DenseTRF: Texture-Aware Unsupervised Representation Adaptation for Surgical Scene Dense Prediction

Guiqiu Liao[1,2], Matjaž Jogan[1,2], and Daniel A. Hashimoto[1,2,3]

[1] GRASP Laboratory, University of Pennsylvania
[2] PCASO Laboratory, Department of Surgery, University of Pennsylvania
[3] Department of Computer and Information Science, University of Pennsylvania
Guiqiu.Liao@pennmedicine.upenn.com

Abstract. Dense prediction tasks in surgical computer vision, such as segmentation and surgical zone prediction, can provide valuable guidance for laparoscopic and robotic surgery. However, these models often suffer from distribution shifts, as training datasets rarely cover the variability encountered during deployment, leading to poor generalization. We propose DenseTRF, a self-supervised representation adaptation framework based on texture-centric attention. Our method leverages slot attention to learn texture-aware representations that capture invariant visual structures. By adapting these representations to the target distribution without supervision, DenseTRF significantly improves robustness to domain shifts. The framework is implemented through conditioning dense prediction on slot attention and model merging strategies. Experiments across multiple surgical procedures demonstrate improved cross-distribution generalization in comparison to state-of-the-art segmentation models and test-distribution adaptation methods for dense prediction tasks. Code shared at: https://github.com/PCASOlab/Dense-TRF.

Keywords: Unsupervised Adaptation · Dense Prediction · Slot Attention

1 Introduction

Dense prediction tasks in surgical video analysis, such as tissue segmentation, instrument tracking, and surgical zone prediction, are critical for computer-assisted interventions and real-time intraoperative guidance [22,15]. Nevertheless, robust deployment remains difficult due to pronounced domain shifts across operating environments caused by variations in anatomy, appearance, viewpoint, illumination, and procedural phase [14,5].

Contemporary surgical scene understanding approaches predominantly rely on supervised learning with densely annotated datasets [19,2,22]. Although these

Accepted to 29th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2026)

methods achieve strong performance when training and test distributions are aligned, their accuracy degrades under domain shift [5]. This limitation is particularly pronounced in surgery, where dense annotation is costly and labor-intensive, often restricting training data to a limited number of procedures and institutions [9,15]. Consequently, models that perform well on development datasets often fail to generalize across clinical settings and procedure types.

Recent advances in foundation vision models [1,11,21] have demonstrated remarkable generalization capabilities by leveraging large-scale pretraining. However, as we show in this work, even these powerful representations benefit from targeted adaptation when deployed on specialized surgical domains with limited annotated data. Test-distribution adaptation methods [17,16,23,24,26] offer a promising direction by adjusting model representations using only unlabeled target data, yet existing approaches often struggle with the texture-rich yet geometrically variable nature of surgical scenes.

A complementary line of research has explored object-centric representations through slot-based attention mechanisms [13,20]. Recent work has further extended slot learning to foundation-model feature spaces, enabling unsupervised object discovery using pretrained representations such as DINO [20,3]. These advances demonstrate that slot attention can capture semantically meaningful structures without supervision. Concurrently, adaptation-oriented extensions have begun to emerge, including slot-based test-time adaptation approaches [17] and cross-domain slot learning frameworks [12]. Despite these developments, the role of object-centric representations for unsupervised test-distribution adaptation in surgical dense prediction remains largely unexplored. Notably, representations that group visual features by appearance rather than spatial configuration may be particularly advantageous in surgical scenes, where tissue textures exhibit consistent statistical properties even as shapes deform across patients and procedural stages.

