Instituto de Matemática
Pura e Aplicada

2D and 3D Pose Estimation

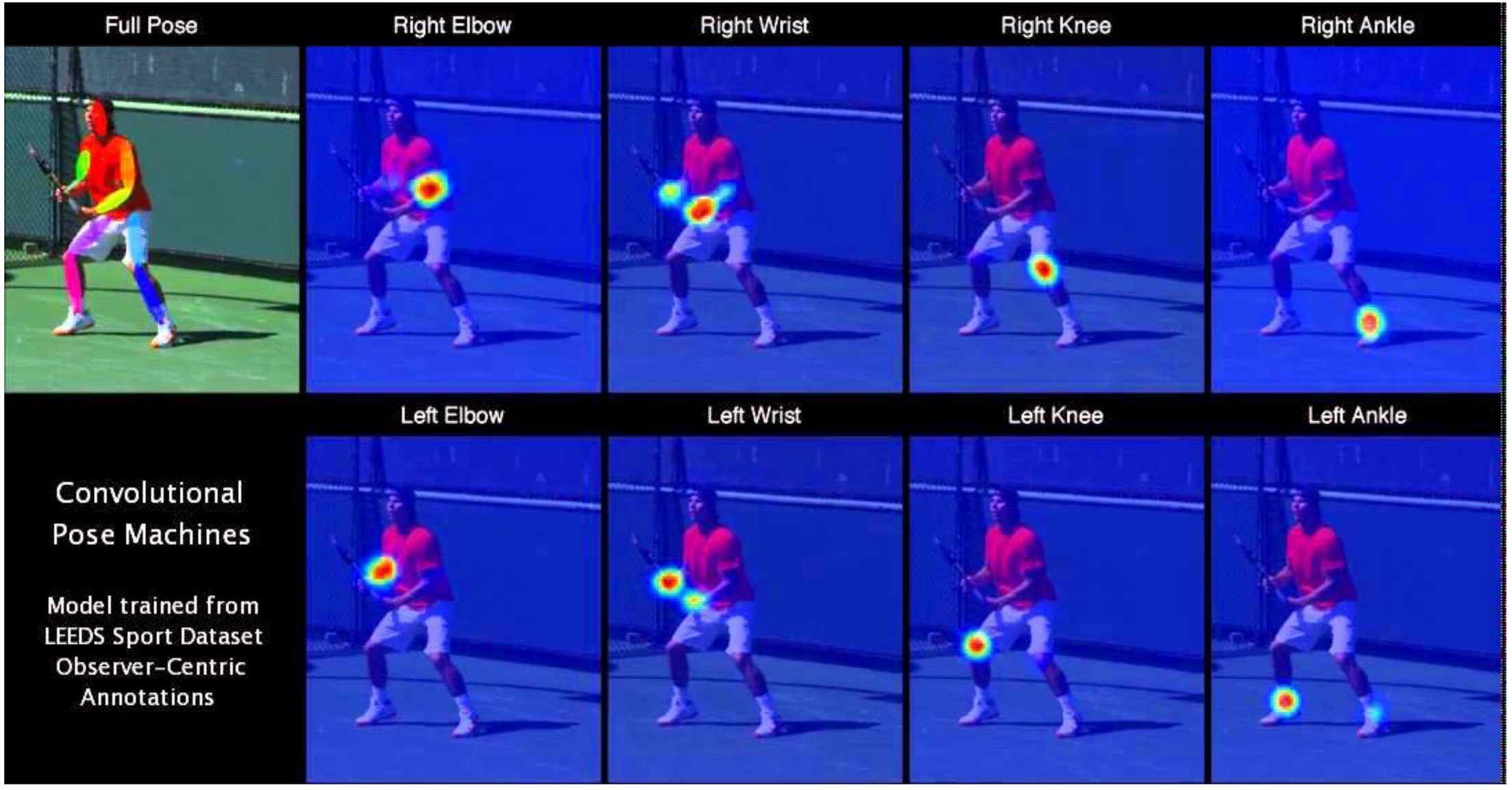
Visgraf Vision and
Graphics
Laboratory

Luiz José Schirmer Silva
Djalma Lúcio Soares da Silva

Pose Estimation

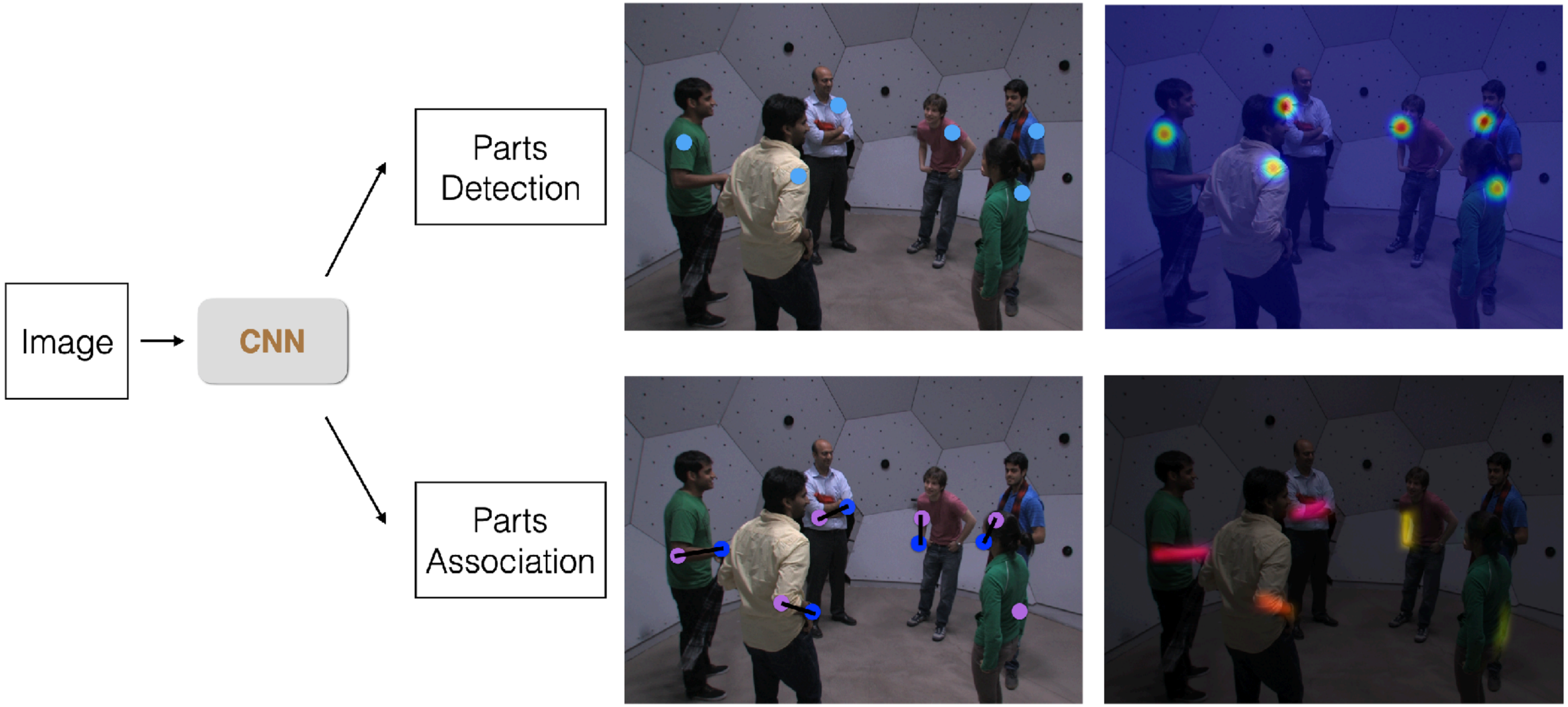
Customer 3
Aipoly

Convolutional Pose Machines

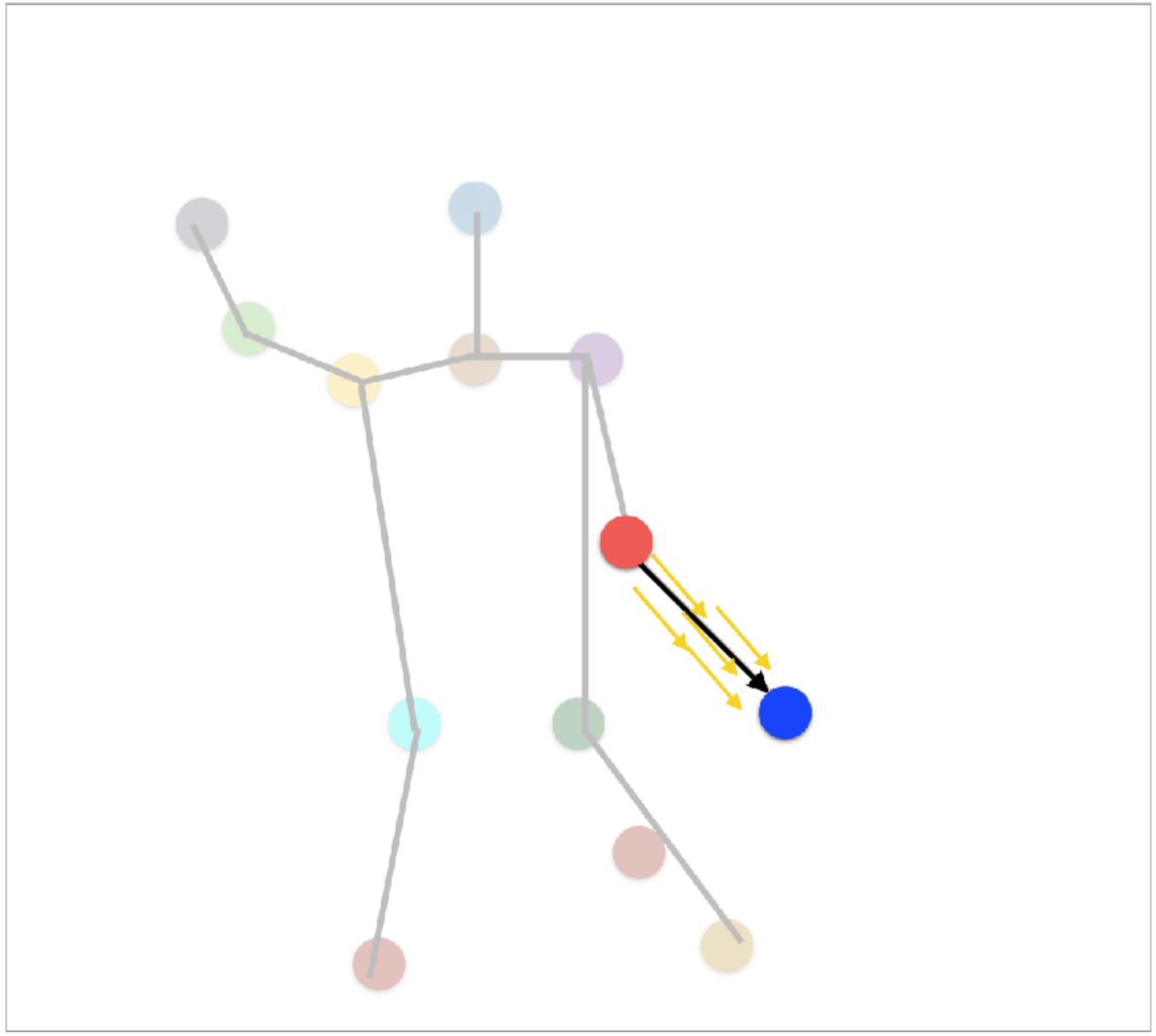

Full Pose
Right Elbow
Right Wrist
Right Knee
Right Ankle
Left Elbow
Left Wrist
Left Knee
Left Ankle
Convolutional Pose Machines
Model trained from LEEDS Sport Dataset Observer-Centric Annotations

