STIXELNEXT

Toward Monocular Low-Weight Perception for Object Segmentation and Free Space Detection

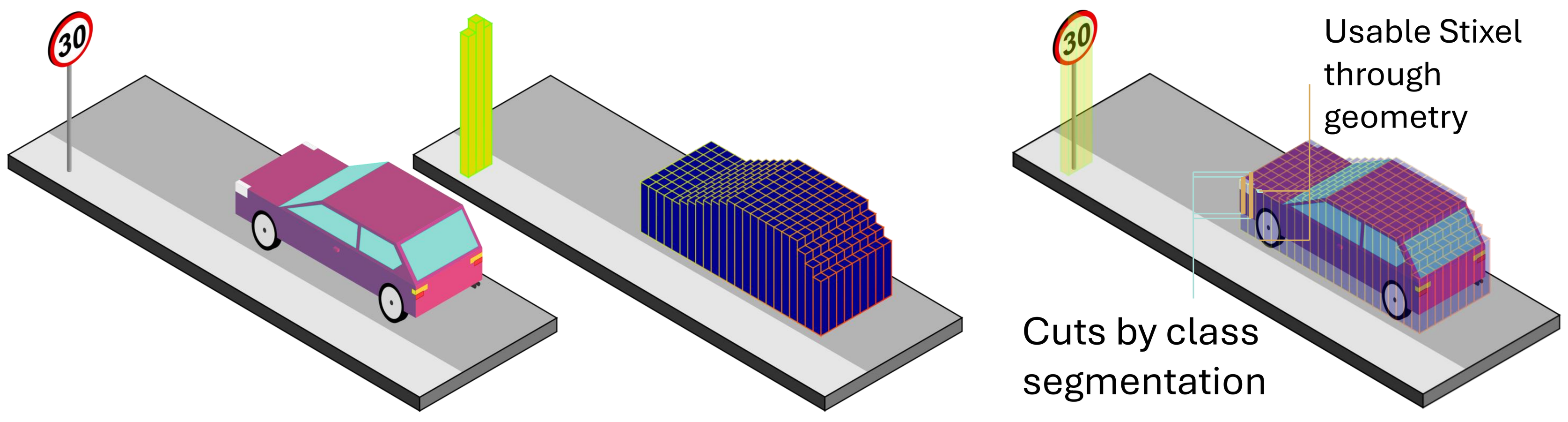
Marcel Vosshans, Omar Ait-Aider, Youcef Mezouar and Markus Enzweiler

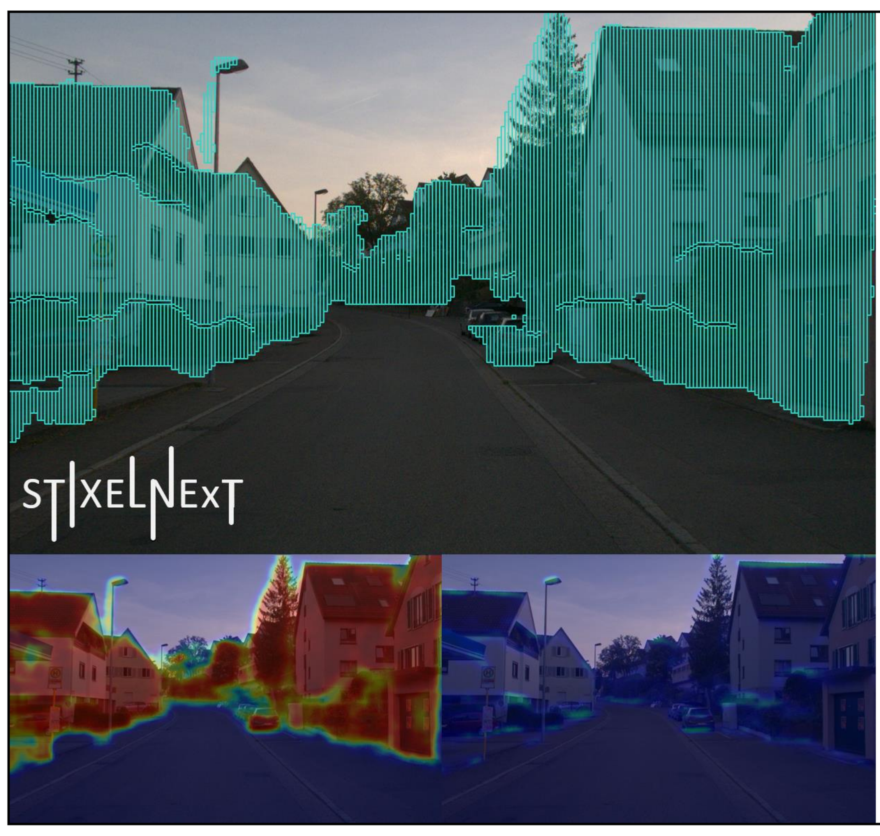
Jeju IV2024 35th IEEE Intelligent Vehicles Symposium

