Illumination Normalization by Partially Impossible Encoder-Decoder Cost Function

Steve Dias Da Cruz [1,2,3]
steve.dias-da-cruz@iee.lu

Bertram Taetz [3]
bertram.taetz@dfki.de

Thomas Stifter [1]
thomas.stifter@iee.lu

Didier Stricker [2,3]
didier.stricker@dfki.de

[1] IEE S.A. [2] University of Kaiserslautern [3] German Research Center for Artificial Intelligence

Abstract

*Images recorded during the lifetime of computer vision based systems undergo a wide range of illumination and environmental conditions affecting the reliability of previously trained machine learning models. Image normalization is hence a valuable preprocessing component to enhance the models' robustness. To this end, we introduce a new strategy for the cost function formulation of encoder-decoder networks to average out all the unimportant information in the input images (*e.g. *environmental features and illumination changes) to focus on the reconstruction of the salient features (*e.g. *class instances). Our method exploits the availability of identical sceneries under different illumination and environmental conditions for which we formulate a partially impossible reconstruction target: the input image will not convey enough information to reconstruct the target in its entirety. Its applicability is assessed on three publicly available datasets. We combine the triplet loss as a regularizer in the latent space representation and a nearest neighbour search to improve the generalization to unseen illuminations and class instances. The importance of the aforementioned post-processing is highlighted on an automotive application. To this end, we release a synthetic dataset of sceneries from three different passenger compartments where each scenery is rendered under ten different illumination and environmental conditions: https://sviro.kl.dfki.de*

1. Introduction

Recording a sufficient amount of images to train and evaluate computer vision algorithms is usually a time consuming and expensive challenge. This is aggravated when the acquisition of images under various lightning and weather conditions needs to be considered as well. Notwithstanding the aforementioned data collection challenges, the performance of many machine learning algorithms suffer from changing illumination or environmental conditions, *e.g.* SLAM [10], place recognition [21], localization and classification [19], semantic segmentation [1], 3D human pose estimation [23] and facial expression recognition [24]. Since it is impracticable to wait for different weather conditions, day times and seasons to record images under as many variations as possible, it would be beneficial to train machine learning models to become invariant with respect to illumination and the exterior environment. Particularly for safety critical applications, as is common in the automotive industry, it would be of interest to reduce the amount of different illumination conditions necessary to guarantee reliable inference of machine learning models. Improvements on the aforementioned invariances would reduce the amount of mileage and images needed to be recorded and hence reduce the financial risk and time investment while improving the overall robustness of the deployed system.

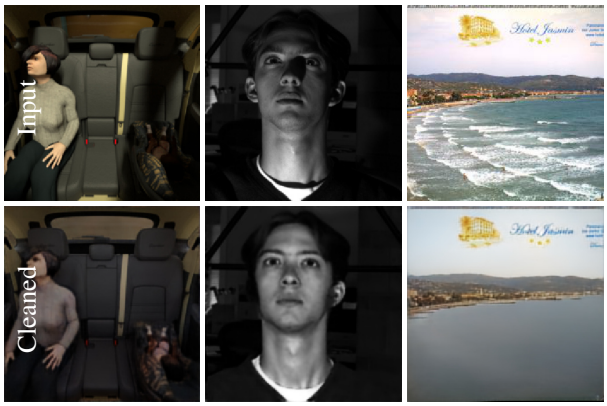


Figure 1: Results for the introduced partially impossible cost function. The input to the encoder-decoder model (first row) is transformed such that illumination and environmental features are averaged out and removed (second row).

We aim to transform the input image by removing illumination and environmental features instead of computing more robust and invariant feature descriptors like SIFT [17] or enforcing illumination invariance in deep neural networks through data augmentation. We achieve this by exploiting the availability of sceneries under different illumination and/or environmental conditions. We will introduce a partially impossible reconstruction loss in Section 3.1 which enforces similarity in the latent space of encoder-decoder models implicitly, in opposition to an explicit constraint [2, 31]. In contrast to shadow removal [29, 22] or relighting [28, 30], our method removes all the illumination and environmental features together. Our method is neither limited to a specific application where prior knowledge, *e.g*. about faces [30, 26], needs to be included, nor does it need shadow and shadow-free image pairs [29, 22] to define a ground truth target. We highlight its applicability on multiple datasets and provide evidence for the usefulness of collecting images under these more challenging conditions. Example results on multiple datasets are shown in Fig. 1.

