Long-Term Temporally Consistent Unpaired Video Translation from Simulated Surgical 3D Data

Dominik Rivoir[1,2], Micha Pfeiffer[1], Reuben Docea[1], Fiona Kolbinger[1,3],
Carina Riediger[3], Jürgen Weitz[2,3], Stefanie Speidel[1,2]

[1]NCT/UCC Dresden, Germany, [2]CeTI, TU Dresden, [3]University Hospital Dresden

{dominik.rivoir, micha.pfeiffer, reuben.docea, stefanie.speidel}@nct-dresden.de

{fiona.kolbinger, carina.riediger, juergen.weitz}@uniklinikum-dresden.de

Abstract

Research in unpaired video translation has mainly focused on short-term temporal consistency by conditioning on neighboring frames. However for transfer from simulated to photorealistic sequences, available information on the underlying geometry offers potential for achieving global consistency across views. We propose a novel approach which combines unpaired image translation with neural rendering to transfer simulated to photorealistic surgical abdominal scenes. By introducing global learnable textures and a lighting-invariant view-consistency loss, our method produces consistent translations of arbitrary views and thus enables long-term consistent video synthesis. We design and test our model to generate video sequences from minimally-invasive surgical abdominal scenes. Because labeled data is often limited in this domain, photorealistic data where ground truth information from the simulated domain is preserved is especially relevant. By extending existing image-based methods to view-consistent videos, we aim to impact the applicability of simulated training and evaluation environments for surgical applications. Code and data: http://opencas.dkfz.de/video-sim2real.

1. Introduction

One of the most promising applications of GAN-based image translation [14, 47] is the transfer from the simulated domain to realistic images as it presents great potential for applications in computer graphics. More importantly, unpaired translation [52] (*i.e.* no image correspondences between domains required during training) enables the generation of realistic data while preserving ground information from the simulated domain which would otherwise be difficult to obtain (*e.g.* depth maps, optical flow or semantic segmentation). This synthetic data can then facilitate training or evaluation in settings where labeled data is limited.

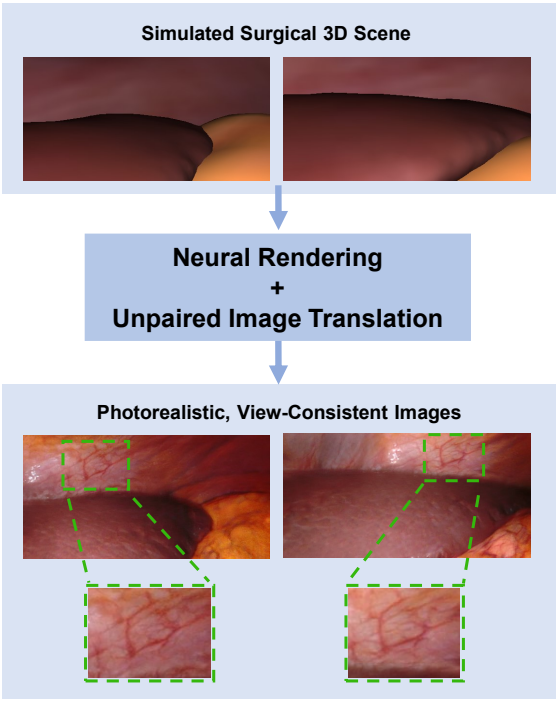

Figure 1. By combining unpaired image translation with a neural rendering approach, we produce photorealistic and view-consistent renderings of simulated surgical scenes. Note that fine details like vessels are rendered consistently across viewpoints although they were not manually modelled in the simulated domain.

The availability of realistic, synthetic data is especially crucial in the field of computer-assisted surgery (CAS) [26, 5]. CAS aims at providing assistance to the surgical team (*e.g.* visualizing target structures or prediction of complications) by means of analyzing available sensor data. In minimally-invasive surgery, where instruments and a camera are inserted into the patient's body through small ports, video is the predominant data source. Intelligent assis-

tance systems are especially relevant here, since performing surgery through small ports and limited view is extremely challenging. However, two major factors which currently limit the impact of deep learning in CAS are the lack of (a) labeled training data and (b) realistic environments for evaluation [25]. For instance, evaluating a SLAM (Simultaneous Localization and Mapping) algorithm [33, 43] on laparoscopic video data poses several problems since the patient's ground truth geometry is typically not accessible in the operating room (OR) and recreating artificial testing environments with realistic and diverse patient phantoms is extremely challenging. Other CAS applications which could benefit from temporally consistent synthetic training data include action recognition, warning systems, surgical navigation and robot-assisted interventions [26, 25].

Previous research has shown the effectiveness of synthetic, surgical images as training data for downstream tasks such as liver segmentation [37, 41]. However, their applications are still limited since many challenges in CAS include a temporal component. Using the previous example of evaluating a SLAM algorithm, realistic as well as temporally consistent video sequences would have to be generated in order to provide a useful evaluation environment.

Unpaired video translation has recently garnered interest in various non-surgical specialties [3, 10, 7, 9, 34, 51]. Most approaches thereby condition the generator on previous translated frames to achieve smooth transitions, *i.e.* short-term temporal consistency. However, they are fundamentally not designed for long-term consistency. Intuitively, when an object entirely leaves the field of view, consistent rendering cannot be ensured when it returns since the previous frame contains no information regarding the object's appearance. Even when the model is conditioned on multiple frames, the problem persists in longer sequences.

In the special case of translating from a simulated environment, however, the underlying geometry and camera trajectories are often available. Point correspondence between views are thus known and can be used to ensure globally consistent translations. The relatively new research area of neural rendering [45] aims at using the knowledge of the underlying 3D scene for image synthesis but has mainly been studied in supervised settings to date [45, 24, 42, 46, 32].

