INTRODUCTION

Goal & Contributions

⊲ Generating synthetic but photo-realistic videos of people for training CNNs.
⊲ Demonstrating advantages of this data for training human parts segmentation and depth estimation.

Motivation

⊲ Manual labeling of 3D human attributes is impractical and pixel-level annotation is expensive.
⊲ Synthetic data comes with rich ground truth.

SURREAL DATASET: Synthetic hUmans foR REAL tasks

Graphics pipeline for synthetic humans generation
⊲ SMPL body models posed with CMU MoCap and rendered on static background.
⊲ 1K clothings, 4K body shapes, 70K background images, random light and camera.
⊲ Ground truth segmentation, depth, optical flow, surface normals, 2D/3D pose.

HUMAN PARTS SEGMENTATION AND DEPTH: TRAINING

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

![Hourglass Network](hourglass.png)

• Align depth map to zero at pelvis joint
• Quantize depth into 19 depth bins (classes)
• Cross-entropy loss for classifying input pixels
• One output channel per body part and for the background
• Cross-entropy loss for classifying pixels as one of the parts or background behind

RESULTS: Synthetic

Evaluation metrics
⊲ Segmentation: IOU, pixel accuracy
⊲ Depth: RMSE, st-RMSE

<table>
<thead>
<tr>
<th>Training data</th>
<th>IOU</th>
<th>Pixel accuracy</th>
<th>RMSE</th>
<th>st-RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synth+Real</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RESULTS: Freiburg Sitting People

![Freiburg Sitting People](freiburg.png)

<table>
<thead>
<tr>
<th>Training data</th>
<th>IOU</th>
<th>Pixel accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synth+Real</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RESULTS: Human3.6M

![Human3.6M](human3.6m.png)

<table>
<thead>
<tr>
<th>Training data</th>
<th>IOU</th>
<th>Pixel accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synth+Real</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

QUALITATIVE: MPII Human Pose

Challenging cases
⊲ Multi-person
⊲ Occlusion
⊲ Object interaction
⊲ Extreme poses
⊲ Clothing

DESIGN CHOICES

CONCLUSIONS

It is possible to learn from synthetic images of people.
⊲ CNNs trained on synthetic people can generalize.

Code & Data available!

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

gil.varol@inria.fr

IEEE 2017 Conference on Computer Vision and Pattern Recognition

2017 CVPR