AR-CP: Uncertainty-Aware Perception in Adverse Conditions with Conformal Prediction and Augmented Reality For Assisted Driving

Achref Doula, Max Mühlhäuser, and Alejandro Sanchez Guinea
Technical University of Darmstadt, Germany
doula@tk.tu-darmstadt.de, Max@informatik.tu-darmstadt.de, sanchez@tk.tu-darmstadt.de

Abstract

Deep learning models are pivotal in enhancing driver assistance systems and improving environmental perception. However, the tendency of neural networks towards overconfident predictions poses a risk of inaccurate predictions, potentially compromising driver safety in adverse conditions. To mitigate this issue, we introduce AR-CP, an uncertainty-aware framework designed to augment driver perception in scenarios characterized by adverse weather and insufficient lighting, through the integration of conformal prediction and augmented reality (AR). Our framework initiates with a conformal prediction step that produces an uncertainty-aware prediction set including potential object classes at a predefined probability level. Subsequently, AR is used to provide a simplified and informative visualization of the closest common parent class of the classes in the prediction set, thereby reducing the likelihood of misinformation. We provide a principled formulation and theoretical analysis of our framework. We evaluate AR-CP on the ROAD dataset, a large dataset containing different difficult situations that induce high uncertainty during prediction time. The results show that our framework outperforms state-of-the-art approaches in providing smaller prediction sets while holding the theoretical guarantees, ensuring an uncertainty-aware prediction, and reducing user confusion. We conduct an immersive user study with 15 participants to investigate the effects of our concept on the quality of perception, situation awareness, and mental load of participants. The results show that our concept facilitates a safer driving experience while holding the mental load low and the situation awareness high.

1. Introduction

Deep learning models have markedly advanced driver assistance systems by enhancing environmental perception and scene understanding. Driver assistance systems, relying on object detection and classification, are crucial for improving safety, reducing accidents, and elevating the driving experience [8, 9]. However, despite the substantial progress, deep learning models tend to yield overconfident predictions [23, 52], which poses safety risks, particularly in driving contexts under adverse weather conditions or in poor lighting, where the reliability of sensory data is already compromised. Prior efforts have primarily aimed at boosting detection and classification precision, often neglecting the management of prediction uncertainty. This oversight results in systems with a high risk of confident but incorrect detections. Addressing this gap requires strategies that not only enhance accuracy but also effectively handle the inherent uncertainties in deep learning predictions for more reliable driver assistance.

Conformal prediction (CP) [50] provides a straightforward mathematical framework to measure deep learning model uncertainty. Unlike traditional neural networks that predict a single class, CP generates a set of possible labels, ensuring the true label is included with a probability of $1 - \alpha$. Here, α represents a user-specified tolerance for error. The prediction set output of CP is particularly beneficial in scenarios where decision-making must consider the reliability of the information provided by the model, e.g., in human-AI teams [2], which is crucial in assisted driving.

In this paper, we employ CP to improve driver assistance under conditions of high uncertainty. We introduce AR-CP, a novel methodology that merges CP with augmented reality (AR) to aid drivers in accurately perceiving their surroundings during challenging scenarios like bad weather or poor lighting. AR-CP involves two key steps. Initially, we utilize a conformalized detection model to pinpoint objects of interest in the scene, such as pedestrians, vehicles, and cyclists, for which the model exhibits high uncertainty. The conformalized model generates a prediction set that ensures the inclusion of the true object class, with a user-specified probability, offering an alternative to potentially inaccurate single predictions. Subsequently, we develop a taxonomy that categorizes classes from the most general to the most specific. The taxonomy helps identify the nearest common parent class within the conformal prediction set for uncer-

tain detections, which is then overlaid on the detected object using AR. The generated representation benefits the driver in three ways: it raises awareness of uncertain detections, enhancing situation awareness; it simplifies the visualization with a clear, easy-to-see overlay; and it provides a general idea of the detected object without the false precision of uncertain detections.

We provide a principled formulation and theoretical analysis of our framework. We evaluate AR-CP on the ROAD dataset [44], a large dataset containing different difficult situations that induce high uncertainty during prediction time. The results show that our framework outperforms state-of-the-art approaches in providing smaller prediction sets while holding the theoretical guarantees, ensuring an uncertainty-aware prediction, and reducing user confusion. We conduct an immersive user study with 15 participants to investigate the effects of our concept on the quality of perception, situation awareness, and mental load of participants. The results show that our concept facilitates a safer driving experience while holding the mental load low and the situation awareness high.

2. Related Work

2.1. Uncertainty Visualization

Understanding and visualizing uncertainty plays a pivotal role in decision-making, a concept extensively explored within decision theory [47]. Uncertainty, typically modeled as probabilistic distributions, affects outcomes in scenarios involving human interaction with predictive models. Effective visualization of uncertainty can improve user trust [25, 58], decision quality [25, 30], and equalize decision-making capabilities across different expertise levels [27]. Uncertainty visualization techniques fall into three categories [39]: (1) graphical annotations like error bars and distributions [10, 15, 26, 32, 40, 59], (2) visual encoding using color, position, or transparency [34], and (3) hybrid methods that merge these strategies [16, 38].

In autonomous driving, effective uncertainty communication through visual cues—ranging from facial expressions for awareness, scale/bar representations for vehicle autonomy level, to peripheral lights for workload reduction—has been shown to enhance driver interaction and trust [3, 25, 29, 46].

