Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes

Christian Wojek, Stefan Walk, Stefan Roth, Konrad Schindler, Bernt Schiele

Abstract—Following recent advances in detection, context modeling and tracking, scene understanding has been the focus of renewed interest in computer vision research. This paper presents a novel probabilistic 3D scene model that integrates state-of-the-art multi-class object detection, object tracking and scene labeling together with geometric 3D reasoning. Our model is able to represent complex object interactions such as inter-object occlusion, physical exclusion between objects, and geometric context. Inference in this model allows to jointly recover the 3D scene context and perform 3D multi-object tracking from a mobile observer, for objects of multiple categories, using only monocular video as input. Contrary to many other approaches our system performs explicit occlusion reasoning and is therefore capable of tracking objects that are partially occluded for extended periods of time, or objects that have never been observed to their full extent. In addition, we show that a joint scene tracklet model for the evidence collected over multiple frames substantially improves performance. The approach is evaluated for different types of challenging onboard sequences. We first show a substantial improvement to the state-of-the-art in 3D multi-people tracking. Moreover, a similar performance gain is achieved for multi-class 3D tracking of cars and trucks on a challenging dataset.

Index Terms—Scene Understanding, Tracking, Scene Tracklets, Tracking-by-Detection, MCMC

1 Introduction

Robustly tracking objects from a moving observer is an active research area due to its importance for driver assistance, traffic safety, and autonomous navigation [1, 2]. Dynamically changing cluttered backgrounds, varying lighting conditions in the outdoor environment, (partial) object occlusion and the low viewpoint of vehicle-mounted cameras all contribute to the difficulty of the problem. Furthermore, to support navigation, object locations should be estimated in a global 3D coordinate frame rather than in image coordinates.

The main goal of this paper is to address this important and challenging problem by proposing a new *probabilistic 3D scene model* (see Fig. 2 for an overview). Our model builds upon several important lessons from previous research: (1) robust tracking performance is currently best achieved with a *tracking-by-detection* framework [2–4]; (2) short term evidence aggregation, typically termed *tracklets* [4–6], allows for increased tracking robustness; (3) the objects should not be modeled in isolation, but in their *3D scene context*, which puts strong constraints on the position and motion of tracked objects [1, 7]; (4) *multi-cue combination* of scene labels and object detectors allows to strengthen weak detections, but also to prune inconsistent false detections [7]. While all these different components have been shown to boost performance individually, it appears that these components have not yet been integrated in a single system.

Having a full 3D scene model allows us to determine the visibility of each individual object in the scene, which in turn enables to predict which parts of the object are sufficiently visible and thus detectable. This allows us to define a complete 3D scene likelihood that tightly integrates full and partial human detectors within a 3D scene tracking framework. Thus, our model is capable of reasoning about object-object occlusion and can recognize objects when they are partially occluded for extended periods of time and even when they have never been fully visible.

Fig. 1: Example results with our multi-frame 3D inference and explicit occlusion reasoning for onboard vehicle and pedestrian tracking with overlaid horizon estimate for different public state-of-the-art datasets (all results at 0.1 FPPI).

As our experiments show, the proposed probabilistic 3D scene model significantly outperforms the current state-of-the-art. Fig. 1 shows example results for two different types of challenging onboard sequences. Our system is able to robustly track a varying number of targets in 3D world coordinates in

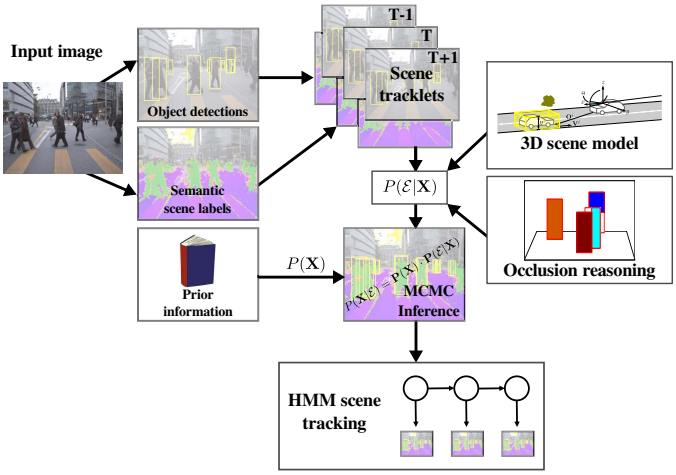


Fig. 2: Overview on our system. For each input frame we run an object detector and extract semantic scene labels. Object hypotheses are fused to short-term tracklets and put into a strong 3D scene model with explicit occlusion reasoning. MCMC inference allows tractable inference in our Bayesian scene model while HMM scene tracking ensures long-term associations.

highly dynamic scenes. This enables us to use a single camera only instead of relying on stereo cameras as in previous work (e.g., [1, 2]). Additionally, we improve performance under full and partial object-object occlusion, which occurs frequently in scenes of realistic complexity and still severely challenges state-of-the-art systems.

Despite using only monocular input, the proposed model allows to constrain object detections to geometrically feasible locations and enforces physically plausible 3D dynamics. This improves object detection results by pruning physically implausible false positives and strengthening weak detections along an object's trajectory. We demonstrate that accumulating scene evidence over a small number of frames with help of a 3D scene model significantly improves performance. As exact inference is intractable we employ reversible-jump Markov Chain Monte Carlo (RJMCMC) sampling for approximate inference. Finally, we show that further improvement can be achieved by performing long-term data association with a Hidden Markov Model (HMM).

2 Related Work

Our work builds on recent advances in scene understanding by pixel-wise labeling, 3D scene analysis and tracking.

The use of scene context has been investigated in the computer vision literature in several ways. Torralba [8] proposes to employ Gabor filter-bank responses in a bottom-up fashion in order to gain prior information on likely 2D object positions. More recently, Shotton et al. [9] use a strong joint-boosting classifier with context reasoning based on a CRF framework to provide a local, per-pixel classification of image content. Tu et al. [10] use MCMC sampling techniques to combine top-down discriminative classifiers with bottom-up generative models for 2D image understanding.

For traffic scene understanding Ess et al. [11] and Brostow et al. [12] particularly address the use of context. [11] uses 2D Walsh-Hadamard filter-bank responses together with stereo depth information to infer traffic situations, while [12] leverages 3D point clouds to improve 2D scene segmentation. Common to these approaches is the goal of 2D image understanding. Our work includes scene labeling as a cue, but its ultimate goal is to obtain a 3D model of the observed world.

Similar to our work Hoiem et al. [7] and Ess et al. [1] employ a 3D scene model. [7] combines image segmentation and object detections in order to infer the objects' positions in 3D. Gupta et al. [13] advance Hoiem's work by adding physical constraints to yield a blocks world model. Both works, however, are limited to single images and do not exploit temporal information available in video. [1] extends [7] for video input, but requires a stereo camera setup to achieve robust tracking of pedestrians from a mobile platform. Similarly, [2] tracks pedestrians for driver assistance applications and employs a stereo camera to find regions of interest and to suppress false detections. Franke et al. [14] employ 6D scene flow computed from a stereo setup to model the vicinity for applications in mobile traffic safety while Geiger et al. [15] use a mobile stereo setup to infer the layout of urban traffic junctions. Note, however, that stereo will yield only little improvement in the far field, because a stereo rig with a realistic baseline will have negligible disparity. Thus, further constraints are needed, since appearance-based object detection is unreliable at very small scales. Therefore, we investigate the feasibility of a monocular camera setup for mobile scene understanding. Other systems that use monocular sequences are Shashua et al. [16] and Choi and Savarese [17]. Contrary to their work, we tightly couple our scene model and the hypothesized positions of objects with the notion of scene tracklets, and exploit constraints given by a-priori information (e.g., approximate object heights and camera pitch). Our experiments show that these short-term associations substantially stabilize 3D inference and improve robustness beyond what has previously been reported. Interestingly, even though 3D scene information is available in many of the above systems, only Ess et al. [18] (an extension to [1]) uses stereo disparity to extend tracks under occlusion using occlusion maps. [17] only models track interaction but does not perform occlusion reasoning. As our experiments show, the proposed tightly integrated scene tracklet model with explicit occlusion handling outperforms these approaches [1, 17], and is able to detect and associate also partially visible objects.

Most tracking-by-detection systems for full scene understanding mentioned above use full-object detectors, which allow to deal with occlusions only when the object is visible enough to be detectable by the respective full-object detector. In many scenes, however, certain objects are never fully visible making these approaches prone to fail. To address this, we propose to use multiple partial detectors that allow to also detect and track objects that are never fully visible. To our knowledge, the only other work that uses partial detectors for tracking from a mobile observer is by Xing et al. [19], which uses only two partial and one full detector without a strong 3D scene model to infer expected occlusion. For tracking humans in a static camera surveillance setting, Shet et al. [20] and Wu and Nevatia [21] employ a combination of body part detectors supported by a 2D scene model. In contrast, this paper uses six partial detectors for explicit occlusion reasoning to infer

the 3D scenes recorded from a moving monocular camera with the goal to eventually support autonomous navigation.

A larger number of approaches for explicit occlusion handling is reported in the object detection literature. Enzweiler et al. [22] learned local head, torso and leg detectors combined in a mixture-of-experts framework and leveraging stereo and flow cues. Wang et al. [23] performed occlusion handling in a modified SVM framework, while Lin et al. [24] adapted a boosting cascade to cope with partial occlusion. Kwak et al. [25] infer occluded regions with a patch classifier and improve tracking performance. Winn and Shotton [26] employed CRFs to couple segmentation and detection under partial occlusion. Finally, Vedaldi and Zisserman [27] employed structured output regression to detect partially truncated and misaligned multi-aspect objects. Similarly, Gao et al. [28] leverage segmentation-based reasoning in a structured output framework to disable occluded parts of a bounding box. For human pose estimation, Sigal and Black [29] proposed an approach to model self-occlusion of body-parts. However, none of these approaches models the interaction of different objects. The goal in our work, in contrast, is an explicit model that builds on 3D scene analysis to address complex object-object occlusions. By explicitly leveraging 3D scene information for occlusion reasoning our approach is able to improve in challenging scenes with partial and full occlusions, which are a major error source for tracking in many systems.

Tracking-by-detection, with an offline learned appearance model, is a popular approach for tracking objects in challenging environments. Breitenstein et al. [30], for instance, track humans based on a number of different detectors in image coordinates. Similarly, Okuma et al. [3] track hockey players in television broadcasts. Huang et al. [6] track people in a surveillance scenario from a static camera, grouping detections in neighboring frames into *tracklets*. Similar ideas have been exploited by Kaucic et al. [5] to track vehicles from a helicopter, and by Li et al. [31] to track pedestrians with a static surveillance camera. Kaucic et al. [5] even leverage scene segmentation to predict occlusion. Nonetheless, their approach only allows to temporally extend tracks of fully visible objects, but does not allow to detect and track objects that are only partially visible. Moreover, none of these tracklet approaches exploit the strong constraints given by the size and position of other objects, and instead build up individual tracks for each object. Others [32–35] propose to use network flows or phrase tracking as a linear optimization problem to overcome challanges such as occlusion. These works however are usually targeted at offline or batch processing. Due to the resulting lag these methods are less favoured for processing on a mobile observer.

