# *Learning to Parse Object from Virtual Data: "You only annotate once".*

Alan Yuille

Bloomberg Distinguished Professor

JOHNS HOPKINS UNIVERSITY

JOHNS HOPKINS UNIVERSITY

# *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 their3D 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.

JOHNS HOPKINS UNIVERSITY

$599 Image source: turbosquid.com $799

JOHNS HOPKINS UNIVERSITY

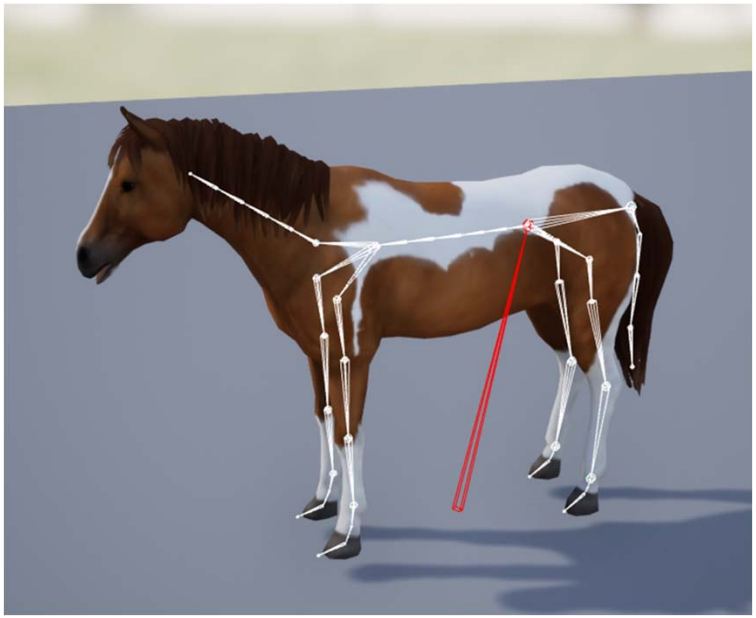
# *The Value of Simulated Data*

- ***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”.

JOHNS HOPKINS UNIVERSITY

# Illustration: Animal Parsing

- Annotate Joints of a Synthetic Animal.  Goal: parse real animal.
- J. Mu et al. CVPR. 2020.



JOHNS HOPKINS UNIVERSITY

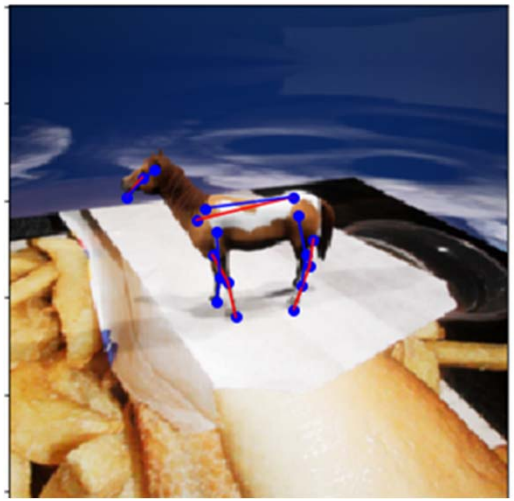
# History of this project

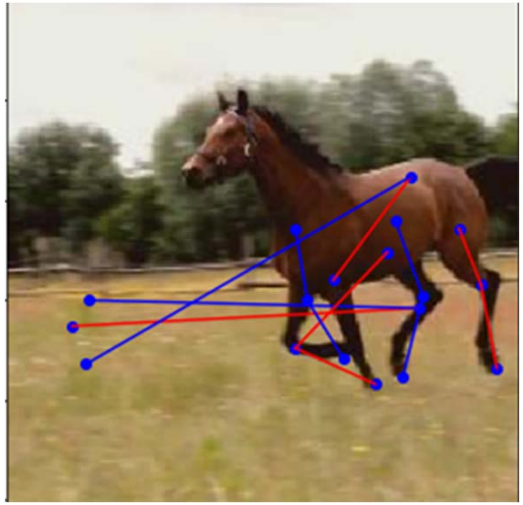
- Stage 1: Use synthetic data as if it was real data (naïve). *Failed due to the big domain gap between real and synthetic stimuli.*
- Stage 2: Use diversity of lighting/viewpoint/texture/background to help solve the domain gap. *Success by combining diversity with learning from simulation.*
- Stage 3: Use properties of synthetic data to scale up to multiple objects and multiple tasks. *Possible by exploiting the synthetic annotations.*

