Are All Objects Equal? Deep Spatio-Temporal Importance Prediction in Driving Videos

Eshed Ohn-Bar and Mohan Manubhai Trivedi
University of California, San Diego
La Jolla, CA 92093-0434
{eohnbar,mtrivedi}@ucsd.edu

Abstract

Understanding intent and relevance of surrounding agents from video is an essential task for many applications in robotics and computer vision. The modeling and evaluation of contextual, spatio-temporal situation awareness is particularly important in the domain of intelligent vehicles, where a robot is required to smoothly navigate in a complex environment while also interacting with humans. In this paper, we address these issues by studying the task of on-road object importance ranking from video. First, human-centric object importance annotations are employed in order to analyze the relevance of a variety of multi-modal cues for the importance prediction task. A deep convolutional neural network model is used for capturing video-based contextual spatial and temporal cues of scene type, driving task, and object properties related to intent. Second, the proposed importance annotations are used for producing novel analysis of error types in image-based object detectors. Specifically, we demonstrate how cost-sensitive training, informed by the object importance annotations, results in improved detection performance on objects of higher importance. This insight is essential for an application where navigation mistakes are safety-critical, and the quality of automation and human-robot interaction is key.

Keywords- *Spatio-temporal object analysis, vision-based behavior analysis, intelligent and automated vehicles, human-centric artificial intelligence, contextual robotics, driver perception modeling, object detection.*

1. Introduction

There is a great need for smarter and safer vehicles [1, 2]. Large resources in both industry and academia have been allocated for the development of vehicles with a higher level of autonomy and advancement of human-centric artificial intelligence (AI) for driver assistance. Understanding, modeling, and evaluation of situational awareness tasks, in particular the understanding of the behavior and intent of agents surrounding a vehicle, is an essential component in the development of such systems [3, 4, 5, 6]. Human drivers continuously depend on situation awareness when making decisions. In particular, the observation that attention given by human drivers to surrounding road occupants varies based on a task-related, scene-specific, and object-level cues motivates our study of human-centric object recognition.

A model of driver perception of the scene requires reasoning over spatio-temporal saliency, agent intent, potential risk, as well as past and possible future events. For instance, consider the on-road scene in Fig. 1. Obstacle avoidance requires robust recognition of all obstacles in the scene, yet surrounding obstacles are not all equal in terms of relevance to the driving task and attention required by a driver. Given the specific scene in Fig. 1, a subset of the road occupants (remote, occluded, or low-relevance objects) was consistently annotated at a lower importance level by human annotators when considering the driving task. On the other hand, a pedestrian intending to cross and a cyclist at the ego-lane were consistently annotated at higher importance levels for the driving task. The input to the modeling/annotation task is a video, and the output is a per-frame, object-level importance score. This level of contextual reasoning is essential for an intelligent robot required to navigate in the world, as well as communicate with and understand humans. This work is concerned with training recognition algorithms that can perform such complex reasoning. In order to better understand the aforementioned observations and issues, we propose to study a notion of on-road object importance, as measured in a spatio-temporal context of driving a vehicle. The contributions of our study are as follows.

1.1. Contributions

Modeling object importance: The main contribution of this work is in the study of which cues are useful for

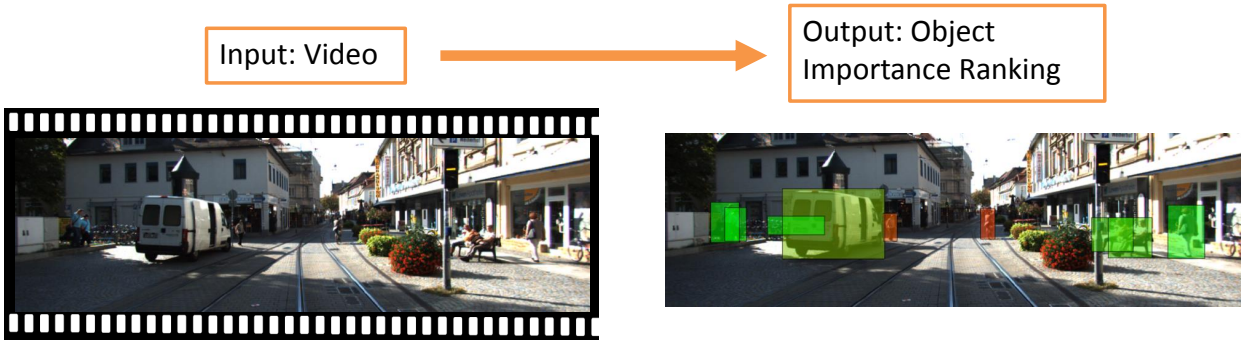


Figure 1: What makes an object salient in the spatio-temporal context of driving? Given a video, this work aims to rank agents in the surrounding scene by relevance to the driving task. Furthermore, the notion of importance defined in this work allows a novel evaluation of vision algorithms and their error types. The importance score (averaged over subjects' annotations) for each object are shown, colored from high to moderate to low.

on-road object importance ranking. Specifically, a set of spatio-temporal object attributes are proposed for capturing attention, agent intent, and scene context. The analysis is performed in the context of autonomous driving on KITTI videos [7], but may also be useful to other application domains in computer vision requiring spatio-temporal analysis and human perception modeling, including saliency modeling [8, 9], robotics [10], and ego-centric vision [11].

Importance-guided performance metrics: The collected dataset is used to produce new evaluation insights for vision tasks. In particular, the annotations are used to highlight dataset bias in object detection for autonomous driving. As highly important objects are rare, we experimentally demonstrate existing training and testing procedures to be biased towards certain object characteristics, thereby hindering insights from comparative analysis. Furthermore, the object importance annotations are used to train cost-sensitive, attention-aware object detection models. The proposed importance-guided training procedure is shown to result in models which produce less errors when objects of higher importance are concerned - a useful insight for the safety-critical application considered in this study.

2. Motivation and Related Research Studies

Importance analysis: Importance ranking essentially involves modeling context. Capturing spatial image context has been heavily studied [12, 13]. Berg *et al*. [14] measure object-level importance in an image by the likelihood of the object to be mentioned by a person describing it. Temporal context implies movement modeling [15], understanding of what an agent can do, intends to do, or how multiple agents may interact [16]. Lee *et al*. [17] studies object importance regression in long-term ego-centric videos using gaze, hand-object interaction, and occurrence frequency cues, but no human importance annotations are employed. Mathialagan *et al*. [18] performs single image importance prediction of people with linear regression over pose, occlusion, and distance features. On the other hand, we pursue spatio-temporal importance ranking as it relates to a perceived driving environment by a driver. The task of on-road object importance modeling may also be somewhat correlated with general visual saliency [8, 19], but the latter is often not studied for a driving task.

