International Journal of Computer Assisted Radiology and Surgery (2021) 16:849–859
https://doi.org/10.1007/s11548-021-02383-4

ORIGINAL ARTICLE

Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation

Manish Sahu[1] · Anirban Mukhopadhyay[2] · Stefan Zachow[1]

Received: 18 March 2021 / Accepted: 16 April 2021 / Published online: 12 May 2021

Abstract

Purpose Segmentation of surgical instruments in endoscopic video streams is essential for automated surgical scene understanding and process modeling. However, relying on fully supervised deep learning for this task is challenging because manual annotation occupies valuable time of the clinical experts.

Methods We introduce a teacher–student learning approach that learns jointly from annotated simulation data and unlabeled real data to tackle the challenges in *simulation-to-real* unsupervised domain adaptation for endoscopic image segmentation.

Results Empirical results on three datasets highlight the effectiveness of the proposed framework over current approaches for the endoscopic instrument segmentation task. Additionally, we provide analysis of major factors affecting the performance on all datasets to highlight the strengths and failure modes of our approach.

Conclusions We show that our proposed approach can successfully exploit the unlabeled real endoscopic video frames and improve generalization performance over pure simulation-based training and the previous state-of-the-art. This takes us one step closer to effective segmentation of surgical instrument in the annotation scarce setting.

Keywords Surgical instrument segmentation · Simulation-based learning · Self-supervision · Consistency learning · Self-ensembling · Unsupervised domain adaptation

Introduction

A faithful segmentation of surgical instruments in endoscopic videos is a crucial component of surgical scene understanding and realization of automation in computer- or robot-assisted intervention systems.[1] A majority of recent approaches address the problem of surgical instrument segmentation by training deep neural networks (DNNs) in a fully-supervised scheme. However, the applicability of such supervised approaches is restricted by the availability of a sufficiently large amount of real videos with clean annotations. The annotation process (especially pixel-wise) can be prohibitively expensive (see Fig. 1) because it takes valuable time of medical experts.

An alternative direction to mitigate the dependency on annotated video sequences is to utilize synthetic data for the training of DNNs. Recent advances in graphics and simulation infrastructures have paved the way to automatically create a large number of photo-realistic simulated images with accurate pixel-level labels [14,23]. However, the DNNs trained purely on simulated images do not generalize well on real endoscopic videos due to the domain shift/bias issue [30,32]. We hypothesize that a DNN's bias towards recognizing textures rather than shapes [4] results in a significant drop of performance when the DNNs are trained on simulation (rendered) data and applied to real environments. This is mainly because the heterogeneity of information within a real surgical scene is heavily influenced by factors such as lighting conditions, motion blur, blood, smoke, specular reflection, noise etc. However, simulation data only mimic shapes of instrument and patient-specific organs [10].

✉ Manish Sahu
sahu@zib.de

Anirban Mukhopadhyay
anirban.mukhopadhyay@gris.tu-darmstadt.de

Stefan Zachow
zachow@zib.de

[1] Zuse Institute Berlin (ZIB), Berlin, Germany

[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany

[1] EndoVis Sub-challenges—2015, 2017, 2018, 2019 [https://endovis.grand-challenge.org].

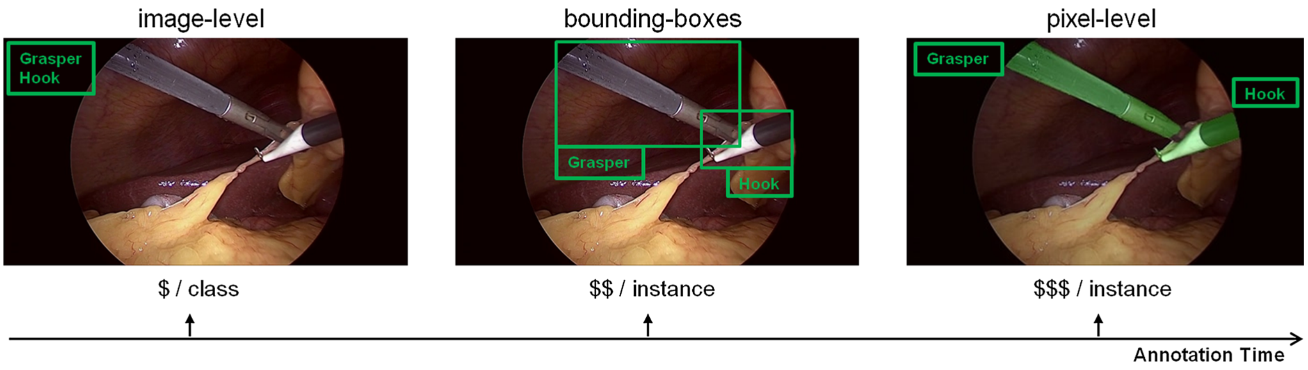

Fig. 1 Fully supervised deep learning is unrealistic for instrument segmentation due to a significantly high annotation effort

The research problem of learning a task-specific representation from the annotated data in a source domain (e.g. simulation) that generalizes on a different but related target domain (e.g. real) is commonly referred to as visual domain adaptation [33]. Unsupervised domain adaptation (UDA) is a specific scenario of domain adaptation where the annotations for the target domain are not available during learning [34]. Here, the primary goal is to learn domain-invariant feature representations for addressing the domain shift/bias [14,35]. For instance, Pfeiffer et al. [23] utilized an *image-to-image* translation approach, where the simulated images are translated into realistic looking ones by mapping image styles (texture, lighting) of the real data using a Cycle-GAN. In contrast, we argued for a shape-focused joint learning from simulated and real data in an end-to-end fashion and introduced a consistency-learning-based approach *Endo-Sim2Real* [27] to align the DNNs on both domains. We showed that similar performance can be obtained by employing a non-adversarial approach while improving the computational efficiency (with respect to training time). However, similar to perturbation-based consistency learning approaches for image classification, [11,17] *Endo-Sim2Real*, being a consistency learning approach at its core, suffers from so-called confirmation bias [29]. This is caused by noise accumulation or erroneous learning during the training stages, which may result in a degenerate solution [6].

In this work, we introduce the teacher–student learning paradigm to the task of surgical instrument segmentation in endoscopic videos. Our proposed approach tackles the erroneous learning by improving the pseudo-label generation procedure for the unlabeled data and facilitate stable training of DNNs while maintaining computational efficiency. Through quantitative and qualitative analysis, we show that our proposed approach outperforms the previous *Endo-Sim2Real* approach across three data sets. Moreover, the proposed approach leads to a stable training without loosing computational efficiency.

