BL-UDA: Towards Unsupervised Domain-Adaptive Surgical Instrument Segmentation with Source Box Labels

Ziyuan Zhao
Nanyang Technological University
Singapore, Singapore
S210088@e.ntu.edu.sg

Yuhua Wang
National University of Singapore
Singapore, Singapore
e0950550@u.nus.edu

Yifang Yin
Institute for Infocomm Research, A*STAR
Singapore, Singapore
yin_yifang@i2r.a-star.edu.sg

Yichen Zhang
National University of Singapore
Singapore, Singapore
zhang.yichen@u.nus.edu

Xulei Yang
Institute for Infocomm Research, A*STAR
Singapore, Singapore
yang_xulei@a-star.edu.sg

Jun Cheng
Institute for Infocomm Research, A*STAR
Singapore, Singapore
cheng_jun@a-star.edu.sg

Roger Zimmermann
National University of Singapore
Singapore, Singapore
rogerz@comp.nus.edu.sg

Cuntai Guan
Nanyang Technological University
Singapore, Singapore
CTGUAN@NTU.EDU.SG

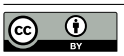
S. Kevin Zhou
University of Science and Technology of China
Suzhou, China
s.kevin.zhou@gmail.com

Abstract

Recent advances in unsupervised domain adaptation (UDA) by adapting the model from one domain to another unseen domain have shown considerable promise in improving surgical instrument segmentation performance across domains. However, existing UDA methods primarily rely on pixel-wise labels, which are always difficult to collect due to the labor-intensive annotation process. In this work, we aim to relax the dependence on pixel-level supervision and investigate a challenging UDA setting - source box annotations, where weak supervision and domain shifts coexist. To achieve this, we introduce a novel unsupervised domain adaptation framework, BL-UDA, which leverages bounding box annotations for surgical instrument segmentation across domains. By utilizing the Segment Anything Model (SAM) for pseudo label generation from box annotations, our method effectively bridges object-level and pixel-level domain adaptation. The proposed BL-UDA framework comprises a teacher-student network with entropy minimization for object detection and an entropy-based label selection strategy for generating box prompts to SAM, facilitating pixel-level domain adaptation. Extensive experiments on the EndoVis 2017 and 2018 datasets demonstrate the superiority of BL-UDA over existing UDA methods, significantly mitigating domain shifts and addressing weak supervision challenges with minimal annotation requirements.

CCS Concepts

• **Applied computing** → **Life and medical sciences**; • **Computing methodologies** → **Computer vision**.



ICMR '26, Amsterdam, Netherlands

ACM ISBN 979-8-4007-2617-0/26/06
https://doi.org/10.1145/3805622.3810665

Keywords

Domain adaptation, surgical instrument segmentation

ACM Reference Format:
Ziyuan Zhao, Yuhua Wang, Yifang Yin, Yichen Zhang, Xulei Yang, Jun Cheng, Roger Zimmermann, Cuntai Guan, and S. Kevin Zhou. 2026. BL-UDA: Towards Unsupervised Domain-Adaptive Surgical Instrument Segmentation with Source Box Labels. In *International Conference on Multimedia Retrieval (ICMR '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3805622.3810665

1 Introduction

In recent years, computer and robotic assisted interventions have been promising in the minimally invasive surgery (MIS) completed with small incisions, making the surgical process safer, more precise, and less invasive [17]. The success of robot-assisted minimally invasive surgery (RAMIS) highly relies on surgical scene understanding to provide context awareness, where surgical instrument segmentation plays an important role in providing pixel-wise awareness of instruments for supporting many downstream tasks, including surgery monitoring [5], surgical navigation [25] and skill assessment [11], consequently improving surgical safety and reliability. Despite the significant progresses achieved by deep learning-based methods in surgical instrument segmentation [1, 18, 22, 27], these supervised learning approaches are developed based on the i.i.d assumption, ignoring that surgical scenarios may come from different domains, *e.g.,* hospitals, patient population and procedures. As a result, these models trained on one domain will suffer from substantial performance degradation when applied to previously unseen domains, due to the domain shifts [4, 29–31].