ESSLINGEN UNIVERSITY
UCA UNIVERSITÉ

About Stixel (Stick + Pixel)

A Stixel is a medium representation of the digital world. Unlike a voxel, certain assumptions allow Stixels to simplify the representation to the essence of needed perception information. Stixels are defined by two main rules: they stand on the ground, and the ground has a constant slope [1].





INTRODUCTION

Object Segmentation

We present a novel object segmentation method using monocular images, eliminating manual labeling (automatic LiDAR annotations) and enabling quick training with minimal data.

The corresponding heat maps (lower, left: object occupancy, right: object differentiation). The image on top shows both the passable area and the segmented objects.

Automatic Ground Truth

Our model uses LiDAR solely for ground truth generation during training and operates exclusively on monocular images thereafter. Building on the StixelNet [2] concept, we improved the ground truth generation process by breaking it into sub-problems and solving them step-by-step.

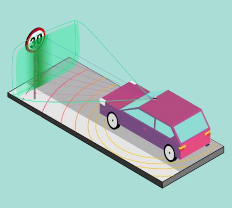
1. LiDAR-camera projection
2. Detect ground with RANSAC [3] – Stixel bottom
3. Hidden point removal [4]
4. Cluster objects (angle wise) based on inclination with DBSCAN [5] and identify Stixel top