We propose a novel approach for unpaired video translation which utilizes the available information of the simulated domain's geometry to achieve long-term temporal consistency. A state-of-the-art image translation model is extended with a neural renderer which learns global texture representations. This way, information can be stored in 3D texture space and can be used by the translation module from different viewpoints. *I.e.* the model can learn the position of details such as vessels and render them consistently (Fig. 1). To ensure texture consistency, we introduce a lighting-invariant view-consistency loss. Furthermore, we employ methods to ensure that labels created in the simulated domain remain consistent when translating them to realistic images. We show experimentally that our final generated video sequences retain detailed visual features over long time distances and preserve label consistency as well as optical flow between frames.

2. Related Work

2.1. Unpaired Image and Video Translation

Image-based GANs [14, 38] have gathered much attention showing impressive results as unconditioned generative models [38, 6, 17, 18, 19] or in conditional settings such as image-to-image translation [47, 8, 36, 23]. However, their real-world applications are limited, since the content of generative models is difficult to control and supervised image translation requires corresponding image pairs, which are often not available. The introduction of unpaired translation through cycle consistency [52] hence widened their applicability and impact. Since then, several extensions have been proposed, *e.g.* shared content spaces [21], multimodality [20, 16], few-shot translation [22] or replacing cycle consistency with contrastive learning [35]. From an application standpoint, several works [37, 30, 41, 39, 29, 50] have shown the effectiveness of leveraging synthetic training data for surgical applications.

There have been several attempts at extending unpaired translation to videos where generated sequences have to be temporally smooth in addition to being realistic in individual frames [3, 10, 7, 9, 34, 51]. Bansal *et al.* [3] tackle this problem by introducing a temporal cycle consistency loss and Engelhardt *et al.* [10] use a temporal discriminator to model realistic transitions between frames. Several recent approaches estimate optical flow to ensure temporal consistency in consecutive frames [7, 9, 34, 51]. While there have been steady improvements in generating smooth transitions between frames, these models fail to capture long-term consistency. We aim to overcome this by adding a neural rendering component to our model. To our best knowledge, no successful solutions for long-term consistent video translation in the unpaired setting have been published to date.

2.2. Physically-grounded Neural Rendering

While unpaired visual translation methods are also sometimes categorized as neural rendering, the term most commonly refers to image synthesis approaches which incorporate knowledge of the underlying physical world [45]. By introducing differentiable components to rendering pipelines, neural representations of 3D shapes [32, 42, 53, 24], lighting [44, 31, 1, 2], textures [46] or view-dependent appearance [32] can be learned from image data for applications like novel view synthesis, facial re-enactment or relighting. Most closely related to our work, Thies *et al.* [46]

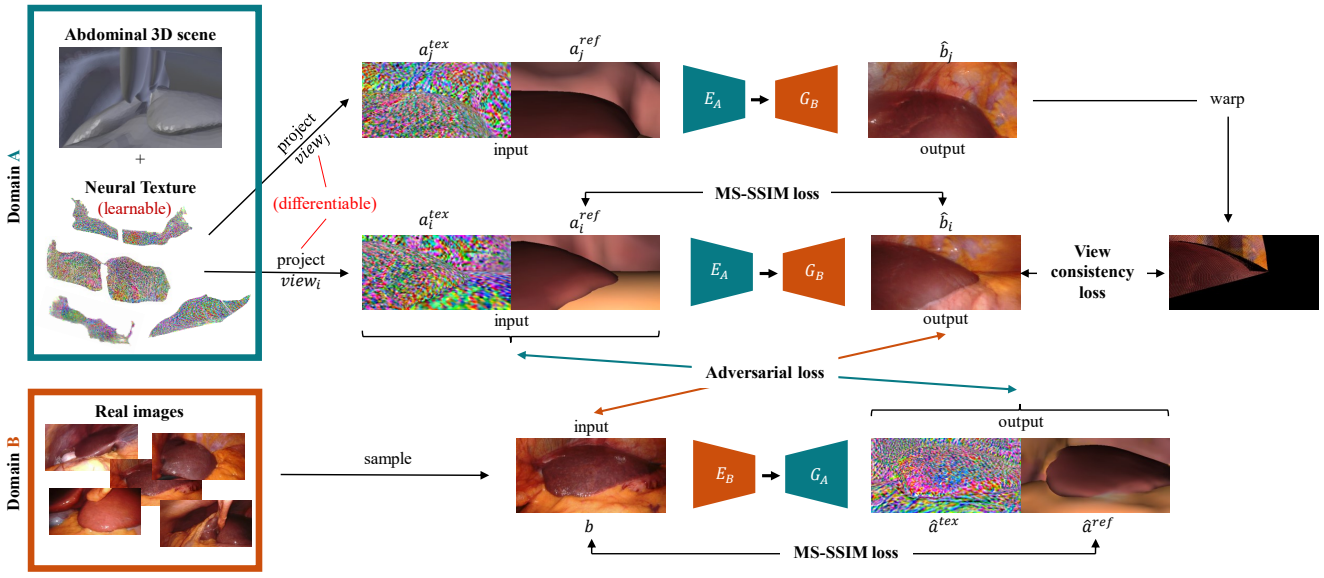

Figure 2. We combine unpaired image translation with neural rendering for view-consistent translation from simulated to photorealistic surgical videos. The model's key concept is a learnable, implicit representation of the scene's global texture. During training, texture features are projected into image space as a_i^{tex} which, combined with a simple rendering a_i^{ref}, serve as input to the unpaired image translation module. To encourage long-term temporally consistent translation, we warp two translated views into a common pixel space and employ our lighting-invariant consistency loss. Also note that the projected texture maps are part of the translation cycle, *i.e.* transfer from B to A includes the prediction of a reference image $\hat{a}^{ref}$ as well as a texture map $\hat{a}^{tex}$.

introduce a *deferred neural renderer* with *neural textures*, where implicit texture representations are learned from image sequences with a ground truth 3D model and camera poses. In contrast to their work, however, our model is built in an unsupervised setting where no correspondence between the simulated 3D data and real images is available. Finally, Alhaija *et al*. [2] propose a *deferred neural renderer* for unpaired translation from fixed albedo, normal and reflection maps to realistic output. However, since the texture representation is not learned, this work is more closely related to image-to-image translation.