JOHNS HOPKINS UNIVERSITY

# 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).





JOHNS HOPKINS UNIVERSITY

# 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)

JOHNS HOPKINS UNIVERSITY

# 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.

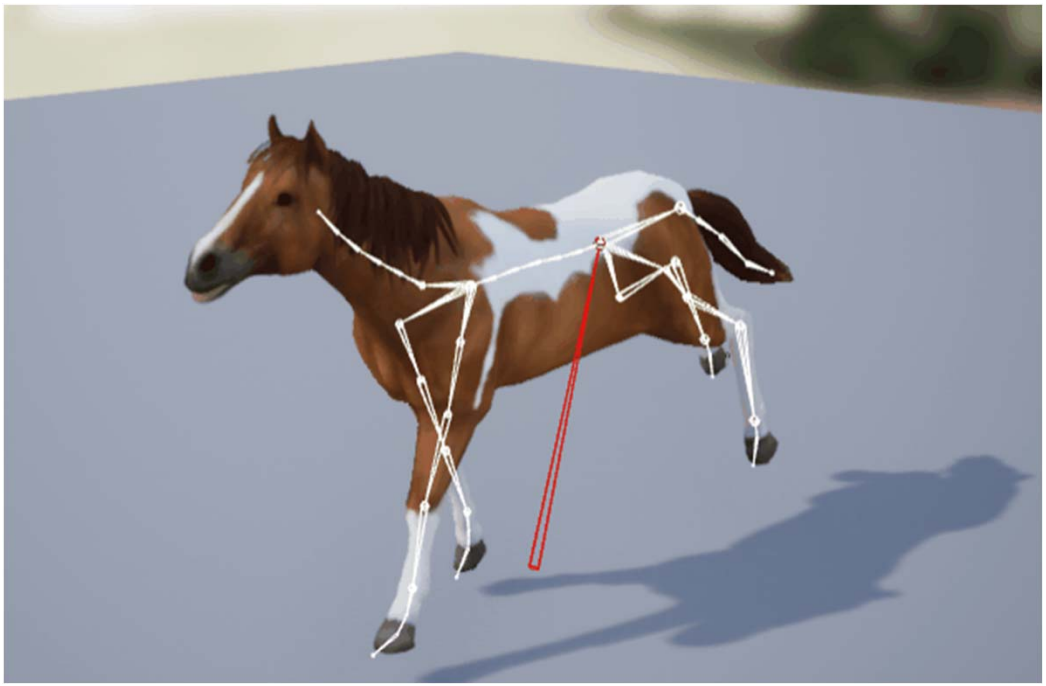
JOHNS HOPKINS UNIVERSITY

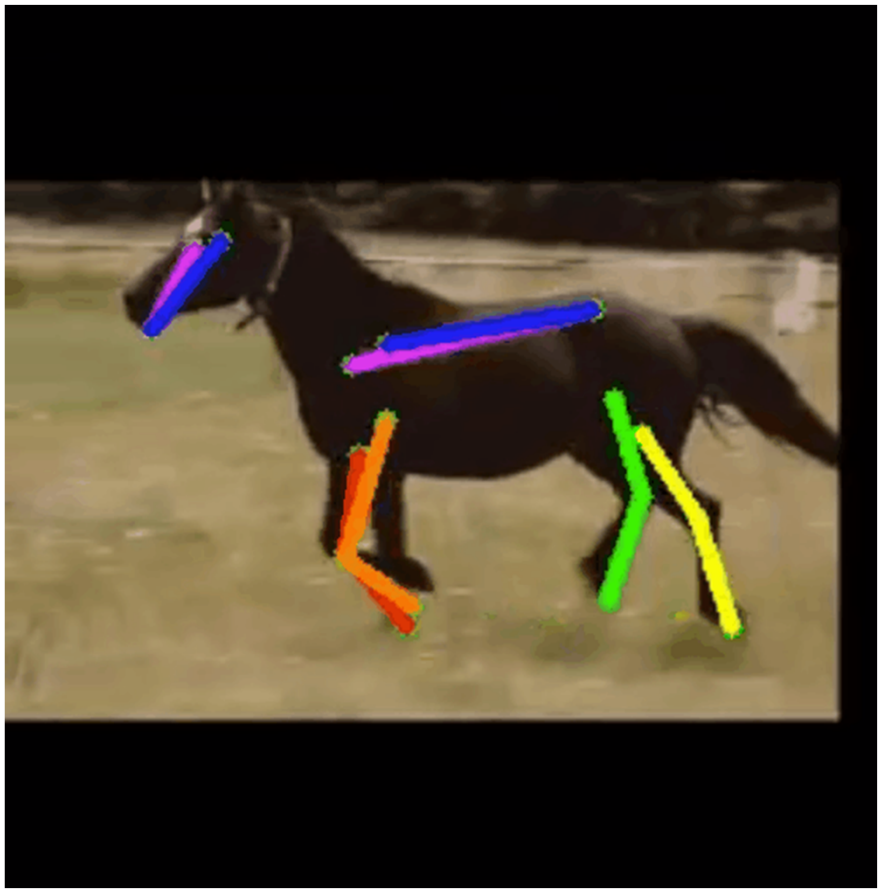
# How to improve performance?

