Adversarial Inverse Graphics Networks

Katerina Fragkiadaki

Carnegie Mellon University
3D representations

- depth map
- surface normals
- 3D mesh
- 3D point cloud
- 3D voxel occupancy
2D-to-3D synthesis

Hard to collect 3D annotations on real images/videos

Can we improve 2D-to-3D synthesis with unlabelled data?
Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Cao et al.
2D keypoints

Generator
Adversarial Inverse Graphics Networks (AIGNs)

Parameter-free decoder

Generator

2D keypoints

camera projection

reconstruction loss

Discriminator

unpaired 3D poses

Tung at al. 2017
<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Discuss</th>
<th>Eat</th>
<th>Greet</th>
<th>Phone</th>
<th>Photo</th>
<th>Pose</th>
<th>Purchase</th>
<th>Sit</th>
<th>SitDown</th>
<th>Smoke</th>
<th>Wait</th>
<th>Walk</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward2Dto3D</td>
<td>75.2</td>
<td>118.4</td>
<td>165.7</td>
<td>95.9</td>
<td>149.1</td>
<td>154.1</td>
<td>77.7</td>
<td>176.9</td>
<td>186.5</td>
<td>193.7</td>
<td>142.7</td>
<td>99.8</td>
<td>74.7</td>
<td>128.9</td>
</tr>
<tr>
<td>3Dinterpr [33]</td>
<td>56.3</td>
<td>77.5</td>
<td>96.2</td>
<td>71.6</td>
<td>96.3</td>
<td>106.7</td>
<td>59.1</td>
<td>109.2</td>
<td>111.9</td>
<td>111.9</td>
<td>124.2</td>
<td>93.3</td>
<td>58.0</td>
<td>88.6</td>
</tr>
<tr>
<td>Monocap [39]</td>
<td>78.0</td>
<td>78.9</td>
<td>88.1</td>
<td>93.9</td>
<td>102.1</td>
<td>115.7</td>
<td>71.0</td>
<td>121.0</td>
<td>118.2</td>
<td>102.5</td>
<td>82.6</td>
<td>75.62</td>
<td>92.3</td>
<td></td>
</tr>
<tr>
<td>AIGN (ours)</td>
<td><strong>53.7</strong></td>
<td><strong>71.5</strong></td>
<td><strong>82.3</strong></td>
<td><strong>58.6</strong></td>
<td><strong>86.9</strong></td>
<td><strong>98.4</strong></td>
<td><strong>57.6</strong></td>
<td><strong>104.2</strong></td>
<td><strong>100.0</strong></td>
<td><strong>112.5</strong></td>
<td><strong>83.3</strong></td>
<td><strong>68.9</strong></td>
<td><strong>57.0</strong></td>
<td><strong>79.0</strong></td>
</tr>
</tbody>
</table>

Table 1. **3D reconstruction error** in H3.6M using ground-truth 2D keypoints as input.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Discuss</th>
<th>Eat</th>
<th>Greet</th>
<th>Phone</th>
<th>Photo</th>
<th>Pose</th>
<th>Purchase</th>
<th>Sit</th>
<th>SitDown</th>
<th>Smoke</th>
<th>Wait</th>
<th>Walk</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward2Dto3D</td>
<td>80.2</td>
<td>92.4</td>
<td>102.8</td>
<td>92.5</td>
<td>115.5</td>
<td>79.9</td>
<td>119.5</td>
<td>136.7</td>
<td>136.7</td>
<td>144.4</td>
<td>109.3</td>
<td>94.2</td>
<td>80.2</td>
<td>104.6</td>
</tr>
<tr>
<td>3Dinterpr [33]</td>
<td>78.6</td>
<td><strong>90.8</strong></td>
<td>92.5</td>
<td>89.4</td>
<td>108.9</td>
<td>112.4</td>
<td>77.1</td>
<td><strong>106.7</strong></td>
<td>127.4</td>
<td>139.0</td>
<td>103.4</td>
<td>91.4</td>
<td>79.1</td>
<td>98.4</td>
</tr>
<tr>
<td>AIGN (ours)</td>
<td><strong>77.6</strong></td>
<td>91.4</td>
<td><strong>89.9</strong></td>
<td><strong>88</strong></td>
<td><strong>107.3</strong></td>
<td><strong>110.1</strong></td>
<td><strong>75.9</strong></td>
<td>107.5</td>
<td><strong>124.2</strong></td>
<td><strong>137.8</strong></td>
<td><strong>102.2</strong></td>
<td><strong>90.3</strong></td>
<td><strong>78.6</strong></td>
<td><strong>97.2</strong></td>
</tr>
</tbody>
</table>

