Real-time Style Transfer for Videos to Enhance the Realism of Simulation of Laparoscopic Surgeries

1st Marine Shao
Centre For Print Research
University of the West of England
Bristol, United-Kingdom
marine.shao@uwe.ac.uk

2nd James Clark
Royal Cornwall Hospitals NHS Trust
Truro, United-Kingdom

3rd David Huson
Centre For Print Research
University of the West of England
Bristol, United-Kingdom

4th Jon Hardeberg
Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway

***Abstract*—Surgical simulation has repeatedly proven its potential, but is limited by the lack of realism to the surgical experience. For laparoscopic surgery simulation, an image processing method known as style transfer can improve the realism by transferring the style of a picture from surgery onto the video of the simulator using a stylisation network. In this article, we propose an adjustable style transfer algorithm for videos to improve the realism of silicone based models. The results show that our method can successfully implement multiple stylisations in real-time onto the video of the simulator while maintaining temporal consistency and overall smoothness of the video. Comparing to other style-transfer methods, our technique can offer multiple stylisations and, except from the method of Johnson *et al.* [6], achieves better realness score when evaluated by surgeons. This method can also be trained on a generic database in only 3h43 contrary to the other image processing techniques such as image translation which require specific training datasets.**

***Index Terms*—Surgical simulation, Laparoscopy, Style transfer, Virtual Reality**

I. Introduction

Outside of the operating theatre, surgeons have traditionally trained using the animal or human cadaveric model [1]. This has raised a significant number of issues not only relating to the clear ethical issues but also the expense and access. Neither model however is ideal for training, with limitations in anatomy in the case of the animal model, or pliability of the tissues in the case of the human cadaveric model in part caused by the embalming process [2].

Surgical simulation aims to provide surgeons with a model upon which they can train in any environment at any time. These simulators have proven their potential by showing that surgeons can improve their performances after using them [3]. The two main types of surgical simulators available are virtual-reality based or physical simulators. Virtual-reality simulators can provide visual realism and training for complex surgeries; however, they often lack tactile feedback and are expensive. Physical simulators are generally simpler, as such they tend to aim at the novice trainee. They can provide tactile feedback but are not very realistic visually and are also often single-use only [4].

The laparoscopic procedure is a surgical technique using instruments controlled by the surgeon guided by frames captured from a video endoscope inserted inside the patient. In this article, we propose to improve the simulation of laparoscopic procedures by combining the assets of physical simulation and virtual-reality. The approach is to offer a physical model that provides tactile feedback while improving the visual realism by performing style transfer on the video of the surgery.

This paper first summarises the related work on style transfer then presents the proposed method to enhance visual realism of surgical simulation. The following sections provide a description of the experiments and of the results.

II. Related Work

A. Style Transfer

Style transfer was first described by Gatys in 2016 with a neural algorithm that could implement an artistic style onto an image [5]. The idea of the method was to feed a style image and a content image into an optimisation algorithm which aimed to create an output image with the content of the content image and the style of the style image. The optimisation minimises the difference, or loss, of content and of style between the two images to create this output.

Gatys' algorithm uses an optimisation method to create the output image from a white-noise image. The optimisation process is time-consuming and can process only one content image at a time. Previous research developed quicker solutions. Johnson *et al.* [6] proposed to train a feed-forward Convolutional Neural Network on a dataset of content images with one style image; during the training, the optimisation of the content and style losses is back-propagated on the parameters of the network. This method can replace the long optimisation and enables real-time stylisation of multiple content images;

Funded by the Horizon 2020 programme of the European Union. Grant number 814158 (ApPEARS).

978-1-6654-6623-3/22/$31.00 © 2022 IEEE

which means that the processing time is inferior to 40ms to ensure smooth stylisation with at least 25 frames per seconds.

