A Data-driven Regularization Model for Stereo and Flow

Donglai Wei
MIT
donglai@csail.mit.edu

Ce Liu
Microsoft Research
celiu@microsoft.com

William T. Freeman
MIT
billf@mit.edu

Abstract

Data-driven techniques can reliably build semantic correspondence among images [16, 15]. In this paper, we present a new regularization model for stereo or flow through transferring the shape information of the disparity or flow from semantically matched patches in the training database. Compared to previous regularization models based on image appearance alone, we can better resolve local ambiguity of the disparity or flow by considering the semantic information without explicit object modeling [1, 5].

We incorporate this data-driven regularization model into a standard Markov Random Field (MRF) model, inferred with a gradient descent algorithm [14] and learned with a discriminative learning approach [19]. Compared to prior state-of-the-art methods, our full model achieves comparable or better results on the KITTI stereo and flow datasets [7], and improves results on the Sintel Flow dataset [4] under an online estimation setting.

1. Introduction

Stereoscopic vision and dense motion estimation are crucial in 3D reconstruction. In order to regularize local matching evidence, Markov Random Field (MRF) models have been widely used to enforce *smoothness constraints* on scene properties (disparity or flow). Most regularization models [21, 23, 31] assume that the discontinuities of scene properties coincide with the discontinuities of the image appearance, which works well for examples in the lab settings (Figure 1a-c). However, in real-world scenarios where the local matching evidence is insufficient, these models suffer from the ambiguous correlation between appearance and scene properties (Figure 1d-f), since regions with the same appearance can have different scene properties.

Our first key idea is to infer shape information from contextually matched regions in a database. As more labeled data become available, data-driven approaches, for example, by means of dense correspondences [16], can transfer information such as semantic labels [15] and high-resolution images [25] from an existing database to the query image. Here, we want to transfer shape information. For example, a less textured patch can have a similar disparity along the horizontal direction if from a road, or the vertical direction if from a wall. By checking if it is consistently

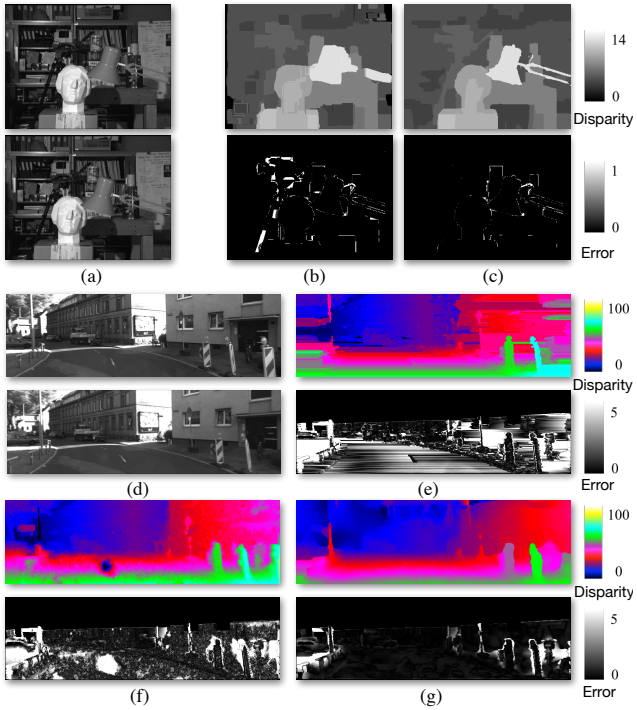


Figure 1: Motivation for a data-driven regularization model. In a lab setting, e.g., (a) stereo images from Middlebury dataset [20], (b) a patch-based matching model [3] performs well and achieves a state-of-the-art result with (c) an additional bilateral regularization model [21] by refining the disparity boundary. In a real-world scenario, e.g.e (d) stereo images from KITTI dataset [7], neither (e) the matching model nor (f) the regularization model above performs well. We introduce (g) a data-driven regularization model to make use of the semantic parsing of the scene, which transfers disparity information from regions with similar context in the training data. For the disparity (top row) and error map (bottom row) in (b-c, e-g), we use the default color code from the corresponding dataset.

