IPMI 2021
Rønne, Bornholm

Future Frame Prediction for Robot-assisted Surgery

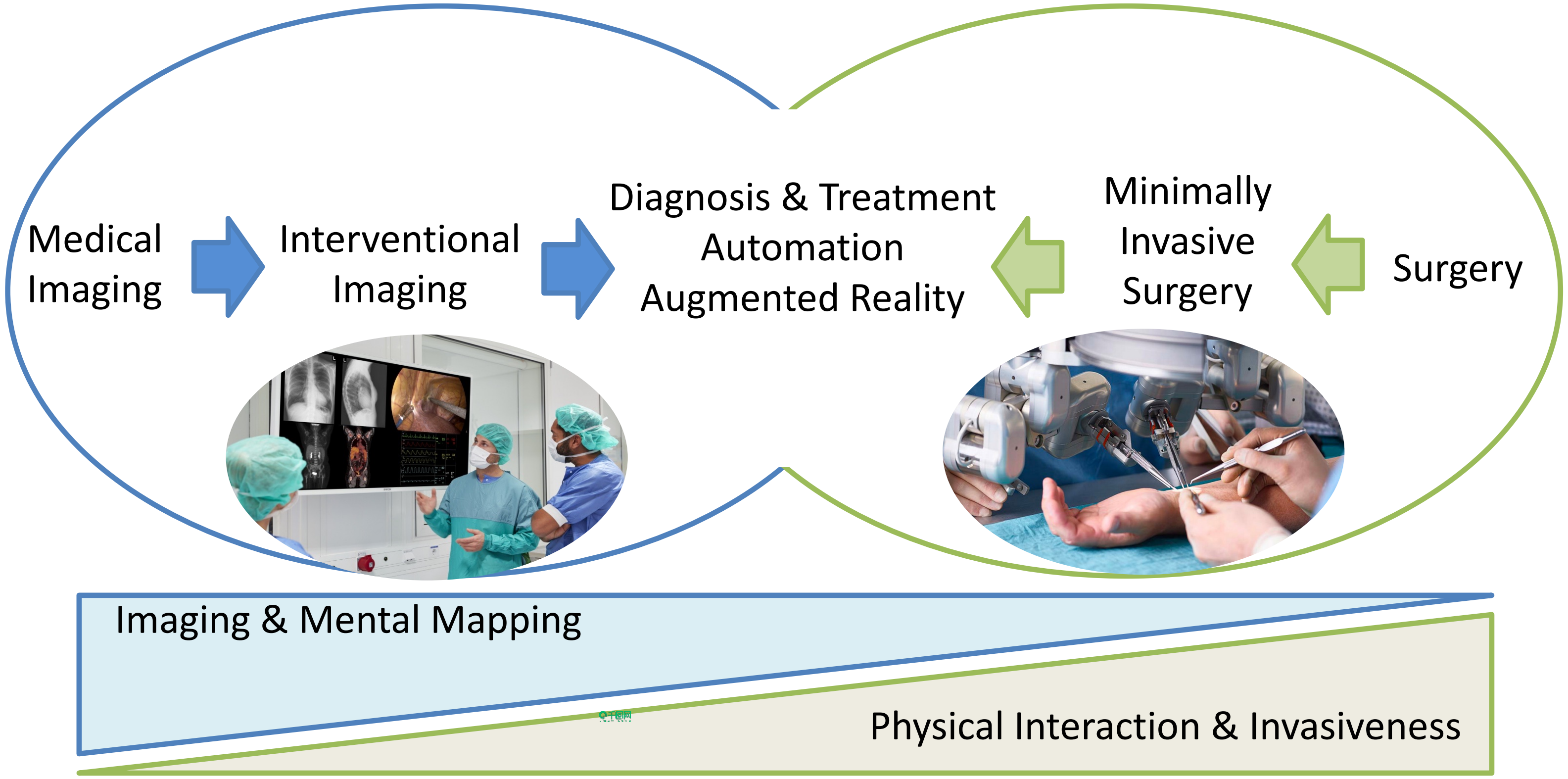
Xiaojie Gao[1], Yueming Jin[1], Zixu Zhao[1], Qi Dou[1,2], and Pheng-Ann Heng[1,2]

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China

[2] T Stone Robotics Institute, CUHK, Hong Kong, China

Introduction

- Computer assisted interventions (CAI) is essential for modern operating rooms (OR) to promote **surgical therapy and structure repair**.
- Its typical applications include **context-aware systems**, robot-assisted surgeries, surgeon training, etc.



