MULTI-VIEW CAMERA SYNTHESIS USING CONVOLUTIONAL NEURAL NETWORK

A Dissertation

By: Valeria Olyunina

Supervisors: Matthew Moynihan, Prof Aljosa Smolic

August 2019

MOTIVATION & OBJECTIVES

- **Goal:** Given 2 multi-view camera images generate an accurate in-between image. Particular focus on human performers in a multi-view camera setup

- **Motivation:**
  - Improve 3D reconstruction with photogrammetry methods (Structure-from-Motion, Shape-from-Silhouette)
  - Viewpoint synthesis for Free-viewpoint and 360° video

- **Objectives:**

  Primary: Train a **neural network** capable of producing accurate interpolated images from multi-view cameras, **affectively synthesising a synthetic camera**

  Secondary:

  1) Produce a **multi-view dataset** suitable for NN training
  2) Investigate neural networks suitable for the task
  3) Investigate quality measurements suitable for mulit-view interpolated images

BACKGROUND

Temporal video interpolation using neural networks (CNN, GAN):

- Deep optic flow based
- Phase-based: *PhaseNet by Mayer et al., 2018*
- Pixel-motion based:

  - Adaptive Separable Convolution by S. Niklaus et al, 2017 and 2018

  - Deep-voxel flow by Zhang et al., 2018

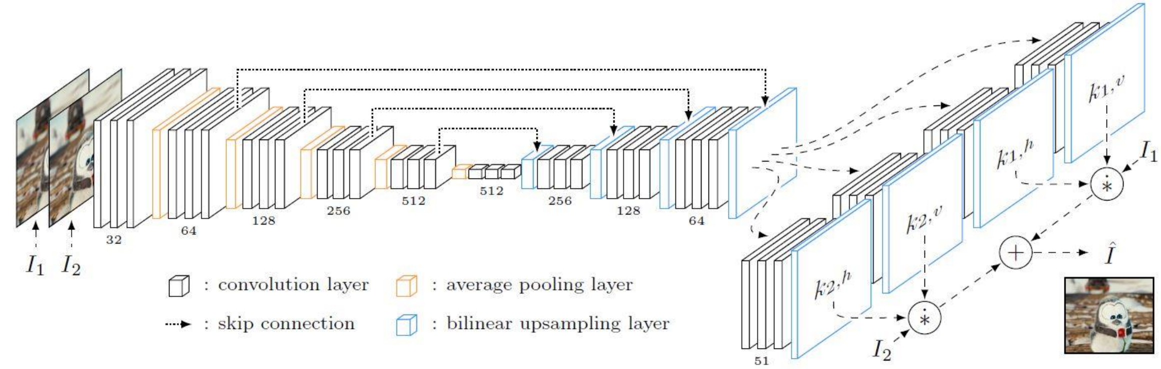
  - GANs: Multi-Scale Frame-Synthesis Network by Hu et al., 2017

Viewpoint synthesis: involves 3D reconstruction

- *Artifact Reduction based Multi-view generation: Lee et al., 2017*

Novelty: Multi-view generation using neural networks directly, without prior 3D reconstruction, viewpoint estimation and depth camera

METHODOLOGY – NN ARCHITECTURE



Fully convolutional NN:

- Down-sampling

(18 layers deep)

- Up-sampling with skip connections
- Kernel estimation – 1D horizontal and vertical
- Picture construction via adaptive separable convolution

Loss: L1 and SSIM, Adamax optimizer

Trained 3 networks – L1 with 51-px kernel, L1 with 71-px kernel, SSIM-loss 71-px

METHODOLOGY – DATASET GENERATION

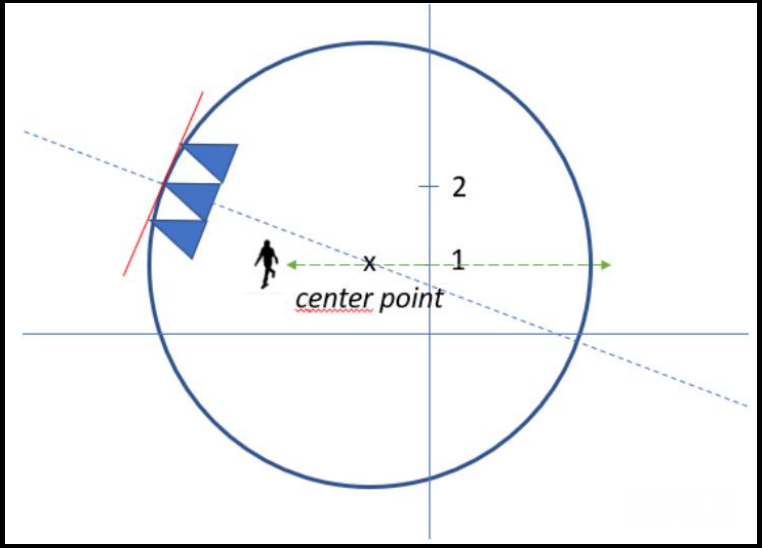
- Real multi-view datasets deemed insufficient for training
- Synthetic dataset:

  - Generated with Python module in Blender - bpy

  - Based on human models: SMPL male and female

  - Textures, mocap from SURREAL

  - Random multiview camera parameters:
    - Distance to center point (3-5 m)
    - Yaw around the center point (0 – 360 degrees)
    - Camera height above ground (0.9 – 1.9 meters)
    - Random roll from pointing to center point (-10 to 10 °)
    - Left and Right camera offset from ground truth (5 – 50 cm)



DATASET EXAMPLES

EVALUATION

2D measures (ground truth to interpolated image)

- SSIM
- PSNR
- Silhouette difference:
    - False negative – matter will be removed in reconstruction
    - False positive – matter will be added in reconstruction

3D measure – Hausdorff distance

Perform SfS reconstruction based on silhouettes with 12 real cameras vs 24 cameras (12 real + 12 synthetic images)

Sample 3534 Frame 9

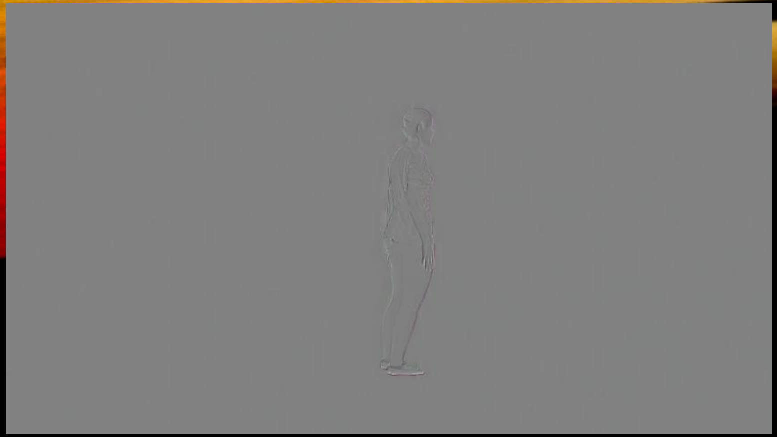
Model: 71- SSIM

Camera offset: 0.286m

Dist to subject: 4.61m

Pixel dist: 11 pxls

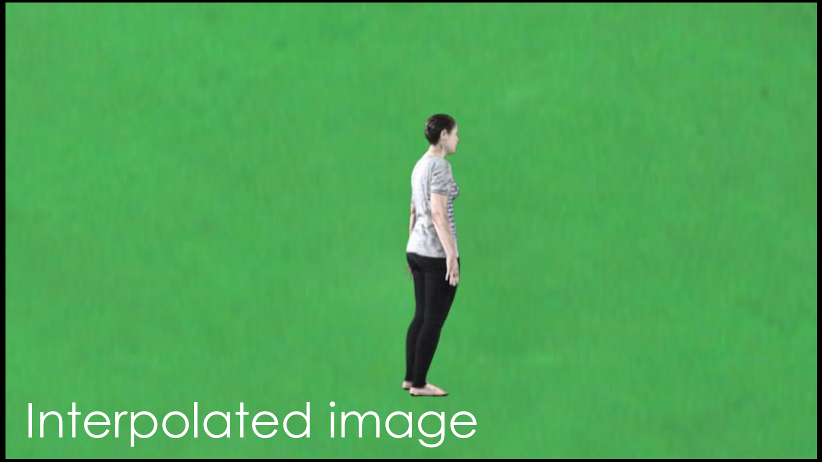
SSIM: 0.9970, PSNR: 31.26



Left

Ground Truth

Right



Interpolated image

Video

Sample 3517 Frame 5

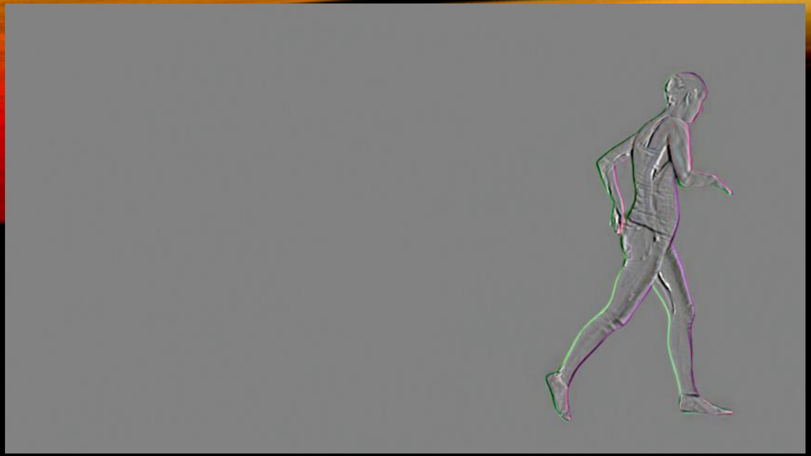
Model: 71- SSIM

Camera offset: 0.347m

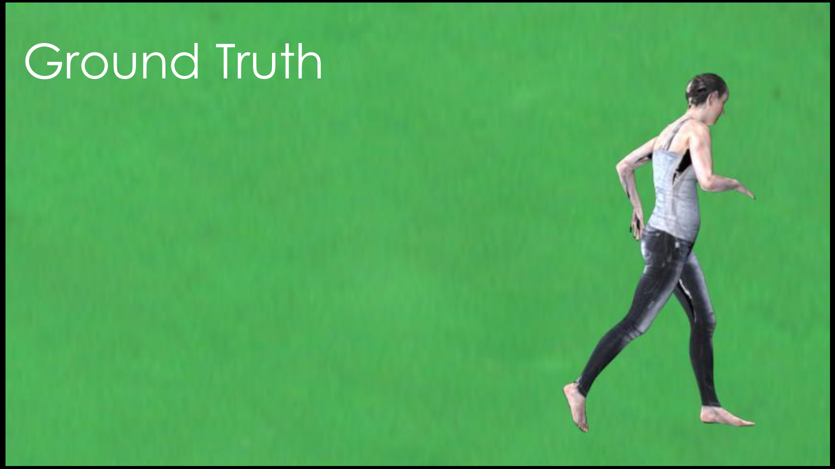
Dist to subject: 3.95 m, Pixel dist: 21 pxls

