Future Slot Prediction for Unsupervised Object Discovery in Surgical Video

Guiqiu Liao[1], Matjaz Jogan[1], Marcel Hussing[2], Edward Zhang[1,2], Eric Eaton[2], and Daniel A. Hashimoto[1,2]

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

Abstract. Object-centric slot attention is an emerging paradigm for unsupervised learning of structured, interpretable object-centric representations (slots). This enables effective reasoning about objects and events at a low computational cost and is thus applicable to critical healthcare applications, such as real-time interpretation of surgical video. The heterogeneous scenes in real-world applications like surgery are, however, difficult to parse into a meaningful set of slots. Current approaches with an adaptive slot count perform well on images, but their performance on surgical videos is low. To address this challenge, we propose a dynamic temporal slot transformer (DTST) module that is trained both for temporal reasoning and for predicting the optimal future slot initialization. The model achieves state-of-the-art performance on multiple surgical databases, demonstrating that unsupervised object-centric methods can be applied to real-world data and become part of the common arsenal in healthcare applications[1].

Keywords: Surgical video · Object-centric learning · Self-supervision

1 Introduction

Unsupervised object-centric learning seeks to bind features into modular latent representations of objects in visual scenes. Building on the foundations of variational auto-encoders [8], disentangled representation learning [5], iterative attention mechanisms [16] and contrastive learning [10], Slot Attention (SA) emerged as a dominant paradigm for object-centric learning [12,17,3,13,7,23]. In SA, an attention mechanism is used to aggregate image pixels or features into slots, iteratively refining a representation optimized to map pixels to objects as in a set prediction problem using a reconstruction objective.

In the temporal domain, recent efforts added object dynamics to self-supervised objectives [9,19,2,18]. However, the permutation-invariant nature of set prediction in SA poses challenges for the temporal consistency of slot allocations. Surgical videos are a domain where these methods typically fail due to

[1] Code and models are publicly available at: https://github.com/PCASOlab/Xslot.

the non-uniform motion and intermittent visibility of instruments and anatomical structures. Recurrent architectures exploiting temporal feature similarity [9,19,24] and methods based on optical flow [9] capture only limited, short-term causal interactions, and tend to drift over long sequences, struggling to capture long-range dependencies. Methods that batch process entire videos [18] increase computational demands. Slot-BERT [11] adopts the pre-training paradigm of large language models (LLMs), treating slot embeddings as word tokens to improve temporal coherence. All the aforementioned methods use a fixed number of slots per dataset, which can lead to over- or under-decomposition of the scene. Adaptive slot attention [3] improves object discovery by dynamically adjusting the number of slots to account for scene variability, yet it is ineffective on complex, rapidly changing surgical videos.

To address this, we propose a new architecture with a dynamic slot initialization module that predicts the appropriate number and allocation of slots in future frames. Inspired by next-token prediction in large language models, the Dynamic Temporal Slot Transformer (DTST) processes sequences of slot embeddings and learns to predict future slots via a masked auto-encoding objective, effectively performing bidirectional temporal reasoning. Importantly, future slot prediction is jointly learned both as an initial slot allocation policy and as a post-refinement mechanism, further ensuring consistency and stability. Our model can operate on larger context window with minimal computational overhead by predicting slots in a latent space. With a two-stage training objective we achieve convergence and robustness across varying scenes. We validate our approach on multiple surgical video datasets, achieving state-of-the-art performance in unsupervised segmentation and localization of surgical instruments.

2 Methods

Fig. 1a shows an overview of our approach. The model processes videos of arbitrary length and iteratively operates on a buffered latent embedding of length T, which is also the length of the attention window. A frame sequence $\{I_t\}_{t=1}^{T} \in \mathbb{R}^{W\times H\times C\times T}$, where each $I_t \in \mathbb{R}^{W\times H\times C}$ represents a frame at time step t, is encoded to obtain features $X \in \mathbb{R}^{N\times D_{feature}\times T}$. Following a recurrent iterative attention step we obtain the slots representation $S = \{s_t\}_{t=1}^{T} \in \mathbb{R}^{K\times D_{slot}\times T}$, where s_t are the latent space slots that embed objectness of image I_t. Slots S are then fed to DTST encoder and a merger module that aggregates the between-slots information and calculate S_{merged}, allocating the redundant slots to new objects that entered the scene, removing the slots of objects that exit the scene, and merging multiple slots of different parts of the same object. A slot decoder then recurrently maps each element s_t of S_{merged} back to the video encoding space $X_{recon} \in \mathbb{R}^{N\times D_{feature}\times T}$. Simultaneously, the object segmentation masks for each slot are reconstructed. The objective is to minimize the reconstruction loss

