9-Accurate detection of MOCAP Suit MI01722, D&L-2017

Prof. Rahul Garg (CSE) Prof. Varsha Singh (HUSS) Prof. Prathosh AP (DEE)

Shubham Mittal (2018EE10957) Dishant Dhiman (2019CS10347)

09 – 07 -2021

Contents

- Project Aim
- Earlier Work
- Detection of Motion Capture (MOCAP) Suit
- Future Work

Project Aim

- To track yoga asanas and generate skeleton in real time
- To provide accurate measurement of keypoints (or body joints) in real time
- Using a camera and special novel dress called MOCAP Suit

Earlier Work

- **Kinect**
- **2D Deep Learning Pipeline** (like **AlphaPose)**

  Tracking the Yoga asanas can be challenging due to unusual body extensions and high amount of self-occlusion in the frames.

  Both Kinect SDK and 2D Deep Learning pipelines (openPose, alphaPose) fails in such cases.

Earlier Work

Kinect

AlphaPose

MOCAP Suit

Algorithm:

- Suit comprises of red and blue patterns with different amounts of intensity.
- Thus, a 3x3 matrix as captured by a simple high resolution camera will be sufficient to uniquely label a point.

Applications:

- This is a real time high accuracy tracking which will improve the current motion capture process.
- This has high applications in sports, animation industry, filming, posture studies, recommender and selection systems.
- Cost of present day tracking is at least 1.5 lakh, with high precision goniometers and sensors. Our model aims to bring this down drastically along while increasing number of tracked points
- We also plan to submit a paper on the progress we have made so far.

Artificial Dataset

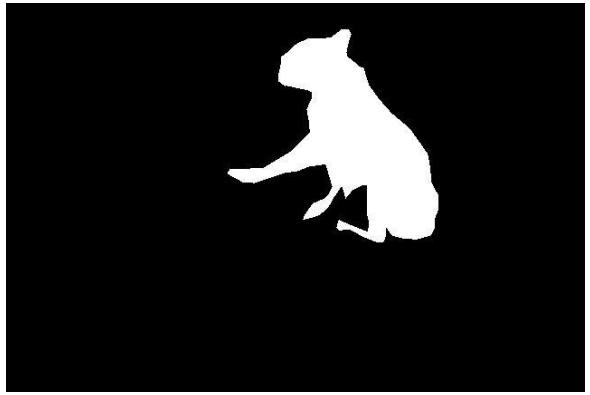
Why create Artificial Dataset:

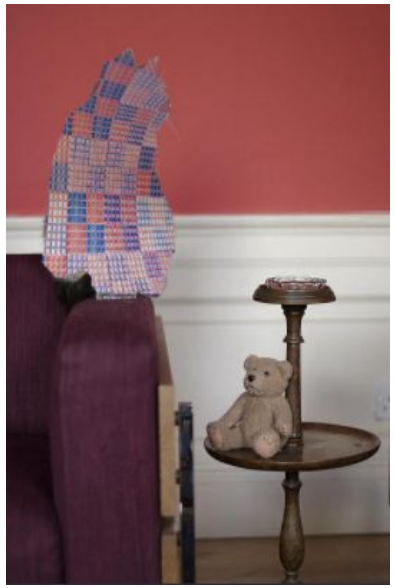
- DL models require a good amount of data to train upon.
- The data for CheckerBoard detection was not available openly and thus we had to create the data from scratch.

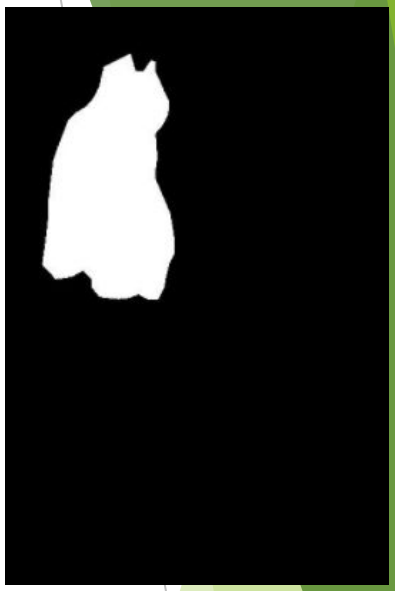
How was the data created:

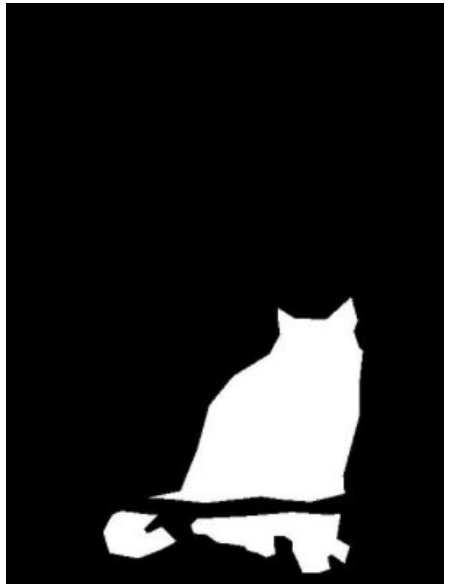
- We leverage the available COCO dataset (83K/13GB) to generate the checkerboard dataset.
- The annotations were used and the patterns were placed on subjects like humans, vehicles, animals, etc in the COCO dataset.
- Total images generated after this process were 23500.

Dataset-

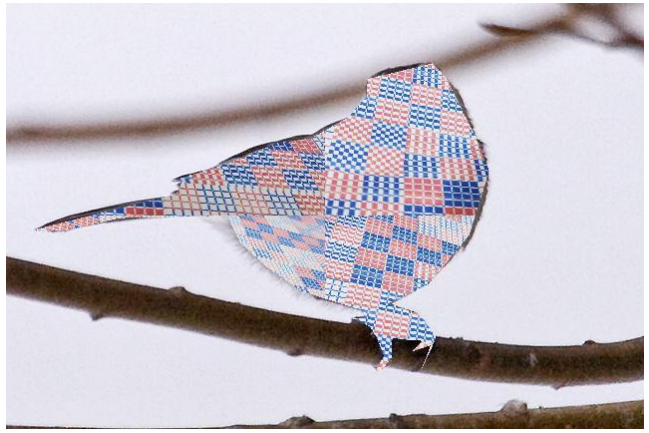


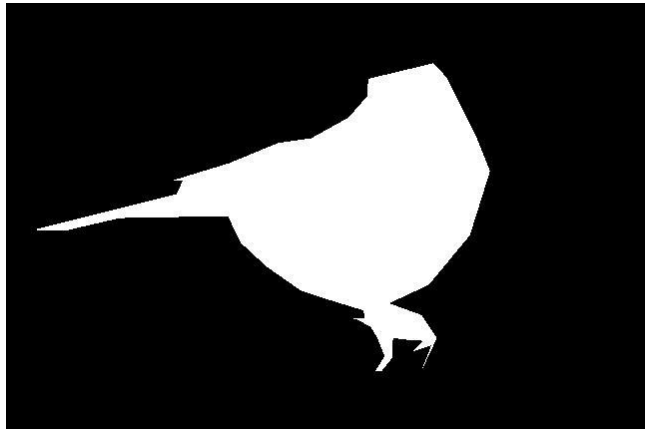


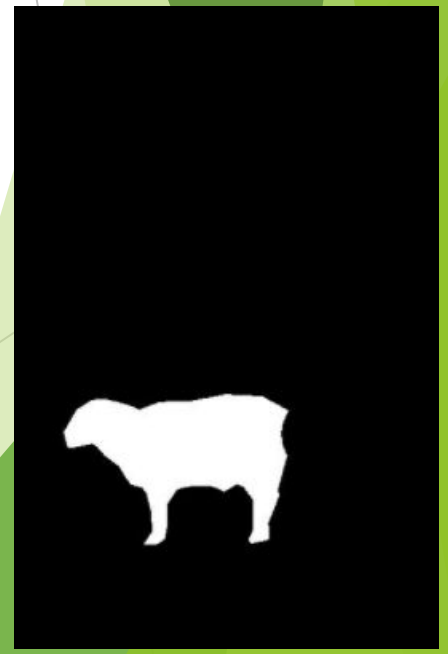




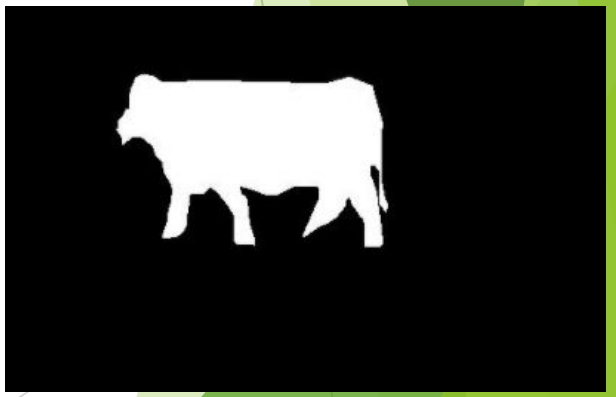
