IJCARS manuscript No.
(will be inserted by the editor)

The Effects of Different Levels of Realism on the Training of CNNs with only Synthetic Images for the Semantic Segmentation of Robotic Instruments in a Head Phantom

Saul Alexis Heredia Perez · Murilo Marques Marinho* · Kanako Harada · Mamoru Mitsuishi

Received: December 27, 2019 / Accepted: April 23, 2020

Funding: This work was supported by JSPS KAKENHI Grant Number JP19K14935.

Conflict of interest: The authors declare that they have no conflict of interest.

Abstract ***Purpose:***

The manual generation of training data for the semantic segmentation of medical images using deep neural-networks is a time-consuming and error-prone task. In this paper, we investigate the effect of different levels of realism on the training of deep neural-networks for semantic segmentation of robotic instruments. An interactive virtual-reality environment was developed to generate synthetic images for robot-aided endoscopic surgery. In contrast with earlier works, we use physically-based rendering for increased realism.

Methods:

Using a virtual reality simulator that replicates our robotic setup, three synthetic image databases with an increasing level of realism were generated: flat, basic, and realistic (using the physically-based rendering). Each of those databases was used to train 20 instances of a UNet-based semantic-segmentation deep-learning model. The networks trained with only synthetic images were evaluated on the segmentation of 160 endoscopic images of a phantom. The networks were compared using the Dwass-Steel-Critchlow-Fligner non-parametric test.

Results:

Our results show that the levels of realism increased the mean intersection-over-union (mIoU) of the networks on endoscopic images of a phantom ($p<0.01$). The median mIoU values were 0.235 for the flat dataset, 0.458 for the basic, and 0.729 for the realistic. All the networks trained with synthetic images outperformed naive classifiers. Moreover, in an ablation study, we show that the mIoU of physically-based rendering is superior to texture mapping ($p<0.01$) of the instrument (0.606), the background (0.685), and the background and instruments combined (0.672).

Conclusion:

Using physical-based rendering to generate synthetic images is an effective approach to improve the training of neural networks for the semantic segmentation of surgical instruments in endoscopic images. Our results show that this strategy can be an essential step in the broad applicability of deep neural-networks in semantic segmentation tasks and help bridge the domain gap in machine learning.

Keywords Deep learning · Semantic segmentation · Photorealistic rendering

Authors are with Department of Mechanical Engineering, The University of Tokyo, Tokyo, Japan

*Murilo Marques Marinho (Corresponding author) Tel.: +81-3-5841-6357 ORCID: 0000-0003-2795-9484 E-mail: murilo@nml.t.u-tokyo.ac.jp

Saul Alexis Heredia Perez ORCID: 0000-0002-1812-8649 E-mail: saul@nml.t.u-tokyo.ac.jp
· Kanako Harada ORCID: 0000-0002-0221-7890 E-mail: kanako@nml.t.u-tokyo.ac.jp
· Mamoru Mitsuishi E-mail: mamoru@nml.t.u-tokyo.ac.jp

1 Introduction

Hundreds of procedures on actual patients are required before a doctor can graduate from surgical residency [11]. For example, before a surgeon can take the examination to be certified by the "American Board of Surgery," the surgeon needs to have at least 850 operative procedures in their five years of residence. Even

though some of the initial procedures are supervised, surgical skill positively affects surgical outcomes [19]. At the same time that training with actual patients is required, we must minimize the exposure of patients to the risks of inexperience [19, 22].

In this context, we have been developing a versatile robotic system, called SmartArm [13], which can assist in many surgical scenarios. In addition, our group and collaborators have been working on the development of artificial phantoms with high anatomical-fidelity, for instance, the Bionic-Brain [14]. The Bionic-Brain is a head phantom used to train for neurosurgery. Most recently, the SmartArm system has been successfully validated in endoscopic dura-mater suturing using the Bionic-Brain, through an endonasal approach [13].

Endonasal suturing of the dura mater is a complex task because of the narrow workspace, being challenging even for expert neurosurgeons. Our initial results [13] indicate that using the SmartArm robot through the endonasal approach is feasible and can reduce task time and increase accuracy when compared to doing it manually. To further increase the accuracy of our robotic system and to streamline the training of users, the precise positional information of the robotic instruments is required. In an earlier work [12], we have shown that the instruments' position obtained from the robot's encoders, even after careful offline calibration, can still be inaccurate by a few millimeters. By using the endoscopic image, which is readily available during surgery, we aim to perform an online calibration of the robotic systems, increasing its safety and accuracy. The first step towards this direction is to be able to accurately segment the robotic instruments in the endoscopic images. Before the SmartArm can go through a certification process to perform clinical trials and eventually be introduced in the operating theater to operate inside a human patient, it is essential to study first the robotic instrument calibration using phantoms inside a controlled setup. The segmentation of clinical images is out of the scope of this work.

