CoStoDet-DDPM: Collaborative Training of Stochastic and Deterministic Models Improves Surgical Workflow Anticipation and Recognition

Supplementary Material

7. More Dataset for Recognition

7.1. Datasets and Inplementation Details

Following the Rebuttal phase and the Area Chair's suggestions, we include two additional datasets in the Supplementary Material of the final version to further validate the effectiveness of CoStoDet-DDPM in surgical phase recognition. The *CATARACTS* dataset consists of 50 cataract surgery videos, each ranging from 6 to 40 minutes in duration, and annotated with 19 fine-grained surgical steps. We follow BNPitfalls [48] protocol and resample the videos to 5 FPS, using a 25/25 split for training and testing. The *OphNet-APTOS* dataset [1], a subset of OphNet [22], originates from the APTOS Big Data Competition Phase Recognition of Surgical Videos Using ML and contains cataract surgery videos. We use the official training set (401 videos) and validation set (95 videos), which are annotated with 35 surgical phases. Videos are resampled to 1 FPS for consistency.

The experimental settings for both CATARACTS and OphNet-APTOS are kept consistent with those used in the main paper, except that the learning rate for CATARACTS is modified to $5e-5$.

7.2. Comparison with State-of-the-arts

We compare our method with previous SOTAs on CATARACTS and OphNet-APTOS, the results are shown in Table 5. The results demonstrate that our co-training scheme CoStoDet-DDPM achieves more accurate recognition across a greater number of phases or steps. Specifically, our method yields improvements of 2.5% and 0.9% in Jaccard index on the two datasets, respectively, highlighting its robustness and generalization capability.

Datasets	Methods	Accuracy ↑	Precision ↑	Recall ↑	Jaccard ↑
CATARACTS	SV-RCNet [24]	81.3	66.0	57.0	47.2
	3D-CNN [12]	80.1	66.2	55.7	45.9
	TeCNO [6]	79.0	62.6	56.9	45.1
	TMRNet [26]	80.6 ± 10.2	63.1	54.6	45.7
	Trans-SVNet [14]	77.8	61.3	55.0	43.8
	Dual Pyramid [4]	84.2	**69.3**	**66.4**	53.7
	BNPitfalls [48]	83.3	66.8	61.8	50.3
	DACAT [61]	85.7 ± 10.9	63.8	64.8	53.5
	Ours	**87.1 ± 10.5**	67.9	65.4	**56.2**
OphNet-APTOS	TeCNO-ResNet50 [6]	67.8	46.0	42.6	-
	TeCNO-ViT-Base [6]	67.7	45.7	44.1	-
	BNPitfalls [48]	72.9 ± 14.7	50.6	48.2	35.3
	DACAT [61]	73.2 ± 13.7	52.8	48.5	38.1
	Ours	**75.5 ± 14.5**	**54.0**	**50.2**	**39.0**

Table 5. Comparison results (%) with SOTA on CATARACTS and OphNet-APTOS. **Bold**: Best results. Underline: Second-best results.

8. More Ablation Study

8.1. Observation Encoder of DDPM

We explored the impact of different observation encoders in the DDPM. As shown in Table 6, it is evident that incorporating temporal processing with LSTM, using longer sequences, and simultaneously performing both tool and phase tasks are beneficial for the results. Most importantly, incorporating the task loss term $\mathcal{L}_{Task}$ significantly improves performance.

Encoder	seq	Tool	Phase	*Smooth*
ResNet18-GN	32	1.35/0.63/2.61	-	-
ConvNeXt	32	1.17/0.49/2.46	-	-
ConvNeXt	64	1.02/0.46/2.16	-	-
ConvNeXt-LSTM	64	0.97/0.29/2.05	-	-
ConvNeXt-LSTM	64	0.97/0.34/2.03	0.63/0.19/1.05	$0.133+0.079$
Ours ($w/\ \mathcal{L}_{Task}$)	64	0.93/0.34/1.75	0.62/0.23/0.86	$0.095+0.077$

Table 6. The impact of different observation encoders on DDPM performance. The results are evaluated on Cholec80 with $h=5$ min. The columns present $wMAE/outMAE/eMAE$, while $Smooth$ is reported as $Smooth_{Tool}+Smooth_{Phase}$. All inferences use DDIM with 16 steps. ResNet18-GN denotes a model where Batch Normalization is replaced with Group Normalization.