The contributions of our work are as follows:

1. We formalise the consistency-based unsupervised domain adaptation framework to identify the confirmation bias problem of *Endo-Sim2Real* and propose a teacher–student learning paradigm to address this problem.
2. We evaluate our work on three different datasets with varying degrees of the domain gap to show consistent improvement in the performance generalization capability of the DNN across the datasets and in presence of unseen instruments or multiple instrument combinations.
3. We provide a thorough quantitative and qualitative analysis to show the strengths and limitations of our approach. In particular, identification of the failure modes with respect to specific cases and scenarios in order to provide valuable insights into addressing the remaining performance gap.

Related work

Research on instrument segmentation for endoscopic procedures is dominated by supervision-based approaches ranging from full supervision [5], semi/self-supervision [25], and weak supervision [12] up to multi-task [16] and multi-modal learning [15]. Some recent works also explored unsupervised approaches [7,18], however, for the sake of brevity, we will only focus on approaches that employ learning from simulation data for unsupervised domain adaptation.

Within the context of domain adaptation in surgical domains, Mahmood et al. [20] proposed an adversarial-based transformer network to translate a real image to a synthetic image such that a depth estimation model trained on synthetic images can be applied to the real image. On the other hand, Rau et al. [24] proposed a conditional Generative Adversarial Network (GAN)-based approach to estimate depth directly from real images. Other works have argued for translating synthetic images to photo-realistic images by using domain mapping via style transfer [19,21], for instance by using Cycle-GAN based *unpaired image-to-image* translation [9,22] and utilize annotations from synthetic environment for deep learning tasks.

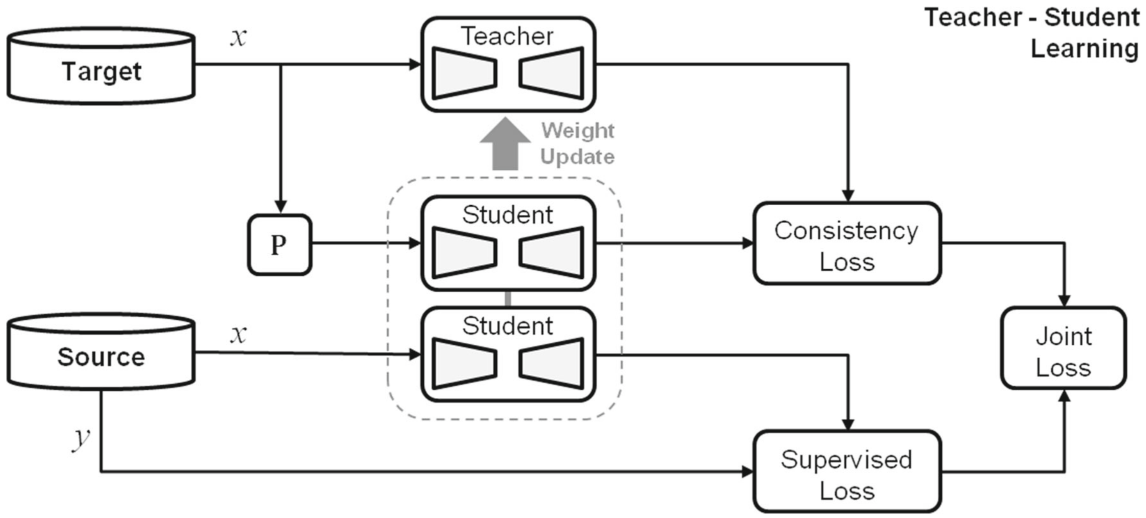

Fig. 2 Proposed teacher–student learning approach comprising supervised learning from source (simulation) data as well as consistency learning from unlabeled target (real) data

Pfeiffer et al. [23] proposed an unpaired *image-to-image* translation approach *I2I* that focuses on reducing the distribution difference between the source and the target domain by employing a Cycle-GAN-based style transfer. Afterwards, a DNN is trained on the translated images and its corresponding labels. On the other hand, *Endo-Sim2Real* [27] utilizes similarity-based joint learning from both simulation and real data under the assumption that the shape of an instrument remains consistent across domains as well as under semantic preserving perturbations (like adding pixel-level noise or transformations).

This work is in line with *Endo-Sim2Real* and focuses on end-to-end learning for unsupervised domain adaptation. However, we formalise the consistency-based UDA to identify the confirmation bias problem and unstable training of *Endo-Sim2Real* approach and address it by employing a teacher–student paradigm. This facilitates stable training of the DNN and enhances its performance generalization capability.

Method

Our proposed teacher–student domain-adaptation approach (see Fig. 2) aims to bridge the domain gap between source (simulated) and target (real) data by aligning a DNN model to both domains. Given:

- a source domain $D_s = (X_s, Y_s)$ associated with a feature space $|X_s|$ and a label space $|Y_s|$ and containing n_s labeled samples $\{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ where, $x_i \in X_s$ and $y_i \in Y_s$ denote the i-th pair of image and label data, respectively
- a target domain $D_t = (X_t)$ associated with a feature space $|X_t|$ and a label space $|Y_t|$ and containing n_t unlabeled samples $\{x_i^t\}_{i=1}^{n_t}$ where, $x_i \in X_t$ denote the i-th image of the unlabeled data

the goal of unsupervised domain adaptation is to learn a DNN model that generalizes on the target domain D_t. It is important to note that although the simulation and real endoscopic scene may appear similar, the label space between source- and target-domains generally differ (i.e. $Y_s \neq Y_t$), representing for example different organs or different instrument types. Since we are focusing on binary instrument segmentation, the label categories are twofold (i.e. $Y_s = Y_t = \{\text{“instrument”}, \text{“background”}\}$). For the sake of simplicity, we refer to the source domain D_s as labeled simulated domain D_L^{Sim} and to the target domain D_t as unlabeled real domain D_{UL}^{Real}.

