# Stereo Depth Estimation with Learned Regularizers

Jonathan Dumanski and Joana Mizrahi

**Abstract**—We study stereo depth estimation as an inverse problem: given a rectified image pair, recover a dense disparity map by minimizing a cost-volume data term plus a regularizer. We compare handcrafted regularizers (Total Variation via ADMM, Semi-Global Matching) against learned priors implemented via Half-Quadratic Splitting (HQS). A key finding is that the choice of denoiser training distribution is critical: a Gaussian denoiser fails at inference due to domain mismatch, while a DnCNN trained on (WTA $\rightarrow$ GT) disparity pairs from the synthetic vKITTI dataset generalizes well to real KITTI data. Our method, HQS-WTA-L1, achieves EPE 6.79 px on 20 KITTI 2015 frames, outperforming the SGM baseline (7.63 px) while matching its bad-pixel rate. As an additional experiment, a denoiser trained on SGM outputs and applied as a post-processing step reduces SGM error by 45% (EPE 4.19 px), demonstrating the broader applicability of domain-matched learned priors.

**Index Terms**—Stereo matching, disparity estimation, plug-and-play priors, half-quadratic splitting, learned regularizers.

## 1 Introduction

Stereo depth estimation recovers a dense disparity map $d$ from a rectified image pair. By epipolar geometry, matching pixels lie on the same horizontal scanline, reducing the search to a 1D problem per row. The simplest approach, Winner-Takes-All (WTA), independently selects the minimum-cost disparity at each pixel:

$$d_{\text{WTA}}(u, v) = \arg\min_{d} C(u, v, d),$$

where $C(u, v, d)$ is a cost volume measuring the dissimilarity between the left pixel at $(u, v)$ and the right pixel at $(u - d, v)$. WTA is fast and parallelizable, but it ignores spatial coherence entirely, producing noisy disparity maps with streaking artifacts, especially in textureless regions.

Regularization addresses this by incorporating a spatial prior into the estimation:

$$d^* = \arg\min_{d} \underbrace{\sum_{u,v} C(u, v, d(u, v))}_{\text{data term}} + \lambda \underbrace{R(d)}_{\text{regularizer}} . \quad (1)$$

Handcrafted regularizers such as Total Variation (TV) or the Semi-Global Matching (SGM) smoothness penalty [1] are computationally efficient and interpretable, but they impose fixed assumptions on disparity structure that may not reflect all real-world scenes.

The Plug-and-Play (PnP) framework [2] offers a principled alternative: any off-the-shelf denoiser $\mathcal{D}$ can serve as an implicit learned prior within an iterative solver. Prior work has used Gaussian denoisers for image reconstruction tasks, but applying PnP to stereo matching reveals a failure mode: a denoiser trained on Gaussian noise is poorly matched to the structured, non-Gaussian artifacts in WTA disparity maps. We show that simply training the denoiser on actual (WTA, GT) disparity pairs (domain matching) resolves this problem and yields better results.

Our main contributions are:

- A systematic comparison of WTA, TV-ADMM, HQS with a Gaussian denoiser, and HQS with a domain-matched denoiser on 20 KITTI 2015 frames.
- An analysis of why HQS is more stable than ADMM for this task, and how domain mismatch causes Gaussian PnP to fail catastrophically.
- Empirical evidence that our domain-matched HQS method, trained entirely on synthetic vKITTI data, generalizes to real KITTI and outperforms SGM on endpoint error (EPE 6.79 vs. 7.63 px).
- A bonus result: a denoiser trained on SGM outputs and applied as a post-processing step reduces SGM error by 45% (EPE 4.19 px).

## 2 Related Work

### 2.1 Classical Stereo Methods

Scharstein and Szeliski [3] provide a comprehensive taxonomy of stereo methods. Local methods aggregate costs within a support window; global methods minimize an energy functional (1) using graph cuts or belief propagation [4]. SGM [1] achieves a practical middle ground by running 1D dynamic programming (DP) along 8 directions and summing the path costs. Its P1/P2 penalty structure allows smooth disparity slopes cheaply while penalizing sudden jumps, effectively approximating 2D regularization. SGM remains a strong baseline and is still used as post-processing for modern learned matching costs [5].

