Towards Effective Synthetic Data Sampling for Domain Adaptive Pose Estimation

Isha Dua[*] **Arjun Sharma**[*] **Shuaib Ahmed** **Rahul Tallamraju**
Mercedes-Benz Research and Development India
{isha.dua, arjun.a.sharma, shuaib.ahmed, rahul.tallamraju}@mercedes-benz.com

Abstract

In this paper, we investigate a synthetic data sampling approach towards unsupervised domain adaptation (UDA) for pose estimation. UDA is characterized by a labeled source domain and an unlabeled target domain. We observe that recent work in UDA for pose estimation fails to generalize across poses in target data, despite having support for such poses in the source data. We hypothesize that this failure to generalize is due to a lack of uniform support across poses of varying complexity in the source domain. Motivated by this challenge, we aim to sample and train with the source domain data to improve the domain adaptation performance on a target domain. The proposed sampling strategy sorts the source domain samples based on a difficulty score, which reflects the lack of uniform support across varying pose complexity in the source domain. The difficulty score is a reconstruction error obtained from training an auto-encoder on the source domain poses. We categorize the dataset into closely related groups using this score. Selectively training from all or some of these groups help us to better utilize the source pose distribution. Finally, current pose estimation evaluation metrics do not effectively measure the ability of the model to learn the geometry of pose. We evaluate our approach qualitatively and quantitatively on benchmark datasets. Our sampling strategy outperforms existing state-of-the-art for domain adaptation.

1 Introduction

Pose estimation is an important problem in computer vision with applications ranging from autonomous driving [12, 11], motion capture [2], and robotics [13]. To this end, the synthesis of labeled training data using modern computer graphics[9, 10] is becoming increasingly relevant, allowing the creation of vast datasets under controlled conditions. These synthetic datasets reduce the need for laborious manual keypoint annotations. However, pose estimation models trained on rendered synthetic data suffer from a domain gap problem, arising from the differences in appearance, viewpoint, lighting, and environmental factors. This can significantly affect the performance of model to new, unseen domains. A specific gap more relevant to pose estimation is the 'pose distribution gap', a gap between available poses in the source and target domains. Pose distribution gap under unsupervised domain adaptation (UDA) has been scarcely studied in the literature, which is being addressed in this paper. With careful analysis of the state-of-the-art UDA models namely RegDA[5] and UDAPE [7], we noticed a significant limitation that they struggle to generalize across poses in the target domain despite similar poses being present in the source domain. As shown in Figure 1, for the ground truths, there are similar poses present in source domain SURREAL [10], though with difference in 3D orientation. However, the RegDA model fails to predict these poses accurately. A similar behaviour was observed with the UDAPE model. We hypothesize, that this failure to generalize originates from the inherent bias of model towards poses with low complexity, redundancy and lack of uniform support across poses of varying complexity in the source domain. These observations motivated us to

[*] Both authors contributed equally.

NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research.

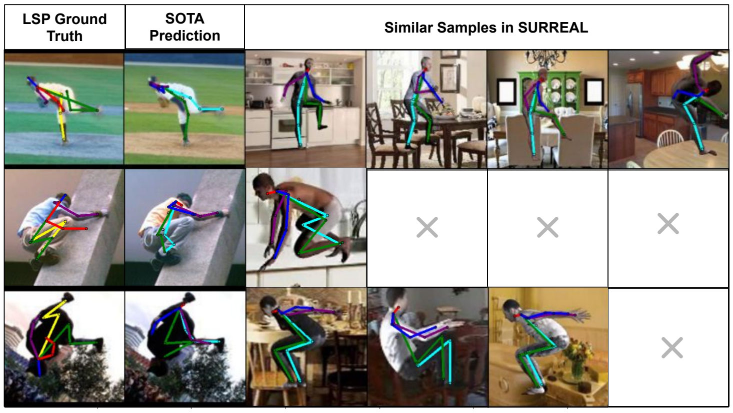


Figure 1: LSP [6] ground truth, prediction by SOTA: RegDA [5] and similar poses in SURREAL [10]. Even with a diverse range of scenarios in SURREAL, we observe that the RegDA model encounter difficulties in achieving effective generalization. This challenge appears to exhibit from an uneven representation of poses in the source domain.

investigate methods to categorize the source domain poses across varying levels of complexity and subsequently sample from this categorization to improve the generalization of pose estimation.

