Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Sijie Zhu

B.S., University of Science and Technology of China, 2015
M.S., University of Chinese Academy of Sciences, 2018

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

14 November, 2022, 1:00 P.M.
MSB 336

DISSERTATION COMMITTEE

Professor Chen Chen, *Chair*, *chen.chen@crcv.ucf.edu*
Professor Mubarak Shah, *shah@crcv.ucf.edu*
Professor Yanjie Fu, *yanjie.fu@ucf.edu*
Professor Mingjie Lin, *milin@ucf.edu*

DISSERTATION RESEARCH IMPACT

Cross-view image geo-localization aims to determine the locations of street-view query images by searching in a GPS-tagged reference image database from aerial view. One fundamental challenge is the dramatic view-point/domain difference between the street-view query images and aerial-view reference images. Recent works have made great progress on bridging the domain gap with advanced deep learning techniques and geometric prior knowledge, i.e. the query is aligned at the center of one aerial-view reference image (spatial alignment) and the orientation relationship between the two views is known (orientation alignment). However, such prior knowledge of the geometry correspondence of the two views is usually not available for real-world scenarios.

In this dissertation, we push cross-view image geo-localization toward real-world application with more realistic settings, higher accuracy, lower computational cost and better understanding/interpretation.

SELECTED PUBLICATIONS

1. **TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization,** Sijie Zhu, Mubarak Shah, Chen Chen, *IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2022.*
2. **VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval,** Sijie Zhu, Taojiannan Yang, Chen Chen, *IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2021.*
3. **Visual Explanation for Deep Metric Learning,** Sijie Zhu, Taojiannan Yang, Chen Chen, *IEEE Transactions on Image Processing (**TIP**), 2021.*
4. **Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation,** Sijie Zhu, Taojiannan Yang, Chen Chen, *Winter Conference on Applications of Computer Vision (**WACV**), 2021.*
5. **GALA: Toward Geometry-and-Lighting-Aware Object Search for Compositing,** Sijie Zhu, Zhe Lin, Scott Cohen, Jason Kuen, Zhifei Zhang, Chen Chen, *European Conference on Computer Vision (**ECCV**), 2022.*
6. **3D Human Pose Estimation with Spatial and Temporal Transformers,** Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, Zhengming Ding, *IEEE International Conference on Computer Vision (**ICCV**), 2021.*
7. **MutualNet: Adaptive ConvNet via Mutual Learning from Different Model Configurations,** Taojiannan Yang, Sijie Zhu, Matias Mendieta, Pu Wang, Ravikumar Balakrishnan, Minwoo Lee, Tao Han, Mubarak Shah, Chen Chen, *IEEE Transactions on Pattern Analysis and Machine Intelligence (**PAMI**), 2021.*
8. **GradAug: A New Regularization Method for Deep Neural Networks,** Taojiannan Yang, Sijie Zhu, Chen Chen, *Thirty-fourth Conference on Neural Information Processing Systems (**NeurIPS**), 2020.*

DISSERTATION

TOWARD REAL-WORLD CROSS-VIEW IMAGE GEO-LOCALIZATION

Learning to match street-view query and aerial-view reference images for geo-localization is very challenging due to the dramatic the appearance difference between the two views. Existing works generally adopt a two-branch CNN (Convolutional Neural Network) model, which does not encode geometric or positional information explicitly, so they reply on manually designed geometric transformation (i.e. polar transform) or GAN generation to reduce the view-point gap with prior knowledge. However, such the geometric prior knowledge depends on strong assumptions about the two views, i.e. the spatial and orientation alignment, and they may not hold for real-world scenarios. Real-world applications also require efficiency and explanation, which are not explicitly studied in this field.

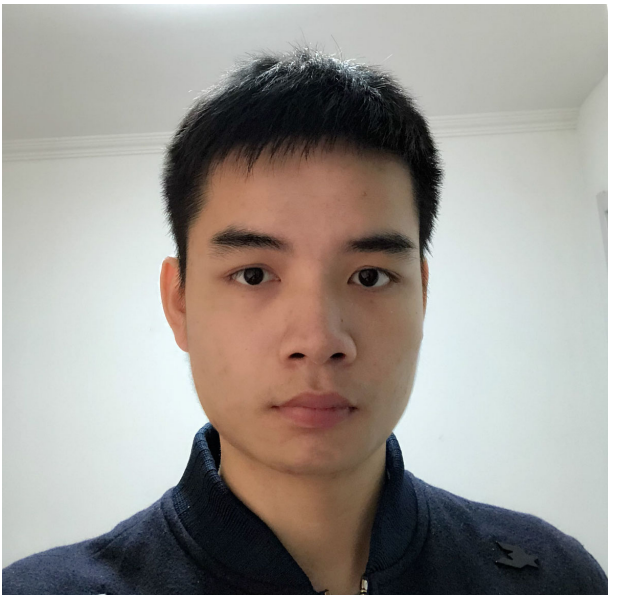
In this dissertation, we push this field toward real-world application in the following aspects: 1) Exploring more realistic scenarios, i.e. w/o spatial or orientation alignment. 2) Achieving better performance. 3) Studying the efficiency. 4) Introducing visual explanation. Current academic setting assumes perfect alignment on both spatial location and orientation, which is not the case of real-world scenario. Our exploration on realistic setting makes the setting of this field more complete. Our proposed real-world solution and visual explanation significantly boost the performance, efficiency, and interpretation ability, which are critical for practical applications. All these advancements together make this task closer to real-world use.

In Chapter 3, we for the first time consider spatial alignment issues in realistic scenario, where the query location is arbitrary in the AOI (area of interest) and the aerial-view reference images are densely sampled to provide a seamless coverage on AOI. Any arbitrary query is guaranteed to be covered at the central area of one positive reference aerial -view image and may be covered at the edge area of another three reference images (i.e. semi-positive images). Based on this real-world setting, we introduce a new dataset as a real-world testbed for other researchers and practitioners in this field.

Next, in Chapter 4, we provide a comprehensive revisiting and analysis about the orientation alignment issue for cross-view image geo-localization and point out that different orientation settings could lead to unfair comparison. Given that geometric prior knowledge is not available, we investigate another direction, i.e. better metric learning techniques, to improve the performance without orientation alignment. We further provide the visual explanation analysis to show the different behaviors of models trained w/ and w/o orientation alignment.

In Chapter 5, we provide the first quantitative study on visual explanation for deep metric learning and its applications including cross-view image geo-localization. We show the limitation of classification-based methods like GradCAM and propose a point-specific activation map to provide a point-to-point correspondence between two images. The proposed method can provide a better orientation estimation for cross-view image geo-localization than GradCAM, and it also shows better qualitative and quantitative results for other metric learning tasks.

In Chapter 6, we propose the first pure transformer based solution for all the mentioned challenging real-world scenarios where spatial and orientation alignment may not be available. The method does not rely on geometric prior knowledge (i.e. polar transform) and achieves state-of-the-art performance on different scenarios with less computational cost. We also propose to remove these uninformative regions based on attention, which can reduce computation cost or improve performance by reallocating the saved computation to other aspects, e.g. higher resolution.



Sijie Zhu

1993	Born in Hubei, China
2011-2015	B.S., University of Science and Technology of China, Hefei
2015-2018	M.S., University of Chinese Academy of Science, Beijing
2019-2021	Ph.D. Student, University of North Carolina at Charlotte, NC
2021	Research Internship, Adobe Research
2022	Research Internship, ByteDance Inc.
2021-2022	Ph.D., University of Central Florida, Orlando, FL