In this paper we contribute a probabilistic scene model that allows to jointly infer the camera parameters and the position of *all* objects in 3D world coordinates by using only monocular video and odometry information. Increased robustness is achieved by extending the tracklet idea to *entire scenes* toward the inference of a global scene model.

Realistic, but complex models for tracking including ours are often not amenable to closed-form inference. Thus, several approaches resort to MCMC sampling. Khan et al. [36] track ants and incorporate their social behavior by means of an MRF. Zhao et al. [37] use MCMC sampling to track people from a static camera. Isard and MacCormick [38] track people in front of relatively uncluttered backgrounds from a static indoor camera. All three approaches use rather weak appearance models, which prove sufficient for static cameras. Our model employs a strong object detector and pixel-wise scene labeling to cope with highly dynamic scenes recorded from a moving platform.

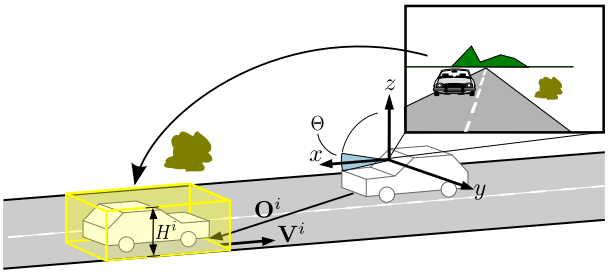


Fig. 3: Visualization of the 3D scene state $\mathbf{X}$ in the world coordinate system. Note that the camera is mounted to the vehicle on the right.

3 Single-Frame 3D Scene Model

We begin by describing our 3D scene model for a *single image*, which aims at combining available prior knowledge with image evidence in order to reconstruct the 3D positions of all objects in the scene. For clarity, the time index t is omitted when referring to a single time step only. Variables in image coordinates are printed in lower case, variables in 3D world coordinates in upper case; vectors are printed in bold face.

The posterior distribution for the 3D scene state $\mathbf{X}$ given image evidence $\mathcal{E}$ is defined in the usual way, in terms of a prior and an observation model:

$$P(\mathbf{X}|\mathcal{E}) \propto P(\mathcal{E}|\mathbf{X})P(\mathbf{X}) \tag{1}$$

The 3D state $\mathbf{X}$ consists of the individual states of all objects $\mathbf{O}^i$, described by their relative 3D position $(O_x^i, O_y^i, O_z^i)^\top$, speed $(V_x^i, V_y^i, V_z^i)^\top$ w.r.t. the observer and by their height H^i. For the single-frame model described in this section we are not able to estimate speed and thus we drop the according state vector entries. Moreover, $\mathbf{X}$ includes the internal camera parameters $\mathbf{K}$ and the camera orientation $\mathbf{R}$.

The goal of this work is to infer the 3D state $\mathbf{X}$ from video data of a monocular, forward facing camera (see Fig. 3). While in general this is an under-constrained problem, in robotic and automotive applications we can make the following assumptions that are expressed in the prior $P(\mathbf{X})$: The camera undergoes no roll and yaw w.r.t. the platform (cf. Fig. 4), its intrinsics $\mathbf{K}$ are constant and have been calibrated off-line, and the speed and turn rate of the platform are estimated from odometer readings. Furthermore, the platform as well as all objects of interest are constrained to stand on a common ground plane (i.e., $O_z^i = 0$ and $V_z^i = 0$). Note that under these assumptions the ground plane in camera-centric coordinates is fully determined by the pitch angle Θ. As the camera is rigidly mounted to the vehicle, it can only pitch

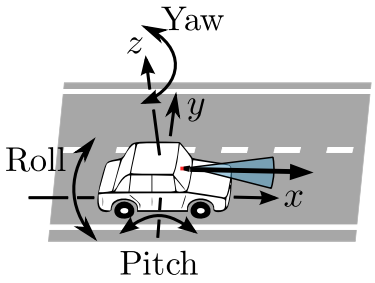


Fig. 4: For applications in robotics and automotive safety the rotation angles of a vehicle mounted camera with respect to the environment are heavily constrained.

a few degrees. To avoid degenerate camera configurations, the pitch angle is therefore modeled as normally distributed around the pitch of the resting platform as observed during calibration: $\mathcal{N}(\Theta; \mu_\Theta, \sigma_\Theta)$. We note that the choice of the normal distribution is likely to be inaccurate in particular due to the bounded range of angles. However, as realistic values of σ_Θ are relativly small this is no practical limitation. This prior allows deviations arising from acceleration and braking of the observer. This is particularly important for the estimation of distant objects as, due to the low camera viewpoint, even minor changes in the pitch may cause a large error for distance estimation in the far field.

Moreover, we assume the height of all scene objects to follow a normal distribution $\mathcal{N}(H^i; \mu_H^{c_i}, \sigma_H^{c_i})$ around a known mean value $\mu_H^{c_i}$, which is specific for the respective object class c_i. This helps to prune false detections that are consistent with the ground plane, but are of the wrong height (e.g., background structures such as street lights). We note that for the height of people a normal distribution is likely to be inaccurate but use it nevertheless with manually set parameters due to the lack of better data. The overall prior is thus given as:

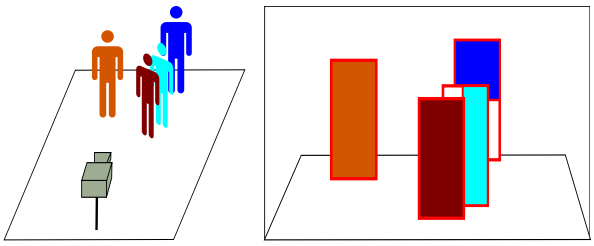
$$P(\mathbf{X}) \propto \mathcal{N}(\Theta; \mu_\Theta, \sigma_\Theta) \cdot \prod_i \mathcal{N}(H^i; \mu_H^{c_i}, \sigma_H^{c_i}) \qquad (2)$$

Next, we turn to the observation model $P(\mathcal{E}|\mathbf{X})$. The image evidence $\mathcal{E}$ is comprised of a set of potential object detections and a scene labeling, i.e., category labels densely estimated for every pixel. As we will see in the experiments, the combination of these two types of image evidence is beneficial, because object detections give reliable but rather coarse bounding boxes, and low level cues enable more fine-grained data association by penalizing inconsistent associations and supporting consistent, but weak detections.

For each object i our model fuses object appearance given by the object detector confidence, geometric constraints, and local evidence from bottom-up pixel-wise labeling:

$$P(\mathcal{E}|\mathbf{X}) \propto \prod_i \Psi_G\big(\mathbf{O}^i, \Theta; \mathbf{d}^{a(i)}\big) \cdot \Psi_D\big(\mathbf{d}^{a(i)}\big) \cdot \Psi_L^i\big(\mathbf{X}; \mathbf{l}\big) \quad (3)$$

Here, $a(i)$ denotes the association function, which assigns a candidate object detection $\mathbf{d}^{a(i)}$ to every 3D object hypothesis $\mathbf{O}^i$. Note that the associations between objects and detections are not fixed in advance, but established as part of the MCMC sampling procedure (see Sec. 4.1). The following subsections will discuss Ψ_G, Ψ_D and Ψ_L^i in more detail.



(a) 3D scene with severe object-object occlusions (b) 2D projection onto the camera's image plane

Fig. 5: We approximate objects by their bounding boxes and project them onto the image. By leveraging the depth order obtained from the 3D scene model we are able to estimate occluded object regions. Only partial detectors belonging to visible regions contribute to the mixture model (cf. Eqn. 7).

3.1 Scene geometry

The geometry potential Ψ_G models how well the estimated 3D state $\mathbf{O}^i$ satisfies the geometric constraints due to the ground plane specified by the camera pitch Θ. Denoting the projection of the 3D position $\mathbf{O}^i$ to the image plane as $\mathbf{o}^i$, the distance between $\mathbf{o}^i$ and the associated detection $\mathbf{d}^{a(i)}$ in x-y-$scale$-space serves as a measure of how much the geometric constraints are violated. We model Ψ_G using a Gaussian

$$\Psi_G(\mathbf{O}^i, \Theta; \mathbf{d}^{a(i)}) = \mathcal{N}(\mathbf{o}^i; \mathbf{d}^{a(i)}, \boldsymbol{\sigma}_G + \bar{\boldsymbol{\sigma}}_G) \,, \qquad (4)$$

where we split the kernel bandwidth into a constant component $\boldsymbol{\sigma}_G$ and a scale-dependent component $\bar{\boldsymbol{\sigma}}_G$ to account for inaccuracies that arise from the scanning stride of the sliding-window detectors.

3.2 Detector appearance

For the detector appearance potential we will explore one variant with explicit occlusion reasoning and one without. Our formulation with occlusion reasoning requires several different partial object detectors while the formulation without occlusion reasoning only requires a full object detector and is therefore faster to compute. We will first introduce our model without occlusion reasoning [39] and then extend the formulation to the more complex and richer model [40].

3.2.1 *Without explicit occlusion reasoning*

When we build our model on a single full object detector the appearance potential Ψ_D appearance score of detection $\mathbf{d}^{a(i)}$ for object i into the positive range. Contrary to our previous work [39] we consistently use a hard-clipping function that clips the score to a minimum value ξ:

$$\Psi_D\big(\mathbf{d}^{a(i)}\big) = \begin{cases} \text{score}(\mathbf{d}^{a(i)}) & \text{if score}(\mathbf{d}^{a(i)}) \geq \xi \\ \xi & \text{if score}(\mathbf{d}^{a(i)}) < \xi \end{cases} \qquad (5)$$

Hard-clipping the scores in the above way yields the same or better performance than previously used soft-clipping [41] or sigmoid [42] mappings.

3.2.2 *Explicit occlusion reasoning*

To perform explicit occlusion reasoning that can cope with partial object visibility, we will use a bank of detectors,

Algorithm 1 Efficient visible area computation for rectangular regions: $\mathbf{r}$ - rectangle for which the number of visible pixels is computed; m - maximum tested object depth.
Since the intersection and AREA can be computed quickly for rectangles, this algorithm is faster in practice than a dense pixel-wise occlusion map which is often used for arbitrary shapes.

