Learning to Parse Object from Virtual Data: “You only annotate once”.

Alan Yuille
Bloomberg Distinguished Professor
Parsing: Beyond Humans

• Almost all work on object parsing has been done on humans.
• This is because there exist many annotated datasets of humans – with 2D joints annotated and also 3D positions. This has resulted in effective algorithms methods for detecting 2D joints and estimating their 3D structure.
• Do we need large datasets with joints labeled for other animals?
• Instead we can use computer graphics models of animals (e.g., horses and tigers) to give training data for joints. Then perform domain transfer by self-supervision to obtain models that work on real world images.
• Jiteng Mu et al. CVPR. 2020. Oral Presentation.
The Value of Simulated Data

• *Simulated data can help, but there are three important findings.*

• It is important to do *domain adaptation* – e.g., adversarial training. Combining synthetic with real data naively works poorly.

• It is better to have *diverse stimuli* (texture, lighting, data augmentation) than realistic synthetic stimuli.

• You *can annotate many properties of the synthetic object* – e.g., part positions, key-points, and 3D structure. You only need to annotate the synthetic model and then you can render it with different viewpoints, poses, texture, lighting, noise, etc. “You only annotate once”.
Illustration: Animal Parsing

• Annotate Joints of a Synthetic Animal. Goal: parse real animal.
• J. Mu et al. CVPR. 2020.
History of this project

- Stage 1: Use synthetic data as if it was real data (naïve). *Failed due to the big domain gap between real and synthetic stimuli.*
- Stage 2: Use diversity of lighting/viewpoint/texture/background to help solve the domain gap. *Success by combining diversity with learning from simulation.*
- Stage 3: Use properties of synthetic data to scale up to multiple objects and multiple tasks. *Possible by exploiting the synthetic annotations.*
Stage 1: Naïve Strategy does not work

• Train using synthetic data only (left image).
• Works well on synthetic data, but very badly on real data (right image).
• (The deep network features are too different in real and synthetic).
How to Improve Performance?

• Try better synthetic data?
• Buy more realistic (expensive) models and make realistic backgrounds?

• This is intuitive, but does not work well.

• Results are terrible. By contrast, Training with Real Data gives (78.98 PCK@0.05 for keypoint detection)
Stage 2: Realism versus Diversity tradeoff

• These realistic synthetic models are expensive.
• They lack diversity – only one horse, only one tiger.

• Instead:
  • (I) Increase diversity by randomizing texture, lighting, background. (25.33 PCK@0.05)
  • (II) Data augmentation – adding Gaussian noise, rotating the images. (60.85 PCK@0.05)
• Recall Training with Real Data achieves 78.98 PCK@0.05.
How to improve performance?

- Training with Real Only (78.98)
- Train with diverse synthetic data:
  - More realistic model, realistic background (intuitive, but not work)
  - Texture Randomization (25.33)
  - Data Augmentation, rotation, gaussian noise (60.84)
- Add self-supervised training on real data:
  - Domain adaptation
    - synthetic +unlabeled real data, adversarial training (62.33)
    - synthetic +unlabeled real data, semi-supervised training (70.77) No real annotations!
    - synthetic +labeled real data, (82.43 > 78.98) Combining real with synthetic does best.
Animal keypoint video (2)
Animal keypoint video (2)
Stage 3: Scale Up – extend to new tasks.

Scale Up -- extend to more categories

“You only annotate once” (for each object category) but same diversity and learning strategies still apply.
Scale Up: extend to different domains.

Better Domain Generalization: line drawings, pictures.
Recap: History of this project

• Stage 1: Use synthetic data as if it was real data (naïve). Failed due to the big domain gap between real and synthetic.
• Stage 2: Use diversity to help solve the domain gap. Success by combining diversity with self-supervision on real data.
• Stage 3: Use properties of synthetic data to scale up to multiple objects and multiple domains. Exploit the synthetic annotations.

