SafeSO: Interpretable and Explainable Deep Learning Approach for Seat Occupancy Classification in Vehicle Interior

Joanna Jaworek-Korjakowska, Aleksander Kostuch, Pawel Skruch
AGH University of Science and Technology, Krakow, Poland
Department of Automatic Control and Robotics

jaworek,kostuch,pawel.skruch@agh.edu.ppl

Abstract

Classification of seat occupancy in in-vehicle interior remains a significant challenge and is a promising area in the functionality of new generation cars. As majority of accidents are related to the driver errors the consequences of not wearing, or improperly wearing, a seat belt are clear. The NHTSA reports that 47% of the 22,215 passenger vehicle occupants killed in 2019 were not wearing seat belts. To address this problem we propose a deep learning based framework to classify seat occupancy into seven most important categories. In this study, we present an interpretable and explainable AI approach that takes advantage of pre-trained networks including ResNet152V2, DenseNet121 and the most recent EfficientNetB0-B5-B7 to calculate the feature vectors followed by an adjusted densely-connected classifier. Our model provides an interpretation of its results through the identification of object parts without direct supervision and their contribution towards classification. We explore and propose two new statistical metrics including HGD_{score} and $HGDA_{score}$ which are based on the multivariate Gaussian distribution for assessing heatmaps without using human-annotated object parts to quantify the interpretability of our network. We demonstrate that the calculated statistical metrics lead to an interpretable model that correlates with the framework accuracy and can flexibly analyze heatmaps at any resolution for different user needs. Furthermore, extensive experiments have been performed on the SVIRO database [7] including 7,500 sceneries for BMW X5 model which confirm the ability of the developed framework based on the EfficientNetB5 architecture to classify seat occupancy into seven main categories with 79.87% overall accuracy as well as 95.92% recall and 90.32% specificity for empty seats recognition, which is a state-of-the-art result in this domain.

1. Introduction

In-cabin (or vehicle interior) sensing can be considered as a novel and promising approach to enhance the functionality of new generation cars. It is already proved to be useful in driver state monitoring and occupant detection systems and its application is gradually spreading from luxury marques to mass-produced models. In the near future, fusion of exterior and interior sensing can provide additional features necessary to achieve higher levels of autonomy. As a reminder, the Society of Automotive Engineers (SAE) distinguishes six level of autonomy [28], from 0, which means no automation to full automation at level 5. It should be noted that most of the currently produced models are at level 2 in this scale.

The perceptual system for the vehicle interior analyzes raw data provided by a sensor to detect and classify seat occupancy states, adults, children, animals, other objects such as belts or infant seat, driver's state and behavior. The information about the seat occupancy, for example, how many people are in the vehicle and at which seat positions they are located, can be used to remind the driver of passengers in the back seat when the driver is still in the vehicle. In particular this is important in case of children remaining in the vehicle as there are many cases when children died from heatstroke after being left in the vehicle. The child presence detection is one of the feature on the road map of the Euro NCAP standard. The detection of seat occupancy when connected with belt recognition can significantly enhance the safety of the vehicle's occupants. As majority of accidents are related to the driver errors the consequences of not wearing, or improperly wearing, a seat belt are clear. The NHTSA reports that 47% of the 22,215 passenger vehicle occupants killed in 2019 were not wearing seat belts [1]. In addition, information about what seats are occupied by passengers can be combined with an automatic emergency call system in case of accident situations. The perceptual system is exposed to many types of uncertainties caused by

Figure 1. Examples of the SVIRO database for X5 model regarding the seat occupancy scenarios: a) empty seat, b) infant in the infant seat, c) child in child seat, d) adult, e) everyday object, f) empty infant seat, and g) empty child seat [7].

environmental conditions (mainly lighting conditions) and characteristics of sensors. Moreover, the environment being monitored by the sensors contains actually infinite number of possible scenarios. So the design of efficient algorithms for detection and classification with high accuracy and precision in such working environment can be considered as a really challenging and difficult task which is not solved completely yet. Furthermore, as deep networks are increasingly used in the autonomous driving domain to automate data analysis, detection and classification, their decision-making process remains largely unclear and is difficult to explain to the end user. An interpretable and explainable approach can provide answers to many questions, which are crucial in autonomous driving and provide extra solutions.