Table 2. **3D reconstruction error** in H3.6M using detected 2D keypoints as input.
AIGNs for Image-to-Image translation
AIGNs for Image-to-Image translation
AIGNs for Image-to-Image translation
AIGNs for Image-to-Image translation
AIGNs for Image-to-Image translation
AIGNs for Image-to-Image translation
AIGNs for Image-to-Image translation
Male -> Female
CV researcher -> Tom Cruise
Female to male
Female to male
To older
To younger
To younger
AIGNs for plastic surgery

Masking

...

...

[Images of face with and without a mask]
ALIGNs for plastic surgery

Masking

... → Generator → ...
ALIGNs for plastic surgery

Masking

Generator

Masking

reconstruction loss
AIGNs for plastic surgery

Masking Generator Masking

reconstruction loss

Discriminator
To bigger lips
To bigger lips
2D-to-3D synthesis

Recover a human 3D mesh from 2D videos

Can we improve with unlabelled data?

Self-supervised learning of motion capture, Tung et al. NIPS 2017
3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

Pose $\theta$

Shape $\beta$

3D mesh

SMPL($\theta, \beta$)

SMPL: A Skinned Multi-Person Linear Model  Loper et al. SIGGRAPH Asia 2015
3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

SMPL: A Skinned Multi-Person Linear Model  Loper et al. SIGGRAPH Asia 2015
3D human shape model

**SMPL** [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

Pose \( \theta \)

Shape \( \beta \)

3D mesh

\( \text{SMPL}(\theta, \beta) \)
3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh

SMPL( $\theta$, $\beta$ )
3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

\[
\text{SMPL}(\theta, \beta)
\]

Pose \(\theta\)  
Shape \(\beta\)  

3D mesh
SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.
3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

SMPL: A Skinned Multi-Person Linear Model  Loper et al. SIGGRAPH Asia 2015
SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D human shape model

SMPL: A Skinned Multi-Person Linear Model  Loper et al. SIGGRAPH Asia 2015
RGB - to - 3D mesh

**Inputs:**
- RGB frame
- 2D keypoint heatmaps
**Inputs:**
- RGB frame
- 2D keypoint heatmaps

**Outputs:**
- SMPL parameters \((\beta, \theta)\)
Our model

**Inputs:**
- RGB frame
- 2D keypoint heatmaps

**Outputs:**
- SMPL parameters $(\beta, \theta)$
- Camera parameters $(R, T)$
Self-supervised reprojection losses

Frame t

Keypoint re-projection
Self-supervised reprojection losses

Frame $t$

Keypoint re-projection

Segmentation re-projection
Self-supervised reprojection losses

Keypoint re-projection
Segmentation re-projection
Motion re-projection

Frame t

Flownet 2.0: Evolution of optical flow estimation with deep networks. Ilg at al., 2016
Visibility-aware reprojection

Visible parts

Occluded parts

Camera
Supervised training

Synthetic data: SURREAL dataset

Learning from Synthetic Humans, Varol et al. CVPR 2017
## Results

### Per-Joint Error

<table>
<thead>
<tr>
<th>Method</th>
<th>Per-Joint Error (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>optimization</td>
<td>562.4</td>
</tr>
<tr>
<td>supervised pretrained</td>
<td>125.6</td>
</tr>
<tr>
<td>Supervised+self-supervised</td>
<td>98.4</td>
</tr>
</tbody>
</table>
Results
Thank you!

- Adversarial Inverse Graphics Networks, Tung et al., ICCV 2017
- Self-supervised learning of motion capture, Tung et al. NIPS 2017