$$\mathcal{L}_{\text{recon}} = \|X_{\text{recon}} - X\|^2 \tag{1}$$

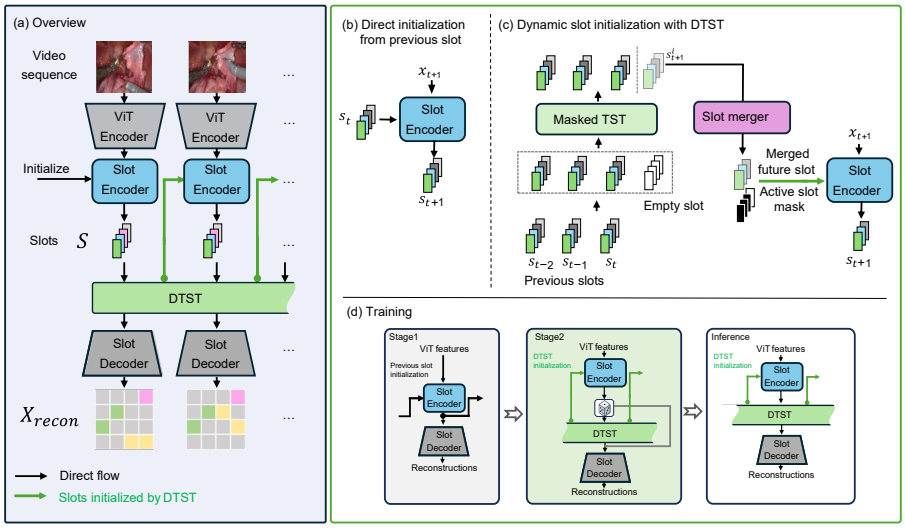

Fig. 1. (a) An overview of the proposed approach using DTST. (b) Direct initialization of slots at time $t+1$ from slot predictions at time t. (c) Dynamic slot initialization using DTST. (d) The overall training pipeline.

which measures the distance between the reconstructed and original features X [17]. Feature-level reconstruction is more robust to variations in scene appearance than image reconstruction.

2.1 Slot encoder

The slot encoder f_{SA} [12] groups image features $x \in \mathbb{R}^{N \times D_{feature}}$ into K spatial groupings:

$$s := f_{\text{SA}}(x, s^{i}) \tag{2}$$

where $s^{i} \in \mathbb{R}^{K \times d_{slot}}$ is the initialization slot. K is a predefined constant to determine the size of attention query size.

2.2 Slot decoder

Multilayer Perceptron (MLP) broadcast decoders efficiently decode features and slot masks [22,17]. Each of the K slots is broadcast to match the number of spatial patches N, resulting in N tokens for each slot. A learnable positional encoding is added to each token; these are then processed independently using a shared MLP to output reconstructed features $\hat{x}_k$ and associated alpha masks α_k indicating the slot's attentive region. The final reconstruction $x \in \mathbb{R}^{N \times D_{\text{feature}}}$ is obtained by a weighted sum:

$$x = \sum_{k=1}^{K} \hat{x}_k \odot m_k \ , \quad m_k = \text{softmax}_k(\alpha_k) \tag{3}$$

where $\odot$ denotes element-wise multiplication. This simple design is very efficient: as the MLP is shared across slots and positions, m_k is directly inferred as a set of non-overlapping object segmentation masks.

2.3 Pre-training

We pre-train the encoder and decoder using a simple slot assignment via a RNN-like computation, as illustrated in Fig. 1 b. The first frame's slots $s_0^{i=0}$ are initialized randomly by sampling from a standard Gaussian. Slots $s_t^{i=0} = \hat{s}_{t-1}$ for t-th frame are initialized by using the slot prediction from the prior frame, encouraging stability of slot identity across frames.

