Deep Classification-driven Domain Adaptation for Cross-Modal Driver Behavior Recognition

Simon Reiß*, Alina Roitberg*, Monica Haurilet and Rainer Stiefelhagen
Karlsruhe Institute of Technology, Germany

*authors contributed equally to this work and are listed alphabetically

***Abstract*— We encounter a wide range of obstacles when integrating computer vision algorithms into applications inside the vehicle cabin, *e.g.* variations in illumination, sensor-type and -placement. Thus, designing domain-invariant representations is crucial for employing such models in practice. Still, the vast majority of driver activity recognition algorithms are developed under the assumption of a *static* domain, *i.e.* an identical distribution of training- and test data. In this work, we aim to bring driver monitoring to a setting, where domain shifts can occur at any time and explore generative models which learn a *shared representation* space of the source and target domain. First, we formulate the problem of unsupervised domain adaptation for driver activity recognition, where a model trained on labeled examples from the source domain (*i.e.* color images) is intended to adjust to a different target domain (*i.e.* infrared images) where only unlabeled data is available during training. To address this problem, we leverage current progress in image-to-image translation and adopt multiple strategies for learning a joint latent space of the source and target distribution and a mapping function to the domain of interest. As our long-term goal is a robust cross-domain classification, we enhance a Variational Auto-Encoder (VAE) for image translation with a classification-driven optimization strategy. Our model for classification-driven domain transfer leads to the best cross-domain recognition results and outperforms a conventional classification approach in color-to-infrared recognition by 13.75%.**

I. Introduction and Related Work

Driver behavior analysis opens doors to a more natural, convenient and safe human-vehicle interaction [1], [2]. Especially in automated driving systems, increased freedom leads to driver engagement in more complex and distractive activities [3]. Recognizing what humans are doing inside the vehicle cabin encourages new technologies that enhance the comfort *e.g.* by adjusting the light when the human is reading or lowering the music volume when making a phone call. As today no driving system has achieved full automation (SAE level 5, as in [4]), an even more vital safety application arises in the transition period of conditional or high automation (SAE levels 3 and 4), where drivers are still required to be alert and provide inputs in uncertain situations. When the driver is in the midst of a secondary activity, the vehicle can detect the distraction and preemptively signal the human regarding upcoming passages *e.g.* requiring to take over.

While the rise of deep learning has led to significant progress of driver action recognition in a *static* scenario, such models are notably bad in handling changes in sensor type or -placement [5]. In this work, we aim to develop models which are able to classify data from domains different from the ones seen during training and study driver activity recognition in the *cross-modal* setting. Our final goal is cross-modal knowledge transfer: given an existing model trained on annotated data from the *source* domain, we aim to adjust it to classify data examples from a different *target* domain where new annotations would be too costly to acquire (overview in Figure 1). To enable such a knowledge transfer, we leverage current progress in image-to-image translation and learn a joint latent space of both domains in an unsupervised manner.

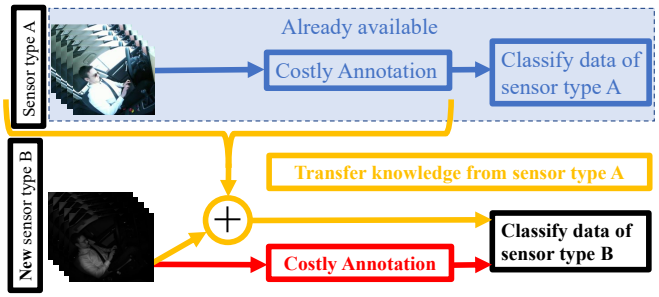

Fig. 1: In order to circumvent the expensive annotation process for a new sensor type, we explore the knowledge transfer from an existing sensor into the recognition pipeline.

Activity Recognition A widespread way to capture driver behavior is by observing the person through cameras inside the vehicle cabin [5], [1], [6], [7], [2]. Fueled by the recent advancements in computer vision, the field of activity recognition [8], [9] experienced tremendous improvements, but also inherited its challenges, *e.g.* sensitivity to domain shifts [5]. There are different strategies to deal with these problems, for example, varying illumination can be addressed by using different camera types, such as near-infrared or depth sensors [10]. While combining multiple cameras consistently leads to improvement in recognition results [10], [11], [5], introducing a novel sensor into the setup often requires costly data collection, annotation and model re-training for the new modality type. Could we skip the costly annotation of the new data and instead transfer the already existing knowledge to our domain? Such a *domain adaptation for cross-modal driver behavior recognition*, as described in Figure 1, is the key goal of this paper.

