Auxiliary Task-Guided CycleGAN for Black-Box Model Domain Adaptation

Michael Essich, Markus Rehmann and Cristóbal Curio
Cognitive Systems Group, Reutlingen University, Germany
{michael.essich, markus.rehmann, cristobal.curio}@reutlingen-university.de

Abstract

The research area of domain adaptation investigates methods that enable the transfer of existing models across different domains, e.g.*, addressing environmental changes or the transfer from synthetic to real data. Especially unsupervised domain adaptation is beneficial because it does not require any labeled target domain data. Usually, existing methods are targeted at specific tasks and require access or even modifications to the source model and its parameters which is a major drawback when only a black-box model is available. Therefore, we propose a CycleGAN-based approach suitable for black-box source models to translate target domain data into the source domain on which the source model can operate. Inspired by multi-task learning, we extend CycleGAN with an additional auxiliary task that can be arbitrarily chosen to support the transfer of task-related information across domains without the need for having access to a differentiable source model or its parameters. In this work, we focus on the regression task of 2D human pose estimation and compare our results in four different domain adaptation settings to CycleGAN and RegDA, a state-of-the-art method for unsupervised domain adaptation for keypoint detection.*

1. Introduction

Deep learning has shown to be tremendously successful in complex tasks such as natural language processing [3], computer vision [37, 7], or content generation [28] and is a key technology for autonomous driving [10, 30]. However, if the training data on which such algorithms are trained diverges from the data they are supposed to operate on, which is commonly referred to as domain shift, performance degradation is to be expected. Thinking about the dynamic, diverse, and open world we are living in, it is clearly not possible to cover every possible scenario in a training dataset, showing the necessity to explicitly account for domain shifts. The active research area of domain adaptation (DA) works on methods to compensate for domain shifts and to improve model performance across domains.

DA can be attributed to transfer learning (TL), more specific to transductive TL [26], where only labeled data from the source domain is available and existing knowledge is intended to be transferred between the source and target domain under the assumption that the tasks do not differ. We can further distinguish between semi-supervised DA, when some labeled data is available in the target domain, and unsupervised domain adaptation (UDA), when there is no labeled data available in the target domain [8]. In this work, we follow a UDA approach, since UDA requires no time-consuming and expensive data labeling and therefore promises the greatest benefit. Moreover, while earlier shallow DA methods account for the domain shift by, *e.g.*, instance re-weighting [31] or simple feature augmentation [9], deep DA methods are considered more promising due to better DA performance [8, 29]. For this reason, we follow an unsupervised deep DA approach for cross-sensor adaptation based on CycleGAN [43], a generative adversarial network (GAN) [14] for unpaired image-to-image translation.

Existing DA methods are usually targeted and optimized for specific tasks or network architectures, *e.g.*, image segmentation [15, 42] or keypoint detection [41, 20, 11], and as a major downside require modifications and thus access to the model and its parameters for which domain adaptation is to be performed. There is only little work on DA of black-box models, *i.e.*, only having access to the non-differentiable predictions and having no access to the model parameters, and those methods are mainly targeted at image classification [40, 23]. We also assume a black-box model and, in contrast to existing work, keep our UDA approach more general and do not rely on a specific task or architecture. Moreover, we focus on black-box UDA for regression instead of classification and thus train and evaluate our approach on the challenging task of 2D human pose estimation. For this purpose, we created a motion capture-based dataset consisting of paired real and synthetic images, cf. Fig. 1. While our approach does not rely on or make use of paired data, it allows us to evaluate the model's performance with and without DA on scenes with the same content. We have the same human pose configurations across domains,

but the domain shift is either caused by different sensors, *i.e*., synthetic and real RGB images or synthetic RGB and synthetic depth images, or by variations in the person's appearance, *i.e*., clothing.

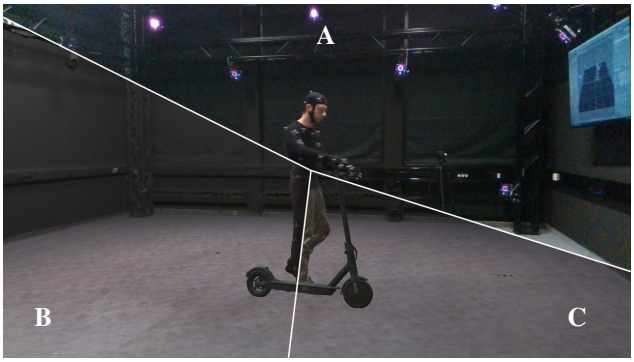

Figure 1: A sample frame of our paired dataset showing three different domains A) real RGB sensor data with motion capture suit, B) synthetic RGB sensor data with motion capture suit, C) synthetic RGB sensor data with casual clothing

We use the Transfer-Learning-Library developed by Jiang *et al*. [19], an open-source library that includes various TL methods and reference models for various tasks, and extend it with our proposed method to conduct our experiments.

