.* [3] use domain adaptation and domain-adversarial neural networks to utilize synthetic training data in an end-to-end learning approach in order to learn robotic grasping. Bewley *et al.* [2] use image-to-image translation to transfer a vision-based driving policy from simulation to reality. Instead of explicit representations such as a semantic segmentation, their end-to-end approach uses more implicit representations. In contrast to our



**Fig. 1.** The general workflow of our proposed method: Images and corresponding masks are generated in simulation. The background of the generated images is replaced with a random image from the COCO dataset. MUNIT image-to-image translation is applied to these enhanced images, which are now the intermediate dataset.

proposed method, both approaches have in common that they utilize the image-to-image translation in an end-to-end approach. Flexibility in the postprocessing plays an important role in RoboCup settings, which is not given with the current state of the art of end-to-end approaches. Using intermediate representations such as pixel-accurate labels allows a higher flexibility and versatility. A dataset using such low level representations can be used both for semantic segmentation and for high level classification and object detection.

*Deep learning approaches for robot vision in RoboCup* were first applied in the humanoid league, where Schnekenburger *et al.* [20] use a segmentation to detect different field features such as line intersections and objects. All classes except for the detection of other robots performed satisfactory. Van Dijk *et al.* [27] propose a novel model architecture without any residual connections for a semantic segmentation, which was able to be executed in close to real-time on a typical smartphone CPU, but lacks the ability to properly detect complex features or multiple classes at once. In contrast to the humanoid league, the SPL, which uses the SoftBank NAO robots, is more constrained in terms of hardware. Nevertheless, Hess *et al.* [8] trained a simple classifier model to demonstrate the feasibility of using synthetic images in such a scenario. Szemenyei *et al.* [24] propose a novel, small semantic segmentation model that uses images generated with *UERoboCup* for pre-training the model and eventually tune it with real images. However, this approach does not focus on the more extreme conditions that games in environments with natural lighting require. Furthermore, Poppinga *et al.* [15] proposed a robot detection framework for which data obtained in a simulation was used to learn additional features that are hard to label manually, such as robot distances.

### 3 Data Generation

In this section, after briefly describing the background and the tools generating the simulated data, we give insights into the Multimodal Unsupervised Image-to-Image Translation (MUNIT) framework [10], which we use for the sim-to-real image translation, and describe the online data augmentation methods we used.

#### 3.1 RoboCup and UERoboCup

While RoboCup provides a benchmark to quantitatively compare the progress over time, the community is small and lacks labeled training data for deep learning approaches. Efforts were made to create a community-driven database for labeled real images [7], but the data required for specific tasks and pixel-accurate labels are rarely available.

UERoboCup is an application based on *Unreal Engine* that allows generating game situations with multiple robots from the view of a specific robot [8] and is capable of generating pixel-accurate semantic annotation. To provide a more accurate representation of the environment, we added additional labels relevant to the SPL context, such as the penalty mark and the goal bar as well as an adapted appearance of the robots to that of the latest NAO robot generation. In addition, we increased the variation in the generated images to reflect reality more properly. This is achieved by a variable camera pitch instead of a fixed one and the definition of a skeleton for the previously static robot mesh, which allows dynamic robot poses. This procedure is referred to as domain randomization [26]. Furthermore, the robot skeleton can be used for a more detailed segmentation of individual robot limbs, allowing, for instance, a pose detection. We also export meta data such as the extrinsic camera parameters and the poses of robots in the standardized JSON format. Such information is difficult to annotate manually, therefore learning further characteristics, such as the distance to another robot, as shown in [15], mainly relies on synthetic data.

We generated a set of 10000 images and labels with UERoboCup. A generated image with the corresponding converted segmentation mask can be seen in the overview in Fig. 1.

#### 3.2 Image Post Processing

UERoboCup only creates images of a plain RoboCup scene taking place in a white room. However, during an actual game, the background is cluttered with a wide range of different objects, such as people walking around. In order to make potential deep learning applications understand the concept of unwanted background clutter, we replace the background with structured images, similar to [26]. This can be considered another variation of domain randomization. We use images from the COCO test set [14]. Even though these images do not exactly represent how a scene would look like at a RoboCup event, we found that they are well-suited to help deep learning applications differentiate between relevant foreground and irrelevant background clutter.

