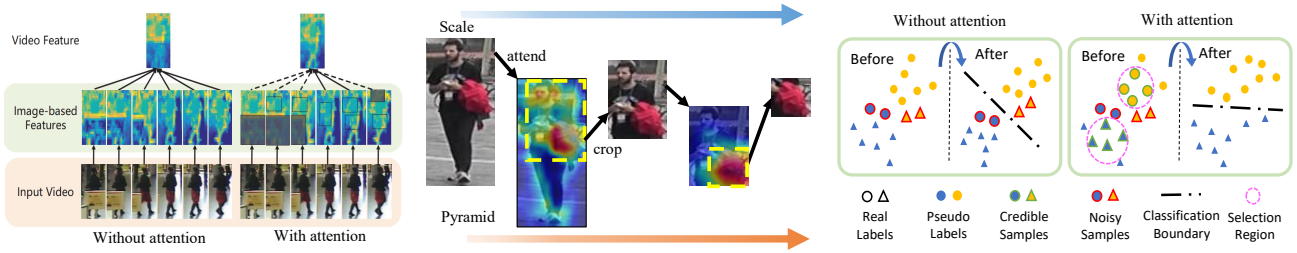
RESEARCH STATEMENT

Guangyi Chen (chen-gy16@mails.tsinghua.edu.cn)

Over the last decade, the great progress has been achieved in the field of computer vision by the success of deep learning. Despite the powerful representation ability of deep learning, it is still challenging to imagine and reason like human. Therefore, my research aims at developing robust and explainable visual understanding models by imitating the cognitive process of human brain. I believe that the brain-inspired models such as attention, memory, imagination, and reasoning has great potential in visual understanding. There are two principal directions I have explored: 1) attention learning, and 2) causal learning. In the following, I will explain my research on these two directions and corresponding applications to different computer vision tasks such as video understanding, person re-identification, fine-grained recognition, and human trajectory prediction.



(a) Spatial-temporal attention (b) Attention pyramid (c) Attention in the sampling

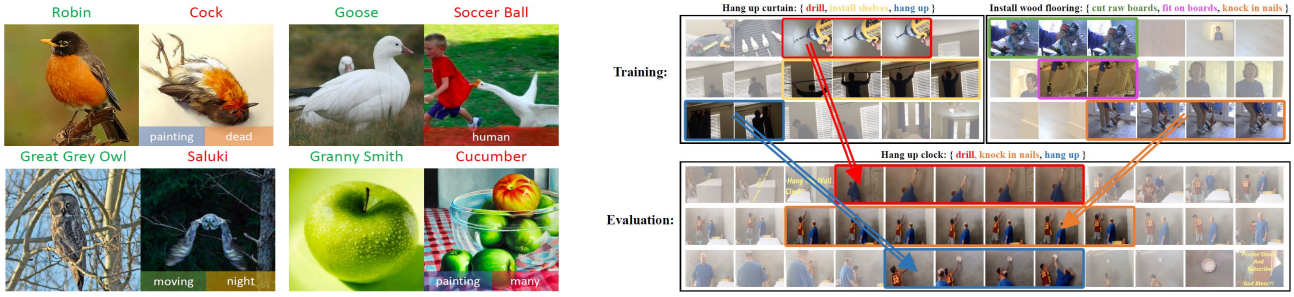
Figure 1: The proposed attention learning methods: (a) Learning spatial-temporal attention for video understanding; (b) Learning attention pyramid for multi-scale saliences; (c) introducing the attention in the sampling process.

Attention Learning

Attention mechanism plays an important role in the human conscious processing to abstract out the irrelevant information. I have been working on exploring the attention mechanism in the computer vision systems, which aims to facilitate high-performance recognition by discovering discriminative regions and mitigating the negative effects brought by diverse visual appearance, cluttered backgrounds, occlusions, pose variations, etc. My research on attention learning mainly focused on three challenges: designing effective attention models [3–5, 7] (especially for videos containing complex spatial-temporal clues [3–5]), learning attentions in a weakly-supervised manner [1, 10], and introducing the attention model in the sampling process of metric learning [2, 6].