Mallya *et al*. [27] propose a model for long-term consistent, paired video translation. They estimate information of the underlying physical world (depth, optical flow, semantic segmentation) to render globally consistent videos. This is currently the closest attempt at combining neural rendering with GAN-based translation. However, the paired data setting is a major hurdle for real-world applications.

We rather aim at bringing unpaired translation and neural rendering closer together. We believe that requiring knowledge of the simulated 3D geometry in an unpaired setting is often less restrictive than paired video translation, which requires rich ground truth in the real domain.

3. Method

We propose a method for unpaired, view-consistent translation from the domain of simulated surgical scenes A to the domain of realistic surgical images B. The method consists of three components: learnable neural textures, an unpaired image translation module and a lighting-invariant view-consistency loss (Fig. 2).

1) For a given viewpoint $view_i$ of the simulated scene, learnable features are projected from the neural texture tex to their pixel locations in the image plane and form a spatial feature map a_i^{tex}:

$$a_i^{tex} = project(tex, view_i) \tag{1}$$

2) Additionally, a simple but unrealistic rendering a_i^{ref} of the same view is used as a prior for translation. Combined, $(a_i^{tex}, a_i^{ref}) \in A$ serves as input to the unpaired image translation module to get the fake image $\hat{b}_i \in B$:

$$\hat{b}_i = translate_\theta(a_i^{ref}, a_i^{tex}) \tag{2}$$

Errors can be backpropagated into tex since $project(\cdot)$ is differentiable and enables the model to learn the global texture representation tex end to end with the network parameters θ of the translation module.

3) To ensure globally consistent rendering, pairs of translated views $\hat{b}_i, \hat{b}_j$ are sampled during training, warped into a common pixel space and constrained using our lighting-invariant view-consistency loss.

$$\hat{b}_i \xleftrightarrow[consist.]{view} \hat{b}_j \tag{3}$$

The main insight of our model is that neural textures allow the model to learn global information about the scene independent of time-point and view, *e.g.* material properties or locations of details such as vessels. After projecting texture features into the image plane, the translation module serves as a deferred renderer to synthesizing realistic images. Since the translation module operates on individual views, view-dependent effects such as specular reflections or changing lighting conditions are synthesized here. We jointly learn one neural texture for each of the 7 simulated scenes and common translation networks for all scenes.

3.1. Neural Texture & Projection Mechanism

For true long-term consistency, we require a method which can store information independent of time-point or view. To do so, we use a learnable, global texture tex, named *neural texture* by Thies et al. [46]. For each object in the scene (liver, gallbladder, ligament, abdominal wall and fat), tex contains learnable, spatial feature maps as an implicit texture representation. At each spatial location (texel), N features are learned and enable the model to learn consistent tissue properties or locations of details such as vessels. The shape of tex is $O \times P \times H \times W \times N$ with $O = 5$ objects, $P = 6$ projection planes of size $H \times W = 512 \times 512$ per object and $N = 3$ learnable texture features per texel.

To learn tex end-to-end with the translation module, we only require a differentiable projection mechanism (Eq. 1) which maps features from the global texture tex into the image plane for a given view $view_i$. The resulting image-sized feature map a_i^{tex} serves as an input to the translation module and thus errors can be propagated back into tex.

$$a_i^{tex}[x, y] = \sum_{x_p, y_p}^{tri(s)} (n_s^T \cdot n_p)^2 \cdot tex[o, p, x_p, y_p] \quad (4)$$

We define the projection into a_i^{tex} by means of ray casting [49], triplanar mapping [13] and bilinear interpolation [12] (Fig. 3 and Eq. 4). For each pixel (x, y), we cast a *ray* onto its $3D$ surface point $s \in \mathbb{R}^3$ in the scene and determine the object o it belongs to. The neural texture $tex[o]$ of an object consists of 6 axis-aligned texture planes surrounding the mesh. Through *triplanar mapping* $tri(s)$, we obtain one texture coordinate (x_p, y_p) for each of the three planes $p \in \{1..P\}$ which face s (Fig. 3). Texture features are weighted by the dot-product of plane and surface normal n_p, n_s to obtain the aggregated features $a_i^{tex}[x, y]$ in pixel space. Since texture planes are discrete grids, we use *bilinear interpolation* to obtain texture features $tex[o, p, x_p, y_p]$ from arbitrary, continuous locations. Hence, a total of 12 texels contribute to one pixel (4 discrete texels for each of the 3 plane coordinates). For details, see the supplementary. Note that triplanar mapping was chosen for its simplicity but could easily be replaced by other UV mappings.

