I’ll Have the “CNN-Three-Ways” Please!
Automated Identification of Human Gait Abnormalities

Adam Gotlin[1], Apurva Pancholi[2], Umang Agarwal[3]
{agotlin,apurva03,uagarwal}@stanford.edu

We would like to extend a special thanks to **Lukasz Kadzinski** of the Stanford Mobilize Center, who was our advisor for this project, as well as the **CS 229 teaching team**.

Background

Problem Statement: Several prominent pathologies (Cerebral Palsy, Parkinson’s and Alzheimer’s) can manifest themselves in an abnormal walking gait. Gait Deviation Index (GDI) indicates the extent of gait pathology and is currently measured through a cumbersome and expensive marker-based motion capture process.

Model Inputs/Outputs: Patient video is captured by commodity devices and analyzed by machine learning algorithm to predict GDI.

Approach: We leverage DensePose to featurize each frame of video which is then passed through a machine learning model to minimize root mean square error (RMSE) of GDI prediction.

Results: The highest performing model passed each frame in a 10-frame video through a 2D CNN, then passed the featurized frame-vectors into an LSTM for GDI prediction. **The final model predicted GDI with an RMSE of 3.6.**

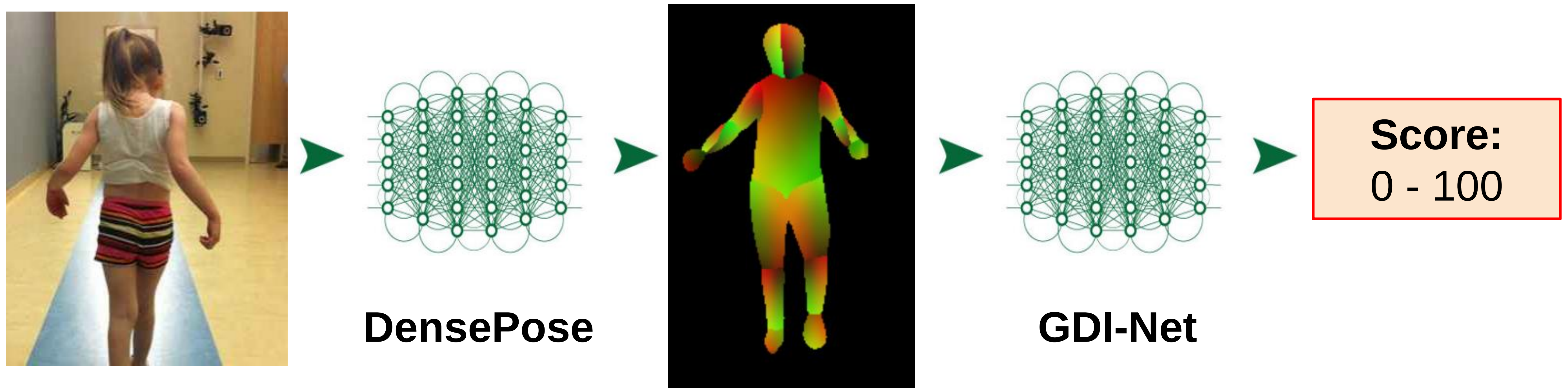
Data Processing

We use 3,249 videos of patients walking in a room at Gillette Children’s Hospital. The videos have a resolution of 640x480 and are processed using DensePose, which maps all human pixels of an RGB image to the 3D surface of the human body.

DensePose-RCNN finds dense correspondence by partitioning the human body surface, assigning each pixel to a body partition and determining the pixel location in 2D parameterization (UV coordinates) of the body part.

The parametric surface model that DensePose fits is the Skinned Multi-Person Linear (SMPL) model.

Methods



Methods Detailed:

- Monocular video footage of patient gait is captured by physician during a session in a motion analysis lab
- Video footage is processed by DensePose and transformed to IUV coordinates
- Frame(s) are sent through a deep neural network consisting of spatial and temporal components
- GDI predictions are made for each batch; loss is calculated using mean squared error compared to physician’s score
- Overall model performance is judged on RMSE of GDI prediction and tuned accordingly