- 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.*

JOHNS HOPKINS UNIVERSITY

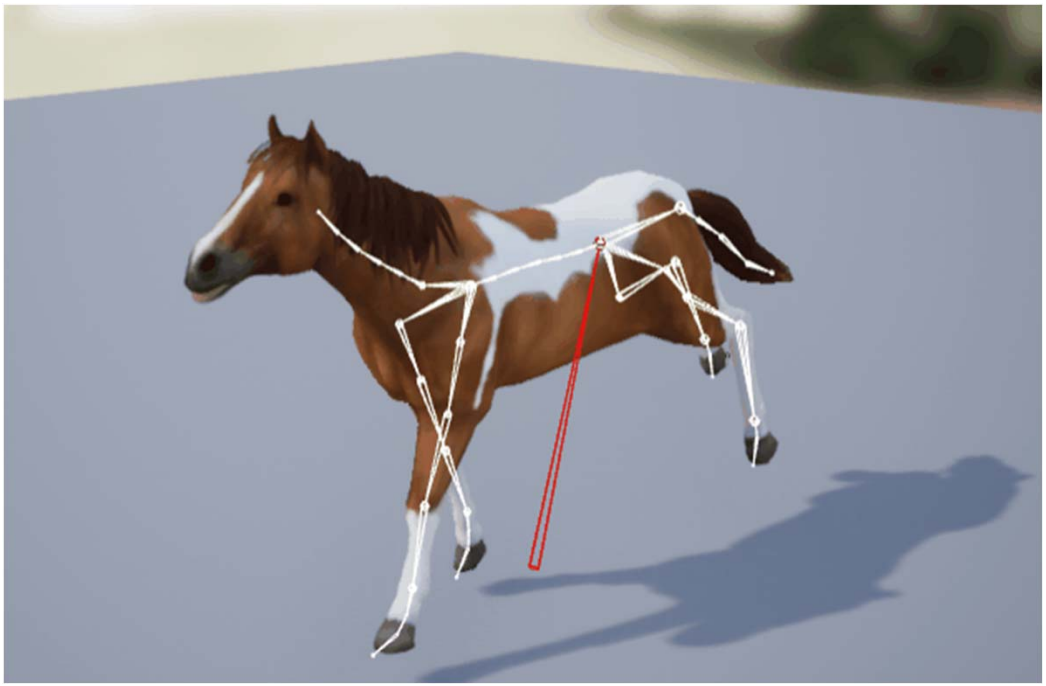
# Animal keypoint video (2)

