

# End-to-End Learning of Geometry and Context for Deep Stereo Regression

Alex Kendall      Hayk Martirosyan      Saumitro Dasgupta      Peter Henry  
Ryan Kennedy      Abraham Bachrach      Adam Bry

Skydio Research

{alex, hayk, saumitro, peter, ryan, abe, adam}@skydio.com

## Abstract

*We propose a novel deep learning architecture for regressing disparity from a rectified pair of stereo images. We leverage knowledge of the problem’s geometry to form a cost volume using deep feature representations. We learn to incorporate contextual information using 3-D convolutions over this volume. Disparity values are regressed from the cost volume using a proposed differentiable soft argmin operation, which allows us to train our method end-to-end to sub-pixel accuracy without any additional post-processing or regularization. We evaluate our method on the Scene Flow and KITTI datasets and on KITTI we set a new state-of-the-art benchmark, while being significantly faster than competing approaches.*

## 1. Introduction

Accurately estimating three dimensional geometry from stereo imagery is a core problem for many computer vision applications, including autonomous vehicles and UAVs [2]. In this paper we are specifically interested in computing the disparity of each pixel between a rectified stereo pair of images. To achieve this, the core task of a stereo algorithm is computing the correspondence of each pixel between two images. This is very challenging to achieve robustly in real-world scenarios. Current state-of-the-art stereo algorithms often have difficulty with textureless areas, reflective surfaces, thin structures and repetitive patterns. Many stereo algorithms aim to mitigate these failures with pooling or gradient based regularization [15, 23]. However, this often requires a compromise between smoothing surfaces and detecting detailed structures.

In contrast, deep learning models have been successful in learning powerful representations directly from the raw data in object classification [28], detection [17] and semantic segmentation [31, 3]. These examples demonstrate that deep convolutional networks are very effective for understanding semantics. They excel at classification tasks when

supervised with large training datasets. We observe that a number of these challenging problems for stereo algorithms would benefit from knowledge of global semantic context, rather than relying solely on local geometry. For example, given a reflective surface of a vehicle’s wind-shield, a stereo algorithm is likely to be erroneous if it relies solely on the local appearance of the reflective surface to compute geometry. Rather, it would be advantageous to understand the semantic context of this surface (that it belongs to a vehicle) to infer the local geometry. In this paper we show how to learn a stereo regression model end-to-end, with the capacity to understand wider contextual information.

Stereo algorithms which leverage deep learning representations have so far been largely focused on using them to generate unary terms [48, 32]. Applying cost matching on the deep unary representations performs poorly when estimating pixel disparities [32, 48]. Traditional regularization and post processing steps are still used, such as semi global block matching and left-right consistency checks [23]. These regularization steps are severely limited because they are hand-engineered, shallow functions, which are still susceptible to the aforementioned problems.

This paper asks the question, can we formulate the entire stereo vision problem with deep learning using our understanding of stereo geometry? The main contribution of this paper is an end-to-end deep learning method to estimate per-pixel disparity from a single rectified image pair. Our architecture is illustrated in Figure 1. It explicitly reasons about geometry by forming a cost volume, while also reasoning about semantics using a deep convolutional network formulation. We achieve this with two key ideas:

- We learn to incorporate context directly from the data, employing 3-D convolutions to learn to filter the cost volume over height  $\times$  width  $\times$  disparity dimensions,
- We use a soft argmin function, which is fully differentiable, and allows us to regress sub-pixel disparity values from the disparity cost volume.

Section 3 introduces this model. In Section 4 we evaluate our model on the synthetic Scene Flow dataset [36] and set a new state-of-the-art benchmark on the KITTI 2012 and

Figure 1: Our end-to-end deep stereo regression architecture, GC-Net (Geometry and Context Network).

2015 datasets [14, 35]. Finally, in Section 4.3 we present evidence that our model has the capacity to learn semantic reasoning and contextual information.

## 2. Related Work

The problem of computing depth from stereo image pairs has been studied for quite some time [5]. A survey by Scharstein and Szeliski [39] provides a taxonomy of stereo algorithms as performing some subset of: matching cost computation, cost support aggregation, disparity computation and optimization, or disparity refinement. This survey also described the first Middlebury dataset and associated evaluation metrics. The KITTI dataset [14, 35] is a larger dataset from data collected from a moving vehicle with LIDAR ground truth. These datasets first motivated improved hand-engineered techniques for all components of stereo, of which we mention a few notable examples.