Designing effective attention models The videos contain complex spatial-temporal clues and heavy negative misalignments. As shown in Figure 1 (a), I have explored to develop the spatial-temporal attention model [3] to learn the robust representations by jointly mining the salient clues of videos in both spatial and temporal domain. Moving ahead, motivated by the inherent consistencies between spatial and temporal clues, I have proposed to attend the 3D regions by treating the video as a unified 3D bin [4]. Besides, as shown in Figure 1 (b), in order to mine the salient clues in different scales, I have developed an attention pyramid network with the designed "split-attend-merge-stack" principle [7]. Results show the robust representation ability brought by these attention models.

Learning attentions in a weakly-supervised manner Despite the widespread use, the challenge of how to learn effective attention is still barely studied. Most existing methods learn the visual attention in a weakly-supervised manner, i.e., the attention modules are simply supervised by the final



(a) Examples in FairBench1K (b) An example in GAIN

Figure 2: The examples of benchmarks for fairness and generalizability: (a) The examples with out-of-distribution attributes; (b) An example of how to generate out-of-distribution instructional task.

loss function, without a powerful supervisory signal to guide the training process. Hence, I proposed a self-critical attention learning method [1] which learns a critic to measure the attention quality and provide supervisory signals. Furthermore, I explored the attention learning in a causal perspective and proposed a counterfactual attention learning [10] to analyze the effects of learned visual attention with counterfactual causality. These methods improve the learning process of attention and achieve better performance on fine-grained recognition.

Introducing the attention model in the sampling process. The effects of samples have a large variance in the training process, where easy samples hardly produce effective gradient while hard samples caused by noise labels may mislead the model. To mine valuable samples, I have introduced an attention model in the sampling process. I have proposed a deep meta metric learning method, which formulates the metric learning process in a meta learning perspective and apples attention to mine representative samples in the set to learn the set-based distance. Besides, as shown Figure 1 (c), I have also proposed to mine credible samples for unsupervised domain adaptation to avoid the misleading from noise labels.

Causal Learning

Existing computer vision systems are good at learning what the object is (i.e., visual recognition) or where it is (i.e. object detection and segmentation), yet bad at explaining why it is that. These systems learn to predict based on likelihood, instead of the underlying causation. Recently, I have been working on applying the tools of causal inference for computer vision systems to alleviate the negative effects brought by confounding data bias and enhance the model's generalizability, fairness, and explainability. My research focuses on benchmarking the fairness and generalizability by causal inference [11, 13] and mitigating the data bias by counterfactual comparison [9, 10, 12].

Benchmarking the fairness and generalizability. Despite the remarkable progress on computer vision thanks to the deep learning techniques, computer vision models still perform unfavorably for out-of-distribution samples due to the dataset bias. This fact reveals that traditional metrics like classification accuracy may overestimate the capacity and reliability of the models. Towards a more comprehensive evaluation, I have introduced a large real-world dataset to benchmark fairness of image recognition models [11]. As shown in Figure 2 (a), the examples with out-of-distribution attributes require the fairness of models with regard to different protected attributes. Besides, I have also introduced a dataset to benchmark the generalizability of instructional video analysis models [13]. Figure 2 (b) illustrates an example that an out-of-distribution task are generated with in-distribution training

samples. I furthermore proposed a method to enhance the generalizability by cutting off excessive contextual dependency with sampling intervention.

Mitigating the data bias by counterfactual intervention. Limited by the dataset scale and annotation level, the models are always inevitably misled by the data bias to focus on the spurious correlations in training data. To mitigate the influence of confounding data bias, I have proposed to learn the model by counterfactual intervention. Counterfactual intervention is of critical importance in causal inference, which encourages the model to involve consideration of an alternate version of a past event. I have proposed to conduct counterfactual attention by imagining non-existent attention maps, and maximize the effect of attention model by comparison [10]. Besides, I proposed a retrieval-based counterfactual intervention method which retrieves videos from an off-line video pool as counterfactual example to optimize the unintentional action localization model [12].

Future Research Plans

My research goal is developing robust and explainable visual understanding models, which imitates the cognitive process of human to understand the world. I'm a strong believer that brain-inspired model may be a potential solution to achieve general machine intelligence. Beyond the attention model used for observing, I believe causal effect and commonsense knowledge are key for human reasoning. Motivated by this, my future research plans are organized around two topics: 1) general causal learning and 2) commonsense knowledge.

