Slanted Stixels: Representing San Francisco's Steepest Streets

Daniel Hernandez-Juarez*†[1]
http://www.cvc.uab.es/people/dhernandez/
Lukas Schneider*[3]
lukas.schneider@daimler.com
Antonio Espinosa[1]
antoniomiguel.espinosa@uab.es
David Vázquez[1,2]
dvazquez@cvc.uab.es
Antonio M. López[1,2]
antonio@cvc.uab.es
Uwe Franke[3]
uwe.franke@daimler.com
Marc Pollefeys[4]
marc.pollefeys@inf.ethz.ch
Juan C. Moure[1]
juancarlos.moure@uab.es

[1] Universitat Autònoma de Barcelona
Barcelona, Spain
[2] Computer Vision Center
Barcelona, Spain
[3] Daimler AG, R&D
Böblingen, Germany
[4] ETH Zürich
Zürich, Switzerland

Abstract

In this work we present a novel compact scene representation based on Stixels that infers geometric and semantic information. Our approach overcomes the previous rather restrictive geometric assumptions for Stixels by introducing a novel depth model to account for non-flat roads and slanted objects. Both semantic and depth cues are used jointly to infer the scene representation in a sound global energy minimization formulation. Furthermore, a novel approximation scheme is introduced that uses an extremely efficient over-segmentation. In doing so, the computational complexity of the Stixel inference algorithm is reduced significantly, achieving real-time computation capabilities with only a slight drop in accuracy. We evaluate the proposed approach in terms of semantic and geometric accuracy as well as run-time on four publicly available benchmark datasets. Our approach maintains accuracy on flat road scene datasets while improving substantially on a novel non-flat road dataset.

1 Introduction

Autonomous vehicles, advanced driver assistance systems, robots and other intelligent devices need to understand their environment; this requires both geometric (distance) and semantic (classification) information. This data must be represented in a very compact model

* Both authors contributed equally.
† Work performed during an internship at Daimler AG.

and must be computed in real-time that can serve as building block of higher-level modules, such as localization and planning.

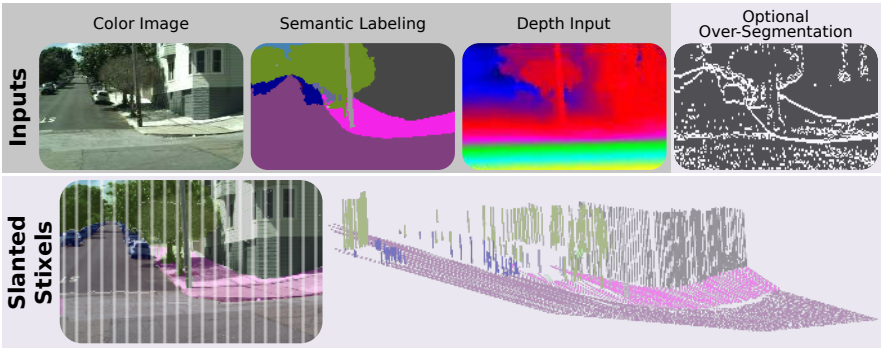


Figure 1: The proposed approach: pixel-wise semantic and depth information serve as input to our Slanted Stixels, a compact semantic 3D scene representation that accurately handles arbitrary scenarios as *e.g.* San Francisco. The optional over-segmentation in the top-right yields significant speed gains nearly retaining the depth and semantic accuracy.

The Stixel World has been successfully used for representing traffic scenes, as introduced in [20]. The intelligent vehicles community has shown an increasing interest in this model over the last years [2, 3, 4, 6, 11, 13, 15, 23]. It defines a compact medium-level representation of dense 3D disparity data obtained from stereo vision using rectangles (Stixels) as elements. These are classified either as *ground*-like planes, upright *objects* or *sky*, which are the geometric primitives found in man-made environments. This converts millions of disparity pixels to hundreds or thousands of Stixels. At the same time, most task-relevant scene structures, such as free space and obstacles, are adequately represented.

A recent work [23] fuses geometric and semantic information in an extended Stixel model, which is able to provide a richer yet compact representation of the traffic scene. Our work extends [23] by incorporating a new depth model that takes arbitrary kinds of slanted objects and non-flat roads into account. The induced extra computational complexity is reduced in this paper by incorporating an over-segmentation strategy that nearly retains the accuracy and can be applied to any Stixel model proposed so far.

