Peijun Bao (Patrick)

(+65) 91486347 · peijun001@e.ntu.edu.sg · homepage

Research Interest

My research interests are centered around computer vision and machine learning, with a particular focus on the multimodal understanding of video and language. I have published six first-author papers on this topic, including five in top-tier conferences (one at ECCV and four at AAAI) and one in a flagship conference (ICMR).

Education

Nanyang Technological University Sep 2021 – Present

- PhD student in computer science
- Supervisor: Alex Kot (SAEng / IEEE Fellow) Er Meng Hwa (SAEng / IEEE Fellow)

Peking University Sep 2018 – June 2021

- Master student in computer science

Northwestern Polytechnical University Sep 2014 – June 2018

- Bachelor student at Honors College, majoring in computer science
- GPA: 90.5/100, Rank: 1/27

Publication

- Zero-Shot Spatio-Temporal Video Grounding with Expectation-Maximization Multimodal Modulation,
  Peijun Bao, Zihao Shao, Wenhan Yang, Boon Poh Ng, Alex Kot
  ECCV 2024 (oral, top 2%) [pdf], [code]
- Omnipotent Distillation with LLMs for Weakly-Supervised Natural Language Video Localization: When Divergence Meets Consistency,
  Peijun Bao, Zihao Shao, Wenhan Yang, Boon Poh Ng, Meng Hwa Er, Alex Kot
  AAAI 2024 [pdf]
- Local-Global Multi-Modal Distillation for Weakly-Supervised Temporal Video Grounding,
  Peijun Bao, Yong Xia, Wenhan Yang, Boon Poh Ng, Meng Hwa Er, Alex Kot
  AAAI 2024 [pdf]
- DAG: A Large-Scale Domain Adaptation Benchmark for Video Grounding,
  Peijun Bao, Chenqi Kong, Zihao Shao, Wenhan Yang, Boon Poh Ng, Alex Kot
  Under review
- Cross-Modal Label Contrastive Learning for Unsupervised Audio-Visual Event Localization,
  Peijun Bao, Wenhan Yang, Boon Poh Ng, Meng Hwa Er, Alex Kot
  AAAI 2023 (oral) [pdf]
- Learning Sample Importance for Cross-Scenario Video Temporal Grounding,
  Peijun Bao, and Yadong Mu
  ICMR 2022 (oral) [pdf]
- Dense Events Grounding in Video,
  Peijun Bao, Qian Zheng and Yadong Mu
  AAAI 2021 (oral) [pdf], [code]

Skills

- Proficient in using pytorch and python.
- Familiar with tensorflow, c, c++ and matlab.
- Reviewer for CVPR, AAAI, ACM Multimedia, IJCAI, WACV, TMM