2.2. Conformal Prediction

Conformal prediction is a robust framework for estimating uncertainty across a range of tasks, including classification [18, 57], regression [12, 43], segmentation [49], and information retrieval [17], regardless of the underlying model. It excels in providing prediction sets for classification tasks, ensuring these sets contain the true label within a predetermined error threshold. CP's applicability extends to various critical fields requiring accurate uncertainty measures, such as healthcare diagnostics [33, 57], autonomous robot navigation [14, 31], and time series analysis [45, 54, 56]. The framework's reliability has prompted research into improving its adaptability [11, 42] and optimizing the size of prediction sets [1].

2.3. Vehicle-Driver Interaction

Interaction modalities between cars and drivers encompass a variety of methods through which drivers connect with their vehicles and receive information. These include visual, auditory, and haptic/gesture interactions [6, 7, 13, 19, 21, 28, 41, 53]. Of these, vision-based interactions stand out for their congruence with human perceptual processes and their ability to convey detailed information efficiently, ensuring the vehicle's surroundings are communicated effectively without overwhelming the driver [8, 9]. The customizability of vision-based interfaces accommodates diverse driving scenarios and preferences while emphasizing safety by reducing distractions and enabling quick access to information [35]. While integrating multiple modalities can be beneficial [5], vision-based communication is paramount for seamless driver-vehicle interaction.

In our paper, we opt for vision-based interaction, particularly through an augmented reality Windshield Display System (WSD). This approach aligns with our goal of merging real-time road conditions with our visualization concepts, striking a balance between enhancing the driver's view and minimizing distractions. We consider WSDs a promising avenue for future visualization technologies in vehicles.

2.4. Virtual Reality For Autonomous Driving Experiments

The evaluation of user interfaces in autonomous vehicles, especially those that incorporate elements of uncertainty, necessitates a balance between creating a realistic testing environment and ensuring the safety of the participants [36]. To meet these requirements, researchers have been innovating with various levels of simulation fidelity. This includes simple desktop configurations with gaming wheels and screens [37] to more complex arrangements that employ panoramic visuals and actual vehicles [22, 55]. In our work, we leverage the advantages of VR environments to carry out experiments involving autonomous driving that would otherwise be considered too hazardous. Our experimental framework features a fully immersive VR setting, enabling interaction via a Head-Mounted Display (HMD) and a steering wheel.

3. Background

In this section, we delve into Conformal Prediction with a focus on its application to the task of classification, which is directly relevant to our paper.

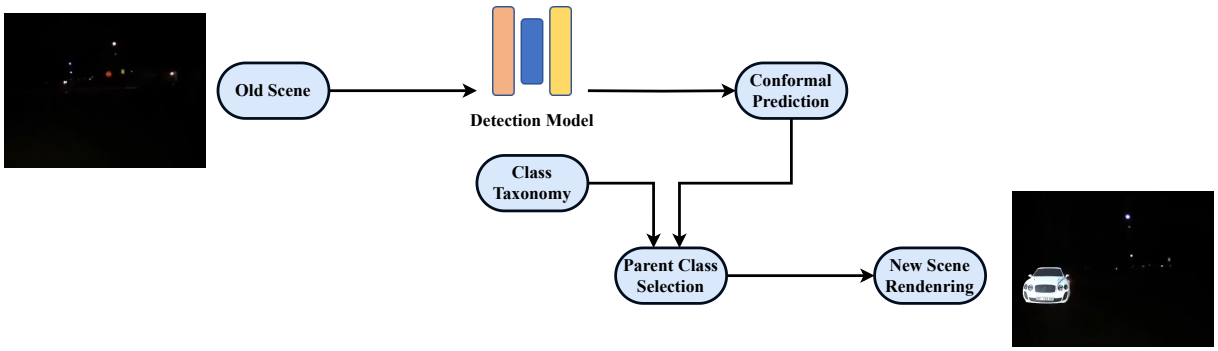


Figure 1. Architecture of the AR-CP framework. AR-CP receives as input a scene with potential prediction errors. A detection model performs conformal prediction and generates uncertainty-aware prediction sets that are guaranteed to include the true object label with a user-defined probability $1-\alpha$. The nearest common parent concept of all the classes in the prediction set is determined using the class taxonomy. Finally, a visual representation of the nearest common parent concept is rendered in the windshield display.

Consider a classifier, denoted as g_θ, that has been previously trained on a dataset D_{train}. The model, g_θ, provides probability estimates for each class, yielding outputs such as $g_\theta(X) \in [0,1]^K$, where X represents an input image and K the total number of class categories. Utilizing a calibration set $D_{cal} = (X_i, Y_i), i \in [1, n]$, which consists of *exchangeable* unseen data pairs drawn from the same distribution as D_{train}, CP aims to generate prediction sets $C(X_{test}) \subset 1, .., K$ for new data samples. These sets are designed to be valid, meaning they include the true label Y_{test} with probability $1-\alpha$, where α is the user-predefined error rate. This concept is known as *marginal coverage* and is formulated as:

$$\mathbb{P}[Y_{test} \in C(X_{test})] \geq 1 - \alpha \tag{1}$$

The construction of prediction sets in CP is facilitated by a *non-conformity score* $S(X, Y)$, which quantifies the discrepancy between the predictions for a new, unseen data point and those made on the training dataset. The score allows for the ranking of D_{cal} elements, from which the empirical $1-\alpha$ quantile $\hat{q}$ is derived. For any new test instance X_{test}, with an unknown Y_{test} at inference time, the prediction set $C(X_{test})$ is defined as $C(X_{test}) = Y : S(X, Y) \leq \hat{q}$. The calibration process guides the construction of prediction sets and can be formalized in Theorem 3.1:

Theorem 3.1 (Conformal Prediction [50])**.** *Assuming D_{train}, D_{cal}, and X_{test} are sets of exchangeable random variables from the same distribution, and given a non-conformity score S and an error rate α, the prediction set $C(X_{test})$ defined as:*

$$C(X_{test}) = \{Y \in \mathcal{Y} : S(X_{test}, Y)) \leq \hat{q}\} \tag{2}$$

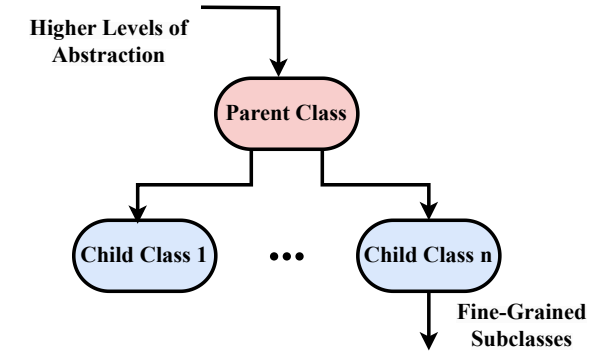
where $\hat{q}$ represents the $1-\alpha$ quantile of S over D_{cal}, adheres to the marginal coverage condition outlined in Equation 1.

4. AR-CP: Conformal Scene Augmentation

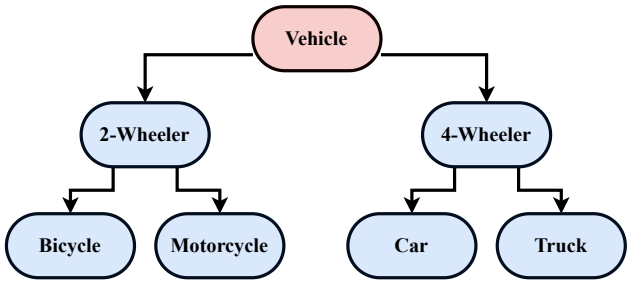
4.1. Approach Overview

Accurate perception of the environment under adverse conditions—such as heavy rain, fog, or dim lighting—presents a significant challenge for drivers. Several approaches [8, 9] propose to assist the driver in perceiving its environment through using deep learning models applied to object detection and rendering the results of the detection in the windshield display. However, the simple utilization of deep learning models for object detection is not guaranteed to provide trustworthy predictions, especially in the case where the situation is out of distribution, or represents situations or objects that stem from distributions different from the data used to train the model.

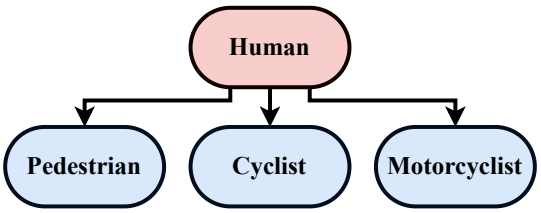
We propose to alleviate this perceptive confusion while keeping the theoretical coverage guarantees of conformal prediction by rendering the nearest common parent of the classes in the prediction set, following the process depicted in Figure 1. To do this, we build a class taxonomy $\mathcal{T}$ that structures the semantic knowledge about the possible agents that might appear in an urban environment. The leaf nodes in $\mathcal{T}$ represent the fine-grained classes on which the model is trained. As shown in Figure 2, classes with similar semantic properties are grouped and represented by a concept with higher abstraction. For example, the classes *car* and *truck* can be grouped in a higher abstraction class called *4-wheeled vehicle*. Consequently, an object having the class "*car*" is also an object of the class "*Vehicle*", since every car is a vehicle. Combining this property with conformal prediction provides a strong theoretical basis to output a single prediction, that is uncertainty-aware, and importantly with lower bound guarantees on the class probability, ensuring safety, ease of use, and validity. For a particular object X_i in the scene, the conformolized deep learning predictor g_{CP} generates a prediction set $C(X)$ containing a set of



(a) Class taxonomy used to model the semantic knowledge in AR-CP.



(b) Example of the sub-taxonomy used for the class *vehicle*.



(c) Example of the sub-taxonomy used for the class *human*.

Figure 2. Taxonomy structure used for the nearest common parent class selection in AR-CP.

leaf nodes classes that contains the true object class with a probability $1-\alpha$. We use the class taxonomy $\mathcal{T}$ to look for the nearest common parent node of the objects $P(C)$ and overlay its representation on the object of interest on the windshield display of the car.

4.2. Theoretical Guarantees of AR-CP

In addition to the simplicity of our approach, it can be proven that replacing the prediction set with the nearest common parent of all the classes in the prediction set holds the coverage guarantees provided by conformal prediction. That is, the rendered object represents the true label of the object with a probability of at least $1-\alpha$. In the following, we demonstrate that AR-CP holds the theoretical coverage guarantees of CP while reducing the set size to at most 1. We formalize this property in Theorem 4.1 and provide a proof.

