Autoencoder for Synthetic to Real Generalization: From Simple to More Complex Scenes

Steve Dias Da Cruz*†‡, Bertram Taetz‡, Thomas Stifter*, Didier Stricker†‡
*IEE S.A.
†University of Kaiserslautern
‡German Research Center for Artificial Intelligence
Email: steve.dias-da-cruz@iee.lu, bertram.taetz@dfki.de, thomas.stifter@iee.lu, didier.stricker@dfki.de

***Abstract*—Learning on synthetic data and transferring the resulting properties to their real counterparts is an important challenge for reducing costs and increasing safety in machine learning. In this work, we focus on autoencoder architectures and aim at learning latent space representations that are invariant to inductive biases caused by the domain shift between simulated and real images showing the same scenario. We train on synthetic images only, present approaches to increase generalizability and improve the preservation of the semantics to real datasets of increasing visual complexity. We show that pre-trained feature extractors (e.g. VGG) can be sufficient for generalization on images of lower complexity, but additional improvements are required for visually more complex scenes. To this end, we demonstrate a new sampling technique, which matches semantically important parts of the image, while randomizing the other parts, leads to salient feature extraction and a neglection of unimportant parts. This helps the generalization to real data and we further show that our approach outperforms fine-tuned classification models.**

I. Introduction

The generation of synthetic data constitutes a cost efficient way for acquiring machine learning training data together with exact and free annotations. Notwithstanding this obvious advantage, bridging the gap between synthetic and real data remains an open challenge, in particular for camera based applications. Learning from synthetic data is an important tool in robotics: for example, to train a quadrupedal robot on synthetic data by incorporating proprioceptive feedback [1], to train a robot hand to solve real Rubik's cubes by learning the model in a simulation only [2] or by translating the real world input data into synthetic data for a reinforcement learning agent [3] and to *make the robot feel at home*. In view of safety critical applications, synthetic data can provide the means to reduce costs related to acquiring samples for edge cases, or which are difficult to obtain since they are too dangerous, e.g. accidents. We focus on learning invariances empirically on synthetic data, which should transfer to real data, opposed to constructing invariances as in equivariant neural networks [4].

We investigate the case of single independent images for which consistency between frames and physical interactions cannot be taken advantage of. The latter is commonly used by reinforcement learning methods [1]. We focus on training on synthetic data only and limit ourselves to autoencoder models which provide interesting properties due to their bottleneck design. The low-dimensional latent space of autoencoders can be subject to metric constraints [5], allows for scene decomposition [6] and it is believed that latent factor disentanglement can be useful for downstream tasks [7]. We assess to what extend we can generalize to real images and we highlight which design choices improve the autoencoder models performance with respect to accuracy and reconstruction quality. To this end, we first develop a method using features of pre-trained classifiers and show that we achieve better results on MPI3D [8] to generalize from synthetic (toy or realistic) to real images compared to Autoencoder, Variational Autoencoder (VAE) [9], β-VAE [10] and FactorVAE [11]. Although successful, we highlight that insights and design choices on a simple dataset do not necessarily transfer to real applications of higher visual complexity. To improve generalization, we propose to use the partially impossible reconstruction loss (PIRL) [12] (matching semantically important parts while randomizing the other parts) and we propose a novel variation thereof. We extensively show that our variation is the driving force for the improved generalization capacities. Additionally, we induce structure in the latent space by a triplet loss regularization. We evaluate and justify the benefits of the different design choices on an automotive application focusing on occupancy classification in the vehicle interior. The challenge of training in a single vehicle interior and transferring results between different vehicle interiors has been investigated [13]. The latter and similar industrial applications suffer from the limited availability and variability of training data. A successful transfer from synthetic to real data would avoid the necessity of collecting real data for each vehicle interior: the invariances could be learned and improved on synthetic data only.

II. Related Works

There have been successful applications of reinforcement learning systems being trained in a simulated environment and deployed to a real one, for example by combining real and synthetic data during training [14], [15], [16], [17]. However, these approaches can take into account temporal information and action-reaction causalities while in this work we use independent frames only. A good overview on reinforcement learning based simulation to real transferability is provided in [18]. Another line of research uses generative adversarial networks (GAN) to make synthetic images look like real images or vice versa [19], [20]. This requires both synthetic

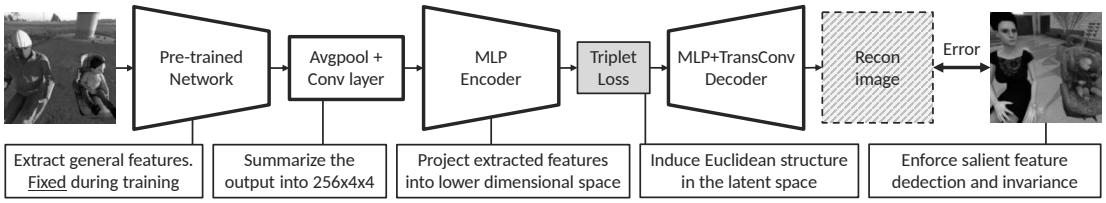


Fig. 1. Impossible Instance Extractor Triplet Autoencoder (II-E-TAE) model architecture.

