This article has been accepted to the International Conference on Information Processing in Computer-Assisted Interventions, 2024.

Optimizing Latent Graph Representations of Surgical Scenes for Zero-Shot Domain Transfer

Siddhant Satyanaik[a,*], Aditya Murali[a,1,*], Deepak Alapatt[a], Xin Wang[d], Pietro Mascagni[b,c], Nicolas Padoy[a,b]

[a]*ICube, University of Strasbourg, CNRS, France*
[b]*IHU Strasbourg, France*
[c]*Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy*
[d]*West China Hospital of Sichuan University, Chengdu, China*

Purpose: Advances in deep learning have resulted in effective models for surgical video analysis; however, these models often fail to generalize across medical centers due to domain shift caused by variations in surgical workflow, camera setups, and patient demographics. Recently, object-centric learning has emerged as a promising approach for improved surgical scene understanding, capturing and disentangling visual and semantic properties of surgical tools and anatomy to improve downstream task performance. In this work, we conduct a multi-centric performance benchmark of object-centric approaches, focusing on Critical View of Safety assessment in laparoscopic cholecystectomy, then propose an improved approach for unseen domain generalization.

Methods: We evaluate four object-centric approaches for domain generalization, establishing baseline performance. Next, leveraging the disentangled nature of object-centric representations, we dissect one of these methods through a series of ablations (e.g. ignoring either visual or semantic features for downstream classification). Finally, based on the results of these ablations, we develop an optimized method specifically tailored for domain generalization, LG-DG, that includes a novel disentanglement loss function.

Results: Our optimized approach, LG-DG, achieves an improvement of 9.28% over the best baseline approach. More broadly, we show that object-centric approaches are highly effective for domain generalization thanks to their modular approach to representation learning.

Conclusion: We investigate the use of object-centric methods for unseen domain generalization, identify method-agnostic factors critical for performance, and present an optimized approach that substantially outperforms existing methods.

Keywords: Surgical Video Analysis, Domain Adaptation, Object-Centric Learning, Graph Neural Networks

1. Introduction

Surgical video analysis is a rapidly growing field that aims to mine critical information from unstructured surgical videos at scale, which can then be used to enhance various aspects of surgical practice. Over the past decade, many works have harnessed advances in deep learning to achieve proficiency in tasks like automated phase recognition [20], tool and anatomy segmentation [4, 16], and fine-grained action recognition [17, 6]. Yet, practical deployment of these models requires effective *domain generalization*, or in other words, generalization to distinct hospital environments characterized by variations in surgical workflow, instrumentation, camera setups, and patient demographics.

A straightforward approach to this problem is to simply build training datasets comprising annotated videos from numerous surgical centers; however, this is often impractical due to multifaceted difficulties and privacy concerns regarding data collection and sharing [8]. Moreover, even if these concerns can be overcome, annotating these datasets and ensuring consistency across centers is a critical bottleneck, particularly for fine-grained tasks like semantic segmentation and action recognition. As a result, many works in medical and surgical computer vision have approached multicentric generalization as a domain adaptation problem, aiming to adapt models trained for one center (source domain) to other centers (target domains) [19, 21, 10, 22]. In this context, the diverse aforementioned data and annotation availability scenarios can be broadly divided into three different problem settings: (1) Fully Supervised/Supervised Domain Adaptation (SDA), where data and annotations are available for the source and target domain, (2) Unsupervised Domain Adaptation (UDA), where data from the target domain is available but not annotations, and (3) Domain Generalization (DG), where no information about the target domain is known.

While several works have focused on UDA, the DG setting is relatively under-explored; nevertheless, it is especially relevant in the surgical domain, where collecting and sharing data is still extremely challenging. Consequently, in this work, we focus on the latter, specifically aiming to improve performance by leveraging *object-centric classification* approaches. Object-centric methods learn image representations that are

*These authors contributed equally to this work.
[1]Corresponding author: murali@unistra.fr