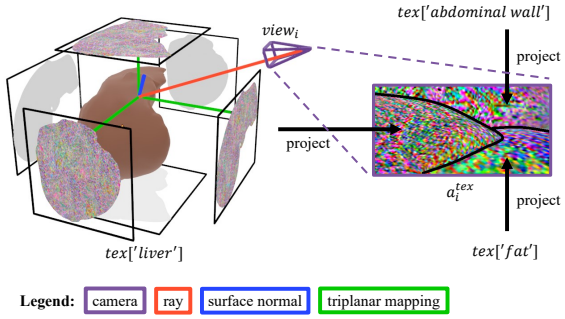

Figure 3. Neural texture projection. A ray (red) is cast for every pixel to the scenes's surface. Triplanar mapping (green) is used to map surface points to learnable texture planes (with bilinear interpolation in texture space). This differentiable mapping allows us to backpropagate errors from image space to global texture space.

3.2. Unpaired Image Translation Module

Our translation module is a deterministic, style-less variant of Pfeiffer et al.'s model [37], which itself is based on MUNIT [16]. The model enforces cycle consistency as well as a shared content space through interchangeable encoders E_A, E_B and decoders G_A, G_B for each domain [21].

Given a projected texture map and reference image $(a_i^{tex}, a_i^{ref}) \in A$, the encoder E_A extracts a domain-independent content code c^{a_i} and decoder G_B predicts a fake image $\hat{b}_i \in B$ from c^{a_i}. E_B then reconstructs content $c_{rec}^{a_i}$ and G_A translates back to domain A to complete the cycle. Additionally, the input is directly reconstructed through $(a_{i,rec}^{tex}, a_{i,rec}^{ref}) = G_A(E_A(a_i^{tex}, a_i^{ref}))$. Translation from B to A is done analogously. Finally, Multi-Scale Discriminators [47] D_A, D_B distinguish fake and real images.

$$\begin{pmatrix} a_i^{tex} \\ a_i^{ref} \end{pmatrix} \xrightarrow{E_A} c^{a_i} \xrightarrow{G_B} \hat{b}_i \xrightarrow{E_B} c_{rec}^{a_i} \xrightarrow{G_A} \begin{pmatrix} a_{i,cyc}^{tex} \\ a_{i,cyc}^{ref} \end{pmatrix} \quad (5)$$

We use the LS-GAN loss [28] as adversarial loss L_{adv}, and $L1$ losses for L_{cyc}, L_{rec}, L_c to ensure cycle consistency as well as image and content reconstruction. Finally, we enforce a Multi-Scale Structural-Similarity loss [48, 37] L_{ssim} on the brightness of a_i^{ref} and $\hat{b}_i$ as well as b and $\hat{a}^{ref}$ to encourage label-preserving translation. Details on networks and losses can be found in the supplementary.

$$L_{translation} = L_{adv} + L_{cyc} + L_{rec} + L_c + L_{ssim} \quad (6)$$

3.3. View Consistency Loss

To enforce view consistency, two random views i, j of the same simulated scene are sampled and translated during each training iteration. Using the knowledge about the scene's geometry, the second view is warped into the pixel space of the first view and consistent rendering is enforced through a pixel-wise view-consistency loss. In minimally-invasive surgery, however, the only source of light is a lamp

mounted on the camera. This results in changing light conditions whenever the field of view is adjusted and the image center typically being brighter than its surroundings. This poses an additional challenge for view-consistency. Therefore, we propose to minimize the angle between RGB vectors instead of a channel-wise loss. For a pair of translated views $\hat{b}_i, \hat{b}_j$, the loss is defined as

$$L_{vc} = \frac{1}{|M_{\hat{b}_i\hat{b}_j}|} \sum_{(x,y)}^{M_{\hat{b}_i\hat{b}_j}} cos^{-1}\left(\frac{\hat{b}_i^{xy} \cdot w_i(\hat{b}_j)^{xy}}{\|\hat{b}_i^{xy}\|\|w_i(\hat{b}_j)^{xy}\|}\right), \quad (7)$$

where $(x, y) \in M_{\hat{b}_1\hat{b}_2}$ are the pixel locations in $\hat{b}_i$ that have a matching pixel in $\hat{b}_j$. $\hat{b}_i^{xy}$ is the RGB vector at this location. $w_i(\cdot)$ is the warping operator into $\hat{b}_i$'s pixel space. Note that the angle between vectors u, v can be computed by $cos^{-1}((u \cdot v)/(\|u\|\|v\|))$. This enforces consistent hue in corresponding locations while allowing varying brightness.

$$L_{total} = L_{translation} + \lambda L_{vc} \quad (8)$$

Equation 8 shows the final loss function. To avoid an imbalance between domains A and B, $L_{translation}$ is not enforced on $\hat{b}_j$ and errors from L_{vc} are only backpropagated through $\hat{b}_i$ and not $\hat{b}_j$. λ is initialized with 0 and set to 20 after 10^4 training steps to avoid forcing consistency on unrefined translations in early stages of training. Complete training details can be found in the supplementary.

3.4. Data

For the domain of real images B, we collected 28 recordings of robotic, abdominal surgeries from the University Hospital Carl Gustav Carus Dresden and manually selected sequences which contain views of the liver. The institutional review board approved the usage of this data. Frames were extracted at 5fps, resulting in a total of 13,334 training images. During training, images are randomly resized and cropped to size 256x512.

For the simulated domain A, we built seven artificial abdominal 3D scenes in Blender containing liver, liver ligament, gallbladder, abdominal wall and fat/stomach. The liver meshes were taken from a public dataset (3D-IRCADb 01 data set, IRCAD, France) while all other structures were designed manually. For each scene, we generated 3,000 random views of size 256x512, resulting in a total of 21,000 training views. To evaluate temporal consistency, we manually created seven 20-second sequences at 5fps which pan over each scene with varying viewpoints and distances.