In recent years, the quality of the semantic segmentation of medical instruments in endoscopic images has been dramatically improved by the use of convolutional neural-networks [1] (CNNs). CNNs [21] are neural-networks in which at least one of the layers performs a convolution operation. Like other types of neural networks, CNNs are composed of several layers, each of which has trainable parameters and can perform different mathematical operations. Similarly to other types of machine learning, CNNs are trained by feeding pairs of input and expected outputs and calculating an error signal commonly called loss. Then, the derivative of the trainable parameters of the network with respect to the loss can be calculated through back-propagation using stochastic-gradient descent (or similar methods). By iterating through the training dataset, the network parameters are slowly moved towards a "good enough" local minimum of the loss. A network with a large number of trainable parameters can learn complex non-linear functions [10].

A large amount of data is required to train the large scale networks required to solve complex tasks. In the case of the semantic segmentation of endoscopic images, the training data are often composed of endoscopic images paired with their manually segmented versions. In this context, the manual generation of large amounts of data becomes a time-consuming and error-prone task. Moreover, a change in the design of the instrument might invalidate all prior manually annotated data. Synthetically generated data is then a reasonable alternative [23, 25, 8], because after a suitable virtual environment is developed, a change in the instrument's structure can be readily taken into account by changing the computer model of the instrument.

In [8], the authors evaluated the detection of objects using Faster R-CNN [18], trained using purely synthetic data of randomly placed objects. Although no photorealistic rendering was employed, the model trained with synthetic data outperformed object detectors trained purely on real data. Other works, such as [7], focused in the detection of objects with reflectance materials, using ray casting to generate photorealistic images. They found that using a combination of photorealistic images and domain optimization has the potential to train robust object detectors on synthetic data that can be successfully applied to real-world images.

Closer to the application sought in this work, a surgical simulation was employed in [25] to create a dataset to train deep learning models for surgical instrument detection in cataract operations. The surgical simulator they made can be considered an animation instead of an interactive environment. In our work, we focus on the real-time interactive simulation and photorealistic rendering to generate a synthetic database. Additionally, we analyze the effect of the rendering quality in the training of the deep learning models.

1.1 Statement of contributions

With the prior discussion in mind, we aim to improve on the state-of-the-art in two ways. (1) Develop a VR environment used to study the effects of increasing levels of realism in synthetic images on the generalization of the trained CNNs to real endoscopic images of a phantom. We hypothesize that higher levels of realism will have increased validation performance. Our methodology for

generating photorealistic images is discussed in detail. Moreover, (2) the data is made available to the scientific community[1] in hopes that it can become a useful benchmark in the validation of networks trained with only simulated data. The proposed training methodology uses the U-Net architecture [20] than can be executed in a relatively affordable 8GB graphics GPU[2].

To the best of the authors' knowledge, this is the first work to address the different levels of realism on the training of CNNs with only synthetic images.

2 Materials and methods

Our real setup for endonasal robotic surgery [13] is briefly described in Section 2.1. The proposed methodology is based on the development of a VR-based simulator described in Section 2.2, based on our real robotic setup. The simulator can render images with three levels of realism. The synthetic images are used to train the deep-learning model described in Section 2.3. The workflow of this study is depicted in the Figure 1. Our aim is to evaluate what are the effects of the different levels of realism on the semantic segmentation accuracy of the deep learning model on real images of the head phantom.

2.1 Real images database

Our robot-assisted setup for endonasal suturing consists of two robotic arms (DENSO VS050, DENSOWAVE, Japan) to which we attach dexterous flexible robotic instruments [2]. In our prior work introducing the SmartArm system [13], we validated our endonasal setup on an anatomically-correct phantom (Bionic-Brain, Medridge, Japan) [14].

In this work, we manually segmented 160 frames obtained from five videos recorded during our validation trials [13]. Samples of the manual segmentation are shown in Fig. 2. The ground-truth images were to be used only for validation of the machine learning model and not for training. This way, we can be sure that our validation data is not contaminated by examples used during training.

[1] Refer to `https://github.com/mmmarinho/levels-of-realism-ijcars2020` for more information.

[2] The initial version of the code is submitted along the manuscript. For a more up-to-date version, refer to`https://github.com/mmmarinho/levels-of-realism-ijcars2020`.