In this work, we focus on the automotive application of occupant classification in the vehicle interior rear bench to demonstrate our proposed method's applicability. To this end, we release a synthetic dataset for occupant classification in three vehicle interiors where each scenery is rendered under ten different illumination and environmental conditions. We will demonstrate the benefits of combining an encoder-decoder based approach for illumination and environmental feature removal together with a triplet loss regularizer in the latent space. The latter improves the nearest neighbour search on test samples and hence the reliability and generalization to unseen samples. We quantitatively assess this improvement based on the classification accuracy. Our key contributions can be summarized as follows:

- We introduce a partially impossible reconstruction cost function in encoder-decoder models to remove illumination and environmental features,
- We highlight the importance of a triplet loss regularizer in the latent space of encoder-decoder models to improve generalization to unseen sceneries,
- We release the SVIRO-Illumination dataset, which contains 1500 sceneries (once with people only and once with child and infant seats) from three vehicle interiors, where each scene is rendered under 10 different illumination and environmental conditions.

2. Related Work

Datasets: Recording identical, or similar, sceneries under different lightning or environmental conditions is a challenging task. Large scale datasets for identical sceneries under different lightning conditions are currently scarce. The Deep Portrait Relighting Dataset [32] is based on the CelebA-HQ [14] dataset and contains human faces under different illumination conditions. However, the re-illumination has been added synthetically. Regarding the latter constraint, we instead used the Extended Yale Face Database B [8], which is a dataset of real human faces with real illumination changes. While cross-seasons correspondence datasets prepared according to [16] and based on RobotCar [18] and CMU Visual Localization [4] could potentially be used for our investigation, the correspondences are usually not exact enough to have an identical scene under different conditions. Moreover, dominantly visible changing vehicles on the streets induce a large difference in the images. Alternative datasets such as St. Lucia Multiple Times of Day [9] and Nordland [21] suffer from similar problems. However, these datasets stem from the image correspondence search, place recognition and SLAM community. We adopt the Webcam Clip Art [15] to include a dataset for the exterior environment with changing seasons and day times instead. The latter contains webcam images of outdoor regions from different places all over the world.

Consistency in latent space: Existing encoder-decoder based methods try to represent the information from multiple domains [2] or real-synthetic image-pairs [31] identically in the latent space by enforcing some similarity constraints, *e.g*. the latent vectors should be close together. However, these approaches often force networks to reconstruct some (or all) of the images correctly in the decoder part. Forcing an encoder-decoder to represent two images (*e.g*. same scenery, but different lightning) identically in the latent space, yet simultaneously forcing it to reconstruct both input images correctly implies an impossibility: The decoder cannot reconstruct two different images using the same latent space. Antelmi *et al*. [2] adopted a different encoder-decoder for each domain, but as illumination changes are continuous and not discrete, we cannot have a separate encoder or decoder for each possible illumination.

Shadow removal and relighting: Recent advances in portrait shadow manipulation [30] try to remove shadows cast by external objects and to soften shadows cast by the facial features of the subjects. While the proposed method can generalize to images taken in the wild, it has problems for detailed shadows and it assumes that shadows either belong to foreign or facial features. Most importantly, it assumes facial images as input and exploits the detection of facial landmarks and their symmetries to remove the shadows. Other shadow removal methods [29, 22] are limited to simpler images. The backgrounds and illumination are usually quite uniform and they contain a single connected shadow. Moreover, the availability of shadow and shadow-free image pairs provides the means of a well defined ground truth. However, this is is not possible for more complex scenes and illumination conditions for which a ground truth is not

available or even impossible to define. Image relighting [28, 32] could potentially be used to change the illumination of an image to some uniform illumination. However, as noted in [28, 30] relighting struggles with foreign or harsh shadows. While it is possible to fit a face to a reference image [26], this option is limited to facial images as well.

3. Method

We will introduce our proposed partially impossible cost function for encoder-decoder networks to exploit the availability of identical sceneries under different lightning conditions. We will suggest to extend our method by applying a triplet loss regularizer in the latent space to improve generalization. This induces some useful properties such that more robust and reliable results on unseen test samples can be achieved by adopting the nearest neighbour search.

3.1. Partially impossible reconstruction loss

Our proposed partially impossible reconstruction cost function can be applied to any encoder-decoder neural network architecture. Instead of considering the standard autoencoder reconstruction loss defined as the difference between the input image and the decoder reconstruction, we formulate an alternative reconstruction loss based on the decoder reconstruction and a new variation of the input image.