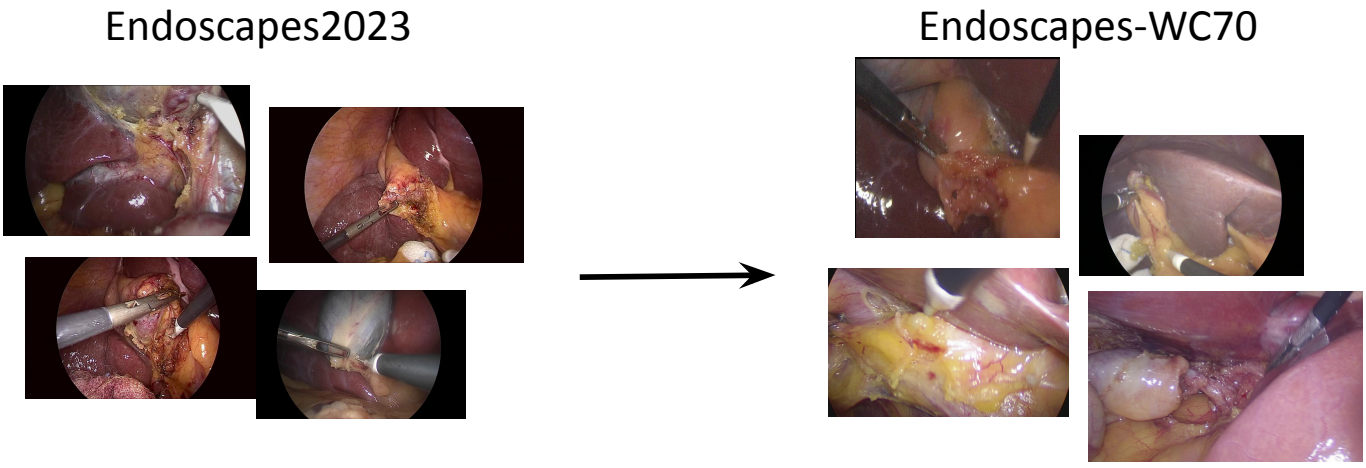

Fig. 1. Qualitative examples to illustrate the visual domain gap between Endoscapes2023 and Endoscapes-WC70. The images differ in aspect ratio, color distribution, and field of view that could be caused by variations in the laparoscope used and in surgical workflow.

strongly conditioned on the objects in the scene, often by including an object detection step prior to classification. These representations can then be finetuned for different downstream tasks, enabling improved performance. Our key observation is that such methods could be highly effective for domain generalization because they explicitly enforce downstream predictions to be conditioned on detected objects rather than extraneous factors like lighting or background artifacts.

To explore this hypothesis, we begin by thoroughly benchmarking domain adaptation performance of four different object-centric methods, starting with the fully supervised and SDA settings to establish ceiling performances, then moving to the DG setting. We focus on Critical View of Safety (CVS) prediction, for two reasons: (1) it is a challenging fine-grained task that has been successfully tackled using object-centric methods [9, 13, 12], and (2) exploration of multicentric generalization in the context of such fine-grained tasks is limited, raising the need for further investigation. To conduct a multicentric analysis, we employ the Endoscapes2023 dataset [11], collected in Strasbourg, France, and additionally introduce **Endoscapes-WC70**, a dataset of 70 laparoscopic cholecystectomy videos collected in Sichuan, China annotated with CVS criteria and segmentation masks. Finally, we propose an object-centric approach for improved domain generalization, **LG-D**omain**G**en **(LG-DG)**, that extends the best single-center method, LG-CVS [13], with a disentanglement loss function to learn more robust representations and thereby improve domain generalization.

In our experiments, we find that object-centric models generally outperform non-object-centric classifiers for domain generalization, and pinpoint critical attributes across methodologies for effective domain generalization. Our proposed approach incorporates these various principles into a single method, thereby outperforming existing approaches.

Our contributions can be summarized as follows:

1. We study 4 object-centric methods in the context of domain generalization, focusing on Critical View of Safety prediction.
2. We propose an improved latent graph-based object-centric approach, LG-DG, that substantially outperforms existing approaches for domain generalization.
3. We introduce the Endoscapes-WC70 dataset, which comprises 7,690 images from 70 videos annotated with CVS criteria, of which 510 are additionally annotated with segmentation masks.

2. Related Work

2.1. Domain Generalization

Annotating surgical data is extremely expensive due to the need for expert knowledge. To alleviate this burden, the research community has actively explored various domain adaptation methodologies, such as unsupervised domain adaptation (UDA) [19, 21], where target data but not labels are used to refine source domain-trained models, and semi-supervised domain adaptation (SSDA) [10, 1], where some of the target data also contains labels. A related research area is federated learning (FL), which restricts data sharing among domains, but allows decentralized model training using all available data. Federated learning has been explored extensively in the medical imaging community [18], and more recently for surgical phase recognition [8].

Unlike UDA, SSDA, and FL, Domain Generalization (DG) eliminates the need for target domain data altogether. DG methods aim to build a single model that can generalize to unseen target domains. Popular approaches include model ensembling, data augmentation/generation to enhance the training set with estimations of out-of-domain data, and most commonly, representation learning-based approaches that seek to learn domain-invariant representations, generally through custom model or loss function design. [3, 2]

Our work focuses on the under-explored area of Domain Generalization in surgical computer vision, falling under the umbrella of Representation Learning-based DG. Unlike previous works, we investigate the use of object-centric representations, which are already disentangled feature representations,