2.2 Synthetic images database

For the production of the synthetic training database images, we developed an interactive VR simulator that replicates the physical simulation of the dura mater suturing by the SmartArm surgical robotic system, as shown in Fig. 3. The position of the simulated tools, the point of view of the virtual camera, and the lighting conditions were set consistently with the physical head phantom, allowing the generation of rendered images that resemble the real operating conditions. The dynamics simulation of tools and the soft bodies were modeled using rigid elements connected by soft joints [4], leading to a rigid multibody system that is simulated in real-time using rigid body dynamics and solved by the PhysX engine (NVIDIA, USA).

Our simulation system is capable of interactively simulating the needle driving and knot tying on a simulated dura mater membrane. Both the VR simulator and the SmartArm surgical robotic system are controlled using the same haptic interfaces providing force feedback. The objects' pose data from the VR simulator is recorded at 30 Hz, so that the whole scene can be reconstructed offline by a different software created using the Unity 3D (Unity Technologies, USA) game engine, to synthesize the frame images while applying different shading techniques. Under this approach, the ground truth segmentation is obtained with little effort by rendering the instruments using a solid white color and using a uniform black color for the background and the other objects.

In 3D computer graphics, rendering is the process of producing a graphical output from the objects described inside a virtual scene, typically modeled as a set of triangle meshes. Unlike raytracing, in real-time rendering, the triangles are projected to the virtual camera's image plane and discretized into pixels during a process called rasterization. Shading is the process through which the final color of each pixel is determined based on variables such as the camera point of view, surface curvature, the incidence of the light, and the material's color properties of the virtual objects that originated such pixels. This process involves complex computations coded into a program called shader that runs in parallel on the GPU. By extension, the term 'shader' is also used to refer to the implemented technique. In this work, the term 'renderer' will also be used for the same effect. It is possible to achieve different visual effects using a shader, ranging from artistic styles such as toon shading to realistic-looking images using lighting models to determine the resulting color considering the interaction of the simulated light sources with

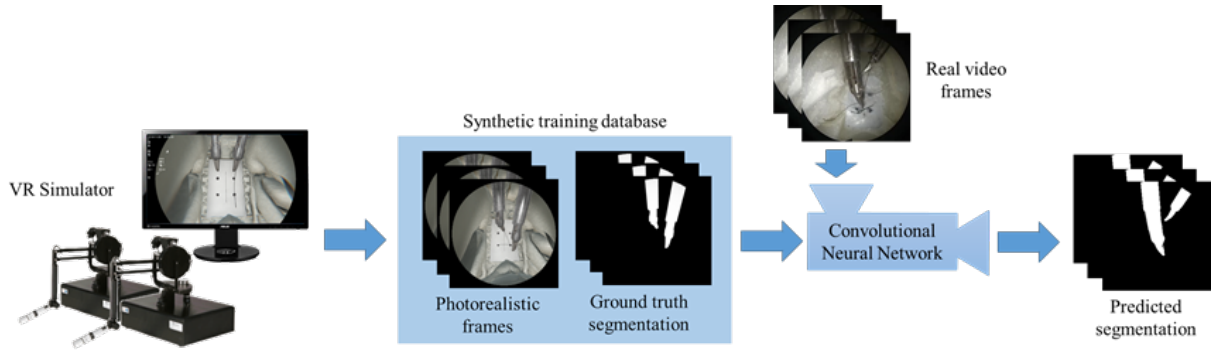

Fig. 1 An overview of the proposed methodology in this study. First, the synthetic training database is generated using data from VR simulation. CNN is then trained using synthetic images. Finally, the trained CNN is tested on real images.

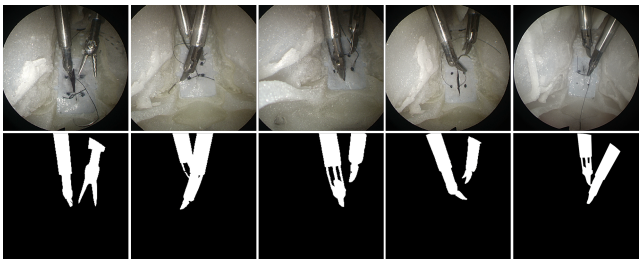

Fig. 2 Samples of manually segmented images. Note that the topology of the annotations is complex and might have holes through which the background is visible. Moreover, different trials had different camera placements and light conditions, which affect the visibility of the instruments and the appearance of the background.

