

Sixun DONG 董思勋

EMAIL: dongsx@shanghaitech.edu.cn Mobile: +86-136-4924-4805
Homepage: <https://ironieser.github.io/> Github: <https://github.com/Ironieser>

ABOUT ME

I am a third-year graduate student at [ShanghaiTech University](#), supervised by Prof. [Shenghua Gao](#). Before that, I received my Bachelor's degree in 2020 from [Dalian University of Technology](#). My research interests lie in video understanding and weakly supervised learning, including human activity recognition and video representation learning. I am also focusing on multi-modal learning.

EDUCATION

- JULY 2024 Master of Computer Science at [ShanghaiTech University](#), CHINA
(expected) **Major:** Computer Vision & Deep Learning
- JULY 2020 Bachelor Degree in Process Equipment and Control Engineering (Major)
[Dalian University of Technology](#), CHINA
- JULY 2020 Bachelor Degree in Computer Science (Dual Degree)
[Dalian University of Technology](#), CHINA

ACADEMIC EXPERIENCE

† = Co-first author

- PRESENT **Improve video understanding ability of large visual language modal.**
[Sixun Dong](#), et al., Advisor: Prof. Shenghua Gao ShanghaiTech University
- OCT. 2023 Working on video understanding by leveraging pre-trained large language models. To address the limitations of large visual language models in understanding temporal information in videos, we collected instruction data and utilized supervised fine-tuning techniques on large language models to enhance their video comprehension capabilities. This project is currently in progress.
- MAY. 2023 **Weakly Supervised Video Representation Learning with Unaligned Text for Sequential Videos.**
[\[CVPR 2023\]](#) [\[Paper\]](#) [\[Code\]](#) ShanghaiTech University
[Sixun Dong](#)[†], Huazhang Hu[†], Dongze Lian, Weixin Luo, Yicheng Qian, Shenghua Gao.
- APR. 2022 Worked on weakly supervised sequential video understanding where the accurate time-stamp level text-video alignment is not provided. By borrowing ideas from CLIP, we aggregated frame-level features for video representation and encoded the texts corresponding to each action and the whole video, respectively.
 - Proposed a novel weakly supervised video representation learning pipeline with unaligned text for sequential videos.
 - Designed a multiple granularity contrastive learning loss that uses the fact that video actions happen sequentially in the temporal domain to generate pseudo frame-sentence correspondence.
 - Extensive experiments on video sequence verification and text-to-video matching showed the effectiveness of our proposed approach.
- APR. 2022 **TransRAC: Encoding Multi-scale Temporal Correlation with Transformers for Repetitive Action Counting.**
[\[CVPR 2022 Oral\]](#) [\[Paper\]](#) [\[Code\]](#) ShanghaiTech University
Huazhang Hu[†], [Sixun Dong](#)[†], Yiqun Zhao, Dongze Lian, Zhengxin Li*, Shenghua Gao*.
- SEP. 2021 Worked on repetitive action counting(RAC). Specifically, the previous works focus on performing RAC in short videos, which is tough for dealing with longer videos in more realistic scenarios, such as interruption during the actions or inconsistent action cycles.
 - Collected a new repetitive action counting dataset with fine-grained annotations.
 - Encoded multi-scale temporal correlation with transformers that can consider both performance and efficiency.
 - Designed a density map regression-based method to predict the action period.
 - Our approach yielded better performance with sufficient interpretability and achieved SoTA results.