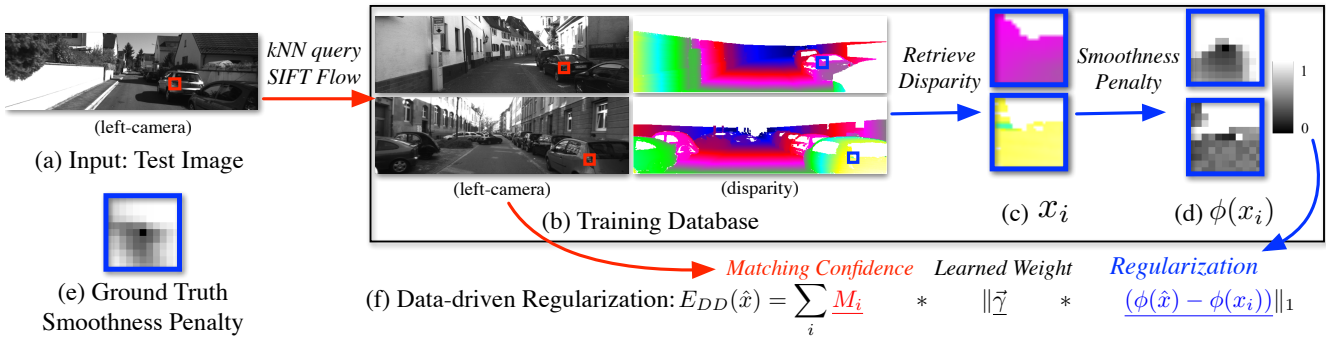


Figure 2: Illustration of our data-driven regularization model (stereo example). For (a) a test left-camera image, we first retrieve similar images with their underlying disparity based on the GIST descriptor from (b) the training database. For each test image patch (red) from (a), we find its matched training image patches through the SIFT flow algorithm. Given (c) the disparity of these matched patches, we apply a nonlocal smoothness penalty function (i.e. subtracting the disparity in the center) and (d) the results for the training patches are similar to (e) that for the ground truth test patch. Our data-driven regularization penalizes the difference between these smoothness penalty results, weighted by the matching confidence and the pixel position within the patch.

matched with patches from walls or roads in the database, we obtain stronger shape prior for this uninformative patch.

Our second key idea is to represent the shape information as the *relative* relationship of scene properties instead of their absolute values. Most of the existing data-driven applications transfer the *data-term* of the MRF model, namely transferring the exact values of the scene from similar patches in similar images. Such a *data-term transfer* approach seems to work well for problems not constrained by data, such as image hallucination [25] and depth synthesis [12]. However, for accuracy-demanding problems such as stereo and flow, the data-term transfer approach greatly limits the re-usability of scene properties in the training data. For example, if we want to regularize the disparity value of a car, the data-term transfer approach requires matched cars to have similar positions, while the relative-relationship transfer approach only requires similar local shape.

In this paper, we make two main technical contributions.

MRF Model We build a joint model for both the test and training examples to incorporate the data-driven module.

Data-driven Method We transfer the relative relationship of scene properties instead of the data value, which is applicable to accuracy-demanding tasks like stereo and flow.

2. Related Work

Markov Random Field Model for Visual Reconstruction Markov random fields (MRF) have been widely used in low-level vision for visual reconstruction, such as image denoising stereo, optical flow, etc. As the local information from the observation can be noisy or ambiguous, the MRF model imposes regularization to produce a spatially smooth estimate for the scene property. Filter-based penalty functions [10, 13, 27, 23] and subspace-based penalty functions [11, 8] are often used. To adaptively model the correlation between the observed image and the scene property for estimation, new regularization functions, like weighted median filters [22], bilateral filters [21], and linear regression [31], have been proposed.

Data-driven Techniques for Non-parametric Modeling In low-level vision tasks, scene properties, like image appearance [6, 9, 24, 25], depth [12, 2], and motion [6], have been transferred as different proposals for local evidence from the matched patches. In this paper, we use the SIFT flow algorithm to find training patches with similar context with each test patch to build a "semantic prior" that is modeled explicitly in [1, 5] for visual reconstruction.

3. Pipeline

Our pipeline to build the data-driven regularization model is illustrated in Figure 2. Given an image patch from the test example, we first retrieve patches in the training database that have similar semantic information (Step 1), and then extract the shape information from their disparity values to regularize the estimate (Step 2). We incorporate this regularization model into a traditional stereo/flow model and produce an estimate based on the full model.

Step 1: Retrieve Semantic-Similar Patches For each test left-camera image in Figure 2a, we first use the GIST descriptor to retrieve training images and their disparity with similar scene structures, as shown in Figure 2b. We apply SIFT flow to align the test image with retrieved images and find contextual matches (shown as red boxes) for the patch from the test image.

Step 2: Regularize Shape Information Given the disparity of these matched patches (Figure 2c), we represent the local shape information as the response from a nonlocal

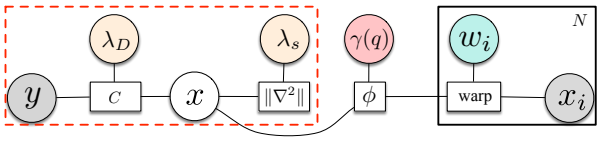


Figure 3: Graphical representation of our stereo/flow model. Inside the red box (dash line boundary), we show a traditional model which consists of the data term $E_D(x, y)$ and smoothness term $E_S(x)$. We introduce a data-driven regularization term $E_{DD}(x, \{x_i, w_i\})$ which encourages the output of penalty function ϕ for the test scene property x to be similar to that for the training scene property x_i warped by w_i.

smoothness penalty function, which subtracts the disparity at the center pixel. The results for matched training patches (Figure 2d) are similar to that for the ground truth disparity of the test patch (Figure 2e). Although the local appearance of these patches is different, their semantic labels as being part of a car suggest them to have similar local shape information while their absolute disparity values differ.