Let $\mathcal{X}$ be the set of all training images and x_k be the kth scene of the training data. For each scene we have n images, where each image represents the same scene under different lightning and/or environmental conditions. We denote by x_k^j the jth image out of the n images for scene k. Hence, the training data can be expressed as $x_k^j \in \mathcal{X}$ for $k \in [0, N]$ and $j \in [0, n]$, where N is the total number of unique scenes. Moreover, $x_k^j \in x_k$ for $j \in [0, n]$. Denote by $\mathcal{X}_m \subset \mathcal{X}$ a subset containing m number of sceneries from all the sceneries available in the training data. During training, the batches iterate over the x_k and for each x_k we randomly select $a, b \in [0, n], a \neq b$ to get $x_k^a, x_k^b \in x_k$. Finally, x_k^a is considered input to the encoder-decoder network and x_k^b is considered as the target for the reconstruction loss. The aforementioned method is illustrated in Fig. 2. The reconstruction loss can hence be formulated as

$$\mathcal{L}_R(\mathcal{X}_m; \theta, \phi) = \sum_{k=0}^{m} \mathrm{r}\left(\mathrm{h}_\theta(\mathrm{g}_\phi(x_k^a)), x_k^b\right), \qquad (1)$$

where g_ϕ is the encoder and h_θ the decoder. The reconstruction loss $\mathrm{r}(\cdot, \cdot)$ is computed between the reconstruction of the input image x_k^a and an image of the same scene under different environmental conditions x_k^b. In this work, we consider for the reconstruction loss the structural similarity index (SSIM) [5]: $\mathrm{r}(a, b) = 1 - \mathrm{SSIM}(a, b)$, but alternative image comparison functions can be considered as well.

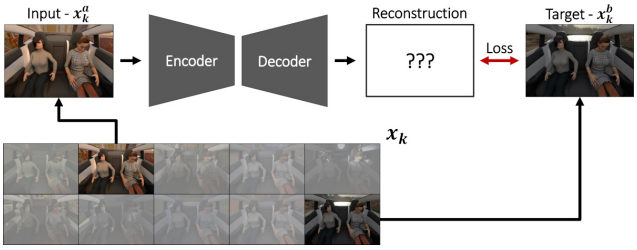


Figure 2: For each scenery x_k, we randomly select two images x_k^a and x_k^b under different lightning and environmental conditions. One is considered as input x_k^a to the encoder, the other one as target image x_k^b for the reconstruction.

Our cost function formulation implies a partially impossible task to solve. The input image x_k^a does not convey enough information to perfectly reconstruct the same scene under different environmental conditions x_k^b in its entirety. While x_k^a contains, usually, all the information of the objects in the scene, it does not contain any information about the illumination or environmental condition of x_k^b. However, both images are similar enough such that the encoder-decoder model can learn to focus on what is important, *i.e.* the salient features (*e.g.* people). Consequently, the only possibility for the neural network to minimize the loss is to focus on the objects in the scene which remain constant and neglect all the lightning and environment information, because the input images do not include information on how to handle it correctly. This implies that the neural network implicitly learns to focus the reconstruction on the people, objects and vehicle interior and to average out all the other information which changes between the similar scenes, *e.g.* the illumination and environment. This can be observed in Fig. 5 where we compare the reconstruction of similar sceneries after training: all background information and lightning conditions has either been removed or replaced by constant values. The encoder learns to remove the illumination information. The decoder is light invariant and cannot produce different illuminations, since the information has already been removed in the latent space representation.

Our proposed method is not limited to having the same scenery under different illumination conditions. One could use different augmentation transformations on the same input image x_k to form x_k^a and x_k^b and hence create the images on the fly. Alternatively, one could apply a *reverse* denoising approach where only x_k^b is augmented and x_k^a is the clean input image. See Fig. S1 in the supplementary material for an example for both approaches.

3.2. Triplet loss and nearest neighbour search

While the aforementioned method works well on the training data, generalizing to unseen test images remains a challenging task if no additional precautions are taken.

The illumination is still removed from test samples, but the reconstruction of the objects of interest can be less stable. As training data is limited, the encoder-decoder network is mostly used as a compression method instead of a generative model. Consequently, generalizing to unseen variations cannot trivially be achieved. Example of failures are plotted in Fig. 6 and Fig. 10: it can be observed, that the application on test images can cause blurry reconstructions. It turns out that the blurry reconstruction is in fact a blurry version of the reconstruction of its nearest neighbour in the latent space (or a combination of several nearest neighbours). An example of a comparison of the five nearest neighbours for several encoder-decoder models is shown in Fig. 9.

Consequently, instead of reconstructing the encoded test sample, it is more beneficial to reconstruct its nearest neighbour. However, applying nearest neighbour search in the latent space of a vanilla autoencoders (AE) or variational autoencoders (VAE) will not provide robust results. This is due to the fact that there is no guarantee that the learned latent space representation follows an L^2 metric [3]. As the nearest neighbour search is (usually) based on the L^2 norm, the latter will hence not always work reliably.