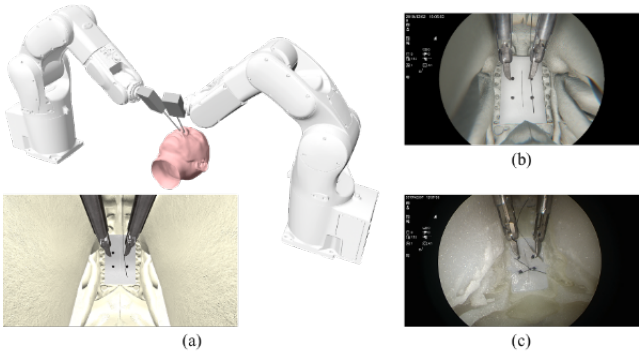

Fig. 3 a) Simulated scenario with two robot arms equipped with thin tools operating inside the nasal cavity. b) A screenshot of the VR simulation and c) a video frame from the physical simulation.

the optical properties of the virtual surfaces mathematically.

Physically-based rendering (PBR) comprises techniques aiming to obtain photorealistic images by accurately modeling the physical nature of light and its interaction with the material's surfaces in the real world [17]. Real-time PBR uses microfacets theory [24] to statistically model rough surfaces, and the principle of energy conservation (a surface never reflects more light than it receives) to achieve realistic diffuse and specular reflection. In PBR, the optical properties of materials are modeled using two intuitive parameters, namely roughness and metalness [16]. The roughness parameter controls how smooth or rough is a surface: while a smooth surface reflects light in a specular way, rough surfaces diffusely scatter the light. The metalness parameter (also known as metallic or metallicity) controls the metal-like appearance of material by modulating the reflectivity of the surface. By adjusting these two parameters, the appearance of different materials can be approximated. A particular material is then described by its base color (albedo), roughness, and metalness properties encoded as a bi-dimensional map or texture. High-frequency surface details such as scratches on the metallic instruments and the porous surface of the synthetic bones can be approximated using normal mapping [9].

To analyze the effect of realism of the synthetic database in the training of the network, we defined three levels of realism. The first level corresponds to a 'flat' shading without any gradient using a solid color matching the average color at each object. The second level, called 'basic' shading, was obtained using only diffuse lighting [3], producing intensity gradients according to the light incidence and the surface normal. The third level of realism uses the PBR implementation by Unity engine to achieve photorealistic rendering (Fig. 4), especially of the robotic instruments with plausible specular highlights and metal-like surface reflections. We formed a database consisting of 10,376 color images of 256x256 pixels for each shading condition with their respective ground truth segmentation. Both color images and the respective segmentation were converted to 8-bit grayscale images before its use as input for the machine-learning model.

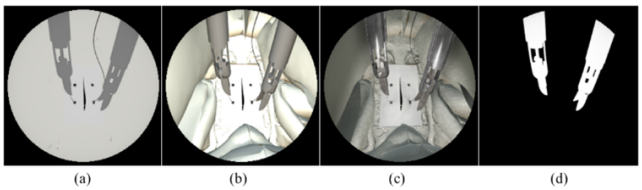

Fig. 4 Different rendering qualities applied to the same simulation frame: a) flat shader (solid color), b) basic shader (Phong), c) realistic shader (physically-based rendering), d) the corresponding ground truth segmentation.

2.3 CNN architecture and loss function

For the semantic segmentation of the robotic instruments, we used an architecture based on the U-Net [20], implemented in Tensorflow 2.0[3] in Python. The U-Net has a symmetric encoder-decoder structure. The encoder reduces the size of the image after each consecutive convolution while increasing the number of learnable filters, in a total of five stages. The decoder increases the size of the image while reducing the number of learnable filters. This strategy has been shown to be quite effective in the semantic segmentation of biomedical images [20] and is the base of many architectures being used in the semantic segmentation of medical instruments [1].

The input of our network was a single-channel greyscale image, $I(x,y) \in [0,1] \subset \mathbb{R}$, with height $h = 256$ and width $w = 256$ pixels, i.e. $x, y \in [1,256] \subset \mathbb{N}$. The size is a configurable variable that does not increase the number of filters required by the network but increases the memory requirement and runtime.

The output of the network was a single-channel grayscale image with the same size as the input, $O(x,y) \in [0,1] \subset \mathbb{R}$. The intensity of each pixel corresponded to the level of confidence of the network in that pixel being part of a robotic instrument. For example, $O(x,y) = 0$ means the network is confident that the pixel was part of the background, and $O(x,y) = 1$ that the pixel was part of a robotic instrument. We used a threshold of 0.5 to decide whether the pixel was from the instrument or from the background.

