Characterizing 3D Visual World in Parsimony via Exploring Structured Visual Geometries

Nan Xue
(xuenan@ieee.org)

1 Research Objectives

Our 3D world is constructed by shapes, with humans innately skilled in observing and interpreting these shapes. We, as human beings, focus on the shape information, uncovering and utilizing the structural regularities within, and abstracting the world using geometric primitives. For instance, 3D

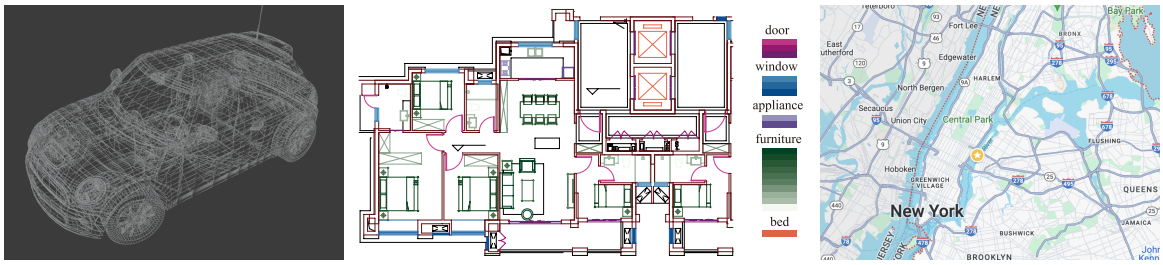

Figure 1: Shapes are everywhere in our daily life to represent a wide range of stuffs in the world, from the objects to our living indoor scenes, and the world in abstract yet expressive visual languages.

artists create clean, complete, and precise content using a minimal array of these primitives to depict the world, demonstrating their potency in capturing the essence of our environment. Similarly, we, as humans, possess the ability to gauge dimensions and spatial relationships, such as parallelism and orthogonality, with just our sight. This capability enables us to navigate confidently through structurally complex environments like indoor corridors and parking lots and to simplify our living spaces into maps with symbolic representations, as shown in Fig. 1.

The simplicity and efficiency of geometric primitives — including points, lines, curves, and planes, have always fascinated me for their remarkable ability to represent the complexity of our world in a parsimonious manner. My research is thus motivated by, and I always believe

> *The essence of creating a human-like perception system extends beyond mere pixel analysis, which should involve delving into shape information and structural regularities within visual data, whether concerning objects, indoor scenes, or larger worldly contexts.*

As a computer vision researcher, I am fascinated in thinking vision problems from the perspective of visual boundaries in both 2D and 3D cases, as they are directly related to the shape information while keeping the sparsity and compactness of shape representations. A research direction I have embarked on and aim to further develop involves devising learnable approaches to: (1) perceive geometric structures from raw pixel data in a set of points, lines, wireframe graphs, and planes; (2) organize them into structured representations; and (3) discover new computational approaches for reconstructing the world with the utmost structural regularities from 2D images. My research ambition is to inversely compute a CAD-like 3D world representation from images and video frames, pursuing parsimony, efficiency, and precision, using the simplest geometric primitives in a structured way, and eventually build a system to *program our visual world with codes*. I would like to name my research as *Structured Visual Geometry*, short in SVG.

2 Highlighted Contributions

My research journey began with exploring line segment representation for image matching [1] and has since broadened to include various representations such as points [1, 2], edge pixels [3], line segments [4, 5, 6], wireframes [7, 8, 9], and object masks [10, 11] for 2D image characterization. Building on insights gained from exploring 2D boundaries and shapes, my focus shifted toward structured 3D reconstruction and dynamic 3D reconstruction. Utilizing 2D boundaries, shapes, and

