Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer

Jun Xie[1] Martin Kiefel[2] Ming-Ting Sun[1] Andreas Geiger[2]
[1]University of Washington [2]MPI for Intelligent Systems Tübingen
{junx,mts}@uw.edu {martin.kiefel,andreas.geiger}@tue.mpg.de

Abstract

Semantic annotations are vital for training models for object recognition, semantic segmentation or scene understanding. Unfortunately, pixelwise annotation of images at very large scale is labor-intensive and only little labeled data is available, particularly at instance level and for street scenes. In this paper, we propose to tackle this problem by lifting the semantic instance labeling task from 2D into 3D. Given reconstructions from stereo or laser data, we annotate static 3D scene elements with rough bounding primitives and develop a model which transfers this information into the image domain. We leverage our method to obtain 2D labels for a novel suburban video dataset which we have collected, resulting in 400k semantic and instance image annotations. A comparison of our method to state-of-the-art label transfer baselines reveals that 3D information enables more efficient annotation while at the same time resulting in improved accuracy and time-coherent labels.

1. Introduction

The revolutionary success of high-capacity deep learning architectures [23, 26, 52] may flag the beginning of a paradigm shift in computer vision. Rather than developing methods for solving a certain task, future research could be directed towards teaching a "universal program" (e.g., a deep network) a mapping from input to output space. One fundamental question arising in this context is how the required ground truth labels for training these models can be generated at very large scales (i.e., $>$ 100k images). While for some tasks large annotated datasets are already available today (e.g., image classification [34]), other tasks such as semantic segmentation of street scenes lack this information as human annotation is labor-intensive. We refer to this phenomenon as the **curse of dataset annotation** (Fig. 1).

One option to circumvent this problem is to exploit auxiliary tasks for which large annotated datasets are available. While generalization to the target domain can be achieved to some extent, discriminative cues which solve the auxiliary problem will dominate the learned representation [51].

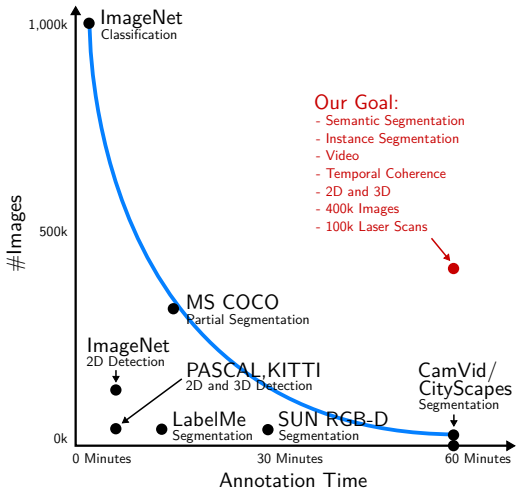


Figure 1: **The Curse of Dataset Annotation.**

A second option is the creation of synthetic datasets. Unfortunately, our community still lacks rich generative image formation models which are able to produce realistic and diverse imagery from the true underlying distribution of the 3D world we live in. In this paper, we therefore propose an alternative approach which leverages additional 3D information to simplify the 2D annotation task.

Recently, applications such as autonomous cars and humanoid robots have attracted significant attention. For research in these applications, a street view video dataset with dense semantic labels will be very useful. Motivated by those needs, our work focuses on the challenging task of semantic and instance video annotation of street scenes for which pixelwise labeling requires up to 60 minutes per image for a human annotator as acknowledged in [2]. Inspired by the easy usage of 3D modeling tools (Blender, SketchUp) we propose to annotate scenes directly in 3D and then transfer this knowledge back into the image domain. The required 3D information can be obtained from various sources including structure-from-motion (SfM), stereo or laser scanners. This approach has several advantages over labeling in 2D: First, objects often project into several images of the video sequence, thus lowering annotation efforts considerably. Further, the obtained 2D instance annotations

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.401

IEEE computer society
ALIGN – EXHIBIT 1013

are temporally coherent as they are associated with a single object in 3D. And finally, our 3D annotations might be useful by themselves for reasoning in 3D [14, 50] or to enrich 2D annotations with approximate 3D geometry.

Unfortunately, obtaining dense and accurate 2D labels from sparse noisy point clouds and coarse 3D annotations is a challenging task by itself. Towards solving this problem, we propose a non-local multi-field CRF model which reasons jointly about semantic and instance labels of all 3D points and all pixels in the image as illustrated in Fig. 2. This approach offers several advantages over methods which reason purely in 2D [1, 44]: Occluders and occludees which exhibit complex boundaries when projected onto the image plane (e.g., tree in front of a building) are often easier to separate in 3D. Besides, our approach is not affected by missing labels due to occlusions or drift in optical flow. Further, our model allows to specify a tractable semantic instance loss for principled and efficient end-to-end parameter learning. And finally, the probabilistic nature of our model allows for estimating label uncertainties which can be used to increase label accuracy when only a subset of the pixels require a label. In summary, we make the following two contributions in this paper:

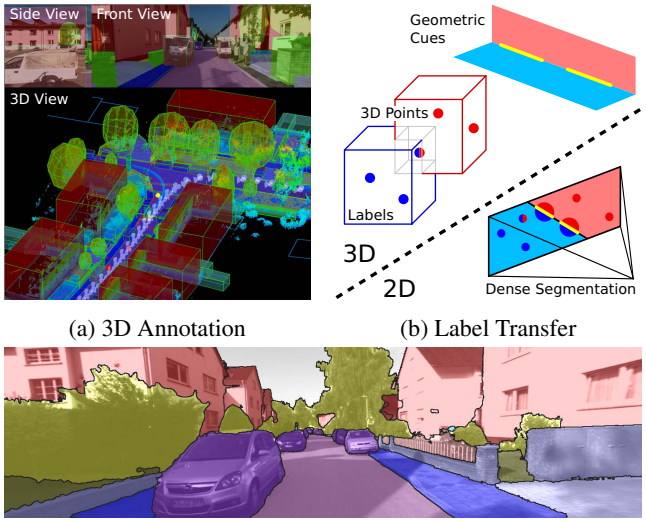
- We present a novel geo-registered dataset of suburban scenes recorded by a moving platform. The dataset comprises over 400k images and over 100k laser scans, and we provide semantic 3D annotations for all static scene elements.
- We propose a method which is able to transfer these labels from 3D into 2D, yielding pixelwise semantic instance annotations. We demonstrate the potential of our approach in ablation studies and with respect to several 2D and 3D baselines.

