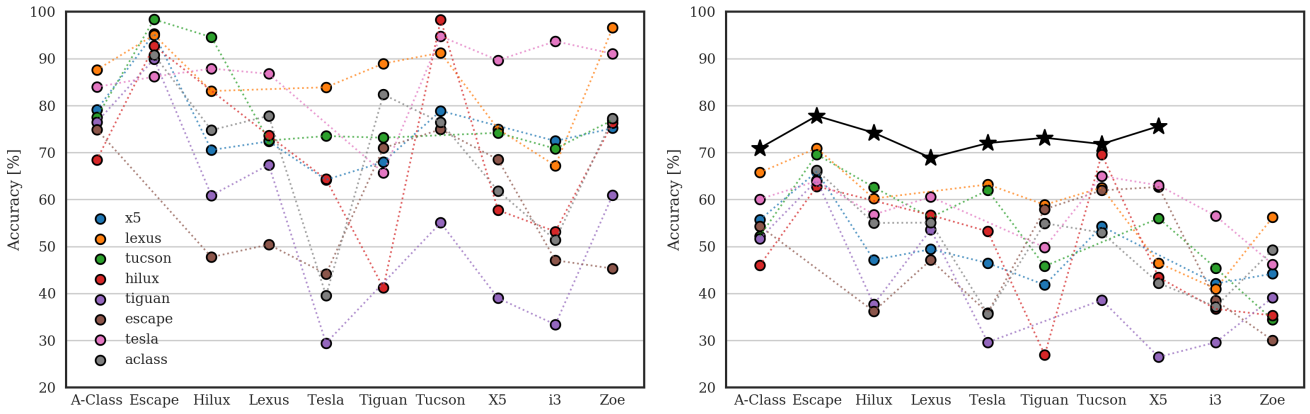
Autoencoder Based Inter-Vehicle Generalization for In-Cabin Occupant Classification

Supplementary Material

1. Ablation study: Performance on different vehicles when trained on different vehicles



(a) Training images - unknown vehicles

(b) Test images - unknown vehicles

Figure S1: Additional results for Fig. 4 of the main paper. We trained an individual MS-SSIM autoencoder on each of the eight vehicles. The resulting models were evaluated on the training images (left) and the test images (right) of the nine vehicles not seen during training. The different colors represent the vehicles each model was trained on. The performances on the test images of the vehicle the models were trained on (★) are plotted as well. The fluctuation for the generalization to unknown class instances is smaller than for the generalization between different vehicles. Although generalizing to new class instances is hard, this gives hint that the changing background has a large influence on the models' robustness.