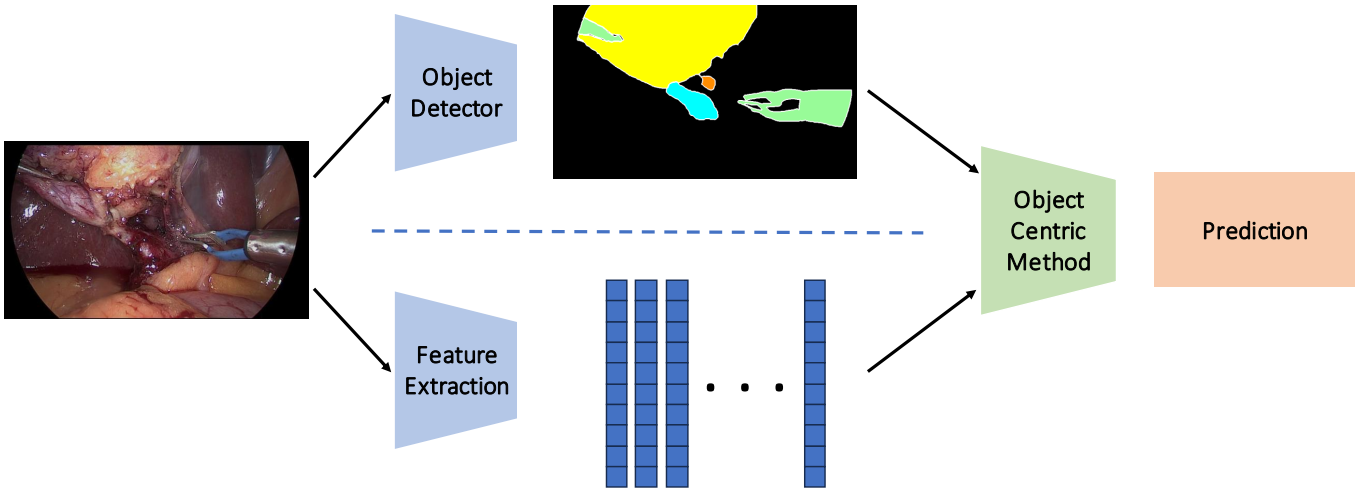

Fig. 2. An illustration of a generic Object-Centric method.

and additionally propose an auxiliary learning objective to further boost feature disentanglement, thereby improving generalization.

2.2. Object-Centric Learning

Object-centric learning focuses on learning scene representations in terms of the present objects, then using these representations for downstream tasks. It has been gaining prominence in the surgical domain, having been applied for activity recognition [5, 14] and phase recognition [7, 12], action triplet recognition [17], scene captioning [15], and most relevantly, CVS prediction [13, 12].

While these works focus on solving specific downstream tasks, we are instead interested in leveraging object-centric learning for improved domain generalization.

3. Methods

In this section, we begin by detailing the datasets used in our multicentric study and describe the CVS prediction downstream task. Then, we describe the four object-centric approaches that we investigate. Finally, we describe our methodology for optimizing LG-CVS [13], a state-of-the-art object-centric method for CVS prediction, to develop our proposed approach: LG-**D**omain**G**en, or LG-DG. This consists of preliminary ablations to understand elements of object-centric methods that aid domain generalization, followed by the development of a disentanglement loss function to boost latent graph robustness, and, as a result, domain generalization performance.

3.1. Datasets and Downstream Task

The Critical View of Safety, or CVS, is a measure to assess dissection quality and exposure of key anatomical structures; proper achievement of the CVS is strongly associated with reduced adverse outcomes, such as bile duct injury. CVS consists of three different binary criteria, detailed in [11], making CVS prediction a multi-label classification problem.

Table 1. Frame-Level Achievement Rates (%) of each CVS Criterion.

Criterion	Endoscapes2023			Endoscapes-WC70		
	Train	Val	Test	Train	Val	Test
C1: Two Structures	15.6	16.3	24.0	2.1	2.0	3.2
C2: HCT Dissection	11.2	12.5	17.1	7.1	7.3	11.8
C3: Cystic Plate	17.9	16.7	27.1	8.4	9.0	14.9

To enable a multi-centric study, we use two datasets: (1) Endoscapes2023, introduced in [11] and collected in Strasbourg, France, and (2) Endoscapes-WC70, which we introduce here, collected in Sichuan, China. Endoscapes2023 contains 58585 frames from 201 laparoscopic cholecystectomy videos; we use two subsets of Endoscapes2023: Endoscapes-CVS201, which contains 11090 frames annotated with CVS, and Endoscapes-Seg201, which contains segmentation masks for 1933 of the previous 11090 frames. We adopt the official dataset splits, using 120 videos for training, 41 for validation, and 40 for testing. Meanwhile, we construct Endoscapes-WC70 by collecting 70 laparoscopic cholecystectomy videos from the West China Hospital, Sichuan University, Sichuan, China; following the protocol of Endoscapes2023 [11], we annotate CVS at 5 second intervals and segmentation masks at 30 second intervals. Finally, we split the 70 videos into 40 training, 15 validation, and 15 test videos, applying stratified sampling based on video-level CVS achievement as done in [13]. Table 1 shows the per-criterion CVS achievement rates for the two datasets. Endoscapes-WC70 is particularly imbalanced with respect to CVS achievement; we discuss the impact of this imbalance in our results.

