What Makes an On-road Object Important?

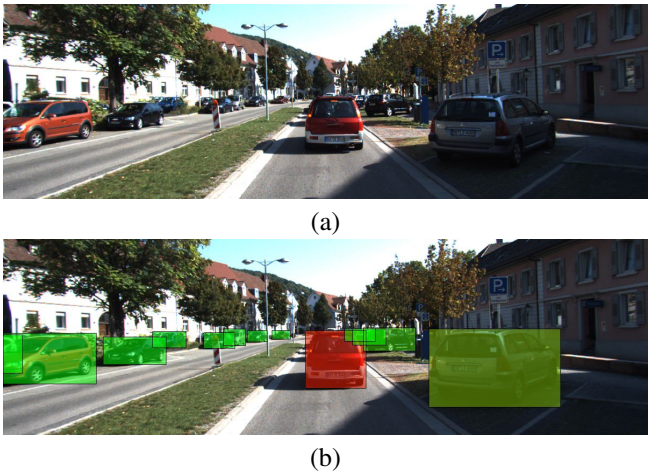
Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California San Diego
{eohnbar, mtrivedi}@ucsd.edu

***Abstract*—Human drivers continuously attend to important scene elements in order to safely and smoothly navigate in intricate environments and under uncertainty. This paper develops a human-centric framework for object recognition by analyzing a notion of object importance, as measured in a spatio-temporal context of driving a vehicle. Given a video, a main research question in this paper is - which of the surrounding agents are most important? The answer inherently requires complex reasoning over the current driving task, object properties, scene context, intent, and possible future actions. Therefore, we find that various spatio-temporal cues are relevant for the importance classification task. Furthermore, we demonstrate the usefulness of the importance annotations in evaluating vision algorithms (specifically, for the task of object detection) in an application where trust in automation is imperative and errors are costly. Finally, we show that importance-guided training of object detection models results in improved detection performance of surrounding objects of higher importance. Hence, such models may be better suited for use in representing safety-critical situations, predicting surrounding agents' intentions, and in human-robot interactivity. The dataset and code will be made publicly available.**

I. Introduction

Consider the image in Fig. 1(a). Can you try to detect all of the vehicles in the image? Some of the vehicles in the proximity of the ego-vehicle are easy to detect while other vehicles with increased distance, occlusion, and truncation may be more challenging. On the road, human drivers attend to only a subset of the on-road occupants which are most relevant to the current navigation task, given the situational spatio-temporal context. Motivated by this phenomenon, the goal of this paper is to learn object importance ranking models which can automatically predict the importance score of objects in a driving scene.

Next, please re-consider the image in Fig. 1(a). Although knowledge of all obstacles is important for obstacle avoidance while navigating, some objects may require more of a driver's attention than others (Fig. 1(b)). For instance, remote objects may be less relevant than near ones to the immediate navigation task, but other object properties may also play a role. The answer to the question of what makes an object important inherently requires reasoning over the current driving task, object properties, scene context, intent, previous experience, and *expected future actions* [2]. Humans rely on situational awareness in order to continuously analyze the scene and its salient elements [2], [3], and intelligent and self-driving vehicles are expected to perform similar reasoning skills for smooth and safe navigation (as well as earn the trust of their users). This study is also concerned with learning human-centric models for scene perception, which can be valuable to a wide range of applications in computer vision including saliency [4]–[7], robotics [8], and ego-centric vision [9], [10].