We make our code, dataset and annotations publicly available at http://www.cvlibs.net/projects/label_transfer.

2. Related Work

In this section, we first review semi-supervised video annotation methods, followed by an overview over existing semantic and instance segmentation datasets.

Methods: Compared to annotating individual images [16, 25, 48], video sequences offer the advantage of temporal coherence between adjacent frames. Label propagation techniques exploit this fact by transferring labels from a sparse set of annotated keyframes to all unlabeled frames based on color and motion information. While in some works a single foreground object is assumed [19, 42], here we focus on methods which can handle multiple object categories. Towards this goal, [2, 8] proposed a coupled Bayesian network based on video epitomes and semantic regions to propagate label information between two annotated keyframes.



(a) 3D Annotation (b) Label Transfer

(c) 2D Annotation

Figure 2: **3D to 2D Label Transfer:** (a) We annotate all objects in 3D using bounding primitives. (b) Our model then transfers this information into 2D by jointly reasoning about 3D geometric cues, sparse 3D points, as well as image pixels. (c) This allows us to infer temporally consistent semantic instance annotations for every frame in the video.

To better account for errors in label propagation, [31] proposed a hierarchy of local classifiers for this task and [1] leveraged a mixture-of-tree model for temporal association. The problem of selecting the most promising key frames for annotation has been considered in [44].

In contrast to the aforementioned methods which propagate labels in 2D, in this paper we propose to annotate directly in 3D and then project these annotations into the 2D domain. While this approach requires a source of 3D information (e.g., SfM, stereo, laser), it is able to produce more accurate semantic and temporally consistent instance annotations. Further, our experiments indicate that annotation in 3D is more time efficient than labeling in 2D as scene elements can be separated more easily and often project into many images of the input video sequence.

There exists little work on 3D to 2D label transfer. A notable exception is the approach of Chen et al. [9], where annotations from KITTI [12] as well as 3D car models are leveraged to infer separate figure-ground segmentations for all vehicles in the image. In comparison, our approach reasons jointly about all objects in the scene and also handles categories for which CAD models or 3D point measurements are unavailable (e.g., "Tree", "Sky"). In the context of street view image segmentation, Xiao et al. [47] present a hybrid method where annotated 3D points from structure-from-motion are projected onto superpixels in the image and users interactively correct wrong predictions with 2D scribbles. However, as no occlusion reasoning

is performed, their method can only be applied to scenes with little variations in depth (e.g., facades). Other methods [6, 27, 29, 30, 32] which model the interaction between image pixels and 3D points focus primarily on improving classification performance or efficiency by exploiting multiple input modalities while our goal is to transfer ambiguous 3D primitive labels to every pixel in the image.

Datasets: While some datasets such as PASCAL VOC [11] or MS COCO [24] provide semantic labels for a subset of pixels in the image, here we focus on datasets with dense semantic annotations. Most of these datasets provide only a small number ($\sim$ 1k) of accurately annotated indoor [39] or outdoor [15, 38] images. A notable exception is LabelMe [35] with more than 10k images labeled using crowdsourcing techniques. Compared to the smaller datasets, however, not all images are densely annotated, quality varies heavily amongst annotators, and polygons have been chosen over pixels as more efficient but less accurate representation.

A number of works have also considered the annotation of video sequences [7, 40, 46]. In [46], eight RGB-D sequences of indoor scenes have been manually annotated using an interactive tool which propagates 2D polygons from one frame to another. The recently proposed SUN RGB-D dataset [40] provides labeled 2D polygons as well as 3D cuboids for 10k RGB-D images captured indoors. For street scenes, less annotated data is available [3, 29, 30, 33, 43]. While KITTI [13] provides semantic information only for a few object categories[1], CamVid [7] offers pixel-accurate labels, but without instances and for a very limited number of frames. Very recently, the Cityscapes dataset [10] has been proposed with 5k manually annotated individual 2D images of street scenes[2]. Our dataset differs from Cityscapes in that we provide temporally coherent semantic instance annotations at a much larger scale as well as omnidirectional imagery, 3D laser scans and 3D annotations which might also be directly useful for reasoning in 3D. While [10] focuses on inner-city scenes, our dataset comprises mainly suburban areas, thus both datasets complement each other.

3. Method

In this work, we are interested in generating semantic instance annotations for urban scenes at large scale by transferring labels from sparse 3D point clouds into the images. In particular, we focus on static scene elements which dominate suburban scenes. Dynamic objects could be handled via 3D models [9, 28] but as our dataset comprises little dynamic objects we leave this extension for future work. This section describes our data collection efforts, our 3D annotation process, as well as the proposed label transfer model.

[1] http://www.cvlibs.net/datasets/kitti/

[2] http://www.cityscapes-dataset.net/

3.1. Data Collection

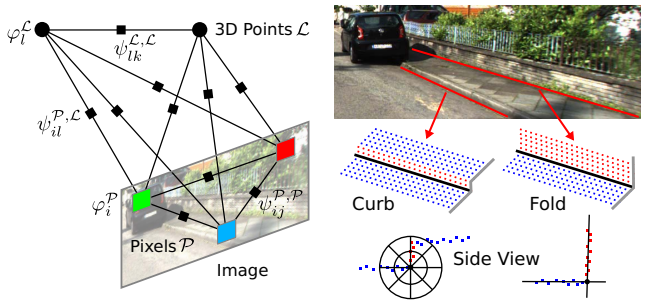
For our data collection, we equipped a station wagon with one 180° fisheye camera to each side and a 90° perspective stereo camera (baseline 60 cm) to the front. Furthermore, we mounted a Velodyne HDL-64E and a SICK LMS 200 laser scanning unit in pushbroom configuration on top of the roof. This setup is similar to the one used in KITTI [12, 13], except that we gain a full 360° field of view due to the additional fisheye cameras and the pushbroom laser scanner while KITTI only provides perspective images and Velodyne laser scans with a 26.8° vertical field of view. Compared to omnidirectional camera systems [36, 37] our setup benefits from increased resolution. Approximate localization is provided by an IMU/GPS measurement unit.