Our work yields an improved Stixel representation that accounts for non-flat roads, outperforming the original Stixel model in this context while keeping the same accuracy on flat road scenes. An overview of our approach is shown in Fig. 1.

2 Related work

Our proposed approach introduces a novel Stixel-based scene representation that is able to account for non-flat roads, *c.f.* Fig. 1. We also devise an approximation to compute Stixels faster. Therefore, we see three categories of related publications.

The first group is comprised by road scene models. In most cases, occupancy grid maps are used to represent the surrounding of the vehicle [7, 17, 18, 24]. Typically a grid in bird's eye perspective is defined and used to detect occupied grid cells. These grids and the Stixel World both represent the 2D image in terms of column-wise stripes allowing to capture the

camera data in a polar fashion. However, the Stixel inference in the image domain differs significantly from classical grid-based approaches.

The second category includes different Stixel-based methods. Stixels were originally devised to represent the 3D scene as observed by stereoscopic [2, 20] or monocular imagery [15]. Our proposal is based on [23]: they use semantic cues in addition to depth to extract a Stixel representation. However, they are limited to flat road scenarios due to constant height assumption. In contrast, our proposal overcomes this drawback by incorporating a novel plane model together with effective priors on the plane parameters.

Finally, the third category consists of fast methods for Stixel computation. Some methods [1, 2, 13] model the scene with a single Stixel per column: they can be faster but provide an incomplete world model, *e.g.* they cannot represent a pedestrian and a building in the same column. A recent work [3] uses edge-based disparity maps to compute Stixels: this method is also fast but gives inferior accuracy compared to the original Stixels [21]. The FPGA implementation from [17] runs at 25 Hz. Finally, [11] present an embedded GPU implementation that runs at 26 Hz for Stixel widths of 5 px computed using an SGM stereo algorithm also implemented on GPU [10]. In contrast, we propose a novel algorithmic approximation that is hardware agnostic. Accordingly, it could also benefit of the aforementioned approaches.

Our main contributions are: (1) a real-time, compact and robust Stixel representation that incorporates a novel depth model to accurately represent arbitrary kinds of slanted surfaces in a sound probabilistic formulation; (2) a novel over-segmentation input to the main Stixel segmentation algorithm that significantly reduces the run-time of the method while nearly retaining its accuracy; (3) a new challenging synthetic dataset with non-flat roads that includes pixel-level semantic and depth ground-truth and allows to evaluate the accuracy of competing algorithms in such scenarios. This dataset will be made publicly available[1] with this paper; (4) an in-depth evaluation in terms of run-time as well as semantic and depth accuracy carried out on this novel dataset and several real-world benchmarks. Compared to the existing state-of-the-art approach, the depth accuracy is substantially improved especially in non-flat road scenarios.

3 Stixel Model

The Stixel world is a segmentation of image columns into stick-like super-pixels with class labels and a 3D planar depth model. This joint segmentation and labeling problem is carried out via optimization of the column-wise posterior distribution $P(S_{:} \mid M_{:})$ defined over a Stixel segmentation $S_{:}$ given all measurements $M_{:}$ from that particular image column. In the following, we drop the column indexes for ease of notation. We obtain Stixel widths > 1 as illustrated *e.g.* in Fig. 1 by down-sampling of the inputs.

A Stixel column segmentation consists of an arbitrary number N of Stixels S_i, each representing four random variables: the Stixel extent via bottom V_i^{b} and top V_i^{t} row, as well as it's class C_i and depth model D_i. Thereby, the number of Stixels itself is a random variable that is optimized jointly during inference. To this end, the posterior probability is defined by means of the unnormalized prior and likelihood distributions $P(S \mid M) = \frac{1}{Z}\tilde{P}(M \mid S)\,\tilde{P}(S)$ transformed to log-likelihoods via $P(S = s \mid M = m) = -\log(e^{-E(s,m)})$.