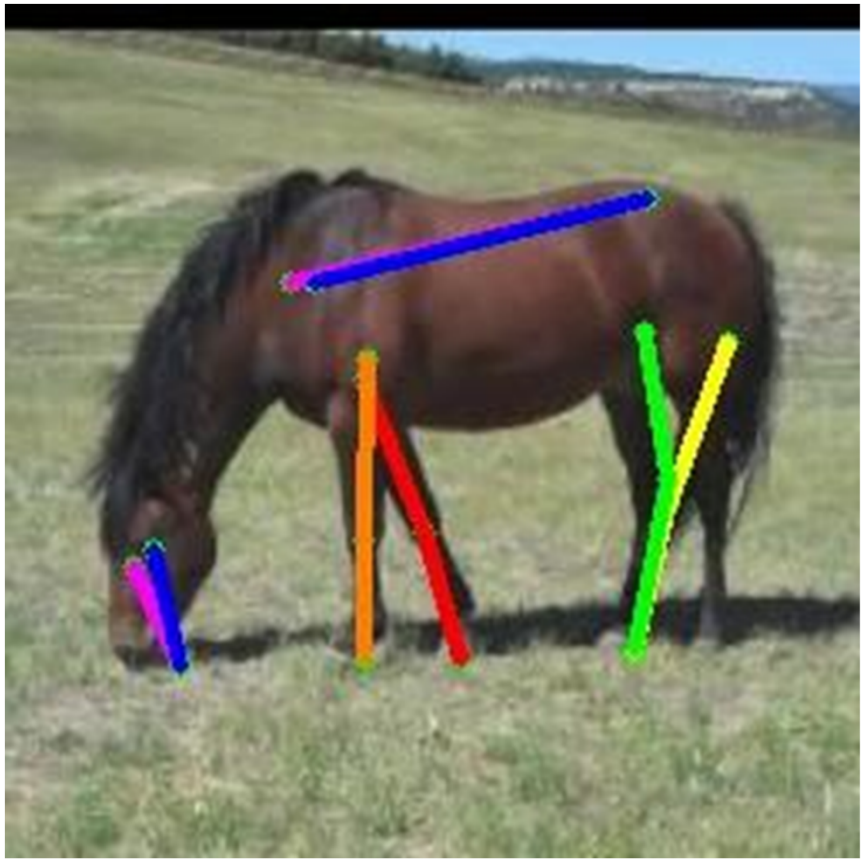




JOHNS HOPKINS UNIVERSITY

# Animal keypoint video (2)

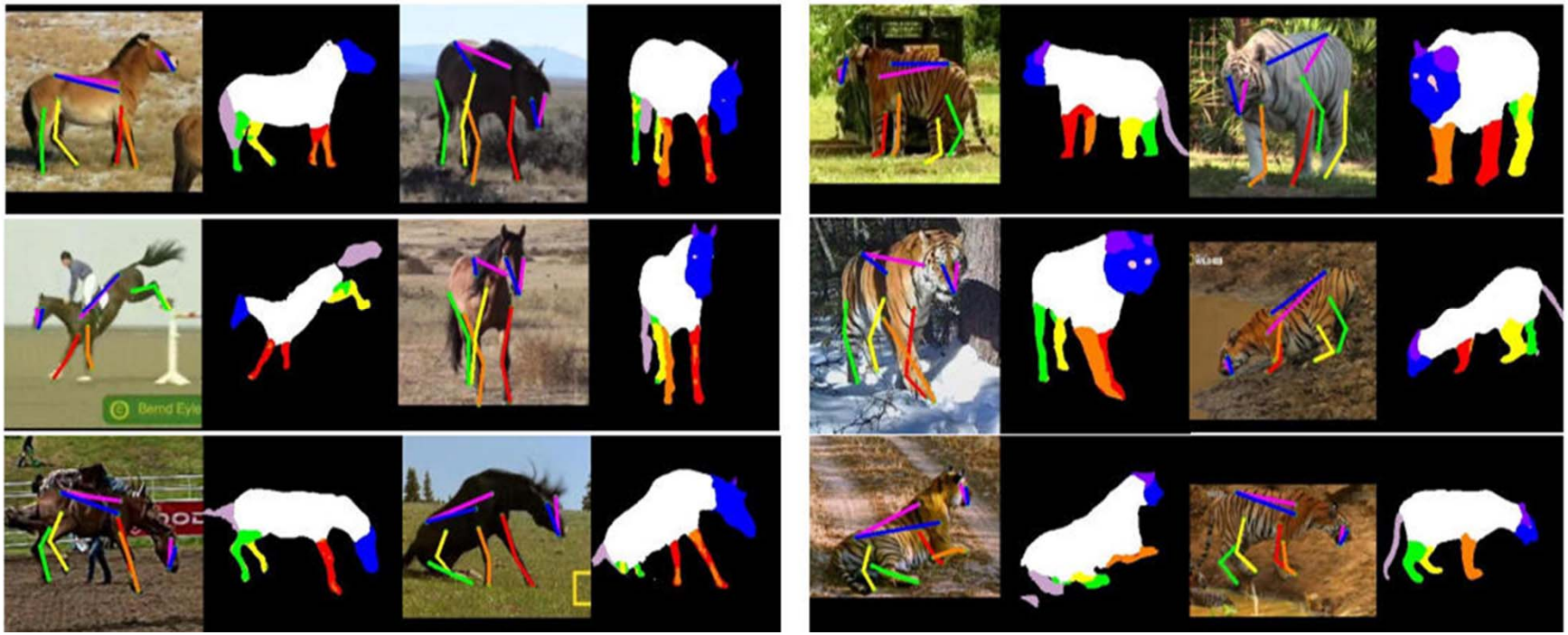




JOHNS HOPKINS UNIVERSITY