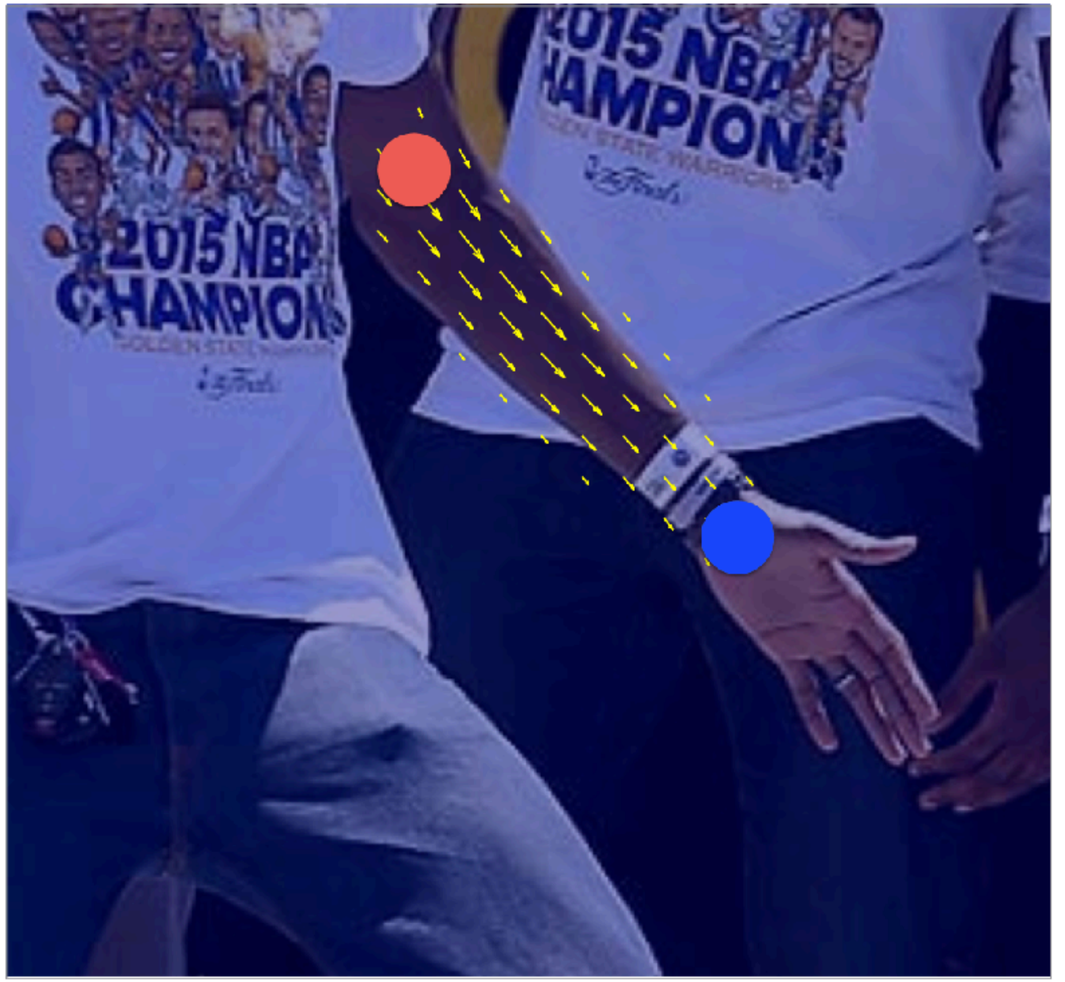
OpenPose

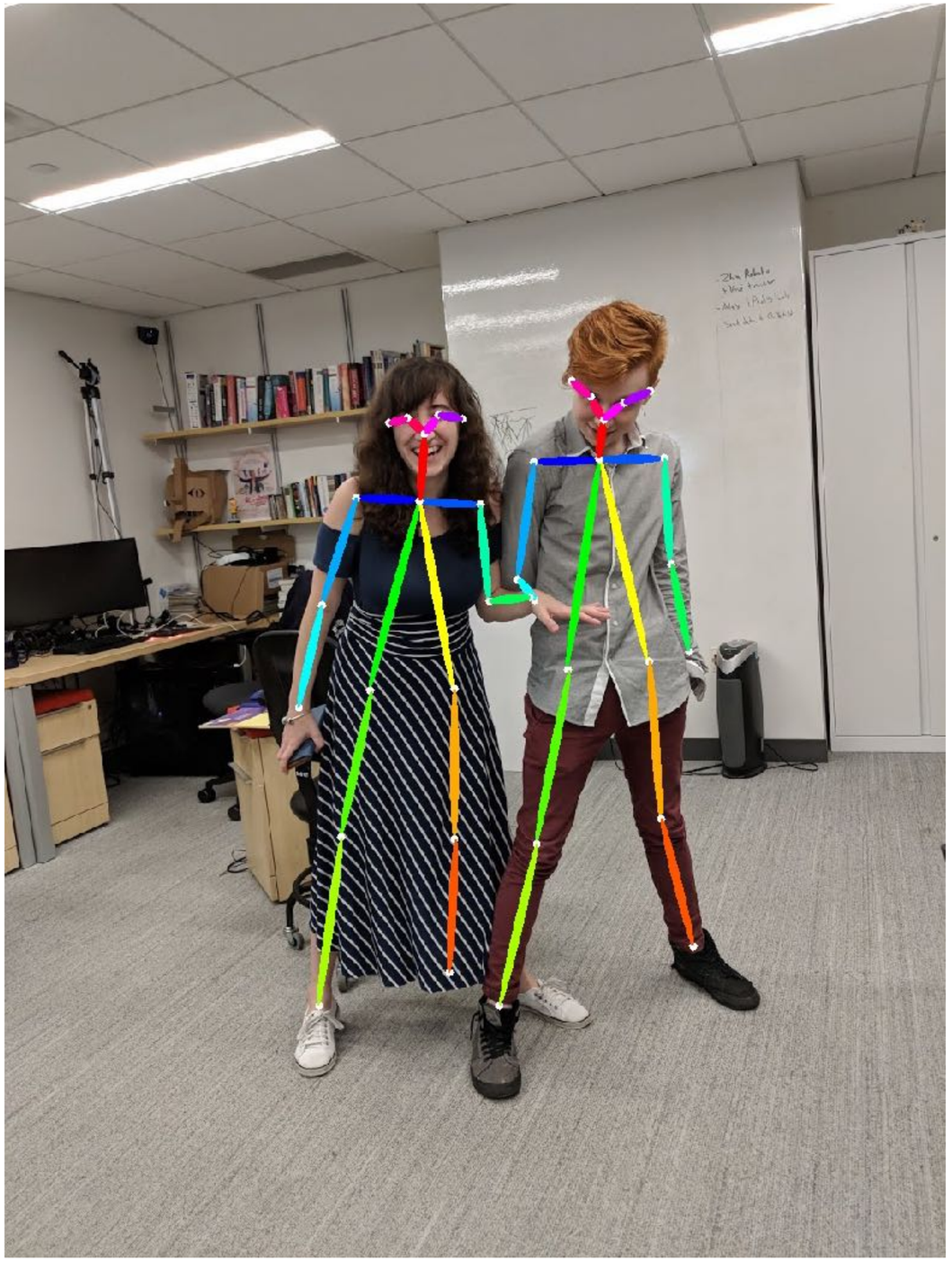
Novelty: Jointly Learning Parts Detection and Parts Association



Part Affinity Fields for Part-to-Part Association







The state-of-art approach on multi-person pose estimation does not achieve the desired performance in terms of frames per second

Also CNNs are over-parametrizated and it leads to a great amount of redundancy and increasing computation requirements

Our Proposal

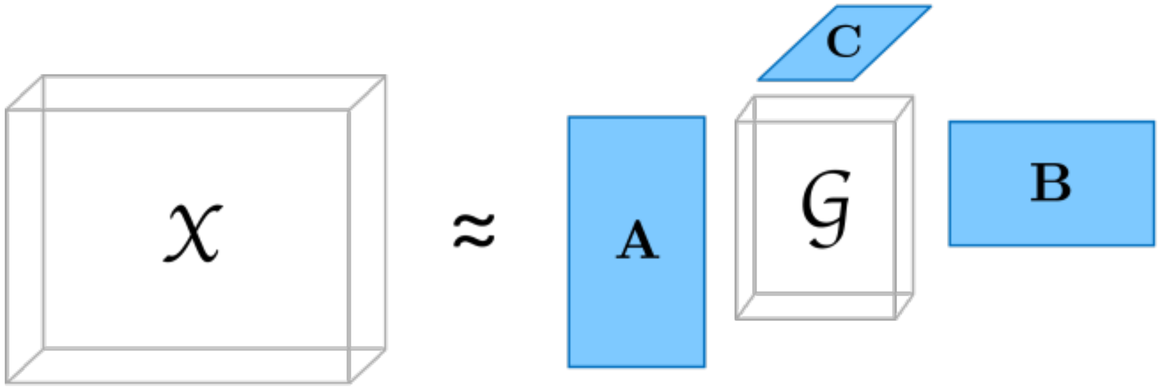
A streamlined architecture with tensor decomposition for pose estimation with improved processing time;

leveraging the redundancy in the parameters of Convolutional Pose Machine;

improve accuracy using attention mechanisms.

Our Technique

Tensor Decomposition



Tucker decompositon, a.k.a High Order SVD

- rising interest in exploring efficient architectures for Neural Networks;
- Mobilenets, Factorized Convolutions, Flattened Convolutional Neural Networks;
- Main theory related to tensor decompositions and multilinear algebra methods;
- A 2d convolutional layer can be described as a 4-dimensional matrix;
- High Order SVD.

CONVOLUTION

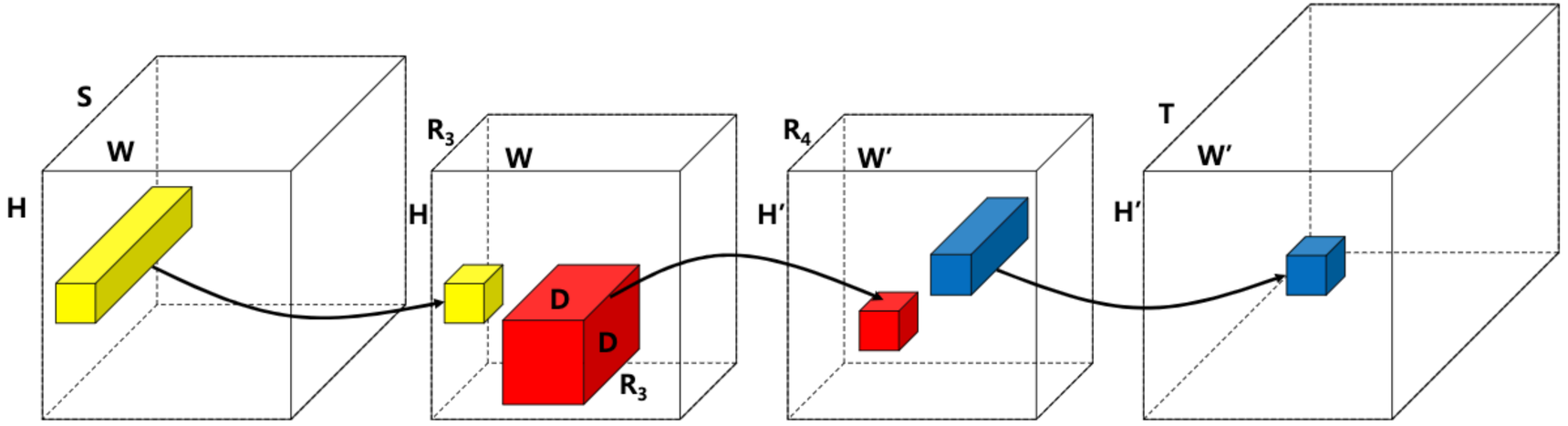
$$con(x, y, z) = \sum_i \sum_j \sum_z \Theta(i, j, k, z)\ \mathscr{W}(x - i, y - j, k)$$

where θ is a kernel of size i,j,k,z (for example, a kernel with size i x j = 3x3, with k=3 channels (RGB), and z=128 convolution filters) and W an input tensor with size XYK, as we refer typically as an image, for example, with X and Y the image dimensions and K the number of channels

