CS7643: Group Project Proposal

Monocular Depth and Normal Estimation

Team Name: DepthSense

March 24, 2025

Project Summary

This project proposes developing a deep learning model for estimating depth and surface normals from monocular images, producing mutually-refined dense depth and normal maps for enhanced accuracy. The project's unique contribution lies in combining transformer-based models with synthetic-to-real transfer learning, explicitly targeting improved robustness and generalization.

Motivated by a shared interest in computer vision, the project addresses key challenges in 3D perception, self-supervised learning, and visual transformers, with broad applicability to tasks including object monitoring, visual odometry, and SLAM (Simultaneous Localization And Mapping). Ethical risks associated with this technology are minimal but include potential misuse in surveillance applications; clear guidelines and ethical constraints are recommended to mitigate these concerns.

Approach

The project primarily leverages DINOv2 [1] as the base ViT (Vision Transformer) model, while following and reimplementing the architecture defined in GeoNet [2], and the training setup described in Depth Anything v2 [3]. The implementation extends GeoNet by introducing specific architecture modifications and leveraging advanced training techniques from Depth Anything v2. This ensures a meaningful extension rather than simple reproduction. Synthetic-to-real transfer learning is explicitly employed to address domain gaps between synthetic training datasets and real-world evaluation benchmarks.

The synthetic and real-world datasets we are working with will provide an RGBD (red-green-blue-depth) image as a label. We will augment this label to additionally include a surface normal vector for each pixel. This can be computed from the depth and camera intrinsics, by first converting the pixels to a point cloud with the depth readings, and then fitting a plane against its neighbors for each point using the cross product of nearby points. This will extract the surface normal for us, which is then added to the dataset as an additional label feature. We will ignore camera extrinsics, opting to use the camera's frame for simplicity. We are hoping that making surface normal data available to the model will make it easier to train, and as a stretch goal, will improve benchmark accuracy on real world datasets.

Potential anticipated challenges include domain shifts between synthetic and real-world data, difficulties in convergence during self-supervised training, and computational limitations posed by transformer model complexity. These challenges will be addressed through iterative refinement of models, systematic hyperparameter tuning, dataset massaging, and leveraging high-quality synthetic-to-real transfer learning.

Resources and Related Work

Main influences include:

- DINOv2 [1], defining the choice of base vision model.
- GeoNet [2], formally defining the joint depth and normal estimation problem and pertinent loss functions.
- Depth Anything v2 [3], outlining the training setup for self-supervised training using synthetic images and unlabeled real-world images.

Additional references relevant for deeper understanding:

- Joint Prediction of Depths, Normals and Surface Curvature [4]
- Designing Deep Networks for Surface Normal Estimation [5]
- Joint Graph-based Depth Refinement and Normal Estimation [6]

Datasets

Experiments use high-quality synthetic images as primary training resources for computing depth and normal maps on the teacher model, then performing self-supervised learning with student models on unlabeled real-world images. The following datasets are specifically considered:

Synthetic

- Hypersim [7]: High-quality photorealistic synthetic indoor scenes.
- Virtual KTTI [8]: Synthetic outdoor driving scenes providing varied visual data.

Real World

- NYU Depth v2 [9]: Widely used real-world indoor depth benchmark dataset.
- DIODE [10]: Dataset with diverse real-world scenes for benchmarking depth estimation.

Team Members

- Lee, YoungJin <ylee904@gatech.edu>
- Morales Bacquerie, Luis Alejandro <lbacquerie3@gatech.edu>
- Yang, Albert H <ayang434@gatech.edu>

References

[1] DINOv2, `https://arxiv.org/pdf/2304.07193`

[2] GeoNet, `https://xjqi.github.io/geonet.pdf`

[3] Depth Anything v2, `https://arxiv.org/pdf/2406.09414`

[4] Joint Prediction of Depths, Normals and Surface Curvature, `https://arxiv.org/pdf/1706.07593`

[5] Designing Deep Networks for Surface Normal Estimation, `https://www.cs.cmu.edu/~xiaolonw/papers/deep3d.pdf`

[6] Joint Graph-based Depth Refinement and Normal Estimation, https://arxiv.org/pdf/1912.01306

[7] Hypersim Dataset, https://mikeroberts3000.github.io/papers/hypersim/

[8] Virtual KTTI Dataset, https://paperswithcode.com/dataset/virtual-kitti

[9] NYU Depth v2 Dataset, https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html

[10] DIODE Dataset, https://diode-dataset.org/