The stylisation of a video is possible with the previous method by implementing the algorithm on each frame. However, with the algorithm being designed to stylise single images, it does not take into account temporal consistency. Huang *et al.* [7] proposed a method to stylise videos in real-time without temporal fluctuations by calculating the temporal loss, a loss term that represents the temporal inconsistencies. The method also implements a Total Variation (tv) loss to limit the spatial inconsistencies by comparing neighbouring pixels.

B. Image processing and surgical simulation

Image processing has been used in surgical simulation before. Luengo *et al.* [8] improved the realism of virtual-reality based training of eye surgeries. In this study, Luengo *et al.* [8] modifies the style of the simulator by implementing different styles from a video of a surgery. The strength of the algorithm is that it can implement more than one style on the image and can change the style selection for different parts of the image.

Another example is the implementation of Engelhardt *et al.* [9], which is based on Generative Adversarial Networks (GANs). Their method can improve the style of physical simulators made of silicone. Their implementation of tempCycleGAN included a temporal consistency factor which allows fluid stylisation of videos. To avoid artefacts in the generated images, they also developed cross-domain conditional GANs which generates more consistent and realistic outputs [20]. GANs have also shown potential to improve the realism of virtual simulators with the aim of creating datasets of surgical images for the training of neural networks [18].

C. Adjustable style transfer

One of the drawbacks from the style transfer techniques is that they use weights predefined before the training of the algorithm. To explore the results from varying weights, a new neural network needs to be trained each time which is time-consuming. Babaeizadeh and Ghiasi [10] developed an adjustable style transfer method where two networks are trained at the same time to offer the possibility of changing the weights after the training. During the training of the algorithm, the weights are variable inputs instead of fixed parameters.

III. PROPOSED METHOD

The goal of this article is to enable real touch of instruments with a physical simulator, but with control over the visual appearance through an adjustable style transfer method. The adjustable style transfer method must be able to offer both multiple stylisation with only one training session of the algorithm and the possibility of stylising videos.

With the method of Babaeizadeh and Ghiasi [10], it is possible to stylise an image with multiple stylisations by adjusting the hyper-parameters after training of the algorithm; however, their algorithm is trained on images and not on videos. For this reason, it does not provide temporal stability between frames when stylising a video. Temporal stability is an important feature for our application, because the surgeons will be training on videos of the surgical simulator. The method from Huang *et al.* [7] provides temporal stability; however, it can achieve only one stylisation after the training of the algorithm. Having one stylisation only is limiting because it requires to restart the optimisation process to achieve other stylisations which is time-consuming. Furthermore, different surgeons might not agree on which stylisation is the most realistic, which generates the need to be able to provide quickly multiple types of stylisation. At last, having multiple stylisations also allows for variety in between different training sessions on the surgical simulator which makes it more appealing and challenging to surgeons.

A. Loss Networks

The loss network includes the content loss, the style loss, and the temporal loss to ensure that our method can perform the stylisation of videos while maintaining temporal consistency between frames.

1) Content Loss: The content loss can asset the difference of content between the content image C and the output image O. It is based on the filters of the different layers (l) of the VGG19 network. The size of the filter depends on the layer where it is situated within the neural network. Each filter targets a special feature inside of the image at different scales. If we consider F_{ij} the contribution of the i^{th} filter situated in the layer l on the position j of the image, then the following function calculates the difference of content between the output image and the content image:

$$L_{content}(C, O, l) = \frac{1}{2}\sum_{ij}(F_{ij}(C) - F_{ij}(O)) \tag{1}$$

2) Style Loss: The Gram matrix is a tool defined to capture the style of an image; it is the covariance of the contribution of the different filters at a given layer *l*:

$$G_{ij}(O) = \sum_{k} F_{ik}(O)F_{jk}(O) \tag{2}$$

The following function calculates the loss of style between the generated image and the style image S at a layer l by comparing their Gram matrices:

$$L_{style}(S, C, O) = \sum_{ij}(G_{ij}(S) - G_{ij}(O)) \tag{3}$$

3) Temporal Loss: The temporal loss is defined as:

$$L_{temporal}(O^t, O^{t-1}) = \frac{1}{D}\sum_{i=[1,D]} c_i(O_i^t - f(O_i^{t-1})), \tag{4}$$

where *t* is the time of the frame, *D* is the dimension of the output calculated by $D = H \times W \times N$, where *H* and *W* are the dimensions of the output and *N* is the number of channels, and *f* is a function that warps the stylised output at time *t-1* to time *t* according to the optical flow field that was estimated between the content images at time *t-1* and time *t*. The parameter *c* is between 0 and 1 and defines the per-pixel confidence of the optical flow.