Our proposed (and previous) approach learns by jointly minimizing the supervised loss L_{sl} for the labeled simulated data-pair as well as the consistency loss L_{cl} for the unlabeled real data. A core component of the joint learning approach is unsupervised consistency learning, where a supervisory signal is generated by enforcing the DNN f_θ (parameterized with network weights θ) to produce a consistent output for an unlabeled input x and its perturbed form $\mathcal{P}(x)$.

$$\min_{\theta} \;\; \mathcal{L}_{sl} + \mathcal{L}_{cl} \;\; \{\underbrace{f_\theta(x)}_{\tilde{\mathbf{y}}},\, f_\theta(\mathcal{P}(x))\} \tag{1}$$

Here in Eq. 1, the DNN prediction $\tilde{\mathbf{y}}$ for unperturbed data x acts as a *pseudo-label* for perturbed data $\mathcal{P}(x)$ to guide the learning process. Therefore, the *Endo-Sim2Real* scenario can be interpreted as a *student-as-teacher* approach where the DNN acts as both a teacher that produces pseudo-labels and a student that learns from these labels. Since the DNN predictions may be incorrect or noisy during training [17], this *student-as-teacher* approach leads to so-called the confirmation bias [29], which reinforces the student to overfit to the incorrect pseudo-labels generated by the teacher and prevents learning new information. This issue is especially prominent during early stages of the training, when the DNN still lacks the correct interpretation of the labels. If the unsupervised consistency loss (L_{cl}) outweighs the supervised loss (L_{sl}), the learning process is not effective and leads to a sub-optimal performance. Therefore, the consistency loss is typically employed with a temporal weighting function $w(t)$ such that the DNN learns prominently from the supervised loss during the initial stages of the learning and gradually shifts towards unsupervised consistency learning in the later stages.

Although the temporal ramp-up weighting function in *Endo-Sim2Real* helps to reduce the effect of the confirmation bias during joint learning, the DNN still learns directly from the incorrect pseudo-labels generated by the teacher.

$$\min_{\theta} \left\{ \underbrace{\sum_{(x_i, y_i) \in D_L^{\text{Sim}}} \mathcal{L}_{sl}\left(f_\theta\left(x_i\right), y_i\right)}_{\text{supervised simulation}} + w(t) * \underbrace{\sum_{x_i \in D_{\text{UL}}^{\text{Real}}} \mathcal{L}_{cl}\left(f_{\theta'}\left(x_i\right), f_\theta\left(\mathcal{P}(x_i)\right)\right)}_{\text{unsupervised real}} \right\}$$

$$\theta_t^{'} = (\alpha \cdot \theta_{t-1}^{'} + (1-\alpha) \cdot \theta_t) \tag{2}$$

In this work, we address this major drawback of the *Endo-Sim2Real* approach by improving the pseudo-label generation procedure of the unlabeled consistency learning. To this end, the teacher network is de-coupled from the student network and redefined ($f_\theta \longrightarrow f_{\theta'}$) to generate reliable targets to enable the student to gradually learn meaningful information about the instrument shape. In order to avoid separate training of the teacher model, the same architecture is used for the teacher and its parameters are updated as a temporal average [29] of the student network's weights.

At each training step t, the student f_θ is updated using *gradient-descent* while the teacher $f_{\theta'}$ is updated using student network weights, where the smoothing factor α controls the update rate of the teacher. A pseudo code of our proposed teacher–student learning approach is provided in Algorithm 1.

Algorithm 1 Teacher–student Algorithm for Domain Adaptation

```
Model: $f_\theta$                                   ▷ Student with trainable parameters $\theta$
Model: $f_{\theta'}$                                ▷ Teacher with trainable parameters $\theta^{'}$
Data: $D_L^{Sim}(x, y)$                             ▷ Labeled simulated samples
Data: $D_{UL}^{Real}(x)$                            ▷ Unlabeled real samples
Require: $\alpha$                                   ▷ Update rate of teacher
Require: $w(t)$                                     ▷ Temporal weight of consistency loss
Ensure: $\theta^{'} \leftarrow \theta$              ▷ Initialize weights
loop JOINT LEARNING($D_L^{Sim}$, $D_{UL}^{Real}$)   ▷ $(D_s \rightarrow D_L^{Sim}, D_t \rightarrow D_{UL}^{Real})$
    Supervised Loss
    $\{(x_i, y_i)\}_{i=1}^B \sim D_L^{Sim}(x, y)$                       ▷ Sample mini-batch
    $\{(x_i^a, y_i^a)\}_{i=1}^B = \{\mathcal{A}(x_i, y_i)\}_{i=1}^B$   ▷ Augment batch
    $L_{sl} = \{f_\theta(x_i^a), y_i^a\}_{i=1}^B$                      ▷ Supervised loss
    Consistency Loss
    $\{x_i\}_{i=1}^B \sim D_{UL}^{Real}(x)$                            ▷ Sample mini-batch
    $\{x_i^p\}_{i=1}^B = \{\mathcal{P}(x_i)\}_{i=1}^B$                 ▷ Perturb batch
    $\{\tilde{y}_i\}_{i=1}^B = \{f_{\theta'}(x_i^p)\}_{i=1}^B$         ▷ Pseudo Segmentation
    $L_{cl} = \{f_\theta(x_i), \tilde{y}_i\}_{i=1}^B$                  ▷ Unsupervised loss
    Joint Loss
    $L = L_{sl} + w(t) \cdot L_{cl}$                                   ▷ Joint loss
    $g_\theta = \nabla_\theta L$                                       ▷ Compute gradients
    $\theta \leftarrow Update(\theta, g_\theta)$                       ▷ Update student (gradient descent)
    $\theta^{'} \leftarrow (\alpha \cdot \theta^{'} + (1-\alpha) \cdot \theta)$   ▷ Update teacher
end loop
return $\theta^{'}$                                                    ▷ Learned model
```

Experimental setup

Data

Simulation [23] data contain $20K$ rendered images acquired via 3-D laparoscopic simulations from the CT scans of 10 patients. The images describe a rendered view of a laparoscopic scene with each tissue having a distinct texture and a presence of two conventional surgical instruments (grasper and hook) under a random placement of the camera (coupled with a light source).

Cholec [27] data contain around $7K$ endoscopic video frames acquired from 15 videos of the Cholec80 dataset [31]. The images describe the laparoscopic cholecystectomy scene with seven conventional surgical instruments (grasper, hook, scissors, clipper, bipolar, irrigator and specimen bag). The data provide segmentations for each instrument type, however, the specimen bag is considered as a counterexample that is treated as background during evaluation, following the definition of an instrument in RobustMIS challenge [26].

EndoVis [1] data consist of 300 images from six different in-vivo 2D recordings of complete laparoscopic colorectal surgeries. The data provide binary segmentations of instruments for validation where images describe an endoscopic scene containing seven conventional instruments (including hook, traumatic grasper, ligasure, stapler, scissors and scalpel) [5].

