Learning from Synthetic Humans
Gül Varol¹, Javier Romero², Xavier Martin¹, Naureen Mahmood²,
Michael J. Black², Ivan Laptev¹, Cordelia Schmid¹

¹ INRIA ² MPI
Goal

- Generating **synthetic** but **photo-realistic** videos of **people** for training CNNs.

- Demonstrating advantages of this data for training:
  1. Human parts **segmentation**
  2. Human **depth** estimation
Motivation

• The annotation for 2D human pose is expensive to collect and difficult to extend.

• Manual labeling of 3D human pose, depth and motion is impractical.

• Synthetic data comes with rich ground truth.
Challenges

• Domain adaptation

• Multi-person

• Extreme poses

• Occlusion

• Object interaction
A body with random 3D shape is configured in a random pose and a 2D image is rendered from a random camera with random lighting by compositing the human model with random texture on top of a random static scene image.

Together with the RGB image, 2D/3D pose, surface normals, optical flow, depth image, and segmentation map for body parts are generated.
SURREAL Dataset

• CAESARS dataset for human body shapes

• LSUN dataset for static background images

• CAESARS dataset and another collection of 3D scans for body textures (clothes)

• CMU dataset for MoCap sequences (marker data)

<table>
<thead>
<tr>
<th></th>
<th>#subjects</th>
<th>#sequences</th>
<th>#clips</th>
<th>#frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>115</td>
<td>1,964</td>
<td>55,001</td>
<td>5,342,090</td>
</tr>
<tr>
<td>Test</td>
<td>30</td>
<td>703</td>
<td>12,528</td>
<td>1,194,662</td>
</tr>
<tr>
<td>Total</td>
<td>145</td>
<td>2,607</td>
<td>67,582</td>
<td>6,536,752</td>
</tr>
</tbody>
</table>
SURREAL Dataset

https://www.youtube.com/watch?v=SJ0vw6CzS7U
Tasks

- Human parts segmentation
- Human depth estimation
Approach - Segmentation

We build on the stacked hourglass network architecture introduced originally for 2D pose estimation problem, extend it for segmentation.

2D pose

- head
- left arm
- right foot

MSE for regressing heatmaps

Segmentation

- left arm
- backg.
- torso

15 output channels (14 parts + backg.)

Softmax error for classifying pixels as one of the parts
Approach - Depth

Depth is continuous. However we are interested in the global pose of the person instead of the precise surface. We discretize depth of a person in 20 values and pose depth estimation as a classification problem.

We align depth maps so that the pelvis depth falls on the center of the axis and quantize the depth into 19 bins (9 behind and 9 in front of the pelvis).
Experiments - Datasets

• SURREAL
  • validation on synthetic test set for segmentation and depth

• Freiburg Sitting People
  • segmentation dataset

• Human3.6M
  • MoCap dataset with RGB videos
  • we generate ground truth for segmentation and depth

• MPII Human Pose
  • 2D pose dataset
  • no ground truth
  • qualitative results for segmentation and depth
Experiments - Evaluation Metrics

• Segmentation
  • Pixel accuracy
  • IOU (intersection over union)

• Depth
  • RMSE (root mean squared error)
  • st-RMSE (scale and translation invariant RMSE)
  • pose-RMSE (RMSE evaluated on joint locations)
  • st-pose-RMSE
Experiments - SURREAL Dataset

<table>
<thead>
<tr>
<th>Input</th>
<th>$\text{Pred}_{\text{segm}}$</th>
<th>$\text{GT}_{\text{segm}}$</th>
<th>$\text{Pred}_{\text{depth}}$</th>
<th>$\text{GT}_{\text{depth}}$</th>
</tr>
</thead>
</table>

**Segmentation**

- **IOU** 69.13 %
- **Accuracy** 80.61 %

**Depth**

- **RMSE** 72.9 mm
- **st-RMSE** 56.3 mm
Experiments - Freiburg Sitting People Dataset