Image-to-Image Translation and Domain Adaptation This work is influenced by the image-to-image translation task (*i.e.* mapping an image from a source domain to a

different target space [12]), which experienced steep progress since the emergence of Generative Adversarial Networks (GANs) [13]. To this end, Zhu *et al*. introduced the concept of cycle-consistency [12], that entails the transfer back to the original representation employing a second GAN. At the same time, Liu *et al*. explored the idea of a shared-latent space, that aims to learn a joint representation of both distinct domains [14]. Image-to-image translation methods have been successfully applied for unsupervised domain adaptation in fields such as digit recognition, semantic segmentation and person re-identification [15], [16], [17], [18], [19].

We adopt and extend these image-to-image translation paradigms to handle domain changes inside the vehicle cabin, which, to our best knowledge, is explored for the first time in context of driver observation. We further present a novel **CL**a**S**sification-driven model for **UN**supervised **I**mage **T**ranslation *CLS-UNIT*. Our model is based on a Variational Auto-Encoder (VAE) for domain adaptation [14], which we enhance with an additional classification-driven loss influenced by a similar strategy employed successfully in previous semantic segmentation works [16]. To evaluate our idea, we explore two settings, in which the test data is captured by a sensor different to the one used during supervised training: classification of (1) *near-infrared* and (2) *depth* videos with annotated examples only available for *color* data. Our *CLS-UNIT* model consistently outperforms the baselines and other image-to-image translation approaches.

II. Deep Classification-driven Cross-Modal Translation for Driver Observation

We address the problem of *cross-modal driver activity recognition*, which aims at inferring the correct driver behavior from a different modality type than the one seen during training. Next, we define the task of unsupervised domain adaptation for driver behavior recognition (Section II-A) and describe our proposed strategy for leveraging generative image-to-image translation models on our task (Section II-B). Finally, we describe the modules of our proposed classification-driven architecture (Section II-C).

A. Cross-Modal Driver Activity Recognition

Conventional action recognition research aims at assigning an activity label $y \in Y$ to new input data $x \in X$ [8]. Thereby, both training and evaluation samples are generated by the same underlying probability distribution $x \sim p_{data}$. In *cross-modal action recognition*, on the other hand, test and training data are sampled from distinct probability distributions. Formally, our training set comprises labeled instances from the *source* domain: (x_s, y_s), with $x_s \in X_s$ and $y_s \in Y$, and unlabeled data from the *target* domain $x_t \in X_t$. Our goal is to classify each instance x_t^{test} in the target domain X_t from the test set.

In this work, we aim to develop models for *cross-modal driver activity recognition*. We explore two mapping functions: learning to transfer (1) from source to target $m_{s\to t} : X_s \to X_t$ and (2) from target to source $m_{t\to s} : X_t \to X_s$. After we learn these mapping-functions (see Section II-B and Section II-C) the prediction from the new domain for an instance x_t^{test} is computed as follows:

(1) The function $m_{s\to t}$ can be directly used on the labeled training data. That is, we translate the labeled source examples x_s into the target domain, which we use for training a classifier $c_t : X_t \to Y$ on $(m_{s\to t}(x_s), y_s)$.
(2) Another strategy is to exploit $m_{t\to s}$ to convert an instance x_t^{test} from the target domain of our test set into the source domain. More precisely, a classifier $c_s : X_s \to Y$ trained on (x_s, y_s) is subsequently used to yield the class-prediction for $m_{t\to s}(x_t^{test})$.

B. Neural Video Translation

In this section, we examine how to learn the mapping-functions for video frame transfer from the *source* domain (*e.g*. RGB) to the *target* domain (*e.g*. NIR) and vice versa. As we deal with an unsupervised setting, *i.e*. there are no labels for target domain, we also lack access to pairwise registered videos between the two modalities. We therefore leverage the concept of cycle-consistency, which allows us to learn the mapping without the available ground-truth pairs.