Using this setup, we recorded several suburbs of a mid-size city corresponding to over 400k images and 100k laser scans. We estimated all vehicle and camera poses using structure-from-motion [17]. More specifically, we minimize 3D reprojection errors based on all feature matches while regularizing against the GPS solution. This results in accurate georegistered camera poses. While our label transfer approach does not assume geolocalization, geospatial information[3] can facilitate the 3D annotation task.

3.2. Annotation

We augmented our dataset with 3D annotations in the form of bounding primitives, i.e., we placed cuboids and ellipsoids around objects in 3D and assigned a semantic label to each of them. More specifically, we asked a group of annotators to tightly enclose the 3D points belonging to an object by the respective primitive. For this purpose, we developed a 3D annotation tool based on WebGL (see Fig. 2a) which visualizes the colored point clouds (obtained by projecting the 3D points back onto multiple images), two camera views, and provides tools to facilitate navigation and annotation. To enable efficient annotation, our primitives are rough approximations of the true object shapes and thus are allowed to overlap in 3D (see Fig. 2b). For stuff categories (e.g., "Road", "Sidewalk", "Grass") we allow users to draw 2D polygons in bird's eye view which are then extruded into 3D to better approximate the shape and to facilitate annotation. Ambiguities are resolved using our label transfer method described in the following section. Annotating a single batch comprising 200 laser scans and 800 images required about 3 hours. While the focus of this paper is on annotating static scene elements which cover the majority of pixels in general, our annotation GUI could be extended to a keyframe based dynamic 3D video annotation tool which visualizes point clouds and images over time akin to the annotation utility developed for labeling the KITTI dataset [12, 13].

[3] http://www.openstreetmap.org/



(a) Graphical Model

(b) 3D Curbs and Folds

Figure 3: **Label Transfer Model.** (a) Factor graph representation of our model. (b) 3D structures such as folds and curbs are leveraged to improve segmentation boundaries between the categories "Road", "Sidewalk" and "Wall".

3.3. Model

Given sparse point clouds and 3D annotations, we are interested in generating dense semantic instance annotations for all images. Towards this goal, we propose a CRF model which reasons jointly about the labels of the 3D points and all pixels in the image, leveraging the calibration and registration described in Section 3.1. Note that our 3D annotations are sparse and noisy, i.e., 3D points can carry none, one or multiple labels due to overlapping bounding primitives in 3D. The algorithm described in this section is designed to resolve these situations and infers marginal estimates for all 3D points and pixels in the image. In order to make our approach more robust in regions where appearance is not discriminative, we investigate additional geometric cues of the 3D point cloud such as 3D surface folds and curbs (see Fig. 3b). If detected, these cues can provide accurate boundaries between semantic classes in the image.

More formally, let $\mathcal{P}$, $\mathcal{L}$ and $\mathcal{F}$ denote the set of image pixels, sparse 3D points from laser/stereo, and detected 3D fold or curb segments, respectively. For each pixel $i \in \mathcal{P}$ and each 3D point $l \in \mathcal{L}$, we specify random variables s_i and s_l taking values from the set of semantic (or instance) labels $\{1, \dots, S\}$, where S denotes the number of classes. For instance inference, we assign a unique ID to each object which projects into the image. Thus, semantic and instance inference can be treated equally under our model and we will refer to both as "semantic labels" in the following.

Let $\mathbf{s} = \{s_i | i \in \mathcal{P}\} \cup \{s_l | l \in \mathcal{L}\}$ denote the set of semantic labels. Dropping all dependencies on the image and point cloud for clarity we specify our CRF in terms of the following Gibbs energy function:

$$E(\mathbf{s}) = \sum_{i \in \mathcal{P}} \varphi_i^{\mathcal{P}}(s_i) + \sum_{l \in \mathcal{L}} \varphi_l^{\mathcal{L}}(s_l) + \sum_{m \in \mathcal{F}} \sum_{i \in \mathcal{P}} \varphi_{mi}^{\mathcal{F}}(s_i) \quad (1)$$
$$+ \sum_{i,j \in \mathcal{P}} \psi_{ij}^{\mathcal{P},\mathcal{P}}(s_i, s_j) + \sum_{l,k \in \mathcal{L}} \psi_{lk}^{\mathcal{L},\mathcal{L}}(s_l, s_k) + \sum_{i \in \mathcal{P}, l \in \mathcal{L}} \psi_{il}^{\mathcal{P},\mathcal{L}}(s_i, s_l)$$

with unary potentials $\varphi(\cdot)$ and pairwise potentials $\psi(\cdot)$. For notational clarity, we omit all conditional dependencies on the input images, 3D points and 3D annotations.

Pixel Unary Potentials: The pixel unary potentials $\varphi_i^{\mathcal{P}}(s_i)$ encode the likelihood of pixel i taking label s_i

$$\varphi_i^{\mathcal{P}}(s_i) = w_1^{\mathcal{P}}(s_i)\, \xi_i^{\mathcal{P}}(s_i) - w_2^{\mathcal{P}}(s_i) \log p_i^{\mathcal{P}}(s_i) \quad (2)$$

where $w_1^{\mathcal{P}}$ and $w_2^{\mathcal{P}}$ denote learned feature weights. Our first constraint $\xi_i^{\mathcal{P}}(s_i)$ determines the set of admissible labels and is obtained by projecting the 3D bounding primitives (which are an upper bound on the objects' extent) into the image. We formulate the constraint via a binary feature $\xi_i^{\mathcal{P}}(s_i) \in \{0, 1\}$ which takes 0 for pixel i if its ray passes through a primitive of class s_i, and 1 otherwise.

In addition, we leverage appearance information by projecting all non-occluded sparse 3D points into all adjacent frames of the image sequence and training a pixel-wise classifier [38] based on these projections. This results in a per-pixel probability distribution over semantic labels $p_i^{\mathcal{P}}(s_i)$. The intuition behind this feature is that regions of the same semantic class are similar in adjacent frames and thus yield highly discriminative cues for the current frame.