Medical Imaging

Interventional Imaging

Diagnosis & Treatment
Automation
Augmented Reality

Minimally Invasive Surgery

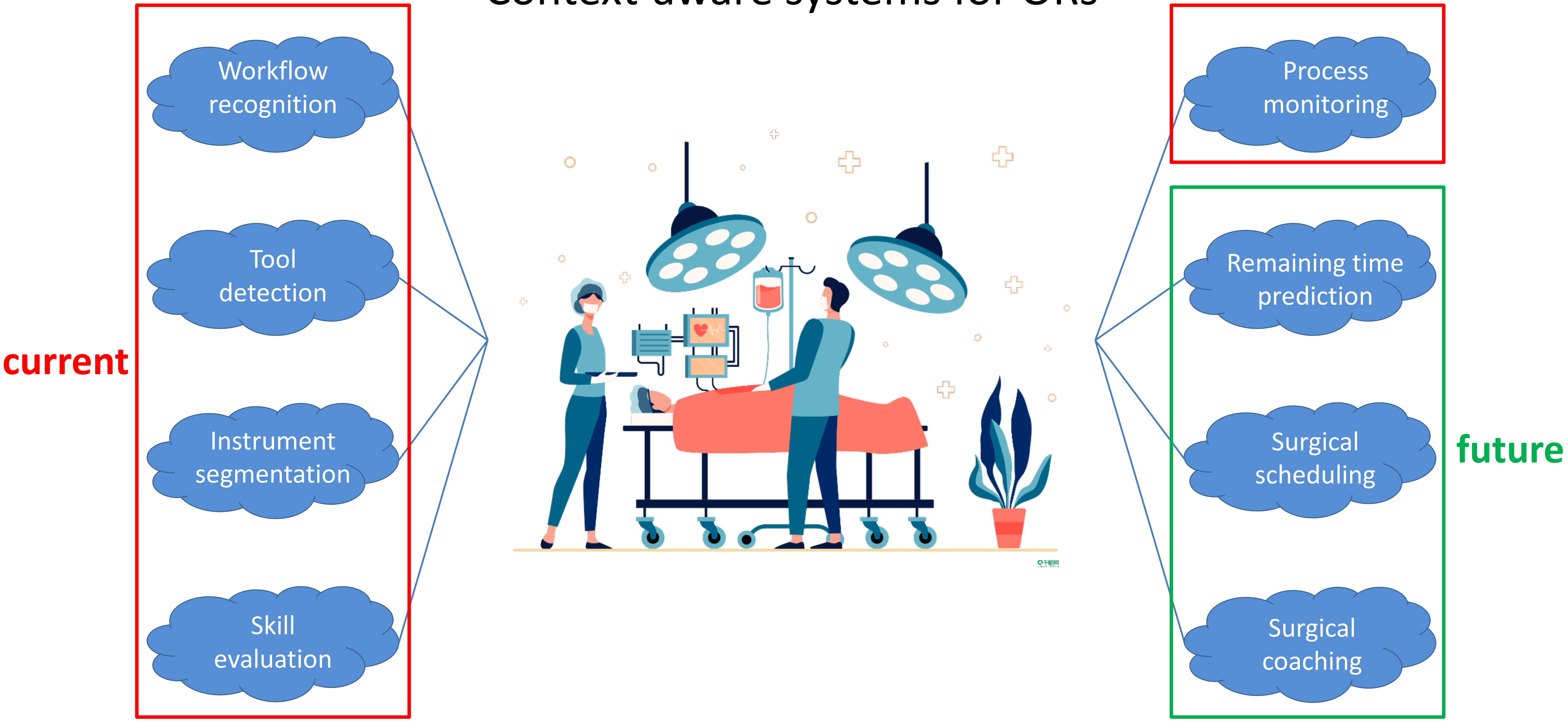
Surgery

Imaging & Mental Mapping

Physical Interaction & Invasiveness

Introduction

Context-aware systems for ORs



current

Workflow recognition

Tool detection

Instrument segmentation

Skill evaluation

Process monitoring

Remaining time prediction

Surgical scheduling

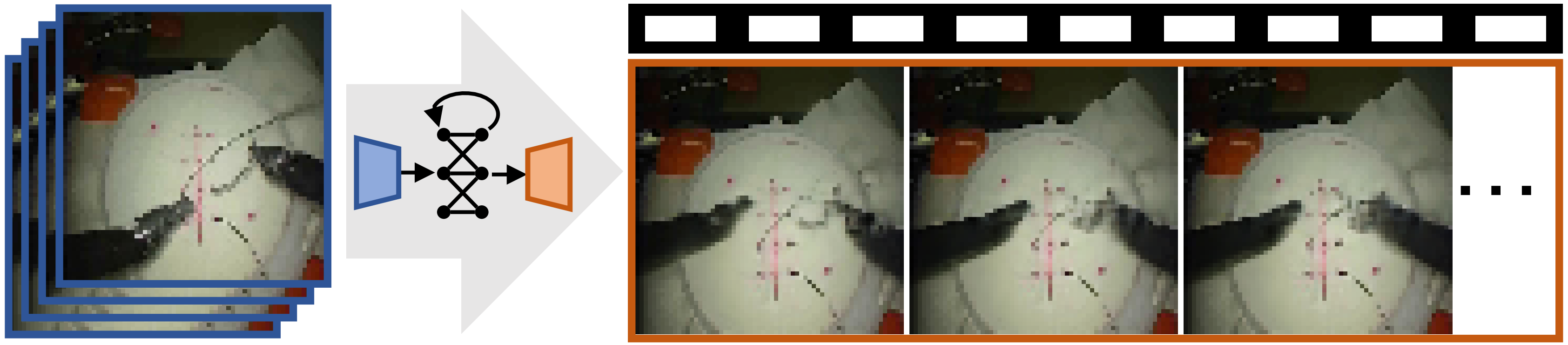
Surgical coaching

future

Problem definition

Future frame prediction

Given a video clip $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t_p}]$ with each frame $\mathbf{x}_t \in \mathbb{R}^{w \times h \times c}$, we aim to generate the following frames $\left[\hat{\mathbf{x}}_{t_p+1}, \hat{\mathbf{x}}_{t_p+2}, \ldots, \hat{\mathbf{x}}_T\right]$ in the pixel space.

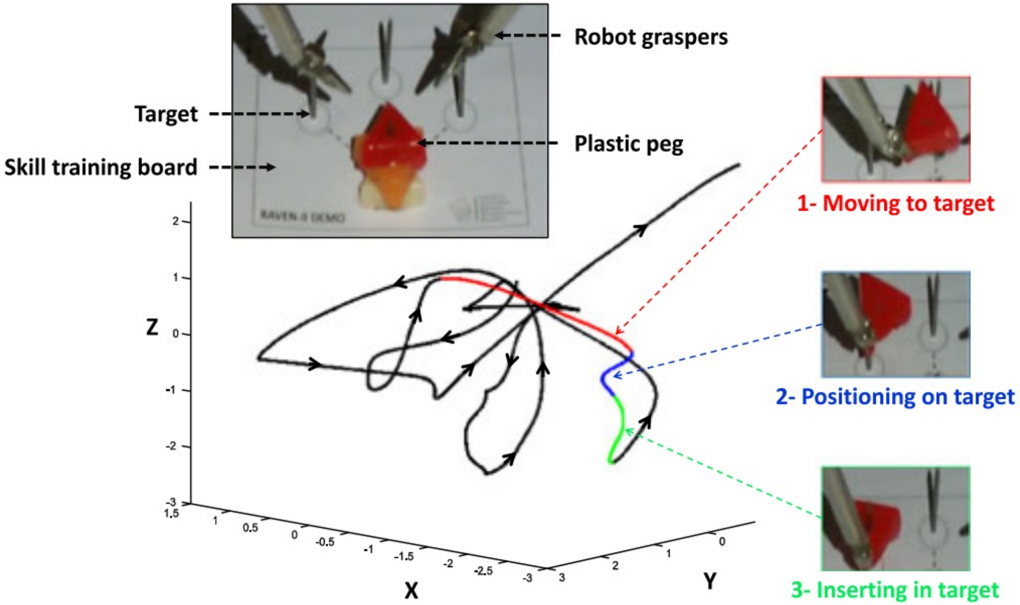


Unique characteristics of surgical videos

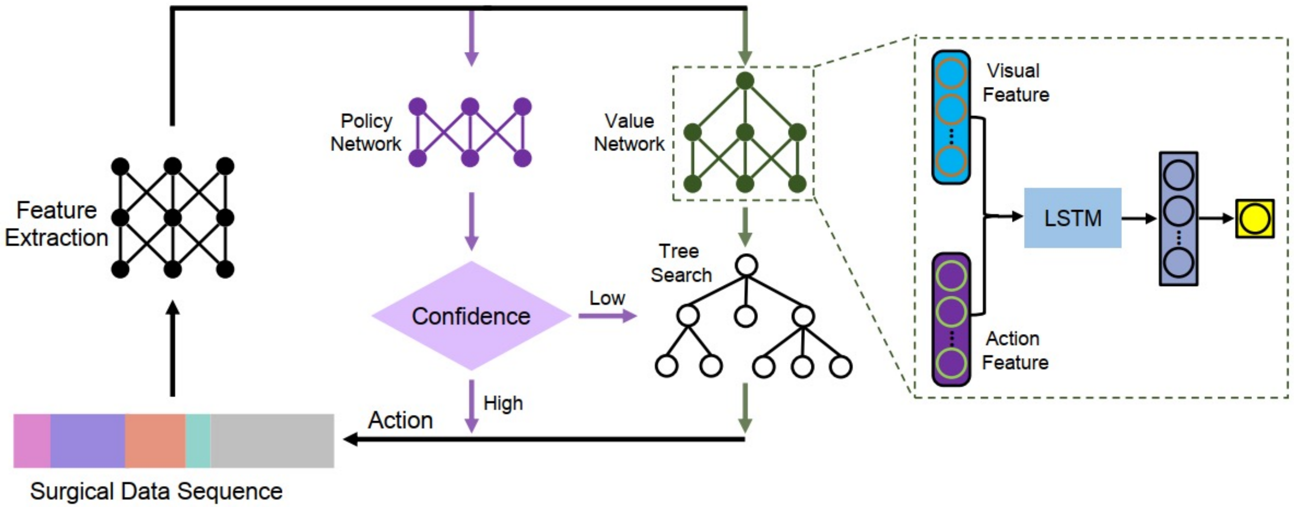
- Long time duration without clear boundaries among different phases
- Meaningful movements depending on goals
- Complicated dual-arm interactions with limited inter-class variance

Motivation

- Providing **image-guided navigation** rather than semantic descriptions
- Enabling **reasoning of decision-making** to promote surgical gesture recognition
- A **higher-level interpretation** of surgeries for realizing surgical automation



Despinoy et al. 2015



Gao et al. 2020

Despinoy F, et al. Unsupervised trajectory segmentation for surgical gesture recognition in robotic training. IEEE Transactions on Biomedical Engineering, 2015.
Gao et al. Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search. ICRA, 2020.