We introduce DenseTRF (Dense Texture-Centric Representation Adaptation Framework), an object-centric approach for unsupervised test-distribution adaptation in surgical dense prediction. Our key insight is that slot attention can be steered toward texture-aware representations that capture invariant visual structures across domains. DenseTRF integrates three components: (i) conditioning dense prediction on unsupervised slot representations learned via reconstruction, (ii) a periodic model-merging strategy that balances target specialization with generalization, and (iii) evaluation across three surgical datasets (Thoracic, POEM and Cholec) under extreme low-data regimes (1–5% annotations). Results demonstrate consistent improvements over state-of-the-art segmentation and foundation-model baselines. Through comprehensive ablation studies, we demonstrate the contribution of each component and show that our representation adaptation strategy and object-centric representations work in synergy, unlocking flexible and scalable applications of dense prediction in surgical domains.

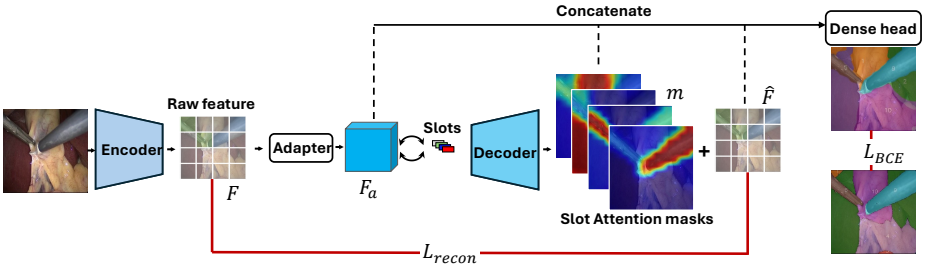

Fig. 1. Architecture of the proposed network, where the dense head is conditioned on object-centric representations learned via Slot Attention and optimized using both reconstruction and supervised dense prediction objectives.

2 Method

2.1 Network architecture

Our dense prediction network is built upon object-centric representations learned via Slot Attention (SA). Given an input image, we extract features using a pre-trained foundation model encoder [21], yielding $F \in \mathbb{R}^{H\times W\times C_r}$. A lightweight MLP adapter g projects these into adapted features $F_a = g(F) \in \mathbb{R}^{H\times W\times C_a}$, which are passed to a SA encoder [13] that iteratively refines K slot latent vectors $\{s_k\}_{k=1}^{K}$. Following DINOSAUR [20], MLP decoders reconstruct the original features from the slots, producing per-slot feature maps $\hat{F}_k$ and alpha masks α_k:

$$\hat{F} = \sum_{k=1}^{K} \hat{F}_k \odot m_k, \quad m_k = \mathrm{softmax}_k(\alpha_k), \quad \mathcal{L}_{\mathrm{recon}} = \|F - \hat{F}\|^2 \tag{1}$$

This learning objective encourages the slots to capture coherent object-like regions, producing interpretable attention masks as a byproduct. We initially considered using only the adapted features F_a as input to the dense prediction head. However, we observed that this underutilized the rich object-centric structure captured by the slots. Therefore, we concatenate three sources of information: (i) the adapter output features, (ii) the reconstructed features, and (iii) the SA masks m_k, forming a combined representation per spatial location:

$$z_{ij} = \left[f_{ij}^{\mathrm{ada}},\ \hat{f}_{ij}^{\mathrm{recon}},\ m_{1,ij}, \ldots, m_{K,ij} \right] \tag{2}$$

This is fed into a lightweight MLP classifier whose logits are upsampled to the input resolution. The model is trained jointly with:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \lambda\, \mathcal{L}_{\mathrm{recon}} \tag{3}$$

DenseTRF is motivated by the need to combine generalizable foundation features with target-specific adaptation under surgical domain shift. The adapter provides a lightweight task-specific projection, while Slot Attention encourages structured grouping of coherent appearance patterns.