Theorem 4.1 (coverage guarantees of AR-CP)**.** *Let $\mathcal{T}$ be a class taxonomy and $\mathcal{N}$ a semantic function returning the nearest common parent concept of a set of classes. Let $C(X)$ be the prediction set obtained by a conformal prediction procedure for a pre-trained model g_{CP} as described in Theorem 3.1. Let $\mathcal{Y}_{parent}$ be the nearest common parent class for the classes in the prediction set $C(X)$. Then we have:*

$$P[Y_{test} \text{ is } \mathcal{Y}_{parent}] \geq 1-\alpha \tag{3}$$

Proof. The set $C(X)$ is constructed based on Theorem 3.1 and we have:

$$P[Y_{test} \in C(X_{test})] \geq 1-\alpha$$

and we have $\mathcal{Y}_{parent} = \mathcal{N}(C(X_{test}))$. This means that $\forall Y \in C(X_{test}), Y$ is $\mathcal{Y}_{parent}$. Since $Y_{test} \in C(X_{test})$ based on Theorem 3.1, then Y_{test} is $\mathcal{Y}_{parent}$. Then, we have

$$P[Y_{test} \in C(X_{test})] \geq 1-\alpha \Longrightarrow$$
$$P[Y_{test} \text{ is } \mathcal{Y}_{parent}] \geq 1-\alpha$$

□

We describe the AR-CP in Algorithm 1. In our approach, we focus solely on objects with high uncertainty, defined as those whose prediction set size $|C(X_i)|$ exceeds a certain threshold t. This limitation allows AR-CP to specifically highlight objects that might be critical yet difficult to see, minimizing unnecessary visual distractions for the driver.

Algorithm 1: AR-CP

Input : $\mathcal{T}$, which is the class taxonomy.
g which is the model.
$\{X_i\}$, which is the set of detected objects.
D_{cal}, which is the calibration set.
α, which is the user-defined error rate.
t, which is the threshold on the set size.
Output: visualization of higher concept

Function AR−CP (X_{test}, D_{cal}, $\mathcal{T}$,α) **:**
  $g_{CP} = Calibrate(g, D_{cal})$
  for X_i *in* $\{X_i\}$ **do**
    $C(X_i) \leftarrow$CP(g_{CP}, X_i)
    if $|C(X_i)| > t$ **then**
      $\mathcal{Y}_{i,parent} \leftarrow$N(C($X_i$))
      Render_Windshield ($\mathcal{Y}_{i,parent}$)
  end for

5. Evaluation

We evaluate different aspects of AR-CP. First, we quantitatively evaluate the efficiency of our conformal prediction framework in reducing the set size while holding the theoretical coverage guarantees. For this, we adopt two widely

Score	Method	$\alpha = 0.1$		$\alpha = 0.2$		$\alpha = 0.3$		$\alpha = 0.4$		$\alpha = 0.5$	
		Set Size	Coverage	Set Size	Coverage	Set Size	Coverage	Set Size	Coverage	Set Size	Coverage
Softmax	Standard	6.03	0.89	5.01	0.80	3.94	0.69	3.06	0.59	2.22	0.50
	AR-CP	**1.00**	0.89	**1.00**	0.80	**1.00**	0.69	**1.00**	0.59	**1.00**	0.50
APS	Standard	6.16	0.90	5.13	0.80	4.07	0.71	3.17	0.60	2.35	0.51
	AR-CP	**1.00**	0.90	**1.00**	0.80	**1.00**	0.71	**1.00**	0.60	**1.00**	0.51
RAPS	Standard	1.22	0.97	1.11	0.96	1.06	0.95	1.04	0.95	1.01	0.94
	AR-CP	**1.00**	0.97	**1.00**	0.96	**1.00**	0.95	**1.00**	0.95	**1.00**	0.94

Table 1. Results of the experiments on the ROAD dataset [44] for $\alpha = [0.1, 0.2, 0.4, 0.5]$. The prediction sets for the task agent classification are generated using AR-CP with 3 different scoring functions: $1 - Softmax$, APS [42], and RAPS [1]. Bold designates better performance for set size.

used metrics: coverage rate and set size. Second, we conduct a user study with 15 participants to investigate the effects of the visualization generated by AR-CP on several cognitive aspects of the user, since our concept is mainly designed to be applied to increase the safety of the assisted driving experience.

5.1. Conformal Prediction

5.1.1 Dataset and Model

We evaluate AR-CP on the Road Event Awareness Dataset (ROAD) [44] since it provides diverse scenes of urban environments with highly dynamic agents taken under different weather and lighting conditions (e.g., sunny, night, overcast, snow), which facilitate the evaluation of our framework under a variety of scene difficulties. The ROAD dataset is composed of 22 videos and includes 560K bounding boxes and 1.7M instances of individual labels, which makes it suitable to evaluate our procedure thoroughly. Furthermore, the ROAD dataset provides a rich range of agents (11 classes), locations (15 classes), and actions (23 classes).

We use 3D-RetinaNet [44] as the underlying model for our processing. 3D-RetinaNet is proposed as the baseline model for the ROAD dataset. This model is composed of a backbone and a pyramidal network to generate classes and bounding boxes for the agents and outputs triplets containing the agent class, the location class, and the agent class. Since we require a separate output for the agent classification task, we modify the architecture such that each classification task is performed by a separate head, providing softmax scores.

5.1.2 Metrics

We use 2 metrics to evaluate our approach.

Coverage. Coverage designates the rate at which the predicted sets include the ground truth value. A CP-based approach is valid if the coverage is approximately greater or equal to $1 - \alpha$.

Set Size. We measure the average set size provided by our CP. As providing the full set of labels would be a trivial output for CP-based methods to achieve the coverage guarantees, a smaller set size demonstrates better predictive efficiency.