4. Experiments

To establish that our method produces both realistic and long-term consistent outputs, we need to evaluate the quality of individual images as well as consistency between consecutive or non-consecutive frames. Thus, we establish several baselines and evaluate them using various metrics. We place a special focus on both detailed and temporally consistent translation, since correct re-rendering of details such as vessels is crucial for obtaining realistic videos.

4.1. Baselines

SSIM-MUNIT: This is Pfeiffer *et al.*'s [37] model for surgical image translation trained on our dataset of real and synthetic images b and a^{ref}. It corresponds to our image translation module but with added styles and noise injected into generator input. We remove these components in our model since they are disadvantageous for view consistency.

ReCycle and SSIM-ReCycle: We compare to Bansal *et al.*'s unpaired video translation approach ReCycle-GAN [3] which is trained on triplets of consecutive video frames to maintain temporal consistency. We use the variant with additional non-temporal cycles (https://github.com/aayushbansal/Recycle-GAN). Additionally, we implement a variant with MS-SSIM loss for label preservation.

OF-UNIT: State-of-the-art unpaired video translation models condition the generator on translations from previous time-steps to ensure short-term temporal consistency. Many methods thereby warp the previous image by estimating optical flow (OF) and achieve incremental improvement through better OF estimation [7, 9, 34]. We argue, however, that even perfect OF is not enough for long-term consistency and can even have detrimental effects as we show later. To demonstrate this, we build a variant of our model which uses ground-truth OF to warp the previous translation, *i.e.* it can potentially produce perfect transitions between frames. We replace the input (a^{tex}, a^{ref}) of the encoder E_A with $(w(\hat{b}_{prev}), a^{ref})$, where $\hat{b}_{prev}$ is the generated frame of the previous time step and w is the perfect warping operator using ground-truth optical flow. Thus, *OF-UNIT* serves as an upper-bound for the state of the art in unpaired video translation, where OF has to be estimated and is therefore imperfect.

Ours w/o vc and Ours w/o tex, vc: Finally, we ablate our model by removing first the view-consistency loss and then also neural textures. The second model corresponds to SSIM-MUNIT without styles or noise in the generator.

4.2. Metrics

Realism: We compare the realism of models through the commonly used metrics *Frechet Inception Distance (FID)* [15] and *Kernel Inception Distance (KID)* [4] for which we sample 10,000 random images from the set of real and generated training images, each. Further, we train a U-Net variant for liver segmentation on a dataset of 405 laparoscopic images from 5 patients and report the *Dice* score when evaluated on all 21,000 generated images. This metric measures both realism and label preservation.

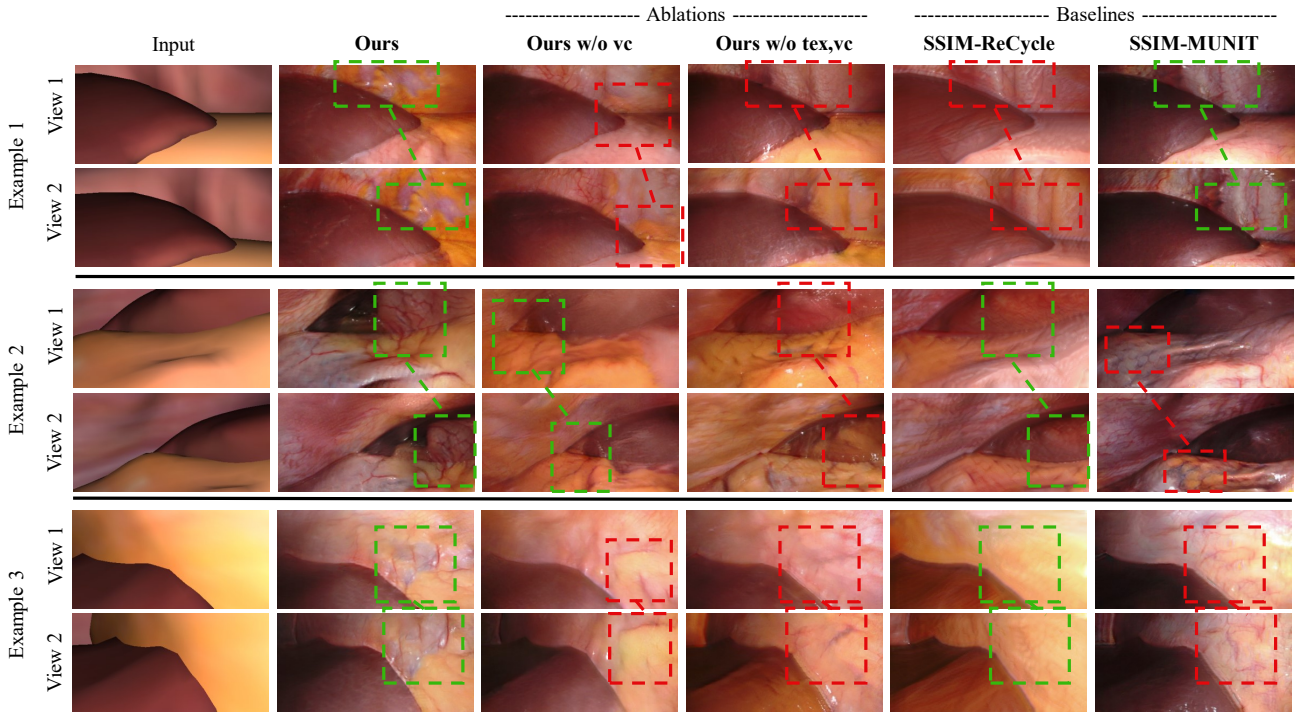

Figure 4. Qualitative comparison: View-consistent areas are marked green, inconsistent ones red. Our model can render fine-grained details consistently across views. *SSIM-ReCycle* often produces consistent outputs but lacks detail and realism. *SSIM-MUNIT* produces realistic but flickering results. Quality and consistency can best be judged in videos at http://opencas.dkfz.de/video-sim2real.