RobustMIS [26] data consist of around $10K$ images acquired from 30 surgical procedures of three different types of colorectal surgery (10 rectal resection procedures, 10 proctocolectomy procedures and 10 procedures of sigmoid resec-

Table 1 List of source (simulation) and target (real) datasets used during evaluation, where [videos (#) | empty frames (%)] reflects the number of videos and percentage of frames with no instrument, respectively

Dataset	Training	Testing	Instruments
Simulation	20,000 (10 | 33%)	n.a.	Two
Cholec15	5034 (10 | 12%)	2136 (5 | 13%)	Six + specimenbag
EndoVis	160 (4 | 0%)	140 (6 | 0%)	Seven
RobustMIS	5983 (16 | 17%)	4057 (14 | 20%)	Six + trocar
Stage 1		663 (2 | 15%)	–
Stage 2		514 (2 | 13%)	–
Stage 3		2880 (10 | 23%)	–

tion procedures). An instrument is defined as an elongated rigid object that is manipulated directly from outside the patient. Therefore, grasper, scalpel, clip applicator, hooks, stapling device, suction and even trocar is considered as an instrument while non-rigid tubes, bandages, compresses, needles, coagulation sponges, metal clips etc. are considered as counterexamples as they are indirectly manipulated from outside [26]. The data provide instance level segmentations for validation, which are performed in three different stages with an increasing domain gap between the training- and the test-data. Stage 1 contains video frames from 16 cases of the training data, stage 2 has video frames of two proctocolectomy and rectal surgeries each, and stage 3 has video frames from 10 sigmoid resection surgeries

It is important to note that the domain gap increases not only in the three stages of testing in Robust-MIS dataset, but also from *Simulation* towards *Real* datasets (EndoVis < Cholec < Robust-MIS) as the definition of instrument (and/or counterexample) changes along with other factors (Table 1).

Implementation

We have redesigned the implementation of the *Endo-Sim2Real* framework in view of a teacher–student approach. To ensure a direct and fair comparison, we employ the same *TerNaus11* [28] as a backbone segmentation model. Also, we utilize the best performing perturbation scheme (i.e. applying one of the *pixel-intensity* perturbation[2] followed by one of the *pixel-corruption* perturbation[3]) and the loss function (i.e. *cross-entropy* and *jaccard*) of *Endo-Sim2Real* for evaluation. All simulated input images and labels are first pre-processed with a stochastically-varying circular outer mask to give them the appearance of real endoscopic images.

We use a batch size of 8 for 50 epochs and apply weight decay ($1e-6$) as standard regularization. During consistency training, we use a time-dependent weighting function, where the weight of the unlabeled loss term is linearly increased over the training. The teacher model is updated with α (0.95) at each training step.

During evaluation of a dataset, we use an image-based dice score and average over all images to obtain a global dice metric for the dataset. For computation of the dice score, we exclude the cases where both the prediction and ground truth images are empty. However, we include cases with false positives for the empty images and set it to zero. So the dice score for empty ground-truth images (without any instrument) is either zero and considered in case of any false positives or undefined and not considered in case of correct prediction. Also, we report all the results as an average performance of three runs throughout our experiments.

Results and discussion

This section provides a quantitative comparison with respect to the state-of-the-art approaches to demonstrate the effectiveness of our approach. Moreover, we perform quantitative and qualitative analyses on three different datasets with varying degrees of the domain gap to highlight the challenges in simulation-to-real unsupervised domain adaptation. Particularly, we identify the failure modes with respect to specific cases and scenarios in order to provide valuable insights into addressing the remaining performance gap.

Comparison with *baseline* and *state-of-the-art*

In these experiments, we first highlight the performance of the two baselines: the *lower baseline* (supervised learning purely on simulated data) and the *upper baseline* (supervised learning purely on annotated real data) in Table 2. The substantial performance gap between the baselines indicates the domain gap between simulated and real data. Secondly, we compare our proposed teacher–student approach with other unsupervised domain adaptation approaches, i.e. the domain style transfer approach (*I2I*) and the plain consistency-based joint learning approach (*Endo-Sim2Real*) on the *Cholec* dataset. The empirical results show that *Endo-Sim2Real*

[2] *pixel-intensity*: random brightness and contrast shift, posterisation, solarisation, random gamma shift, random HSV color space shift, histogram equalization and contrast limited adaptive histogram equalization.

[3] *pixel-corruption*: gaussian noise, motion blurring, image compression, dropout, random fog simulation and image embossing.

Table 2 Quantitative comparison using DSC [mean (std)] [empty frames (%)] reflects the percentage of empty frames

Approach	EndoVis	Cholec	R-MIS (S 1)	R-MIS (S 2)	R-MIS (S 3)
Simulation only	.42 (.29)	.30 (.30)	.30 (.28)	.36 (.30)	.20 (.24)
I2I [23]	n.a.	.68 (.30)	n.a.	n.a.	n.a.
Endo-Sim2Real [27]	.76 (.17)	.68 (.31)	.57 (.33)	.60 (.32)	.51 (.35)
Teacher–student	**.80** (.16)	**.72** (.30)	**.61** (.32)	**.65** (.31)	**.58** (.34)
Real only	.93 (.06)	.86 (.24)	.82 (.27)	.83 (.25)	.78 (.30)
Empty frames	0%	13%	15%	13%	23%

The two baselines: simulation only and real only means training only on the simulated data and training on the annotated real data, respectively (i.e. no adaptation). The Wilcoxon signed-rank test for *Endo-Sim2Real* and our work results in *p-value* $\ll 0.01$. Bold values represents performance score

works similar to *I2I*, while our proposed approach outperforms both of these approaches. Later, we evaluate our approach on two additional datasets and show that it consistently outperforms *Endo-Sim2Real*. These experiments demonstrate that the generalization performance of the DNN can be enhanced by employing unsupervised consistency learning on unlabeled data. Finally, the performance gap with the upper baseline calls for identification of the issues needed to bridge the remaining domain gap.

