This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Gaussian Spatial Priors for Anatomy-Aware Object Detection in Surgical Videos

Yunfan Li[1], Artem Shmelev[2], and Himanshu Gupta[1]

[1] Stony Brook University, Stony Brook NY 11794, USA
yunfan.li@stonybrook.edu
[2] Stony Brook University Hospital, Stony Brook NY 11794, USA

Abstract. Detecting anatomical structures in surgical video is essential for intraoperative safety frameworks such as the Critical View of Myopectineal Orifice (CVMPO) in inguinal hernia repair. While prominent structures like the Cooper's Ligament and Triangle of Doom are reliably detected by standard methods, smaller structures such as the epigastric vessels remain challenging due to their visual ambiguity and intermittent visibility. We observe that the spatial relationship between structures is anatomically constrained, and propose a Gaussian Spatial Prior (GSP) module that encodes this relationship as a compact, parametric bias injected into the self-attention of a DAB-DETR decoder. The prior is computed offline from training annotations as a small set of frozen Gaussian parameters and recomputed at each decoder layer using the iteratively refined reference points. On a dataset of inguinal hernia repair videos with 5-fold cross-validation, GSP improves dependent class detection by +33.5% (AP_{50}) over DAB-DETR and +53.9% over YOLOv26, while also improving anchor detection by +6.0%. These gains are statistically significant across all folds ($p = 0.012$, paired $t-$test).

Keywords: Surgical Scene Understanding · Gaussian Spatial Prior.

1 Introduction

Inguinal hernia repair is one of the most common surgical procedures, with approximately 800,000 operations performed annually in the United States [3]. Despite well-documented advantages of minimally invasive surgery (MIS)—including faster recovery, reduced pain and wound complications, and the ability to diagnose bilateral and occult hernias through a single posterior mesh reinforcement [4]—the MIS approach accounts for only a minority of repairs, with open mesh repair (Lichtenstein-Amid) remaining the dominant technique [3]. The MIS approach carries a significant learning curve, estimated at 50–100 procedures for laparoscopic repair [4, 6] and 20–35 for robotic-assisted repair [7]. Surgeons early in this curve experience elevated rates of complications and recurrence, frequently attributable to insufficient mesh overlap of the myopectineal orifice (MPO), inadequate parietalization of cord structures, and missed anatomical defects [8, 9].

To mitigate these failure modes, Daes and Felix [10] proposed the Critical View of the Myopectineal Orifice (CVMPO), a standardized framework encompassing the identification of key anatomical landmarks and the completion of essential dissection steps prior to mesh placement. Subsequent work by Claus et al. [11] and Marmolejo et al. [12] augmented CVMPO with scoring systems anchored to five triangular zones delineated by an inverted-Y configuration of anatomical structures. Analogous to the Critical View of Safety (CVS) in laparoscopic cholecystectomy [13], CVMPO aims to reduce adverse events by enforcing systematic anatomical verification. Yet unlike CVS, which has received substantial attention from the computer vision community [16–19], automated CVMPO assessment remains largely unexplored.

The only prior work addressing AI-based CVMPO assessment is that of Takeuchi et al. [14], who applied a standard DETR [1] model to identify anatomical landmarks from the first seven CVMPO steps, achieving a mean average precision (mAP) of 51.2% at the frame level. Their approach applies off-the-shelf detectors without incorporating the structured spatial relationships between landmarks, leaving considerable room for improvement—particularly in a safety-critical surgical context.

A fundamental challenge in surgical anatomy detection is that many anatomical structures are small, lack distinctive visual features, and are intermittently visible due to occlusion by tissue or instruments. Standard object detectors treat each class independently, relying on content-driven attention to learn inter-object relationships entirely from data. When training data is scarce—as is inherent to rare surgical conditions and novel clinical formulations—this purely data-driven approach is insufficient. However, the spatial layout of posterior groin anatomy is well characterized: given the positions of consistently visible structures such as Cooper's Ligament (CL) and the Triangle of Doom (DOOM), the expected locations of less prominent landmarks follow predictable, anatomically constrained distributions.