In this paper, we address this problem by providing statistical metrics to explain and assess the deep model's decision. Specifically, we are interested in explaining classification decisions based on the heatmaps generated by the Grad-CAM algorithm that shows how important each image pixel is for the network's prediction [29]. We are proposing statistical metrics without using human-annotated object parts to quantify the interpretability of the deep network. Furthermore, in this paper we present a convolutional neural network based architecture (SO-CNN) to classify seat occupancy in vehicle interior into seven categories including infant in infant seat, child in child seat, adult, everyday object, or empty infant seat, child seat or seat, respectively (Fig. 1).We reuse the pre-trained CNN models including ResNet152V2, DenseNet121 and the most recent EfficientNetB0-B5-B7 for feature extraction which is followed by an adjusted densely-connected classifier. Experiments have been performed on the SVIRO database including dataset in-depth visualization and analysis as well as deep learning architecture adjustment and performance verification [7]. We further employ the Grad-CAM algorithm to generate heatmaps and calculate the proposed new statistical metrics including HGD_{score} and $HGDA_{score}$ which are based on the multivariate Gaussian distribution to conduct the multi-task learning model interpretability.

The main contributions of the present paper can be summarized as follows:

- In this paper, we present a CNN based approach for the seat occupancy classification in vehicle interior into seven main categories based on the adjusted pre-trained EfficientNetB5 network architecture.
- We propose a new approach for model interpretability based on Grad-CAM heatmaps analysis. The statistical metrics HGD_{score} and $HGDA_{score}$ are based on the values of the density function for a non axis-aligned multivariate Gaussian distribution and its probability to quantify the interpretability without using human-annotated object parts.
- We perform an in depth analysis of the SVIRO dataset benchmarked for the classification of seat occupancy for each individual seat.
- We perform extensive experiments and compare the outcomes of state-of-the-art pre-trained models including ResNet152V2, DenseNet121 as well as EfficientNetB0-B5-B7. We visualize the feature distribution extracted by each architecture.
- We compare and estimate the correlation between prediction and proposed statistical metrics dedicated for model interpretability.

1.1. Related works

The problem of detection and classification of the seat occupancy in the vehicle interior can be accomplished using the information provided by cameras, radars or ultrasonic sensors. Raw data provided by these sensors constitutes an input to the perception algorithms aimed to monitor and interpret what is happening both inside and outside of the vehicle. In the design and development process of the vision-based systems for the automotive industry one of the most important issue is related to the system's performance, safety, reliability and interpretability. This is in particular valid to the systems that utilize machine learning components. It is clear that testing in detail such a system is

impossible as the number of important scenarios is actually infinite. The process of selecting just a few of the many possible scenarios is a difficult and challenging task and currently is most often based on qualitative best engineering judgment. One of the most challenging task is proposing safe AI solutions regarding standardization, transparent training as well as model evaluation including interpretability for automated driving.

Deep model interpretability: Recently, different methods have been developed to visualize and interpret deep learning architectures using the gradient-based or its variants including DeconvNet [21] and Saliency Maps[30]. Grad-CAM uses the gradients of a target concept, flowing only into the final convolutional layer to produce a coarse localization map [29]. Also different statistical metrics have been proposed to analyze CNN features including analysis of properties of CNN [34] or quantification of the generality versus specificity of neurons [39]. Ribeiro et al. proposed the LIME method which is an explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the predictions [26]. Extracted image regions that were responsible for each network output in order to analyse the network behaviour were used in [16]. Another statistical metric include pointing game introduced in [40] or RISE approach that estimates importance empirically by probing the model with randomly masked versions of the input image and obtaining the corresponding output[24]. However, most of these methods require experts to manually identify discriminative image regions for the label prediction for each testing image.

Deep models in autonomous driving: There are ongoing studies on the design of neural network based architectures for detection and classification tasks. A review of deep learning approaches for object detection including some experimental analysis can be found in [41]. Besides object detection there are also deep learning methods designed for specific tasks, such as driver gaze estimation [22, 38], gaze and eye tracking [14]. Within the last years several networks have been developed that might be considered as key building blocks for dedicated applications. Family of R-CNN (Region Based CNNs), YOLO (You Only Look Once), SDD (Single Shot Detectors), RetinaNet [13], PSMNet [4], NASNet, Inception, MobileNet, EfficientDet [36], VGG19 [31], ResNet50, Exception, DenseNet121 [11], EfficientNets [35] are current state- of-the-art models in computer vision benchmarks.

Seat occupancy frameworks: There are some commercial solutions and patents developed by the automakers and automotive suppliers for the seat occupancy detection and classification, however there is not a lot of research papers that present efficient algorithms in this area alongside with the experimental analysis. The work [6] describes a stereo vision system capable to detect the passenger presence and its location in the vehicle interior. Several typical configurations such as empty seat, adult present and baby seat can be distinguished by the developed system. In [19] Near-Infrared Camera (NIR) has been used to detect the passenger presence on front and rear seats. Images recorded by thermal camera have been used in [23] to train CNNs to detect number of passengers in the vehicle. To the best of our knowledge this is the first attempt to classify seat occupancy in vehicle interior based on scenarios into seven most important categories using deep learning approach.