Output Branch	λ	Tool	Phase	*Smooth*
$\mathcal{D}$	1	0.95/0.31/2.04	0.63/0.20/1.03	$0.093+0.073$
$\mathcal{T}$	1	0.94/0.32/2.04	0.61/0.20/1.03	$0.031+0.026$
$\mathcal{D}$	16	**0.91**/0.36/1.83	**0.58**/0.21/**0.83**	$0.057+0.044$
$\mathcal{T}$	16	**0.91**/0.35/1.85	**0.58**/0.20/0.85	$0.028+0.021$
$\mathcal{D}$	32	0.93/0.34/**1.75**	0.62/0.23/0.86	$0.095+0.077$
$\mathcal{T}$	32	**0.91**/**0.29**/1.77	0.59/**0.18**/0.88	$\mathbf{0.025+0.020}$

Table 7. Comparison of results between $\mathcal{D}$ and $\mathcal{T}$ with different λ. The evaluation is conducted on Cholec80 with $h=5$ min. The columns report $wMAE/outMAE/eMAE$, while $Smooth$ is presented as $Smooth_{Tool}+Smooth_{Phase}$. All experiments are conducted using CoStoDet-DDPM, where $\mathcal{D}$ performs inference with DDIM (16 steps), and the clip sequence length is set to 64.

8.2. Historical Time Span of Labels in DDPM

We also investigated the historical span λ of labels, as shown in the Table 7. Due to memory and time constraints, we explored $\lambda=1$, 16, and 32. When $\lambda=1$, only the current frame label is used. The results show that incorporating historical information improves the anticipation per-

Method	$wMAE \downarrow$ / $outMAE \downarrow$ / $eMAE \downarrow$					
	Tool			Phase		
	$h = 2$ min	$h = 3$ min	$h = 5$ min	$h = 2$ min	$h = 3$ min	$h = 5$ min
Add	0.39/0.05/0.94	0.56/0.13/1.31	0.95/0.33/2.07	0.27/0.05/0.58	0.37/0.09/0.68	0.60/0.19/0.99
Concat	0.41/0.05/0.99	0.56/0.13/1.29	0.91/0.30/1.94	0.28/0.05/0.59	0.37/0.09/0.68	0.59/0.21/0.89
Att-UNet	0.40/0.06/0.99	0.57/0.11/1.36	0.96/0.36/2.08	0.29/0.05/0.59	0.39/0.10/0.72	0.60/0.22/1.04
FiLM (Ours)	0.39/0.05/0.93	0.56/0.09/1.28	0.91/0.29/1.77	0.27/0.05/0.52	0.39/0.09/0.67	0.59/0.18/0.88

Table 8. The impact of denoising network architecture and conditional feature fusion.

Method	$wMAE \downarrow$ / $outMAE \downarrow$ / $eMAE \downarrow$					
	Tool			Phase		
	$h = 2$ min	$h = 3$ min	$h = 5$ min	$h = 2$ min	$h = 3$ min	$h = 5$ min
BNP	0.40/0.06/1.06	0.58/0.13/1.42	0.96/0.41/2.11	0.29/0.06/0.63	0.39/0.10/0.78	0.61/0.24/0.99
MaskAE	0.41/0.13/0.97	0.64/0.26/1.47	1.17/0.66/2.27	0.29/0.10/0.62	0.45/0.20/0.76	0.79/0.44/1.21
CURL	0.41/0.07/1.01	0.58/0.16/1.44	1.01/0.52/2.05	0.27/0.06/0.55	0.39/0.11/0.74	0.63/0.32/0.91
GC	0.40/0.08/1.03	0.59/0.19/1.35	0.98/0.44/2.04	0.29/0.08/0.60	0.39/0.13/0.66	0.60/0.24/0.93
RandomMask	0.41/0.05/0.99	0.56/0.11/1.29	0.91/0.28/1.86	0.30/0.06/0.60	0.38/0.09/0.70	0.60/0.17/0.90
FiLM (Ours)	0.39/0.05/0.93	0.56/0.09/1.28	0.91/0.29/1.77	0.27/0.05/0.52	0.39/0.09/0.67	0.59/0.18/0.88

Table 9. The impact of co-training methods.

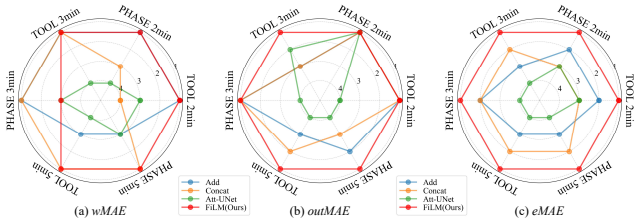

Figure 10. The impact of denoising network architecture and conditional feature fusion. The scale represents rankings, with detailed data provided in the Table 8.