2.4 Dynamic future slot prediction

After pretraining, a Dynamic temporal slot transformer (DTST) and a slot merger are enabled to facilitate slot interactions across frames. DTST is based on the same transformer architecture as in language models [21,15], replacing word tokens with the video slot embedding vectors $S \in \mathbb{R}^{K \times d_{slot} \times T}$. Importantly, temporal positional embeddings and random masking-out of slots also helps predict missing masked slots, which will be crucial for dynamic slot initialization.

Slot merger. Unlike previous methods including Slot-BERT [11], which use all available slots for reconstruction, we employ a slot merger that resolves the problem of over-grouping or under-grouping given a fixed slot count K. Inspired by slot merging using clustering [1], we design a method that works directly on slot similarity. Given a sequence of slot representations $S \in \mathbb{R}^{K \times d_{\text{slot}} \times T}$, we compute the cosine similarity between each pair of slots $s_i, s_j \in \mathbb{R}^{d_{\text{slot}}}$ as

$$\text{sim}(s_i, s_j) = \frac{s_i \cdot s_j}{\|s_i\| \, \|s_j\| + \epsilon} \tag{4}$$

Thresholding similarity scores results in a binary merge mask identifying similar slots that are then averaged, giving an updated slot representation $\tilde{S}_{merge}$. In addition, a binary mask $m_s \in \{0,1\}^{T \times K}$ indicates which slots remain valid after merging. Our slot-merger module is trained within the unsupervised pipeline with a drop-path mechanism, enabling the DTST and the decoder to adapt to varying slot counts. In contrast to Slot-Roll-Out [25] which requires sparse supervision, our method performs unsupervised merging of redundant slots and integrates object parts solely based on appearance and dynamics.

Dynamic temporal slot transformer. DTST is used for bidirectional temporal reasoning using random masking in training. As illustrated by Fig. 1 c, the DTST module is also used to initialize the recurrent estimation of future slots. To do this, we reformulate the module as next slot prediction by appending empty slots, initialized as zero vectors, to the most recent slot buffer. The DTST module then predicts the future slot using current slot history and a single slot mask for last position. We also apply the slot merger to these predictions. This design offers two key advantages: 1) Existing slots are initialized more accurately, improving the tracking of moving objects; and 2) Adjusting the number of active

slots with the slot merger prior to the competitive SA mechanism allows the slot encoder to form more optimal groupings.

To reinforce model convergence, as shown in Fig. 1 d, we allow the model to stochastically bypass the dynamic DTST before the decoding step while keeping it active in slot initialization. This design allows the module to specialize both as a next-slot predictor and as a temporal reasoner, aiding the decoder in reconstructing both refined (without the slot merger) and original slots, resulting in a more robust training framework. The resulting training framework gains robustness by co-learning the context and the future predictions; aligning thus with modern LLM practices [15] that are more effective than in-context training of architectures like SlotBERT [11].

3 Experiments and results

3.1 Datasets and metrics

We evaluate performance on four surgical video datasets spanning three surgery types: (1) an Abdominal surgery dataset based on the public MICCAI 2022 SurgToolLoc Challenge Data [26] that includes 24,542 training clips and 100 testing clips from animal, phantom, and simulator surgeries; (2) a proprietary Thoracic Robotic Surgery dataset that includes 1,730 training clips and 264 testing clips annotated with instruments sampled at 6 FPS from 40 robotic lung lobectomy videos; (3) a cholecystectomy dataset based on Cholec80 [20] that consists of 5,296 training and 100 testing 30-frame clips sampled at 1 FPS from cholecystectomy videos, where the testing clips are from the CholecSeg8K [6] semantic segmentation dataset; and (4) EndoVis 2017 with 480 clips sampled at 1 FPS (five-frame sequences from abdominal surgery) that is used for zero-shot evaluation. We evaluate our approach by comparing the slot masks produced by the decoder with ground-truth instrument masks using video mean best overlap (mBO-V) [14], frame mean best overlap (mBO-F), mean best Hausdorff Distance (mBHD), and foreground adjusted rand index (FG-ARI) [4]. We also use CorLoc [3] to evaluate the instrument bounding box localization accuracy. mBO-F, mBHD, FG-ARI and CorLoc measure overlap with ground truth while mBO-V additionally evaluates the temporal consistency of instrument segmentation.