Ground Truth

Generation

Ground Truth

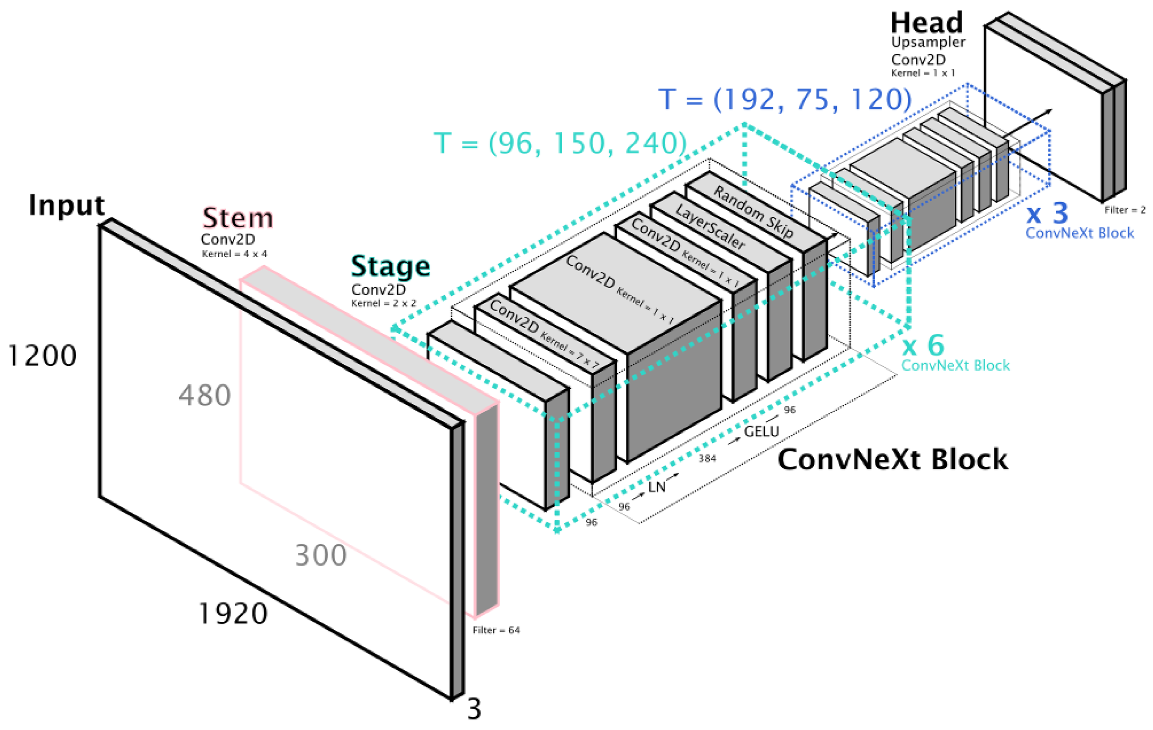
Setup



We created a Dataset with a dense LiDAR and Stereo data for the baseline.

StixelNExT

The StixelNExT architecture, derived from ConvNeXT [6], features two output channels in its architecture. One layer detects the presence of objects through a heatmap, while the other divides the objects into individual instances via postprocessing. We evaluated our approach using public datasets like KITTI and primarily on our own custom-created dataset.



EXPERIMENTS

Loss Function

We used Binary Cross Entropy (BCE) loss similar to object detection:

$$L_{BCE}(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} L_i$$

$$L_i = y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

Additionally, we added a column summarizing term to enforce confidence:

$$L_{Sum}(y,\hat{y}) = -\frac{1}{N}\sum_{u=1}^{m}\sum_{v=1}^{n} T_{uv}$$

Finally, we added both weighted losses:

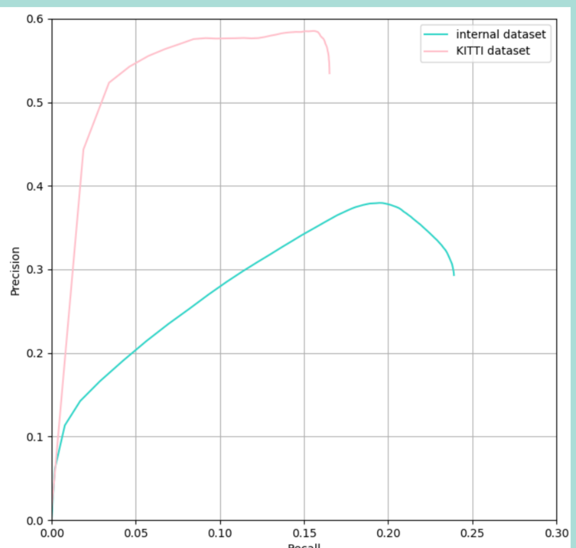
$$L(y,\hat{y}) = \alpha \cdot L_{BCE_{occ}} + \beta \cdot L_{Sum_{occ}} + \gamma \cdot L_{BCE_{cut}}$$

StixelNExT

A neural network for predicting multi-layer Stixels from a monocular camera and a novel method for automated LiDAR-based ground truth generation.