5.1.3 Baselines

To demonstrate the validity of our approach, we use AR-CP with one baseline scoring function that is equal to $1 - Softmax$, and 2 state-of-the-art CP scores. The first conformal score is adaptive predictive sets (APS) [42], which is a scoring technique known to improve conditional coverage. The second approach is regularized adaptive predictive sets (RAPS) [1], which is known to generate relatively smaller predictive sets.

5.1.4 Comparison with State-of-the-Art

We conduct our evaluation with varying degrees of confidence levels, operationalized through the parameter α ranging from 0.1 to 0.5 with a step increment of 0.1. The results, as detailed in Table 1, underscore the efficacy of our proposed AR-CP framework in reducing the prediction set size across different non-conformity scores, namely softmax, APS, and RAPS, which are critical in ensuring robust prediction performance in uncertain scenarios.

For a granular analysis, the set size for softmax, APS, and RAPS is reduced by up to 83.41%, 83.76%, and 18.03%, respectively. Such results are indicative of AR-CP's capability to significantly compact the prediction set size while maintaining the integrity of predictions. Notably, the impact of AR-CP on APS is the most significant, elucidating APS's inherent design to cater to class-wise coverage, which is especially beneficial in handling the imbalances present within the dataset. This design propensity of APS typically leads to larger set sizes, which our framework effectively counteracts. The differential effect of AR-CP on the non-conformity scores suggests a nuanced interaction between the nature of the score and the AR-CP set reduction capabilities. Particularly, the substantial set size reduction achieved with APS underscores AR-CP's adaptability in reducing prediction sets in the presence of class

(a) The "*None*" study condition. (b) The "*BB-CP*" study condition. (c) The "*AR-CP*" study condition.

Figure 3. Conditions used in the user study: the *None* condition with no visual assistance, the *BB-CP* condition with a bounding box and the full prediction set of CP, and *AR-CP* rendering the nearest common parent concept in the prediction set.

imbalances, a common challenge in machine learning models applied to dynamic and complex environments like driving assistance systems. Moreover, the consistent holding of coverage guarantees across all α values and non-conformity scores reinforces the reliability of AR-CP. This reliability is crucial, given that the assurance of coverage guarantees underpins the framework's ability to provide uncertainty-aware predictions that do not compromise on safety or accuracy.

5.2. User Study

We conduct a user study to investigate the effects of our uncertainty-aware scene augmentation on different cognitive aspects of the user. We create an immersive setup to simulate a vehicle with a heads-up display. We perform our evaluation on 4 different situations that represent challenging conditions for the vehicle and the driver. The study scenarios are chosen from 3 datasets: the ROAD dataset [44], the Gated2Gated dataset [51], and the Seeing Through Fog dataset [4].

5.2.1 Participants

For the study, we recruited 15 participants (8 male and 7 female) aged between 18 and 34 ($M = 25.4, SD = 3.137$). All participants held a driving license. Half of the participants were university students (50%), 37.5% worked in technical and scientific fields like engineering and biology, and 12.5% of the participants had non-technical occupations. The participants had uneven knowledge about autonomous driving, namely, 35% had no knowledge about autonomous cars, 35% had limited knowledge and 35% were very familiar with the topic. When asked about experience with machine learning methods, 12.5% stated to be experts. 75% had medium knowledge about machine learning, and 12.5% never heard of it in the past. All participants voluntarily took part in our study and no compensation was paid.

5.3. Study Design

We design the study to be a within-subject study with the driver visualizations as the unique independent variable. The visualizations we consider in the study are depicted in Figure 3, and described as follows:

No Visualization (None). In this condition, participants are shown the environment as seen by the driver, with no augmentation or visual cues. This condition represents the situation of drivers who are using traditional vehicles with no visual driver assistance.

Bounding Box with full CP set (BB-CP). In this condition, participants are presented with the bounding box with the full prediction set generated by a conformal prediction process.

AR-CP. This visualization method renders a model of the detected object onto the object, giving the user immediate visual information without having to read anything.

5.4. Implementation and Setup

For the user study, we generate 4 scenarios, as shown in Figure 4, including challenging and safety-critical situations. To ensure that the study environment resembles real-life situations, we create the scenarios using video sequences from 3 datasets: the ROAD dataset [44], the Gated2Gated dataset [51], and the Seeing Through Fog dataset [4]. Every scenario duration is 3 minutes. We use 3D-RetinaNet [44] with the same modifications described in Section 5.1.1. We use a 4-level class taxonomy, where the highest concept is "*thing*", and consider only the concepts "*vehicle*" and "*human*", as detailed in Figure 2. This design choice is motivated by the fact that the classes "*vehicle*" and "*human*" are predominant in autonomous driving benchmarks and datasets. For each scene, we choose agents for which the

(a) Snow and frost (b) Heavy rain (c) Heavy snow dust. (d) Night/extreme low light

Figure 4. Driving situations used during the study. The situations showcase challenging driving conditions inducing high uncertainty.

