You Only Annotate Once, and maybe never

Alan Yuille

Bloomberg Distinguished Professor

Depts. Cognitive Science and Computer Science

Johns Hopkins University

Why I believe in learning with little supervision. The Perspective from Human Vision.

- Human infants learn vision without direct supervision. And, despite a few recent claims, the human visual system remains the gold standard for general purpose vision.
- There is an enormous literature on how infants learn vision. Different visual abilities arise at different times in a stereotyped sequence.
- Infants learn by actively interacting with and exploring the world. They are not merely passive acceptors of stimuli. They are more like tiny scientists who understand the world by performing experiments and seeking causal explanations for phenomena.

Arterberry and Kellman. “Development of Perception in Infancy” (2016).
Gopnik et al. “The Scientist in the Crib”. (2000)

The Perspective from Computer Vision

- The current evaluation paradigm for computer vision assumes finite annotated datasets which are balanced for training and testing.
- This is limited for several reasons:
- (1) It is hard/impossible to provide annotations for many visual tasks. This biases researchers to work on problems for which annotated datasets exist. My students say "we can't work on this problem because there isn't an annotated dataset". *Fortunately my wife writes an unsupervised algorithm to solve the problem.*
- (2) In real world situations, balanced training and testing datasets do not exist and it is impractical to create them.
- (3) Current datasets are finite-sized, of necessity, and fail to capture the complexity of the real world. They are biased and contain corner cases ("almost everything is a corner case" – professional annotator).
- (4) Fundamentally, the world is combinatorially complex.

A.L. Yuille and C. Liu. "Deep Networks: What Have They Ever Done for Vision?". Arxiv. 2018.

To a New Evaluation Paradigm

- We need to move towards a new paradigm where we separate learning/training.
- *We should train with very little annotated data (rest of talk).*
- We should test over an infinite set of images by studying the worst cases and allowing our "worst enemy" to test our algorithm. An *Adversarial Examiner* who adaptively selects a sequence of test images to probe the weaknesses of your algorithm. *Don't test an algorithm on random samples. Would a professor test students by asking them random questions?*
- M. Shu, C. Liu, W. Qiu, & A.L. Yuille. Identifying Model Weakness with Adversarial Examiner. AAAI. 2020.

I will now give three examples of learning with little, or zero, supervision.

- Part 1. Learning Geometry: by loss functions and exploring the world.
- Part 2. Learning Image Features and Architectures.
- Part 3. Learning to Parse Animals using weak Prior Models. “You Only Annotate Once”.

Part 1: Learning Geometry. Unsupervised Learning by Loss Functions

- Problem: it is hard to obtain datasets with annotated optical flow.
- Solution: unsupervised optical flow (e.g., Zhe Ren et al. 2017).
- Key Idea: use a loss function based on classical optical flow algorithms (local smoothness of motion) to supervise a deep network in an unsupervised manner. Not quite as effective as supervised optical flow, on datasets where annotation is possible, but more general.
- When Zhe Ren visited my group I had a deja vue moment. The algorithm is like an obscure paper in 1995 by Stelios Smirnakis and myself on using neural networks to learn models for image segmentation.
- Very good work by Stelios: *but bad timing, bad choice of publication venue, and bad advertising (no twitter or NYT).* So Stelios had to become a doctor.
- He is now an Associate Professor in the Harvard Medical School.

Learning Geometry by Exploring the World.

- How can an infant learn about the world?
- (I) The infant learns to estimate correspondence between images. This gives the ability to estimate optical flow and stereo correspondence.
- (II) The infant moves in a world where there is a static background and a few moving objects. The infant learns to estimate 3D depth by factorizing the (estimated) correspondence into 3D depth and camera/infant motion. Hence the infant estimates depth of the background scene.
- (III) The infant uses the estimated depth to train deep networks to estimate depth from single images. And to estimate stereo depth.
- (IV) The infant detects objects moving relative to the background (inconsistency between factorized correspondence and optical flow) and uses rigidity and depth from single images to estimate shape of these moving objects.
- Note: in practice, it is more complicated. There are a series of papers on this topic (USC, Baidu, etc.) with nice results on KITTI and other datasets.
- My group is only tangentially involved. Chenxu Luo

was an intern with ex-group member Peng Wang (Baidu).

Part 2. Unsupervised Learning of Features and Neural Architectures.