Require: $\mathbf{O}^1, \ldots, \mathbf{O}^m$ sorted in increasing depth
function VISIBLEAREA($\mathbf{r}$, m)
  $v^r \leftarrow$AREA($\mathbf{r}$)
  for $k = 1 \ldots m-1$ **do**
    $\mathbf{o}^k \leftarrow$PROJECT($\mathbf{O}^k$)
    if $\mathbf{r} \cap \mathbf{o}^k \neq \varnothing$ **then**
      if $k \neq 1$ **then**
        $v^r \leftarrow v^r -$VISIBLEAREA($\mathbf{r} \cap \mathbf{o}^k, k$)
      else
        $v^r \leftarrow v^r -$AREA($\mathbf{r} \cap \mathbf{o}^k$)
      end if
    end if
  end for
  return v^r
end function

$v_p^i \leftarrow$ VISIBLEAREA($\mathbf{o}_p^i, i$)/AREA($\mathbf{o}_p^i$)

each trained to detect parts of the objects of interest (see Section 6 for details), in addition to the full-object detector above. To incorporate these local part detections robustly we perform occlusion handling by explicitly leveraging 3D scene information. Thus, we rephrase the above definition of Eqn. 3 to incorporate object visibility. For each part p (we also refer to the full object detector as a *part* in the following) we compute its projection's expected visibility v_p^i based on the global 3D scene model (cf. Fig. 5). Assuming that the camera views the scene along the x-axis (cf. Fig. 3) and that the objects are sorted with increasing depth, we can formally express a part's visibility as:

$$v_p^i = \text{AREA}\left(\mathbf{o}_p^i \setminus \bigcup_{j<i} \mathbf{o}^j\right) / \text{AREA}(\mathbf{o}_p^i), \quad \text{s.t. } \forall_j O_x^j < O_x^i \qquad (6)$$

where AREA($\mathbf{o}_p^i$) denotes the image area in pixels covered by the projection of the p^{th} part of $\mathbf{O}^i$. Alg. 1 gives an efficient algorithm for obtaining v_p^i for rectangular projections $\mathbf{o}_p^i$. As detectors tend not to respond for parts with low visibility due to the lack of occluded samples in the training data, we discard part detections when the visible area v_p^i is below a certain threshold v_{min}[1]. We define our multi-detector observation likelihood with explicit occlusion handling as a mixture-of-experts [43] where the experts are the part detectors and the weights are proportional to the visible area v_p^i of those parts:

$$P(\mathcal{E}|\mathbf{X}) \propto \prod_i \frac{1}{\sum_q [v_q^i > v_{min}] \cdot v_q^i} \cdot \sum_p \Big([v_p^i > v_{min}] \cdot \qquad (7)$$
$$v_p^i \cdot \Psi_G\big(\mathbf{O}^i, \Theta; \mathbf{d}_p^{a(i,p)}\big) \cdot \Psi_D\big(\mathbf{d}_p^{a(i,p)}\big) \Big) \cdot \Psi_L^i(\mathbf{X}; \mathbf{l})$$

Now, $a(i,p)$ is an association function that assigns candidate part detections $\mathbf{d}_p^{a(i,p)}$ (at most one for each part p) to every 3D object hypothesis $\mathbf{O}^i$. $[expr]$ denotes the Iverson bracket ($[expr]=1$ if $expr$ is true, and 0 otherwise) and is used here to enable and disable partial detectors based on their expected visibility. In case a detector is not firing despite a sufficiently large estimated visibility ($v_p^i > v_{min}$) we use a minimum appearance score to compensate missing evidence. Ψ_D and Ψ_G are defined as for the single detector likelihood (cf. Eqn. 5 and Eqn. 4), but use the associated part detector's estimate for the full object extent instead of the full object detector. As an alternative for the (visibility) weighted mixture model above we have also tried an unweighted mixture model, a mixture model based on learned weights for the partial detectors and a model that assumes independent partial detectors and thus computes the geometric mean, but found the mixture-of-experts model above to perform best or equally well. In particular the option that assumes independent partial detectors suffered from missing detections or low scores of a single detector.

Hypotheses clustering. To enable efficient inference we cluster $a(i,p)$ agglomeratively into groups of possible associations. Starting from an association function $a(i,p)$ that only associates full object detections, we iteratively add associations to part detections when those overlap sufficiently for the respective object part. In each iteration we add the part detection with the highest overlap that has previously not been matched. Part detections that cannot be matched to an existing cluster lead to an additional, new cluster.

Regarding the comparability of detector scores we found empirically that SVM margins and boosting likelihoods on true positive detections tend to be larger for better performing detectors. For SVMs this is probably due to the fact that we train all detectors on the same training set, and thus scores are implicitly normalized by scaling the SVM margin to 1. Therefore an implicit detector weighting is learned during SVM training and no further provision to balance SVM scores is required. For boosting we use the same number of weak classifiers for each part detector. Due to the lower discriminative power of the partial object detectors the estimated weak classifiers' weights are lower and thus the scores of weaker partial detectors are implicitly down-weighted.

3.3 Semantic scene labels

The scene labeling potential Ψ_L^i describes how well the projection $\mathbf{o}^i$ matches the bottom-up pixel labeling. For each pixel j and each class c the labeling yields a classification score $l^j(c)$. Here the labeling scores obtained from boosting [44] are normalized pixel-wise by means of a $softmax$ transformation in order to obtain positive values (cf. Fig. 6).

1. In our implementation we use $v_{min} = 0.75$.

(a) Input image (b) Normalized scene labeling likelihood for *pedestrian*

Fig. 6: Scene labeling likelihoods are normalized locally with a softmax transform to yield positive values.

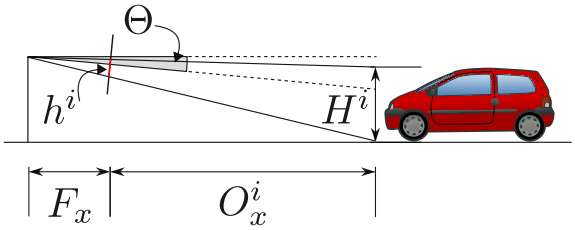


Fig. 7: Employing the theorem of intersecting lines we can derive the distance to object i along the ground plane O_x^i in viewing direction as $O_x^i = \frac{H^i F_x}{h^i \cos(\theta)}$. For small pitch values Θ, $\cos(\Theta)$ can be discarded and thus $O_x^i \approx \frac{H^i F_x}{h^i}$. Here, F_x denotes the focal length in x-direction which is assumed to be calibrated off-line while h^i is object i's projection height on the image plane.

It is important to note that this cue demands 3D scene modeling: To determine the set of pixels that belong to each potential object, one needs to account for inter-object occlusions, and hence know the objects' depth ordering. Given that ordering, we proceed as follows: each object is back-projected to a bounding box $\mathbf{o}^i$, and that box is split into a visible region $\boldsymbol{\delta}^i$ and an occluded region $\boldsymbol{\omega}^i$ (see Alg. 1 for a fast algorithm when object projections are approximated with bounding boxes). The object likelihood is then defined as the ratio between the cumulative score for the expected label e and the cumulative score of the pixel-wise highest scoring background class label $k \neq e$, evaluated over the visible part of $\mathbf{o}^i$:

$$\Psi_L^i(\mathbf{X};\mathbf{l}) = \left(\frac{\sum_{j \in \boldsymbol{\delta}^i} l^j(e) + \tau}{\epsilon|\boldsymbol{\omega}^i| + \sum_{j \in \boldsymbol{\delta}^i} l^j(k) + \tau} \right)^{\alpha}, \tag{8}$$

where the constant τ corresponds to a weak Dirichlet prior; $\epsilon|\boldsymbol{\omega}^i|$ avoids highly occluded objects to have a large influence with little available evidence; and α balances the relative importance of detector score, scene geometry consistence and pixel label likelihood.

4 Inference Framework

To perform inference in the above model, we simulate the posterior distribution $P(\mathbf{X}|\mathcal{E})$ in a Metropolis-Hastings MCMC framework [45]. At each iteration s new scene samples $\mathbf{X}'$ are proposed by different *moves* from the proposal density $Q(\mathbf{X}';\mathbf{X}^{(s)})$. The proposal's posterior is evaluated and the *acceptance ratio* is computed as

$$r = \frac{P(\mathbf{X}'|\mathcal{E})}{P(\mathbf{X}^{(s)}|\mathcal{E})} \frac{Q(\mathbf{X}^{(s)};\mathbf{X}')}{Q(\mathbf{X}';\mathbf{X}^{(s)})}. \tag{9}$$

The proposal is accepted with probability $\min(1; r)$. We assign $\mathbf{X}^{(s+1)} \leftarrow \mathbf{X}'$ if the proposal is accepted; otherwise the last state is retained, $\mathbf{X}^{(s+1)} \leftarrow \mathbf{X}^{(s)}$. Since our goal is to sample from the equilibrium distribution, we discard the samples from an initial burn-in phase. Note that the normalization of the posterior does not have to be known, since it is independent of $\mathbf{X}$ and therefore cancels out in the posterior ratio. Importantly, $P(\mathbf{X}|\mathcal{E})$ is not comparable across scene configurations with different numbers of objects. We address this with a reversible jump MCMC framework [46].

4.1 Proposal moves

Proposal moves change the current state of the Markov chain. We employ three different move types: *diffusion moves* to update the last state's variables, *add moves* and *delete moves* to change the state's dimensionality by adding or removing objects from the scene. Add and delete moves are mutually reversible and trans-dimensional. At each iteration, the move type is selected randomly with fixed probabilities $q_{\text{Add}}, q_{\text{Del}}$ and q_{Dif}.

Diffusion moves change the current state by sampling new values for the state variables. At each diffusion move, object variables are updated with a probability of $q_{\mathbf{O}}$, while Θ is updated with a probability of q_{Θ}.

To update objects we draw the index i of the object to update from a uniform distribution and then update either $\mathbf{O}^i$ or $\mathbf{V}^i$. Proposals are drawn from a multi-variate normal distribution with diagonal covariance centered at the previous state.

To update the camera pitch Θ, proposals are generated from a mixture model. The first mixture component is a broad normal distribution centered at the calibrated pitch for the motionless platform. For the remaining mixture components, we assume distant objects associated with detections at small scales to have the class' mean height and use $\mathbf{d}^{a(i)}$ to compute their distance by means of the theorem of intersecting lines (cf. Fig. 7). Then the deviation between the detected bounding box and the object's projection in the image allows one to estimate the camera pitch. This relationship can directly be derived from the perspective projection equation. Note that one property of this relationship is that uncertainty in the distance estimate (arising from the deviation in height from the class mean) for distant cars translates to a lower uncertainty in the pitch value than for close cars. We place one mixture component around each pitch computed this way and assign mixture weights proportional to the detection scores to put more weight on more likely objects.