TABLE II
EVALUATION ON THE TEST SPLIT OF THE INTERNAL DATASET (2040 SAMPLES) AND THE KITTI DATASET (964 SAMPLES) WITH $IoU = 0.5$ AND VARYING THRESHOLD t

	Precision	Recall	F1
internal dataset			
StixelNExT, $t = 0.32$	0.2931	**0.2391**	0.2546
StixelNExT, $t = 0.45$	0.3348	0.2291	**0.2639**
StixelNExT, $t = 0.68$	**0.3796**	0.1948	0.2507
Stereo	0.1431	0.2325	0.1691
KITTI dataset			
StixelNExT, $t = 0.42$	0.5347	**0.1651**	0.2505
StixelNExT, $t = 0.58$	0.5589	0.1644	**0.2522**
StixelNExT, $t = 0.70$	**0.5854**	0.1559	0.2447
Stereo	0.1265	0.0857	0.1011



EXPERIMENTS

The Naive Evaluation

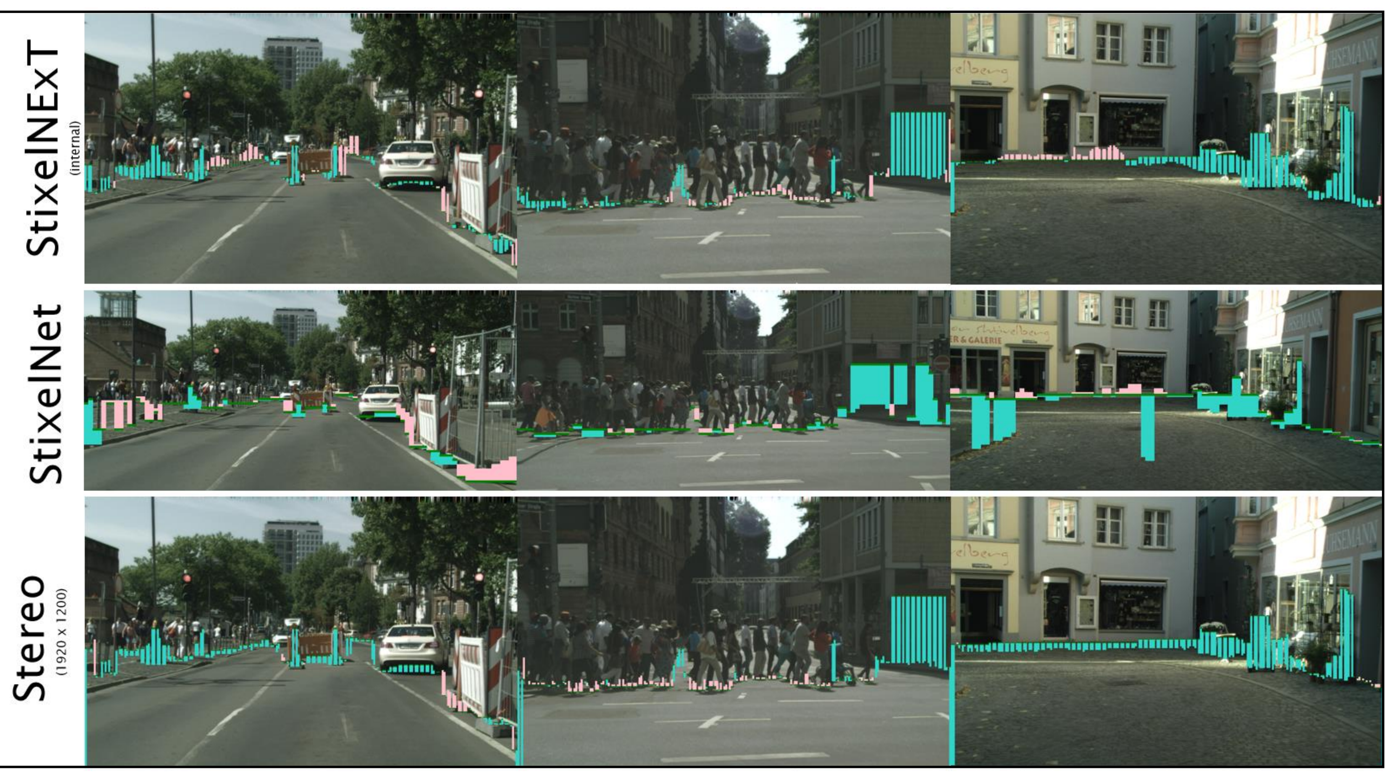
We evaluated our model using an Intersection over Union (IoU) metric, treating Stixels as one-dimensional objects similar to bounding boxes. See conclusions graphic for samples.