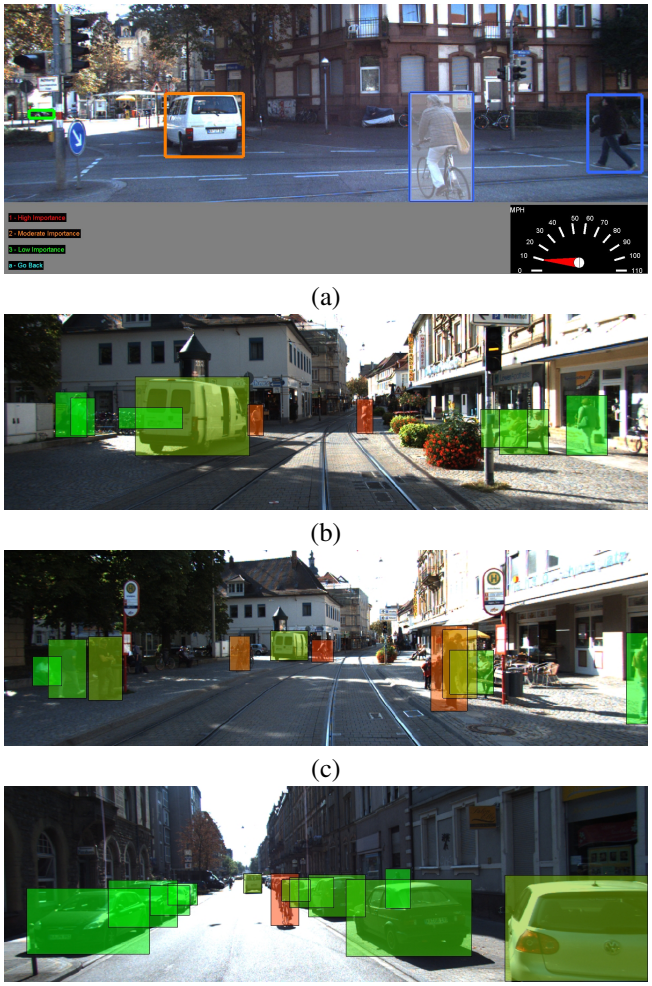
(a) (b)

Fig. 1: This paper studies on-road object recognition under a notion of importance. (a) An example frame from a driving video used in this study (taken from the KITTI dataset [1]). Annotators/models observe the video and produce object-level importance annotations as relevant to the navigation task. (b) Example frame with overlaid object-level importance (averaged over subjects), colored from high to moderate to low importance, provided by human annotators.

A. Contributions

Our paper makes the following contributions in the important field of computer vision and pattern recognition for intelligent vehicles.

Dataset and cue analysis: We collect object-level importance annotations on KITTI videos [1] from a varied subject pool. The dataset is used to learn importance ranking models which provide insights into what makes an object important. The proposed prediction task allows for evaluating a complex spatio-temporal reasoning skill in real-world settings, and a wide range of cues is shown to impact spatio-temporal prediction of importance. We emphasize that this study is not concerned with ethical issues in autonomous driving, but with obtaining a better understanding of the limitations and requirements for on-road object recognition, safe navigation, and human-centric AI.



(a)

(b)

(c)

(d)

Fig. 2: (a) The interface used to obtain object-level importance ranking annotations. The cyclist is highlighted as it is the currently queried object to annotate, colored boxes have already been annotated with an importance level by the annotator, and blue boxes are to be annotated. (b-d) Example frames in the dataset with overlaid object-level importance (averaged over subjects). Object type, state, position, orientation, scene context, and other cues are all shown to play a role in determining an importance score by annotators.

Importance-guided training and evaluation of models: Our analysis demonstrates that the large majority of road occupants on the KITTI benchmark [1] are consistently categorized as low important objects. As highly important objects are rare, the importance annotations give rise to novel performance metrics and error/bias analysis in an application where errors made by an algorithm are safety-critical. Specifically for the task of object detection, we demonstrate that importance-guided training can significantly improve detection of highly important objects. As such joint detection-attention models are encouraged to emphasize saliency and contextual cues in detection, they can provide more useful output for driver assistance and safe navigation tasks. Interestingly, we show that such an important insight would have been difficult to identify with traditional detection performance evaluation metrics.

II. Importance Annotations Collection

The KITTI dataset [1], [11] is widely used for evaluating a variety of vision tasks for autonomous driving, including object detection and tracking. We supplement the object annotations with an object importance label. In addition to image-level tracklet annotations of road occupants (pedestrians, cyclists, and vehicles), KITTI provides a rich set of other modalities and semantic object attributes, including a 3D box and orientation (annotated in the LIDAR data), IMU/ego-vehicle dynamics, GPS, and an occlusion state. The sensorization and annotation makes KITTI an ideal dataset for studying cues related to importance in driving settings.