and real images, whereas we focus on training on synthetic images only. Part of our methodology is related to domain randomization [21], where the environment is being randomized, but the authors deployed this to object detection and the resulting model needs to be fine-tuned on real data. A similar idea of freezing the layers of a pre-trained model was investigated for object detection [22], but neither with a dedicated sampling strategy nor in the context of autoencoders. Another work focuses on localization and training on synthetic images only [23], though the applicability is only tested on simple geometries. Although, we start our investigations on the simple dataset MPI3D, we increase the visual complexity by incorporating human models and child seats. Others rely on the use of real images during training for the minimization of the synthetic to real gap for autoencoders [24], [25]. Recent advances on synthetic to real image segmentation [26], [27], [28] on the VisDA [29] dataset show a promising direction to overcome the gap between synthetic and real images, however, this cannot straightforwardly be compared against the investigation in this work, particularly, since we are focusing on autoencoder models and their generative nature. While our cost function variation is based on a previous work [12], we show that our approach improves generalization while needing less demanding training data such that it can easily be applied to any commonly recorded classification dataset (i.e. no variations of the same scene are needed).

III. Method

Consider N_s sceneries and N_v variations of the same scenery, e.g. same scenery under different illuminations, with different backgrounds or under different data augmentation transformations. Let $\mathcal{X} = \{X_i^j \,|\, 1 \leq i \leq N_v, 1 \leq j \leq N_s\}$ denote the training data, where each $X_i^j \in \mathbb{R}^{C \times H \times W}$ is the ith variation of scene j consisting of C channels and being of height H and width W. Let $X^j = \{X_i^j \,|\, 1 \leq i \leq N_v\}$ be the set of all variations i of scenery j and $\mathcal{Y} = \{Y^j \,|\, 1 \leq j \leq N_s\}$ be the corresponding target classes of the scenes of $\mathcal{X}$. Notice that the classes remain constant for the variations i of each scene j. In the following, we will present the final model architecture as illustrated in Fig. 1 and we provide evidences for each design choice in Section IV.

A. Model Architecture: Extractor Autoencoder

By an abuse of terminology, we will refer to our method as a variation of vanilla autoencoders, although an encoder-decoder formulation would strictly speaking be more correct, because the goal will not be to reconstruct the input image exactly. We propose to apply ideas from transfer learning and use a pre-trained classification model to extract more general features from the input images. Instead of using the images itself, the extracted features are used as input. Our autoencoder consists of a summarization module which reduces the number of convolutional filters. This is fed to a simple MLP encoder which is then decoded by a transposed convolutional network. We refer to this model as *extractor autoencoder* (E-AE). Let e_ϕ be the encoder, d_θ the decoder and ext_ω be a pre-trained classification model, referred to as *extractor*. For ease of notation, we define $\mathrm{e}_\phi(\mathrm{ext}_\omega(\cdot)) = \mathrm{ee}_{\phi,\omega}(\cdot)$. The model, using the vanilla reconstruction loss, can be formulated as

$$\mathcal{L}_R(X_i^j;\theta,\phi) = \mathrm{r}\left(\mathrm{d}_\theta(\mathrm{e}_\phi(\mathrm{ext}_\omega(X_i^j))), X_i^j\right) \quad (1)$$

$$= \mathrm{r}\left(\mathrm{d}_\theta(\mathrm{ee}_{\phi,\omega}(X_i^j)), X_i^j\right), \quad (2)$$

where $\mathrm{r}(\cdot,\cdot)$ computes the error loss between target and reconstruction. We use the structural similarity index measure (SSIM) [30] and binary cross entropy (BCE). Model details are provided in the appendix Section S2-A.

B. Sampling Strategy: Partial Impossible

An additional improvement to the autoencoder training approach is a dedicated sampling strategy for which we provide two variations. The first one is the partially impossible reconstruction loss (PIRL) as introduced for illumination normalization [12]. As our results will show, this also helps the transfer between synthetic and real images. For sampling the individual elements of a batch, we randomly select for each scene two images, one as input and the other one as target. This sampling strategy preserves the semantics while varying the unimportant features such that the model needs to focus on what remains constant. For random $a, b \in [0, N_v]$ and $a \neq b$:

$$\mathcal{L}_{R,I}(X_a^j;\theta,\phi) = \mathrm{r}\left(\mathrm{d}_\theta(\mathrm{ee}_{\phi,\omega}(X_a^j)), X_b^j\right). \quad (3)$$

We refer to using the PIRL by prepending an *I*, e.g. I-E-AE.