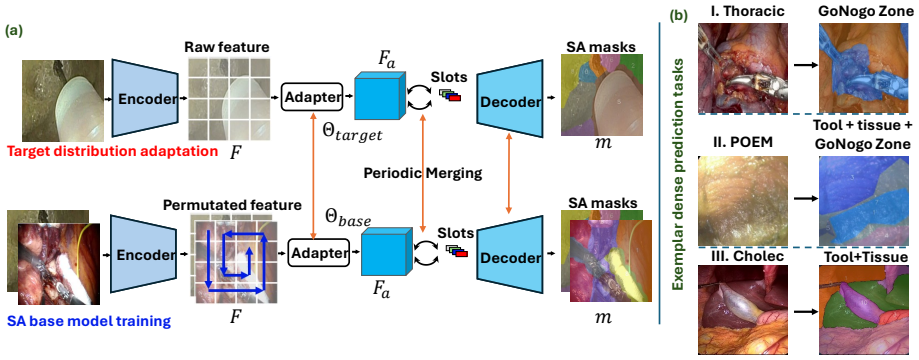

Fig. 2. (a) Slot Attention (SA) adaptation via periodic model merging, which anchors the adapted model to the base representation during specialization. (b) Example dense prediction tasks across Thoracic, POEM, and Cholec surgical datasets.

2.2 Unsupervised test-distribution adaptation

When SA is trained on a mixture of diverse visual domains, the learned slots must explain substantial appearance variability, which can reduce their ability to specialize to a specific deployment scenario. This is especially relevant in surgical settings, where a model's success will be measured by its performance within a particular surgical case and procedure, with video information that exhibits relatively consistent visual statistics compared to all possible surgeries.

To address this challenge, we propose an unsupervised test-distribution adaptation strategy that enables SA to retain broad generalization while refining its representation for the target test domain. Our key observation is that when SA is trained jointly on data of any given domains, its attention masks learn to optimize reconstruction loss of the images of training domains. This has an advantage over training exclusively on a single domain, but will lead to poor generalization due to representation drift and bias toward a narrow set of concepts [12].

To balance specialization and generalization, we adopt a periodic model merging strategy inspired by continual learning approaches [27]. As shown in Figure 2, specifically, we maintain two SA branches that share the same architecture:

- **Base branch:** trained on a broad multi-domain image pool.
- **Target branch:** trained solely on unlabeled target-domain images.

After each adaptation round, the parameters of the two branches are merged through weighted averaging:

$$\Theta^{(r+1)} = \frac{1}{2}(\Theta^{(r)}_{base} + \Theta^{(r)}_{target}) \tag{4}$$

where Θ denotes the parameters of the adapter, iterative competitive SA module, and slot decoder. The merged model is then broadcast back to both branches to initialize the next round.

This periodic synchronization serves as an anchoring mechanism that prevents target-domain over-specialization while still allowing progressive refinement of slot representations toward the deployment distribution.

In addition, we design the SA module to learn texture-aware representations that capture invariant visual structures rather than object geometry. We adopt patch order permutation from SPOT [10] to weaken spatial bias and encourage texture-centric slot grouping. This relaxation promotes more robustness when tissue shapes deform or when the camera viewpoint changes, and decouples slot position from appearance.

3 Experiments and Results

3.1 Datasets and Metrics

We evaluate our approach on three surgical video datasets. (1) **Thoracic**: 30 sequences (2K unlabeled frames each) collected from robot-assisted lung malignancy performed at the Toronto General Hospital between 2014 and 2023, with 688 annotated go/no-go zones for training and 169 for test. (2) **POEM**: 40K total Peroral endoscopic myotomy (POEM) images curated from Hospital of the University of Pennsylvania and Massachusetts General Hospital, 12K training pool, with 120 test images containing sparse go/no-go zone annotations. (3) **Cholec80** [22] (24K frames) for unsupervised pretraining and **CholecSeg8K** [8] (8K annotated frames): 7K training pool (avoiding temporal overlap), 1k for testing. For all datasets, we use only 1%–5% of training pool labels. We report Intersection over Union (IoU), DICE, and Hausdorff Distance (HD in pixels) [15] for region overlap and boundary accuracy.

