Multi-modal Localization and Point Cloud Reconstruction in Challenging Industrial Environments

CS231N Final Project - Milestone 1

Zihan(Monica) Geng, Eyas Taifour, Taarush Grover

{ zihangen, etaifour, taarush } @stanford.edu

Image provided by the Hilti 2023 SLAM dataset challenge website..
Augmented with ChatGPT to extend the left wall.

Problem Motivation

3D reconstruction alone fails in dark, textureless industrial environments. Can fusing IMU data into SOTA model (VGGT) improve point cloud reconstruction in low-texture, low-light industrial environments?

3D Reconstruction models - specifically VGGT which we focus on - includes inductive biases that prevent it to generalize to out of distribution scenes such as construction sites and underground rooms, which are characterized by faint illumination, repetitive surfaces, and the lack of clear visual landmarks.These conditions cause image-only model to produce poor depth estimates and inaccurate camera poses, which in turn yields poor results.

The Model: VGGT is a state-of-the-art Feed-Forward Network built on a DinoV2 backbone. It outputs:

- Camera poses and intrinsics (R/t, K matrices)
- z-Depth estimates per-pixel (Based on a DPT architecture)
- z-Depth Confidence scores per-pixel
- 2D correspondance tracking (Keypoint tracking, based on CoTracker architecture)

Which can be used further downstream for SLAM, Occupancy grids, Gaussian splats priors, Semantic analysis, etc..

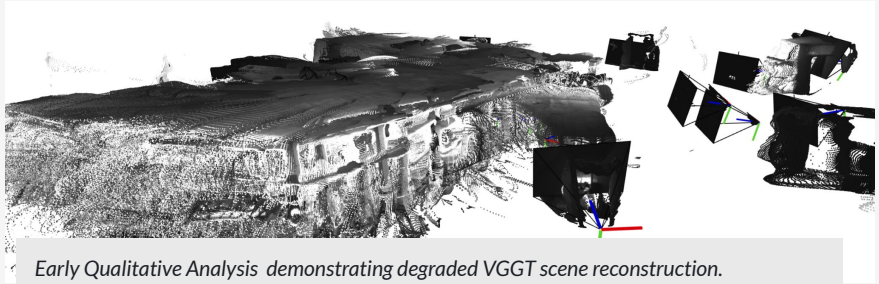
The Problem: VGGT deals with stationary images, disregarding temporal information (motion) or readily available external data points such as 6 DOF IMU. VGGT fails at building high confidence depth maps for textureless / industrial settings. The Hilti 2023 dataset images are B&W

Our Hypothesis: IMU could form a prior for the point cloud reconstruction, and as such could compensate for the loss of data when image-features are uninformative on their own. BW images can be augmented with additional representations

Dataset / Evaluation Setup

The Hilti SLAM Challenge 2023 Dataset (link)

Sample image from the dataset: site1_handheld1/alphasense_cam0. Cropped.



Early Qualitative Analysis demonstrating degraded VGGT scene reconstruction.

A real-world benchmark dataset collected at active construction sites, designed for testing localization and mapping systems in GPS-denied, challenging industrial environments. The data is captured using a handheld sensor rig across multiple floors, underground levels, and staircases.

Sensor	Data	Rate
Cameras (5)	B/W images 720x540,	10Hz
IMU	6 DOF	200Hz
LiDAR	32-channel point clouds	10Hz

Baseline: Comparison of VGGT forward-pass (frozen) vs. LiDAR GT (disregarding world frame): chamfer distance for images only

Experiments overview:

Input: RGB images + IMU (experimenting different image deformation/distortion/fusion/augmentation)

Output: Predicted point clouds

Ground truth: Raw LiDAR scans per frame

Key Metric: Chamfer Distance,

Additional Metrics: ATE, RPE

Note: Hilti does not provide camera pose ground truths for the complete trajectory, therefore we will focus on reconstruction at the local frame instead of world frame first,

Plan / Progress

Analyze VGGT to understand the characteristics of failure in industrial settings, and run experiments to augment the feature representation of the scene with IMU data. We evaluate against the provided LiDAR ground truth.

A. **Model & Data validation (ROS to PNG, IMU data, etc) -** Confirmed we can run the models/ have access to data

B. **Exploratory Data Analysis** (Validate image distortion/IMU integration/Frequency balance/Fill in GT gaps if needed (Colmap)

C. Augment GT:

In addition to the provided LiDAR Point clouds, we will also attempt at building a global PC with SfM (Colmap)
And a VIO pipeline (VINS-MONO)
LiDAR runs at a slower refresh rate than cameras (10 Hz vs 40 Hz). We will explore methods to fill the gaps (if deemed required). Alternatively, we will consider using a batch of 4 sequential images as input + their IMU data to produce 1 LiDAR point cloud

D. Root Cause Analysis

Vanilla VGG ablation: Quantitative & Qualitative evaluation of VGGT on Hilti data (images only) - Characterize the failure of VGGT on existing data, comparing point cloud output to LiDAR GT and calculating Chamfer distance. Explore Attention maps (Hilti vs. VGGT sample dataset) to determine possible source of the construction failure

Perform confidence sweep - (Drop 3D points < X% to find the confidence threshold that results in highest Chamfer distance scores)

Analyse Correspondence and tracking accuracy

Explore the features generated by Dinov2- If deemed the source of the 3R problem, consider augmenting them using external sources such as CLIP, FiLM, or fusing IMU data (exploring early/ late fusion, or other techniques discussed in video understanding).