Generative Adversarial Networks (GANs) We model the mapping functions $m_{s\to t}$ and $m_{t\to s}$ with generator networks, which use convolution layers to translate the frames. Using a generator alone is a viable solution but has two drawbacks: (1) it is prone to learn a transfer to a single instance point (*e.g*. mapping to sepia in case of image colorization), and (2) it requires paired ground-truth data, which is impractical in many applications. To address this, we employ two discriminator networks (for each of our mapping functions) and design two GAN models [13]. The discriminators D_{X_s} and D_{X_t} are neural networks that learn to decide if the samples stem from the probability distribution of the source or target domain respectively or if they were produced by the generators. The architecture for source-to-target mapping of the images is trained by minimizing the L_{GAN} loss [12]:

$$\begin{aligned} L_{GAN}^{s\to t} = \mathbb{E}_{x_t\sim p_{data,t}}[\log D_{X_t}(x_t)] \\ +\mathbb{E}_{x_s\sim p_{data,s}}[\log(1 - D_{X_t}(m_{s\to t}(x_s)))]. \end{aligned} \quad (1)$$

The loss comprises: (1) a target-based loss, which penalizes the discriminator for not classifying data sampled from the target domain correctly; and (2) a loss that includes both the generator and the discriminator, in such a way that they oppose each other during training. While the generator produces data intending to fool the discriminator, the discriminator learns to distinguish between the synthesized and real instances. For the inverse direction (*i.e*. from target to source), the loss is computed by interchanging our two domains in Equation 1.

Cycle-consistency paradigm When using the loss from Equation 1 as-is, we do not enforce the generator to use the input map for fooling the discriminator. Thus, the model can fool the discriminator by producing previously unseen noise. To enforce the generator to keep relevant information in the translation process, we employ the cycle-consistency paradigm [12] and include the cyc-loss in our minimization:

$$L_{cyc}^{s\to t} = \mathbb{E}_{x_s\sim p_{data,s}}[\|m_{t\to s}(m_{s\to t}(x_s)) - x_s\|_1] \quad (2)$$

where $\|\cdot\|_1$ denotes the L_1 distance. This term encourages the network to retain information from the input image by encouraging the mapping to reproduce the original sample.

Semantic consistency loss We augment the cycle-consistency loss with additional semantic information extracted from our labeled *source* data, similarly to [16]. To this intent, we design a classifier $c : X_s \cup X_t \to Y$ for fusing the class-information into the training procedure:

$$\begin{aligned} L_{sem} = \mathbb{E}_{(x_s,y_s)\sim p_{data,s}}[CE(c(m_{s\to t}(x_s)), y_s)] \\ +\mathbb{E}_{x_t\sim p_{data,t}}[CE(c(m_{t\to s}(x_t)), \hat{y_t}(x_t))]), \end{aligned} \quad (3)$$

where $\hat{y_t}(x_t) = \text{argmax}(c(x_t))$ infers a label of a target instance using our classifier c. The cross-entropy loss denoted with $CE(\cdot,\cdot)$ is calculated with respect to the classification result of c on the instance mapped into the other domain.

Classification-driven loss Overall, the building blocks of the final loss, as employed in [16], cover the adversarial loss for realistic image reconstruction and a cycle-consistency loss to compensate for the lack of paired data. Moreover, a semantic consistency loss takes advantage of the labeled source training data and enforces similar classification scores of images before and after the translation. The final loss therefore not only aims to realistically map between the domains, but is also *classification-driven* as it computes the loss by summing the previously defined terms:

$$L_{CLS} = L_{GAN}^{s\to t} + L_{GAN}^{t\to s} + L_{cyc}^{s\to t} + L_{cyc}^{t\to s} + L_{sem}. \quad (4)$$