3.2 Training Details

Slot Attention pretraining. Base branch: LR 4×10^{-4}, weight decay 10^{-5}, with SPOT patch-order permutation and alternating batch schedule.
Test-distribution adaptation. Target branch: LR 1×10^{-4}, weight decay 10^{-5}, trained only on unlabeled target images with reconstruction loss (no feature order permutation).
Dense prediction head. LR 1×10^{-4}, no weight decay. Loss: BCE + $\lambda \mathcal{L}_{\text{recon}}$ ($\lambda = 0.1$). Three-phase training: (i) head only (1k iter), (ii) full model joint, (iii) remove $\mathcal{L}_{\text{recon}}$ (final 1k iter). Total 5k iterations. Early stopping monitored but was never triggered.

All methods, including the proposed approach and competing state-of-the-art baselines, are trained independently across five runs with different random seeds. DenseTRF remains lightweight in practice: both training and inference can be performed with less than 2 GB of GPU memory.

Table 1. Comparison of mDICE scores and model parameter sizes across different methods.

Model	Para [MB]	Thoracic	POEM	Cholec
Mask2former[2]	421	68.14 ± 0.89	32.55 ± 1.17	36.73 ± 3.35
Segman[4]	202	55.52 ± 4.46	27.99 ± 1.07	43.70 ± 1.75
Swin-UNETR[6]	88	68.18 ± 0.43	31.70 ± 0.32	48.83 ± 1.32
Unet[19] + Resnet[7]	127	71.90 ± 0.70	41.55 ± 0.74	52.79 ± 3.12
Unet[19] + MIT[25]	251	76.50 ± 0.17	46.37 ± 0.43	59.56 ± 2.19
DenseTRF (ours)	388	$\mathbf{81.35 \pm 0.33}$	$\mathbf{51.00 \pm 0.79}$	$\mathbf{67.40 \pm 1.00}$

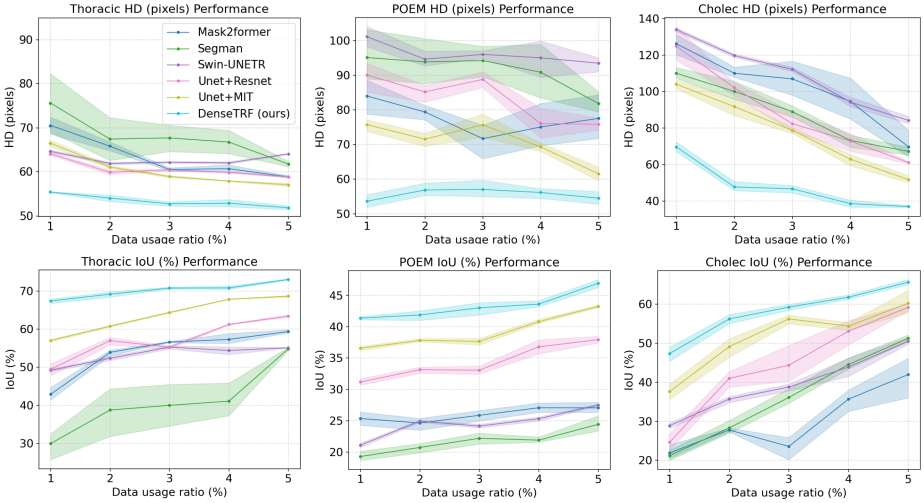

Fig. 3. Comparison with state-of-the-art methods under varying training data ratios using IoU (%) and Hausdorff Distance (HD, pixels) metrics.

3.3 Comparison to state-of-the-art segmentation architectures

We compare DenseTRF against five baseline methods including Mask2Former [2], SegMan [4], Swin-UNETR [6], UNet [19]+ResNet [7], and UNet+MIT [25] across three surgical datasets (Thoracic, Per Oral Endoscopic Myotomy [POEM], and Cholec80) under increasing data-usage ratios from 1% to 5%, evaluating both IoU (%) and Hausdorff Distance (HD, pixels) and DICE score. As shown in Figure 3, across all datasets and supervision levels, DenseTRF consistently demonstrates superior dense prediction accuracy, achieving higher IoU scores while simultaneously reducing boundary error as reflected by lower HD values. The performance gap is particularly pronounced in the low-data regime (1–2%), where DenseTRF maintains stable improvements over competing architectures, indicating enhanced data efficiency and robustness under limited annotation availability.