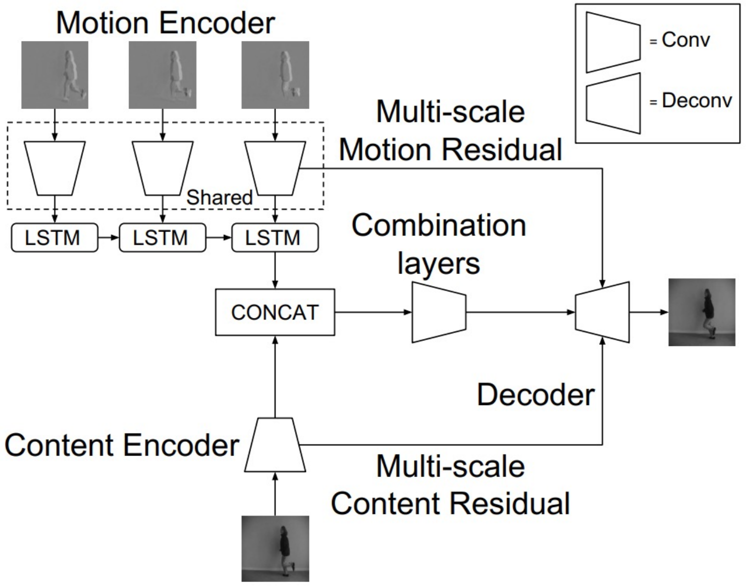
Related work

Deterministic methods modeling only one outcome

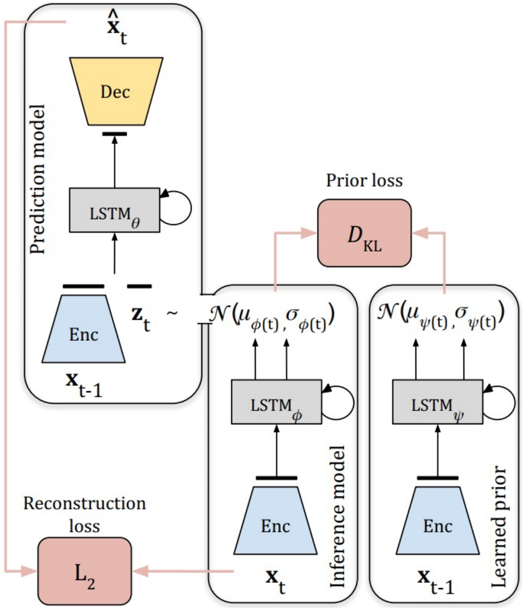
[Oh et al. NurIPS 2015; Finn et al. NurIPS 2016; Villegas et al. ICLR 2017; Villegas et al. ICML 2017; Denton et al. NurIPS 2017; Jin et al. CVPR 2020]

Stochastic methods without considering dual-arm cases

[Babaeizadeh et al. ICLR 2018; Denton et al. ICML 2018; Tulyakov et al. CVPR 2018; Kumar et al. ICLR 2020]



Villegas et al. 2017



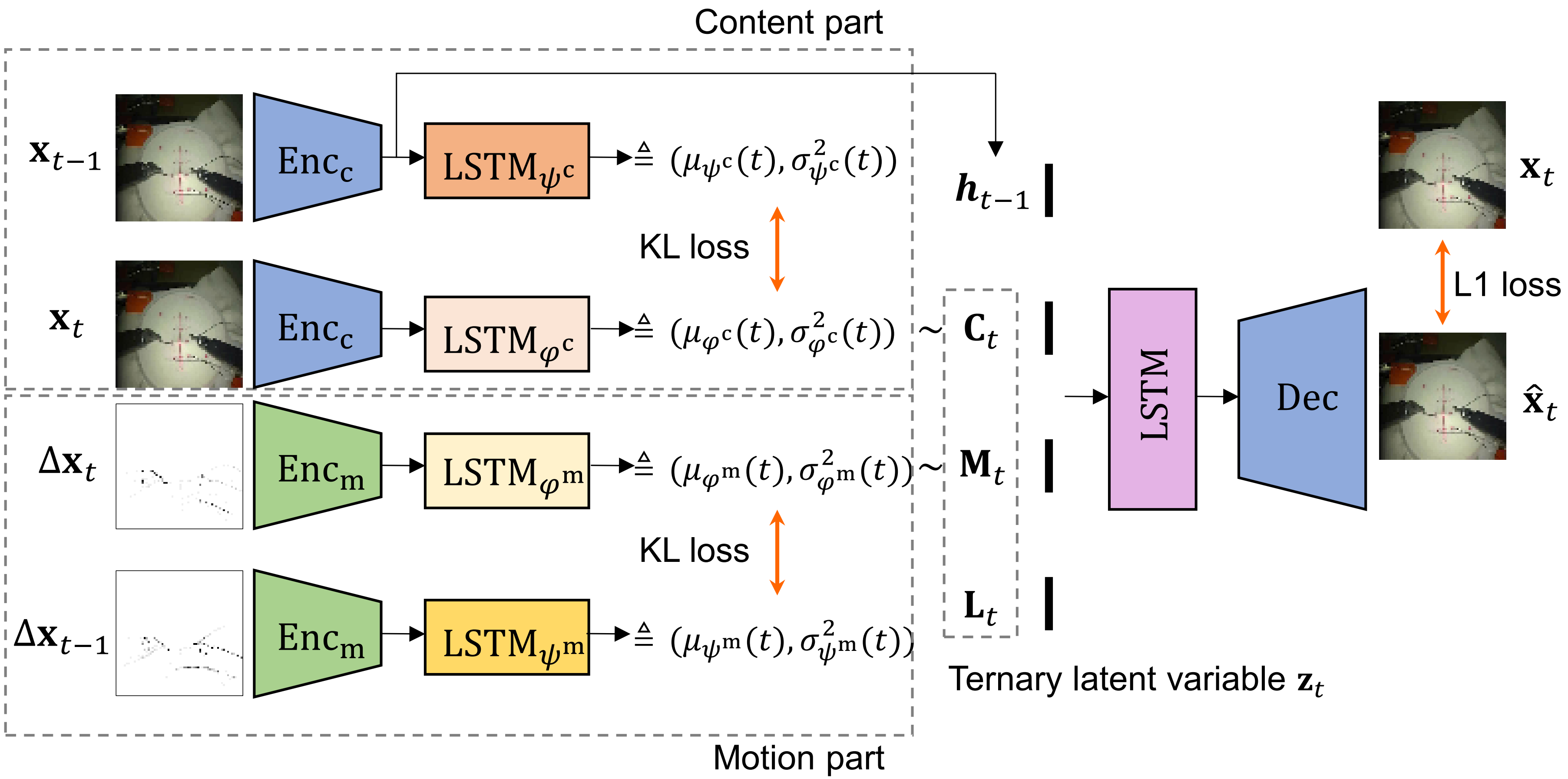
Denton et al. 2018

$$p_\theta(\mathbf{x}_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) \rightarrow \hat{\mathbf{x}}_t$$

$$\mathbf{z}_t \sim \mathcal{N}(\mu, \sigma^2)$$

Proposed method

Ternary Prior Guided Variational Autoencoder (TPG-VAE)