Shared-latent space models Instead of estimating the mapping functions directly, shared-latent space models [14] take a detour through an intermediate representation shared by both the source- and target domain. That is, the direct mapping function $m_{s\to t}$ is divided into a convolutional encoder $m_{s\to\ell}$ and a decoder network $m_{\ell\to t}$. This encoder-decoder setup condenses the input to a compact latent representation and is often implemented using Variational Auto-Encoders (VAEs) [14]. Thus, two VAEs underlie $m_{s\to\ell}$, $m_{\ell\to s}$ and $m_{t\to\ell}$, $m_{\ell\to t}$ respectively. To encourage the encoder networks $m_{s\to\ell}$ and $m_{t\to\ell}$ to have a *common representation space*, the parameters are shared throughout the later layers. An additional regularization constraint that is frequently applied to VAEs, restricts the output of the encoder to follow a standard normal distribution, *i.e.* $z \sim p_{st}(\cdot)$, where $p_{st}(z) = \mathcal{N}(z;\ 0, I)$. The outcomes are penalized when deviating from standard normal distribution via the KL-divergence between $p_{st}(\cdot)$ and the latent parameters (averages and deviations). The final latent representation of an image therefore encompasses sampling from this estimated distribution. The encoder models the distributions by estimating mean vectors of unit Gaussians $\mathcal{N}(z_s;\ m_{s\to\ell}, I)$ and $\mathcal{N}(z_t;\ m_{t\to\ell}, I)$, of which the outputs are used in the decoders. Finally, we encourage the latent representation to follow a standard distribution and the reconstructed image to resemble the input data as follows:

$$\begin{aligned} L_{VAE}^{s,\ell} = \lambda_1 KL(\mathcal{N}(z_s;\ m_{s\to\ell}(x_s), I)\ ||\ p_{st}(z)) - \\ \lambda_2 \mathbb{E}_{z_s\sim\mathcal{N}(\cdot;\ m_{s\to\ell}(x_s),I)}[\log p_{m_{\ell\to s}}(x_s;\ z_s)] \end{aligned} \quad (5)$$

where $KL(\cdot||\cdot)$ computes the Kullback-Leibler divergence between two probability distributions. We model $p_{m_{\ell\to s}}(x_s;\ z_s)$ as a Laplacian distribution which when minimizing its log-likelihood is equivalent to minimizing the absolute distance between the original image and its reconstruction using $m_{\ell\to s}$. A matching loss $L_{VAE}^{t,\ell}$ sets up the second VAE of the framework.

In addition to this, two GANs are employed to ensure that the decoder networks produce samples that fit into their assigned domains. This is established by augmenting our framework with additional discriminator networks for each domain, leading to a loss similar to (1).

To ensure that the mapping of an image to the other domain and back into the original domain results in the input image, a variant of the cycle-consistency paradigm is added as follows:

$$\begin{aligned} L_{vaecyc}^{s\to t\to s} = \lambda_3 KL(\mathcal{N}(z_s;\ m_{s\to\ell}(x_s), I)||p_{st}(z)) \\ +\lambda_3 KL(\mathcal{N}(z_t;\ m_{t\to\ell}(m_{s\to t}(x_s)), I)||p_{st}(z)) \\ -\lambda_4 \mathbb{E}_{z_t\sim\mathcal{N}(\cdot;\ m_{t\to\ell}(m_{s\to t}(x_s)),I)}[\log p_{m_{\ell\to s}}(x_s;\ z_t)], \end{aligned} \quad (6)$$

with a cross-domain mapping of $m_{s\to t}(x_s) = \mathbb{E}_{z_s\sim\mathcal{N}(\cdot;\ m_{s\to\ell}(x_s),I)}[m_{\ell\to t}(z_s)]$. The hyperparameters $\lambda_{\{1-4\}}$ provide measures to weight the different components of the losses L_{VAE}, L_{GAN} and L_{vaecyc} that are all optimized in both directions composing the loss for [14]. This shared-latent space framework is referred to as UNIT.

