Towards Effective Synthetic Data Sampling for Domain Adaptive Pose Estimation

Isha Dua*, Arjun Sharma*,
Shuaib Ahmed, Rahul Tallamraju

Problem

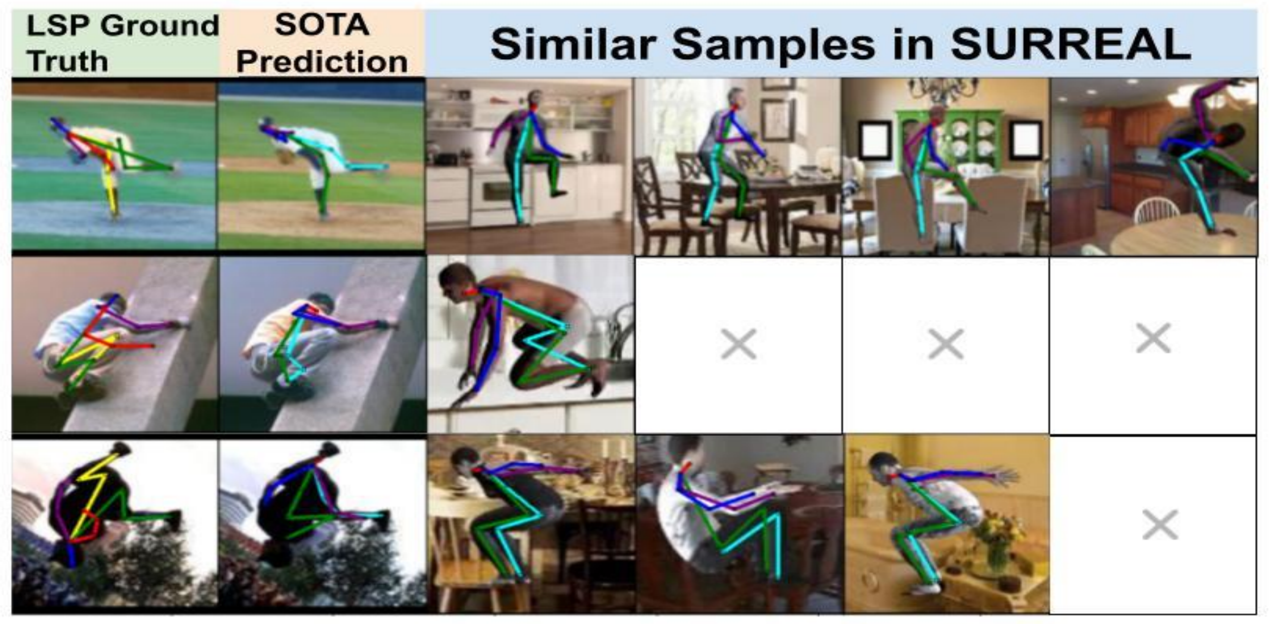


Figure: The figure shows that even with a diverse range of poses or variation in surreal, the state-of-the-art (SOTA) model encounter difficulties in achieving effective generalization on target domain (LSP). This challenge appears to exhibit from an uneven representation of poses in source domain (SURREAL).

- Current state-of-the-art (SOTA) models fails to generalize on target domain despite having support for similar poses in source domain.
- We hypothesize that the failure is due to lack of uniform support across poses of varying complexity in the source domain as shown in above figure.

Proposed Solution

- We propose a novel method that score the source domain poses using an auxiliary deep learning model, categorize based on this score and sample from these categories for domain adaptation.
- The proposed sampling strategy sorts the source domain samples based on a difficulty score. The difficulty score variation reflects the lack of uniform support across varying pose complexity in source domain as shown in the plot below.