### 2.2 Deep Learning for Stereo Matching

Zbontar and LeCun [5] replaced handcrafted matching costs with a siamese CNN but retained SGM for spatial regularization. End-to-end methods such as DispNet [6], PSMNet [7], and RAFT-Stereo [8] learn both matching and aggregation jointly, achieving state-of-the-art benchmark performance. These approaches can be difficult to interpret and typically require large paired training sets, making it harder to incorporate explicit priors or data terms.

- *J. Dumanski and J. Mizrahi are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305.*

### 2.3 Plug-and-Play Priors

Venkatakrishnan et al. [2] introduced PnP: within ADMM or proximal splitting, the proximal operator for the regularizer is replaced by a denoising step. Zhang et al. [9] proposed DnCNN, a batch-normalized CNN denoiser that has become a standard PnP building block for image reconstruction.

### 2.4 Disparity Post-Processing

Confidence-weighted filtering, hole-filling, and bilateral filtering are commonly applied after SGM. Learned refinement networks [10] have been proposed to correct residual errors. Our SGM+denoiser experiment is related but differs in using a DnCNN trained specifically on (SGM, GT) pairs from synthetic data.

## 3 Method

### 3.1 Census Cost Volume

We use the Census transform [11] as our matching cost. For each pixel $(u, v)$, a $7 \times 7$ binary descriptor encodes which neighbors have higher intensity than the center pixel. The matching cost between left pixel $(u, v)$ and right pixel $(u - d, v)$ is the Hamming distance between their Census descriptors, normalized to $[0, 1]$ by dividing by 49. An $11 \times 11$ sum-pooling window aggregates the raw costs before optimization. We compute costs for $d \in \{0, 1, \ldots, 191\}$ with $D_{\max} = 192$.

### 3.2 TV-ADMM

Total Variation promotes piecewise-smooth disparity maps:

$$d^* = \arg\min_d \sum_{u,v} C(u, v, d(u, v)) + \lambda \|\nabla d\|_1. \tag{2}$$

We solve this via ADMM with splitting variable $z = d$, enforcing $d = z$ as a constraint so that the two terms decouple. The $d$-subproblem is separable per pixel: for each pixel we select the disparity minimizing the census cost plus a quadratic penalty toward $z$, which is a simple lookup over the cost volume. The $z$-subproblem is the proximal operator of $\lambda\|\nabla \cdot\|_1$:

$$z \leftarrow \arg\min_z \ \lambda \|\nabla z\|_1 + \tfrac{\mu}{2}\|z - (d + u)\|^2, \tag{3}$$

which we solve with Chambolle's algorithm [12]. The dual variable $u$ is updated as $u \leftarrow u + d - z$ at each iteration.

### 3.3 Half-Quadratic Splitting (HQS)

HQS introduces auxiliary variable $z$ to decouple the data term from the prior:

$$d \leftarrow \arg\min_d \sum_{u,v} C(u, v, d(u, v)) + \tfrac{\mu}{2}\|d - z\|^2 \tag{4}$$

$$z \leftarrow \mathcal{D}(d). \tag{5}$$

The $d$-update (4) decouples per pixel: for each $(u, v)$, we solve $\min_d C(u, v, d) + \frac{\mu}{2}(d - z_{u,v})^2$. The $z$-update (5) applies the denoiser to the current disparity map.

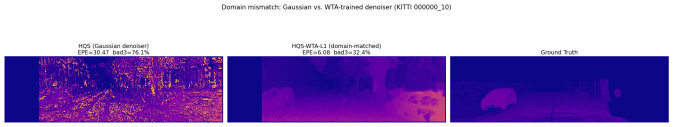


Fig. 1. Domain mismatch illustrated on KITTI frame 000000_10. A Gaussian-trained denoiser (left) degrades the disparity map relative to WTA, while the domain-matched WTA-trained denoiser (center) produces a substantially cleaner result. Right: ground truth.

### 3.4 HQS vs. ADMM.

In ADMM, a dual variable $u$ accumulates residuals: $u \leftarrow u + d - z$. The denoiser input becomes $d + u$, which can grow outside $[0, D_{\max}]$, potentially taking the denoiser out of its training distribution. HQS has no dual variable: the denoiser always receives $d \in [0, D_{\max}]$ directly, keeping it in-distribution. We observe that ADMM with a Gaussian DnCNN diverges in practice; HQS remains stable across all 20 iterations we use. We attribute this to the distribution shift introduced by the dual variable accumulation, though a full theoretical analysis is left to future work.

