

# Zhexiao Xiong

+1-314-319-2407 | [x.zhexiao@wustl.edu](mailto:x.zhexiao@wustl.edu) | [in](#) LinkedIn | [G](#) Github | [G](#) Personal-Webpage





## BIOGRAPHY

I am a third-year CS Ph.D. candidate at Washington University in St. Louis(WashU), advised by **Prof. Nathan Jacobs**. My research lies broadly in computer vision and multi-modal learning, especially generative models and AIGC-related topics, including controllable & personalized image/video generation and editing, and the combination of vision language models(VLMs) with image/video generation. I am also interested in generative models for 3D vision such as neural rendering and cross-view & novel view synthesis.

## EDUCATION

- **Washington University in St. Louis** 2022.08 – 2027.05(Expected)  
Ph.D. Candidate in Computer Science  
St. Louis, MO, USA  
Advisor: **Prof. Nathan Jacobs**
- **Tianjin University** 2018.09 – 2022.06  
B.S. in Electrical and Information Engineering  
Tianjin, China

## WORK EXPERIENCE

- **Bosch Research**  2025.06 – 2025.09  
Research Intern  
Sunnyvale, CA, USA  
◦ Researched on video generation world models for autonomous driving.
- **OPPO US Research Center**  2024.05 – 2024.08  
Research Intern  
Palo Alto, CA, USA  
◦ Researched on text-guided 3D Scene Generation, use Large-language model(LLM)-based dreaming and move-and-lookaround strategy to generate both geometric and semantic consistent 3D scene.
- **OPPO Research Institute**  2022.02 – 2022.05  
Research Intern  
Beijing, China  
◦ Researched on image matting, proposed a framework to use human pose as guidance to achieve whole body matting.
- **Institute of Automation, Chinese Academy of Sciences, Beijing, China(CASIA)**  2021.01 – 2022.01  
Research Intern  
Beijing, China  
◦ Researched on model compression and network pruning, especially the application on Vision Transformers.








## SELECTED PUBLICATIONS

C=CONFERENCE, J=JOURNAL, P=PRE-PRINT

- [P.1] **Zhexiao Xiong**, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, Nathan Jacobs. **GroundingBooth: Grounding Text-to-Image Customization**. *Arxiv Pre-print*.
- [P.2] Feng Qiao, **Zhexiao Xiong**, Eric Xing, Nathan Jacobs. **GenStereo: Towards Open-World Generation of Stereo Images and Unsupervised Matching**. *Arxiv Pre-print*.
- [P.3] Feng Qiao, **Zhexiao Xiong**, Xinge Zhu, Yuexin Ma, Qiumeng He, Nathan Jacobs. **MCPDepth: Omnidirectional Depth Estimation via Stereo Matching from Multi-Cylindrical Panoramas**. *Arxiv Pre-print*.
- [C.1] **Zhexiao Xiong**, Zhang Chen, Zhong Li, Yi Xu, Nathan Jacobs. **PanoDreamer: Consistent Text to 360-Degree Scene Generation**. In *CVPR Workshops (CV4Metaverse)*, 2025.
- [C.2] Wanzhou Liu\*, **Zhexiao Xiong\***, Xinyu Li, Nathan Jacobs. **DeclutterNeRF: Generative-Free 3D Scene Recovery for Occlusion Removal**. In *CVPR Workshops (CV4Metaverse)*, 2025.
- [C.3] **Zhexiao Xiong**, Feng Qiao, Yu Zhang, Nathan Jacobs. **StereoFlowGAN: Co-training for Stereo and Flow with Unsupervised Domain Adaptation**. In *British Machine Vision Conference(BMVC)*, 2023.
- [C.4] Xin Xing, **Zhexiao Xiong**, Abby Stylianou, Srikumar Sastry, Liyu Gong, Nathan Jacobs. **Vision-Language Pseudo-Labels for Single-Positive Multi-Label Learning**. In *CVPR Workshops(CVPRW)*, 2024.
- [C.5] Subash Khanal, Eric Xing, Srikumar Sastry, Aayush Dhakal, **Zhexiao Xiong**, Adeel Ahmad, Nathan Jacobs. **PSM: Learning Probabilistic Embeddings for Multi-scale Zero-Shot Soundscape Mapping**. In *ACM Multimedia(ACM MM)*, 2024.
- [J.1] **Zhexiao Xiong**, Xin Xing, Scott Workman, Subash Khanal, Nathan Jacobs. **Mixed-View Panorama Synthesis using Geospatially Guided Diffusion**. *Transactions on Machine Learning Research(TMLR)*, 2025.
- [J.2] Nanfei Jiang, **Zhexiao Xiong**, Hui Tian, Xu Zhao, Xiaojie Du, Chaoyang Zhao, Jinqiao Wang. **PruneFaceDet: Pruning lightweight face detection network by sparsity training**. *Cognitive Computation and Systems*, 2022.

