Zhipeng Bao

(+1)917-318-5651 | zbao@cs.cmu.edu | https://zpbao.github.io

Education

Carnegie Mellon University Pittsburgh, PA
PhD in Robotics *Aug. 2022 – 2025 (Expected)*

- **Advisor** Prof. Martial Hebert
- **Thesis** *Bridging Generative and Discriminative Learning with Diffusion Models*
- **Committee** Martial Hebert, Deva Ramanan, Jun-Yan Zhu, Alexei Efros (UCB), Yu-Xiong Wang (UIUC)

Master of Science in Robotics GPA:4.15/4.0 *Aug. 2019 – Aug. 2021*

- **Advisor** Prof. Martial Hebert
- **Thesis** *Introducing Generative Models to Facilitate Multi-Task Visual Learning*

Tsinghua University Beijing, China
Bachelor of Electronic Engineering GPA:3.53/4.00 *Aug. 2015 – July 2019*

- **Thesis** *Text-To-Speech Synthesis with Limited Data*
- **Honor** Comprehensive Excellence Award of Tsinghua University (2018)

Australian National University Canberra, Australia
Exchange Program GPA:6.33/7.00 *Feb. 2018 – June 2018*

- **Coursework** Computer Vision, Artificial Intelligence, Data Mining

Selected Publications

Anurag Bagchi, **Zhipeng Bao**, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. *Open-vocabulary Referring Object Segmentation with Video Diffusion Models.* Under Review.

Zhipeng Bao, Anurag Bagchi, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. *Unlock the Potential of Video Diffusion Models beyond Generation.* Under Review.

Shuhong Zheng, **Zhipeng Bao**, Martial Hebert, and Yu-Xiong Wang. *SUNDiff: Bridging Generative and Discriminative Learning with Diffusion Models.* Under Review.

Yunze Man, Shuhong Zheng, **Zhipeng Bao**, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. *Lexicon3D: Probing Visual Foundation Models for Complex 3D Scene Understanding.* Under Review.

Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. *Separate-and-Enhance: Compositional Finetuning for Text-to-Image Diffusion Models.* SIGGRAPH 2024.

Shuhong Zheng*, **Zhipeng Bao***, Martial Hebert, and Yu-Xiong Wang. *Multi-task View Synthesis with Neural Radiance Fields.* ICCV 2023.

Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. *Object Discovery from Motion-guided Tokens.* CVPR 2023.

Mingtong Zhang, Shuhong Zheng, **Zhipeng Bao**, Yuxiong Wang and Martial Hebert. *Beyond RGB: Scene-Property Synthesis with Neural Radiance Fields.* WACV 2023.

Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. *Discovering Objects that Can Move.* CVPR 2022.

Zhipeng Bao, Yuxiong Wang and Martial Hebert. *Generative Modeling for Multi-task Visual Learning.* ICML 2022.

Zhipeng Bao, Yuxiong Wang and Martial Hebert. *Bowtie Networks: Generative Modeling for Joint Few-shot Recognition and Novel-View Synthesis.* ICLR 2021.

Zhipeng Bao, Shaodi You, Lin Gu and Zhenglu Yang. *Single-Image Facial Expression Recognition Using Deep 3D Re-Centralization.* ICCV 2019 workshops.

Research Experience

Unlocking the Potential of Video Diffusion Models beyond Generation Jan 2024 – Sep 2024
Carnegie Mellon University. Advisor: Prof. Martial Hebert *Pittsburgh, PA*

- Pioneered the investigation of video diffusion models for tasks beyond generation, examining both Text-to-Video (T2V) and Image-to-Video (I2V) models
- Our findings affirm that T2V models provide a motion-aware and object-centric feature representation, establishing them as strong candidates for diverse video understanding tasks. The T2V features consistently outperform the image diffusion features by large margins for all the tasks
- Introduce a novel and training-free methodology to adapt I2V diffusion models for applications beyond conventional video generation, including video style transfer and video label generations
- Our paper is under review at ICLR 2025