The matching cost is a measure of pixel dissimilarity for potentially corresponding image locations [25], of which absolute differences, squared differences, and truncated differences are examples. Local descriptors based on gradients [16] or binary patterns, such as CENSUS [45] or BRIEF [7, 22], can be employed. Awareness of the image content can more heavily incorporate neighboring pixels possessing similar appearance, under the assumption that they are more likely to come from the same surface and disparity. A survey of these techniques is provided by Tombari et al. [43]. Local matching costs may also be optimized within a global framework, usually minimizing an energy function combining a local data term and a pairwise smoothness term. Global optimization can be accomplished using graph cuts [27] or belief propagation [26], which can be extended to slanted surfaces [6]. A popular and effective approximation to global optimization is the *Semi-Global Matching* (SGM) of Hirschmüller [24], where dynamic programming optimizes a pathwise form of the energy function in many directions.

In addition to providing a basis for comparing stereo algorithms, the ground truth depth data from these datasets

provides the opportunity to use machine learning for improving stereo algorithms in a variety of ways. Zhang and Seitz [52] alternately optimized disparity and Markov random field regularization parameters. Scharstein and Pal [38] learn conditional random field (CRF) parameters, and Li and Huttenlocher [29] train a non-parametric CRF model using the structured support vector machine. Learning can also be employed to estimate the confidence of a traditional stereo algorithm, such as the random forest approach of Haeusler et al. [19]. Such confidence measures can improve the result of SGM as shown by Park and Yoon [37].

Deep convolutional neural networks can be trained to match image patches [46]. A deep network trained to match  $9 \times 9$  image patches, followed by non-learned cost aggregation and regularization, was shown by Žbontar and LeCun [47, 49] to produce then state-of-the-art results. Luo et al. presented a notably faster network for computing local matching costs as a multi-label classification of disparities using a Siamese network [33]. A multi-scale embedding model from Chen et al. [9] also provided good local matching scores. Also noteworthy is the *DeepStereo* work of Flynn et al. [12], which learns a cost volume combined with a separate conditional color model to predict novel viewpoints in a multi-view stereo setting.

Mayer et al. created a large synthetic dataset to train a network for disparity estimation (as well as optical flow) [34], improving the state-of-the-art. As one variant of the network, a 1-D correlation was proposed along the disparity line which is a multiplicative approximation to the stereo cost volume. In contrast, our work does not collapse the feature dimension when computing the cost volume and uses 3-D convolutions to incorporate context.

Though the focus of this work is on binocular stereo, it is worth noting that the representational power of deep convolutional networks also enables depth estimation from a single monocular image [10]. Deep learning is combined with a continuous CRF by Liu et al. [30]. Instead of supervising training with labeled ground truth, unlabeled stereo pairs can be used to train a monocular model [13].

In our work, we apply no post-processing or regularization. We explicitly reason about geometry by forming a fully differentiable cost volume and incorporate context from the data with a 3-D convolutional architecture. We don't learn a probability distribution, cost function, or classification result. Rather, our network is able to directly regress a sub-pixel estimate of disparity from a stereo image pair.

### 3. Learning End-to-end Disparity Regression

Rather than design any step of the stereo algorithm by hand, we would like to learn an end-to-end mapping from an image pair to disparity maps using deep learning. We hope to learn a more optimal function directly from the data. Additionally, this approach promises to reduce much of the engineering design complexity. However, our intention is not to naively construct a machine learning architecture as a black box to model stereo. Instead, we advocate the use of the insights from many decades of multi-view geometry research [20] to guide architectural design. Therefore, we form our model by developing differentiable layers representing each major component in traditional stereo pipelines [39]. This allows us to learn the entire model end-to-end while leveraging our geometric knowledge of the stereo problem.

Our architecture, GC-Net (Geometry and Context Network) is illustrated in Figure 1, with a more detailed layer-by-layer definition in Table 1. In the remainder of this section we discuss each component in detail. Later, in Section 4.1, we present quantitative results justifying our design decisions.

#### 3.1. Unary Features

First we learn a deep representation to use to compute the stereo matching cost. Rather than compute the stereo matching cost using raw pixel intensities, it is common to use a feature representation. The motivation is to compare a descriptor which is more robust to the ambiguities in photometric appearance and can incorporate local context.

In our model we learn a deep representation through a number of 2-D convolutional operations. Each convolutional layer is followed by a batch normalization layer and a rectified linear non-linearity. To reduce computational demand, we initially apply a  $5 \times 5$  convolutional filter with stride of two to subsample the input. Following this layer, we append eight residual blocks [21] which each consist of two  $3 \times 3$  convolutional filters in series. Our final model architecture is shown in Table 1. We form the unary features by passing both left and right stereo images through these layers. We share the parameters between the left and right towers to more effectively learn corresponding features.