Our contributions can be summarized as follows:

- We analyze the performance of CycleGAN [43] for unsupervised cross-sensor adaptation of a keypoint detection model, *i.e*., human pose estimation, across four different settings with varying domain shift.
- We show that unsupervised cross-sensor adaptation can be greatly improved by two simple modifications to CycleGAN, namely switching to a cyclical learning rate [32] and adding a task-related auxiliary loss inspired by multi-task learning [6, 27, 35, 22] and self-supervision [18], even under the assumption that we only have access to a black-box model but not to its parameters.
- We compare our method to a recent approach for unsupervised domain adaptation for keypoint detection (RegDA [20]) and conclude by emphasizing the necessity for explicitly addressing sensor domain shift.

2. Related Work

Human pose estimation is crucial for the safety of autonomous systems, *e.g*., in the area of autonomous driving or collaborative robots. While early deep learning-based methods such as DeepPose [34] directly regressed the 2D coordinates of human body joints in an image, recent fully convolutional approaches usually generate heatmaps, where joint positions are retrieved with a non-differentiable argmax operation [25, 39]. By replacing argmax with an integral operation, training can be performed in an end-to-end manner [33]. Latest 3D pose estimation methods are able to jointly predict 2D and 3D poses as well as head and body orientation from images [4]. Furthermore, we can distinguish between bottom-up and top-down human pose estimation. In the more commonly used top-down approaches, an additional object detection step is required to obtain bounding boxes of persons in an image. Pose estimation is performed afterward on every detected bounding box, as done in [25, 39]. In contrast to top-down approaches, bottom-up approaches simultaneously predict the poses of multiple persons in a single step. In the case of OpenPose [5], part affinity fields are predicted to associate joints with body parts and individuals.

There are already many different methods and architectures just for the task of human pose estimation that would need to be addressed in terms of domain adaptation. Therefore, the applicability of our UDA method is not targeted at a specific task or architecture, although in this work we focus on 2D human pose estimation and show the feasibility of our approach at the example of the pose estimation method proposed by Xiao *et al*. [39].

Unsupervised domain adaptation is the setting where labeled source but only unlabeled target domain data is available and the goal is to transfer a model trained on the source domain to the target domain. We focus on deep domain adaptation (DDA) which is commonly categorized into discrepancy-, adversarial- and reconstruction-based methods [8, 38].

Discrepancy-based methods aim to learn domain invariant features by reducing the discrepancy of intermediate network layers on the source and target domain, *e.g*. deep adaptation networks [24] use a multi-kernel variant of maximum mean discrepancy.

Adversarial-based methods use a discriminator, *i.e*., a domain classifier, that learns to classify data into source and target domain. The model is encouraged to learn domain invariant features through an adversarial goal, *i.e*., to fool the discriminator to misclassify the domains (domain confusion). It can further be distinguished between non-generative and generative methods [8, 38]. Ganin and Lempitsky [12] propose a non-generative adversarial-based method and introduce a gradient reversal layer combined with a domain classifier to learn domain-invariant features, which is also referred to as domain adversarial neural network (DANN) [13]. The loss for the main task, *i.e*., classification, is minimized using labeled source data, while the features are forced to be discriminative for the source and unlabeled target domain by maximizing the domain classi-

fication loss through a gradient reversal layer prepended to the domain classifier. In the generative case, an additional generator is added leading to a GAN [14] architecture. Such models can be used to, *e.g.*, generate domain-specific images from ground-truth semantic segmentation masks [17].

Reconstruction-based methods assume that reconstructing the source or target data is beneficial to encode domain-specific features. CycleGAN [43] for unpaired image-to-image translation introduces a cycle consistency loss that forces the network to transfer the image style while preserving the image content. The cycle consistency loss is the L1 loss between the original image and the reconstructed image, *i.e.*, after translation to the opposite domain and back again. CyCADA [15] extends CycleGAN with additional semantic consistency, task, and feature losses for DA of a semantic segmentation model.

These DA methods usually require access to the model and its parameters because they rely on (intermediate) activations or model weights. Although generative adversarial- and reconstruction-based approaches, in particular, can be used without having access to the (differentiable) model and its parameters, additional task-specific losses are often based on the model's output [42, 15]. In contrast, Zhang *et al*. [40] focus their work on UDA of black-box source models for image classification. They use the source model to predict noisy labels for the target domain, estimate the noise rate, select good samples and train a new model for the target domain. This procedure is repeated with the updated model for the target domain. While [40] requires predicted soft labels from the source model, DINE [23] can perform UDA of black-box models with hard labels.

We follow the idea of DA of black-box models but focus on a regression task, more specifically the 2D human pose estimation model proposed by Xiao *et al*. [39], in contrast to classification as done in [40, 23]. Our approach is close to [11], where DA for human pose estimation is performed by domain translation between synthetic and real depth data. However, we keep our method as general as possible and hence use a reconstruction-based method without making task-specific assumptions such that it can be easily adapted to other tasks. We compare our approach to RegDA [20], a UDA method for 2D keypoint detection recently proposed by Jiang *et al*. RegDA is an adversarial-based method that trains an adversarial regressor to maximize and a feature regressor to minimize disparity on the target domain. Thus, the feature regressor learns domain invariant features. Ground false poses, which are required for the adversarial training procedure of RegDA, are generated under the assumption that wrong keypoint predictions are most likely located at the position of other keypoints. In contrast to RegDA, our method does not require access to the source model and its parameters but only to its predictions and is, therefore, suitable for black-box DA.

