




# Improving Surgical Training Phantoms by Hyperrealism: Deep Unpaired Image-to-Image Translation from Real Surgeries

Sandy Engelhardt<sup>1,3</sup> , Raffaele De Simone<sup>2</sup>, Peter M. Full<sup>2</sup>,  
Matthias Karck<sup>2</sup>, and Ivo Wolf<sup>3</sup>

<sup>1</sup> Department of Simulation and Graphics & Research Campus STIMULATE,  
Magdeburg University, Magdeburg, Germany

s.engelhardt@hs-mannheim.de

<sup>2</sup> Department of Cardiac Surgery,  
Heidelberg University Hospital, Heidelberg, Germany

<sup>3</sup> Faculty of Computer Science,  
Mannheim University of Applied Sciences, Mannheim, Germany

**Abstract.** Current ‘dry lab’ surgical phantom simulators are a valuable tool for surgeons which allows them to improve their dexterity and skill with surgical instruments. These phantoms mimic the haptic and shape of organs of interest, but lack a realistic visual appearance. In this work, we present an innovative application in which representations learned from real intraoperative endoscopic sequences are transferred to a surgical phantom scenario. The term *hyperrealism* is introduced in this field, which we regard as a novel subform of surgical augmented reality for approaches that involve real-time object transfigurations. For related tasks in the computer vision community, unpaired cycle-consistent Generative Adversarial Networks (GANs) have shown excellent results on still RGB images. Though, application of this approach to continuous video frames can result in flickering, which turned out to be especially prominent for this application. Therefore, we propose an extension of cycle-consistent GANs, named *tempCycleGAN*, to improve temporal consistency. The novel method is evaluated on captures of a silicone phantom for training endoscopic reconstructive mitral valve procedures. Synthesized videos show highly realistic results with regard to (1) replacement of the silicone appearance of the phantom valve by intraoperative tissue texture, while (2) explicitly keeping crucial features in the scene, such as instruments, sutures and prostheses. Compared to the original CycleGAN approach, *tempCycleGAN* efficiently removes flickering between frames. The overall approach is expected to change the future design of surgical training simulators since the generated sequences clearly demonstrate the feasibility to enable a considerably more realistic training experience for minimally-invasive procedures.

**Keywords:** Generative Adversarial Networks  
 Minimally-invasive surgical training · Augmented reality  
 Mitral valve simulator · Surgical skill

## 1 Introduction

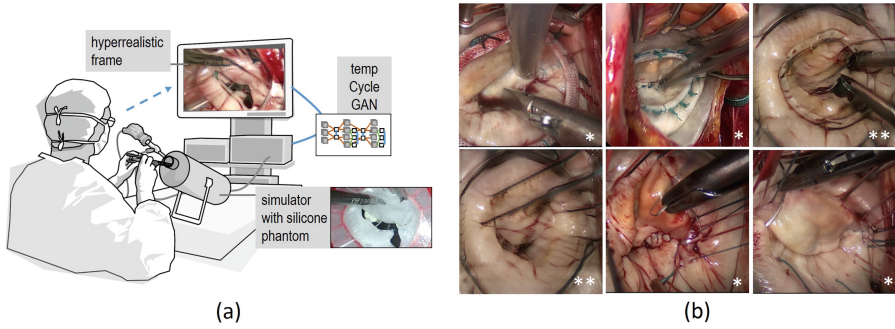
Surgery is a discipline that requires years of training to gain the necessary experience, skill and dexterity. With increasingly minimally invasive procedures, in which the surgeon’s vision often solely relies on endoscopy, this is even more challenging. Due to the lack of appropriately realistic and elaborate endoscopic training methods, surgeons are forced to develop most of their skills in patients, which is truly undesirable. Current training methods rely on practising suturing techniques on *ex-vivo* organs (‘wet labs’), virtual simulators or physical phantoms under laboratory conditions (‘dry labs’). Training on authentic tissue is associated with organizational efforts and costs and is usually not accessible to the majority of the trainees. Virtual simulators overcome these requirements, but are often less realistic due to the lack of blood, smoke, lens contamination and patient-specificity. Physical phantoms, e.g. made from silicone, suffer also from these drawbacks, but they provide excellent haptic feedback and tissue properties for stitching with authentic instruments and suture material [1, 2]. However, their uniform appearance does not reflect the complex environment of a surgical scene. We tackle this issue by proposing a system that is able to map patterns learned from intraoperative video sequences onto the video stream captured during training with silicone models to mimic the intraoperative domain. Our vision for a novel training simulator is to display real-time synthesized images to the trainee surgeon while he/she is operating on a phantom under restricted direct vision, such as illustrated in Fig. 1a.