3.2 Results

Comparison to state-of-the-art. We compare our method to other unsupervised object representation learning methods for video including SAVi [9], STEVE [19], Slot-Diffusion [23], Video-Saur [24], and Slot-BERT [11]. The only slot attention method in the literature that implements dynamic slot allocation is AdaSlot [3] which was developed for images. We thus reimplement AdaSlot for videos using its original slot prediction layer and an extended prediction layer (AdaSlot-ext) to account for the increased complexity of surgical scenes. As shown in Table 1, our method achieves superior performance across all datasets.

Table 1. Comparison to state-of-the-art unsupervised object-centric learning methods on abdominal, thoracic and cholecystectomy datasets.

	Method	mBO-V	mBO-F	mBHD (↓)	FG-ARI	CorLoc
Abdominal	SAVi[9]	29.4±0.2	33.2±0.1	81.72±0.99	36.6±0.2	40.0±0.5
	STEVE[19]	27.9±0.2	31.5±0.1	139.93±0.62	34.3±0.1	17.0±0.4
	Slot-Diffusion[23]	37.5±0.1	42.2±0.1	70.54±0.25	46.3±0.0	42.0±0.2
	Video-Saur[24]	46.3±0.4	50.1±0.4	53.90±1.28	55.1±0.5	60.0±1.9
	Adaslot[3]	38.2±0.9	41.0±0.6	66.58±0.88	47.9±0.5	55.1±0.4
	Adaslot-ext	36.5±0.5	39.9±0.8	65.03±1.86	46.9±0.9	55.7±0.4
	Slot-BERT[11]	48.9±0.2	52.8±0.2	43.40±0.53	58.2±0.3	70.7±0.8
	Ours	**50.8±0.2**	**54.1±0.5**	**40.21±0.66**	**62.8±0.3**	**71.7±1.0**
Thoracic	SAVi[9]	26.5±0.0	30.7±0.1	119.76±0.17	23.1±0.0	27.3±0.7
	STEVE[19]	28.9±0.0	33.0±0.0	135.41±0.54	26.5±0.0	24.6±0.3
	Slot-Diffusion[23]	31.1±0.1	39.9±0.0	84.56±0.15	31.4±0.1	31.7±0.5
	Video-Saur[24]	21.9±0.1	15.7±0.1	139.46±0.23	11.9±0.1	13.0±0.1
	Adaslot[3]	23.0±0.5	28.4±0.6	116.15±1.16	21.7±0.2	25.1±1.6
	Adaslot-ext	33.7±0.3	46.2±0.2	80.51±0.56	37.6±0.3	44.5±0.3
	Slot-BERT[11]	34.0±0.1	40.5±0.1	92.37±0.43	33.8±0.1	34.8±0.7
	Ours	**38.2±0.1**	**52.2±0.1**	**68.17±0.65**	**41.6±0.1**	**54.5±0.3**
Cholecystectomy	SAVi[9]	24.8±0.0	29.4±0.0	96.51±0.21	20.1±0.0	26.7±0.3
	STEVE[19]	26.2±0.1	32.3±0.0	101.55±0.19	24.9±0.1	30.1±0.4
	Slot-Diffusion[23]	27.7±0.2	34.6±0.1	94.70±0.37	26.3±0.2	38.5±0.7
	Video-Saur[24]	30.1±0.3	43.1±0.1	75.82±0.24	34.2±0.1	34.1±0.8
	Adaslot[3]	17.5±0.9	17.8±0.3	95.76±1.22	22.4±0.8	19.0±0.2
	Adaslot-ext	31.6±0.5	31.1±0.3	60.51±0.39	42.4±0.1	33.7±1.0
	Slot-BERT[11]	28.8±0.3	27.8±0.4	64.82±0.90	37.0±0.6	35.3±2.0
	Ours	**33.2±0.8**	**31.9±0.5**	**57.32±0.88**	**43.1±0.9**	**41.2±0.4**