# Stage 3: Scale Up –extend to new tasks.

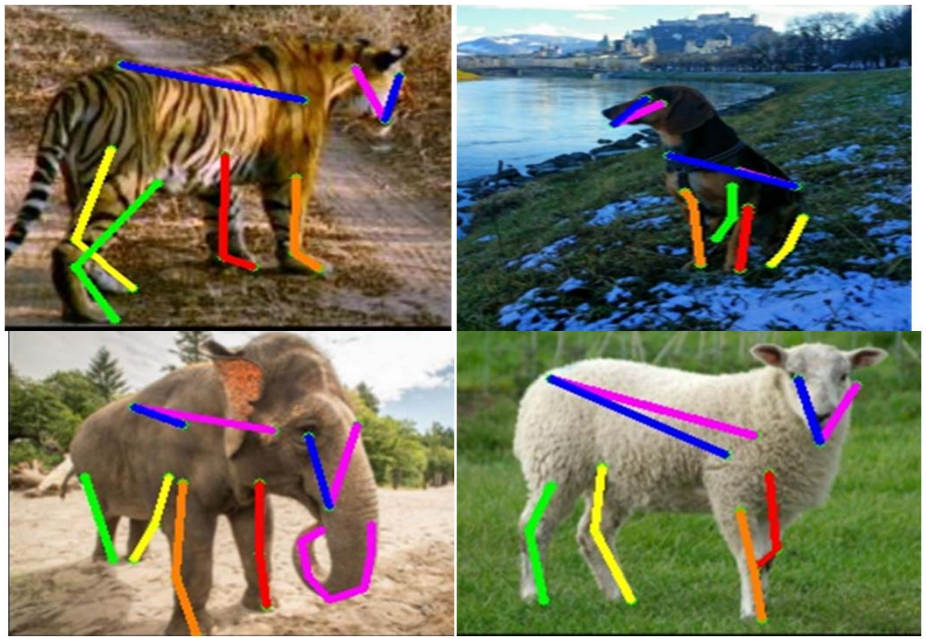
New Visual Task: Part Segmentation: Identify head, torso, legs, tails.
Same diversity plus learning strategy.