### 3.3 MUNIT

MUNIT assumes images from two different domains  $x_1 \in \mathcal{X}_1$  and  $x_2 \in \mathcal{X}_2$  and, given samples drawn from the two marginal distributions  $p(x_1)$  and  $p(x_2)$ , without access to the joint distribution  $p(x_1, x_2)$ , estimate the conditionals  $p(x_1|x_2)$  and  $p(x_2|x_1)$  with the image-to-image translation models  $p(x_{1 \rightarrow 2}|x_2)$ , where  $x_{1 \rightarrow 2}$  is a sample resulting from translating  $x_1$  to  $\mathcal{X}_2$  [10]. Due to the unsupervised nature of MUNIT, no explicit labeling has to be performed on any of the sample images. The learned mapping is multimodal, thus multiple different images with different styles from domain  $x_1 \in \mathcal{X}_1$  can be applied to the same image  $x_2 \in \mathcal{X}_2$  and each time a different image with the style from domain  $x_1 \in \mathcal{X}_1$  is computed.

### 3.4 Style Transfer

MUNIT requires a test set and a training set of images for the two classes  $\mathcal{X}_1$  and  $\mathcal{X}_2$ . For this application, the two classes are real images recorded by a robot’s camera  $\mathcal{X}_{\text{real}}$  and simulated images created with UERoboCup  $\mathcal{X}_{\text{sim}}$ . For the real domain, we select images from previous RoboCup events and from ImageTagger [7], accumulating to an overall of 885 images in the training set and 155 images in the test set. We found that a large variance in the training images is essential for MUNIT to generate useful results. Since the images generated by UERoboCup are random, any subset with about the same size can be used. Note that in this subset, we already replaced the background with a random image from the COCO dataset. In the simulated domain, we used 1000 images for training and 200 for testing. We train on an NVidia Titan V with MUNIT default settings for 70000 epochs. Due to memory limitations, we slightly decreased the amount of the generator and discriminator filters. We noticed a convergence after 50000 epochs, with no further significant improvements from that point. To generate the processed images, we took three random style images from the set of real images and applied them to each image generated by UERoboCup, resulting in 30000 different images. An example can be seen in Fig. 2.

### 3.5 Data Augmentation

Data augmentation is a type of data regularization and helps to avoid overfitting. Additionally, data augmentation can be used to enhance the size and quality of datasets with warping or oversampling methods [21].

We used the following online data augmentations during training: Vertical image flipping, Gaussian noise, multiply, add (RGB and HSV), simplex noise, motion blur, contrast normalization, and simulated sun patches.

It is essential that the model learns to handle extreme environmental situations with patches of light and shadow. Since this is not captured with sufficient variance in our dataset, we introduce an additional domain randomization method during online data augmentation by simulating patches on the field that



**Fig. 2.** After changing the background in the simulated image to a random image from the COCO dataset (left), we apply the image-to-image translation learned by MUNIT to it (middle). In addition to the offline augmentation obtained with the sim-to-real translation, we perform an online augmentation during training (right). Beside standard augmentations such as blur, we also introduce a simple simulation of sun light patches on the field.

are illuminated by the sun. We implement this augmentation by generating multiple random polygons consisting of three to six points in the frame and multiply these areas with a random factor. A sample augmentation is shown in Fig. 2.

## 4 Semantic Segmentation

Designing models that are capable of being executed in real-time on platforms with limited resources is difficult due to computing constraints. In our application, the performance baseline for the design of the model is the NAO’s camera frame rate of 30 fps. We allocate one CPU core for the processing of the frames of one camera, which means that the model must process each frame at the same rate. As inference framework, we use *CompiledNN* [25].

The model architecture is based on the U-Net architecture [18] and incorporates features from MobileNet [9]. The U-Net architecture consists of an encoder and a decoder with multiple residual connections between the encoder and the decoder. The reduced model from this work uses only two residual connections and downscales the image twice. MobileNet was designed to be deployed to mobile devices and is therefore optimized to run with limited computing power. This is mainly realized by replacing convolutions with separable convolutions [9]. For an additional acceleration, pooling layers are replaced with convolutions with a corresponding stride, as proposed in [23]. We use batch normalization for regularization [11] and LeakyReLU [28] as activation functions. The resulting model has 12909 learnable parameters and is shown in Tab. 1.

Our proposed model is capable of performing multi-class predictions (in our case ball, goal posts, lines, and robots). Due to limitations in *CompiledNN*, no softmax activation is applied to the last layer. This means that the model detects all mentioned classes independently from each other. Note that a binary cross entropy or something similar must be used as the loss function. Furthermore, it can be desirable to have multiple independent classifications outputs. For instance, if the model is not certain if the arm of a robot is a similarly looking goal

**Table 1.** Our proposed segmentation model architecture. The scale refers to the tensor size relative to the input size. N is the amount of how often the layer in the row is repeated and F represents the amount of filters.