- There is work on learning visual features by exploiting a range of signals of techniques –rotation, colorization, jigsaw puzzle.
- Unsupervised features are very useful. E.g., (i) to enable a simple classifier for classification given these features as input, (ii) to perform domain transfer, (iii) even to model how an infant learns image features?
- But what about learning the neural architecture? There is much recent work on Neural Architecture Search (NAS). But can this be learnt in an unsupervised manner?
- Yes! Chenxi Liu et al. “Are Labels Necessary for Neural Architecture Search”? Arxiv. 2020.

Signals to Exploit

In this project, we rely on self-supervised objectives

- We will use “unsupervised” and “self-supervised” interchangeably
- These objectives were originally developed to transfer learned weights
- We study their ability to transfer learned architecture

Rotation

Colorization

Jigsaw Puzzle

Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." In ICLR. 2018.
Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." In ECCV. 2016.
Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." In ECCV. 2016.

Signals to Exploit

In this project, we rely on self-supervised objectives

- We will use “unsupervised” and “self-supervised” interchangeably
- These objectives were originally developed to transfer learned weights
- We study their ability to transfer learned architecture

Using these self-supervised objectives, we conduct two sets of experiments of complementary nature

- Sample-Based
- Search-Based

Sample-Based Experiments

Experimental design:

- Sample 500 unique architectures from a search space
- Train them using Rotation, Colorization, Jigsaw Puzzle, and (supervised) Classification
- Measure rank correlation between pretext task accuracy and target task accuracy

Advantage:

- Each network is trained and evaluated individually

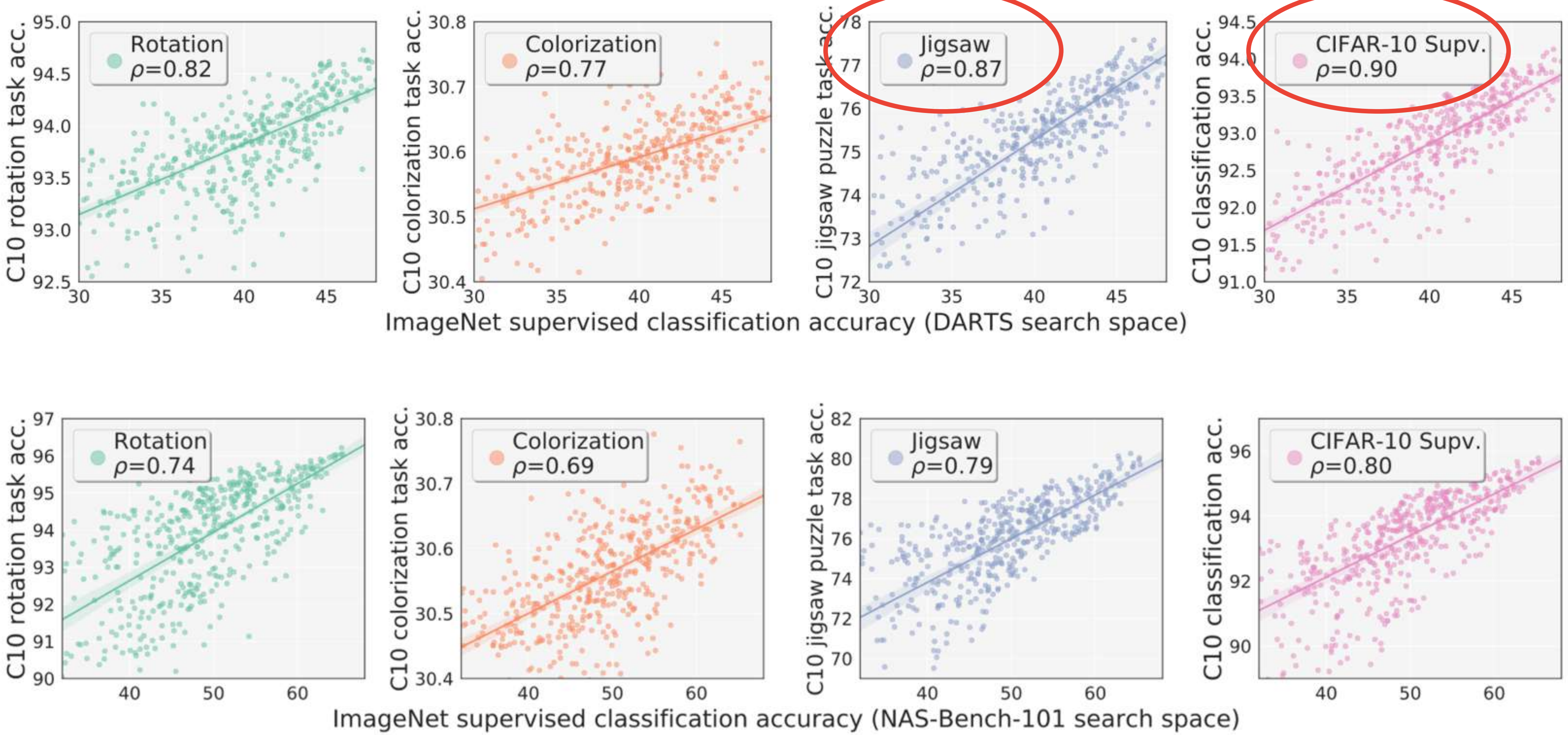
Disadvantage:

- Only consider a small, random subset of the search space

Sample-Based Experiments

Correlation is high!

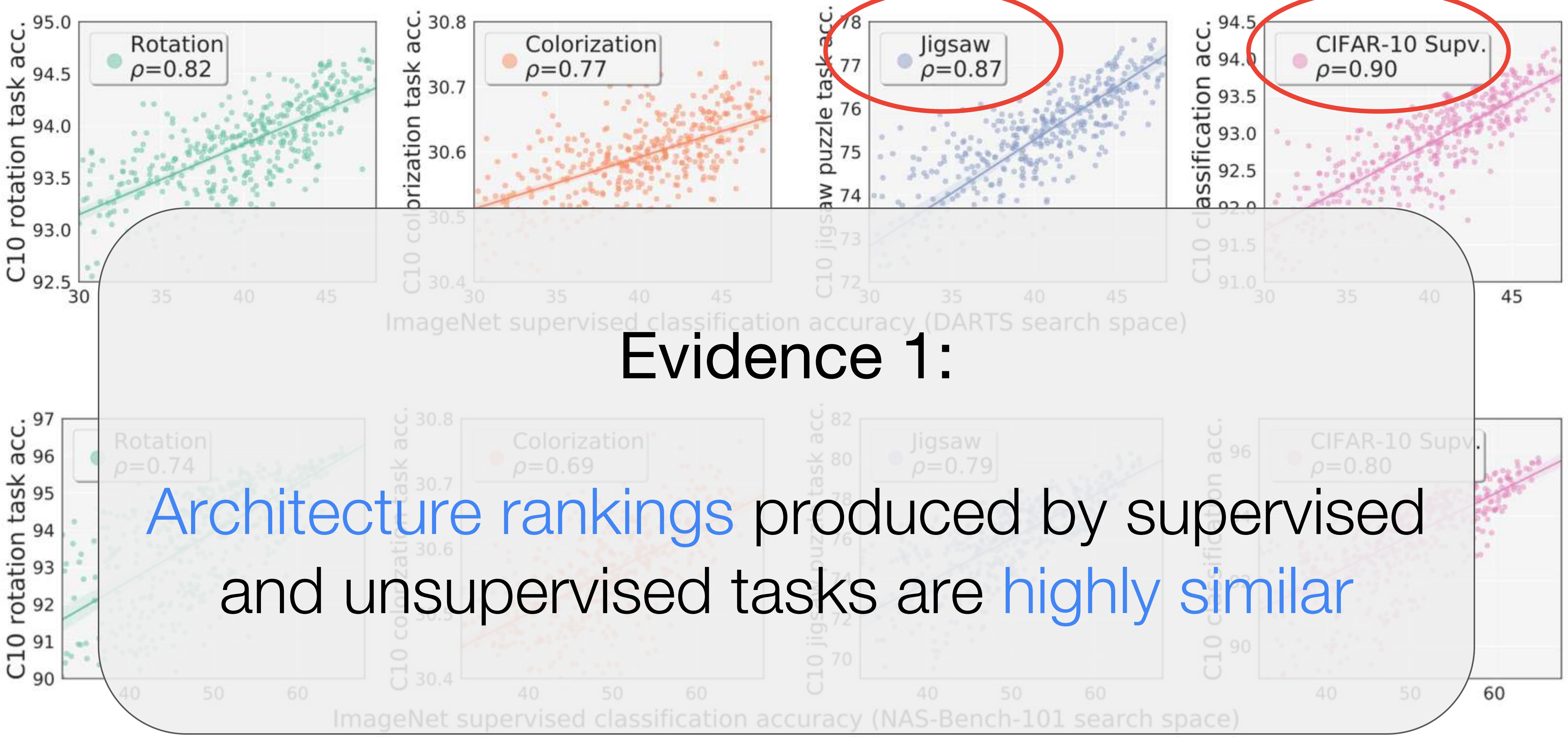
Commonly used proxy in NAS



Sample-Based Experiments

Correlation is high!