The mean DICE score across datasets is presented in Table 1, where DenseTRF achieves the highest performance on Thoracic, POEM, and Cholec80 with mDICE values of 81.35 ± 0.33, 51.00 ± 0.79, and 67.40 ± 1.00, respectively. In compar-

Table 2. Comparison with foundation model-based adaptation under low-data regime.

Dataset	Model	Para [MB]	DICE ↑	IoU ↑	HD ↓
Thoracic	SAM[11]	415	73.57 ± 0.47	59.89 ± 0.48	60.06 ± 0.16
	DINO-v1[1]	346	74.77 ± 0.12	61.53 ± 0.14	58.90 ± 0.16
	CLIP[18]	393	77.76 ± 0.32	65.16 ± 0.43	57.76 ± 1.37
	DINO-v3[21]	342	77.66 ± 0.39	65.48 ± 0.47	56.98 ± 0.46
	Slot-TTA[17]	384	73.20 ± 0.54	59.99 ± 0.57	70.73 ± 1.00
	DenseTRF (ours)	388	**79.32 ± 0.48**	**67.35 ± 0.58**	**55.35 ± 0.09**
POEM	SAM[11]	-	38.38 ± 0.35	31.10 ± 0.30	78.12 ± 2.53
	DINO-v1[1]	-	39.96 ± 0.34	32.73 ± 0.30	71.36 ± 0.94
	CLIP[18]	-	43.10 ± 0.40	35.81 ± 0.39	70.41 ± 2.25
	DINO-v3[21]	-	46.55 ± 0.67	39.06 ± 0.62	63.82 ± 4.61
	Slot-TTA[17]	-	29.67 ± 0.51	25.11 ± 0.44	61.85 ± 3.67
	DenseTRF (ours)	-	**49.03 ± 0.43**	**41.34 ± 0.27**	**53.60 ± 1.82**
Cholec	SAM[11]	-	43.36 ± 1.55	34.90 ± 1.14	116.19 ± 1.70
	DINO-v1[1]	-	46.67 ± 2.21	37.18 ± 1.82	114.11 ± 0.90
	CLIP[18]	-	54.28 ± 2.13	44.00 ± 2.00	108.02 ± 4.08
	DINO-v3[21]	-	55.49 ± 1.98	46.39 ± 2.03	94.24 ± 3.07
	Slot-TTA[17]	-	37.40 ± 0.81	30.53 ± 0.78	88.92 ± 2.54
	DenseTRF (ours)	-	**57.83 ± 1.97**	**47.29 ± 1.83**	**69.66 ± 2.61**

ison to the best among SOTA, UNet+MIT, we improve by 4.85 mean DICE for Thoracic ($p < 0.05/5$, with Bonferroni correction $m = 5$), 4.63 for POEM ($p < 0.01$), and 7.84 for Cholec ($p < 0.01$). Collectively, these results demonstrate that DenseTRF provides consistent gains across diverse surgical domains, highlighting its effectiveness in improving dense prediction performance under limited ground truth supervision signal access.

3.4 Compare to foundation model based adaptation

Following the comparison with state-of-the-art segmentation architectures, we further evaluate DenseTRF against representative foundation models, including SAM [11], DINO-v1 [1], DINO-v3 [21] , and CLIP [18], as well as the slot-based adaptation baseline (Slot-TTA) [17]. To ensure a fair comparison, all foundation models are equipped with the same lightweight MLP dense prediction head used in our method, and training is conducted under an extreme low-data regime using only 1% of labeled samples. Each experiment is repeated across five independent runs.

