Stereo Depth Estimation with Learned Regularizers

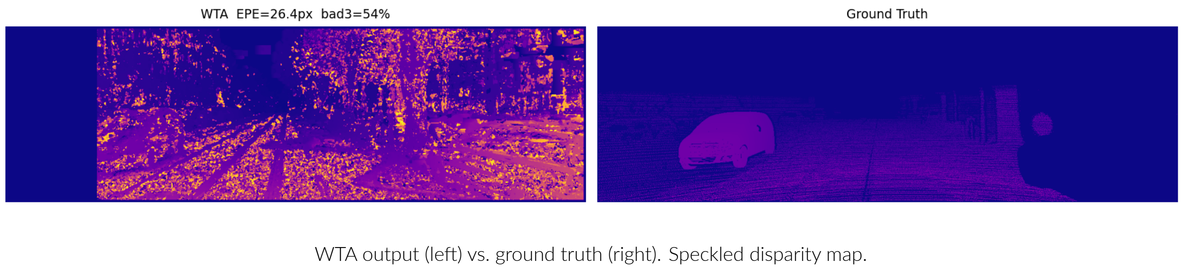
Jonathan Dumanski Joana Mizrahi

Stanford University — EE 367: Computational Imaging, Winter 2026

Motivation

Given a rectified stereo pair, recover a dense disparity map d. Naive Winner-Takes-All (WTA: minimum cost value selected at each pixel) is fast but noisy: ambiguous in textureless regions, occlusions, and reflective surfaces.

Input: rectified left/right image pair (KITTI 2015 [5]).



WTA output (left) vs. ground truth (right). Speckled disparity map.

Question: Can a *learned* prior over disparity maps beat handcrafted regularizers like SGM or TV?

Problem Formulation & Methods Compared

$$d^* = \arg\min_d \underbrace{\sum_{u,v} C(u,v,d(u,v))}_{\text{data term (fixed)}} + \lambda \underbrace{R(d)}_{\text{regularizer (varied)}}$$

Method	Data term	Regularizer	Optimizer
WTA	Census	none	argmin
TV-ADMM	Census	$\|\nabla d\|_1$	ADMM
HQS + Gauss	Census	Gaussian DnCNN	HQS
SGM (baseline)	Internal	DP smoothness	DP
HQS-WTA-L1 **(ours)**	Census	WTA-trained DnCNN	HQS
SGM+denoiser **(ours)**	Internal	SGM-trained DnCNN	post-proc.

Census Transform

Encodes each pixel by comparing its intensity to neighbors in a local window, resulting in a binary sequence. The similarity cost between 2 pixels is then calculated using the Hamming distance.

$$C(u,v,d) = \mathrm{Ham}\big(\mathrm{Cen}_L(u,v),\ \mathrm{Cen}_R(u-d,v)\big)$$

Since this metric encodes relative intensities, it is robust to illumination changes, which can be useful in situations where the brightness of objects is dependant on the viewing angle.

Related Work

SGM [1]: DP smoothness in 8 directions. Strong baseline; fails at fine boundaries.
Zbontar & LeCun [2]: CNN-learned matching cost; still uses SGM post-processing.
Plug-and-Play / HQS [3]: pre-trained denoiser as implicit prior.

Gap: existing PnP work uses Gaussian denoisers, mismatched to stereo noise structure.

References

[1] Hirschmüller, *IEEE TPAMI* 2008. [2] Zbontar & LeCun, *JMLR* 2015. [3] Venkatakrishnan et al., *GlobalSIP* 2013. [4] Zhang et al., vKITTI, *CVPR* 2016. [5] Geiger et al., KITTI, *IJRR* 2013.

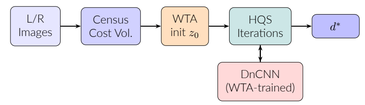
Method: HQS with Domain-Matched Denoiser

Half-Quadratic Splitting (HQS)

Introduce auxiliary variable z; alternate between data and prior:

$$d \leftarrow \arg\min_d C(d) + \tfrac{\mu}{2}\|d - z\|_2^2, \qquad z \leftarrow \mathcal{D}(d)$$

Unlike ADMM, no dual variable u: the denoiser always sees $d \in [0, D_{\max}]$ directly, staying in-distribution. ADMM accumulates u and diverges.



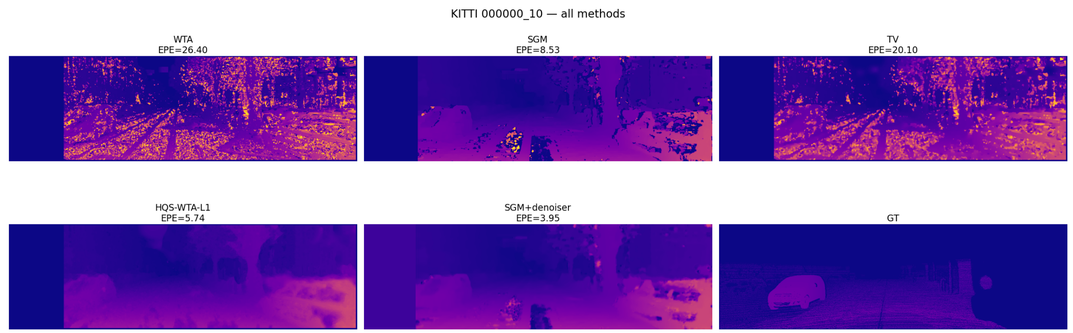
Domain Mismatch & Denoiser

Gaussian denoisers fail because WTA artifacts are structured, not Gaussian. We train **DnCNN** (7× Conv-BN-ReLU + linear output, 64 ch, L1 loss, no residual) on **(WTA → GT)** pairs from **vKITTI** [4] (synthetic driving, perfect GT depth), matching the inference distribution. Best val MAE: **3.08 px**.

Experimental Results

Evaluated on **KITTI 2015** [5] (real driving data, LiDAR ground truth), unseen during training. Avg. over 20 frames. EPE ↓; bad3 = % pixels with error > 3 px ↓.

Method	EPE↓	bad3%↓
WTA	20.19	44.2
TV-ADMM	14.58	41.9
HQS (Gaussian)	20.93	51.0
SGM (baseline)	7.63	22.6
HQS-WTA-L1 **(ours)**	6.79	23.2
SGM+denoiser **(ours)**	**4.19**	**21.0**



All methods on KITTI 000000_10. Top: WTA, SGM, TV. Bottom: HQS-WTA-L1, SGM+denoiser, GT.

HQS-WTA-L1 beats SGM on EPE (6.79 vs. 7.63) using a learned prior inside the optimization loop.

SGM+denoiser (bonus): training a separate denoiser on (SGM → GT) pairs and applying it as a one-shot post-processing step halves SGM error (EPE 4.19).