Commonly used proxy in NAS



Evidence 1:

Architecture rankings produced by supervised and unsupervised tasks are highly similar

Search-Based Experiments

Experimental design:

- Take a well-established NAS algorithm (DARTS)
- Replace its search objective with Rotation, Colorization, Jigsaw Puzzle
- Train from scratch the searched architecture on target data and task

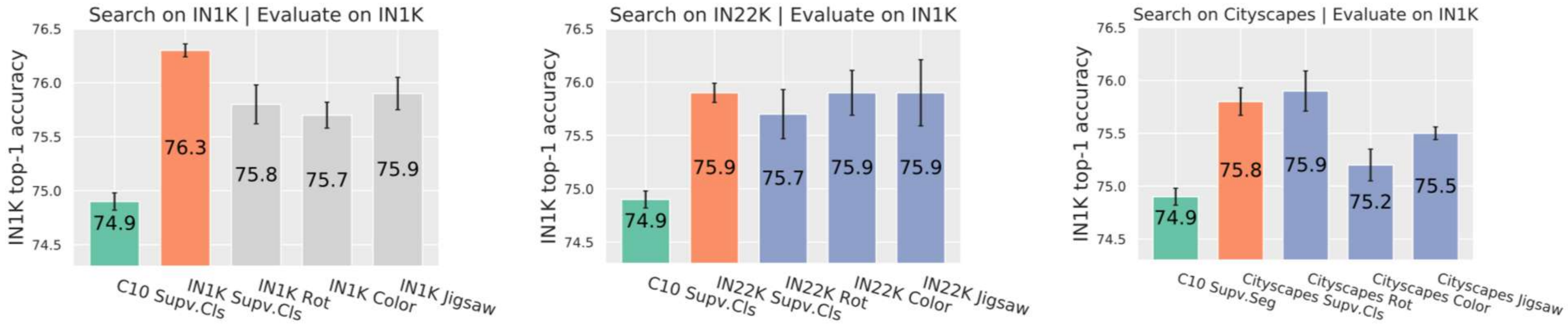
Advantage:

- Explore the entire search space

Disadvantage:

- Training dynamics mismatch between search phase and evaluation phase

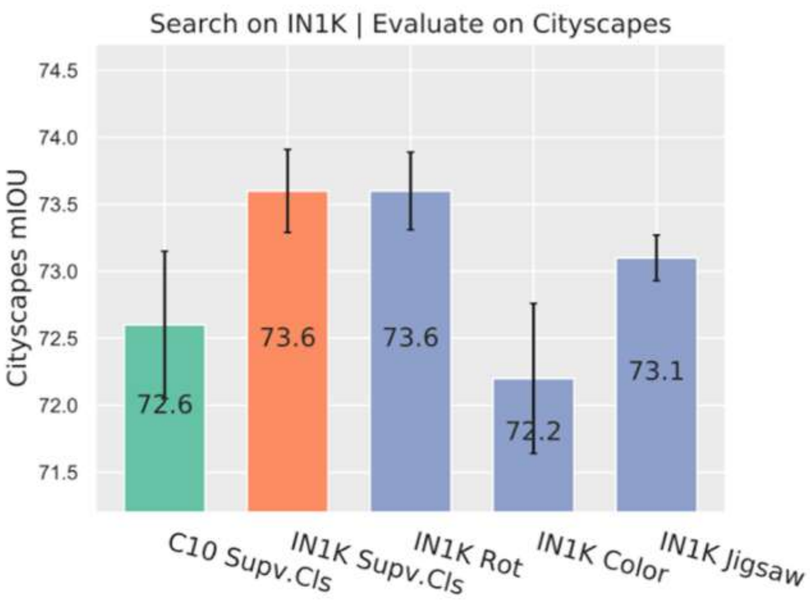
Search-Based Experiments: ImageNet Classification