General causal learning. Beyond the conventional statistical learning, I believe the causal learning has a great potential to allow the models to support intervention, planning, and reasoning. Many efforts have been made to develop causal learning to improve machine learning methods. However, existing causal learning methods in computer vision systems always have two challenges. First, the causal learning relies on high-level semantic representation. Hence, I plan to extend the causal learning into representation learning based on low-level observations. It will improve not only the performance of model prediction but also the explainability of feature representation. Second, most of causal learning methods are limited by the strong human prior of structural causal model. Inspired by the self-supervised learning, I plan to adaptively discover the causal relations by the structure in the data and learn the causations without the prior of causal model. Though there is a long way to go, I believe both two research directions are towards the goal of general causal learning.

Commonsense knowledge. I believe the commonsense is another key for computer vision systems to understand the real world like human. It is desired to build the commonsense knowledge graph in the field of computer vision. It is still challenging when climbing this ladder. In [12], I built an off-line video pool as the commonsense to analyze the intention of human action. I also explored the attribute disentangling of visual objects in [8] and [11]. Motivated by these experiences, I plan to build commonsense knowledge graph with the intrinsic relations among attributes. Visual objects will be associated by reasoning the relations in the semantic attributes. Moving forward, I plan to build a large real world dataset by collecting the visual objects and the relations among them. I expert this dataset can promote future in-depth research on commonsense knowledge in computer vision.

There is still a long way to go towards these goals. I am excited about and ready for these new challenges.

References

[1] Guangyi Chen, Chenze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou, "Self-Critical Attention Learning for Person Re-identification", *ICCV*, 2019.

[2] Guangyi Chen, Tianren Zhang, Jiwen Lu, and Jie Zhou, "Deep Meta Metric Learning", *ICCV*, 2019.

[3] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou, "Spatial-Temporal Attention-aware Learning for Video-based Person Re-identification", *TIP*, 2019.

[4] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou, "Learning Recurrent 3D Attention for Video-based Person Re-identification", *TIP*, 2020.

[5] Guangyi Chen*, Yongming Rao*, Jiwen Lu, and Jie Zhou, "Temporal Coherence or Temporal Motion: Which is More Critical for Video-based Person Reidentification?" *ECCV*, 2020.

[6] Guangyi Chen, Yuhao Lu, Jiwen Lu, and Jie Zhou, "Deep Credible Metric Learning for Unsupervised Domain Adaptation Person Re-identification" *ECCV*, 2020.

[7] Guangyi Chen, Tianpei Gu, Jiwen Lu, Jinan Bao, and Jie Zhou, "Person Re-identification via Attention Pyramid" *in submission to TIP*, 2020.

[8] Guangyi Chen, Weilin Huang, Jiwen Lu, Jinan Bao, and Jie Zhou, "Learning Attribute-Disentangled Embeddings for Person Re-identification" *in submission to TIP*, 2020.

[9] Guangyi Chen, Junlong li, Jiwen Lu, and Jie Zhou, "Human Trajectory Prediction via Counterfactual Analysis" *in submission to ICCV*, 2021.

[10] Yongming Rao*, Guangyi Chen*, Jiwen Lu, and Jie Zhou, "Counterfactual Attention Learning for Fine-grained Recognition" *in submission to ICCV*, 2021.

[11] Yongming Rao*, Guangyi Chen*, Wenliang Zhao*, Jiwen Lu, and Jie Zhou, "FairBench1K: Benchmarking Fairness of Image Recognition Models" *in submission to ICCV*, 2021.

[12] Jinglin Xu*, Guangyi Chen*, Nuoxing Zhou, and Jiwen Lu, "Unintentional Action Localization via Counterfactual Examples" *in submission to ICCV*, 2021.

[13] Junlong Li*, Guangyi Chen*, Yansong Tang, Jinan Bao, Jiwen Lu, and Jie Zhou, "GAIN: Benchmarking Generalizability of Instructional Video Analysis Models" *in submission to ICCV*, 2021.