Factorized Kernels

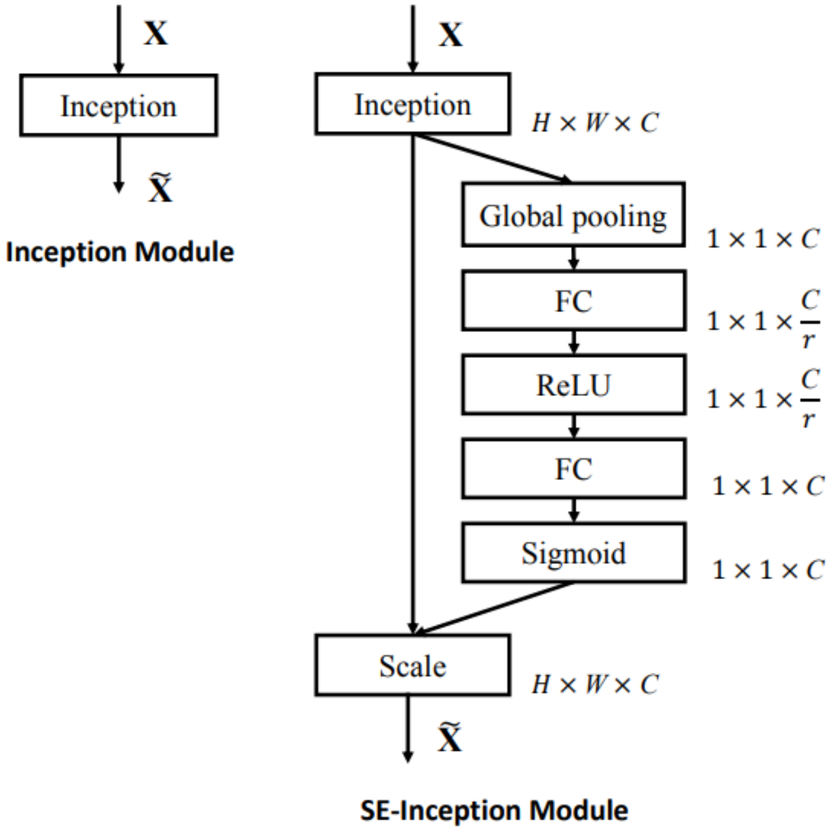
$$\text{Approx} = \sum_{r=1}^{R3} \sum_{r=1}^{R4} C_{i,j,r3,r4} U^{K}_{r3}(k) U^{z}_{r4}(z)$$

where $U_{k,r3}$ and $U_{Z,r4}$ are the factor matrices of sizes KxR3 and ZxR4, respectively, and C is s a core tensor of size (I,J,R3,R4)



Tucker2 decomposition:
Yellow tube and blue tube correspond to 1×1 pointwise convolution
Red tube is a DXD convolution

SE - Networks



- weakness of convolutional neural networks is because convolution operations consider only local neighbor-hoods thus missing global information
- introduce an attention block for CNNs that improves channel interdependencies
- Almost no extra computational cost;
- we fully capture channel-wise dependencies