4) Total Loss: The total loss is defined as:

$$\begin{aligned} L_{total}(S,C,O,t) =& \alpha_{temporal} L_{temporal}(O^t, O^{t-1}) \\ &+ \sum_l (\alpha_{style}(l) L_{style}(S,C,O) \\ &+ \alpha_{content}(l) L_{content}(C,O,l)), \end{aligned} \tag{5}$$

where $\alpha_{content}$ and α_{style} are parameters depending on the number of layers. The optimisation algorithm minimises the total loss to create the final output image. The contributions of the losses are weighted by $\alpha_{content}$, α_{style}, and $\alpha_{temporal}$. The weights define the relative importance of the content vs. the style in the final stylisation. If $\alpha_{content}$ is significantly higher than α_{style}, then the stylised image will be very similar to the input image with small style variation; if $\alpha_{content}$ is very small comparing to α_{style} then the style variation will be very strong and the content might vary a lot from the initial image.

B. Real-time style transfer with adjustable loss

To be able to modify the weights after training, a conditioner network is implemented following the same model as Babaeizadeh and Ghiasi [10]. Using this method, the weights are no longer pre-defined parameters of the network, but inputs that are changeable after training. This results in three inputs which are the content image, the style image, and the weights.

To understand the impact of the weights on the stylisation, we use the same technique as Babaeizadeh and Ghiasi [10] which is conditional instance normalisation. This method implements a conditioner network in addition to the stylisation network; it can condition the activation of the stylisation network with the weight inputs $\alpha = [\alpha_{content}, \alpha_{style}]$ to achieve a normalised activation z instead of the standard activation x. Normalizing the stylisation allows to adjust the weight after training by using them as an additional input parameter.

$$z = \gamma_\alpha \frac{x - \mu}{\sigma} + \beta_\alpha, \tag{6}$$

where μ and σ are the mean and standard deviations of the activation at the l^{th} layer across spatial axes. γ_α and β_α are the learned mean and standard deviations; they are calculated with the conditioner network. The architecture of the network is shown in Fig. 1.

We can note that $\alpha_{temporal}$ is not included in the conditioner network because it only ensures the temporal consistency and not the stylisation; for this reason, it is not interesting to modify its impact after training. The other weights $\alpha_{content}$ and α_{style} are four-dimensional vectors, which include one component for each layer of the VGG19 network.

IV. EXPERIMENTS

A. Implementation details

We used the same architecture as Babaeizadeh and Ghiasi [10] for the networks. The method is tested by transferring the style of frames from videos of real surgeries [11], [12] onto a silicone based simulator. Both networks were trained with these frames as the stylisation images and using the DAVIS

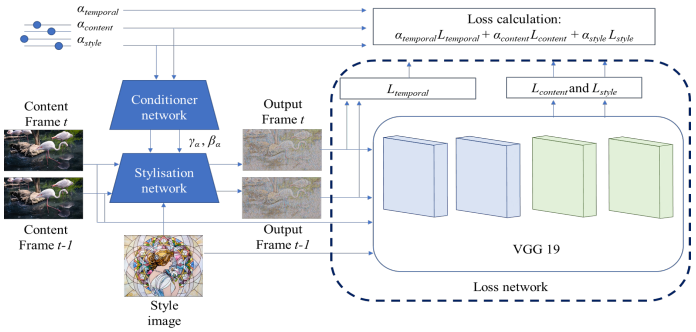

Fig. 1. Architecture of the model. It consists of three parts: a stylising network, a conditioner network, and a loss network. The weight vector α is passed through the conditioner network which will calculate γ_α and β_α. The two stylised frames are passed through the trained image classifier VGG19 to calculate L_{style} and $L_{content}$. The two stylised frames are also compared to calculate $L_{temporal}$. The stylisation network and the conditioner network are jointly trained by minimising this weighted sum of the different losses.