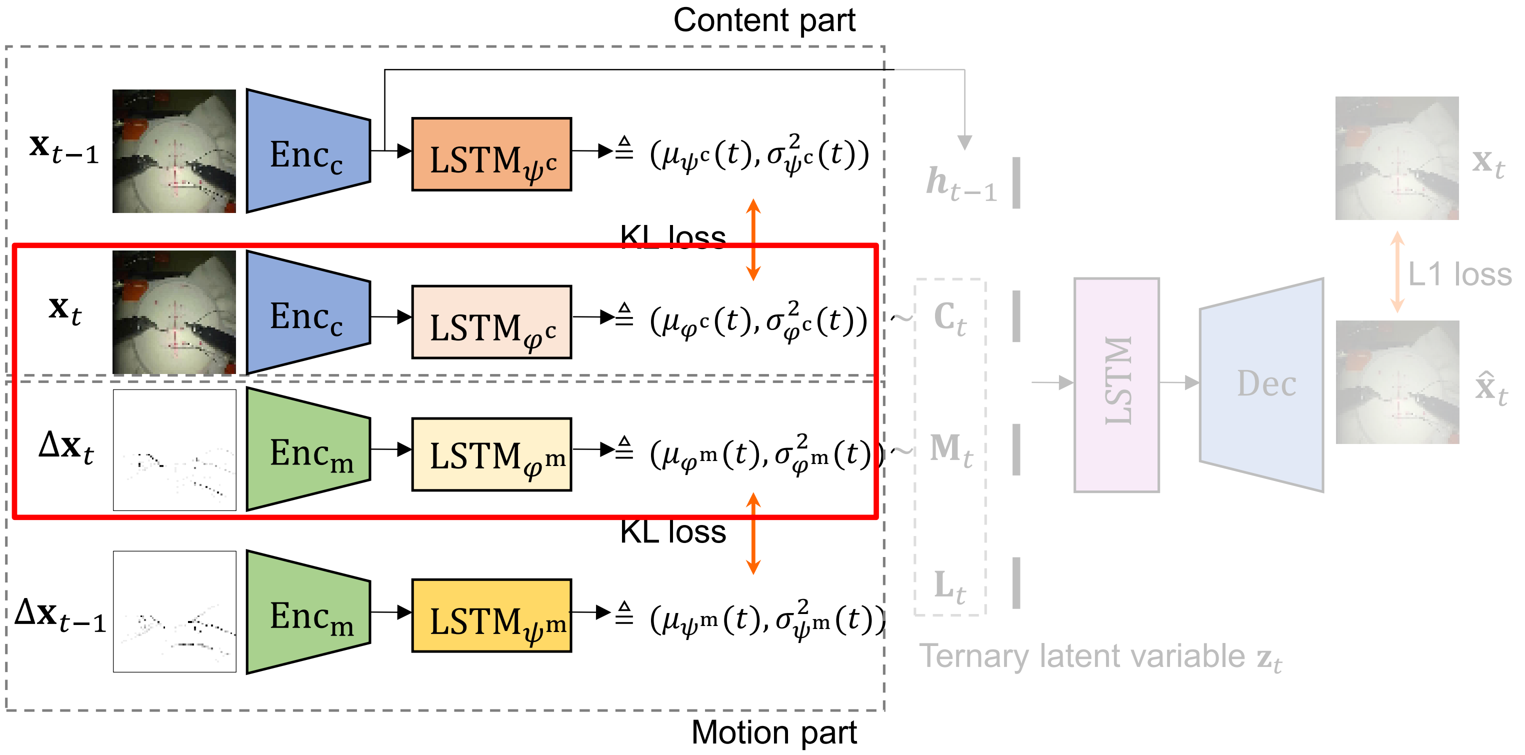


Content part
$\mathbf{x}_{t-1}$
$\mathrm{Enc_c}$
$\mathrm{LSTM}_{\psi^{\mathrm{c}}}$
$\triangleq (\mu_{\psi^{\mathrm{c}}}(t), \sigma^2_{\psi^{\mathrm{c}}}(t))$
KL loss
$\mathbf{x}_t$
$\mathrm{Enc_c}$
$\mathrm{LSTM}_{\varphi^{\mathrm{c}}}$
$\triangleq (\mu_{\varphi^{\mathrm{c}}}(t), \sigma^2_{\varphi^{\mathrm{c}}}(t))$
$\Delta\mathbf{x}_t$
$\mathrm{Enc_m}$
$\mathrm{LSTM}_{\varphi^{\mathrm{m}}}$
$\triangleq (\mu_{\varphi^{\mathrm{m}}}(t), \sigma^2_{\varphi^{\mathrm{m}}}(t))$
KL loss
$\Delta\mathbf{x}_{t-1}$
$\mathrm{Enc_m}$
$\mathrm{LSTM}_{\psi^{\mathrm{m}}}$
$\triangleq (\mu_{\psi^{\mathrm{m}}}(t), \sigma^2_{\psi^{\mathrm{m}}}(t))$
Motion part
$\boldsymbol{h}_{t-1}$
$\sim$ $\mathbf{C}_t$
$\sim$ $\mathbf{M}_t$
$\mathbf{L}_t$
Ternary latent variable $\mathbf{z}_t$
LSTM
Dec
$\mathbf{x}_t$
L1 loss
$\hat{\mathbf{x}}_t$

Proposed method

Decomposed video encoding network

- VGG net and LSTM are combined to exact spatio-temporal features from videos.
- The posterior distributions from content and motion are modeled as Gaussian distribution.



Content distribution $\mathbf{C}_t$:

$$\mathbf{h}_t = \mathrm{Enc_c}(\mathbf{x}_t),$$

$$\mu_{\varphi^c}(t), \sigma^2_{\varphi^c}(t) = \mathrm{LSTM}_{\varphi^c}(\mathbf{h}_t)$$

Motion distribution $\mathbf{M}_t$:

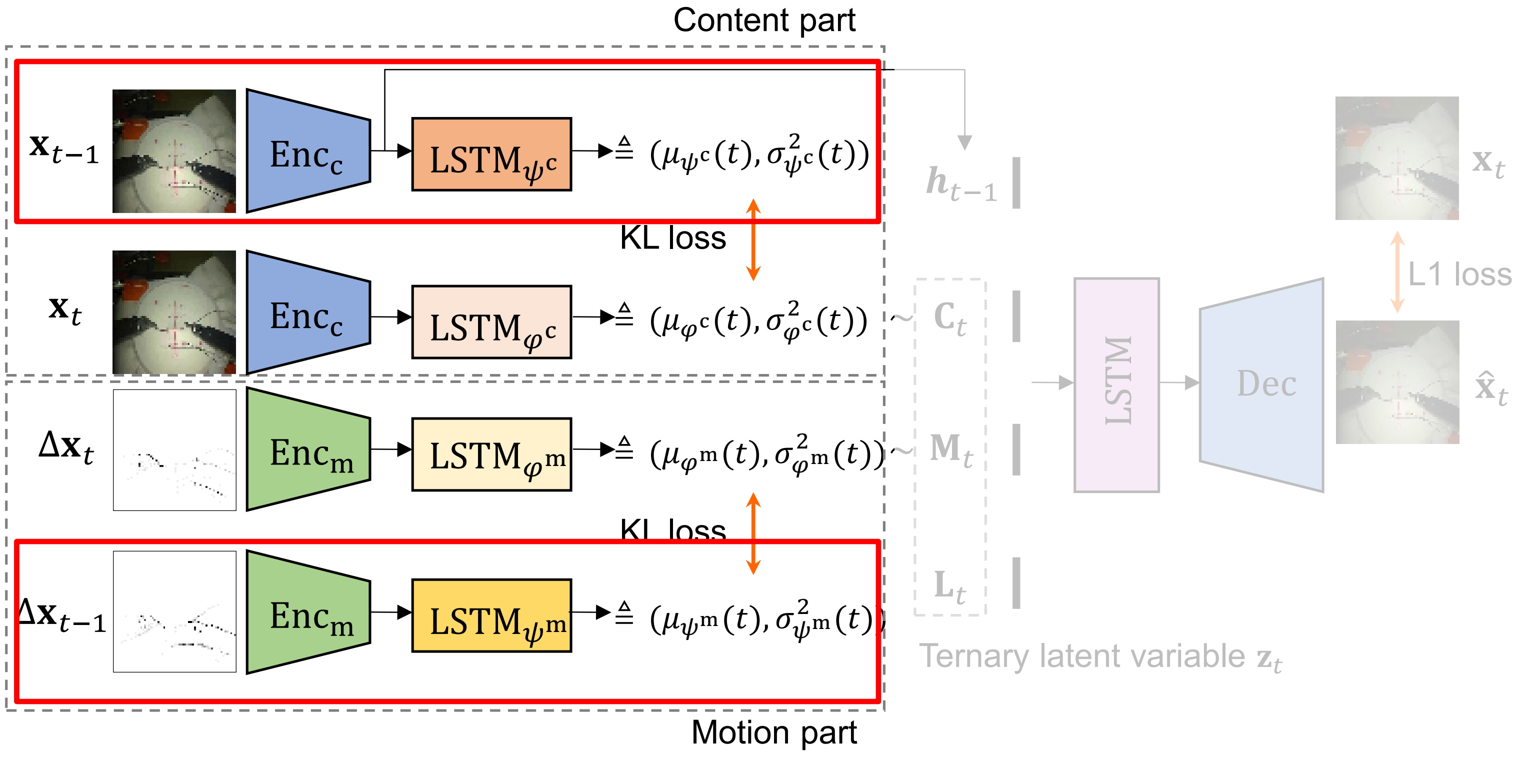
$$\mathbf{h}'_t = \mathrm{Enc_m}(\Delta\mathbf{x}_t),$$