Layer	Scale	N	F
Input	1	1	3
$3 \times 3$ -SConv2D, BN, LeakyReLU	1	1	8
$3 \times 3$ -SConv2D, BN, LeakyReLU	1	1	8
$3 \times 3$ -SConv2D, BN, LeakyReLU	1/2	1	8
$3 \times 3$ -SConv2D, BN, LeakyReLU	1/2	2	16
$3 \times 3$ -SConv2D, BN, LeakyReLU	1/4	1	16
$3 \times 3$ -SConv2D, BN, LeakyReLU	1/4	6	24
Up2D	1/2	1	24
Concat	1/2	1	40
$3 \times 3$ -SConv2D, BN, LeakyReLU	1/2	3	16
Up2D	1	1	16
Concat	1	1	24
$3 \times 3$ -SConv2D, BN, LeakyReLU	1	3	8
$3 \times 3$ -SConv2D, BN, LeakyReLU	1	1	5

post, the likelihood of both outputs would be small with a softmax activation. By keeping the output layers independent from each other, the model can classify the arm of a robot both as part of a robot but also as a goal post. This can be validated in the postprocessing by applying additional domain knowledge and thus discarding false positives. Note that ideally, the model is able to capture this from context, but we found that increasing the model size more leads to an unacceptably low inference time. The maximum amount of features is limited by the complexity of the model. We achieved good results with predicting five and less classes, but it is to be expected that a larger amount of independent output predictions yields worse predictions, if the base model is not adapted.

We perform different experiments to find a working model architecture and look for a compromise between detection accuracy and inference time. The run-time was measured on a single CPU core of the NAO v6 (Intel Atom E3845@1.91 GHz [22]) using CompiledNN [25]. The results for different image resolutions are shown in Tab. 2. As the maximum feasible input resolution for real-time operation is 14 ms, the original input image is subsampled accordingly. The resulting aliasing should not affect the neural model as it should learn to ignore it.

**Table 2.** Inference time on the NAO v6 with CompiledNN

Size [px × px]	40 × 32	80 × 64	108 × 80	120 × 88	160 × 120	320 × 240
Duration [ms]	1.6	6.7	11.2	14.0	27.3	116.0

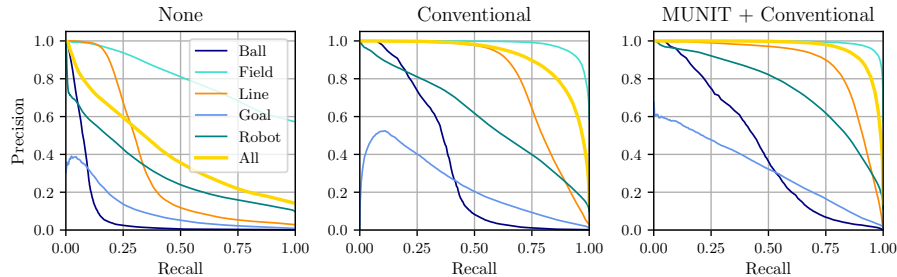
## 5 Results

In order to evaluate the performance of the model and the synthetic dataset, we manually labeled 348 images, which we also used in the MUNIT augmentation resembling varying environment conditions with the five classes field, lines, robots, goal post, and ball. We evaluate the mean Average Precision (mAP) metric. In order to evaluate the effect of the different augmentation techniques, we use five different configurations of our dataset to train the previously described model:

1. no augmentation at all (only the raw images generated by UERoboCup are used, this is the baseline)
2. conventional augmentation without sun (all the augmentations described in Sec. 3.5 except for the sun patch augmentation)
3. conventional augmentation (including the sun patch augmentation)
4. only the image-to-image translation augmentation performed with MUNIT
5. the image-to-image translation combined with all previously described conventional augmentations

We train all five models with a subset of 8000 images and 2000 images as test set with the same augmentations in a batch size of 128 using the Adam optimizer [13] with an initial learning rate of 0.1. We decay the learning rate by 0.5, if the loss on the validation set does not decrease for 10 epochs and we terminate training after 20 epochs without improvement, which was usually reached in less than 100 epochs.