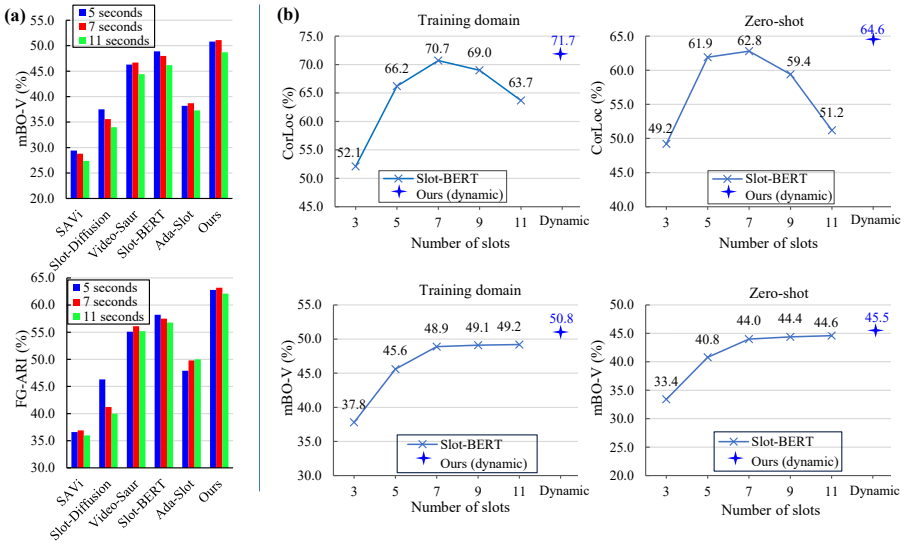

Fig. 2. (a) Quantitative comparison to five state-of-the-art methods on videos of lengths 5, 7 and 11 seconds. (b) Comparison of our dynamic future slot prediction to Slot-BERT with different static slot counts in different domains.

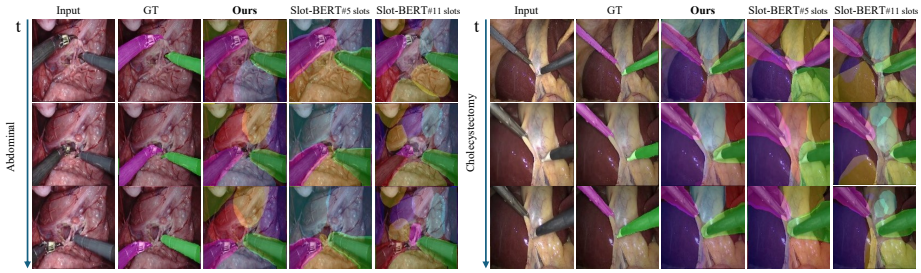

Fig. 3. Qualitative results of unsupervised segmentation using dynamic future slot prediction for abdominal and cholecystectomy surgery videos, compared to Slot-BERT with a fixed number of slots (5 and 11 slots). Ground Truth (GT) on instruments is provided as reference.

Table 2. Ablation study.

			Same domain								Zero-shot			
			Short				Long							
DTST	**Merger**	**X-slot**	mBO-V	mBHD(↓)	FG-ARI	CorLoc	mBO-V	mBHD(↓)	FG-ARI	CorLoc	mBO-V	mBHD(↓)	FG-ARI	CorLoc
			45.6	44.5	59.7	61.7	42.5	44.3	58.5	60.1	39.4	53.5	53.1	48.9
✓			46.2	44.6	60.1	62.2	43.7	44.0	59.9	60.2	40.6	51.7	55.1	49.0
	✓		45.4	41.8	62.8	70.8	45.1	42.7	62.5	70.0	43.4	49.5	57.2	59.1
✓		✓	49.5	**40.0**	**63.2**	66.3	45.9	40.4	61.4	63.9	43.4	47.0	56.5	55.8
✓	✓		49.2	41.0	63.1	71.7	46.5	41.2	**63.1**	72.0	44.3	48.4	**58.0**	60.9
✓	✓	✓	**50.8**	40.2	62.8	71.7	**48.7**	**39.5**	62.1	**73.4**	**45.5**	**44.8**	57.9	**64.6**

The improvement is particularly large for the thoracic dataset, which is also the smallest, suggesting that our method requires significantly less training data. On the abdominal dataset, the CorLoc score of 71.7 approaches the level of performance of supervised methods. Statistical significance of this improvement is confirmed by one tailed t-test (p-values < 0.01 with Bonferroni correction).