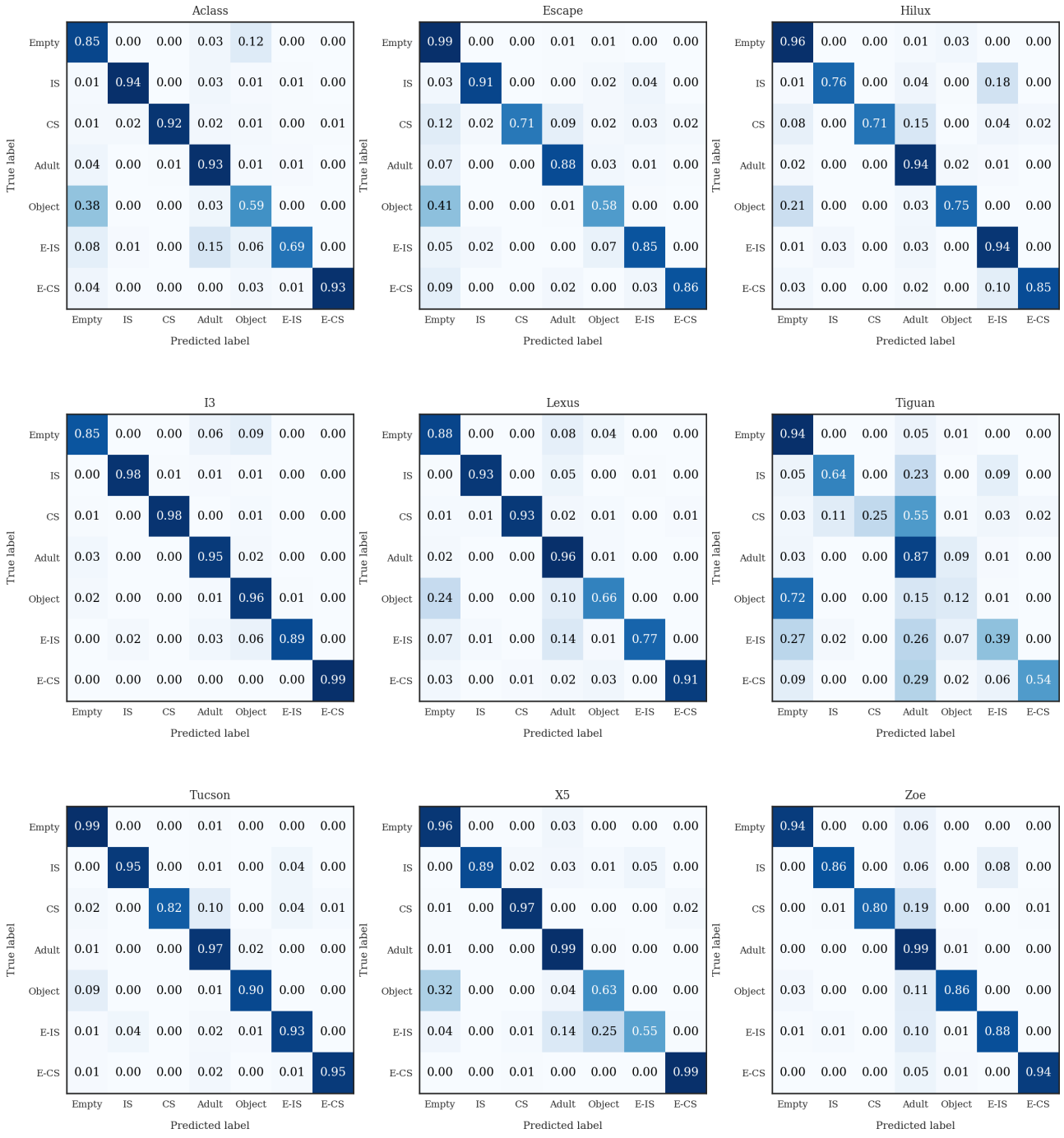


Figure S2: Additional results for Fig. 5. We report the confusion matrices for all the remaining vehicles together with the one presented in the main paper. The model trained on the Tesla vehicle behaves differently on the various vehicles such that no guarantees can be provided without additional precautions.

2. Hyperparameter study

Table S1: We tested the best hyperparamters for the autoencoder on the classification models for a fair comparison between both models' architectures. The classification models were trained on the Tesla and evaluated on all training and test images of the unknown vehicles. We tested the models when trained from scratch or when fine-tuning all layers from a pre-trained one. There is no clear winner, but we decided to not use weight decay as it performs slightly better.

Model	Weight decay	Approach	Train Accuracy	Test Accuracy
VGG-16	0	Scratch	78.3	50.6
		Fine-tune	94.0	61.8
	0.01	Scratch	82.5	47.0
		Fine-tune	92.9	61.4
DenseNet-121	0	Scratch	76.2	47.3
		Fine-tune	88.8	64.5
	0.01	Scratch	62.8	42.7
		Fine-tune	84.5	59.2
MobileNet	0	Scratch	76.7	49.4
		Fine-tune	91.5	60.6
	0.01	Scratch	79.8	50.8
		Fine-tune	91.7	58.9
ResNet-50	0	Scratch	80.1	51.9
		Fine-tune	88.0	57.4
	0.01	Scratch	75.0	46.8
		Fine-tune	84.2	51.9
ResNet-18	0	Scratch	75.4	47.8
		Fine-tune	86.3	57.2
	0.01	Scratch	80.4	52.9
		Fine-tune	86.8	54.9
SqueezeNet	0	Scratch	69.6	46.1
		Fine-tune	83.3	50.2
	0.01	Scratch	71.3	45.3
		Fine-tune	83.2	50.6

Table S2: Comparison of different γ values to weight the classification loss accordingly with the MSE reconstruction loss. The autoencoder was trained on the Tesla vehicle and evaluated on the training and test images of all unknown vehicles.

	γ							
	1	25	50	75 (used)	100	125	150	175
Training Accuracy	61.6	76.7	77.9	83.5	82.8	82.5	83.6	82.7
Test Accuracy	47.2	55.2	51.2	58.2	54.8	55.5	55.8	58.1

3. Additional detailed results for all models trained on all vehicles and tested on training and test images of all vehicles

Table S3: Additional results for Table 1 of the main paper. Comparison of the accuracies (in percentage) across the different vehicles for several classification models and autoencoders with different reconstruction losses. The classification models were trained from ***scratch*** (S) or ***fine-tuned*** (F) and the autoencoders with (AE) and without (AE-W) max-unpooling were trained from scratch only. The models were trained on the augmented training images of the ***Tesla*** vehicle and tested on the ***test*** images of all vehicles not seen during training. The last column is the mean performance across all nine unknown vehicles: the autoencoders outperform all classification models when trained from scratch.

