Research Statement

Zhiwen Fan (zhiwenfan@utexas.edu)

The future of visual computing is rapidly evolving towards seamlessly blending digital content with our physical world. 3D learning is driving transformative advances across robotics, medical imaging, extended reality (XR), architecture, and design, fundamentally reshaping human-computer interaction. In these domains, next-generation AI systems must efficiently perceive and interact with complex physical environments while recreating them in digital space. This fusion of virtual and physical realms promises to enhance human perception, productivity, and creativity in unprecedented ways.

My overarching research objective is to develop ***generalizable 3D foundation models*** that serves as cornerstones for immersive and spatial computing technologies. This model leverages few-shot and end-to-end learning from web-scale data, enabling AI systems to achieve efficient environmental representation, understanding, and safer interaction. My innovations in **few-shot, end-to-end, and semantic 3D learning** have established new benchmarks through four key breakthroughs:

1. Unifying geometric principles and generative priors for state-of-the-art 3D assets creation
2. Self-supervised learning frameworks for reconstructing 3D from minimal 2D image data.
3. Advanced contextual understanding that adapts to complex real-world environments
4. Real-time training and rendering capabilities essential for responsive, safe interactions

▷ **Research Thrusts:** My research works are upon *few-shot 3D learning with geometric principles and generative priors*, *ultra-efficient 3D reconstruction and rendering*, and *a unified 3D foundation model* that bridges the gap between virtual and physical worlds. These advances enable transformative applications that seamlessly integrate digital information with our physical environment, paving the way for safer and more responsible AI technologies while catalyzing the next generation of computing platforms.

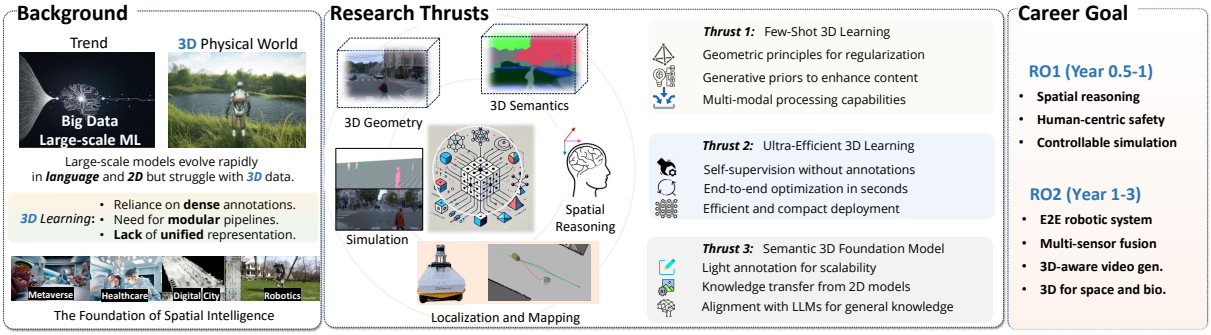

Figure 1: Overview of research background, three primary research thrusts, and long-term career goal.

Thrust 1: Few-Shot 3D Learning with Geometric Principles and Generative Priors

Creating photorealistic digital environments often requires dense scene capture with precisely annotated poses, which are frequently unavailable. My research addresses this fundamental **challenge of data sparsity**, particularly the scarcity of **camera views** or **pose parameters**. To overcome these constraints, I combine **Geometric Principles** with **Generative Priors** in novel ways. Geometric principles provide a strong foundation of structural constraints and spatial relationships inherent in the physical world. Complementing this, generative priors, learned from large datasets, can fill in missing information by leveraging statistical patterns in shape and appearance. By harnessing both the deterministic nature of geometry and the probabilistic power of generative models, my synergistic approach allows for designing architectures that learn effectively from limited data.

My research demonstrates the power of geometric principles to address few-shot challenges in two key projects. The **Cascade Cost Volume** [7] for multi-view depth maps decomposes depth estimation into multiple stages, progressively refining the depth search space based on preliminary estimates and utilizing a fine-grained feature pyramid. Central to this approach is cascade homography warping, which calculates matching costs across multiple stages, enabling high-resolution cost volume formulation and subsequent depth map estimation up to full resolution. Cascade Cost Volume is the first framework to enable city-scale vision-based reconstruction, with continuously improved accuracy from additional annotated 3D data. It achieves state-of-the-art performance in GPU memory efficiency, runtime, and depth accuracy and has become **one of the most widely adopted approaches (cited by 800+)** for

multi-view stereo (MVS), dense simultaneous localization and mapping (SLAM), and human modeling. Constructing volumes from image features can also enhance pose estimation. **Cas6D** [10] utilizes a 3D volume to estimate poses from Top-K candidates, progressively narrowing the pose search space to the optimal result with the lowest matching cost. Cas6D effectively addresses common failures in sparse-view scenarios.