Results in Table 1 are calculated on five-second clips. Fig 2(a) demonstrates that our method is consistently better on longer clips (7 and 11 seconds) measured by mBO-V and FG-ARI. Longer clips (29 seconds each) of object discovery and segmentation are available in the Supplemental Material.

Comparison with tuned slot counts. Fig. 2(b) illustrates our method's performance in localization (CorLoc) and video segmentation (mBO-V) compared to the second-best method, Slot-BERT, across different slot counts. On the training domain (Abdominal), Slot-BERT's localization accuracy declines as the number of slots increases despite a slight improvement in video segmentation. In contrast, our method outperforms other methods in both localization and video-level segmentation, all without a fixed slot count. Additionally, our method leads in zero-shot transfer to the EndoVis dataset, achieving the highest CorLoc (64.6) and mBO-V (45.5), thus demonstrating a superior generalization.

Fig 3 further demonstrates the advantage of dynamic slot allocation and advanced temporal reasoning: A fixed slot count used by Slot-BERT tends to over-segment (11 slots) or under-segment (5 slots), while our method finds a balanced slot count that better delineates instruments and tissues.

Table 3. Influence of slot similarity threshold.

Matrics	0.70	0.80	0.85	0.90	0.95	0.99
CorLoc	58.6±1.5	72.3±1.1	**73.0±0.9**	71.7±1.0	70.3±0.6	70.2±0.2
mBO-V	38.9±0.7	48.9±0.4	49.4±0.2	50.8±0.2	50.9±0.2	**51.6±0.5**
FG-ARI	42.8±1.1	59.9±0.6	61.0±0.6	62.8±0.3	62.8±0.7	**63.1±0.6**

Ablation study. Table 2 shows the contributions of different model components: DTST, slot merger (Merger), and their application to next-slot prediction (X-slot) to performance in same domain for a short (5 sec) and long (11 sec) sequence, and in a zero-shot transfer to the EndoVis dataset. Individually, DTST improves temporal consistency (e.g., mBO-V increases from 45.6 to 46.2 in short sequences), while Merger significantly enhances object localization (e.g., CorLoc improves from 60.1 to 70.0 in long sequences). Next slot prediction further boosts performance, particularly in zero-shot settings (CorLoc increases from 60.9 to 64.6). The full model (DTST + Merger + X-slot) achieves the best overall performance, with notable improvements in temporal consistency as reflected by mBO-V. These results highlight the complementary strengths of these components in achieving robust object-centric learning.

Effect of cosine-similarity threshold. We tested robustness to the slot similarity threshold (Table 3). Both mBO-V and FG-ARI improve consistently with higher thresholds, peaking at 0.99, suggesting stricter similarity refines slot assignments. Meanwhile, CorLoc peaks at 0.85 (73.0) and declines for stricter thresholds, as higher thresholds reduce the likelihood of merging multiple slots belonging to the same object. While there's a trade-off between localization and segmentation, thresholds above 0.80 yield stable performance, showing threshold selection is more permissive than slot count selection. Our method also allows a different number of active slots in each frame which is a clear advantage over fixed slot count approaches.

Computational efficiency. The forward time per frame on a single NVIDIA RTX A6000 GPU is 5.6 ms, supporting real-time downstream tasks. While slot merging and future slot prediction add 3.9 ms overhead compared to Slot-BERT (1.7 ms), this remains minor. Even with a 20× larger context window, latency stays under 100 ms thanks to latent-space temporal reasoning.

4 Conclusion

We proposed a novel approach for temporal reasoning and future slot initialization in Slot Attention (SA) that improves object discovery in video and shows promising results on the complex task of surgical video understanding. The SA model is coupled with our DTST module and a slot merger to extract object slots and predict object masks. We train DTST stochastically, making it efficient for both post-attention and for next slot initialization. Experiments show that our approach performs robustly across multiple surgical video domains and video lengths, highlighting its adaptability and generalization.

Slot attention jointly embeds appearance and location in the latent space, allowing slot tracking when objects are partially or temporarily occluded. However, extreme overlaps of objects might be still difficult to disentangle. E.g., instruments with longer, uniform shafts will be correctly outlined after initial occlusion (due to location bias), but might split beyond the context window (i.e., >11 sec). Exploring solutions with a longer context window and a specialized training set is part of our future work.