Add moves add a new object $\mathbf{O}^{N+1}$ to the chain's last state, where N is the number of objects contained in $\mathbf{X}^{(s)}$. As this move is trans-dimensional (i.e., the number of dimensions of $\mathbf{X}^{(s)}$ and $\mathbf{X}'$ do not match) special consideration needs to be taken when the posterior ratio $\frac{P(\mathbf{X}'|\mathcal{E})}{P(\mathbf{X}^{(s)}|\mathcal{E})}$ is evaluated. In particular, $P(\mathbf{X}^{(s)}|\mathcal{E})$ needs to be made comparable in the state space of $P(\mathbf{X}'|\mathcal{E})$. To this end, we assume a constant probability $\bar{P}(\mathbf{O}^{N+1})$ for each object to be part of the background. Hence, posteriors of states with different numbers of

objects can be compared in the higher dimensional state space by transforming $P(\mathbf{X}^{(s)}|\mathcal{E})$ to

$$\hat{P}(\mathbf{X}^{(s)}|\mathcal{E}) = P(\mathbf{X}^{(s)}|\mathcal{E})\bar{P}(\mathbf{O}^{N+1}) \tag{10}$$

To efficiently explore high density regions of the posterior we use the detection scores in the proposal distribution. A new object index n is drawn from the discrete set of all K detections $\{\bar{\mathbf{d}}\}$, which are not yet associated with an object in the scene, according to $Q(\mathbf{X}';\mathbf{X}^{(s)}) = \frac{\psi_D(\bar{\mathbf{d}}^n)}{\sum_k \psi_D(\bar{\mathbf{d}}^k)}$. The data association function is updated by letting $a(N+1)$ associate the new object with the selected detection. For distant objects (i.e., detections at small scales) we instantiate the new object at a distance given through the theorem of intersecting lines and the height prior (cf. Fig. 7), whereas for objects in the near-field a more accurate 3D position and object height can be estimated from the ground plane and camera calibration. In order to avoid discontinuities in choosing one of these two instantiation methods we use a sigmoid-like weighting function to softly blend the two cases.

Delete moves remove an object $\mathbf{O}^n$ from the last state and move the associated detection $\mathbf{d}^{a(n)}$ back to pool of unclaimed detections $\{\bar{\mathbf{d}}\}$. Similar to the add move, the proposed lower dimensional state $\mathbf{X}'$ needs to be transformed. The object index n to be removed from the scene is drawn uniformly among all objects currently in the scene, thus $Q(\mathbf{X}';\mathbf{X}^{(s)}) = \frac{1}{N}$. Consequently, the acceptance ratios for add and delete moves are:

$$r_{\text{Add}} = \frac{P(\mathbf{X}'|E)}{\hat{P}(\mathbf{X}^{(s)}|E)} \frac{q_{\text{Del}}}{q_{\text{Add}}} \frac{\sum_k \psi_D(\bar{\mathbf{d}}^k)}{\psi_D(\bar{\mathbf{d}}^n)(N+1)} \tag{11}$$

$$r_{\text{Del}} = \frac{\hat{P}(\mathbf{X}'|E)}{P(\mathbf{X}^{(s)}|E)} \frac{q_{\text{Add}}}{q_{\text{Del}}} \frac{N\psi_D(\mathbf{d}^{a(n)})}{\psi_D(\mathbf{d}^{a(n)}) + \sum_k \psi_D(\bar{\mathbf{d}}^k)} \tag{12}$$

4.2 Projective 3D to 2D marginalization

In order to obtain a score for a 2D position $\mathbf{u}$ (including scale) from our 3D scene model, the probabilistic framework suggests marginalizing over all possible 3D scenes $\mathbf{X}$ that contain an object that projects to that 2D position:

$$P(\mathbf{u}|\mathcal{E}) = \int \max_i \left(\left[\mathbf{u} = \mathbf{o}^i \right] \right) P(\mathbf{X}|\mathcal{E}) \, \mathrm{d}\mathbf{X} \, , \tag{13}$$

with $[expr]$ being the Iverson bracket: $[expr] = 1$ if the enclosed expression is true, and 0 otherwise. Hence, the binary function $\max_i([\cdot])$ detects whether there exists *any* 3D object in the scene that projects to image position $\mathbf{u}$. The marginal is approximated with samples $\mathbf{X}^{(s)}$ drawn using MCMC:

$$P(\mathbf{u}|\mathcal{E}) \approx \frac{1}{S} \sum_{s=1}^{S} \max_i \left(\left[\mathbf{u} = \mathbf{o}^{i,(s)} \right] \right) \, , \tag{14}$$

where $\mathbf{o}^{i,(s)}$ denotes the projection of object $\mathbf{O}^i$ of sample s to the image, and S is the number of samples. In practice $\max_i([\cdot])$ checks whether any of the 3D objects of sample s projects into a small neighborhood of the image position $\mathbf{u}$.

5 Multi-frame Model and Inference

So far we have described our scene model for a single image in static scenes only. For the extension to video streams we pursue a two-stage tracking approach. First, we extend the model to neighboring frames by using greedy data association. Second, the resulting *scene tracklets* are used to extend our model towards long-term data association by performing *scene tracking* with an HMM.

5.1 Multi-frame 3D scene tracklet model

To apply our model to multiple frames, we first use the observer's estimated speed V_{ego} and turn (yaw) rate to roughly compensate the camera's ego-motion. Next, we use a coarse dynamic model for all moving objects to locally perform association, which is refined during tracking. For initial data associations objects that move substantially slower than the camera (e.g., people) are modeled as standing still, $\mathbf{V}^i = 0$. For objects with a similar speed (e.g., cars and trucks), we distinguish those moving in the same direction as the observers from the oncoming traffic with the help of the detector's class label. The former are expected to move with a similar speed as the observer, $V_x^i = V_{ego}$, whereas the latter are expected to move with a similar speed, but in opposite direction, $V_x^i = -V_{ego}$. The camera pitch Θ_t can be assumed constant for small time intervals.

For a given frame t we associate objects and detections as described in Sec. 4.1. In adjacent frames we perform data association by finding the detection with maximum overlap to each predicted object while requiring a minimum overlap. Missing evidence is compensated by choosing a small constant as minimal detection likelihood anywhere in the image. Due to the potentially fast moving camera and our very coarse dynamic model associations can only be established very locally within a few frames. In this setting the overlap of estimated and detected object bounding box proved to be sufficient. We note that for other applications more elaborate schemes are certainly possible (e.g. [31]). We define the scene tracklet posterior as

$$P(\mathbf{X}_t|\mathcal{E}_{-\delta t+t:t+\delta t}) \propto \prod_{r=t-\delta t}^{t+\delta t} P(\hat{\mathbf{X}}_r|\mathcal{E}_r), \tag{15}$$

where $\hat{\mathbf{X}}_r$ denotes the predicted scene configuration using the initial dynamic model just explained. Note, that our definition of tracklet uses evidence from a constant number of neighboring frames to instantiate tracklets of constant length $2\delta t$. These tracklets are then used as evidence for inference at time t. However, note that at time $t+1$ new tracklets are instatiated and thus associations in between a pair of frames may vary. This deviates from previous work (e.g. [5, 6, 31, 47–50]) that typically builds tracklets of varying length from stable tracks in a first step and links those to long-term tracks in a second step.

5.2 Long term data association with scene tracking

While the above model extension to scene tracklets is feasible for small time intervals, it does not scale well to longer sequences, because greedy data association in combination with a simplistic motion model will eventually fail. Moreover,

the greedy formalism cannot handle objects leaving or entering the scene.

We therefore introduce an explicit data association variable $\mathcal{A}_t$, which assigns objects to detections in frame t. With this explicit mapping, long-term tracking is performed by modeling associations over time in a hidden Markov model (HMM). Inference is performed in a sliding window of length w to avoid latency as required by an online setting:

$$P(\mathbf{X}_{1:w}, \mathcal{A}_{1:w} | \mathcal{E}_{-\delta t+1:w+\delta t}) = P(\mathbf{X}_1 | \mathcal{A}_1, \mathcal{E}_{-\delta t+1:1+\delta t}) \prod_{k=2}^{w} P(\mathcal{A}_k | \mathcal{A}_{k-1}) P(\mathbf{X}_k | \mathcal{A}_k, \mathcal{E}_{-\delta t+k:k+\delta t}) \quad (16)$$

The emission model is the scene tracklet model from Sec. 5.1, but with explicit data association $\mathcal{A}_k$. The transition probabilities are defined as $P(\mathcal{A}_k|\mathcal{A}_{k-1}) \propto {P_e}^{\eta} {P_l}^{\lambda}$. Thus, P_e is the uniform probability for an object to enter the scene, while P_l denotes the uniform probability for an object to leave the scene. To determine the number η of objects entering the scene, respectively the number λ of objects leaving the scene, we again perform frame-by-frame greedy maximum overlap matching (see Sec. 5.1 for details). In Eq. (16) the marginals $P(\mathbf{X}_k, \mathcal{A}_k | \mathcal{E}_{-\delta t+1:w+\delta t})$ can be computed with the sum-product algorithm. Finally, the probability of an object being part of the scene is computed by marginalization over all other variables (cf. Sec. 4.2):

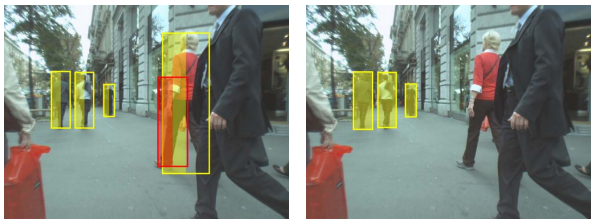
$$P(\mathbf{u}_k | \mathcal{E}_{-\delta t+1:w+\delta t}) = \sum_{\mathcal{A}_k} \int \max_i \left([\mathbf{u}_k = \mathbf{o}_k^i] \right) P(\mathbf{X}_k, \mathcal{A}_k | \mathcal{E}_{-\delta t+1:w+\delta t}) \, \mathrm{d}\mathbf{X}_k \quad (17)$$

In practice we approximate the integral with MCMC samples as above, however this time only using those that correspond to the data association $\mathcal{A}_k$. Note that the summation over $\mathcal{A}_k$ only requires to consider associations that occur in the sample set.

6 Datasets and Implementation Details

For our experiments we use four datasets: (1) *ETH-Loewenplatz*, *ETH-Linthescher*, *ETH-PedCross2* were introduced by [1] and [51] respectively to benchmark pedestrian tracking from a moving observer; and (2) a multi-class dataset we recorded in [39] with an onboard camera to specifically evaluate the challenges targeted by our work including realistic traffic scenarios with a large number of small objects, objects of interest from different categories, and higher driving speed.