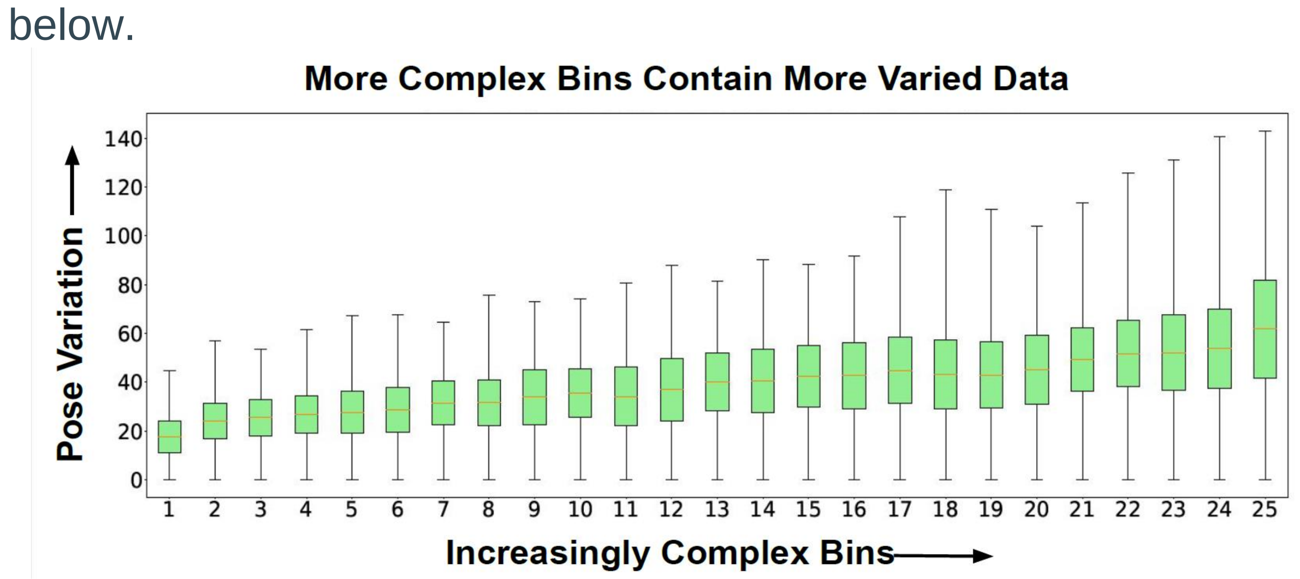


Figure: The box plot represents the pose variation observed within each bin. An increase in mean and variance is observed, upon moving towards more complex bins. Higher variance indicates more diversity of poses within the bin and higher mean indicates less support for poses.

- The difficulty score is a reconstruction error obtained from training an auto-encoder on the source domain poses. The dataset is categorized into closely related groups based on this score.
- We utilize these groups selectively for training to better utilize the source pose distribution for more generalized domain adaptation.
- The proposed sampling strategy outperforms the state-of-the-art model for all the tasks on human pose estimation and hand pose estimation.

Methodology

The proposed architecture consists of a **Pose Variational Auto Encoder** and a **domain adaptation model** which is based on a student-teacher architecture proposed in state-of-the-art UDAPE model.

