CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Recent Advances in Vision-and-Language Pre-training

Zhe Gan, Linjie Li, Chunyuan Li, Jianwei Yang, Pengchuan Zhang,
Lijuan Wang, Zicheng Liu, Jianfeng Gao

Microsoft

Meta

Tutorial Agenda

Morning Session		
9:00 - 9:15	**Opening Remarks**	Lijuan Wang
9:15 - 10:00	**Overview of Image-Text Pre-training**	Jianfeng Wang
10:00 - 10:15	**Coffee Break & QA**	
10:15 - 11:00	**Unified Image-Text Modeling**	Zhengyuan Yang
11:00 - 11:45	**Advanced Topics in Image-Text Pre-training**	Zhe Gan
11:45 – 12:00	**Q & A**	
Afternoon Session		
13:00 - 13: 30	**Overview of Video-Text Pre-training**	Kevin Lin
13:30 - 14:00	**Learning from Multi-channel Videos: Methods and Benchmarks**	Linjie Li
14:00 - 14: 30	**Advanced Topics in Video-Text Pre-training**	Chung-Ching Lin
14:30 - 14:45	**Coffee Break & QA**	
14:45 - 15: 15	**VLP for Image Classification**	Jianwei Yang
15:15 - 15:45	**VLP for Object Detection**	Pengchuan Zhang
15:45 - 16:15	**Benchmarks for Computer Vision in the Wild**	Chunyuan Li
16:15 - 17:00	**VLP for Text-to-Image Synthesis**	Chenfei Wu
17:00 - 17:15	**Q & A**	

