Huilin ZHOU, Qihan REN, Junpeng ZHANG, Quanshi ZHANG, 2025. Towards the first principles of explaining DNNs: interactions explain the learning dynamics. *Frontiers of Information Technology & Electronic Engineering*, 26(7):1017-1026.
https://doi.org/10.1631/FITEE.2401025

Towards the first principles of explaining DNNs: interactions explain the learning dynamics

Key words: First-principles explanation; Theory of equivalent interactions; Two-phase dynamics of interactions; Learning dynamics

Corresponding author: Quanshi ZHANG
E-mail: zqs1022@sjtu.edu.cn
ORCID: https://orcid.org/0000-0002-6108-2738

Motivation

- Why does this work matter?

1. DNNs perform well, but lack interpretability—critically in high-stake applications like medicine or autonomous driving.

2. Existing explainable artificial intelligence (XAI) methods are mostly empirical, lacking a rigorous theoretical foundation.

3. There is no consensus on a “first-principles explanation” that unifies knowledge, generalization, robustness, and learning dynamics.

- This paper’s aims

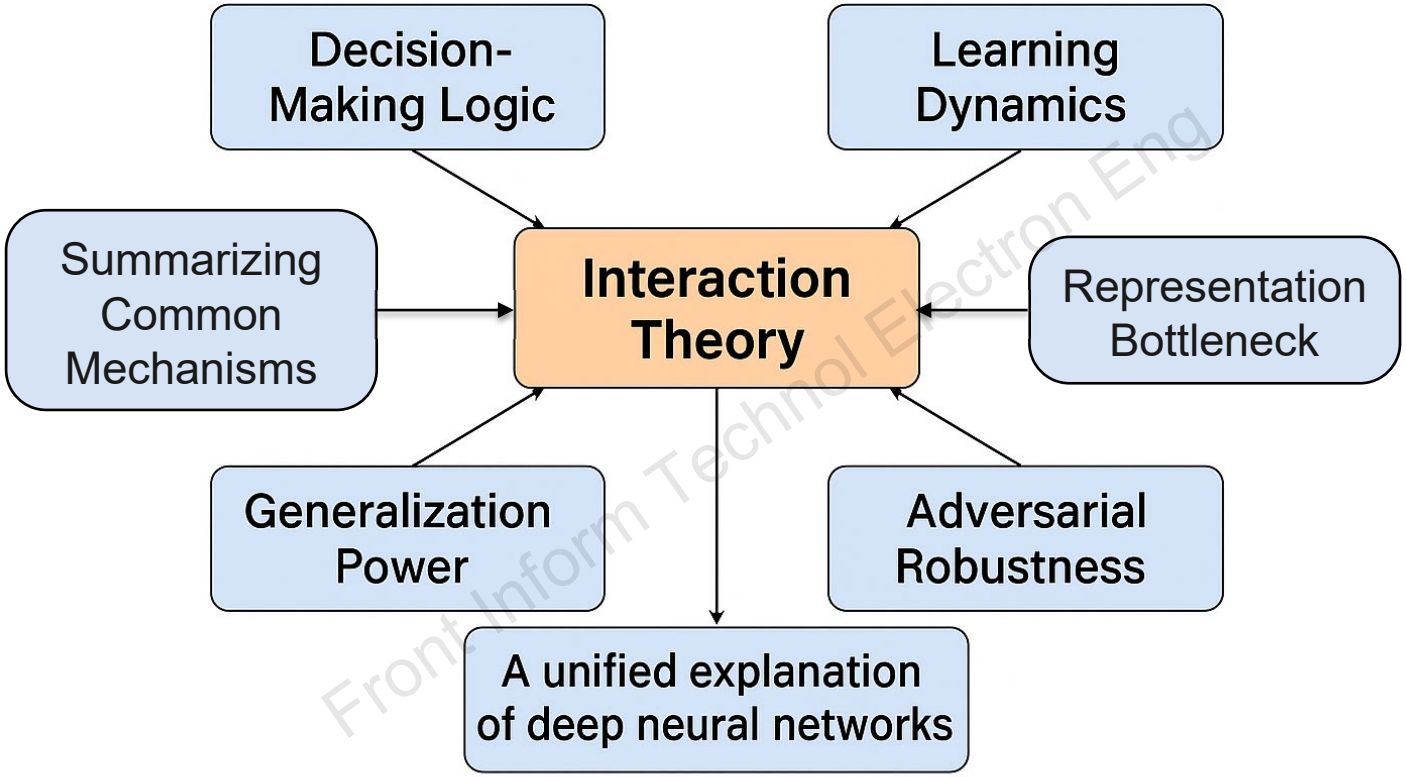
1. Explore whether the “interaction theory” can serve as the basis for such a foundational explanation.