Layer Description		Output Tensor Dim.
Input image		$H \times W \times C$
<b>Unary features (section 3.1)</b>		
1	$5 \times 5$ conv, 32 features, stride 2	$\frac{1}{2}H \times \frac{1}{2}W \times F$
2	$3 \times 3$ conv, 32 features	$\frac{1}{2}H \times \frac{1}{2}W \times F$
3	$3 \times 3$ conv, 32 features	$\frac{1}{2}H \times \frac{1}{2}W \times F$
4-17	add layer 1 and 3 features (residual connection) (repeat layers 2,3 and residual connection) $\times 7$	$\frac{1}{2}H \times \frac{1}{2}W \times F$
18	$3 \times 3$ conv, 32 features, (no ReLU or BN)	$\frac{1}{2}H \times \frac{1}{2}W \times F$
<b>Cost volume (section 3.2)</b>		
Cost Volume		$\frac{1}{2}D \times \frac{1}{2}H \times \frac{1}{2}W \times 2F$
<b>Learning regularization (section 3.3)</b>		
19	3-D conv, $3 \times 3 \times 3$ , 32 features	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times F$
20	3-D conv, $3 \times 3 \times 3$ , 32 features	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times F$
21	From Cost Volume: 3-D conv, $3 \times 3 \times 3$ , 64 features, stride 2	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
22	3-D conv, $3 \times 3 \times 3$ , 64 features	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
23	3-D conv, $3 \times 3 \times 3$ , 64 features	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
24	From 21: 3-D conv, $3 \times 3 \times 3$ , 64 features, stride 2	$\frac{1}{4}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
25	3-D conv, $3 \times 3 \times 3$ , 64 features	$\frac{1}{4}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
26	3-D conv, $3 \times 3 \times 3$ , 64 features	$\frac{1}{4}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
27	From 24: 3-D conv, $3 \times 3 \times 3$ , 64 features, stride 2	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
28	3-D conv, $3 \times 3 \times 3$ , 64 features	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
29	3-D conv, $3 \times 3 \times 3$ , 64 features	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
30	From 27: 3-D conv, $3 \times 3 \times 3$ , 128 features, stride 2	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 4F$
31	3-D conv, $3 \times 3 \times 3$ , 128 features	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 4F$
32	3-D conv, $3 \times 3 \times 3$ , 128 features	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 4F$
33	$3 \times 3 \times 3$ , 3-D transposed conv, 64 features, stride 2 add layer 33 and 29 features (residual connection)	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2F$
34	$3 \times 3 \times 3$ , 3-D transposed conv, 64 features, stride 2 add layer 34 and 26 features (residual connection)	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
35	$3 \times 3 \times 3$ , 3-D transposed conv, 64 features, stride 2 add layer 35 and 23 features (residual connection)	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2F$
36	$3 \times 3 \times 3$ , 3-D transposed conv, 32 features, stride 2 add layer 36 and 20 features (residual connection)	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times F$
37	$3 \times 3 \times 3$ , 3-D trans conv, 1 feature (no ReLU or BN)	$D \times H \times W \times 1$
<b>Soft argmin (section 3.4)</b>		
Soft argmin		$H \times W$

Table 1: **Summary of our end-to-end deep stereo regression architecture, GC-Net.** Each 2-D or 3-D convolutional layer represents a block of convolution, batch normalization and ReLU non-linearity (unless otherwise specified).

#### 3.2. Cost Volume

We use the deep unary features to compute the stereo matching cost by forming a cost volume. While a naive approach might simply concatenate the left and right feature maps, forming a cost volume allows us to constrain the model in a way which preserves our knowledge of the geometry of stereo vision. For each stereo image, we form a cost volume of dimensionality  $height \times width \times (max\ disparity + 1) \times feature\ size$ . We achieve this by concatenating each unary feature with their corresponding unary from the opposite stereo image across each disparity level, and packing these into the 4D volume.

Crucially, we retain the feature dimension through this operation, unlike previous work which uses a dot product style operation which decimates the feature dimension [32]. This allows us to learn to incorporate context which can operate over feature unaries (Section 3.3). We find that forming a cost volume with concatenated features improves performance over subtracting features or using a distance metric. Our intuition is that by maintaining the feature unaries, the network has the opportunity to learn an absolute representation (because it is not a distance metric) and carry this through to the cost volume. This gives the architecture the capacity to learn semantics. In contrast, using a distance metric restricts the network to only learning relative

(a) Soft ArgMin

(b) Multi-modal distribution

(c) Multi-modal distribution with prescaling

Figure 2: **A graphical depiction of the soft argmin operation** (Section 3.4) which we propose in this work. It is able to take a cost curve along each disparity line and output an estimate of the argmin by summing the product of each disparity’s softmax probability and it’s disparity index. (a) demonstrates that this very accurately captures the true argmin when the curve is uni-modal. (b) demonstrates a failure case when the data is bi-modal with one peak and one flat region. (c) demonstrates that this failure may be avoided if the network learns to pre-scale the cost curve, because the softmax probabilities will tend to be more extreme, producing a uni-modal result.

representations between features, and cannot carry absolute feature representations through to cost volume.