SE - Networks

CHANNEL-WISE ATTENTION

$$z_c = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} X_c(i,j),$$

S queeze Operation

$$\alpha = sigmoid(W_2(\beta(W_1(z)))).$$

capture channel-wise dependencies

$$\mathcal{X}^{'} = \alpha \circ \mathcal{X}$$

Scale

SPATIAL ATTENTION

$$\beta = sigmoid(W\mathcal{X}),$$

Pointiwise convolution with sigmoid

$$x^{'}_{i,j} = \beta_{i,j} * x_{i,j},$$

Scale

SE - Networks

CHANNEL-WISE ATTENTION

$$z_c = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} X_c(i,j),$$

Squeeze Operation

$$\alpha = sigmoid(W_2(\beta(W_1(z)))).$$

capture channel-wise dependencies

$$\mathcal{X}' = \alpha \circ \mathcal{X}$$

Scale

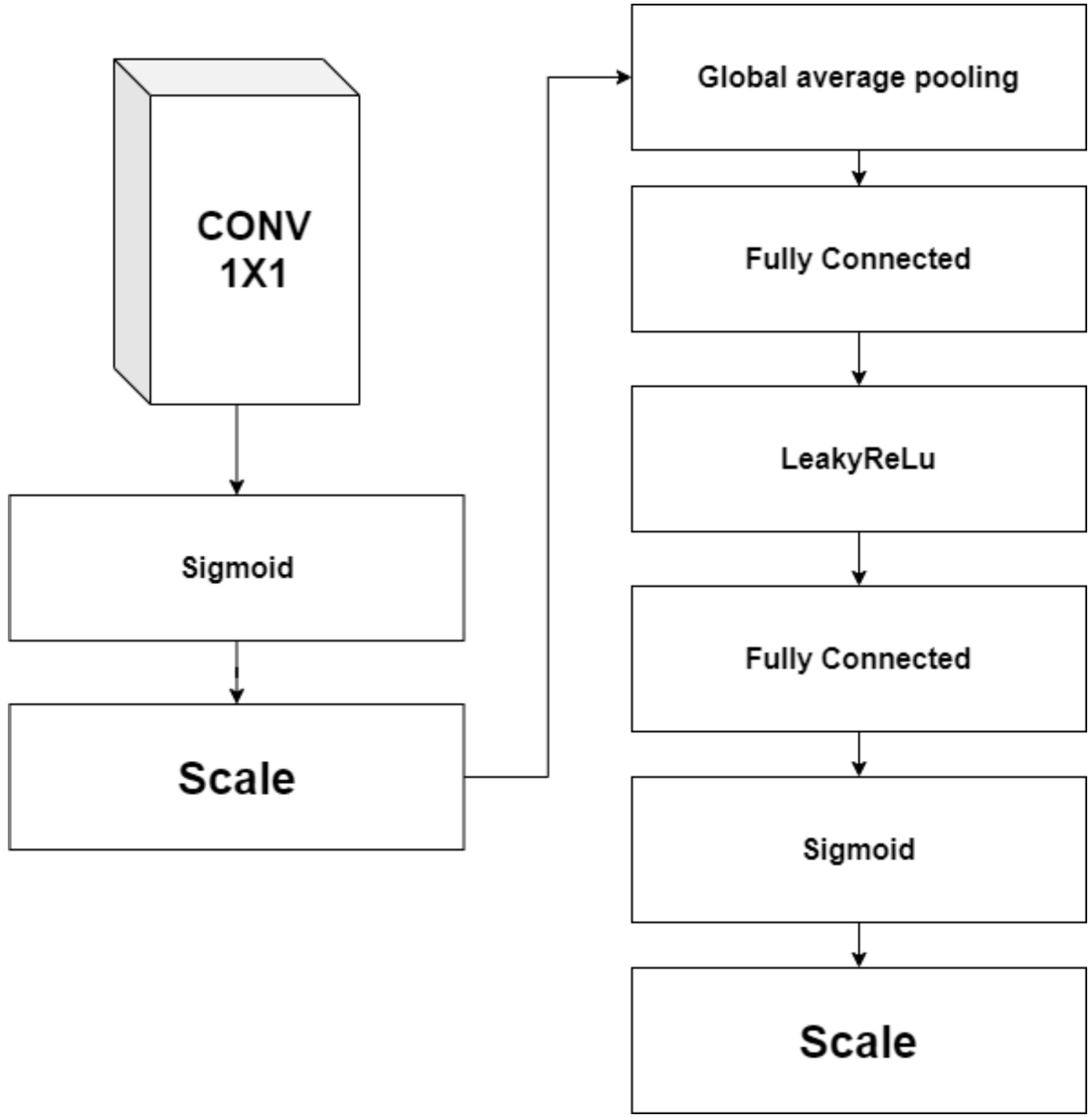
SPATIAL ATTENTION

$$\beta = sigmoid(W\mathcal{X}),$$

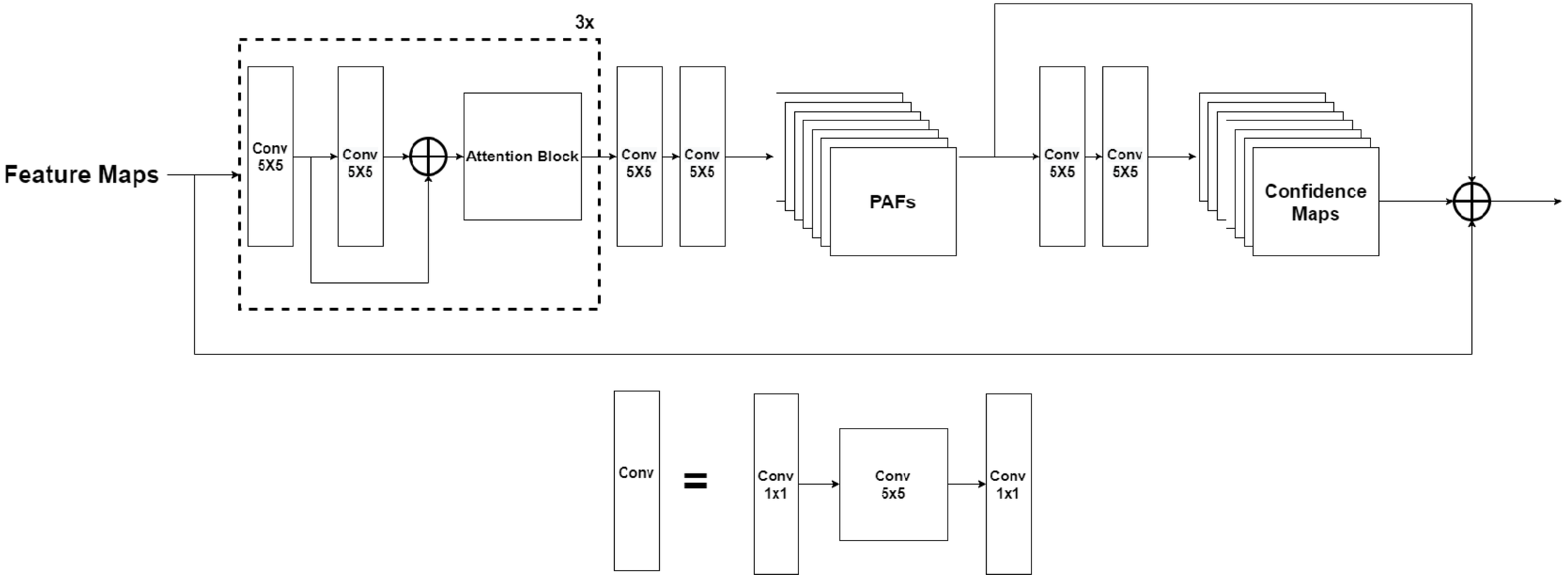
Pointiwise convolution with sigmoid

$$x'_{i,j} = \beta_{i,j} * x_{i,j},$$

Scale



Architecture



Each Factorized Layer: 9.3 less operations than regular convolutions

Architecture

$$f^i_{key} = \sum_{k=1}^{nkeys} \sum_{p} 1 - e^{-w(p)\left\|key^i_k - key^*_k\right\|_2},$$

$$f^i_{paf} = \sum_{k=1}^{paf} \sum_{p} 1 - e^{-w(p)\left\|paf^i_k - paf^*_k\right\|_2},$$

$$loss = \sum_{i=0}^{6} f^i_{key} + f^i_{paf}$$

- Our model were trained over 120 epochs
- We use the Adam Optimizer
- Learning rate varying from 8e–5 to 7.38e–6
- We use a kernel distance function for our loss
- We weight the loss functions spatially to address a practical issue that some datasets do not completely label all people. Here, key and paf represents the inferred maps for each keypoints and part affinty field respectively and the w is a binary mask were w(p) = 0 when the annotation is missing at the pixel p

Results

Comparison with OpenPose (COCO dataset)

OpenPose

Precision: 0.78 Recall: 0.807

Our

Precision: 0.743 Recall: 0.702

Tests in GPU and CPU

Image Resolution	OpenPose	Ours
328 x 193	~55,08 ms	~ 45 ms
656 x 368	~120.224 ms	~ 51.26 ms
984 x 579	~174,75 ms	~ 80.72 ms
1312 x 772	~320,18 ms	~ 112 ms

TABLE I

CNN PROCESSING TIME FOR OPENPOSE AND OUR MODEL. WE VARY THE SCALE OF THE INPUT TESTED BY A FACTOR OF 0.5 CONSIDERING 4 IMAGE SCALES. WE DO THIS ONLY FOR THE NETWORK INPUT, THE FINAL RESULT IS SCALED IN THE ORIGINAL IMAGE SIZE.

Device	OpenPose	Ours
AMD Ryzen 7 1700 3.5GHZ	0.3 FPS	13 FPS
Mac Pro Intel Core I7 2.7GHZ	0.1 FPS	6 FPS

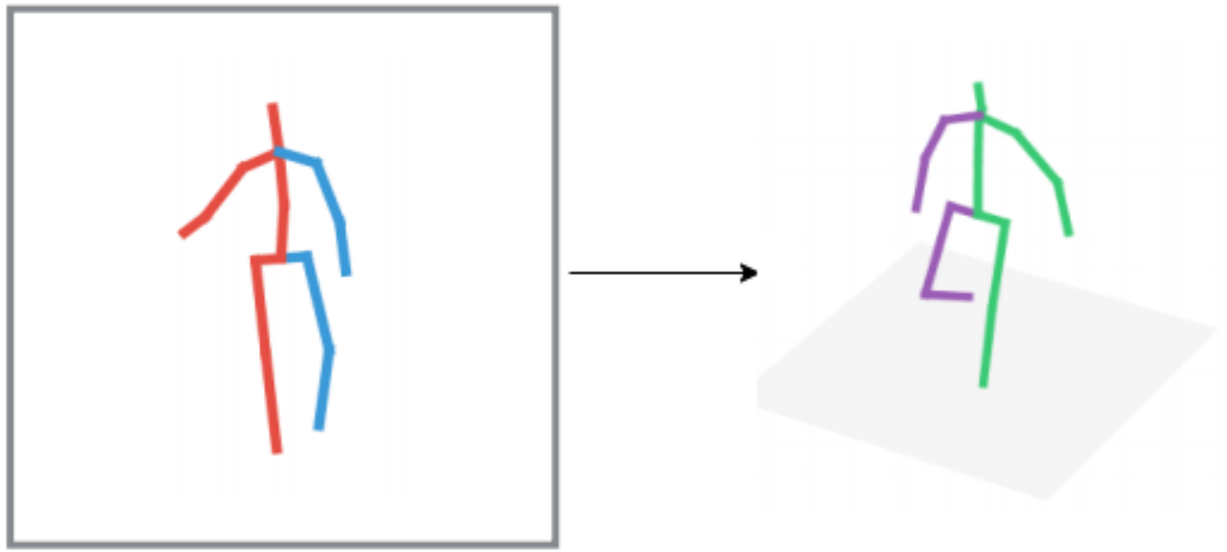
TABLE II

CNN PROCESSING TIME FOR OPENPOSE AND OUR MODEL IN CPU.