C. Classification-driven domain transfer learning with VAEs

While UNIT provides a meaningful mapping from the source to the target domain, it has issues with preserving the class information. In contrast, CyCADA [16] is able to preserve the semantic information during the mapping procedure, but encounters difficulties bridging between the source- and target domain. We aim to capture the advantages of both techniques and introduce a novel **CLaS**sification-driven model for **UN**supervised **I**mage **T**ranslation *CLS-UNIT* (overview in Figure 2). Our model enhances the VAE-based UNIT network for learning a shared-latent space with a *classification-driven loss*, which has previously been successfully used in semantic segmentation models [16]. Our ***CLS**-UNIT* loss is defined as:

$$\begin{aligned} L_{CLS-UNIT} = \lambda_{cls} L_{sem} + \lambda_{unit}(L_{GAN}^{s\to t} + L_{GAN}^{t\to s} \\ + L_{VAE}^{s,\ell} + L_{VAE}^{t,\ell} \\ + L_{vaecyc}^{s\to t\to s} + L_{vaecyc}^{t\to s\to t}) \end{aligned} \quad (7)$$

where λ_{unit} and λ_{cls} are parameters for weighting the losses, which we set empirically using the validation data to 0.6 and 0.4 respectively. In the *NIR-to-color* testbed, larger λ_{unit} causes the translation to be more colorful while resulting in

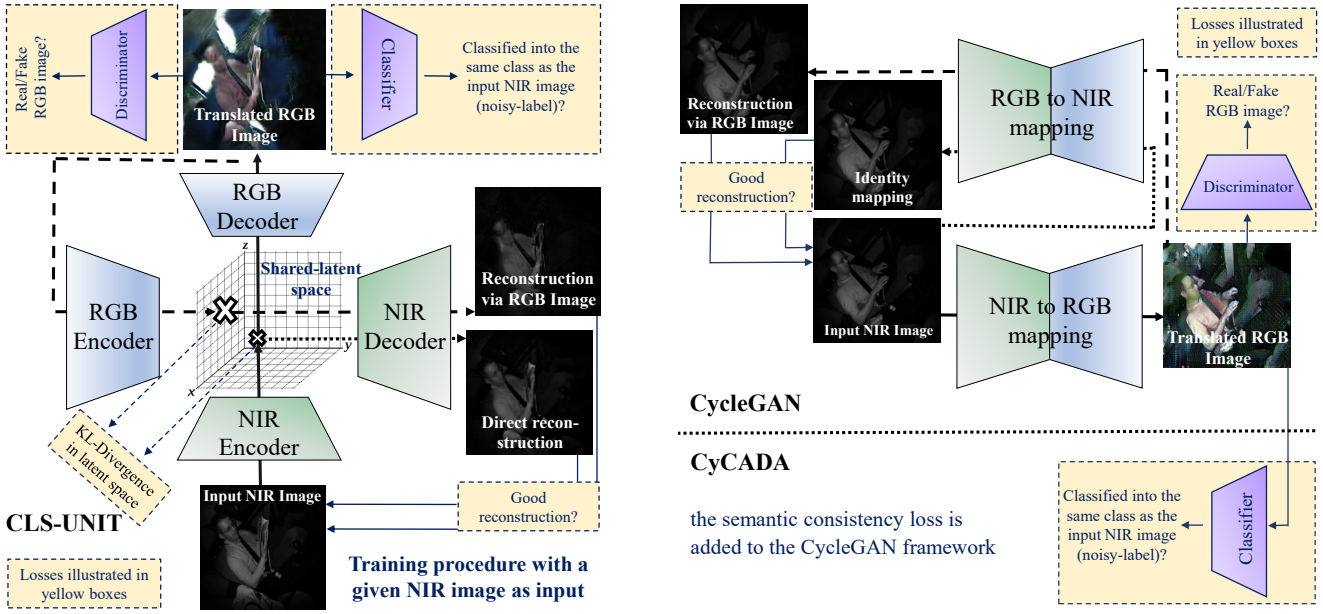

Fig. 2: Overview of our *CLS-UNIT* architecture (left) and other evaluated image-to-image translation models (CycleGAN and CyCADA) on the right. The main difference between CyCADA and CycleGAN is the semantic consistency loss. Our *CLS-UNIT* model extends the conventional UNIT model with the classification-driven loss in a similar way. The training procedure for learning a mapping function from the *target* to *source* domain is depicted for all models.

a blurry image. Choosing a higher value for λ_{cls} leads to the preservation of structure in the mapped images lowering the emphasis on faithful colorization.