2. Analyze interaction theory’s ability to explain representation, unify empirical methods, and model DNN learning dynamics.

Main idea

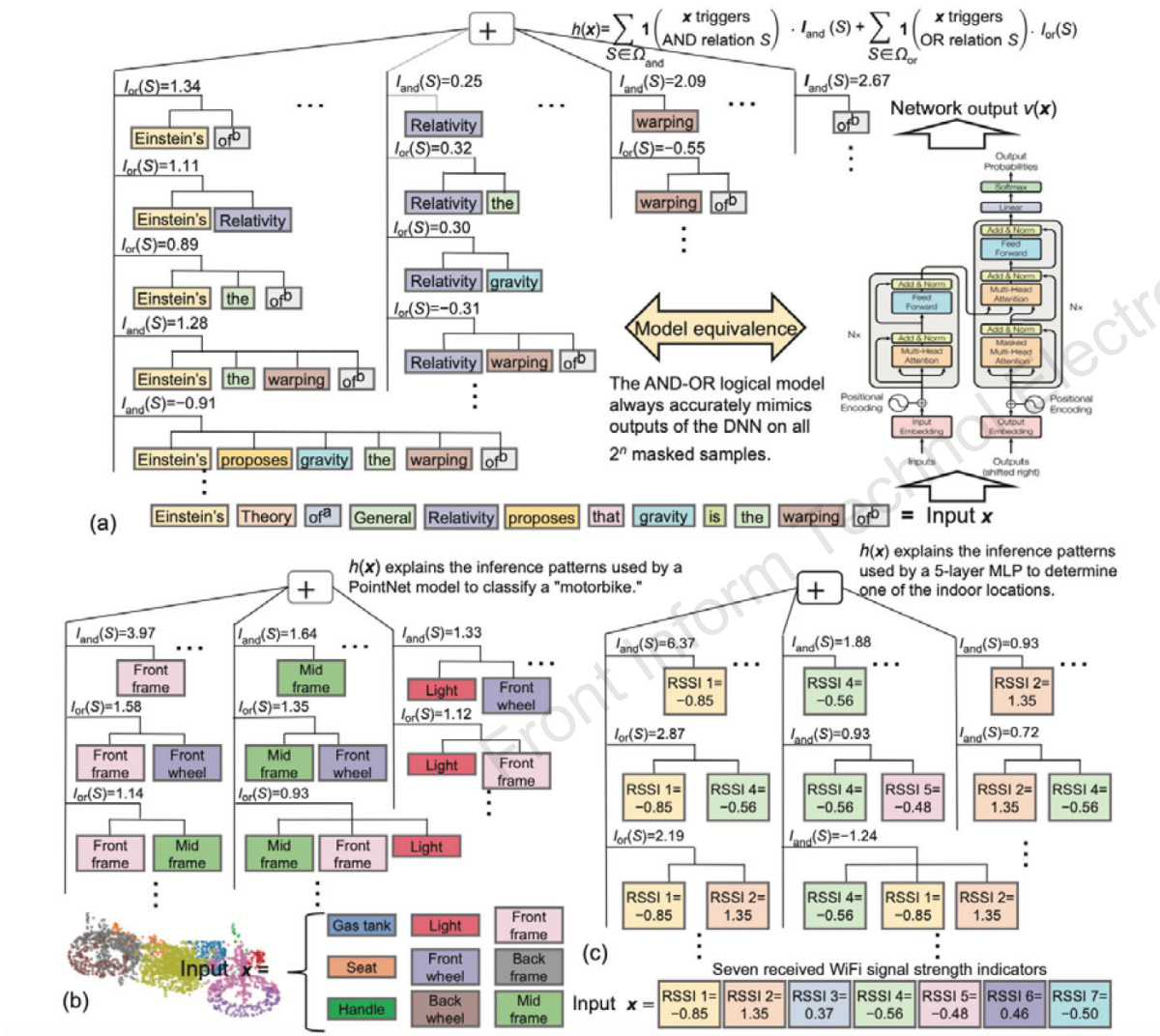
- Convert DNN logic into symbolic AND/OR interactions.
- Form a theoretical axiomatic system explaining DNN behaviors.
- Cover generalization, robustness, representation bottleneck, and learning dynamics.

Framework



Overview of the interaction theory

Explaining the decision logic of DNNs



AND-OR logical model that explains the inference patterns used by a DNN

The inference of a DNN on a certain input sample is equivalent to a surrogate logical model that uses a small number of AND-OR interactions for inference. Each interaction corresponds to an AND relationship or an OR relationship between a set of input variables.