EXPERIMENTS

The Fairer Evaluation

We included a third dataset with pixel-precise segmentation and stereo data, like the Cityscapes dataset, categorizing semantic labels into passable areas and obstacles. Stereo [7], StixelNet [2], and StixelNExT were used to detect free space.

	Score Σ [%]	σ [%]
StixelNExT @ $t = 0.45$, internal	91.070	8.023
StixelNExT @ $t = 0.58$, KITTI	91.661	4.967
StixelNet, KITTI	**94.370**	3.566
Stereo $(1200x1920)$ px	91.054	**2.825**
Stereo $(370x800)$ px	88.528	9.664



CONCLUSIONS

Our work with StixelNExT demonstrates 2D multi-layer Stixel localization in images, achieved through efficient training with LiDAR data and no manual labeling. Although the current model lacks depth prediction, we have seen preliminary successes and continue to focus on this area.

CONCLUSIONS

Future Works

Our project established a baseline comparison as the first step, with the next step involving the addition of depth estimation. Future research will focus on adding end-to-end monocular depth estimation (like e.g. [8]) to StixelNExT, potentially boosting its capabilities significantly and enable Advanced Driver Assistance Systems (ADAS) tasks.

IIS Department of Computer Science and Engineering, Institute for Intelligent Systems, University of applied Science Esslingen, Germany

INSTITUT PASCAL — Institut Pascal ISPR (Image, Systems of Perception, Robotics), Université Clermont Auvergne INP / CNRS, France

[1] H. Badino, U. Franke, and D. Pfeiffer, "The Stixel World - A Compact Medium Level Representation of the 3D-World," in Pattern Recognition, J. Denzler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 5748, pp. 51–60.
[2] D. Levi, N. Garnett, and E. Fetaya, "StixelNet: A Deep Convolutional Network for Obstacle Detection and Road Segmentation," in British Machine Vision Conference 2015. Swansea: British Machine Vision Association, 2015, pp. 109.1–109.12.
[3] M. A. Fischler and R. C. Bolles, Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, in Readings in Computer Vision. Elsevier, 1987, pp. 726±740.
[4] S. Katz, A. Tal, and R. Basri, Direct visibility of point sets, in ACM SIGGRAPH 2007 papers. San Diego California: ACM, Jul. 2007, p. 24.
[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A density based algorithm for discovering clusters in large spatial databases with noise, in kdd, vol. 96, 1996, pp. 226±231.
[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, A ConvNet for the 2020s, in Conference on Computer Vision and Pattern Recognition (CVPR), 2022, publisher: arXiv Version Number: 2.
[7] gishi523, Multi-Stixel-World, GitHub repository, 2019. [Online]. Available: https://github.com/gishi523/multilayer-stixel-world as implementation of D. Pfeiffer and U. Franke, Towards a Global Optimal Multi-Layer Stixel Representation of Dense 3D Data, in British Machine Vision Conference 2011. Dundee: British Machine Vision Association, 2011, pp. 51.1±51.12.
[8] Shao, Shuwei, Zhongcai Pei, Weihai Chen, Qiang Liu, Haosong Yue, and Zhengguo Li. "Sparse Pseudo-LiDAR Depth Assisted Monocular Depth Estimation." IEEE Transactions on Intelligent Vehicles 9, no. 1 (January 2024): 917–29.