$$\mu_{\varphi^m}(t), \sigma^2_{\varphi^m}(t) = \mathrm{LSTM}_{\varphi^m}(\mathbf{h}'_t)$$

Proposed method

Decomposed video encoding network

The prior distributions of $\mathbf{C}_t$ and $\mathbf{M}_t$ are modelled by two LSTMs to enable latent code generation when no more future information is available during testing.

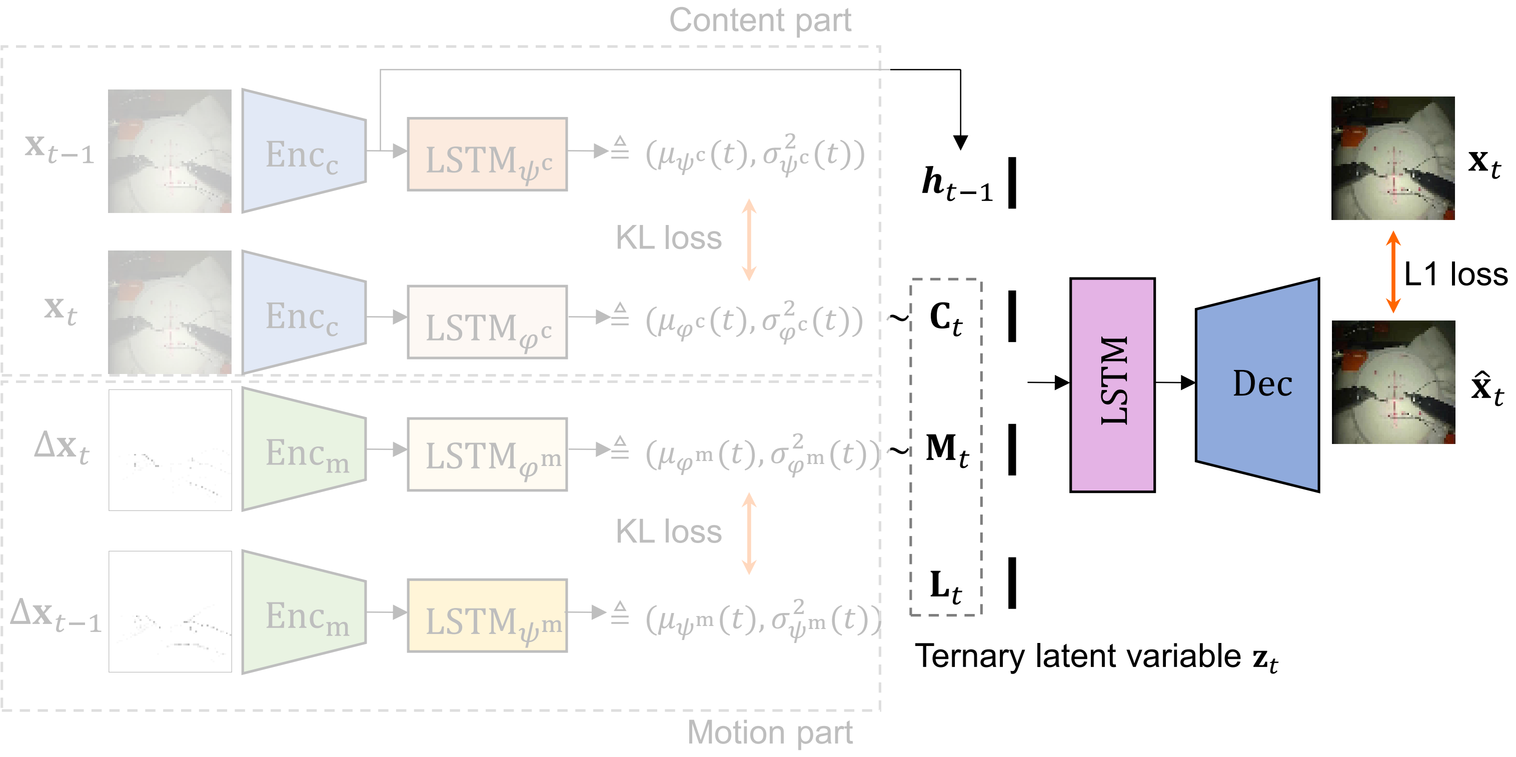


$$\mu_{\psi^c}(t), \sigma^2_{\psi^c}(t) = \mathrm{LSTM}_{\psi^c}(\mathbf{h}_{t-1}),$$
$$\mu_{\psi^m}(t), \sigma^2_{\psi^m}(t) = \mathrm{LSTM}_{\psi^m}(\mathbf{h}'_{t-1})$$

Proposed method

Ternary latent variable

- The class label information $L_t \in \{0,1\}^{n_l}$ is made as the non-learned part of $\mathbf{z}_t$.
- By referring to $\mathbf{z}_t$, the next frame could be generated using an LSTM and a decoder.



