Semi-Supervised Adaptation for Skeletal Data Based Human Action Recognition †

Haitao Tian[1*], Pierre Payeur[1]

School of Electrical Engineering and Computer Science,
University of Ottawa, Ottawa, Canada

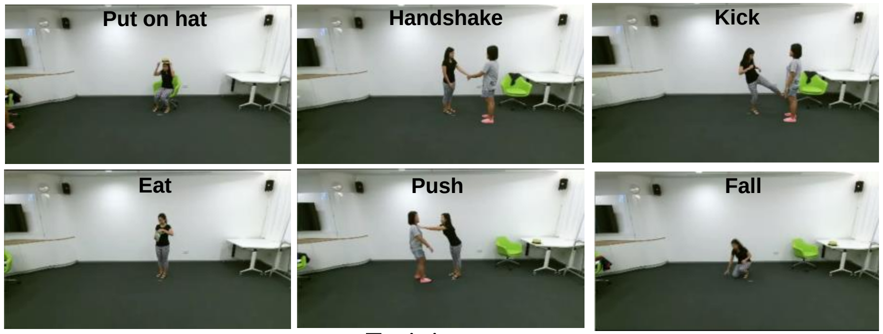
† Presented at 10th International Electronic Conference on Sensors and Applications

* Correspondence: Haitao Tian, *htian026@uottawa.ca*

Introduction

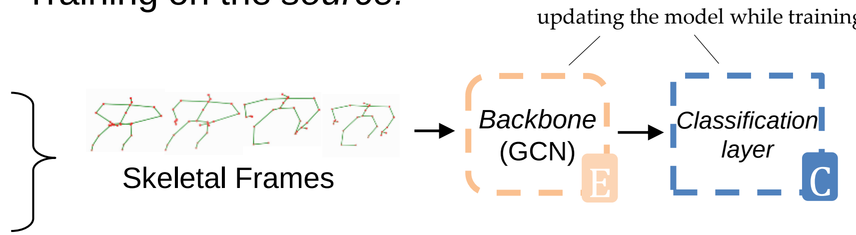
uOttawa

Source Training Benchmark [1]



Training set

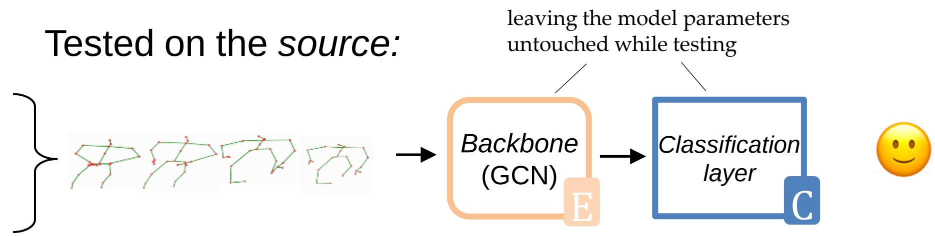
Training on the *source:*



- Training a human action recognition model on a skeletal based benchmark, such as *NTU RGB+D [1]*, is easy to realize

Test set

Tested on the *source:*

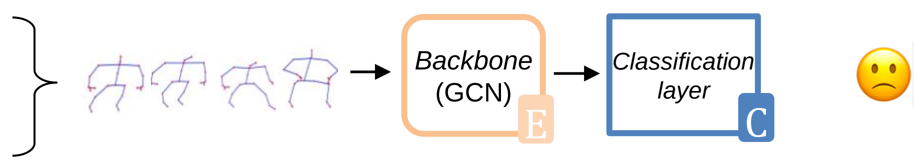


- Model evaluation is promising on the same benchmark.

Target Deployment Environment (lack of training data)

Test set

Tested on the *target:*



- The trained model fails on inference in the target test environment due to data domain shift in imaging configurations (e.g., variations in camera views, heights, and orientations)

[1] Shahroudy, A., Jun L., Tian-Tsong N., Gang W. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." In CVPR. 2016.

Supervised Adaptation *vs.* Semi-supervised Adaptation

Solutions to a learn capable human action recognition model adaptive to a target environment ?

- **Supervised learning** conducts extra fine-tuning rounds with full supervision of the data collected under the target imaging configuration.
  However, the collection and annotation of a large volume of skeleton data for fine-tuning is extremely time-consuming and troublesome, and it could even be prone to subjectivity while in-volving different subjects performing the same actions.

- **Semi-supervised learning** pretrains a model from the benchmark data set, then fine-tunes the pretrained model on a small number of labeled data samples in the target domain.