2. Methodology

The automated determination of seat occupancy is performed automatically according to the flowchart presented in Fig. 2 were the first part is responsible for determining the category of seat occupancy into seven main categories while the second part for the model interpretability and explainability so the decision can be assessed and understood. As the deep learning architecture is designed to perform classification on each individual seat according to the SVIRO dataset, in the first step, the images need to be split into three rectangles such that each seat can be classified individually [7]. In the second step we take advantage of the pre-trained networks including the ResNet152V2, DenseNet121 and EfficientNetB0-B5-B7 architectures for feature extraction stage. We adjust the classifier on top of deep convolutional neural networks which has a three layer structure containing global average pooling, dropout and dense layers followed by Softmax function. The deep learning architecture is fine-tuned with the publicly available SVIRO dataset. As a result, we generate classification outcomes and employ the Grad-CAM algorithm to generate heatmaps and calculate the statistical metrics to conduct the multi-task learning model interpretability.

2.1. SVIRO dataset specification

Verification of automotive vision systems requires large, variable and diverse datasets in order to assure proper reliability and safety levels alongside with the expectation for high accuracy and precision of classification algorithms. KITTI [8], nuScenes[3], Audi [9], Waymo [33], SVIRO [7], U2Eyes[25], CBSR NIR Face Dataset [17] are data collections available for research purposes. It should be also emphasized that for the vehicle interior public circulation of the data is limited due to the General Data Protection Regulation (GDPR). In our research we take advantage of the Synthetic Dataset for Vehicle Interior Rear Seat Occupancy (SVIRO) to classify people and objects in passenger compartment [7]. The dataset is based on 10 different vehicle interiors and 25.000 sceneries in total. In this research we use the BMW X5 model which consists of 7.500 sceneries. The dataset contains detailed description for following cat-

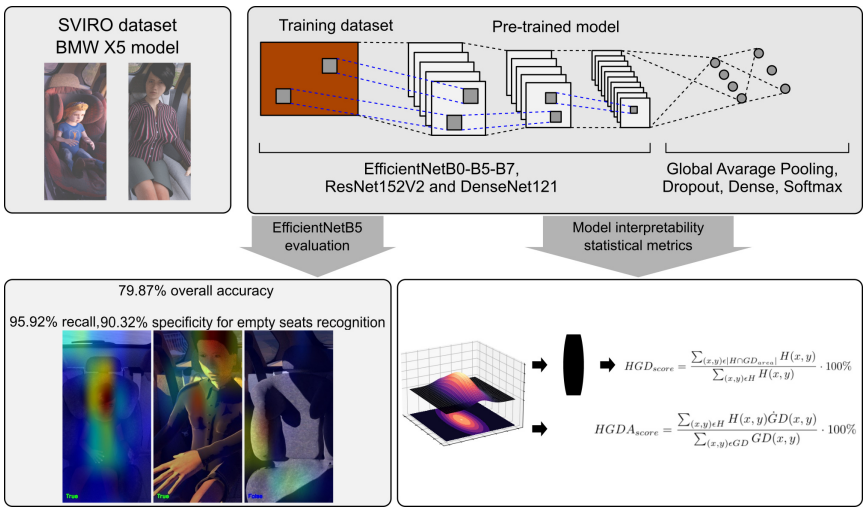


Figure 2. The streamline of our proposed framework based on adjusted CNN architecture and model interpretability. Classification is performed on each individual seat from the SVIRO dataset. For feature extraction we use the pre-trained deep learning models. The classification is based on the adjusted densely-connected classifier. We employ the extracted features to conduct the multi-class classification task. Finally, we perform the model evaluation and interpretation and calculate the proposed metrics HGD_{score} and $HGDA_{score}$.

egories regarding the seat occupancy: infant in infant seat, child in child seat, adult, everyday object, empty infant seat, and empty child seat. In Table 1 we present the database with distribution between training and testing set regarding the number of images for each of the classes.

To understand the complexity of the classification problem we perform the SVIRO dataset analysis through visualization of the data distribution using two dimensionality reduction techniques including the t-distributed Stochastic Neighbor Embedding technique (t-SNE) [37] and recently proposed Uniform Manifold Approximation and Projection (UMAP) method [20]. UMAP is a learning technique using Riemannian manifold distribution for dimension reduction and t-SNE is an unsupervised method that minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the embedding [18]. Fig. 3 shows the visualisation of the SVIRO dataset distribution in terms of seat occupation and seat occupancy classification using t-SNE on PCA-reduced data and UMAP techniques.