### 3.3. Learning Context

Given this disparity cost volume, we would now like to learn a regularization function which is able to take into account context in this volume and refine our disparity estimate. The matching costs between unaries can never be perfect, even when using a deep feature representation. For example, in regions of uniform pixel intensity (for example, sky) the cost curve will be flat for any features based on a fixed, local context. We find that regions like this can cause multi modal matching cost curves across the disparity dimension. Therefore we wish to learn to regularize and improve this volume.

We propose to use three-dimensional convolutional operations to filter and refine this representation. 3-D convolutions are able to learn feature representations from the height, width and disparity dimensions. Because we compute the cost curve for each unary feature, we can learn convolutional filters from this representation. In Section 4.1 we show the importance of these 3-D filters for learning context and significantly improving stereo performance.

The difficulty with 3-D convolutions is that the additional dimension is a burden on the computational time for both training and inference. Deep encoder-decoder tasks which are designed for dense prediction tasks get around their computational burden by encoding sub-sampled feature maps, followed by up-sampling in a decoder [3]. We extend this idea to three dimensions. By sub-sampling the input with stride two, we also reduce the 3-D cost volume

size by a factor of eight. We form our 3-D regularization network with four levels of sub-sampling. As the unaries are already sub-sampled by a factor of two, the features are sub-sampled by a total factor of 32. This allows us to explicitly leverage context with a wide field of view. We apply two  $3 \times 3 \times 3$  convolutions in series for each encoder level. To make dense predictions with the original input resolution, we employ a 3-D transposed convolution to up-sample the volume in the decoder. The full architecture is described in Table 1.

Sub-sampling is useful to increase each feature’s receptive field while reducing computation. However, it also reduces spatial accuracy and fine-grained details through the loss of resolution. For this reason, we add each higher resolution feature map before up-sampling. These residual layers have the benefit of retaining higher frequency information, while the up-sampled features provide an attentive feature map with a larger field of view.

Finally, we apply a single 3-D transposed convolution (deconvolution), with stride two and a single feature output. This layer is necessary to make dense prediction in the original input dimensions because the feature unaries were sub-sampled by a factor of two. This results in the final, regularized cost volume with size  $H \times W \times D$ .

### 3.4. Differentiable ArgMin

Typically, stereo algorithms produce a final cost volume from the matching cost unaries. From this volume, we may estimate disparity by performing an argmin operation over the cost volumes disparity dimension. However, this operation has two problems:



Model	> 1 px	> 3 px	> 5 px	MAE (px)	RMS (px)	Param.	Time (ms)
<i>1. Comparison of architectures</i>							
Unaries only (omitting all 3-D conv layers 19-36) w Regression Loss	97.9	93.7	89.4	36.6	47.6	0.16M	0.29
Unaries only (omitting all 3-D conv layers 19-36) w Classification Loss	51.9	24.3	21.7	13.1	36.0	0.16M	0.29
Single scale 3-D context (omitting 3-D conv layers 21-36)	34.6	24.2	21.2	7.27	20.4	0.24M	0.84
<b>Hierarchical 3-D context (all 3-D conv layers)</b>	16.9	9.34	7.22	2.51	12.4	3.5M	0.95
<i>2. Comparison of loss functions</i>							
GC-Net + Classification loss	19.2	12.2	10.4	5.01	20.3	3.5M	0.95
GC-Net + Soft classification loss [32]	20.6	12.3	10.4	5.40	25.1	3.5M	0.95
<b>GC-Net + Regression loss</b>	16.9	9.34	7.22	2.51	12.4	3.5M	0.95
<b>GC-Net (final architecture with regression loss)</b>	16.9	9.34	7.22	2.51	12.4	3.5M	0.95

Table 2: **Results on the Scene Flow dataset [36]** which contains 35,454 training and 4,370 testing images of size  $960 \times 540$ px from an array of synthetic scenes. We compare different architecture variants to justify our design choices. The first experiment shows the importance of the 3-D convolutional architecture. The second experiment shows the performance gain from using a regression loss.