The state-of-the-art performance demonstrated on multiple surgical databases already opens possibilities for downstream applications that require interpretable surgical video understanding with low computing demands. A larger training dataset drastically improves performance (see results for the abdominal dataset). Thus, incorporating more diverse training data could pave the way for broader applications in healthcare and beyond.

Acknowledgements. This work was supported by the Linda Pechenik Montague Investigator Award and the American Surgical Association Foundation Fellowship.

Disclosure of interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aydemir, G., Xie, W., Güney, F.: Self-supervised object-centric learning for videos. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS (2023)
2. Elsayed, G.F., Mahendran, A., van Steenkiste, S., Greff, K., Mozer, M.C., Kipf, T.: Savi++: Towards end-to-end object-centric learning from real-world videos. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems, NeurIPS (2022)
3. Fan, K., Bai, Z., Xiao, T., He, T., Horn, M., Fu, Y., Locatello, F., Zhang, Z.: Adaptive slot attention: Object discovery with dynamic slot number. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR (2024)
4. Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., Lerchner, A.: Multi-object representation learning with iterative variational inference. In: Proceedings of the 36th International Conference on Machine Learning ICML. Proceedings of Machine Learning Research (2019)
5. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations, ICLR (2017)
6. 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 (2020)
7. Jiang, J., Deng, F., Singh, G., Ahn, S.: Object-centric slot diffusion. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS (2023)

8. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR (2014)
9. Kipf, T., Elsayed, G.F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., Jonschkowski, R., Dosovitskiy, A., Greff, K.: Conditional object-centric learning from video. In: The Tenth International Conference on Learning Representations, ICLR (2022)
10. Kipf, T.N., van der Pol, E., Welling, M.: Contrastive learning of structured world models. In: 8th International Conference on Learning Representations, ICLR (2020)
11. Liao, G., Jogan, M., Hussing, M., Nakahashi, K., Yasufuku, K., Madani, A., Eaton, E., Hashimoto, D.A.: Slot-BERT: Self-supervised object discovery in surgical video. ArXiv preprint (2025)
12. 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, NeurIPS (2020)
13. Mansouri, A., Hartford, J.S., Zhang, Y., Bengio, Y.: Object centric architectures enable efficient causal representation learning. In: The Twelfth International Conference on Learning Representations, ICLR (2024)
14. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (1) (2016)
15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (8) (2019)
16. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (2017)
17. 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)
18. Singh, G., Wang, Y., Yang, J., Ivanovic, B., Ahn, S., Pavone, M., Che, T.: Parallelized spatiotemporal slot binding for videos. In: Forty-first International Conference on Machine Learning, ICML (2024)
19. Singh, G., Wu, Y., Ahn, S.: Simple unsupervised object-centric learning for complex and naturalistic videos. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems, NeurIPS (2022)
20. 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 (1) (2016)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS (2017)
22. Watters, N., Matthey, L., Burgess, C.P., Lerchner, A.: Spatial broadcast decoder: A simple architecture for learning disentangled representations in VAEs. ArXiv preprint (2019)
23. Wu, Z., Hu, J., Lu, W., Gilitschenski, I., Garg, A.: SlotDiffusion: Object-centric generative modeling with diffusion models. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS (2023)

24. Zadaianchuk, A., Seitzer, M., Martius, G.: Object-centric learning for real-world videos by predicting temporal feature similarities. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS (2023)
25. Zhao, Z., Wang, J., Horn, M., Ding, Y., He, T., Bai, Z., Zietlow, D., Simon-Gabriel, C., Shuai, B., Tu, Z., Brox, T., Schiele, B., Fu, Y., Locatello, F., Zhang, Z., Xiao, T.: Object-centric multiple object tracking. In: IEEE/CVF International Conference on Computer Vision, ICCV (2023)
26. Zia, A., Bhattacharyya, K., Liu, X., Berniker, M., Wang, Z., Nespolo, R., Kondo, S., Kasai, S., Hirasawa, K., Liu, B., et al.: Surgical tool classification and localization: results and methods from the MICCAI 2022 SurgToolLoc challenge. ArXiv preprint (2023)