Human-centric evaluation: It is known that driver experience level (usually measured in years) significantly impacts safe driving partly due to improved identification and prediction of other road occupants' intentions [2]. As computer vision datasets become more realistic and complex, one way to evaluate such prior knowledge and complex modeling of spatio-temporal events (involving object recognition, scene context modeling, etc.) is using the proposed set of importance metrics (similar metrics have been devised for other machine learning and vision tasks, such as object segmentation and image captioning [20, 21]). Human-centric metrics provide a rich tool for understanding the human in the loop, from modeling human drivers in general to a specific driver perception and style, and is of great use to development effective driver assistance and human-computer cooperation. Conveying intents by autonomous driving vehicles to other road occupants is also an important task relevant to our study, as it may require understanding of how humans perceive a scene.

Importance metrics for on-road object detection: We employ the importance annotations in order to perform a finer-grained evaluation of object detection. At a high level, two object detectors may potentially have similar detection performance while differing in ability to detect important objects. A dataset bias could further hinder such an insight. Algorithms for visual recognition of objects has seen tremendous progress in recent years, most notably on the

Figure 2: This study is motivated by the fact that not all objects are equally relevant to the driving task. As shown in example frames from the dataset with overlaid object-level importance score (averaged over subjects), drivers' attention to road occupants varies based on task-related, scene-specific, and object-level cues.

ILSVRC [22, 23, 24], PASCAL [25, 26], Caltech [27], and KITTI datasets [28], yet low cost, camera-based object detection with low false positives over many hours of video in a wide variety of possible environmental conditions is still not solved. Therefore, better understanding and evaluation of the limitations of state-of-the-art object detection algorithms is essential. We believe current metrics employed for generic object detection are limited for the study of on-road object detection as detailed below. We emphasize that this study is not concerned with ethical issues in autonomous driving, but instead with deeper understanding of requirements and limitations for safe navigation and human-centric AI on an object detection and classification level.

Are all objects equal? It may not be surprising that the **answer is no**, even in existing evaluation protocols for object detection. Some objects posing certain visual challenges are notoriously more difficult to detect than others. Objects of small size, heavy occlusion, or large truncation are partially or entirely excluded from existing evaluation (and training) on PASCAL, Caltech, and KITTI. Yet in the

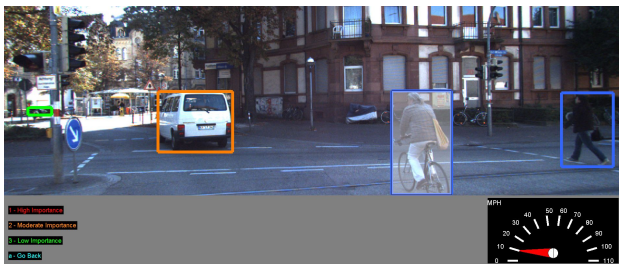
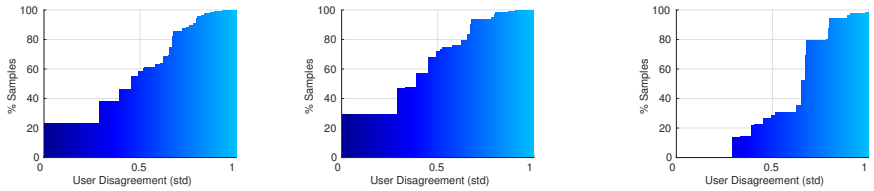


Figure 3: The interface used to obtain object-level importance ranking annotations. The cyclist is highlighted as it is the currently queried object to annotate, colored boxes have already been annotated with an importance level by the annotator, and blue boxes are to be annotated.

context of driving, such instances may be the most relevant under safety-critical events! Existing evaluation metrics are often inconsistent regarding these visual challenges, and reflect a certain bias [29, 30, 31, 32] where importance is measured differently from in the driving domain. We experimentally demonstrate the impact of such bias in evaluation on KITTI (Section 5). Furthermore, importance-based metrics normalize evaluation curves differently than ones based on object appearance properties (properties which may be distributed differently across datasets), and so it has the potential of offering complementary insights. For instance, consider scenarios of dense scenes with tens of road occupants that are heavily occluded or are across a barrier (e.g. highway settings). As annotation of such scenes is challenging and evaluation of objects across a barrier may not be necessary for development and evaluation of algorithmic recognition performance, the importance-centric framework only consider a handful of agents which are of higher importance. As large numbers of objects in KITTI (Fig. 2) were generally annotated at low relevance to the driving task, the proposed annotations could be used to provide deeper understanding of existing object detectors in a domain where errors are costly and the type of errors made should well understood. It will be shown in Section 5 that training detectors without a notion of importance can have a biasing effect on the output of the detector itself. Our approach is also biologically plausible, as human drivers do not generally pay attention to all objects in the scene (Fig. 2), but are skillful at recognition and analysis of only a subset of relevant objects. On the other hand, vision algorithms are evaluated on a large portion of low importance vehicle samples, which may skew analysis and insights.

3. Dataset and Annotations

The KITTI dataset [28, 7] was chosen due to richness of object-level annotation and sensor data. As video data is essential for the notion of importance, we utilize a subset of



(a) Entire Dataset (b) Low Importance Objects (c) High Importance Objects

Figure 4: A cumulative histogram obtained by varying the disagreement requirement (standard deviation among subject labels), until 100% of the data is included. While disagreement exists, a subset of highly important and highly non-important objects shows consistency (see Sec. 3 for discussion).

the raw data recordings with the provided 3D annotations of pedestrians, cyclists, and vehicles. The annotations include bird's eye view orientation and tracklet IDs. The dataset contains synchronized GPS, LIDAR, and vehicle dynamics, useful for studying the dynamics of a variety of cues as they relate to perceived object importance.

Importance annotations: Experiments were done in a driving simulator with KITTI videos shown on a large screen using the interface in Fig. 3. Subjects watched each short video twice, and every 10^{th} frame was annotated by querying for an integer between 1-3 (1 being high and 3 being low importance). 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. Three levels were chosen for simplifying the annotation process - two levels of importance (yes or no) is too restrictive as there is no way of handling ambiguous cases. On the other hand, a continuous ranking score may have been used, but such a task may lead to a large confusion among subjects and for guessing, which we aimed to reduce.

Although subjective in nature, the task of importance ranking is performed by all drivers every day. Out of a total of 18 subject, high correlation between subject driving experience, age, and annotation output was demonstrated. Interestingly, the annotation task resulted in a clear relationship between annotation output and subject driving experience (measured in years). For the interested reader, subject analysis can be found in the Appendix 9. Consistency analysis (Fig. 4) of the annotators output demonstrates that many instances in the low importance class have high agreement among the subjects. On the other hand, the moderate and high importance classes contain higher variation.