To this end, we incorporated a triplet loss [13] in the latent space of the encoder-decoder model (TAE) instead. Using the same notations, the triplet loss can be defined as

$$\mathcal{L}_T(\mathcal{X}_m;\phi) = \sum_{k=0}^{m} \max\Big(0, \big\|\mathrm{g}_\phi(x_k^{a,a}) - \mathrm{g}_\phi(x_{k_1}^{a,p})\big\|^2 - \big\|\mathrm{g}_\phi(x_k^{a,a}) - \mathrm{g}_\phi(x_{k_2}^{a,n})\big\|^2 + \alpha\Big), \tag{2}$$

where $x_k^{a,a}$ is the anchor using scenery k, $x_{k_1}^{a,p}$ is the positive sample using a different scenery k_1 and $x_{k_2}^{a,n}$ is the negative sample using another scenery k_2. An illustration of the nearest neighbour inference is given in Fig. 3 and for the triplet loss in Fig. S2. The triplet loss acts as a regularizer and due to its definition, it will also induce an L^2 norm in the latent space [20, 6, 3]. This effect is highlighted in Fig. 9, where we compare the nearest neighbours of the AE, VAE and TAE. To take full advantage of the triplet selection, we also modified the reconstruction loss (1) such that it is computed for each of the triplet samples:

$$\mathcal{L}_R(\mathcal{X}_m;\theta,\phi) = \sum_{k=0}^{m} \mathrm{r}\left(\mathrm{h}_\theta(\mathrm{g}_\phi(x_k^{a,a})), x_k^{b,a}\right) + \mathrm{r}\left(\mathrm{h}_\theta(\mathrm{g}_\phi(x_{k_1}^{a,p})), x_{k_1}^{b,p}\right) + \mathrm{r}\left(\mathrm{h}_\theta(\mathrm{g}_\phi(x_{k_2}^{a,n})), x_{k_2}^{b,n}\right), \tag{3}$$

where we take for each input image $x^{a,\cdot}$ a different random output image $x^{b,\cdot}$. Consequently, the total loss is defined as

$$\mathcal{L}(\mathcal{X}_m;\theta,\phi) = \mathcal{L}_R(\mathcal{X}_m;\theta,\phi) + \mathcal{L}_T(\mathcal{X}_m;\phi). \tag{4}$$

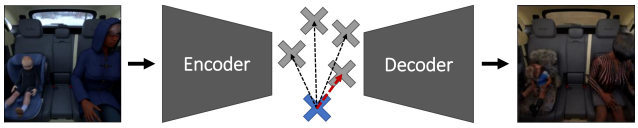


Figure 3: During inference, we choose the nearest neighbour (red arrow) of the latent space vector of the input image (blue cross) from all the training latent space vectors (gray crosses). This vector can be used to reconstruct a clean image or as classification prediction by using its label.

Figure 4: Example scenery from SVIRO-Illumination. The same scenery under eight (out of ten) different illumination and external environments. Left seat: adult passenger, middle seat:empty and right seat: infant seat with a baby.

4. Experiments

We will present an analysis of the aforementioned properties, problems and improvements on the SVIRO-Illumination dataset to highlight the benefit of our design choices. We will present results on two additional publicly available datasets to show the applicability of our proposed cost function to other problem formulations as well.

4.1. Training details

We center-cropped the images to the smallest image dimension and then resized it to a size of 224x224. We used a batch size of 16, trained our models for 1000 epochs and did not perform any data augmentation. We used the Adam optimizer and a learning rate of 0.0001. Image similarity between target image and reconstruction was computed using SSIM [5]. We used a latent space of dimension 16. The model architecture is detailed in Table S1 in the supplementary material: it uses the VGG-11 architecture [27] for the encoder part and reverses the layers together with nearest neighbour up-sampling for the decoder part. However, our proposed cost function is not limited to the model's architecture choice. We used PyTorch 1.6, torchvision 0.7 and pytorch-msssim 2.0 [11] for all our experiments.

4.2. SVIRO-Illumination

Based on the recently released SVIRO dataset [7], we created additional images for three new vehicle interiors. For each vehicle, we randomly generated 250 training and 250 test scenes where each scenery was rendered under 10

different illumination and environmental conditions. We created two versions: one containing only people and a second one including additionally occupied child and infant seats. We used 10 different exterior environments (HDR images rotated randomly around the vehicles), 14 (or 8) human models, 6 (or 4) children and 3 babies respectively for the training and test split. The four infant and two child seats have the same geometry for each split, but they use different textures. Consequently, the models need to generalize to new illumination conditions, humans and textures. There are four possible classes for each seat position (empty, infant seat, child seat and adult) leading to a total of $4^3 = 64$ classes for the whole image. Examples are shown in Fig. 4 and Fig. S3-S5 in the supplementary material.

