Luowei Zhou

2720 152nd Ave NE, Redmond, WA 98052, USA
luoweizhou.github.io | luozhou@microsoft.com | (734)757-0923

RESEARCH INTERESTS

Computer vision and its relations to natural language and deep learning, with a focus on problems in video understanding such as video captioning, grounding, retrieval, question answering, activity recognition, unsupervised representation learning, and learning from multimodal data.

WORK EXPERIENCE

Microsoft   Bellevue, WA
Researcher in Multimodal AI Team   *May 2020 – Present*

University of Michigan, EECS   Ann Arbor, MI
Graduate Student Research Assistant (GSRA) with Dr. Jason J. Corso   *May 2016 – April 2020*
Graduate Student Instructor (GSI) with Dr. Justin Johnson's deep vision class   *Sept. 2019 – Dec. 2019*

Microsoft Research   Redmond, WA
Research Intern with Dr. Hamid Palangi and Dr. Jianfeng Gao   *May 2019 – Aug. 2019*

Facebook AI Research   Menlo Park, CA
Research Intern with Dr. Marcus Rohrbach   *May 2018 – Aug. 2018*

Salesforce Research   Palo Alto, CA
Deep Learning Research Intern with Dr. Caiming Xiong   *May 2017 – Aug. 2017*

EDUCATION

University of Michigan   Ann Arbor, MI
Ph.D. Degree in Robotics (Computer Vision)   *Sept. 2015 – April 2020*
Master's Degree in Robotics (Computer Vision)   *Sept. 2015 – April 2017*
- *Courses:* Advanced Computer Vision, Natural Language Processing, Machine Learning, Optimization
- *Academics:* Curriculum GPA: **4.00/4.00**

Nanjing University   Nanjing, Jiangsu, China
Bachelor's Degree in Automation   *Sept. 2011 – June 2015*
- *Courses:* Computer Vision, Artificial Intelligence, Advanced Programming Language, Data Structure
- *Academics:* Overall GPA: **91.8/100**, Major GPA: **93.0/100**

DOCTORAL DISSERTATION

L. Zhou, *"Language-Driven Video Understanding"*, University of Michigan Deep Blue 2020.

SELECTED PUBLICATIONS AND TALKS (see all at [Google Scholar](#))

L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, *"Unified Vision-Language Pre-Training for Image Captioning and VQA"*, AAAI 2020. (**spotlight)**

Media coverages: [MSR](#), [VentureBeat](#), and [Medium](#). [Code](#). *AR: 20%; h5: 95*

L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, *"Grounded Video Description"*, CVPR 2019. (**oral**) [Code](#). [Dataset](#). *AR: 5.6%; h5: 188*

L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, *"End-to-End Dense Video Captioning with Masked Transformer"*, CVPR 2018. (**spotlight**) [Code](#). *AR: 9%; h5: 158*

L. Zhou, C. Xu, and J. J. Corso, "*Towards Automatic Learning of Procedures from Web Instructional Videos"*, AAAI 2018. (**oral**) [Code](#). [Dataset](#). *AR: 11%; h5: 56*

J. Lei, L. Li, **L. Zhou**, Z. Gan, T. Berg, M. Bansal, and J. Liu, *"Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling"*, CVPR 2021, in submission (initial review: **SA/SA/WA**).

M. Zhou, **L. Zhou**, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, *"UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pretraining"*, CVPR 2021, in submission (initial review: **WA/WA/B**).

H. Huang, **L. Zhou**, W. Zhang, J. J. Corso, and C. Xu, *"Dynamic Graph Modules for Modeling Object-Object Interactions in Activity Recognition"*, BMVC 2019. *AR: 30%; h5: 42*

L. Zhou, N. Louis, and J. J. Corso, *"Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction"*, BMVC 2018. [Code](#). [Dataset](#). *AR: 30%; h5: 42*

L. Zhou et al, *"Multi-agent Reinforcement Learning with Sparse Interactions by Negotiation and Knowledge Transfer"*, IEEE Transactions on Cybernetics 2017, 47 (5): 1238 - 1250. *SCI IF: 7.38; h5: 73*