Dataset-



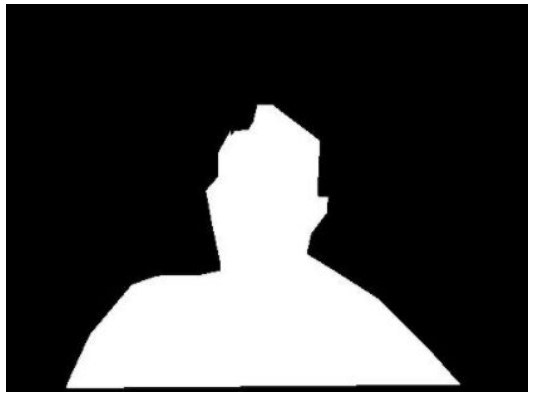


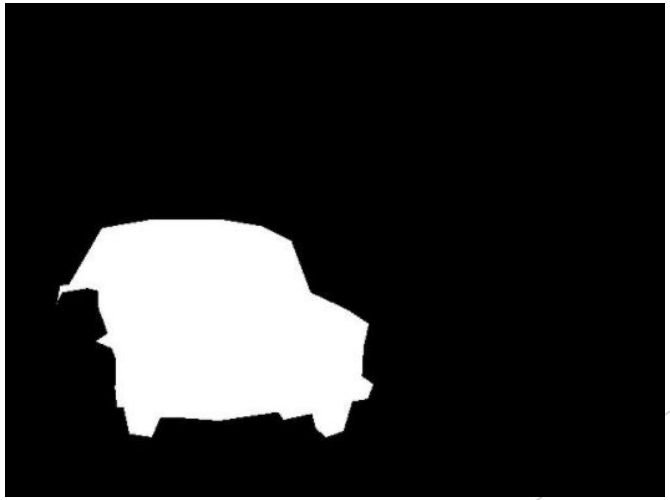


Dataset-



Dataset-

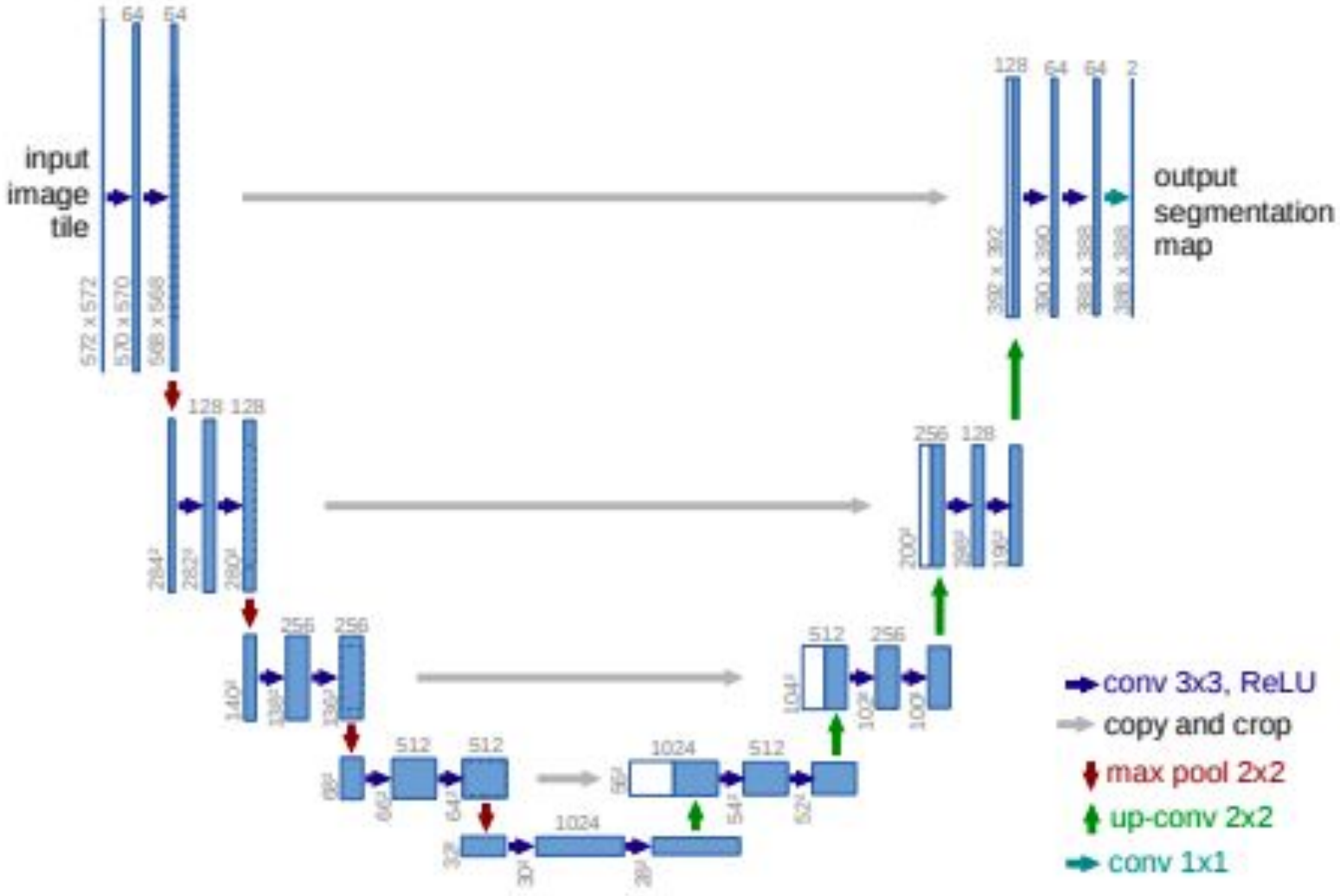




Deep Learning Model to extract MOCAP Suit

- Out of the 23500 images generated:
  - 16450 images were used as training set.
  - 4700 images were used as validation set.
  - 2350 images were used as test set.
- Multiple models were tried: **UNet**, **Attention UNet**, **ResUnet++** and the accuracy received for UNet model was the highest.