The **likelihood** term $E_{data}(\cdot)$ thereby rates how well the measurements m_v at pixel v fit

[1] http://synthia-dataset.net

to the overlapping Stixel s_i

$$E_{data}(s,m) = \sum_{i=1}^{N} E_{stixel}(s_i,m) = \sum_{i=1}^{N} \sum_{v=v_i^b}^{v_i^t} E_{pixel}(s_i,m_v) \ . \tag{1}$$

This pixel-wise energy is further split in a semantic and a depth term

$$E_{pixel}(s_i,m_v) = E_{disp}(s_i,d_v) + w_l \cdot E_{sem}(s_i,l_v) \, . \tag{2}$$

The semantic energy favors semantic classes of the Stixel that fit to the observed pixel-level semantic input [23]. The parameter w_l controls the influence of the semantic data term. The depth term is defined by means of a probabilistic and generative sensor model $P_v(\cdot)$ that considers the accordance of the depth measurement d_v at row v to the Stixel s_i

$$E_{disp}(s_i,d_v) = -\log(P_v(D_v = d_v \mid S_i = s_i)) \ . \tag{3}$$

It is comprised of a constant outlier probability p_{out} and a Gaussian sensor noise model for valid measurements with confidence c_v

$$P_v(D_v \mid S_i) = \frac{p_{\text{out}}}{Z_U} + \frac{1-p_{\text{out}}}{Z_G(s_i)} e^{-\left(\frac{c_v(d_v-\mu(s_i,v))}{\sigma(s_i)}\right)^2} \tag{4}$$

that is centered at the expected disparity $\mu(s_i,v)$ given the depth model of the Stixel. Z_U and $Z_G(s_i)$ normalize the distributions.

This paper introduces a new plane depth model that overcomes the previous rather restrictive constant depth and constant height assumptions for *object* respectively *ground* Stixels. To this end, we formulate the depth model $\mu(s_i,v)$ using two random variables defining a plane in the disparity space that evaluates to the disparity in row v via

$$\mu(s_i,v) = b_i \cdot v + a_i \, . \tag{5}$$

Note that we assume narrow Stixels and thus can neglect one plane parameter, *i.e.* the roll.

The **prior** captures knowledge about the segmentation: the Stixel segmentation has to be consistent, *i.e.* each pixel is assigned to exactly one Stixel. A model complexity term favors solutions composed of fewer Stixels by invoking costs for each Stixel in the column segmentation S. Furthermore, prior assumptions as "objects are likely to stand on the ground" and "sky is unlikely below the road surface" are taken into account. The interested reader is referred to [6] for more details. The Markov property is used so that the prior reduces to pair-wise relations between subsequent Stixels. Accordingly, the prior is computed as

$$E_{prior}(s) = \sum_{i=2}^{N} E_{pair}(s_i,s_{i-1}) + E_{first}(s_1) \ . \tag{6}$$

In this paper, we propose a new additional prior term that uses the specific properties of the three geometric classes. We expect the two random variables A,B representing the plane parameters of a Stixel to be Gaussian distributed, *i.e.*

$$E_{plane}(s_i) = \left(\frac{a-\mu_{c_i}^a}{\sigma_{c_i}^a}\right)^2 + \left(\frac{b-\mu_{c_i}^b}{\sigma_{c_i}^b}\right)^2 - \log(Z) \ . \tag{7}$$

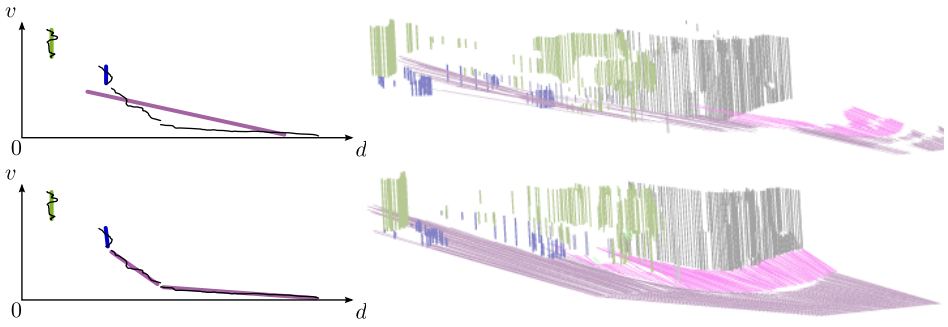


Figure 2: Comparison of original [23] (top) and our slanted (bottom) Stixels: due to the fixed slant in the original formulation the road surface is not well represented as illustrated on the left. The novel model is capable to reconstruct the whole scene accurately.

This prior favors planes in accordance to the expected 3D layout corresponding to the geometric class. *E.g. object* Stixels are expected to have an approximately constant disparity, *i.e.* $\mu^b_{object} = 0$. The expected road slant μ^a_{ground} can be set using prior knowlede or a preceeding road surface detection. Note that the novel formulation is a strict generalization of the original method, since they are equivalent, if the slant is fixed, *i.e.* $\sigma^b_{object} \to 0, \mu^b_{object} = 0$.