In order to analyze the correlations and overlapping areas between the categories in the dataset, we have calculated statistical metrics which give us a better understanding of the problem. We have calculated the intra-class and inter-class ratio (IntraC, InterC) based on the Euclidean distance. We analyze the Silhouette Coefficient Score (S), which is given by [27]:

$$S = \frac{b - a}{max(a, b)} \tag{1}$$

where a is the average distance in the cluster and b is the minimal average distance to the next cluster. Additionally, the Davies-Bouldin index has been calculated that signifies the average similarity between clusters as a measure that compares the distance between clusters with the size of the clusters themselves and is defined as [5]:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{2}$$

where

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{3}$$

and s_i is the average distance between each point of cluster i and the centroid of that cluster, d_{ij} is the distance between cluster centroids and k is the number of clusters.

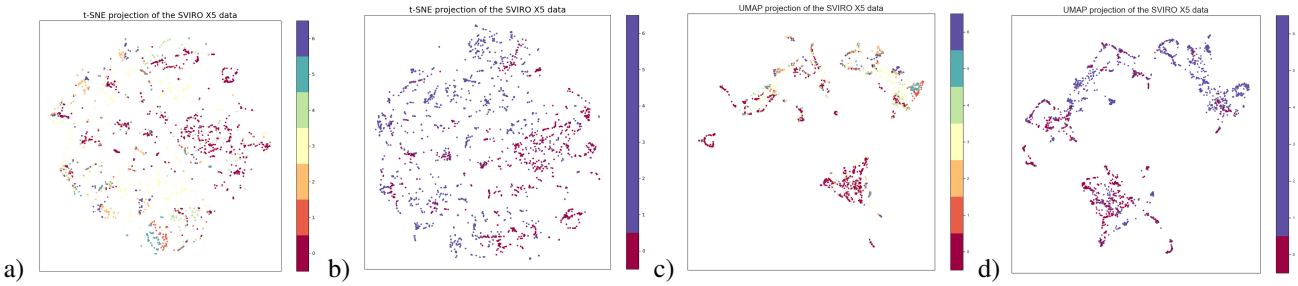
Table 3 contains the results of statistical analysis of the SVIRO dataset. We observe that the complexity of the underlying classification task is very high as the Silhouette score confirms that the classes are overlapping with samples very close to the decision boundary of the neighbouring clusters. However the high CH score as well as $IntraC_{WAvg}$ indicates the possibility of separating the data into seven main categories.

2.2. Proposed deep learning approach

Due to our limited and imbalanced dataset we take advantage of the transfer learning concept which indicates the effectiveness of reusing pre-trained CNN architectures to extract the feature representation. We use several state-of-the-art architectures including ResNet152V2 [10], DenseNet121 [11] as well as the newest EfficientNetB0-B5-B7 [35]. EfficientNet models which have been introduced in 2019 by Tan *et al.* are based on the inverted bottleneck residual blocks of MobileNetV2 and squeeze-and-excitation blocks. They use a compounding scaling method

Table 1. Statistical analysis of the dataset with respect to the training and testing datasets associated with the seven main seat occupancy categories.

	Empty	Infant in infant seat	Child in child seat	Adult	Everyday object	Empty infant seat	Empty child seat	Total Nb.
SVIRO Database	3010	480	669	1126	962	542	711	7.500
Training dataset	2400	371	569	892	767	418	583	6.000
Testing dataset	610	109	100	234	195	124	128	1.500



a) b) c) d)

Figure 3. Visualization of data distribution within the SVIRO dataset where a) t-SNE visualization for seven categories, b) t-SNE visualization for seat occupancy vs empty, c) UMAP visualization for seven categories, d) UMAP visualization for seat occupancy vs empty.

which scales width, depth, and resolution together instead of scaling only one model attribute. The EfficientNetB0 architecture has been proposed by a multi-objective neural architecture search which optimizes both accuracy and floating-point operations. Furthermore, a new activation function, Swish, has been proposed which shows superior performance for deeper networks. Swish is a multiplication of a linear and a sigmoid activation [35]:

$$Swish(x) = x \cdot sigmoid(x) \tag{4}$$

On top of the base, we have adjusted a three layer classifier containing global average pooling with batch normalization, additional dropout layer which randomly sets input units to 0 with frequency of rate 0.2 at each step during training time as a regularization technique for reducing overfitting [32], and dense layer with the number of neurons corresponding to the number of seven classes followed by Softmax activation function for the predict a multinomial probability distribution.

2.3. Training details

Preprocessing: The entire image taken by the virtual rear seat camera has been cropped into 3 pieces of the same size, each area is representing one seating position. Apart from resizing the input image no other preprocessing methods were used. As the dataset has a decent class balance and the variety of scenarios is high no data augmentation was needed to be applied. Furthermore, different lightning conditions, textures and models' positions were randomized during its development.