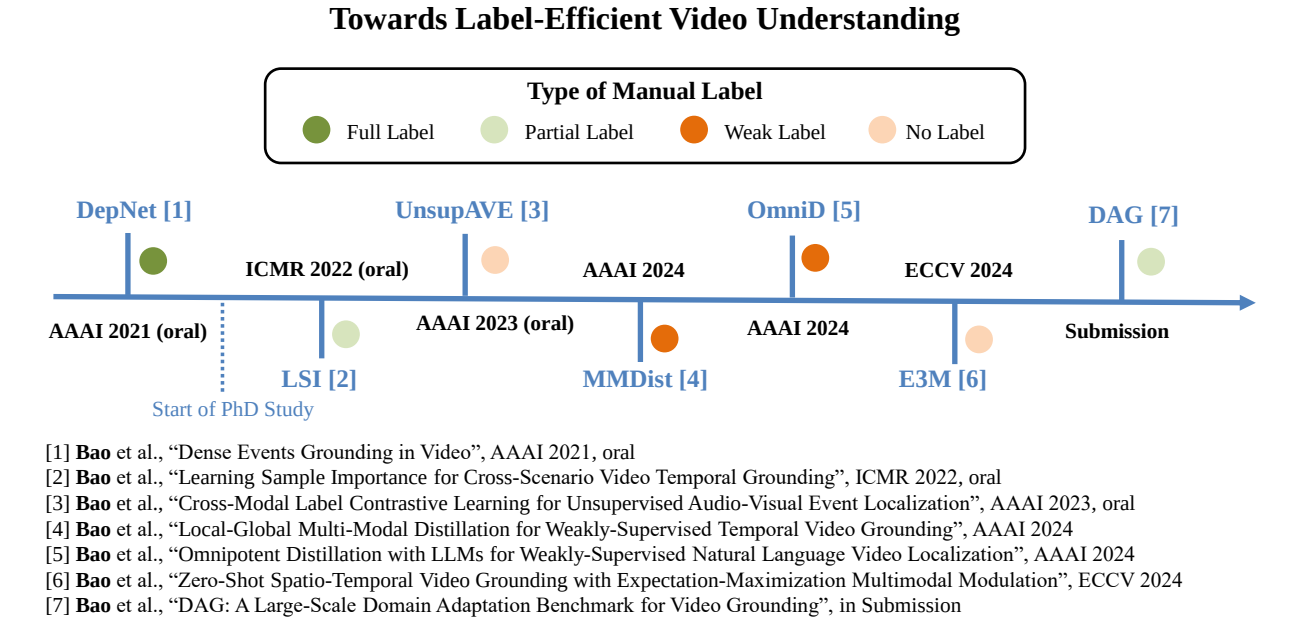
AWARDS

- Nanyang Research Scholarship, Nanyang Technological University
- Merit Student, Northwestern Polytechnical University
- Honors Scholarship, Northwestern Polytechnical University

ACADEMIC REFERENCES

- Alex Kot (PhD supervisor), eackot@ntu.edu.sg
  Professor of Nanyang Technological University, Fellow of SAEng and IEEE
- Er Meng Hwa (PhD co-supervisor), emher@ntu.edu.sg
  Professor of Nanyang Technological University, Fellow of SAEng and IEEE

RESEARCH ROUTE



[1] **Bao** et al., "Dense Events Grounding in Video", AAAI 2021, oral
[2] **Bao** et al., "Learning Sample Importance for Cross-Scenario Video Temporal Grounding", ICMR 2022, oral
[3] **Bao** et al., "Cross-Modal Label Contrastive Learning for Unsupervised Audio-Visual Event Localization", AAAI 2023, oral
[4] **Bao** et al., "Local-Global Multi-Modal Distillation for Weakly-Supervised Temporal Video Grounding", AAAI 2024
[5] **Bao** et al., "Omnipotent Distillation with LLMs for Weakly-Supervised Natural Language Video Localization", AAAI 2024
[6] **Bao** et al., "Zero-Shot Spatio-Temporal Video Grounding with Expectation-Maximization Multimodal Modulation", ECCV 2024
[7] **Bao** et al., "DAG: A Large-Scale Domain Adaptation Benchmark for Video Grounding", in Submission

Figure 1: Overview of my research journey. All listed papers are published with me as the first author.

PROJECT LIST

Omnipotent Distillation with LLMs for Video Grounding

- Current video grounding models encounters two significant challenges: 1) limited input distribution, namely that the limited writing styles of the language query, annotated by human annotators, hinder the model's generalization to real-world scenarios with diverse vocabularies and sentence structures; 2) the incomplete ground truth, whose supervision guidance is insufficient
- To overcome these challenges, we propose an omnipotent distillation algorithm with large language models (LLM). The distribution of the input sample is enriched to obtain diverse multi-view versions while a consistency then comes to regularize the consistency of their results for distillation.
- Our experiments demonstrate substantial performance improvements adaptively to ground diverse kinds of language queries.
- This work was accepted by AAAI 2024.

Zero-Shot Spatio-Temporal Video Grounding with Expectation-Maximization Multimodal Modulation

- To eliminate the annotation costs for spatio-temporal video grounding, we make a first exploration to tackle this task in a zero-shot manner.

- Our method dispenses with the need for any training videos or annotations; instead, it localizes the target object by leveraging pre-trained vision-language models and optimizing within the video and text query during the test time.
- To enable spatio-temporal comprehension, we introduce a multimodal modulation that integrates the spatio-temporal context into both visual and textual representation.
- Experiments validate that our zero-shot approach achieves superior performance in comparison to several state-of-the-art methods with stronger supervision.
- This work was accepted by ECCV 2024.

Dense Events Grounding in Video

- In this work, we explore a novel setting of temporal grounding dubbed as video paragraph grounding. Dense events grounding aims to jointly localize temporal moments of multiple events described in the paragraph.
- Our main motivating fact is that multiple events to be grounded in a video are often semantically related and temporally coordinated according to their order appearing in the paragraph. This fact sheds light on devising a more accurate visual grounding model.
- Based on above motivation, we propose Dense Events Propagation Network (DepNet) for dense events grounding. With a novel aggregation-and-propagation mechanism, DepNet can effectively exploit both the temporal order and semantic relations among dense events.
- This work was accepted by AAAI 2021 (oral). The new task proposed in this paper *i.e.,* video paragraph grounding, is followed by a list of works such as [CVPR24], [CVPR23], [CVPR22], [AAAI24], [ACM MM24], [CVIU24], [EMNLP22] and etc.