SSIM: 0.96985, PSNR: 18.35,

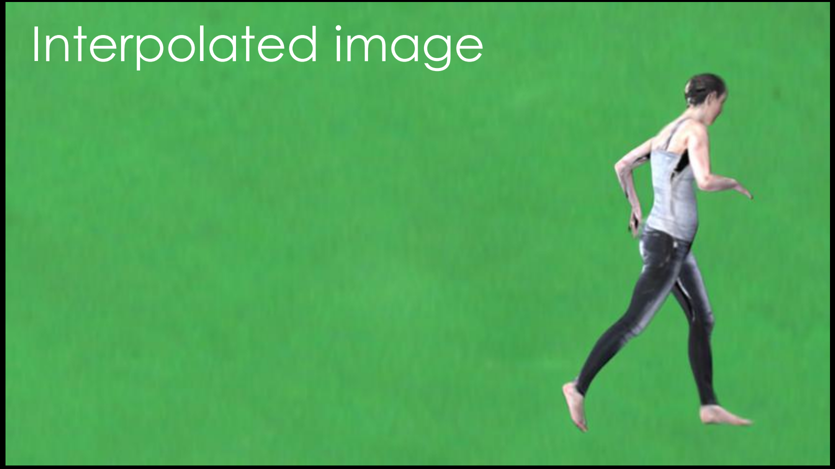
FN 3.9%, FP 6.3%



Left



Ground Truth



Interpolated image

Right

Video

Sample 3506 Frame 2

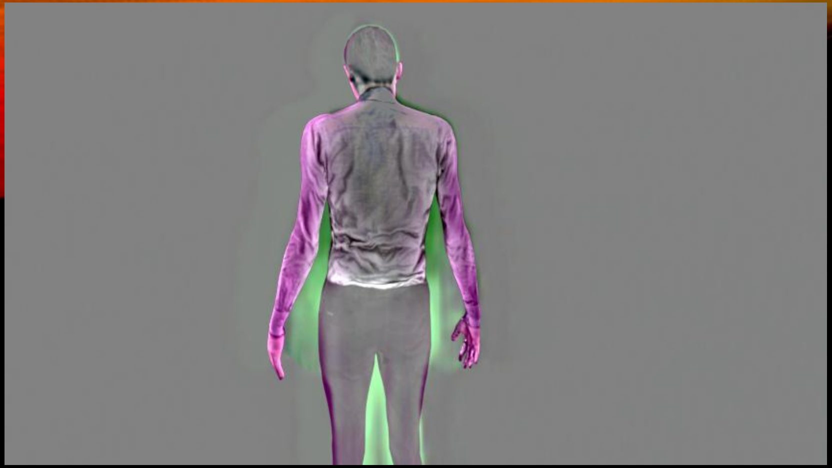
Model: 71- SSIM

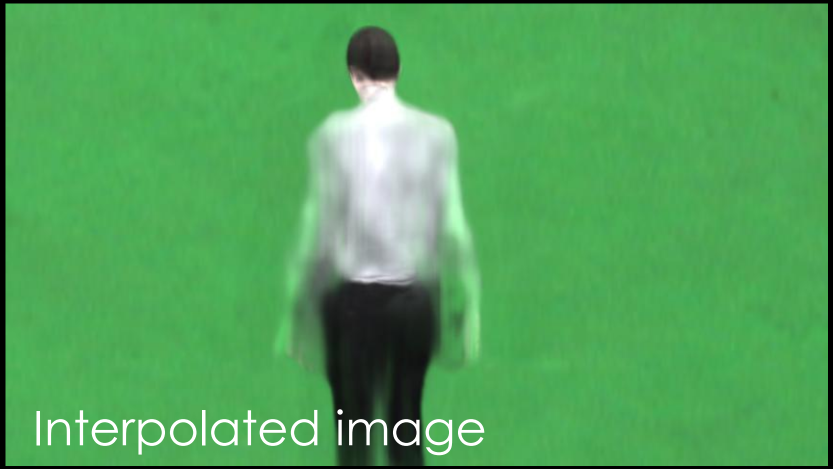
Camera offset: 0.489m

Dist to subject: 2.35m, Pixel dist: 87 pxls

SSIM: 0.91939, PSNR: 16.53,

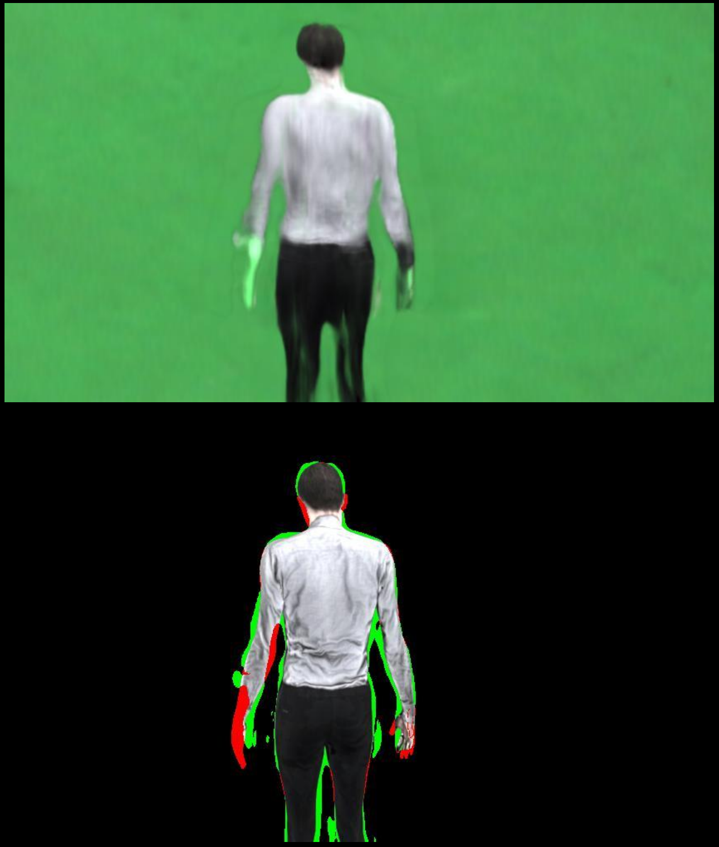
FN 4.5%, FP 19.0%



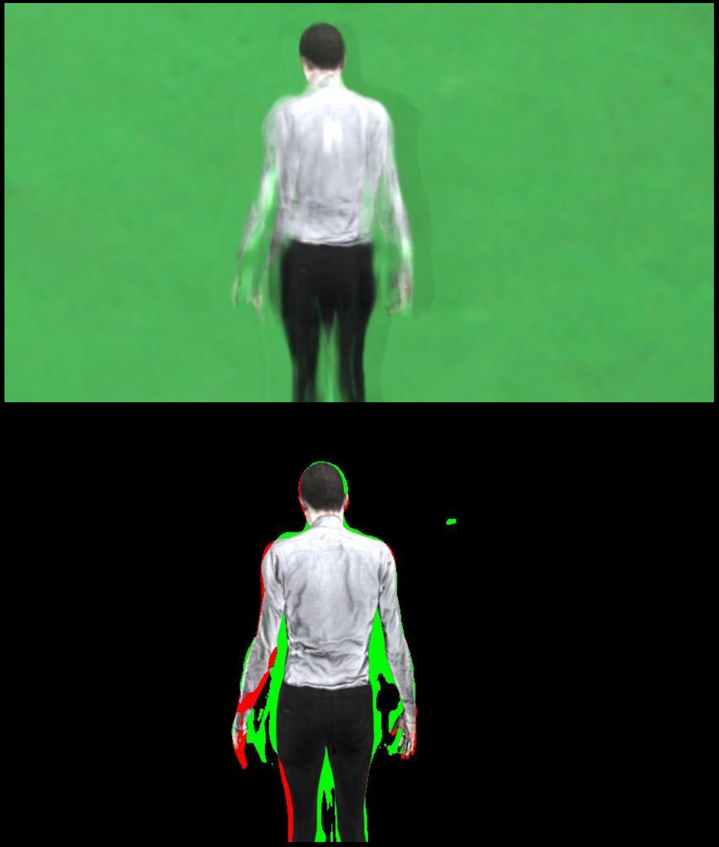


DEMO MODEL DIFFERENCES (ON A BAD EXAMPLE)