The loss function was the binary crossentropy

$$L(I,O) = \frac{1}{w.h} \sum_{j=1}^{h} \sum_{i=1}^{w} \left(I(i,j) \log\left(O(i,j)\right) + (1 - I(i,j)) \log\left(1 - O(i,j)\right) \right)$$

which is a standard loss function for binary classification problems.

Following the recommendations in the initial U-Net publication [20], the network was trained using stochastic gradient descent.

[3] `https://www.tensorflow.org/`

2.4 Data augmentation

Three types of data augmentation were used. The first was an affine transformation, i.e. a rotation of the image about its center in the interval between -45 and 45 degrees, followed by a translation between ± 20 pixels in the horizontal and vertical directions. The second was a brightness augmentation between adding $\pm 50\%$ to the intensity of the whole image. The last was a random uniform additive noise in the interval $[-0.1, 0.1] \subset \mathbb{R}$ for each pixel. No other augmentation strategy was used.

3 Experiments

With the photorealistic rendering strategies described in Section 2, three datasets were generated, namely *flat*, *basic*, and *realistic*. Each dataset was composed of 10376 images of the same simulated endonasal suturing inside the head phantom. The only difference between them was the rendering strategy. Given that we wanted to evaluate the effects of the rendering strategies on the generalization of the network to real images, no real images were used for training. The learning rate was 0.02. Each network was trained for 40000 iterations, and the learning rate was reduced by a factor of two after every 15000 iterations. For the network to fit into an 8GB graphics GPU, we used a batch size of eight. With these settings, each network took about four hours of training on an NVIDIA 2070 RTX GPU.

To study the performance of each network trained only with simulated images in the semantic segmentation of real images, after every 1000 iterations, each network was evaluated in the semantic segmentation of the manually-annotated dataset composed of 160 images sampled from five trials of robot-assisted endonasal suturing [13]. Using this methodology, we trained 20 networks for each dataset to provide a reasonable number of samples for statistical inference.

The metrics for the validation were the loss (binary cross-entropy) and the mean intersection-over-union (mIoU). We compared the median mIoU of all groups using the Kruskal-Wallis test. Post-hoc pairwise comparisons were made using the Dwass-Steel-Critchlow-Fligner [6] (DS) all-pairs comparison test. The DS test is a non-parametric test that evaluates the statistical significance in the differences of the medians of two distributions. It has been shown to be more suitable for comparisons of unknown distributions with unequal variances [5]. We used the two-tailed confidence level of 95%.

We hypothesized that increased levels of realism in the simulated images used for training would increase the performance of the network in the semantic segmentation of real images. We also expected the networks to

overfit to their training dataset composed of only simulated images.

3.1 Ablation study

In addition to the main experiments, we conducted an ablation study to compare the PBR-based 'realistic' shader with texture mapping. For this ablation study, we generated three extra synthetic databases. The first database was generated by mapping a snapshot of the head phantom as the background texture of the VR simulation. We regard this experimental condition as 'photo BG'. For the second database, the background was rendered using PBR and the instruments were rendered by mapping a real photo of the instruments as the instrument's textures. We call this case as 'photo I'. For the last database, we used the photos to render both the background and the instruments. These conditions are exemplified in Fig. 5.

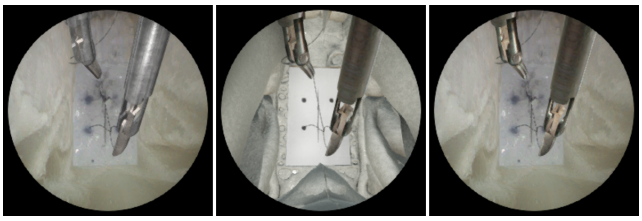

Fig. 5 Additional rendering styles for ablation studies of the proposed methodology. From left to right: using a photo as background, applying texturing mapping to instruments using photos, and both background and instruments are rendered using texture mapping using real photos.

4 Results

The best mIoU in the semantic segmentation of real images for each network was logged during training. The median of those values for each dataset is shown in Fig. 7. The loss and mIoU during training are shown in Fig. 8. The results of the ablation study are shown in Fig. 9.