	Trained on Tesla. Tested on vehicle:									
Model	A-Class	Escape	Hilux	Lexus	Tiguan	Tucson	X5	i3	Zoe	Mean
VGG-16 (F)	64.9	67.7	55.1	67.1	50.4	65.1	67.7	62.5	55.4	61.8
DenseNet (F)	61.4	69.5	67.0	74.9	57.7	57.3	77.0	66.8	49.1	64.5
MobileNet (F)	54.8	57.8	62.2	68.4	54.4	58.8	72.1	63.4	53.2	60.6
ResNet-50 (F)	47.5	61.8	51.7	61.6	53.7	60.9	70.1	61.2	48.1	57.4
ResNet-18 (F)	54.6	68.8	50.3	64.9	50.6	55.3	70.9	59.8	39.5	57.2
SqueezeNet (F)	47.3	55.1	40.7	47.1	56.5	51.6	62.0	54.1	37.3	50.2
VGG-16 (S)	46.9	56.2	42.7	59.9	39.2	61.7	57.7	49.7	41.5	50.6
DenseNet (S)	42.3	55.6	40.2	44.9	47.1	53.9	58.6	48.2	34.6	47.3
MobileNet (S)	55.1	48.1	49.9	55.2	46.0	53.1	53.9	40.5	43.1	49.4
ResNet-50 (S)	53.1	56.5	47.7	50.9	41.8	56.9	60.6	49.9	49.7	51.9
ResNet-18 (S)	39.5	56.9	43.4	46.0	38.0	54.3	60.3	54.8	37.2	47.8
SqueezeNet (S)	47.0	51.0	37.3	42.7	50.9	49.1	59.3	45.5	32.0	46.1
AE - SSIM	52.9	61.3	53.0	49.8	50.9	62.1	66.5	58.9	47.8	55.9
AE - MS-SSIM	60.1	63.9	56.8	60.5	49.8	65.0	63.0	56.5	46.2	58.0
AE - Perceptual	51.0	59.3	47.9	56.7	49.3	61.1	59.5	56.9	47.3	54.3
AE - MSE	57.8	61.4	57.4	61.1	50.5	62.4	61.7	61.7	49.8	58.2
AE-W - SSIM	54.9	60.3	50.3	59.5	49.7	61.4	62.5	53.6	49.3	55.7
AE-W - MS-SSIM	50.4	59.9	54.3	61.2	48.5	60.7	63.1	53.7	47.8	55.5
AE-W - Perceptual	62.5	59.6	57.1	61.1	52.3	60.5	63.8	58.5	49.7	58.3
AE-W - MSE	61.3	60.4	53.3	62.1	50.6	61.8	65.1	59.6	49.5	58.2

Table S4: Additional results for Table 1. The autoencoders without (AE-W) max-unpooling were trained from scratch on the augmented training images of the ***Tesla*** vehicle and tested on the ***training*** images of all vehicles not seen during training.

	Trained on Tesla. Tested on vehicle:									
Model	A-Class	Escape	Hilux	Lexus	Tiguan	Tucson	X5	i3	Zoe	Mean
AE-W - SSIM	81.6	86.0	85.9	85.0	65.7	92.0	88.7	90.0	91.5	85.1
AE-W - MS-SSIM	73.9	82.7	82.1	84.1	58.3	89.8	88.5	88.4	84.4	81.4
AE-W - Perceptual	81.1	80.7	83.9	82.5	61.1	91.3	83.4	90.3	89.6	82.7
AE-W - MSE	81.7	81.3	78.1	83.3	61.1	93.0	81.8	88.8	91.6	82.3

Table S5: Additional results for Table 2 of the main paper. The models were trained on different vehicles and then evaluated on the ***test*** images of all unknown vehicles. We compare the accuracy (in percentage) for different models presented in this work. The classification models are ***fine-tuned*** (F) and AE stands for autoencoder with max-unpooling and AE-W without max-unpooling. The main paper report results for classification and autoencoders when trained from ***scratch*** (S), which are here repeated for ease of comparison.