Generative Adversarial Networks (GANs) demonstrate tremendous progress in the field of image-to-image translation with regard to both perceptual realism and diversity. Recently, methods have been proposed using Convolutional Neural Networks (CNNs) for deep image synthesis with paired [3] and even unpaired natural images, namely DualGAN [4] and CycleGAN [5]. These networks translate an image from one domain X to another target domain Y. The key to the success of GANs is the idea of an adversarial loss that forces the generated images to be, in principle, indistinguishable from real images, which is particularly powerful for image generation tasks. However, current solutions do not take time consistency of a video stream into account. While each frame of a generated video looks quite realistic on its own, the whole sequence lacks consistency.

In order to increase realism for endoscopic surgical training on physical phantoms, we propose the concept of *hyperrealism*.<sup>1</sup> We define hyperrealism as a new paradigm on the Reality-Virtuality continuum [6] as a concept closer to ‘full reality’ in comparison to other applications where artificial overlays are super-

---

<sup>1</sup> The term is related to the homonymous art form, where an excessive use of details is used to create an exaggeration of reality which cannot be seen by the human eye.



**Fig. 1.** (a) Vision: Augmentation of the minimally invasive training process with real-time generated *hyperrealistic* frames. (b) Visual comparison of real intraoperative frames from mitral valve surgery (\*) and generated fake images (\*\*).

imposed on a video frame. In a hyperrealistic environment, those parts of the physical phantom that look unnatural are replaced by realistic appearances.

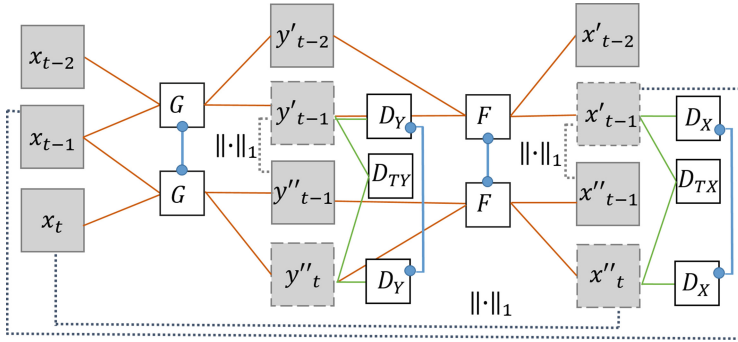
The extended CycleGAN network, named *tempCycleGAN*, learns to translate an image stream from the source domain of *phantom data* to a target domain *intraoperative surgeries* and vice versa in the absence of paired endoscopic examples. The network’s main task is to capture specific characteristics of one image set and to figure out how these characteristics could be translated into the other image domain. We evaluate the approach for the specific application of training mitral valve repair, where the network has to learn (1) how to enhance the silicon’s surface appearance, at the same time not altering its shape, (2) not to replace other important features in the scene, such as surgical instruments, sutures, needles and prostheses.

## 2 Methods

We build upon the CycleGAN model proposed by Zhu et al. [5]. The goal of CycleGANs is to obtain mapping functions between two domains  $X$  and  $Y$  given unpaired training samples,  $\{x_i | i = 1..N\} \in X$  and  $\{y_j | j = 1..M\} \in Y$ . Mappings in both directions are learned by two generator networks,  $G : X \rightarrow Y$  and  $F : Y \rightarrow X$ . The generators are trained to produce output that cannot be distinguished from real images of the target domain by adversarially trained discriminators  $D_Y$  and  $D_X$ .

### 2.1 Temporal Cycle GAN

Temporal consistency requires to include preceding time steps into the learning process. The proposed advancement *tempCycleGAN* processes the current time step  $x_t$  and its two predecessors  $x_{t-1}$  and  $x_{t-2}$ , as shown in Fig. 2. In general,