Importance annotations: Experiments were done in a driving simulator with KITTI videos shown on a large screen. A set of 8 videos was selected for annotation. Subjects initially watched each video twice, and then annotated every 10^{th} frame by providing an object-level integer between 1-3 (1 being high and 3 being low importance). A visualization of the annotation interface is shown in Fig. 2. The choice of three importance levels was chosen in order to reduce ambiguity as much as possible without overly restricting the experiments. For instance, allowing for only two levels (yes/no) of importance is somewhat restrictive as there may be ambiguous cases where a decision can't be confidently made. Furthermore, it is reasonable to expect that some objects will fall under a middle between high and low importance. On the other hand, a continuous ranking score may have been used, but could have lead to large confusion among subjects and more guessing, which we aimed to reduce. Subjects were asked to imagine driving under similar situations, and mark objects by the level of attention and relevance they would've given the object under real driving.

Fig. 3 visualizes the output provided by the subjects. Although the instructions were kept fixed among the experiments, the resulting annotations contained variations due to the task's inherent subjectivity. Upon a closer inspection, we found a strong correlation between annotation output and subject driving experience. Interestingly, greater driving experience implied a higher percentage of moderate and high importance annotations (Fig. 3(b)). Subjects reported inspecting object orientation (useful for knowledge of future activities [12]), relative position, existence of traffic barriers, and relationship among objects for determining the importance label. We note that despite the subjective nature of the task, a subset of the annotated objects does contain high consistency. For instance, scenarios of dense scenes with tens of road occupants that are heavily occluded or are across a barrier consistently generated low importance annotations. This applies to a large number of objects in KITTI. Three importance classes are obtained and used throughout the paper by taking the median vote among subjects for each sample. The dataset used in the experiments

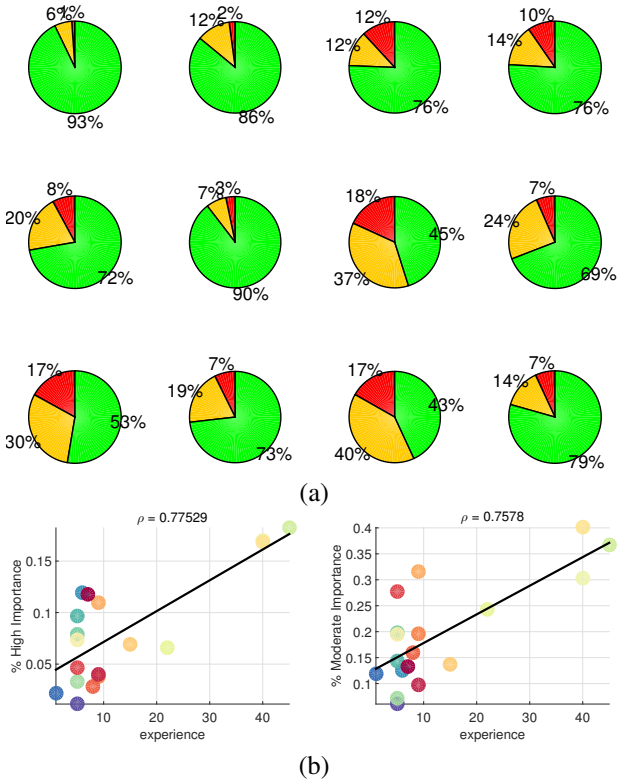


Fig. 3: (a) Object type distribution per subject (a subset of 12 out of the 18 total), color coded from high to moderate to low importance. (b) Annotations were shown to correlate with subject's driving experience (in years). Each subject is represented by a differently colored dot in the plot.

contains 17,635 object annotations, out of which 15,057 are vehicles (cars, vans, and trucks), 1,452 pedestrians, and 562 cyclists. Out of these totals, there were high/moderate/low importance (by median vote among subjects) 293/2159/12,605 vehicles, 143/524/785 pedestrians, and 267/147/148 cyclists. High important objects were generally shown to be not occluded and within 40 meters or less of the ego-vehicle. As can been observed from the collected statistics in Fig. 3, objects ranked as high importance are rare.

III. Object Importance Model

We continue with our aim to better understand object importance in the context of driving using a classification study. As mentioned in Section II, KITTI contains a set of object attributes which can be investigated for relevance to importance prediction. These are obtained either from annotation or from additional modalities (LIDAR, GPS, etc.). Nonetheless, the ultimate goal is training visual prediction models for object importance. Hence, we experiment with two types of models, $M_{attributes}$ and M_{visual}, which we'll define next. The weights for the models are learned using a logistic regression classifier.