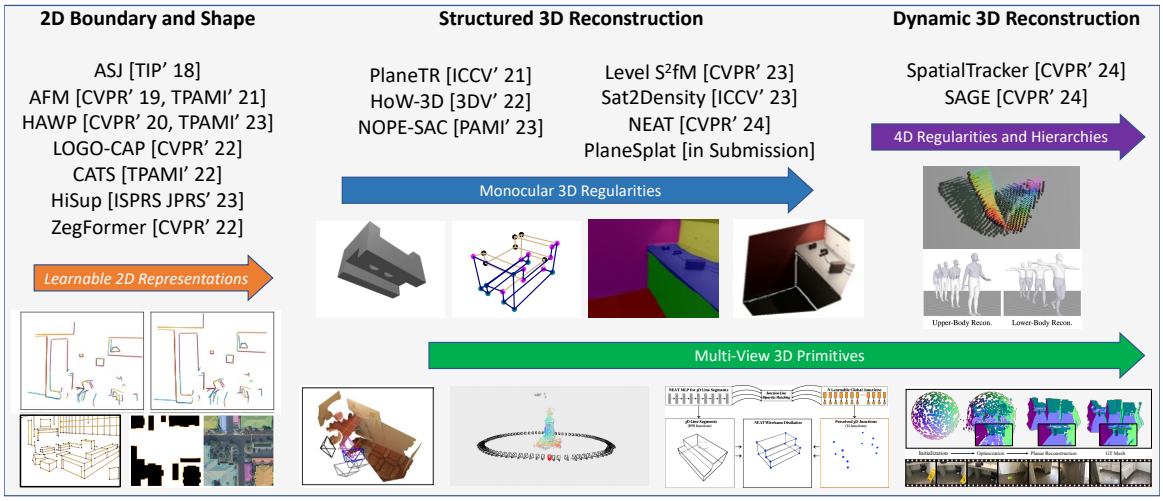

Figure 2: An illustrative summary of contributions.

inherent shape regularities, I proposed novel approaches for structured single-view 3D reconstruction in 3D wireframes [12] and 3D planes [13], sparse two-view 3D reconstruction in indoor [14] and outdoor [15] settings, surface-regularized neural Structure from Motion [16], and rendering-driven structured 3D reconstruction in wireframes [17] and planar primitives [18]. More recently, my studies have incorporated dynamics in monocular video and temporal sensory data for 3D reconstruction, aiming to reconstruct dynamic scenes [19] by tracking 2D pixels in 3D space and to infer full-body motion from sparse sensory observations [20], all from the perspective of discovering shape-related regularities and hierarchies. Fig. 2 provides an illustrative summary of my contributions, ranging from the learning of 2D boundary and shape from images, to the structured and dynamic 3D reconstruction.

In what follows, I will elaborate on the technical contributions from three parts among my publications.

2.1 Image Boundaries in 2D Line Segments and Wireframes

My research has extensively focused on boundaries in 2D images, with specific focuses on line segment detection and wireframe parsing through deep learning, inspired by the shortcomings of ASJ in specific scenarios. Notable contributions include the introduction of the Attraction Field representation and the enhancement of wireframe parsing with the Holistic Attraction Field representation, detailed in our works (Attraction Field representation [4, 6] and Holistic Attraction Field representation [7, 8]).

Attraction Field Representations. These efforts have led to the development of differentiable representations for sparse boundary geometries, facilitating advanced learning from data. The idea of attraction field representations also inspired us to rethink the detection of edge pixels, in which we tackled the localization ambiguity in learning-based edge detectors and introduced a context-aware tracing strategy (CATS) for crisp edge detection [3]. Our advancements, particularly with HAWPv1 [7], was leading the state-of-the-art for wireframe parsing and line segment detection for a long period. The evolution to HAWPv2 and HAWPv3 [8] further validates our innovative representations, showcasing superior performance in both fully-supervised and self-supervised learning contexts. HAWPv3, in particular, excels in learning line segments' inductive biases via the HAT field representation and shows promise for wireframe parsing in images beyond the training set. These contributions lay a robust foundation for geometric structure perception in 2D, encouraging further exploration into 3D perception and reconstruction from varied image views.