**Fig. 2.** Training setup of the  $X \rightarrow Y \rightarrow X$  cycle of the proposed tempCycleGAN network (reverse cycle accordingly) using temporal pairs: the generators  $G$ ,  $F$  and the temporal discriminators  $D_{T\{X,Y\}}$  take the current frame and a single preceding frame. Each run of  $G$  (and  $F$ ) synthesizes outputs for both frames. In the application of the generator, the frame of interest is the second output ( $y'_{t-1}$  and  $y''_t$ , respectively). The temporal discriminators are trained on these frames of interest, thus, the generator  $G$  needs to run twice to generate the two frames of interest ( $y'_{t-1}$  and  $y''_t$ ) for  $D_{TY}$  (for  $F$  and  $D_{TX}$  accordingly). L1-distances (dotted lines) between matching time frames are used in the loss function to further enforce time consistency.  $D_{\{X,Y\}}$ : discriminators with 1 input; blue connections: shared weights.

it is possible to use more preceding frames and to adjust the network architecture accordingly. In the following, the general concepts of tempCycleGAN are explained and a detailed description of the setup is provided subsequently.

Temporal discriminators  $D_{TX}$  and  $D_{TY}$  (one for each domain) are introduced that take consecutive frames and try, as usual, to distinguish real from generated data. The idea is that flickering would allow the discriminators to easily identify generated data. Thus, the generators are forced to avoid flickering to successfully cheat their adversarial temporal discriminators. The generators need at least one preceding frame as additional input to be able to create a temporal consistent output for the current frame. In the current setup two frames are used as input for both the generators and the temporal discriminators.

To define a cycle consistency loss that is symmetric in  $G$  and  $F$ , we let the generators  $G$  and  $F$  create as many output frames as they get input frames. For example,  $G(x_{t-1}, x_t)$  creates outputs  $y'_{t-1}$  and  $y''_t$ . Only the output for the latest frame ( $y''_t$  in the example) is the frame of interest used in the actual output video. Consecutive frames of interest (shown as dashed boxes in Fig. 2) are evaluated by the temporal discriminators. Thus, the temporal discriminators are provided with inputs of multiple runs of the generator. For example,  $D_{TY}$  takes  $y'_{t-1}$  and  $y''_t$  as input, where  $y'_{t-1}$  is the frame of interest of  $G(x_{t-2}, x_{t-1})$  and  $y''_t$  is the frame of interest of  $G(x_{t-1}, x_t)$ . To enforce consistency between frames of matching time and domain, L1 distances to the respective frames are used as additional terms in the loss function.

## 2.2 Network Architectures

The network architectures of the generators and discriminators are largely the same as in the original CycleGAN approach [5]. A TensorFlow implementation provided on GitHub<sup>2</sup> was used as the basis and extended with the new temp-CycleGAN blocks. All discriminators take the complete input images, which is different from the  $70 \times 70$  PatchGAN approach by Zhu et al. [5]. The temporal discriminators have 6 ( $2 \times \text{RGB}$ ) instead of 3 input channels. For the generators, 8 instead of 9 residual blocks are used, because experiments on our data showed better results for this configuration.

## 3 Experiments

The commercial minimally invasive mitral valve simulator (MICS MVR surgical simulator, Fehling Instruments GmbH & Co. KG, Karlstein, Germany) was extended with patient-specific silicone valves. Details on 3D-printed mold and valve production are elaborated on in a previous work [2]. An expert segmented mitral valves with different pathologies, such as posterior prolaps and ischemic valves on the end-systolic time step from echocardiographic data. From these virtual models, 3D printable molds and suitable annuloplasty rings were automatically generated with stitching holes using 3 different low to medium cost 3D-printers, varying material (polylactide, acrylonitrile butadiene styrene) in various colors (e.g. white, beige, red, orange). From these molds, 10 silicone valves were cast that could be anchored in the simulator on a printed valve holder. We asked an expert and a trainee to apply mitral valve repair techniques (annuloplasty, triangular leaflet resection, neo-chordae implantation) on these valves and captured the training process endoscopically.

### 3.1 Data and Training of Network

In total, approx. 330,000 video frames from the training procedures in full HD resolution were captured. Valves shown in videos for training were not used for testing. For training, three continuous frames after each 120th frame from a subset of approx. 160,000 frames was sampled retrospectively, such that the set comprised 1300 small sequences. Furthermore, training material for the target domain from 3 endoscopic mitral valve repair surgeries was captured. In total, approx. 320,000 frames were acquired during real surgery. For training, three continuous frames after each 240th frame were sampled retrospectively and 1294 small sequences were used for training. All streams were captured with 30fps. The scenes are highly diverse, as the valve’s appearance drastically changes over time (e.g. due to cutting of tissue, implanting sutures and prostheses, fluids such as blood and saline solution), see Fig. 1b. All frames were square-cropped and re-scaled to  $286 \times 286$ . Data augmentation was performed by random cropping of a  $256 \times 256$  region and random horizontal flipping. For all the experiments,