In this work, we propose to encode this domain knowledge as a parametric spatial prior and inject it directly into the detection architecture. Specifically, we augment DAB-DETR [2], a transformer-based detector that uses explicit 4D anchor box coordinates as queries, with a Gaussian Spatial Prior (GSP) module. The GSP encodes compact Gaussian distributions over normalized spatial offsets between anatomical class pairs from training annotations, and biases the decoder's self-attention toward spatially plausible query configurations at every refinement layer.

We evaluate our approach on a dataset of robotic transabdominal preperitoneal (TAPP) inguinal hernia repair videos annotated with bounding boxes for 3 anatomical structures. To our knowledge, this is the first work to (1) introduce the problem of automated anatomical landmark detection for CVMPO assessment to the computer vision community, and (2) propose a parametric spatial prior mechanism for Transformer-based surgical anatomy detection. Our experiments demonstrate that the GSP substantially improves detection performance

over the standard DAB-DETR baseline, with particular gains on the dependent anatomical classes whose detection benefits most from spatial context.

2 Methods

Our method augments a transformer-based object detector with a structured spatial prior derived from the anatomical layout of surgical scenes. We first describe the partitoining of anatomical classes into anchors and dependents (Section 2.1), and then review the base DAB-DETR architecture whose explicit anchor box queries provide the spatial interface for our prior (Section 2.2). Section 2.3 introduces the Gaussian Spatial Prior (GSP) module, detailing the offline computation of prior distributions, the evaluation of pairwise spatial compatibility at inference, and the injection of the resulting bias into the decoder's self-attention. Finally, we summarize the training procedure in Section 2.4.

2.1 Anchor-Dependent Class Structure

We focus on anatomical structures relevant to the Critical View of Myopectineal Orifice (CVMPO) safety framework for inguinal hernia repair. Of those, Cooper's Ligaments (CL) and Triangle of Doom (DOOM) are two important structures for surgeons to understand the scene and assess the progress of dissection. Moreover, they are consistently detectable and visually prominent across frames. We designate these as anchor classes $\mathcal{A} = \{a_1, a_2\}$. The remaining classes, which tend to be less distinct and intermittently visible, constitute the dependent classes $\mathcal{D}$.

This partitioning reflects a structural property of the inguinal hernia repair scene: given the positions of anchor structures, the locations of dependent structures follow predictable spatial distributions. We encode these distributions as parametric Gaussian priors and inject them as an inductive bias into the detector.

2.2 Base DAB-DETR Architecture

We build on DAB-DETR [2], which directly uses 4D box coordinates $\mathbf{a}_q = (x_q, y_q, w_q, h_q)$ as queries in the Transformer decoder and dynamically updates them layer by layer. Each query comprises a dual representation: a positional query (the anchor box) and a content query (a decoder embedding). Using box coordinates as positional queries provides explicit positional priors that improve query-to-feature similarity, while also enabling modulation of the positional attention map using the anchor width and height. This formulation is central to our approach, as it exposes interpretable spatial coordinates to which anatomical priors can be directly applied.

Encoder. A ResNet-50 backbone extracts features from the input frame. The flattened feature map is processed by a stack of self-attention layers with learned positional encoding.

Decoder. Each decoder layer applies self-attention among object queries and cross-attention to encoder memory. Query positions are derived from sinusoidal embeddings of the current anchor box coordinates, and the width and height components modulate the cross-attention map to achieve scale-aware feature extraction. The anchor boxes are updated layer by layer through residual refinements predicted by a shared bounding box head, progressively approaching the target objects.

2.3 Gaussian Spatial Prior

Self-attention in DETR-based decoders is purely content-driven: inter-object spatial relations must be learned entirely from data. In surgical anatomy detection, where annotated data is scarce and spatial arrangement of structure is anatomically constrained, this imposes an unnecessary learning burden. We address this with a Gaussian Spatial Prior (GSP) module that biases decoder self-attention toward spatially plausible query configurations. The GSP operates in three stages: offline prior computation, pairwise spatial compatibility scoring, and attention bias injection.

