Biography

My name is Wenliang Guo. I hold a Bachelor of Engineering degree from Xidian University, China, with a major in tele-communications engineering. Currently, I am pursuing a Master of Science degree at Columbia University, US, with an expected graduation date in February 2024. My academic background has provided me with a solid foundation in mathematics, signal processing, and coding skills. My primary research interests are 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. In this work, my responsibilities included coding, experimentation, and manuscript writing.

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, with 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 applies 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.

[1] **Wenliang Guo**, 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.

Current Research

Now I am working 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. Our research focuses on multi-modality and video understanding. Our recent work centers on procedure planning in instructional videos. My responsibilities include developing codebase and experimenting.

Procedure planning is an essential and fundamental reasoning ability for embodied AI. It aims to arrange a sequence of instructional action steps to achieve a specific goal, given the image observations at both the beginning and end of a procedure. Recent works succeeded in sequence modeling of action steps with only sequence-level annotations during training. However, simply representing an action by its name is unreliable and non-interpretable because all actions' names are highly abstracted human language that only contains linguistic but not visual semantics. An AI model tends to use language-priors mined from a large amount of training data to predict the next action, without being able to understand the motivation and insight behind the action itself. Unlike current planning algorithms, human action is driven by the visual state of the object and the goal. For instance, to make a steak, seeing that the current state of the steak is cooked and intact, we perform the action of cutting the steak to divide it into pieces for consumption.

Following the above motivation, our work [3] pointed out that the state change does matter for procedure planning in instructional videos which is overlooked in most existing works. We proposed to establish a more structured state space by investigating the causal relations between action steps and visual states in procedures. Specifically, we explicitly represent each action step as state changes and track the state changes in procedures. For step representation, we leveraged the commonsense knowledge in large language models (LLMs) to describe the state changes of action steps via our designed chain-of-thought prompting. For state change tracking, we align visual state observations with language state descriptions via cross-modal contrastive learning, and explicitly model the intermediate states of the procedure using LLM-generated state descriptions. Experimental results demonstrated that our method significantly improved the performance.

Future Plan

My past research experiences have equipped me with valuable skills in literature review, data analysis, and coding. They also fueled my passion and strengthened my resolve to conduct research in this field.

In the future, my ultimate research goal is to develop human-aligned embodied AI. It will be able to perceive its surroundings autonomously at a fine-grained level, fuse multi-modal signals, and reason compositionally. I believe the future embodied AI should be an aggregator of current computer vision and multi-modal models, which is what crucially distinguishes it from large fundamental models. To this end, I aspire to continue diving into computer vision research, including topics such as video/image understanding, generative modeling, robotics, and scene reconstruction. I am also interested in building causal trustworthy visual systems, which enable further alignment of AI with human reasoning patterns.

Combined with my research experience, I think I could start with instructional video understanding. It is an interesting topic and still leaves many open questions. For example, I can investigate the action representations that allow AI systems to understand the causal relationships of actions, or the way of utilizing and updating external knowledge to enable AI systems to learn new actions in a human-like manner. I may also study related multi-modal tasks based on instructional videos, including video frame retrieval, action grounding, and open vocabulary detection, which greatly aligns with my interests.

[3] Yulei Niu, **Wenliang Guo**, Long Chen, Xudong Lin, and Shih-Fu Chang. "SCHEMA: State CHangEs MAtter for Procedure Planning in Instructional Videos", https://openreview.net/forum?id=abL5LJNZ49. (***Under review, submitted to ICLR 2024***)