3.1 Inference

The sophisticated energy function defined in Sec. 3 is optimized via Dynamic Programming as in [20]. However, we also have to optimize jointly for the novel depth model. When optimizing for the plane parameters a_i, b_i of a certain Stixel s_i, it becomes apparent that all other optimization parameters are independent of the actual choice of the plane parameters. We can thus simplify

$$\underset{a_i,b_i}{\operatorname{argmin}} E(s,m) = \underset{a_i,b_i}{\operatorname{argmin}} E_{stixel}(s_i,m) + E_{plane}(s_i) \ . \tag{8}$$

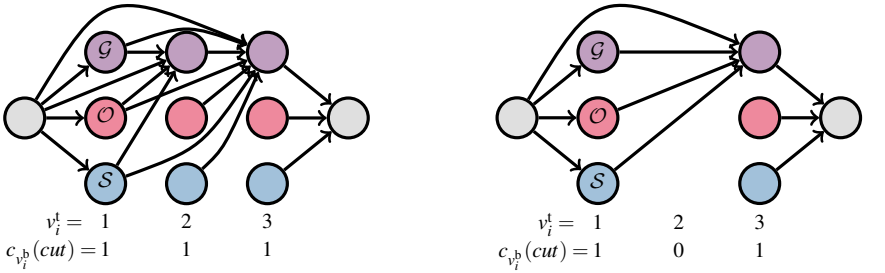
Thus, we minimize the global energy function with respect to the plane parameters of all Stixels and all geometric classes independently. We can find an optimal solution of the resulting weighted least squares problem in closed form, however, we still need to compare the Stixel disparities to our new plane depth model. Therefore, the complexity added to the original formulation is another quadratic term in the image height.

3.2 Stixel Cut Prior

The Stixel inference described so far requires to estimate the costs for each possible Stixel in an image, although many Stixels could be trivially discarded, *e.g.* in image regions with homogeneous depth and semantic input. We propose a novel prior that can be easily used to significantly reduce the computational burden by exploiting hypothesis generation. To this end, we formulate a new prior similar to [4], but instead of Stixel bottom and top probabilities we incorporate generic likelihoods for pixels being the cut between two Stixels.

We leverage this additional information adding a novel term for a Stixel s_i

$$E_{cut}(s_i) = -\log\left(c_{v_i^b}(cut)\right) \ , \tag{9}$$



(a) Stixel graph representation

(b) Pruned graph using Stixel cut prior

Figure 3: Stixel inference illustrated as shortest path problem on a directed acyclic graph: the Stixel segmentation is computed by finding the shortest path from the source (left gray node) to the sink (right gray node). The vertices represent Stixels with colors encoding their geometric class, *i.e.* **g**round, **o**bject and **s**ky. Only the incoming edges of ground nodes are shown for simplicity. Adapted from [6].

where $c_{v_i^b}(cut)$ is the confidence for a cut at v_i^b, thus $c_{v_i^b}(cut) = 0$ implies that there is no cut between two Stixels at row v.

As described in [19], we can use a recursive definition of the optimization problem to design the Dynamic Programming solution scheme. In order to simplify our description we use a special notation to refer to Stixels: $ob_b^t = \{v^b, v^t, object\}$. Similarly, OB^k is defined as the minimum aggregated cost of the best segmentation from position 0 to k. The Stixel at the end of the segmentation associated with each minimum cost is denoted as ob^k. We next show a recursive definition of the problem:

$$OB^k = min \begin{cases} E_{data}(ob_0^k) + E_{prior}(ob_0^k) \\ E_{data}(ob_x^k) + E_{prior}(ob_x^k, ob^{x-1}) + OB^{x-1} \forall x \in cuts, x \leq k \\ E_{data}(ob_x^k) + E_{prior}(ob_x^k, gr^{x-1}) + GR^{x-1} \forall x \in cuts, x \leq k \\ E_{data}(ob_x^k) + E_{prior}(ob_x^k, sk^{x-1}) + SK^{x-1} \forall x \in cuts, x \leq k \end{cases} \tag{10}$$

We only show the case for object Stixels, but the other cases are solved similarly. Also, GR^k and SK^k stand for ground and sky respectively. The base case problem, *i.e.* segmenting a column of the single pixel at the bottom, is defined: $OB^0 = E_{data}(ob_0^0) + E_{prior}(ob_0^0)$. Our method trusts that all the optimal cuts will be included in our over-segmentation (*cuts* in Eq. (10)), therefore, only those positions are checked as Stixel start and end. This reduces the complexity of the Stixel estimation problem for a single column to $\mathcal{O}(h' \times h')$, where h' is the number of over-segmentation cuts computed for this column, h is image height and $h' \ll h$.