Tests with COCO Dataset

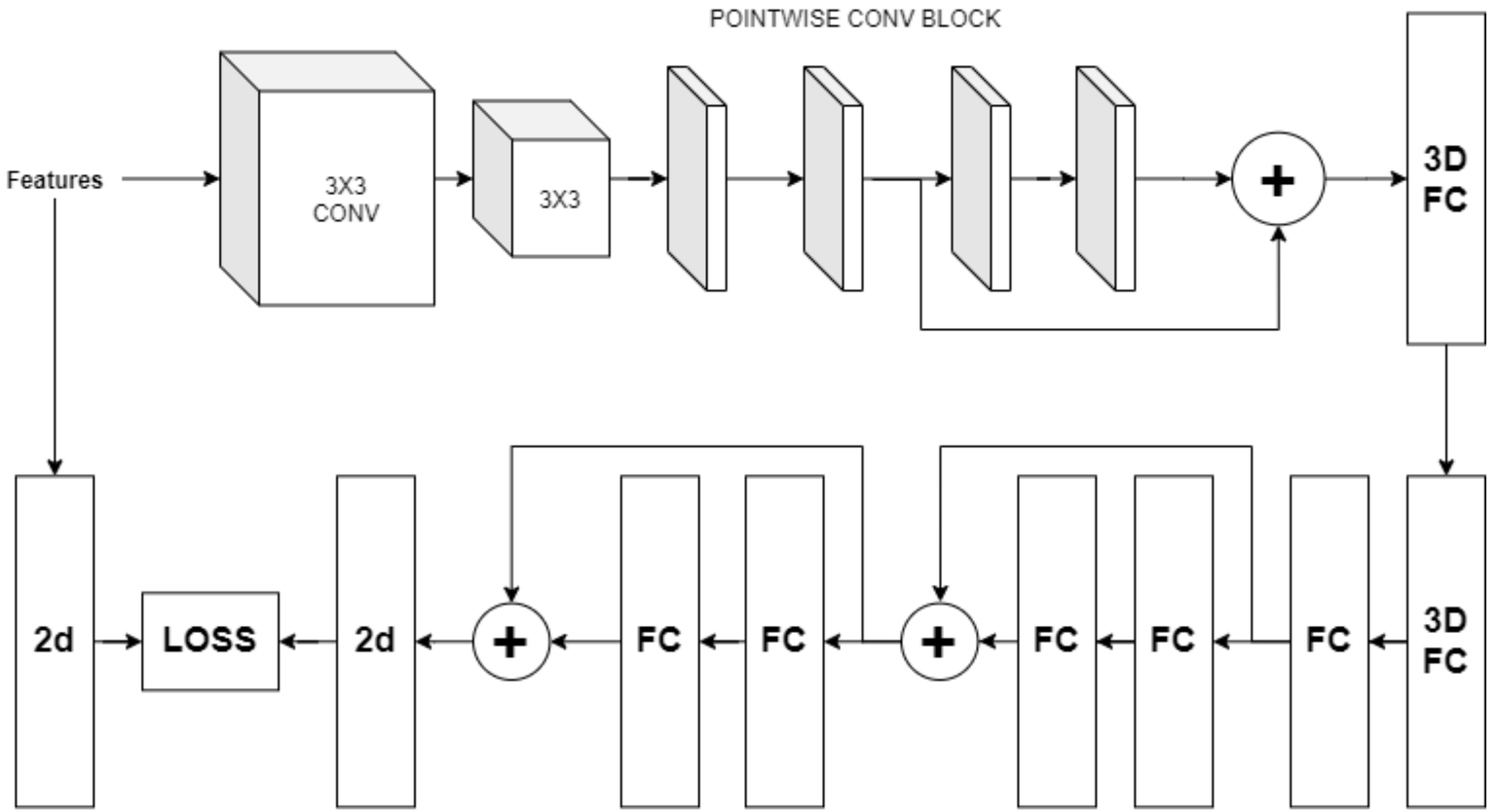
3D Pose Estimation

2D to 3D

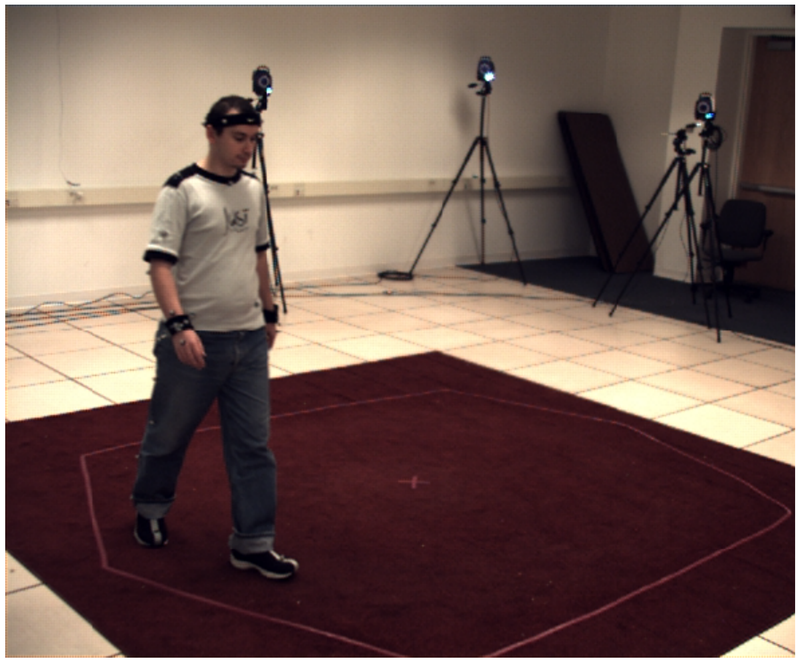


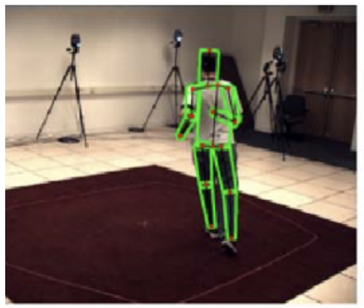
$$f^* = \min_f \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f(\mathbf{x}_i) - \mathbf{y}_i\right).$$