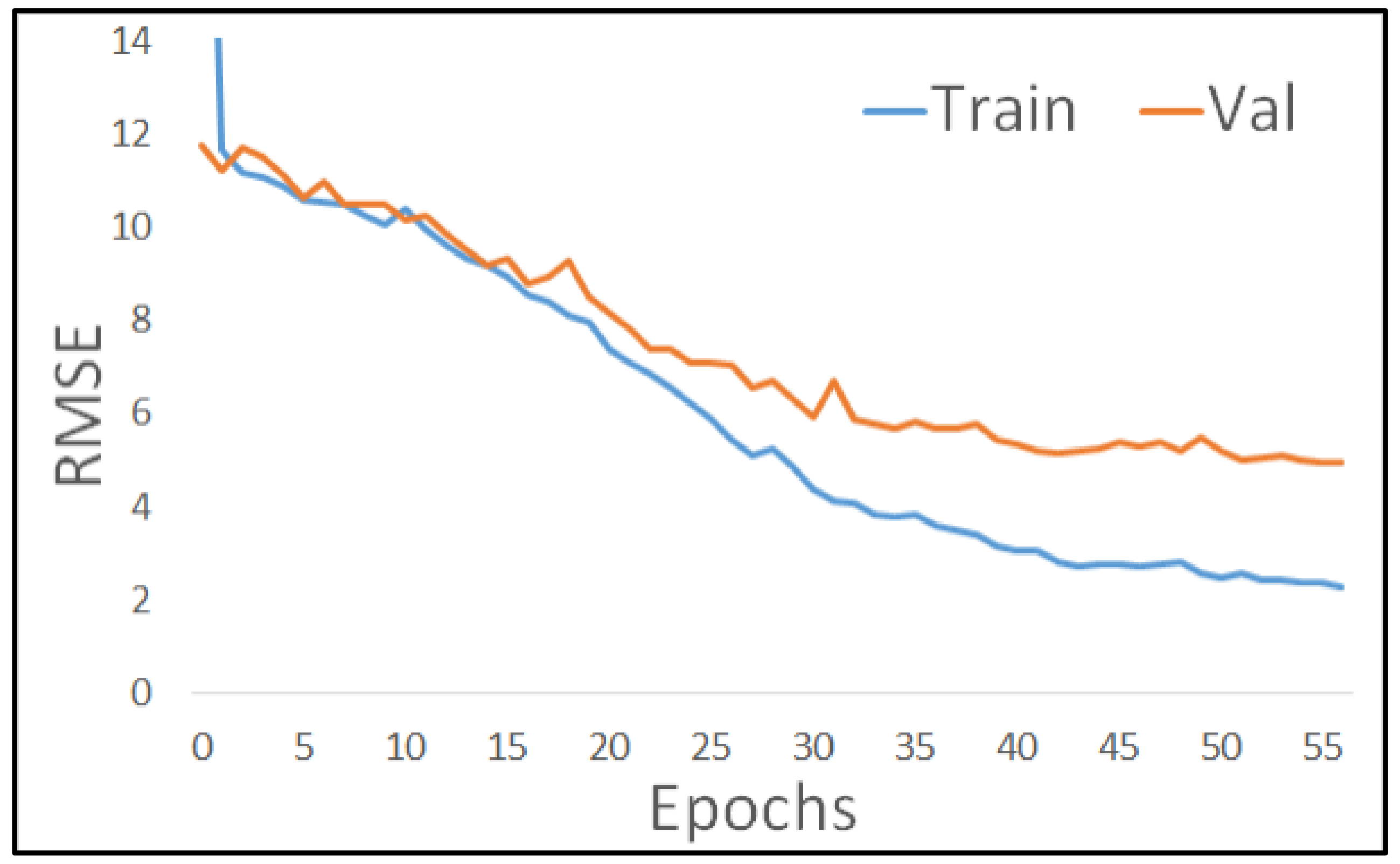
Results

Experiments

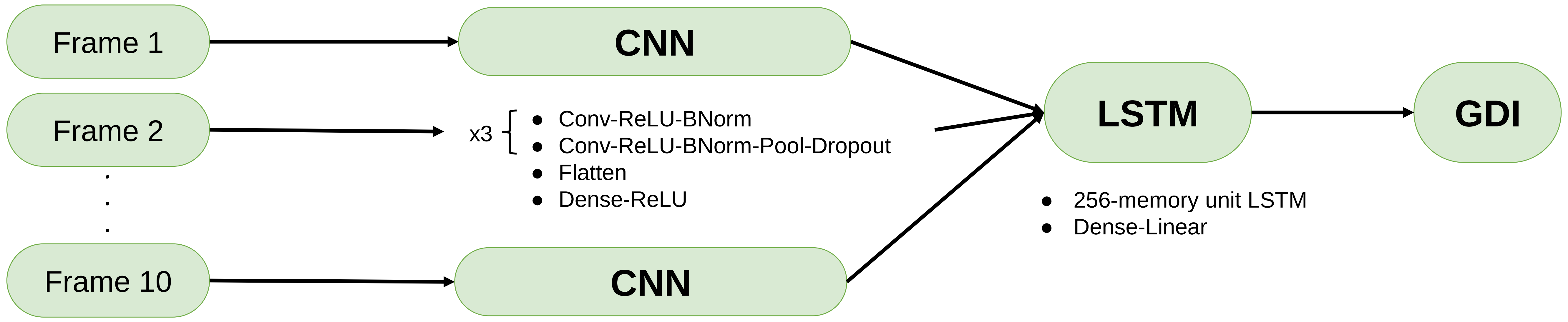
Models	Input Type	Training Error	Validation Error
Guess Mean	Frame	-	13.7
Linear Regression	Frame	-	13.0
VGG16	Frame	10.1	9.5
CNN	Frame	8.1	8.2
CNN + LSTM	**Video**	**2.3**	**4.9**
CNN + 1D-CNN	Video	4.2	11.3

Note: Due to massive data volume, some models were built on randomized subsets of data to avoid memory overload