To address this challenge, significant efforts have been made towards achieving unsupervised domain adaptation by leveraging labeled source domain and unlabeled target domain for minimizing the domain shift, thereby improving model generalization on the

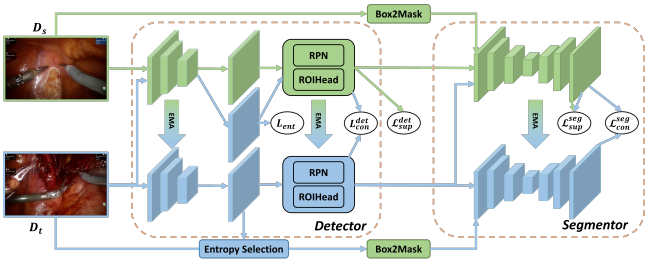


Figure 1: Overview of BL-UDA, which consists of: (1) cross-domain object detection, (2) reliable target-domain sample selection, (3) box-to-pixel pseudo-label generation (Box2Mask), and (4) cross-domain semantic segmentation.

target domain. Existing UDA approaches mainly focus on image translation [3, 21, 23, 28] and feature alignment [12, 13, 20] from different perspectives, including adversarial learning [3, 21, 23, 35], self-training [14, 20], and graph representation learning [12, 13]. Notably, Liu *et al.* [12] proposed a graph-based UDA framework, by constructing the prototypical graph and leveraging domain-common knowledge to effectively align feature distributions across domains, thereby achieving promising adaptation performance in surgical instrument segmentation. However, for unsupervised domain adaptation, massive annotations from source domain are required for model training and adaptation, which would inevitably increase annotation time and costs, especially for surgical instrument segmentation, where the fully annotated surgical videos are very limited due to the great workload of pixel-wise annotations [33, 34].

In comparison to pixel-wise annotations, cheap and weak annotations, *e.g.*, bounding boxes can be more easily obtained in most clinical cases, as well as surgical videos. To leverage box annotations, weakly supervised learning has been widely investigated in various segmentation tasks [9, 19, 24], mostly utilizing the box annotations as prior information for pseudo label generation. However, current methods are limited by the quality of pseudo labels and do not account for cross-domain scenarios. These motivate us to advocate for investigating a practical, challenging, and distinct UDA setting where only box annotations are available in the source domain, alongside existing domain shifts and weak supervision.

Recently, segmentation foundation models, such as the Segment Anything Model (SAM) [8] have received large attention on various segmentation tasks with their strong capabilities in generating precise object masks using spatial prompts, *e.g.*, bounding boxes. It is observed that SAM-generated pseudo labels usually possess precise boundaries, which can facilitate weakly supervised segmentation and unsupervised domain adaptation. Motivated by this, we leverage SAM to generate pseudo labels for weakly supervised segmentation in a cross-domain setting. Specifically, we propose BL-UDA, a novel unsupervised domain adaptation framework based on source-domain box annotations, to address both weak supervision and domain shifts in surgical instrument segmentation. As illustrated in Fig. 1, we first construct a teacher-student network with entropy minimization for cross-domain object detection, the goal of which is to bridge the domain gap at object level. Then, we design an entropy-based sample selection strategy to identify target-domain samples with reliable pseudo box labels, which are then used as box prompts for SAM. Finally, we construct a teacher-student segmentation network for cross-domain instrument segmentation using SAM-generated pseudo labels from both domains, where mutual teacher–student learning is employed to achieve pixel-level domain adaptation. Our method leverages SAM as a bridge between object-level DA and pixel-level DA, achieving unsupervised domain adaptation only using source box labels. Extensive experiments and ablation analysis on the EndoVis Instrument Segmentation 2017 and 2018 benchmarks validate the effectiveness of the proposed BL-UDA in addressing both weak supervision and domain shifts, outperforming existing UDA methods by a large margin.