Offline Prior Computation. From training annotations, we compute parametric Gaussian distributions over the relative spatial offset between each dependent-anchor class pair. For each pair (d_m, a_n) with $d_m \in \mathcal{D}, a_n \in \mathcal{A}$, we collect co-occurring instances across training data and compute normalized offsets:

$$\delta x_{mn} = \frac{x^{(d)} - x^{(a)}}{w^{(a)}}, \delta y_{mn} = \frac{y^{(d)} - y^{(a)}}{h^{(a)}} \tag{1}$$

where $(x^{(d)}, y^{(d)})$ denotes the center coordinate of a dependent instance and $(x^{(a)}, y^{(a)}, w^{(a)}, h^{(a)})$ denotes the center and dimensions of an anchor instance. Normalization by anchor box size renders the offsets scale-invariant. We then fit a 2D axis-aligned Gaussian per class pair:

$$\mu_{mn} = (\overline{\delta x_{mn}}, \overline{\delta y_{mn}}), \sigma_{mn} = (\mathrm{std}(\delta x_{mn}), \mathrm{std}(\delta y_{mn})) \tag{2}$$

This yields $|\mathcal{D}| \times |\mathcal{A}|$ Gaussians, each described by four scalars $(\mu_x, \mu_y, \sigma_x, \sigma_y)$. Standard deviations are clamped to a minimum of 0.1 to prevent degenerate distributions. These parameters are pre-computed and stored as fixed buffers, and not updated during training.

Pairwise Spatial Compatibility. At each decoder layer, given current reference points $\{\mathbf{r}_i\}_{i=1}^{N_q}$ for all N_q queries, we compute spatial compatibility for every query pair (i, j). The normalized offset from query j to query i is:

$$\Delta x_{ij} = \frac{x_i - x_j}{w_j}, \Delta y_{ij} = \frac{y_i - y_j}{h_j} \tag{3}$$

Each offset $\Delta_{ij} = (\Delta x_{ij}, \Delta y_{ij})$ is evaluated against all $|\mathcal{D}| \times |\mathcal{A}|$ Gaussians. The score of the k-th distribution is:

$$g_k(\Delta_{ij}) = \exp(-\frac{1}{2} \sum_{d \in \{x,y\}} (\frac{\Delta d_{ij} - \mu_k^d}{\sigma_k^d})^2) \tag{4}$$

Scores are aggregated via a class-agnostic maximum over all distributions:

$$s_{ij} = \max_{k=1,\cdots,|\mathcal{D}|\times|\mathcal{A}|} g_k(\Delta_{ij}) \tag{5}$$

This selects the best-matching anatomical relationship for each query pair without requiring class prediction at this stage.

Attention Bias Injection. The compatibility scores are converted to an additive attention bias. We first zero-center each row:

$$\hat{s}_{ij} = s_{ij} - \frac{1}{N_q} \sum_{j'=1}^{N_q} s_{ij'} \tag{6}$$

Since softmax is shift-invariant, a uniform bias has no effect on attention weights; zero-centering ensures that only relative spatial compatibility influences the distribution. Offsets consistent with known anatomical relationships receive positive bias, while implausible offsets receive negative bias.

The centered bias is scaled by a hyperparameter λ and shared across all attention heads:

$$\mathbf{B} = \lambda \cdot \hat{\mathbf{S}} \in \mathbb{R}^{N_q \times N_q} \tag{7}$$

$\mathbf{B}$ is added to the self-attention logits prior to softmax:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{B})\mathbf{V} \tag{8}$$

Because DAB-DETR iteratively refines reference points, the bias is recomputed at every decoder layer using the updated positions, allowing the prior to progressively sharpen as predictions converge.

2.4 Training

We train the model end-to-end using the standard DETR bipartite matching loss, comprising focal loss for classification, L1 loss for box regression, and generalized IoU loss. The Gaussian prior parameters remain fixed while only the base DAB-DETR parameters are optimized. We use 5-fold cross-validation partitioned at the video level to prevent patient overlap between training and validation splits.