4. MRF with Data-driven Regularization

We here build a fully generative MRF model for each test example and all training examples jointly, which incorporates naturally the data-driven regularization. For a test example, let y denote its observation (such as stereo pairs or adjacent frames) and x denote its underlying scene (such as disparity or flow). Traditional stereo and flow models aim to minimize

$$\begin{aligned} -\log P(x|y) &\propto -\log\Big\{P(y|x)P(x))\Big\} \\ &\propto -\log P(y|x) - \log P(x) \\ &= E_D(x, y) + E_S(x), \end{aligned}$$

where E_D and E_S are called the data term and smoothness term respectively.

Here, we assume that the training database has k scene properties $\{x_i\}_{i=1}^k$ which are "generated" from x through the dense correspondence field $\{w_i\}_{i=1}^k$. Conditioning on these training information $z = \{x_i, w_i\}_{i=1}^k$, the objective function becomes

$$\begin{aligned} -\log P(x|y, \boldsymbol{z}) &\propto -\log\Big\{P(y|x, z)P(z|x)P(x)\Big\} \\ &\propto -\log\Big\{P(y|x)P(x)\Big\} - \log P(z|x) \\ &= E_D(x, y) + E_S(x) + E_{DD}(x, \{x_i, w_i\}_{i=1}^k), \end{aligned}$$

where E_{DD} is the new data-driven smoothness term regularizing x to have the similar smoothness property to scenes from the training database. Note that our new data-driven regularization term can be added to any traditional formulation of MRF model.

4.1. Data-driven Regularization Term

Our first step is to regularize the smoothness property of x to be similar to that of the matched regions in x_i. We represent the local smoothness property with a weighted non-local penalty function

$$\phi_p(t, q) = \gamma(q)\Big(t(p+q) - t(p)\Big), \tag{1}$$

where $\gamma(q)$ is the weight based on the relative position q and is set to 0 for $p+q \notin \Omega_p$. Given the correspondence field w_i from the test scene property x to a training scene property x_i, every pixel p in x is matched to the pixel $p+w_i(p)$ in x_i. Thus, at each pixel x, we want to minimize the difference between its penalty response and that from $p+w_i(p)$ within the local patches

$$\sum_q \|\phi_p(x, q) - \phi_{p+w_i(p)}(x_i, q)\|_1 \tag{2}$$

The second step is to adaptively weigh the regularization term at each pixel based on its matching quality. We define the matching quality with training example $\{x_i, w_i\}$ at pixel p as

$$M_i(p) = \exp\{-m(w_i(p))\} \tag{3}$$

where $m(w_i(p))$ is the pre-computed matching cost between corresponding patches in SIFT flow. See the supplementary material for the justification for such metric through the context matching accuracy. For matched regions with high $M_i(p)$, which tend to have similar semantic context [16], we increase the regularizaiton weight for their smoothness property difference, and vice versa. If there are few good matches found for a test example, then our model will fall back to the baseline model, as little shape information can be transferred from the database.

Combining Eq (2,3), we have the data-driven smoothness term for a set of training examples $\{x_i, w_i\}_{i\in s_i}$

$$\begin{aligned} E_{DD}(x,&\{x_i, w_i\}_{i\in s_i}) = \sum_i E_{DD}(x, x_i, w_i) \\ &= \sum_i \sum_p M_i(p) \sum_q \gamma(q) \| \Big(x(p+q) - x(p)\Big) \\ &\quad - \Big(x_i(p + w_i(p) + q) - x_i(p + w_i(p))\Big)\Big\|_1 \end{aligned} \tag{4}$$

4.2. Traditional Data Term and Smoothness Term

As a baseline, we choose a simple formulation for the data term and smoothness term in our system.

Data Term $y = \{y^a, y^b\}$ denotes the two input images (left and right images for stereo, adjacent frames for flow). At each pixel p, we define $C(p, x(p))$ as the matching cost for the disparity/flow value $x(p)$ with the Centralized-Sum-of-Absolute-Difference (CSAD) metric recommended in [26]

$$C(p, x(p)) \triangleq \sum_{q \in \Omega_p} \| \left(y^a(q) - y^a(p)\right) - \left(y^b(q + x(p)) - y^b(p + x(p))\right) \|_1, \quad (5)$$

where Ω_p is the local patch around p. We choose the continuous MRF formulation and approximate the matching cost with a quadratic function centered at the initial estimate $x'(p)$

$$C(p, x(p)) \approx \lambda_D(p) \|x - x'(p)\|_2,$$

where $\lambda_D(p)$ is fitted from Eq (5). Our MRF data term can be written as the sum of the matching costs at all pixels:

$$E_D(x, y) = \sum_p \lambda_D(p) \|x - x'(p)\|_2. \quad (6)$$

Smoothness Term We define a smoothness term to regularize the second-order gradient of the scene property x as

$$E_S(x) = \lambda_S \sum_p \|x(p-1) - 2x(p) + x(p+1)\|_1, \quad (7)$$

where λ_S is the weight parameter and L1 norm is used for robustness. Similar to that in [27], this smoothness term encourages disparity or flow to be piecewise planar.