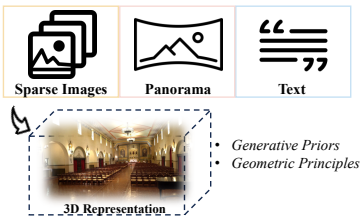

Figure 2: **Few-shot** 3D asset creation from sparse-view images [7, 11, 19], panoramas, or text [15]. Predefined poses are not required [10]

In response to the challenge of **few-shot view** reconstruction, my research addresses limitations in two complementary approaches: Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3D-GS), which originally require hundreds of captures to reconstruct a 3D scene. My work **SinNeRF** [18] and its follow-up **NeuralLift-360** [19] overcome this limitation by creating 3D assets from a single view, leveraging strong **generative diffusion priors** to inpaint unseen regions, integrated with **geometric regularizations** from monocular depth estimation trained with massive RGB-D data pairs, to guide 3D optimization in occluded 3D spaces where image observations are unavailable. Then, with the emergence of 3D Gaussian Splatting, which achieves real-time rendering and high-resolution output quality, we introduced Few-Shot Gaussian Splatting (**FSGS**) [11]. FSGS reconstructs 3D environments from only three images by employing Gaussian densification in empty spaces based on proximity scores to enhance scene detail and by leveraging monocular depth model priors to regularize scene surfaces. FSGS accelerates rendering speed by three orders of magnitude compared to NeRF baselines and surpasses all previous methods in content creation quality and efficiency.

My newest work, **DreamScene360** [15], further pushes the boundaries of 3D scene generation by creating immersive scenes with full 360° coverage from text prompts. DreamScene360 leverages the generative power of 2D diffusion models and LLM prompt self-refinement to produce high-quality, globally coherent panoramic images, which are then transformed into 3D representation. By imposing semantic and geometric constraints on both synthesized and input camera views, it optimizes in 3D to reconstruct unseen regions, resulting in globally consistent 3D scenes with a full 360° perspective. These few-shot 3D learning works streamline the process of creating digital worlds from limited data sources, making the customizable creation of digital assets more accessible to users.

Thrust 2: Ultra-Efficient 3D Reconstruction and Rendering

Thrust 1 significantly pushed the boundaries of 3D reconstruction with either few-shot camera poses or sparse views. However, what if *both limitations occur simultaneously*, as is often the case in practical scenarios? Thrust 2 addresses this even more ambitious and practically compelling challenge: **3D reconstruction directly from sparse, pose-free views**. Traditional pipelines involve a complex series of steps for accurate pose estimation and dense 3D reconstruction, each potentially introducing errors that propagate through the system. My research significantly simplifies this process by developing efficient **end-to-end** frameworks that connect input images with 3D representations, jointly optimizing both camera parameters and 3D structures under self-supervision, as illustrated in Figure 3.

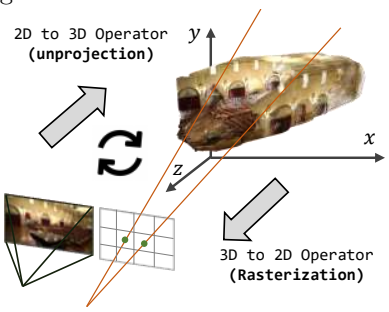

Figure 3: **End-to-end** 3D reconstruction from 2D images with self-supervision [6].

In **InstantSplat** [6], we introduce a groundbreaking approach leveraging the end-to-end 3D system, and **synergizing** the power of geometric priors with the capabilities of large **foundation models**. InstantSplat utilizes pixel similarity as an objective function between 2D images and the rasterized 2D projections from the optimized 3D representation. It incorporate a pretrained DUSt3R model [17] to initialize dense points, while subsequently optimizing camera parameters and scene attributes using Gaussian Bundle Adjustment efficiently. InstantSplat reduces large-scale 3D reconstruction time **from 33 minutes to just 7 seconds** with **as few as 3 unposed views**, while significantly improving visual quality and pose metrics. It is also generally compatible with other point-based representations such as Mip-Splatting or 2D Gaussian Splatting. My ongoing research, **VideoLifter**, develops novel algorithms to transform **in-the-wild videos of any length** into high-quality 3D representations. VideoLifter uses a hierarchical framework to track features across overlapping video segments, aligning them into a globally consistent 3D structure.