Analysis on *EndoVis*

Among the three datasets, our proposed approach performs best for EndoVis as shown in Table 2. A visual analysis of the low performing cases in Fig. 3 highlights factors such

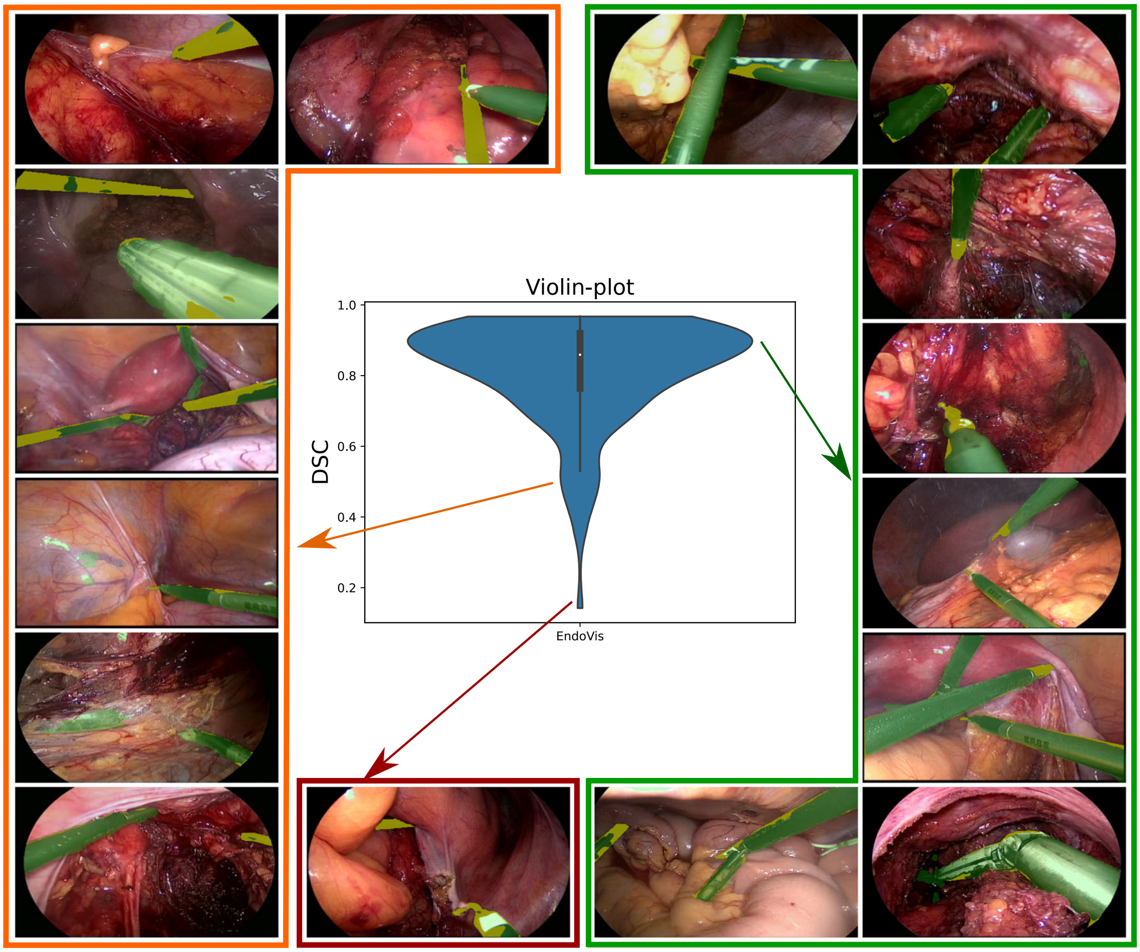

Fig. 3 Qualitative analysis on EndoVis dataset. The green color in the images represents the network predictions while the yellow color represents under-segmentation

as false detection on specular reflection, under-segmentation for small instruments, tool-tissue interaction and partially occluded instruments. These factors can in part be addressed by utilizing the temporal information of video frames [13].

Analysis on *Cholec*

We performed an extensive performance analysis of our proposed approach on the Cholec dataset as instrument-specific labels are available for it (in comparison to EndoVis). To understand the distinctive performance aspects for the Cholec dataset, we compare the segmentation performance across different instrument co-occurrence in Fig. 4. A similar range of dice scores highlights that the performance of our approach is less impacted by the presence of multiple tool combinations in an endoscopic image.

However, it also clearly shows that the segmentation performance of our approach drops when the specimen bag and its related co-occurrences are present (as seen in the respective box plots in Fig. 4). A visual analysis highlights false detection on the reflective surface of the specimen bag.

Apart from the previously analyzed performance degrading factors in the EndoVis dataset, other major factors affecting the performance are as follows:

∗ *Out of distribution cases* such as a non-conventional tool-shape-like instrument: specimen bag (see box-plots for labelsets with specimen bag in Fig. 4).
∗ *False detection for scenarios* such as an endoscopic view within the trocar, instrument(s) near the image border or under-segmentation for small instruments.
∗ *Artefact cases* such as specular reflection. The impact of other artefacts such as blood, smoke or motion blur is lower.

Although our proposed approach struggles to tackle these artefacts and out of distribution cases, addressing these performance degrading factors is itself an open research problem [2].

Analysis on *RobusMIS*

We analyzed the performance of our approach on images with a different number of instruments in the RobustMIS dataset. We found that the performance is not significantly affected by the presence of multiple tools (see Table 3). A low performance for a single visible instrument is attributed to small, stand-alone instruments across image boundary. Apart from the factors in the Cholec dataset, other real-world performance degrading factors in RobustMIS include: presence of other out-of-distribution cases such as non-rigid tubes, bandages, needles etc.; presence of corner cases

