TIANJIAO DING

[Homepage: ding.tj] ◇ [Google Scholar] ◇ [tjding@seas.upenn.edu] ◇ +1 (323) 828 8830 ◇ Philadelphia, PA, USA

EDUCATION

University of Pennsylvania — Expected: 2026
Ph.D. in Computer and Information Science
Advisor: [Dr. René Vidal]   Collaborators: [Dr. Benjamin D. Haeffele], [Dr. Yi Ma]

Johns Hopkins University — 2023
M.S.E. in Applied Mathematics and Statistics

ShanghaiTech University — 2018
B.E. in Computer Science
Advisor: [Dr. Manolis C. Tsakiris]   Awards: Graduation with honor, Academic Excellence Scholarship 2014 & 2016

PROJECT HIGHLIGHTS

I enjoy using rigorous mathematics and practical implementations to address emerging challenges in machine learning.

Trustworthy ML

· Proposed zero-shot plug-and-play methods that visualize and steer the generation of LLMs [3] and diffusion models [1] via sparse coding on the latent activations using interpretable concepts; Achieved state-of-the-art performance on LLM alignment tasks (detoxification, improving faithfulness & sentiment) with LLaMA2 and image editing tasks with StableDiffusion 1.5; Released the first large-scale concept dataset of 40,000 concepts and 1,200,000 context sentences [3].

· Proposed algorithms that are provably robust to gross corruptions in the input data (noise [13], outliers [12, 11, 7, 6], false connections among clusters [9]) with applications in image clustering and robust geometric 3D vision.

Efficient ML

· Proposed variational forms of spectral functions, which enable transformer architectures with linear time and space complexity and competitive performance on long-range modelling tasks [2], as well as efficient representation learning frameworks [10].

· Proposed linearly convergent [13,7] and greedy algorithms [4] for extracting low- and high-dimensional structures in data.

· Proposed 3D reconstruction methods that efficiently estimate wireframes [14], motion parameters [12,11,7], and 3D surfaces [13].

Unsupervised ML

· Proposed unsupervised algorithms for clustering data lying on multiple low-dimensional manifolds, with state-of-the-art clustering performance on CIFAR and ImageNet [5, 8] and theoretical guarantees on no-false-connections [4].

· Proposed unsupervised methods that learn diverse and disentangled representations [8], which allows for finetuning vision-language model CLIP to yield better image-to-image/text retrieval and clustering performance [5].

PUBLICATIONS

1. Concept Lancet: Representation Decomposition and Transplant for Diffusion-Based Image
   Jinqi Luo, **TD**, Kwan Ho Ryan Chan, Hancheng Min, Chris Callison-Burch, and René Vidal
   Under Review

2. Token Statistics Transformer: Linear-Time Attention via Variational Rate Reduction
   Reviewer ratings: 8,8,8,6 out of 10, top 162/11670 ≈ 1.4% of all submissions
   Ziyang Wu, **TD**, Yifu Lu, Druv Pai, Jingyuan Zhang, Weida Wang, Yaodong Yu, Yi Ma, and Benjamin D. Haeffele
   Under Review

3. PaCE: Parsimonious Concept Engineering for Large Language Models
   Jinqi Luo*, **TD***, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and René Vidal
   [NeurIPS 2024] [pdf] [code] [video]

4. Geometric Analysis of Nonlinear Manifold Clustering
   Reviewer ratings: 8,7,6 out of 10, top 117/3650 ≈ 3% of posters

Nimita Shinde*, **TD***, Daniel P. Robinson, and René Vidal

[NeurIPS 2024] [pdf]

5. Image Clustering via the Principle of Rate Reduction in the Age of Pretrained Models

Tianzhe Chu*, Shengbang Tong*, **TD***, Xili Dai, Benjamin D. Haeffele, René Vidal, and Yi Ma

[ICLR 2024] [pdf] [code]

6. HARD: Hyperplane ARrangement Descent

Oral presentation

TD*, Liangzu Peng*, and René Vidal

[CPAL 2024] [pdf] [code]

7. Outlier-Robust Orthogonal Regression on Manifolds

TD*, Liangzu Peng*, and René Vidal

[OpenReview 2023]

8. Unsupervised Manifold Linearizing and Clustering

CPAL spotlight

TD, Shengbang Tong, Kwan Ho Ryan Chan, Xili Dai, Yi Ma, and Benjamin D. Haeffele

[ICCV 2023] [pdf]

9. Understanding Doubly Stochastic Clustering

Oral presentation

TD, Derek Lim, René Vidal, and Benjamin D. Haeffele

[ICML 2022] [pdf]

10. Efficient Maximal Coding Rate Reduction by Variational Forms

Christina Baek, Ziyang Wu, Kwan Ho Ryan Chan, **TD**, Yi Ma, and Benjamin D. Haeffele

[CVPR 2022] [pdf]

11. Boosting RANSAC via Dual Principal Component Pursuit

Yunchen Yang, Xinyue Zhang, **TD**, Daniel P. Robinson, René Vidal, Manolis C. Tsakiris

[arXiv 2021]

12. Robust Homography Estimation via Dual Principal Component Pursuit

TD, Yunchen Yang, Zhihui Zhu, Daniel P. Robinson, René Vidal, Laurent Kneip, and Manolis C. Tsakiris

