Supplementary Material: TWiST

1 Dataset Description

For our experiments, we used a subset of the **CholecT45** subset of the CholecT50 dataset [Nwoye and Padoy, 2022]. Cholec45 consists of 45 laparoscopic cholecystectomy videos annotated with surgical action triplets. A total of approximately 100.9K frames sampled at 1 frame per second were labeled with around 151K triplet instances involving multiple instrument, verb, and target classes. We used a total of 15 videos from the dataset (limited due to computational constraints).

The dataset was split into training and testing sets as follows:

- **Testing videos** (5 videos):

  {"VID92", "VID96", "VID103", "VID110", "VID111"}

- **Training videos** (10 videos):

  {"VID01", "VID02", "VID04", "VID05", "VID06", "VID08", "VID10", "VID12", "VID13", "VID14"}

The training set provides only binary presence labels for triplets and their components without spatial annotations, enabling weakly-supervised learning for instrument localization and triplet-box association.

2 Further Experimental Results and Ablations

2.1 Additional Results on Backbone Variants

We further evaluated our proposed TWiST framework with different backbone networks (VGG16, VGG19, and InceptionV3). Results are summarized in Table 1. While these models exhibit weaker performance compared to our main ResNet-50 based TWiST, the comparison highlights the robustness of our design choices. Notably, VGG-based backbones achieve moderate recognition but struggle in localization, while InceptionV3 performs poorly across all metrics.

Table 1: Performance comparison of backbone variants: VGG16, VGG19, InceptionV3, and ResNet50 (Ours). Values are presented as percentages.

Model	mAP_i	mRec_i	mAP_{ivt}	mRec_{ivt}	mPre_{ivt}
TWiST (VGG16)	1.95	4.61	0.23	0.68	2.07
TWiST (VGG19)	4.91	9.15	0.59	1.35	4.31
TWiST (InceptionV3)	0.27	0.06	0.01	0.00	0.13
TWiST (Ours)	**12.07**	**19.15**	**1.74**	**3.14**	**7.92**

2.2 On Methods with Supervision on External Datasets

Some methods shown in Nwoye et al. [2023], such as Distilled-Swin-YOLO and ResNet-CAM-YOLOv5, were not included in our main quantitative comparison table in the paper because they are trained in a fully supervised way on other surgical datasets for instrument localization(using bounding box annotations).

- **Distilled-Swin-YOLO** leverages Cholec80 robotic instrument segmentation, Robotic Scene Segmentation, and EndoVis Instrument datasets for teacher-student distillation and pseudo-label generation.
- **ResNet-CAM-YOLOv5** pre-trains its YOLOv5 component on COCO, CholecSeg8k, HeiCo, LapChole, and EndoVis Instrument datasets to obtain strong localization and then fine-tunes on the CholecT45 dataset.

In contrast, our TWiST framework is fully weakly supervised on Cholec45 only, without relying on external bounding box annotations or segmentation datasets. Direct metric comparison would therefore be unfair; however, we note that despite this stricter training setup, our method still achieves comparable mAP for instruments compared to Distilled-Swin-YOLO, as shown in Table 2.

Table 2: Performance comparison of ResNet-CAM-YOLOv5, Distilled-Swin-YOLO, and TWiST (Ours). Values are presented as percentages.

Model	mAP_i	mRec_i	mAP_{ivt}	mRec_{ivt}	mPre_{ivt}
ResNet-CAM-YOLOv5	**41.87**	**49.30**	**4.49**	**7.87**	**11.74**
Distilled-Swin-YOLO	17.28	30.37	2.74	6.16	9.26
TWiST (Ours)	12.07	19.15	1.74	3.14	7.92