Existing efforts to categorize the pose distribution, use heuristics based on joint distance and k-means clustering [3], incorporate model confidence on a pose into the loss computation [1]. However, these methods fail to accurately represent the skeletal geometry as they account for each joint independently but do not model the plausibility of the overall pose. Moreover, in domains like object detection [16] clusters of objects are determined using a difficulty heuristic and then sampled for increasingly difficult objects during training. In contrast to the above methods, we introduce a novel method, to score the source domain poses using an auxiliary deep learning model, categorize based on this score and sample from these categories for domain adaptation. We train the auxiliary scoring model as a variational auto-encoder whose input and output supervision is the same 2D skeleton. The inability of the model to faithfully reconstruct an input pose during testing reflects a lack of support, variability and complexity of the pose in the source distribution. Specifically the reconstruction error enables us to score and categorize the source pose distribution into closely related groups. We rank these groups using this score and sample a subset from them to systematically train the model. This enables the model to improve its utilization of the source pose distribution and makes it more sample efficient.

We showcase through detailed experiments, on UDA benchmarks (Surreal [10] to LSP [6], Surreal to Human3.6M [4], RHD [17] to H3D [15]) and comparisons with the state-of-the-art, that our proposed approach provides consistent improvement for domain adaptation in pose. Furthermore, we demonstrate qualitatively and quantitatively that the proposed evaluation heuristic captures the skeletal geometry preserving capabilities of the pose estimation model. To summarize, the contributions of our work are as follows:

- We examine how the complexity, support and variability of poses in the source domain during domain adaptation impacts the performance of pose estimation in the target domain. To the best of our knowledge this has not been analyzed previously in literature.
- We propose, (a) a variational auto-encoder based scoring and categorization mechanism to analyze the complexity, support and variability of poses in a dataset, (b) this mechanism enables us to group and sample data used from the source domain for domain adaptation.

2 Proposed Approach

For our experiment a labeled source dataset $S = \{(x_s^i, y_s^i)\}_{i=1}^{N}$, where x_s is the source image, y_s is the corresponding keypoint annotation and N is the number of source images and an unlabeled target dataset $T = \{(x_t^i)\}_{i=1}^{M}$, where x_t is the target image and M is the number of target images, we train a keypoint regression model for the pose estimation using the source dataset and transfer the learned knowledge to the target domain in an UDA setting. We categorize poses into distinct complexity levels and selectively sample from these categories to train a domain adaptation model. In this section, we first provide an overview of the proposed approach in Figure 2. Subsequently, we

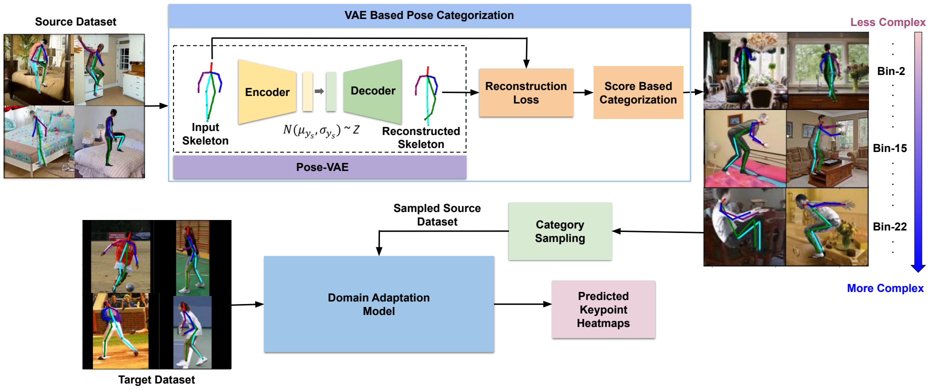


Figure 2: Pose Variational Autoencoders (VAE) is utilized to categorize the source data into k bins in the order of increasing complexity. These categories are then strategically sampled to create a representative set. Together with the target dataset, we train a domain adaptation model for pose estimation.

explain the scoring of poses using a variational auto-encoder in Section 2.1. Next, we propose score based categorization in Section 2.2. We then discuss domain adaptation methods that we use in this work in Section 2.3. Finally we propose a heuristic EvalPose to geometrically compare poses.