Learning Curve



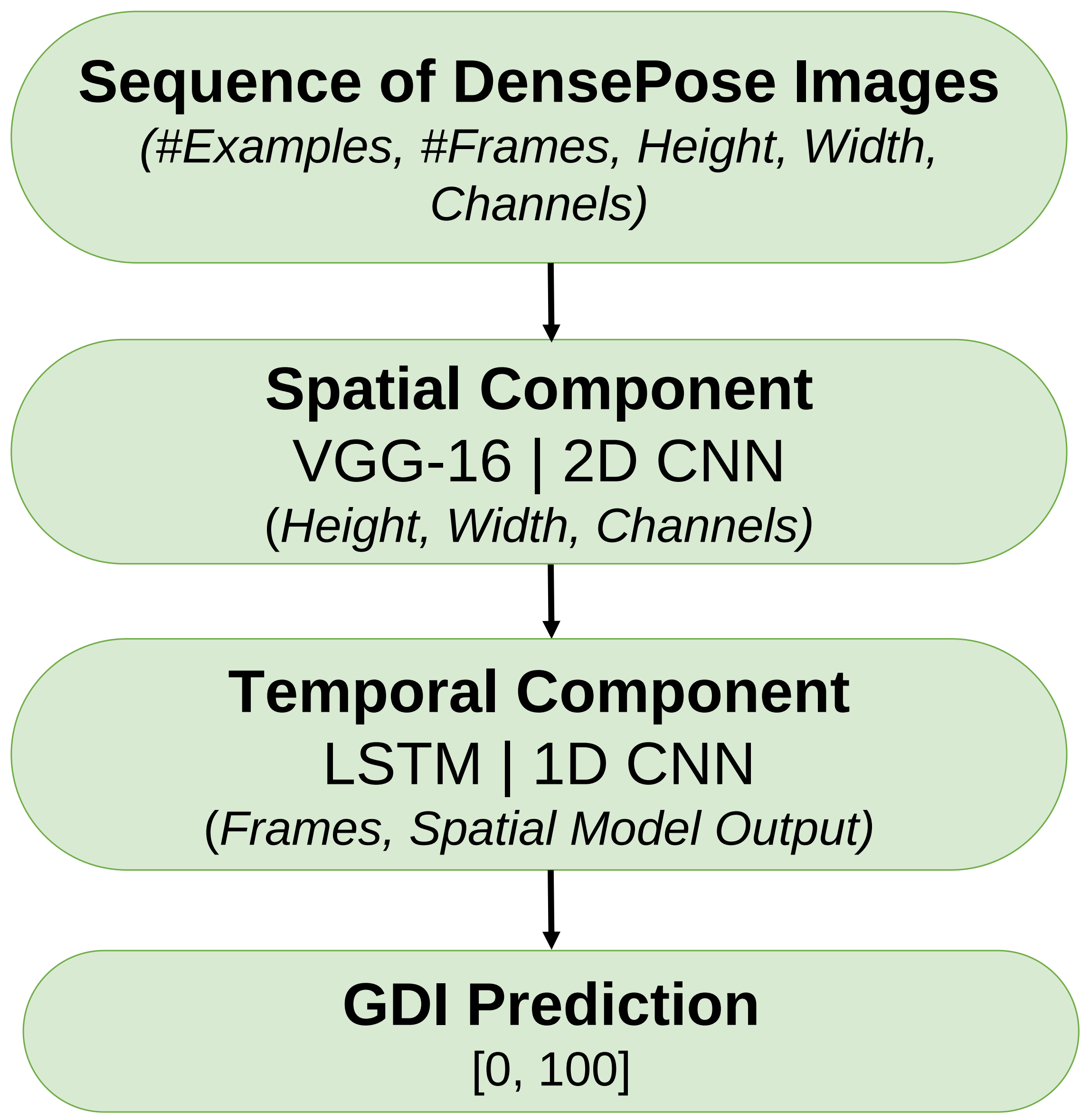
CNN + LSTM Architecture



- Conv-ReLU-BNorm
- Conv-ReLU-BNorm-Pool-Dropout
- Flatten
- Dense-ReLU

- 256-memory unit LSTM
- Dense-Linear

GDI-Net



Sequence of DensePose Images
(#Examples, #Frames, Height, Width, Channels)

Spatial Component
VGG-16 | 2D CNN
(Height, Width, Channels)

Temporal Component
LSTM | 1D CNN
(Frames, Spatial Model Output)

GDI Prediction
[0, 100]

Future Work

Our results are promising, and can be enhanced by:

- **Featurizing** input data to spatial X, Y, Z components using the SMPL human body model
- Incorporating **3D convolution blocks** in earlier layers to capture lower level temporal features
- Leveraging **chunking** and other memory-saving methods to train on a larger dataset
- Training a **classification model** and take probability-weighted average of class values to calculate GDI
- Implementing **grid search** for systematic hyperparameter tuning

References

Hanson, Nick. “Kids Health Matters.” Gillette Children's Specialty Care, 1 Feb. 2017, www.gillettechildrens.org/khm/topics/gait-analysis.

Rıza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. arXiv:1802.00434, 2018.

Author Affiliations:
[1]Dept. of Mechanical Engineering
[2]Stanford Center for Professional Development
[3]Dept. of Aeronautics & Astronautics