As shown in table 2, across all datasets, DenseTRF consistently achieves the best performance. On the Thoracic dataset, DenseTRF improves DICE to 79.32%, surpassing the strongest foundation baseline (CLIP/DINO-v3) by approximately 1.6–1.7 points while also yielding the lowest boundary error (HD). Similar trends are observed on POEM, where DenseTRF reaches 49.03% DICE and reduces HD by over 10 pixels compared to DINO-v3, demonstrating robust adaptation under severe data scarcity. On the Cholec dataset, DenseTRF again delivers the highest DICE (57.83%, $p < 0.05$) and substantially improves boundary quality, reducing HD by more than 24 pixels relative to DINO-v3. Slot-TTA tends to struggle across all domains because it relies on a high-quality slot decoder; although its mask-classification and remixing approach performs well on

Table 3. Ablation study across different datasets.

	Thoracic		POEM		Cholec	
Method	DICE ↑	HD ↓	DICE ↑	HD ↓	DICE ↑	HD↓
w/o SA	77.86 ± 0.87	57.41 ± 1.22	44.95 ± 0.40	64.86 ± 3.40	56.94 ± 1.55	95.71 ± 2.09
w SA w/o ada.	76.34 ± 0.87	60.43 ± 1.10	42.76 ± 0.16	63.08 ± 2.93	53.33 ± 0.93	101.30 ± 1.71
w/o concat	77.86 ± 0.09	60.88 ± 0.32	42.57 ± 0.26	66.76 ± 1.23	55.07 ± 0.74	102.72 ± 1.06
Our full	$\mathbf{79.32 \pm 0.48}$	$\mathbf{55.35 \pm 0.09}$	$\mathbf{49.03 \pm 0.43}$	$\mathbf{53.60 \pm 1.82}$	$\mathbf{57.83 \pm 1.97}$	$\mathbf{69.66 \pm 2.61}$

synthetic images, it does not transfer effectively to the variability of real-world images.

These results indicate that simply attaching a segmentation head to strong foundation representations is insufficient for reliable dense prediction under distribution shift. In contrast, DenseTRF effectively refits texture-aware grouping to the target domain, yielding consistent gains in both region overlap and boundary accuracy across diverse surgical procedures.

3.5 Ablation study

We conduct ablation studies to analyze the contribution of each component in DenseTRF, including SA representation learning, test-distribution adaptation, and multi-source representation concatenation. All variants share the same backbone and training protocol to ensure a controlled comparison.

w/o SA: Our full architecture trained without the SA reconstruction loss and without initializing from a pretrained SA. This baseline underperforms the full model on all datasets, with notably lower Dice on POEM and Cholec, and higher HD on Thoracic and Cholec, e.g., Thoracic DICE drops from 79.32% to 77.86% and Cholec DICE from 57.83% to 56.94%.

w SA w/o ada: The model is initialized with a pretrained SA model but is not adapted to the test distribution during training. Interestingly, this often leads to worse performance compared to w/o SA, especially on Thoracic (DICE drops from 77.86 to 76.34) and POEM (DICE 42.76 vs. 44.95). This indicates that simply incorporating a pretrained SA without task-specific adaptation can be detrimental, likely due to a domain shift.

w/o concat: This variant uses the same training and initialization strategy as the full model, but instead of concatenating the adapted feature with the SA mask and reconstructed feature, it relies solely on the adapted feature. The results show a clear drop in performance compared to the full model across all metrics, particularly in boundary quality (Cholec HD: 102.72 vs 69.66, $p < 0.01$), highlighting the importance of fusing multi-modal cues (feature, mask, reconstruction) for accurate segmentation.

These ablations show that DenseTRF's gains arise from the combination of slot grouping, target-domain adaptation, and feature fusion. Pretrained slots alone are insufficient, while the full model improves both overlap and boundary accuracy, highlighting the value of adapted slot representations for surgical dense prediction.