3. Method

In this work, we follow the notations and definitions of Pan and Yang [26] and Csurka [8], *i.e.*, a domain is denoted by $\mathcal{D} = \{\mathcal{X}, P(X)\}$, where $\mathcal{X}$ is a feature space and $X = \{x_1, ..., x_n\} \in \mathcal{X}$. A task is denoted by $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$, where $\mathcal{Y}$ is a label space and $Y = \{y_1, ..., y_n\} \in \mathcal{Y}$.

Given unlabeled data X_s from source (s) domain $\mathcal{D}_s$ and unlabeled data X_t from target (t) domain $\mathcal{D}_t$ as well as the black-box model f_s trained on source domain data, our goal is to find a model $G_{t\to s}$ that maps X_t to X_s so that f_s can be successfully applied on translated data from domain $\mathcal{D}_t$, *i.e.*, $f_s(G_{t\to s}(x_t)) = \hat{y}$. Our method extends CycleGAN [43] with additional task specific auxiliary losses to further support domain translation and uses a cyclical learning rate. Details of our method are described in Subsection 3.1. Furthermore, we introduce a dataset specifically targeted to study sim-to-real DA in Subsection 3.2. Our simulation is, therefore, temporally synchronized to real RGB sensor and motion capture data. In addition, we calibrate the RGB sensor into our motion capture space which enables the generation of synchronized as well as paired real and synthetic sensor data based on motion-captured, 3D scanned, and animated actors.

3.1. Auxiliary Task-Guided CycleGAN

We follow [43] and define a least square GAN loss as well as a cycle and identity loss to learn a mapping between $\mathcal{D}_s$ and $\mathcal{D}_t$, cf. Eq. 1-3.

$$\begin{aligned}\mathcal{L}_{GAN}(G_{s\to t}, D_t^{GAN}, X_t, X_s) = \\ \mathbb{E}_{x_t\sim X_t}[(D_t^{GAN}(x_t) - 1)^2] \\ + \mathbb{E}_{x_s\sim X_s}[D_t^{GAN}(G_{s\to t}(x_s))^2]\end{aligned} \tag{1}$$

$\mathcal{L}_{GAN}(G_{t\to s}, D_s^{GAN}, X_s, X_t)$ is similarly applied for the opposite domain mapping.

$$\begin{aligned}\mathcal{L}_{cyc}(G_{s\to t}, G_{t\to s}, X_s, X_t) = \\ \mathbb{E}_{x_s\sim X_s}[||G_{t\to s}(G_{s\to t}(x_s)) - x_s||_1] \\ + \mathbb{E}_{x_t\sim X_t}[||G_{s\to t}(G_{t\to s}(x_t)) - x_t||_1]\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{id}(G_{s\to t}, G_{t\to s}, X_t, X_s) = \\ \mathbb{E}_{x_t\sim X_t}[||G_{s\to t}(x_t) - x_t||_1] \\ + \mathbb{E}_{x_s\sim X_s}[||G_{t\to s}(x_s)) - x_s||_1]\end{aligned} \tag{3}$$

While this mapping enables, *e.g.*, image-to-image translation and style-transfer, it is not designed for DA because it does not make use of any task-related information. To

support the transfer of task-related information under the assumption of only having access to a black-box model f_s, we define additional auxiliary losses inspired by multi-task learning [6, 27, 35, 22] and self-supervision [18], *i.e.*, by forcing our discriminators D_s and D_t to predict auxiliary tasks besides distinguishing between real and translated images as shown in Fig. 2. We refer to the discriminators' auxiliary task prediction heads as D_s^{aux} and D_t^{aux} and to the discrimination heads as D_s^{GAN} and D_t^{GAN}, respectively. Our auxiliary task is an image generation task and thus can be arbitrarily chosen, although it has to be task-related to support the transfer of relevant information. Except for the number of channels, the height and width of the output of D_s^{aux} and D_t^{aux} correspond to the input image, *i.e.*, we use a resolution of 256×256 pixels.

For the task of human pose estimation, we have chosen to transform the 2D joint positions predicted by the pose estimation model with $f_{aux}(f_s(X_s))$ into an auxiliary task by placing a filled circle at the position of each joint. All joints and an additional skeleton are stacked across the channel dimension, and Gaussian blur is applied, cf. Fig. 2. The auxiliary task encourages the backbones of the discriminators D_s^{aux} and D_s^{GAN} as well as D_t^{aux} and D_t^{GAN} to encode features that allow to distinguish real and generated images while being descriptive enough to generate the human pose related auxiliary task. The generators are forced to fool the discriminators; thus, they need to learn how the auxiliary task (human pose) relates to the images, *i.e.*, the persons in images, and how to generate them. The transfer of the auxiliary task across domains is further supported by D_{aux} through the task of distinguishing auxiliary tasks predicted from real images and ones predicted from generated images. Due to DA of a black-box model, we assume that the features learned by D_s^{aux} and D_t^{aux} are in line with the features the prediction of the human pose estimation model is based on. We refer to Section 4 for the results of our method. For performance reasons, we compute $f_{aux}(f_s(X_s))$ once for the whole dataset and cache the resulting auxiliary task data for reuse.