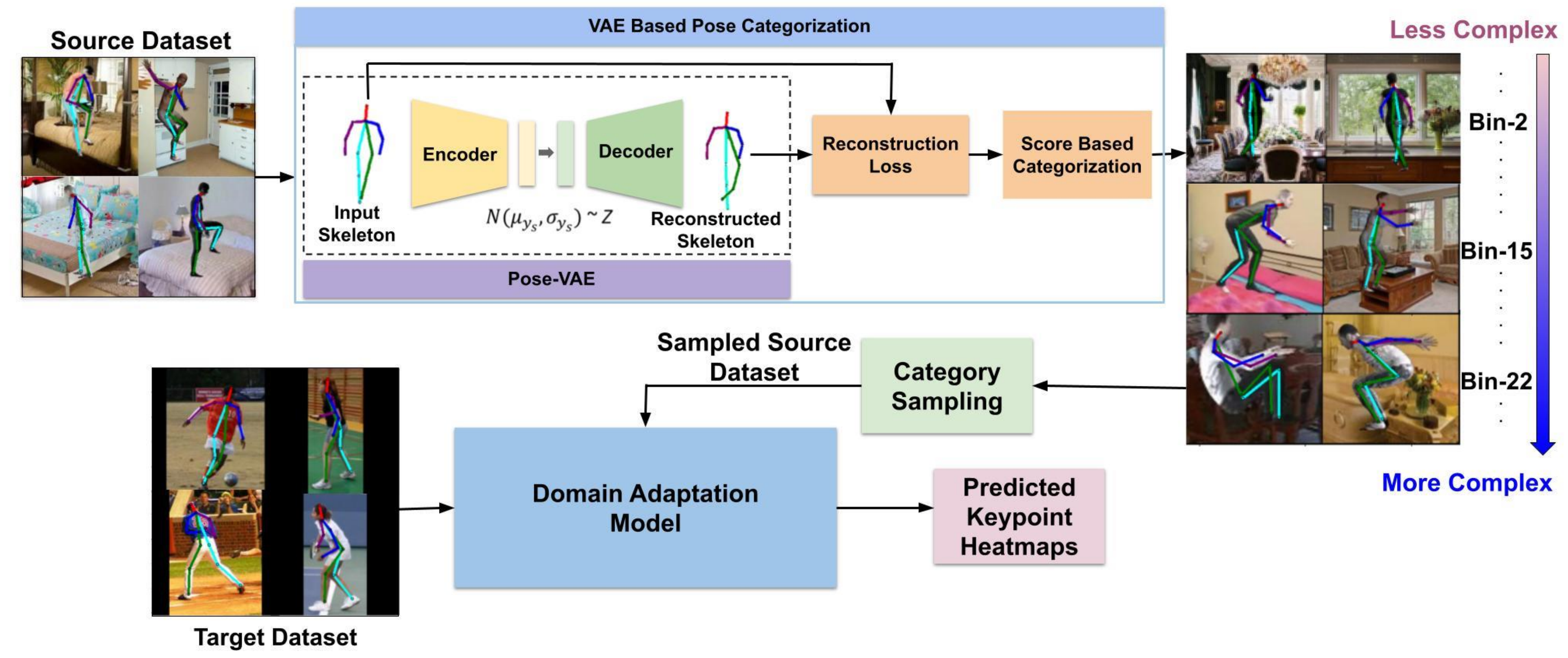


Figure: In the proposed architecture, Pose Variational Autoencoders (VAE) is utilized to categorize the source data into *k* bins in the order of increasing complexity. These categories are then strategically sampled to create a representative set. Together with the target dataset, we train a domain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

Pose VAE:

VAE is trained to estimate the complexity of poses in source dataset. The input and output supervision for the VAE is the same set of 2D keypoints. The loss used for training the VAE is:

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

Score Based Categorization:

Based on Pose VAE reconstruction error, we score the complexity of each pose using the equation:

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

- The score is further used to categorize the source poses into fixed number of bins. Higher number bin contains complex poses with high reconstruction score. Grouped bins are shown in the image below for few bins - 1, 9, 17, and 25.
- Heatmap shows that higher bins have large pose variations in the dataset.