3D Point Unary Potentials: The 3D point unary potentials $\varphi_l^{\mathcal{L}}(s_l)$ encode the likelihood of 3D point l taking label s_l:

$$\varphi_l^{\mathcal{L}}(s_l) = -w^{\mathcal{L}}(s_l)\, \xi_l^{\mathcal{L}}(s_l) \quad (3)$$

where $\xi_l^{\mathcal{L}}(s_l)$ denotes a feature which takes 0 if the 3D point l lies within a 3D primitive of class s_l, and 1 otherwise. As the "sky" class can't be modeled with primitives we set $\xi_l^{\mathcal{L}}(s_l)$ to 0 if s_l takes the label "sky". Additionally, we create "virtual sky points" at infinity for all pixels whose ray doesn't intersect any 3D primitive. Note that these pixels must correspond to sky regions as we assume that each object is completely contained in one or several bounding 3D primitive(s).

Geometric Unary Potentials: We encourage label changes at curbs or folds which we detect in 3D using plane fitting as described in the supplementary document. Given the projections into 2D, we introduce the following constraint:

$$\varphi_{mi}^{\mathcal{F}}(s_i) = w^{\mathcal{F}} \frac{[\mathbf{p}_i \in \mathcal{R}_m \wedge \nu_m(\mathbf{p}_i) \neq s_i]}{\exp\{dist(\mathbf{p}_i, \boldsymbol{\pi}_m)\}} \quad (4)$$

Here, $[\cdot]$ is the Iverson bracket, $\mathbf{p}_i$ denotes the 2D location of pixel i and $\mathcal{R}_m$ represents a 2D disc around curb or fold segment m projected into 2D (yielding a line segment $\boldsymbol{\pi}_m$) as illustrated in Fig. 4. ν_k is a function which takes as input a pixel location and returns the semantic label predicted by fold m. More specifically, we project the 3D fold into 2D and compute the majority label at its two sides from the sparse projected 3D points. The denominator in Eq. 4 ensures a penalty decay towards the disc boundaries.

Figure 4: **Geometric Unary Potentials.** Left: We encourage label changes at 3D curbs or folds after projection into the image domain. Right: This constraint ($\varphi^{\mathcal{F}}_{mi}$) is implemented by pixel unary potentials inside each minimum bounding disc $\mathcal{R}_m$ around each 2D curb or fold segment m.

Pixel Pairwise Potentials: Our dense pairwise term encourages semantic label coherence and connects all pixels in the image via Gaussian edge kernels

$$\psi^{\mathcal{P},\mathcal{P}}_{ij}(s_i,s_j) \;=\; w^{\mathcal{P},\mathcal{P}}_1(s_i,s_j)\exp\left\{-\frac{\|\mathbf{p}_i-\mathbf{p}_j\|^2}{2\,\theta^{\mathcal{P},\mathcal{P}}_1}\right\}$$
$$+\,w^{\mathcal{P},\mathcal{P}}_2(s_i,s_j)\exp\left\{-\frac{\|\mathbf{p}_i-\mathbf{p}_j\|^2}{2\,\theta^{\mathcal{P},\mathcal{P}}_2}-\frac{\|\mathbf{c}_i-\mathbf{c}_j\|^2}{2\,\theta^{\mathcal{P},\mathcal{P}}_3}\right\} \quad (5)$$

where $\mathbf{p}_i$ is the 2D location of pixel i and $\mathbf{c}_i$ denotes its color value. Further, $w^{\mathcal{P},\mathcal{P}}_1$ and $w^{\mathcal{P},\mathcal{P}}_2$ are learned pairwise feature weights and $\theta^{\mathcal{P},\mathcal{P}}$ parametrizes the kernel width.

3D Pairwise Potentials: Similarly, we apply a Gaussian edge kernel to encourage label consistency between 3D points based on their 3D location and surface normals

$$\psi^{\mathcal{L},\mathcal{L}}_{lk}(s_l,s_k) \;=\; w^{\mathcal{L},\mathcal{L}}(s_l,s_k) \quad (6)$$
$$\times\;\exp\left\{-\frac{\|\mathbf{p}^{3d}_l-\mathbf{p}^{3d}_k\|^2}{2\,\theta^{\mathcal{L},\mathcal{L}}_1}-\frac{(n_l-n_k)^2}{2\,\theta^{\mathcal{L},\mathcal{L}}_2}\right\}$$

where $\mathbf{p}^{3d}_l$ is the 3D location of point l and n_l denotes the vertical (up) component of its normal. We use the normal's z-component as it is the most discriminative cue for indicating label changes between horizontal (e.g., road, sidewalk) and vertical (e.g., side of car, wall) surfaces. We estimate the respective normals using principle component analysis in a local neighborhood around each 3D point.

2D/3D Pairwise Potentials: Finally, we encourage coherence between all 3D points and the image pixels

$$\psi^{\mathcal{P},\mathcal{L}}_{il}(s_i,s_l) = w^{\mathcal{P},\mathcal{L}}(s_i,s_l)\exp\left\{-\frac{\|\mathbf{p}_i-\boldsymbol{\pi}_l\|^2}{2\,\theta^{\mathcal{P},\mathcal{L}}}\right\} \quad (7)$$

where $\boldsymbol{\pi}_l$ denotes the projection of the 3D laser or stereo point l onto the image plane. Importantly, we project only points into the image which are likely to be visible. We determine these points by meshing the 3D point cloud using the ball-pivoting method of Bernardini et al. [4], considering only 3D points in front of the mesh. We also tried state-of-the-art multi-view reconstruction approaches [20] for mesh generation, but obtained better results with the described meshing approach.

3.4. Learning and Inference

This section describes inference and parameter estimation in our label transfer model.

Inference: At test time, we are interested in estimating the marginal distribution of each semantic or instance label in $\mathbf{s}$ under our model, specified by the Gibbs distribution defined in Eq. 1. The most likely configuration can then be estimated by variable-wise maximization of these marginals. As our graphical model is loopy, exact inference in polynomial time is intractable. We resort to variational inference and approximate the probability distribution on $\mathbf{s}$ by replacing it with a factorized mean field distribution $Q(\mathbf{s}) = \prod_{i\in\mathcal{P}\cup\mathcal{L}} Q_i(s_i)$. This mean field approximation can be computed efficiently using bilateral filtering [22]. As our model comprises three sets of densely connected variables (namely $\mathcal{P}$, $\mathcal{L}$ and $\mathcal{P}\leftrightarrow\mathcal{L}$), we exploit the algorithm of [21,45] which generalizes [22] to multiple fields.