Additionally, my work addresses the challenge of deploying Gaussian Splatting on resource-constrained devices such as smartphones and headsets. **LightGaussian** [12] introduces a general framework to re-

duce redundancy in Gaussian Splats through visibility-based pruning, attribute distillation, and vector quantization, drawing inspiration from both machine learning and computer graphics. It achieves over 15× compression and 60% faster rendering while preserving reconstruction quality across multiple frameworks. These approaches facilitate immersive, free-view exploration across extensive reconstructed scenes, seamlessly merging digital and physical realities. This integration enables high-resolution, photorealistic human-computer interactions directly derived from everyday captures.

Thrust 3: Semantic 3D Foundation Model: Reconstruction, Understanding and Beyond

Thrust 3 aims to enhance the interaction between digital content and the physical world by developing **semantic 3D foundation models**. These models not only reconstruct scenes but also embed a deep understanding of each primitive. By integrating semantics with geometric representations [13], I strive to create more engaging experiences using intelligent 3D models with versatile capabilities. My goal is to develop a unified deep model that supports a wide range of downstream reconstruction and interaction tasks. This approach aims to enable applications from intelligent scene manipulation to advanced understanding and interaction, paving the way for sophisticated vision systems with spatial and physical awareness.

Our work on **Feature 3DGS** [20] extends the capabilities of 3D Gaussian Splatting to support a wide range of 3D functions, including open-vocabulary 3D semantic segmentation, language-guided 3D editing, and promptable 3D segmentation similar to "Segment Anything". By learning a structured, lower-dimensional 3D feature field from 2D foundation models and then upsampling it through a lightweight convolutional decoder, we achieve greater efficiency in feature field distillation, with faster training and rendering speeds compared to previous approaches.

My latest work, **Large Spatial Model [14] (LSM)**, introduces **the first real-time framework capable of predicting 3D geometry, appearance, and semantics from unposed images in a single pass**. LSM utilizes a scalable Transformer-based framework with cross-view and cross-modal attention to regress semantic anisotropic 3D Gaussians **in real-time**. LSM perceives the physical world by learning from data, and lifts 2D pre-trained models into 3D for consistent 3D open-vocabulary scene understanding and manipulation. To support this, LSM employs **novel view synthesis** during training as a foundational 3D task, constructing accurate 3D representations by establishing correspondences among input images and generating new images without the precomputed camera poses. This method demonstrates significant scalability across large datasets, necessitating only minimal 3D annotations to progressively refine the representation.

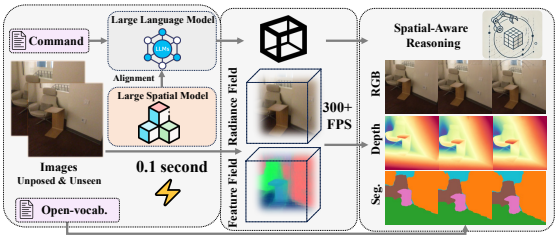

Figure 4: **Semantic 3D Foundation Model [14]:** supports both reconstruction and understanding, and enhances spatial awareness for other ML models.

Central to LSM is its capability to encode 3D scenes into compact, low-dimensional latent spaces, facilitating the **reconstruction of complete 3D scenes** from these embeddings. Through extensive training on large 3D and video datasets, LSM's latent space inherently cultivates **spatial awareness** from multiple images, obviating the need of preprocessing for 3D data. Building on this principle, our ongoing project, Geometric Language Model (GLM), combines **spatial awareness** in LSM with the **general reasoning capabilities** of large language models (LLMs) to enhance spatial reasoning and planning in intelligent machines, paving the way for models that integrate physical laws into spatially-aware, feed-forward, and real-time interactions.

Future Research Agenda

My career vision is to advance spatial intelligence through next-generation 3D learning algorithms. Building on my expertise in Few-shot 3D Learning, End-to-end 3D Contextual Understanding, and Generative AI, I will pursue four interconnected research directions that address critical challenges in autonomous systems, scientific discovery, and human-AI interaction. Through collaborations with industry and academic partners in robotics, healthcare, and hardware, I aim to establish foundational capabilities for the next-generation computing systems.