4.2.1 Reconstruction results

For the triplet loss sampling, we chose the positive sample to be of the same class as the anchor image (but from a different scenery) and the negative sample to differ only on one seat (*i.e*. change only the class on a single seat w.r.t. the anchor image). Images of three empty seats do no contain any information which could mislead the network, so to make it more challenging, we did not use them as negative samples.

After training, the encoder-decoder model learned to remove all the illumination and environmental information from the training images. See Fig. 5 for an example on how images from the same scenery, but under different illumination, are transformed. Sometimes, test samples are not reconstructed reliably. However, due to the triplet constraint and nearest neighbour search, we can preserve the correct classes and reconstruct a clean image: see Fig. 6 for an example. The reconstruction of the test image latent vector produces a blurry person, which is usually a combination of several nearest neighbours. The reliability of the class preservations is investigated in Section 4.2.3 based on the classification accuracy. We want to emphasize that the model is not learning to focus the reconstruction to a single training image for each scenery. In Fig. 7 we searched for the closest and furthest (w.r.t. SSIM) input images of the selected scenery w.r.t to the reconstruction of the first input image. Moreover, we selected the reconstruction of all input images which is furthest away from the first one to get an idea about the variability of the reconstructions inside a single scenery. While the reconstructions are stable for all images of a scenery, it can be observed that the reconstructions are far from all training images. Hence, the model did not learn to focus the reconstruction to a single training sample, but instead learned to remove all the unimportant information from the input image. The shape and features of the salient objects are preserved as long as their position remains constant in each image, *e.g*. see Fig. 11 for vehicles being removed if not contained in each image. The texture of the salient objects is uniformly lit and smoothed out.

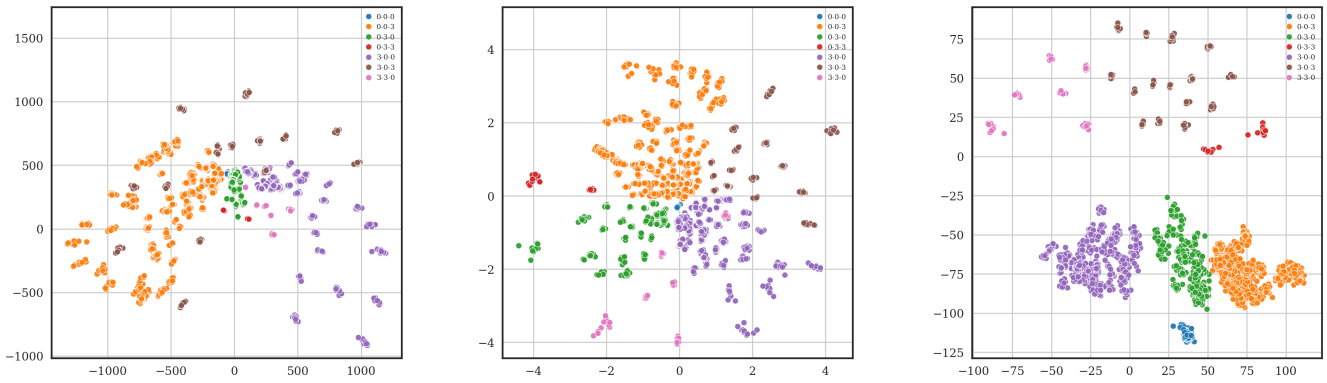
Figure 5: The encoder-decoder model transforms the input images of a same scenery (first row) into a cleaned version (second row) by removing all illumination and environment information (see the background through the window)

Figure 6: The test image (first row) cannot be reconstructed perfectly (second row). However, choosing the nearest neighbour in the latent space and reconstructing the latter leads to a class preserving reconstruction (third row).

Figure 7: The reconstruction of the first scenery input image (first recon) is compared against the furthest reconstruction of all scenery input images (max recon). First recon is also used to determine the closest and furthest scenery input images. The encoder-decoder model does not learn to focus the reconstruction to a training sample.

4.2.2 AE vs. VAE vs. TAE

For visualization purposes, we trained a vanilla autoencoder (AE), variational autoencoder (VAE) and triplet autoencoder (TAE) on the SVIRO-Illumination dataset with people and empty seats only. For simplicity of visualiza-



(a) Autoencoder (b) Variational Autoencoder (c) Triplet Autoencoder

Figure 8: Comparison of training data latent space distributions for different regularizers in the latent space of encoder-decoder models. Different colors represent different classes. For each seat position, we either have 0 (empty) or 3 (adult) such that an image is a composition of three labels, e.g. for 3-0-3 an adult is sitting left and right. Some classes are under-represented and some samples are clustered close together: those are identical sceneries under different lightning conditions.