3.2. Baseline Models

We evaluate 4 different object-centric classification methods: LG-CVS [13], DeepCVS [9], LayoutCVS (an ablation of DeepCVS introduced in [13]), and ResNet50-DetInit (also introduced in [13]); to ensure fair comparisons, we train a single Mask-RCNN object detector to be used by all the methods. We additionally evaluate a vanilla ResNet50

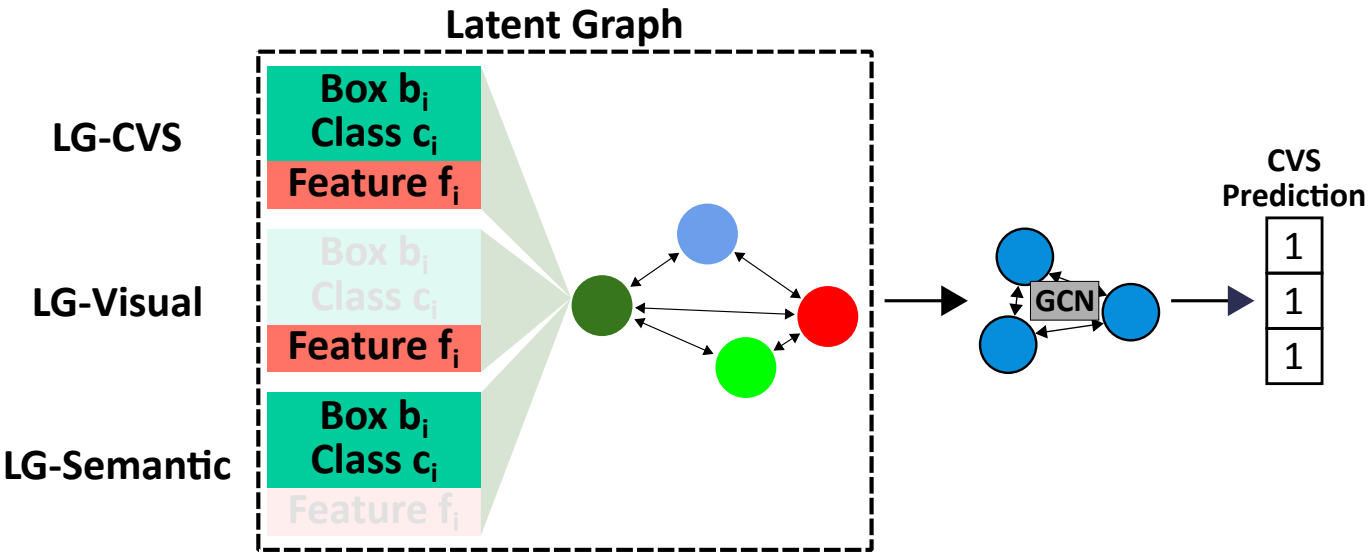

Fig. 3. Three different examples of the masked latent graph $\hat{G}_{\text{CVS}}$ that is passed to the downstream classification head ϕ_{CVS} for CVS prediction. Each node in the graph corresponds to an object in the image. The masking operation, while pictured for a single node, is applied to all nodes.

model to isolate improvements in generalization brought by object-centric modeling. We briefly describe each approach below:

ResNet50: We finetune an ImageNet-pretrained ResNet50 classifier for CVS prediction, requiring no additional labels in the source domain. This baseline allows us to study whether object-centric approaches as a whole can enable better generalization.

DeepCVS [9]: DeepCVS is a two-stage model that consists of first segmenting an image, then concatenating the predicted segmentation mask with the original image and processing the result with a convolutional neural network to predict CVS. In doing so, the downstream CVS predictions are based on both semantic information (identity and location of various anatomical structures, encoded as a segmentation mask), and visual information (from the original image). For fair comparisons, we use the adapted and better performing version of DeepCVS presented in [13].

LayoutCVS: LayoutCVS is identical to DeepCVS except that it does not concatenate the original image with the output of the object detector. Including this model allows us to examine the varying robustness of semantic and visual features to shifts in input domain for downstream classification, especially through comparisons with DeepCVS.