2.3 Metric Definitions

For clarity, we report results on the five key metrics used in the CholecTriplet benchmark:

$$\text{mAP}_i = \int_0^1 p(r)\,dr \quad \text{detection mean average precision for instruments (IoU-based, instrument ID correct)} \tag{1}$$

$$\text{mRec}_i = \frac{TP}{TP+FN} \quad \text{detection mean recall for instruments (averaged over IoU thresholds)} \tag{2}$$

$$\text{mAP}_{ivt} = \int_0^1 p(r)\,dr \quad \text{detection mean average precision for triplets (IoU-based, triplet ID correct)} \tag{3}$$

$$\text{mRec}_{ivt} = \frac{TP}{TP+FN} \quad \text{detection mean recall for triplets (averaged over IoU thresholds)} \tag{4}$$

$$\text{mPre}_{ivt} = \frac{TP}{TP+FP} \quad \text{detection mean precision for triplets (averaged over IoU thresholds)} \tag{5}$$

Together, these metrics balance localization accuracy, triplet recognition correctness, and overall reliability. **Note:** The evaluation metrics reported above were calculated on the test split using the `eval.ai` platform, where the official CholecT2022 challenge was hosted.

3 Hyperparameters

3.1 Loss Functions

We train two separate models, each optimized with its own objective function.

3.1.1 Model 1: Instrument Count Regression

We model instrument count prediction as a multi-output regression task, since each of the six surgical instrument classes may appear with multiple instances per frame.

Let x denote an input frame and $y = [y_1, y_2, \ldots, y_6]^\top$ represent the corresponding ground-truth counts for each instrument class. A ResNet-50 backbone with a final fully connected regression head, $f_\theta(x)$, predicts $\hat{y} = f_\theta(x)$. The model is trained using mean squared error (MSE):

$$\mathcal{L}_{\text{count}}(y, \hat{y}) = \frac{1}{6}\sum_{k=1}^{6} (\hat{y}_k - y_k)^2 ,$$

where y_k and $\hat{y}_k$ are the ground truth and predicted counts, respectively, for instrument class k.

3.1.2 Model 2: Triplet Presence Detection

The second model predicts the presence of instruments, verbs, targets, and triplets. We adopt a weighted binary cross-entropy (BCE) formulation, where a positive-class weight λ scales only the positive term of the loss. For a single sample and prediction, the loss is:

$$\mathcal{L}(y, \hat{y}) = -\Big(\lambda \cdot y \cdot \log(\hat{y} + \epsilon)\Big) \; - \; \Big((1-y) \cdot \log(1 - \hat{y} + \epsilon)\Big),$$

where

- $y \in \{0, 1\}$ is the ground truth label,
- $\hat{y} \in (0, 1)$ is the predicted probability after sigmoid activation,
- $\epsilon = 10^{-6}$ is a stability constant,
- $\lambda > 0$ is a positive-class weight.

Different λ values are assigned per prediction head:

$$\lambda_{\text{instrument}} = 2.0, \quad \lambda_{\text{verb}} = 1.5, \quad \lambda_{\text{target}} = 1.2, \quad \lambda_{\text{triplet}} = 1.8$$

The total training objective is the sum of the weighted BCE losses from the four prediction heads:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{instrument}} + \mathcal{L}_{\text{verb}} + \mathcal{L}_{\text{target}} + \mathcal{L}_{\text{triplet}}$$

Summary:

- Model 1 (Instrument Count Regression): MSE for instrument count prediction.
- Model 2 (Triplet Presence Detection): weighted BCE for instruments, verbs, targets, and triplets.

3.2 CAM Activation Thresholds

For instrument localization, class activation maps (CAMs) generated using Grad-CAM were normalized to the range $[0, 1]$ and thresholded at a fixed value of 0.5 to maximize localization metrics.

References

Chinedu Innocent Nwoye and Nicolas Padoy. Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235*, 2022.

Chinedu Innocent Nwoye, Tong Yu, Saurav Sharma, Aditya Murali, Deepak Alapatt, Armine Vardazaryan, Kun Yuan, Jonas Hajek, Wolfgang Reiter, Amine Yamlahi, et al. Cholectriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. *Medical Image Analysis*, 89:102888, 2023.