Learning: We employ empirical risk minimization in order to learn the parameters in our model, considering the univariate logistic loss, defined as $\Delta(s) = -\log(P(s))$ where $P(\cdot)$ denotes the marginal distribution at the respective site. Let us subsume all model parameters into $\Theta = \{w^{\mathcal{P}}_1, w^{\mathcal{P}}_2, w^{\mathcal{L}}, w^{\mathcal{F}}, w^{\mathcal{P},\mathcal{P}}_1, w^{\mathcal{P},\mathcal{P}}_2, w^{\mathcal{P},\mathcal{L}}, w^{\mathcal{L},\mathcal{L}}\}$. We define our minimization objective $f(\Theta)$ as the regularized univariate logistic loss:

$$f(\Theta) \;=\; \sum_{n=1}^{N}\sum_{i\in\mathcal{P}} -\log\left(Q_{n,i}(s^*_{n,i})\right) + \lambda\, C(\Theta) \quad (8)$$

Here, N is the number of training images, $s^*_{n,i}$ denotes the ground truth semantic label and $Q_{n,i}(\cdot)$ the approximate marginal at pixel i in image n, calculated via mean field approximation. $C(\Theta)$ is a quadratic regularizer on the parameter vector Θ. We whiten all features and use a single value λ which we select via cross-validation on the training set. For learning the instance segmentation parameters we exploit the same loss $f(\Theta)$ as for semantic segmentation. For instance segmentation, we assign unique labels to each individual object, e.g., different cars will be assigned different labels even if they occlude each other. In order to associate 2D ground truth instances with 3D instances we project all visible 3D points into the image and find a consensus via the majority vote which gave good results in practice. As the number of instances per semantic class varies between images, we learn intra- and inter-class pairwise potentials using parameter tying. We optimize the objective function $f(\Theta)$ using stochastic gradient descent and obtain $\partial Q/\partial\Theta$ using auto differentiation. We make use of the ADADELTA algorithm [49] with decay parameter 0.95 and $\epsilon = 10^{-8}$, and randomly sample a batch of 16 training images at each iteration for which all gradients can be computed in parallel.

4. Experimental Evaluation

In this section, we first evaluate our method in ablation studies and with respect to several label transfer baselines. Besides, we exploit the uncertainty in our predictions to increase accuracy for semi-dense predictions. Finally, we show some qualitative results of our method. As input to our method, we accumulate all laser measurements in a common world coordinate system and augment them with 3D points from stereo matching [18]. To reduce outliers, we consider only points up to 15 m distance, and apply left-right as well as forward-backward consistency checks over 5 frames. We fuse all 3D points into one global point cloud and remove all points which are closer than 5 cm to their nearest neighbor. For evaluation, we manually annotated 160 images from 8 different suburbs with dense pixel-wise ground truth. From the 160 frames, 120 frames have been labeled in equidistant steps of 5 frames for comparison with 2D label transfer methods. We learn the parameters in our and the baseline models using 2-fold cross validation at the sequence level to avoid any bias caused by the correlation of adjacent frames within a sequence. The kernel width parameters in our model have been chosen empirically as detailed in our supplementary document.

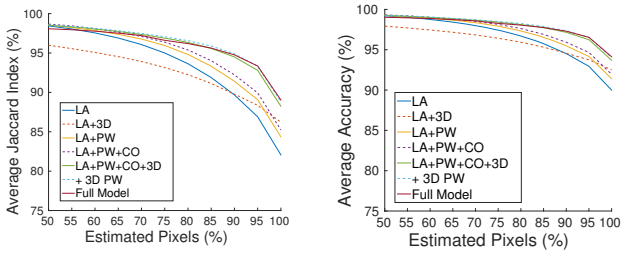
4.1. Quantitative Evaluation

This section presents our quantitative evaluation. We compare our method with respect to several baselines on the semantic and instance segmentation tasks.

Semantic Segmentation: For evaluating semantic segmentation performance, we map the 27 semantic labels in our 3D annotations to the most frequently occurring 14 categories (see supplementary material). We report the frequency of these classes in the supplementary material. We measure overall performance by the average Jaccard Index (JI) weighted by the class frequency and the average pixel accuracy (Acc).

The upper half of Table 1 shows results of several 2D to 2D label transfer methods on all 120 equidistantly labeled frames. Here, the task is to predict the center frame from two annotated images (± 5 frames corresponding to 0.5 seconds of driving or $\sim$ 5 meters travel distance). Our first baseline ("Label Prop.") is the label transfer approach presented in [44]. To ensure that all baselines have access to the same information, we do not select frames in an active fashion but use equidistantly spaced labels for all methods (the driving speed during recording was nearly constant). We construct a second baseline ("Sparse Track. + GC") using the feature tracking approach of [41] to propagate semantic labels from the two closest labeled frames to the target frame. To densify the label map, we apply graph cuts (GC) with contrast sensitive edge potentials [5].

In order to evaluate the value of 3D information, we implemented a third baseline ("3D Prop. + GC") which works



(a) Jaccard Index (b) Accuracy

Figure 5: **Performance wrt. Estimated Pixels.** This figure shows the average Jaccard Index (a) and the average accuracy (b) when estimating only a fraction of the pixels which is selected according to the uncertainty in our predictions.

similar to the previous one, but replaces the sparse tracking part with correspondences obtained by transferring pixels of the two closest labeled frames to the target image via the visible vertices of our 3D mesh followed by graph cuts propagation. Finally, we train the segmentation model of Krähenbühl et al. [22] ("Fully Conn. CRF") on all annotated adjacent frames of the *test sequence*.

From the 2D label transfer baselines, the mesh transfer method which uses projected 3D information performs best. Furthermore, and maybe surprisingly, the image-specific fully connected CRF model performs on par or even better than special purpose label transfer methods. According to our experiments, this is caused by the fact that optical flow (as used in [41,44]) often fails for street scenes like ours due to large displacements, perspective distortions, textureless regions and challenging lighting conditions. On the other hand, the fully connected model performs weaker for less frequent or textureless classes such as "Trailer" or "Box".