5. Inference and Learning

Our model is a standard high-order continuous MRF and we use gradient descent algorithms for inference and learning. Combining Eq (4,6,7), we have the full energy model

$$E(x) = E_D(x, y) + E_S(x) + E_{DD}(x, \{x_i, w_i\}_{i \in s_i}), \quad (8)$$

where $t_{i,pq} = x_i(p + w_i(p) + q) - x_i(p + w_i(p))$ is the smoothness penalty value at position (p, q) for the warped training example x_i

5.1. Inference

We make two approximation to the data-driven regularization model for efficiency, which leads to the pipeline steps described in Section 3.

Relevant Subset of Training Examples Given a test image, many training images have few good matches and consequently small matching confidence M and small contribution to the Data-driven regularization term E_{DD}. We use GIST descriptor [17] to find K nearest training examples.

Dense Appearance Correspondence Instead of jointly estimating the dense scene correspondence $\{w_i\}$ and scene property $\{x_i\}$, we approximate $\{w_i\}$ by the dense appearance correspondence with the SIFT flow algorithm [16].

We infer the full MRF model according to Eq (8) with the standard gradient descent method. As the object function is nonlinear, we use the iterative fixed point method to find an incremental displacement $\Delta x^{(k)}$ at iteration k that satisfies

$$0 = \frac{\partial}{\partial \Delta x^{(k)}} E(x^{(k)} + \Delta x^{(k)}, y)$$

See the supplementary material for more details. In practice, we start from an initial estimate of the scene property, which comes from either the coarse-to-fine optimization scheme of our baseline model or external algorithms.

5.2. Learning

The full energy function Eq (8) has two sets of weight parameters: λ_S for the smoothness term and $\{\gamma(q)\}$ for the data-driven regularization. We do a grid search for λ_S and below we describe the discriminative learning approach [19] to optimize $\{\gamma_q\}$.

One standard evaluation metric for stereo and flow is 0-1 penalty, which is not differentiable. We approximate it an exponential function $\rho_L = 1 - \exp(-(t/2)^2)$ to penalize the difference between the ground truth disparity x_{gt} and the MAP estimate x^*. Below is the objective function for our MRF learning to optimize w.r.t. parameter γ

$$L(x^*(\gamma), x_{gt}) = \rho_L(x^*(\gamma) - x_{gt}).$$

We use a steepest-gradient descent optimization method. See the supplementary material for the details of calculating the gradient $\frac{\partial L}{\partial \gamma}|_{x^*}$ and the parameter learning results.

6. Experiments on Stereo

In this section, we first evaluate our proposed model on the recent KITTI stereo benchmark [7], where our performance is comparable to the state-of-the-art methods (only using the stereo pair images). Then, with the ground truth disparity from the training dataset, we further analyze the importance of the matching quality and transferred regularization for our model.

The stereo experiments below use the following system configuration. For the data-driven pipeline, we use 11-NN and GIST features for image retrieval, and SIFT flow for scene matching with 7× 7 patch size. See the supplementary material for the details of parameter selection from the validation data. The runtime for our entire system is 1 min per example, where most of the computation time is spent on the data-driven module: 0.1 s for image retrieval and 50 s for SIFT flow.

6.1. Benchmark Results on KITTI Stereo Dataset

The KITTI stereo dataset provides 194 pairs of stereo images for training and 195 pairs for testing. Recorded with a calibrated stereo rig on a car driving around the city, the KITTI stereo dataset has rich scene structures from rural areas to downtown regions.

Quantitative Result The best available initialization on the benchmark is from StereoSLIC [29], which is used by the state-of-the-art PCBP-SS algorithm [28]. For a fair comparison, we use the same initialization and our entry name

	> 2 pixels		> 3 pixels		> 4 pixels		> 5 pixels	
	Non-Out	All	Non-Out	All	Non-Out	All	Non-Out	All
DDS-SS (Ours)	5.91 %	6.96 %	3.83 %	**4.59 %**	2.90 %	**3.49 %**	2.36 %	**2.83 %**
PCBP-SS [28]	**5.19 %**	**6.75 %**	**3.40 %**	4.72 %	**2.62 %**	3.75 %	**2.18 %**	3.15 %
StereoSLIC [29]	5.76 %	7.20 %	3.92 %	5.11 %	3.04 %	4.04 %	2.49 %	3.33 %

Table 1: Test result on the KITTI stereo benchmark. Ours is comparable with the state-of-the-art algorithm [28] and the smallest error rate is marked in bold. "All" is the evaluation on the entire region, while "Non-Out" is the evaluation on the non-occluded region only.