<sup>2</sup> <https://github.com/LynnHo/CycleGAN-Tensorflow-PyTorch-Simple>.

the consistency loss was weighted with  $\lambda = 10$  [5]. The Adam solver with a batch size of 1 and a learning rate of 0.0001 without linear decay to zero was used. Similar to Zhu et al. [5], the objective was divided by 2 while optimizing  $D$ , which slows down the rate at which  $D$  learns relative to  $G$ . Discriminators are updated using a history of 50 generated images rather than the ones produced by the latest generative networks [5]. The tempCycleGAN network was trained for 40, 60, 80, 100 epochs to find the visually most attractive results. In analogy, the original CycleGAN networks was trained either with 1 input frame or 3 continuous frames.

### 3.2 Evaluation

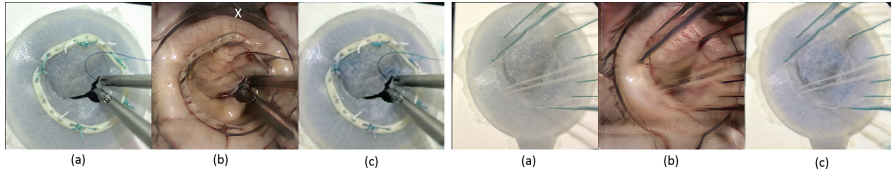
The most important factors for the proposed application are related to perception i.e. how real the generated intraoperative videos appear to an expert with years of experience in mitral valve surgery. Furthermore, reliability plays a crucial role, as the appearance of the scene should be transferred into the target domain, while neither the shape of objects should be altered, nor additional parts should be added or taken away.

*Realness:* An expert was asked to score the visual quality of eleven 10s mini videos synthesized from the test set phantom frames. For assessment, the “realness score” was used, as proposed by Yi et al. [4], ranging from 0 (totally missing), 1 (bad), 2 (acceptable), 3 (good), to 4 (compelling). We decided against the conduction of a Visual Turing Test, as some shape-related features in the scene (a personalized ring shape instead of a standard commercial ring was used in the experiments) would have been easily identified by an expert surgeon.

*Reliability:* The result of tempCycleGAN was comprehensively compared in terms of faithfulness to the input phantom images using 39 randomly selected synthesized frames from the test set (16 showing an annuloplasty and 23 showing a triangular leaflet resection). Predefined criteria relevant for surgery were assessed, e.g. whether the instruments or all green and white sutures are completely visible or whether artifacts disturb the surgical region of interest. 12 of 39 input frames show two instruments and three frames only a single instrument. Stitching needles held by the instrument are used in six frames and a prosthetic ring is visible in 29 frames. On average, 4.9 white and 4.9 green sutures are observable in the phantom frames.

## 4 Results

The result of tempCycleGAN was visually most attractive after 60 epochs. Examples are provided in Fig. 3. Model training of the tempCycleGAN took 18 h for 60 epochs on a single NVIDIA GeForce Titan Xp GPU. Compared to the original CycleGAN [5] trained with a single frame or three consecutive frames, tempCycleGAN produces results with no flickering, contains fewer artifacts, and



**Fig. 3.** tempCycleGAN results for two examples shown left and right, where (a) shows the real phantom  $x_t$ , (b) shows corresponding synthesized intraoperative images  $y_t''$  and (c) shows the re-synthesized phantom image  $x_t''$ . X marks a synthesized atrial retractor.

better preserves content structures in the inputs and capture features (e.g., texture and/or color) of the target domain (see supplemental material videos). The tempCycleGAN approach was even capable of learning where semantically to insert blood (between the leaflets) or an atrial retractor in the scene (Fig. 3). However, it produced slightly blurred instruments and sutures.

The average “realness score” of the 11 mini-videos assessed by the expert surgeon was 3.3 (5 × category “compelling”, 4 × “good”, 2 × “acceptable”). Longer versions of a compelling scene (first scene) and an acceptable scene (second scene) are provided in the supplemental video showing a ring annuloplasty<sup>3</sup>. The valve’s texture was assessed as very realistic in general by the surgeon. Some instruments and rings appeared blurry and had minor artifacts (e.g. the projection of the sewing cuff of the original ring onto the printed ring appeared incomplete), which led to a lower realness score. However, for most of the scenes this was not crucial because relevant image regions were not effected.