- it is discrete and is unable to produce sub-pixel disparity estimates,
- it is not differentiable and therefore unable to be trained using back-propagation.

To overcome these limitations, we define a *soft argmin*<sup>1</sup> which is both fully differentiable and able to regress a smooth disparity estimate. First, we convert the predicted costs,  $c_d$  (for each disparity,  $d$ ) from the cost volume to a probability volume by taking the negative of each value. We normalize the probability volume across the disparity dimension with the softmax operation,  $(\cdot)$ . We then take the sum of each disparity,  $d$ , weighted by its normalized probability. A graphical illustration is shown in Figure 2 and defined mathematically in (1):

$$\text{soft argmin} := \sum_{d=0}^{D_{\max}} d \times (-c_d) \quad (1)$$

This operation is fully differentiable and allows us to train and regress disparity estimates. We note that a similar function was first introduced by [4] and referred to as a soft-attention mechanism. Here, we show how to apply it for the stereo regression problem.

However, compared to the argmin operation, its output is influenced by all values. This leaves it susceptible to multi-modal distributions, as the output will not take the most likely. Rather, it will estimate a weighted average of all modes. To overcome this limitation, we rely on the network’s regularization to produce a disparity probability distribution which is predominantly unimodal. The network can also pre-scale the matching costs to control the peakiness (sometimes called temperature) of the normalized post-softmax probabilities (Figure 2). We explicitly omit batch normalization from the final convolution layer in the unary tower to allow the network to learn this from the data.

<sup>1</sup>Note that if we wished for our network to learn probabilities, rather than cost, this function could easily be adapted to a soft argmax operation.

### 3.5. Loss

We train our entire model end-to-end from a random initialization. We train our model with supervised learning using ground truth depth data. In the case of using LIDAR to label ground truth values (e.g. KITTI dataset [14, 35]) these labels may be sparse. Therefore, we average our loss over the labeled pixels,  $N$ . We train our model using the absolute error between the ground truth disparity,  $d_n$ , and the model’s predicted disparity,  $\hat{d}_n$ , for pixel  $n$ . This supervised regression loss is defined in (2):

$$\text{Loss} = \frac{1}{N} \sum_{n=1}^N |d_n - \hat{d}_n| \quad (2)$$

In the following section we show that formulating our model as a regression problem allows us to regress with sub-pixel accuracy and outperform classification approaches. Additionally, formulating a regression model makes it possible to leverage unsupervised learning losses based on photometric reprojection error [13].

## 4. Experimental Evaluation

In this section we present qualitative and quantitative results on two datasets, Scene Flow [36] and KITTI [14, 35]. Firstly, in Section 4.1 we experiment with different variants of our model and justify a number of our design choices using the Scene Flow dataset [36]. In Section 4.2 we present results of our approach on the KITTI dataset and set a new state-of-the-art benchmark. Finally, we measure our model’s capacity to learn context in Section 4.3.

For the experiments in this section, we implement our architecture using TensorFlow [1]. All models are optimized end-to-end with RMSProp [42] and a constant learning rate of  $1 \times 10^{-3}$ . We train with a batch size of 1 using a  $256 \times 512$  randomly located crop from the input images. Before training we normalize each image such that the pixel intensities range from  $-1$  to  $1$ . We trained the network (from a random initialization) on Scene Flow for approximately 150k iterations which takes two days on a single NVIDIA

	>2 px		>3 px		>5 px		Mean Error		Runtime (s)
	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	
SPS-st [44]	4.98	6.28	3.39	4.41	2.33	3.00	0.9 px	1.0 px	2
Deep Embed [8]	5.05	6.47	3.10	4.24	1.92	2.68	0.9 px	1.1 px	3
Content-CNN [32]	4.98	6.51	3.07	4.29	2.03	2.82	0.8 px	1.0 px	<b>0.7</b>
MC-CNN [50]	3.90	5.45	2.43	3.63	1.64	2.39	0.7 px	0.9 px	67
PBCP [40]	3.62	5.01	2.36	3.45	1.62	2.32	0.7 px	0.9 px	68
Displets v2 [18]	3.43	4.46	2.37	3.09	1.72	2.17	0.7 px	0.8 px	265
GC-Net (this work)	<b>2.71</b>	<b>3.46</b>	<b>1.77</b>	<b>2.30</b>	<b>1.12</b>	<b>1.46</b>	<b>0.6 px</b>	<b>0.7 px</b>	0.9

(a) **KITTI 2012 test set results** [14]. This benchmark contains 194 train and 195 test gray-scale image pairs.