video database for the content videos [13]. This database includes 150 videos of 4.14 GB. At each iteration of the training, the components of α are randomly selected between 0 and 1; then are fed from the conditioner network to the stylisation network; the aim of this random selection is to train the algorithm to perform stylisation with any value of α. The training is conducted on Anaconda Python 3.7 using a Pytorch implementation on a NVIDIA RTX 3090 GPU.

The implementation can process a 480x640 frame in 28 milliseconds. The optimization uses the Adam stochastic gradient descent with a learning rate of 10^3 [14]. To ensure that each loss term is in the same range of order during the optimisation, additional parameters were added into the loss calculation to equilibrate their contributions. These parameters are selected empirically. The batch size is 30 and the number of epochs is 20. Using more epochs does not decrease the overall loss; decreasing the batch size leads to a less smooth output.

B. Creation of a physical model

The silicone simulator was created by designing moulds of the soft tissues using the software Rhinoceros (Robert McNeel & Associates, Washington, USA), and printing them on a Flashforge Creator Pro 3D printer (Flashforge, Zhejiang, China). Organ replicas are made by pouring silicone rubber (Smooth-On Inc., Pennsylvania, USA) into the moulds. The mould mimics the bile duct, the gallbladder, and the liver to train surgeons for a laparoscopic bile duct exploration. An endoscopic camera mimics the choledochoscope. The camera is connected to the computer where the recorded images go through the style transfer algorithm.

C. Evaluation of the outcome

The optimised networks are evaluated on pictures and videos of the physical model, and on a video of the virtual reality simulator LapSim (Surgical Science, Gothenburg, Sweden). In this validation phase, different stylisations are created by varying the component of α between 0 and 1. These

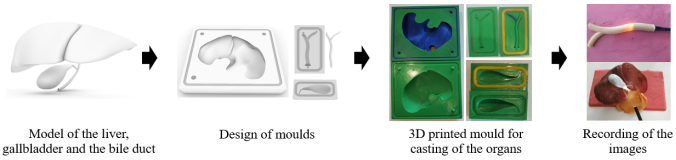

Fig. 2. Steps to create a simulator and to record the images.

stylisations are applied to videos and images of the simulator and are evaluated by the surgeons.

The quantitative evaluation includes a comparison of the scores of the algorithm with other methods from the state-of-the-art; the scores include the training time, the processing time, the optimisation of the different losses, and the temporal consistency. The temporal consistency is tested on two following frames according to two evaluation metrics.

The first evaluation metric evaluates the percentage of difference between two following frames Δ. To do so, it uses a difference score able to compare two images I_1 and I_2, and defined using the following procedure:

1. The difference map between the images I_1 and I_2 is calculated as the absolute value of the pixel-by-pixel difference between the images; it is then converted to a grayscale image,

2. A difference score is attributed to the difference map; this difference score is calculated using the histogram of the image. Histograms are frequency distribution of the intensity values that occur in the image; for instance, *h(i)* defines the number of pixels in the image with the intensity value *i*. The difference score is defined as follow:

$$\text{difference}(I_1,\, I_2) = \sum_i h(i) \times i, \tag{7}$$

where h is the histogram of the difference map calculated between I_1 and I_2. Then, Δ is defined by:

$$\Delta = \frac{\text{difference(two following frames)}}{\text{difference(black image, white image)}} \times 100 \tag{8}$$

The second evaluation metric is the temporal stability [15]; it calculates the temporal inconsistencies between two following frames using the temporal loss function defined during the training of the algorithm in equation 4.