Zhe Gan
Microsoft

Pengchuan Zhang
Meta

Zhengyuan Yang
Microsoft

Kevin Lin
Microsoft

Linjie Li
Microsoft

Chunyuan Li
Microsoft

Jianfeng Wang
Microsoft

Jianwei Yang
Microsoft

Chung-Ching Lin
Microsoft

Chenfei Wu
Microsoft

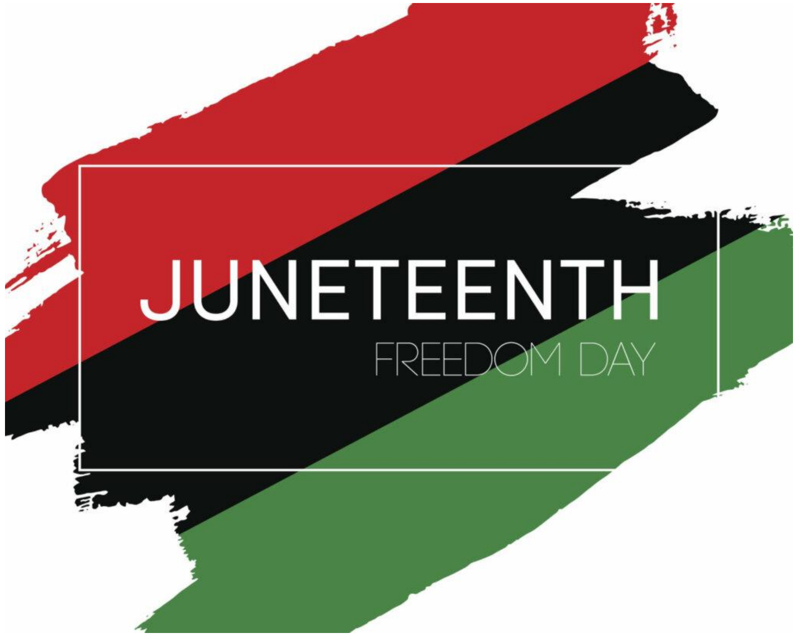
Lijuan Wang
Microsoft

Zicheng Liu
Microsoft

Jianfeng Gao
Microsoft

Tutorial website: https://vlp-tutorial.github.io/

Juneteenth Holiday & Father’s day

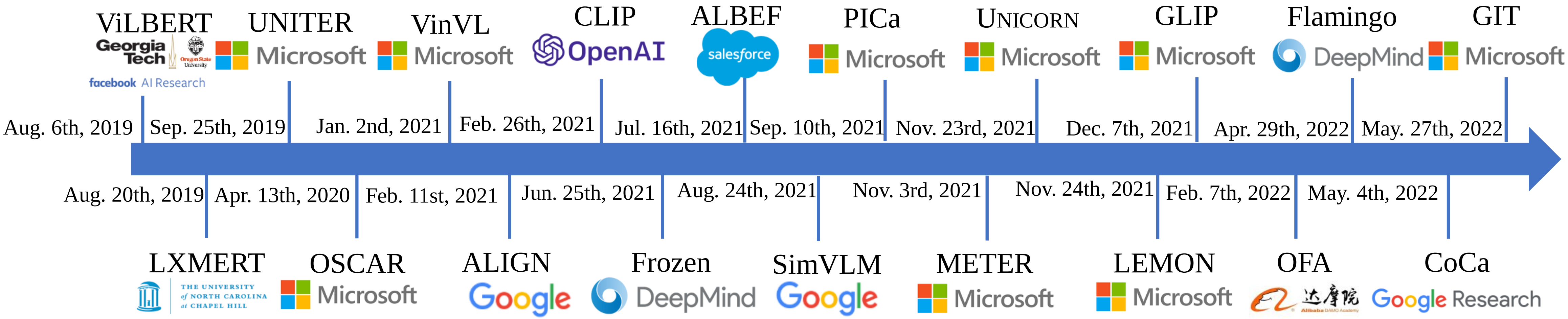


https://cvpr2022.thecvf.com/recognizing-juneteenth

Figure credit: https://wisconsinexaminer.com/brief/biden-signs-law-making-juneteenth-a-new-federal-holiday ; https://wesleychoice.org/12-interesting-facts-about-fathers-day/

Evolution of Image-Text Pre-training

- All these early models depend on pre-trained object detectors to extract visual features offline.
- Newer end-to-end VLP achieves stronger performance with model and data scaling.
- Upscaled VLP models demonstrate new capabilities such as in-context learning and multimodal few shots.



ViLBERT — Georgia Tech, Oregon State University, facebook AI Research — Aug. 6th, 2019

LXMERT — The University of North Carolina at Chapel Hill — Aug. 20th, 2019

UNITER — Microsoft — Sep. 25th, 2019

OSCAR — Microsoft — Apr. 13th, 2020

VinVL — Microsoft — Jan. 2nd, 2021

ALIGN — Google — Feb. 11st, 2021

CLIP — OpenAI — Feb. 26th, 2021

Frozen — DeepMind — Jun. 25th, 2021

ALBEF — salesforce — Jul. 16th, 2021

SimVLM — Google — Aug. 24th, 2021

PICa — Microsoft — Sep. 10th, 2021

METER — Microsoft — Nov. 3rd, 2021

UNICORN — Microsoft — Nov. 23rd, 2021

LEMON — Microsoft — Nov. 24th, 2021

GLIP — Microsoft — Dec. 7th, 2021

OFA — 达摩院 Alibaba DAMO Academy — Feb. 7th, 2022

Flamingo — DeepMind — Apr. 29th, 2022

CoCa — Google Research — May. 4th, 2022

GIT — Microsoft — May. 27th, 2022

Session 1: Overview of Image-Text Pre-training

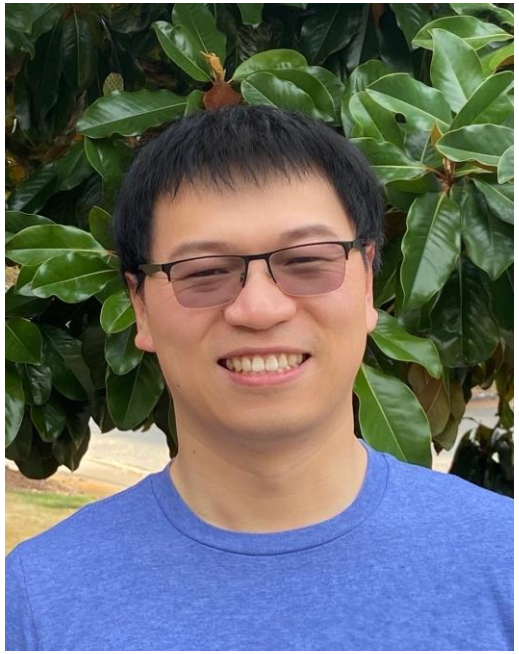
Time:

10:15 – 11:00 AM

Presenter:

Jianfeng Wang (Microsoft)

Dr. Jianfeng Wang is a Principal Researcher at Microsoft Azure AI. His research interest is computer vision and vision-language intelligence. He received his B. Eng. degree and PhD degree from the University of Science and Technology of China in 2010 and 2015, respectively.



Session 2: Unified Image-Text Modeling

Time:

10:15 – 11:00 AM

Presenter:

Zhengyuan Yang (Microsoft)

Dr. Zhengyuan Yang is a Researcher at Microsoft Azure AI. He received the Ph.D. degree in Computer Science at the University of Rochester (UR) in 2021, and the BE degree in electrical engineering from the University of Science and Technology of China (USTC) in 2016. His research interests are vision+language and multi-modal learning.

Session 3: Advanced Topics in Image-Text Pre-training

Time:

11:00 – 11:45 AM

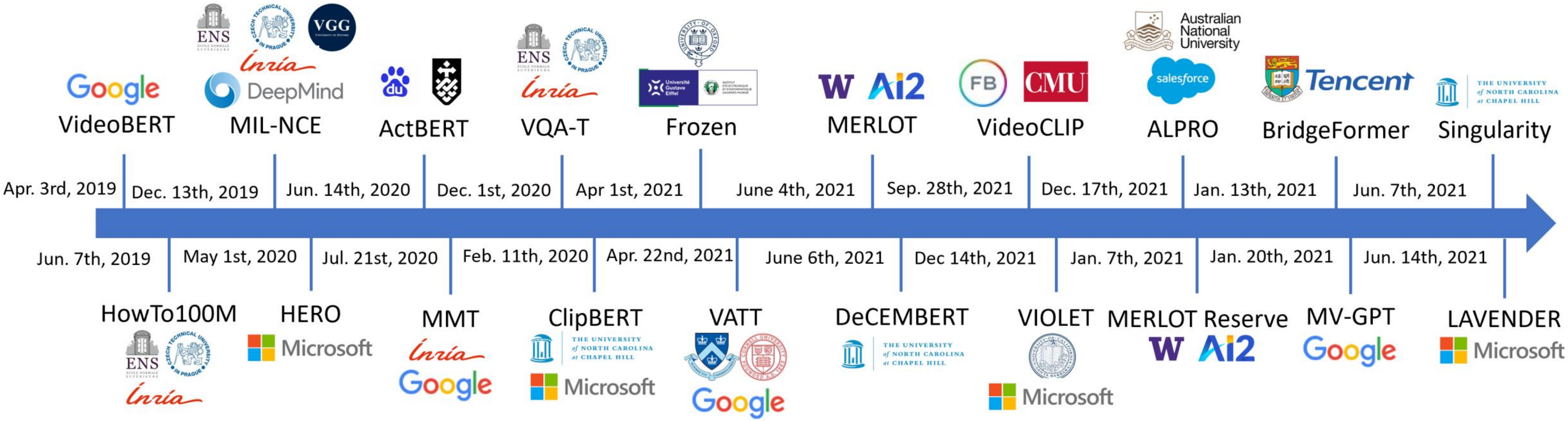
Presenter:

Zhe Gan (Microsoft)

Dr. Zhe Gan is a Principal Researcher at Microsoft Azure AI, primarily working on Vision-and-Language Multimodal Intelligence. He received the PhD degree from Duke University in 2018. Before that, he received the Master's and Bachelor's degree from Peking University in 2013 and 2010, respectively. Together with his co-authors, he has received the Best Student Paper Honorable Mention Awards at CVPR 2021 and WACV 2021, respectively.