### 3.5 DnCNN Denoiser Architecture

We use DnCNN [9]: 8 convolutional layers with 64 channels, batch normalization after each intermediate layer, ReLU activations, and a final linear convolutional layer. The output predicts the clean disparity map directly (not the noise residual). We train with L1 loss rather than MSE, as L1 encourages sharper predictions at depth boundaries, which benefits the bad3 metric.

#### *3.5.1 Domain Mismatch*

A DnCNN trained on clean disparities with additive Gaussian noise ($\sigma \sim \mathcal{U}[2, 25]$ px) yields EPE 20.93 on KITTI: *worse* than WTA (EPE 20.19). We attribute this to domain mismatch: WTA artifacts are structured (cost-volume streaks, textureless-region ambiguities, occlusion errors) and have little resemblance to Gaussian noise. A denoiser trained on Gaussian noise may therefore lack a meaningful prior over disparity maps, causing it to degrade rather than improve the estimate (see Figure 1).

#### *3.5.2 Domain-Matched Training*

We train the denoiser on (WTA disparity, GT disparity) pairs from **vKITTI 2.0.3** [13], which is a photorealistic synthetic driving dataset with perfect ground-truth depth. Figure 2 shows example training pairs. WTA disparities are computed from vKITTI's stereo pairs using our Census cost volume, yielding 1'160 stereo pairs. Ground-truth depth maps are converted to disparity via $d = fB/z$ with $fB = 386.2$ px·m. We sample random $64 \times 64$ patches from regions with $> 70\%$ valid ground truth.

**Training details:** 60 epochs, Adam optimizer, learning rate $10^{-3}$ with cosine annealing, batch size 32. Best validation MAE: **3.08 px**.

### 3.6 SGM+Denoiser

As a separate bonus experiment, we train a second DnCNN on (SGM output, GT) pairs from the same vKITTI data.

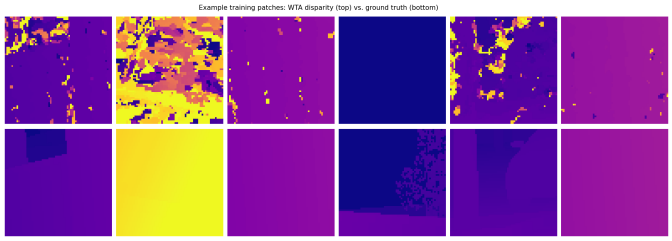


Fig. 2. Example training pairs: WTA disparity (top) and ground-truth disparity (bottom) for random $64 \times 64$ patches from vKITTI. The structured nature of WTA artifacts (speckle, streaking, textureless-region errors) is clearly different from Gaussian noise.


Fig. 3. Top: KITTI 2015 stereo pair with sparse LiDAR ground truth (evaluation). Bottom: vKITTI 2.0 stereo pair with dense synthetic ground truth (training). Note the density difference: KITTI GT is sparse (LiDAR points only) while vKITTI GT is fully dense, enabling richer supervision.

SGM disparities are computed using OpenCV `StereoSGBM` on each vKITTI stereo pair. This denoiser is applied as a single forward pass on the output of SGM (no iterative optimization is used). Best validation MAE: **3.01 px**.

### 3.7 SGM Baseline

We use OpenCV `StereoSGBM_create` with $D_{\max} = 192$, block size 11, P1 $= 8 \times 11^2 = 968$, P2 $= 32 \times 11^2 = 3872$, 3-way mode. Note that SGM uses its own SAD-based matching cost, not our Census cost volume. This means Census-based methods (WTA, TV, HQS) and SGM are not fully comparable; SGM's internal cost may be more informative in textureless regions.

## 4 Analysis & Evaluation

### 4.1 Experimental Setup

All learned denoisers are trained on vKITTI 2.0.3. Evaluation is on **20 frames from KITTI 2015** [14]: real driving data with LiDAR ground truth, entirely unseen during training. Figure 3 shows representative stereo pairs and ground truth from both datasets.

Metrics:

- **EPE**: mean absolute disparity error (px) over all valid pixels.
- **bad3**: percentage of valid pixels with error $> 3$ px.

HQS hyperparameter $\mu = 0.01$ (selected by grid search on 5 held-out frames; 20 iterations; initialized with $z_0 = d_{\text{WTA}}$).