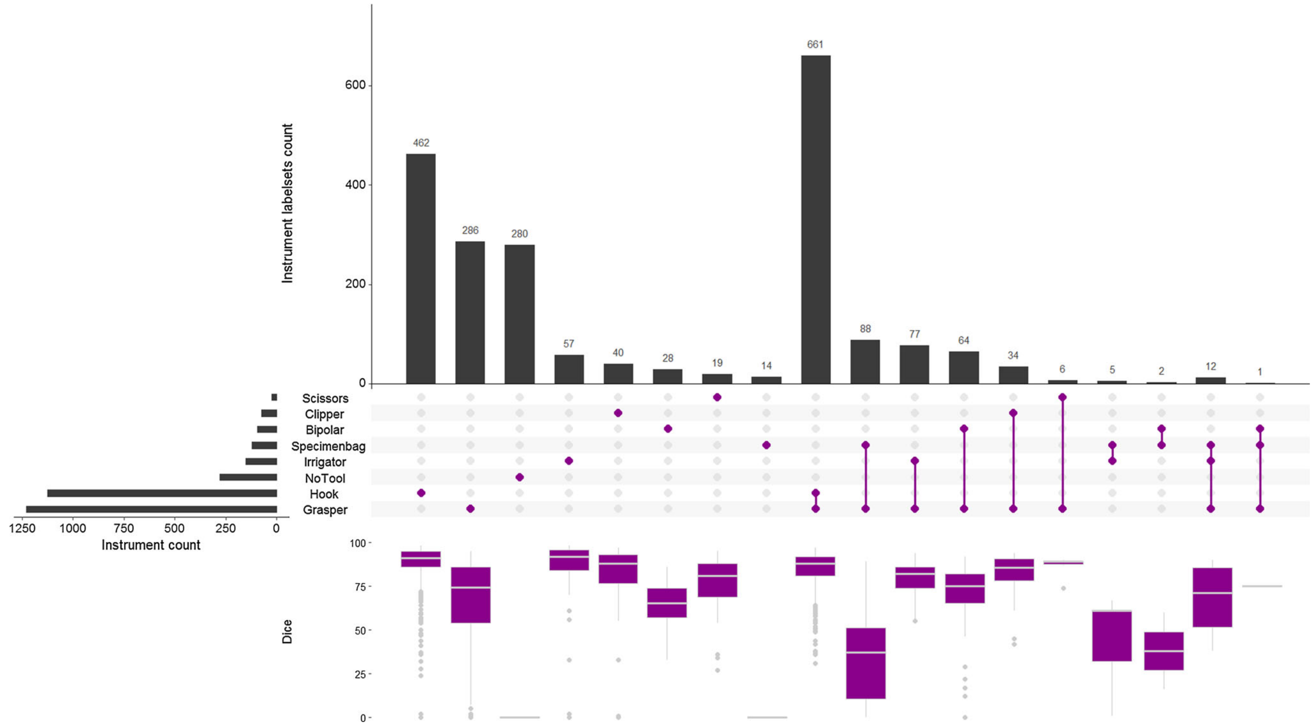

Fig. 4 Visualization of the relation between tool co-occurrence and segmentation quality for the Cholec dataset. Please note that the dice score is zero for no tool cases and specimen bag as it is treated as background

Table 3 Quantitative results for multi-instrument presence in RobustMIS dataset using DSC [mean (std)]

No.	1	2	3	4	5	6	7
#	1802	1173	229	31	3	–	1
DSC	.61 (.34)	.72 (.22)	.73 (.18)	.76 (.11)	.76 (.13)	–	.77 (.00)

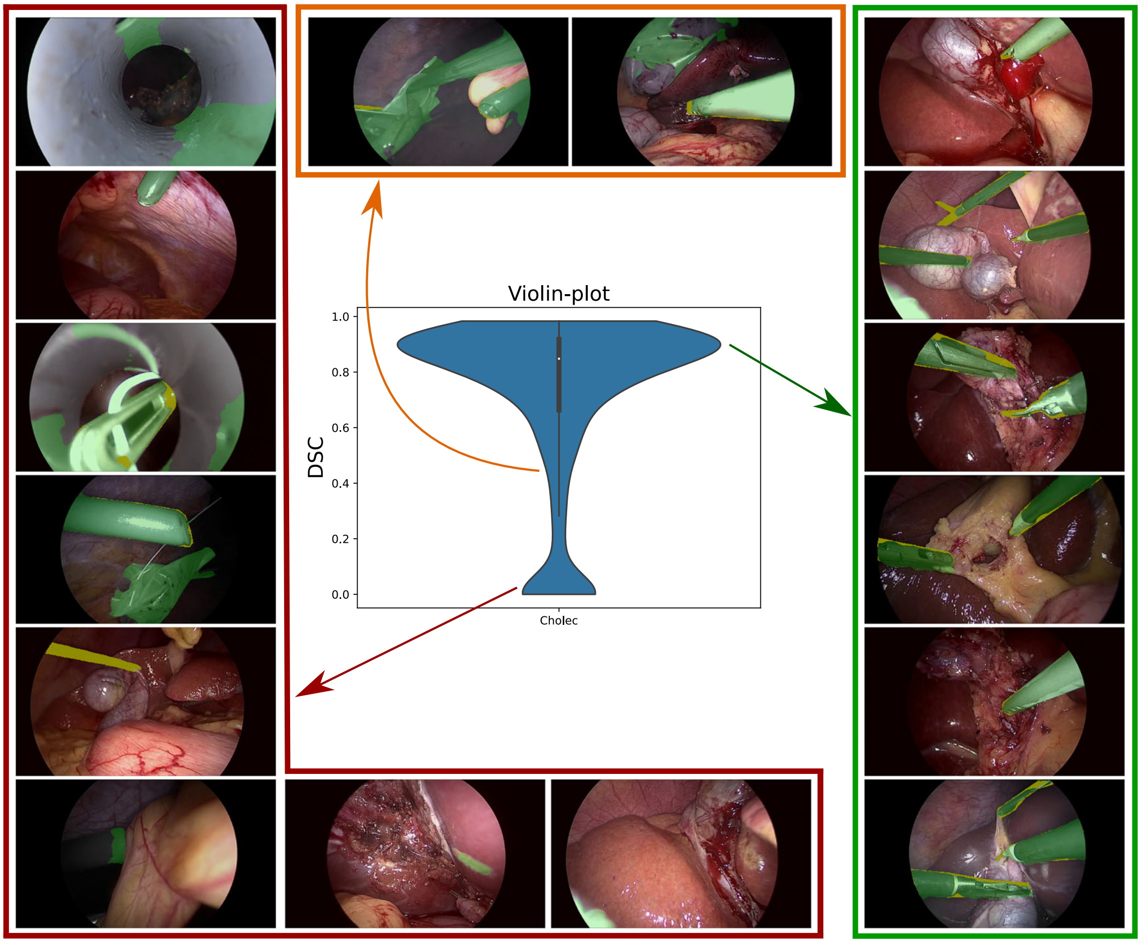

Fig. 5 Qualitative analysis on the Cholec dataset. The green color in the images represents the network predictions while the yellow color represents under-segmentation

such as trocar-views and specular reflections producing a instrument-shape-like appearance. These failure cases highlight a drawback of our approach, which works under the assumption that the shape of the instrument remains consistent between the domains. Therefore, our approach may not be able to produce faithful predictions in case instruments with different shapes are encountered in the real domain (compared to instruments in simulation) or counterexamples with instrument like appearance.

Impact of empty ground-truth frames

The performance of our teacher–student approach is negatively affected by the video frames that do not contain instruments. This is because the dice score is assigned to zero when the network predicts false positives (as seen in Figs. 5 and 6) in instrument-free video frames. A direct relation of this effect can be seen in Table 2 where the dice score across the datasets decreases as the number of empty frames increases (in %) from EndoVis to RobustMIS. It suggests that utilizing false detection techniques in the current framework can help in enhancing the generalization capabilities.