The overall dataset used in the experiments contains 17,635 object annotations, including 15,057 vehicles (cars, vans, and trucks), 1,452 pedestrians, and 562 cyclists. In the existing metrics on KITTI for object detection, test samples are categorized into three levels of difficulty based on object properties of height, occlusion, and truncation. 'Easy' test settings include non-occluded samples with height above 40 pixels and truncation under 15%, 'moderate' settings in-

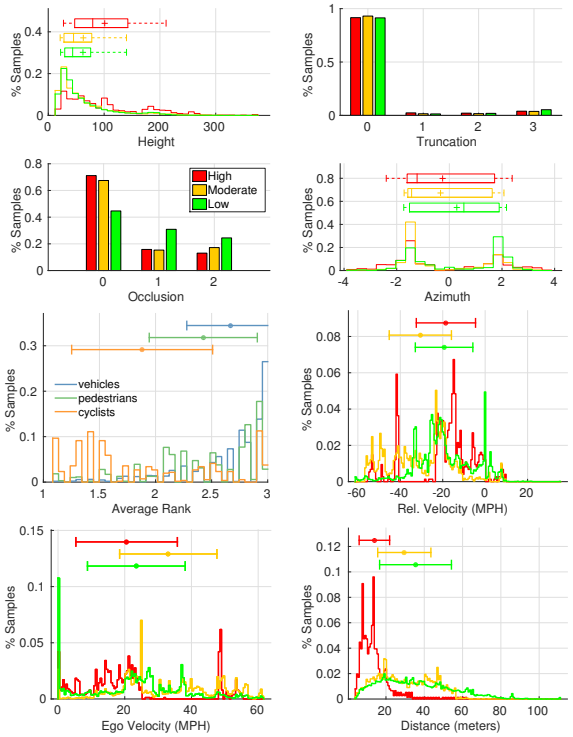
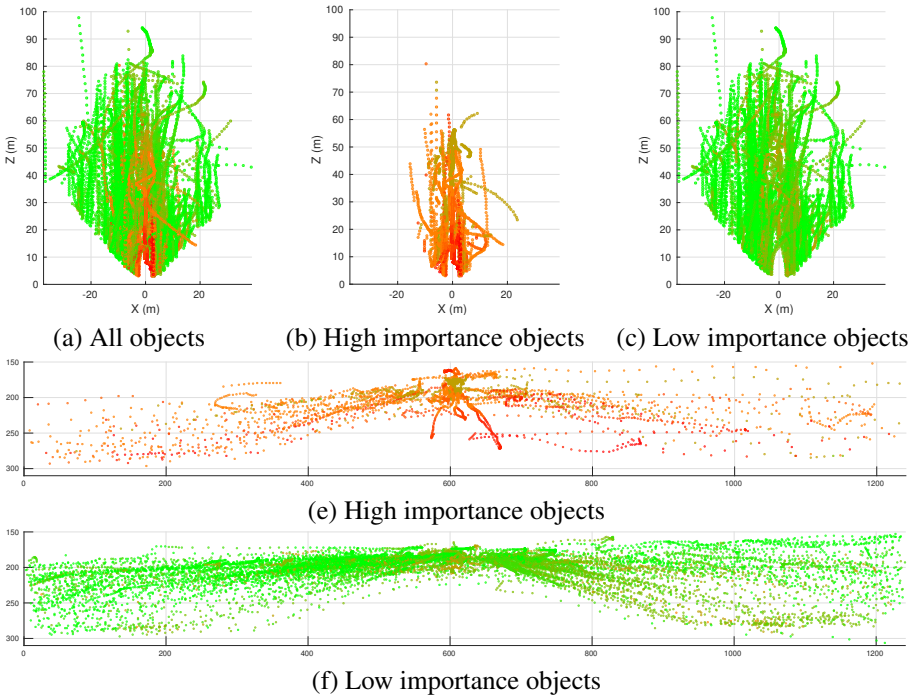


Figure 5: Object statistics corresponding to three classes of object importance in the dataset.

clude partially-occluded samples with height above 25 pixels and truncation under 30%, and 'hard' settings include heavy occlusion samples with height above 25 pixels and truncation under 50%. In the same spirit, we introduce three importance classes by taking the median vote among subjects for each object instance, from high, moderate, to low importance. Out of the totals, there were high/moderate/low importance 293/2159/12,605 vehicles, 143/524/785 pedestrians, and 267/147/148 cyclists. Subjects reported a variety of reasons for importance annotations, from the existence of a barrier in traffic, head orientation cues for pedestrians (also studied in [33, 34, 35]), and spatio-temporal relationships between different objects. The annotations and code



(a) All objects (b) High importance objects (c) Low importance objects

(e) High importance objects

(f) Low importance objects

Figure 6: Distribution of object positions in top-down view (a)-(c) and image plane (d)-(e). Each instance is colored according to average importance ranking, from high to moderate to low importance.

will be made publicly available. In addition to the three importance class, regression of the average importance score will also be studied.

Dataset properties: The dataset statistics are depicted in Fig. 5. When analyzing highly important objects, these are shown to be non-occluded samples within 40 meters or less of the ego-vehicle. Most vehicles are categorized as moderate or low importance, which is to be expected as KITTI contains many parked and stationary vehicles. Truncation percentage statistics binned to a histogram are approximately evenly distributed. Fig. 6(a-c) demonstrates that objects in the proximity of the vehicle may have any level of importance annotation, suggesting other cues besides position alone are necessary for the importance ranking task. In the image plane, Fig. 6(e) demonstrates the distribution of the position in the image plane for high importance objects.

4. Object Importance Model

In this section, we formulate the object importance models which will provide insights into what causes some objects to be perceived as more important than others. To that end, we propose two types of models, differing by the type of features employed for scoring an object instance importance level. All model weights are learned using a logistic regression model.

4.1. Object attributes model, $\mathbf{M}_{attributes}$

KITTI provides several high quality object-level attributes extracted from ground truth information and multi-modal sensor data. The attributes allow for an explicit analysis of the relationship between different object properties and importance ranking. For an instance s and class importance c, we train the following prediction model,

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

where the features used in the $M_{attributes}$ model are defined below.

2D object features: For each sample, the 3D object box annotation is projected to the image plane for obtaining a set of 2D object properties. 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: As shown in Fig. 6, distance from the ego-vehicle is correlated with annotated importance levels. Other 3D object properties, such as orientation, may provide hints as to what an on-road occupant is doing or intends to do. The $\phi_{3D-obj} \in \mathbb{R}^6$ features are composed of the left-right (lateral) and forward-backward (longitudinal) range coordinates (x, z) given by the LIDAR, Eu-

clidean distance from the ego-vehicle, orientation in bird's eye view, and object velocity components, $|V|$ and $\angle V$.