$$\mathbf{z}_t = [\mathbf{C}_t, \mathbf{M}_t, \mathbf{L}_t]$$

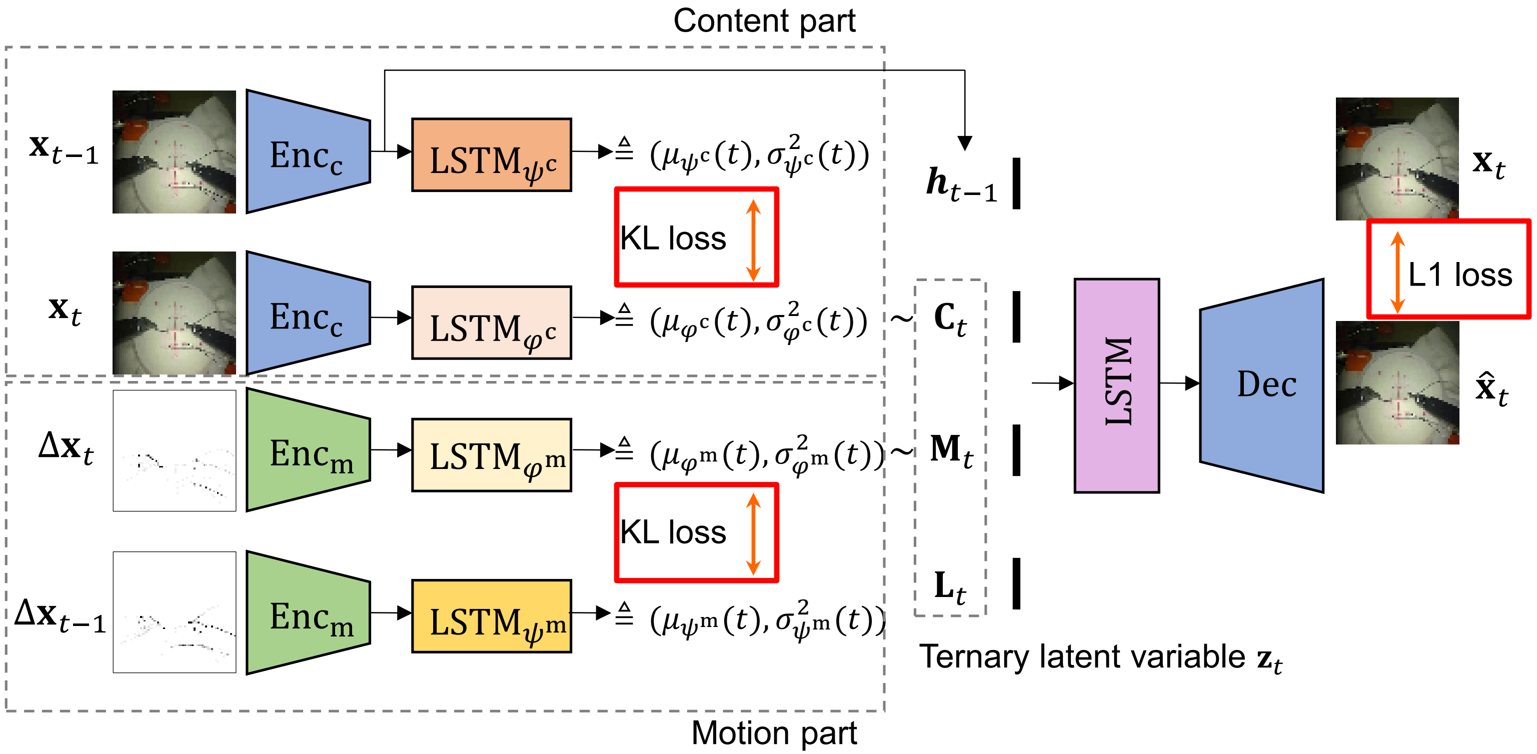
$$\begin{aligned}\mathbf{g}_t &= \mathrm{LSTM}_\theta(\mathbf{h}_{t-1}, \mathbf{z}_t),\\ \hat{\mathbf{x}}_t &= \mathrm{Dec}(\mathbf{g}_t).\end{aligned}$$

Proposed method

Learning process

The model is trained by maximizing the following variational lower bound:

$$\mathcal{L} = \sum_{t=1}^{T} [\|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_1 + \beta D_{\mathrm{KL}}(\mathcal{N}_{\varphi^{\mathrm{c}}}(t)||\mathcal{N}_{\psi^{\mathrm{c}}}(t)) + \beta D_{\mathrm{KL}}(\mathcal{N}_{\varphi^{\mathrm{m}}}(t)||\mathcal{N}_{\psi^{\mathrm{m}}}(t))]$$



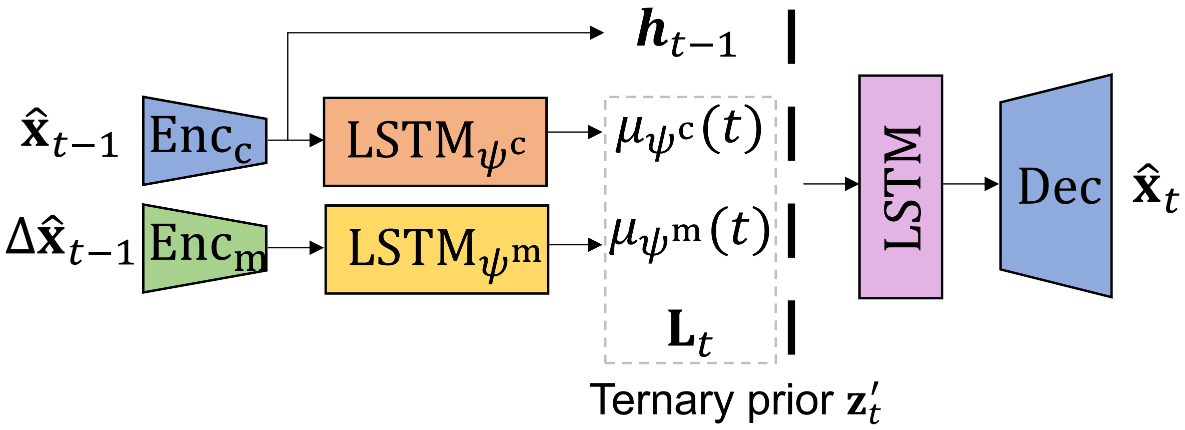
Content part
$\mathbf{x}_{t-1}$
$\mathrm{Enc_c}$
$\mathrm{LSTM}_{\psi^{\mathrm{c}}}$
$\triangleq (\mu_{\psi^{\mathrm{c}}}(t), \sigma^2_{\psi^{\mathrm{c}}}(t))$
KL loss
$\mathbf{x}_t$
$\mathrm{Enc_c}$
$\mathrm{LSTM}_{\varphi^{\mathrm{c}}}$
$\triangleq (\mu_{\varphi^{\mathrm{c}}}(t), \sigma^2_{\varphi^{\mathrm{c}}}(t))$
$\Delta\mathbf{x}_t$
$\mathrm{Enc_m}$
$\mathrm{LSTM}_{\varphi^{\mathrm{m}}}$
$\triangleq (\mu_{\varphi^{\mathrm{m}}}(t), \sigma^2_{\varphi^{\mathrm{m}}}(t))$
KL loss
$\Delta\mathbf{x}_{t-1}$
$\mathrm{Enc_m}$
$\mathrm{LSTM}_{\psi^{\mathrm{m}}}$
$\triangleq (\mu_{\psi^{\mathrm{m}}}(t), \sigma^2_{\psi^{\mathrm{m}}}(t))$
Motion part
$\boldsymbol{h}_{t-1}$
$\sim$ $\mathbf{C}_t$
$\sim$ $\mathbf{M}_t$
$\mathbf{L}_t$
Ternary latent variable $\mathbf{z}_t$
LSTM
Dec
$\mathbf{x}_t$
L1 loss
$\hat{\mathbf{x}}_t$

Proposed method

Inference style

- The expectations of the prior distribution without sampling are used to produce the most likely prediction.
- Choosing the best generation after sampling several times is not practical for online prediction scenarios.

$$\mathbf{z}'_t = [\boxed{\mu_{\psi^{\mathrm{c}}}(t), \mu_{\psi^{\mathrm{m}}}(t)}, \mathbf{L}_t]$$


$\boldsymbol{h}_{t-1}$
$\hat{\mathbf{x}}_{t-1}$
$\mathrm{Enc_c}$
$\mathrm{LSTM}_{\psi^{\mathrm{c}}}$
$\mu_{\psi^{\mathrm{c}}}(t)$
$\Delta\hat{\mathbf{x}}_{t-1}$
$\mathrm{Enc_m}$
$\mathrm{LSTM}_{\psi^{\mathrm{m}}}$
$\mu_{\psi^{\mathrm{m}}}(t)$
$\mathbf{L}_t$
LSTM
Dec
$\hat{\mathbf{x}}_t$
Ternary prior $\mathbf{z}'_t$