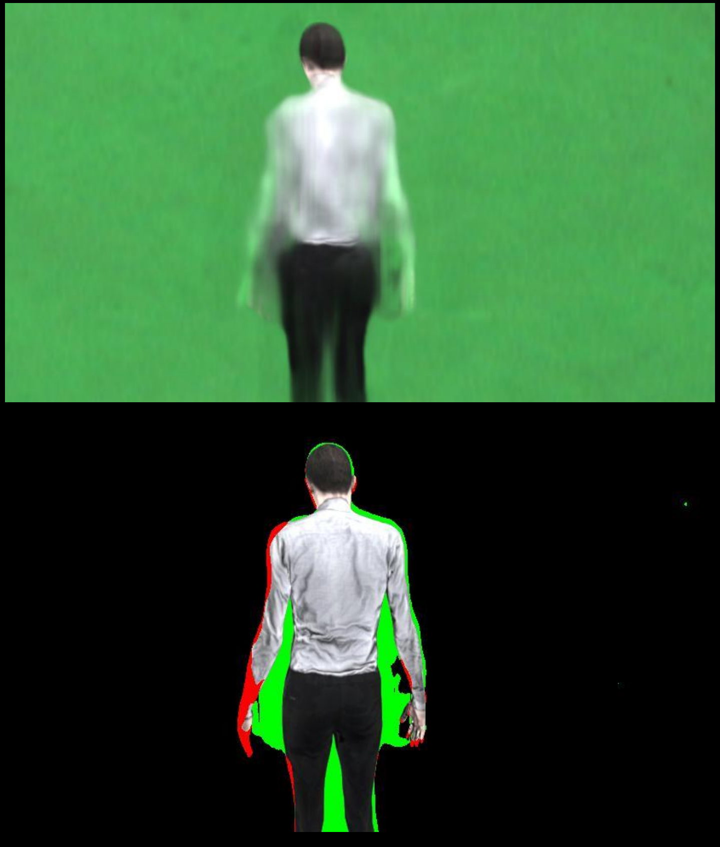
51-kernel L1 loss 30 epochs



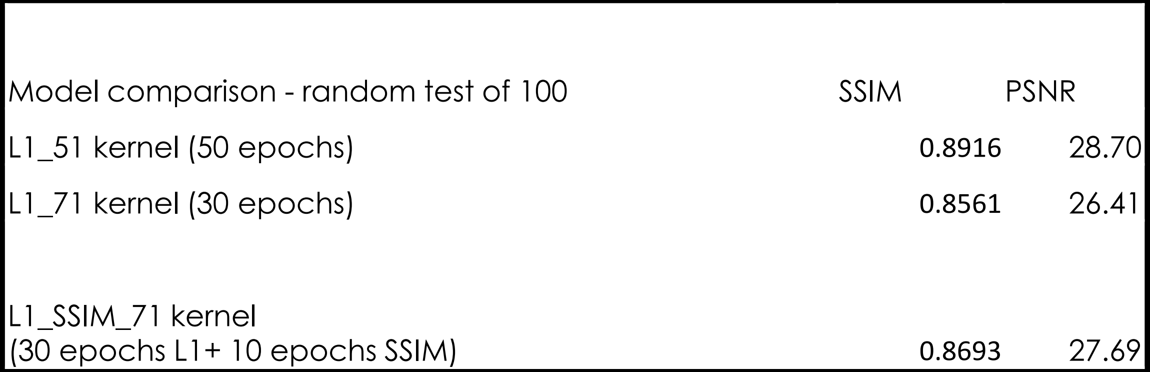
71-kernel L1 loss 30 epochs



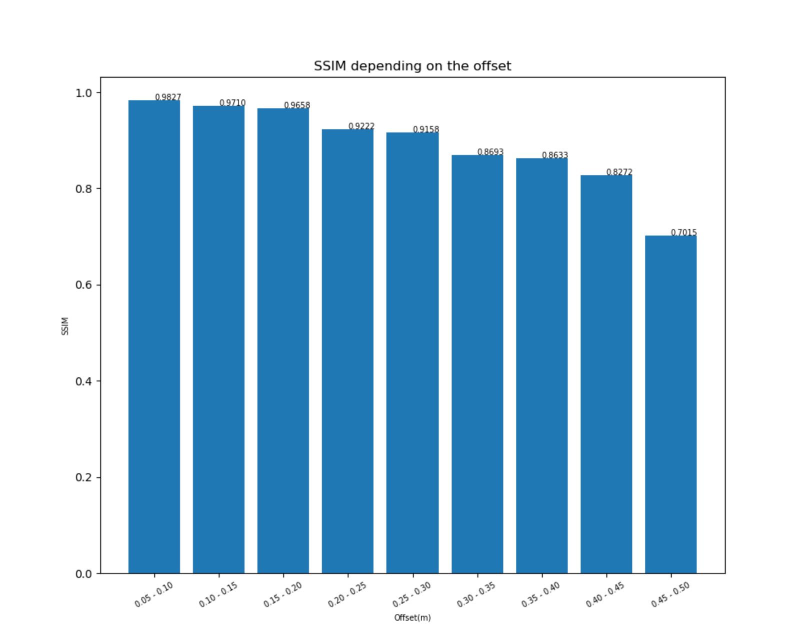
71-kernel + SSIS loss for 10 eps



MODEL COMPARISON

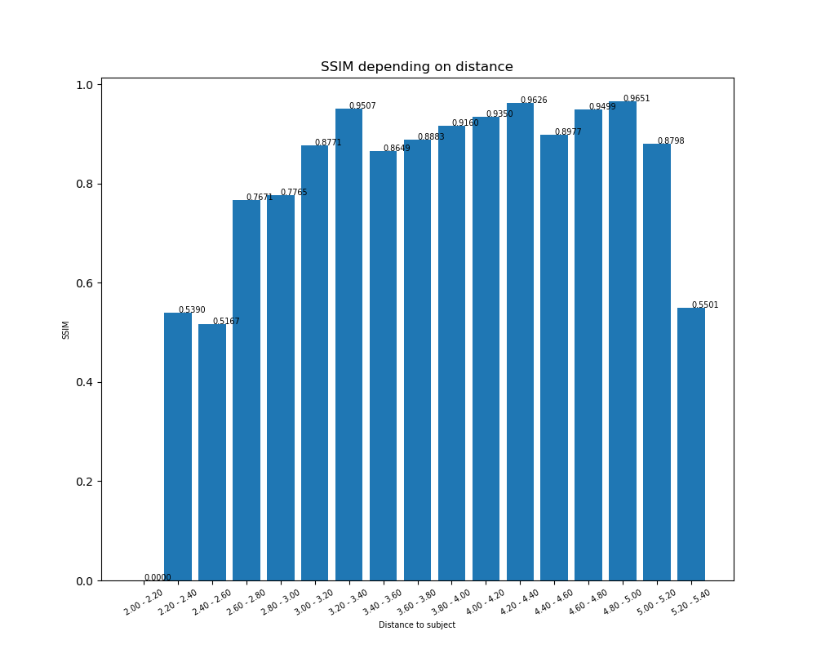


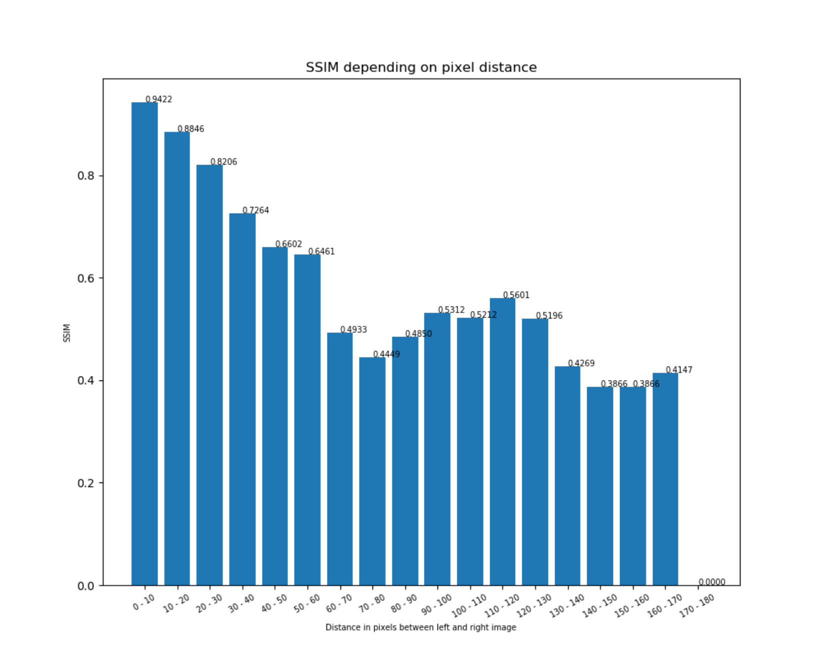
Model comparison - random test of 100	SSIM	PSNR
L1_51 kernel (50 epochs)	0.8916	28.70
L1_71 kernel (30 epochs)	0.8561	26.41
L1_SSIM_71 kernel (30 epochs L1+ 10 epochs SSIM)	0.8693	27.69

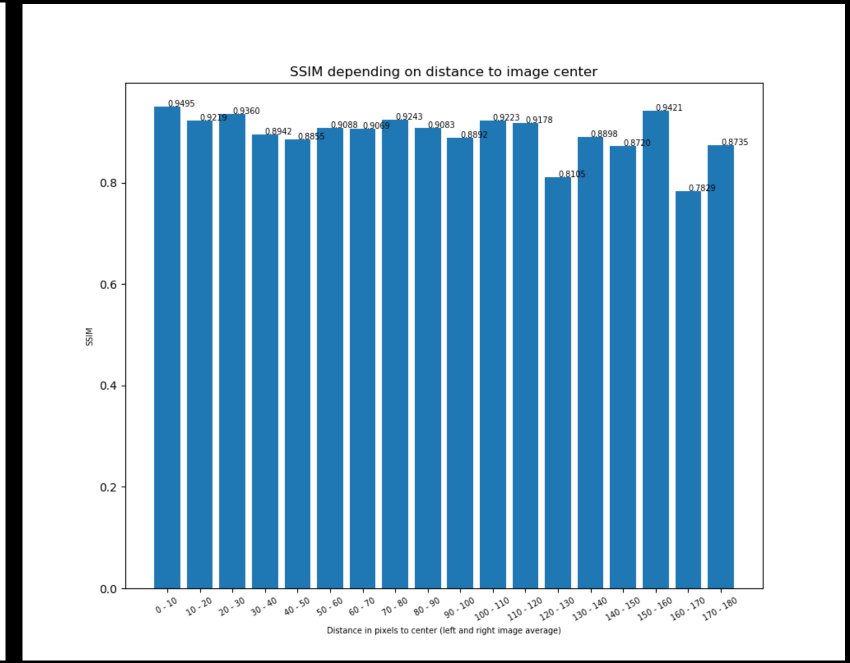


SSIM KERNEL 51 – LOSS L1 (50 EPOCHS)

- Offset: >0.95 if left and right cameras within 40cm, >0.9 within 60 cm
- Distance to subject: best in middle range (3-5 m) same as training data
- Pixel distance between left and right: deteriorates rapidly with increase
- Pixel distance to center: no dependence..