Duality between Boundaries and Shape Masks. Because the boundary representations are dual to the object masks for object-level representation, we are also interested in the object masks in the context of segmentation tasks, pursuing the label-efficient generalizable learning paradigms. We delved into the versatility of semantic-agnostic shape information for the object masks in both natural images [11, 21] and satellite images [10] to accomplish the goal of highly-accurate segmentation tasks. In ZegFormer [11] and HGFormer [21], we showcase that the class-agnostic grouping of pixels into shapes create a great chance to capture the shape information of object masks, thus facilitating the image segmentation tasks in zero-shot and domain generalization settings. For the instance

segmentation, we focused on the building instances in satellite images, and made contributions on the polygonal building extraction. In HiSup [10], we made efforts on closing the gap between the mask-based representation and the polygonal representation, exploited the hierarchical supervision (of bottom-level vertices, mid-level line segments, and high-level regional masks) and proposed a novel interaction mechanism of feature embedding sourced from different levels of supervision signals to obtain reversible building masks for polygonal mapping of buildings. Those studies demonstrated the importance of shape information, leveraged the versatility of shape attributes, and made connections between the object-level representations to the different-level of boundary geometries.

2.2 Structured 3D Reconstruction from Structured Visual Geometry

The research fall into this category can also be dating back to the junction matching paper [1], in which I came across challenges of reconstructing texture-less indoor scenes within the SfM frameworks that are based on local keypoints. I dreamed to, use structured visual geometry in 2D images, to develop the structure-aware 3D reconstruction approaches. In this section, the contributions are lie in three aspects, the structured 3D regularities [13, 12] in monocular 3D vision, the estimation of camera poses and scene geometry in correspondence-challenging multi-view setting [14, 16, 15], the multi-view 3D reconstruction in structured 3D geometric primitives [17, 18].

Monocular 3D Regularities. Single-view/Monocular cues indeed include 3D regularities such as the near-far relationship and vanishing points for infinity. We studied the monocular 3D regularities in neural designs, by using line segments and wireframes, aiming at discovering more general geometric relationship between the 2D and 3D primitives. The work PlaneTR [13], developed the first transformer-based solution that utilize the discrete tokenization of line segments, to infer 3D plane parameters alongside the image context, and resulted in holistic 3D reconstruction with state-of-the-art performance. The work HoW-3D [12] mimics our human visual systems to infer the invisible geometry of 3D wireframe from the visible line segments, greatly simplified the requirements of object-level 3D wireframe reconstruction in the minimal requirement of using only one image.

Two-View Correspondences. Two-view correspondences are crucial for multi-view 3D reconstruction, yet pose challenges in sparse-view configurations, especially in indoor and outdoor scenes like satellite-ground image pairs. Inspired by the human ability to intuitively identify correspondences, we made significant contributions to this area. For indoor scenes, Nope-SAC [14] leverages single-view 3D planes for relative camera pose estimation and holistic planar 3D reconstruction. For outdoor scenes, Sat2Density [15] uses a parameterized density field as a neural primitive to address viewpoint changes, demonstrating that terrain surfaces can be inferred accurately even without explicit 3D information as the supervision. These studies highlight that, by incorporating structural regularities in 2D/2.5D within a learning paradigm, the challenges of sparse-view correspondences can be effectively addressed using data. The recent work on two-view correspondences, such as Dust3R (Wang *et al.,* CVPR' 24), also supports these findings.