The results were evaluated by surgeons through an online survey. Consent was obtained from each participant. The participants were shown the initial content images and the processed images using our method and methods from the state-of-the-art. The initial content images are images as well as videos from the physical simulator and from the LapSim simulator. The order of the images was randomized so that the participants did not know which method they were evaluating. The surgeons are asked to grade the outputs according to the "realness score" proposed by Yi *et al.* [16]. The assessment ranges from 0 (totally missing), 1 (bad), 2 (acceptable), 3 (good), to 4 (compelling); it evaluates the quality of the images, the realism of the colors, and the fluidity of the videos.

The impact of the stylisation vector is also evaluated by analysing images where all weights but one are set to zero.

V. Results

Our method was implemented and compared to the methods from Babaeizadeh and Ghiasi [10], Huang *et al.* [7], and Johnson *et al.* [6]. The methods were implemented using two style images from surgery, one from the view on the gallbladder and one from the view inside of the bile duct. The content images are the DAVIS database for our method and for the method of Huang *et al.* [7] and COCO database for the other methods [17].

A. Quantitative evaluation

Table 1 and Fig. 3 show that training time and generating time from the method of Huang *et al.* [7] are significantly longer; however, the implementation is on TensorFlow on CPU and not on Pytorch on GPU. The generating time using the method of Babaeizadeh and Ghiasi [10] and our method is significantly longer than the method of Johnson *et al.* [6] which could be explain by the utilisation of two networks; however, the generating time for the three methods is below 40ms which allows for real-time stylisation of the videos for all the methods. Using the first temporal metric, our method generates significantly smoother videos than methods with no temporal loss optimisation; however, using the second temporal metric the results are not statistically significant, but the average of our method is lower than the average of the method of Babaeizadeh and Ghiasi [10].

TABLE I
COMPARISON OF THE SCORES BETWEEN OUR METHOD AND THE METHODS FROM THE STATE-OF-THE-ART; THE GENERATING TIME AND TEMPORAL CONSISTENCIES ARE EVALUATED ON EACH FRAME OF A SEQUENCE OF 209 FRAMES OF 480X640 PIXELS.

Method	Babaeizadeh and Ghiasi [10]	Huang *et al.* [7]	Johnson *et al.* [6]	Ours
Training time	3h47	8h09	1h11	3h43
Number of stylisations	∞	1	1	∞
Generating time (ms)	27.1 ± 0.2	328.2 ± 0.4	21.3 ± 0.1	27.9 ± 0.1
Temporal metric 1 (%)	1.7 ± 0.1	1.2 ± 0.1	1.9 ± 0.1	1.5 ± 0.1
Temporal metric 2 (x1000)	54.4 ± 15.7	43.6 ± 25.8	40.1 ± 28.0	42.9 ± 23.8

The impacts of the stylisation vector and of the style image on the processing time are also evaluated by stylising a video using two style images and 14 types of stylisation vectors for each style image; the results show no significant impact.

B. Qualitative evaluation

A qualitative evaluation of the results shows that the algorithm can successfully implement different stylisations in real-time onto the video of the simulator while maintaining temporal consistency and overall smoothness of the image.

Eight expert surgeons assessed the "realness score" of the images and videos. Table 2 shows that all methods improve the realism. The average "realness score" is higher with our method than with the other methods except from the one of

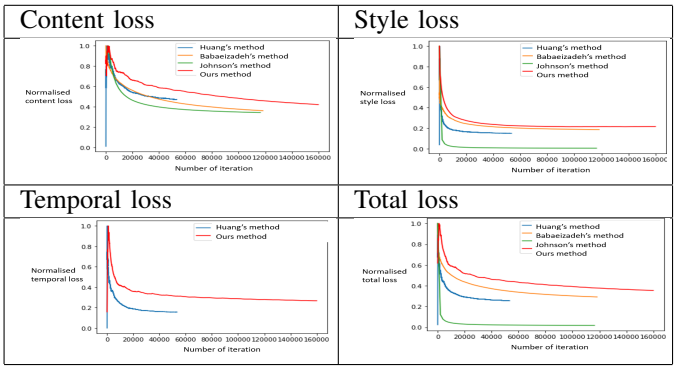

Fig. 3. Comparison of the optimisation of the normalised losses between our method and the methods from the state-of-the-art.