Explaining the representation capacity of DNNs

- The complexity of interactions determines the adversarial robustness of DNNs.

  High-order interactions → adversarial & overfitting.

  Compared to a standard DNN, a Bayesian neural network (BNN) is more likely to avoid encoding complex interactions.

- The complexity of interactions determines the generalization power of DNNs.

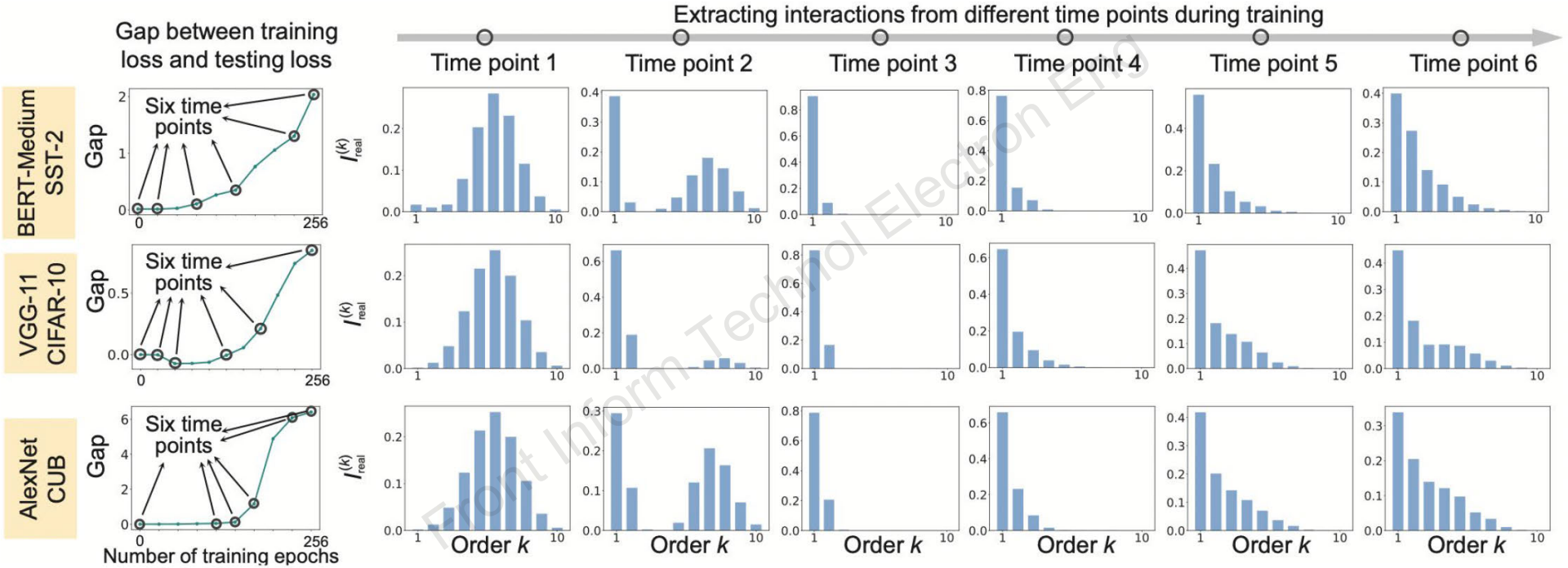
  Low-order interactions → generalize better.

  BNN enhances generalization by simplifying interactions.

- The complexity of interactions determines the representation bottleneck of DNNs.