A. Object attributes model, $M_{attributes}$

For an instance s and class importance $c \in \{1, 2, 3\}$, we train the following prediction model

$$M_{attributes}(s) = \boldsymbol{w}_{c,2D-obj}^T \phi_{2D-obj}(s) + \boldsymbol{w}_{c,3D-obj}^T \phi_{3D-obj}(s) + \boldsymbol{w}_{c,ego}^T \phi_{ego}(s) + \boldsymbol{w}_{c,temporal}^T \phi_{temporal}(s) \quad (1)$$

where each term is defined below.

2D object features: After projection of the annotated 3D object box to the image plane, the $\phi_{2D-obj} \in \mathbb{R}^4$ features are the concatenation of the height in pixels, aspect ratio, occlusion state (either none, partial, and heavy occlusion) and truncation percentage.

3D object features: $\phi_{3D-obj} \in \mathbb{R}^6$ is composed of the 3D left-right and forward-backward range coordinates (x, z) given by the LIDAR, Euclidean distance from the ego-vehicle, orientation in bird's eye view, and object velocity components, $|V|$ and $\angle V$.

Ego-vehicle features: Since ego-vehicle speed may impact which objects are considered important and this information is presented to the annotators, the attribute model includes ego-vehicle velocity and orientation features $\phi_{ego} = [\text{ego}|V|,\ \text{ego}\angle V]$.

Temporal attributes: We hypothesize that past information has an influence on driver expectation as to what is happening and will happen next. Hence, temporal evolution of attributes may contain relevant information for importance ranking. This assumption is captured in $\phi_{temporal}$, where the object and ego-vehicle attributes described above are concatenated over a past time window. Additionally, max-pooling over the time window and the Discrete Cosine Transform (DCT) coefficients [13] are also included in computing $\phi_{temporal}$.

B. Visual cues model, M_{visual}

Given the 2D bounding box annotation of objects in video, we propose a visual prediction model for mapping an image region to an importance class,

$$M_{visual}(s) = \boldsymbol{w}_{c,obj}^T \phi_{obj}(s) + \boldsymbol{w}_{c,spatial}^T \phi_{spatial}(s) + \boldsymbol{w}_{c,temporal}^T \phi_{temporal}(s) \quad (2)$$

Object visual features: For $\phi_{obj} \in \mathbb{R}^{4096}$ features, we employ the activations of the last fully connected layer of the OxfordNet (VGG-16) [14] convolutional network. The network was pre-trained on the ImageNet dataset [15] and fine-tuned on KITTI using Caffe [16].

Spatial context features: In order to better capture spatial context, such as relationship with other objects in the scene or occlusion state, each object instance is padded by an amount of $\times 1.75$ for generating $\phi_{spatial} \in \mathbb{R}^{4096}$.

Temporal context features: Given an object tracklet of box positions and a temporal window, temporal $\phi_{temporal}$ features are extracted. The previous object and spatial context features, ϕ_{obj} and $\phi_{spatial}$, are computed over a time window, concatenated, and max-pooled.

IV. Experimental Evaluation of Importance Models

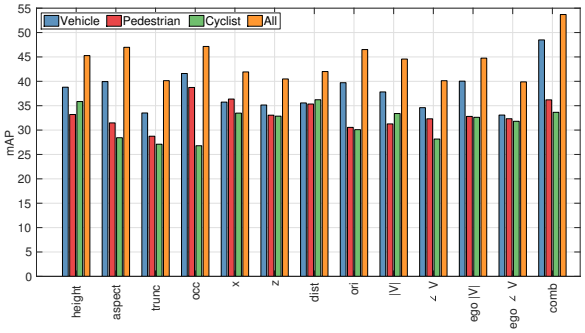
For each of the three importance classes, a precision-recall curve is computed and the area under the curve (AP) provides a summary metric for classification quality of the model. Finally, AP is averaged over the three important classes (mAP) for an overall classification quality metric (higher mAP value implies better classification performance). We note that by treating each importance class as a separate class, the proposed evaluation procedure handles the large imbalance in sample size of the importance classes. The dataset is split approximately in half for a 2-fold cross validation. The split is such that no samples from the same video are used for both training and testing.

