XINYU LI

7324066689 ⋄ lixinyu.arthur@outlook.com ⋄ https://www.arthurlxy.com

EDUCATION

Rutgers University *Sep 2013 - Feb 2018*
Ph.D. in Computer Engineering
University of Electronic Science and Technology of China *Sep 2009 - June 2013*
B.S. in Communication Engineering

EXPERIENCE

AGI, Amazon 2022 - present
SeniorApplied Scientist/Principal Applied Scientist *Bellevue , WA*

· I am leading the visual understanding efforts of the Amazon Nova models, with a focus on video understanding and cross-modal understanding, where our model achieves state-of-the-art comparable performance in tasks such as video summarization, VQA, temporal retrieval, and video grounding.
· I am leading the Amazon AGI encoder and multi-modal embedding project, where we develop quad-modal embedding models for fast cross-modal retrieval.
· Leading research in video understanding and multimodal understanding, with publications in top-tier conferences including CVPR, ICCV, ECCV, NeurIPS, and ICLR.

AML, Tiktok/ByteDance 2021 - 2022
Senior Research Scientist *Bellevue, WA*

· I served as the tech lead for ByteDance's MM foundation model, focusing on visual-acoustic understanding. Our model serves as the backbone for various key functionalities, including compliance, search, and recommendation.
· Led research on video understanding and multimodal understanding, with publications in top-tier conferences such as CVPR, ICASSP, and ICLR.

AWS AI, Amazon 2018 - 2021
Applied Scientist/Senior Applied Scientist *Seattle, WA*

· I served as a research scientist and later as the tech lead for Rekognition Video, where we developed models for key AWS services, including compliance, tracking, shot/scene segmentation, and ad insertion.
· Led research on video understanding, with publications in top-tier conferences such as CVPR, ICCV, ECCV, and NeurIPS.

Multimedia Lab, Rutgers Sep 2013 - Feb 2018
Graduate Research Assistant *New-brunswick, NJ*

· Multimodality-based multilabel action recognition and action detection. The prototype system is installed in an actual trauma room at the Children's National Medical Center. Publications in CVPR, ACM MM, Sensys and Ubicomp, etc..
· Visual-acoustic human emotion recognition and sentiment analysis; publications in ACL, ACM MM, COLING.

SELECTED PUBLICATION

Selected publications in past 5 years, full list of publication can be find at Google Scholar
* denotes equally contributed.

1. Amazon Artificial General. "The Amazon Nova family of models: Technical report and model card."

2. Amazon Artificial General. "Amazon Nova Premier: Technical report and model card."

3. Xianhang Li, Peng Wang, Xinyu Li, Heng Wang, Hongru Zhu, and Cihang Xie. "Efficient Video-MAE via Temporal Progressive Training." In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 2659-2668. 2025.

4. Yicheng Wang, Zhikang Zhang, Jue Wang, David Fan, Zhenlin Xu, Linda Liu, Xiang Hao, Vimal Bhat, Xinyu Li. "GEXIA: Granularity Expansion and Iterative Approximation for Scalable Multi-grained Video-language Learning." WACV 2025.

5. Seon-Ho Lee, Jue Wang, David Fan, Zhikang Zhang, Linda Liu, Xiang Hao, Vimal Bhat, Xinyu Li. "Now You See Me: Context-Aware Automatic Audio Description." WACV 2025.

6. Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, Xinyu Li. "Video Token Merging for Long Video Understanding." NeurIPS 2024.

7. David Fan, Jue Wang, Shuai Liao, Zhikang Zhang, Vimal Bhat, Xinyu Li. "Text-Guided Video Masked Autoencoder." ECCV 2024.

8. David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, Xinyu Li. "Motion-Guided Masking for Spatiotemporal Representation Learning." ICCV 2023.