The computational complexity reduction becomes apparent in Fig. 3a. As stated in [6], the inference problem can be interpreted as finding the shortest path in a directed acyclic graph. Our approach prunes all the vertices associated with the Stixel's top row not included according to the Stixel cut prior, *c.f.* Fig. 3b.

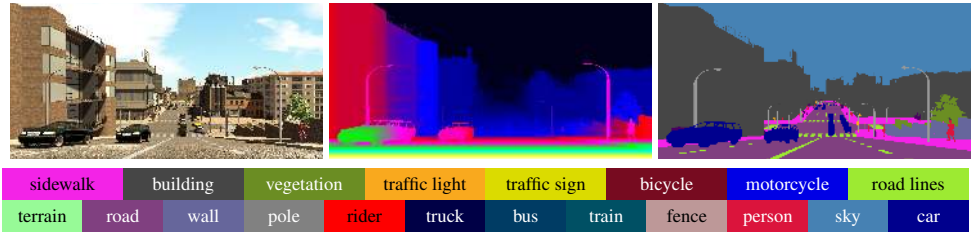


Figure 4: The SYNTHIA-SF Dataset. A sample frame (left) with its depth (center) and semantic labels (right).

4 Experiments

This section assesses the accuracy and run-time of our proposal. A previous concern is to verify that our method maintains the accuracy for scenes containing only flat roads and represents non-flat roads better. For that purpose, we evaluate on both synthetic and real data, *c.f*. Sec. 4.1. We introduce metrics and baselines as well as a Stixel over-segmentation method and other details in Sec. 4.2. Finally, quantitative and qualitative results are reported, *c.f*. Sec. 4.3.

4.1 Datasets

As our Stixel model represents geometric and semantic information, we measure the accuracy of our method for both. Therefore, we evaluate on the -to the best of our knowledge- only such dataset: an annotated subset of KITTI [14] called *Ladicky*. It consists of 60 images with a resolution of 0.5 MP that we all use for evaluation. We follow the suggestion of the original author to ignore the three rarest object classes, leaving a set of 8 classes. Following [23], we use additional publicly available semantic annotations on other parts of KITTI for training. All in all, we have a training set of 676 images, where we harmonized the object classes used by the different authors.

We also evaluate disparity accuracy on the training data of the well-known stereo challenge KITTI 2015 [9]. This dataset comprises a set of 200 images with sparse disparity ground truth obtained from a laser scanner. However, there is no suitable semantic ground truth available for this dataset. Furthermore, we evaluate the semantic accuracy on Cityscapes [5], a highly complex dataset with dense annotations of 19 classes on ca. 3000 images for training and 500 images for validation that we used for testing.

Unfortunately, all the above datasets were generated in flat road environments, and they only help us validate that we are not decreasing our accuracy for this case. In order to compare the accuracy of competing algorithms on non-flat road scenarios, a new dataset is required. Therefore, we introduce a new synthetic dataset inspired by [22]. This dataset has been generated with the purpose of evaluating our model, but it contains enough information to be useful in additional related tasks, such as object recognition, semantic and instance segmentation, among others. SYNTHIA-San Francisco (SYNTHIA-SF) consists of photo-realistic frames rendered from a virtual city and comes with precise pixel-level depth and semantic annotations for 19 classes (see Fig. 4). This new dataset contains 2224 images that we use to evaluate both depth and semantic accuracy.

4.2 Experiment Details

We evaluate our proposed method in terms of semantic and depth accuracy using two **metrics**. The depth accuracy is obtained as the outlier rate of the disparity estimates, the standard metric used to evaluate on KITTI benchmark [9]. An outlier is a disparity estimation with an absolute error larger than 3 px or a relative deviation larger than 5% compared to the ground truth. The semantic accuracy is evaluated with average Intersection-over-Union (IoU) over all classes, also a standard measure for semantics [8]. We also provide Frame-rate (Hz) to ensure our system is capable of real-time performance and number of Stixels per image to quantify the complexity of the obtained representation. All Stixel run-times are obtained using a multi-threaded implementation on standard consumer hardware: Intel i7-6800K. For the FCN output, a Maxwell NVidia Titan X is used.