ResNet50-DetInit: We initialize a ResNet50 classifier using trained object detection model weights, then finetune for CVS prediction, thus utilizing the fine-grained scene information encoded implicitly in the object detector's learned visual features. This baseline allows us to isolate the impact of visual information, but unlike the ResNet50 baseline, the visual information is sufficiently fine-grained for effective CVS prediction.

LG-CVS: A two-stage approach for object-centric classification: in the first stage, an encoder Φ_{LG} encodes an image I as a latent graph G, wherein nodes represent objects (tools and anatomical structures) and edges capture 2D geometric relationships between objects. These nodes and edges contain object/relation-specific visual features as well as semantic features (bounding box coordinates and class probabilities). The latent graph G is then passed to a GNN-based classification head ϕ_{CVS} for downstream CVS prediction, and an auxiliary decoder $\phi_{\mathcal{R}}$ that reconstructs the original image given G and a backgroundized version of I (foreground regions replaced with noise). This auxiliary reconstruction objective helps learn more discriminative object features, improving downstream performance.

3.3. Investigating Feature Disentanglement in the Latent Graph

A key facet of LG-CVS is its disentanglement of object-level semantic and visual properties, encoded in the latent graph nodes, and image-level visual properties, captured by the backbone feature map. While these properties contribute synergistically in the single-center setting [13], we note that this may not be the case for domain generalization, where errors in individual modes of information can derail model predictions.

To study this effect, we conduct a series of ablations where we mask one or more information categories from the latent graph for both training and testing before passing to the downstream classification head ϕ_{CVS} and the auxiliary reconstruction head $\phi_{\mathcal{R}}$. Figure 3 illustrates the most extreme ablation settings, while Table 2 lists the various latent graph configurations that we evaluate.

CVS Head Ablations. In this first study, we experiment with different masking combinations, but only apply the masking

before the downstream head, leaving the input to the reconstruction head intact. By doing so, we aim to identify whether (1) certain feature categories are more domain-invariant then others and (2) the different feature categories contribute synergistically when the model is exposed to domain shift. Concretely:

$$G = \Phi_{\text{LG}}(I);\ \mathcal{M}(G, c) = \hat{G}_{\text{CVS}};$$

$$\hat{y} = \phi_{\text{CVS}}(\hat{G}_{\text{CVS}});\ \hat{I} = \phi_{\mathcal{R}}(G), \quad (1)$$

where $\mathcal{M}$ is a masking function that replaces the original values with Gaussian noise, c is the feature category to mask, $\hat{G_{\text{CVS}}}$ is the masked latent graph, and $\hat{y}$ is the final CVS prediction.

Reconstruction Head Ablations. We follow the same process as above, but leave the input to the CVS head intact while masking the input to the reconstruction head, thereby optimizing the auxiliary learning objective to maximize domain-invariance in the learned latent graph representations:

$$G = \Phi_{\text{LG}}(I);\ \mathcal{M}(G, c) = \hat{G}_{\mathcal{R}};\ \hat{y} = \phi_{\text{CVS}}(G);\ \hat{I} = \phi_{\mathcal{R}}(\hat{G}_{\mathcal{R}}). \quad (2)$$

Notably, LG-CVS does not include image visual features in the reconstruction input, instead constructing an object-centric feature layout L_{feat} from G alone before decoding with ϕ_{decoder}. For completeness, we introduce a mechanism to incorporate the image features H_{backbone} into L_{feat}, which uses predicted object locations to arrange per-object features into a spatial grid, built on simple concatenation:

$$\hat{L}_{\text{feat}} = [H_{\text{backbone}}; L_{\text{feat}}];\ \hat{I} = \phi_{\text{decoder}}(\hat{L}_{\text{feat}}), \quad (3)$$

where we spatially resize H_{backbone} to match the dimensions of $\hat{L}_{\text{feat}}$.

3.4. Disentanglement Loss

While the feature disentanglement ablation studies can shed light on the effect of domain shift on different types of features, ultimately, we are interested in creating a model that uses all available information. Our key observation is that the masking function $\mathcal{M}$ can also serve as a form of data augmentation during training, and provide robustness to cases where one or more feature categories in G are inaccurate. To do so, we introduce an auxiliary *disentanglement loss function* $\mathcal{L}_{\text{DIS}}$, that is a linear combination of binary cross entropy loss terms each using the prediction from a different masked graph $\hat{G}$. Concretely:

$$\mathcal{L}_{\text{DIS}} = \lambda_{\text{sem}}\mathcal{L}_{\text{CVS}}(\hat{y}_{\text{sem}}, y) + \lambda_{\text{viz}}\mathcal{L}_{\text{CVS}}(\hat{y}_{\text{viz}}, y) + \lambda_{\text{img}}\mathcal{L}_{\text{CVS}}(\hat{y}_{\text{img}}, y), \quad (4)$$

where $\hat{y}_{\text{sem}} = \phi_{\text{CVS}}(\hat{G}_{\text{sem}})$, $\hat{y}_{\text{viz}} = \phi_{\text{CVS}}(\hat{G}_{\text{viz}})$, and $\hat{y}_{\text{img}} = \phi_{\text{CVS}}(\hat{G}_{\text{img}})$, y is the ground-truth CVS label, and λ_{sem}, λ_{viz}, and λ_{img} are loss weighting terms.