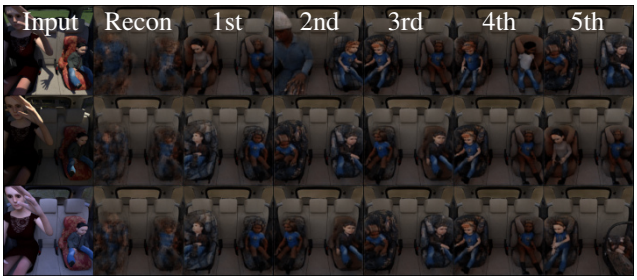
tion, we chose a latent space dimension of 2 for the model definition. After training, we computed the latent space representation for all training samples and plotted the resulting distributions in Fig. 8. The triplet based encoder-decoder model separates and clusters the classes best. Some small clusters are due to under-represented classes, for which the model clusters images from the same scenery under different illuminations together. The AE uses a large range of possible values in the latent space and both the AE and VAE contain wrong classes inside other clusters. The test distribution is plotted in Fig. S6 in the supplementary material and highlights the additional benefit of the TAE for potential outlier detection. Moreover, we show in Fig. S7 and Fig. S8 that a 2-dimensional principal component analysis and T-SNE projection of a 16-dimensional latent space provides even further benefits when a TAE is used. The same models were trained with a latent space dimension of 16 including occupied child and infant seats. The classification results obtained by nearest neighbour search are compared against several other models in Section 4.2.3. The TAE outperforms the other encoder-decoder models w.r.t. accuracy.

We needed to adjust the weight in the loss for the KL divergence (regularizer w.r.t. Gaussian prior) to $\beta = 0.001$ for training the VAE and prevent mode collapses. This is due to the background of the vehicle interior which is dominant in all training samples and remains similar.

It is important to note that the comparison between AE, VAE and TAE is not entirely fair, because the latter implicitly uses labels during the positive and negative sample selection. Nevertheless, for the problem formulations at hand, it is beneficial to collect the classification labels considering the additional advantage of the induced L^2 norm in the latent space and improved classification accuracy.

4.2.3 Classification results

We further compared the classification accuracy of our proposed method together with the nearest neighbour search against vanilla classification models when the same training data is being used. This way, we can quantitatively estimate the reliability of our proposed method against commonly used models. To this end, we trained baseline classification models (ResNet-50 [12], VGG-11 [27] and MobileNet V2 [25]) as pre-defined in torchvision on SVIRO-Illumination. For each epoch, we randomly selected one $x_k^j \in \mathcal{X}$ for each scenery x_k. The classification models were either trained for 1000 epochs or we performed early stopping with a 80:20 split on the training data. We further fine-tuned pre-trained models for 1000 epochs. The triplet based autoencoder model is being trained exactly as before. During inference, we take the label of the nearest training sample as the classification prediction. The random seeds of all libraries were fixed for all experiments and cuDNN was used in deterministic mode. Each setup was repeated 5 times with 5 different (but the same ones across all setups) seeds. Moreover, the experiments are repeated for all three vehicle interiors. The mean classification accuracy over all 5 runs together with the variance is reported in Table 1. Our proposed method significantly outperforms vanilla classification models trained from scratch and the models' performances undergo a much smaller variance. Moreover, our proposed method outperforms fine-tuned pre-trained classification models, despite the advantage of the pre-training of these models. Additionally, we trained the encoder-decoder models using the vanilla reconstruction error between input and reconstruction, but using the nearest neighbour search as a prediction. Again, including our proposed reconstruction loss improves the models' performance significantly.



(a) Autoencoder

(b) Variational Autoencoder

(c) Triplet Autoencoder

Figure 9: Comparison of the reconstruction of the 5 nearest neighbours (columns 3 to 7) for different encoder-decoder latent spaces (a), (b) and (c). The reconstruction (second column) of the test sample (first column) is also reported. The triplet regularization is by far the most reliable and consistent one across all 5 neighbours. Notice the class changes across neighbours for the AE and VAE models.

4.3. Extended Yale Face Database B

The Extended Yale Face Database B [8] contains images of 28 human subjects under 9 poses. For each pose and human subject, the same image is recorded under 64 illumination conditions. We considered the full-size image version instead of the cropped one and used 25 human subjects for training and 3 for the testing. We removed some of the extreme dark (no face visible) illumination conditions. Example images from the dataset are plotted in Fig. 10.

For the triplet sampling we chose as a positive sample an image with the same head pose and for the negative sample an image with a different head pose. We report qualitative results of a trained model in Fig. 10. The model is able to remove lightning and shadows from the training images, but the vanilla reconstruction on test samples can be blurry. We are not using the center cropped variant of the dataset, which makes the task more complicated, because the head is not necessarily at the same position for different human subjects. Nevertheless, the model is able to provide a nearest neighbour with a similar head pose and head position.

