EE367 Project Proposal

Stereo Depth Estimation with Learned Regularizers

Jonathan Dumanski Joana Mizrahi

1 Motivation

Stereo depth estimation recovers a per-pixel depth map from a pair of rectified left and right images. The standard approach computes a matching cost $C(u, v, d)$ for every pixel (u, v) and candidate disparity d (essentially asking "how similar does this pixel look in the left vs right image if I assume depth d?"), and then selects the best disparity per pixel independently.

The problem is that this is ill-posed. For many pixels (textureless surfaces, reflective objects, occluded regions), many different disparity values explain the images equally well, leading to noisy and inconsistent depth maps. Classical methods like Semi-Global Matching (SGM) [1] handle this by adding a handcrafted smoothness penalty, but they still fail on the hard cases most common in autonomous driving.

In this project we treat stereo matching as an inverse problem, explicitly adding a prior over what valid depth maps look like, and ask whether replacing the hand-crafted smoothness prior with a learned one improves results, especially on those hard cases.

2 Related Work

Scharstein and Szeliski [2] provide a comprehensive taxonomy of classical stereo algorithms, establishing the standard pipeline and benchmarks still used today. SGM [1] remains a strong classical baseline, enforcing a smoothness constraint via dynamic programming.

Zbontar and LeCun [3] showed that replacing the handcrafted patch similarity in the cost function with a learned CNN significantly improves matching quality, while still relying on classical SGM post-processing. This shows that learned components can be cleanly inserted into the classical pipeline without replacing it entirely, which motivates our approach of replacing just the regularizer.

Venkatakrishnan et al. [4] introduced Plug-and-Play (PnP) priors: rather than specifying an explicit regularizer, you train a denoiser on examples of your signal of interest and use it as an implicit prior during iterative optimization. Applied to disparity maps, this means training a small network on real KITTI depth maps so it learns what realistic depth maps look like, then using it to regularize the stereo matching optimization.

3 Project Overview

We pose stereo matching as an optimization problem that finds the disparity map d that both explains the observed images well and looks like a realistic depth map:

$$d^* = \arg\min_d \sum_{u,v} C(u, v, d(u, v)) + \lambda R(d) \tag{1}$$

Here $C(u, v, d)$ is the matching cost (how dissimilar pixel (u, v) in the left image looks compared to pixel $(u - d, v)$ in the right image), computed using standard SSD patch matching. $R(d)$ is a regularizer that penalizes disparity maps that look unrealistic (e.g. abrupt random jumps in depth). C is fixed across all methods; $R(d)$ is what we vary. We compare three methods on the KITTI 2015 stereo benchmark [5]:

1. **Baseline (SGM):** OpenCV's Semi-Global Matching, a well-established classical method with a hand-crafted smoothness penalty.
2. **TV prior:** Iterative gradient descent with $R(d) = \|\nabla d\|_1$, a standard total variation smoothness prior that can be computed analytically.
3. **Learned prior:** Replace $R(d)$ with a small neural network trained on KITTI disparity maps. The network learns the gradient of $\log p(d)$. Optimization alternates between a gradient step on the data term and a step using this network, following the PnP framework [4].

This gives a clean ablation: same data term throughout, handcrafted prior vs learned prior.

Open Questions

We would welcome feedback from our mentor on the following:

- **Scope:** Is the regularizer comparison alone sufficient in scope, or would you recommend also varying the data term C (e.g. replacing SSD with a learned patch similarity as in Zbontar & LeCun [3])?
- **Learned prior architecture:** We plan to train a small denoiser network on KITTI disparity patches. Is a lightweight UNet reasonable, or would you suggest something simpler or different?
- **Uncertainty:** We are considering using variance across multiple optimization runs as a proxy for per-pixel depth uncertainty. Is this a reasonable approach?

Goals

- Implement and evaluate all three methods on KITTI 2015
- Show qualitative improvement of learned prior on failure mode examples (if there is improvement)
- Quantify uncertainty of our model if possible

4 Timeline

Days	Milestone
1–3	KITTI setup, SGM baseline, evaluation pipeline
4–6	TV prior optimization loop
4–6	Score network training on KITTI disparity patches
7–9	Learned prior integrated into optimization loop
10–11	Ablations, failure mode analysis, uncertainty quantification
12–14	Report, poster, cleanup

References

[1] H. Hirschmüller, "Stereo Processing by Semiglobal Matching and Mutual Information", *IEEE TPAMI*, 2008.

[2] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", *IJCV*, 2002.

[3] J. Zbontar and Y. LeCun, "Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches", *JMLR*, 2015.

[4] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-Play Priors for Model Based Reconstruction", *GlobalSIP*, 2013.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The KITTI Dataset", *IJRR*, 2013.