L_{aux} ensures that D_s^{aux} and D_t^{aux} jointly learn to predict the auxiliary task on source and target domain data, respectively, and is defined as follows:

$$\begin{aligned}\mathcal{L}_{aux}(f_s, f_{aux}, G_{s\to t}, G_{t\to s}, D_s^{aux}, D_t^{aux}, X_s, X_t) = \quad (4)\\ \mathbb{E}_{x_s\sim X_s}[(D_s^{aux}(x_s) - f_{aux}(f_s(x_s)))^2]\\ + \mathbb{E}_{x_s\sim X_s}[(D_t^{aux}(G_{s\to t}(x_s)) - f_{aux}(f_s(x_s)))^2]\\ + \mathbb{E}_{x_t\sim X_t}[(D_s^{aux}(G_{t\to s}(x_t)) - D_t^{aux}(x_t))^2]\\ + \mathbb{E}_{x_t\sim X_t}[(D_t^{aux}(G_{s\to t}(x_s)) - D_s^{aux}(x_s))^2]\end{aligned}$$

While the auxiliary task for the source domain is given by $f_{aux}(f_s(X_s))$, we assume that D_t^{aux} can be learned

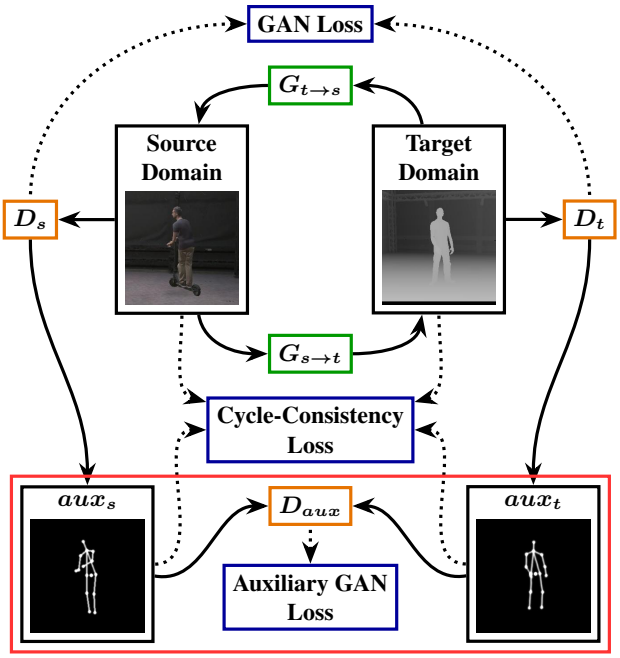

Figure 2: Our method extends CycleGAN with additional auxiliary GAN and cycle consistency losses (highlighted red box). The identity loss and the loss for training the auxiliary task based on a source model's output are not rendered here for simplicity.

leveraging the continuously improving $G_{s\to t}(X_s)$. To further support the transfer of the auxiliary task from source to target domain, we introduce the discriminator D_{aux} and define $\mathcal{L}_{auxGAN}$:

$$\begin{aligned}\mathcal{L}_{auxGAN}(G_{s\to t}, D_t^{aux}, D_{aux}, X_t, X_s) = \quad (5)\\ \mathbb{E}_{x_t\sim X_t}[(D_{aux}(D_t^{aux}(x_t)) - 1)^2]\\ + \mathbb{E}_{x_s\sim X_s}[D_{aux}(D_t^{aux}(G_{s\to t}(x_s)))^2]\end{aligned}$$

$\mathcal{L}_{auxGAN}(G_{t\to s}, D_s^{aux}, D_{aux}, X_s, X_t)$ is similarly applied. Finally, we enforce cycle consistency on the auxiliary task:

$$\begin{aligned}\mathcal{L}_{auxCyc}(G_{s\to t}, G_{t\to s}, D_s^{aux}, D_t^{aux}, X_s, X_t) = \quad (6)\\ \mathbb{E}_{x_s\sim X_s}[(D_s^{aux}(G_{t\to s}(G_{s\to t}(x_s))) - D_s^{aux}(x_s))^2]\\ + \mathbb{E}_{x_t\sim X_t}[(D_t^{aux}(G_{s\to t}(G_{t\to s}(x_t))) - D_t{}^{aux}(x_t))^2]\end{aligned}$$