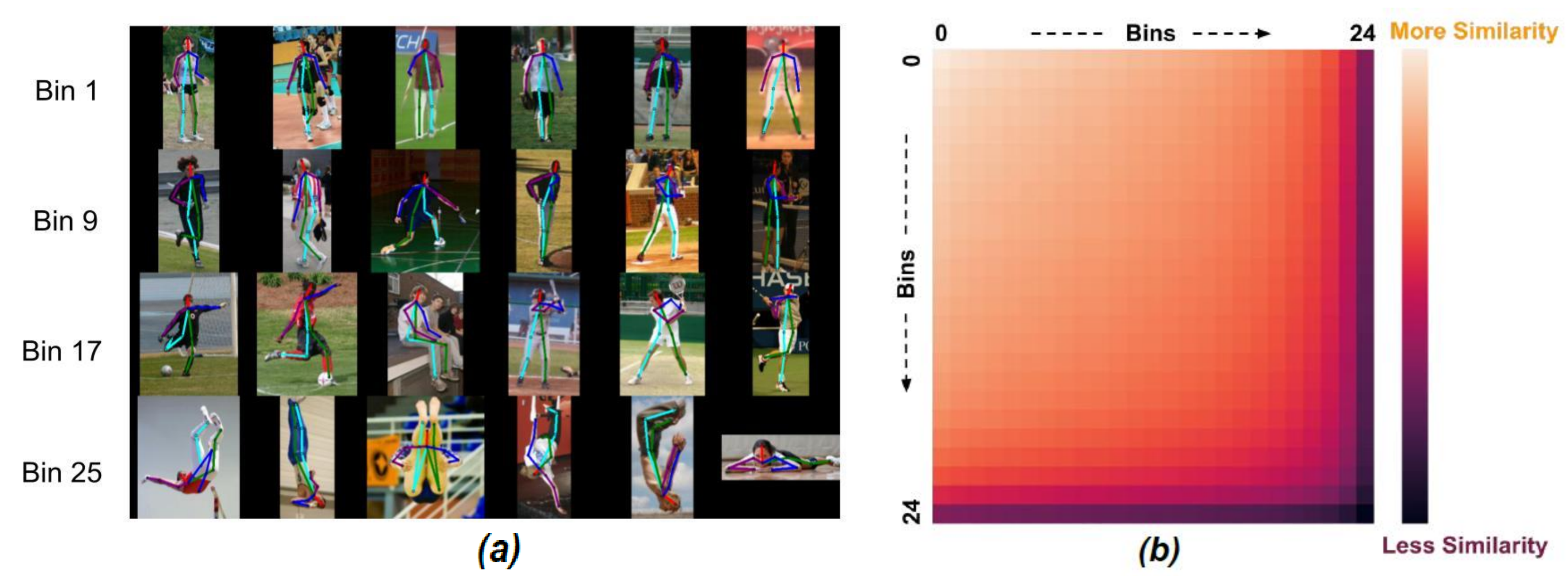


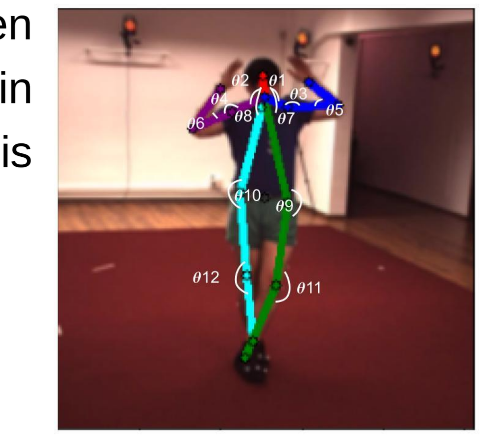
Figure: (a) Figure shows sampled images from the categorized bins of LSP arranged in the order of increasing complexity. This shows that the reconstruction error from VAE is an effective in grouping the dataset based on pose complexity and rarity. (b) Heatmap plot illustrates the correlation of poses among different bins in the SURREAL dataset. Darker values signify less similarity and high pose variation. We observed that the last few bins cover high variation in poses and have more information for training a pose estimation model.

EvalPose

EvalPose is a similarity metric between poses computed based on angles chosen in a kinematic graph as shown in figure. It is scale, translation and rotation invariant.

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

For each pose, $\Theta = [\theta_1 \ldots \theta_j]$ is a set of angles computed across all triplets defined in the kinematic graph.



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: The figure shows the kinematic tree and angles derived from the selected set of keypoint triplets within the kinematic graph.

Results

Quantitative Results:

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: PCK@0.05 score on task SURREAL -> LSP. Sld: Shoulder, Elb: Elbow. We observe that our method (RegDA, UDAPE) + VAE-HM outperforms the SOTA models.

	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**

Table: PCK@0.05 and EvalPose score on benchmark tasks SURREAL -> LSP, SURREAL -> Human 3.6M and Rendered Hand Pose (RHD) -> Hand-3D-Studio (H3D).

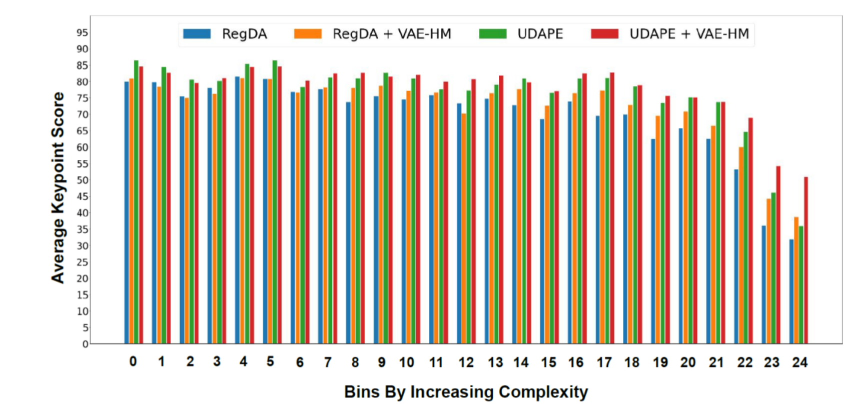


Figure: Plot shows that (RegDA, UDAPE) + VAE-HM models show significant improvement in later bins compared to the SOTA models.