	All Pixels			Non-Occluded Pixels			Runtime (s)
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	
MBM [11]	4.69	13.05	6.08	4.33	12.12	5.61	0.13
ELAS [15]	7.86	19.04	9.72	6.88	17.73	8.67	0.3
Content-CNN [32]	3.73	8.58	4.54	3.32	7.44	4.00	1.0
DispNetC [34]	4.32	<b>4.41</b>	4.34	4.11	<b>3.72</b>	4.05	<b>0.06</b>
MC-CNN [50]	2.89	8.88	3.89	2.48	7.64	3.33	67
PBCP [40]	2.58	8.74	3.61	2.27	7.71	3.17	68
Displets v2 [18]	3.00	5.56	3.43	2.73	4.95	3.09	265
GC-Net (this work)	<b>2.21</b>	6.16	<b>2.87</b>	<b>2.02</b>	5.58	<b>2.61</b>	0.9

(b) **KITTI 2015 test set results** [35]. This benchmark contains 200 training and 200 test color image pairs. The qualifier ‘bg’ refers to background pixels which contain static elements, ‘fg’ refers to dynamic object pixels, while ‘all’ is all pixels (fg+bg). The results show the percentage of pixels which have greater than three pixels or 5% disparity error from all 200 test images.

Table 3: Comparison to other stereo methods on the test set of **KITTI 2012 and 2015 benchmarks** [14, 35]. Our method sets a new state-of-the-art on these two competitive benchmarks, out performing all other approaches.

Titan-X GPU. For the KITTI dataset we fine-tune the models pre-trained on Scene Flow for a further 50k iterations. For our experiments on Scene Flow we use F=32, H=540, W=960, D=192 and on the KITTI dataset we use F=32, H=388, W=1240, D=192 for feature size, image height, image width and maximum disparity, respectively.

#### 4.1. Model Design Analysis

In Table 2 we present an ablation study to compare a number of different model variants and justify our design choices. We wish to evaluate the importance of the key ideas in this paper; using a regression loss over a classification loss, and learning 3-D convolutional filters for cost volume regularization. We use the synthetic Scene Flow dataset [36] for these experiments, which contains 35,454 training and 4,370 testing images. We use this dataset for two reasons. Firstly, we know perfect, dense ground truth from the synthetic scenes which removes any discrepancies due to erroneous labels. Secondly, the dataset is large enough to train the model without over-fitting. In contrast, the KITTI dataset only contains 200 training images, making the model is susceptible to over-fitting to this very small dataset. With tens of thousands of training images we do not have to consider over-fitting in our evaluation.

The first experiment in Table 2 shows that including the 3-D filters performs significantly better than learning unaries only. We compare our full model (as defined in Table 1)

to a model which uses only unary features (omitting all 3-D convolutional layers 19-36) and a model which omits the hierarchical 3-D convolution (omitting layers 21-36). We observe that the 3-D filters are able to regularize and smooth the output effectively, while learning to retain sharpness and accuracy in the output disparity map. We find that the hierarchical 3-D model outperforms the vanilla 3-D convolutional model by aggregating a much large context, without significantly increasing computational demand.

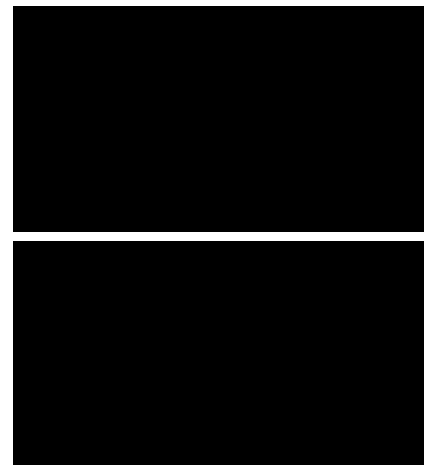
The second experiment in Table 2 compares our regression loss function to baselines classifying disparities with hard or soft classification as proposed in [32]. Hard classification trains the network to classify disparities in the cost volume as probabilities using cross entropy loss with a ‘one hot’ encoding. Soft classification (used by [32]) smooths this encoding to learn a Gaussian distribution centered around the correct disparity value. In Table 2 we show our regression approach outperforms both hard and soft classification. This is especially noticeable for the pixel accuracy metrics and the percentage of pixels which are within one pixel of the true disparity, because the regression loss allows the model to predict with sub-pixel accuracy.

#### 4.2. KITTI Benchmark

In Table 3 we evaluate the performance of our model on the KITTI 2012 and 2015 stereo datasets [14, 35]. These consist of challenging and varied road scene imagery col-

(a) KITTI 2012 test data qualitative results. From left: left stereo input image, disparity prediction, error map.