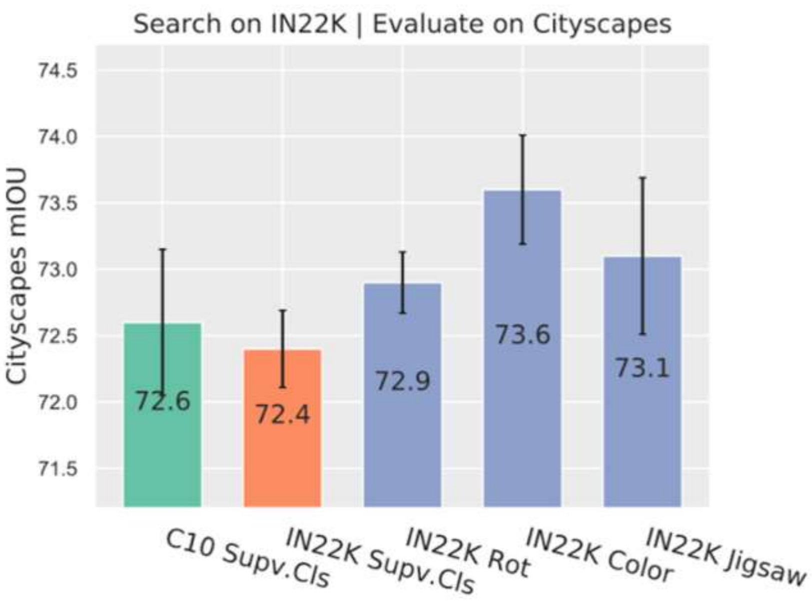


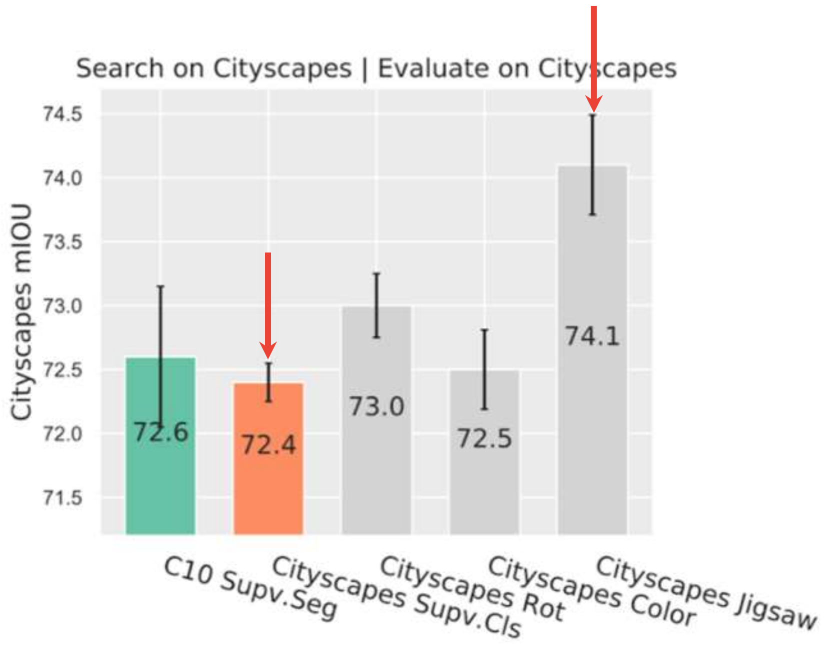
- UnNAS is better than the commonly used CIFAR-10 supervised proxy
- UnNAS is comparable to (supervised) NAS across search tasks and datasets
- UnNAS even outperforms the state-of-the-art (75.8) which uses a more sophisticated algorithm

Xu, Yuhui, et al. "Pc-darts: Partial channel connections for memory-efficient differentiable architecture search." In ICLR. 2020.

Search-Based Experiments: Cityscapes Sem. Seg.