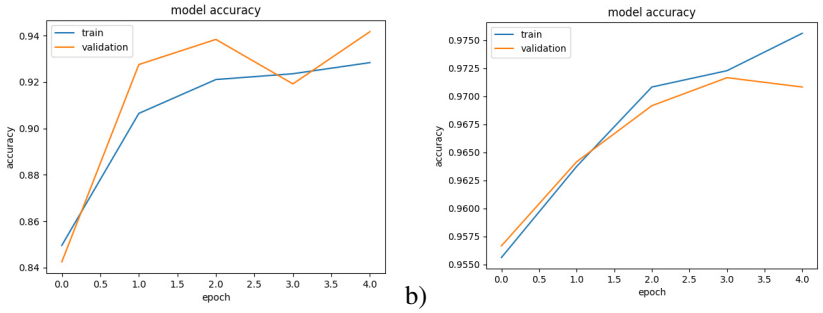
Training: Training process consisted of two separate phases where first we trained the added top layers and then the entire network (excluding batch normalization layers). The goal of this transfer-learning approach is to adapt the new and randomly filled top layers to the class set defined in the database. After this phase the rest of the model is unfrozen and trained again. The original datasets used for training are very different than SVIRO base, thus according to definition [12] batch normalization parameters (γ and β) should not be updated. Otherwise it would necessitate optimizing every other weight from the start, which means the loss of previously trained feature extraction functionality.

Hyperparameter choice: For each training phase of the pre-trained architectures including ResNet152V2, DenseNet121 and three different types of EfficientNet: B0, B5 and B7, we deployed randomized search (RandomizedSearchCV) for optimizing hyperparameters including epoch count, optimizer and batch size [2]. The algorithm selected 20 random parameter sets from the predefined range. We tested the batch size and number of epochs in the range of 8-64 and 5-50 respectively and several optimizers including RMSprop, SGD, Adadelta, Adam and Adamax. Initial learning rate has been set at different levels for each phase, 10^{-2} for the first and 10^{-4} for the second phase. The validation set was used at the initial stage of classifier selection and for empirical evaluation of the model's behaviour during and after training. In Fig. 4 we show the average training and validation accuracy for the EfficientNetB5 architecture, which is the most efficient and has achieved the highest accuracy score.

3. Experimental results

3.1. Effectiveness of the proposed framework

We test our proposed framework on the BMW X5 model from the SVIRO dataset. The adjusted pre-trained models have been trained in an end-to-end fashion to classify the images into seven categories based on the seat occupancy. The architectures have been trained independently in order to optimise the hyperparameters and the proposed



a) b)

Figure 4. Accuracy score plot for 5 epochs of training and validation in both training phases during EfficientNetB5 training. In the first phase, only the top layers are being updated, whereas in the following one, the entire network excluding batch normalization layers.

pre-trained CNN models have been evaluated on the test set. The test results have been calculated 5 times and averaged. Following performance metrics including accuracy, precision, recall and F1-score have been calculated. Table 2 presents the evaluation metrics for each network architecture for the best set of training hyperparameters. EfficientNetB5 achieved 79,9% accuracy, 76,8% precision, 69,3% recall and 67,8% F1-score which were the best results when trained with 5+5 epochs, batch size of 32 and Adam optimizer [15]. It can be seen from Table 2 that the EfficentNetB5 model achieves the highest classification accuracy in the classification task.

Furthermore, in Fig. 5 we present the confusion matrix for the multi-task classification problem. The EfficientNetB5 network achieved 95.92% recall and 90.32% specificity for empty seats recognition, which is a state-of-the-art result in this domain.

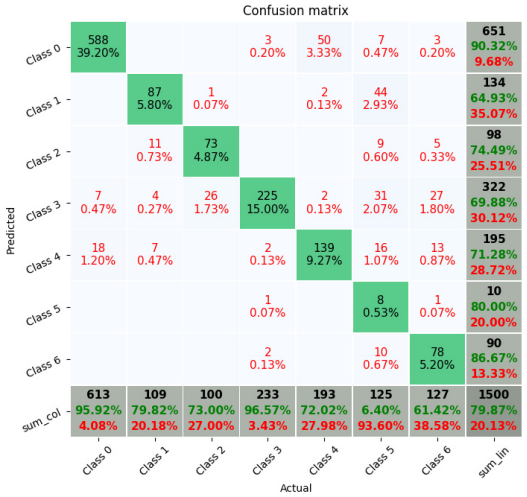
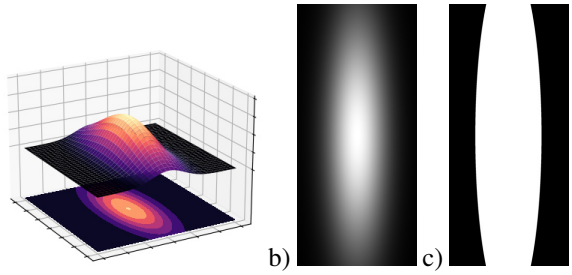


Figure 5. Confusion matrix for the multi-task classification problem where the class number is according to Fig. 1.