2.1 Pose VAE

We train a variational auto-encoder to estimate the complexity of a pose in the source dataset similar to [8]. The input and output supervision for the VAE is the same set of 2D keypoints. These keypoints are flattened out and passed as input to the VAE. The loss function used to train the VAE is mentioned below in equation 1.

$$L_{VAE} = \sum_{i=0}^{K} \|y_s^i - \hat{y}_s^i\|^2 + \lambda KL[\mathcal{N}(\mu_s, \sigma_s), \mathcal{N}(0, I)] \tag{1}$$

Here, y_s^i denotes a keypoint location for a pose sampled from the source domain, K is the number of keypoints in a pose, $\hat{y}_s^i$ is the prediction of VAE decoder for the i^{th} keypoint, KL denotes the Kulback-Leibler divergence loss, $\mathcal{N}$ is a normal distribution with μ_s and σ_s as mean and standard deviation of the VAE's latent space. The trained VAE enables us to compute the reconstruction error for a given pose. If the reconstruction error is high, the model is unable to faithfully reconstruct a pose. This implies lack of support, complexity and variability of the skeleton in the source pose distribution. This provides a useful mechanism to score poses in the source domain.

2.2 Score based Categorization

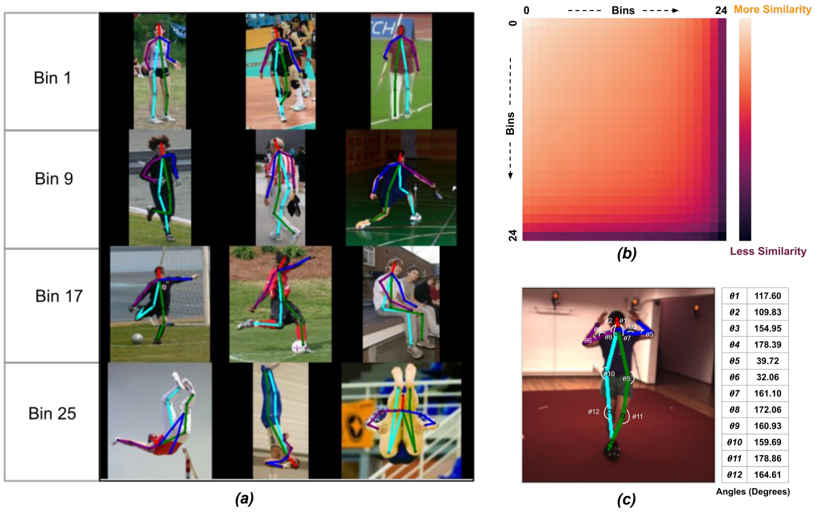
Given an input 2D pose, we compute the score reflecting the complexity of such a pose as the reconstruction error of the VAE as given by the equation 2.

$$score = \sum_{i=0}^{K} \|y_s^i - \hat{y}_s^i\|^2 \tag{2}$$

Using this score, we categorize the source poses into fixed number of bins. To achieve this, we sort all poses from the least to the largest reconstruction error. Intuitively, we observe that the score defined in equation 2 reflects the complexity or the difficulty of the pose. Figure 3(a) visualizes qualitatively how the complexity of the pose increases across the bins for LSP [6].

2.3 Category Sampling and Domain Adaptation

Using the pose VAE 2.1 and the score based categorization 2.2, we sample a subset of categories which gives us more information about the source domain poses. Specifically, about complexity and rarity of a pose. After selecting a subset we train a domain adaptation model to study the impact of the selected categories. Since our approach is general and applies across any domain adaptation method, we incorporate the process of category selection into two state-of-the-art UDA methods namely, UDAPE [7] and RegDA [5]. UDAPE follows a student-teacher learning paradigm with additional consistency losses applied on pseudo-labels for UDA. Additionally, it bridges the appearance level gap using a bi-directional style transfer method. RegDA is an adverserial domain adaptation method that builds on the concept of margin disparity discrepancy (MDD) [14] to human pose estimation. It follows a three step training strategy, starting from source domain training on the whole model,



θ1	117.60
θ2	109.83
θ3	154.95
θ4	178.39
θ5	39.72
θ6	32.06
θ7	161.10
θ8	172.06
θ9	160.93
θ10	159.69
θ11	178.86
θ12	164.61

Angles (Degrees)

Figure 3: (a) Sample images from categorized bins of LSP [6], arranged in order of increasing bin complexity. (b) Heatmap plot illustrates the correlation among different bins in SURREAL dataset. Darker values signify less similarity and high pose variation. We observe that the last few bins cover high variation in poses and have more information for training a pose estimation model. (c) Kinematic tree and angles derived from selected set of keypoint triplets within the kinematic graph.

an adverserial maximization step and a minimization step on the model head and the backbone. We observe that our proposed category sampling improves both these methods. For instance, for UDA benchmark SURREAL to LSP, we sample the categories corresponding to high reconstruction error. Intuitively, this part of the dataset has high variation in poses and is more informative for training a pose estimation model. We support this intuition with a visualization in Figure 3(b) which showcases the similarity across poses in each of the categories. The computation of this similarity is given in 2.4. The intuition of sampling high reconstruction error categories may seem simple but is effective in improving the domain adaptation performance as supported by the results in Table 2.