Table 1: Mean accuracy and variance over 5 repeated training runs on each of the three vehicle interiors. F = fine-tuned pre-trained model, ES=early stopping with 80:20 split, NS=no early stopping and V=vanilla reconstruction loss. Our proposed reconstruction loss improves the encoder-decoder vanilla version and with the nearest neighbour search outperforms all other models significantly.

	Vehicle		
Model	Cayenne	Kodiaq	Kona
MobileNet-ES	62.9 ± 3.1	71.8 ± 4.3	73.0 ± 0.8
VGG11-ES	64.4 ± 35	74.0 ± 19	75.5 ± 5.7
ResNet50-ES	72.3 ± 3.7	77.9 ± 35	76.6 ± 9.9
MobileNet-NS	72.7 ± 3.8	77.0 ± 4.1	77.4 ± 2.2
VGG11-NS	74.1 ± 5.8	71.2 ± 14	78.4 ± 2.6
ResNet50-NS	76.2 ± 18	83.1 ± 1.1	82.0 ± 3.2
MobileNet-F	85.8 ± 2.0	90.6 ± 1.2	88.6 ± 0.6
VGG11-F	90.5 ± 2.0	90.3 ± 1.2	89.2 ± 0.9
ResNet50-F	87.9 ± 2.0	89.7 ± 6.1	88.5 ± 1.0
AE-V	74.1 ± 0.7	80.1 ± 1.8	73.3 ± 0.9
VAE-V	73.4 ± 1.3	79.5 ± 0.6	73.0 ± 0.9
TAE-V	$\underline{90.8} \pm 0.3$	$\underline{91.7} \pm 0.2$	$\underline{89.9} \pm 0.6$
AE (ours)	86.8 ± 0.3	86.7 ± 1.5	86.7 ± 0.9
VAE (ours)	81.4 ± 0.5	86.6 ± 0.9	85.9 ± 0.8
TAE (ours)	$\underline{\mathbf{92.4}} \pm 1.5$	$\underline{\mathbf{93.5}} \pm 0.9$	$\underline{\mathbf{93.0}} \pm 0.3$

4.4. Webcam Clip Art

The Webcam Clip Art [15] dataset consists of images from 54 webcams from places all over the world. The images are recorded continuously such that a same scenery is available for different day times, seasons and weather conditions. For each of the 54 regions, we selected randomly 100 sceneries. Example images are provided in Fig. 11.

For the triplet sampling, we chose as positive sample an image from the same location and for the negative sample an image from a different location. Each landscape

Figure 10: Examples for Extended Yale Face Database B. All the illumination information is removed from the training samples (first row) to form the reconstruction (second row). The test samples (third row) cannot always be reconstructed reliably (fourth row). However, by reconstructing the nearest neighbour (fifth row) the head pose and position of the head can be preserved and the illumination removed.

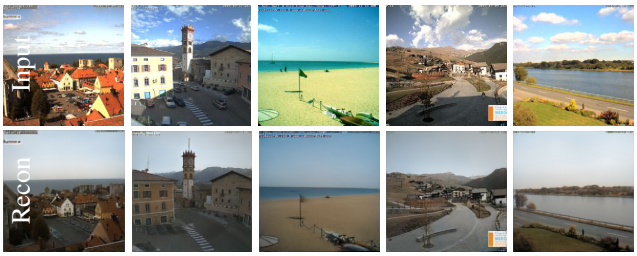


Figure 11: Examples for the Webcam Clip Art dataset. The encoder-decoder model removes the environmental features from the images (first row) to form the output images (second row). Vehicles and people are removed from the scenery and skies, rivers and beaches are smoothed out.

and building arrangement undergoes unique shadow, illumination and reflection properties. The generalization to unknown places under unknown illumination conditions is thus too demanding to be deduced from a single input image. Hence, we do not provide a test evaluation and report results on training samples only in Fig. 11. The model removes the illumination variations and shadows from the images. Moreover, rivers, oceans and skies as well as beaches are smoothed out. Most of the people and cars are removed and replaced by the actual background of the scenery.

5. Limitations

Our proposed method works well on the training data, which can be sufficient for some applications, *e.g.* when a fixed dataset is available on which some post-processing needs to be done only. Since the generalization to test images can be achieved by a nearest neighbour search, the latter will only be useful for a subset of machine learning tasks. Our method preserves the classes for a given problem formulation, which will be fine for classification and object detection. Although our method preserves even head poses (e.g. Fig. 10) when it is dominantly present in the training images, our approach will likely not preserve complex human poses (e.g. Fig. 6) or detailed facial landmarks, because the body poses and key features are not necessarily preserved by the nearest neighbour search. Future work should try to incorporate constraints such that the poses and landmarks of test samples are preserved as well.