Multi-View 3D Reconstruction with Structured Regularities. The neural fields for volume rendering greatly advanced the multi-view 3D reconstruction. While many studies have demonstrated the ability of neural fields in different tasks, the most essential aspect, the implicit matching of scene geometry was not fully discovered, in particular with the structure visual geometry of points, lines and planes. We fill the empty in the realm of volume rendering using structured visual geometry of boundaries, focusing on the interaction between the neural surfaces (in SDF) and the explicit geometries towarding to the structured 3D reconstruction. In Level-S^2fM [16], we pioneered the first neural Structure-from-Motion (SfM) solution, leveraging inductive biases between SDFs and 2D keypoint correspondences. This approach uses continuous surfaces for top-down regularization, aligning triangulated 3D points from 2D correspondences with the surface's zero-level set, and redefines SfM as pose-free rendering. We further refined this with inductive biases between ray tracing and sphere tracing, revising traditional SfM components like triangulation within neural surfaces. Building on this, NEAT [17], developed atop HAWPv3 [8], offers a matching-free neural strategy for 3D wireframe reconstruction from 2D wireframes. NEAT's rendering-distilling approach addresses line-segment matching limitations, achieving a compact 3D wireframe representation from multi-view images. This technique significantly enhances initialization for 3D reconstruction tasks, demonstrating its efficacy with 20 times fewer points than conventional methods. Our latest endeavor [18] explores structured 3D reconstruction from planar primitives without explicit matching, through a differentiable rendering process that optimize discrete 3D planes based on 2D image loss from a random initialization. This results in a holistic, dense, and accurate 3D scene representation, advancing indoor scene surface representation without relying on surface marching from SDF.

2.3 Dynamic 3D Reconstruction from Structured Regularities and Hierarchies

Dynamic motions are common in daily life but pose challenges for traditional 3D reconstruction systems, which rely on the assumption of static scenes. Understanding dynamic motions is crucial for the next generation of 3D reconstruction systems from casual captures. In our recent work, we made significant advancements in understanding dynamic motions, including tracking any point in videos [19] and capturing body motion using sparse sensory data from head-mounted devices [20]. In SpatialTracker [19], we developed a feedforward neural method to track pixels as 3D points, creating long-range 4D trajectories from monocular videos by learning local rigidity with an As Rigid As Possible (ARAP) loss. Its point-level representation's universality allows SpatialTracker to perform well on real-world data without tuning, enabling straightforward downstream applications like camera pose and rigidity estimation. In SAGE [20], we explored structured regularity and hierarchies in body motion, inspired by the SMPL representation. We discovered a significant pattern of disentanglement, where the unobservable lower-body motion is highly influenced by the upper-body motion. A latent-diffusion-based solution was designed to learn the distributional trajectories of body joints, first predicting upper-body motion from sparse observations at the head and hands, then conditioning the generated upper-body motion on the unknown lower-body motion. This intuitive disentanglement highlights the importance of structures in vision problems and demonstrates that other non-trivial separations may not be as effective for modeling body motions.

3 Ongoing Initiatives and Future Plan

My future research will continue to build upon the foundational work in Structured Visual Geometry (SVG), aiming to advance the field of computer vision by further enhancing the ability of machines to perceive, understand, and reconstruct the 3D world in the gist of compact, structured and informative, from casually-captured visual data (in single-view and multi-view images, as well as monocular videos).

3.1 Monocular 3D Regularities from 2D Wireframes in Self-Supervised Learning

Single-still images are the most accessible visual data modality, containing a wealth of information about shape and boundaries. While we have explored learning boundaries from images and monocular 3D regularities, the scale of data has been limited due to constrained resources in data and computation. Recently, we have been focusing on *learning the boundaries of shapes* using the self-supervised HAWPv3 model [8]. We have observed the potential for scaling up the holistic attraction fields using 1 million unlabeled images. Furthermore, we found that the heuristic designs in line verification are already encompassed within the holistic attraction field representation itself. This has led to a new version of the HAWP models in a purely one-stage fashion, revisiting some of our initial observations in AFM [6]. More importantly, we discovered that traditional designs from the pre-deep-learning era can be learned based on the holistic attraction field representation to achieve better performance. Issues such as false alarms of line segments and wireframes, as well as incomplete detection results, will no longer be a concern. Building upon these known findings, we will have a powerful tool to delve into monocular 3D regularities with a large-scale and geometric-oriented pre-trained model. This approach can address problems including, but not limited to, monocular metric depth and normal estimation, vanishing point estimation, and gravity direction estimation. With the geometrically-induced monocular 3D cues, we believe, the many prequisitions in 3D reconstruction would be largely alleviated or even eliminated.