[CVPR 2020] [pdf] [code] [video]

13. Noisy Dual Principal Component Pursuit

Oral presentation

Tianyu Ding*, Zhihui Zhu*, **TD**, Yunchen Yang, Daniel P. Robinson, Manolis C. Tsakiris, and René Vidal

[ICML 2019] [pdf] [code]

14. Learning to Parse Wireframes in Images of Man-Made Environments

Over 214 citations as of Dec 2024

Kun Huang, Yifan Wang, Zihan Zhou, **TD**, Shenghua Gao, and Yi Ma

[CVPR 2018] [pdf] [code]

TALKS

Parsimonious Representations in Modern AI

Warren & ASSET Center Research Mixer, UPenn	2024

Unsupervised Manifold Linearizing and Clustering

Conference on Parsimony and Learning (CPAL), Hong Kong University	2024
Vision Lab Retreat, UPenn	2023
Third Workshop on Seeking Low-Dimensionality in Deep Neural Networks (SlowDNN), MBZUAI	2023

Hyperplane Arrangement Descent

Conference on Parsimony and Learning (CPAL), Hong Kong University	2024

Doubly Stochastic Clustering: Algorithms, Theory, and Applications

Guest lectures invited by Dr. Chun-Guang Li, Online	2022 & 2023
Mathematical Institute for Data Science (MINDS) Retreat, JHU	2022
International Conference on Machine Learning (ICML), Maryland	2022

TEACHING

As a teaching assistant:

ESE 6450 Deep Generative Models	Fall 2024, UPenn
SI 132 Linear Algebra for Information Science	Spring 2020, ShanghaiTech
SI 232 Subspace Learning	Fall 2019, ShanghaiTech
SI 231 Matrix Analysis	Fall 2018, ShanghaiTech

As a tutor:

AP Calculus, Physics, Computer Science	2017 - 2020, Shanghai American School
Math and Physics of the Chinese College Entrance Exam	2014 - 2017

PROFESSIONAL EXPERIENCE

Research Assistant, Vision, Dynamics and Learning Lab at JHU and UPenn, USA — 2020 - now

· Advisor: [Dr. René Vidal]   Collaborators: [Dr. Benjamin D. Haeffele], [Dr. Yi Ma]

· Conducted detailed study of leveraging low-dimensionality of data for unsupervised learning including clustering [9,8,6,5,4], outlier rejection [7,6], representation learning [10,8,5], generative modelling [3,1], and neural architecture design [2].

· Developed algorithms that are provably robust to gross corruptions in the input data (outliers [12, 11, 7, 6], false connections among clusters [9,4]) with applications in image clustering and robust geometric 3D vision.

· Proposed to learn a linear representation for high-dimensional data lying on low-dimensional manifolds without labels [8], which yields state-of-the-art clustering performance on CIFAR & ImageNet and improves the representation of CLIP [5].

· Developed zero-shot methods that steers generative models (LLMs/diffusion) via sparse coding on the latent activations for detoxification, improving truthfulness and sentiment, and concept-based editing [3, 1].

· Proposed variational forms of spectral functions for 1) transformer architectures with linear time and space complexity and competitive performance on long-range modelling tasks [2] and 2) efficient representation learning frameworks [10].

Research Assistant, Tsakiris Lab at ShanghaiTech, China — 2017 - 2020

· Advisor: [Dr. Manolis C. Tsakiris]   Collaborators: [Dr. Laurent Kneip], [Dr. René Vidal]

· Conducted detailed study of globally optimal subspace learning methods for outlier-robust 3D vision problems, e.g., 3D surface estimation [13], 2/3-view structure from motion [12,11], rolling shutter camera motion estimation.

· Proposed to estimate motion parameters via a non-convex non-smooth ℓ^1 optimization problem, which provably tolerates $O(\#\text{inliers}^2)$ outliers [13], and performs competitively with state-of-the-art RANSAC algorithms while using an order of magnitude less running time on large-scale SLAM datasets [12,13,11]. Proved the uniqueness of solution via abstract mathematical methods [12].

Research Assistant, 3D Vision Group at ShanghaiTech, China — May - Dec. 2017

· Advisor: [Dr. Yi Ma]

· Proposed an algorithm that extracts a wireframe representation from images of man-made scenes for piece-wise planar 3D reconstruction [14], which outperforms state-of-the-art methods, e.g., LSD and MCMLSD.

Software Engineer Intern, ShanghaiTech, China — May - Dec. 2017

· Worked on a container-based GPU computing cluster, serving as the infrastructure for high-performance computing tasks at ShanghaiTech University.

Software Engineer Intern, ShanghaiTech, China — Summer 2016

· Used Rust (few courses taught it by then, e.g., Stanford's CS140e) to implement a concurrent web crawler that cloned > 100k items with guarantees on memory and thread safety.

Software Engineer Intern, iHuman Institute, Shanghai, China — Summer 2015

· Designed and implemented a mobile client for a wearable medical monitoring device.

TECHNICAL SKILLS

Languages	Python, Matlab, C/C++, Rust, Shell, LaTeX; English, Shanghainese, Mandarin
Dev Tools	PyTorch, Numpy, Sklearn, Wandb, Git, OpenMP, OpenGL
Mathematics	Matrix Analysis, Optimization, High-dimensional Probability