Methods	lr, wd	Accuracy ↑	Precision ↑	Recall ↑	Jaccard ↑
BNPitfalls	$1e-4, 1e-2$	93.7 ± 4.1	88.7	87.7	79.2
$w/\ \mathcal{D}$	$1e-4, 1e-2$	92.2 ± 6.6	87.0	86.6	77.2
$w/\ \mathcal{D}$	$1e-3, 1e-2$	$\mathbf{94.1 \pm 3.5}$	**88.9**	**89.2**	**80.8**
$w/\ \mathcal{D}$	$1e-3, 1e-6$	91.3 ± 6.7	84.0	86.3	74.7
DACAT	$1e-5, 1e-2$	94.1 ± 4.3	**89.2**	88.5	80.6
$w/\ \mathcal{D}$	$1e-5, 1e-2$	$\mathbf{94.5 \pm 3.6}$	88.4	**90.2**	**81.4**
$w/\ \mathcal{D}$	$1e-5, 1e-6$	94.4 ± 3.6	88.1	89.8	80.9

Table 10. The impact of different learning rate (lr) and weight decay (wd) for recognition results on Cholec80.

formance, and the effects of $\lambda = 16$ and $\lambda = 32$ are nearly identical.

8.3. Denoising Network Architecture and Conditional Feature Fusion

We explored the impact of feature fusion methods and denoising network architectures within $\mathcal{D}$. Specifically, we investigated three approaches: (1) directly adding the conditional features at each layer (Add), (2) concatenating the conditional features with the noisy features before inputting them into the U-Net (Concat), and (3) using Attention U-Net (Att-UNet) as the denoising network. As shown in Figure 10, FiLM and U-Net are more effective for co-training, leading to improved performance.

8.4. Concrete Experiment Results

We present more detailed results of two ablation experiments: Denoising Network Architecture and Conditional Feature Fusion Sec. 8.3, as shown in Table 8, and Different Collaborative Learning Strategies Sec. 4.3, as shown in Table 9.

9. More Discussion

9.1. Sensitivity to Hyperparameter for Recognition

For phase recognition, we find that CoStoDet-DDPM is sensitive to the learning rate (lr) and weight decay (wd) during training, as shown in Table 10. For BNPitfalls, a higher learning rate than the original one is required to fully lever-

age the effect of $\mathcal{D}$. For DACAT, the original lr and wd settings are sufficient. Due to time and resource constraints, we have not yet fully explored the optimal parameters for the recognition task, and further investigation is needed in the future.

9.2. Modal Complexity Analysis

We take the phase recognition task as an example to analyze the model complexity during both training and inference when integrating our method with different SOTA baselines, *i.e.*, BNPitfalls (BNP) and DACAT. All experiments are conducted on an NVIDIA GeForce RTX 4090 24GB GPU, and the results are shown in Table 11. Although the introduction of FiLM-UNet increases the number of model parameters, the additional complexity remains limited due to the compact size of the conditional feature $\mathbf{c}_t \in \mathbb{R}^{512}$. Importantly, since DDPM is discarded during inference, the overall complexity and speed remain consistent with the original SOTA methods.

Moreover, Table 11 presents a detailed analysis of the complexity when employing our $\mathcal{D}$ under different DDIM sampling steps (1-step, 16-step, and 100-step). In summary, our co-training framework introduces only a modest increase in training cost, while providing significant performance gains.

Methods	Param (M)	FLOPs (G)	GPU (GB)	FPS
BNP	30.4	7.1 / -, -, -	8.5 / 0.9	91.0
Ours (BNP)	280.0	7.1 / 9.5, 45.2, 245.1	15.0 / 0.9	91.0
DACAT	62.6	14.2 / -, -, -	13.9 / 1.2	39.5
Ours (DACAT)	312.2	14.2 / 20.6, 56.3, 256.3	23.6 / 1.2	39.5

Table 11. All complexity measurements are conducted on the same hardware platform: an NVIDIA GeForce RTX 4090 24GB GPU. FLOPs are reported separately for inference using the Task branch only ($\mathcal{T}$) and the Diffusion branch ($\mathcal{D}$) under DDIM sampling with 1, 16, and 100 steps. GPU usage is measured for both the training stage (on batched frames) and the inference stage (on a single frame).