3 Experiments

3.1 Dataset and Evaluation Protocol

We evaluate on a proprietary dataset of laparoscopic inguinal hernia repair videos annotated with bounding boxes for 3 key anatomical structures for CVMPO assessment: Cooper's Ligaments (CL), Triangle of Doom (DOOM), and Epigastric Vessels (EPI). The choice of these 3 classes is non-trivial: the relative positions of the three establish the orientation of the entire surgical scene, enabling surgeons to locate other visible or latent anatomical structures.

In total, we selected 619 frames from 119 IHR videos, and performed 5-fold cross validation partitioned at the video level to prevent patient-level data leakage. The frames were manually selected from later stages of the surgery before mesh placement. Dataset statistics are shown in Fig. 1 and Table 4. We use standard COCO-style Average Precision at IoU thresholds 0.50:0.95 (AP) and 0.50 (AP_{50}), reported both per class and as the mean across classes (mAP, mAP_{50}).

3.2 Implementation Details

All DETR-based models use a ResNet-50 backbone pretrained on COCO 2017 dataset. We train for 50 epochs using AdamW with a learning rate of $1e^{-4}$ and weight decay of $1e^{-4}$. The learning rate is reduced by a factor of 10 at epoch 40. The Gaussian prior parameters are computed from training annotations in each fold and stored as frozen buffers. All other training settings are default per original methods. We compare our method with both CNN-based and Transformer-based baselines, and we vary the prior strength $\lambda \in \{1, 3, 5, 7\}$.

3.3 Results

Table 1 and Table 2 present the main detection results. We report per-class AP and AP_{50} alongside mean metrics to distinguish effects on anchor versus dependent classes. We note that for IHR scenes, AP_{50} is more clinically meaningful than strict mAP: the goal is to reliably localize structures to guide the surgeons' attention, not to produce pixel-precise boundaries. Furthermore, unlike general objects, anatomical structures like ligaments and vessels are non-rigid and visually homogeneous, where precise bounding boxes often cannot be determined. All baselines exhibit the same pattern: adequate anchor detection but poor EPI performance. YOLOv26 has the highest mAP, mainly due to strong performance on anchor classes, while having the lowest AP_{50} on EPI. This confirms that strong overall detection does not address the unique challenge of dependent class in IHR scenes.

While DAB-DETR underperforms vanilla DETR, the introduction of GSP consistently improves upon DAB-DETR, achieving best mAP^{50} result at $\lambda = 3$, a +9.3% relative improvement over DAB-DETR and +9.5% over YOLOv26. The improvement is most pronounced on the dependent class: EPI AP_{50} rises to

Table 1. Mean detection performance (mean ± std over 5 folds). †Paired t-test vs. DAB-DETR, $p < 0.05$ (one-sided). **Bold** = best; underlined = second best.

Method	mAP	mAP_{50}
YOLOv26	**.262**±.019	.592±.023
DETR	.253±.023	.615±.044
DAB-DETR	.238±.030	.593±.057
GSP (λ=3)	.256±.021†	**.648**±.034†

Table 2. Per-class detection performance.

	CL (anchor)		DOOM (anchor)		EPI (dependent)	
Method	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
YOLOv26	.335±.044	.793±.070	**.383**±.047	.800±.070	.068±.026	.184±.059
DETR (50ep)	**.340**±.019	.829±.053	.352±.038	.796±.061	.066±.033	.222±.104
DAB-DETR	.316±.039	.802±.047	.331±.042	.766±.084	.067±.034	.212±.095
GSP (λ=3)	.342±.037	**.835**±.054	.349±.026	**.827**±.056	**.078**±.020	**.283**±.070

0.283, surpassing DAB-DETR by +33.5% and YOLOv26 by +53.9%. Crucially, the prior also benefits anchor detection — combined anchor AP_{50} rises from 0.784 (DAB-DETR) to 0.831 (+6.0%), indicating that the spatial prior regularised the full query set toward spatially cohherent configurations. GSP outperforms DAB-DETR on mAP and mAP_{50} in all five folds, and paired $t-$tests confirm significance ($p = 0.012$ for both; one-sided). The prior also reduces cross-fold variability for EPI, stabilizing detection across patient-level splits.