Qualitative Results:

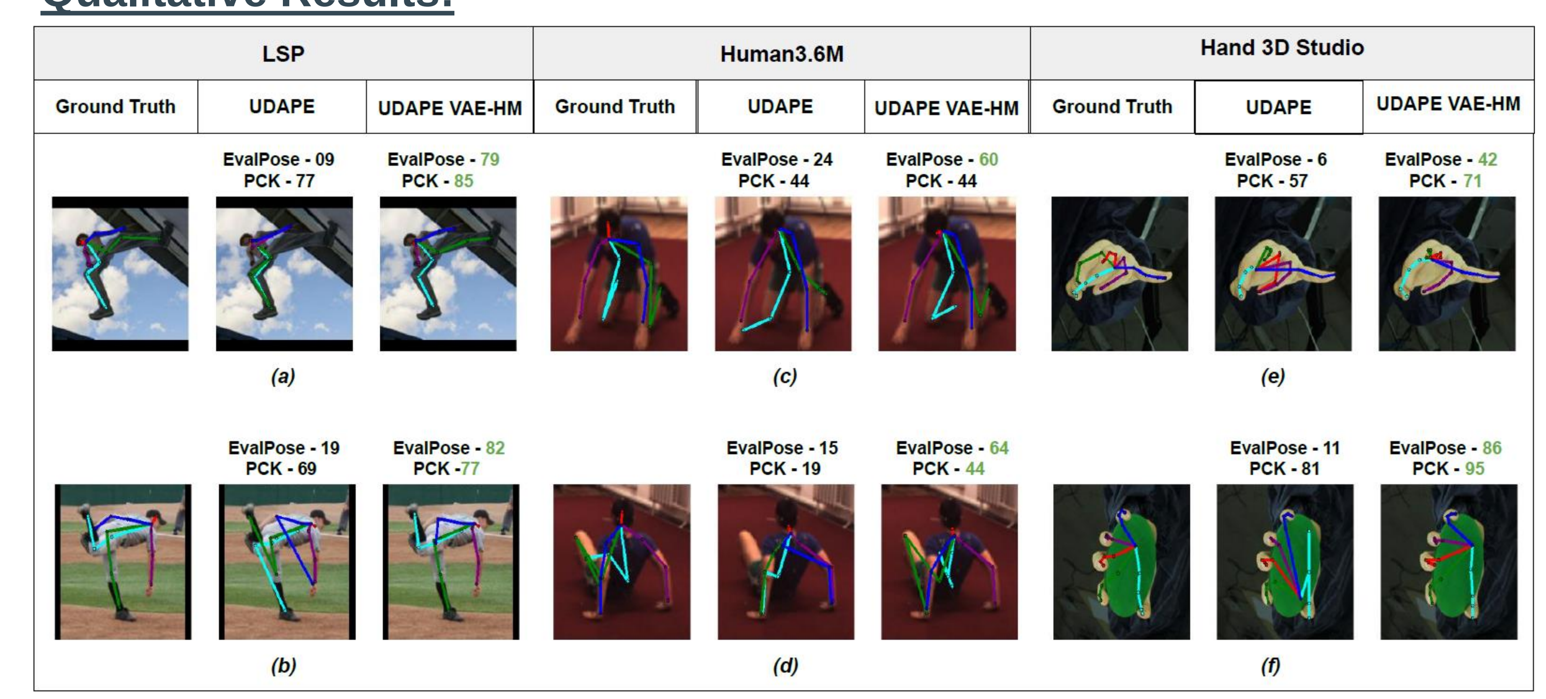


Figure: It shows that UDAPE + VAE-HM demonstrates better performance on highly complex samples compared to the state-of-the-art UDAPE model. Better performance is measured using the evaluation metric PCK and EvalPose score.

References

Varol, Gul and Romero, Javier and Martin, Xavier and Mahmood, Naureen and Black, Michael J. and Laptev, Ivan and Schmid, Cordelia (2017) "Learning from Synthetic Humans" In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Donghyun Kim and Kaihong Wang and Kate Saenko and Margrit Betke and Stan Sclaroff (2022) "A Unified Framework for Domain Adaptive Pose Estimation." in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII. .

Sam Johnson and Mark Everingham (2010) "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation" in British Machine Vision Conference

* Equal Contributors