Ego-vehicle features: Ego-vehicle parameters can be used in order to capture contextual settings relevant to the importance ranking task. For instance, if the ego-vehicle is traveling at low speeds, the surrounding radius in which objects may be considered relevant decreases. For that reason, ego-vehicle speed information is displayed during the annotation process as shown in Fig. 3. Hence, the attribute model includes ego-vehicle velocity magnitude and orientation features, $\phi_{ego} = [\text{ego}|V|,\ \text{ego}\angle V]$.

Temporal attributes: The total aforementioned 2D object, 3D object, and ego-vehicle features can be used to represent an object and certain contextual information in a given frame. Nonetheless, the temporal evolution of such properties may also provide useful information in representing past, present, and potential future actions, and consequently impact importance ranking. This assumption is captured in $\phi_{temporal}$, which is computed using the aforementioned object and ego-vehicle attributes but over a past time window. Specifically, $\phi_{temporal}$ is obtained by concatenating the attributes over the time window. In addition, we add the values after a max-pooling operation over the time window, as well as the Discrete Cosine Transform (DCT) coefficients [15].

We note that $M_{attributes}$, while utilizing the extensive KITTI multi-modal data and annotations, is not intended to be exhaustive. Additional attributes can potentially be considered, such as object-object relationships attributes, object-lane relationship attributes, as well as scene-type attributes (although these are not currently provided with KITTI and will need to be extracted/annotated). The objective of $M_{attributes}$ is in gaining explicit insight into the role of object attributes which are known to contain little noise on importance ranking. Furthermore, $M_{attributes}$ is of use when comparing to a visual, video-only importance prediction model, which will be presented next. For instance, limitations in the visual prediction model will be analyzed using $M_{attributes}$. On the other hand, the visual model can implicitly encode attributes missing from $M_{attributes}$, such as spatial relationships among objects, scene types, and more.

4.2. Visual prediction model, $\mathbf{M}_{visual}$

Our main task is the visual prediction of object importance. Given a 2D bounding box annotation, M_{visual} learns a mapping from an image region to an importance class using

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

where the feature components of the visual prediction model are defined next.

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) [36] convolutional network. The network was pre-trained on the ImageNet dataset [23] and fine-tuned on KITTI using Caffe [37].

Spatial context features: In order to capture spatial context, such as relationship with other objects in the scene, lane information, scene type information, or better capture object properties (e.g. occlusion, truncation, orientation), each object instance is padded by a factor of $\times 1.75$ for generating $\phi_{spatial} \in \mathbb{R}^{4096}$.

Temporal context features: Similarly to in $M_{attributes}$, we hypothesize the human annotators reason over spatio-temporal cues in the videos shown to them when determining object relevance to a driving task. In order to test the hypothesis and provide insights into the importance ranking task, the per-frame visual descriptors, ϕ_{obj} and $\phi_{spatial}$, are employed for computing a $\phi_{temporal}$ component. Specifically, given an object tracklet with 2D box positions and a temporal window, the previous object and spatial context features are computed over a time window, concatenated, and max-pooled.

5. Importance Metrics for Object Detection

As described in Section 2, there are potential issues with applying traditional object detection metrics to on-road object detection analysis. In addition to the importance ranking task described in Sections 4.1 and 4.2, we provide further insights into the proposed importance dataset by studying the importance annotations in the context of object detection. Specifically, we study the usefulness of importance-based metrics in evaluating object detectors. For instance, as the majority of vehicles in KITTI were consistently ranked with lower importance to the immediate driving task, the rarity of objects of higher importance may result in a bias both in training and evaluation. First, training may rather emphasize visual attributes found in the most common objects. Second, evaluation using traditional metrics may not reveal such a bias. In order to demonstrate this phenomenon and motivated by work on specializing convolutional networks (ConvNets) [38], we train object detectors which are specialized at detecting objects of higher importance.

The experiments employ the Faster R-CNN framework [39] with two training procedures, one importance-agnostic and one importance-guided. Following Fast R-CNN [40], the framework trains a network with two sibling output layers. The first output layer predicts a discrete probability distribution per each image region, $p = (p_0, \ldots, p_K)$ over $K+1$ object categories, using a softmax over the $K+1$ outputs of a fully connected layer. The second layer out-

Model	mAP (%)	MAE	$\text{MAE}_{\gamma=2.25}$
$M_{visual}(\phi_{obj})$	51.06	0.2648	0.5392
$M_{visual}(\phi_{obj} + \phi_{spatial})$	55.53	0.2611	0.5007
$M_{visual}(\phi_{obj} + \phi_{temporal})$	53.30	0.2507	0.4765
$M_{visual}(\phi_{obj} + \phi_{spatial} + \phi_{temporal})$	56.34	0.2447	0.4625
$M_{attributes}$ (without $\phi_{temporal}$)	53.70	0.2440	0.3853
$M_{attributes}$ (with $\phi_{temporal}$)	**60.35**	**0.2148**	**0.2914**

Table 1: Summary of the classification experiments using the two proposed importance prediction models.

puts bounding-box regression offsets for the 4 coordinates of the image region. For each training region labeled with a ground-truth class u and a ground-truth bounding-box regression target v, we use the following multi-task loss

$$L(p, u, \gamma, t^u, v) = L_{cls}^{IG}(p, u, \gamma) + \lambda_{loc}[u \geq 1]L_{loc}(t^u, v) \quad (3)$$

such that $L_{cls}^{IG}(p, u, \gamma) = -\alpha_\gamma \log p_u$ is the log loss for true class u. The weight factor α_γ is added, defined as

$$\alpha_\gamma = \begin{cases} \lambda & \gamma \leq 2.25 \\ 1/\lambda & \text{otherwise} \end{cases} \quad (4)$$

to allow cost-sensitive importance-guided training, where γ is the average importance score of the current sample. The cost-sensitive training allows steering the objective function optimization by increasing mis-classification penalty on objects with higher importance. The second task loss, L_{loc}, is the sum of the smooth L1 loss function over the 4 box coordinates as defined in [40]. L_{loc} is computed for samples of non-background class ($[u \geq 1]$) only. In the experiments, we set $\lambda_{loc} = 1$ and $\lambda = 10$. We note that setting $\alpha = 1$ for all γ results in the commonly used, importance-agnostic training procedure.

6. Experimental Evaluation

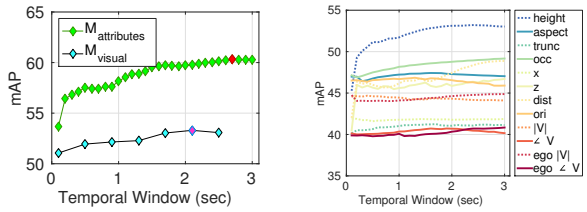
6.1. Importance Prediction Models

A total of 8 videos is employed in the experiments, with a 2-fold validation split. Results using the two importance models are shown in Table 1. In each experiment, classification is done for each importance class, a Precision-Recall (PR) curve is calculated, and the area under the curve (AP) is averaged (mAP) over the classes for an overall performance summary, so that higher mAP value implies better classification performance. For a second evaluation metric, we regress the average importance score for each object instance and compute the mean absolute error (MAE). Due to the large imbalance in the distribution of the importance scores, we show overall MAE on all samples as well as MAEγ which is computed over a subset of samples with an average importance score less than or equal to γ. Setting $\gamma = 2.25$ allows for computing the MAE only on objects of higher importance, excluding objects considered of lower importance (with average importance score of more than 2.25).