C. Sampling Strategy: Partial Impossible Class Instance

We propose a novel variation to further improve this strategy by sampling a target image of a different scene, but of the same class. This should cause the model to learn invariances with respect to certain class variations which are not important for the task at hand, e.g. clothes, human poses, textures. This sampling variation is reflected in the reconstruction loss by

$$\mathcal{L}_{R,II}(X_a^j;\theta,\phi) = \mathrm{r}\left(\mathrm{d}_\theta(\mathrm{ee}_{\phi,\omega}(X_a^j)), X_b^k\right), \quad (4)$$

(a) Vanilla: Eq. (2) (b) I-PIRL: Eq. (3) (c) II-PIRL: Eq. (4)

Fig. 2. Different input-target pairs for the reconstruction loss.

for random $a, b \in [0, N_v]$, $j \neq k$ and $Y^j = Y^k$. We refer to this method as impossible class instance sampling marked by prepending *II*, e.g. II-E-AE. It is important to notice that our novel variation can easily be applied to any common dataset. The sampling variations are visualized in Fig. 2.

D. Structure in the Latent Space: Triplet Loss

The final adjustment to our training strategy is the incorporation of the triplet loss regularization in the latent space [5] to induce structure. This can be integrated by

$$\mathcal{L}_T(X_a^j;\phi) = \max\Big(0, \left\|\mathrm{ee}_{\phi,\omega}(X_a^j) - \mathrm{ee}_{\phi,\omega}(X_b^k)\right\|^2 - \left\|\mathrm{ee}_{\phi,\omega}(X_a^j) - \mathrm{ee}_{\phi,\omega}(X_c^l)\right\|^2 + 0.2\Big), \quad (5)$$

for random $a, b, c \in [0, N_v]$, $j \neq k \neq l$ and $Y^j = Y^k \neq Y^l$. We refer to this model as *triplet autoencoder* (TAE) either with or without using the PIRL. We can sample impossible target instances for the positive and negative triplet samples such that the total loss becomes (for some α and β):

$$\mathcal{L}(X_a^j;\theta,\phi) = \alpha\mathcal{L}_T(X_a^j;\phi) + \beta\left(\mathcal{L}_{R,II}(X_a^j;\theta,\phi) + \mathcal{L}_{R,II}(X_b^k;\theta,\phi) + \mathcal{L}_{R,II}(X_c^l;\theta,\phi)\right). \quad (6)$$

IV. Experiments

This section is organized in observations, formulated as subsections, which are built on one another and contain results highlighting the improvements. This provides explanations for the design choices leading to our final model architecture and cost function formulations presented in Section III. Improvements regarding the transfer to real images when only being trained on synthetic images are assessed qualitatively based on reconstruction quality and latent space structure and quantitatively on classification accuracy. All experiments use the same hyperparameters whenever possible. Training details are provided in the appendix and in our implementation (link).

We perform a baseline evaluation on MPI3D [8], which provides simple and realistic renderings and real counterparts. We reduced the dataset to contain only the large objects. For a higher visual complexity, we use as synthetic images the SVIRO [31] dataset. TICaM [32] is used to evaluate the performance on a real dataset of a similar application. The latter datasets are grayscale images from the vehicle interior and consider the task of classification (empty, infant, child or adult) for each seat position. The design choices made on MPI3D and the available synthetic images are not sufficient to obtain a good transferability to real images from the vehicle interior. Hence, we release an additional dataset, see Section IV-E and S1-D in the appendix. We introduce step by step modifications to the autoencoder architecture leading to steady quantitative and qualitative improvements. MPI3D and the vehicle interior share interesting properties: they have almost identical backgrounds and the environment is more tractable than many computer vision datasets. The transfer from SVIRO to TICaM is further complicated by new unseen attributes, e.g. steering wheel. An additional ablation study shows that our novel variation of PIRL is the driving force for the improved generalization capacity. Finally, to be in line with common benchmark datasets, we show that our design choices also improve the transfer from training on MNIST [33] to generalizing to real images of digits [34].

A. Autoencoders struggle on real images when trained on synthetic images

In the first, albeit naïve experiment we assumed that due to the bottleneck of autoencoders, the latter should generalize to some extent to real images when trained on synthetic ones. We trained convolutional autoencoders (AE) on the toy and realistic MPI3D images, respectively, and evaluated the resulting models on the real recordings. The first row of Fig. 3b shows the reconstruction of real images when trained on the realistic synthetic images: the model preserves some of the semantics. The model fails to perform senseful reconstructions when trained on toy images, see Fig. 3c.