4 Conclusion

We presented DenseTRF, a texture-aware test-distribution adaptation framework for surgical dense prediction that combines slot-based object-centric representations with unsupervised periodic model merging. By specializing texture-centric slots to the target domain without annotations and fusing encoder features, reconstructions, and attention masks, DenseTRF effectively mitigates distribution shift. Experiments on Thoracic, POEM, and Cholec datasets demonstrate consistent improvements over state-of-the-art segmentation and foundation model baselines under extreme low-data regimes. Ablation results further validate the contribution of each component. Overall, DenseTRF highlights object-centric learning as a promising strategy for robust domain-adaptive surgical scene understanding.

Acknowledgements. This work was supported by the Linda Pechenik Montague Investigator Award, the American Surgical Association Foundation Fellowship, and Penn AI fellowship. We also acknowledge Kenta Nakahashi and Surgical AI Research Academy of University Health Network for sharing the thoracic dataset.

Disclosure of interests. The authors have no competing interests to declare.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9630–9640 (2021)
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 1280–1289 (2022)
3. Didolkar, A., Zadaianchuk, A., Goyal, A., Mozer, M., Bengio, Y., Martius, G., Seitzer, M.: On the transfer of object-centric representation learning. In: The Thirteenth International Conference on Learning Representations (2025)
4. Fu, Y., Lou, M., Yu, Y.: Segman: Omni-scale contextual modeling for semantic segmentation. In: CVPR (2024)
5. Hashimoto, D.A., Rosman, G., Rus, D., Meireles, O.R.: Artificial intelligence in surgery: promises and perils. Annals of surgery **268**(1), 70–76 (2018)
6. Hatamizadeh, A., et al.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: MICCAI (2022)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778 (2016)
8. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. ArXiv preprint **abs/2012.12453** (2020)

9. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics **7**(1), 29 (2016)
10. Kakogeorgiou, I., Gidaris, S., Karantzalos, K., Komodakis, N.: SPOT: self-training with patch-order permutation for object-centric learning with autoregressive transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 22776–22786 (2024)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3992–4003 (2023)
12. Liao, G., Jogan, M., Eaton, E., Hashimoto, D.A.: Forla: Federated object-centric representation learning with slot attention. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
13. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
14. Madani, A., Namazi, B., Altieri, M.S., Hashimoto, D.A., Rivera, A.M., Pucher, P.H., Navarrete-Welton, A., Sankaranarayanan, G., Brunt, L.M., Okrainec, A., et al.: Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Annals of surgery **276**(2), 363–369 (2022)
15. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods **21**(2), 195–212 (2024)
16. Nguyen, M., Wang, A.Q., Kim, H., Sabuncu, M.R.: Adapting to shifting correlations with unlabeled data calibration. In: European Conference on Computer Vision. pp. 230–246. Springer (2024)
17. Prabhudesai, M., Goyal, A., Paul, S., van Steenkiste, S., Sajjadi, M.S.M., Aggarwal, G., Kipf, T., Pathak, D., Fragkiadaki, K.: Test-time adaptation with slot-centric models. In: International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. vol. 202, pp. 28151–28166 (2023)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. vol. 139, pp. 8748–8763 (2021)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
20. Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C., He, T., Zhang, Z., Schölkopf, B., Brox, T., Locatello, F.: Bridging the gap to real-world object-centric learning. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 (2023)
21. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. ArXiv preprint **abs/2508.10104** (2025)

22. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
23. Wang, D., Shelhamer, E., Liu, S., Olshausen, B.A., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
24. Wang, Q., Fink, O., Gool, L.V., Dai, D.: Continual test-time domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 7191–7201 (2022)
25. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 12077–12090 (2021)
26. Xie, Q., Li, Y., He, N., Ning, M., Ma, K., Wang, G., Lian, Y., Zheng, Y.: Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. IEEE Transactions on Medical Imaging **43**(1), 4–14 (2024)
27. Yang, E., Tang, A., Shen, L., Guo, G., Wang, X., Cao, X., Zhang, J.: Continual model merging without data: Dual projections for balancing stability and plasticity. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)