3.4 Ablation: Prior Strength

Table 3 illustrates the effect of prior strength λ on detection performance. Anchor detection is stable across all λ, confirming that the prior does not disturb the detection of anchors, while modulating dependent class detection. Performance peaks at $\lambda = 3$ on most metrics, which we select for comparison.

4 Conclusion

We introduced the Gaussian Spatial Prior, a parameter-free module that injects anatomical spatial knowledge into the self-attention of a DAB-DETR decoder. By encoding the spatial relationship between anchor and dependent structures as a compact set of Gaussian distributions, the GSP biases query interactions toward anatomically plausible configurations without modifying the detector architecture or adding learned parameters. On laparoscopic inguinal hernia repair videos, this improves detection of the challenging epigastric vessels by +33.5%

Table 3. Effect of prior strength λ (mean over 5 folds). Anchor AP_{50} is the average of CL and DOOM AP_{50}. **Bold** = best per column.

λ	Anch. AP_{50}	EPI AP	EPI AP_{50}	mAP	mAP_{50}
1	.821	**.081**	.257	.255	.633
3	**.831**	.078	**.283**	**.256**	**.648**
5	.821	.059	.180	.248	.607
7	.816	.074	.257	.251	.630

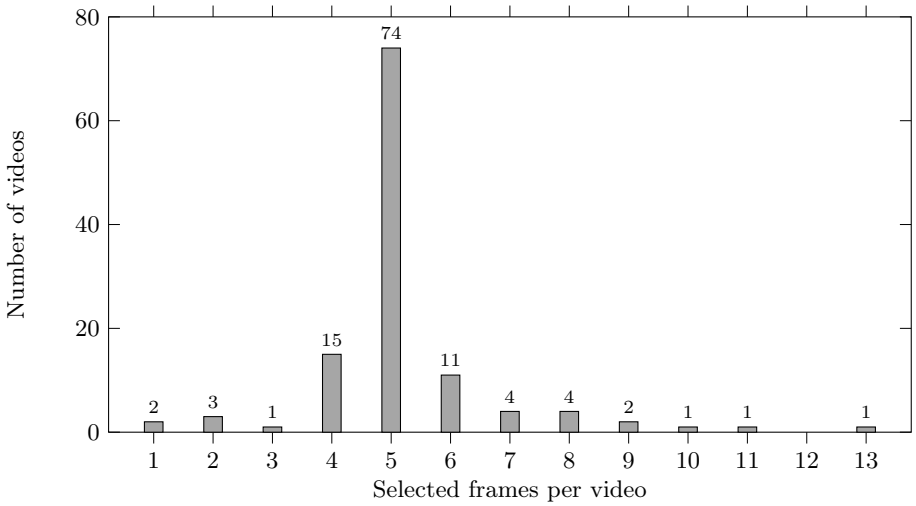


Fig. 1. Distribution of the 619 selected frames across 119 videos. The dataset contains a mean of 5.20 frames per video (median: 5; range: 1–13).

AP50 over DAB-DETR and +53.9% over YOLOv26, with gains that are statistically significant and consistent across all cross-validation folds. Future work will extend the approach to additional dependent classes and procedures, explore graph-based modeling of higher-order spatial relationships, and integrate detection with downstream CVPMO criteria assessment toward a complete intraoperative safety system.

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

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV), pp. 213–229. Springer, Cham (2020)

Table 4. Class distribution in the 3-class dataset. Instance percentages are relative to all 1,142 bounding boxes; frame percentages are relative to the 586 annotated frames.