	Trained on vehicle								
Model	A-Class	Escape	Hilux	Lexus	Tesla	Tiguan	Tucson	X5	Mean
VGG-16 (F)	61.3	62.6	54.1	58.1	61.8	48.5	66.3	58.3	58.9
DenseNet-121 (F)	59.2	58.1	60.9	67.4	64.5	49.2	68.4	52.0	60.0
MobileNet (F)	51.8	47.9	58.3	57.0	60.6	45.5	53.0	51.2	53.2
ResNet-50 (F)	49.1	49.7	52.5	49.1	57.4	50.0	60.0	49.2	52.1
ResNet-18 (F)	56.6	52.3	56.2	59.5	57.2	51.2	58.7	42.4	54.3
SqueezeNet (F)	55.0	51.1	54.5	56.4	50.2	53.4	51.4	45.7	52.2
VGG-16 (S)	51.9	49.3	49.7	60.6	50.6	44.6	45.9	52.0	50.6
DenseNet-121 (S)	42.4	49.4	47.6	48.2	47.3	35.5	46.6	40.0	44.6
MobileNet (S)	46.5	50.9	45.1	53.6	49.4	44.7	47.9	47.4	48.2
ResNet-50 (S)	48.3	47.1	41.3	48.5	51.9	35.7	44.7	43.1	45.1
ResNet-18 (S)	47.6	49.4	45.6	52.8	47.8	40.9	49.9	46.8	47.6
SqueezeNet (S)	52.4	48.8	48.4	51.2	46.1	43.4	50.6	39.2	47.5
AE - SSIM	54.8	52.0	49.1	58.2	55.9	42.1	50.4	53.3	52.0
AE - MS-SSIM	49.8	47.2	47.8	58.4	58.0	41.2	53.8	50.0	50.8
AE - Perceptual	52.5	56.0	45.2	61.2	54.3	39.1	46.1	52.9	50.9
AE - MSE	45.3	49.5	51.1	59.0	58.2	46.8	44.5	55.0	51.2
AE-W - SSIM	49.9	53.5	51.5	61.6	55.7	41.9	53.6	59.1	53.3
AE-W - MS-SSIM	51.6	53.7	48.5	57.6	55.5	43.2	52.1	52.0	51.8
AE-W - Perceptual	52.3	53.1	51.2	59.7	58.3	44.8	55.1	53.9	53.6
AE-W - MSE	55.0	52.9	49.9	60.6	58.2	46.2	47.6	52.9	52.9

Table S6: Additional results for Table 2 of the main paper. The models were trained on different vehicles and then evaluated on the ***training*** images of all unknown vehicles. We compare the accuracy (in percentage) for different models presented in this work. The classification models are trained from ***scratch*** (S) and ***fine-tuned*** (F) and AE stands for autoencoder with max-unpooling and AE-W without max-unpooling. The main paper compared performances on the test images.

	Trained on vehicle								
Model	A-Class	Escape	Hilux	Lexus	Tesla	Tiguan	Tucson	X5	Mean
VGG-16 (F)	86.0	90.7	86.8	92.2	94.0	75.4	97.0	83.5	88.2
DenseNet-121 (F)	85.5	86.6	94.0	93.9	88.8	75.8	93.6	81.2	87.4
MobileNet (F)	80.8	77.2	90.2	86.5	91.5	77.1	85.6	81.5	83.8
ResNet-50 (F)	82.9	78.6	80.7	80.4	88.0	74.1	89.4	81.4	81.9
ResNet-18 (F)	84.3	84.4	86.2	93.1	86.3	75.1	87.6	75.7	84.1
SqueezeNet (F)	84.2	77.3	83.0	84.3	83.3	76.9	80.7	74.3	80.5
VGG-16 (S)	76.3	80.5	82.4	89.5	78.3	66.9	78.5	82.5	79.4
DenseNet-121 (S)	65.6	76.2	70.6	77.4	76.2	53.1	76.0	64.5	70.0
MobileNet (S)	69.8	69.1	69.2	76.8	76.7	60.3	65.9	73.9	70.2
ResNet-50 (S)	70.9	71.3	63.2	74.0	80.1	51.6	68.8	72.6	69.1
ResNet-18 (S)	72.4	72.9	69.1	83.9	75.4	59.6	74.8	72.8	72.6
SqueezeNet (S)	78.5	72.9	68.3	76.3	69.6	60.6	72.7	64.2	70.4
AE - SSIM	76.2	70.5	72.0	82.6	83.5	59.7	73.5	80.0	74.8
AE - MS-SSIM	70.2	58.2	69.5	85.4	86.6	56.9	79.1	75.1	72.6
AE - Perceptual	78.6	77.5	66.1	86.3	84.7	55.9	69.7	77.8	74.6
AE - MSE	68.1	70.0	72.0	81.5	83.5	63.8	69.2	77.1	73.2
AE-W - SSIM	68.3	71.4	72.4	85.2	85.1	56.5	76.4	86.4	75.2
AE-W - MS-SSIM	71.9	73.5	69.4	81.6	81.4	58.4	77.2	76.8	73.8
AE-W - Perceptual	72.3	74.6	71.9	85.9	82.7	60.2	77.3	78.5	75.4
AE-W - MSE	75.1	71.5	68.8	86.7	82.3	59.4	68.7	76.0	73.6