2 Method

Consider two domains: source domain $\mathcal{D}_s = \left\{\left(x_i^s, y_i^s\right)\right\}_{i=1}^{N}$ with N labeled samples, and target domain $\mathcal{D}_t = \left\{\left(x_i^t\right)\right\}_{i=1}^{M}$ with M unlabeled samples. In a traditional UDA setting, ground truth pixel-wise annotations are provided in the source domain, *i.e.*, $y_i^s \in \{0,1\}^{H\times W\times C}$, where H, W, and C represent the height, width and the number of classes, respectively. In comparison, only ground truth bounding boxes in the source domain are provided in our setting, *i.e.*, $b_i^s \in \mathbb{R}^{k\times 5}$, where k is the number of bounding boxes comprising of four coordinates and one class label, which is cheaper but more challenging. We aim to leverage source samples with box annotations and unlabeled target samples to jointly address weak supervision and domain shift, thereby achieving unsupervised domain-adaptive surgical instrument segmentation. Fig. 1 depicts an overview of the proposed BL-UDA architecture.

2.1 Addressing domain gap for object detection

To harness the potential of fine bounding box annotations in the source domain, we first propose to close the domain gap on the object detection level by leveraging the mean-teacher framework based on self-ensembling. Specifically, we construct a student model f_θ^{od} learned by standard gradient updating and an architecturally identical teacher model $f_{\theta'}^{od}$ updated using the exponential moving average (EMA) weights of the student model, *i.e.*, $\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$, where α is the EMA decay rate which is used to regulate the impact of student model parameters. To improve the quality of pseudo labels for the target domain, we feed source images with strong augmentations to the student model, while feeding target images with strong and weak augmentations to the student and teacher model individually, following [16]. To this end, the teacher model is expected to generate reliable pseudo labels without being affected by heavy augmentation. We optimize the student model using labeled source data x^s via the supervised loss defined as:

$$L_{sup}^{det}(x^s, b^s) = L_{cls}^{rpn}(x^s, b^s) + L_{reg}^{rpn}(x^s, b^s) + L_{cls}^{roi}(x^s, b^s) + L_{reg}^{roi}(x^s, b^s),$$

which combines the following losses: the Region Proposal Network loss L^{rpn} for proposing candidate regions and the Region of Interest loss L^{roi} for predicting bounding boxes adjustments. Both loosely perform bounding box regression (reg) and classification (cls). Here, we employ binary cross-entropy loss for the classification and l_1 loss for the regression. To effectively utilize unlabeled target data, we generate the refined pseudo labels b^t on the target domain by

setting a confidence threshold T on the predicted boxes with the teacher model and exclude duplicated boxes with non-maximum suppression for each class. Then we calculate the consistency loss to update the student model, denoted as follows:

$$L_{con}^{det}(x^t, b^t) = L_{cls}^{rpn}(x^t, b^t) + L_{cls}^{roi}(x^t, b^t).$$

Moreover, we introduce the entropy loss using the Shannon Entropy for maximizing the prediction certainty in the target domain to generate highly confident pseudo labels. To be more specific, for an input target image x_i^t, we calculate the entropy loss based on the feature maps produced by an additional 1×1 convolutional layer connecting the student feature encoder E_s , denoted as follows:

$$L_{ent}(x_i^t) = \sum_{h,w} \frac{-1}{\log(C)} \sum_{c=1}^{C} P_{x_i^t}(h, w, c) \log P_{x_i^t}(h, w, c),$$

where $P_{x_i^t}(h, w, c)$ denotes the predicted probability of class c for the pixel at position (h, w) after a softmax layer. Finally, the total loss for training the proposed cross-domain object detection model is written as:

$$L_{od} = L_{sup}^{det} + \lambda_{con}^{det} L_{con}^{det} + \lambda_{ent}^{det} L_{ent}, \quad (1)$$

where λ_{con}^{det} and λ_{ent}^{det} are the weights of the associated losses. With the proposed method, we can effectively minimize the domain gap for the object detection task, and generate reliable bounding boxes on the target domain for the following cross-domain segmentation.