Overview of the Proposed Method

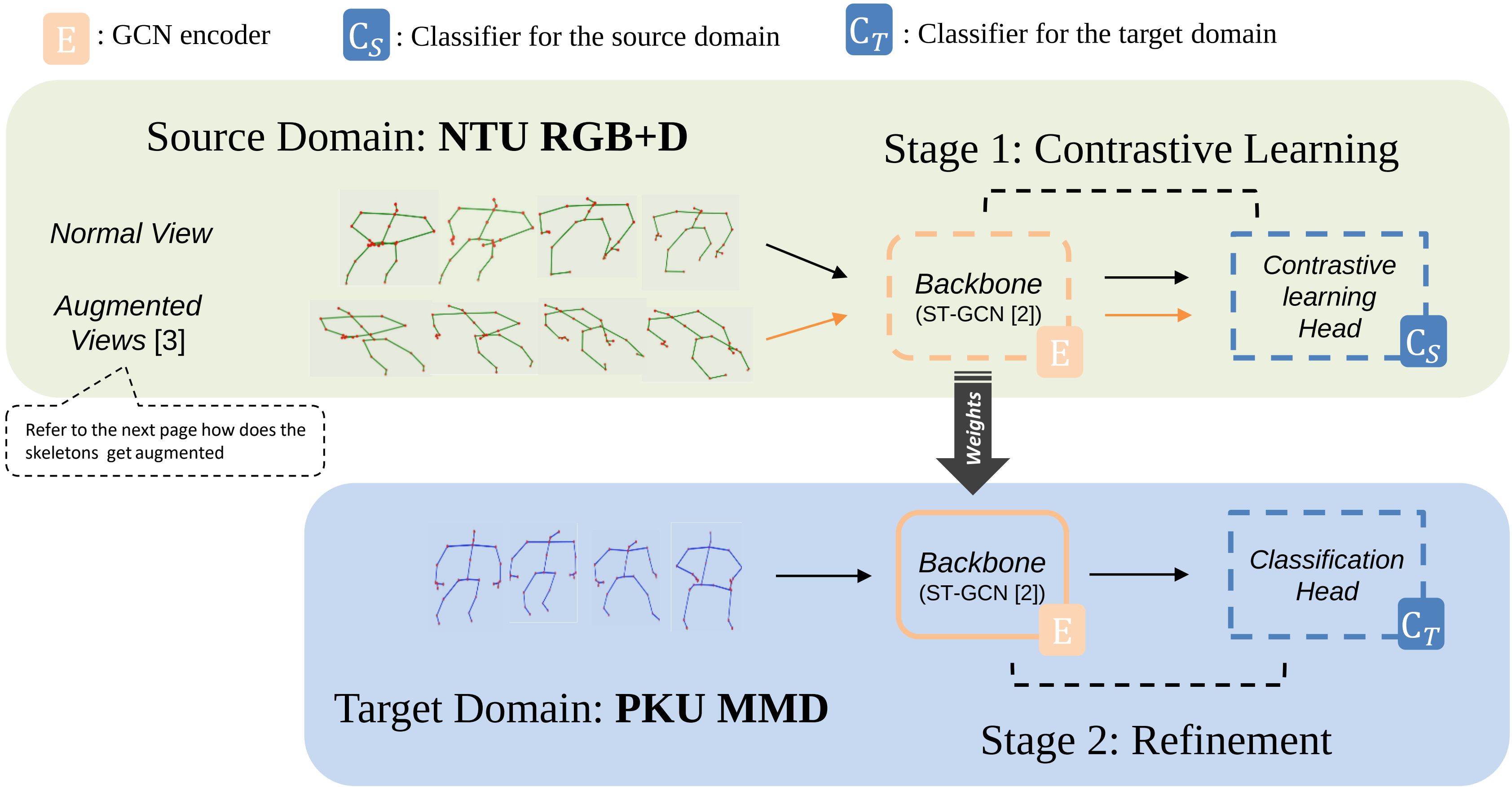


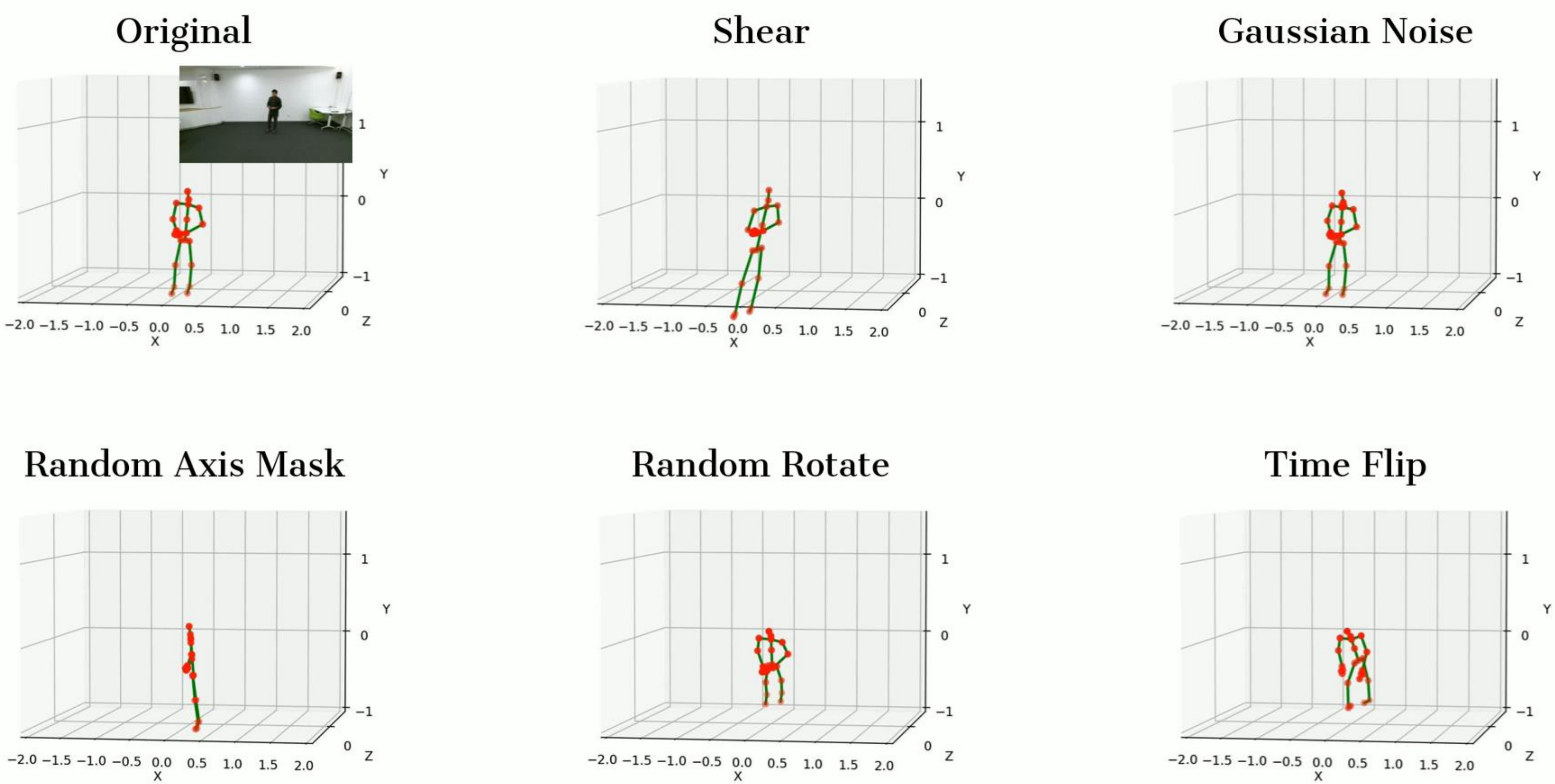
Figure 1. Framework of the proposed semi-supervised adaptation strategy. In the first stage, the training data (unlabeled) from the source domain is utilized to contrastive learning after data augmentation. The learned backbone is recycled in stage 2 for refining over the data samples (labeled) in the target domain.

[2]Yan, S., Xiong, Y. and Lin, D. "Spatial temporal graph convolutional networks for skeleton-based action recognition." In AAAI. 2018.
[3] Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T. and Ding, R. "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition." In AAAI 2022.

Skeleton Augmentations

- The Extremely Augmented Skeleton (EAS) scheme [3] enriches the input space in both spatial and temporal dimensions

It is a video demo



ECSA-10

Experimental Results

uOttawa

Table 1. Comparative performance of the proposed method against full supervised training.

Model	On benchmark (NTU RGB+D)	On target domain (PKU-MMD)
Source only	**69.07 %**	57.32 %
Adaptation	34.41%	**75.74%**

Table 2. Performance gains related to different proportions of the target data samples for $\mathbf{C}_T$ refinement training.

Percentage of data use	5%	20%	30%	50%	70%	100%
Accuracy	71.60%	77.33%	81.03%	82.25%	83.28%	85.06 %

Table 3. Performance of Fully supervised learning vs. Semi-supervised learning.

	Full Supervision		**Semi-supervision**
	NTU RGB+D & 10% PKU-MMD (fine tuning)	NTU RGB+D & 10% PKU-MMD (combined)	NTU RGB+D & 10% PKU-MMD (fine-tuning)
Accuracy	45.97%	62.61%	75.74%

[4] Liu, C., Hu, Y., Li, Y., Song, S. and Liu, J. "PKU-MMD: A large scale benchmark for skeleton-based human action understanding." In Proceedings of the Workshop on VASCC, pp. 1-8. 2017.