Architecture

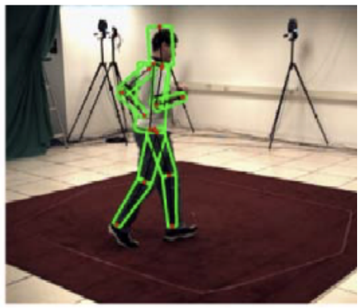


Datasets – HumanEVA and Human 3.6

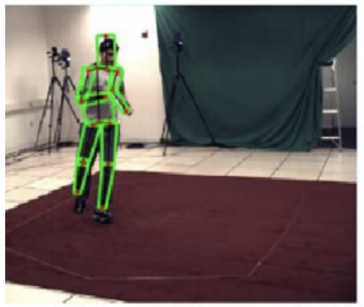




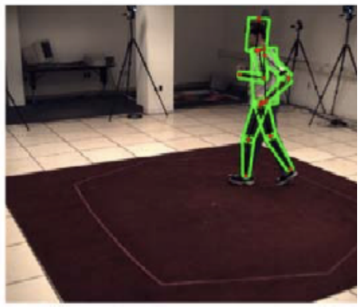
C1



C2

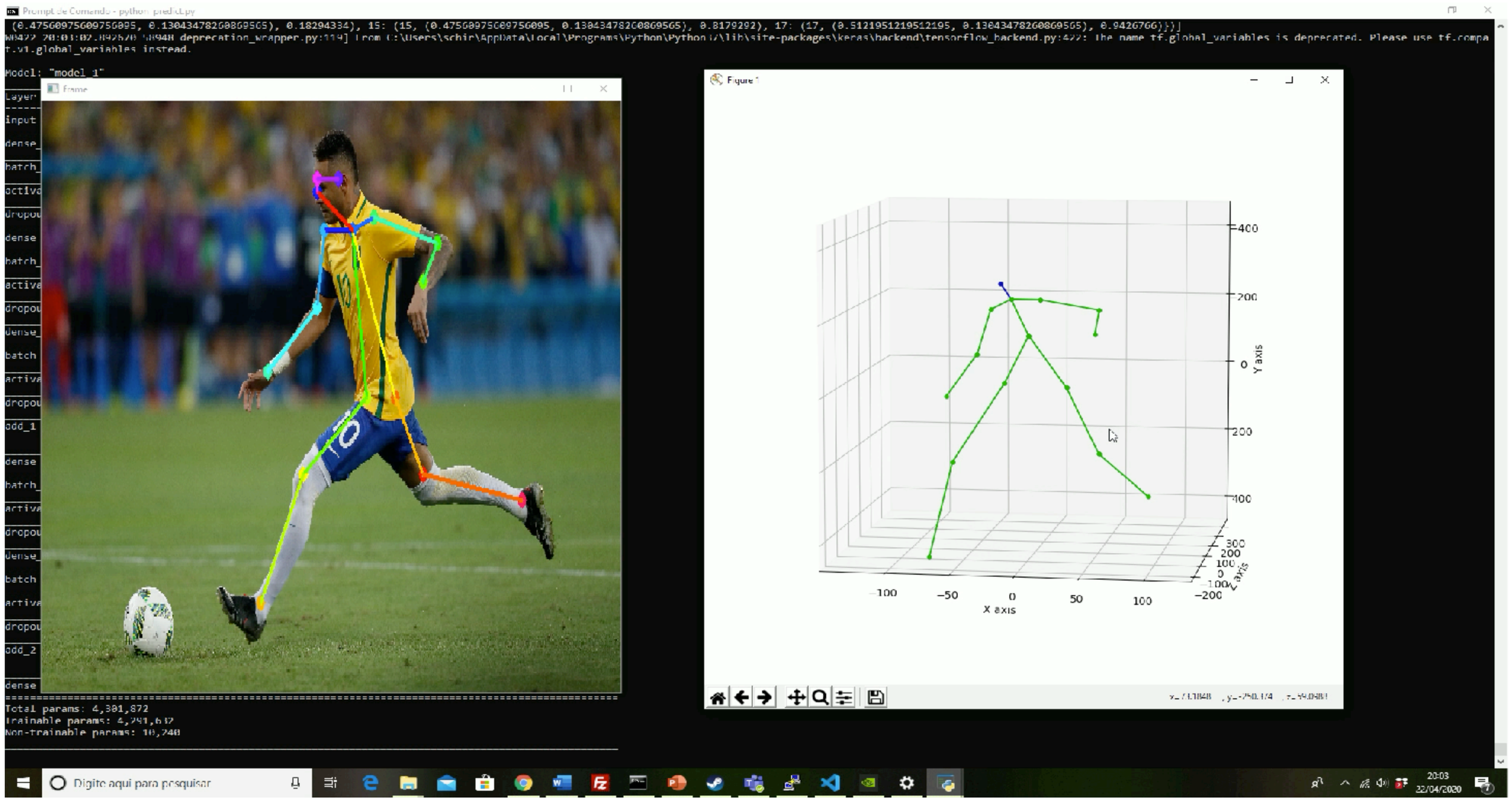


C3

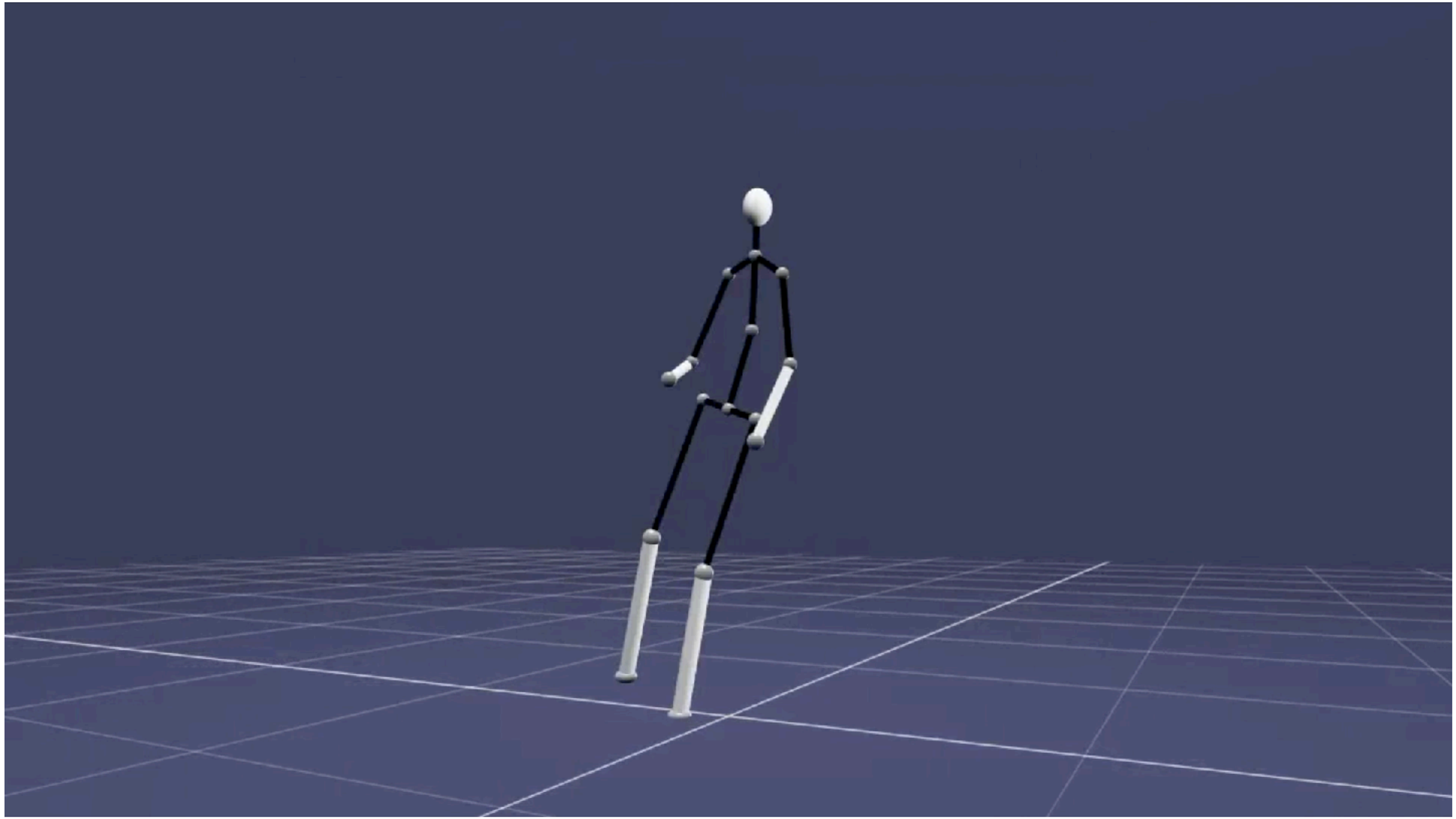


C4

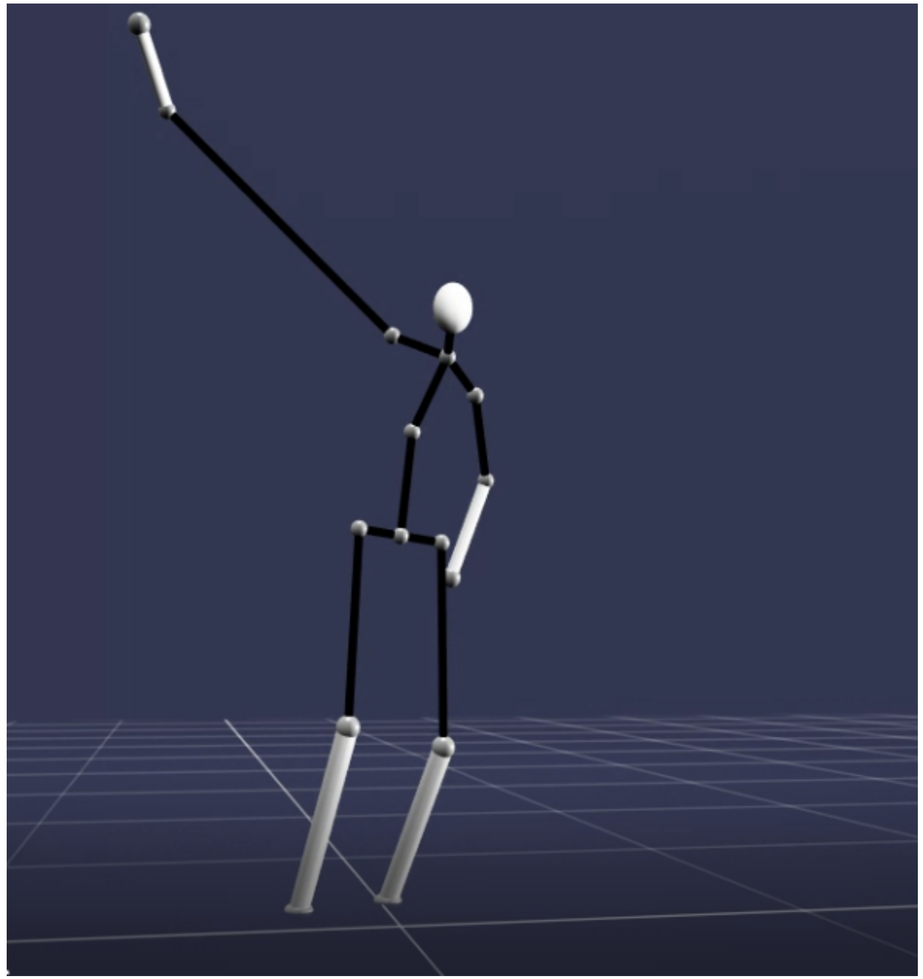
Tests

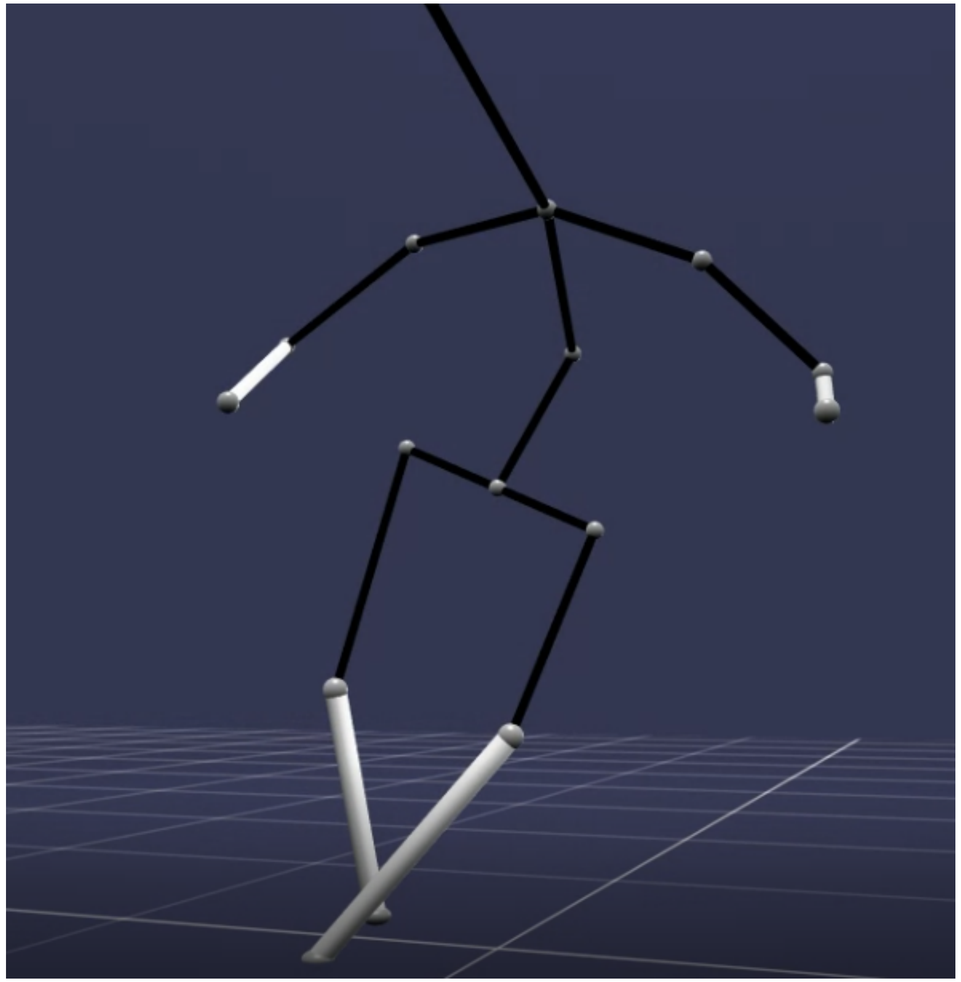

Prompt de Comando - python predict.py
Model: "model_1"
Frame
Figure 1
Y axis
X axis
Z axis
Total params: 4,381,872
Trainable params: 4,291,672
Non-trainable params: 90,200
Digite aqui para pesquisar
20:03
22/04/2020

Tests



Tests





Texture Capture and 3D model

An end-to-end framework for 3D capture and human digitization with a single RGB camera

Luiz Schirmer, Djalma da Silva, Luiz Velho and Hélio Lopes

Thank You!

lschirmer@inf.puc-rio.br

References

[1496Senst2016] SENST, T.; GEISTERT, J. ; SIKORA, T.. Robust local optical flow: Long-range motions and varying illuminations. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING, p. 4478–4482, Phoenix, AZ, USA, Sept. 2016. IEEE. IEEE Catalog Number: CFP16CIP-USB ISBN: 978-1-4673-9960-9 DOI:10.1109/ICIP.2016.7533207.

[babenko2009visual] BABENKO, B.; YANG, M.-H. ; BELONGIE, S.. Visual tracking with online multiple instance learning. In: 2009 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 983–990. IEEE, 2009.

[bradski2008learning] BRADSKI, G.; KAEHLER, A.. Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc.", 2008.

[burbano20163d] BURBANO, A.; VASILIU, M. ; BOUAZIZ, S.. 3d cameras benchmark for human tracking in hybrid distributed smart camera networks. In: PROCEEDINGS OF THE 10TH INTERNATIONAL CONFERENCE ON DISTRIBUTED SMART CAMERA, p. 76–83. ACM, 2016.

[cao2017realtime] CAO, Z.; SIMON, T.; WEI, S.-E. ; SHEIKH, Y.. Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR, 2017.

[chui2017kalman] CHUI, C. K.; CHEN, G. ; OTHERS. Kalman filtering with Real-Time Applications. Springer, 2017.