.

3.2. Model interpretability

In Table 3 we compare the statistical analysis of the SVIRO dataset and feature vectors extracted from the last layer of the pre-trained architectures in terms of data separability into seven main categories. We can observe that $IntraC_{WAvg}$ values for EfficientNetB5 is relatively high as well as the S score which is the highest for all classes. That indicates good separability and confirms the best performance for this network. To improve the explainability of the model, we used the Grad-CAM visualization algorithm [29], which creates a heatmap that shows which parts of the input image contributed most to the classification.



a) b) c)

Figure 6. CNN model interpretability: a) 3D surface plot of bivariate Gaussian distribution, b) GD mask for $HGDA_{score}$, c) GD_{area} mask for HGD_{score}.

Moreover, we propose two statistical metrics HGD_{score} and $HGDA_{score}$ which are based on the multivariate Gaussian distribution for assessing heatmaps without using human-annotated object parts to quantify the interpretability of our network. The multivariate normal distribution is a generalization of the univariate normal distribution to two or more variables and can be defined for a k-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)^{\mathrm{T}}$ with the following notation:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5}$$

The probability density function (pdf) of the d-dimensional multivariate normal distribution is given by:

$$f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|\,(2\pi)^d}} e^{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)'} \tag{6}$$

Table 2. Obtained results for state-of-the-art pre-trained CNN architectures with obtained optimal training hyperparameters.

Architecture	Optimal training hyperparameters			Metrics			
	Optimizer	Batch Size	Epochs	Accuracy	Precision	Recall	F1
ResNet50	Adam	64	5+5	0.558	0.451	0.420	0.427
DenseNet121	Adam	32	5+5	0.537	0.469	0.412	0.372
EfficientNetB0	Adam	32	5+5	0.739	0.730	0.598	0.622
EfficientNetB7	Adam	32	5+5	0.647	0.721	0.657	0.619
EfficientNetB5	**Adam**	**32**	**5+5**	**0.799**	**0.768**	**0.693**	**0.678**

Table 3. Statistical analysis on the separability of the SVIRO dataset as well as deep learning methods based on the visualization of the UMAP dimension reduction technique, where the intra-class distance has been calculated for each category: 0- infant in infant seat, 1- child in child seat, 2- adult, 3- everyday object, 4- empty infant seat, 5- empty child seat, and 6- empty seat.

Method	Metrics									
	IntraC$_0$	IntraC$_1$	IntraC$_2$	IntraC$_3$	IntraC$_4$	IntraC$_5$	IntraC$_6$	IntraC$_{WAvg}$	S	DB
SVIRO dataset	568.45	499.85	420.23	567.13	545.09	461.23	409.67	524.07	-0.064	7.05
EffcientNetB0	361.77	268.03	270.41	290.72	343.65	214.89	246.76	327.61	-0.113	7.38
EffcientNetB5	743.80	521.71	599.65	580.07	723.70	375.31	535.74	661.84	-0.115	8.10
EffcientNetB7	892.60	729.35	608.85	822.82	1023.12	662.42	728.54	874.80	-0.114	8.39
ResNet152	336.74	301.30	264.83	348.13	361.15	268.51	323.35	328.93	-0.077	8.20
DenseNet121	182.19	145.12	141.26	167.11	157.39	135.84	122.77	164.01	-0.104	7.35

where it is parametrized in terms of the mean vector and the covariance matrix, denoted by μ and Σ respectively. x and μ are 1-by-d vectors and Σ is a d-by-d symmetric, positive definite matrix.

Based on the assumption that the most important elements responsible for the classification process lie in the center of the picture we propose two different masks which are generated on the basis of the probability density function with $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 3.5 & 0 \\ 0 & 1 \end{pmatrix}$ .The parameters have been obtained experimentally based on generated bounding boxes for a set of analysed images. Based on the calculated probability density function we propose two masks which are used to determine the significance of the location of the heatmap in the image. The first map GD_{area} is a binary map obtained by including the central 90% of the probability of the normal distribution while leaving out a total of $\alpha = 5\%$ in each tail, of the normal distribution. The HGD_{score} is defined as the sum of intensity pixels in the heatmap within the GD_{area} divided by the sum of all pixels in the heatmap. The formula is given by:

$$HGD_{score} = \frac{\sum_{(x,y)\epsilon|H \cap GD_{area}|} H(x,y)}{\sum_{(x,y)\epsilon H} H(x,y)} \cdot 100\% \quad (7)$$

The second map is the bivariate normal distribution GD with high values in the center of the image and low values at the edges. The second metric The $HGDA_{score}$ is defined as the sum of intensity pixels in the heatmap multiplied by the GD mask and divided by the sum of all pixels in the GD mask. The formula is given by:

$$HGDA_{score} = \frac{\sum_{(x,y)\epsilon H} H(x,y)GD(x,y)}{\sum_{(x,y)\epsilon GD} GD(x,y)} \cdot 100\% \quad (8)$$

In Table 4 we present the results for calculated statistical metrics. We observe that the HGD_{score} achieves high values for correctly classified images when the heatmap has the highest activation in the center of the image. Values of $HGD_{score} < 50\%$ correlate to high classification error. Based on the analysis of the HGD_{score} metric we can observe that 64% of classification errors have been based on wrong CNN network region attention while for over 89% of correctly classified examples the network has focused on the area in the centre of the image.