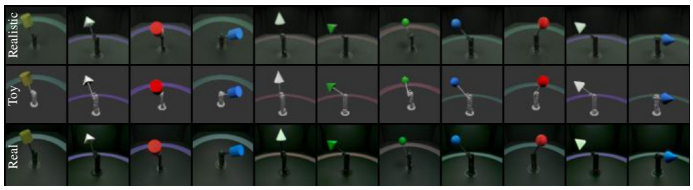
B. Autoencoders overfit to the synthetic distribution

A consequence of the results of the previous section is the assumption that the autoencoder overfits to the synthetic distribution and takes into consideration some artefacts (e.g. rendering noise). We followed the idea of the MPI3D authors [8] and trained Variational Autoencoder (VAE) [9], β-VAE [10] and FactorVAE [11] on the same data as before using the BCE reconstruction loss. The results in the second (β-VAE with $\beta = 8$) and third (FactorVAE with $\gamma = 50$) row of Fig. 3b show that the models reconstruct real images better and more of the semantics are preserved. If trained on toy renderings, the representation gap is too large, causing the reconstruction of the real images to be bad: see Fig. 3c.

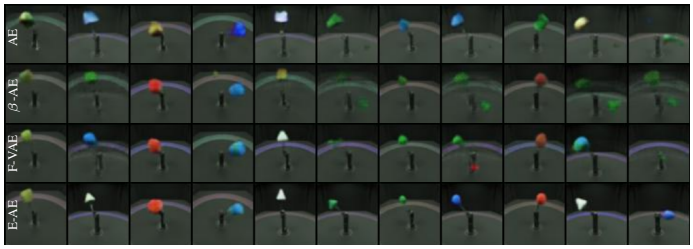
C. More general input features improve reconstructions

A small gap between the synthetic and real distribution can potentially be closed by a dedicated data augmentation approach to avoid overfitting to synthetic artefacts. Nevertheless, an abstraction from toy to real images cannot be achieved by means of simple data transformations or model constraints (e.g. denoising autoencoder). To this end we propose to use a pre-trained feature extractor as presented in Section III and as defined by Eq. (2). We used the VGG-11 model pre-trained on Imagenet as the extractor if not stated otherwise.

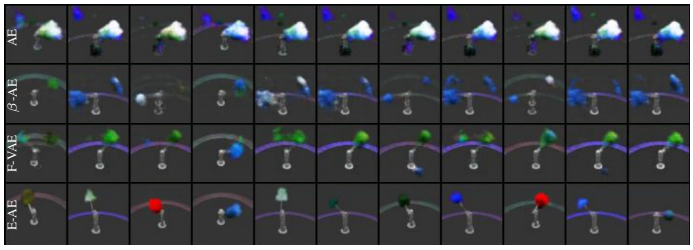
The results from the fourth row of Fig. 3b and Fig. 3c, respectively, show that the proposed modifications enable the model to generalize to real images when trained on synthetic ones. Much more of the semantics are preserved even when the model was only trained on toy images. Our method produces semantically more correct and less noisy



(a) Synthetic realistic and toy data used for training respectively, as well as real data used as input after training for evaluation.

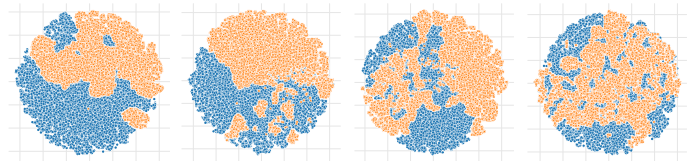


(b) Reconstruction of real data when being trained on realistic data.



(c) Reconstruction of real data when being trained on toy data.

Fig. 3. Reconstruction of unseen real data for different autoencoders: Autoencoder (AE), β Variational Autoencoder (β-VAE), FactorVAE (F-VAE), Extractor Autoencoder (E-AE). Our methods preserves the semantics best.



(a) AE (b) β-VAE (c) FactorVAE (d) E-AE

Fig. 4. t-SNE projection of the 10 dimensional latent space representation of the realistic training (blue circle) together with the real (orange cross) images. Autoencoder (AE), β Variational Autoencoder (β-VAE), FactorVAE and Extractor Autoencoder (E-AE). The extractor approach is the only method clustering both synthetic and real images together.

reconstructions compared to the VAE and FactorVAE baseline results. Additional qualitative improvements are highlighted by visualizing the latent space: both the 10-dimensional training (synthetic) and test (real) data latent spaces are projected together into a 2-dimensional representation using t-SNE. In Fig. 4 we can observe that VAE and FactorVAE improve the representation of real and synthetic images in the same region in the latent space, however, only partially, indicating a different representation for real and synthetic images. When using E-AE, real and synthetic images are represented more similarly in the latent space and the clusters are completely overlapping. Even when trained on the toy dataset, the latent

TABLE I

WE REPORT THE SSIM AND LPIPS [35] NORM BETWEEN THE RECONSTRUCTIONS OF THE REAL IMAGES (UNKNOWN) AND THE CORRESPONDING SYNTHETIC (SYNTH.) TRAINING IMAGES (REALISTIC (R) OR TOY (T)) OR INPUT IMAGES (REAL). WE REPORT THE MEAN OF THE NORMS ACROSS THE DATASET: FOR SSIM LARGER ↑ AND FOR LIPIPS SMALLER ↓ IS BETTER. E-AE PERFORMS BEST.