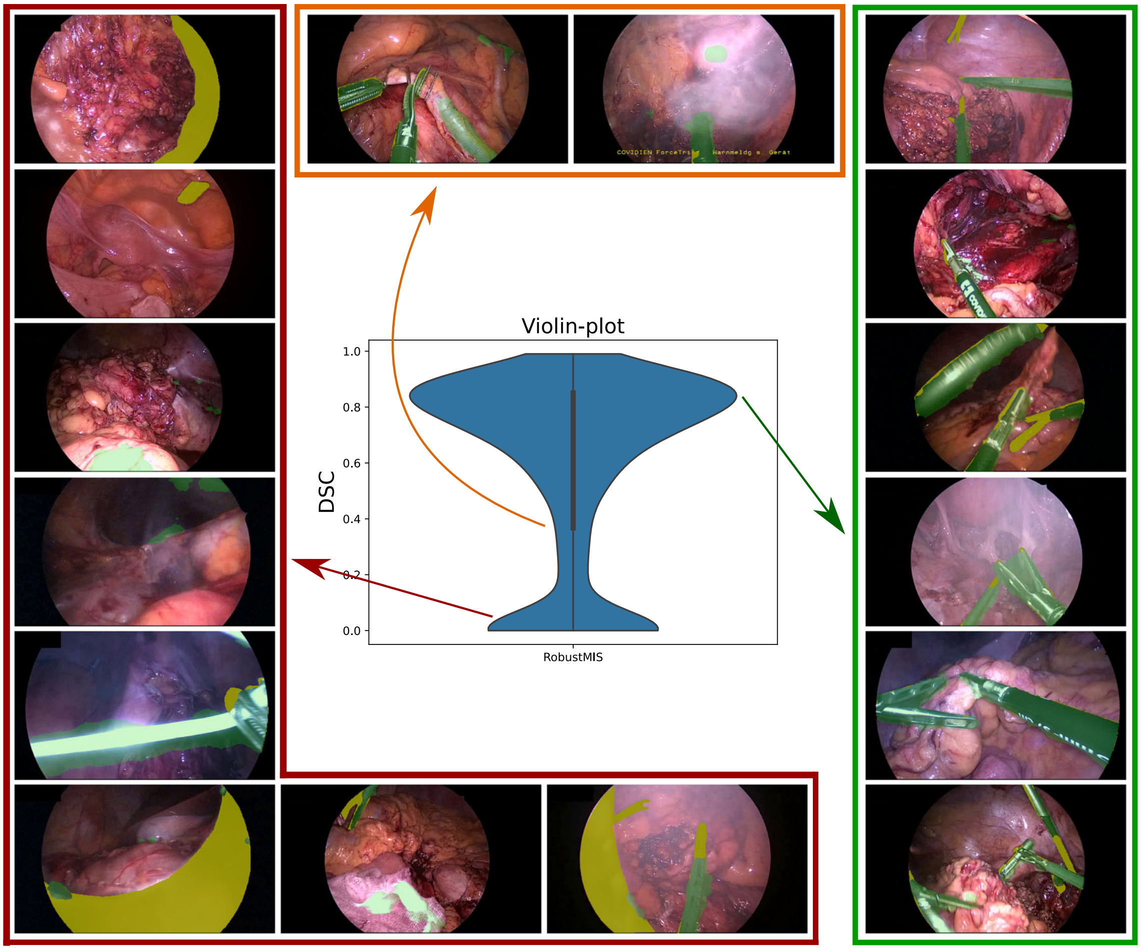

Fig. 6 Qualitative analysis on the RobustMIS dataset. The green color in the images represents the network predictions while the yellow color represents under-segmentation

Conclusion

We introduce teacher–student learning to address the confirmation bias issue of the *EndoSim2Real* consistency learning. This enables us to tackle the challenging problem of the domain shift between synthetic and real images for surgical tool segmentation in endoscopic videos. Our proposed approach enforces the teacher model to generate reliable targets to facilitate stable student learning. Since the teacher is a moving average model of the student, the extension does not add computational complexity to the current approach.

We show that the proposed teacher–student learning approach generalizes across three different datasets for the instrument segmentation task and consistently outperforms the previous state-of-the-art. For a majority of images (see high peak in Figs. 3, 5 and 6), the segmentation predictions are usually correct with small variations across the instrument boundary. Moreover, a thorough analysis of the results highlight interpretable failure modes of simulation-to-real deep learning as the domain gap widens progressively.

Considering the strengths and limitations of our teacher–student enabled simulation-to-real unsupervised domain adaptation approach, the framework admits multiple straightforward extensions to bridge the remaining domain gap:

∗ Implementing techniques to suppress false detection for empty frames, instruments near the image border and specular reflections, for instance by utilizing temporal information [13] of video frames.
∗ Improving physical properties of simulation to capture instrument-tissue interaction, considering the variations in predictions across instrument boundaries.
∗ Extension towards semi-supervised domain adaptation or real-to-real unsupervised domain adaptation by utilizing labels from target (real) data for the endoscopic instrument segmentation task.
∗ Employing this approach in conjunction with other self-supervised or adversarial domain mapping approaches such as *I2I* [27].

Being flexible, end-to-end and unsupervised with respect to the target domain, our approach can be adapted to other imaging modalities or learning tasks which utilize joint learning from labeled and unlabeled data. For instance, it can be extended towards other domain-adaptation tasks, such as depth estimation [20,24] or instrument pose estimation [3,8] by exploiting depth maps from the simulated virtual environments.

The heavy reliance of current approaches on manual annotation and the harsh reality of surgeons sparing time for the annotation process propels simulation-to-real domain adaptation as the obvious problem to address in surgical data science. The proposed approach ushers annotation-efficient surgical data science for the operating room of the future.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was funded by the German Federal Ministry of Education and Research (BMBF) under the project COMPASS (grant no. - 16 SV 8019).

Declarations

Conflict of interest The authors state no conflict of interest.