[cichocki2009nonnegative] CICHOCKI, A.; ZDUNEK, R.; PHAN, A. H. ; AMARI, S.-I.. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons, 2009.

[de2000multilinear] DE LATHAUWER, L.; DE MOOR, B. ; VANDEWALLE, J.. A multilinear singular value decomposition. SIAM journal on Matrix Analysis and Applications, 21(4):1253–1278, 2000.

References

[ge2018real] GE, L.. Real-time 3D hand pose estimation from depth images. PhD thesis, 2018.

[gupta2017indoor] GUPTA, T.; LI, H.. Indoor mapping for smart cities—an affordable approach: Using kinect sensor and zed stereo camera. In: 2017 INTERNATIONAL CONFERENCE ON INDOOR POSITIONING AND INDOOR NAVIGATION (IPIN), p. 1–8. IEEE, 2017.

[howard2017mobilenets] HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M. ; ADAM, H.. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[jin2014flattened] JIN, J.; DUNDAR, A. ; CULURCIELLO, E.. Flattened convolutional neural networks for feedforward acceleration. arXiv preprint arXiv:1412.5474, 2014.

[kanazawa2018end] KANAZAWA, A.; BLACK, M. J.; JACOBS, D. W. ; MALIK, J.. End-to-end recovery of human shape and pose. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 7122–7131, 2018.

[kim2009realtime] KIM, J.-S.; HWANGBO, M. ; KANADE, T.. Realtime affinephotometric klt feature tracker on gpu in cuda framework. In: 2009 IEEE 12TH INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS, ICCV WORKSHOPS, p. 886–893. IEEE, 2009.

[kim2015compression] KIM, Y.-D.; PARK, E.; YOO, S.; CHOI, T.; YANG, L. ; SHIN, D.. Compression of deep convolutional neural networks for fast and low power mobile applications. arXiv preprint arXiv:1511.06530, 2015.

[kolda2009tensor] KOLDA, T. G.; BADER, B. W.. Tensor decompositions and applications. SIAM review, 51(3):455–500, 2009.

[lin2014microsoft] LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P. ; ZITNICK, C. L.. Microsoft coco: Common objects in context. In: EUROPEAN CONFERENCE ON COMPUTER VISION, p. 740–755. Springer, 2014.