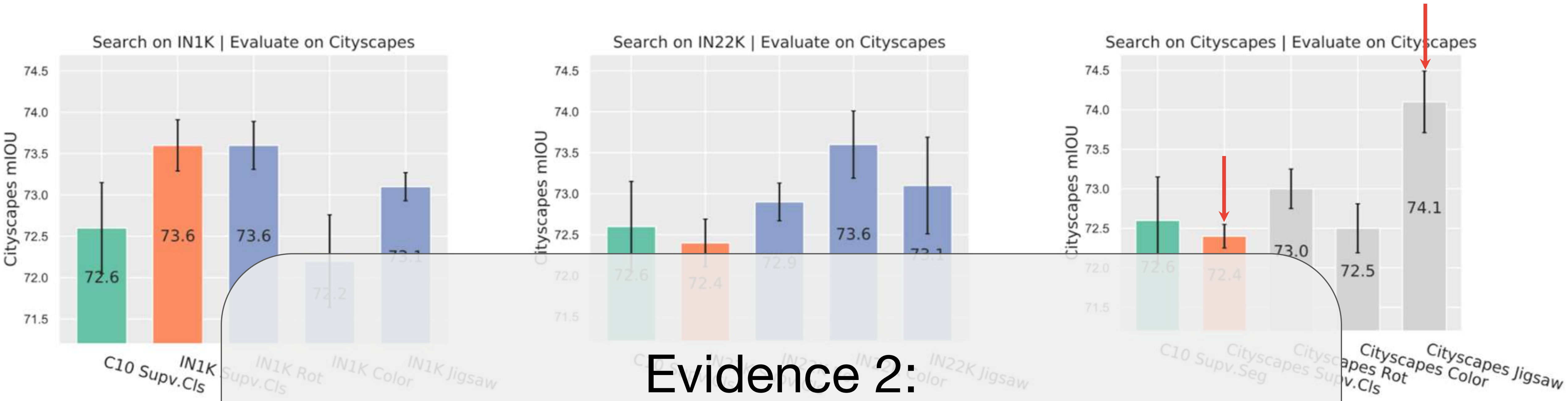






- UnNAS is better than the commonly used CIFAR-10 supervised proxy
- UnNAS is comparable to (supervised) NAS across search tasks and datasets
- Even a case where UnNAS is clearly better than supervised NAS

Search-Based Experiments: Cityscapes Sem. Seg.