Method	Data	Realism			Temp. Consistency			
		FID ↓	KID ↓	Dice ↑ %	OF ↓	ORB-1 ↑ % (# per pair)	ORB-5 ↑ % (# per pair)	ORB-10 ↑ % (# per pair)
SSIM-MUNIT [37]	img	28.3	.0132	**59.2**	8.64	60.5% (32.5)	36.1% (15.7)	19.1% (7.0)
ReCycle [3]	vid	61.5	.0454	40.7	8.89	69.6% (16.3)	43.9% (7.0)	23.5% (2.7)
SSIM-ReCycle	vid	80.6	.0622	50.9	8.75	88.9% (13.3)	67.4% (6.2)	43.2% (2.8)
OF-UNIT	vid	**26.8**	.0125	57.7	8.53	**93.5%** (32.4)	59.0% (11.1)	30.7% (4.4)
OF-UNIT (revisit)	vid	-	-	-	8.91	69.9% (15.7)	43.8% (7.3)	24.7% (3.4)
Ours w/o tex,vc	img	27.3	**.0114**	56.8	*8.43*	81.7% (31.2)	51.3% (13.3)	29.5% (6.2)
Ours w/o vc	vid	27.0	.0134	55.2	8.35	88.3% (27.9)	66.8% (14.6)	44.5% (7.5)
Ours	vid	**26.8**	.0124	57.1	**7.62**	91.8% (**49.7**)	**73.0%** (**27.2**)	**49.6%** (**13.9**)

Table 1. Quantitative results with best scores printed **bold**. For metrics *ORB-1*, *ORB-5* and *ORB-10*, we report the accuracy of feature matches and the total number of correct matches per image pair, indicating both consistency as well as level of detail.

Temporal Consistency: We introduce two metrics to evaluate the temporal consistency of the sequences generated from each scene. Firstly, we measure the mean absolute error for the estimated optical flow *OF* of consecutive translated frames $\hat{b}_t, \hat{b}_{t+1}$ and their corresponding simulated reference images a_t^{ref}, a_{t+1}^{ref} by $mean(|OF(a_t^{ref}, a_{t+1}^{ref}) - OF_{GF}(\hat{b}_t, \hat{b}_{t+1})|)$ where $OF(a_t^{ref}, a_{t+1}^{ref})$ is the ground truth optical flow of the synthetic scene and $OF_{GF}(\hat{b}_t, \hat{b}_{t+1})$ is the optical flow estimated by the Gunnar-Farneback method [11] on the generated frames. As argued by Chu *et al.* [9], this is better than the more common RGB error on warped images, since the latter favors blurry sequences. Secondly, the metrics *ORB-1, ORB-5 and ORB-10* measure how consistently image features are rendered. For *ORB-1*, we compute all ORB feature [40] matches in consecutive frames and determine whether the matched feature points correspond the the same 3D location. We report the accuracy of matches as well as the average number of correct matches per image pair. A blurry but consistent sequence might yield a high accuracy, so the number of matches gives additional information on

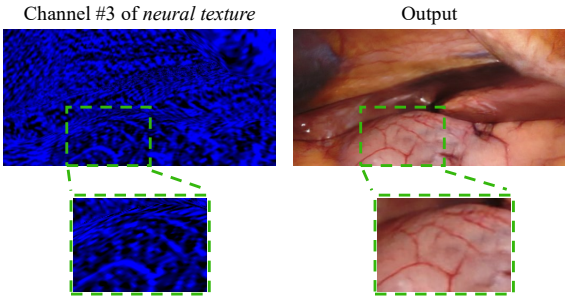

Figure 5. Details are stored in our *neural textures*. We found the 3rd feature channel often to correspond to vessels.

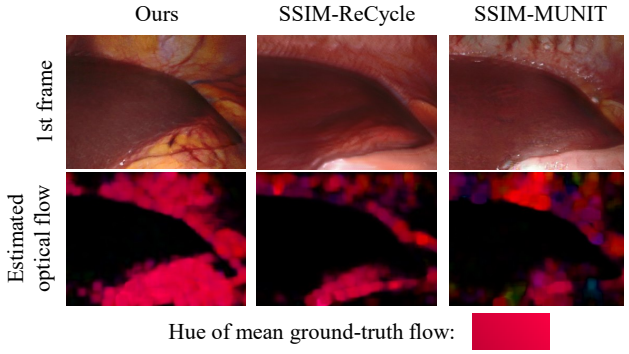

Figure 6. Estimated optical flow in a scene with camera motion (where hue indicates the direction of the flow). In our results, consistent motion is detected on textured surfaces while blurriness or flickering lead to poor flow estimates in other models.

how detailed the results are. A match is considered correct if its distance is smaller than 1mm in the underlying 3D scene. To investigate consistency beyond consecutive frames, we do the same with pairs that are 5 and 10 frames apart (*i.e.* 1 and 2 sec.) as *ORB-5* and *ORB-10*.

