Introduction

My name is Wenliang Guo. I hold a Bachelor of Engineering degree from Xidian University, China, with a major in telecommunications engineering. My academic background has provided me with a solid foundation in mathematics, signal processing, and coding skills. Currently, I am pursuing a Master of Science degree at Columbia University, USA, with an expected graduation date in December 2023. My primary research interests lie in computer vision, particularly in multi-modality, representation learning, and their applications in various video and image understanding tasks. I am highly enthusiastic about exploring new frontiers in this field and am open to a wide range of vision-related topics.

Research Experience

During my undergraduate studies, I had the privilege of being mentored by Prof. Xiao Xiao as a student researcher. Together, we worked on a project about image recognition on mobile platforms, delving into image segmentation for various scenarios and focusing our research on feature scales. This experience guided me to the world of computer vision.

My first projects concentrated on blur detection, which aims to realize pixel-level discrimination between clear and blurred areas due to defocusing or object motion. It is meaningful for the development of photographic systems because the blur detection algorithm produces fine-grained blur information, which not only benefits the image post-processing by saving a great deal of labor costs while maintaining high discrimination accuracy, but also brings opportunities to improve the image quality, such as facilitates image tuning through system hardware and software.

Our work [1] introduced a pyramid-pooling encoder and a nested U-Net decoder to enhance multi-scale feature extraction efficiently without significantly increasing parameters. Channel attention is also integrated into our model to increase the weight of informative features. My responsibilities included coding, experimentation, and manuscript writing. In retrospect, the innovation of my first work seems to be limited, but it laid a strong foundation for my future research endeavors and honed my problem-solving, experimental design, and academic writing skills.

Motivated by this experience, I furthered my exploration into image segmentation, leading a research project on building extraction in remote sensing images. This task is essential in applications such as regional administration, disaster prevention, and map services. From the perspective of modern computer vision, building extraction is one of the applications of image segmentation, while its specific challenge being that remote sensing images are often high-resolution, leading to buildings covering large pixel-areas. Previous CNN-based algorithms faced limitations in extracting large-scale semantic features due to the local receptive field of the convolution kernel, resulting in missing or incorrect building segmentation. Therefore, our motivation was to develop a model capable of efficiently extracting large-scale semantic features.

Inspired by the success of Vision Transformer (ViT), we aimed to enhance models' representation learning abilities using ViT. Related work mainly focused on fusing the ViT and U-Net at the network structure level in different ways. However, despite their improved final accuracy, they lacked flexibility, generalization to other vision tasks, and interpretability of learned representations. In contrast, we proposed a simple, yet effective encoding booster based on the Swin Transformer and a hierarchical fusion of features extracted by the Transformer and U-Net at different scales, fully exploiting their advantages in large-scale feature extraction and high localization accuracy [2]. Our approach is highly flexible, scalable, and interpretable. It is applicable to various downstream tasks and can be easily integrated into models for improving

representation learning. My contributions to this work encompassed problem specification, idea proposal, experimentation, and drafting.

Current Research

My current role as a research assistant in the Digital Video and Multimedia (DVMM) Lab at Columbia University, under the supervision of Dr. Yulei Niu and Prof. Shih-fu Chang, focuses on multi-modality and video understanding. Our recent work [3] centers on procedure planning in instructional videos, which aims to arrange a sequence of instructional steps to achieve a specific goal, given the image observations at both the beginning and end of the procedure. Procedure planning is an essential and fundamental reasoning ability for embodied AI systems and is crucial in complicated real-world problems like robotic navigation.

Recognizing that observations captured from instructional videos may contain noisy or irrelevant image content, direct use or fine-tuning of a pre-trained visual encoder can lead to low-quality features and a degradation in the model's performance. We introduced language state descriptions generated by a large language model to guide the encoder in learning better and more interpretable vision-language representations aligned with visual state observations. Additionally, we decomposed procedure planning into subproblems of subgoal decomposition and action step prediction, leveraging the generated state descriptions as external memory for cross-modal Transformers. Experimental results demonstrated that our method significantly improved the performance. My responsibilities in this project included codebase development and experimentation.

Our current research direction involves enhancing the reasoning abilities of embodied AI systems for procedure planning. A trustworthy embodied AI system should be able to reason reliably, interpretably, and transparently based on causal factors, even in unknown environments or with unseen inputs. To achieve this, it should support counterfactual reasoning, enabling users to explore "what-if" scenarios and understand the potential outcomes of different interventions or decisions. Therefore, we plan to introduce counterfactual planning to multi-modal procedure planning, allowing the model to generate or revise plans based on a given goal while adhering to additional counterfactual conditions.

Future Plan

My past research experiences, spanning low-level vision tasks to multi-modality, have enriched my knowledge, and equipped me with valuable skills in literature review, problem-solving, and data analysis. They also fueled my passion and strengthened my resolve to conduct research in this field.

In the future, I aspire to continue contributing to computer vision-related research, encompassing topics such as video/image understanding, compositionality, generative models, robotic, scene reconstruction, VR/AR, autonomous driving, and low-level tasks. I am also interested in applying visual models to causal trustworthy AI systems.

[1] Guo, Wenliang, Xiao Xiao, Yilong Hui, Wenming Yang, and Amir Sadovnik. "Heterogeneous attention nested U-shaped network for blur detection." IEEE Signal Processing Letters 29 (2021): 140-144.

[2] Xiao, Xiao, Wenliang Guo, Rui Chen, Yilong Hui, Jianing Wang, and Hongyu Zhao. "A swin transformer-based encoding booster integrated in u-shaped network for building extraction." Remote Sensing 14, no. 11 (2022): 2611.

[3] Preprint: Yulei Niu, Wenliang Guo, Long Chen, Xudong Lin, and Shih-Fu Chang. "State-Enhanced Procedure Planning in Instructional Videos", https://openreview.net/forum?id=5FeRi11ARd.