[lin2017recurrent] LIN, M.; LIN, L.; LIANG, X.; WANG, K. ; CHENG, H.. Recurrent 3d pose sequence machines. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 810–819, 2017.

[loper2015smpl] LOPER, M.; MAHMOOD, N.; ROMERO, J.; PONS-MOLL, G. ; BLACK, M. J.. Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG), 34(6):248, 2015.

[luo2018lstm] LUO, Y.; REN, J.; WANG, Z.; SUN, W.; PAN, J.; LIU, J.; PANG, J. ; LIN, L.. Lstm pose machines. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 5207–5215, 2018.

[masson2017holo] MASSON, T.; PERLIN, K. ; OTHERS. Holo-doodle: an adaptation and expansion of collaborative holojam virtual reality. In: ACM SIGGRAPH 2017 VR VILLAGE, p. 9. ACM, 2017.

[mehta2017vnect] MEHTA, D.; SRIDHAR, S.; SOTNYCHENKO, O.; RHODIN, H.; SHAFIEI, M.; SEIDEL, H.-P.; XU, W.; CASAS, D. THEOBALT, C..Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG), 36(4):44, 2017.

[ramakrishna2014pose] RAMAKRISHNA, V.; MUNOZ, D.; HEBERT, M.; BAGNELL, J. A. ; SHEIKH, Y.. Pose machines: Articulated pose estimation via inference machines. In: EUROPEAN CONFERENCE ON COMPUTER VISION, p. 33–47. Springer, 2014.

[ranjan2019hyperface] RANJAN, R.; PATEL, V. M. ; CHELLAPPA, R.. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41 (1):121– 135, 2019.

[ruggero2017benchmarking] RUGGERO RONCHI, M.; PERONA, P.. Benchmarking and error diagnosis in multi-instance pose estimation. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, p. 369–378, 2017.

[sandler2018mobilenetv2] SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A. ; CHEN, L.-C.. Mobilenetv2: Inverted residuals and linear bottlenecks. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 4510–4520, 2018.

[schwarcz20183d] SCHWARCZ, S.; POLLARD, T.. 3d human pose estimation from deep multi-view 2d pose. In: 2018 24TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION (ICPR), p. 2326–2331. IEEE, 2018.