Explore the effect of camera tokens / Papers such as Fisheye3R enhanced VGGT by modifying the camera tokens only. We will explore if such a strategy is applicable to our problem statement

E. Evaluate strategies to resolve the issue

Evaluate early/late fusions (IMU), including pre-integration of IMU (a la VINS-MONO)

Evaluate attaching probes to VGGT (to keep the backbone frozen) vs. (least favourably) re pre-training.

Related Works

VGGT: Visual Geometry Grounded Transformer

A FFN that outputs camera poses/point clouds (depth/confidence/tracking) using transformer blocks. The innovations are (1) Dinov2 feature extraction (2) "**Alternating Attention"** Transformers (frame-wise attn. + global attn.) as well as (3) camera tokens that learn R, t. The representation is then split across multiple heads ((4) Camera pose: 4 Transformer blocks), (5) Dense Prediction based on DPT, and (6) 2D tracking info based on CoTracker. Multiple variants where built on VGGT to resolve drawbacks (quadratic mem. requirement, streaming, Multi-view images, Wide FOV, etc) but none fuse 6DOF to correct camera pose/enhance features in textureless environments. VGGT builds on previous work most notably VGGSfM

VINS-Mono: A Robust & Versatile Monocular Visual Inertial State Estimator

Tightly coupled monocular Visual Intertial Odometry architecture. We will likely adopt their approach to fuse IMU into VGGT. We will use VINS-MONO output as a baseline to compare the LiDAR GT against a VIO pipeline. Instead of re-integrating IMU measurements every time the optimizer updates a state, **VINS-MONO contribution is the preintegration of IMU which collapses the inertial readings between two consecutive image frames into a single relative-motion factor** whose Jacobians can be analytically re-linearized when bias estimates change. This way we could make the 400 Hz IMU cohabit the 40 Hz image stream

On Geometric Understanding and Learned Priors in Feed-forward 3D Reconstruction Models

This paper interrogates how FFN 3R like VGGT actually arrive at their 3D predictions. The question they ask is "**do these networks learn epipolar geometry, or do they rely on learned priors arising from large training setups?** They conclude that they do learn geometry indeed (they probe representations at different layers for indicators) They find the network effectively learns the camera **F** matrices. The paper provides ablations for textureless scenes

LiDAR-VGGT: LiDAR Guided Visual Geometry Grounded Transformer

It combines VGGT image-based reconstruction with LiDAR geometry instead of relying on RGB.
This is relevant since **it shows a direct path for sensor-aware VGGT:** use external depth/geometry to correct scale, pose, and point-cloud failures in difficult scenes. While we will not rely on fusing LiDAR to input data, we will look at different methods to fuse external features to enhance pose estimations

Risks

- Our dataset's images are B&W, but VGGT uses a DinoV2 backbone that was not trained solely on B&W. We will run analysis to understand the impact this has on VGGT, and experiment with approaches to augment the feature map of B&W images, or alternatively enhance the correspondence tracking if we deem it the source of the issue.
- Multi view images are provided but we are still considering 1 image stream so far: we will probably explore how multi view images affect the outcome
- Images have wide FOV while VGGT was trained on perspective images / pinhole cameras: we will probably undistort the images using openCV (we've experimented a bit with this)
- We're unsure if the IMU is properly calibrated. Initial integration over 6DOF data show the operator walking at 20 m/s, which is gravely erroneous
- LiDAR runs at 10 Hz, while cameras are at 40 Hz (we are limited to ¼ GT)
- The rig operator (the human holding the rig) can be seen in one the cameras - we might need to mask the operator using a segmentation technique, potentially SAM
- We might not have enough compute/time to finetune VGGT if we need to (~48GB, Batch size=1, float16)
- Our data is collected in one location only (different times/illuminations/rooms). It might still not generalize to out of domain distributions

References (1/2)

- Ashish Devadas Nair, Julien Kindle, Plamen Levchev, & Davide Scaramuzza. (2024). Hilti SLAM Challenge 2023: Benchmarking Single + Multi-session SLAM across Sensor Constellations in Construction.
- René Ranftl, Alexey Bochkovskiy, & Vladlen Koltun. (2021). Vision Transformers for Dense Prediction.
- Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, & Christian Rupprecht. (2023). CoTracker: It is Better to Track Together.
- Jianyuan Wang, Nikita Karaev, Christian Rupprecht, & David Novotny. (2023). Visual Geometry Grounded Deep Structure From Motion.
- Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, & David Novotny. (2025). VGGT: Visual Geometry Grounded Transformer.
- Tong Qin, Peiliang Li, & Shaojie Shen (2017). VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator.
- Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, & Christian Rupprecht. (2025). On Geometric Understanding and Learned Priors in Feed-forward 3D Reconstruction Models.
- Lijie Wang, Lianjie Guo, Ziyi Xu, Qianhao Wang, Fei Gao, & Xieyuanli Chen. (2025). LiDAR-VGGT: Cross-Modal Coarse-to-Fine Fusion for Globally Consistent and Metric-Scale Dense Mapping.

References (2/2)

- Schönberger, J., & Frahm, J.M. (2016). Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR).*
- Schonberger, J., Zheng, E., Pollefeys, M., & Frahm, J.M. (2016). Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, & Piotr Bojanowski. (2023). DINOv2: Learning Robust Visual Features without Supervision.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, & Ilya Sutskever. (2021). Learning Transferable Visual Models From Natural Language Supervision.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, & Aaron Courville. (2017). FiLM: Visual Reasoning with a General Conditioning Layer.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, & Ross Girshick. (2023). Segment Anything.