- The accuracy metric used was F1 Score.
- The code ran for 300 epochs and the final scores obtained by UNet on the sets were:
  - Training set: 96.47%
  - Validation set: 96.3%
  - Test set: **96.79%**

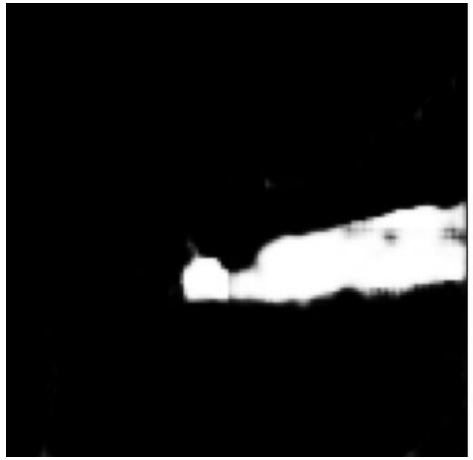
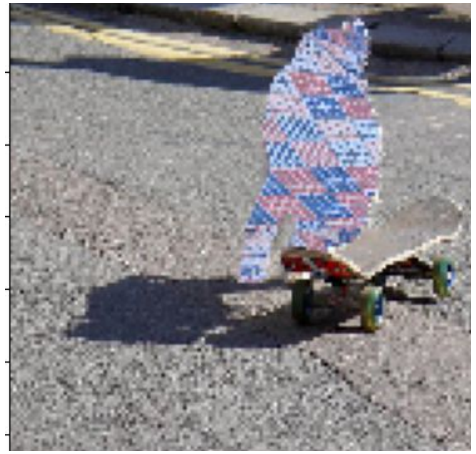
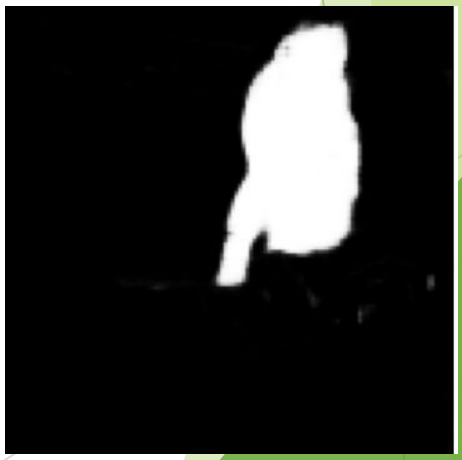
DL Model : UNet



UNet results





Note: The Original Images were 128*128. Zooming in on this slides leads to pixelation of images.