Evaluation of $M_{attributes}$: Table 1 shows the performance of the attributes-based model. We note that for the experiments in Table 1, training and evaluation is done in an object class agnostic manner, only considering the importance class/score of samples. We note that due to the high-level features used in $M_{attributes}$, it should be considered as a strong baseline, achieving mAP of 53.70% and 60.35% without and with temporal features extraction, respectively. Temporal features are shown to be essential for both importance classification and regression of objects of higher importance. As shown in Fig. 7, a past time window of up to 2.7 seconds is shown to contain beneficial information for importance classification with $M_{attributes}$, while performance saturates for M_{visual} with a $\sim$2 seconds window.

Next, we further analyze the impact of different components in $M_{attributes}$ in order to better understand what makes an on-road object important. Fig. 8 depicts the relationship between individual attributes and importance prediction. Performance using the combination of all of the object attributes is shown as 'comb', which provides the best importance ranking results. Analysis is shown for each object category separately, as well as for training a single importance prediction model over all object types in an object class agnostic manner. Highest mAP for importance classification of vehicles is achieved using the object attributes of occlusion, aspect ratio, orientation, and height in the image plane. Because occlusion by another object often implies lower relevance to the driving task, occlusion state is shown to be a particularly useful cue. Similarly, orientation and aspect ratio may capture traffic flow direction and planned future actions. Ego-vehicle velocity magnitude is also shown to have high relationship with importance ranking, serving as a frame-level contextual cue. For the pedestrian object class, high impact attributes are also the distance and position in 3D. The cyclist object class follows similar trends, yet reliable conclusions are more difficult to draw as it contains a small number of samples.

Fig. 7(b) isolates the benefit that each individual attribute provides as the time window for the feature computation in-



(a) (b)

Figure 7: Cue analysis with the importance models. (a) Classification accuracy when varying the time window used for computing $\phi_{temporal}$ in both models. (b) Classification accuracy with each of the attributes in $M_{attributes}$ with an increasing temporal window used for a temporal feature extraction.

creases. Results are shown when considering an object class agnostic model. Fig. 7(b) highlights the importance of temporal feature extraction for several high-level semantic cues, including past occlusion and truncation, distance change from the ego-vehicle, lateral movement, and object size in the image plane. Certain attributes, such as ego-vehicle parameters, are shown to benefit from a larger past temporal window. This is to be expected, as ego-vehicle information serves as a general frame-level contextual cue. Fig. 10 shows the PR curves used to compute the final performance summary in Table 1. Fig. 10 demonstrates the significant impact of temporal attribute cues in classifying importance class for different object types, improving classification performance in almost every case. The smaller, cyclist object class contains large annotation inconsistencies, in particular within the moderate importance class, leading to poor performance for all of the importance prediction models. A larger dataset could resolve such issues. Furthermore, additional insights may be gained by subject-specific modeling and evaluation, which is left for future work.

Evaluation of M_{visual}: Table 1 shows the performance summary of different components in the visual importance prediction model. Contrasting with $M_{attributes}$, simply using the object region features ϕ_{obj} results in a reduction of 2.64% mAP points to 51.06% mAP. This is expected, as $M_{attributes}$ employs clean annotation and other sensor data. The MAE in prediction average importance score also suffers, in particular on objects of higher importance. Addition of the spatial context component, $\phi_{spatial}$, results in a large performance improvement of 4.47% mAP points, as well as a noticeable reduction in MAEγ. The analysis demonstrates the importance of contextual information in modeling object importance. We've also experimented with schemes of feature extraction from the entire image for capturing scene information, but no additional benefit was shown.

As with $M_{attributes}$, incorporation of a temporal feature extraction component, $\phi_{temporal}$, to M_{visual} results in a

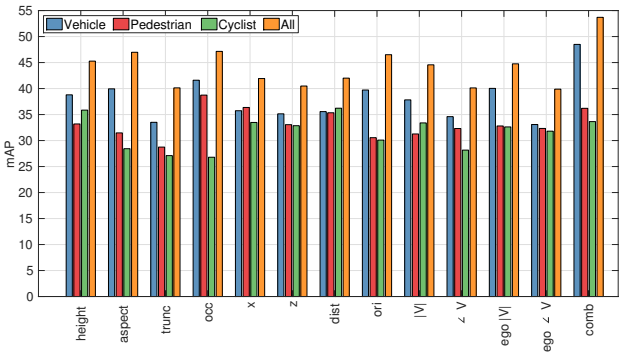


Figure 8: Object importance classification results using each attribute in $M_{attributes}$ separately, as well as with a combination of all attributes ('comb'). Results are shown for training and evaluation on each object class separately, as well as in an object class agnostic manner ('All'). No temporal feature extraction is used in these experiments.

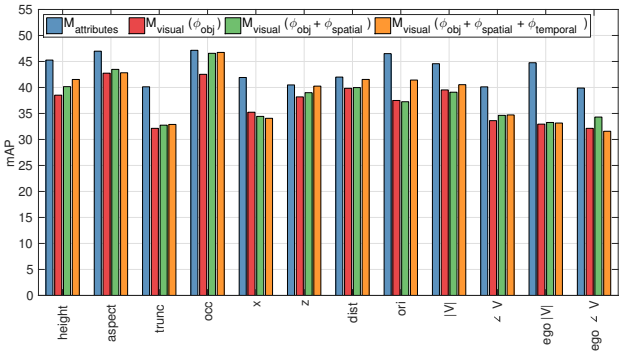


Figure 9: Regressing each attribute using various feature combinations in M_{visual} and consequently using the attribute for importance class classification allows for explicit analysis of the limitations of M_{visual}.

further performance improvement, although to a lesser extent (56.34% mAP). As shown in Fig. 7, the improvement plateaus beyond a ~2 seconds past window. When comparing performance among the two models, both in classification and regression, the M_{visual} model is significantly outperformed by $M_{attributes}$ (in particular on objects of higher importance). The results in Table 1 motivate further study of models suitable for capturing spatio-temporal visual cues [41, 42, 43, 44], which can be a future study.