2.2 Addressing domain gap for semantic segmentation

Inspired by the great success of Segment Anything Model (SAM) in weakly supervised semantic segmentation [7], we propose to leverage SAM to address the weak supervision for generating dense mask by using box annotations as prompts (*i.e.,* "box2mask") and use the generated masks as pseudo labels for optimizing the segmentation framework. Specifically, we feed source data x^s to SAM using the ground truth box annotations b^s as prompts and generate segmentation masks $\hat{y}^s$, *i.e.,* $\hat{y}^s = \text{SAM}(x^s, b^s)$. Similarly, we can apply the same process on pseudo target box annotations generated from the object detection model. However, the pseudo box quality on object detection as prompts would influence the SAM-generated pseudo mask quality. In this regard, we propose an entropy-based sample selection strategy to select some high-quality pseudo box annotations for target pseudo mask generation with SAM. More specifically, we select the top $K\%$ target samples for pseudo mask generation by sorting the entropy scores (see Eq. 1) in ascending order by the teacher model in object detection, *i.e.,* $\hat{y}^t = \text{SAM}(x^t_{\text{top-k}}, b^t_{\text{top-k}})$. Then, we construct the student-teacher network based on self-ensembling for cross-domain segmentation, where we feed SAM-annotated data from both domains to the student model f_ϑ^{seg} for supervised learning, and feed remaining unlabeled target data to both student f_ϑ^{seg} and EMA teacher $f_{\vartheta'}^{seg}$ for unsupervised consistency learning. Finally, the training objective of the student model is written as:

$$L_{seg} = L_{sup}^{seg} + \lambda_{con}^{seg} L_{con}^{seg}, \quad (2)$$

where L_{sup}^{seg} is the combination of cross-entropy loss and dice loss, and we employ the MSE loss for L_{con}^{seg} with the weights λ_{con}^{seg}. Within the BL-UDA framework, we leverage SAM to distill knowledge from object detection model to the segmentation model and close the domain gaps at both levels sequentially, thereby achieving more accurate and robust box-supervised domain adaptive segmentation.

Table 1: Performance comparison on Endovis17 → Endovis18 for surgical instrument detection.

Method	Scissor		Needle Driver		Froceps		mAP	mAP_{50}
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}		
Source only	21.48	42.62	24.45	49.77	38.68	62.22	28.21	51.53
Oracle	66.80	94.73	13.99	24.03	53.58	80.08	44.79	66.28
MT [26]	52.54	73.13	16.83	19.36	52.17	74.43	40.51	55.64
AT [10]	51.67	76.95	15.54	27.56	45.53	68.46	37.58	57.66
BL-UDA, Ours	57.93	80.61	16.13	29.51	53.67	76.33	42.57	62.15

3 Experiments

Dataset and evaluation metrics. We evaluate the effectiveness of our proposed framework using the *EndoVis*17 [2] and *EndoVis*18 [1] datasets. Following [13], we partition the *EndoVis*18 dataset into 596 images for testing and 1, 639 images for training. Our evaluation metrics encompass the use of labels for forceps, needle drivers, and scissors. Bounding boxes are generated for each connected region on the semantic labels of the *EndoVis*17 dataset. We utilize SAM [8] to generate pseudo masks from the predicted bounding boxes in the target domain with $K = 50$. For evaluation metrics, we adopt the average precision at thresholds of 50% (AP_{50}) and the mean average precision (mAP) for object detection. Moreover, the mean intersection over union (mIoU) and Dice similarity coefficient (Dice) are employed for evaluating segmentation performance following [12].
Implementation details. We adopt the Faster R-CNN framework as our primary object detection model following [10]. For hyperparameters, we set the consistency loss weight λ_{con}^{det} to 0.5, the entropy loss weight λ_{ent} to 0.1, and the confidence threshold T to 0.7. We initially "warmed up" the student detection model using source labels for 500 iterations, followed by 36, 000 iterations on both domains. During training, we set a learning rate of 0.0025 with a Stochastic Gradient Descent (SGD) optimizer. The EMA decay rate is set to 0.9996. For semantic segmentation, we follow [15], employing DeeplabV3+ with a ResNet101 as the segmentation backbone. The images are cropped to dimensions of 256×320 pixels as inputs. For hyperparameters, the weight for consistency λ_{con}^{seg} is set as 1.5. We train the segmentation model for 80 epochs, starting using an initial learning rate of $5e-4$ with polynomial learning rate decay $(1 - \frac{iter}{iter_{max}})^{0.9}$, where $iter$ and $iter_{max}$ represent the current and maximum number of iterations individually.