Semantic Stixels [23] serves as **baseline**, because they achieve state-of-the-art results in terms of Stixel accuracy. We provide the accuracy of our new disparity model, *c.f.* Sec. 3. Finally, we also evaluate our reduced complexity approach, *c.f.* Sec. 3.2, both for our model and Semantic Stixels, *Ours (fast)* and *[23] (fast)*, respectively.

As **input**, we use disparity images D obtained via semi-global-matching (SGM) [12] and pixel-level semantic labels L computed by a fully convolutional network (FCN) [16]. We use the same FCN model used in [23] without retraining to allow for comparison. For the same reason, we set Stixel width to 8 px. The same downscaling is applied in the v-direction. The rest of the parameters are taken from [23]. We use the known camera calibration to obtain expected μ^a_{ground} and μ^b_{ground}. For objects, we set $\sigma^b_{object} \to 0, \mu^b_{object} = 0$ because the disparity is too noisy for the slanted object model. Finally, for Sky Stixels it does not make sense to have slanted surfaces, therefore, we set: $\mu^a_{sky} = 0, \mu^a_{sky} = 0, \sigma^a_{sky} \to 0, \sigma^b_{sky} \to 0$.

In order to improve the efficiency of our approach, we use a preceding **Stixel cut prior generation**. Our goal is to show that we can speed up the Stixel computation with a very simple and fast approach. Accordingly, we opt for [13] to generate our over-segmentation: a very low-level method based on simple mathematical concepts like the use of local and strict extrema of the disparity map to find points of interest. This method first performs an *Extreme points detection* step that generates possible Stixel cuts and subsequently filters them. As we are only interested in cut hypothesis we only use the first step. We also use semantic segmentation to generate cuts, when there is a change in semantic class we assume there is a cut.

4.3 Results

The quantitative results of our proposals and baselines as described in Sec. 3 are shown in Table 1. The first observation is that all variants are compact representations of the surrounding, since the complexity of the Stixel representation is small compared to the high resolution input images, *c.f.* the last row in the Table 1.

Second, our method achieves comparable or slightly better results on all datasets with flat roads. This indicates that the novel and more flexible model does not harm the accuracy in such scenarios.

The third observation is that the proposed *fast* variant improves the run-time of both the original Stixel approach by up to 2x and the novel Stixel approach by up to 7x with only a slight drop in depth accuracy. The benefit increases with higher resolution input images. This is due to the mean density of Stixel cuts in our over-segmentation for SYNTHETIC-SF of 13% with standard deviation of 2, which is equivalent to a 8x reduced vertical resolution.

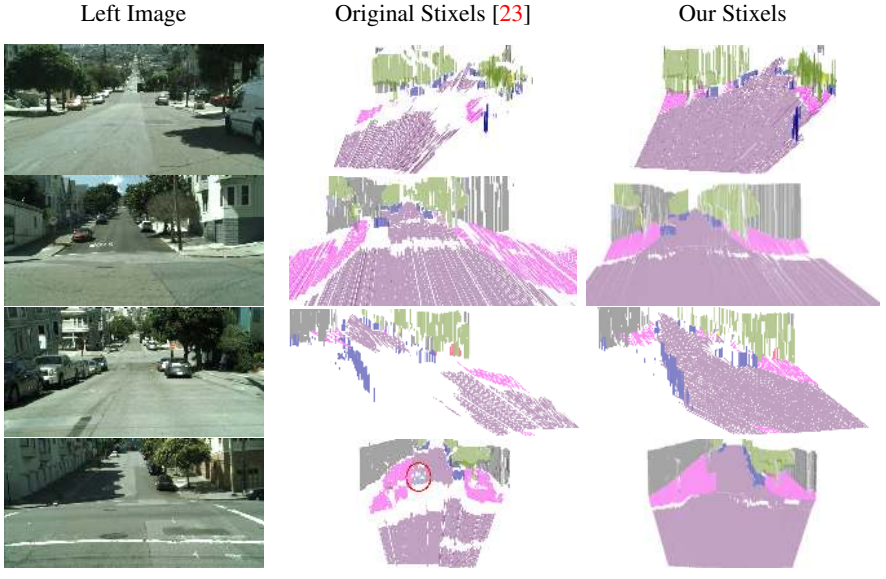


Figure 5: Exemplary outputs on real data: in all scenes with non-flat roads our model correctly represents the scene, while retaining accuracy on objects. The last line shows a failure case, where our approach classifies the road as sidewalk due to erroneous semantic input. However, the original approach reconstructs a wall in this case, highlighted by a red circle. This could lead *e.g.* to an emergency break.