2.4 Pose Geometry Comparison

To compute the similarity of a pair of poses, we first define a custom kinematic graph defining the skeleton structure. We then compute the angles within a chosen set of keypoint triplets in the graph as shown in Figure 3(c). Next, the similarity between the pair of poses is computed as a dot product of the resulting set of angles. The equation below summarizes the computation, we refer to this as EvalPose.

$$EvalPose = \frac{\Theta_1 \Theta_2}{|\Theta_1||\Theta_2|} \tag{3}$$

Here, $\Theta = [\theta_1 \ldots \theta_j]$ is a set of angles computed across all triplets defined in the kinematic graph. Each θ_j is computed as $\theta_j = \cos^{-1} \frac{\mathbf{v}_1 . \mathbf{v}_2}{|\mathbf{v}_1||\mathbf{v}_2|}$. v_1 and v_2 are keypoint connection vectors between pairs in the triplets defined in the kinematic graph. This heuristic is scale, translation and rotation invariant. We utilize this to compute the skeletal similarity between the ground truth and prediction.

3 Experiments and Results

In this section, we present results of our synthetic data sampling approach.

Datasets. We experiment our proposed methods on two diverse tasks: (1) human body pose estimation and (2) hand pose estimation. For human pose estimation, we used *SURREAL*[10] as source domain and *Leeds Sports Pose* [6] (LSP) as well as *Human3.6M* [4] (H3.6M) as target domain. On the other hand, for hand pose estimation we used *Rendered Hand Pose Dataset* [17] (RHD) as source domain whereas *Hand-3D-Studio dataset* [15] (H3D) as target domain. We now explain the training and testing test used for experimentation. **Training Set:** The datasets are stratified into easy-to-hard categories based on pose complexity scores (Section 2.2). These categories are employed in two UDA experimental setups: (1) *VAE-HM* (Hard Mining) utilizes only the challenging subset in source domain adaptation. (2) *VAE-CL* (Curriculum Learning) progressively integrates easy, medium, and hard datasets, following a curriculum approach. **Validation Set:** Each dataset

Table 1: PCK@0.05 score on task SURREAL → LSP. Sld: shoulder, Elb: Elbow. We observe that our method VAE-HM outperforms the UDAPE model across all the evaluation metrics.

Method	PCK						
	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source only	51.5	65.0	62.9	68.0	68.7	67.4	63.9
Oracle	95.3	91.8	86.9	95.6	94.1	93.6	92.9
RegDA	62.7	76.7	71.1	81	80.3	75.3	74.6
RegDA VAE-CL (Ours)	**65.1**	**77.2**	**72.6**	78.5	78.7	**77.8**	**75.0**
RegDA VAE-HM (Ours)	60.1	**79.9**	**75.1**	**81.1**	**80.8**	**79.5**	**76.1**
UDAPE	**69.2**	84.9	83.3	85.5	84.7	84.3	82.0
UDAPE*	68.1	83.1	82.3	83.8	83.1	83.0	80.6
UDAPE VAE-CL (Ours)	69.0	**85.2**	**84.0**	**85.8**	**84.9**	84.3	**82.2**
UDAPE VAE-HM (Ours)	68.5	**86.2**	**84.7**	84.8	**85.8**	**85.6**	**82.6**

Table 2: Avg PCK@0.05 and EvalSket score on benchmark tasks , SURREAL → Human3.6M (H3.6M) and Rendered Hand Pose (RHD) → Hand-3D-Studio (H3D).

	SURREAL→ LSP		SURREAL→ $H3.6M$		RHD→ $H3D$	
Method	Avg PCK	EvalPose	Avg PCK	EvalPose	Avg PCK	EvalPose
RegDA*	74.6	54.5	75.6	**66.5**	**68.6**	**69.7**
RegDA VAE-HM (Ours)	**76.1**	**58.2**	**77.8**	64.8	67.8	68.4
UDAPE*	80.6	61.5	**78.3**	**77.9**	79.6	78.5
UDAPE VAE-CL (Ours)	82.2	62.6	77.2	**77.9**	77.6	77.6
UDAPE VAE-HM (Ours)	**82.6**	**63.5**	**78.3**	77.8	**79.8**	**78.8**

is categorized into 25 bins of equal sizes using the category based sampling approach mentioned in section 2.3. This categorization helps establish our key insights on the domain adaptation model's generalization capabilities, across increasing pose complexities, on our VAE-HM and VAE-CL experimental settings.