3.1 Cross-domain object detection performance

We implement MT [26] and AT [10] as comparison methods. MT is a mean-teacher framework originally developed for semi-supervised learning, while AT presents a cross-domain adaptive teacher designed for object detection. We also include a source-only baseline as a lower bound, where the model is trained on the source domain and directly evaluated on the target domain without adaptation, as well as an oracle baseline as an upper bound, where the model is both trained and evaluated on the target domain. All methods are implemented using the Faster R-CNN backbone for fair comparison.

Table 2: Performance comparison on Endovis17 → Endovis18 for surgical instrument segmentation.

Method	Label	Scissor		Needle Driver		Froceps		mIoU
		IoU	Dice	IoU	Dice	IoU	Dice	
Source only	Mask	45.97	55.33	22.08	34.98	38.92	51.72	35.66
SePIG (MICCAI'21)	Mask	65.11	70.86	19.21	29.84	61.36	74.76	48.56
IGNet (TMI'21)	Mask	70.65	74.87	37.60	46.92	52.97	66.00	53.74
SimT (CVPR'22)	Mask	76.24	-	39.83	-	58.94	-	58.34
SeCo (NeurIPS'24)	Mask	78.40	-	41.20	-	61.20	-	60.40
Source OD + SAM	BBox	45.56	54.03	36.04	58.78	39.18	50.68	40.26
MT [26] + SAM	BBox	73.17	83.85	45.59	56.86	63.56	**74.82**	60.77
AT [10] + SAM	BBox	72.19	84.89	41.55	**63.13**	54.28	65.24	56.01
BL-UDA, Ours	BBox	**83.94**	**85.72**	**48.97**	57.89	**67.07**	69.78	**66.66**

Table 1 presents the object detection results achieved by different methods, where we observe a notable performance drop from "oracle" to "source only" (from 44.79% to 28.21% in mAP), indicating the domain gap across domains. Moreover, both MT and AT significantly improve domain adaptive object detection performance compared with the source-only baseline. Our proposed method further outperforms MT and AT by 2.1% ∼ 5.0% in mAP and 4.5% ∼ 6.5% in mAP_{50}, which indicates that our method can generate more accurate bounding boxes for the following segmentation task.

3.2 Cross-domain segmentation performance

Next, we compare our method against four state-of-the-art approaches in domain-adaptive surgical instrument segmentation, namely SePIG [13], IGNet [12], SimT [6], and SeCo [32]. It is worth noting that these methods are supervised with pixel-wise annotations (Mask). Moreover, we integrate SAM on top of various detection models trained using only bounding-box annotations for comparison. Specifically, we first perform domain-adaptive object detection using MT or AT, followed by mask extraction using SAM. Table 2 shows the performance comparison of different methods. We observe that combining SAM with the source object detection (OD) model improves the lower-bound segmentation performance (*i.e.*, source only) and narrows the domain gap to some extent, suggesting that the domain discrepancy may be smaller at the box level than at the pixel level, particularly when facilitated by foundation models such as SAM. Furthermore, when combining SAM with OD-UDA methods (MT and AT), better segmentation performance than SePIG and IGNet can be achieved, which indicates that closing domain gap for object detection followed by SAM can effectively address the domain gap for segmentation.