### 4.2 Quantitative Comparison

Table 1 reports results averaged over 20 KITTI 2015 frames.

Several findings stand out. First, the Gaussian HQS denoiser (EPE 20.93) performs *worse* than WTA (EPE 20.19), consistent with the domain mismatch hypothesis. Second, TV-ADMM (EPE 14.58) improves substantially over WTA but remains well below SGM. Third, **HQS-WTA-L1 (EPE 6.79) outperforms SGM (EPE 7.63)** by 11% in EPE while achieving a comparable bad3 rate (23.2% vs. 22.6%). This is notable given that HQS-WTA-L1 uses a Census cost volume whereas SGM uses its own internal SAD-based cost. Fourth, SGM+denoiser (EPE 4.19, bad3 21.0%) achieves the best results overall, reducing SGM's EPE by 45% with a single forward pass.

TABLE 1
Results on 20 KITTI 2015 frames. All methods except SGM use the Census cost volume.

| Method | EPE (px) ↓ | bad3 (%) ↓ |
|---|---|---|
| WTA | 20.19 | 44.2 |
| TV-ADMM | 14.58 | 41.9 |
| HQS (Gaussian denoiser) | 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** |

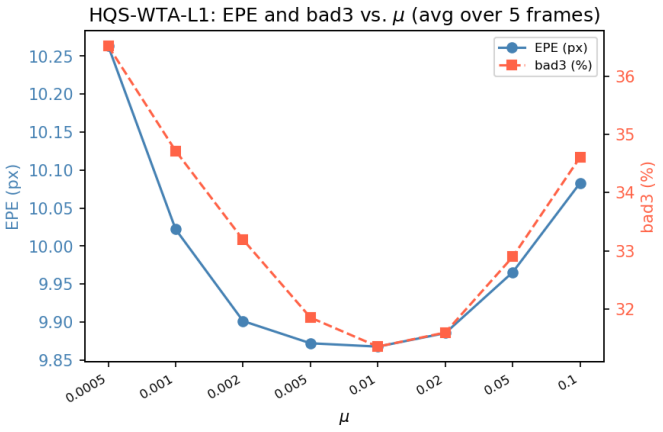


Fig. 4. EPE and bad3 vs. $\mu$ for HQS-WTA-L1, averaged over 5 KITTI frames. Both metrics are minimized near $\mu = 0.01$.

### 4.3 Ablation: L1 vs. MSE Loss

On frame 000000_10, L1 loss (EPE 6.27 px, bad3 33.8%) outperforms MSE (EPE 6.48 px, bad3 34.6%). L1 avoids over-smoothing at depth boundaries by not squaring large residuals, which is especially important for the bad3 metric.

### 4.4 Ablation: HQS Hyperparameter $\mu$

A grid search over $\mu \in \{0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1\}$ on 5 frames shows that $\mu = 0.01$ is optimal. Larger $\mu$ gives the denoiser too much weight, leading to over-smoothing; smaller $\mu$ under-regularizes. Figure 4 shows EPE and bad3 as a function of $\mu$.

## 5 Results

### 5.1 Qualitative Comparison

Figure 5 shows disparity maps for selected methods on KITTI frame 000000_10. WTA produces a speckled, spa-