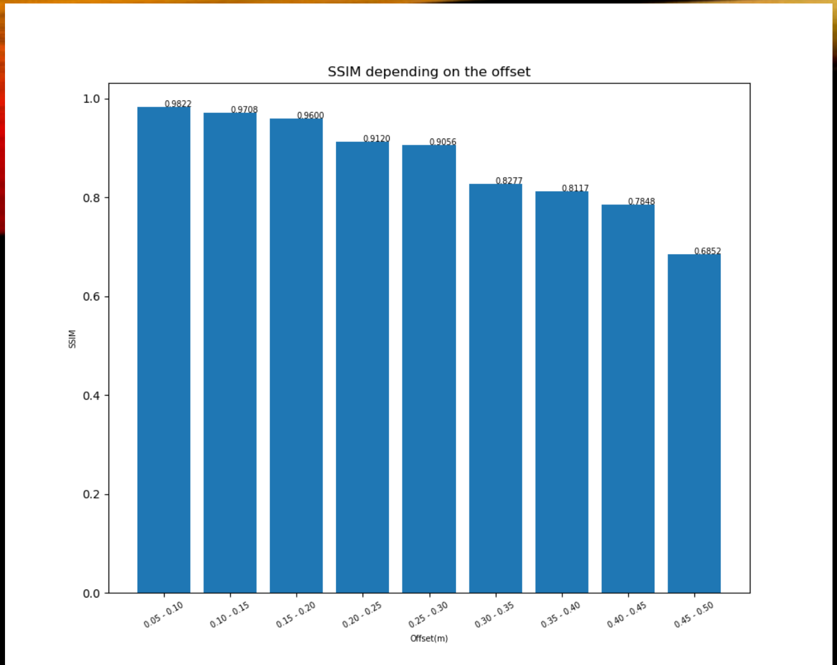


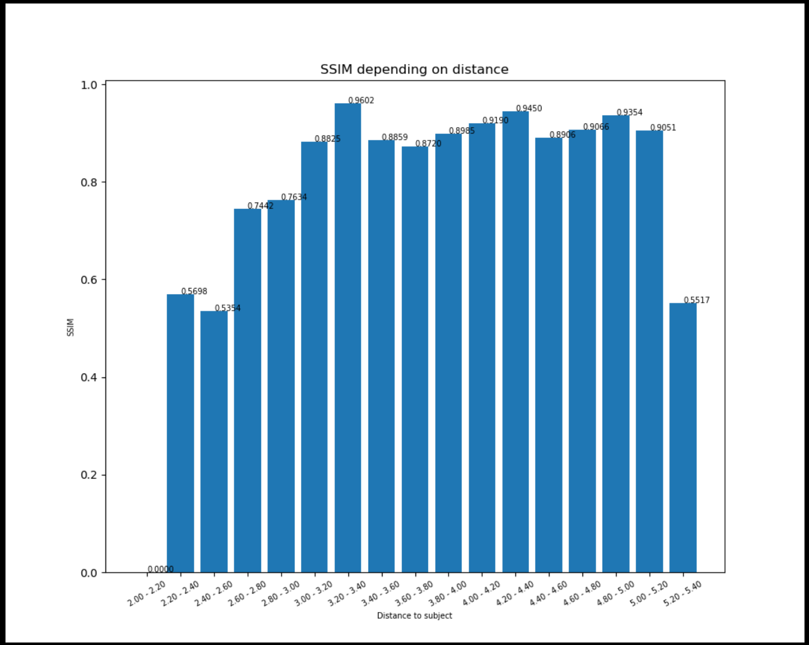
SSIM KERNEL 71 – LOSS SSIM (30EPOCHS)

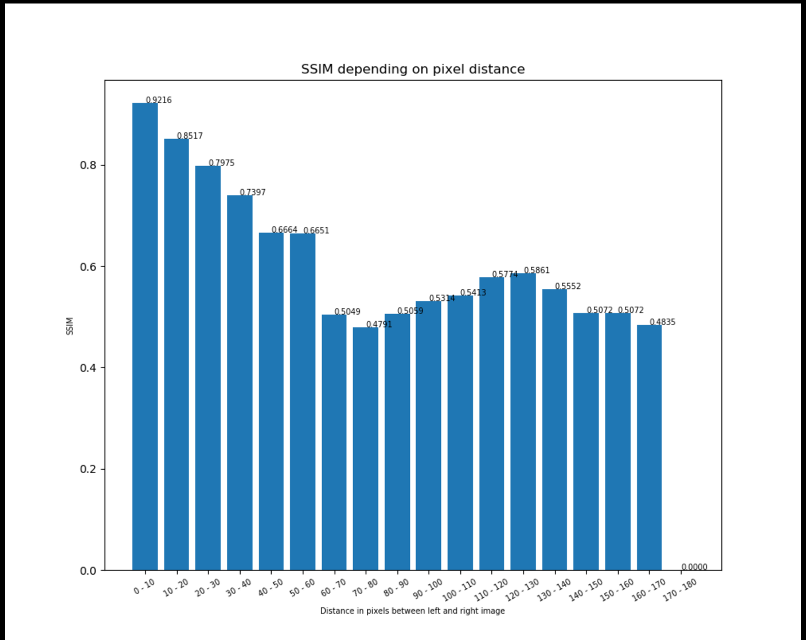
…(training further)

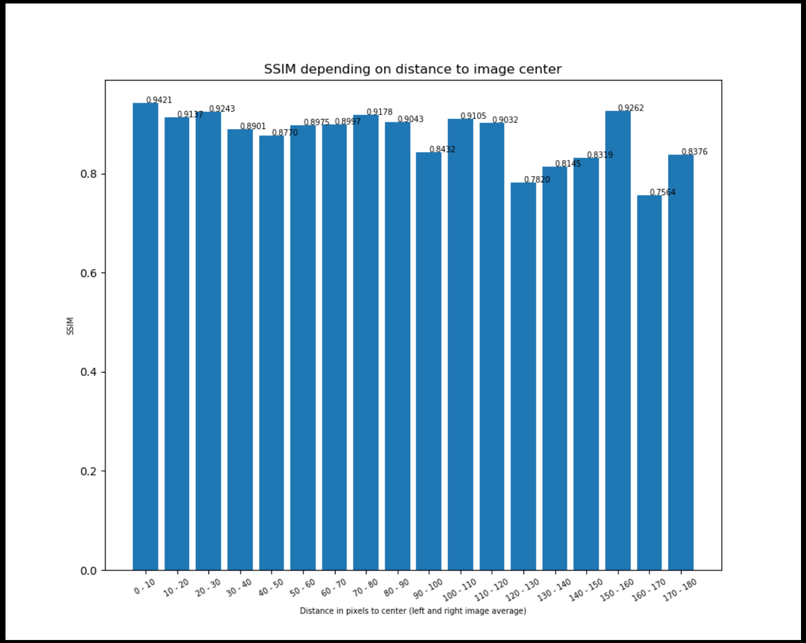
>0.95 for pixel distances upto 60 pixels