Johnson *et al.* [6]. The high score from the method of Johnson *et al.* [6] could be due to a stronger stylisation. The evaluation of the realism conducted by the surgeons seems to correlate to the second temporal evaluation metric; which shows the importance of the smoothness of the video on realism.

One surgeon commented that the colours from the different images are all acceptable because they represent the variation one sees *in vivo*. This illustrates the benefit of the adjustable style transfer that can offer the variety of real-life. This is also visible when surgeons gave similar scores to videos stylised by our algorithm using different stylisation vectors, resulting in different stylisations. Furthermore, the surgeons also made comments such as that two images are good but the best would be something in between; with this method, it is possible to adjust the outcome to this kind of comments by modifying α.

TABLE II
COMPARISON OF THE "REALNESS SCORES" BETWEEN OUR METHOD AND METHODS FROM THE STATE-OF-THE-ART.

Method	Intial	Babaeizadeh and Ghiasi [10]	Huang *et al.* [7]	Johnson *et al.* [6]	Ours
Average "Realness score"	1.5	1.6	1.8	2.2	2.1

Fig. 4 shows that each layer changes different features of the image; the last layer generates more contrast between the features making the smaller details such as the blood vessels more visible, while the first layer generates a smoother texture. Each layer creates a different type of stylisation; however, there is not one stylisation better than the others, each surgeon might find different results more appealing. The choice of the stylisation vector also has an impact on the colours.

The texture of the soft tissues is sometimes blurry and presents a pronounced texture which does not seem very realistic compared to the smooth aspect of real tissues. The choice of style image could explain this limitation; on the style image there is no large neat and smooth area, which could prevent the algorithm to learn the stylisation of smooth surfaces. Furthermore, the light is reflected on multiple small points in the style image, generating a lot of contrast on a small scale; the algorithm could interpret that as a part of the style of the image and try to recreate this high level of contrast on the output image, resulting in a less convincing stylisation.

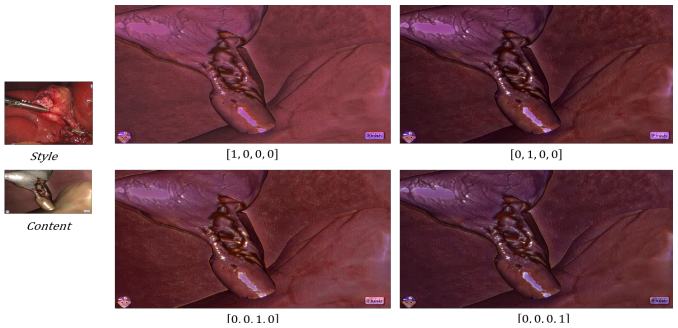

Fig. 4. Results from adjusting the input vector α in real-time after training. The targeted weight α_i is set to 1, while maintaining the others at 0. Each image differs in style which is visible with the variations of colours and contrast; deeper layers highlight more details.