Finally, we observe that our novel model is able to accurately represent non-flat scenarios in contrast to the original Stixel approach yielding a substantially increased depth accuracy by more than 17%. We also improve in terms of semantic accuracy, which we address to the joint semantic and depth inference that benefits of a better depth model.

The observations from the quantitative evaluation are confirmed also in the qualitative results, *c.f.* Fig. 5.

5 Conclusions

This paper presented a novel depth model for the Stixel World that is able to represent non-flat roads and slanted objects in a compact representation that overcomes the previous restrictive constant height and depth assumptions respectively. Moreover, a novel approximation is introduced in order to reduce the computational complexity significantly by only checking reasonable Stixel cuts. This representation change is required for difficult environments that are found in the real world. We showed in extensive experiments on several related datasets that our depth model is able to better represent those scenes and our approximation is able to reduce the run-time drastically with only a slight drop in accuracy.

Table 1: Quantitative results of our methods compared to [23], raw SGM and FCN. Significantly best results highlighted in bold.

Metric	Dataset	SGM	FCN	[23]	Ours	[23] (fast)	Ours (fast)
Disp Error (%)	Ladicky	16.6	-	17.3	16.9	18.5	17.8
	KITTI 15	11.5	-	10.9	11.0	11.8	11.7
	SYNTHIA-SF	11.0	-	30.9	**12.9**	33.9	15.4
IoU (%)	Ladicky	-	69.8	63.5	63.4	63.9	63.7
	Cityscapes	-	60.8	65.7	65.8	65.7	65.8
	SYNTHIA-SF	-	45.9	46.0	**48.5**	46.9	**48.5**
Frame-rate (Hz)	KITTI	55	47.6	113	61	120	116
	Cityscapes	22	15.4	20.9	6.6	36.6	27.5
	SYNTHIA-SF	21	13.9	19.4	4.7	38.9	33.1
# Stixels (10^3)	KITTI	226	226	0.6	0.6	0.6	0.6
	Cityscapes	1 k	1 k	1.4	1.5	1.3	1.4
	SYNTHIA-SF	1 k	1 k	1.5	1.7	1.2	1.3

6 Acknowledgements

This work has been partially supported by Ministerio de Economía y Competitividad MINECO-Spain under contract TIN2014-53234-C2-1-R and TRA2014-57088-C2-1-R, the Generalitat de Catalunya projects 2014-SGR-1506 and 2014-SGR-1562, we also thank CERCA Programme / Generalitat de Catalunya, NVIDIA for the donation of the systems used in this work and SEBAP for the internship funding program. Finally, we thank Francisco Molero, Marc García, and the SYNTHIA team for the dataset generation.

References

[1] Hernán Badino, Uwe Franke, and David Pfeiffer. The stixel world - A compact medium level representation of the 3D-world. In *Pattern Recognition, 31st DAGM Symposium, Jena, Germany, September 9-11, 2009. Proceedings*, pages 51–60, 2009. doi: 10.1007/978-3-642-03798-6_6. URL http://dx.doi.org/10.1007/978-3-642-03798-6_6.

[2] Rodrigo Benenson, Radu Timofte, and Luc J. Van Gool. Stixels estimation without depth map computation. In *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, pages 2010–2017, 2011. doi: 10.1109/ICCVW.2011.6130495. URL http://dx.doi.org/10.1109/ICCVW.2011.6130495.

[3] Dexmont Alejandro Pena Carrillo and Alistair Sutherland. Fast obstacle detection using sparse edge-based disparity maps. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 66–72. IEEE, 2016.

[4] Marius Cordts, Lukas Schneider, Markus Enzweiler, Uwe Franke, and Stefan Roth. Object-level priors for stixel generation. In *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings*, pages 172–183, 2014. doi: 10.1007/978-3-319-11752-2_14. URL http://dx.doi.org/10.1007/978-3-319-11752-2_14.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223, 2016. doi: 10.1109/CVPR.2016.350. URL http://dx.doi.org/10.1109/CVPR.2016.350.