4. Visual comparison of different autoencoder cost functions and reconstruction for different vehicles used during training

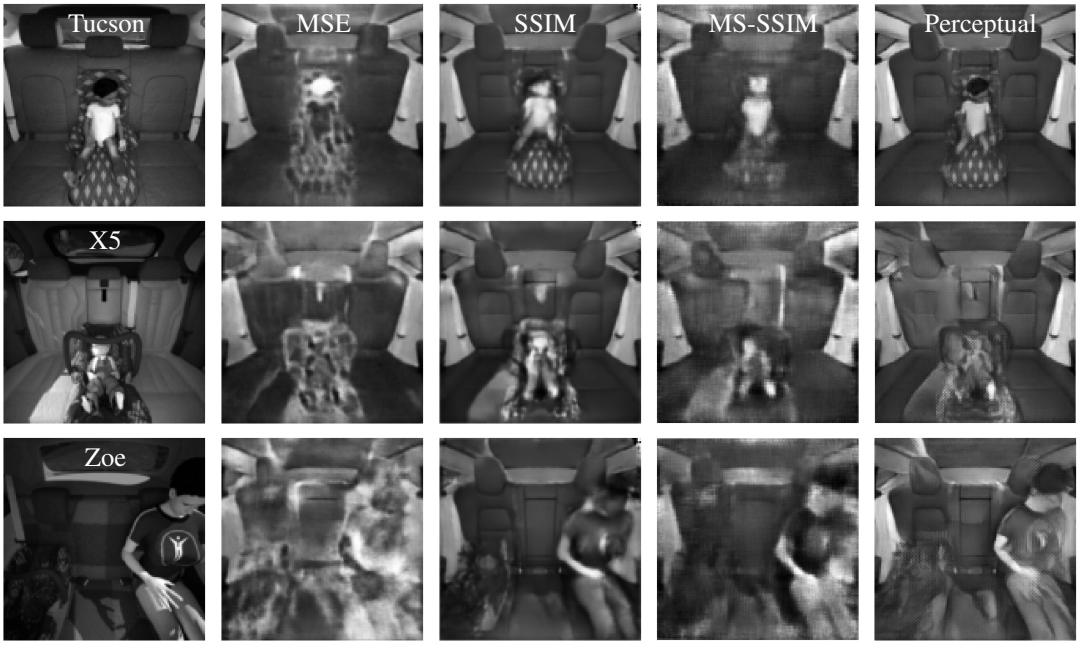


Figure S3: Results for the remaining vehicles for the autoencoder domain transformation presented in Fig. 6. The first column contains input training images from unknown vehicles of the SVIRO dataset. The other columns show the corresponding transformations by the autoencoder for different cost functions used: MSE, SSIM, MS-SSIM and perceptual loss.

Figure S4: Reconstruction examples for autoencoders trained using the nearest neighbour up-sampling instead of max-unpooling with indices. The first column contains input training images from unknown vehicles. The other columns show the corresponding transformations by the autoencoder for different cost functions used: MSE, SSIM, MS-SSIM and perceptual loss. The model was trained on the Tesla vehicle. The results are blurrier and less stable as for max-unpooling.

Figure S5: Reconstruction comparison for the same scenery when applied to autoencoders trained on different vehicles using the perceptual loss. Columns are the reconstructions for a same scenery by different autoencoders. Rows are the autoencoders trained on different vehicles.

Figure S6: Additional examples for Fig. S5 for the remaining vehicles. The last two columns are reconstructions for the i3 and Zoe on which we did not train.

5. Comparison of different autoencoder cost functions regarding the transferability to real infrared images

Figure S7: Comparison on real images of the effect of different reconstruction cost functions (MSE, SSIM, MS-SSIM and perceptual) used to train the autoencoder model. Each model was trained on the real X5 images and then evaluated on real Sharan images (first column). The model trained using the SSIM loss achieves the best details.

Figure S8: The same experiment as in Fig. S7, but here the models are trained on real Sharan images and evaluated to real X5 images (first column).

Figure S9: Comparison of the transferability to real images when different reconstruction cost functions (MSE, SSIM, MS-SSIM and perceptual) are used to train the autoencoder model. Each model was trained on the Tesla images and then evaluated on real Sharan images (first column). The model trained using the perceptual loss achieves by far the best transformations.

Figure S10: The same models as in Fig. S9 applied to real X5 images (first column).