			SSIM ↑		LPIPS ↓	
	Model	Variant	Synth.	Real	Synth.	Real
T	AE	SSIM	0.56	0.42	0.35	0.40
T	VAE	BCE	0.50	0.33	0.34	0.42
T	β-VAE	BCE, $\beta = 4$	0.53	0.38	0.31	0.44
T	β-VAE	BCE, $\beta = 8$	0.71	0.48	0.26	0.37
T	FactorVAE	BCE, $\gamma = 10$	0.66	0.45	0.26	0.39
T	FactorVAE	BCE, $\gamma = 50$	0.71	0.51	0.22	0.35
T	E-AE (ours)	SSIM	0.90	0.58	0.10	0.2
R	AE	SSIM	0.83	0.62	0.20	0.24
R	VAE	BCE	0.74	0.61	0.20	0.23
R	β-VAE	BCE, $\beta = 4$	0.81	0.64	0.18	0.20
R	β-VAE	BCE, $\beta = 8$	0.79	0.64	0.19	0.21
R	FactorVAE	BCE, $\gamma = 10$	0.88	0.68	0.15	0.19
R	FactorVAE	BCE, $\gamma = 50$	0.78	0.64	0.16	0.18
R	E-AE (ours)	SSIM	0.92	0.70	0.08	0.14

space representation for synthetic and real images produced by E-AE overlaps partially as visualized in the appendix Fig. S2. Finally, we report in Table I a quantitative evaluation between the reconstructions of the real images against their synthetic training counterparts across all dataset images for different norms. We compute the same metrics between the real input images and their reconstruction to measure whether the semantics are being preserved : in all cases E-AE performs best. Additional results can be found in the appendix in Table S5 and reconstructions of synthetic input images in Fig S3. The latter shows that all models perform similarly well on the training data, hence the training was successful, but our proposed design choices generalize best to the real images.

D. It works for visually simple images - More is needed on more complex data

Since the method introduced in the previous section achieved good results, even when being trained on toy images, we wanted to apply it to images of higher visual complexity, e.g. a vehicle interior. We trained the same model architecture, but with a 64-dimensional latent space, on images from the Tesla vehicle from SVIRO and the Kodiaq vehicle from SVIRO-Illumination, respectively, and evaluated the model on the real TICaM images. Examples of the resulting model's reconstructions are plotted in Fig. 5 (b) and in the appendix Fig. S4. In both cases only blurry human models are reconstructed, which is similar to the mode collapse in the first row of Fig. 3c. We concluded that more robust features are needed.

E. PIRL helps generalization

As defined in Eq. (3), a partially impossible reconstruction loss (PIRL) for autoencoders has proven to work well for image normalization [12]. We hypothesized that the same approach could lead to a better generalization to real vehicle interiors. We applied this strategy to variations of the same

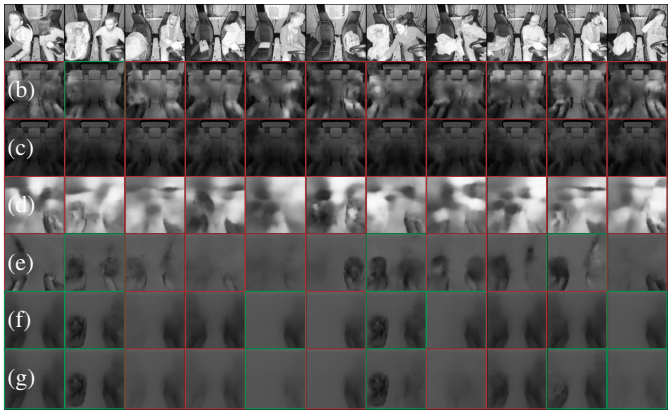


Fig. 5. Reconstructions of unseen real data (a) from TICaM: (b) E-AE and (c) I-E-AE trained on Kodiaq SVIRO-Illumination, (d) E-AE, (e) I-E-AE, (f) II-E-AE and (g) II-E-TAE trained on our new dataset. A red (wrong) or green (correct) box highlights whether the classes are preserved.

scene under different illumination conditions, but realized that the learned invariances are not suitable for the transfer between synthetic and real. An example is provided in Fig. 5 (c) where we trained on the Kodiaq images from SVIRO-Illumination.