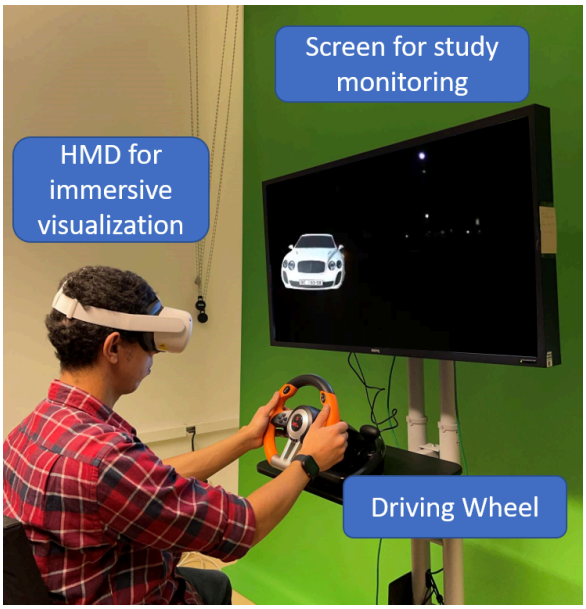


Figure 5. Setup used to conduct the user study. Participants are immersed in the environment using a head-mounted display (HMD) and interact with the scene using a driving wheel. An external screen is used to visualize the participant view in real-time.

model has the highest uncertainty, i.e., the agents with the largest set size using a vanilla conformal prediction procedure. To increase the level of immersion of the participants during the study, we opt for a study setup that resembles a real car. The study setup is composed of a gaming driving wheel and a head-mounted display (HMD) that is worn by participants to facilitate an intuitive interaction with the situation. The setup of the user study is depicted in Figure 5. We use Unreal Engine 4[1] to create the scenarios and combine them with our visualizations.

5.5. Metrics

To evaluate the utility of AR-CP, we use the following metrics:

Identification Score. We report the rate of correct classifications reported by participants when asked to count the number of instances of a particular class, during the scenario. This metric assesses the quality of perception of the users given the study conditions.

Mental Load. We measure the mental load of the participants using the mental workload subscale of the NASA TLX questionnaire [24].

Situation Awareness. To evaluate the impact of AR-CP on the situation awareness of the participants, we use the situation awareness technique questionnaire (SART) [48].

5.6. Procedure

The task of the participants is to count the number of instances of a particular class that appeared during the scenarios. The classes that we considered are cars, cyclists, pedestrians, and trucks. The study starts with obtaining formal consent from the participants and a short demographic questionnaire. Participants also indicate whether they have a driving license and how is their knowledge about assisted driving. After a short introduction to the task, participants visualized the situations using the HMD. After each situation, participants indicate the number of class instances that they were tasked to count. Afterward, participants answered a questionnaire about mental load, situation awareness, trust in the machine, perceived model understandability, and situation awareness. This process is repeated until all situations and visualization modalities are covered by the participants. To counterbalance the learning effects that may arise from repeating the same situation but with different visualization modalities, we use a balanced Latin square ordering of the situations throughout the study. The study ends with a semi-structured interview conducted with participants, in order to acquire more insights about possible aspects of the visualizations that are not covered by the questionnaires. The study took around 35 minutes per participant.

5.7. Results

We analyze the results of the study with a repeated measure one-way ANOVA, with the visualization modality as the unique independent variable. The results of the user study are depicted in Figure 6.

The results of the study show that AR-CP allowed the participants to have the highest identification score

[1] Unreal Engine 4: https://www.unrealengine.com/en-US

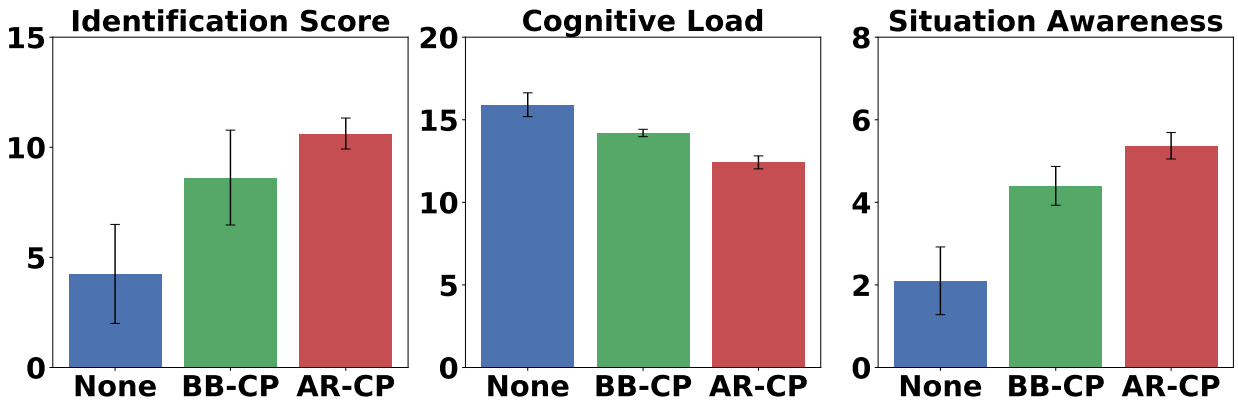


Figure 6. Study results.

($M = 10.26, SD = 0.70$) compared to the BB-CP ($M = 8.62, SD = 2.15$) and the None ($M = 4.25, SD = 2.24$) conditions. This result shows that AR-CP facilitates a more accurate identification of road elements in challenging situations with high model uncertainty. The capability of communicating a non erroneous prediction despite high uncertainty, is a crucial feature for the safety of users of highly automated vehicles and a factor to increase trust, a necessary factor for large adoption. The results of the TLX questionnaire show that AR-CP induced the lowest mental load ($M = 12.42, SD = 0.39$) compared to the BB-CP condition ($M = 14.20, SD = 0.39$) and the None condition ($M = 15.91, SD = 0.72$), with a significant statistical difference ($p \leq 0.05$). This is an expected result since the None condition does not provide any visual assistance to the driver which requires a higher concentration, and the BB-CP condition presents the full list of semantic classes that are present in the prediction set. Although the full list provides the driver with better visibility of the possible classes that represent the object, it induces a higher mental load making it harder to decide in time-critical situations. AR-CP represents a more suitable representation making a tradeoff between visibility and perception of uncertainty, and concise and easy-to-understand visual representations. AR-CP induced a higher situation awareness ($M = 5.37, SD = 0.32$) compared BB-CP ($M = 4.40, SD = 0.47$) and the None ($M = 2.10, SD = 0.82$) conditions, with a statistically significant difference ($p \leq 0.05$). The result demonstrates how AR-CP provides a better-perceived situation interpretability and directed attention to conduct a safe driving experience.