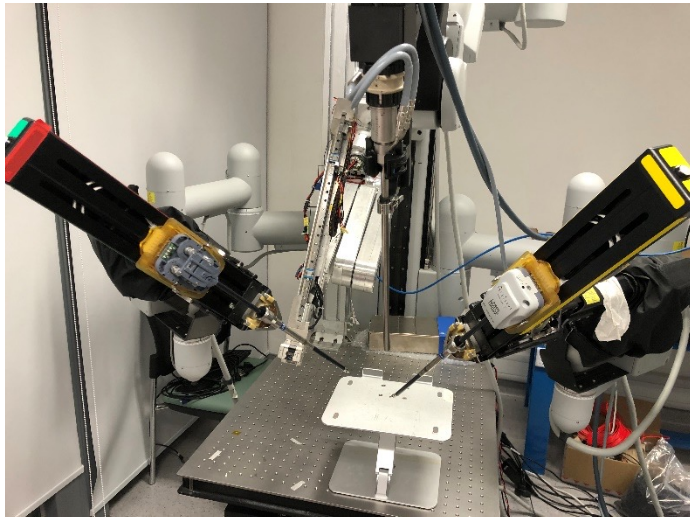
Experiments

Datasets

- The suturing task of the JIGSAWS dataset using *da Vinci* surgical system
- Video frames resized to 64×64
- Positioning needle (G2), pushing needle through tissue (G3), transferring needle from left to right (G4), and pulling suture with left hand (G6)
- Training set (470 sequences), testing set (142 sequences)

Evaluation metrics

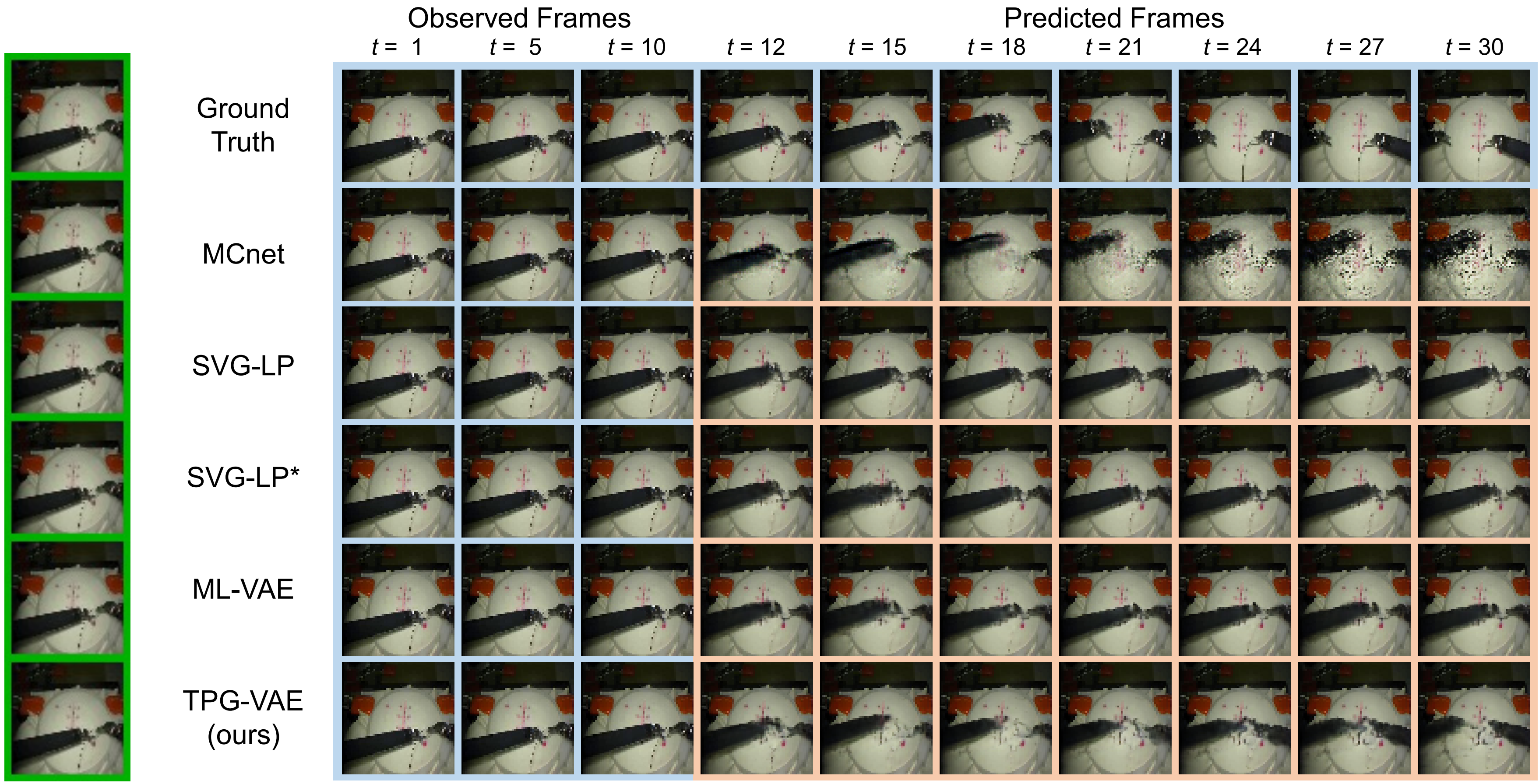
- VGG Cosine Similarity
- Peak Signal-to-Noise Ratio (PSNR)
- structural similarity (SSIM)



da Vinci surgical system

Experiments

Qualitative evaluation

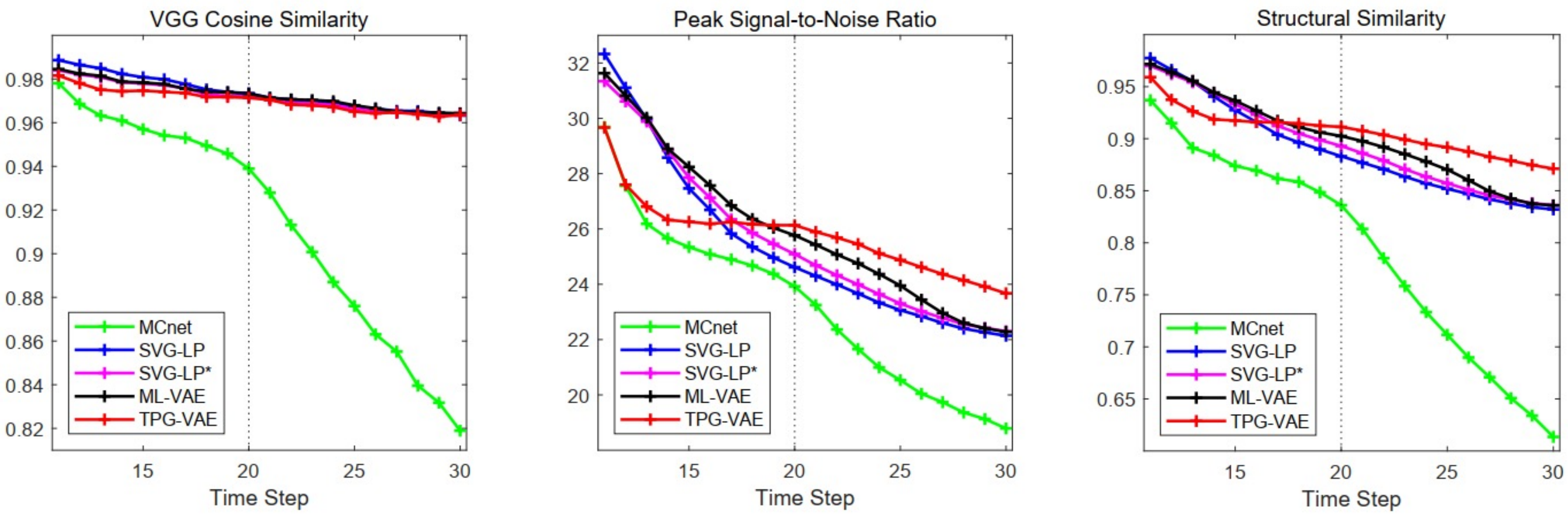


Qualitative results showing the gesture of G2 among different models.