Open-vocabulary RVOS with Video Diffusion Models March 2024 – Sep 2024

Carnegie Mellon University. Advisor: Prof. Martial Hebert *Pittsburgh, PA*

- Designed a novel and effective referring object segmentation (RVOS) model with pretrained video diffusion models for open-vocabulary videos
- Collected an evaluation dataset containing 200 videos that span stuff categories, object state-change, large ego-motion, etc
- Our method demonstrated promising results for open-vocabulary videos, not only for normal foreground objects but for stuff categories as well
- Our paper is under review at ICLR 2025

Probing Visual Foundation Models for Complex 3D Scene Understanding Jan 2024 – May 2024

Carnegie Mellon University. Advisor: Prof. Martial Hebert and Prof. Yu-Xiong Wang *Pittsburgh, PA*

- Present a comprehensive study that probes various visual encoding models for 3D scene understanding, identifying the strengths and limitations of each model across different scenarios
- Our evaluation spans *seven* vision foundation encoders, including image-based, video-based, and 3D foundation models. We evaluate these models in *four* tasks: Vision-Language Scene Reasoning, Visual Grounding, Segmentation, and Registration, each focusing on different aspects of scene understanding
- Key findings include: DINOv2 demonstrates superior performance, video models excel in object-level tasks, diffusion models benefit geometric tasks, and language-pretrained models show unexpected limitations in language-related tasks
- Paper under review at NeurIPS 2024 with strong reviews

Bridging Generative and Discriminative Learning with Diffusion Models Sep 2023 – March 2024

Carnegie Mellon University. Advisor: Prof. Martial Hebert and Prof. Yu-Xiong Wang *Pittsburgh, PA*

- Proposed SUNDiff, a unified framework that seamlessly integrates generative and discriminative learning based on diffusion models
- Introduced a novel self-improving mechanism that progressively boosts both data generation and discriminative learning
- Our method demonstrated consistent performance improvements across various discriminative backbones and high-quality data generation under both realism and usefulness
- Our paper is under review at ICLR 2025

Multi-task View Synthesis with Neural Radiance Fields Sep 2022 – March 2023

Carnegie Mellon University. Advisor: Prof. Martial Hebert and Prof. Yu-Xiong Wang *Pittsburgh, PA*

- Proposed a novel problem, multi-task view synthesis (MTVS), which formulates multi-task visual learning as a set of view synthesis tasks
- Proposed MuvieNeRF, a unified framework that leverages cross-view and cross-task information for the proposed MTVS problem
- Comprehensive experimental evaluations demonstrate that MuvieNeRF shows promising results for MTVS, and greatly outperforms conventional discriminative models, owing to the proposed CVA and CTA modules
- Our paper was accepted to ICCV 2023

Beyond RGB: Scene Analysis by Synthesis with Neural Radiance Fields June 2021 – June 2022

Carnegie Mellon University. Advisor: Prof. Martial Hebert and Prof. Yu-Xiong Wang *Pittsburgh, PA*

- Introduced a novel problem of “scene analysis by synthesis” that exploits generative modeling for a variety of scene understanding tasks
- Proposed an implicit representation-based model, SS-NeRF, that extended NeRF to simultaneously render novel-view images and their corresponding view-dependent and view-independent scene properties
- Evaluated SS-NeRF on several realistic datasets and reached comparable results with heuristic methods. Explored the applications of SS-NeRF with data augmentation and auto-labeler
- Our paper was accepted to WACV 2023

Generative Modeling for Multi-task Visual Learning June 2020 – June 2021

Carnegie Mellon University. Advisor: Prof. Martial Hebert *Pittsburgh, PA*

- Considered a novel problem of learning a shared generative model for various visual perception tasks, and proposed a general framework named MGM, by coupling a discriminative multi-task network with a generative network
- Evaluated MGM model on two standard multi-task benchmarks, the experimental results showed MGM consistently outperformed both SOTA single-task and multi-task approaches
- Further proposed a joint learning mechanism for MGM, which improved the performance of all the tasks by large margins. Studied the scalability of MGM framework to more visual tasks
- Our paper was accepted to ICML 2022