Time-independence: Finally, we show the pitfalls of previous approaches which condition on previous time steps. We extend each test sequence by running it first forward and then backward such that each view is visited twice with varying temporal distance. I.e. given a sequence $1, \dots, T$, we extend it to $1, \dots T, T, \dots, 1$ similar to Mallya *et al.*'s [27] evaluation. We then compute the same metrics *OF* and *ORB-1*. But instead of comparing frame t to its successor $t+1$, we use the time point in the extended sequence which corresponds to its successor, namely $2T - t$. For all methods except *OF-UNIT*, this is equivalent to the original metric since they depend only on the current view. Analogously for *ORB-5* and *ORB-10*, we compare to time points $2T - t - 4$ and $2T - t - 9$, respectively. We denote these experiments as *OF-UNIT (revisit)*.

5. Results

5.1. Realism

Table 1 shows our model achieves similar FID and KID scores as image-based approaches (*SSIM-UNIT* and *Ours w/o tex,vc*) while strongly outperforming video-based methods *ReCycle* and *SSIM-ReCycle*. We hypothesize that their temporal cycle-loss favors blurry images since they are easier to predict for the temporal prediction model. Fig. 4 supports this hypothesis as our and image-based models show more detailed and realistic translations than *SSIM-ReCycle*. For *OF-UNIT*, similar realism scores to ours are expected, since it uses the same translation module.

Further, we evaluate a pretrained liver segmentation network on the generated data. Again, our model yields comparable results to image-based methods while outperforming *ReCycle* and *SSIM-ReCycle*. This indicates that our results are not only realistic but content of the simulated domain is also translated correctly. The gap between *ReCycle* and *SSIM-ReCycle* additionally shows the importance of the MS-SSIM loss for label-preservation. Example 2 in Fig. 4 shows a failure case of our model where a stomach-like texture with vessels is rendered on the liver. Introducing *neural textures* supposedly improves the sharpness and level of detail in translations but increases the model's freedom to change content in the scene. The quantitative results, however, suggest that this is only a minor effect.

5.2. Temporal Consistency

Using the established ORB feature detector, we evaluate how consistently visual features are re-rendered in following frames of generated video sequences. We report how often detected feature matches are correct as well as the number of correct matches per pair of frames. For neighboring frames, our model achieves an accuracy of 91.8%, outperforming all baselines except *OF-UNIT*. However, this is not surprising since the latter uses the perfectly warped previous frame as input. For larger frame distances, however, our model outperforms *OF-UNIT*, showing its superiority w.r.t. long-term consistency. Additionally, the absolute number of correct matches per image pair is drastically higher than in *OF-UNIT* and other models even for neighboring frames. This indicates that our *neural textures* not only enable consistent translation but also encourage more detailed rendering. Fig. 4 shows several translated views with detailed as well as consistent textures. In Fig. 5, we show how the location of vessels is stored in the *neural texture*.

We observe that other methods fail to generate detailed as well as temporally consistent sequences. While *SSIM-MUNIT* produces detailed translations (indicated by the high number of matches), it achieves the lowest accuracies. Oppositely, video-based *ReCycle* and *SSIM-ReCycle* produce more consistent but less detailed renderings, indicated by their high accuracy but low number of correct matches.

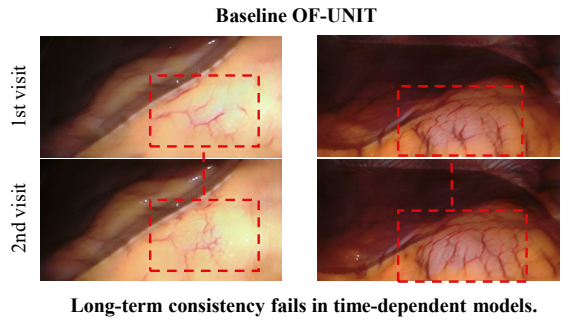

Figure 7. When revisiting a previous view, time-dependent models such as OF-UNIT fail to render textures consistently. Our model maintains consistency independently of the duration between visits by storing information in texture-space.

Note that *SSIM-MUNIT* induces flickering since noise is injected into the generator. Temporal consistency can already be strongly improved by removing this component (*Ours w/o tex,vc*). Adding *neural textures* without enforcing view-consistency (*Ours w/o vc*) further improves results.

Evaluating temporal consistency through optical flow (OF) supports our previous findings. This metric measures both temporal consistency as well as level of detail, since Gunnar-Farneback flow often fails on smooth surfaces. Image- and other video-based methods yield high errors, since the former tend to produce detailed but flickering sequences, while the latter often generate blurry but consistent views (Fig. 6). By learning textures in 3D space, our model achieves both detailed and consistent renderings.

5.3. Time-independence

We have seen that the time-dependent baseline *OF-UNIT* achieves very consistent transitions between frames and still achieves respectable results for larger frame distances. However, if the second frame is replaced with the same view revisited at a later point of the sequence, then performance drastically degrades. This is because the model does not have the capacity to *remember* the appearance of areas which have left the field of view. It even underperforms compared to its unconditioned variant *Ours w/o tex,vc*. We hypothesize that dependence on the previous trajectory actually encourages appearance changes over time (Fig 7). We believe time-independence is therefore an important feature for achieving long-term consistency, even in non-static scenes. With our approach, moving objects as well as deformations can potentially be handled by moving or deforming the neural texture accordingly.

5.4. Lighting-invariant View Consistency

We proposed an angle-based loss for view consistency which only keeps the hue of corresponding areas consistent. Fig. 8 shows that our angle loss allows for more realistic

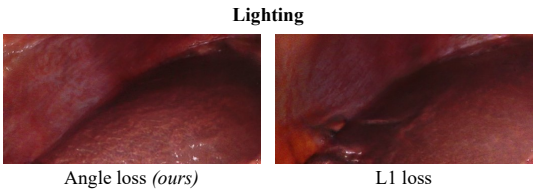

Figure 8. Our angle loss allows the translation module to adjust brightness of areas according to the current view. In real images, the center is often brightest since the light source is mounted on the camera.

lighting since the translation module can change brightness according to the current view. On the other hand, an L1 loss enforces static brightness from arbitrary viewpoints. This results in incorrect lighting like in the left image where the light appears to come from the bottom right. More examples can be found in the supplementary material.