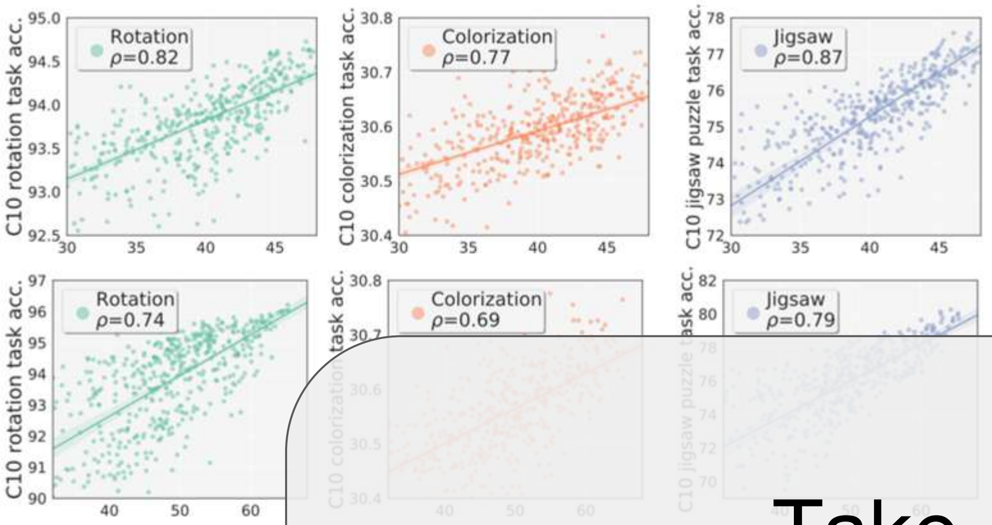


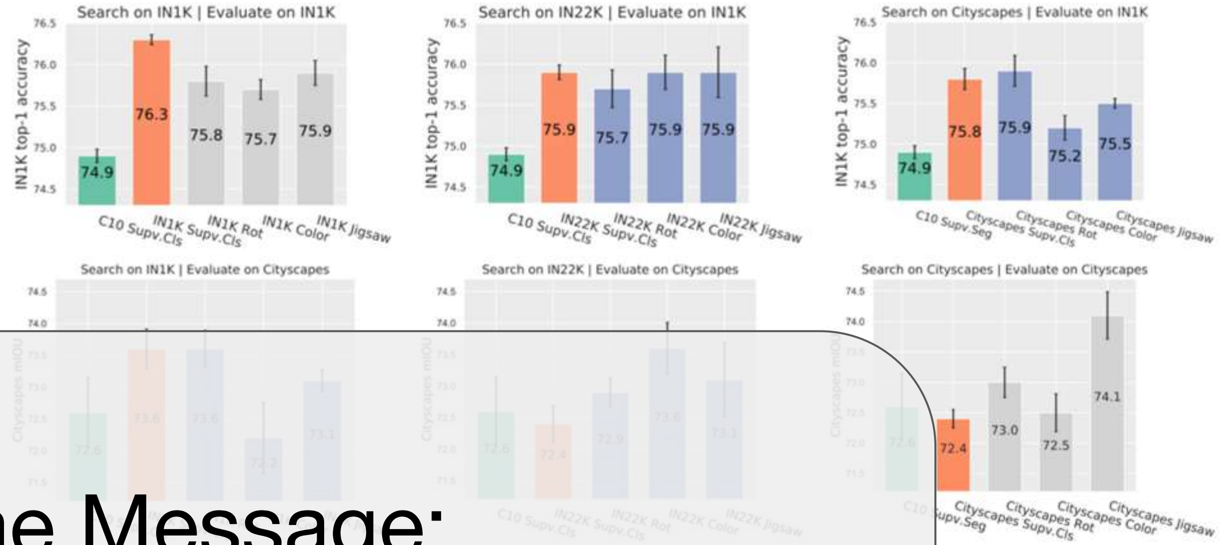
- UnNAS is better than the commonly used CIFAR-10 supervised proxy
- UnNAS is comparable to (supervised) NAS across search tasks and datasets
- Even a case where UnNAS is clearly better than supervised NAS

Evidence 2:

UnNAS architectures are comparable in performance to their supervised counterpart

Evidence 1 + Evidence 2





Take-Home Message:

To perform NAS *successfully*,
labels are *not necessary*

Part 3. Learning to Parse Animals with Weak Prior Knowledge: “You Only Annotate Once”.

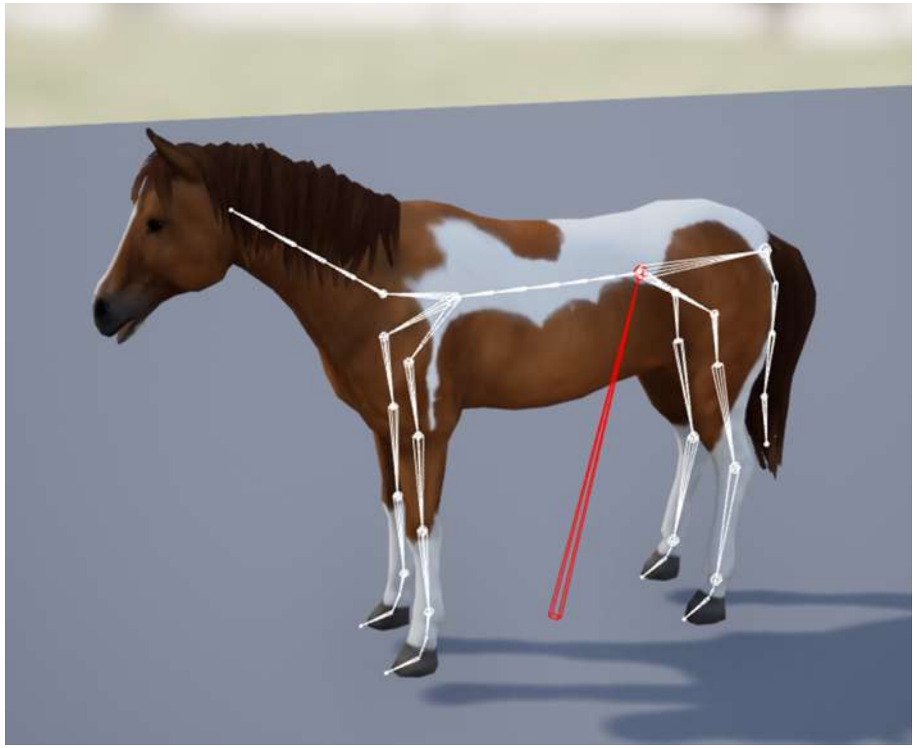
- Infants play with toys.
- An infant can play with a toy horse, or a toy dog.
- The infant can explore what geometric configurations it can take (without breaking) and identify the key-points where it bends.
- The infant can see the toy horse from different viewpoints and under different lighting conditions.
- The infant can paint the horse, or smear food on it, to see how the appearance changes.
- In short, the infant can build a computer graphics model of the horse. The infant has “annotated a horse once”.
- How can this help the infant detect and parse real world horses?