5 Discussion

The networks trained with the flat renderer had the lowest mIoU ($p < 0.01$). After about 2000 iterations, the validation mIoU stagnated near 0.28, as shown in Fig. 8. As shown in Fig 8, the loss on the validation dataset diverged while the loss on the training dataset converged. This divergence indicates that the networks

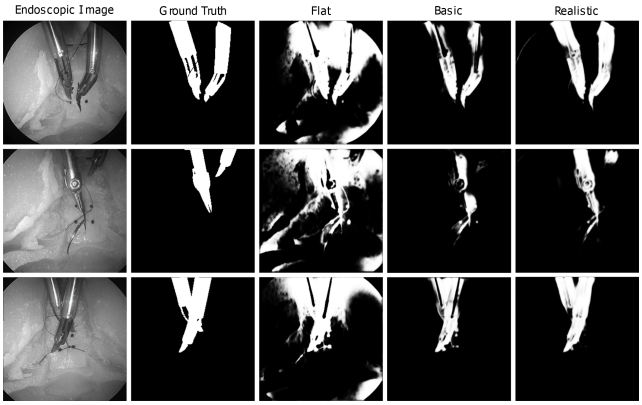

Fig. 6 Results of the semantic segmentation of one real endoscopic image by a network trained in each of the three rendering conditions. The network trained by the flat renderer seems to be mostly working as an intensity threshold and had many false positives. The network trained with the basic renderer successfully segmented the shape of the tool but failed to segment the reflecting parts of the shafts of the instruments. The network trained with the realistic renderer was able to segment the input image with high accuracy, even in the presence of partial occlusion, although some misclassifications are still visible.

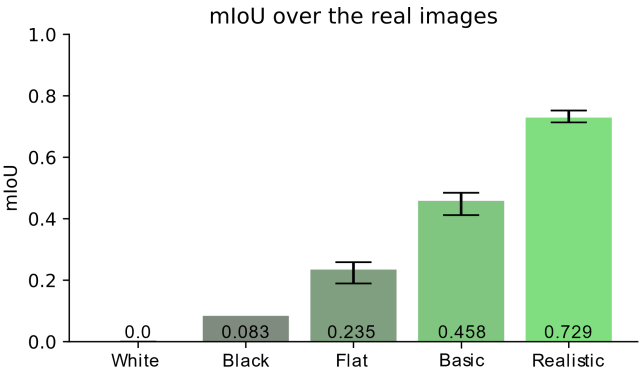

Fig. 7 Medians and 95% confidence intervals of the best mean IoU achieved by training on each dataset synthetic image dataset. "Black" was a naive classifier that classified all pixels of the image as background. "White" was a naive classifier that classified all pixels of the image as the instrument. Pair-wise comparisons using the Dwass-Steel-Critchlow-Fligner test for medians showed that all pair-wise differences were statistically significant ($p < 0.01$).

trained with the flat renderer overfit to the training data and were unable to generalize well to real images. The flat renderer was still 183% better than a naive renderer that classifies all pixels as being part of an instrument, which means that some information from the simulated images was still useful.

The networks trained with the basic renderer showed a 94% better mIoU than those trained with the flat renderer ($p < 0.01$). The mIoU over real images stabilized near 0.40. The validation loss had a convergent pattern in the first 5000 steps but was followed by a divergent pattern in the following steps. This indicates that the

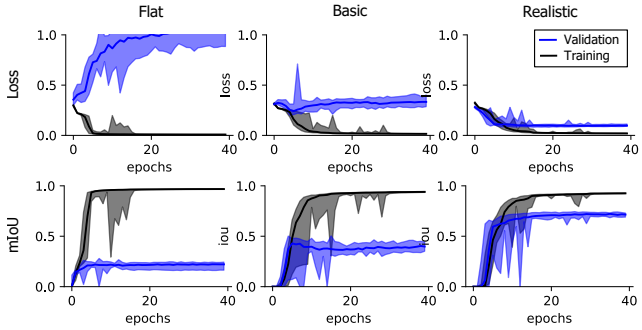

Fig. 8 Loss and mIoU during training. The solid black line corresponds to the median on the training data (simulation), and the solid blue line corresponds to the test data (real images). The regions around the lines are the 95% confidence intervals for the medians.