We concluded that, for learning more general features by applying the PIRL, we needed input-target pairs where both images are of the same scene, but differ in the properties we want to become invariant to: the dominant background. To this end we created 5919 synthetic scenes where we placed humans, child and infant seats as if they would be sitting in a vehicle interior, but instead of a vehicle, the background was replaced by selecting randomly from a pool of available HDRI images. Each scene was rendered using 10 different backgrounds. Examples from the dataset are shown in Fig. S1 in the appendix. During training, we randomly select two images per scene and use one as input and the other as target, i.e. as defined in Eq. (3). When applied to real images, see Fig. 5 (e), the model better preserves the semantics of the real images: the model starts to reconstruct child seats and not people only, anymore. We also trained a model without the PIRL to show that the success is not due to the design choice of the dataset: in Fig. 5 (d) the model performs worse.

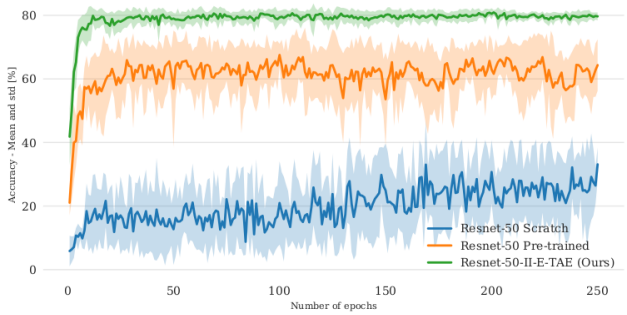
Finally, we extended this idea further with our novel PIRL loss variation: instead of taking the same scene with a different background as target image, we randomly selected a different scene of the same class, e.g. if a person is sitting at the left seat position, we take another image with a person on the left seat, potentially a different person with a different pose. This approach is formulated in Eq. (4). While this leads to a blurrier reconstruction, which is expected because the autoencoder needs to learn an average class representation, the classes are preserved more robustly and the reconstructions look better than before, see Fig. 5 (f). This additional randomization improves classification accuracy as discussed in Sections IV-G and V. A visualization of the different input-target pairs can be found in Fig. 2 and the dataset can been downloaded (link).

F. Structure in the latent space helps generalization

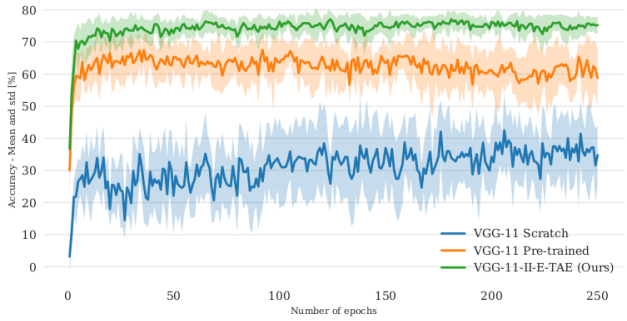
The final improvement is based on the assumption that structure in the latent space should help the model performance. Class labels are included by formulating a triplet loss regularization to the latent space representation as defined by Eq. (5): images of the same class should be mapped closely together and images of different classes pushed away. The triplet loss induces a more meaningful L^2-norm in the latent space [12] such that a k-nearest neighbour (KNN) classifier can be used in the next section. As the results of Fig. 5 (g) and in the appendix show, these final improvements, together with the previous changes, yield the semantically most correct reconstructions. In the appendix we show that due to the triplet loss the nearest neighbour of (g) makes sense and yields a clearer reconstruction. The triplet loss without the PIRL is not sufficient and in Section V we show that the II-PIRL loss is the driving force for the improved performance.

G. KNN with triplet loss out-performs classification models

We investigated whether the qualitative improvements also transfer to a quantitative improvement. We took the most basic approach: we combined the E-TAE with a k-nearest neighbour classifier in the latent space and used our new dataset for training. We retrieve the latent space vectors for all flipped training images as well and used only a single image per scene (i.e. not all 10 variations). We choose $k = \sqrt{N} = 115$, where N is the size of the training data together with its flipped version [36]. The model should classify occupancy (empty, infant, child or adult) for each seat position and we used the same hyperparameters for all methods and variations thereof. We froze the same layers of the pre-trained models for fine-tuning the later layers in case of classification models or to train our autoencoder using it as an extractor. We evaluated the model performance after each epoch on the real TICaM images (normal and flipped images of the training and test splits) for both the autoencoder and the corresponding classification model. This provides a measure on the best possible result for each method, but is of course not a valid approach for model selection. We report in Fig. 6 the training results for seeds 1 to 10 and summarize the training performance by plotting the mean and standard deviation per epoch per method. Our approach converges more robustly and consistently to a better mean accuracy. For each experiment, we retrieve the best accuracy across all epochs and compute the mean, standard deviation and maximum of these values across all runs: these statistics are reported in Table II. See the appendix for training from scratch and Densenet-121 results. The model weights corresponding to the epochs selected by the previous heuristics were applied on the SVIRO dataset to verify whether the learned representations are universally applicable to other vehicle interiors. For SVIRO, we used the training images and excluded all images containing empty child seats or empty infant seats, treated everyday objects as background. The results show that our E-AE significantly outperforms the classification models across three different pre-trained models and across all datasets. A consistent improvement for the



(a) Resnet-50



(b) VGG-11

Fig. 6. Training performance distribution for each epoch over 250 epochs. II-E-TAE is compared against training the corresponding extractor from scratch or fine-tuning the layers after the features used by the extractor.