6. Discussion and Future Work

As the results of our evaluation demonstrate, AR-CP approach represents a promising starting point for designing human-centric driving assistants. The uncertainty-aware aspect of our approach facilitates an interpretable and safer interaction between the user and the vehicle. However, this work presents some limiting factors that we use as guidelines for future works.

In our implementation, we adopt a 4-level deep taxonomy and consider only the classes "*vehicle*" and "*human*". This choice is based on available taxonomies and most frequent object classes in large-scale urban perception benchmarks, such as SemanticKitti [20], and the ROAD dataset [44]. However, it is crucial to investigate the quality of AR-CP with different taxonomy depths and sizes and more classes, such as "*traffic signs*" and "*vegetation*".

In our user study, we use VR as a realistic study environment. While VR represents a suitable trade-off between high levels of immersion and safety, it is important to see how the results would generalize in real-life conditions.

7. Conclusion

In this paper, we propose AR-CP, a novel uncertainty-aware framework for driver assistance that innovatively combines conformal prediction with augmented reality to enhance environmental perception and safety. By combining conformal prediction and augmented reality, AR-CP ensures more reliable detection in difficult conditions such as adverse weather and poor lighting. Our evaluation on the ROAD dataset demonstrates that AR-CP outperforms existing methods in providing small prediction sets while maintaining the theoretical coverage guarantees. Results from our immersive user study further show that AR-CP improves situation awareness and reduces mental load, significantly contributing to a safer driving experience.

Acknowledgement

This work has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center. This work has been partially funded by the German Federal Ministry of Education and Research (BMBF) through the Software Campus Project.

References

[1] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2, 5

[2] Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2457–2463. ijcai.org, 2022. 1

[3] Johannes Beller, Matthias Heesen, and Mark Vollrath. Improving the driver–automation interaction: An approach using automation uncertainty. *Human factors*, 55(6):1130–1141, 2013. 2

[4] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[5] Michael Braun, Nora Broy, Bastian Pfleging, and Florian Alt. A design space for conversational in-vehicle information systems. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–8, 2017. 2

[6] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019. 2

[7] Marine Capallera, Peïo Barbé-Labarthe, Leonardo Angelini, Omar Abou Khaled, and Elena Mugellini. Convey situation awareness in conditionally automated driving with a haptic seat. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, pages 161–165, 2019. 2

[8] Mark Colley, Christian Bräuner, Mirjam Lanzer, Marcel Walch, Martin Baumann, and Enrico Rukzio. Effect of visualization of pedestrian intention recognition on trust and cognitive load. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 181–191, 2020. 1, 2, 3

[9] Mark Colley, Max Rädler, Jonas Glimmann, and Enrico Rukzio. Effects of scene detection, scene prediction, and maneuver planning visualizations on trust, situation awareness, and cognitive load in highly automated vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–21, 2022. 1, 2, 3

[10] Michael Correll and Michael Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151, 2014. 2

[11] Alex Derhacobian, John Guibas, Linden Li, and Bharath Namboothiry. Adaptive prediction sets with class conditional coverage. 2

[12] Nicolas Dewolf, Bernard De Baets, and Willem Waegeman. Valid prediction intervals for regression problems. *Artificial Intelligence Review*, 56(1):577–613, 2023. 2

[13] Patrizia Di Campli San Vito, Edward Brown, Stephen Brewster, Frank Pollick, Simon Thompson, Lee Skrypchuk, and Alexandros Mouzakitis. Haptic feedback for the transfer of control in autonomous vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 34–37, 2020. 2

[14] Anushri Dixit, Lars Lindemann, Skylar X Wei, Matthew Cleaveland, George J Pappas, and Joel W Burdick. Adaptive conformal prediction for motion planning among dynamic agents. In *Learning for Dynamics and Control Conference*, pages 300–314. PMLR, 2023. 2

[15] Achref Doula, Lennart Schmidt, Max Mühlhäuser, and Alejandro Sanchez Guinea. Visualization of machine learning uncertainty in ar-based see-through applications. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 109–113. IEEE, 2022. 2

[16] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018. 2

[17] Adam Fisch, Tal Schuster, Tommi S. Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2

[18] Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pages 3329–3339. PMLR, 2021. 2

[19] Markus Funk, Vanessa Tobisch, and Adam Emfield. Nonverbal auditory input for controlling binary, discrete, and continuous input in automotive user interfaces. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020. 2

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 8

[21] Michael A Gerber, Mohammad Faramarzian, and Ronald Schroeter. Inception of perception—augmented reality in virtual reality: Prototyping human–machine interfaces for automated driving. In *User Experience Design in the Era of Automated Driving*, pages 477–503. Springer, 2022. 2

[22] David Goedicke, Jamy Li, Vanessa Evers, and Wendy Ju. Vr-oom: Virtual reality on-road driving simulation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018. 2