## PROJECTS

- **Physically Coherent Video Generation** 2025.01 – present  
◦ Propose a framework that leverages vision language model(VLM)'s understanding and reasoning ability to achieve video generation with physically-coherent motion.  
◦ Use VLM as test-time verifier during test-time scaling to sample physically-coherent and action-aligned object motion and trajectory.

- Grounded text-to-image Customization** 2024.01 – 2024.09  
*Collaboration with Adobe Research* 
  - Proposed a framework that achieved zero-shot instance-level spatial grounding on both foreground subjects and background objects in the text-to-image customization task, enabling the customization of multiple subjects.
  - Our work is the first work to achieve a joint grounding on both subject-driven foreground generation and text-driven background generation.
  - Results show the effectiveness of our model in text-image alignment, identity preservation, and layout alignment.
- Text to 360-Degree Scene Generation** 2024.05 – 2024.11
  - Proposed a holistic text to 360-degree scene generation pipeline, which achieved consistent text-to-360-degree scene generation with customized trajectory-guided scene extension.
  - Introduced semantically guided novel view synthesis into the refinement of 3D-GS optimization, reducing artifacts and improving geometric consistency.
- Mixed-View Panorama Synthesis Using Geospatially-Guided Diffusion** 2023.05 – 2023.11  

  - Introduced the task of mixed-view panorama synthesis, where the goal is to synthesize a novel panorama given a small set of input panoramas and a satellite image of the area.
  - Introduced an approach that utilizes diffusion-based modeling and an attention-based architecture for extracting information from all available input imagery.
- Open-World Generation of Stereo Images and Unsupervised Matching** 2024.09 – 2025.03  

  - Proposed GenStereo, a novel diffusion-based framework for open-world stereo image generation with applications in unsupervised stereo matching.
- Co-training for Stereo and Flow with Unsupervised Domain Adaptation** 2023.01 – 2023.05  

  - Built an end-to-end joint learning framework to combine unsupervised domain translation with optical flow estimation and stereo matching in the absence of real ground truth optical flow and disparity.
  - Applied novel constraints on the cycle domain translation process to achieve cross-domain translation with global and local consistency.
  - Employed task-specific multi-scale feature warping loss and iterative feature warping loss during the training phase to regulate the training process in both spatial and temporal dimensions.
- Vision-Language Pseudo-Labels for Single-Positive Multi-Label Learning** 2022.11 – 2023.05  

  - Proposed a novel approach called Vision-Language Pseudo-Labeling (VLPL), which uses a vision-language model to suggest strong positive and negative pseudo-labels, and outperforms the current SOTA methods by 5.5% on Pascal VOC, 18.4% on MS-COCO, 15.2% on NUS-WIDE, and 8.4% on CUB-Birds.
- Pruning Lightweight Face Detection Network by Sparsity Training** 2021.01 – 2022.01  

  - Performed the network training with sparsity regularization on channel scaling factors of each layer, and then removed the connections and the corresponding weights with the near-zero scaling factors after the sparsity training.
  - Applied the proposed pruning pipeline on a state-of-the-art face detection method, EagleEye, and got a shrunken model which has a reduced number of computing operations and parameters.
  - Achieved 56.3% reduction of parameter size with almost no accuracy loss on WiderFace dataset.
- Mobile AI 2021 Real-Time Camera Scene Detection Challenge** 2021.01 – 2021.03  
*Mobile AI Workshop @ CVPR 2021* 
  - Used two-stage fine-tuning method to improve the accuracy and the model pruning method to improve the model's efficiency.
  - Used the float32-to-int8 quantization and model pruning methods to optimize our model.

## SERVICES

- **Reviewer:** CVPR(2025), ECCV(2024), NeurIPS(2024,2025), ICML(2025), ICLR(2025)
- **Teaching Services (WashU):** CSE 559A Computer Vision (**Teaching Assistant/Grader**)

## TECHNICAL SKILLS

**Programming:** Python, C/C++, Java, Matlab

**Deep Learning Frameworks:** Pytorch, Tensorflow

**Research Frameworks:** Diffusion models, Transformer, GAN, 3DGS, NeRF, CNN, CLIP

**Languages:** English, Chinese