The different precision-recall curves are visualized in Fig. 3. The mAPs for all models and all classes can be seen in Tab. 3. We report the precision-recall curves for the individual classes as well as micro-averaged for all classes. Note that the computed mAP represents the classification performance of each individual pixel over all test images and not the classification performance of the object instances.



**Fig. 3.** Precision-recall curves for a model trained without any augmentation (left), a model trained with conventional online augmentation as described before (middle), and a model trained with sim-to-real image translation and conventional augmentation (right). A significant improvement can be seen due to the augmentation with MUNIT.



**Table 3.** All mAP values for all tested models and all classes

	None	Conventional without sun	Conventional	MUNIT	MUNIT and conventional
Ball	0.0843	0.3927	0.3440	<b>0.4577</b>	0.4277
Field	0.8037	0.9846	0.9835	0.9877	<b>0.9917</b>
Line	0.3404	0.7866	0.7952	0.8745	<b>0.8779</b>
Goal	0.1024	0.1418	0.2367	<b>0.3554</b>	0.3207
Robot	0.3059	0.5361	0.5983	0.6529	<b>0.7478</b>
All	0.4203	0.9140	0.9165	0.9536	<b>0.9647</b>

## 6 Discussion

As expected, the model without any augmentation performs worst with an overall mAP of 0.4203, which shows that reality gap is an issue for plain images generated in UERoboCup. The ball and the goals are rarely detected with a slightly higher mAP for robots and lines. Since a quantitative comparison with related works [20, 27, 24] is not possible due to different metrics that do not capture the problem well, this model is used as baseline of what is possible with purely synthetic data generated in UERoboCup.

Just by performing basic augmentation (all augmentations mentioned above except for the sun patch augmentation), the performance of the model increases drastically to an overall mAP of 0.914. Most classes receive a significant increase in mAP, particularly the robot, line, and ball classes. The goal classes’ mAP only increase slightly. Adding the sun patch augmentation increases the overall mAP again slightly to 0.9165. The sun patches augmentation helps improving the line, goal, and robot classification, but results in a drop in the ball class, while the field prediction stays about the same. This demonstrates that the sun patch augmentation in fact helps.

When considering the model trained solely with data augmented with MUNIT, a clear rise in overall mAP to 0.9536 can be seen, with an improvement for all classes. Adding conventional augmentation to the MUNIT augmentation results in a slight overall mAP increase to 0.9647. While the ball and goal classes’ mAPs drop again slightly, all other classes’ performances increase. This shows that our proposed method yields the desired result. Particularly highly unbalanced classes benefit from the image-to-image translation augmentation.

The ball class is consistently the one with the lowest mAP, which is likely due to highly unbalanced training samples (with the ball being significantly smaller than any other objects). The same applies to the goal post class. The ball is one of the most challenging objects to detect in different lighting conditions since it is small, very close to the ground and throws shadows due to its spherical shape. Lastly, the model operates on a small resolution which makes it impossible to detect the ball at large distances.

Despite its little size with only 12909 trainable parameters (opposed to 300000 parameters in [20]), our proposed model trained with our synthetically generated training data seems to perform better than the approaches proposed



**Fig. 4.** Examples demonstrating the final performance of the neural network on an easy sample with constant lighting (left) and two hard samples with extreme lighting conditions (middle and right). The best performing model was trained on the full dataset (30000 images) with a termination patience of 40 epochs. The classes are encoded as follows: field (green), line (white), robot (pink), ball (red), goal post (blue), and background (black). None of these images were used for the MUNIT training. Despite the extreme lighting conditions, the segmentation performs reasonably well and even underrepresented classes such as the ball are mostly detected successfully.

by van Dijk *et al.* [27] and by Szemenyei *et al.* [24]. A quantitative comparison is difficult due to the differing capabilities of the models. While our model operates at a low resolution of only  $120 \times 88$  pixels, van Dijk *et al.* operates at QVGA resolution and Szemenyei *et al.* use QQVGA resolution while Schnekenburger *et al.* [20] use an image of  $640 \times 512$  as input. In contrast to van Dijk *et al.*, our model is successful in predicting multiple classes at once. In contrast to Szemenyei *et al.*, our model seems to predict a subjectively more precise multi-class classification for each independent prediction, which Szemenyei *et al.* solve with an expensive label propagation. Due to the same reasons, a runtime comparison is difficult, as van Dijk *et al.* and Schnekenburger *et al.* utilize GPUs for the inference. With 14 ms, our model is faster than the fastest model proposed by Szemenyei *et al.* (22 ms + 170 ms label propagation).