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 1

[24] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006. 7

[25] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. Presenting system uncertainty in automotive uis for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*, pages 210–217, 2013. 2

[26] Alex Kale, Sarah Lee, Terrance Goan, Elizabeth Tipton, and Jessica Hullman. Metaexplorer: Facilitating reasoning with epistemic uncertainty in meta-analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023. 2

[27] Susan S Kirschenbaum, J Gregory Trafton, Christian D Schunn, and Susan B Trickett. Visualizing uncertainty: The impact on performance. *Human Factors*, 56(3):509–520, 2014. 2

[28] Andreas Korthauer, Clemens Guenther, Andreas Hinrichs, Wen Ren, and Yiwen Yang. Watch your vehicle driving at the city: Interior hmi with augmented reality for automated driving. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020. 2

[29] Alexander Kunze, Stephen J Summerskill, Russell Marshall, and Ashleigh J Filtness. Conveying uncertainties using peripheral awareness displays in the context of automated driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 329–341, 2019. 2

[30] Jared LeClerc and Susan Joslyn. The cry wolf effect and weather-related decision making. *Risk analysis*, 35(3):385–395, 2015. 2

[31] Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 2023. 2

[32] Le Liu, Alexander P Boone, Ian T Ruginski, Lace Padilla, Mary Hegarty, Sarah H Creem-Regehr, William B Thompson, Cem Yuksel, and Donald H House. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE transactions on visualization and computer graphics*, 23(9):2165–2178, 2016. 2

[33] Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12008–12016, 2022. 2

[34] Alan M MacEachren, Robert E Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE transactions on visualization and computer graphics*, 18(12):2496–2505, 2012. 2

[35] Andrii Matviienko, Damir Mehmedovic, Florian Müller, and Max Mühlhäuser. " baby, you can ride my bike" exploring maneuver indications of self-driving bicycles using a tandem simulator. *Proceedings of the ACM on Human-Computer Interaction*, 6(MHCI):1–21, 2022. 2

[36] Andrii Matviienko, Florian Müller, Dominik Schön, Paul Seesemann, Sebastian Günther, and Max Mühlhäuser. Bikear: Understanding cyclists' crossing decision-making at uncontrolled intersections using augmented reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022. 2

[37] David Michalík, Miroslav Jirgl, Jakub Arm, and Petr Fiedler. Developing an unreal engine 4-based vehicle driving simulator applicable in driver behavior analysis—a technical perspective. *Safety*, 7(2):25, 2021. 2

[38] Mahsa Mirzargar, Ross T Whitaker, and Robert M Kirby. Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE transactions on visualization and computer graphics*, 20(12):2654–2663, 2014. 2

[39] Lace Padilla, Matthew Kay, and Jessica Hullman. Uncertainty visualization. *Computational Statistics in Data Science*, pages 405–421, 2022. 2

[40] Lace M Padilla, Sarah H Creem-Regehr, Mary Hegarty, and Jeanine K Stefanucci. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications*, 3(1):1–25, 2018. 2

[41] Andreas Riegler, Andreas Riener, and Clemens Holzmann. Content presentation on 3d augmented reality windshield displays in the context of automated driving. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 543–552. IEEE, 2022. 2

[42] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020. 2, 5

[43] Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021. 2

[44] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Salman Khan, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, et al. Road: The road event awareness dataset for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):1036–1054, 2022. 2, 5, 6, 8

[45] Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. *Advances in neural information processing systems*, 34:6216–6228, 2021. 2

[46] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[47] MT Taghavifard, K Khalili Damghani, and R Tavakkoli Moghaddam. Decision making under uncertain and risky situations. In *Enterprise Risk Management Symposium Monograph Society of Actuaries*, 2009. 2

[48] Richard M Taylor. Situational awareness rating technique (sart): The development of a tool for aircrew systems design. In *Situational awareness*, pages 111–128. Routledge, 2017. 7

[49] Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive inference with feature conformal prediction. *11th International Conference on Learning Representations, ICLR 2023*, 2023. 2

[50] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. 1, 3

[51] Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, and Felix Heide. Gated2gated: Self-supervised depth estimation from gated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2811–2821, 2022. 6

[52] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020. 1

[53] Xingwei Wu, Coleman Merenda, Teruhisa Misu, Kyle Tanous, Chihiro Suga, and Joseph L Gabbard. Drivers' attitudes and perceptions towards a driving automation system with augmented reality human-machine interfaces. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020. 2

[54] Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR, 2021. 2

[55] Dohyeon Yeo, Gwangbin Kim, and Seungjun Kim. Toward immersive self-driving simulations: Reports from a user study across six platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12, 2020. 2

[56] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022. 2

[57] Yizhe Zhang, Shuo Wang, Yejia Zhang, and Danny Z Chen. Rr-cp: Reliable-region-based conformal prediction for trustworthy medical image classification. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 12–21. Springer, 2023. 2

[58] Jieqiong Zhao, Yixuan Wang, Michelle V Mancenido, Erin K Chiou, and Ross Maciejewski. Evaluating the impact of uncertainty visualization on model reliance. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[59] Brian J Zikmund-Fisher, Holly O Witteman, Mark Dickson, Andrea Fuhrel-Forbis, Valerie C Kahn, Nicole L Exe, Melissa Valerio, Lisa G Holtzman, Laura D Scherer, and Angela Fagerlin. Blocks, ovals, or people? icon type affects risk perceptions and recall of pictographs. *Medical decision making*, 34(4), 2014. 2