(b) KITTI 2015 test data qualitative results. From left: left stereo input image, disparity prediction, error map.



(c) Scene Flow test set qualitative results. From left: left stereo input image, disparity prediction, ground truth.

Figure 3: **Qualitative results.** By learning to incorporate wider context our method is often able to handle challenging scenarios, such as reflective, thin or texture-less surfaces. By explicitly learning geometry in a cost volume, our method produces sharp results and can also handle large occlusions.

- (a) Left stereo input image
- (b) Predicted disparity map
- (c) Saliency map (red = stronger saliency)
- (d) What the network sees (input attenuated by saliency)

Figure 4: **Saliency map visualization** which shows the model’s effective receptive field for a selected output pixel (indicated by the white cross). This shows that our architecture is able to learn to regress stereo disparity with a large field of view and significant contextual knowledge of the scene, beyond the local geometry and appearance. For example, in the example on the right we observe that the model considers contextual information from the vehicle and surrounding road surface to estimate disparity.

lected from a test vehicle. Ground truth depth maps for training and evaluation are obtained from LIDAR data. KITTI is a prominent dataset for benchmarking stereo algorithms. The downside is that it only contains 200 training images, which handicaps learning algorithms. For this reason, we pre-train our model on the large synthetic dataset, Scene Flow [36]. This helps to prevent our model from over-fitting the very small KITTI training dataset. We hold out 40 image pairs as our validation set.

Table 3a and 3b compare our method, GC-Net (Geometry and Context Network), to other approaches on the KITTI 2012 and 2015 datasets, respectively<sup>2</sup>. Our method achieves state of the art results for both KITTI benchmarks, by a notable margin. We improve on state-of-the-art by 9% and 22% for KITTI 2015 and 2012 respectively. Our method is also notably faster than most competing approaches which often require expensive post-processing. In Figure 3 we show qualitative results of our

<sup>2</sup>Full leaderboard: [www.cvlibs.net/datasets/kitti/](http://www.cvlibs.net/datasets/kitti/)

method on KITTI 2012, KITTI 2015 and Scene Flow.

Our approach outperforms previous deep learning patch based methods [48, 32] which produce noisy unary potentials and are unable to predict with sub-pixel accuracy. For this reason, these algorithms do not use end-to-end learning and typically post-process the unary output with SGM regularization [11] to produce the final disparity maps.

The closest method to our architecture is DispNetC [34], which is an end-to-end regression network pre-trained on SceneFlow. However, our method outperforms this architecture by a notable margin for *all* test pixels. DispNetC uses a 1-D correlation layer along the disparity line as an approximation to the stereo cost volume. In contrast, our architecture more explicitly leverages geometry by formulating a full cost volume by using 3-D convolutions and a soft argmin layer, resulting in an improvement in performance.

### 4.3. Model Saliency

In this section we present evidence which shows our model can reason about local geometry using wider contextual information. In Figure 4 we show some examples of the model’s saliency with respect to a predicted pixel’s disparity. Saliency maps [41] shows the sensitivity of the output with respect to each input pixel. We use the method from [51] which plots the predicted disparity as a function of systematically occluding the input images. We offset the occlusion in each stereo image by the point’s disparity.

These results show that the disparity prediction for a given point is dependent on a wide contextual field of view. For example, the disparity on the front of the car depends on the input pixels of the car and the road surface below. This demonstrates that our model is able to reason about wider context, rather than simply  $9 \times 9$  local patches like previous deep learning patch-similarity stereo methods [50, 32].

## 5. Conclusions

We propose a novel end-to-end deep learning architecture for stereo vision. It is able to learn to regress disparity without any additional post-processing or regularization. We demonstrate the efficacy of our method on the KITTI dataset, setting a new state-of-the-art benchmark.

We show how to efficiently learn context in the disparity cost volume using 3-D convolutions. We show how to formulate it as a regression model using a soft argmin layer. This allows us to learn disparity as a regression problem, rather than classification, improving performance and enabling sub-pixel accuracy. We demonstrate that our model learns to incorporate wider contextual information.

For future work we are interested in exploring a more explicit representation of semantics to improve our disparity estimation, and reasoning under uncertainty with Bayesian convolutional neural networks.