Analysis of $M_{attributes}$: Table I shows the results of the classification experiments with varying combinations of features and an object class-agnostic importance prediction model. As $M_{attributes}$ employs clean annotation (occlusion state, truncation level using ground truth 3D boxes, etc.) and sensor data, it provides a strong classification baseline. In addition to providing a comparison with M_{visual}, it is also useful for analysis of cue relevance. Fig. 4(a) shows the classification power of each of the features and a combination of all, which leads to a performance of 53.70% mAP. We note that the training and testing sets are different for each of the object classes in Fig. 4(a), either limited to a specific object class or over all objects shown in the 'All' category. When training a class-agnostic model, all object classes are treated as one and selected cues for importance prediction must be learned such that they generalize over the different object classes. For the vehicle object class, occlusion state is shown to be a strong cue when determining an object's importance class. This is to be expected, as occlusion by another object may generally reduce the object's importance class. Nonetheless, a combination of the occlusion state with the other attributes provides a better representation of the object's importance class (increasing mAP from 47.15% to 53.70%). Vehicle orientation is also shown to be particularly useful when considering the vehicle object class, as it relates to traffic direction and future action. 2D object properties, such as height in pixels, are important as they are related to distance perception. For the pedestrian object class, occlusion state, longitudinal placement, and distance from the ego-vehicle are all shown to be particularly useful in predicting importance. The patterns for the cyclist object class are less clear as it contains a smaller number of object samples, yet similar observations hold.

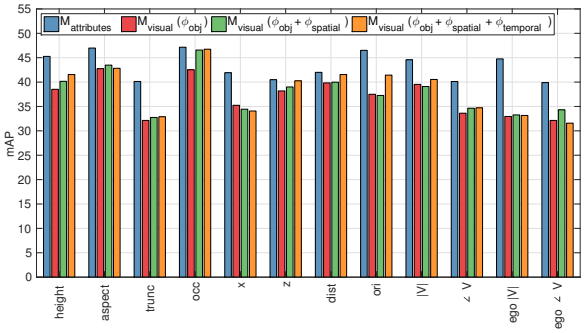
Table I also demonstrates the impact of temporal cue modeling for importance class prediction. When considering a class-agnostic model, the temporal window features are shown to improve performance over a window of up to 2.5 seconds. Incorporation of $\phi_{temporal}$ leads to a significant classification improvement, from 53.70% to 60.35%. Therefore, a good temporal context model is essential for the task of importance prediction. On the vehicle object class, incorporation of temporal attribute features leads to the largest improvement, from 48.49% to 56.18%. For the pedestrian object class, the performance significantly improves as well, from 36.20% to 40.09%. The improvement for the cyclist object class is minor, but obtaining further instances is required for drawing reliable conclusions.

TABLE I: Summary of the classification experiments using the two models discussed in Section III.

Model	mAP (%)
$M_{visual}(\phi_{obj})$	51.06
$M_{visual}(\phi_{obj} + \phi_{spatial})$	55.53
$M_{visual}(\phi_{obj} + \phi_{temporal})$	53.30
$M_{visual}(\phi_{obj} + \phi_{spatial} + \phi_{temporal})$	56.34
$M_{attributes}$ (w/o $\phi_{temporal}$)	53.70
$M_{attributes}$ (with $\phi_{temporal}$)	**60.35**



(a)



(b)

Fig. 4: (a) Impact of each attribute (and all combined) on object importance classification using $M_{attributes}$. (b) Explicit limitation analysis by employing M_{visual} with various feature combinations for individual attribute regression and consequent importance class classification. See Section III-A for attribute definitions.

Analysis of M_{visual}: The analysis of $M_{attributes}$ revealed which cues are useful for importance prediction. The visual prediction model must also be able to accurately capture object attributes, including object occlusion state, orientation, and distance from the ego-vehicle. The ϕ_{obj} baseline, of importance classification using a local object window, provides a performance of 51.06%. While the network is able to capture

TABLE II: Evaluation of object detection models (average-precision) using the proposed set of importance metrics and the Faster-RCNN framework (FRCN) [17]. '**IG**' stands for importance-guided cost-sensitive training.