Evolution of Video-Text Pre-training

- More advanced models trained end2end on up to 1B video-text pairs with applications to video-text/video-only or even image-text tasks
- Learning from multi-channel videos, leveraging information from vision + language + audio
- Unified architecture across modalities and tasks



VideoBERT — Apr. 3rd, 2019

MIL-NCE — Dec. 13th, 2019

ActBERT — Jun. 14th, 2020

VQA-T — Dec. 1st, 2020

Frozen — Apr 1st, 2021

MERLOT — June 4th, 2021

VideoCLIP — Sep. 28th, 2021

ALPRO — Dec. 17th, 2021

BridgeFormer — Jan. 13th, 2021

Singularity — Jun. 7th, 2021

HowTo100M — Jun. 7th, 2019

HERO — May 1st, 2020

MMT — Jul. 21st, 2020

ClipBERT — Feb. 11th, 2020

VATT — Apr. 22nd, 2021

DeCEMBERT — June 6th, 2021

VIOLET — Dec 14th, 2021

MERLOT Reserve — Jan. 7th, 2021

MV-GPT — Jan. 20th, 2021

LAVENDER — Jun. 14th, 2021

Session 4: Overview of Video-Text Pre-training

Time:

13:00 – 13:30 PM

Presenter:

Kevin Lin (Microsoft)

Dr. Kevin Lin is an Applied Scientist at Microsoft Azure AI, working on computer vision and vision-language multimodal intelligence. Prior to Microsoft, he obtained his PhD from the University of Washington in 2020, and his MS from National Taiwan University in 2014.

Session 5: Learning from Multi-channel Videos: Methods and Benchmarks

Time:

13:30 – 14:00 PM

Presenter:

Linjie Li (Microsoft)

Linjie Li is a Senior Researcher at Microsoft Azure AI. She obtained her Master’s degree in computer science from Purdue University in 2018. Her current research interests include Vision-and-Language pre-training and self-supervised learning. Linjie (co-)organizes multiple tutorials on vision+language at CVPR 2020/2021/2022. Her recent work ClipBERT for video+language pre-training is nominated for the Best Student Paper Award at CVPR 2021.

Session 6: Advanced Topics in Video-Text Pre-training

Time:

13:00 – 13:30 PM

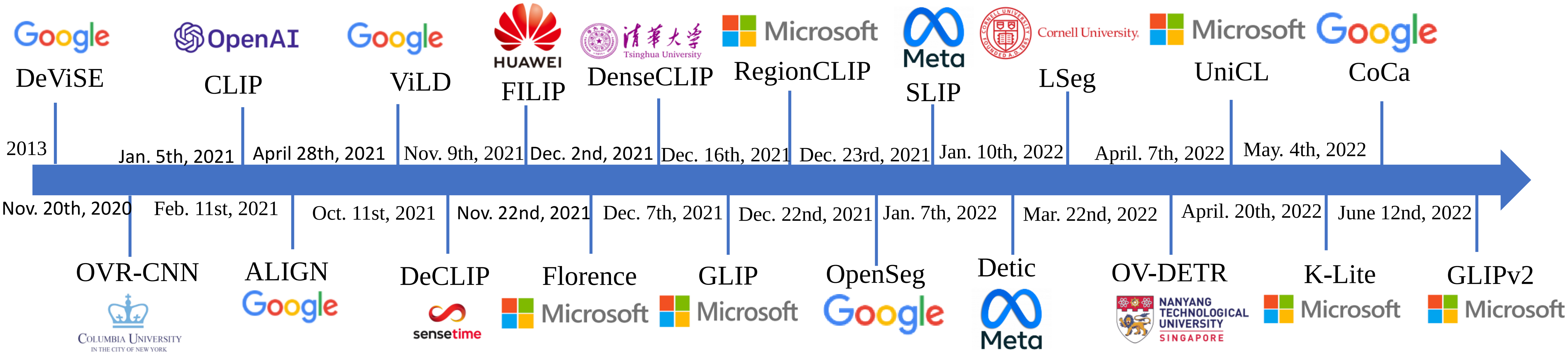
Presenter:

Chung-Ching Lin (Microsoft)

Dr. Chung-Ching Lin is a Principal Researcher at Microsoft Azure AI. Before joining Microsoft, Chung-Ching was a Research Staff Member with IBM T.J. Watson Research, where he began his career after earning his Ph.D. from Georgia Institute of Technology. His research interests include computer vision and machine learning, with specific interests in video understanding, representation learning, and vision and language.

Evolution of VLP for Vision

- Vision has gradually transit from canonical classification problem to a vision-language alignment problem.
- Contrastive vision-language pretraining supports both representation learning and zero-shot/few-shot transferring learning.
- Vision tasks at image-level, region-level and pixel-level can be unified into a single vision-language interface using contrastive learning and/or generative pretraining.



Google DeViSE 2013

OpenAI CLIP Jan. 5th, 2021

Google ViLD April 28th, 2021

HUAWEI FILIP Nov. 9th, 2021

清華大學 Tsinghua University DenseCLIP Dec. 2nd, 2021

Microsoft RegionCLIP Dec. 16th, 2021

Meta SLIP Dec. 23rd, 2021

Cornell University LSeg Jan. 10th, 2022

Microsoft UniCL April. 7th, 2022

Google CoCa May. 4th, 2022

Nov. 20th, 2020 OVR-CNN Columbia University in the City of New York

Feb. 11st, 2021 ALIGN Google

Oct. 11st, 2021 DeCLIP sensetime

Nov. 22nd, 2021 Florence Microsoft

Dec. 7th, 2021 GLIP Microsoft

Dec. 22nd, 2021 OpenSeg Google

Jan. 7th, 2022 Detic Meta

Mar. 22nd, 2022 OV-DETR Nanyang Technological University Singapore

April. 20th, 2022 K-Lite Microsoft

June 12nd, 2022 GLIPv2 Microsoft

Session 7: VLP for Image Classification

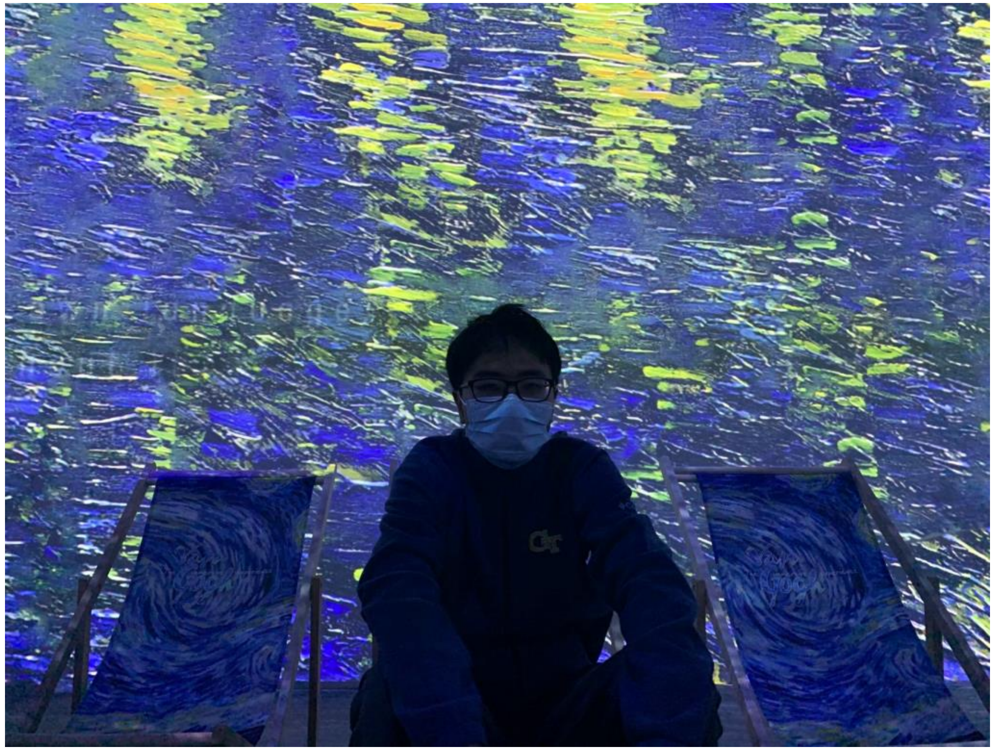
Time:

14:45 – 15:15 PM

Presenter:

Jianwei Yang (Microsoft)

Dr. Jianwei Yang is a Senior Researcher at Microsoft Research. His main research interests span in computer vision, vision-language and embodied AI. He obtained his Ph.D. in CS at Georgia Institute of Technology.



Session 8: VLP for Object Detection

Time:

15:15 – 15:45 PM

Presenter:

Pengchuan Zhang (Meta AI)

Dr. Pengchuan Zhang is currently an AI research scientist at Meta AI and an affiliate assistant professor at University of Washington. Before joining Meta, he was a Principal Researcher at Microsoft Research. His research interests are mainly in the areas of computer vision, multimodal intelligence, and theoretical foundations for deep learning.

Session 9: Benchmarks for Computer Vision in the Wild

Time:

15:45 – 16:15 PM

Presenter:

Chunyuan Li (Microsoft)

Dr. Chunyuan Li is currently a Principal Researcher at Microsoft Research. His recent research focuses on large-scale pre-training in computer vision and natural language processing. He obtained his PhD in machine learning at Duke University.

Session 10: VLP for Text-to-Image Synthesis

Time:

16:15 – 17:00 PM

Presenter:

Chenfei Wu (Microsoft)

Dr. Chenfei Wu is currently a Senior Researcher at Microsoft Research. His research interests include visual-language understanding and generation. Recently, he focuses on language guided high-quality image/video generation.

Tutorial Agenda

Morning Session		
9:00 - 9:15	**Opening Remarks**	Lijuan Wang
9:15 - 10:00	**Overview of Image-Text Pre-training**	Jianfeng Wang
10:00 - 10:15	**Coffee Break & QA**	
10:15 - 11:00	**Unified Image-Text Modeling**	Zhengyuan Yang
11:00 - 11:45	**Advanced Topics in Image-Text Pre-training**	Zhe Gan
11:45 – 12:00	**Q & A**	
Afternoon Session		
13:00 - 13: 30	**Overview of Video-Text Pre-training**	Kevin Lin
13:30 - 14:00	**Learning from Multi-channel Videos: Methods and Benchmarks**	Linjie Li
14:00 - 14: 30	**Advanced Topics in Video-Text Pre-training**	Chung-Ching Lin
14:30 - 14:45	**Coffee Break & QA**	
14:45 - 15: 15	**VLP for Image Classification**	Jianwei Yang
15:15 - 15:45	**VLP for Object Detection**	Pengchuan Zhang
15:45 - 16:15	**Benchmarks for Computer Vision in the Wild**	Chunyuan Li
16:15 - 17:00	**VLP for Text-to-Image Synthesis**	Chenfei Wu
17:00 - 17:15	**Q & A**	

Zhe Gan
Microsoft

Pengchuan Zhang
Meta

Zhengyuan Yang
Microsoft

Kevin Lin
Microsoft

Linjie Li
Microsoft

Chunyuan Li
Microsoft

Jianfeng Wang
Microsoft

Jianwei Yang
Microsoft

Chung-Ching Lin
Microsoft

Chenfei Wu
Microsoft

Lijuan Wang
Microsoft

Zicheng Liu
Microsoft

Jianfeng Gao
Microsoft

Tutorial website: https://vlp-tutorial.github.io/