End-to-End Robotic Learning. While current large-scale AI models demonstrate exceptional capabilities in 2D and language tasks, they lack the spatial understanding required for precise and diverse robotic operations. Humans, by contrast, can intuitively learn tasks from video demonstrations and integrate vision, sound, and touch to interact effectively with environment. My future research aims to bridge this

gap by leveraging web-scale egocentric videos to track motion and reconstruct objects, environments, and interactions which could serve as key knowledge targeted for robotic tasks. By distilling large-scale human activity videos, I aim to develop intelligent generalist robots capable of perceiving, planning, and acting in 3D with human-like awareness and adaptability.

Building on my expertise in multi-view geometry for scene and object reconstruction [6, 7, 10] and the interpretation of human behavior from in-the-wild videos [2], I plan to scale data collection efforts using web-scale egocentric human activity videos to create fine-grained, annotated datasets across diverse scenarios with 6-DoF object poses, and human motions. This work will serve as the foundation for training embodied robots capable of understanding human-environment interactions and generalizing their manipulation policies to dynamic, real-world conditions. To develop unified foundational models for robotic tasks, I will focus on spatial awareness and multimodal perception, enabling robots to interpret 3D geometry from visual inputs and integrate visual, tactile, and auditory modalities. Drawing on my prior work in end-to-end 3D learning [14], multimodal SLAM [4], symbolic processing [5], and multi-task learning [3,9], I aim to collaborate with experts in robotics and hardware to develop platforms that enable robust generalization and energy-efficient performance across diverse robotic applications.

Human-Centric Safety Simulation for Autonomous Systems. Advancing spatial intelligence in autonomous systems demands a rigorous focus on human safety, particularly for vulnerable road users such as pedestrians and cyclists. Ensuring safety in rare and long-tail scenarios remains a critical challenge, as current state-of-the-art generative models and simulators often fall short in accurately modeling and simulating the unpredictable behaviors of non-rigid actors.

Building on my expertise in human modeling [2], scene reconstruction [6], and generative methods [16], I aim to create simulation platforms that capture the complexity of human behaviors and intensions, collaborated with industrial partners, such as Nvidia Research. These platforms will integrate advanced machine learning models to represent non-rigid actors, including their expressions, intentions, and dynamic actions. The simulator will also incorporate high-quality geometric surface modeling and generative priors for fine-grained realism and precise control.

3D Consistent Video Generation. Recent advancements in generative AI, particularly in video generation, hold tremendous potential as infinite data engines in digital media production, extended reality, and customized manipulation demonstrations. However, the scarcity of 3D annotations significantly limits these models' ability to ensure 3D consistency in generated videos.

In contrast, human-captured videos inherently exhibit 3D consistency and physical plausibility. My research vision is to embed 3D consistency into video generation models by leveraging a self-supervised pretraining paradigm that captures static geometry and tracks dynamic motion from in-the-wild real-world videos. Building on my previous work in 3D generation [15, 18], this approach aims to enable video generation models to produce 3D-consistent, cross-frame coherent videos, and could generate accurate geometric representations with only a small amount of high-quality 3D data for fine-tuning. I will release datasets and models to accelerate research in generative modeling for autonomy and digital media.

Bandwidth-Constrained 3D Reconstruction. Advancing spatial intelligence for scenarios with strict bandwidth constraints, such as constructing navigable 3D site models for space exploration or reconstructing protein structures using cryo-electron microscopy (cryo-EM) in biological imaging, poses significant challenges. These involve reconstructing 3D structures from limited or noisy 2D observations of textureless planar surfaces and visualizing the 3D organization of proteins and biomolecules from noisy, randomly oriented 2D cryo-EM images.

To address these issues, I will build on my expertise in neural reconstruction methods [6, 11], lunar simulation [1], and medical imaging [8] and bridge 3D learning with the constraints of real-world scenarios. My research seeks to solve critical scientific and engineering problems while fostering interdisciplinary engagement across fields.