Class	Instances (%)	Annotated frames (%)
CL	555 (48.6%)	551 (94.0%)
DOOM	458 (40.1%)	456 (77.8%)
EPI	129 (11.3%)	129 (22.0%)
All classes	1,142	586

2. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In: International Conference on Learning Representations (ICLR) (2022)
3. Ehlers, A.P., Lai, Y.L., Hu, H.M., Howard, R., Davidson, G.H., Waljee, J.F., Dimick, J.B., Telem, D.A.: Five year trends in surgical technique and outcomes of groin hernia repair in the United States. Surgical Endoscopy **37**(6), 4818–4823 (2023)
4. HerniaSurge Group: International guidelines for groin hernia management. Hernia **22**(1), 1–165 (2018)
5. Haladu, N.U., Alabi, A., Brazzelli, M., Imamura, M., Ahmed, I., Ramsay, G., Scott, N.W.: Open versus laparoscopic repair of inguinal hernia: an overview of systematic reviews of randomised controlled trials. Surgical Endoscopy **36**(7), 4685–4700 (2022)
6. Sivakumar, J., Chen, Q., Hii, M.W., Cullinan, M., Choi, J., Steven, M., Crosthwaite, G.: Learning curve of laparoscopic inguinal hernia repair: systematic review, meta-analysis, and meta-regression. Surgical Endoscopy **37**(4), 2453–2475 (2023)
7. Solaini, L., Cavaliere, D., Rocco, G., Avanzolini, A., Di Pietrantonio, D., Ercolani, G.: Differences in the learning curve of robotic transabdominal preperitoneal inguinal hernia repair according to surgeon's robotic experience. Hernia **27**(5), 1123–1129 (2023)
8. Felix, E., Scott, S., Crafton, B., Geis, P., Duncan, T., Sewell, R., McKernan, B.: Causes of recurrence after laparoscopic hernioplasty: a multicenter study. Surgical Endoscopy **12**(3), 226–231 (1998)
9. Sato, M., Nozawa, M., Watanabe, T., Onoda, T., Matsuyama, A., Shiiya, N., Wada, H.: Insufficiency of prosthetic posterolateral overlap related to recurrence after laparoscopic transabdominal preperitoneal inguinal hernioplasty, as assessed by video review. BMC Surgery **20**(1), 27 (2020)
10. Daes, J., Felix, E.: Critical view of the myopectineal orifice. Annals of Surgery **266**(1), e1–e2 (2017)
11. Claus, C., Furtado, M., Malcher, F., Cavazzola, L.T., Felix, E.: Ten golden rules for a safe MIS inguinal hernia repair using a new anatomical concept as a guide. Surgical Endoscopy **34**(4), 1458–1464 (2020)
12. Marmolejo, A., Farell, J., Ruiz Funes, A.P., Ayala, S., Sánchez, A., Navarro, C.A., Ramírez, N.A., García, L., Daes, J.: Critical view of the myopectineal orifice: a scoring system to objectively evaluate transabdominal preperitoneal inguinal hernia repair. Surgical Endoscopy **36**(7), 5094–5103 (2022)
13. Strasberg, S.M., Brunt, M.L.: Rationale and use of the critical view of safety in laparoscopic cholecystectomy. Journal of the American College of Surgeons **211**(1), 132–138 (2010)
14. Takeuchi, M., Collins, T., Lipps, C., Haller, M., Uwineza, J., Okamoto, N., Nkusi, R., Marescaux, J., Kawakubo, H., Kitagawa, Y., Gonzalez, C.: Towards automatic

verification of the critical view of the myopectineal orifice with artificial intelligence. Surgical Endoscopy **37**(6), 4525–4534 (2023)
15. Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M.: YOLO26: key architectural enhancements and performance benchmarking for real-time object detection. arXiv preprint arXiv:2509.25164 (2025)
16. 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., Alseidi, A.: Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Annals of Surgery **276**(2), 363–369 (2022)
17. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Padoy, N.: Encoding surgical videos as latent spatiotemporal graphs for object and anatomy-driven reasoning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 647–657. Springer, Cham (2023)
18. Li, Y., Gupta, H., Ling, H., Ramakrishnan, I.V., Prasanna, P., Georgakis, G., Sasson, A.: Automated assessment of critical view of safety in laparoscopic cholecystectomy. In: IEEE International Conference on Healthcare Informatics (ICHI), pp. 330–337. IEEE (2023)
19. Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., Dallemagne, B.: Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Annals of Surgery **275**(5), 955–961 (2022)