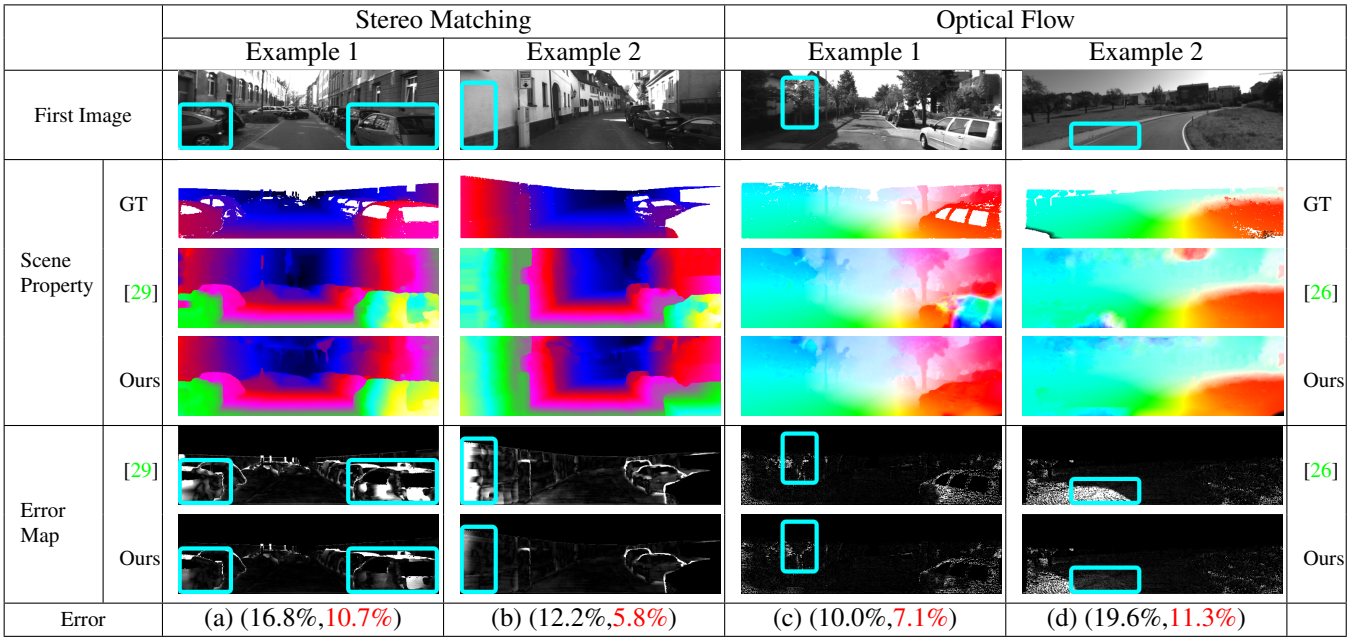


Figure 4: Visual comparison on examples from the KITTI stereo and flow datasets. We initialize our model with [29] (rank 3^{rd}) for stereo and [26] (rank 4^{th}) for optical flow. For each example, we show (top row) its left-camera image, (middle row) color-coded disparity map and (bottom row) error map scaled between 0 (black) and 5 (white). We mark regions with big improvement with cyan rectangles. Note that those regions often correspond to semantic objects like cars, buildings, trees, and roads, where our data-driven regularizer can use the good matches from training examples to make better guesses.

is "DDS-SS" on the benchmark website. Shown in Table 1, our data-driven regularization model significantly improves upon the initialization, by around 10%, and obtains comparable results to that from the slanted-plane MRF regularization model in [28]. In general, PCBP-SS performs better in non-occluded regions while ours improves more in occluded regions. With the metric using a 4 pixel error threshold, our proposed method outperforms [28] by 10% when evaluated on the entire region, and is worse by 10% when evaluated on the non-occluded region.

Qualitative Results We visualize results from the KITTI stereo training dataset, to empirically understand when our system works or fails.

(1) *Success Cases* The images from KITTI benchmark are taken from a moving car and the scene structures are similar to that in the LabelMe Outdoor dataset [16], where the SIFT flow algorithm has been proved reliable for findng contextual matches. In the left two columns in Figure 4, we show two stereo examples. Our model improves the disparity estimate for regions of cars and walls, which are matched well from the training database.