No trend for distance to subject or distance from subject to center of the image



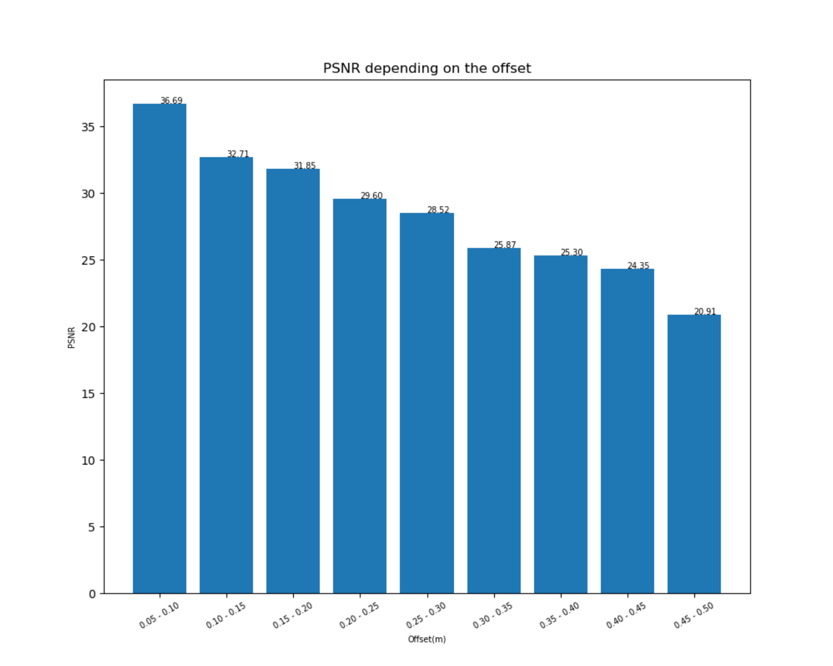


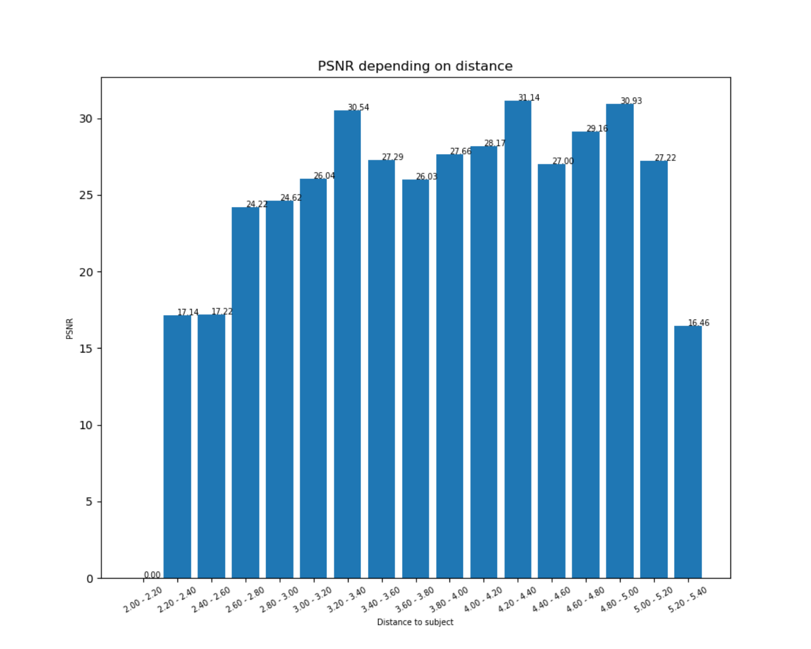


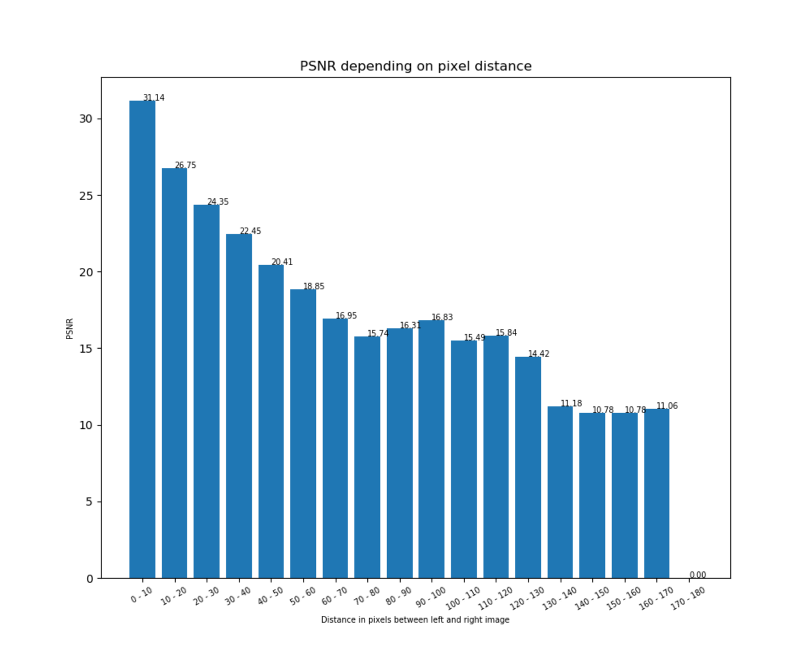


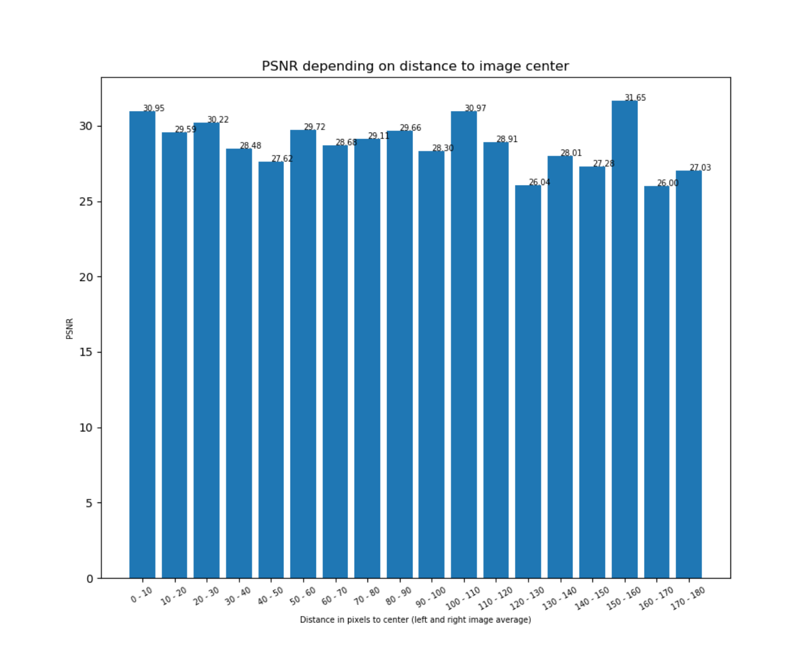
PSNR KERNEL 51 – LOSS L1 (50 EPOCHS)

- Offset: >29 if left and right cameras within 40cm, 23-29 within 80 cm
- Distance to subject: best in middle range (3-5 m) same as training data
- Pixel distance between left and right: <23 already after distance >30 pxl
- Pixel distance to center: no dependence..







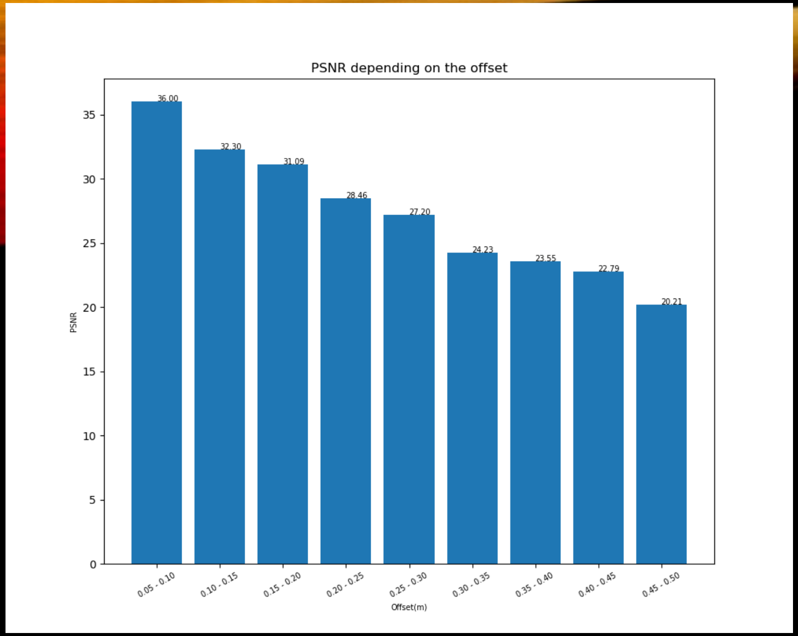


RESULTS: PSNR KERNEL 71 – LOSS SSIM

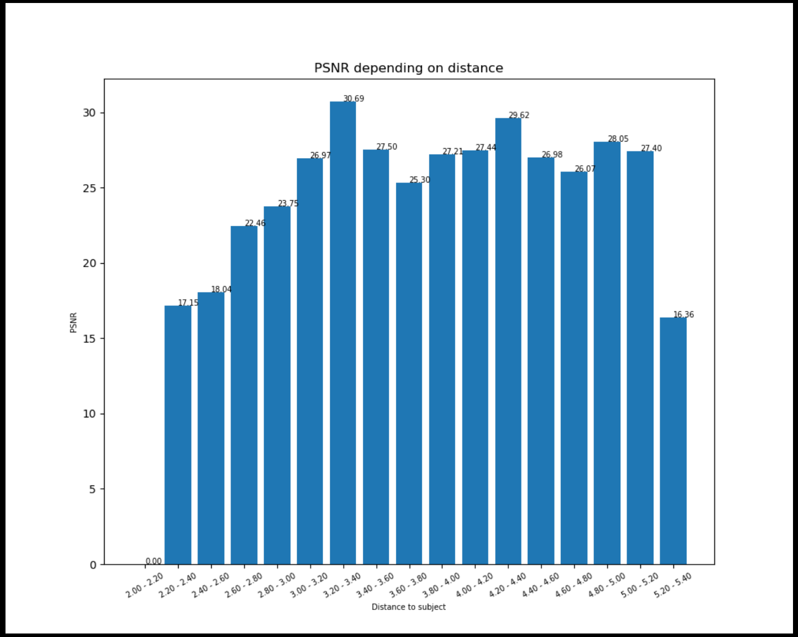
A little better than the other model:

Offset: >29 if left and right cameras within 40cm, 23-29 within 80 cm

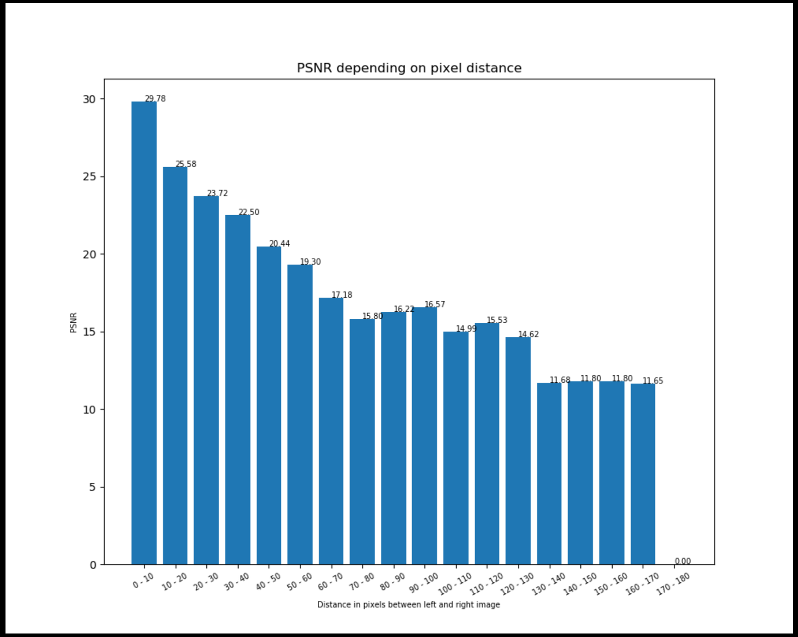
Pixel distance between left and right: <23 already after distance >30 pxl



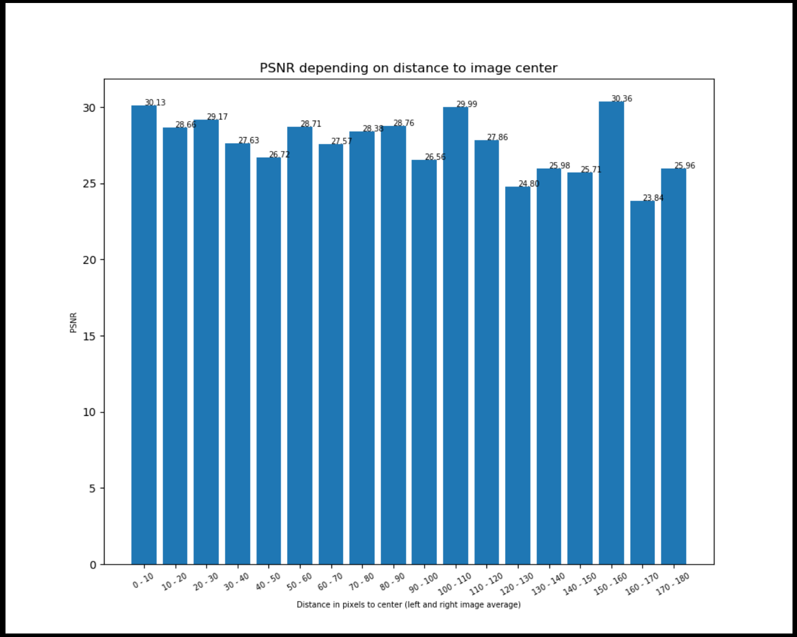
PSNR depending on the offset



PSNR depending on distance



PSNR depending on pixel distance



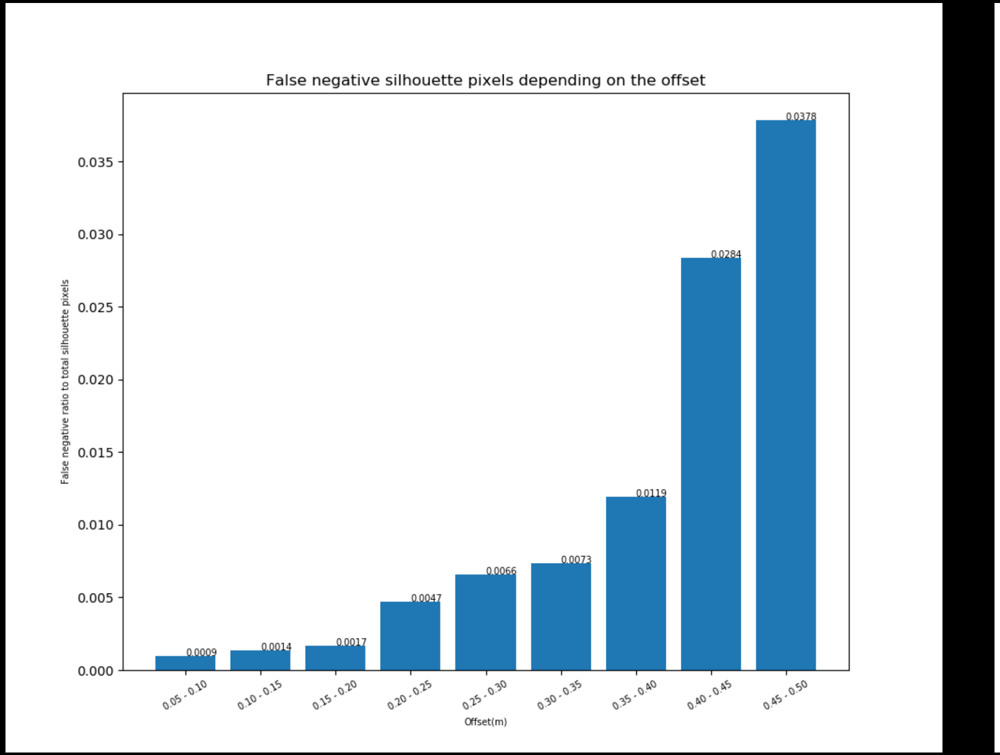
PSNR depending on distance to image center