Our combined loss function is as follows:

$$
\begin{aligned}
\mathcal{L}_{total}(&G_{s\to t}, G_{t\to s}, D_s^{GAN}, D_t^{GAN}, \\
&D_s^{aux}, D_t^{aux}, X_s, X_t, f_{aux}, f_s) = \\
&+ \mathcal{L}_{GAN}(G_{s\to t}, D_t^{GAN}, X_t, X_s) \\
&+ \mathcal{L}_{GAN}(G_{t\to s}, D_s^{GAN}, X_s, X_t) \\
&+ \lambda \cdot \mathcal{L}_{cyc}(G_{s\to t}, G_{t\to s}, X_s, X_t) \\
&+ 0.5\lambda \cdot \mathcal{L}_{id}(G_{s\to t}, G_{t\to s}, X_t, X_s) \\
&+ 10\lambda \cdot \mathcal{L}_{aux}(f_{aux}, f_s, G_{s\to t}, G_{t\to s}, \\
&\qquad D_s^{aux}, D_t^{aux}, X_t, X_s) \\
&+ \mathcal{L}_{auxGAN}(G_{s\to t}, D_t^{aux}, D_{aux}, X_t, X_s) \\
&+ \mathcal{L}_{auxGAN}(G_{t\to s}, D_s^{aux}, D_{aux}, X_s, X_t) \\
&+ 10\lambda \cdot \mathcal{L}_{auxCyc}(G_{s\to t}, G_{t\to s}, D_s^{aux}, D_t^{aux}, \\
&\qquad X_s, X_t)
\end{aligned} \tag{7}
$$

In Eq. 7 we follow [43] and set $\lambda = 10$. For $\mathcal{L}_{aux}$ and $\mathcal{L}_{auxCyc}$ a weighting factor of 10λ resulted in the best DA results in our experiments. Thus, our goal is to solve:

$$
{G^*}_{s\to t}, {G^*}_{t\to s} = \arg \min_{\substack{G_{s\to t} \\ G_{t\to s}}} \min_{\substack{D_s^{aux} \\ D_t^{aux}}} \max_{\substack{D_s^{GAN} \\ D_t^{GAN} \\ D_{aux}}} \mathcal{L}_{total} \tag{8}
$$

In contrast to CycleGAN, we do not use an image buffer and we switched to a cyclical learning rate and found that oscillating between a learning rate of $2 \cdot 10^{-6}$ and 10^{-4} leads to a higher DA performance and faster convergence compared to a linear decaying learning rate. We refer to Section 4 for our results.

3.2. Synchronized Sim-to-Real Dataset

Although we do not need paired cross-domain data for training, it is important for the evaluation of our DA method, especially in the setting of black-box model DA. Our human pose estimation model learns to predict the poses present in the source domain data and is thus biased towards the source domain's feature space. For the general case where we do not have paired cross-domain data for evaluation, we cannot compare source and target domain performance because we cannot attribute performance gains or drops to the domain adaptation method or to poses that are simply not known to the pose estimation model.

To overcome this, we created a dataset consisting of synchronized and paired synthetic and real samples recorded with an RGB sensor as shown in Fig. 1. Human3.6M [16] is close to our dataset, but we focus even more on producing paired data and therefore also created a 3D scanned virtual representation of the motion capture laboratory in addition to 3D scanned actors with the possibility of human-object interaction. To further vary the image content, we recorded our actor interacting with micromobility vehicles, *i.e.*, two different e-scooters. Our overall goal is having control over the domain shift.

For accurate human pose ground truth, we recorded the actor with a Vicon motion capture system. In addition, we calibrated an RGB camera into the motion capture space and used the acquired intrinsic and extrinsic camera parameters to project the human poses from motion capture into camera space. Our Unity-based simulation recreates the real recorded scene as accurately as possible. This includes the 3D scanned environment, the motion-captured actor and objects as well as the recorded camera based on its intrinsic and extrinsic parameters. Timecodes provided by the motion capture system are used in simulation to synchronize the synthetic data generation to the real data, resulting in paired synthetic and real data including ground truth, *i.e.*, human pose, bounding boxes, and additional synthetic depth data.

In total, our dataset contains 43,098 synthetic and real samples each. Despite the synthetic data with an actor wearing a motion capture suit (same as in the real data), we generated synthetic data with another 3D scanned actor wearing casual clothing to introduce an additional domain shift for our experiments.

4. Experiments

For comparison, all of our experiments are conducted with the Transfer-Learning-Library [19] that we have modified for our purposes and extended with our method.

We consider the following four DA settings, cf. Tab 1 and Fig. 3 for a visualized overview, with varying domain shift:

1. From synthetic RGB sensor data with casual clothing to synthetic RGB sensor data with motion capture suit (our dataset).

2. From synthetic RGB sensor data with casual clothing to synthetic DEPTH sensor data with casual clothing (our dataset).

3. From synthetic RGB sensor data with motion capture suit to real RGB sensor data with motion capture suit (our dataset).

4. From SURREAL [36] to LSP dataset [21].

The human pose estimation model is first trained on each of the source domains. We follow the training procedure of RegDA as implemented in the Transfer-Learning-Library which assumes that an epoch consists of 500 batches. We use a batch size of 8 for training the pose estimation model for 70 epochs on our dataset and for 500 epochs on the SURREAL dataset due to a larger dataset size and more

Table 1: Initial pose estimation accuracy. The models were trained on the source domain and the PCKh metric is reported for the source domain as well as the target domain before performing DA.