Limitation analysis of M_{visual}: Comparing the visual-only ranking against the strong baseline $M_{attributes}$ of object attributes reveals insights as to the current limitations in representing object properties with the VGG network. This motivates an explicit limitation study, as shown in Fig. 9. In this experiment, the VGG network is used to regress each object attribute in $M_{attributes}$, and consequently the

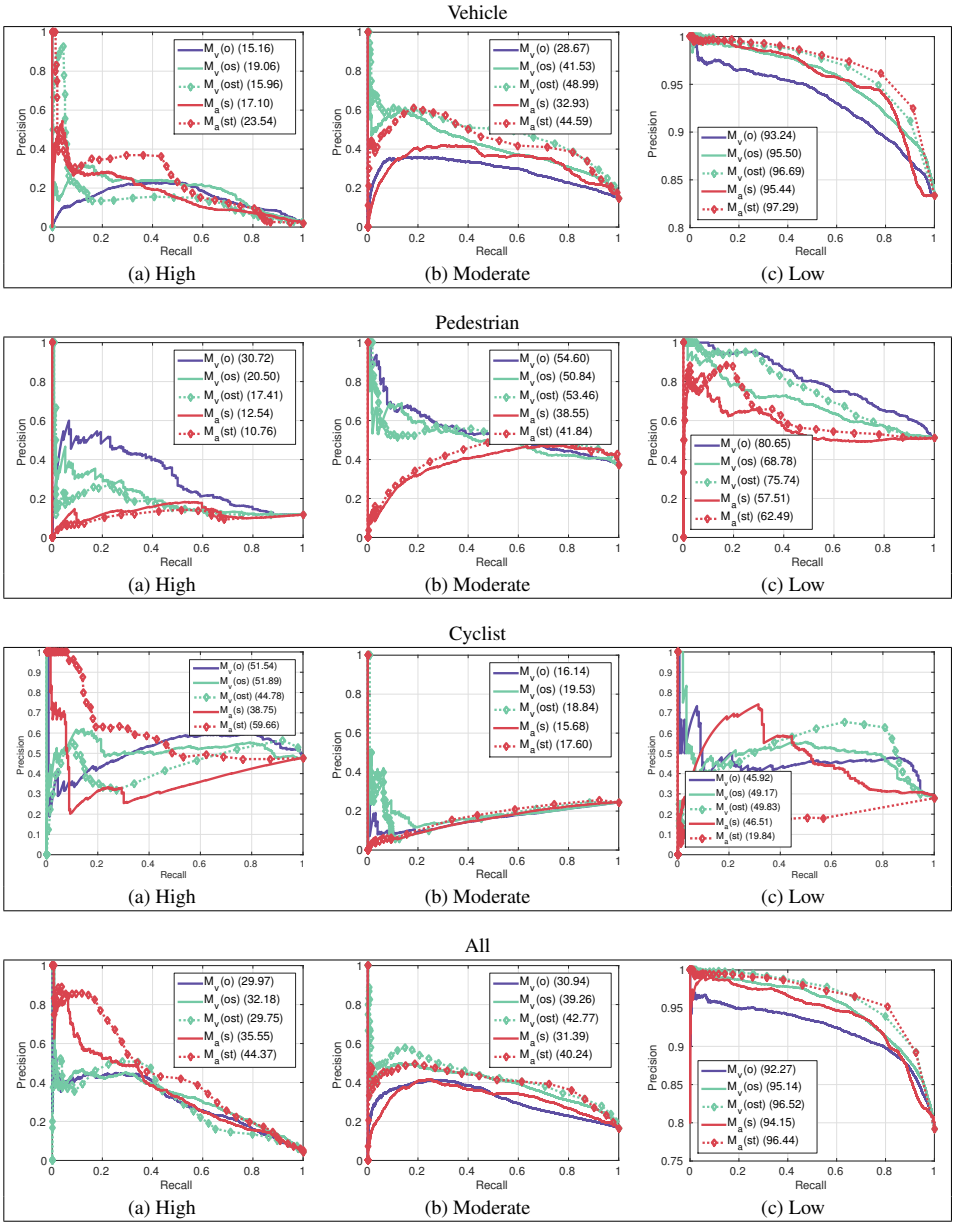


Figure 10: For each object class (rows) and object importance level (columns), we show performance precision-recall curves when employing different models and cue types. For the attributes model (M_a), performance without and with temporal features is shown as 's' and 'st', respectively. Similarly, for the visual model (M_v) performance with ϕ_{obj}, $\phi_{obj} + \phi_{spatial}$, and $\phi_{obj} + \phi_{spatial} + \phi_{temporal}$ is shown as 'o', 'os', and 'ost', respectively. In parenthesis is the area under the curve.

regressed value is used for importance ranking instead of the original value from $M_{attributes}$. The experiment is repeated for different feature combinations in M_{visual}, providing insight into the benefit that different features provide and assist in explaining the current limitations in M_{visual}. Fig. 9 demonstrates that while some object attributes as they relate to object importance are predicted well (such as occlusion state), others (such as orientation, object velocity, or truncation) are lacking. The incorporation of the spatial and temporal context features significantly improves the ability to capture object state, in particular object occlusion state, range, and orientation. On the other hand, explicit regression of object velocity, ego-vehicle parameters, or truncation value is challenging.

6.2. Importance-Guided Object Detection

In the detection experiments, we follow the KITTI evaluation protocol of correct detection at 0.7 overlap for vehicles, and 0.5 for pedestrians and cyclists. All models are first fine-tuned for object detection on KITTI using the publicly available detection benchmark, but excluding frames from videos used in the importance experiments. Next, for each fold in the 2-fold cross validation, we fine-tune faster R-CNN (FRCN) [39] in an importance-agnostic manner and importance-guided manner, as described in Section 5. Results are shown in Table 2 for both the ZF [45] and the VGG [36] network architectures. Table 2 depicts the complementary relationship between the proposed set of importance metrics and traditional test settings (defined in Section 3). For instance, AP values differ among the easy/hard test settings when comparing to high/low importance test settings. In particular, as the low importance class isolates many instances with challenging settings of larger occlusion and smaller height, it exhibits the lowest performance across all metrics. Another observation is the impact of importance-guided training, in particular when performance is measured with importance-based metrics. For instance, importance-guided training with ZF results in a significant 6.11% AP improvement in detection of objects of the high importance class, while such an improvement is not visible in traditional metrics based on object height, occlusion, or truncation. This is due to a dataset bias, as most vehicles in the dataset are of lower importance ranking. A similar observation holds for results using VGG, but to a lesser extent as the larger and deeper VGG model is better at general object detection.