JOHNS HOPKINS UNIVERSITY

# Scale Up -- extend to more categories

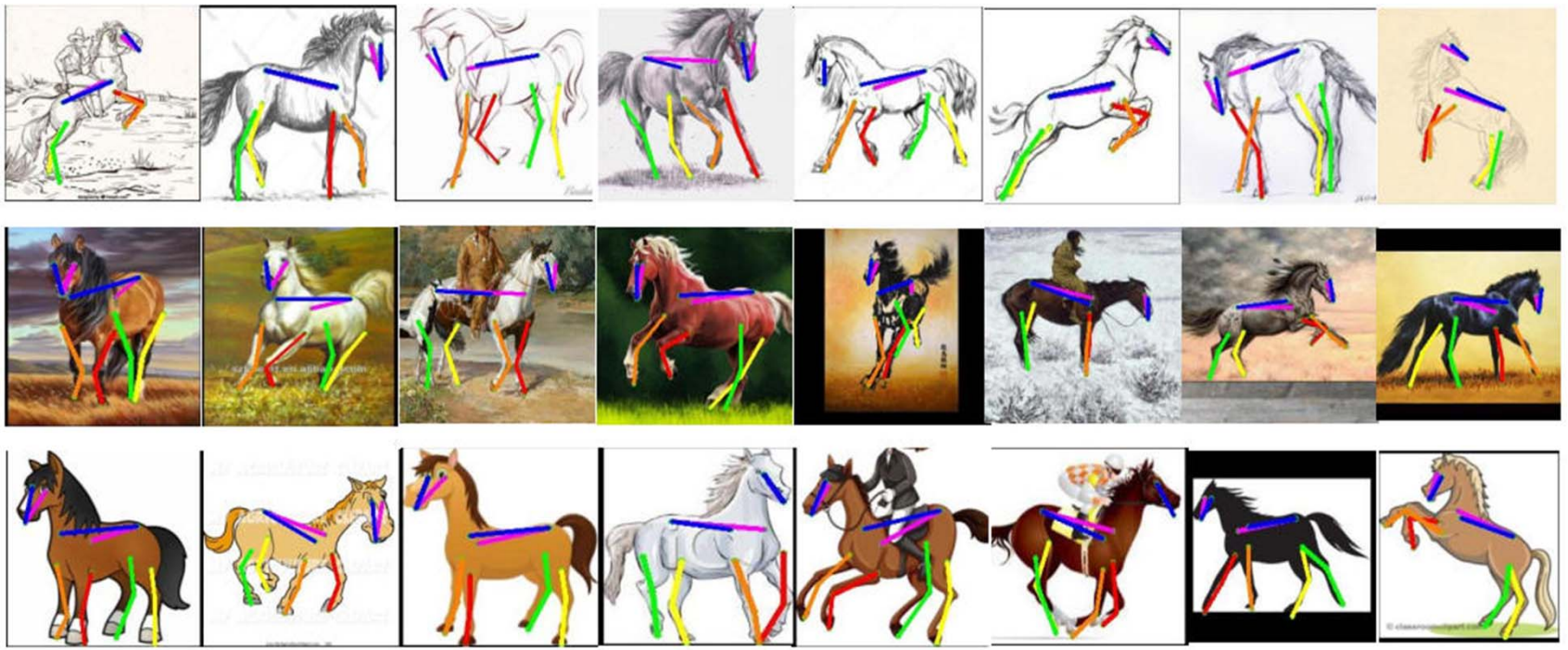
"You only annotate once" (for each object category) but same diversity and learning strategies still apply.



JOHNS HOPKINS UNIVERSITY

# Scale Up: extend to different domains.

Better Domain Generalization: line drawings, pictures.