Experiments

Quantitative evaluation

Quantitative evaluation on the average of the different metrics.



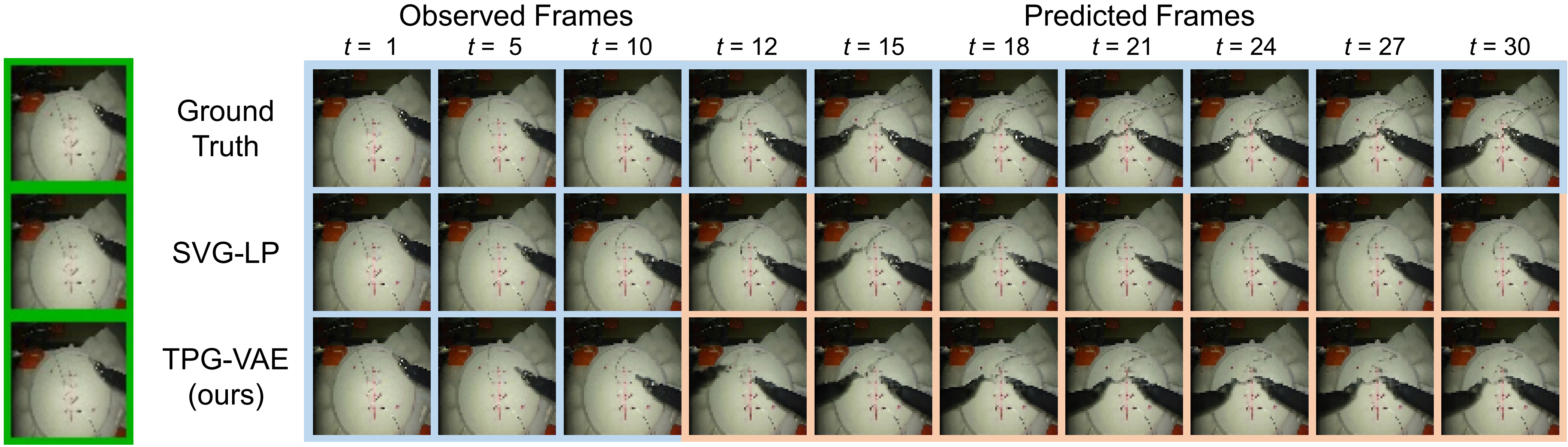
Methods	PSNR				SSIM			
	t=15	t=20	t=25	t=30	t=15	t=20	t=25	t=30
MCnet [1]	25.34±2.58	23.92±2.46	20.53±2.06	18.79±1.83	0.874±0.053	0.836±0.058	0.712±0.074	0.614±0.073
SVG-LP [2]	27.47±3.82	24.62±4.21	23.06±4.20	22.14±4.09	0.927±0.054	0.883±0.080	0.852±0.089	0.832±0.088
SVG-LP*	27.85±3.57	25.09±4.13	23.30±4.30	22.30±4.21	0.933±0.046	0.893±0.072	0.857±0.087	0.836±0.088
M-VAE	27.74±3.67	25.14±4.09	23.24±4.30	22.15±4.10	0.932±0.050	0.894±0.072	0.857±0.088	0.834±0.087
CM-VAE	27.44±3.83	25.09±4.07	23.02±4.19	22.16±4.16	0.927±0.056	0.893±0.075	0.853±0.087	0.834±0.088
CL-VAE	28.00±3.73	25.32±4.15	23.49±4.34	22.24±4.28	0.935±0.042	0.897±0.073	0.862±0.087	0.835±0.088
ML-VAE	**28.24±3.51**	25.77±4.02	23.95±4.26	22.28±4.26	**0.936±0.046**	0.903±0.071	0.870±0.084	0.836±0.088
TPG-VAE (ours)	26.26±3.17	**26.13±3.85**	**24.88±3.68**	**23.67±3.50**	0.917±0.048	**0.911±0.060**	**0.892±0.067**	**0.871±0.071**

[1] Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. In: ICLR (2017)

[2] Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: ICML (2018)

Discussions

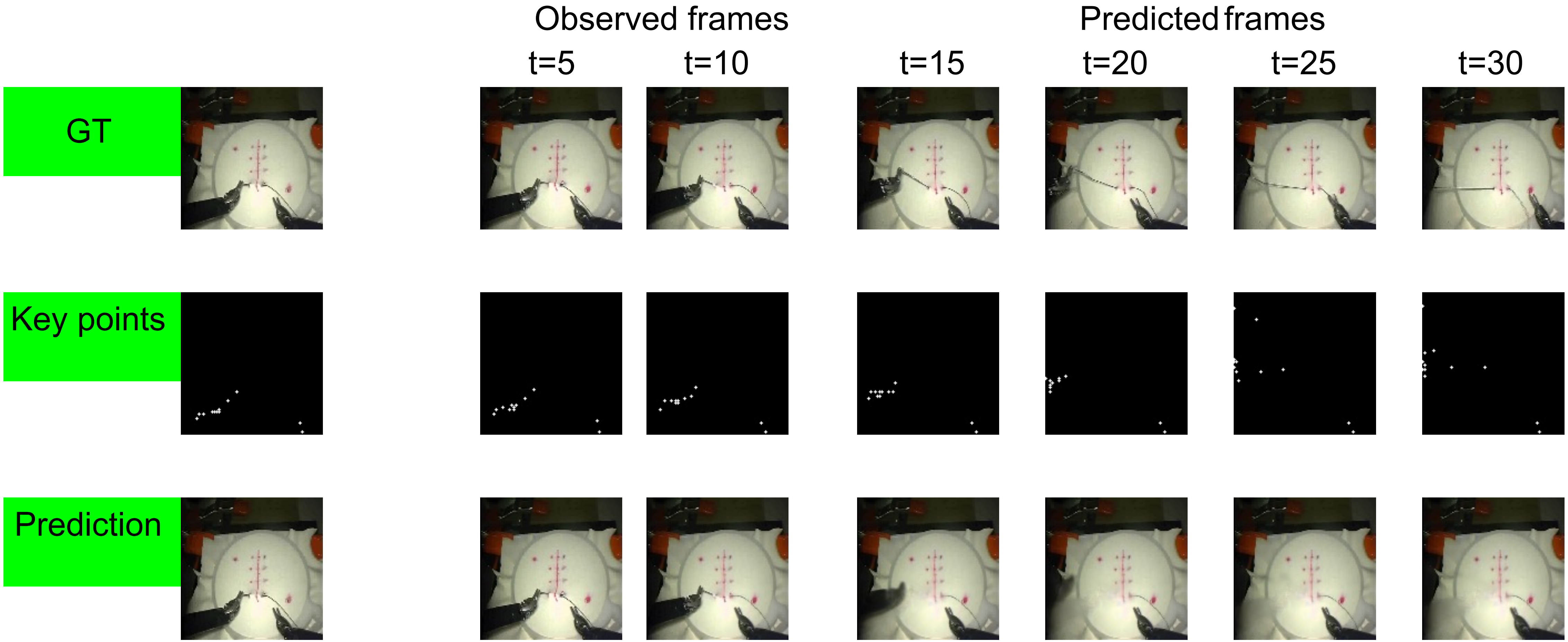
Dual Arm Cases



Qualitative comparison indicating the gesture of G4.

Following work

Decomposing surgical videos via interpretable key Points



Conclusions

- A novel TPG-VAE model to predict future scene in dual arm surgical robots.
- Inherent prior information is used as guidance to generate future scenes.
- SOTA results on the suturing task of the public JIGSAWS dataset.

Broader effect

Thank you!
Q & A