When analyzing results on KITTI, we observed a large number of false positives occurring for both the ZF and VGG models on objects of small height. In addition to the challenge in detecting small objects, we also observed inaccurate annotations in KITTI on small objects. Furthermore, the importance-guided training may be simply emphasizing large objects which are generally of higher importance. Therefore, Table 2 shows results on objects of 25 pixels and up (as proposed by KITTI), as well as on objects of 40 pixels and up. The latter corresponds to varying only occlusion/truncation in the 'moderate' and 'hard' traditional test settings. Comparing the two test settings on objects of 40 pixels and up, we can see that while importance-guided training indeed emphasizes correct detection on larger objects, the importance-based metrics are still able to capture complementary insights to the importance-agnostic metrics. For the pedestrian object class, there is a stronger correlation between the two types of metrics due to a higher proportion of high and moderate importance classes samples. Nonetheless, the general trends of improved performance due to importance-guided training still hold. Due to the small number of cyclists, only the vehicles and pedestrian categories are analyzed. The results demonstrate the feasibility of the proposed metrics both for the training and testing of vision tasks, in particular object detection. We note that as mentioned in [46], training task-specific ConvNets (e.g. for occlusion) does not necessarily result in improvement (and may even reduce overall detection performance). As shown in Table 2, this is not the case with importance classes.

7. Concluding Remarks

This paper studies object recognition under a notion of importance, as measured in a spatio-temporal context of driving a vehicle. Given a driving video, our main research aim was to model which of the surrounding vehicles are most important to the immediate driving task. Employing human-centric annotations allowed for gaining insights as to how drivers perceive different on-road objects. Although perception of surrounding agents is influenced by previous experience and driving style, we demonstrated a consistent human-centric framework for importance ranking. Extensive experiments showed a wide range of spatio-temporal cues to be essential when modeling object-level importance. Furthermore, the importance annotations proved useful when evaluating vision algorithms designed for on-road applications and autonomous driving. Future work includes studying the relationship between gaze dynamics, saliency, and object importance ranking. Furthermore, the dataset can be used in order to study subject-specific modeling which is relevant to cooperative driving and control transitions [42, 47, 48, 1]. Further investigation of the cost-sensitive training procedure [49, 50, 43] may lead to additional insights in the future. Appropriate temporal metrics, such as how quickly an object was classified as important in the video, can also be useful for comparing methods in importance prediction. Cross-dataset generalization and annotations on additional datasets [51, 52, 53] can provide further understanding into models and evaluations for importance prediction. Ideally, annotation of additional datasets can be done more efficiently by employing lessons learned from this work. Evaluation of the sensitivity of the importance models on different times of day, night, weather condition, and diverse traffic scenes are also important next steps. We hope that this study will motivate further developments in spatio-temporal object detection and importance modeling, essential for real-world video applications.

8. Acknowledgments

The authors would like to thank the reviewers and editors for their helpful comments, the subjects who participated in the study for their valuable time and great attitude, our colleagues at the Laboratory for Intelligent and Safe Automobiles for their encouragement, Rakesh Rajaram for helpful

	Traditional Test Settings			Importance Test Settings		
Method	Easy	Mod.	Hard	High	High+Mod.	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

(a) Vehicle, height 25 pixels and up

	Traditional Test Settings			Importance Test Settings		
Method	Easy	Mod.	Hard	High	High+Mod.	Low
FRCN-ZF	89.26	85.69	72.68	71.27	84.46	65.11
FRCN-ZF-IG	91.09	86.74	73.75	76.01	87.59	65.88
ΔAP	+1.83	+1.05	+1.07	+4.74	+3.13	+0.77
FRCN-VGG	95.63	92.74	80.90	85.56	92.29	74.53
FRCN-VGG-IG	94.54	91.70	79.56	86.73	91.44	73.40
ΔAP	-1.09	-1.04	-1.34	+1.17	-0.85	-1.13

(b) Vehicle, height 40 pixels and up

	Traditional Test Settings			Importance Test Settings		
Method	Easy	Mod.	Hard	High	High+Mod.	Low
FRCN-ZF	50.30	45.66	42.91	21.88	30.45	35.03
FRCN-ZF-IG	62.43	57.07	51.97	34.29	47.15	37.67
ΔAP	+12.13	+11.41	+9.06	+12.41	+16.70	+2.64
FRCN-VGG	66.71	61.23	57.96	22.48	44.91	48.67
FRCN-VGG-IG	70.67	64.81	59.47	33.01	53.76	43.53
ΔAP	+3.96	+3.58	+1.51	+10.53	+8.85	-5.14

(c) Pedestrian, height 25 pixels and up

	Traditional Test Settings			Importance Test Settings		
Method	Easy	Mod.	Hard	High	High+Mod.	Low
FRCN-ZF	50.30	47.59	44.75	22.57	32.13	36.45
FRCN-ZF-IG	62.43	58.12	52.98	34.61	48.13	38.39
ΔAP	+12.13	+10.53	+8.23	+12.04	+16.00	+1.94
FRCN-VGG	66.71	63.09	59.74	16.81	47.18	49.91
FRCN-VGG-IG	70.67	67.23	61.80	27.19	56.61	45.46
ΔAP	+3.96	+4.14	+2.06	+10.38	+9.43	-4.45

(d) Pedestrian, height 40 pixels and up

Table 2: Evaluation of object detection (AP) using the proposed set of importance metrics and the Faster-RCNN framework (FRCN) [39]. '**IG**' refers to importance-guided fine-tuning, where correct classification of samples with higher importance annotations is weighted heavier in the training loss.

discussions, and NVIDIA for a hardware donation used for this research. We also thank our sponsors and associated industry partners.

References

[1] 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. 1, 7

[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. 1, 2

[3] A. Doshi and M. M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," in *IEEE Conf. Intell.*

Transp. Syst., 2011. 1

[4] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *IEEE Intl. Conf. Robotics and Automation*, 2016. 1

[5] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer, "Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking," in *IEEE Intelligent Vehicles Symposium*, 2014. 1

[6] E. Ohn-Bar and M. M. Trivedi, "What makes an on-road object important?," in *Intl. Conf. Pattern Recognition*, 2016. 1

[7] 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, 3

[8] A. Borji, Dicky, N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *CVPR*, 2012. 2

[9] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *CVPRW*, 2010. 2

[10] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *HRI*, 2013. 2

[11] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from RGB-D images," in *ICCV*, 2015. 2

[12] T. Li, T. Mei, I. S. Kweon, and X. S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 381–392, 2011. 2

[13] Y. Wang, T. Mei, S. Gong, and X.-S. Hua, "Combining global, regional and contextual features for automatic image annotation," *Pattern Recognition*, vol. 42, no. 2, pp. 259–266, 2009. 2

[14] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi, "Understanding and predicting importance in images," in *CVPR*, 2012. 2

[15] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*, 2014. 2, 4

[16] W. Chen, C. Xiong, R. Xu, and J. J. Corso, "Actionness ranking with lattice conditional ordinal random fields," in *CVPR*, 2014. 2

[17] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *IJCV*, vol. 114, no. 1, pp. 38–55, 2015. 2

[18] C. S. Mathialagan, A. C. Gallagher, and D. Batra, "Vip: Finding important people in images," in *CVPR*, 2015. 2

[19] N. Pugeault and R. Bowden, "Learning pre-attentive driving behaviour from holistic visual features," in *ECCV*, 2010. 2