<table>
<thead>
<tr>
<th>Input</th>
<th>Real</th>
<th>Synth</th>
<th>Real+Synth</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training data</th>
<th>Head IOU</th>
<th>Torso IOU</th>
<th>Legs_{up} IOU</th>
<th>mean IOU</th>
<th>mean Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real+Pascal[21]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64.10</td>
<td>81.78</td>
</tr>
<tr>
<td>Real</td>
<td>58.44</td>
<td>24.92</td>
<td>30.15</td>
<td>28.77</td>
<td>38.02</td>
</tr>
<tr>
<td>Synth</td>
<td>73.20</td>
<td>65.55</td>
<td>39.41</td>
<td>40.10</td>
<td>51.88</td>
</tr>
<tr>
<td>Synth+Real</td>
<td>72.88</td>
<td>80.76</td>
<td>65.41</td>
<td>59.58</td>
<td>78.14</td>
</tr>
<tr>
<td><strong>Synth+Real+up</strong></td>
<td><strong>85.09</strong></td>
<td><strong>87.91</strong></td>
<td><strong>77.00</strong></td>
<td><strong>68.84</strong></td>
<td><strong>83.37</strong></td>
</tr>
</tbody>
</table>

Varol et al. Learning from Synthetic Humans, CVPR’17.
Experiments - Human3.6M Dataset

<table>
<thead>
<tr>
<th>Training data</th>
<th>IOU</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fg+bg</td>
<td>fg</td>
</tr>
<tr>
<td>Real</td>
<td>49.61</td>
<td>46.32</td>
</tr>
<tr>
<td>Synthetic</td>
<td>46.35</td>
<td>42.91</td>
</tr>
<tr>
<td>Synthetic+Real</td>
<td>57.07</td>
<td>54.30</td>
</tr>
</tbody>
</table>
Experiments - Human3.6M Dataset

<table>
<thead>
<tr>
<th>Input</th>
<th>Real</th>
<th>Synth</th>
<th>Real+Synth</th>
<th>GT</th>
<th>Input</th>
<th>Real</th>
<th>Synth</th>
<th>Real+Synth</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td></td>
</tr>
<tr>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td></td>
</tr>
<tr>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td></td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td></td>
</tr>
<tr>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
<td><img src="image28.png" alt="Image" /></td>
<td></td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training data</th>
<th>RMSE</th>
<th>st-RMSE</th>
<th>PoseRMSE</th>
<th>st-PoseRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>96.3</td>
<td>75.2</td>
<td>122.6</td>
<td>94.5</td>
</tr>
<tr>
<td>Synthetic</td>
<td>111.6</td>
<td>98.1</td>
<td>152.5</td>
<td>131.5</td>
</tr>
<tr>
<td>Synthetic+Real</td>
<td>90.0</td>
<td>67.1</td>
<td>92.9</td>
<td>82.8</td>
</tr>
</tbody>
</table>
Experiments - Human3.6M Dataset

https://www.youtube.com/watch?v=bK4tAGOWayE
Experiments - MPII Human Pose Dataset
Experiments - Design choices

Clothing variation

- Depth - RMSE
- Segm - IOU (%)

Amount of data

- Depth - RMSE
- Segm - IOU (%)

MoCap variation

We rendered synthetic data using Human3.6M MoCap.

<table>
<thead>
<tr>
<th>MoCap source</th>
<th>s-CMU</th>
<th>s-H3.6M</th>
</tr>
</thead>
<tbody>
<tr>
<td>depth - RMSE</td>
<td>2.57</td>
<td>2.44</td>
</tr>
<tr>
<td>segm - IOU (%)</td>
<td>42.82</td>
<td>48.11</td>
</tr>
</tbody>
</table>

Tested on real-H3.6M
Conclusions

• It is possible to learn from synthetic images of people.

• We have shown the generalization capability of CNNs trained on synthetic people on two tasks:
  • segmentation,
  • depth estimation.

• The rich ground truth can potentially be used for other tasks.
Thanks

www.di.ens.fr/willow/research/surreal

Data and code are available.