TABLE II

FOR EACH EXPERIMENT, THE BEST ACCURACY ON REAL TICAM IMAGES ACROSS ALL EPOCHS IS TAKEN AND THE MEAN, STANDARD DEVIATION AND MAXIMUM OF THOSE VALUES ACROSS ALL 10 RUNS IS REPORTED. THE MODEL WEIGHTS ACHIEVING MAXIMUM PERFORMANCE PER RUN ON TICAM ARE EVALUATED ON SVIRO.

		TICaM		SVIRO	
Model	Variant	Mean	Max	Mean	Max
VGG	Pretrained	75.5 ± 1.5	78.0	78.7 ± 2.9	84.0
Resnet	Pretrained	78.1 ± 1.7	80.4	83.5 ± 2.7	88.1
VGG	E-TAE	76.7 ± 2.3	81.5	78.6 ± 2.6	82.3
Resnet	E-TAE	83.8 ± 1.3	86.0	85.8 ± 2.4	89.1
VGG	I-E-TAE	79.7 ± 2.1	82.2	80.9 ± 4.0	85.6
Resnet	I-E-TAE	83.5 ± 1.3	85.6	89.2 ± 1.0	90.3
VGG	II-E-TAE	81.0 ± 0.6	82.0	79.1 ± 3.9	84.8
Resnet	II-E-TAE	83.7 ± 0.5	84.5	93.0 ± 0.8	94.1

different modifications is achieved: I-E-TAE outperforms E-TAE and II-E-TAE outperforms I-E-TAE.

V. DISCUSSION AND LIMITATIONS

We want to highlight that most of the contribution to the success of our introduced model variations stems from the novel II variation of the PIRL loss. To this end we trained several types of classifiers in the latent space of different autoencoder model variations and report the results in Table III. The II variation of the PIRL loss largely improves the classification accuracy compared to the I variation. Moreover,

TABLE III

FOR EACH OF THE 10 RUNS PER METHOD AFTER 250 EPOCHS USING THE VGG-11 EXTRACTOR WE TRAINED DIFFERENT CLASSIFIERS IN THE LATENT SPACE: K-NEAREST NEIGHBOUR (KNN), RANDOM FOREST (RFOREST) AND SUPPORT VECTOR MACHINE WITH A LINEAR KERNEL (SVM). MOST OF THE CONTRIBUTION TO THE SYNTHETIC TO REAL GENERALIZATION IS DUE TO THE NOVEL II VARIATION OF THE PIRL.

Variant	KNN	RForest	SVM
E-AE	17.1 ± 6.7	24.2 ± 4.1	40.6 ± 8.5
I-E-AE	18.2 ± 7.3	42.4 ± 6.5	50.1 ± 3.7
II-E-AE	73.2 ± 3.9	68.8 ± 5.7	66.9 ± 6.7
E-TAE	69.2 ± 3.4	66.4 ± 4.0	68.7 ± 2.2

the performance is better compared to the triplet loss variation which uses the label information explicitly as a latent space constraints, compared to the implicit use by the II-PIRL.

The II variation of the PIRL loss implicitly assumes that the classes are uni-modal, i.e. objects of the same class should be mapped onto a similar point in the latent space. This characteristic can either improve generalization or have a detrimental effect on the performance depending on the task to be solved. Under its current form there is no guarantee that, for example, facial landmarks or poses would be preserved. Nevertheless, we believe that extensions of our proposed loss, for example based on constraints (e.g. preservation of poses) could be an interesting direction for future work. It can be observed that our model is not perfect and sometimes struggles: e.g. for more complex human poses (e.g. people turning over). However, we believe that these problems are related to the training data: a more versatile synthetic dataset would probably improve the model performance on more challenging real images.

Finally, we show that improvements reported in this work are not limited to the application in the vehicle interior. To this end, we trained models using the same design choices on MNIST [33] and evaluate the generalization onto real digits [34] in Fig. S6 and Table S7 in the appendix: similar improvements by the different design choices can be observed.