III. Experiments

A. Implementation Details and Evaluation Setup

Dataset As no established evaluation procedure is available for driver monitoring in the cross-modal setting, we adapt the Drive&Act dataset [5] for standard driver activity recognition for our task. The dataset comprises color, NIR- and depth-videos of 15 drivers, which are densely annotated with 34 fine-grained activity labels. As previously described, our training data consists of: (1) *labeled* data in the *source* domain and (2) *unlabeled* recordings in the *target* domain. We select color videos as our source modality and both, NIR and depth as our target domains, resulting in two distinct experimental setups. For our training data, we therefore randomly select color data of 7 drivers with the corresponding activity annotations and unlabeled videos of 3 drivers in the target (*i.e.* NIR or depth) domain. To evaluate our model, we then use NIR and depth footage of the remaining 5 drivers for validation (2 subjects) and testing (3 subjects). As in [5], we divide the recordings in 3s chunks, compute the prediction for each chunk and then use balanced accuracy (*i.e.* average accuracy of each individual class) as our performance metric.

Video embedding scheme We embed the input videos using the I3D network pre-trained on Kinetics [8]. Depending on the mapping strategy (see Section II-A), we either fine-tune the model on the labeled source data (*i.e.* color videos) or on the frames translated in our target domain (*i.e.* NIR or depth).

The network operates on 16 frame clips of a size of 224×224 and is trained using SGD for 200 epochs. We sample 16 frames from the chunks to fit the I3D- and mapping network into memory at a reasonable minibatch size of 8. The training hyperparameters are adopted from [5], *i.e.* we use a learning rate of 0.05 decreased by 0.2 after 50, 100 and 150 epochs, a weight decay of $1e-7$ and a dropout probability of 0.5.

Semantic signal For determining the semantic consistency loss of our mapping network, we use an auxiliary ResNet pretrained on ImageNet [20]. The backpropagated signal flows through the parameters of the mapping network, encouraging it to preserve information about the action semantics. As the auxiliary classifier has not learned useful semantic information early in training, we only backpropagate the signal if its loss falls below a threshold θ (*i.e.* $\theta = 3.4$, just below $\log(\#classes)$, the loss of uniform classification).

Frame sampling scheme As the input to our mapping network is an image, we need a strategy for sampling from the video while training. Selecting the frame from Drive&Act uniformly is problematic, because the class distribution is highly unbalanced [5]. To tackle this, we perform class-wise sampling for the source domain data which draws frames of each class with the same probability. In case of the target domain data (*i.e.* NIR or depth), we draw instances uniformly over all frames as we do not have associated class labels.

Generator and Discriminator Architectures for the mapping networks in CycleGAN- and UNIT-based methods are adapted from [12] and [14]. They were trained for 20 epochs with $10K$ sampled images of size 256×256 per epoch. We use the initial learning rate of 0.0001 for the first 10 epochs,

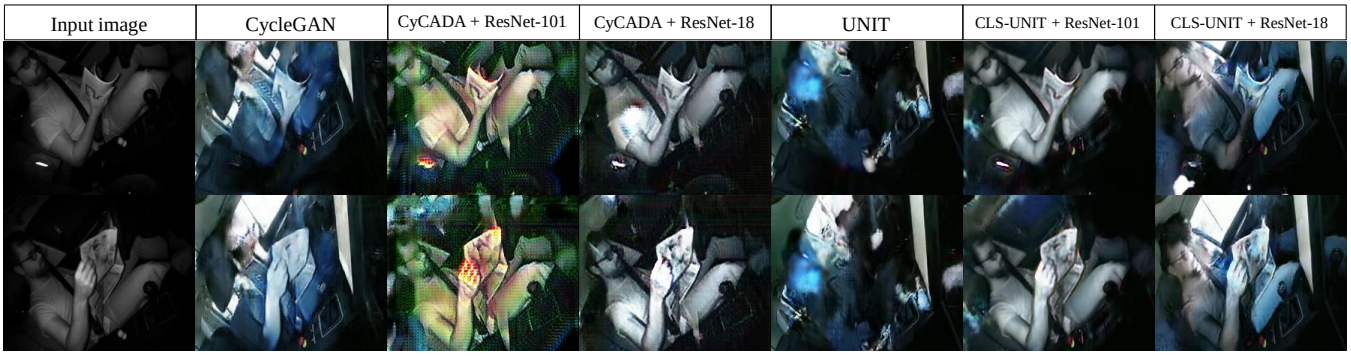

Fig. 3: Example translations of different models from NIR- to color images; the proposed method with a weighting scheme achieves meaningful colorization while preserving the structure and shapes in the input image.