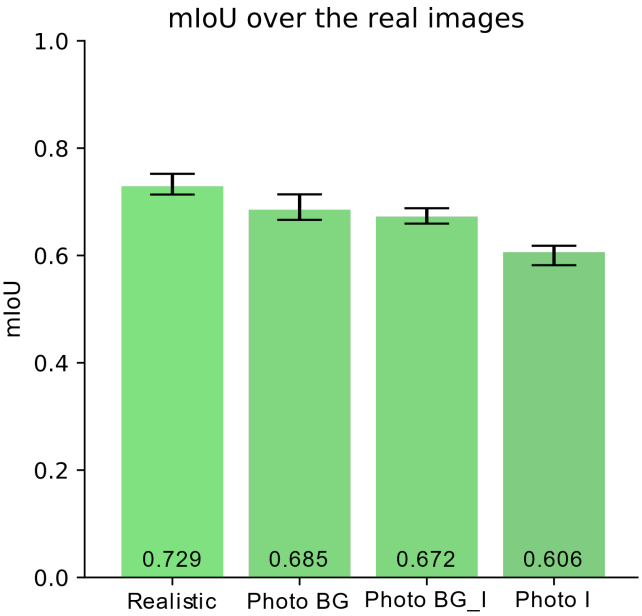

Fig. 9 Comparing with the 'realistic' dataset, the mIoU was about 6% worse ($p < 0.01$) for 'Photo BG', about 8% worse ($p < 0.01$) for 'Photo BG_I', and about 17% worse ($p < 0.01$) for 'Photo I'.

network was overfitting to the training data. Despite this overfitting, the network still showed decent performance in the validation of real images.

The networks trained with the realistic renderer had a 210% increase in mIoU over the networks trained with the flat renderer ($p < 0.01$) and a 59% increase in mIoU over the networks trained with the basic renderer ($p < 0.01$). In addition, the loss function showed convergence throughout the entire training. This convergence indicates that the network might not be overfitting to the training images and that there is a learnable similarity between the real images and the artificial images. In addition, because the training dataset contained samples with partial occlusion of the instruments, the networks performed well on real data containing partial occlusion of the instruments.

Given that the range of movement of the camera inside the head phantom is limited, we opted to fix the camera point throughout the synthetic dataset. Nonetheless, we analyzed the effect of the camera motion in the realistic dataset by applying random translations and rotation of the virtual camera within a certain range. However, this did not improve the mIoU (0.715, $p < 0.01$). This might be due to some of the unnatural camera viewpoints added by automatically moving the camera.

With these results, the main hypothesis of the paper has been successfully tested. More realistic rendering has considerably better performance than simplified rendering. The low level of the realism of the rendering might be one of the factors that affected the classifying power of earlier works [25].

The head phantom accurately resembles human anatomy, but it does not replicate the color or photometric properties of living tissues. With that in mind, we do not expect the CNN trained with synthetic data of the head phantom to work well on clinical background conditions. However, we do expect the presented framework, possibly in conjunction with other techniques such as image translation for domain transfer [15], to be applicable to clinical images when clinical data becomes available.

In the ablation study, we applied texture mapping using real photos of the head phantom and the instruments as an alternative method to generate realistic synthetic frames. Texture mapping allowed the generation of somewhat convincing images, but it failed to accurately reproduce the variability present in the real images. We believe this was mostly due to the lack of ambient occlusion and the contact shadows, as well as the improper metallic appearance of the instruments which did not include reflections and highlights. In all the ablation experiments, the texture-mapping performance was worse compared to the realistic shader using PBR: about 6% worse ($p < 0.01$) for 'Photo BG', about 8% worse ($p < 0.01$) for 'Photo BG_I', and about 17% worse ($p < 0.01$) for 'Photo I'. These results suggest that the variability obtained using PBR might enhance the performance of CNNs trained with synthetic images and help bridge the domain gap in machine-learning research.

6 Conclusions

In this work, we developed a virtual-reality environment that replicates our robot-aided endonasal suturing

setup [13], inside a head phantom [14]. Using the simulator, we created three synthetic image databases with increasing levels of realism. We used the three databases to train deep neural-network models and evaluated the performance of the models–trained only on synthetic images–on the classification of real images. We found that increased levels of realism increased the semantic segmentation of real images.

Ethical approval: This article does not contain any studies with human participants or animals.