Evaluation Metrics. In this paper, we utilize the PCK Metric and the EvalPose heuristic to evaluate our results. **PCK Score:** Percentage of Correct Keypoints Score is a metric used to measure the precision of body joint localization. The predicted joint is correct when the distance between the predicted and the true location is within a certain threshold. In our experiments, we present results using PCK@0.05, which quantifies the proportion of correct predictions within 5% of the image size. A higher PCK value signifies higher prediction accuracy. **EvalPose:** Section 2.4 describes the computation process for pose geometry comparison. We note that PCK alone is not sufficient to analyze subtle variations in joint-wise distances as visualized in Figure 5. Our proposed heuristic offers advantages over PCK as it is both scale and label gap agnostic. It does not need a threshold parameter and helps interpret model's pose geometry understanding in an intuitive way.

Quantitative Results. We present results from several benchmark Unsupervised Domain Adaptation (UDA) experiments, namely (1) SURREAL → LSP, (2) SURREAL → H3.6M, and (3) RHD → H3D. Table 1 showcases the outcomes of the experiments using various methods: (a) Source-only training, (b) Oracle (target-only training), (c) RegDA* [5], (d) UDAPE [7], (e) UDAPE*. '*' indicates our reproduced results of UDAPE from the provided codebase, and our approaches, (f) VAE-HM and (g) VAE-CL trained with RegDA and UDAPE. We observe that our experiments employing hard mining and curriculum learning outperform state-of-the-art UDA approaches, RegDA and UDAPE, on average across all joints. It is also important to note that we only use a fraction (33%) of the source dataset making it sample efficient. Additionally, our proposed category sampling mechanism proves to be invaluable in guiding the generation of synthetic data, offering a promising strategy for future datasets.

Furthermore, Table 2 outlines the results for all the considered benchmark UDA experiments. We highlight that our proposed category sampling approach outperforms the the baseline RegDA and UDAPE experiments for RHD → H3D experiment. We also note that the results for SURREAL → to H3.6M shows is on-par with RegDA and UDAPE baseline experiments despite using only 33% of the source dataset in the case of VAE-HM. We further highlight that VAE-CL is part of on-going work to strategically use and train with source domain data. The mentioned results are part of early work in this direction. Finally, Figure 4 showcases the bins from the category sampling of the LSP dataset. Each bar plot in a bin corresponds to a different method. The first bar plot corresponds to the RegDA UDA method. The second bar plot corresponds to our proposed VAE-HM method

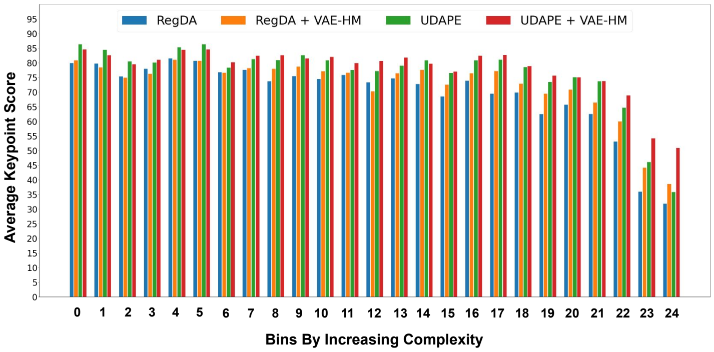


Figure 4: SURREAL → LSP task, we conducted a comparative analysis involving RegDA, RegDA + VAE-HM, UDAPE, and UDAPE + VAE-HM methods. We organized the bins in ascending order of pose complexity. We observe that RegDA + VAE-HM and UDAPE + VAE-HM demonstrate enhancements in the later bins compared to RegDA and UDAPE, respectively.