VI. CONCLUSION

We introduced an autoencoder model which uses a pre-trained classification model as a feature extractor. Our results showed that the resulting model produces superior reconstructions for synthetic to real generalization. However, we highlighted that design choices made on simple datasets do not necessarily transfer to visually more complex tasks. We performed a step-by-step investigation of additional model changes and showcased the improvements of each change. Although a k-nearest neighbour classifier is used in the latent space, our proposed autoencoder model outperforms consistently and more robustly all classification model counterparts.

ACKNOWLEDGMENT

The first author is supported by the Luxembourg National Research Fund (FNR) under grant number 13043281. The second author is supported by DECODE (grant number 01IW21001). This work was partially funded by the Luxembourg Ministry of the Economy (CVN 18/18/RED).

REFERENCES

[1] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, 2020.

[2] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas *et al.*, "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[3] J. Zhang, L. Tai, P. Yun, Y. Xiong, M. Liu, J. Boedecker, and W. Burgard, "Vr-goggles for robots: Real-to-sim domain adaptation for visual control," *IEEE Robotics and Automation Letters (RA-L)*, 2019.

[4] D. W. Romero and M. Hoogendoorn, "Co-attentive equivariant neural networks: Focusing equivariance on transformations co-occurring in data," in *International Conference on Learning Representations (ICLR)*, 2020.

[5] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*, 2015.

[6] M. Engelcke, A. R. Kosiorek, O. P. Jones, and I. Posner, "Genesis: Generative scene inference and sampling with object-centric latent representations," in *International Conference on Learning Representations (ICLR)*, 2020.

[7] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem, "Are disentangled representations helpful for abstract visual reasoning?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[8] M. W. Gondal, M. Wuthrich, D. Miladinovic, F. Locatello, M. Breidt, V. Volchkov, J. Akpo, O. Bachem, B. Schölkopf, and S. Bauer, "On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[10] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations (ICLR)*, 2017.

[11] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning (ICML)*, 2018.

[12] S. Dias Da Cruz, B. Taetz, T. Stifter, and D. Stricker, "Illumination normalization by partially impossible encoder-decoder cost function," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[13] S. Dias Da Cruz, B. Taetz, O. Wasenmüller, T. Stifter, and D. Stricker, "Autoencoder based inter-vehicle generalization for in-cabin occupant classification," in *IEEE Intelligent Vehicles Symposium (IV)*, 2021.

[14] K. Kang, S. Belkhale, G. Kahn, P. Abbeel, and S. Levine, "Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[15] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, "Rl-cyclegan: Reinforcement learning aware simulation-to-real," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] K. Fang, Y. Bai, S. Hinterstoisser, S. Savarese, and M. Kalakrishnan, "Multi-task domain adaptation for deep learning of instance grasping from simulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[17] A. Bewley, J. Rigley, Y. Liu, J. Hawke, R. Shen, V.-D. Lam, and A. Kendall, "Learning to drive from simulation without real world labels," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[18] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020.

[19] D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai, "Retinagan: An object-aware approach to sim-to-real transfer," *arXiv preprint arXiv:2011.03148*, 2020.

[20] A. Carlson, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "Sensor transfer: Learning optimal sensor effect image augmentation for sim-to-real domain adaptation," *IEEE Robotics and Automation Letters (RA-L)*, 2019.

[21] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[22] S. Hinterstoisser, V. Lepetit, P. Wohlhart, and K. Konolige, "On pre-trained image features and synthetic images for deep learning," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[23] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[24] T. Inoue, S. u. Choudhury, G. De Magistris, and S. Dasgupta, "Transfer learning from synthetic to real images using variational autoencoders for precise position detection," in *IEEE International Conference on Image Processing (ICIP)*, 2018.

[25] X. Zhang, Y. Fu, S. Jiang, L. Sigal, and G. Agam, "Learning from synthetic data using a stacked multichannel autoencoder," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2015.

[26] W. Chen, Z. Yu, Z. Wang, and A. Anandkumar, "Automated synthetic-to-real generalization," in *International Conference on Machine Learning (ICML*, 2020.

[27] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[28] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[29] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "Visda: The visual domain adaptation challenge," 2017.

[30] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *arXiv preprint arXiv:1807.02011*, 2018.

[31] S. Dias Da Cruz, O. Wasenmüller, H.-P. Beise, T. Stifter, and D. Stricker, "Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[32] J. S. Katrolia, B. Mirbach, A. El-Sherif, H. Feld, J. Rambach, and D. Stricker, "Ticam: A time-of-flight in-car cabin monitoring dataset," 2021.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[34] T. E. De Campos, B. R. Babu, M. Varma *et al.*, "Character recognition in natural images." *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, 2009.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[36] M. Jirina, M. Jirina, and K. Funatsu, "Classifiers based on inverted distances," in *New fundamental technologies in data mining*. InTech, 2011, vol. 1, pp. 369–387.