The bottom half of Table 1 compares the proposed method with respect to several 3D to 2D label transfer baselines which in contrast to the 2D to 2D label transfer methods exploit our 3D annotations and don't require equidistantly labeled 2D annotations. As evidenced by our results, simply projecting 3D primitives or meshes into the image and smoothing via GC does not perform well due to the crude approximation of the geometry ("3D Primitives + GC"; "3D Mesh + GC"). Better results are obtained when projecting the visible 3D points followed by spatial propagation ("3D Points + GC").

Finally, we observe that all baselines are outperformed by the proposed method (last row) in almost all categories. Importantly, note that the 2D methods require every 10th frame to be labeled, while our method (as well as the other 3D baselines) require 3D annotations in the form of 3D primitives. Assuming 60 minutes annotation time per image, this amounts to 20 hours of annotation time per batch of 200 frames when labeling one 2D image every 10th frame, while the respective 3D annotations for this scene can be obtained in less than 3 hours. Note that labeling each frame

Method	Road	Park	Sdwlk	Terr	Bldg	Vegt	Car	Trler	Carvn	Gate	Wall	Fence	Box	Sky	JI	Acc
Label Prop. [44]	93.4	51.8	73.5	58.3	80.2	69.9	61.5	22.4	42.3	30.6	45.3	45.7	32.5	89.6	74.4	84.4
Sparse Track. + GC [41]	89.6	37.1	69.0	54.2	84.6	79.5	78.2	2.5	35.3	3.2	38.9	32.9	7.0	91.0	77.8	87.3
3D Prop. + GC	91.3	44.5	74.0	62.4	86.2	81.8	81.6	5.2	38.6	12.7	47.4	42.0	15.0	88.8	80.2	88.9
Fully Conn. CRF [22]	88.5	37.8	68.4	55.8	85.5	79.8	76.8	2.5	30.6	2.9	38.3	32.4	0.0	**92.8**	77.9	87.4
3D Primitives + GC	78.7	46.4	43.9	46.5	54.9	55.4	55.1	72.3	54.6	51.0	40.2	52.3	40.2	55.4	56.4	72.1
3D Mesh + GC	92.1	66.4	72.4	66.1	69.1	74.9	87.7	88.9	88.5	61.9	51.4	60.7	30.4	46.4	72.5	82.6
3D Points + GC	93.1	72.4	78.9	72.4	81.9	77.2	88.1	92.4	91.1	70.2	66.8	68.7	62.4	69.9	80.5	89.0
Proposed Method	**95.3**	**80.6**	**86.4**	**81.0**	**90.9**	**86.9**	**91.5**	**94.9**	**91.8**	**73.6**	**78.9**	**79.4**	**73.0**	91.0	**89.2**	**94.2**

Table 1: **Comparison to Label Transfer Baselines on Semantic Segmentation Task.** We compare our method to 2D label transfer baselines (top) and to 3D to 2D label transfer baselines (bottom) on 120 consecutive images. See text for details.

Method	Road	Park	Sdwlk	Terr	Bldg	Vegt	Car	Trler	Carvn	Gate	Wall	Fence	Box	Sky	JI	Acc
LA	92.2	64.6	77.9	67.5	85.2	81.9	81.7	85.7	81.5	46.8	62.1	60.3	49.4	83.1	82.1	90.0
LA+3D	95.0	76.9	85.5	73.3	87.9	84.3	89.4	88.2	90.2	68.8	74.6	74.0	63.7	83.4	86.2	92.5
LA+PW	92.5	68.6	79.5	73.1	87.3	84.2	84.1	89.9	85.9	48.7	66.2	64.9	54.5	86.6	84.4	91.4
LA+PW+CO	93.0	72.7	81.2	73.8	87.7	84.5	85.7	90.9	88.4	57.7	70.4	69.6	57.6	86.9	85.2	92.0
LA+PW+CO+3D	93.2	78.6	85.0	76.3	90.6	86.7	89.1	90.9	92.7	68.5	77.8	78.9	67.8	90.7	88.2	93.7
+ 3D PW	94.9	80.1	85.9	80.0	90.6	87.0	91.2	91.3	93.8	72.6	78.1	78.5	69.3	90.8	88.8	94.0
Full Model	95.4	80.1	87.1	80.0	90.6	87.0	91.2	91.3	93.9	72.6	78.4	78.6	69.4	90.8	89.0	94.1
Full Model (90%)	98.1	92.3	94.7	92.4	95.3	93.5	96.5	95.8	97.6	83.7	90.7	90.7	84.0	94.6	94.9	97.4
Full Model (80%)	98.8	95.3	96.7	94.9	96.8	95.5	97.5	96.4	98.5	86.4	93.7	93.4	87.9	96.4	96.6	98.2
Full Model (70%)	**99.2**	**96.8**	**97.9**	**96.4**	**97.5**	**96.8**	**97.9**	**97.2**	**99.0**	**88.1**	**95.0**	**94.6**	**90.1**	**97.2**	**97.5**	**98.7**

Table 2: **Ablation Study on Semantic Segmentation Task.** This table shows the importance of the different components in our model on all 160 images. The components are abbreviated as follows: LA = local appearance ($p^{\mathcal{P}}$), PW = 2D pairwise constraints ($\psi^{\mathcal{P},\mathcal{P}}$), CO = 3D primitive constraints ($\xi^{\mathcal{P}}$), 3D = 3D points ($\varphi^{\mathcal{L}}$,$\psi^{\mathcal{P},\mathcal{L}}$), 3D PW = 3D pairwise constraints ($\psi^{\mathcal{L},\mathcal{L}}$), Full Model = all potentials including folds. Percentages denote fractions of estimated pixels. See text for details.