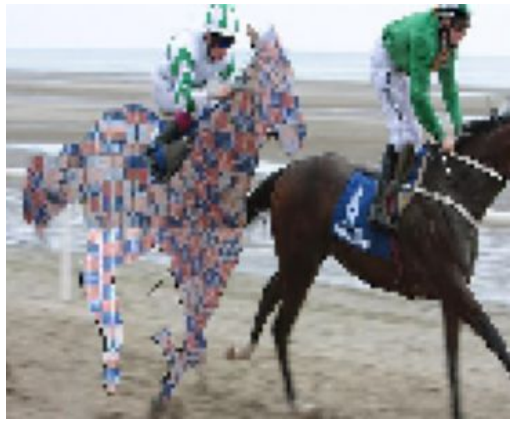
UNet results





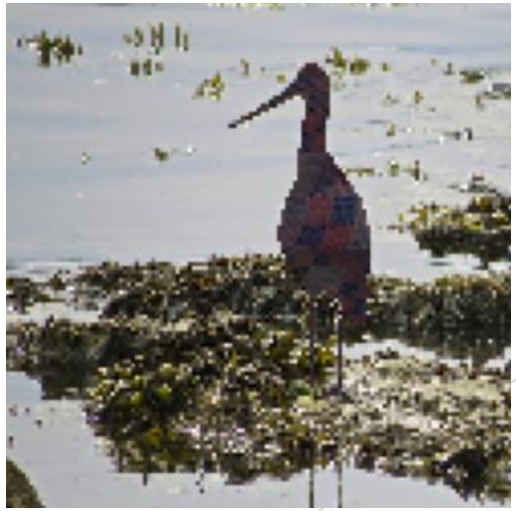



UNet results










UNet results







Future Work

- We aim to detect the key-points, or the joints accurately
- Pipeline : MOCAP Suit segmentation + Keypoint detection
- Cost reduced significantly as compared to previous work (KINECT).

References

- Martinez, J., Hossain, R., Romero, J., & Little, J. J., A simple yet effective baseline for **3d human pose estimation**, 2017
- Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2018). **OpenPose**: realtime multi-person 2D pose estimation using Part Affinity Fields., 2018
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. **U-Net**: Convolutional Networks for Biomedical Image Segmentation, 2015.
- Ozan Oktay, Jo Schlemper , Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. **Attention U-Net**: Learning Where to Look for the Pancreas, 2018.
- Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Dag Johansen, Thomas de Lange, Pal Halvorsen, Havard D. Johansen, **ResUNet++**: An Advanced Architecture for Medical Image Segmentation, 2019.

Appendix i: UNet Model Architecture Details

- The model is using UNet Architecture. The depth of UNet is 5, and the filter size at each depth is 2 ^ (5 + depth), 1 <= depth <= 5. The image passed in has dimensions of 3 * 128 * 128.
- While down sampling, channels are -> 3, 64, 128, 256, 512, 1024
- While up sampling, channels are -> 512, 256, 128, 64, 1
- Images are padded before performing down sampling.
- The activation function used in forwarding is sigmoid.
- At each down sampling step, 2 convolution operations, both 2D with kernel size = 3 each followed by ReLU activation. The first convolution is the one that doubles the channels.

Appendix ii: Attention Unet Model Architecture Details

- The number of channels while downsampling and upsampling are exactly the same as that of the UNet model discussed in Appendix i. The only difference comes at the attention blocks which are present while up sampling the image.
- Each Attention block performs 3 convolution operations with Batch Normalization followed by a ReLU activation function.
- It uses sigmoid activation function after upsampling is complete.

Appendix iii: ResUnet++ Architecture Details

- ResUnet++ model too had the same number of channel as UNet and attention UNet. It has 3 squeeze_excite blocks and 3 residual_conv blocks.
- Each squeeze block performs 2 linear transformations with ReLU activation in between followed by sigmoid activation at the end.
- Each residual_conv block performs 2 conv2D operations with batch normalization and ReLU activation.
- Following the squeeze_excite and residual_conv blocks, There is 1 Atrous Spatial Pyramid Pooling(ASPP) block followed by upsampling with Attention Blocks.
- Uses Sigmoid after upsampling is complete.