PATENTS

Y. Zhou, **L. Zhou**, C. Xiong, and R. Socher, *"Dense Video Captioning",* US10542270B2.

HONORS AND AWARDS

Outstanding Winner Awards (0.2%), Mathematical Contest in Modeling (MCM)	2014
Sienhua New and Tsu Way Shen Memorial Award (Top 1), of University of Michigan	2015
Best Undergrad Thesis (Top 1), of Nanjing University and Jiangsu Province	2015
National Scholarship (1%), of Nanjing University	2012
Red Sun Scholarship, of Nanjing University	2014

OTHER INVITED TALKS

Microsoft   Bellevue, WA
Hosted by Dr. Jingjing Liu   *Oct. 2019*

Facebook AI   Menlo Park, CA
Hosted by Dr. Yatharth Saraf   *Nov. 2019*

NVIDIA Research   Santa Clara, CA
Hosted by Dr. Jan Kautz   *Nov. 2019*

Salesforce Research   Palo Alto, CA
Hosted by Dr. Caiming Xiong   *Nov. 2019*

Amazon AI   Seattle, WA
Hosted by Dr. Joseph Tighe   *Nov. 2019*

Waymo   Mountain View, CA
Hosted by Dr. Dragomir Anguelov   *Nov. 2019*

Tencent AI Lab   Bellevue, WA
Hosted by Dr. Tong Zhang   *Oct. 2019*

NVIDIA AI Lab   Toronto, Ontario, Canada
Hosted by Dr. Sanja Fidler   *Dec. 2018*

SAMSUNG AI Centre   Toronto, Ontario, Canada
Hosted by Dr. Afsaneh Fazly and Dr. Allan Jepson   *Dec. 2018/2020*

PROFESSIONAL ACTIVITIES

Organizer, CVPR 2020 / CVPR 2021 Challenge on ActivityNet-Entities Object Localization ([AEOL](#)), a guest task in the annual ActivityNet Workshop

Co-organizer, CVPR 2020 / CVPR 2021 [Tutorial](#) on Recent Advances in Vision and Language

Co-organizer, CVPR 2018 Workshop on Fine-grained Instructional Video Understanding ([FIVER](#)), with Jason Corso, Josef Sivic, and Ivan Laptev

Co-organizer, UMich Computer Vision Reading Group

Program Committee Member / Reviewer:

- Computer Vision: CVPR 2021/2020/2019, ECCV 2020, ICCV 2021/2019, TPAMI 2020/2019/2018, IJCV 2019, ACM MM 2020/2019, and CVIU 2017
- Language Understanding: ACL 2020, EMNLP 2020, and NAACL (Workshops) 2018
- Machine Learning: NeurIPS 2020, ICML 2020, AAAI 2020, IJCAI 2020, and NIPS 2016
- Robotics: ICRA 2018 and ITS 2019/2018/2017

Volunteer, RSS 2016

RESEARCH EXPERIENCE (open-source projects on [Github](#))

Large-Scale Unified Vision-Language Pre-training   Microsoft Research

Supervisors: Dr. Jianfeng Gao, Dr. Lei Zhang, and Dr. Hamid Palangi   *May 2019 – Nov. 2019*

- Introduced a generic and unified framework for Vision-Language Pre-training (VLP). VLP is pre-trained on millions of image-text pairs automatically mined from the web and fine-tuned for disparate downstream tasks including image captioning and VQA.
- Proposed to use two unsupervised learning objectives for VLP: bidirectional and sequence-to-sequence (seq2seq) masked vision-language prediction.
- Thanks to our vision-language pre-training, both training speed and overall accuracy have been significantly improved on the downstream tasks compared to other model initialization methods.
- Set new SotA on COCO Captions (CIDEr 129), VQA 2.0 (overall 71) and Flickr30k Captions (CIDEr 67 vs previous SotA 62), all from a single model architecture.
- Current focuses: VLP on videos by leveraging a large amount of instructional video data and the associated ASR scripts. Multi-task learning of captioning, QA, and event proposal.