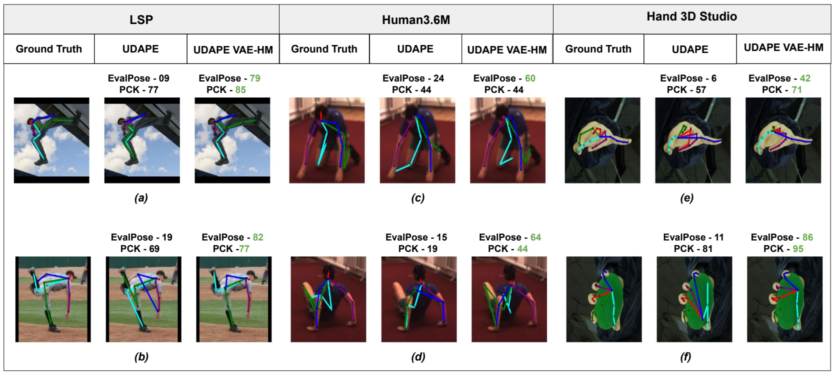


Figure 5: Qualitative Results showing our approach (VAE-HM) against UDAPE model predictions on LSP, Human3.6M and Hand 3D Studio Datasets. Notably, UDAPE VAE-HM demonstrates better performance on highly complex samples supported by the PCK score and EvalSket score.

using RegDA for domain adaptation. Third bar plot is the UDAPE UDA method. The fourth bar plot corresponds to our proposed VAE-HM method using UDAPE for domain adaptaion. The y-axis reflects the average keypoint score, which is computed as the average accuracy per pose, averaged across all the poses in the bin. We note that our proposed VAE-HM method outperforms the baseline in most of the bins. Moreover, we observe that as the bins increase in complexity, the advantage with the VAE-HM is more evident. Because, VAE-HM trains on poses which have high variation and, as evidenced empirically by the results, these poses are representative of the complete source domain pose distribution.

Qualitative Results. Figure 5 visualizes the qualitative improvements in our proposed VAE-HM method with UDAPE used for domain adaptation. Additionally, the corresponding PCK score and EvalPose score are also mentioned approximated to the nearest whole number. We observe that visually, the predicted skeleton from our proposed approach looks more plausible and closer to the groudtruth in each of the examples. Moreover, we note in some cases, Figure 5(b),(c), the PCK score does not show a clear improvement quantitatively, however we highlight that the proposed EvalPose heuristic provides a clear distinction in measurement of the predicted pose quality. Thereby, emphasizing the need for EvalPose as an explainable measure. We provide more such examples in the supplementary for the **Leeds Sports Pose**[6] (LSP) dataset.

4 Conclusion

Through this work we present a method to use the best subset of synthetic data for domain adaptation to improve performance of any existing domain adaptation model. Our strategy works as a plug and play approach which provides relevant and varied source domain distribution to the domain adaptation model. This leads to an increase in performance on the complex poses while retaining performance simpler poses. Further, we have provided a framework to categorize any pose estimation data set into highly related groups based on the skeleton structure.

References

[1] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[2] Yann Desmarais, Denis Mottet, Pierre Slangen, and Philippe Montesinos. A review of 3d human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding*, 212:103–275, 2021.

[3] Jihye Hwang, John Yang, and Nojun Kwak. Exploring rare pose in human pose estimation. *IEEE Access*, 8:194964–194977, 2020.

[4] 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*, 36:1325–1339, 2014.

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

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

[7] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. *ArXiv*, abs/2204.00172, 2022.

[8] Adrian Alan Pol, Victor Berger, Cecile Germain, Gianluca Cerminara, and Maurizio Pierini. Anomaly detection with conditional variational autoencoders. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 1651–1657. IEEE, 2019.

[9] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision – ECCV 2016*, pages 102–118. Springer, 2016.

[10] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017.

[11] Jun-Sang Yoo, Jung, and Seung-Won. Survey on in-vehicle datasets for human pose estimation. In *2022 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–2, 2022.

[12] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. Hum3dil: Semi-supervised multi-modal 3d humanpose estimation for autonomous driving. In *Proceedings of The 6th Conference on Robot Learning, PMLR*, pages 1114–1124, 2022.

[13] Feng Zhang, Xiatian Zhu, and Chen Wang. Comprehensive survey on single-person pose estimation in social robotics. *Int J of Soc Robotics*, 14:1995–2008, 2022.

[14] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International conference on machine learning*, pages 7404–7413. PMLR, 2019.

[15] Zheng Fa Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2478–2482, 2020.

[16] Ziyue Zhu, Qiang Meng, Xiao Wang, Ke Wang, Liujiang Yan, and Jian Yang. Curricular object manipulation in lidar-based object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1125–1135, 2023.

[17] Christiane Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017.