Comparatively, SimT and SeCo are designed for general-purpose domain adaptive semantic segmentation. When applied to surgical instrument segmentation task, their performance can be compromised due to the scarcity of training data. Finally, by integrating our proposed OD-UDA and Seg-UDA with SAM, we achieve the best segmentation performance of 66.66% in mIoU, outperforming previous methods with mask supervision by a large margin. This result affirms the effectiveness of jointly conducting domain adaptation at both levels, namely object detection and segmentation, to close the domain gap. The qualitative segmentation results of various methods are shown in Fig. 2, where we can see that our method can generate more accurate masks with smooth boundaries.

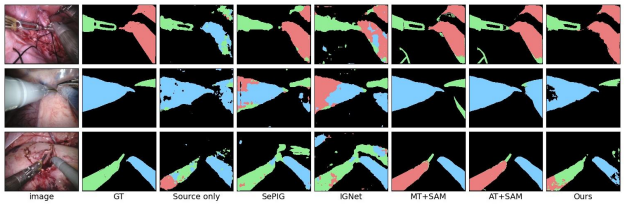


Figure 2: Visual comparison of different methods for instrument segmentation.

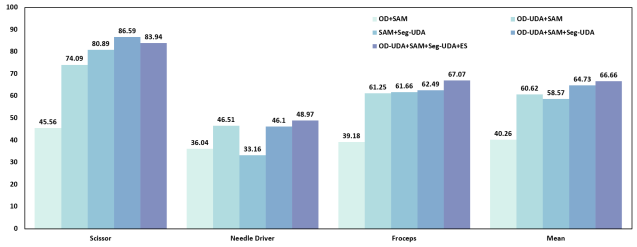


Figure 3: Bar chart of segmentation results (mIOU[%]) on different variants of our proposed approach.

3.3 Ablation Study

We conduct an ablation analysis by removing the key components of the proposed method, as shown in Fig. 3. Specifically, "OD+SAM" conducts SAM after object detection excluding any domain adaptation techniques, receiving a substantial reduction in mean Intersection over Union (mIoU) by 26.4%. For "OD-UDA+SAM", where the Seg-UDA component is removed, there is a 6.04% decrease in mIoU. Conversely, the 'SAM+Seg-UDA' setup, which does not incorporate OD-UDA, results in an 8.09% mIoU reduction. Lastly, in the 'OD-UDA+SAM+Seg-UDA' case, we remove the entropy-based label selection strategy and randomly involve 50% target images into supervised training for Seg-UDA, resulting in a 1.93% performance drop in mIoU compared to our BL-UDA model, demonstrating the effectiveness of different components in the proposed method.

4 Conclusions

We propose a novel UDA framework to jointly address weak supervision and domain shift for surgical instrument segmentation. Facilitated by SAM, our model effectively generates high-quality pseudo labels for domain adaptation at both object detection and segmentation levels, thereby improving domain adaptation performance using only source box annotations. The extensive experimental results and ablation analysis demonstrate that our proposed BL-UDA is effective in addressing weak supervision and domain shifts, outperforming existing UDA methods by a significant margin.

Acknowledgments

This research is supported by the RIE2030 Career Development Fund (Award No. H26-KSR0023), administered by A*STAR, and is also funded by the NUS Artificial Intelligence Institute (NAII) Seed Grant (No. NAII-SG-2025-027).

References

[1] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2020. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190* (2020).

[2] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2019. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426* (2019).

[3] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. 2019. CrDoCo: Pixel-level Domain Transfer with Cross-Domain Consistency. In *IEEE CVPR*.

[4] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. 2020. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *IEEE CVPR*. 4023–4032.

[5] Luis Claudio Gubert, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2020. Context awareness in healthcare: a systematic literature review. *Universal Access in the Information Society* 19 (2020), 245–259.

[6] Xiaoqing Guo, Jie Liu, Tongliang Liu, and Yixuan Yuan. 2022. SimT: Handling open-set noise for domain adaptive semantic segmentation. In *IEEE CVPR*. 7032–7041.

[7] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. 2024. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems* 36 (2024).

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[9] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. 2021. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *IEEE CVPR*. 2643–2652.

