Technische Universität München – Faculty of Informatics
Chair for Computer Aided Medical Procedures (Prof. Nassir Navab)
Practical Course: Machine Learning in Medical Imaging WS23/24

Future Video Generation with Scene Graphs in Medical Imaging

1 General Info

Project Title: Future Video Generation with Scene Graphs in Medical Imaging
Supervisors: Azade Farshad, Yousef Yeganeh
Contact Email: azade.farshad@tum.de, y.yeganeh@tum.de

2 Background and Motivation

Video generation involves synthesizing realistic and diverse video frames from a given input. In medical imaging, video generation can have various applications, such as data augmentation, anomaly detection, and surgical flow simulation. However, most existing methods for video generation rely on pixel-level information and ignore the high-level semantic structure of the scene. Our project seeks to address this by leveraging higher-level inputs, such as text descriptions or scene graphs, to guide video generation. Specifically, scene graphs offer a visual representation of a scene's objects, their attributes, and the relationships between them, encapsulating both semantic and spatial details of images. Ultimately, we aim to produce subsequent video frames based on the initial image and a dynamic scene graph. This graph will detail the evolving changes in the scene as time progresses.

3 Project Abstract

The project consists of the following steps: 1) Optical flow estimation: In this step, we would like to estimate the optical flow between the input image and the next frame. Optical flow is the pattern of apparent motion of pixels in an image, which can indicate the direction and magnitude of the movement of objects [1]. Optical flow estimation can help us capture the temporal dynamics of the scene and provide a smooth transition between frames [3]. This step would involve training a model that would estimate the optical flow from a single image and dynamic scene graphs defining the actions in the scene. 2) Video Generation: We will use SOTA generative models such as Waldo [2], CoDi [6] and Stable Diffusion [5] to predict the next frame conditioned on the input image, the dynamic scene graph and optionally the flow maps from the first step. The dynamic scene graph encodes the semantic information of the scene, such as the objects, their attributes, and their relations. The methods would be trained and evaluated on the CholecT50 dataset [4] that contains videos of laparoscopic cholecystectomy surgery.

4 Technical Prerequisites

- Good background in machine learning and deep learning
- Experienced in PyTorch
- Experienced in Python
- Experience with Generative Models

5 Benefits

- Weekly supervision and discussions
- Possible novelty of the research
- The results of this work are intended to be published in a conference or journal

6 Work packages and Time-plan

* The dates are adopted from the previous year and are not finalized yet.

	Description	# Students	From	To
WP1	Familiarizing with the literature.	4	24.10	31.10
WP2	Implementing the baselines	4	31.10	14.11
WP3	Improving the baselines and validation on relevant datasets	4	14.11	27.11
	Midterm Presentation (Date is not finalized)	4	27.11	05.12
WP4	Implementing the model	4	05.12	19.12
WP5	Finalizing the results and evaluation	4	19.12	07.02
	Final Presentation (Date is not finalized)	4	07.02	14.02

Table 1: Project Timeline

References

[1] Optical Flow. `https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html`, 2023. [Online; accessed 24-October-2023].

[2] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Waldo: Future video synthesis using object layer decomposition and parametric flow prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23229–23241, 2023.

[3] Ao Luo, Fan Yang, Kunming Luo, Xin Li, Haoqiang Fan, and Shuaicheng Liu. Learning optical flow with adaptive graph reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1890–1898, 2022.

[4] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[6] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023.