ETH datasets. These publicly available pedestrian benchmarks[2] were recorded with a moving stereo camera in densely populated pedestrian zones and originally published by Ess et al. [1, 51]. Fig. 11 shows some examples. All three sequences are recorded with a resolution of 640×480 pixels at ≈15 fps. As our system with explicit occlusion reasoning is capable of detecting severely occluded pedestrians that are not contained in the original annotation, we manually extended the annotations (by all pedestrians which are at least 20% visible) to obtain a fair evaluation. *ETH-Loewenplatz* has been recorded from a driving car in urban traffic and contains 802 frames overall of which every 4th frame is annotated resulting in a total of 2665 annotated bounding boxes (2631 original annotations). *ETH-Linthescher* is comprised of 1209 stereo image pairs with a total of 3018 pedestrians of which 2606 were annotated in the original annotation. *ETH-PedCross2* consists of 840 frames recorded at a pedestrian crossing and along a rather narrow sidewalk with frequent occlusions among pedestrians. As this sequence comes without annotations, we annotated pedestrians in every 4^{th} frame similar to *ETH-Linthescher* and *ETH-Loewenplatz*, and included instances that are truncated by the image boundaries. Overall our annotations contain 1635 pedestrians.



(a) Left-half detector (b) Full-object detector

Fig. 8: Sample detections for models trained on the left half of a pedestrian and for the full-object detector. While the models for partial views do not perform as well overall, they are able to provide the scene model with hypotheses for partially occluded pedestrians.

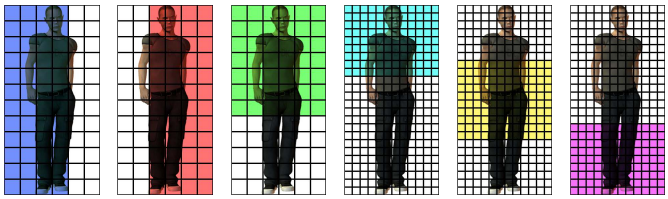
For our experiments we only use the left camera's images as input to our monocular system, and simulate yaw and speed sensor readings based on structure-from-motion results kindly provided by the authors of [1]. For the evaluation we follow their protocol, which only considers pedestrians with an annotation height of at least 60 pixels. Smaller annotations and detections are discarded by *post-filtering* [52].

Multi-class test set. As the above datasets are restricted to pedestrians observed at low driving speeds, we recorded a new multi-class test set in [39] consisting of 674 images. The data is subdivided into 5 sequences and has been recorded at a resolution of 752×480 pixels from a driving car at ≈15 fps. Additionally, ego-speed and turn rate are obtained from the car's ESP module. See Fig. 13 for example images. 1331 front view of cars, 156 rear view of cars, and 394 front views of trucks were annotated in the original set [39] which we extended as well for this paper to include instances which are up to 80% occluded. Thus, the annotation for this analysis contains 1420 front views of cars, 156 rear view of cars and 403 front views of trucks. Vehicles appear over a large range of scales from as small as 20 pixels to as large as 270 pixels. 45% of the objects have a height of ≤ 30 pixels, and are thus hard to detect[3].

Object detectors. To detect potential object instances, we use state-of-the-art object detectors. For pedestrian detection we use our motion feature enhanced variant of the HOG framework [53] and partial detectors trained on the same features as well as the deformable part model (DPM) by Felzenszwalb et al. [54] along with partial HOG [55] detectors. These partial

2. http://www.vision.ee.ethz.ch/~aess/dataset/

3. Data available at http://www.d2.mpi-inf.mpg.de/



(a) Left (b) Right (c) Top (d) Top-HR (e) Mid-HR (f) Bot-HR

Fig. 9: Detector regions.

HOG detectors differ from [55] by an intersection kernel SVM instead of a linear SVM [56], using multiple rounds of retraining to make the training procedure stable [54], and an improved non-maximum suppression scheme [57].

All detectors are trained on pedestrians that are scale normalized for a 128×64 pixel detection window and evaluated on images that are upscaled by a factor of 2 to detect smaller instances. The DPM detector and the six partial HOG detectors that are used in combination with DPM are trained on the INRIA Person dataset [55], while the motion enhanced detectors are trained on TUD-MotionPairs [57]. Overall, we train six kinds of partial detectors as depicted in Fig. 9. The SVM for three of the detectors is trained on the upper, left and right halves of the block grids (rounded up). The upper-body detector uses the top 8×7 blocks, and the left- and right-half body detectors the 15×4 left and right blocks, respectively (for an illustration see Fig. 9(a)–(c)). We also employ three models using a higher resolution detection window (256×128 pixels, resulting in a grid of 31×15 blocks), trained using only rows 3–12 (Top-HR), 11–20 (Mid-HR) and 20–29 (Bot-HR) (see Fig. 9(d)-(f)). These are motivated by the fact that in crowded scenes such as in *ETH-Linthescher* and *ETH-Pedcross2* pedestrians are often quite close to the camera, and thus cannot be detected with a sliding-window detector for the full object. Due to the relatively low average pedestrian height on *ETH-Loewenplatz* we only use the full object and low-resolution partial detectors (Fig. 9(a)-(c)) for this dataset.

It is important to note that training a classifier on parts, i.e. subsets of blocks, is different from using the model learned for the full object and only evaluating on the "visible" subset (which would be theoretically possible for additive kernels or boosted classifiers), especially because during the bootstrapping phase of training the detector finds hard samples for the partial-view models instead of hard samples for the full-object model. Even though a bank of detectors increases the computational load, we stress that the low-level feature representation can be shared among detectors and therefore only the classifiers need to be evaluated. To further reduce the load it may also be possible to adapt the DPM formulation to allow a tighter integration of our partial detectors; this is left for future work. Fig. 8 shows detection examples for the left-half and full-object models. The full-object detector (Fig. 8(b)) is good at spotting fully visible pedestrians, but has problems finding pedestrians that are partially occluded. For these cases partial-view detectors can be beneficial, as they can spot partially occluded pedestrians (Fig. 8(a)). However, they typically also produce more false positives (as in Fig. 8(a)). As these tend to be inconsistent with the 3D scene model, our method can discard them.

For our *MPI-VehicleScenes* test set we employ a multi-class detector based on traditional HOG-features and joint boosting [44] as classifier. It can efficiently detect the four object classes *car front, car back, truck front* or *truck back* that are in contrast to binary SVMs [55] trained jointly. Due to the low-resolution and rigid object structures a part-based DPM detector is unlikely to further improve performance for these classes. Also objects are mostly moving in the same direction as the camera, such that optic flow and motion features will not add much information. For the full object detector we use a 40×40 pixel detection window, but upscale the image by a factor of 2, which turned out to perform better than a 20×20 pixel detection window. Additionally, we train low-resolution left and right partial detectors similar to Fig. 9(a) and (b). Note that for our application it is important to explicitly separate front from back views, because the motion model is dependent on the heading direction. Our detectors were trained on a separate dataset recorded from a driving car, with a similar viewpoint as in the test data.

Scene labeling. Every pixel is assigned to the classes *pedestrian, vehicle, street, lane marking, sky* or *void* to obtain a scene labeling. As features we use the first 16 coefficients of the Walsh-Hadamard transform (WHT) extracted at five scales (4-64 pixels) from each channel of the CIE-Lab color space, along with the pixels' (x, y)-coordinates to account for their location in the image. The WHT is a discrete approximation of the cosine transform and can be computed efficiently [58] – on a modern GPU even in real-time.

The L-channel is mean/variance normalized to cope with global lighting variations, whereas the a- and b-channels are normalized with the gray world assumption to mitigate color shift. We also found normalizing the transformation coefficients with the L_1-norm as in [59] to be beneficial. We then compute mean and variance on 4×4 pixel groups and again classify them with joint boosting. The method directly performs multi-label classification [44], and yields more efficient classifiers because of its capability to share features between classes.

Experimental setup. For both datasets and all object classes we use the same set of parameters for the MCMC sampler: $q_{Add} = 0.1$, $q_{Del} = 0.1$, $q_{Dif} = 0.8$, $q_{\mathbf{O}} = 0.8$, $q_{\Theta} = 0.2$. For our scene tracklet formulation (cf. Sec. 5.1) we consistently use tracklets of length three which are centered at the frame for which we perform inference. For the HMM's sliding window of Eqn. 16 we choose a length of $w = 7$ frames. The sampler draws 3,000 samples for burn-in and 20,000 samples to approximate the posterior and runs without parallelization at 0.3-2 fps on recent hardware. By running multiple Markov chains in parallel we expect a possible speed-up of 1-2 orders of magnitude. As we do not have 3D ground truth to assess 3D performance, we project the results back to the images and match them to ground truth annotations with the PASCAL criterion ($intersection/union > 50\%$).

Baselines. As baselines we report both the performance of the object detectors as well as the result of an extended Kalman filter (EKF) atop the detections. The EKFs track the

objects independently, but work in 3D state space with the same dynamic models as our MCMC sampler. To reduce false alarms in the absence of an explicit model for new objects entering, tracks are declared valid only after three successive associations. Analogous to our system, the camera ego-motion is compensated using odometry. Best results were obtained when the last detection's score was used as confidence measure.

7 Experimental Results

Due to the lack of 3D ground truth we project the estimated 3D models to the image plane and employ detection metrics to report full image performance as miss rate vs. false positives per image (FPPI) (see Fig. 10 and Fig. 12). To perform inference, we approximate the posterior mean using samples and obtain the respective hypotheses scores by projective marginalization as described in Sec. 4.2. Moreover, we use the *log-average miss rate* (LAMR) for an assessment across a large range of false positive rates. The LAMR is defined as the average miss-rate sampled from the lowest false positive rate to a false positive rate of 1 FPPI. Missing samples on curve for high FPPI rates are interpolated by the highest valid FPPI (the right-most sample on the curve). We use equally distant samples in log-space and therefore the log-average miss rate stresses low miss rates at high precision, which is preferable as system output. LAMRs for each curve are reported in parentheses in the respective legend.

We start by analyzing our models' performance on the ETH datasets. For all sequences we first analyze the detector performance for the deformable part model [54] and our own detector [53] and then build the proposed 3D scene models on top of the better performing detector. We first discuss the results in detail for *ETH-Loewenplatz* and then briefly analyze how the results consistently transfer to *ETH-Linthescher* and *ETH-PedCross2*.

7.1 ETH-Loewenplatz

We begin with an analysis of the 3D single-frame and scene tracklet model for which results are shown in Fig. 10(a). On this dataset the MultiFtr+Motion detector achieves an LAMR of 45.0% and outperforms DPM with an LAMR of 56.1%. The substantially worse performance of DPM on this dataset can be explained by the fact that 70.1% of the evaluated instances have a height of 60-100 pixels and therefore parts can not be detected very robustly, which impacts the overall performance. Adding the single-frame 3D scene model as described in Sec. 3 improves the performance to an LAMR of 35.4%. The 3D scene information is able to prune scene inconsistent detections on body parts or on background clutter (cf. Fig. 11(b),(d)) and can further improve the results when temporal information is added by the scene tracklet model, to a LAMR of 34.2%. When we add long-term data association and tracking with a HMM as described in Sec. 5.1 we achieve a LAMR of 45.6% for the single frame model and an LAMR of 37.6% for the scene tracklet model. For both models the performance seems to already saturate when leveraging temporal consistency in the short-term tracklets, while over longer timespans high scoring false detections that are consistent with the scene are reinforced in this setting.