Setting No.	**Source domain**			**Target domain**		
	Dataset	PCKh@0.5	PCKh@0.1	Dataset	PCKh@0.5	PCKh@0.1
1	Our dataset synthetic RGB sensor casual clothing	99.50	62.60	Our dataset synthetic RGB sensor motion capture suit	30.95	1.81
2				Our dataset synthetic DEPTH sensor casual clothing	1.38	0.19
3	Our dataset synthetic RGB sensor motion capture suit	99.62	62.46	Our dataset real RGB sensor motion capture suit	9.26	0.60
4	SURREAL	60.22	0.12	LSP	37.12	5.64

(a) Setting 1 (b) Setting 2 (c) Setting 3 (d) Setting 4

Figure 3: Overview of the four DA settings considered in this work with visualized pose ground truth. Top row: source domains; bottom row: corresponding target domains

variations in the case of the latter. The model of the final training epoch is selected to conduct our DA experiments. We use 60% of our dataset for training (25,858 samples) and 10% for validation (4,309 samples). In the case of the SURREAL dataset (source domain), we use 451,439 samples for training and 3,200 samples for validation. Due to the small size of only 2,000 samples of the LSP dataset (target domain), the whole dataset is used for training and validation.

We report the PCKh [2] metric of each pose estimation model on its corresponding source domain as well as the raw target domain performance before DA in Tab. 1. As expected, the performance on target domain data decreases with increasing domain shift. A larger performance degradation can be noticed when we switch synthetic sensor modalities (from 99.5% to 1.38%) or from a synthetic to a real sensor (from 99.62% to 9.62%) than when we only vary the actor's appearance in the synthetic domain (from 99.5% to 30.95%).

The overall source domain performances of the pose estimation models are higher on our dataset compared to SURREAL while the relative performance degradation on the target domain is smaller in the case of the SURREAL source model applied on the LSP dataset (from 60.22% to 37.12%). In-domain variations are kept low in our dataset, *e.g.*, the actor's appearance does not change inside the same domain, but there are inter-domain variations, *e.g.*, a change in the actor's appearance or in the sensors between source and target domain. This makes it easier for the human pose estimation model to learn the poses, leading to a high source domain performance, with the drawback of overfitting to the

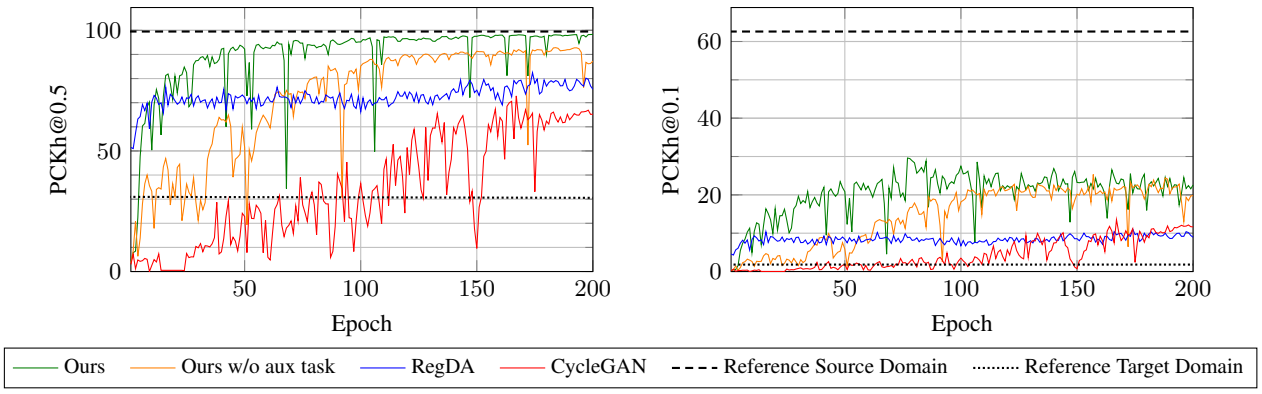

Figure 4: Setting 1: PCKh accuracy when performing domain adaptation on our dataset from synthetic RGB sensor data with casual clothing to synthetic RGB sensor data with motion capture suit.

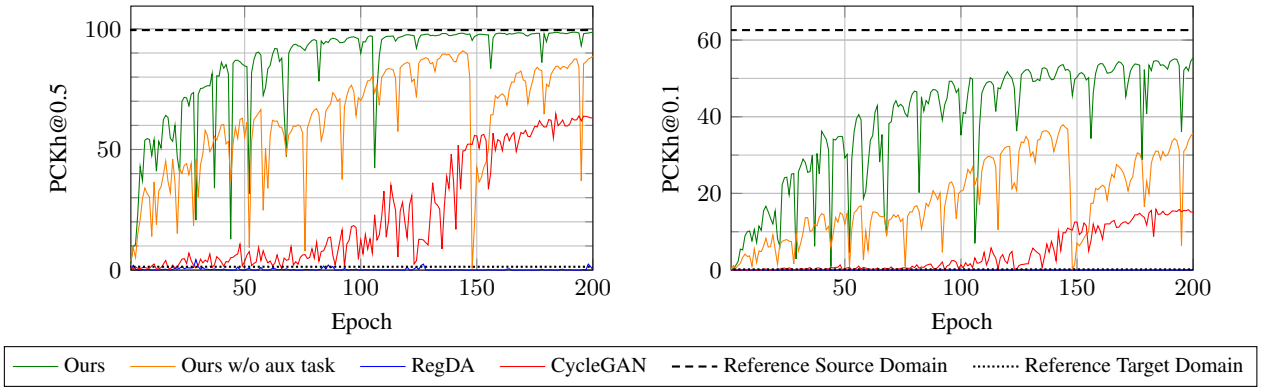

Figure 5: Setting 2: PCKh accuracy when performing domain adaptation on our dataset from synthetic RGB sensor data with casual clothing to synthetic DEPTH sensor data with casual clothing.