3.2 Structured SfM and SLAM from Casual Captures in Indoor Scenes

The current SfM/SLAM systems are mainly built on the keypoints to estimate camera poses and 3D point clouds by consuming a video or multi-view image set. It should be noted, the 3D point clouds, especially for the one by point triangulation, only locally captured the scene geometry in an unstructured manner. On one hand, although we can identify the 3D scene via the unstructured 3D point cloud, it is mainly due to the Gestalt principles for perceptual grouping. On the other hand, in indoor scenes, the texture-less yet structurally-meaningful characteristics of the casually captured multi-view images will probably lead to a challenging scenario with poor correspondences, and then a potential failure to reconstruct the scene. Accordingly, how to use the structurally-meaningful characteristics of indoor images to build the next generation of 3D reconstruction systems is desired. Based on my recent findings on structured 3D reconstruction, as well as the monocular 3D regularities from 2D Wireframes in Sec. 3.1, a seemingly promising direction is to merge the Level-S^2fM [16], NOPE-SAC [14], NEAT [17] and PlaneSplat [18] all in one, using the structured visual geometries in 2D and 3D, to simultaneously estimate camera poses and structured scene geometry in holistic primitives. Many problems such as the efficiency issues, the interaction between different primitives, as well as the generalizable learning without per-scene optimization would be further studied in this

direction. The expectation of this direction is, we can reconstruct/build a structured 3D scene from pose-agnostic video captures by anyone, to eventually form a clear structure for the room, and large-scale buildings, in wireframes as the scene skeleton, and holistic planar representations for the room structures or building interiors, and characterize the objects in indoor scenes using reconstructed CAD models (instead of the retrival ones). We believe, such an ideal organized representation of indoor scenes would provide a structure-induced visual language of the scene, and will facilitate the embodied AI tasks in a different way.

3.3 Scene Language: Towards Symbolic Visual Representations of Visual Worlds

Atop Structured SfM and SLAM, I am interested in further organizing structured visual geometries to elevate the information level from geometric shapes to higher-level semantics, forming a visual language representation based on parsimonious 3D scene representations. This approach marries the recent trends in Artificial Intelligence led by Large Language Models (LLMs). The structured visual geometry naturally provides an organization of visual concepts using points, lines, and planes/faces, which can align with language forms that have their own grammatical rules for organizing characters, words, and sentences. This can address the issue of lacking explicit structures in visual data, potentially providing a powerful tool to abstract the visual world into a *Scene Language*.

Ideally, the *Scene Language* should have the functionalities to: (1) Interpret raw pixel data into 3D structured visual geometry; (2) Tokenize the 3D SVGs using neural networks and make connections with pretrained LLMs; (3) Abstract visual information from the tokenized 3D SVGs and learn knowledge from LLMs; and (4) Leverage the learned knowledge to refine and complement the 3D geometries. In this direction, the structured visual geometry can be seen as a fast perception system, while the scene language is regarded as the "brain" of the visual world, thinking slower. Once deeper knowledge discovery is accomplished by querying LLMs using structured scene geometry, and as the structured visual geometry of boundaries is in explicit form, we can directly perceive the refined SVGs for scene representation.

References

[1] **Nan Xue**, Gui-Song Xia†, Xiang Bai, Liangpei Zhang, and Weiming Shen. Anisotropic-Scale Junction Detection and Matching for Indoor Images. *IEEE Transactions on Image Processing*, 27(1):79–91, 2018.

[2] **Nan Xue**, Tianfu Wu†, Zhen Zhang, and Gui-Song Xia. Learning Local-Global Contextual Adaptation for Fully End-to-End Bottom-up Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Linxi Huan*, **Nan Xue***, Xianwei Zheng†, Wei He, Jianya Gong, and Gui-Song Xia. Unmixing Convolutional Features for Crisp Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 44(10), 2022.