3.5. Training

As in [13, 11], we train all models (except ResNet50) in two-stages, first finetuning a COCO-pretrained Mask-RCNN instance segmentation model on the images with segmentation annotations, then freezing this model and training each object-centric classifier using all images with CVS annotations. To address class imbalance, we train with an inverse frequency-balanced binary cross entropy loss. Finally, when training LG-DG, we set $\lambda_{\text{sem}} = 1$, $\lambda_{\text{viz}} = 0.3$, and $\lambda_{\text{img}} = 0.3$.

Table 2. CVS Classifier Head Inputs Ablation Study in the Domain Generalization Transfer Setting: Endoscapes2023 to Endoscapes-WC70

Feature Type			Performance	
Graph Visual	Graph Semantic	Backbone Image	(mAP)	
✓	✓	✓	27.88 ± 0.37	(LG-CVS [13])
✓	✓	✗	31.37 ± 1.72	
✓	✗	✓	26.31 ± 5.32	(LG-Visual)
✗	✓	✗	33.81 ± 1.32	(LG-Semantic)
✗	✓	✓	**34.14 ± 0.72**	
✓	✗	✗	29.36 ± 3.04	

Table 3. Reconstruction Inputs Ablation Study for LG-CVS in the Domain Generalization Setting, Endoscapes2023 to Endoscapes-WC70.

Feature Type			Performance	
Graph Visual	Graph Semantic	Backbone Image	(mAP)	
✓	✓	✓	27.88 ± 0.37	
✓	✓	✗	23.07 ± 2.22	(LG-CVS [13])
✓	✗	✓	27.23 ± 0.61	
✗	✓	✗	**28.60 ± 2.41**	
✗	✓	✓	28.29 ± 2.36	

4. Experiments and Results

As introduced in Section 1, there are traditionally three paradigms for evaluating domain transfer: **Supervised Domain Adaptation**, **Unsupervised Domain Adaptation**, and **Domain Generalization**. Because object-centric models contain two components: an object detector and a classification model, we also consider an additional evaluation setting: **Partially Supervised Domain Adaptation**, where we assume access to CVS labels in the target domain to train the classification model but not segmentation labels to train the object detector[2]. Here, we freeze the Mask-RCNN detector trained on the source domain and train only the second stage of each object-centric method on the target domain. Finally, we replace the Supervised Domain Adaptation setting with Fully Supervised evaluation, as our primary objective is to establish ceiling performances. We also omit the Unsupervised Domain Adaptation setting, which warrants a more thorough investigation in future work.

In Table 4, we analyze the results of the two latent graph ablation studies: CVS Head and Reconstruction Head (Tables 2 and 3 respectively). Then, we move on to the main experiments in subsection 4.2, where we analyze the domain generalization performance of the four original object-centric methods, a ResNet50 non-object-centric baseline, and our proposed LG-DG (Tables 4 and 5). Lastly, Table 6 shows the domain generalization performance of each object detector for reference. We also show the partially supervised and supervised performances in Tables 4, 5, and 6 to better contextual-

[2]This setting represents a very realistic scenario as collecting dense bounding box or segmentation labels is orders of magnitude more expensive than image-level annotations for classification tasks like CVS.

Table 4. Existing Object-Centric Models evaluated in Various Domain Adaptation Settings: Endoscapes2023 to Endoscapes-WC70.

Model	Performance (mAP)		
	Fully Supervised	Partially-Supervised	Domain Generalization
LG-CVS [13]	33.69 ± 4.60	35.63 ± 3.34	28.06 ± 3.40
DeepCVS	34.21 ± 2.42	**37.83 ± 2.87**	29.54 ± 1.46
LayoutCVS	**35.43 ± 1.68**	35.95 ± 4.11	**30.22 ± 1.92**
ResNet50-DetInit	27.38 ± 7.03	32.34 ± 5.11	23.48 ± 3.13
ResNet50	27.58 ± 2.40	-	12.77 ± 2.71
LG-DG (Ours)	**38.21 ± 2.03**	**44.18 ± 2.48**	**33.34 ± 2.22**

Table 5. Existing Object-Centric Models evaluated in Various Domain Adaptation Settings: Endoscapes-WC70 to Endoscapes2023.