• It took months to go from Stage 1 to Stage 2. It took weeks to go from Stage 2 to Stage 3.
Conclusion

• *Synthetic Data is very helpful but have three messages:*
  • (1) Diversity of Synthetic Data is required. Synthetic alone is not realistic enough.
  • (2) Domain Adaptation is required. Self-supervised learning.
  • (3) Rich annotations on synthetic data: “you only annotate once”.
Backup Slides
Synthetic Animal Project

From Stage 1 to Stage 2: Diversity + Adaptation

<table>
<thead>
<tr>
<th>synthetic + real</th>
<th>Eye</th>
<th>Chin</th>
<th>Shoulder</th>
<th>Hip</th>
<th>Elbow</th>
<th>Knee</th>
<th>Hoove</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>79.04</td>
<td>89.71</td>
<td>71.38</td>
<td>91.78</td>
<td>82.85</td>
<td>80.80</td>
<td>72.76</td>
<td>78.98</td>
</tr>
<tr>
<td>CC-SSL-R</td>
<td>89.39</td>
<td>92.01</td>
<td>69.05</td>
<td>92.28</td>
<td>86.39</td>
<td>83.72</td>
<td>76.89</td>
<td>82.43</td>
</tr>
<tr>
<td>synthetic only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syn</td>
<td>46.08</td>
<td>53.86</td>
<td>20.46</td>
<td>32.53</td>
<td>20.20</td>
<td>24.20</td>
<td>17.45</td>
<td>25.33</td>
</tr>
<tr>
<td>CycleGAN [45]</td>
<td>70.73</td>
<td>84.46</td>
<td>56.97</td>
<td>69.30</td>
<td>52.94</td>
<td>49.91</td>
<td>35.95</td>
<td>51.86</td>
</tr>
<tr>
<td>BDL [26]</td>
<td>74.37</td>
<td>86.53</td>
<td>64.43</td>
<td>75.65</td>
<td>63.04</td>
<td>60.18</td>
<td>51.96</td>
<td>62.33</td>
</tr>
<tr>
<td>CyCADA [16]</td>
<td>67.57</td>
<td>84.77</td>
<td>56.92</td>
<td>76.75</td>
<td>55.47</td>
<td>48.72</td>
<td>43.08</td>
<td>55.57</td>
</tr>
<tr>
<td>CC-SSL</td>
<td>84.60</td>
<td>90.26</td>
<td>69.69</td>
<td>85.89</td>
<td>68.58</td>
<td>68.73</td>
<td>61.33</td>
<td>70.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tiger Accuracy</th>
<th>Eye</th>
<th>Chin</th>
<th>Shoulder</th>
<th>Hip</th>
<th>Elbow</th>
<th>Knee</th>
<th>Hoove</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>96.77</td>
<td>93.68</td>
<td>65.90</td>
<td>94.99</td>
<td>67.64</td>
<td>80.25</td>
<td>81.72</td>
<td>81.99</td>
<td></td>
</tr>
<tr>
<td>95.72</td>
<td>96.32</td>
<td>74.41</td>
<td>91.64</td>
<td>71.25</td>
<td>82.37</td>
<td>82.73</td>
<td>84.00</td>
<td></td>
</tr>
<tr>
<td>Syn</td>
<td>23.45</td>
<td>27.88</td>
<td>14.26</td>
<td>52.99</td>
<td>17.32</td>
<td>16.27</td>
<td>19.29</td>
<td>21.17</td>
</tr>
<tr>
<td>CycleGAN [45]</td>
<td>71.80</td>
<td>62.49</td>
<td>29.77</td>
<td>61.22</td>
<td>36.16</td>
<td>37.48</td>
<td>40.59</td>
<td>46.47</td>
</tr>
<tr>
<td>BDL [26]</td>
<td>77.46</td>
<td>65.28</td>
<td>36.23</td>
<td>62.33</td>
<td>35.81</td>
<td>45.95</td>
<td>54.39</td>
<td>52.26</td>
</tr>
<tr>
<td>CyCADA [16]</td>
<td>75.17</td>
<td>69.64</td>
<td>35.04</td>
<td>65.41</td>
<td>38.40</td>
<td>42.89</td>
<td>48.90</td>
<td>51.48</td>
</tr>
<tr>
<td>CC-SSL</td>
<td>96.75</td>
<td>90.46</td>
<td>44.84</td>
<td>77.61</td>
<td>55.82</td>
<td>42.85</td>
<td>64.55</td>
<td>64.14</td>
</tr>
</tbody>
</table>
Stage 3: More gains?

1. stage 2 Vs stage 3 = 4 months Vs 3 weeks

Segmentation Masks

More categories

Domain Generalization