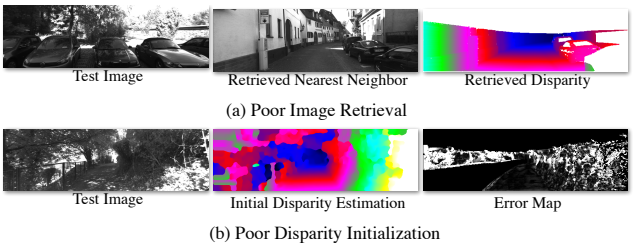
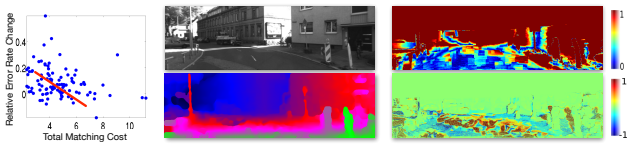


Figure 5: Two failure cases of little improvement on KITTI stereo training examples. For test images (left), (a) GIST-based retrieval finds images (middle) with dissimilar scene structures (right); (b) the initial disparity estimate (middle) has large error (right).

(2) *Failure Cases* Figure 5, shows examples in two typical scenarios, where little improvement is obtained upon



(a) Correlation Plot (b) Test Image and Initial Disparity (c) Matching Cost and Reduced Error

Figure 6: The effect of the matching quality. On the KITTI stereo training data, we show (a) the anti-correlation between the matching quality and our relative ratio of the improvement. We show (b) a left-camera image (top) with its initial disparity estimate (bottom). The (c) well-matched regions (blue) in the matching cost map (top) correspond to regions with improved disparity estimate (red) in the map of reduced disparity error (bottom).

the initialization. In the first scenario shown in Figure 4a, for a test image (left), the retrieved training images (middle) based on the GIST descriptor have dissimilar disparity structures (right). With the high SIFT flow matching cost, the data-driven regularization terms have negligible weight in the MRF model according to Eq (4), which leads to little change from the initialzation. In the second scenario, the initial disparity estimate can have too large an error to be corrected. In Figure 4b, for a left-camera image (left), the initial disparity estimate (middle) is off as a whole for the tree regions, seen from the error map (right). Even with the ground truth shape information, we cannot improve the result due to the big offset from its initial error.

6.2. Breakdown System Analysis

Given the ground truth disparity from the KITTI stereo training data, we perform a breakdown analysis to understand the importance of the (1) matching results of the data-driven technique, and (2) data-driven regularization model in our stereo system. In practice, we randomly split the stereo training data into a training set and a test set with 1:1 ratio.

Correlation with Matching Quality In Eq (3), the SIFT matching cost for a test patch centered at pixel p is denoted as $m_i(p)$. We here evaluate the matching result of a test example with the total matching cost $\sum_p \min_i\{m_i(p)\}$, the sum of the smallest matching cost for each patch.

In Figure 5a, for all examples in the test set, we plot the anti-correlation between the relative error change from the data-driven regularization against the total matching cost. As expected, if a test example retrieves better matches, then it has bigger improvement. We visualize such correlation qualitatively on a test example, shown in Figure 5b-c. In Figure 5b, we show a test image (top) and its initial disparity map (bottom). For each patch from the test image, we show the map of its matching cost in the top row in Figure 5c, within $[0, 1]$ shown in the top of Figure 5c. where regions of roads and buildings are well matched. In the bottom row in Figure 5c, we show the map of reduced disparity error, where positive values (red color) overlaps substantially with regions with the smaller matching cost.

Smoothness Term	Initial	∇	∇^2	$\mathcal{B}$	Ours	GT
Coarse-to-Fine	12.6%	14.0%	11.8%	10.6%	**8.3%**	5.0%
StereoSLIC [29]	5.4%	7.5%	6.8%	6.3%	**4.6%**	3.0%

Table 2: Comparison of different smoothness terms. Given the same data term on KITTI stereo training data, our data-driven smoothness term outperforms others. See the descriptions for these smoothness terms in the text.

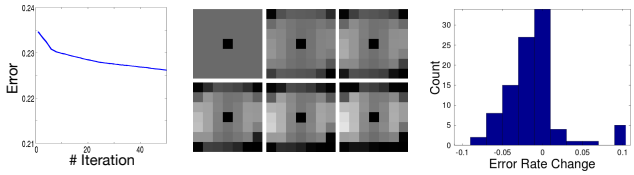
Comparison Against Other Smoothness Terms We fix the data term of the stereo model and compare our data-driven smoothness against the following smoothness terms: first-order smoothness (∇ [13]), second-order smoothness (∇^2 [27]), and the higher-order bilateral smoothness ($\mathcal{B}$ [30]). We also include the upper bound result for our model, which transfers the local shape information from the ground truth disparity map (GT).

In practice, we calculate data terms following Eq (6) with two different initializations: coarse-to-fine scheme of our model and StereoSLIC [29]. The weight parameters for each smoothness term (except for GT) are chosen through the cross-validation procedure. The weight for GT is set the same as the data-driven smoothness, otherwise the error will drop to 0 with increasing weight. We evaluate the test results with a threshold of 3 pixels on the entire disparity map and we show in Table 2 that our data-driven smoothness term consistently outperforms others. The gap between ours and the upper bound performance is due to the performance of the SIFT flow algorithm, which may fail to transfer correct local shape information for some regions.