source domain and thus resulting in low performance on the target domain and low generalization capability. In contrast, the SURREAL dataset contains a lot more data and variations, *e.g.* different persons and scenes, making it harder to learn but also leading to higher generalization capabilities.

In the following experiments, we investigate how DA with CycleGAN, RegDA, and our method behaves in these four different DA settings. For reasons of limited GPU memory, we train CycleGAN and our method with a batch size of 8, while we keep RegDA's default batch size of 32. This is important to mention because the training procedure of the Transfer-Learning-Library assumes that an epoch consists of 500 batches leading to four times more samples per epoch in the case of RegDA compared to the training runs of CycleGAN and our method. To show the benefit of using an auxiliary task, we perform an ablation study in the first and second DA settings where we remove the auxiliary task losses, which is basically CycleGAN with a cyclical learning rate and without an image buffer.

Our method is able to compensate for the sensor domain shift and is very close to source domain performance in the first three DA settings, as shown in Fig. 4-6, outperforming CycleGAN and RegDA. The training process of CycleGAN shows larger oscillations in PCKh accuracy across epochs and unstable performance. Our method also suffers from occasionally appearing performance drops, but the overall performance is more stable as it usually recovers within a few epochs and maintains high accuracy. As our ablation study suggests, a cyclical learning rate already increases DA performance with CycleGAN, yet using our auxiliary task results in about 10% higher PCKh@0.5 accuracy in settings 1 and 2 and a more stable training process. While RegDA struggles with large sensor domain shifts in settings 2 (RGB/depth) and 3 (synthetic/real), it achieves the best DA performance in setting 4 (SURREAL/LSP). Because RegDA depends on the pose estimation's predictions on the target domain to generate a ground false, it is very sensitive to those predictions, which is supported by our results. The higher the target domain performance (settings 1 and 4), the better RegDA's DA performance and vice versa (settings 2 and 3). Therefore, we conclude that it is necessary to explicitly address sensor domain shift as settings 1 to 3

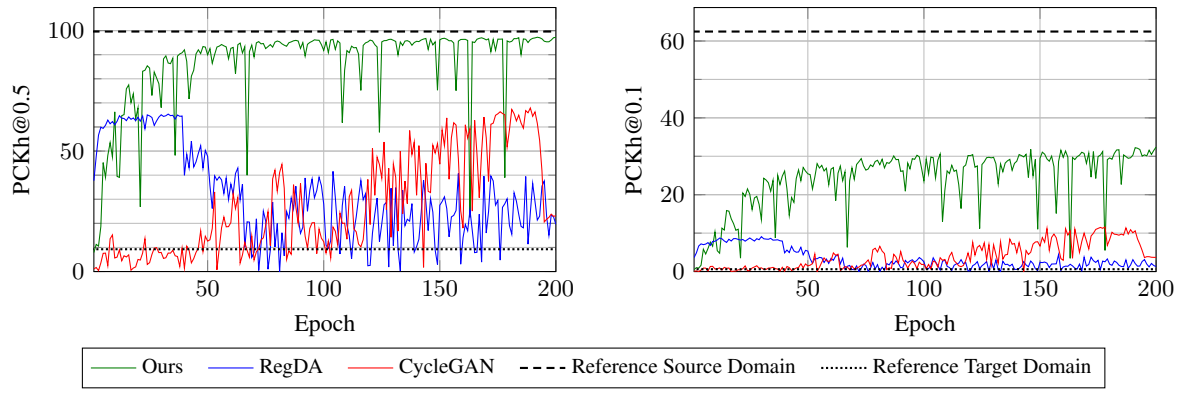

Figure 6: Setting 3: PCKh accuracy when performing domain adaptation on our dataset from synthetic RGB sensor data with motion capture suit to real RGB sensor data with motion capture suit.