Informed consent This study contains patient data from a publicly available datasets.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. (2015) Endovis sub-challenge: instrument segmentation and tracking. https://endovissub-instrument.grand-challenge.org/. Accessed 28 October 2020
2. Ali S, Zhou F, Braden B, Bailey A, Yang S, Cheng G, Zhang P, Li X, Kayser M, Soberanis-Mukul RD, Albarqouni S, Wang X, Wang C, Watanabe S, Oksuz I, Ning Q, Yang S, Khan MA, Gao XW, Realdon S, Loshchenov M, Schnabel JA, East JE, Wagnieres G, Loschenov VB, Grisan E, Daul C, Blondel W, Rittscher J (2020) An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci Rep 10(1):1–15
3. Allan M, Ourselin S, Hawkes DJ, Kelly JD, Stoyanov D (2018) 3-D pose estimation of articulated instruments in robotic minimally invasive surgery. IEEE Trans Med Imaging 37(5):1204–1213
4. Baker N, Lu H, Erlikhman G, Kellman PJ (2018) Deep convolutional networks do not classify based on global object shape. PLoS Comput Biol 14(12):e1006613
5. Bodenstedt S, Allan M, Agustinos A, Du X, Garcia-Peraza-Herrera L, Kenngott H, Kurmann T, Müller-Stich B, Ourselin S, Pakhomov D, Sznitman R, Teichmann M, Thoma M, Vercauteren T, Voros S, Wagner M, Wochner P, Maier-Hein L, Stoyanov D, Speidel S (2018) Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv:1805.02475
6. Chapelle O, Scholkopf B, Zien A (2009) Semi-supervised learning. IEEE Trans Neural Netw 20(3):542
7. Colleoni E, Edwards P, Stoyanov D (2020) Synthetic and real inputs for tool segmentation in robotic surgery. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 700–710
8. Du X, Kurmann T, Chang PL, Allan M, Ourselin S, Sznitman R, Kelly JD, Stoyanov D (2018) Articulated multi-instrument 2-d pose estimation using fully convolutional networks. IEEE Trans Med Imaging 37(5):1276–1287
9. Engelhardt S, De Simone R, Full PM, Karck M, Wolf I (2018) 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. Springer, pp 747–755
10. Engelhardt S, Sharan L, Karck M, De Simone R, Wolf I (2019) Cross-domain conditional generative adversarial networks for stereoscopic hyperrealism in surgical training. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 155–163
11. French G, Mackiewicz M, Fisher M (2018) Self-ensembling for visual domain adaptation. In: International conference on learning representations
12. Fuentes-Hurtado F, Kadkhodamohammadi A, Flouty E, Barbarisi S, Luengo I, Stoyanov D (2019) Easylabels: weak labels for scene segmentation in laparoscopic videos. Int J Comput Assist Radiol Surg 14:1247–1257. https://doi.org/10.1007/s11548-019-02003-2
13. González C, Bravo-Sánchez L, Arbelaez P (2020) Isinet: an instance-based approach for surgical instrument segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 595–605
14. Hoffman J, Wang D, Yu F, Darrell T (2016) FCNs in the wild: pixel-level adversarial and constraint-based adaptation. arXiv:1612.02649
15. Jin Y, Cheng K, Dou Q, Heng PA (2019) Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 440–448
16. Laina I, Rieke N, Rupprecht C, Vizcaíno JP, Eslami A, Tombari F, Navab N (2017) Concurrent segmentation and localization for tracking of surgical instruments. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 664–672
17. Laine S, Aila T (2017) Temporal ensembling for semi-supervised learning. In: International conference on learning representations
18. Liu D, Wei Y, Jiang T, Wang Y, Miao R, Shan F, Li Z (2020) Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 657–667
19. Luengo I, Flouty E, Giataganas P, Wisanuvej P, Nehme J, Stoyanov D (2018) Surreal: enhancing surgical simulation realism using style

transfer. In: British machine vision conference 2018, BMVC 2018, BMVA, pp 1–12

20. Mahmood F, Chen R, Durr NJ (2018) Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans Med Imaging 37(12):2572–2581
21. Marzullo A, Moccia S, Catellani M, Calimeri F, De Momi E (2020) Towards realistic laparoscopic image generation using image-domain translation. Comput Methods Programs Biomed 200:105834
22. Oda M, Tanaka K, Takabatake H, Mori M, Natori H, Mori K (2019) Realistic endoscopic image generation method using virtual-to-real image-domain translation. Healthc Technol Lett 6(6):214–219
23. Pfeiffer M, Funke I, Robu MR, Bodenstedt S, Strenger L, Engelhardt S, Roß T, Clarkson MJ, Gurusamy K, Davidson BR, Maier-Hein L, Riediger C, Welsch T, Weitz J, Speidel S (2019) 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. Springer, pp 119–127
24. Rau A, Edwards PE, Ahmad OF, Riordan P, Janatka M, Lovat LB, Stoyanov D (2019) Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int J Comput Assist Radiol Surg 14(7):1167–1176
25. Ross T, Zimmerer D, Vemuri A, Isensee F, Wiesenfarth M, Bodenstedt S, Both F, Kessler P, Wagner M, Müller B, Kenngott H, Speidel S, Kopp-Schneider A, Maier-Hein K, Maier-Hein L (2018) Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int J Comput Assist Radiol Surg 13(6):925–933
26. Ross T, Reinke A, Full PM, Wagner M, Kenngott H, Apitz M, Hempe H, Mindroc Filimon D, Scholz P, Nuong Tran T, Bruno P, Arbeláez P, Bian GB, Bodenstedt S, Lindström Bolmgren J, Bravo-Sánchez L, Chen HB, González C, Guo D, Halvorsen P, Heng PA, Hosgor E, Hou ZG, Isensee F, Jha D, Jiang T, Jin Y, Kirtac K, Kletz S, Leger S, Li Z, Maier-Hein KH, Ni ZL, Riegler MA, Schoeffmann K, Shi R, Speidel S, Stenzel M, Twick I, Wang G, Wang J, Wang L, Wang L, Zhang Y, Zhou YJ, Zhu L, Wiesenfarth M, Kopp-Schneider A, Müller-Stich BP, Maier-Hein L (2020) Robust medical instrument segmentation challenge 2019. arXiv:2003.10299
27. Sahu M, Strömsdörfer R, Mukhopadhyay A, Zachow S (2020) Endo-sim2real: consistency learning-based domain adaptation for instrument segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 784–794
28. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI (2018) Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 624–628
29. Tarvainen A, Valpola H (2017) Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems, pp 1195–1204
30. Torralba A, Efros AA (2011) Unbiased look at dataset bias. In: CVPR 2011. IEEE, pp 1521–1528
31. Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N (2016) Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36(1):86–97
32. Vercauteren T, Unberath M, Padoy N, Navab N (2020) Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. Proc IEEE 108(1):198–214. https://doi.org/10.1109/JPROC.2019.2946993
33. Wang M, Deng W (2018) Deep visual domain adaptation: a survey. Neurocomputing 312:135–153
34. Wilson G, Cook DJ (2020) A survey of unsupervised deep domain adaptation. ACM Trans Intell Syst Technol (TIST) 11(5):1–46
35. Zhang Y, David P, Gong B (2017) Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE international conference on computer vision, pp 2020–2030