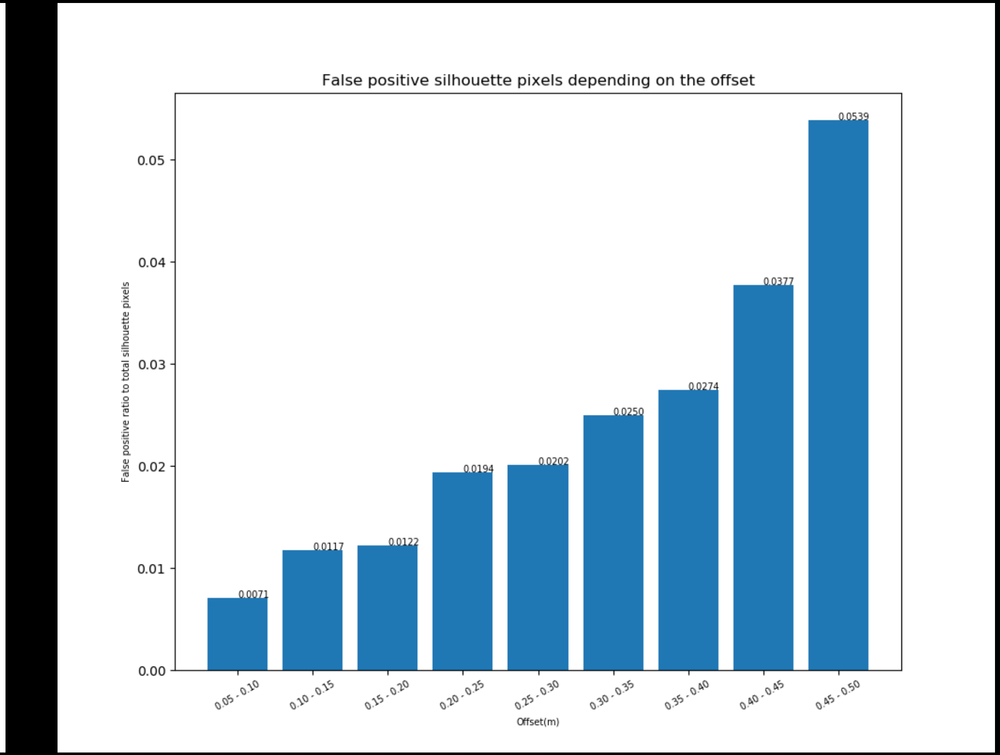
RESULTS:
FN AND FP SILHOUETTE - KERNEL 51 – LOSS L1

False Negative depending on Offset – low upto 70 cm: < 1%

- Upto 3.8% for distances upto 1m

False Positives lower in this model - <2% upto 50cm, then increasing to 5.4%



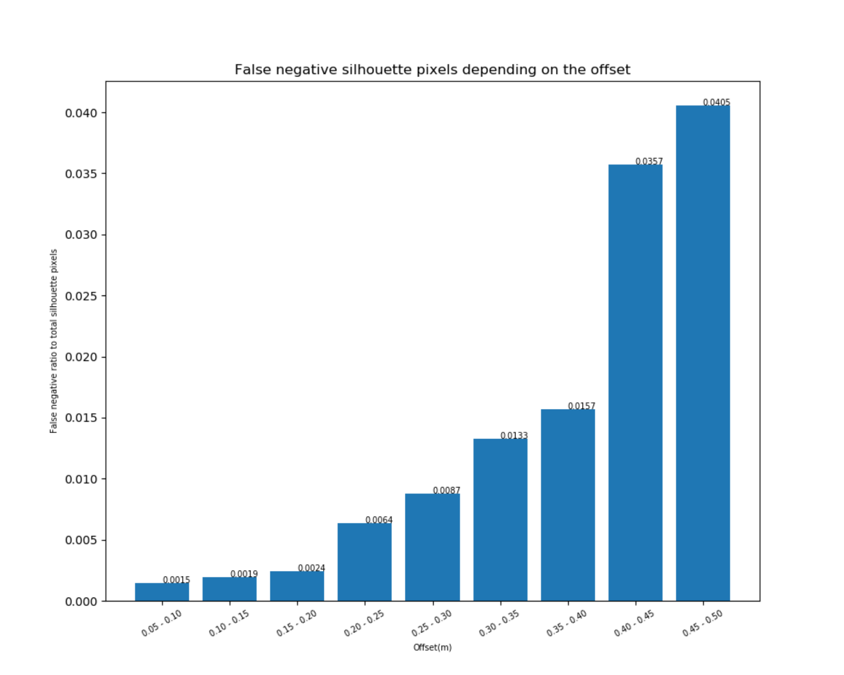


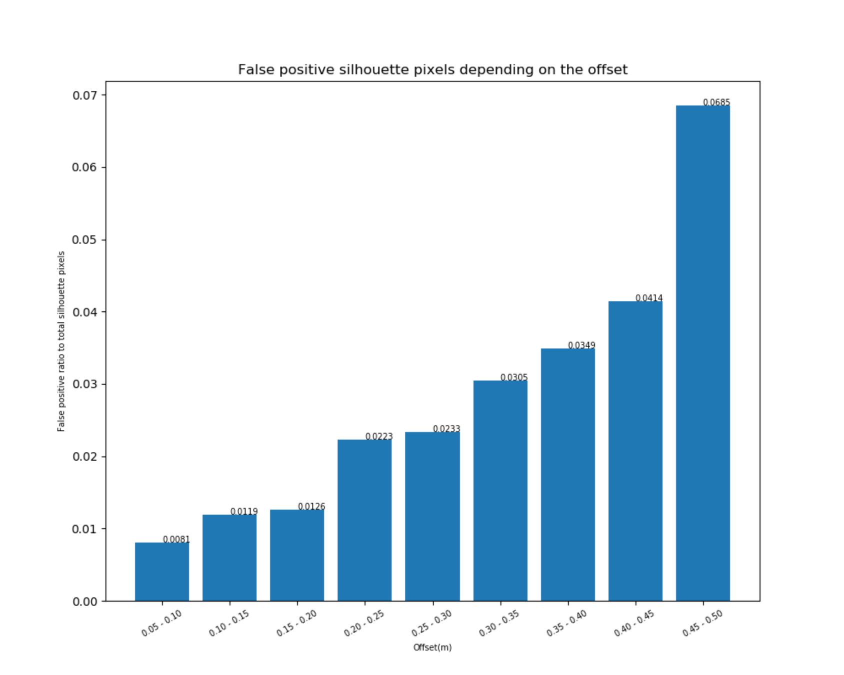
RESULTS:
FN AND FP SILHOUETTE - KERNEL 71 – LOSS SSIM

False Negative depending on Offset – better for small distance:

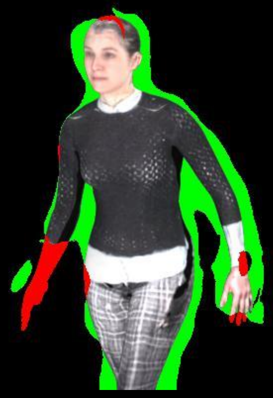
low upto 60 cm: < 1%, upto 4% for distances of 1m

False Positives are higher - <2% upto 40cm, then increasing to 6.5%


False negative silhouette pixels depending on the offset


False positive silhouette pixels depending on the offset

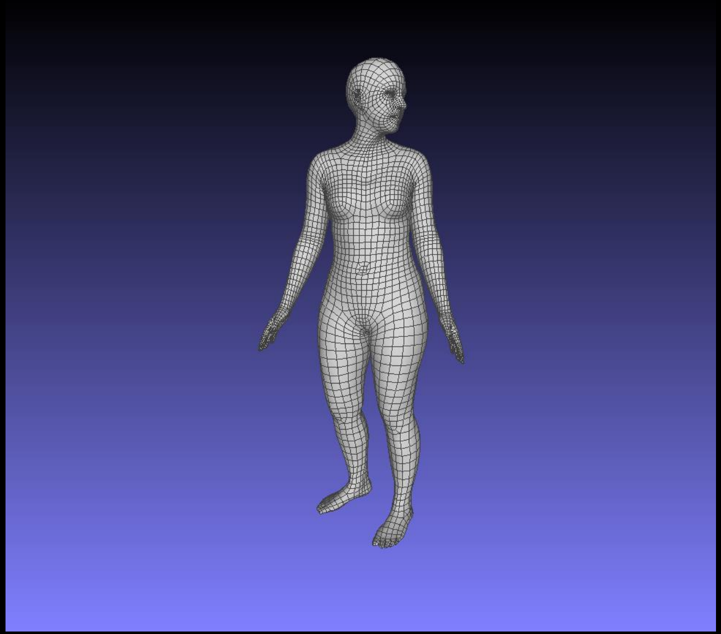
EXAMPLE SILHOUETTES FN AND FP



RESULTS – HAUSDORF DISTANCE KERNEL 51 – LOSS L1

Ground Truth

7567 vertices

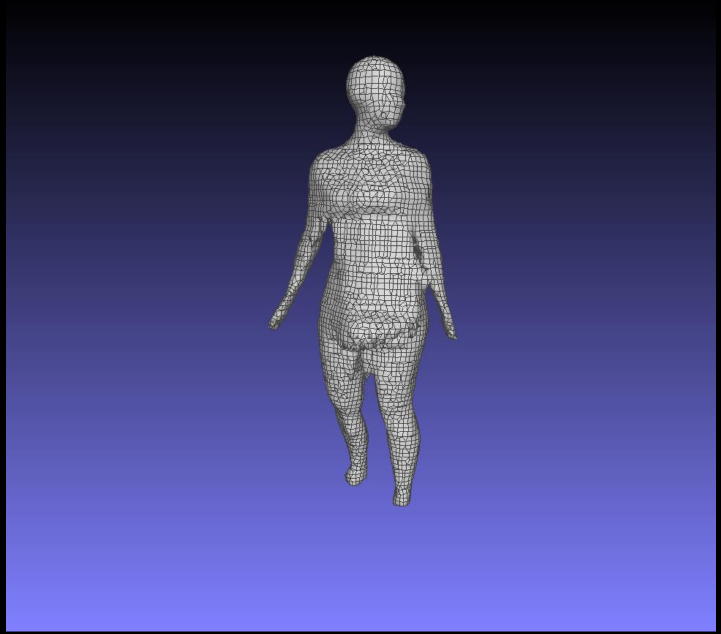


12-cameras test:

Mesh simplified 6314 vert

RMS 0.015301

Mean 0.009057

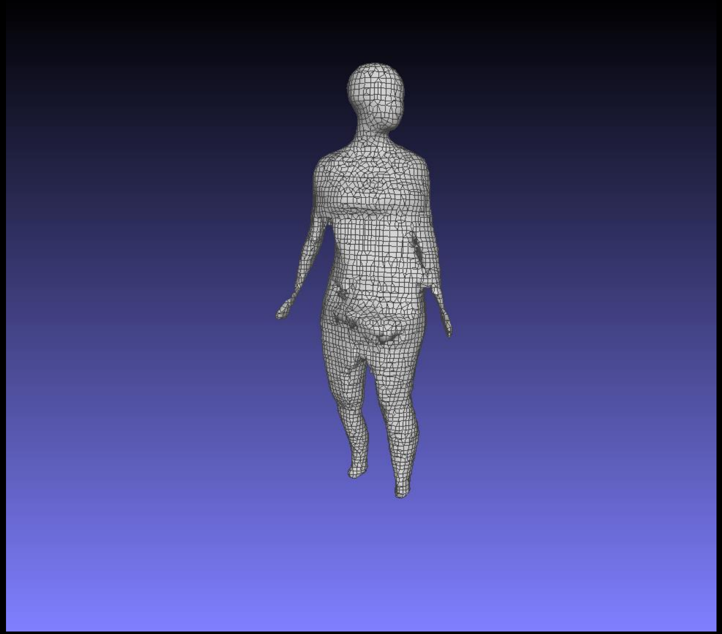


24-cameras test:

Mesh simplified 6200 vert

RMS 0.013312 (better)

Mean 0.010133 (worse)



RESULTS – HAUSDORF DISTANCE KERNEL 71 – LOSS SSIM

Ground Truth

7567 vertices

12-cameras test:

Mesh simplified 6724 vert

RMS 0.015835

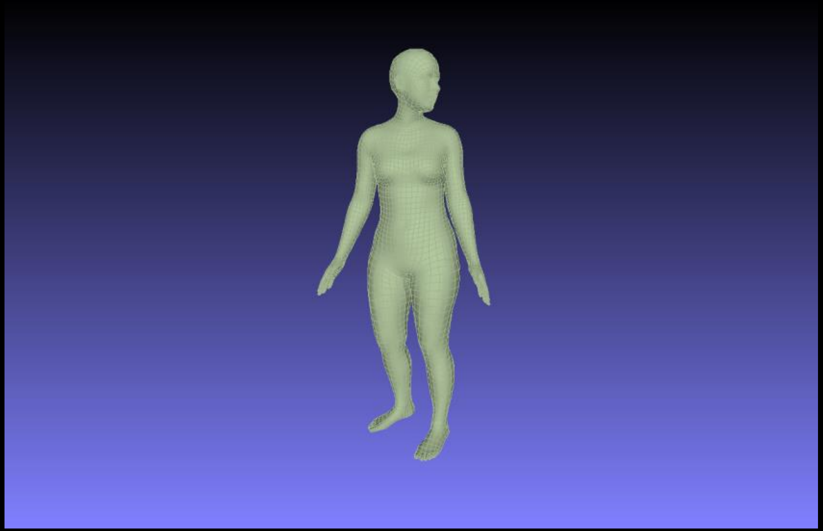
Mean 0.008969

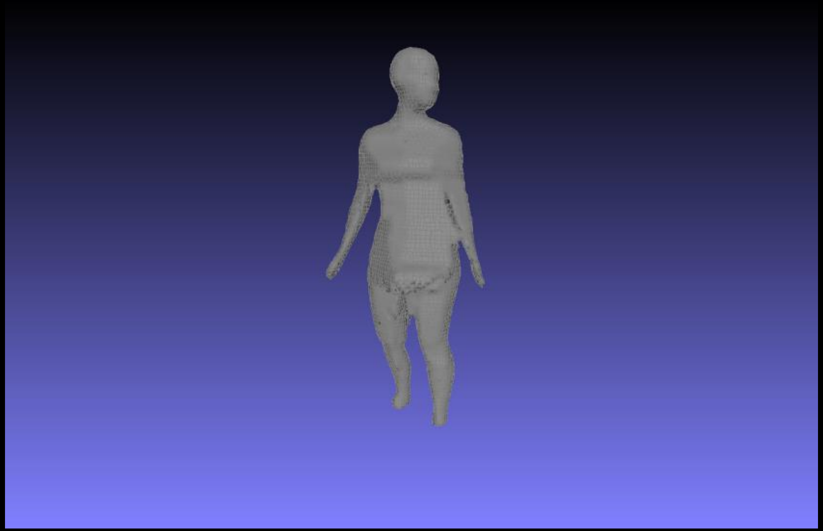
24-cameras test:

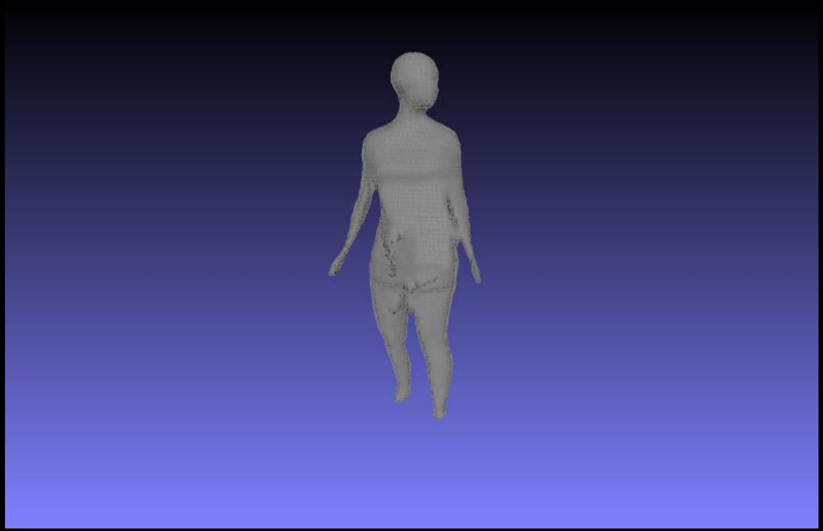
Mesh simplified 6049 vert

RMS 0.013329 (better)

Mean 0.009387 (worse)



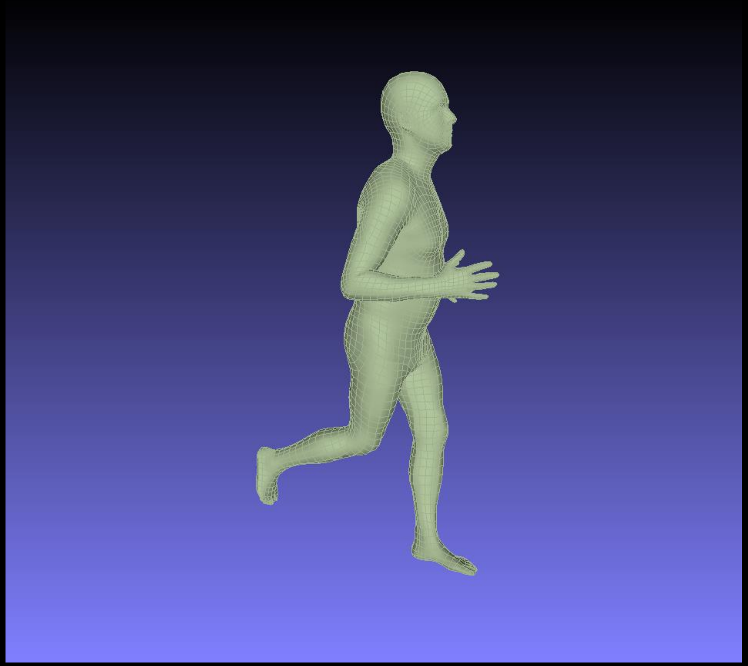




RESULTS – HAUSDORF DISTANCE KERNEL 51 – LOSS L1 (DISAPPEARING LIMBS)

Ground Truth

7567 vertices

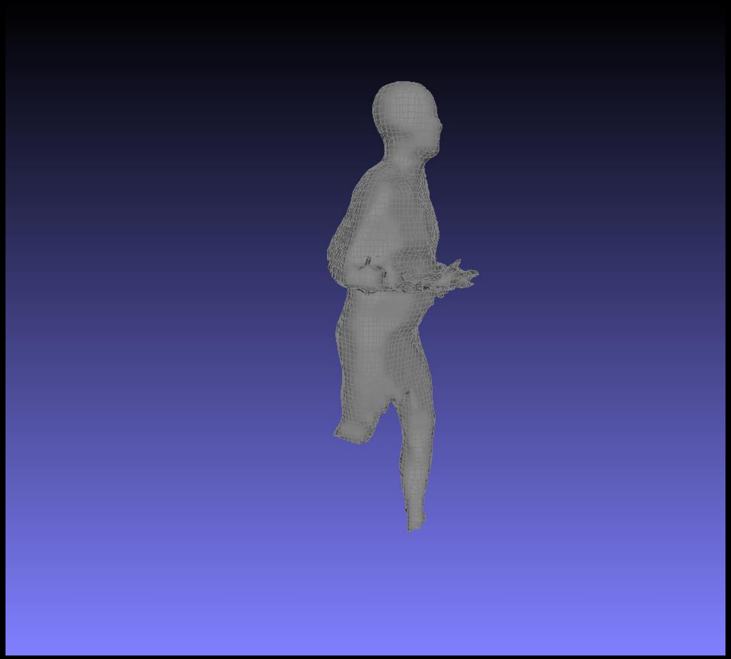


12-cameras test:

Mesh simplified 6326 vert

RMS 0.023363

Mean 0.013658

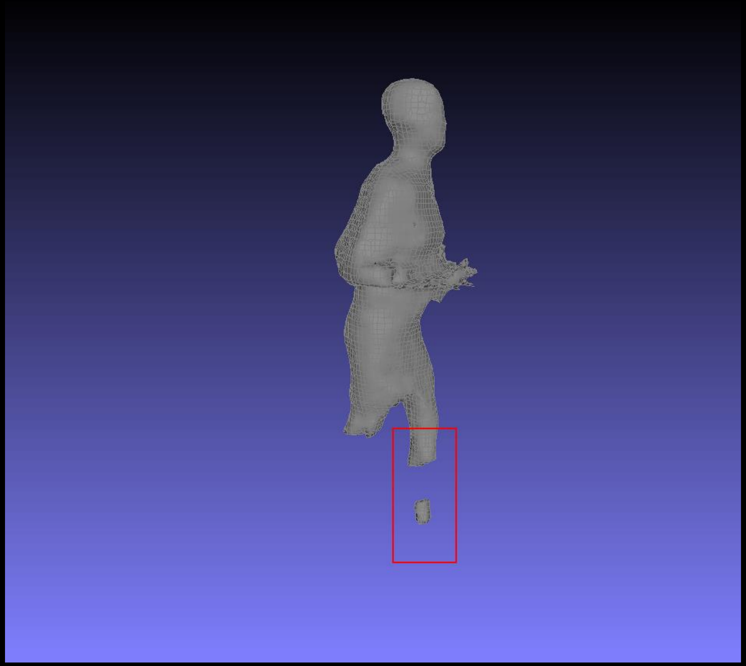


24-cameras test:

Mesh simplified 6112 vert

RMS 0.022285 (better)

Mean 0.014806 (worse)



RESULTS – HAUSDORF DISTANCE KERNEL 71 – LOSS SSIM (BETTER THAN PREV)

Ground Truth

7567 vertices

12-cameras test:

Mesh simplified 6326 vert

RMS 0.023245

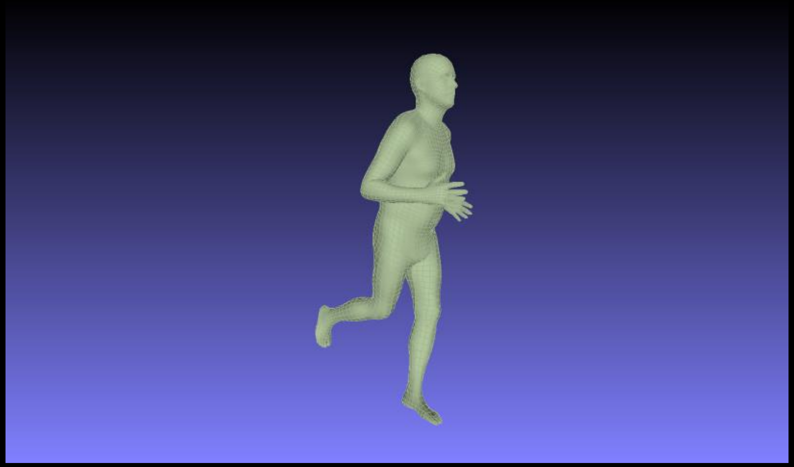
Mean 0.013589

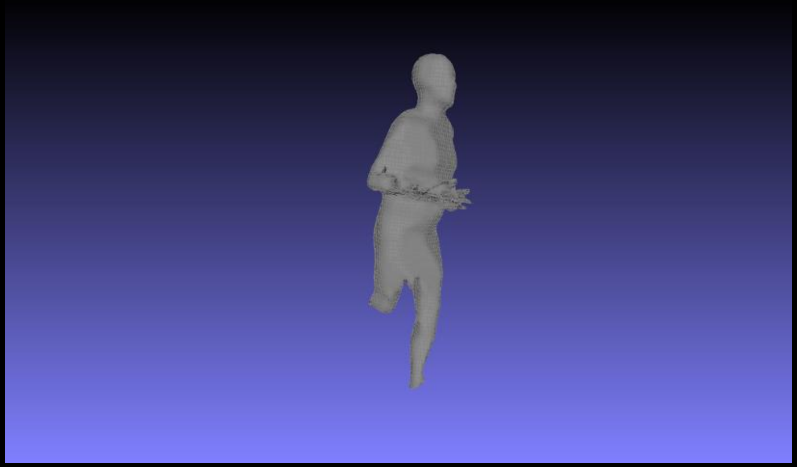
24-cameras test:

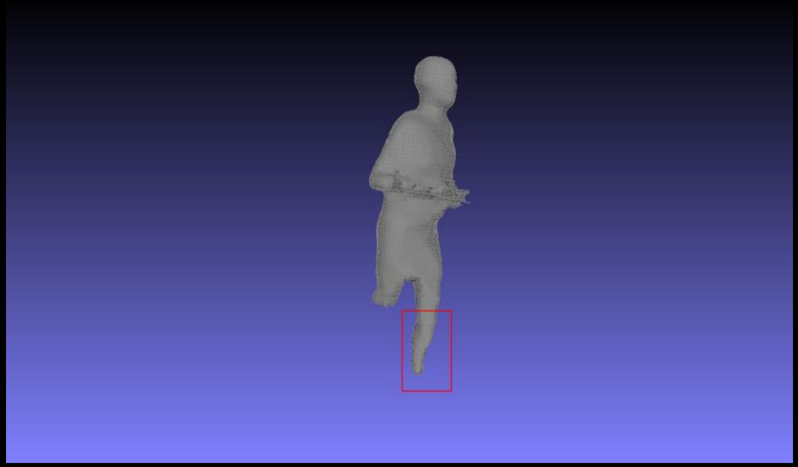
Mesh simplified 6360 vert

RMS 0.021990 (better)

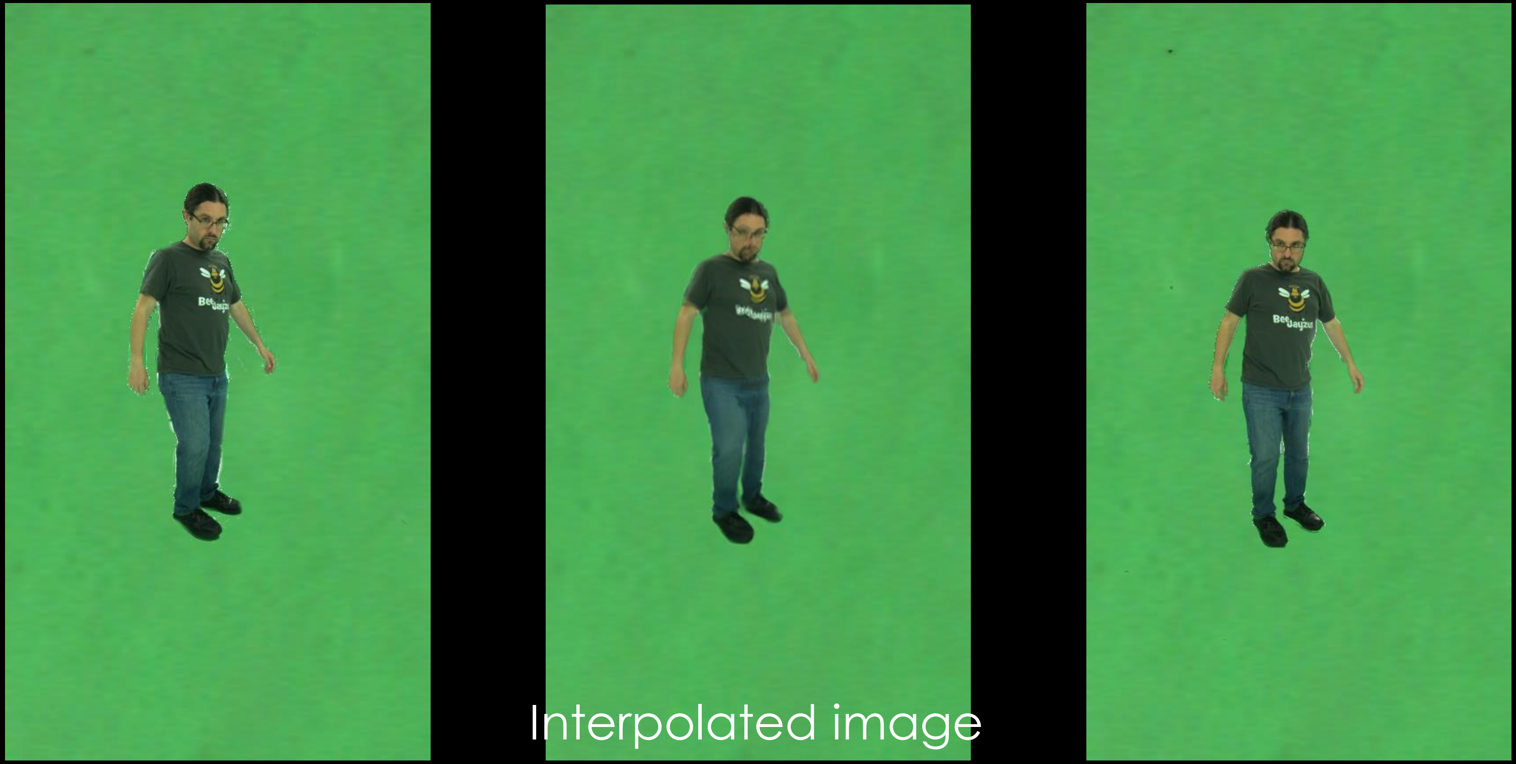
Mean 0.014255 (worse)







REAL DATA TEST (NO GT AVAILABLE)



Interpolated image

CONCLUSION

The approach for generating multi-view images with NN:

1) Works for small distances between images:
   - Camera offset upto 30cm (distance between cameras 60 cm)
   - Pixel distance between left and right image under 20 pixels

2) Can be deployed in Free-viewpoint video for novel points of view as images are believable.

3) 3D reconstruction needs more accurate model with pixel correct images as there is a risk of making the reconstruction worse by removing the voxels, i.e. disappearance of limbs.

4) Real-data - only works with the specific background/ image size. Needs more test

Training on a synthetic dataset is possible

FUTURE WORK

1) Improve dataset:

- Add pitch to left and right camera
- Add varied 3D backgrounds
- Change camera intrinsics
- Improve models – clothes, hands etc
- Use real-life samples!

2) NN Training:

- Train for more epochs
- Train with a larger kernel (> 71-pixel)