[20] D. M. Y. Zhu, Y. Tian and P. Dollár, "Semantic amodal segmentation," *CoRR*, vol. abs/1509.01329, 2015. 2

[21] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *CVPR*, 2015. 2

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012. 2

[23] 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. 2, 4

[24] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *PAMI*, vol. 35, no. 8, pp. 1915–1929, 2013. 2

[25] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *IJCV*, 2010. 2

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014. 2

[27] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *PAMI*, 2014. 2

[28] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012. 2, 3

[29] X. Chen and A. Gupta, "Webly supervised learning of convolutional networks," in *ICCV*, 2015. 3

[30] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011. 3

[31] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "An exploration of why and when pedestrian detection fails," in *ITSC*, 2015. 3

[32] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "Looking at pedestrians at different scales: A multiresolution approach and evaluations," *TITS*, 2016. 3

[33] F. Flohr, M. Dumitru-Guzu, J. Kooij, and D. Gavrila, "A probabilistic framework for joint pedestrian head and body orientation estimation," *TITS*, vol. 16, no. 4, pp. 1872–1882, 2015. 3

[34] J. Kooij, N. Schneider, F. Flohr, and D. Gavrila, "Context-based pedestrian path prediction," in *ECCV*, 2014. 3

[35] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intelligent Transportation Systems*, vol. 8, pp. 413–, 2007. 3

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 4, 7

[37] 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. 4

[38] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI*, 2015. 5

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015. 5, 7, 8

[40] R. Girshick, "Fast r-cnn," in *Intl. Conf. on Computer Vision*, 2015. 5

[41] 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. 6

[42] 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," *Computer Vision and Image Understanding*, vol. 134, pp. 130–140, 2015. 6, 7

[43] 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. 6, 7

[44] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," *CVPRW*, 2015. 6

[45] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014. 7

[46] B. Pepik, R. Benenson, T. Ritschel, and B. Schiele, "What is holding back convnets for detection?," in *GCPR*, 2015. 7

[47] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *ICCV*, 2015. 7

[48] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *CVPRW*, 2010. 7

[49] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004. 7

[50] O. Beijbom, M. Saberian, D. Kriegman, and N. Vasconcelos, "Guess-averse loss functions for cost-sensitive multiclass boosting," in *ICML*, 2014. 7

[51] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *PAMI*, vol. 31, no. 12, pp. 2179–2195, 2009. 7

[52] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classication with partial occlusion handling," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010. 7

[53] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016. 7

9. Appendix

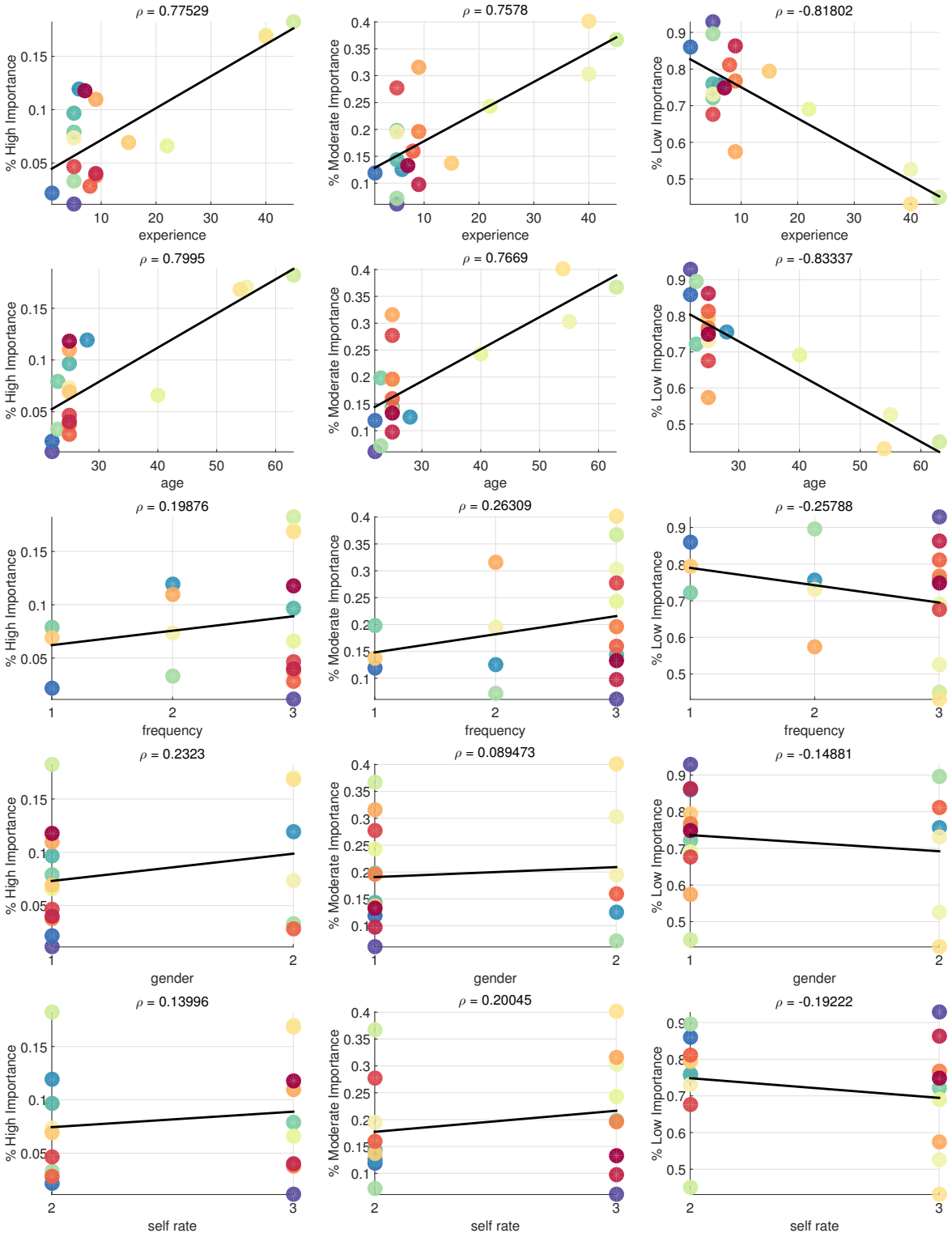


Figure 11: Relationship between importance level (grouped by columns) and subject personal information (grouped by rows). Each subject has been assigned a unique color, and is represented in each figure by a dot. From top row: (1) driving experience in years, (2) age in years, (3) frequency of driving, either 1-rarely, less than once a month, 2-occasionally, about once a week, 3-frequently, more than three times a week, (4) gender 1-male, 2-female, (5) rating of driving skill, 2-intermediate, 3-advanced. We observed a strong relationship between experience in years and importance ranking annotations.