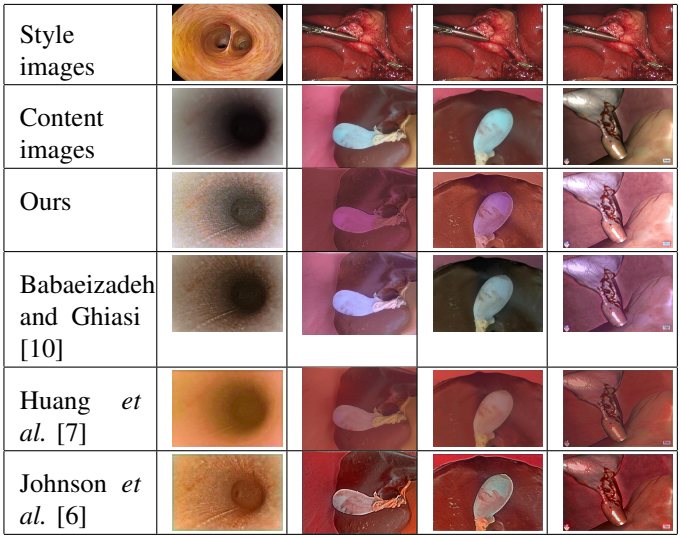

Fig. 5. Comparison of the generated images between our method and the methods from the state-of-the-art for different sets of stylisation and content images. The first row shows the style image, the second row the content images, the following four rows show the results from our method and from the methods from Babaeizadeh and Ghiasi [10], Huang *et al.* [7], and Johnson *et al.* [6] respectively. Each row displays a different stylisation.

The generated images from the different algorithms in Fig. 5 display the variety of stylisations possible. The method of Huang *et al.* [7] creates smoother results, which can be explained by the tv-loss term, while the method of Johnson *et al.* [6] creates a more pronounced stylisation, as visible on the gallbladder with pronounced blood vessels, this can also be useful to recreate the details of real-life tissues; however, both results depends on the hyper-parameters chosen before training and the two algorithms could generate very different results. The method of Babaeizadeh and Ghiasi [10] and our method can generate multiple stylisations; however, the texture is less convincing than the stylisation of Huang *et al.* [7].

In our method, the temporal loss is only evaluated between frames at time *t* and *t-1*, which can lead to inconsistencies between views during the training. Previous work managed to achieve global temporal consistency across views by adding a neural rendering approach to the image translation [19].

Comparing to other image processing techniques used in surgical simulation, our method has the advantages of being quick and easy to implement and of being able to offer diversified stylisations. The method of Engelhardt *et al.* [9] can stylise videos with a good level of realism; in an evaluation among surgeons, they achieved an average realness score of 3.3 using the same scale as ours. However, the method is based on image-to-image translation which requires to train on specific databases including images from the simulator and images from surgery; while our method, based on style transfer, aims to implement a style onto an image while preserving its content. For this reason, the type of content of the training images does not matter and a generic database allows to stylise any type of images.

The stylisation technique from Luengo *et al.* [8] is too complex to be used in real time and does not include a temporal loss, both factors are limiting its use in simulation; however, Luengo *et al.* [8] managed to implement multiple stylisation within one frame, depending on the elements in the frame. This is valuable because the instruments require different stylisation from the tissues. With our method, the instruments are modified using the same stylisation as the tissues and no longer look as realistic. The inconsistencies of objects that are not soft tissues was also a limitation for Engelhardt *et al.*; a solution was to use landmark detection [21].

TABLE III
BENEFITS AND LIMITATIONS OF OUR METHOD COMPARING TO OTHER IMAGE PROCESSING METHODS USED FOR SURGICAL SIMULATION.

Method	Benefits	Limitations
Luengo *et al.* [8]	Multiple styles in one frame Generic dataset	Not real-time No temporal consistency
Engelhardt *et al.* [9]	Realism Temporal consistency	Specific training dataset
Ours	Adjustable style selection Real-time Generic dataset Temporal consistency	Realism One style per frame

VI. CONCLUSION

This paper aims to present a method to adjust the stylisation of videos in real-time. The targeted application is the enhancement of the videos of laparoscopic surgery simulators to improve the visual realism and the training experience of surgeons. The strengths of the method are the possibility to adjust the stylisation to the end-user preferences and the quick and easy implementation requiring only a generic database. The main limitation of the method is that it can only apply on type of stylisation per frame, which can lead to inconsistencies when there are instruments in the frame. Future work should focus on the possibility to adjust to the style selection.