Method	Road	Park	Sdwlk	Terr	Bldg	Vegt	Car	Trler	Carvn	Gate	Wall	Fence	Box	Sky	JI	Acc
LA+3D	94.5	74.7	83.5	73.4	80.7	84.5	86.3	90.8	90.9	66.3	74.7	75.6	63.1	81.9	83.5	91.0
LA+PW+CO	92.8	70.3	79.8	73.9	64.9	84.6	82.2	90.7	87.1	51.7	67.8	66.6	24.7	88.0	78.4	87.4
LA+PW+CO+3D	94.6	78.4	84.2	78.4	86.3	87.6	90.8	93.0	93.3	70.9	77.6	79.4	68.6	91.1	87.5	93.3
+ 3D PW	95.1	**80.6**	85.3	**79.3**	86.4	**87.9**	91.5	93.0	93.6	**73.6**	78.1	79.0	70.4	90.7	87.9	93.5
Full Model	**95.7**	**80.6**	**86.9**	79.2	**86.4**	**87.9**	**91.5**	**93.1**	**93.6**	**73.6**	**78.5**	**79.1**	**70.5**	**90.7**	**88.1**	**93.6**

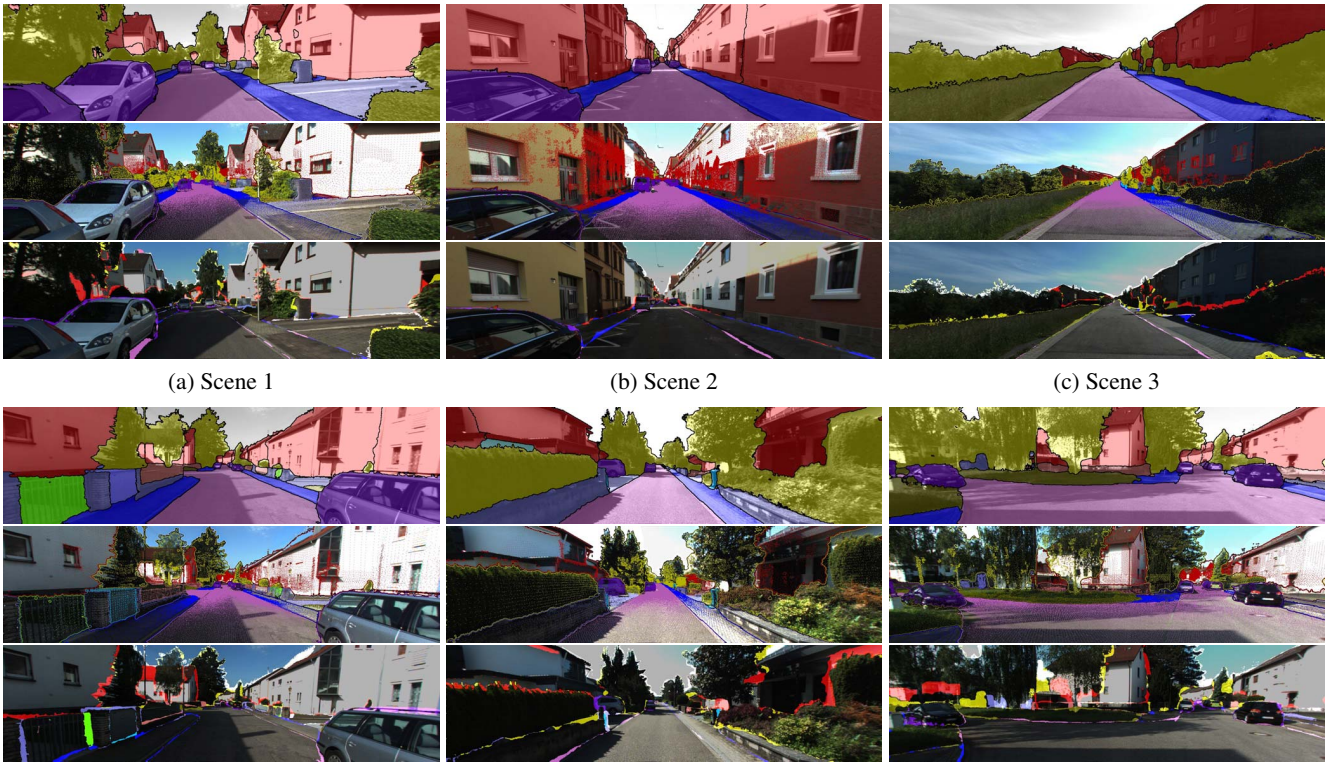
Table 3: **Ablation Study on Instance Segmentation Task** using the same abbreviations as in Table 2. See text for details.

of the sequence manually would require 200 hours. This gain multiplies with the frame rate and the number of cameras (our setup comprises four).

Ablation Study: We evaluate the importance of the individual components of our model in Table 2 (top). Starting with the appearance classifier trained on the projected sparse 3D points ($p^{\mathcal{P}}$), we incrementally add the terms related to the 3D points ($\varphi^{\mathcal{L}}$,$\psi^{\mathcal{P},\mathcal{L}}$), the semantic pairwise term between pixels ($\psi^{\mathcal{P},\mathcal{P}}$), the 3D primitive constraints ($\xi^{\mathcal{P}}$), the 3D pairwise constraints ($\psi^{\mathcal{L},\mathcal{L}}$) and finally the remaining terms ($\varphi^{\mathcal{F}}_{mi}$) as specified in Eq. 1. We note that each component is able to increase performance. As expected, we obtain the largest improvement by reasoning about the relationship between points in 3D and pixels in the image. Integrating 3D fold and curb detections improves road boundaries slightly.

Semi-dense Inference: Often, it is not necessary to label all pixels in every image for training a semantic segmentation model. In this section, we therefore leverage our model's awareness of label uncertainty to estimate semi-dense label maps with high accuracy. To quantify uncertainty, we measure the entropy of the label marginal distribution at every pixel. Sorting all pixels according to their entropy allows us to predict the most certain regions in the image. Table 2 (bottom) and Fig. 5 show our results when predicting only those parts of the image. Note how this helps to boost our performance to 94.9% JI and 97.4% accuracy when predicting at 90% pixel density. In contrast, uncertainty is not directly accessible in most of the baseline models as they are deterministic or rely on MAP estimates. The only exception is the "Fully Conn. CRF" baseline. We provide the corresponding experiment in the supplementary material.