Model	Performance (mAP)		
	Fully Supervised	Partially-Supervised	Domain Generalization
LG-CVS [13]	**64.45 ± 1.30**	55.64 ± 0.59	33.83 ± 1.20
DeepCVS	58.80 ± 2.06	42.84 ± 2.32	**44.30 ± 1.70**
LayoutCVS	58.14 ± 0.76	43.76 ± 0.70	43.26 ± 1.40
ResNet50-DetInit	61.86 ± 2.57	**57.63 ± 0.89**	44.24 ± 0.39
ResNet50	52.27 ± 3.45	-	36.11 ± 0.68
LG-DG (Ours)	**67.25 ± 0.90**	**57.91 ± 2.44**	**47.95 ± 2.40**

Table 6. Per-Class Breakdown of Object Detection Performance (Mask-RCNN, Instance Segmentation mAP) in Target Domain. WC70 refers to Endoscapes-WC70.

Trained On	Tested On	Cystic Plate	Calot Triangle	Cystic Artery	Cystic Duct	Gallbladder	Tool	Avg
WC70	WC70	13.1	3.0	11.9	15.2	**44.7**	**55.2**	**23.8**
Endoscapes2023	WC70	**17.5**	**3.9**	**14.1**	**16.3**	39.5	49.1	23.4
WC70	Endoscapes2023	1.8	6.9	4.4	9.4	43.4	36.9	17.13
Endoscapes2023	Endoscapes2023	**9.1**	**22.5**	**13.4**	**19.2**	**65.8**	**57.6**	**31.3**

ize our results.

4.1. Latent Graph Ablation Studies

For both ablation studies, we use Endoscapes2023 as the source domain and Endoscapes-WC70 as the target domain. Table 2 shows the results of the CVS head ablations (different configurations of $\hat{G}_{\text{CVS}}$); we identify two particular settings of interest: (1) LG-Semantic, which omits all visual information, and (2) LG-Visual, which omits semantics and uses all the visual information.

We observe that, in the domain generalization setting, LG-Semantic attains an mAP of 33.81 mAP, surpassing LG-CVS by 21.2%. This shows that the various feature modalities do not contribute synergistically when posed with domain shift. We attribute the particular effectiveness of LG-Semantic to the fact that it only uses the detected objects for downstream CVS prediction, thereby providing robustness to visual domain gap. Studying the domain generalization performance of the Mask-RCNN object detector (see Table 6) further substantiates this notion: the object detector trained on Endoscapes2023 performs effectively on par with that trained on WC70 when tested on WC70. It even outperforms the latter on the four smaller classes, all of which are critical for CVS assessment. Since the object detector generalizes effectively, LG-Semantic performs well. Meanwhile, LG-Visual performs markedly worse than LG-Semantic, and slightly worse than LG-CVS, reiterating the notion that semantic features are more robust to domain shifts than visual features.

Table 3 shows the results of the reconstruction head ablations (different configurations of $\hat{G}_{\mathcal{R}}$). Broadly, we find that including visual features in the reconstruction input, whether graph visual or backbone image features, has a negative impact on domain generalization performance: masking both graph visual features and backbone image features in $G_{\mathcal{R}}$ yields the highest performance. One potential explanation for this trend is that, when visual features are part of the reconstruction input, they learn to encode highly domain-specific information like color and texture distributions, degrading the overall domain invariance of the latent graph G. On the other hand, when only semantic information is included, the reconstruction objective only enforces encoding general, perhaps even instance-agnostic, properties of the objects, as the semantic features do not encode properties like color and texture by construction.

4.2. Main Experiments

Table 4 and Table 5 summarize CVS prediction performance for all three domain adaptation settings, tested on Endoscapes2023 and Endoscapes-WC70 respectively. To recall: Fully Supervised indicates that both the object detector and object-centric classification model are trained and tested on the target domain; Partially Supervised indicates that the ob-

ject detector is trained on the source domain; Domain Generalization indicates that neither object detector nor object-centric classification model is trained on the target domain. We show both Fully Supervised and Partially Supervised performance to illustrate the ceiling performance, which we use to contextualize domain generalization performance.

LG-DG consistently outperforms all baseline methods across all settings: for Domain Generalization, LG-DG demonstrates a mean increase of 9.28% over the best-performing baseline models, LayoutCVS and DeepCVS respectively. Interestingly, both LayoutCVS and DeepCVS are primarily based on semantic features, with LayoutCVS explicitly ignoring visual information and DeepCVS ineffective at leveraging visual information (shown in [13]); this reinforces our conclusions from the Latent Graph ablation study. Importantly, LG-DG vastly outperforms LG-CVS: 41.74% better when tested on Endoscapes2023 and 18.82% better when tested on Endoscapes-WC70 (average 30.28%). This clearly highlights the effectiveness of our proposed optimizations. Altogether, LG-DG takes great strides in closing the domain gap with no information about the target domain: it is 24.54% lower than the ceiling performance for Endoscapes-WC70 (Partially Supervised LG-DG) and 28.70% lower than the ceiling performance for Endoscapes (Fully Supervised LG-DG).