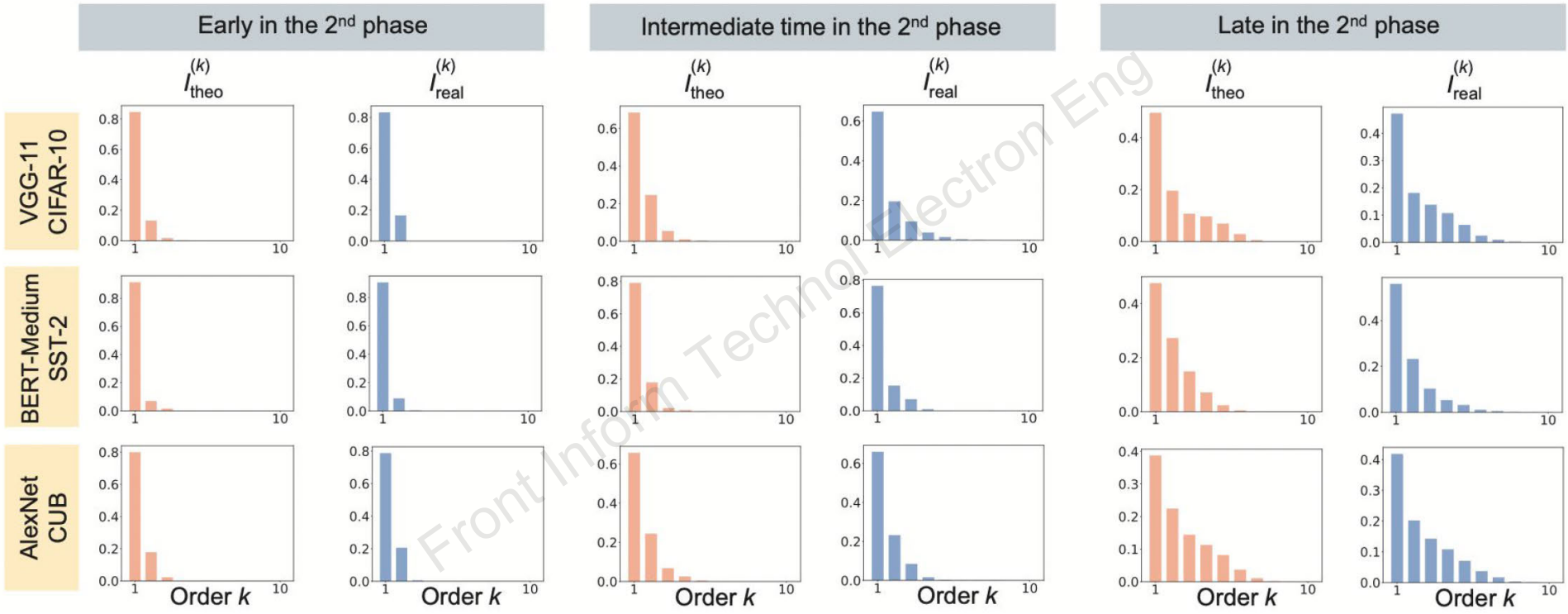
  Compared to the interactions of medium orders, interactions of extremely high and extremely low orders are more likely to be encoded by the DNN.

Two-phase learning dynamics



The two-phase phenomenon of the change of interaction's complexity (Ren QH et al., 2024). Before training: time point 1; first phase: time points 2–3; second phase: time points 3–6