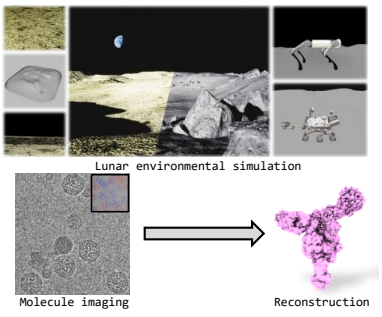

Figure 5: **AI for space and biological imaging.**

Selected publications: NAME denotes the author as Zhiwen Fan's mentee

References

[1] H. Bao, T.-H. Chen, Z. Chen, H. Lou, Y. Ge, **Z. Fan**, M. Pavone, and Y. Wang, "Moonsim: A photorealistic lunar environment simulator," in *submission*.

[2] H. Hu, **Z. Fan**, T. Wu, Y. Xi, S. Lee, G. Pavlakos, and Z. Wang, "Expressive gaussian human avatars from monocular rgb video," *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

[3] R. Sarkar, H. Liang, **Z. Fan**, Z. Wang, and C. Hao, "Edge-moe: Memory-efficient multi-task vision transformer architecture with task-level sparsity via mixture-of-experts," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE, 2023, pp. 01–09.

[4] L. C. Sun, N. P. Bhatt, J. C. Liu, **Z. Fan**, Z. Wang, T. E. Humphreys, and U. Topcu, "Mm3dgs slam: Multi-modal 3d gaussian splatting for slam using vision, depth, and inertial measurements," *International Conference on Intelligent Robots and Systems (IROS) 2024*, 2024.

[5] **Z. Fan**, T. Chen, P. Wang, and Z. Wang, "Cadtransformer: Panoptic symbol spotting transformer for cad drawings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 986–10 996.

[6] **Z. Fan***, W. Cong*, K. Wen*, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, "Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds," *arXiv preprint arXiv:2403.20309*, 2024.

[7] **Z. Fan***, X. Gu*, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2495–2504.

[8] **Z. Fan***, L. Sun*, X. Ding, Y. Huang, and J. Paisley, "Joint cs-mri reconstruction and segmentation with a unified deep network," in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*. Springer, 2019, pp. 492–504.

[9] **Z. Fan***, H*. Liang, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, Z. Wang *et al.*, "M^{3}vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 441–28 457, 2022.

[10] **Z. Fan***, P. Pan*, B. Y. Feng, P. Wang, C. Li, and Z. Wang, "Learning to estimate 6dof pose from limited data: A few-shot, generalizable approach using rgb images," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 1059–1071.

[11] **Z. Fan***, Y. Z. Zhu* andJiang, and Z. Wang, "Fsgs: Real-time few-shot view synthesis using gaussian splatting," in *European Conference on Computer Vision*. Springer, 2025, pp. 145–163.

[12] **Z. Fan***, K. Wang*, K. Wen, Z. Zhu, D. Xu, and Z. Wang, "Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps," *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

[13] **Z. Fan**, P. Wang, Y. Jiang, X. Gong, D. Xu, and Z. Wang, "Nerf-sos: Any-view self-supervised object segmentation on complex scenes," in *The Eleventh International Conference on Learning Representations*.

[14] **Z. Fan***, J. Zhang*, W. Cong, P. Wang, R. Li, K. Wen, S. Zhou, A. Kadambi, Z. Wang, D. Xu *et al.*, "Large spatial model: End-to-end unposed images to semantic 3d," *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

[15] **Z. Fan***, S. Zhou*, D. Xu, H. Chang, P. Chari, T. Bharadwaj, S. You, Z. Wang, and A. Kadambi, "Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting," in *European Conference on Computer Vision*. Springer, 2025, pp. 324–342.

[16] R. Li, P. Pan, B. Yang, D. Xu, S. Zhou, X. Zhang, Z. Li, A. Kadambi, Z. Wang, and **Z. Fan**, "4k4dgen: Panoramic 4d generation at 4k resolution," *arXiv preprint arXiv:2406.13527*, 2024.

[17] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[18] D. Xu, Y. Jiang, P. Wang, **Z. Fan**, H. Shi, and Z. Wang, "Sinnerf: Training neural radiance fields on complex scenes from a single image," in *European Conference on Computer Vision*. Springer, 2022, pp. 736–753.

[19] D. Xu, Y. Jiang, P. Wang, **Z. Fan**, Y. Wang, and Z. Wang, "Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4479–4489.

[20] S. Zhou, H. Chang, S. Jiang, **Z. Fan**, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 676–21 685.