5. Conclusion

We investigate the capabilities of object-centric approaches for domain adaptation specifically in the context of fine-grained classification, using Critical View of Safety prediction to measure performance. We show that object-centric methods are highly effective for Domain Generalization, particularly compared to non-object-centric image classifiers. Then, leveraging the modularity of object-centric approaches, we propose an optimized method, **LG-D**omain**G**en, that includes a novel disentanglement loss to improve robustness to domain shift. Future work should seek to extensively validate these findings on various surgical procedures and tasks and to extend these findings to other settings including UDA and Federated Learning.

6. Acknowledgments

This work was supported by French state funds managed by the ANR within the National AI Chair program under Grant ANR-20-CHIA-0029-01 (Chair AI4ORSafety). This work was granted access to the HPC resources of IDRIS under the allocation AD011013523R1 made by GENCI.

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

Competing Interests: The authors declare no conflict of interest.

Informed Consent: This manuscript does not contain any patient data.

Code Availability: The source code is publicly available at `https://github.com/CAMMA-public/SurgLatentGraph`.

References

[1] Basak H, Yin Z (2023) Semi-supervised domain adaptive medical image segmentation through consistency regularized disentangled contrastive learning. In: MICCAI, Springer, pp 260–270

[2] Chen Z, Pan Y, Ye Y, et al (2023) Treasure in distribution: A domain randomization based multi-source domain generalization for 2d medical image segmentation. In: MICCAI. Springer Nature Switzerland, Cham, pp 89–99

[3] Choi S, Jung S, Yun H, et al (2021) Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: CVPR, pp 11580–11590

[4] Grammatikopoulou M, Flouty E, Kadkhodamohammadi A, et al (2021) Cadis: Cataract dataset for surgical rgb-image segmentation. Medical Image Analysis 71

[5] Hamoud I, Jamal MA, Srivastav V, et al (2023) St(or)^2: Spatio-temporal object level reasoning for activity recognition in the operating room. In: Medical Imaging with Deep Learning

[6] Hao L, Hu Y, Lin W, et al (2023) Act-net: Anchor-context action detection in surgery videos. In: MICCAI, Springer, pp 196–206

[7] Holm F, Ghazaei G, Czempiel T, et al (2023) Dynamic scene graph representation for surgical video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 81–87

[8] Kassem H, Alapatt D, Mascagni P, et al (2022) Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. IEEE Transactions on Medical Imaging

[9] Mascagni P, Vardazaryan A, Alapatt D, et al (2021) Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Annals of Surgery

[10] Mottaghi A, Sharghi A, Yeung S, et al (2022) Adaptation of surgical activity recognition models across operating rooms. In: MICCAI, Springer, pp 530–540

[11] Murali A, Alapatt D, Mascagni P, et al (2023) The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv preprint arXiv:231212429

[12] Murali A, Alapatt D, Mascagni P, et al (2023) Encoding surgical videos as latent spatiotemporal graphs for object and anatomy-driven reasoning. In: MICCAI, Springer, pp 647–657

[13] Murali A, Alapatt D, Mascagni P, et al (2023) Latent graph representations for critical view of safety assessment. IEEE Transactions on Medical Imaging pp 1–1

[14] Özsoy E, Czempiel T, Holm F, et al (2023) Labrad-or: Lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms. arXiv preprint arXiv:230313293

[15] Pang W, Islam M, Mitheran S, et al (2022) Rethinking feature extraction: Gradient-based localized feature extraction for end-to-end surgical downstream tasks. IEEE Robotics and Automation Letters 7(4):12623–12630

[16] Sestini L, Rosa B, De Momi E, et al (2023) Fun-sis: A fully unsupervised approach for surgical instrument segmentation. Medical Image Analysis 85:102751

[17] Sharma S, Nwoye CI, Mutter D, et al (2023) Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. In: MICCAI, Springer, pp 505–514

[18] Sohan MF, Basalamah A (2023) A systematic review on federated learning in medical image analysis. IEEE Access

[19] Srivastav V, Gangi A, Padoy N (2022) Unsupervised domain adaptation for clinician pose estimation and instance segmentation in the operating room. In: Medical Image Analysis

[20] Twinanda AP, Shehata S, Mutter D, et al (2016) Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging 36(1):86–97

[21] Wang Q, Bu P, Breckon TP (2019) Unifying unsupervised domain adaptation and zero-shot visual recognition. In: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1–8

[22] Xu J, Zhang Q, Yu Y, et al (2022) Deep reconstruction-recoding network for unsupervised domain adaptation and multi-center generalization in colonoscopy polyp detection. Computer Methods and Programs in Biomedicine 214:106576