6. Conclusion

We combine neural rendering with unpaired image translation from simulated to photorealistic videos. We target surgical applications where labeled data is often limited and realistic but simulated evaluation environments are especially relevant. Through extensive evaluation and comparison to related approaches, we show that our results maintain the realism of image-based approaches while outperforming video-based methods w.r.t. temporal consistency. We show that optical flow is consistent with the underlying simulated scene and that our model can render fine-grained details such as vessels consistently from different views. Also, data generation can easily be scaled up by adding more simulated scenes. A crucial observation about the model is that it leverages the rich information contained in the simulated domain while requiring only an unlabeled set of images on the real domain. This way, consistent and label-preserving data can be generated without limiting its relevance for real-world applications. Specifically, ground truth which would be unobtainable in surgical settings can be generated (*e.g.* depth, optical flow, point correspondences). This work is a step towards more expressive simulated environments for *e.g.* surgical assistance systems, robotic applications or training aspiring surgeons. While we focus on surgical applications (where access to labeled data is especially restricted), the model can potentially be used for any setting with a simulated base for translation.

Acknowledgements Funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Andreas Geiger, and Carsten Rother. Geometric image synthesis. In *Asian Conference on Computer Vision*, pages 85–100. Springer, 2018. 2

[2] Hassan Abu Alhaija, Siva Karthik Mustikovela, Justus Thies, Matthias Nießner, Andreas Geiger, and Carsten Rother. Intrinsic autoencoders for joint neural rendering and intrinsic image decomposition. In *International Conference on 3D Vision*, 2020. 2, 3

[3] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018. 2, 5, 6

[4] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 5

[5] Sebastian Bodenstedt, Martin Wagner, Beat Peter Müller-Stich, Jürgen Weitz, and Stefanie Speidel. Artificial intelligence-assisted surgery: Potential and challenges. *Visceral Medicine*, 36(6):450–455, 2020. 1

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 2

[7] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. Mocycle-gan: Unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 647–655, 2019. 2, 5

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[9] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020. 2, 5, 6

[10] Sandy Engelhardt, Raffaele De Simone, Peter M Full, Matthias Karck, and Ivo Wolf. Improving surgical training phantoms by hyperrealism: deep unpaired image-to-image translation from real surgeries. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 747–755. Springer, 2018. 2

[11] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003. 6

[12] Jerry D Gibson and Al Bovik. Handbook of image and video processing, 2000. 4

[13] Ben Golus. Normal mapping for a triplanar shader. https://bgolus.medium.com/normal-mapping-for-a-triplanar-shader-10bf39dca05a. Accessed: 2020-08-06. 4

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1, 2

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 5

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2, 4

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 4

[22] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019. 2

[23] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[24] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[25] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts to clinical translation. *arXiv preprint arXiv:2011.02284*, 2020. 2

[26] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017. 1, 2

[27] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 359–378, Cham, 2020. Springer International Publishing. 3, 7

[28] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 4

[29] Aldo Marzullo, Sara Moccia, Michele Catellani, Francesco Calimeri, and Elena De Momi. Towards realistic laparoscopic image generation using image-domain translation. *Computer Methods and Programs in Biomedicine*, 200:105834, 2021. 2

[30] Shawn Mathew, Saad Nadeem, Sruti Kumari, and Arie Kaufman. Augmenting colonoscopy using extended and directional cyclegan for lossy image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4696–4705, 2020. 2

[31] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[33] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[34] Kwanyong Park, Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Preserving semantic and temporal consistency for unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1248–1257, 2019. 2, 5

[35] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 2

[36] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[37] Micha Pfeiffer, Isabel Funke, Maria R Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, Tobias Roß, Matthew J Clarkson, Kurinchi Gurusamy, Brian R Davidson, et al. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019. 2, 4, 5, 6

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016. 2

[39] Anita Rau, PJ Eddie Edwards, Omer F Ahmad, Paul Riordan, Mirek Janatka, Laurence B Lovat, and Danail Stoyanov. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *International journal of computer assisted radiology and surgery*, 14(7):1167–1176, 2019. 2

[40] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 6

[41] Manish Sahu, Ronja Strömsdörfer, Anirban Mukhopadhyay, and Stefan Zachow. Endo-sim2real: Consistency learning-based domain adaptation for instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 784–794. Springer, 2020. 2

[42] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2

[43] Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. *IEEE Robotics and Automation Letters*, 3(4):4068–4075, 2018. 2

[44] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019. 2

[45] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020. 2

[46] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 4

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1, 2, 4

[48] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In

The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee, 2003. 4

[49] A Watt and Mark Watt. Advanced animation and rendering techniques: Theory and practice, 1992. 4

[50] Aji Resindra Widya, Yusuke Monno, Masatoshi Okutomi, Sho Suzuki, Takuji Gotoda, and Kenji Miki. Self-supervised monocular depth estimation in gastroendoscopy using gan-augmented images. In *Medical Imaging 2021: Image Processing*, volume 11596, page 1159616. International Society for Optics and Photonics, 2021. 2

[51] Jiabo Xu, Saeed Anwar, Nick Barnes, Florian Grimpen, Olivier Salvado, Stuart Anderson, and Mohammad Ali Armin. Ofgan: Realistic rendition of synthetic colonoscopy videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 732–741. Springer, 2020. 2

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2

[53] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2