9. Najmeh Sadoughi, Xinyu Li, Avijit Vajpayee, David Fan, Bing Shuai, Hector Santos-Villalobos, Vimal Bhat, Rohith MV. "MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation." ICCV 2023.

10. Chen, Yuxiao, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N. Metaxas, and Hongxia Yang. "Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15095-15104. 2023.

11. Xiaoyu Liu, Hanlin Lu, Jianbo Yuan, and Xinyu Li. "CAT: Causal Audio Transformer for Audio Classification." In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.

12. Li, Xinyu, Yanyi Zhang, Jianbo Yuan, Hanlin Lu, and Yibo Zhu. "Discrete Cosin TransFormer: Image Modeling From Frequency Domain." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5468-5478. 2023.

13. Jiaojiao Zhao*, Yanyi Zhang*, Xinyu Li*, Hao Chen, Shuai Bing, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G.M. Snoek, Joseph Tighe. "TubeR: Tubelet Transformer for Video Action Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2022 Oral.

14. A S M Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, Davide Modolo. "What to look at and where: Semantic and Spatial Refined Transformer for detecting human-object interactions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2022 Oral.

15. Feng Cheng, Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Li, Wei Xia. "Stochastic Backpropagation: A Memory Efficient Strategy for Training Video Models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2022 Oral.

16. Bing Shuai, Xinyu Li, Kaustav Kundu, Joseph Tighe. "Id-Free Person Similarity Learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2022.

17. Xinyu Li*, Yanyi Zhang*, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. "VidTr: Video Transformer Without Convolutions." ICCV 2021.
18. Chunhui Liu*, Xinyu Li*, Hao Chen, and Joseph Tighe "Selective Feature Compression for Efficient Activity Recognition Inference." ICCV 2021.
19. Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu and Stefano Soatto. "Long Short-Term Transformer for Online Action Detection." NeurIPS 2021 (Spotlight).
20. Zhang, Yanyi, Xinyu Li, and Ivan Marsic. "Multi-Label Activity Recognition using Activity-specific Features." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2021.
21. Shuai, Bing, Andrew G. Berneshawi, Xinyu Li, Davide Modolo, and Joseph Tighe. "Multi-object tracking with Siamese track-RCNN." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2021.
22. Zhu, Yi, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. "A Comprehensive Study of Deep Video Action Recognition." arXiv preprint arXiv:2012.06567 (2020).
23. Shuai, Bing, Andrew Berneshawi, Manchen Wang, Chunhui Liu, Davide Modolo, Xinyu Li, and Joseph Tighe. "Application of Multi-Object Tracking with Siamese Track-RCNN to the Human in Events Dataset." In Proceedings of the 28th ACM International Conference on Multimedia, pp. 4625-4629. ACM MM 2020.
24. Li, Xinyu, Bing Shuai, and Joseph Tighe. "Directional temporal modeling for action recognition." In European Conference on Computer Vision, pp. 275-291. Springer, Cham, ECCV 2020.

SELECTED OPEN-SOURCE PACKAGES

GluonCV-Torch Project Link

· GluonCV provides implementations of state-of-the-art (SOTA) deep learning algorithms in computer vision. Available in mxnet and pytorch.

GluonMM Project Link

· GluonCV-Transformer is an open-sourced library based on PyTorch, providing a list of SOTA transformer-based research implementations on various image tasks (image classification, object detection, semantic segmentation), video tasks (video classification, spatio-temporal action detection, long-video reasoning), and multimedia tasks (sound classification, video to text, retrieval, etc.).

TubeR: Tubelet Transformer Project Link

· This repo contains the supported code to reproduce spatio-temporal action detection results of TubeR: Tubelet Transformer for Video Action Detection.

PROFESSIONAL SERVICES

Conference Reviewer: CVPR, ICCV, ECCV, WACV, ICLR, NeurIPS, ICASSP, InterSpeech.
Journal Reviewer: Pattern Recognition Letters, IMWUT, Transaction of Mobile Computing.