[4] **Nan Xue**, Song Bai, Fudong Wang, Gui-Song Xia†, Tianfu Wu, Liangpei Zhang, and Philip H.S. Torr. Learning Regional Attraction for Line Segment Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 43(6):1998–2013, 2021.

[5] Zhu-Cun Xue, **Nan Xue**, Gui-Song Xia†, and Weiming Shen. Learning to Calibrate Straight Lines for Fisheye Image Rectification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] **Nan Xue**, Song Bai, Fudong Wang, Gui-Song Xia†, Tianfu Wu, and Liangpei Zhang. Learning Attraction Field Representation for Robust Line Segment Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] **Nan Xue**, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia†, Liangpei Zhang, and Philip H.S. Torr. Holistically-Attracted Wireframe Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] **Nan Xue**†, Tianfu Wu, Song Bai, Fu-Dong Wang, Gui-Song Xia, Liangpei Zhang, and Philip H.S. Torr. Holistically-Attracted Wireframe Parsing: From Supervised to Self-Supervised Learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2023.

[9] Jiakun Xu, Bowen Xu, Gui-Song Xia, Liang Dong, and **Nan Xue**. Patched Line Segment Learning for Vector Road Mapping. In *AAAI*, 2024.

[10] Bowen Xu*, Jiakun Xu*, **Nan Xue**†, and Gui-Song Xia†. Accurate Polygonal Mapping of Buildings in Satellite Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2023.

[11] Jian Ding, **Nan Xue**, Gui-Song Xia†, and Dengxin Dai. Decoupling Zero-Shot Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] Wenchao Ma, Bin Tan, **Nan Xue**†, Tianfu Wu, Xianwei Zheng, and Gui-Song Xia. HoW-3D: Holistic 3D Wireframe Perception from a Single Image. In *IEEE Conference on 3D Vision (3DV)*, 2022.

[13] Bin Tan*, **Nan Xue***, Song Bai, Tianfu Wu, and Gui-Song Xia†. PlaneTR: Structure-Guided Transformers for 3D Plane Recovery. In *International Conference on Computer Vision (ICCV)*, 2021.

[14] Bin Tan, **Nan Xue**†, Tianfu Wu, and Gui-Song Xia. NOPE-SAC: Neural One-Plane RANSAC for Sparse-View Planar 3D Reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2023.

[15] Ming Qian, Jincheng Xiong, Gui-Song Xia, and **Nan Xue**†. Sat2Density: Faithful Density Learning from Satellite-Ground Image Pairs. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[16] Yuxi Xiao, **Nan Xue**†, Tianfu Wu, and Gui-Song Xia. Level-S^{2}fM: Structure from Motion on Neural Level Set of Implicit Surfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] **Nan Xue**, Bin Tan, Yuxi Xiao, Liang Dong, Gui-Song Xia, and Tianfu Wu. NEAT: Distilling 3D Wireframes from Neural Attraction Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] Bin Tan, Rui Yu, Yujun Shen, and **Nan Xue**†. PlanarSplat: Structured 3D Reconstruction of Indoor Scenes via Learning 3D Planes. In *Submission*, 2024.

[19] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, **Nan Xue**, Sida Peng, Yujun Shen, and Xiaowei Zhou. SpatialTracker: Tracking Any 2D Pixels in 3D Space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Highlight*, 2024.

[20] Han Feng*, Wenchao Ma*, Quankai Gao, Xianwei Zheng, **Nan Xue**†, and Huijuan Xu. Stratified Avatar Generation from Sparse Observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Oral*, 2024.

[21] Jian Ding, **Nan Xue**, Gui-Song Xia†, Bernt Schiele, and Dengxin Dai. HGFormer: Hierarchical Grouping Transformer for Domain Generalized Semantic Segmentation . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.