REFERENCES

[1] A. E. Forte, S. Galvan, F. Manieri, F. Rodriguez y Baena, and D. Dini, "A composite hydrogel for brain tissue phantoms," *Mater. Des.*, vol. 112, pp. 227–238, 2016.

[2] A. Phillips, "Exploring surgical and clinical skills learning in postgraduate and undergraduates," Ph.D. dissertation, University of East Anglia, 2017.

[3] M. M. Maddox, A. Feibus, J. Liu, J. Wang, R. Thomas, and J. L. Silberstein, "3D-printed soft-tissue physical models of renal malignancies for individualized surgical simulation: a feasibility study," *J. Robot. Surg.*, vol. 12, pp. 27–33, 2018.

[4] R. M. Viglialoro, N. Esposito, S. Condino, F. Cutolo, S. Guadagni, M. Gesi, M. Ferrari, and V. Ferrari, "Augmented reality to improve surgical simulation: Lessons learned towards the design of a hybrid laparoscopic simulator for cholecystectomy," *IEEE. Trans. Biomed. Eng.*, vol. 66, pp. 2091–2104, 2019.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *J. Vis.*, vol. 16, p. 326, 2016.

[6] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, Cham, 2016, vol. 9906, pp. 694–711.

[7] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu, "Real-time neural style transfer for videos," in *2017 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7044–7052.

[8] I. Luengo, E. Flouty, P. Giataganas, P. Wisanuvej, J. Nehme, and D. Stoyanov, "Surreal: enhancing surgical simulation realism using style transfer," *CoRR*, vol. abs/1811.02946, 2018.

[9] S. Engelhardt, R. De Simone, P. M. Full, M. Karck, and I. Wolf, "Improving surgical training phantoms by hyperrealism: Deep unpaired image-to-image translation from real surgeries," *Lect. Notes Comput. Sci.*, p. 747–755, 2018.

[10] M. Babaeizadeh and G. Ghiasi, "Adjustable real-time style transfer," *CoRR*, vol. abs/1811.08560, 2018.

[11] T. Koshitani, "Direct cholangioscopy combined with double-balloon enteroscope-assisted endoscopic retrograde cholangiopancreatography," *World J. Gastroenterol.*, vol. 18, p. 3765, 2012.

[12] D. A. Sherwinter, S. R. Subramanian, L. S. Cummings, M. F. Malit, S. L. Fink, and H. L. Adler. (2020) Laparoscopic Cholecystectomy Technique on Medscape. [Online]. Available: https://emedicine.medscape.com/article/1582292-technique#c2

[13] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K. Maninis, and L. Van Gool, "The 2019 DAVIS challenge on (VOS): Unsupervised multi-object segmentation," 2019.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2014.

[15] W.S. Lai, J.B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.H. Yang, "Learning blind video temporal consistency," in *Proceedings of the European Conference on Computer Vision 2018 (ECCV 2018)*, Cham, September 2018, vol. 11219.

[16] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2868-2876.

[17] T.Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Doll´ar, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014.*, Cham, 2015, vol 8693.

[18] M. Pfeiffer, *et al.*, "Generating Large Labeled Data Sets for Laparoscopic Image Processing Tasks Using Unpaired Image-to-Image Translation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019.*, Cham, 2019, vol 11768.

[19] D. Rivoir, M. Pfeiffer, R. Docea, F. Kolbinger, C. Riediger, J. Weitz, S. Speidel, "Long-Term Temporally Consistent Unpaired Video Translation From Simulated Surgical 3D Data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3343-3353.

[20] S. Engelhardt, L. Sharan, M. Karck, R. De Simone, I. Wolf, Cross-Domain Conditional Generative Adversarial Networks for Stereoscopic Hyperrealism in Surgical Training," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019.*, 2019, vol 11768.

[21] L. Sharan *et al.*, "Mutually Improved Endoscopic Image Synthesis and Landmark Detection in Unpaired Image-to-Image Translation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 127-138, Jan. 2022.