Fig. 4: Image-to-image translation, left-to-right: source color image, ground truth depth image (for reference) and *color-to-depth* translation, using our *CLS-UNIT* model.

and linearly decay it afterwards. The weights are initialized using He-initialization [21]. The mapping network and the classification stream with the semantic consistency loss are optimized using Adam [22] with a weight decay of 0.0001.

B. Qualitative and Cross-Modal Recognition Results

Image-to-Image Translation Results While our main goal is to recognize human activities, we also showcase translation examples of the learned mapping functions. In Figure 3, we compare the *NIR-to-color* translation of previously described models. Most of the networks ignore fine structures, such as the hands of a person, in their translations, with the exception of our *CLS-UNIT* model, where such classification-relevant cues are preserved. The *NIR-to-color* mapping schemes in CycleGAN and UNIT that do not employ a classification signal, generate colorful images, however, at the expense of blurring the driver. The balance between retaining details relevant to classification (*e.g.* person- and object-related cues) and meaningful colorization is done best by our *CLS-UNIT* approach with an auxiliary ResNet-18 classifier. An example of the *color-to-depth* translation and the corresponding ground-truth depth map are visualized in Figure 4.

Cross-Modal Recognition Results We demonstrate the effectiveness of our model in Table I. Additionally to prominent image-to-image translation approaches [12], [16], [14], we compare our model to three baselines: (1) a random classifier, (2) the I3D network trained on the source data (*i.e.* color) classifying the data in target domain directly and (3) an I3D trained on source data transformed to grayscale

Translation Model	Direction	Classifier	Val	Test
Baseline Methods				
–	–	Random	3.03	2.94
–	–	Color-I3D	10.97	15.57
–	–	Grayscale-I3D	17.91	17.22
CycleGAN-based Networks				
CycleGAN	NIR→Color	Color-I3D	16.52	15.06
CyCADA + RN-101	NIR→Color	Color-I3D	14.91	12.01
CyCADA + RN-18	NIR→Color	Color-I3D	16.94	22.33
CyCADA + RN-18	Color→NIR	NIR-I3D	29.14	24.58
Shared-Latent Space Models				
UNIT	NIR→Color	Color-I3D	4.11	4.03
CLS-UNIT + RN-101 (ours)	NIR→Color	Color-I3D	14.06	18.35
CLS-UNIT + RN-18 (ours)	NIR→Color	Color-I3D	12.20	16.46
CLS-UNIT + RN-18 + λ (ours)	NIR→Color	Color-I3D	24.88	23.06
CLS-UNIT + RN-18 + λ (ours)	Color→NIR	NIR-I3D	**31.52**	**29.32**

TABLE I: Cross-modal activity recognition results with knowledge transfer from color to NIR. The *translation model* and the *direction* can be derived from the recognition procedure employed (details can be found in Section II-A). RN denotes the ResNet architecture and λ indicates models where λ_{cls} and λ_{unit} were tuned.

directly classifying the target domain data. The third baseline (grayscale) yields a fair *color-to-NIR* evaluation, as NIR data might seem similar to grayscale images with the naked eye.