References

1. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., García-Peraza, L.C., Li, W., Iglovikov, V., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., Azizian, M.: 2017 robotic instrument segmentation challenge. CoRR **abs/1902.06426** (2019). URL `http://arxiv.org/abs/1902.06426`
2. Arata, J., Fujisawa, Y., Nakadate, R., Kiguchi, K., Harada, K., Mitsuishi, M., Hashizume, M.: Compliant four degree-of-freedom manipulator with locally deformable elastic elements for minimally invasive surgery. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE (2019). DOI 10.1109/icra.2019.8793798
3. Bishop, G., Weimer, D.M.: Fast phong shading. ACM SIGGRAPH Computer Graphics **20**(4), 103–106 (1986). DOI 10.1145/15886.15897
4. Budsberg, J., Zafar, N.B., Aldén, M.: Elastic and plastic deformations with rigid body dynamics. In: ACM SIGGRAPH 2014 Talks on - SIGGRAPH14. ACM Press (2014). DOI 10.1145/2614106.2614132
5. Dolgun, A., Demirhan, H.: Performance of nonparametric multiple comparison tests under heteroscedasticity, dependency, and skewed error distribution. Communications in Statistics - Simulation and Computation **46**(7), 5166–5183 (2016). DOI 10.1080/03610918.2016.1146761
6. Douglas, C.E., Michael, F.A.: On distribution-free multiple comparisons in the one-way analysis of variance. Communications in Statistics - Theory and Methods **20**(1), 127–139 (1991). DOI 10.1080/03610929108830487
7. Hartwig, S., Ropinski, T.: Training object detectors on synthetic images containing reflecting materials. arXiv preprint arXiv:1904.00824 (2019)
8. Hinterstoisser, S., Pauly, O., Heibel, H., Marek, M., Bokeloh, M.: An annotation saved is an annotation earned: Using fully synthetic training for object instance detection. CoRR **abs/1902.09967** (2019). URL `http://arxiv.org/abs/1902.09967`
9. Kilgard, M.J.: A practical and robust bump-mapping technique for todays gpus. In: Game Developers Conference 2000, pp. 1–39 (2000)
10. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). DOI 10.1038/nature14539
11. Malangoni, M.A., Biester, T.W., Jones, A.T., Klingensmith, M.E., Lewis, F.R.: Operative experience of surgery residents: Trends and challenges. Journal of Surgical Education **70**(6), 783–788 (2013). DOI 10.1016/j.jsurg.2013.09.015
12. Marinho, M.M., Adorno, B.V., Harada, K., Mitsuishi, M.: Dynamic active constraints for surgical robots using vector-field inequalities. IEEE Transactions on Robotics **35**(5), 1166–1185 (2019). DOI 10.1109/tro.2019.2920078
13. Marinho, M.M., Harada, K., Morita, A., Mitsuishi, M.: Smartarm: Integration and validation of a versatile surgical robotic system for constrained workspaces. The International Journal of Medical Robotics and Computer Assisted Surgery (IJMRCAS) (2019). DOI 10.1002/rcs.2053. (In press)
14. Masuda, T., Kanako, H., Adachi, S., Arai, F., Omata, S., Morita, A., kin, T., Saito, N., Yamashita, J., Chinzei, K., Haswgawa, A., Fukuda, T.: Patients simulator for transsphenoidal surgery. In: 2018 International Symposium on Micro-NanoMechatronics and Human Science (MHS). IEEE (2018). DOI 10.1109/mhs.2018.8886922
15. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4500–4509 (2018)
16. Peddie, J.: Work flow and material standards. In: Ray Tracing: A Tool for All, pp. 65–90. Springer International Publishing (2019). DOI 10.1007/978-3-030-17490-3_5
17. Pharr, M., Humphreys, G., Jakob, W.: Physically Based Rendering. Elsevier LTD, Oxford (2016). URL `https://www.ebook.de/de/product/25867811/matt_pharr_greg_humphreys_wenzel_jakob_physically_based_rendering.html`
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)
19. Reznick, R.K., MacRae, H.: Teaching surgical skills — changes in the wind. New England Journal of Medicine **355**(25), 2664–2669 (2006). DOI 10.1056/nejmra054785
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science, pp. 234–241. Springer International Publishing (2015). DOI 10.1007/978-3-319-24574-4_28
21. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (2017). DOI 10.1109/tpami.2016.2572683
22. van der Sluis, P.C., Ruurda, J.P., van der Horst, S., Goense, L., van Hillegersberg, R.: Learning curve for robot-assisted minimally invasive thoracoscopic esophagectomy: Results from 312 cases. The Annals of Thoracic Surgery **106**(1), 264–271 (2018). DOI 10.1016/j.athoracsur.2018.01.038
23. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: 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). IEEE (2017). DOI 10.1109/iros.2017.8202133
24. Torrance, K.E., Sparrow, E.M.: Theory for off-specular reflection from roughened surfaces. Journal of the Optical Society of America **57**(9), 1105 (1967). DOI 10.1364/josa.57.001105
25. Zisimopoulos, O., Flouty, E., Stacey, M., Muscroft, S., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D.: Can surgical simulation be used to train detection and classification of neural networks? Healthcare Technology Letters **4**(5), 216–222 (2017). DOI 10.1049/htl.2017.0064