Grounded Video Description   Facebook AI Research

Supervisors: Dr. Marcus Rohrbach, Dr. Yannis Kalantidis, and Dr. Xinlei Chen   *May 2018 – Dec. 2018*

- Introduced a large-scale video description and grounding dataset, called [ActivityNet-Entities](#), where we annotated noun phrases (& objects) from sentence descriptions in videos as spatial bounding boxes. ActivityNet-Entities contains over 158k labeled boxes for 52k video clips.
- Proposed a unified framework for video and image description, where a supervised grounding module dynamically detects objects in the scene and provides visual clues to the captioning module.
- Set new SotA performance on video description and image description and demonstrated that our generated sentences are more explainable through grounding.

Fine-grained Instructional Video Understanding   University of Michigan

Supervisor: Prof. Jason Corso   *Sept. 2016 – April 2020*

- Introduced [YouCook2](#) dataset, which contains temporally localized recipe sentence annotations and bounding boxes for 2000 YouTube cooking videos.
- Tackled a series of problems related to instructional video understanding: i) event proposal (AAAI 2018), ii) dense video captioning (CVPR 2018), iii) weakly supervised object grounding from language description (BMVC 2018).
- *Event proposal*: Proposed an event proposal and sequential modeling network that can temporally localize procedure steps in web instructional videos and capture the temporal structure of the video.
- *Dense video captioning*: Caption generation for event proposals. See Page 4 for more details.
- *Weakly supervised object grounding*: Given a video and the corresponding description, localize the objects mentioned from the description in the video as bounding boxes. No box is given for training.
- Current focuses: Graph-based procedure structure learning.

Dense-Captioning Events in Video and Temporal Action Proposal   Salesforce Research
Supervisors: Dr. Caiming Xiong and Dr. Richard Socher   *May 2017 – Aug. 2017*

- Introduced a self-attention-based video captioning model and improved our previously proposed action/event proposal network with carefully-designed Temporal Convolutional Networks.
- Proposed to bridge event proposal and captioning by a differentiable visual mask and achieved state-of-the-art results on dense video captioning.

Text-conditional Visual Captioning with Guiding LSTM   University of Michigan
Supervisor: Prof. Jason Corso   *Mar. 2016 – Nov. 2016*

- Proposed an encoder-decoder image captioner though explicit text-conditional image guidance.
- Extended the work to video captioning by leveraging audio features for the extra guidance.

End-to-End Grasping with Deep Reinforcement Learning   University of Michigan
Supervisor: Prof. Satinder Singh   *Sept. 2015 – Apr. 2016*

- Applied state-of-the-art Deep RL algorithm named Deep Q-network (DQN) to robot grasping tasks.
- Built an API between physics engine MuJoCo and the DQN module.

Research on Multi-Agent Reinforcement Learning with Sparse Interactions   Nanjing University
Supervisors: Prof. Chunlin Chen, Dr. Pei Yang, and Prof. Yang Gao   *Dec. 2014 – Jul. 2015*

- Introduced the concept of equilibrium into traditional sparse-interaction-based MARL algorithms and proposed a knowledge transfer approach to initialize the joint-state Q table.
- Applied the proposed algorithm in a real-world setting, i.e., our intelligent warehouse simulator.

Multi-Robot Task Allocation and Path Planning in Dynamic Environments   Nanjing University
Supervisor: Dr. Pei Yang   *Nov. 2013 – Jul. 2014*

- Proposed a Balanced Heuristic Mechanism to balance task allocation in multi-robot systems.
- Built an intelligent warehouse simulator from scratch using C/OpenGL for the experiments.

PROFICIENCY AND SKILLS

Technical Skills: PyTorch/Torch, Python, C/C++, Linux, Git, LaTeX, Matlab, Caffe, HTML, CSS, JS etc.
Languages: English (proficient) and Mandarin (native)

REFERENCES

Prof. Jason Corso, Professor, University of Michigan, jjcorso@umich.edu

Dr. Marcus Rohrbach, Research Scientist, Facebook AI, mrf@fb.com

Dr. Yannis Kalantidis, Research Scientist, Naver Labs Europe, yannis.kalantidis@naverlabs.com

Prof. Chenliang Xu, Assistant Professor, University of Rochester, chenliang.xu@rochester.edu

Dr. Hamid Palangi, Senior Researcher, Microsoft Research, hpalangi@microsoft.com