MRF Learning Results To estimate the weight parameter $\gamma(q)$, we run 50 steps of gradient descent, which is described in Section 5.2. In Figure 6a, we show that these $\gamma(q)$ learned from the approximated loss function can effectively decrease the desired loss evaluated under the true 0-1 penalty function. By the definition in Eq (1), $\gamma(q)$ reflects the importance of each position in a patch to regularize the disparity value at the center pixel. In Figure 6b, we visualize the weight vector $\gamma(q)$, where brighter pixels have higher value. As the number of iterations increases, $\gamma(q)$ evolves from the initial uniform configuration to a configuration that favors regularization along the horizontal direction. In Figure 6c, we show the change of the absolute error rate on the test set of examples from using uniform weight to the learned parameters. On average, the new parameters only marginally improve the performance on the test set by 0.1%.

7. Experiments on Flow

In this section, we first evaluate our model on the KITTI flow benchmark [7], which is comparable to a recent state-



(a) Learning Error (b) Learned Weight (c) Testing Result

Figure 7: MRF learning result on the first 10 training examples from the KITTI stereo dataset. (a) the average 0-1 loss decreases over the iterations; (b) the weight vector learned by the approximated loss function at iteration {0, 10, 20, 30, 40, 50}; (c) the histogram of the change of the error rate on the test set with the learned parameters, where the improvement is insignificant.

of-the-art method (excluding those using additional frames or restrictive prior information [29]). Then, we show significant improvement on the training sequences from the Sintel flow dataset [4] in an online flow estimation scenario. In the flow experiments below, we use the same MRF model parameters for flow as that for stereo.

7.1. Benchmark Results on KITTI Flow Dataset

The KITTI flow dataset provides 194 pairs of temporal adjacent frames for training and 195 pairs for testing.

Quantitative Result The best available initialization on the benchmark is from DataFlow [26], whose model is similar to ours without the data-driven regularization. Thus, we directly use the coarse-to-fine initialization of our own model and our entry name is "DDS-DF" on the benchmark website. Shown in Table 3, our data-driven regularization model significantly improves upon the initialization by around 10% and obtains comparable results to the recent state-of-the-art method [18]. With the metric using a 4 pixel error threshold, our proposed method outperforms [18] by 1% when evaluated on the entire region, and is worse by 8% when evaluated on the non-occluded region.

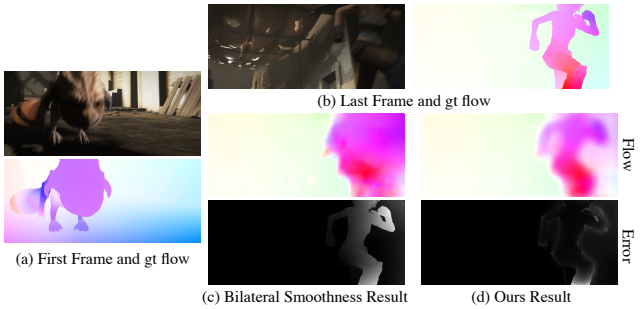
Qualitative Results We visualize results from the KITTI flow training dataset. As the failure mode for flow is similar to that for stereo, we describe two successful examples in the right two columns in Figure 4. Note that our model improves the flow estimate for regions of trees and roads, which are matched well from the training database.

7.2. Results on Sintel Dataset

Smoothness Term	Initial	∇	∇^2	$\mathcal{B}$	Ours
Endpoint Error	8.7	8.5	7.8	7.4	**6.2**

Table 4: Results on the Sintel flow dataset with the same data term but different smoothness terms. Our data-driven smoothness term outperforms others. "Initial" result is obtained by [26].

The Sintel dataset contains CG-generated, naturalistic video sequences that are challenging for large motion amplitude, motion blur and non-rigid motion. However, its 23



(a) First Frame and gt flow (b) Last Frame and gt flow (c) Bilateral Smoothness Result (d) Ours Result

Figure 8: (a) the first frame and its GT flow from a training sequence from the Sintel flow dataset; (b) the last frame and its GT flow from the same sequence; (c) flow estimate with bilateral regularization and its error map; (d) flow estimate with our and its error map. Error maps are scaled between 0 (black) and 5 (white).

training sequences share little similar scene structure with the 12 testing sequences, which breaks the assumption of our model. We here benchmark our model under the online estimation scenario: Given the ground truth flow for the first two frames, the goal is to improve the flow estimate for the later frames. Such a scenario could be useful to avoid the costly computation of the flow estimate from more complicated algorithms for every pair of adjacent frames. In Figure 7a, we show the first frame (top) and its ground truth flow (bottom) from a training sequence in the Sintel dataset. Although the sequence contents change significantly over time, the background scene structures still share a certain similarity. In Figure 7b, we show the last frame (left) and its ground truth flow map (right).