	Traditional Test Settings			**Importance Test Settings**		
Method	Easy	Moderate	Hard	High	High+Moderate	Low
FRCN-ZF	89.26	79.70	64.96	66.89	82.80	58.85
FRCN-ZF-IG	91.09	80.86	66.18	73.00	87.19	59.90
ΔAP	+1.83	+1.16	+1.22	+6.11	+4.39	+1.05
FRCN-VGG	95.63	88.98	74.65	81.73	91.60	69.54
FRCN-VGG-IG	94.54	88.71	74.01	85.13	91.67	69.09
ΔAP	-1.09	-0.27	-0.64	+3.40	+0.07	-0.45

local information regarding occlusion and orientation, it can't fully capture the situational spatio-temporal context. We've experimented with several techniques for cue extraction in order to improve this baseline, and Table I lists two successful choices. First, we experimented with increasing the object box size for including surround context. Although commonly done for the task of object detection [18], [19], the impact on importance ranking needs to be studied. The assumption here is that certain attributes may become more easily recognizable with this increased box size, while also allowing for better scene and object-object relationship representation. Consequently, $\phi_{obj} + \phi_{spatial}$ improve performance over the baseline by a significant 4.47 mAP points, revealing insight as to the type of cues used by human annotators when ranking objects. As temporal features where shown to have a large impact on mAP in the case of $M_{attributes}$, temporal reasoning is expected to benefit the M_{visual} model as well. Incorporation of temporal features results in an improvement of 2.24 mAP points for $\phi_{obj} + \phi_{temporal}$ over the ϕ_{obj} baseline, and 0.81 mAP points when spatial context is included as well. Although the visual prediction model performance are impressive, it falls short of the attribute prediction model (see Fig. 4(b)). This motivates further study of models suitable for capturing spatio-temporal visual cues as a next research step [20]–[23].

V. Importance-Guided Evaluation and Training

Next, we demonstrate the usefulness of the importance annotations in the training and evaluating of object detection models. The analysis provides insights into the collected dataset as well as into the limitations of currently employed evaluation metrics. In particular, we demonstrate how two object detection models which perform similarly when compared with traditional metrics (importance-agnostic) can significantly vary in their ability to detect objects of high importance. Hence, the main goal of this section is error type analysis. Furthermore, the analysis motivates training joint detection-attention models which are better at detecting objects of high-importance, and therefore may be more suitable for use in analyzing the intentions of surrounding agents.

A closer look at the dataset visualization figures (Figs. 1, 2, and 3) demonstrates how many of the KITTI objects are consistently ranked under the low importance category while high importance objects are rare. This fact raises concerns regarding non-biased evaluation of vision tasks.

Motivating importance metrics for object detection: Traditional evaluation on vision datasets (PASCAL [24], Caltech [25], KITTI [1], etc.) separate objects by size, occlusion, and truncation. Specifically for KITTI, three evaluation procedures of 'easy' (above 40 pixels in height, truncation under 15%), 'moderate' (above 25 pixels in height, with partial occlusion, truncation under 30%), and 'hard' (with heavy occlusion and truncation under 50%) are employed. In all the aforementioned datasets, challenging instances of size, occlusion, and truncation are often entirely excluded from training/evaluation, yet these choices are arbitrary in the context of driving where such instances may be potentially relevant to safety-critical events (e.g. a highly truncated vehicle which is overtaking the ego-vehicle). This motivates evaluation on importance classes for providing complementary analysis to the traditional metrics. The importance metrics can also reveal dataset bias, as the rarity of high importance instances may bias models both in training and evaluation. Specifically, in training time, the model may emphasize visual attributes found in the most common objects (e.g. vehicles of low relevance), and evaluation with traditional metrics may not reveal such bias.

Importance-guided training: As a final experiment, we employ the Faster-RCNN [17] framework for training a vehicle detector on KITTI. First, fine-tuning is performed on an auxiliary dataset of KITTI objects taken from videos not used in the importance prediction experiments. Next, we continue fine-tuning on the importance dataset (performed twice for each fold in the cross validation), but modify the loss function from [17] to weigh objects of higher importance more heavily. We refer to this training process as 'importance-guided' training. The process is performed both for the ZF [26] and VGG network architectures. In all test settings, an overlap of 0.7 is required between a predicted and a ground truth box for a true positive. Furthermore, images are up-sampled by a factor of 2.6 in training and testing, which we found necessary for detection of smaller objects.