The I3D model performs an accuracy of 67.76% in the conventional (*i.e.* color-to-color) setting, which may be seen as the upper bound for our cross-modal approach. I3D performance drops to only 15.57%, when applied in our cross-modal setting without any additional transfer, as CNNs per se are highly susceptible to domain shifts. Converting the training images to grayscale clearly helps, as they appear similar to the target IR frames (17.22%), but the recognition rate still remains low. When using CycleGAN-based methods and the mapping direction *NIR-to-color*, only the CyCADA model with an auxiliary ResNet-18 classifier outperforms the grayscale I3D baseline. The conventional UNIT framework

Translation Model	Classifier	Val	Test
Baseline Methods			
–	Random	3.03	2.94
–	Color-I3D	8.21	8.42
–	Grayscale-I3D	10.21	9.50
Shared-Latent Space Model			
CLS-UNIT + ResNet-18 (ours)	Depth-I3D	**17.23**	**17.57**

TABLE II: Cross-modal activity recognition results of our *color-to-depth* mapping, where we show that the transfer scheme increases performance significantly.

could not carry the relevant information for classification through the mapping functions at all and obtains an accuracy slightly better than random. However, our proposed extension of UNIT with a classification-driven loss (*CLS-UNIT*) heavily increases the performance, leading to the best recognition results when using weights λ_{cls} and λ_{unit}. As described in Section II-A we are flexible in choosing the mapping direction for classifying the unfamiliar modality. Utilizing a mapping from *color-to-NIR* we translate the labeled color videos and train a NIR classifier on top of them for our best CycleGAN-based and shared-latent space models. In addition to consistently producing better results (see *color-to-NIR* models in Table I) this scheme eliminates the necessity to compute the translation of incoming data in an online scenario as the classifier directly operates on the target domain. Overall, our model with the recognition rate of 31.52% on validation set and 29.32% on test set surpasses other translation models and baselines by a significant margin.

Our second experiment in the *color-to-depth* setting (Table II) shows consistent recognition results with a clear advantage of our approach. Using a *color-to-depth* transfer function with our *CLS-UNIT* model and then learning to classify the translated frames boosts the native I3D performance by 9.15% (test) and 9.02% (validation).

IV. Conclusion

When employing computer vision models in the uncontrolled driving environment, one quickly faces the problem of domain shifts. This paper addresses the task of cross-domain driver monitoring and has two-fold contributions. First, we formalize the problem of unsupervised cross-domain driver activity recognition and extend the popular Drive&Act testbed with our setting. To provide a challenging benchmark, we implement multiple off-the-shelf image translation models and conduct an extensive analysis of their cross-domain classification performance. Second, we introduce a novel approach for cross-modal activity recognition in context of driver observation. We leverage activity labels of the source domain training data and learn a shared-latent space of both modalities with a VAE-based model extended with an additional *classification-driven loss*. Enhancing the UNIT-VAE model training with the *classification-driven loss* encourages the network to learn a shared representation which reflects the semantic nature of the activity classes and which consistently leads to the best recognition results.

References

[1] P. Gebert, A. Roitberg, M. Haurilet, and R. Stiefelhagen, "End-to-end Prediction of Driver Intention using 3D Convolutional Neural Networks," in *Intelligent Vehicles Symposium (IV)*. Paris, France: IEEE, June 2019.

[2] E. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 90–104, 2016.

[3] F. Naujoks, C. Purucker, and A. Neukum, "Secondary task engagement and vehicle automation–comparing the effects of different automation levels in an on-road experiment," *Transportation research part F: traffic psychology and behaviour*, vol. 38, pp. 67–82, 2016.

[4] S. International, "Automated driving: levels of driving automation are defined in new sae international standard j3016," 2014.

[5] M. Martin*, A. Roitberg*, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles," in *ICCV*. IEEE, October 2019, *equal contribution.

[6] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *International Conference on Pattern Recognition*, 2014, pp. 660–665.

[7] S. Martin, E. Ohn-Bar, A. Tawari, and M. M. Trivedi, "Understanding head and hand activities and coordination in naturalistic driving videos," in *Intelligent Vehicles Symposium*, 2014.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017.

[9] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," no. i, 2016.

[10] A. Roitberg, T. Pollert, M. Haurilet, M. Martin, and R. Stiefelhagen, "Analysis of deep fusion strategies for multi-modal gesture recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[11] A. Roitberg, N. Somani, A. Perzylo, M. Rickert, and A. Knoll, "Multimodal human activity recognition for industrial manufacturing processes in robotic workcells," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[14] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.

[15] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016, pp. 469–477.

[16] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.

[17] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3722–3731.

[18] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4500–4509.

[19] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *CVPR*, 2018, pp. 994–1003.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] ——, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.