T-SNE Visualization

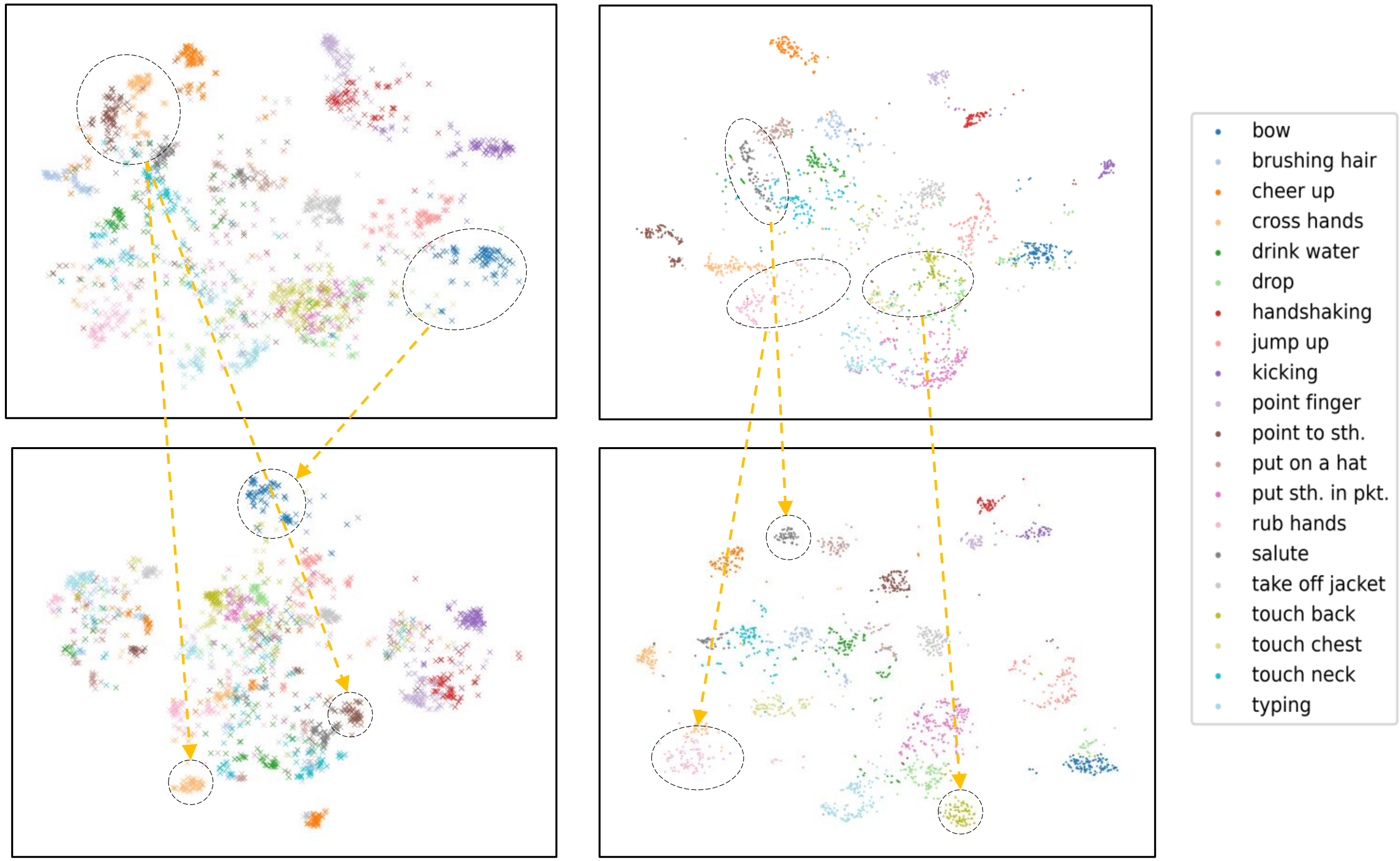


Figure 2. T-SNE visualization on action clusters on the embedding space of ST-GCN (upper row: "Source only"; bottom row: "Adaptation"). Clusters are distinguished by colors where twenty actions are randomly selected among fifty actions for clarity. Left column represents action clusters of the source domain (NTU RGB+D) and right column shows clusters of the target domain (PKU-MMD), respectively.

Conclusion

This work proposes a simple but efficient method to deploy a skeleton data based human action recognition model to a target environment while requiring only a small amount of labeled data from the latter.

- The proposed semi-supervised learning strategy utilizes contrastive learning to pretrain a model that learns key skeletal representations from an unlabeled dataset, then fine-tunes the pretrained model on a small number of labeled data samples in the target domain.
- Experiments are conducted to demonstrate the effectiveness of the proposed strategy. It suggests that the semi-supervised learning method achieves convincing results compared to fully supervised learning that requires voluminous labeled data from both the source and target domains.
- The research also experimentally characterizes a trade-off between data usage and model performance, providing reference to develop and deploy future applications.