Analysis with importance metrics: The results of the experiments for object detection are shown in Table II. Several conclusions can be drawn from the analysis. While importance metrics may be correlated with traditional, importance-agnostic metrics, the two types of metrics contain complementary information. For instance, comparing easy and high importance test settings, or hard and low importance test

settings, the overall AP numbers differ. The low importance category is shown to be particularly challenging. Furthermore, the importance test settings contain vehicles with higher truncation ratio than the hard test settings.

The deeper VGG model is shown to significantly improve over the ZF model in all test settings, but the main takeaway is demonstrated by inspecting the difference in AP (ΔAP) between importance-guided and importance-agnostic training and evaluation. As shown in Table II, importance test settings are essential for measuring improvement in detection performance over objects of higher importance, which is otherwise not clear. For instance, the ZF-IG model achieves an AP of 87.19% on high and moderately important objects, an improvement of 4.39 AP points over the importance-agnostic baseline. On the other hand, the performance improvement when employing traditional test settings is minor. A similar observation holds for the VGG-IG model. Importance-guided training specifically targets visual challenges and appearance patterns of objects of higher importance, thereby significantly improving the detection performance for such objects. This experiment demonstrates the feasibility of employing the collected importance annotations for evaluating vision tasks, as traditional and importance metrics are shown to be complementary.

VI. Concluding Remarks

This paper analyzed the task of video-based importance prediction using a variety of multi-modal spatio-temporal cues. To that end, a human-centric object-level importance annotations dataset was collected for KITTI videos. Two types of models were used to perform in-depth analysis into what types of cues make an object important. In addition to subtle object-level cues, such as object occlusion state and orientation, temporal dynamics of a proposed set of multi-modal situation attributes were shown to be crucial for object importance classification. Moreover, we demonstrated how the collected annotations can be useful when evaluating vision tasks, in particular for object detection performance evaluation. The experimental analysis uncovered the role of dataset bias and motivated training cost-sensitive object detection models which are better at detecting objects of higher relevance to the driving task. In the future, we would like to further study spatio-temporal visual modeling for the purpose of importance prediction [20], [22], analyze generalization and importance in other contextual settings, such as U.S. highways [27], and research the usefulness of the importance annotations for other vision tasks on KITTI. Another future direction is in developing subject-specific scene perception models, as motivated by the demonstrated relationship between subject personal attributes (e.g. driving experience) and importance perception.

VII. Acknowledgments

We thank the subjects who participated in the study, our CVRR colleagues for helpful discussions, and the reviewers for their constructive comments. We are grateful for the support of our associated industry partners.

References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *IJRR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] M. Sivak and B. Schoettle, "Road safety with self-driving vehicles: General limitations and road sharing with conventional vehicles," Tech. Rep. UMTRI-2015-2, University of Michigan Transportation Research Institute, 2015.

[3] E. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE Transactions on Intelligent Vehicles*, 2016.

[4] A. Borji, Dicky, N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *CVPR*, 2012.

[5] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image," in *ICPR*, 1998.

[6] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *CVPRW*, 2010.

[7] S. Muddamsetty, D. Sidibé, A. Trémeau, and F. Mériaudeau, "Spatio-temporal saliency detection in dynamic scenes using local binary patterns," in *ICPR*, 2014.

[8] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *HRI*, 2013.

[9] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from RGB-D images," in *ICCV*, 2015.

[10] Y. Iwashita, A. Takamine, R. Kurazume, and M. Ryoo, "First-person animal activity recognition from egocentric videos," in *ICPR*, 2014.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012.

[12] J. Kooij, N. Schneider, F. Flohr, and D. Gavrila, "Context-based pedestrian path prediction," in *ECCV*, 2014.

[13] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*, 2014.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[15] 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," *IJCV*, pp. 1–42, 2015.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[19] S. Gidaris and N. Komodakis, "Object detection via a multi-region semantic segmentation-aware cnn mode," in *ICCV*, 2015.

[20] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," *CVPR*, 2016.

[21] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems," vol. 134, pp. 130–140, 2015.

[22] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," *ICRA*, 2016.

[23] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," *CVPRW*, 2015.

[24] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *IJCV*, 2010.

[25] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," in *PAMI*, 2012.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[27] J. Dueholm, M. Kristoffersen, R. Satzoda, T. Moeslund, and M. Trivedi, "Trajectories and behaviors of surrounding vehicles using panoramic camera arrays," *IEEE Transactions on Intelligent Vehicles*, 2016.