Research Statement of Zijun Wei

My primary research interests span Computer Vision, Machine Learning, and Human Perception. A major goal of computer vision is to make machines understand the world as humans do. Typical computer vision tasks such as image/video classification, detection or segmentation aim to recognize objective entities such as object categories, locations and shapes. However, in addition to these tasks, human-centered artificial intelligence systems should also be able to recognize the subjective components of human perception in visual data, *i.e.*, predicting regions that attract human attention, recognizing subjective attributes (*e.g.*, implied sentiments) and extracting spatial or temporal regions that capture human interest (video summarization or image cropping). Progress in these studies will improve human computer interaction experiences in various applications. The driver of my research is to **model the subjective components of human perception and develop robust computer vision systems that produce the optimal content for subjective perception**. My research so far has focused on the following components:

1. Region Ranking and Selection for Image Recognition and Attention Prediction. Recognizing the semantic category of an image is challenging because the location of the semantic region that we need to recognize is unknown. However, humans are remarkably good at this because they attend the image regions in a sparse and diverse manner. Inspired by this human attention mechanism, I developed Region Ranking SVM (RRSVM, CVPR2016, see attached CV for reference), an image classification algorithm that incorporates locally sampled information into global decisions. RRSVM achieved state-of-the-art performance on image classification tasks. I further enhanced RRSVM by incorporating the biologically mechanism Inhibition of Return to impose diversity on the selected regions (NIPS2016). The model, while not trained with eye movement data, predicted well where humans will attend. As RRSVM is general, I applied it to classify reading behavior as reading or skimming based on eye movement data (ETRA2019).

2. Automatic image cropping. Automatic photo good taking requires understanding the subjective perceptual choices that would satisfy the human viewer. The task of finding views with good photo composition is simple for humans but very challenging for computer vision algorithms because learning the subjective choices requires large scale annotated datasets. In my CVPR2018 paper I developed a cost-effective crowd sourcing workflow and created the first large scale Comparative Photo Composition dataset, which contains over one million annotated comparative view pairs. I showed that this dataset is essential for training deep networks to generate well-composed photos. A novel knowledge transfer network led to a deep model that runs at real-time and achieved the state-of-the-art performance. I further developed the real-time view proposal network into *SmartEye* (CHI2019), a mobile system that incorporates the personalized composition preferences of users, to interactively help them compose good photos.

3. Video Summarization. A similar problem to image composition in video streams is finding and summarizing interesting segments from videos. Video summarization is challenging as it often requires not only knowledge of individual video clips, but also contextual, often subjective understanding of the entire video and the relationships between video clips. In our Sequence-to-Segments Network NeurIPS 2018 paper I developed a general architecture that generates video summarization by first encoding the entire video information and then decoding segments based on the video. The architecture also achieved state-of-the-art performance in multiple other related tasks, such as video highlighting and human action proposal generation.

4. Fine-grained visual sentiment analysis with natural language processing. The subjective understanding of images by humans is largely revealed by the language they use to describe them. This is particularly true when one wants to associate the sentiment of images based on the words used to describe them. In my current research (submitted to ICCV2019), I developed a bottom-up language-driven approach to learn diverse and fine-grained visual sentiments from the annotation-free text data in the internet. The algorithms naturally introduced a joint visual-text embedding space of sentiment that enables multiple applications such as image retrieval or automatic image annotation.

Future research directions. I am enthusiastic to continue my research on constructing explainable feature space for the subjective components of human perception using supervision from natural sources such as eye movements and natural language. The feature space will not only model the subjective components for computer vision tasks, but also will shed light on understanding the human visual process in the brain.