In Fig. 7 we present four different examples including two child seats, an empty seat and an everyday object (bottle box). First two examples have been incorrectly classified and the heatmap focuses on the surrounding not the seat. The last two examples have been correctly classified which is also confirmed by the heatmaps concentrating in the middle of the image. From these images we can draw several conclusions. Firstly, we observe that the heatmaps and the area of activations highly correlate with the region of interest and the statistical metrics confirm the certainty of the final result. These results provide strong evidence of the importance of differentiating between classification results based on the significant area and surroundings which may lead to higher accuracy, interpretability and classification certainty.

4. Conclusion

In this work, we developed a pre-trained based architecture for safe occupancy classification into seven main categories. The framework is based on the EfficientNetB5 architecture and achieved 79.87% overall accuracy as well as 95.92% recall and 90.32% specificity for empty seats recognition, which is a state-of-the-art result in this domain. Furthermore, we have proposed two new statistical met-

Table 4. Outcome for the proposed statistical metrics calculated for the test dataset including 1,500 images for seven categories. The threshold has been set to 50% for HGD_{score} and to 25% for $HGDA_{score}$.

	Correct Classification (#$HGD_{score} > 50\%$ / #Total Nb.)	Classification Error (#$HGD_{score} < 50\%$ / #Total Nb.)
HGD_{score}	89% (952/1069)	64% (276/431)
	Correct Classification (#$HGDA_{score} > 25\%$ / #Total Nb.)	Classification Error (#$HGDA_{score} < 25\%$ / #Total Nb.)
$HGDA_{score}$	74% (791/1069)	68% (293/431)