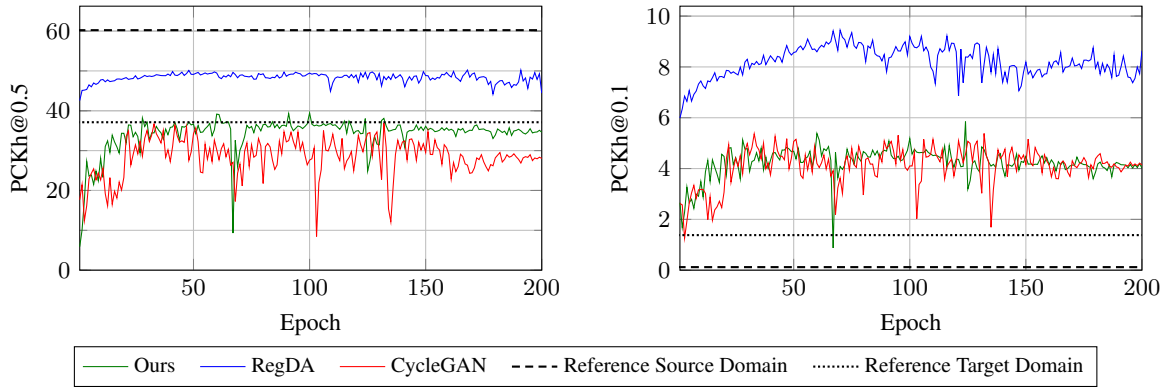

Figure 7: Setting 4: PCKh accuracy when performing domain adaptation from SURREAL to LSP dataset.

can be successfully handled with image or in general sensor translation. Still, sensor translation has its limitations as can be seen in setting 4 where CycleGAN and our approach both stay below the target domain performance of the reference model without DA. This behavior might be attributed to the fact that CycleGAN and thus our method can not handle one-to-many mappings, which occur due to the mismatch in the number of training samples of the SURREAL (451,439) and LSP (2,000) dataset. Although there are methods [1] that can handle one-to-many mappings, this is beyond the scope of this work, *i.e.*, investigating black-box DA for cross-sensor adaptation in settings with controllable domain shift.

5. Conclusion

In this work, we investigated the influence of the domain shift in four different settings on CycleGAN and RegDA for the task of UDA of a model for human pose estimation. For benchmarking, we created a motion capture-based, synchronized, and paired dataset especially targeted at the sim-to-real domain shift for the task of human pose estimation. Furthermore, we proposed to extend CycleGAN with auxiliary tasks, which can be arbitrarily but task-related chosen, inspired by multi-task learning and to use a cyclical learning rate for training that improves performance compared to CycleGAN and RegDA in three out of four DA settings while making our method suitable for black-box DA.

ACKNOWLEDGMENTS

This work is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project ”KI Delta Learning” (grant 19A19013S). The authors would like to thank the consortium for the successful cooperation.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv*, Feb. 2018.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark

and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[4] Dennis Burgermeister and Cristóbal Curio. PedRecNet: Multi-task deep neural network for full 3d human pose and orientation estimation. In *IEEE Intelligent Vehicles Symposium (IV)*, 2022.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, Jan. 2021.

[6] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[7] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, Dec. 2021.

[8] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer International Publishing, 2017.

[9] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[10] You Dingyi, Wang Haiyan, and Yang Kaiming. State-of-the-art and trends of autonomous driving technology. In *2018 IEEE International Symposium on Innovation and Entrepreneurship (TEMS-ISIE)*, pages 1–8, Mar. 2018.

[11] Michael Essich, Dennis Ludl, Thomas Gulde, and Cristobal Curio. Learning to translate between real world and simulated 3d sensors while transferring task models. In *2019 International Conference on 3D Vision (3DV)*. IEEE, Sept. 2019.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 2015. PMLR.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, Jan. 2016.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, July 2014.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[18] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. pages 8784–8794, Seattle, WA, USA, 2020. IEEE.

[19] Junguang Jiang, Baixu Chen, Bo Fu, and Mingsheng Long. Transfer-Learning-library, 2020.

[20] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6776–6785, 2021.

[21] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010.

[22] Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Chen, Rogerio Feris, David Cox, and Nuno Vasconcelos. VALHALLA: Visual hallucination for machine translation. May 2022.

[23] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. DINE: Domain adaptation from single and multiple black-box predictors. In *2022 IEEE/CVF Conference on Computer Vison and Pattern Recognition (CVPR)*, 2022.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, July 2015. PMLR.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 483–499, Cham, 2016. Springer International Publishing.

[26] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct. 2010.

[27] Bernardino Romera Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. Exploiting unrelated tasks in multi-task learning. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 951–959, La Palma, Canary Islands, 2012. PMLR.

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.

[29] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv*, Apr. 2020.

[30] Weisong Shi and Liangkai Liu. *Computing Systems for Autonomous Driving*. Springer International Publishing, 2021.

[31] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, Oct. 2000.

[32] Leslie N. Smith. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, Mar. 2017.

[33] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Computer Vision – ECCV 2018*, pages 536–553. Springer International Publishing, 2018.

[34] Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, June 2014. ISSN: 1063-6919.

[35] Partoo Vafaeikia, Khashayar Namdar, and Farzad Khalvati. A brief review of deep multi-task learning and auxiliary task learning. *arXiv*, July 2020.

[36] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[37] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018:1–13, 2018.

[38] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, Oct. 2018.

[39] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 472–487, Cham, 2018. Springer International Publishing.

[40] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. In *32nd British Machine Vision Conference (BMVC)*, 2021.

[41] Lei Zhang. Transfer adaptation learning: A decade survey. *arXiv*, Mar. 2019.

[42] Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 599–607. Springer International Publishing, 2018.

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017.