The segmentation model was applied to images recorded in extreme lighting conditions. Multiple samples can be seen in Fig. 4.

## 7 Conclusion

We proposed a method for segmenting an image in real-time at a reduced resolution into five different classes by utilizing a deep learning approach. We generated all training data synthetically and evaluated the results on a test set made of manually labeled real data.

We demonstrated our proposed method of generating a dataset with image-to-image translation with publicly available simulation tools. With this dataset, we showed that a small semantic segmentation model is capable of running in real-time on low-end hardware while producing results that outperform related work. In contrast to end-to-end solutions such as [3, 2], which integrate the image augmentation into the model, our approach allows to generate a versatile and high quality dataset, which we share with the community<sup>3</sup>, without the need to have access to high performance GPUs required to generate such datasets or the knowledge to design image-to-image translation models.

## References

1. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), <http://tensorflow.org/>, software available from tensorflow.org
2. Bewley, A., Rigley, J., Liu, Y., Hawke, J., Shen, R., Lam, V.D., Kendall, A.: Learning to drive from simulation without real world labels. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE (2019)
3. Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al.: Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In: 2018 IEEE International Conference on Robotics and Automation (ICRA) (2018)
4. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* 30(2) (2009)
5. Cireşan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation* 22(12) (2010)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. *International journal of computer vision* 88(2) (2010)
7. Fiedler, N., Bestmann, M., Hendrich, N.: ImageTagger: An open source online platform for collaborative image labeling. In: RoboCup 2018: Robot World Cup XXII. Springer (2018)
8. Hess, T., Mundt, M., Weis, T., Ramesh, V.: Large-scale stochastic scene generation and semantic annotation for deep convolutional neural network training in the RoboCup SPL. In: RoboCup 2017: Robot World Cup XXI. Springer (2018)
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient convolutional neural networks for mobile vision applications. *CoRR* abs/1704.04861 (2017)
10. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: *Computer Vision – ECCV 2018*. Springer (2018)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37. ICML’15*, JMLR.org (2015)
12. Jakobi, N., Husbands, P., Harvey, I.: Noise and the reality gap: The use of simulation in evolutionary robotics. In: *Advances in Artificial Life*. vol. 929. Springer (1995)

---

<sup>3</sup> [https://sibylle.informatik.uni-bremen.de/public/datasets/semantic\\_segmentation](https://sibylle.informatik.uni-bremen.de/public/datasets/semantic_segmentation)

13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. *International Conference on Learning Representations* (2014)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: *Computer Vision – ECCV 2014*. Springer (2014)
15. Poppinga, B., Laue, T.: JET-Net: Real-time object detection for mobile robots. In: *RoboCup 2019: Robot World Cup XXIII. Lecture Notes in Artificial Intelligence*, vol. 11531. Springer (2019)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. NIPS’15*, MIT Press (2015)
17. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: *Computer Vision – ECCV 2016*. Springer (2016)
18. Ronneberger, O., P.Fischer, Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. LNCS, vol. 9351. Springer (2015)
19. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
20. Schnekenburger, F., Scharffenberg, M., Wülker, M., Hochberg, U., Dorer, K.: Detection and localization of features on a soccer field with feedforward fully convolutional neural networks (FCNN) for the adult-size humanoid robot Sweaty. In: *Proceedings of the 12th Workshop on Humanoid Soccer Robots, 17th IEEE-RAS International Conference on Humanoid Robots*. Birmingham (2017)
21. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *Journal of Big Data* 6(1), 60 (2019)
22. SoftBank Robotics Documentation: What’s new in NAOqi 2.8? <http://doc.aldebaran.com/2-8/index.html> (2019), retrieved 09/15/19
23. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: *ICLR (workshop track)* (2015), <http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a>
24. Szemenyei, M., Estivill-Castro, V.: Real-time scene understanding using deep neural networks for RoboCup SPL. In: *RoboCup 2018: Robot World Cup XXII*. Springer (2019)
25. Thielke, F., Hasselbring, A.: A JIT compiler for neural network inference. In: *RoboCup 2019: Robot World Cup XXIII. Lecture Notes in Artificial Intelligence*, vol. 11531. Springer (2019)
26. Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2018)
27. van Dijk, S.G., Scheunemann, M.M.: Deep learning for semantic segmentation on minimal hardware. In: *RoboCup 2018: Robot World Cup XXII*. Springer (2019)
28. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. *CoRR* abs/1505.00853 (2015), <http://arxiv.org/abs/1505.00853>