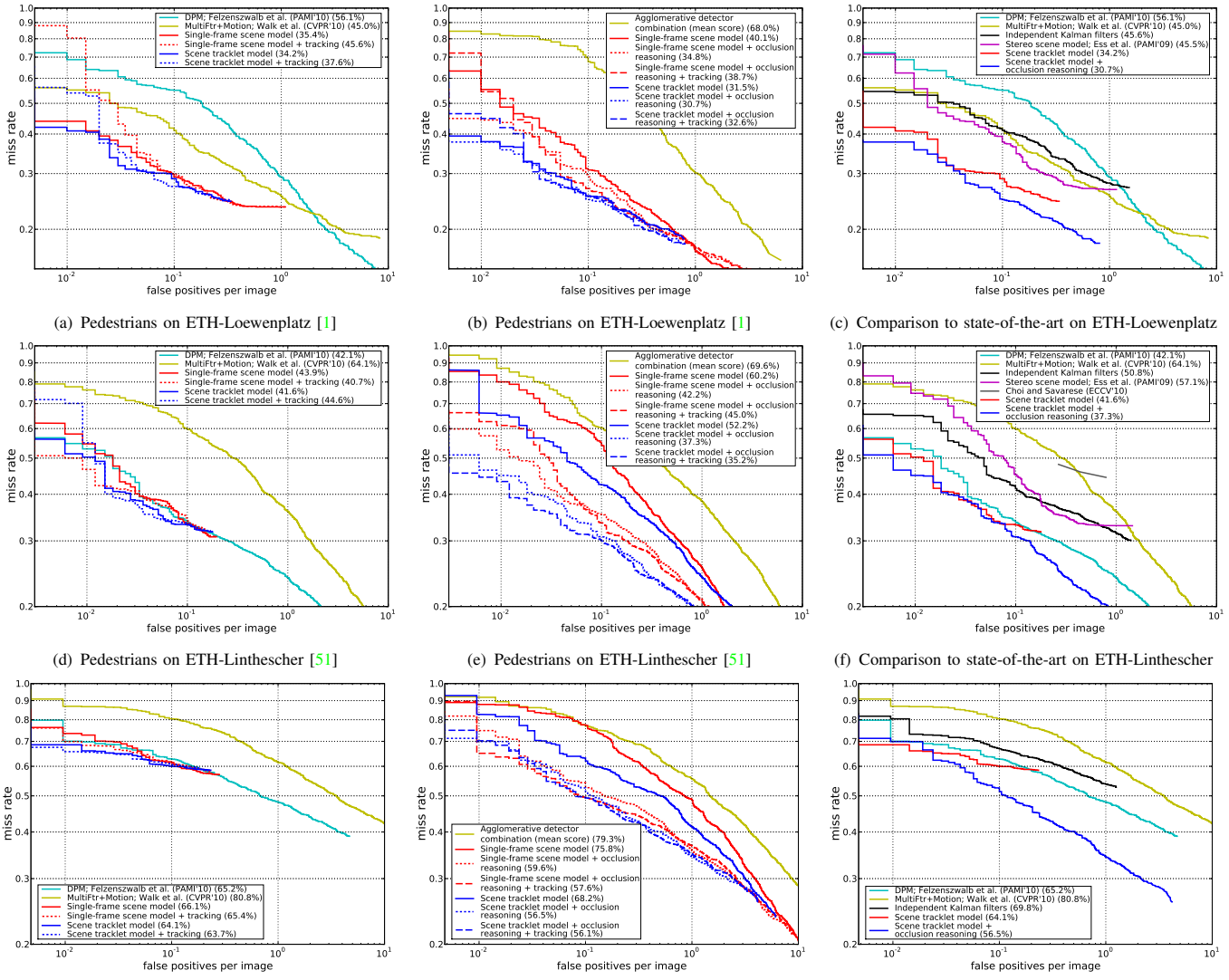
Next, we turn to our model with explicit occlusion reasoning (cf. Fig. 10(b)) as described in Sec. 3.2.2. Here, we use the performance of the agglomeratively clustered detections as an additional baseline. In this setting no 3D scene information is used and the partial object detectors, which by themselves perform worse than the full object detector (see [40] for details), return additional false positive detections and therefore the LAMR increases to 68.0%. When we add our 3D inference to prune inconsistent partial detections performance improves considerably and the single-frame model achieves an LAMR of 40.1% without the proposed occlusion reasoning and 34.8% with occlusion reasoning. Scene tracklet inference additionally aids performance further in the high precision regime and overall the LAMR improves to 31.5%. Additionally adding occlusion reasoning to the scene tracklet inference further improves our results to 30.7% LAMR. Similarly, to the model with a single full object detector adding long term data association with tracking does not further improve performance (38.7% LAMR for single frame model and 32.6% for scene tracklet model).

In Fig. 10(c) we compare our results to the state-of-the-art on this sequence. When we simply add independently operating Kalman filters for each object, the performance slightly decreases compared to the detector to an LAMR of 45.6%. This is due to the fact that very short tracks and tracks with few missing detections due to partial occlusion are lost. The stereo scene model by Ess et al. [1] performs slightly better[4] with an LAMR of 45.5%. Both the Kalman filters and the stereo scene model, are outperformed by our models which achieve the lowest LAMRs on this sequence. We attribute the improved performance to the tight integration of detection and segmentation within the 3D scene tracklet model, which is able to enforce consistency over time, consistency of segmentation and detections, as well as consistency with respect to object-object occlusion.

7.2 ETH-Linthescher

Next, we turn to the results on ETH-Linthescher for which plots are shown in Fig. 10(d)-(f). This sequence has been recorded in a busy pedestrian zone where pedestrians frequently occlude each other and tend to walk in groups (see Fig. 11(c),(d)). Due to the large scale-variation for pedestrians on this sequence the scene labeling algorithm yields only unsatisfying results and can not improve the overall performance. Therefore, we discard the respective potential Ψ_L for this series of experiments. As this dataset is dominated by front and back views of pedestrians, DPM (LAMR 42.1%) outperforms our motion enhanced detector (LAMR 64.1%). For the setup with a single full object DPM detector and without explicit occlusion reasoning our scene tracklet model achieves the best results with an LAMR of 41.6% (cf. Fig. 10(d)). With

4. The original results published in [1] were biased *against* Ess et al., because they did not allow detections slightly <60 pixels to match true pedestrians ≥ 60 pixels, discarding many correct detections. We therefore regenerated all FPPI-curves.



(a) Pedestrians on ETH-Loewenplatz [1] (b) Pedestrians on ETH-Loewenplatz [1] (c) Comparison to state-of-the-art on ETH-Loewenplatz

(d) Pedestrians on ETH-Linthescher [51] (e) Pedestrians on ETH-Linthescher [51] (f) Comparison to state-of-the-art on ETH-Linthescher

(g) Pedestrians on ETH-Pedcross2 [51] (h) Pedestrians on ETH-Pedcross2 [51] (i) Comparison to state-of-the-art on ETH-Pedcross2

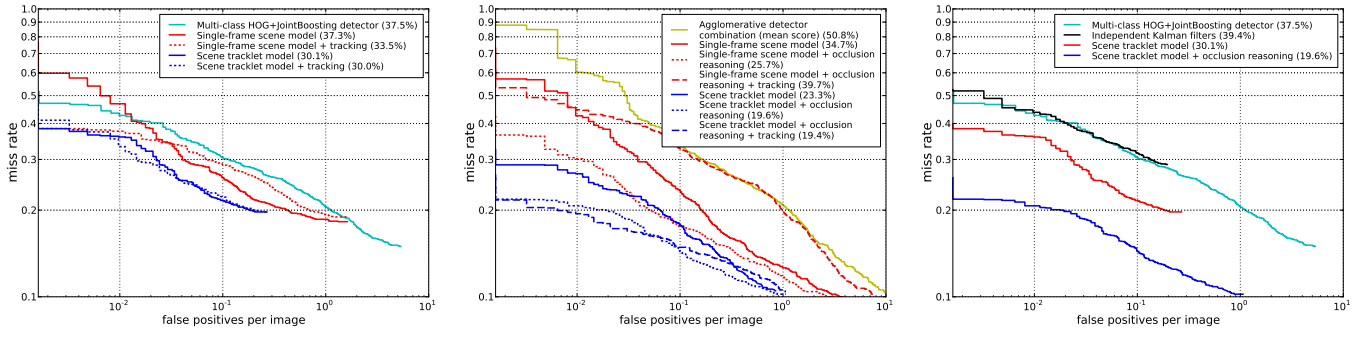
Fig. 10: Results obtained for *pedestrians* on *ETH-Loewenplatz* (first row), *ETH-Linthescher* (second row) and *ETH-Pedcross2* (third row) and comparison to the state-of-the-art (rightmost column). Experiments on ETH-Loewenplatz include the pixel labeling potential Ψ_L while it is being discarded for ETH-Linthescher and ETH-Pedcross2 due to unsatisfying labeling performance. Log average miss rate (for definition see Sec. 7) is reported in the according legend text. Figure best viewed in color.

this model we loose performance in particular for pedestrians that are only fully visible for very short periods of time. Therefore, when we additionally perform occlusion reasoning and use a bank of detectors (cf. Fig. 10(e)), we are able to improve performance up to an LAMR of 37.3%. We compare our models' performance to the state-of-the-art in Fig. 10(f). Similar to our scene tracklet model with full object detector only, independently running Kalman filters are not able to track pedestrians well that are fully visible only briefly, and lose performance compared to the detector, with a LAMR of 50.8%. The stereo scene model by Ess et al. achieves an LAMR of 57.1%. Choi and Savarese [17] report three points on the recall vs. FPPI curve, which are not competitive to our model; the miss rate is about 20% higher at the same error rates. Moreover, we note that [17] reports performance for the original annotations, which do not include all occluded pedestrians. Hence, the performance on the modified annotation set may be slightly worse.