The quantitative assessment of reliability (i.e. comparison of source and target frame) yielded different results for instruments, needles, sutures and silicone surface: Neither instruments, needles nor annuloplasty rings were erroneously added. One generated instrument was classified as ‘not preserved’, since it was partially coalesced with the valve and two (out of six) needles could not be seen in the generated images. Green sutures were better preserved (4.0 of 4.9 sutures per frame) compared to white sutures (2.2 of 4.9 sutures per frame). In 14 frames all green sutures were consistent in both source and target domain, whereas there is no such frame for white sutures. The appearance of the generated frames was evaluated to be ‘overall realistic’ in 82.1%. The quality of the generated valves (shape and tissue texture) was compared to the silicone valve. The visual inspection yielded ‘valve differs completely’ in 2.6%, ‘good alignment but details differ’ in 10.3% and ‘good agreement’ in 87.2%.

## 5 Discussion

According to the widely accepted definition from Azuma [7], Augmented Reality (AR) “allows the user to see the real world, with virtual objects superimposed

<sup>3</sup> <https://youtu.be/qugAYpK-Z4M>.

*upon or composited with the real world. Therefore, AR supplements reality, rather than completely replacing it. Ideally, it would appear to the user that the virtual and real objects coexisted in the same space".* We consider *hyperrealism* as a subform of AR where real, but artificially looking objects (in our case the silicone valves) are transfigured to appear realistically (as in a real surgery). Nothing is added to the scene of the real world, it is just altered to appear more realistic, thus the term *hyperrealistic*. Objects that already appear realistic ideally stay the same (in our case the instruments, sutures, needles).

The idea to use a transformer network to translate a real endoscopic image into a synthetic-like virtual image has been assessed before with the overall aim of obtaining a reconstructed topography [8,9]. We focus on the opposite transformation, synthesizing intraoperative images from real training procedures on patient-specific silicone models. Our scenes are more complex, since they contain e.g. blood and lens contamination in the target domain and moving instruments, sutures and needles in both source and target domains.

Our methodological advancement tempCycleGAN shows a substantially stabilized composition of the synthesized frames in comparison to the original CycleGAN approach. The architecture's extension by two temporal discriminators, temporal paired input frames fed into multiple runs of generators and further L1 distances in the loss function to penetrate inconsistency yields such significantly more stable results. Beyond that, tempCycleGAN reduced the number of artifacts in the reported outcomes, while slightly sacrificing image sharpness. To the best of our knowledge, our approach is the first method for unpaired image-to-image translation addressing the problem of temporal inconsistencies in moving sequences.

**Acknowledgements.** The authors thank Bernhard Preim for his valuable hints and Benjamin Hatscher for making the illustration in Fig. 1a. The work was supported by DFG grant DE 2131/2-1, EN 1197/2-1. The GPU was donated by NVidia small scale grant.

## References

1. Kenngott, H.G., Wünsch, J.J., Wagner, M., et al.: OpenHELP (Heidelberg Laparoscopy Phantom): development of an open-source surgical evaluation and training tool. *Surg. Endosc.* **29**(11), 3338–3347 (2015)
2. Engelhardt, S., et al.: Elastic mitral valve silicone replica made from 3d-printable molds offer advanced surgical training. *Bildverarbeitung für die Medizin* 2018. I, pp. 74–79. Springer, Heidelberg (2018). [https://doi.org/10.1007/978-3-662-56537-7\\_33](https://doi.org/10.1007/978-3-662-56537-7_33)
3. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1125–1134, October 2017
4. Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: unsupervised dual learning for image-to-image translation. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2868–2876, October 2017
5. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *IEEE International Conference on Computer Vision (ICCV)* 2017, pp. 2242–2251 (2017)



6. Milgram, P., Kishino, F.: A taxonomy of mixed reality visual displays. *IEICE Trans. Inf. Syst.* **77**(12), 1321–1329 (1994)
7. Azuma, R.T.: A survey of augmented reality. *Presence Teleoper Virtual Environ.* **6**(4), 355–385 (1997)
8. Visentini-Scarzanella, M., Sugiura, T., Kaneko, T., Koto, S.: Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. *Int. J. Comput. Assist. Radiol. Surg.* **12**(7), 1089–1099 (2017)
9. Mahmood, F., Durr, N.: Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy, arXiv preprint: [arXiv:1710.11216](https://arxiv.org/abs/1710.11216) (2017)