Quantitative Result We follow the same procedure as in Section 6.2 to compare our data-driven smoothness term with other smoothness terms. For comparison, we evaluate the standard Endpoint Error for the flow between the last two frames averaged over all training sequences. Shown in Table 4, our data-driven smoothness term outperforms those popular smoothness terms. In Figure 7c, we show the flow estimate (top) and the error map (bottom) of the bilateral smoothness term, which can't correctly infer the object boundary from color information due to the low contrast. Shown in Figure 7d, our data-driven smoothness term alleviates the problem of over-smoothing flow by transferring the regularization on scene structures from the ground truth flow between the first two frames.

8. Summary

Recent data-driven techniques can reliably estimate, from a single test image, scene properties such as high-resolution appearance or depth. Here, we incorporate a non-parametric approach into a generative MRF model to improve results for stereo and flow estimation. We regularize the estimates based not only on local appearance but also on the scene properties of contextually similar images

	> 2 pixels		> 3 pixels		> 4 pixels		> 5 pixels	
	Non-Out	All	Non-Out	All	Non-Out	All	Non-Out	All
NLTGV-DF [18]	**7.64%**	**14.55 %**	**5.93 %**	**11.96 %**	5.08 %	**10.48 %**	4.50 %	**9.42 %**
DDS-DF (Ours)	8.23 %	16.01 %	6.03 %	13.08 %	**5.03 %**	11.49 %	**4.41 %**	10.41 %
Data-Flow [26]	9.16 %	17.41 %	7.11 %	14.57 %	6.05 %	12.91 %	5.34 %	11.72 %

Table 3: On the KITTI flow test data, ours is comparable to a recent state-of-the-art algorithms and the smallest error rate is marked in bold. "All" means evaluation is on the entire region, while "Non-Out" evaluation is on the non-occluded region.

from a labeled database. This data-driven regularization model can better distinguish the ambiguity between appearance and scene properties, through exploiting the contextual similarities.

Acknowledgements. This work was supported by the NSF CGV 1212849 and the Office of Naval Research Multidisciplinary Research Initiative (MURI) program awards N00014-09-1-1051.

References

[1] S. Y. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction with semantic priors. In *CVPR*, 2013. 1, 2

[2] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. In *BMVC*, 2011. 2

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 1

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 1, 7

[5] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid. Dense reconstruction using 3d object shape priors. In *CVPR*, 2013. 1, 2

[6] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low level vision. *IJCV*, 2000. 2

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 4, 6

[8] C. Hane, B. Zeisl, C. Zach, and M. Pollefeys. A patch prior for dense 3d reconstruction in man-made environments. In *3DIMPVT*, 2012. 2

[9] J. Hays and A. A. Efros. Scene completion using millions of photographs. *SIGGRAPH 2007*, 26(3), 2007. 2

[10] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 1981. 2

[11] K. Jia, X. Wang, and X. Tang. Optical flow estimation using learned sparse model. In *ICCV*, 2011. 2

[12] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *ECCV*. 2012. 2

[13] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, 2001. 2, 6

[14] P. Krähenbühl and V. Koltun. Efficient nonlocal regularization for optical flow. In *ECCV*. 2012. 1

[15] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009. 1

[16] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: dense correspondence across different scenes. In *ECCV*. 2008. 1, 3, 4, 5

[17] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001. 4

[18] R. Ranftl, K. Bredies, and T. Pock. Non-local total generalized variation for optical flow estimation. In *ECCV*. 2014. 7, 8

[19] K. G. Samuel and M. F. Tappen. Learning optimized map estimates in continuously-valued mrf models. In *CVPR*, 2009. 1, 4

[20] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 1

[21] B. M. Smith, L. Zhang, and H. Jin. Stereo matching with nonparametric smoothness priors in feature space. In *CVPR*, 2009. 1, 2

[22] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010. 2

[23] D. Sun, S. Roth, J. Lewis, and M. J. Black. Learning optical flow. In *ECCV*. 2008. 1, 2

[24] L. Sun and J. Hays. Super-resolution from internet-scale scene matching. In *ICCP*, 2012. 2

[25] M. F. Tappen and C. Liu. A Bayesian approach to alignment-based image hallucination. In *ECCV*. 2012. 1, 2

[26] C. Vogel, S. Roth, and K. Schindler. An evaluation of data costs for optical flow. In *GCPR*. 2013. 3, 5, 7, 8

[27] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *TPAMI*, 2009. 2, 4, 6

[28] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *ECCV*. 2012. 4, 5

[29] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *CVPR*, 2013. 4, 5, 6, 7

[30] Q. Yang. A non-local cost aggregation method for stereo matching. In *CVPR*, 2012. 6

[31] S. Zhu, L. Zhang, and H. Jin. A locally linear regression model for boundary preserving regularization in stereo matching. In *ECCV*. 2012. 1, 2