7.3 ETH-PedCross2

This sequence has been recorded on a very crowded sidewalk with pedestrians occluding each other and pedestrians occurring close to the camera such that they can only be seen partially (cf. Fig. 11(f)-(h)). Similarly to *ETH-Linthescher* the scene labeling algorithm suffers from the large scale variation of pedestrians and thus we drop the scene labeling potential for this sequence as well. Like *ETH-Linthescher* this sequence mostly contains pedestrians seen from frontal and back views, hence our motion-enhanced detector (LAMR 80.8%) cannot outperform DPM (LAMR 65.2%). In particular we note that the maximum achieved recall of 61.0% for the full object detectors is not satisfactory. When we add our scene tracklet based 3D reasoning without partial detectors and occlusion reasoning the maximum recall becomes even worse (41.5%) as many detections in this sequence are only observed for single frames, which is too little evidence for

Fig. 11: Sample results for ETH-Loewenplatz ((a),(b)), ETH-Linthescher ((c)-(e)) and ETH-PedCross2 ((f)-(h)). All images are displayed for an error rate of 0.1 FPPI. The leftmost column shows the results of single-frame object detection; the second column adds the single-frame 3D model described in Sec. 3. The third column shows results which are obtained with scene tracklets while the rightmost column additionally adds occlusion reasoning. Our 3D scene model allows to prune scene inconsistent false detections, e.g. on body parts (see (b),(d)) and strengthens weak object detections over time (cf. (a)). Explicit occlusion reasoning allows to track and associate objects even if they have never been fully visible. Yellow bounding boxes indicate *pedestrian* hypotheses. Figure best viewed in color.



(a) Cars front on MPI-VehicleScenes [39]

(b) Cars front on MPI-VehicleScenes [39]

(c) Cars front comparison to state-of-the-art on MPI-VehicleScenes

(d) Trucks front on MPI-VehicleScenes [39]

(e) Trucks front on MPI-VehicleScenes [39]

(f) Trucks front comparison to state-of-the-art on MPI-VehicleScenes

Fig. 12: Results obtained with our system for *car front* (first row) and *truck front* (second row) on *MPI-VehicleScenes*. All scene models were run with the pixel labeling potential Ψ_L. Log average miss rate (for definition see Sec. 7) is reported in the respective legend text. Figure best viewed in color.

the tracklet framework (cf. Fig. 10(g)). Still, pedestrians that are visible for several frames can be successfully tracked and consequently performance in high precision increases and the overall LAMR decreases to 64.1%. When we further strengthen weak evidence by performing occlusion reasoning and use evidence from a bank of detectors (Fig. 10(h)) an improvement can be seen on shorter tracks – even when they are only partially visible throughout the entire track. The LAMR in this setting substantially improves to 56.5%, while the maximum recall is increased to 74.0%. Similarly to the scene tracklet model Kalman filters achieve a substantially lower maximum recall (13.7% loss compared to DPM) with an overall LAMR of 69.8% (cf. Fig. 10(i)). For this dataset we additionally analyzed the performance on partially occluded pedestrians. To that end we annotated all partially occluded pedestrians and performed the evaluation restricted to these instances. Overall 1052 pedestrians were marked as partially occluded out of which DPM [54] detected 40.7%. Our model without explicit occlusion reasoning was able to detect only 19.2%. This low recall compared to the standalone detector is mostly due to the tracklet formulation as discussed above. The proposed model with explicit occlusion reasoning, on the other hand, can solve this shortcoming and achieves a recall almost three times higher (55.0%).

7.4 MPI-VehicleScenes

Finally, we analyze our system's performance on the *MPI-VehicleScenes* dataset (see example results in Fig. 13). Compared to the other datasets the camera as well as all other objects are moving substantially faster, since part of the database has been recorded on rural highways. The two classes of interest on this dataset are *cars front* and *truck front*. We start by analyzing the performance for the detection of frontal cars (cf. Fig. 12(a)) for which our boosted detector achieves a LAMR of 37.5%. In particular the performance for oncoming cars is impaired when they are occluded by another car driving in front of them. Adding the single-frame scene model slightly improves performance to a LAMR of 37.3% by pruning scene inconsistent false detections, mostly responses of the full-object detector on vehicle parts. Scene tracklet inference further improves performance by strengthening hypotheses over time (LAMR of 30.1%). When we add explicit occlusion reasoning (cf. Fig. 12(b)) we are able to increase performance to a LAMR of 19.6%. Especially the performance on partially visible oncoming cars is greatly improved with this model (see for instance Fig. 13(c)). Again we compare our models to Kalman filters as baseline in Fig. 12(c) which achieve an overall LAMR of 39.4% but loose substantial recall. This is due to the fact that partially occluded oncoming cars in the distance are detected only occasionally, and hence no tracks can be set up. For frontal cars we also analyze the impact when model is run with varying potentials (cf. Fig. 14). Most performance is gained by tracklet and occlusion reasoning, but best performance can only be achieved by the full model including bottom-up scene labeling.

Finally, we discuss the performance for *truck front* for which the object detector achieves an LAMR of 69.4%. Compared to the car detector the performance is considerably worse because

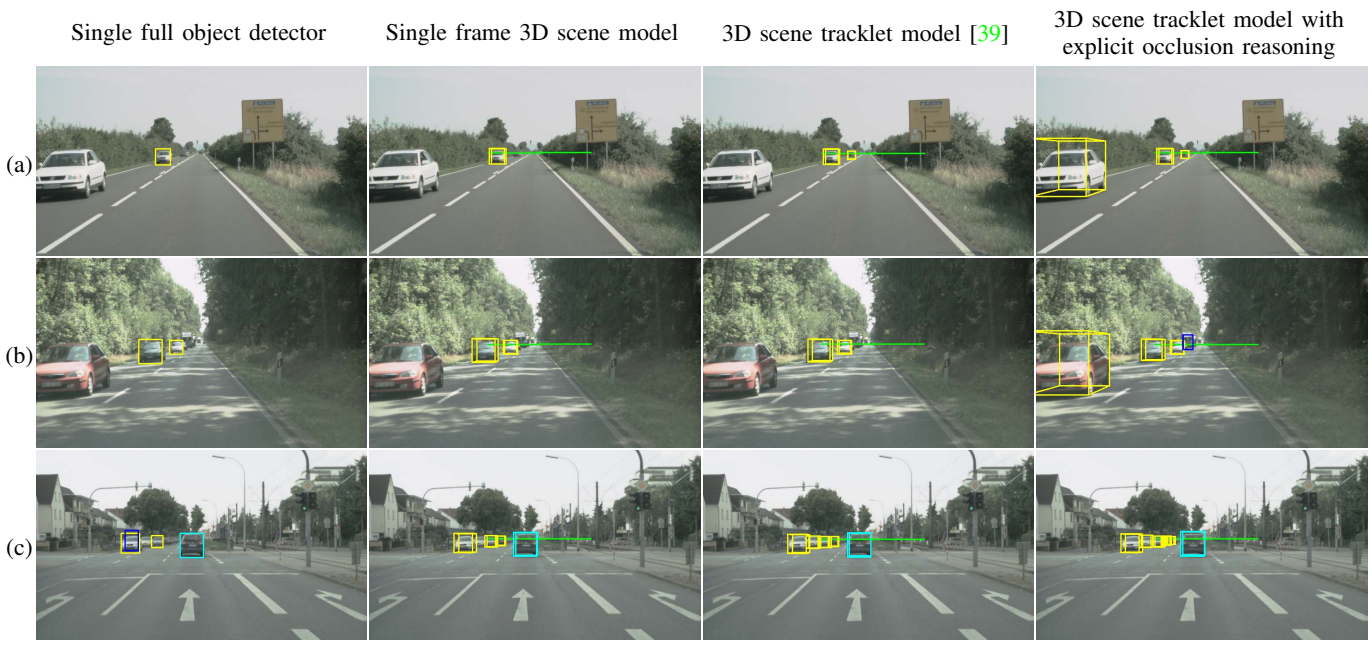


Fig. 13: Example results on MPI-VehicleScenes. All images are displayed for an error rate of 0.1 FPPI. The leftmost column shows the results of single-frame object detection; the second column adds the single-frame 3D model described in Sec. 3. The third column shows results that are obtained with scene tracklets while the rightmost column additionally adds occlusion reasoning. Our tightly integrated 3D scene model strengthens scene consistent object detections over time in particular for far ranges (cf. (a),(b)). Moreover, it allows to prune false detections that typically appear on instances of other similar object classes (cf. (c)). Partial detectors with explicit occlusion reasoning allow to improve performance for partially visible objects even when those are never fully visible. (cf. (b),(c)). Yellow bounding boxes indicate the *car front* class, dark blue *truck front* and light blue *car rear*. Figure best viewed in color.

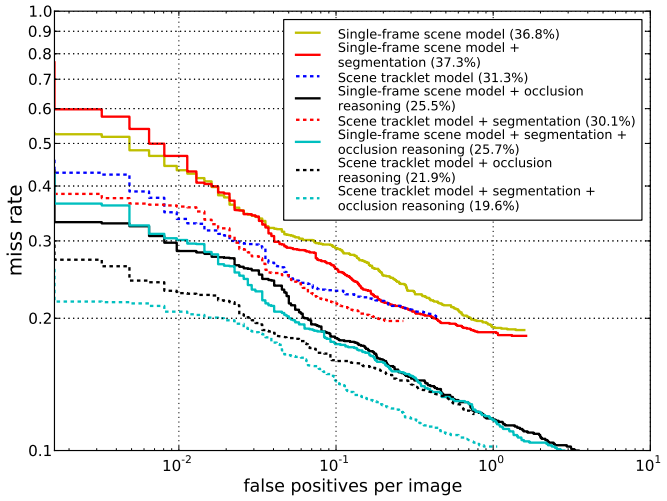


Fig. 14: Results obtained for cars front on *MPI-VehicleScenes* with varying system components

of larger intra-class variability. False detections typically arise as detections on cars (see Fig. 13(c)). The inconsistence with the bottom-up scene labeling allows to resolve close-range false positives and improves performance compared to the detector with an LAMR of 59.0%. The scene tracklets model is able to improve on spurious false detections but also misses true detections which are not continuously detected so that the overall performance remains unchanged (LAMR 59.0%). Adding occlusion reasoning brings a further improvement to 45.3% LAMR. Fig. 12(f) compares our models to Kalman filters, which only achieve a LAMR of 67.5%, for the same reason as on cars.

7.5 Discussion

Overall, our experiments on four datasets and on four different object classes indicate that our 3D scene tracklet model is able to leverage scene context to robustly infer both the 3D scene geometry and the presence of objects in that scene from a monocular camera. This performance is mainly due to the use of a strong *tracking-by-detection* framework which employs *tracklets* on a scene level, thereby leveraging evidence from a number of consecutive frames. The tight coupling with the observation model allows to exploit *3D scene context* as well as to *combine multiple cues* of a detector and from scene labeling. Moreover, our 3D scene model enables us to perform explicit occlusion reasoning. Inferred object visibility can easily be embedded into the integrated formulation as a mixture-of-experts, and it improves performance on all tested sequences, for all four object classes. Long-term tracking with an HMM does not lead to additional gains. In all cases, independent extended 3D Kalman filters cannot significantly improve the output of state-of-the-art object detectors, and are greatly outperformed by the integrated state model. Comparing to other work that integrates detection and scene modeling, we outperform [1] for the case of pedestrians, even though we do not use stereo information, and our 3D scene tracklet model also outperforms the competing monocular approach of [17], which is less tightly integrated.