Instance Segmentation: As time consistent 2D instance



(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5 (f) Scene 6

Figure 6: **Qualitative Results.** Each subfigure shows from top-to-bottom: the input image with inferred semantic instance segmentation, the projected 3D points and inferred semantic segmentation boundaries, as well as the errors with respect to 2D ground truth annotation where colors indicate ground truth labels. See supplementary material and text for details.

ground truth is hard to obtain, most existing 2D label transfer methods focus on the semantic segmentation problem. Therefore, we chose to evaluate instance segmentation performance in an ablation study. We annotated the classes "Building", "Car", "Trailer", "Caravan" and "Box" with instances in our 2D ground truth. While the remaining classes (e.g., "Road", "Sky") do not admit unambiguous instance labels, we also report their performance as our model reasons about all instance and semantic classes jointly. Table 3 shows our results. Note how the instance segmentation results are on par with the semantic segmentations, demonstrating our model's intra-class separation ability. Semi-dense instance results are provided in the supplementary.

4.2. Qualitative Evaluation

Fig. 6 illustrates our dense inference results qualitatively for 6 different scenes in terms of semantic instance segmentation. The last row shows the error maps where colors indicate the true label (see supplementary for color coding). While the proposed method is able to delineate most object boundaries satisfyingly, some challenges remain. Errors occur in low-contrast image regions with overlapping 3D annotations (scene 1: car/road boundary) and in regions where 3D points are absent due to sensor occlusion (scene 4: building roof). Another source of errors are inherent label ambiguities which occur for porous objects such as fences or trees (scene 6: tree boundary) where even 2D ground truth annotation is a hard and ambiguous task. Finally, also manual 2D annotations contain errors, in particular at complex boundaries which are hard to delineate (scene 4: trees, scene 5: hedge). However, note that our semi-dense inference is able to successfully identify those regions as shown in Fig. 5 and our supplementary material.

5. Conclusion

We presented a method for semantic instance labeling of large datasets from annotated 3D primitives. In the presence of 3D data, our method yields better results compared to several state-of-the-art 2D label transfer baselines while lowering annotation time. Furthermore, our method yields temporally consistent instance labels and explicitly exposes label uncertainty. We also proposed a novel dataset comprising 400k images, laser point clouds and annotations for all objects which we make publicly available. In future work, we plan to extend our method to dynamic scenes by joint inference over multiple frames.

References

[1] V. Badrinarayanan, I. Budvytis, and R. Cipolla. Mixture of trees probabilistic graphical model for video segmentation. *International Journal of Computer Vision (IJCV)*, 110(1):14–29, 2014. 2

[2] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 1, 2

[3] J. Behley, V. Steinhage, and A. B. Cremers. Performance of histogram descriptors for the classification of 3d laser range data in urban environments. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2012. 3

[4] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE Trans. on Visualization and Computer Graphics (VCG)*, 5(4):349–359, 1999. 5

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 26:1124–1137, 2004. 6

[6] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using SfM point clouds. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2008. 3

[7] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 1 2009. 3

[8] I. Budvytis, V. Badrinarayanan, and R. Cipolla. Label propagation in complex video sequences using semi-supervised learning. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2010. 2

[9] L.-C. Chen, S. Fidler, A. L. Yuille, and R. Urtasun. Beat the mturkers: Automatic image labeling from weak 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 3

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013. 2, 3

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[14] A. Geiger and C. Wang. Joint 3d object and layout inference from a single rgb-d image. In *Proc. of the German Conference on Pattern Recognition (GCPR)*, 2015. 2

[15] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2009. 3

[16] M. Guillaumin, D. Küttel, and V. Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision (IJCV)*, 110(3):328–348, 2014. 2

[17] L. Heng, B. Li, and M. Pollefeys. Camodocal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry. In *Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS)*, 2013. 3

[18] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, 2008. 6

[19] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[20] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 5

[21] M. Kiefel and P. Gehler. Human pose estimation with fields of parts. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 5

[22] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems (NIPS)*, 2011. 5, 6, 7

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. 1

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 3

[25] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 33(12):2368–2382, 2011. 2

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[27] A. Martinović, J. Knopp, H. Riemenschneider, and L. Van Gool. 3d all the way: Semantic segmentation of urban scenes from start to end in 3d. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[28] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[29] D. Munoz, J. A. Bagnell, and M. Hebert. Co-inference machines for multi-modal scene analysis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 3

[30] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert. Contextual classification with functional max-margin markov networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[31] N. S. Nagaraja, P. Ochs, K. Liu, and T. Brox. Hierarchy of localized random forests for video annotation. In *Proc. of the DAGM Symposium on Pattern Recognition (DAGM)*, 2012. 2

[32] S. T. Namin, M. Najafi, M. Salzmann, and L. Petersson. A multi-modal graphical model for scene analysis. In *Proc. of*

the IEEE Winter Conference on Applications of Computer Vision (WACV), 2015. 3

[33] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. V. Gool. Learning where to classify in multi-view semantic segmentation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 3

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *arXiv.org*, 1409.0575, 2014. 1

[35] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision (IJCV)*, 77:157–173, 2008. 3

[36] M. Schönbein and A. Geiger. Omnidirectional 3d reconstruction in augmented manhattan worlds. In *Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS)*, 2014. 3

[37] M. Schönbein, T. Strauss, and A. Geiger. Calibrating and centering quasi-central catadioptric cameras. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2014. 3

[38] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision (IJCV)*, 81:2–23, 2009. 3, 4

[39] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGB-D images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 3

[40] S. Song, S. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[41] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2010. 6, 7

[42] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg. Motion coherent tracking using multi-label MRF optimization. *International Journal of Computer Vision (IJCV)*, 100(2):190–202, 2012. 2

[43] J. P. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. Torr. Mesh based semantic modelling for indoor and outdoor scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[44] S. Vijayanarasimhan and K. Grauman. Active frame selection for label propagation in videos. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 2, 6, 7

[45] V. Vineet, G. Sheasby, J. Warrell, and P. H. S. Torr. Posefield: An efficient mean-field based method for joint estimation of human pose, segmentation, and depth. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2013. 5

[46] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using sfm and object labels. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 3

[47] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2009. 2

[48] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and i will show you where it is. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[49] M. Zeiler. Adadelta: An adaptive learning rate method. *arXiv.org*, 1212.5701, 2012. 5

[50] H. Zhang, A. Geiger, and R. Urtasun. Understanding high-level semantics by modeling traffic patterns. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 2

[51] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *Proc. of the International Conf. on Learning Representations (ICLR)*, 1412.6856, 2015. 1

[52] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler. segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1