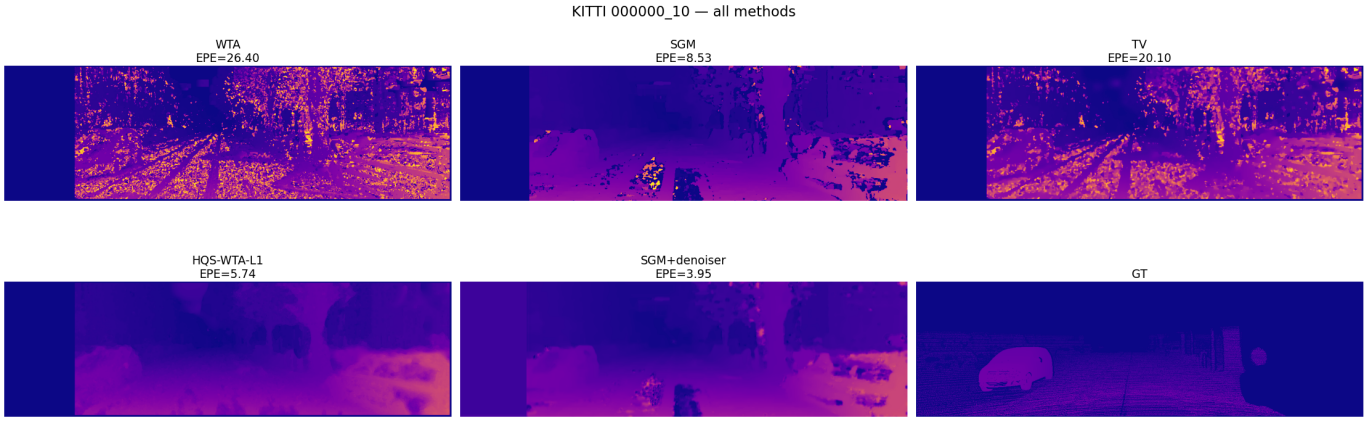


Fig. 5. Selected methods on KITTI 2015 frame 000000_10. Top row (left to right): WTA, SGM, TV-ADMM. Bottom row: HQS-WTA-L1 (ours), SGM+denoiser (ours), ground truth. Color encodes disparity (yellow = close, purple = far).

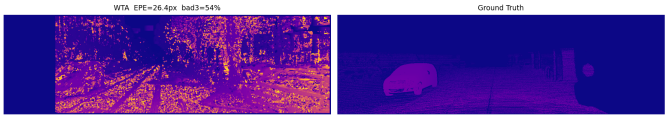


Fig. 6. WTA disparity (left) vs. ground truth (right) on KITTI 000000_10. Speckled artifacts are omnipresent.

tially incoherent map (see also Figure 6). TV-ADMM reduces speckle noise but artifacts still remain. HQS-WTA-L1 produces smoother, more coherent disparity than SGM. SGM+denoiser achieves a clean result, smoothing the raw SGM output.

## 6 Discussion, Limitations, Future Work, and Conclusion

### 6.1 Domain Matching

The most important takeaway from our experiments is that PnP denoiser performance depends on training distribution alignment. The Gaussian PnP result (EPE 20.93) fails not because the framework is flawed, but because the denoiser has no prior relevant to stereo disparity maps. Training on (WTA, GT) pairs encodes an accurate prior that generalizes from synthetic vKITTI to real KITTI data.

### 6.2 HQS Stability

The absence of a dual variable in HQS is a practical necessity when using task-specific denoisers. ADMM's dual variable potentially pushes inputs outside the denoiser's training range, causing divergence.

### 6.3 SGM+Denoiser Interpretation

The large improvement from SGM+denoiser (7.63 $\rightarrow$ 4.19 EPE) suggests that SGM makes systematic, predictable errors that a DnCNN can efficiently learn to correct. The simplicity of this approach (train once on synthetic data, apply as post-processing), makes it attractive for practical deployment.

### 6.4 Limitations

(1) Both denoisers are trained on synthetic data; a larger domain gap for diverse real-world conditions may limit generalization. (2) HQS adds 20 network forward passes at inference.

### 6.5 Future Work

(1) Apply to other cost functions: Census is robust but not the best possible cost; pairing our HQS framework with a learned matching cost could push results further. (2) Confidence weighted HQS that trusts the data term more at low uncertainty pixels. (3) Video/temporal consistency: KITTI is a driving sequence so enforcing temporal smoothness across frames could improve results.

### 6.6 Conclusion

We have shown that domain-matched learned priors can be used to outperform hand-crafted regularizers for PnP stereo matching. A DnCNN trained on (WTA, GT) pairs from synthetic vKITTI, used as the prior in HQS, beats SGM on endpoint error while generalizing to unseen real KITTI data. An additional SGM post-processing denoiser halves SGM's error with a single forward pass, demonstrating that domain-specific learned priors offer a highly practical path to improving classical stereo pipelines.

## References

[1] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[2] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," 2013.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2002.

[4] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," vol. 25, no. 7, 2003, pp. 787–800.

[5] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.

[6] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[8] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *International Conference on 3D Vision (3DV)*, 2021.
[9] K. Zhang, W. Zuo, Y. Chen, M. Devaney, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
[10] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," in *Asian Conference on Computer Vision (ACCV)*, 2018.
[11] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference on Computer Vision (ECCV)*, 1994.
[12] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1–2, pp. 89–97, 2004.
[13] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," 2016.
[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.