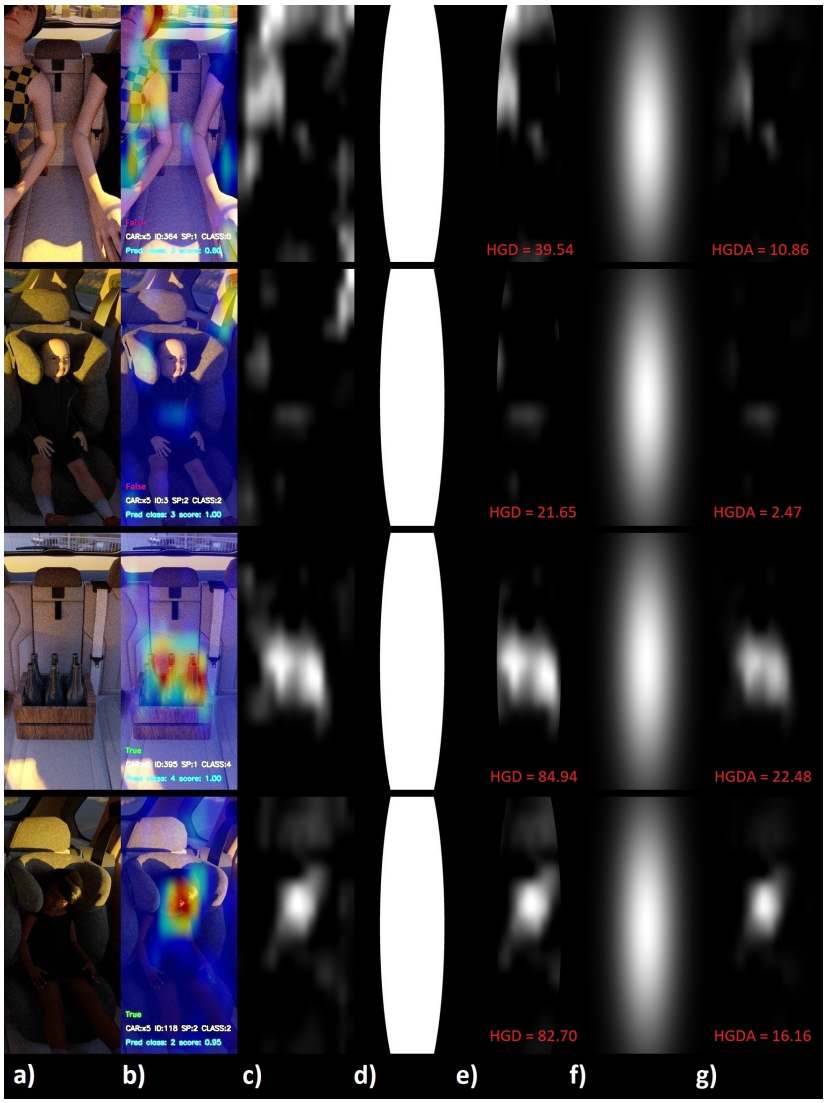


Figure 7. Results of the CNN model interpretability: a) input image, b) Grad-CAM visualization, c) heatmap activations, d) GD_{area} mask for HGD_{score}, e) overlapping area of heatmap, f) GD mask for $HGDA_{score}$, g) overlapping area of heatmap.The two first rows are false prediction cases and the next two are true.

rics which provide an interpretation of its results through the identification of object parts without direct supervision and their contribution towards classification. Starting from the described framework, further research efforts will be firstly addressed to compare and integrate other car models which are available in the SVIRO dataset, improve the network performance through fine-tuning of the layers. Future research will concentrate on the model interpretability and calculation of statistical metrics for other classification tasks.

Acknowledgment

We gratefully acknowledge the funding support of the "Excellence initiative – research university" programme for the AGH University of Science and Technology. The work has been carried out with support of Advanced Engineering group at Aptiv Technical Center Krakow.

References

[1] United States. National Highway Traffic Safety Administration. 1

[2] J. Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, 2012. 5

[3] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3

[4] J.R. Chang and Y.S. Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[5] D. L. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1:224–227, 1979. 4

[6] M. Devy, A. Giralt, and A. Marin-Hernandez. Detection and classification of passenger seat occupancy using stereovision. In *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511)*, pages 714–719, 2000. 3

[7] S. Dias Da Cruz, O. Wasenmüller, H. Beise, T. Stifter, and D. Stricker. Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1, 2, 3

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

[9] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A.S. Chung, L. Hauswald, V.H. Pham, M. Mühlegg, S. Dorn, T. Fernandez, M. Jänicke, S. Mirashi, C. Savani, M. Sturm, O. Vorobiov, M. Oelker, S. Garreis, and P. Schuberth. A2d2: Audi autonomous driving dataset, 2020. 3

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks, 2018. 3, 4

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. 5

[13] H. Kaiming, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[14] M.Q. Khan and S. Lee. Gaze and eye tracking: Techniques and applications in ADAS. *Sensors*, 19(24):5540, 2019. 3

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[16] D. Kumar, A. Wong, and Graham W. Taylor. Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1686–1694, 2017. 3

[17] S. Li, D. Yi, Z. Lei, and S. Liao. The casia nir-vis 2.0 face database. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 348–353, 2013. 3

[18] L. V. D. Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 4

[19] A. Makrushin, M. Langnickel, M. Schott, C. Vielhauer, J. Dittmann, and K. Seifert. Car-seat occupancy detection using a monocular 360° NIR camera and advanced template matching. In *2009 16th International Conference on Digital Signal Processing*, pages 1–6, 2009. 3

[20] L. McInnes, J. Healy, N. Saul, and L. Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. 4

[21] A. Mukherjee, S. Chakraborty, and S. Saha. Detection of loop closure in slam: A deconvnet based approach. *Appl. Soft Comput.*, 80:650–656, 2019. 3

[22] R.A. Naqvi, M. Arsalan, G. Batchuluun, H.S. Yoon, and K.R. Park. Deep learning-based gaze detection system for automobile drivers using a nir camera sensor. *Sensors*, 18(2):456, 2018. 3

[23] F.E. Nowruzi, A.W. El Ahmar, R. Laganiere, and A.H. Ghods. In-vehicle occupancy detection with convolutional networks on thermal images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[24] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 3

[25] S. Porta, B. Bossavit, R. Cabeza, A. Larumbe-Bergera, Gonzalo G. Garde, and A. Villanueva. U2eyes: a binocular dataset for eye tracking and gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[26] M. Tulio Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. 3

[27] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 4

[28] SAE International. *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, Standard J3016\201806.www.sae.org/standards*, 2018. 1

[29] R. R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, D. Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019. 2, 3, 6

[30] K. Simonyan, A. Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014. 3

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[32] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014. 5

[33] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, and B. Caine. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 3

[34] Ch. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014. 3

[35] M. Tan and Q.V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 3, 4, 5

[36] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020. 3

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4

[38] B. Vasli, S. Martin, and M. M. Trivedi. On driver gaze estimation: Explorations and fusion of geometric and data driven approaches. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 655–660, 2016. 3

[39] J. Yosinski, J. Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014. 3

[40] J. Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126:1084–1102, 2017. 3

[41] Z. Zhao, P. Zheng, S. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019. 3