[6] Marius Cordts, Timo Rehfeld, Lukas Schneider, David Pfeiffer, Markus Enzweiler, Stefan Roth, Marc Pollefeys, and Uwe Franke. The stixel world: A medium-level representation of traffic scenes. *Image and Vision Computing*, pages –, 2017. ISSN 0262-8856. doi: http://doi.org/10.1016/j.imavis.2017.01.009. URL http://www.sciencedirect.com/science/article/pii/S0262885617300331.

[7] Vikas Dhiman, Abhijit Kundu, Frank Dellaert, and Jason J Corso. Modern MAP inference methods for accurate and fast occupancy grid mapping on higher order factor graphs. 2014.

[8] Mark Everingham, S. M. Ali Eslami, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. doi: 10.1007/s11263-014-0733-5. URL http://dx.doi.org/10.1007/s11263-014-0733-5.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? the KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[10] Daniel Hernandez-Juarez, Alejandro Chacón, Antonio Espinosa, David Vázquez, Juan Carlos Moure, and Antonio M. López. Embedded real-time stereo estimation via semi-global matching on the GPU. In *International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA*, pages 143–153, 2016. doi: 10.1016/j.procs.2016.05.305. URL http://dx.doi.org/10.1016/j.procs.2016.05.305.

[11] Daniel Hernandez-Juarez, Antonio Espinosa, Juan C. Moure, David Vázquez, and Antonio Manuel López. GPU-accelerated real-time stixel computation. In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 1054–1062, 2017. doi: 10.1109/WACV.2017.122. URL https://doi.org/10.1109/WACV.2017.122.

[12] Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2008. doi: 10.1109/TPAMI.2007.1166. URL http://dx.doi.org/10.1109/TPAMI.2007.1166.

[13] Oana Ignat. Disparity image segmentation for free-space detection. In *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 217–224, Sept 2016. doi: 10.1109/ICCP.2016.7737150.

[14] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014,*

Columbus, OH, USA, June 23-28, 2014, pages 89–96, 2014. doi: 10.1109/CVPR.2014.19. URL http://dx.doi.org/10.1109/CVPR.2014.19.

[15] Dan Levi, Noa Garnett, and Ethan Fetaya. Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 109.1–109.12, 2015. doi: 10.5244/C.29.109. URL http://dx.doi.org/10.5244/C.29.109.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965. URL http://dx.doi.org/10.1109/CVPR.2015.7298965.

[17] Maximilian Muffert, Nicolai Schneider, and Uwe Franke. Stix-fusion: A probabilistic stixel integration technique. In *Canadian Conference on Computer and Robot Vision, CRV 2014, Montreal, QC, Canada, May 6-9, 2014*, pages 16–23, 2014. doi: 10.1109/CRV.2014.11. URL http://dx.doi.org/10.1109/CRV.2014.11.

[18] Dominik Nuss, Ting Yuan, Gunther Krehl, Manuel Stuebler, Stephan Reuter, and Klaus Dietmayer. Fusion of laser and radar sensor data with a sequential monte carlo bayesian occupancy filter. 2015.

[19] David Pfeiffer. *The Stixel World - A Compact Medium-level Representation for Efficiently Modeling Three-dimensional Environments*. PhD thesis, Hu Berlin, 2014.

[20] David Pfeiffer and Uwe Franke. Towards a global optimal multi-layer stixel representation of dense 3D data. In *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*, pages 1–12, 2011. doi: 10.5244/C.25.51. URL http://dx.doi.org/10.5244/C.25.51.

[21] David Pfeiffer, Stefan Gehrig, and Nicolai Schneider. Exploiting the power of stereo confidences. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 297–304, 2013. doi: 10.1109/CVPR.2013.45. URL http://dx.doi.org/10.1109/CVPR.2013.45.

[22] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. 2016.

[23] Lukas Schneider, Marius Cordts, Timo Rehfeld, David Pfeiffer, Markus Enzweiler, Uwe Franke, Marc Pollefeys, and Stefan Roth. Semantic stixels: Depth is not enough. In *2016 IEEE Intelligent Vehicles Symposium, IV 2016, Gotenburg, Sweden, June 19-22, 2016*, pages 110–117, 2016. doi: 10.1109/IVS.2016.7535373. URL http://dx.doi.org/10.1109/IVS.2016.7535373.

[24] Sebastian Thrun. Robotic mapping: A survey. In *Exploring artificial intelligence in the new millennium*. 2002.