JOHNS HOPKINS UNIVERSITY

# 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.*

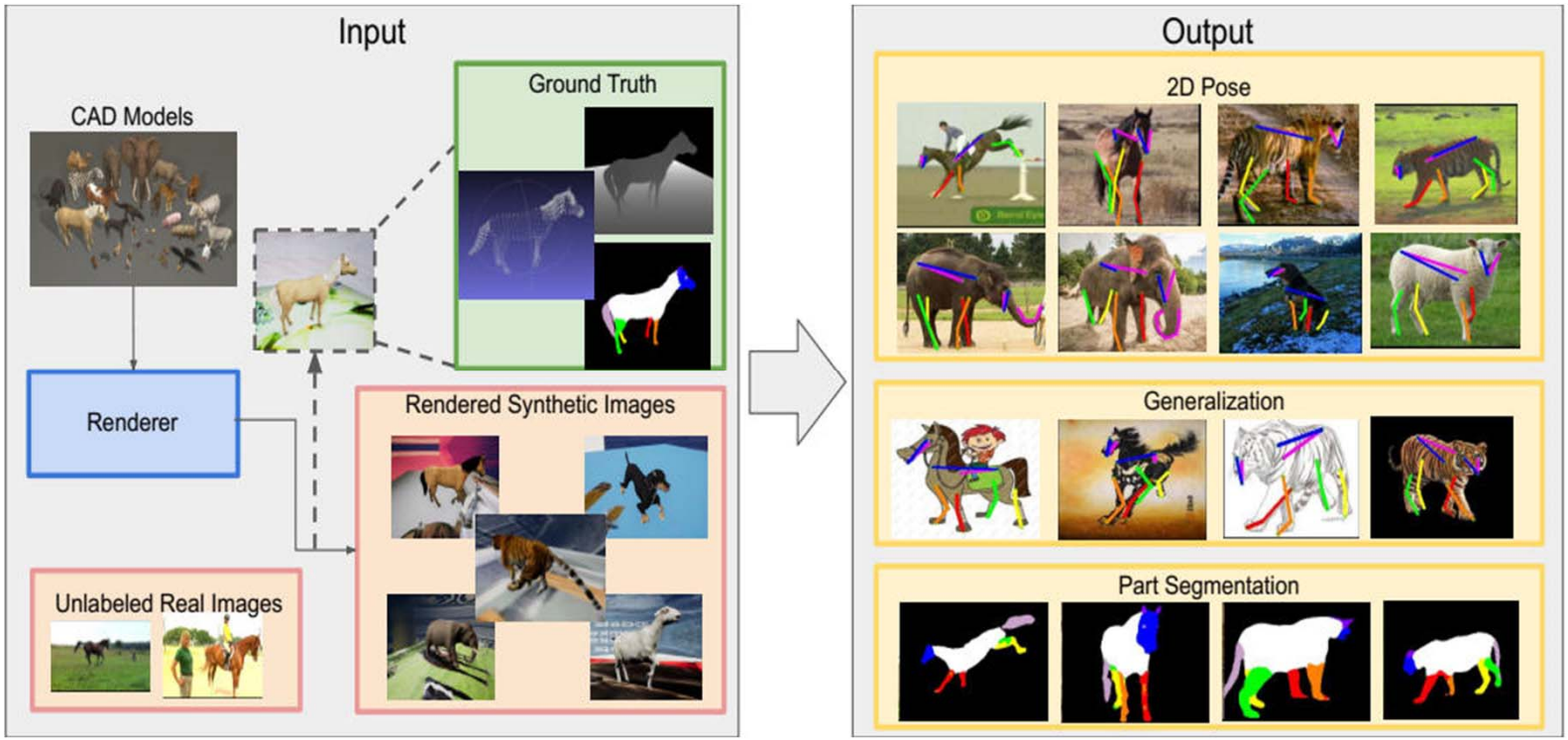
JOHNS HOPKINS UNIVERSITY

# 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

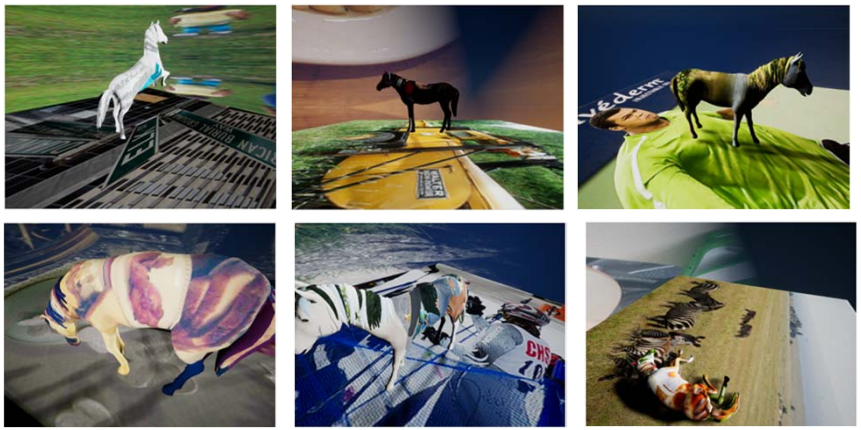
# Synthetic Animal Project



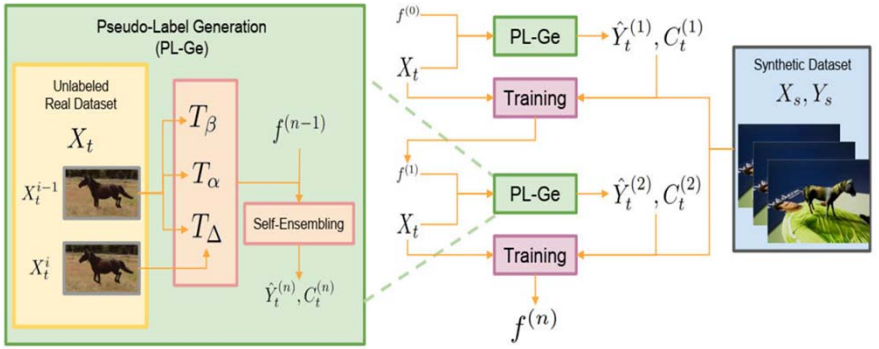

J Mu et al. ‘Learning from Synthetic Animals’. CVPR 2020.