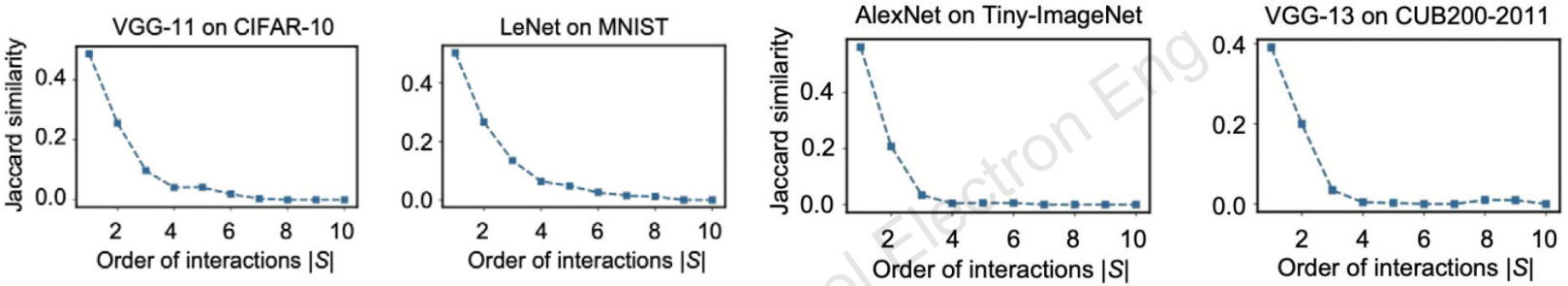
Two-phase learning dynamics



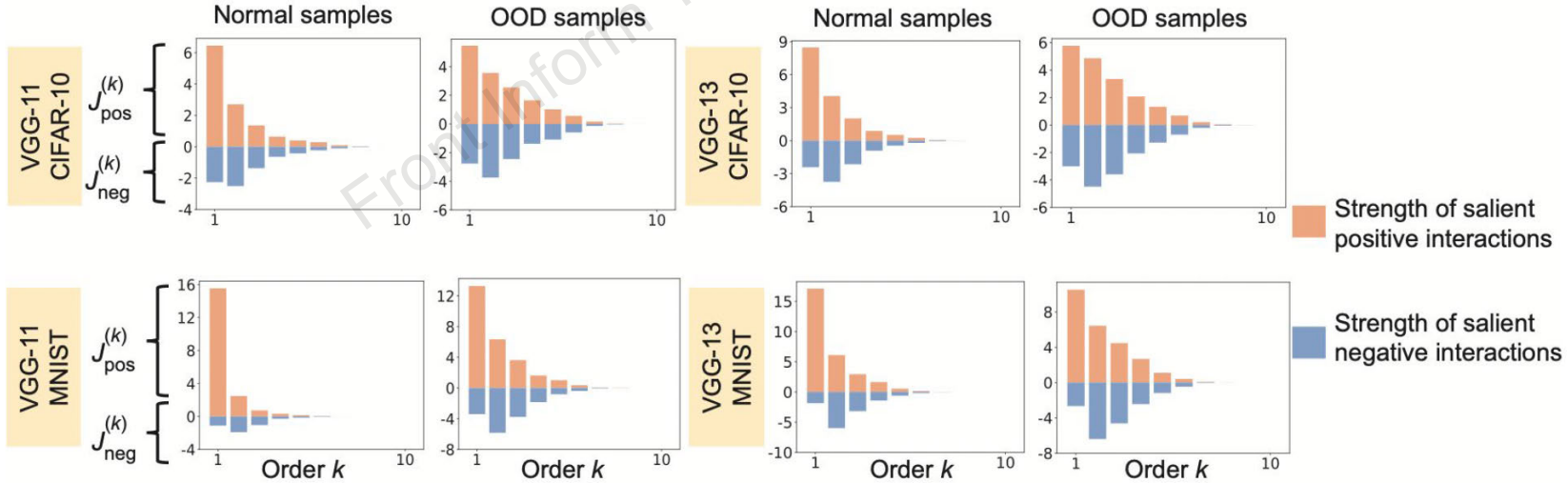
Comparison between the theoretically predicted distribution of interactions across different orders and the real distribution in the second phase (Ren QH et al., 2024)

Using the interactions encoded by the DNN to explain the dynamics of its generalization power

- Low-order interactions generalize better than high-order interactions



The Jaccard similarity between interactions extracted from training samples and interactions extracted from testing samples (Zhang JP et al., 2024)



The distribution of interactions extracted from the original samples and incorrectly labeled samples (Zhang JP et al., 2024). OOD: out-of-distribution

Using the interactions encoded by the DNN to explain the dynamics of its generalization power

The two-phase dynamics of interaction complexity reflects the two-phase dynamics of a DNN's generalization power.

- In the first phase:
  DNN gradually eliminates non-generalizable high-order interactions.
- By the end of this phase:
  DNN primarily encodes generalizable low-order interactions.
- In the second phase:
  DNN begins to learn interactions of gradually increasing orders, which are typically non-generalizable, indicating that the DNN becomes increasingly over-fitted.

Conclusions

We have explored the potential of the interaction theory serving as a first-principles explanation for DNNs. Unlike empirical methods, the interaction theory offers a new axiomatic system that explains the decision-making logic of DNNs through symbolic interaction concepts. These symbolic interaction concepts simultaneously clarify the internal mathematical mechanisms underlying various deep learning phenomena, including generalization power, adversarial robustness, representation bottleneck, and learning dynamics. Additionally, the interaction theory unifies diverse empirical methods, revealing shared mechanisms across 14 attribution methods and 12 transferability-boosting methods.

Huilin Zhou received her BS degree in mathematics from University of Electronic Science and Technology of China in 2019 and her PhD degree in Department of Computer Science of Technology, Shanghai Jiao Tong University, China in 2025. Her research interests include computer vision, machine learning, and explainable AI.

Qihan Ren is a PhD student at Shanghai Jiao Tong University, China. He received his BS degree in computer science from Shanghai Jiao Tong University in 2022. His research interests include computer vision, machine learning, and explainable AI.

Junpeng Zhang is a PhD student at Shanghai Jiao Tong University. He received his BS degree in software engineering from Sun Yat-sen University, China in 2024. His research interests include computer vision, machine learning, and explainable AI.

Quanshi Zhang is an associate professor at Shanghai Jiao Tong University. He received his PhD degree from the University of Tokyo in 2014. From 2014 to 2018, he was a post-doctoral researcher at the University of California, Los Angeles, USA. He won the ACM China Rising Star Award at ACM TURC 2021. He was the speaker of the tutorials on XAI at IJCAI 2020 and IJCAI 2021. He was a co-chair of the workshops towards XAI in ICML 2021, AAAI 2019, and CVPR 2019. His research interests are mainly machine learning and computer vision. In particular, he has made influential research in explainable AI.