Xiaodong Wang

✉ wangxiaodong21s@stu.pku.edu.cn · G 180+ citations · Homepage

Education

Peking University, School of Software and Microelectronics Sep. 2021 – Jul. 2024 (Expected)

Master student in Software Engineering (Advisor: Assoc. Prof. Yuejian Fang)
GPA: **3.66/4.00**
Research interests: Visual generation, Multi-modal pre-training, Domain Adaptation
Major courses: AI Application Development (A+), digital image processing (A+), Machine Learning (A-)

Beijing Information Science and Technology University (BISTU) Sep. 2017 – Jul. 2021

Bachelor in Data Science & Big Data Technology (Advisor: Assoc. Prof. Xiaoming Huang)
GPA: **4.43/5.00** Rank: **1/32**
Major courses: Artificial Intelligence (99), Data Mining (98), Algorithm Design and Analysis (94)

Experience

Microsoft Research Asia, Beijing May. 2022 – Present

- Research intern at Natural Language Computing group, mentored by Dr. Chenfei Wu and Dr. Nan Duan
- Work closely with Dr. Zicheng Liu (IEEE Fellow) and Dr. Lijuan Wang at Microsoft Redmond
- Focus on Visual Content Generation and Multi-modal Pretraining, especially for texts and images

Megvii Technology, Beijing Jun. 2021 – Sep. 2021

- Research intern at Face++ Research
- Focus on Face Detection and Domain Adaptation

Chinese Academy of Sciences, Institute of Computing Technology Dec. 2020 – May. 2021

- Research intern at VIPL group, advised by Prof. Shuhui Wang
- Focus on Domain Adaptation and Transfer Learning

Projects

Multi-modal Large Language Models (Leader; ongoing project) Apr. 2023 – Present

- **Multi-task parameter-efficient fine-tuning (based on Vicuna-7B)**: (1) Stage-1: Aligned language and vision by only training a linear projection in large image-text pairs (CC3M). (2) Stage-2: trained an adapter in LLM to understand different human intentions (image reasoning, image editing, and image generation).
- **Pretraining (based on Llama-13B)**: pretrained Llama-13B and visual encoder in LAION-COCO and LAION-en datasets; only used 2 million images, but achieved zero-shot 91.6 CIDEr in NoCaps, 68.7 BLEU@4 and 191.9 CIDEr in COCO-caption (**SOTA score**).

NUWA-3D (Leader; a project belongs to Microsoft/NUWA team) Sep. 2022 – Jan. 2023

- The first study for diffusion models in 3D Photography. Expanded the diffusion models from the static 2D image out-painting to 3D photography.
- Learned to generate controllable 3D videos only on image datasets and kept a small training/inference gap.
- Proposed a novel self-supervised diffusion model and Masked UNet to learn to predict occluded regions.
- Exceled Stable Diffusion in 2D image out-painting and comparable with SOTAs in 3D photography.

NUWA-XL (A project belongs to Microsoft/NUWA team) Sep. 2022 – Feb. 2023

- The first diffusion over diffusion architecture for text to long videos, supporting parallel generation.
- For generating a 1024-frames video, we only need 26s, whereas AR diffusion models need around 8min.
- Three-level diffusion models from coarse to fine, learned at different granularities of video.
- Personal contributions: designed and implemented the injection for the first and last frames to learn to predict all middle frames from texts; used spatial attention layers to inject frames and masks to hint temporal information.

***Visual ChatGPT* (34K stars in GitHub, 140+ citations)** Feb. 2023 – Mar. 2023

- Opened up the research direction of using large language models (LLMs) as logic control centers and inspired work such as HuggingGPT and AutoGPT.
- Coupled fragmented visual models and proposed a visual processing model that can understand human intentions.
- Used discrete texts to communicate visual signals, allowing LLMs to understand and generate images.
- Personal contributions: responsible for the code implementation of the entire project, introduced the Chinese version interaction, and proposed template API.

Unsupervised Domain Adaptation: a Smoothness Perspective (Leader) Jan. 2022 – Jul. 2022

- Analyzed UDA methods from a new perspective: *smoothness*, which is defined by the intra-class variance.
- Promoted the smoothness of models, by introducing semantic consistency learning in unlabeled data, using strong and weak augmentations for each image separately.
- Proposed method not only reduced intra-class variances, but also increased inter-class variances.
- A plug-and-play module, achieved 73.2% in Office-Home, 86.3% in VisDA-C, and 52.6% in DomainNet.

Selected Honors

• Merit Student of Peking University	2022
• ACCV 2022 Student Travel Grant	2022
• Beijing Outstanding Graduates	2021
• President Scholarship (Highest Student Honor in BISTU)	2020
• Outstanding Student, BISTU	2018 & 2019 & 2020
• National Encouragement Scholarship	2019
• National Scholarship	2018

Skills

- Good at Multi-modal training algorithm design and implementation, diffusion models and LLMs.
- **Language**: Chinese (native), English (fluent, good writing, CET-6: 518, IELTS: prepared)
- **Skills**: Python, PyTorch, Linux, LaTeX, C++

Publications

- **Wang, X.**, Wu, C., Yin, S., Ni, M., Wang, J., Li, L., ... Duan, N. (2023). Learning 3D Photography Videos via Self-supervised Diffusion on Single Images. IJCAI 2023
- **Wang, X.**, Zhuo, J., Zhang, M., Wang, S., Fang, Y. (2022). Revisiting Unsupervised Domain Adaptation Models: A Smoothness Perspective. In Proceedings of the Asian Conference on Computer Vision
- **Wang, X.**, Zhuo, J., Cui, S., Wang, S. (2021). Learning invariant representation with consistency and diversity for semi-supervised source hypothesis transfer. arXiv preprint arXiv:2107.03008
- **Wang, X.**, Huang, X. (2020, October). Background Cleaning and Direction Weight in Salient Object Detection. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV)
- Jia, M., **Wang, X.**, Xu, Y., Cui, Z., Xie, R. (2020). Testing machine learning classifiers based on compositional metamorphic relations. International Journal of Performability Engineering, 16(1), 67
- Ni, M., Wu, C., **Wang, X.**, Yin, S., Wang, L., Liu, Z., Duan, N. (2023). ORES: Open-vocabulary Responsible Visual Synthesis. arXiv preprint arXiv:2308.13785
- Wu, C., Yin, S., Qi, W., **Wang, X.**, Tang, Z., Duan, N. (2023). Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671
- Yin, S., Wu, C., Yang, H., Wang, J., **Wang, X.**, Ni, M., ... Duan, N. (2023). NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation. ACL 2023 ORAL