## References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. **5**
- [2] M. Achtelik, A. Bachrach, R. He, S. Prentice, and N. Roy. Stereo vision and laser odometry for autonomous helicopters in gps-denied indoor environments. In *SPIE Defense, security, and sensing*, pages 733219–733219. International Society for Optics and Photonics, 2009. **1**
- [3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. **1, 4**
- [4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*, 2014. **5**
- [5] S. T. Barnard and M. A. Fischler. Computational stereo. *ACM Computing Surveys*, 14(4):553–572, 1982. **2**
- [6] M. Bleyer, C. Rhemann, and C. Rother. PatchMatch Stereo-Stereo Matching with Slanted Support Windows. *Bmvc*, i(1):14.1–14.11, 2011. **2**
- [7] M. Calonder, V. Lepetit, and C. Strecha. BRIEF : Binary Robust Independent Elementary Features. In *European Conference on Computer Vision (ECCV)*, 2010. **2**
- [8] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 972–980, 2015. **6**
- [9] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 972–980, 2016. **2**
- [10] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *Nips*, pages 1–9, 2014. **2**
- [11] N. Einecke and J. Eggert. A multi-block-matching approach for stereo. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 585–592. IEEE, 2015. **6, 8**
- [12] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. DeepStereo: Learning to Predict New Views from the World’s Imagery. *CVPR*, 2016. **2**
- [13] R. Garg, V. Kumar BG, and I. Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. *ECCV*, pages 1–16, 2016. **2, 5**
- [14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. **2, 5, 6**
- [15] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38. Springer, 2010. **1, 6**
- [16] A. Geiger, M. Roser, and R. Urtasun. Efficient Large-Scale Stereo Matching. *Computer Vision ACCV 2010*, (1):25–38, 2010. **2**
- [17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. **1**
- [18] F. Guney and A. Geiger. Displets: Resolving stereo ambiguities using object knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4165–4175, 2015. **6**
- [19] R. Haeusler, R. Nair, and D. Kondermann. Ensemble Learning for Confidence Measures in Stereo Vision. *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 305–312, 2013. **2**
- [20] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. **3**
- [21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. **3**
- [22] P. Heise, B. Jensen, S. Klose, and A. Knoll. Fast Dense Stereo Correspondences by Binary Locality Sensitive Hashing. *ICRA*, pages 1–6, 2015. **2**
- [23] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’05)*, volume 2, pages 807–814. IEEE, 2005. **1**
- [24] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. **2**
- [25] H. Hirschmüller and D. Scharstein. Evaluation of Cost Functions for Stereo Matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007. **2**
- [26] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. *Proceedings - International Conference on Pattern Recognition*, 3:15–18, 2006. **2**
- [27] V. Kolmogorov and R. Zabih. Computing visual correspondences with occlusions using graph cuts. In *International Conference on Computer Vision (ICCV)*, 2001. **2**
- [28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. **1**
- [29] Y. Li and D. P. Huttenlocher. Learning for stereo vision using the structured support vector machine. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008. **2**
- [30] F. Liu, C. Shen, G. Lin, and I. Reid. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *Pattern Analysis and Machine Intelligence*, page 15, 2015. **2**
- [31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. **1**
- [32] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016. **1, 3, 5, 6, 8**
- [33] W. Luo, A. G. Schwing, and R. Urtasun. Efficient Deep Learning for Stereo Matching. *CVPR*, 2016. **2**
- [34] 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. *CoRR*, abs/1510.0(2002), 2015. **2, 6, 8**
- [35] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. **2, 5, 6**
- [36] 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 International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. **1, 5, 6, 8**
- [37] M. G. Park and K. J. Yoon. Leveraging stereo matching with learning-based confidence measures. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:101–109, 2015. **2**
- [38] D. Scharstein and C. Pal. Learning conditional random fields for stereo. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. **2**

- [39] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002. 2, 3
- [40] A. Seki and M. Pollefeys. Patch based confidence prediction for dense disparity map. In *British Machine Vision Conference (BMVC)*, 2016. 6
- [41] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 8
- [42] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 5
- [43] F. Tombari, S. Mattoccia, L. D. Stefano, and E. Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008. 2
- [44] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*, pages 756–771. Springer, 2014. 6
- [45] R. Zabih and J. Woodfill. Non-parametric Local Transforms for Computing Visual Correspondence. In *Proceedings of European Conference on Computer Vision*, (May):151–158, 1994. 2
- [46] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June(i):4353–4361, 2015. 2
- [47] J. Žbontar and Y. Le Cun. Computing the stereo matching cost with a convolutional neural network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June(1):1592–1599, 2015. 2
- [48] J. Žbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1592–1599, 2015. 1, 8
- [49] J. Žbontar and Y. LeCun. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *CoRR*, abs/1510.0(2002), 2015. 2
- [50] J. Žbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016. 6, 8
- [51] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 8
- [52] L. Zhang and S. M. Seitz. Estimating optimal parameters for FMRFG stereo from a single image pair. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):331–342, 2007. 2