8 Conclusion

We have presented a probabilistic 3D scene model, that enables multi-frame tracklet inference on a scene level in a tracking-

by-detection framework. Our system performs monocular 3D scene geometry estimation in realistic traffic scenes, and leads to more reliable detection of objects such as pedestrians, cars, and trucks. Leveraging the 3D scene model enables us to perform explicit occlusion reasoning and reliably track objects even under partial occlusion or when they have never been fully visible. We exploit information from object (category) detection and low-level scene labeling to obtain a *consistent 3D description of an observed scene*, even though we only use a single camera. Our experimental results show a clear improvement over top-performing state-of-the-art object detectors. Moreover, we significantly outperform basic Kalman filters, a competing monocular system [17], as well as a state-of-the-art stereo-based system [1].

Our experiments underline the observation that objects are valuable constraints for the underlying 3D geometry, and vice versa (cf. [1, 7]), so that a joint estimation can improve detection performance.

For future work it would be interesting to explore the fusion with complementary sensors such as RADAR or LIDAR, which should allow for further improvements.

REFERENCES

[1] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Robust multi-person tracking from a mobile platform," *PAMI*, vol. 31(10), pp. 1831–1846, 2009. 1, 2, 8, 10, 11, 14, 15

[2] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *IJCV*, vol. 73, pp. 41–59, 2007. 1, 2

[3] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *ECCV*, 2004. 1, 3

[4] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *CVPR*, 2008. 1

[5] R. Kaucic, A. G. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs, "A unified framework for tracking through occlusions and across sensor gaps," in *CVPR*, 2005. 1, 3, 7

[6] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *ECCV*, 2008. 1, 3, 7

[7] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *IJCV*, vol. 80, no. 1, pp. 3–15, 2008. 1, 2, 15

[8] A. Torralba, "Contextual priming for object detection," *IJCV*, vol. 53(2), pp. 169–191, 2003. 2

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "*TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV*, 2006. 2

[10] Z. Tu, X. Chen, A. Yuille, and S. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *IJCV*, vol. 63, no. 2, 2005. 2

[11] A. Ess, T. Müller, H. Grabner, and L. Van Gool, "Segmentation-based urban traffic scene understanding," in *BMVC*, 2009. 2

[12] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using SfM point clouds," in *ECCV*, 2008. 2

[13] A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *ECCV*, 2010. 2

[14] U. Franke, C. Rabe, H. Badino, and S. Gehrig, "6D-vision: Fusion of stereo and motion for robust environment perception," in *DAGM*, 2005. 2

[15] A. Geiger, M. Lauer, and R. Urtasun, "A generative model for 3d urban scene understanding from movable platforms," in *CVPR*, 2011. 2

[16] A. Shashua, Y. Gdalyahu, and G. Hayun, "Pedestrian detection for driving assistance systems: Single-frame classification and system level performance," in *Proc. IEEE International Conf. on Intelligent Vehicles (IV)*, 2004. 2

[17] W. Choi and S. Savarese, "Multiple target tracking in world coordinate with single, minimally calibrated camera," in *ECCV*, 2010. 2, 11, 14, 15

[18] A. Ess, K. Schindler, B. Leibe, and L. Van Gool, "Improved multi-person tracking with active occlusion handling," in *ICRA Workshop on People Detection and Tracking*, 2009. 2

[19] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *CVPR*, 2009. 2

[20] V. Shet, J. Neumann, V. Ramesh, and L. Davis, "Bilattice-based Logical Reasoning for Human Detection," in *CVPR*, 2007. 2

[21] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *IJCV*, vol. 82, no. 2, pp. 185–204, 2009. 2

[22] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *CVPR*, 2010. 3

[23] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *ICCV*, 2009. 3

[24] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Fast object detection with occlusions," in *ECCV*, 2004. 3

[25] S. Kwak, W. Nam, B. Han, and J. H. Han, "Learning occlusion with likelihoods for visual tracking," in *ICCV*, 2011. 3

[26] J. Winn and J. Shotton, "The layout consistent random field for recognizing and segmenting partially occluded objects," in *CVPR*, 2006. 3

[27] A. Vedaldi and A. Zisserman, "Structured output regression for detection with partial truncation," in *NIPS*, 2009. 3

[28] T. Gao, B. Packer, and D. Koller, "A segmentation-aware object detection model with occlusion handling," in *CVPR*, 2011. 3

[29] L. Sigal and M. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *CVPR*, 2006. 3

[30] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *ICCV*, 2009. 3

[31] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: HybridBoosted multi-target tracker for crowded scene," in *CVPR*, 2009. 3, 7

[32] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *CVPR*, 2008. 3

[33] A. Andriyenko and K. Schindler, "Globally optimal multi-target tracking on a hexagonal lattice," in *ECCV*, 2010. 3

[34] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR*, 2011. 3

[35] H. B. Shitrit, J. Berclaz, F. Fleuret, , and P. Fua, "Tracking Multiple People under Global Appearance Constraints," in *ICCV*, 2011. 3

[36] Z. Khan, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," *PAMI*, vol. 27(11), pp. 1805–1819, 2005. 3

[37] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *PAMI*, vol. 30, no. 7, pp. 1198–1211, 2008. 3

[38] M. Isard and J. MacCormick, "BraMBLe: A Bayesian Multiple-Blob tracker," in *ICCV*, 2001, pp. 34–41. 3

[39] C. Wojek, S. Roth, K. Schindler, and B. Schiele, "Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes," in *ECCV*, 2010. 4, 8, 12, 13, 14

[40] C. Wojek, S. Walk, S. Roth, and B. Schiele, "Monocular 3d scene understanding with explicit occlusion reasoning," in *CVPR*, 2011. 4, 10

[41] N. Dalal, "Finding people in images and videos," Ph.D. dissertation, Institut National Polytechnique de Grenoble, July 2006. 4

[42] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 2000, pp. 61–74. 4

[43] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991. 5

[44] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *PAMI*, vol. 29, no. 5, pp. 854–869, 2007. 5, 9

[45] W. Gilks, S. Richardson, and D. Spiegelhalter, Eds., *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1995. 6

[46] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, 1995. 6

[47] V. K. Singh, B. Wu, and R. Nevatia, "Pedestrian tracking by associating tracklets using detection residuals," in *Proceedings of the 2008 IEEE Workshop on Motion and Video Computing*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 1–8. 7

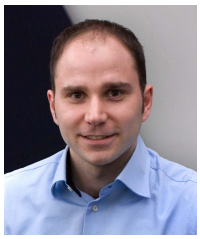
[48] P. M. Jorge, A. J. Abrantes, and J. S. Marques, "On-line tracking groups of pedestrians with bayesian networks," in *6th International Workshop on Performance Evaluation for tracking and Surveillance (PETS)*, 2004. 7

[49] A. G. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-

object tracking through simultaneous long occlusions and split-merge conditions," in *CVPR*, 2006. 7
[50] P. Nillius, J. Sullivan, and S. Carlsson, "Multi-target tracking-linking identities using Bayesian network inference," in *CVPR*, 2006. 7
[51] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *CVPR*, 2008. 8, 11
[52] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, vol. 99, no. PrePrints, 2011. 8
[53] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *CVPR*, 2010. 8, 10
[54] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, pp. 1627–1645, 2010. 8, 9, 10, 13
[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005. 8, 9
[56] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel SVMs is efficient," in *CVPR*, 2008. 9
[57] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *CVPR*, 2009. 9
[58] Y. Hel-Or and H. Hel-Or, "Real-time pattern matching using projection kernels," *PAMI*, vol. 27, pp. 1430–1445, 2005. 9
[59] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination indepence," in *ECCV*, 2002. 9

Christian Wojek received his Diplom degree in Computer Science from the University of Karlsruhe in 2006 and his PhD from TU Darmstadt in 2010. He was awarded a DAAD scholarship to visit McGill University from 2004 to 2005. He was with MPI Informatics Saarbrücken as a postdoctoral fellow from 2010 to 2011 and in 2011 joined Carl Zeiss Corporate Research. His research interests are object detection, scene understanding and activity recognition in particular with application to real world scenarios.

Stefan Walk received his Masters degree in Physics from TU Darmstadt in 2007. He worked as a PhD student at the Computer Science department in Darmstadt until 2011. In 2012, he joined the Institute of Geodesy and Photogrammetry of ETH Zürich. His scientific interests are machine learning and its applications to computer vision, in particular object detection and multi-class classification in adverse imaging conditions.



Stefan Roth received the Diplom degree in Computer Science and Engineering from the University of Mannheim, Germany in 2001. In 2003 he received the ScM degree in Computer Science from Brown University, and in 2007 the PhD degree in Computer Science from the same institution. Since 2007 he is an assistant professor of Computer Science at Technische Universität Darmstadt, Germany. His research interests include probabilistic and statistical approaches to image modeling, motion estimation, human tracking, and object recognition. He received several awards, including an honorable mention for the Marr Prize at ICCV 2005 (with M. Black), the Olympus-Prize 2010 of the German Association for Pattern Recognition (DAGM) and the Heinz Maier-Leibnitz Prize 2012 of the German Research Foundation (DFG).

Konrad Schindler received a Diplomingenieur (M.tech) degree in photogrammetry from Vienna University of Technology, Austria in 1999, and a PhD from Graz University of Technology, Austria, in 2003. He has worked as a photogrammetric engineer in the private industry, and held researcher positions in the Computer Graphics and Vision Department of Graz University of Technology, the Digital Perception Lab of Monash University, and the Computer Vision Lab of ETH Zurich. He became assistant professor of Image Understanding at TU Darmstadt in 2009, and since 2010 has been a tenured professor of Photogrammetry and Remote Sensing at ETH Zurich. His research interests lie in the field of computer vision, photogrammetry, and remote sensing, with a focus on image understanding and 3d reconstruction. He currently serves as head of the Institute of Geodesy and Photogrammetry, and as associate editor for the ISPRS Journal of Photogrammetry and Remote Sensing, and for the Image and Vision Computing Journal.

Bernt Schiele received his masters in computer science from Univ. of Karlsruhe and INP Grenoble in 1994. In 1997 he obtained his PhD from INP Grenoble in computer vision. He was a postdoctoral associate and Visiting Assistant Professor at MIT between 1997 and 2000. From 1999 until 2004 he was an Assistant Professor at ETH Zurich and from 2004 to 2010 he was a full professor of computer science at TU Darmstadt. In 2010, he was appointed scientific member of the Max Planck Society and a director at the Max Planck Institute for Informatics. Since 2010 he has also been a Professor at Saarland University. His main interests are computer vision, perceptual computing, statistical learning methods, wearable computers, and integration of multi-modal sensor data. He is particularly interested in developing methods which work under real-world conditions.