Bowtie Networks for Joint Recognition and View Synthesis Dec. 2019 – June 2020

Carnegie Mellon University. Advisor: Prof. Martial Hebert *Pittsburgh, PA*

- Introduced a novel dual-task of few-shot recognition and novel-view synthesis
- Proposed feedback-based bowtie networks that simultaneously learned 3D geometric and semantic representations with feedback. Addressed the incompatibility issues between different modules by leveraging resolution distillation
- The proposed framework significantly improved both view synthesis and recognition performance, especially in the low-data regime. The model was flexible to incorporate other tasks such as style-guided synthesis
- Our paper was accepted as a poster paper in ICLR 2021

Marker-Free Alignment for Electron Tomographic Projections July 2018 – Dec. 2018

Carnegie Mellon University. Advisor: Prof. Min Xu *Pittsburgh, PA*

- Designed an adaptive deep learning model for electron projection feature extracting, which achieved a robust performance for low-quality images compared with classic features such as SIFT and SURF
- Proposed an iterative algorithm for feature matching and patterns tracking for electron tomographic projections
- Contributed to form a complete pipeline for 3D reconstruction with raw tomographic projections
- Our paper was accepted as an oral paper in ISMB 2019

Industry Experience

GenAI, Meta Menlo Park, CA

Research Intern, Mentor: Dr. Xiaofang Wang *May 2024 – Aug 2024*

- Localized the limitation of current vision-language models in performing recognition-related tasks
- Proposed a novel self-distillation objective to enhance the recognition ability of VLMs without looping additional training data
- Prepared a submission for CVPR 2025

Adobe Research Seattle, WA

Research Intern, Mentor: Dr. Yijun Li *May 2023 – Aug 2023*

- Analyzed the underlying factors for the compositional misalignment in T2I diffusion models: attention overlaps and low activations
- Proposed a compositional fine-tuning algorithm with two novel objectives to tackle the two issues
- Our method not only demonstrated state-of-the-art performance on compositional image generation task, but also expressed promising generation capacity after large-scale training
- Filled a patent and authored a paper at SIGGRAPH 2024

Toyota Research Institute Los Altos, CA

Research Intern, Machine Learning Research Group. Mentor: Dr. Pavel Tokmakov *June 2022 – Sep. 2022*

- Focused on enhancing the performance of the motion-guided object discovery model (see below)
- Introduced an additional token feature space by reconstructing with the quantized discrete features, and also introduced different unsupervised grouping signals for this token space such as MCG signal
- The enhanced model greatly outperforms the pure motion-guided model, and achieved state-of-the-art object discovery performance on realistic driving benchmarks, such as TRI-PD and Waymo
- Our paper was accepted to CVPR 2023

Research Intern, Machine Learning Research Group. Mentor: Dr. Pavel Tokmakov *June 2021 – Nov. 2021*

- Studied the problem of object discovery – separating objects from the background without manual labels
- Scaled the recent frameworks for unsupervised object discovery from toy, synthetic images to complex, real world scenes by simplifying their architecture, and augmenting with a weak learning signal from motion
- Evaluation on a photo-realistic auto-driving dataset and real-world KITTI dataset demonstrated that the proposed approach outperformed both heuristic- and learning-based methods by capitalizing on motion cues
- Authored a paper accepted by CVPR 2022

DATA 61, CSIRO Canberra, Australia

Research Intern, Computer Vision research group. Mentor: Dr. Shaodi You *Feb. 2018 – Sep. 2018*

- Introduced a 3D facial reconstruction method to re-align the still face image, which significantly reduced the influence of orientations and shadings for a wide range of facial expression recognition tasks
- Proposed a novel triple-channel model for single image-based FER task using learning-based features, landmark features and 3D facial features to achieve a reliable expression detection
- Evaluated the proposed model on three real-world databases (CK+, OULU-CASIA, and RAF dataset), with the experimental results proving the proposed model outperformed other state-of-the-art methods
- Authored a paper accepted by ICCV 2019 workshops

Technical Skills

Languages: Python, MATLAB, Java, C/C++, HTML, R

Tools & Frameworks: Git, SVN, Tensorflow, Pytorch, Latex, Keras, SQL, Jupyter Notebook, Linux Operations