[10] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. 2022. Cross-domain adaptive teacher for object detection. In *IEEE CVPR*. 7581–7590.

[11] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. 2021. Towards unified surgical skill assessment. In *IEEE CVPR*. 9522–9531.

[12] Jie Liu, Xiaoqing Guo, and Yixuan Yuan. 2021. Graph-based surgical instrument adaptive segmentation via domain-common knowledge. *IEEE Transactions on Medical Imaging* 41, 3 (2021), 715–726.

[13] Jie Liu, Xiaoqing Guo, and Yixuan Yuan. 2021. Prototypical interaction graph for unsupervised domain adaptation in surgical instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 272–281.

[14] Xiaofeng Liu, Fangxu Xing, Maureen Stone, Jiachen Zhuo, Timothy Reese, Jerry L Prince, Georges El Fakhri, and Jonghye Woo. 2021. Generative self-training for cross-domain unsupervised tagged-to-cine mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 138–148.

[15] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *IEEE CVPR*. 4258–4267.

[16] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. 2020. Unbiased Teacher for Semi-Supervised Object Detection. In *International Conference on Learning Representations*.

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

[18] Zhen-Liang Ni, Xiao-Hu Zhou, Guan-An Wang, Wen-Qian Yue, Zhen Li, Gui-Bin Bian, and Zeng-Guang Hou. 2022. SurgiNet: Pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation. *Medical Image Analysis* 76 (2022), 102310.

[19] Youngmin Oh, Beomjun Kim, and Bumsub Ham. 2021. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *IEEE CVPR*. 6913–6922.

[20] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. 2020. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *IEEE CVPR*. 3764–3773.

[21] 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. 2019. 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: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*. Springer, 119–127.

[22] Tobias Ross, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, et al. 2020. Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299* (2020).

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

[24] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *IEEE CVPR*. 3136–3145.

[25] Leonardo Tanzi, Pietro Piazzolla, Francesco Porpiglia, and Enrico Vezzetti. 2021. Real-time deep learning semantic segmentation during intra-operative surgery for 3D augmented reality assistance. *International Journal of Computer Assisted Radiology and Surgery* 16, 9 (2021), 1435–1445.

[26] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30 (2017).

[27] Yanjie Xia, Shaochen Wang, and Zhen Kan. 2023. A nested u-structure for instrument segmentation in robotic surgery. In *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 994–999.

[28] Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *IEEE CVPR*. 4085–4095.

[29] Yifang Yin, Jinming Cao, Zhenguang Liu, Guanfeng Wang, Shili Xiang, and Roger Zimmermann. 2026. Improving Test-Time Efficiency in Source-Free Semantic Segmentation via Multi-Stage Self-Training. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2026).

[30] Yifang Yin, Wenmiao Hu, Zhenguang Liu, Guanfeng Wang, Shili Xiang, and Roger Zimmermann. 2023. Crossmatch: Source-free domain adaptive semantic segmentation via cross-modal consistency training. In *IEEE ICCV*. 21786–21796.

[31] Yichen Zhang, Yifang Yin, Ying Zhang, Zhenguang Liu, Zheng Wang, and Roger Zimmermann. 2023. Prototypical cross-domain knowledge transfer for cervical dysplasia visual inspection. In *ACM Multimedia*. 1504–1514.

[32] Dong Zhao, Qi Zang, Shuang Wang, Nicu Sebe, and Zhun Zhong. 2024. Connectivity-driven pseudo-labeling makes stronger cross-domain segmenters. *Advances in Neural Information Processing Systems* 37 (2024), 78936–78963.

[33] Ziyuan Zhao, Fangcheng Zhou, Kaixin Xu, Zeng Zeng, Cuntai Guan, and S Kevin Zhou. 2022. LE-UDA: Label-efficient unsupervised domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging* 42, 3 (2022), 633–646.

[34] Ziyuan Zhao, Fangcheng Zhou, Zeng Zeng, Cuntai Guan, and S Kevin Zhou. 2022. Meta-hallucinator: Towards few-shot cross-modality cardiac image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 128–139.

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