# From Stage 1 to Stage 2: Diversity + Adaptation

## Domain Randomization



## Domain Adaptation

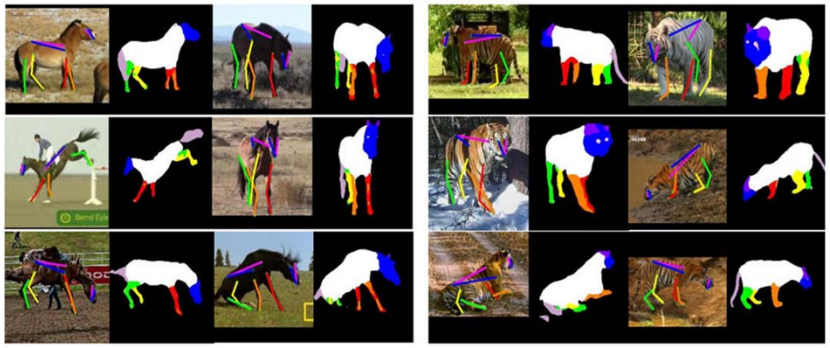



| | Horse Accuracy | | | | | | | | Tiger Accuracy | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Eye | Chin | Shoulder | Hip | Elbow | Knee | Hoove | Mean | Eye | Chin | Shoulder | Hip | Elbow | Knee | Hoove | Mean | |
| *synthetic + real* | | | | | | | | | | | | | | | | | |
| Real | 79.04 | 89.71 | **71.38** | 91.78 | 82.85 | 80.80 | 72.76 | 78.98 | **96.77** | 93.68 | 65.90 | 94.99 | 67.64 | 80.25 | 81.72 | 81.99 | Stage 1 |
| CC-SSL-R | **89.39** | **92.01** | 69.05 | **92.28** | **86.39** | **83.72** | **76.89** | **82.43** | 95.72 | **96.32** | **74.41** | **91.64** | **71.25** | **82.37** | **82.73** | **84.00** | Stage 2 |
| *synthetic only* | | | | | | | | | | | | | | | | | |
| Syn | 46.08 | 53.86 | 20.46 | 32.53 | 20.20 | 24.20 | 17.45 | 25.33 | 23.45 | 27.88 | 14.26 | 52.99 | 17.32 | 16.27 | 19.29 | 21.17 | Stage 1 |
| CycleGAN [45] | 70.73 | 84.46 | 56.97 | 69.30 | 52.94 | 49.91 | 35.95 | 51.86 | 71.80 | 62.49 | 29.77 | 61.22 | 36.16 | 37.48 | 40.59 | 46.47 | |
| BDL [26] | 74.37 | 86.53 | 64.43 | 75.65 | 63.04 | 60.18 | 51.96 | 62.33 | 77.46 | 65.28 | 36.23 | 62.33 | 35.81 | 45.95 | 54.39 | 52.26 | |
| CyCADA [16] | 67.57 | 84.77 | 56.92 | 76.75 | 55.47 | 48.72 | 43.08 | 55.57 | 75.17 | 69.64 | 35.04 | 65.41 | 38.40 | 42.89 | 48.90 | 51.48 | |
| CC-SSL | **84.60** | **90.26** | **69.69** | **85.89** | **68.58** | **68.73** | **61.33** | **70.77** | **96.75** | **90.46** | **44.84** | **77.61** | **55.82** | **42.85** | **64.55** | **64.14** | Stage 2 |

# Stage 3: More gains?

1. stage 2 Vs stage 3 = 4 months Vs 3 weeks
2. More categories? More annotation format? Better domain generalization?

Segmentation Masks



More categories

Domain Generalization

| | Horse | | | | | | Tiger | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Visible Kpts Accuracy | | | Full Kpts Accuracy | | | Visible Kpts Accuracy | | | Full Kpts Accuracy | | |
| | Sketch | Painting | Clipart | Sketch | Painting | Clipart | Sketch | Painting | Clipart | Sketch | Painting | Clipart |
| Real | 65.37 | 64.45 | 64.43 | 61.28 | 58.19 | 60.49 | 48.10 | 61.48 | 53.36 | 46.23 | 53.14 | 50.92 |
| CC-SSL | 72.29 | 73.71 | **73.47** | **70.31** | **71.56** | **72.24** | 53.34 | 55.78 | 59.34 | 52.64 | 48.42 | 54.66 |
| CC-SSL-R | **73.25** | **74.56** | 71.78 | 67.82 | 65.15 | 65.87 | **54.94** | **68.12** | **63.47** | **53.43** | **58.66** | **59.29** |