In practise, it will be challenging to record identical sceneries under different lightning conditions. However, as the Extended Yale Face Database B [8] and Webcam Clip Art [15] dataset have shown, it is also feasible. Since we have highlighted the benefit of the acquisition of said datasets, the investment of recording under similar conditions in practise can be worth for some applications. We believe that future work will develop possibilities to facilitate the data acquisition process. Moreover, the possibility to incorporate images taken for the same scene, but in less perfect conditions, should be explored (e.g. Fig. S1).

6. Conclusion

Our results show the benefit of recording identical sceneries under different lightning and environmental conditions such that unwanted features can be remove by a partially impossible reconstruction loss function: without the need for a ground truth target image. Our method works well for classification and post-processing tasks due to an enhanced nearest neighbour search induced by a triplet loss regularization in the latent space of an encoder-decoder network. We demonstrated the universal applicability of our proposed method, as long as the correct data (i.e. same scenery under different conditions) is available, on three different tasks and datasets. Moreover, our proposed method improves classification accuracy significantly compared to standard encoder-decoder and classification models, even when the latter was a fine-tuned pre-trained model.

Acknowledgement

The first author is supported by the Luxembourg National Research Fund (FNR) under grant number 13043281. The second author is supported by VIDETE (grant number 01IW18002). This work was partially funded by the Luxembourg Ministry of the Economy (CVN 18/18/RED).

References

[1] Naif Alshammari, Samet Akcay, and Toby P Breckon. On the impact of illumination-invariant image pre-transformation for contemporary automotive semantic scene understanding. In *Intelligent Vehicles Symposium (IV)*, 2018.

[2] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In *International Conference on Machine Learning (PMLR)*, 2019.

[3] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] Hernán Badino, D Huber, and Takeo Kanade. Visual topometric localization. In *Intelligent Vehicles Symposium (IV)*, 2011.

[5] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

[6] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodolà. Limp: Learning latent shape representations with metric preservation priors. *arXiv preprint arXiv:2003.12283*, 2020.

[7] Steve Dias Da Cruz, Oliver Wasenmüller, Hans-Peter Beise, Thomas Stifter, and Didier Stricker. Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[8] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2001.

[9] Arren Glover, Will Maddern, Michael Milford, and Gordon Wyeth. FAB-MAP + RatSLAM: Appearance-based SLAM for Multiple Times of Day. In *Conference on Robotics and Automation (ICRA)*, 2010.

[10] Arren J Glover, William P Maddern, Michael J Milford, and Gordon F Wyeth. Fab-map+ ratslam: Appearance-based slam for multiple times of day. In *International Conference on Robotics and Automation (ICRA)*, 2010.

[11] Fang Gongfan. Pytorch ms-ssim. https://github.com/VainF/pytorch-msssim, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*, 2015.

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[15] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2009.

[16] Mans Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. A cross-season correspondence dataset for robust semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] David G Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV*, 1999.

[18] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 2017.

[19] Will Maddern, Alex Stewart, Colin McManus, Ben Upcroft, Winston Churchill, and Paul Newman. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *International Conference on Robotics and Automation (ICRA)*, 2014.

[20] Renqiang Min, David A Stanley, Zineng Yuan, Anthony Bonner, and Zhaolei Zhang. A deep non-linear feature mapping for large-margin knn classification. In *International Conference on Data Mining (ICDM)*, 2009.

[21] Daniel Olid, José M. Fácil, and Javier Civera. Single-view place recognition under seasonal changes. In *IROS Workshop on Planning, Perception, Navigation for Intelligent Vehicle (PPNIV)*, 2018.

[22] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Nadia Robertini, Florian Bernard, Weipeng Xu, and Christian Theobalt. Illumination-invariant robust multiview 3d human motion capture. In *Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[24] Ariel Ruiz-Garcia, Vasile Palade, Mark Elshaw, and Ibrahim Almakky. Deep learning for illumination invariant facial expression recognition. In *International Joint Conference on Neural Networks (IJCNN)*, 2018.

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)*, 2017.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 2019.

[29] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. Portrait shadow manipulation. In *ACM Transactions on Graphics (TOG)*, 2020.

[31] Xi Zhang, Yanwei Fu, Andi Zang, Leonid Sigal, and Gady Agam. Learning classifiers from synthetic data using a multichannel autoencoder. *arXiv preprint arXiv:1503.03163*, 2015.

[32] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *International Conference on Computer Vision (ICCV)*, 2019.