You Only Annotate Once

- Key ideas:
- (I) Take a computer graphics model of a horse, or tiger, and annotate its key-point. You only annotate once.
- (II) Generate a large set of simulated images (with key-points known) with diversity of viewpoint, pose, lighting, texture appearance, and of background.
- (III) Train a model for detecting key-points on these simulated images.
- *But these images are not very realistic and are of a single horse only. Their performance at key-point detection is weak on real images.*
- (IV) Retrain the key-point detection using self-supervised learning on real images of horses including videos.
- *Performance is now much better.*
- *Jiteng Mu, Weichao Qiu, et al. CVPR (oral). 2020.*

Animal Parsing

- Annotate Key Points of a Synthetic Animal. Goal: parse real animal.
- J. Mu, W. Qiu, et al. CVPR. 2020.

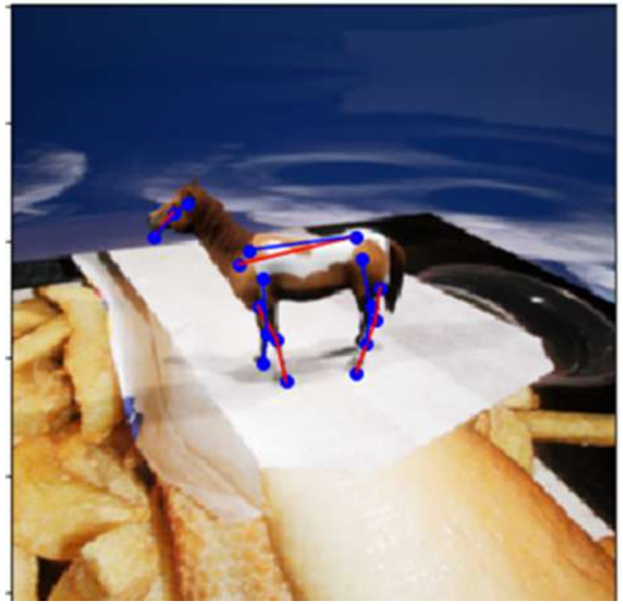


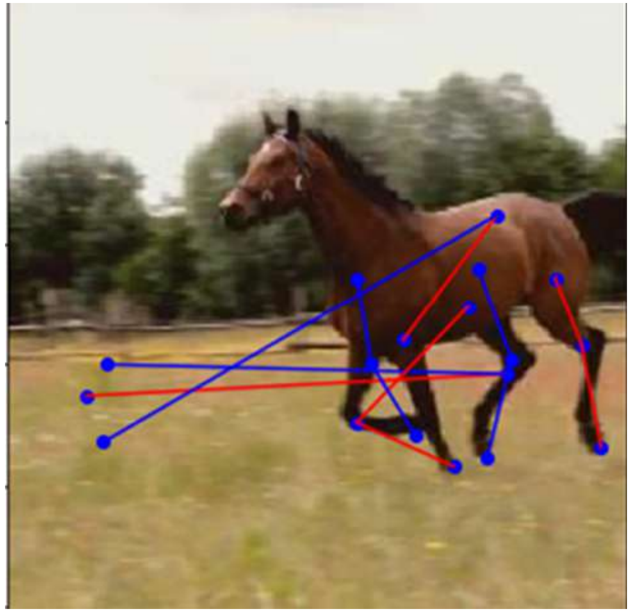
History of this project

- Stage 1: Use synthetic data as if it was real data (naïve). *Failed due to the big domain gap between synthetic and real images.*
- Stage 2: Use diversity 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. *Very fast, by exploiting the synthetic annotations.*

Stage 1: Naïve Strategy does not work

- Train using synthetic data only.
- Works well on synthetic data, but very badly on real data (technically – the deep network features are too different).





How to Improve Performance?

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

- This is intuitive, but we could not get it to work.

- Results are terrible. By contrast, Training with Real Data gives (78.98 PCK@0.05 for keypoint detection)

$599 Image source: turbosquid.com $799

Stage 2: Realism versus Diversity tradeoff

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

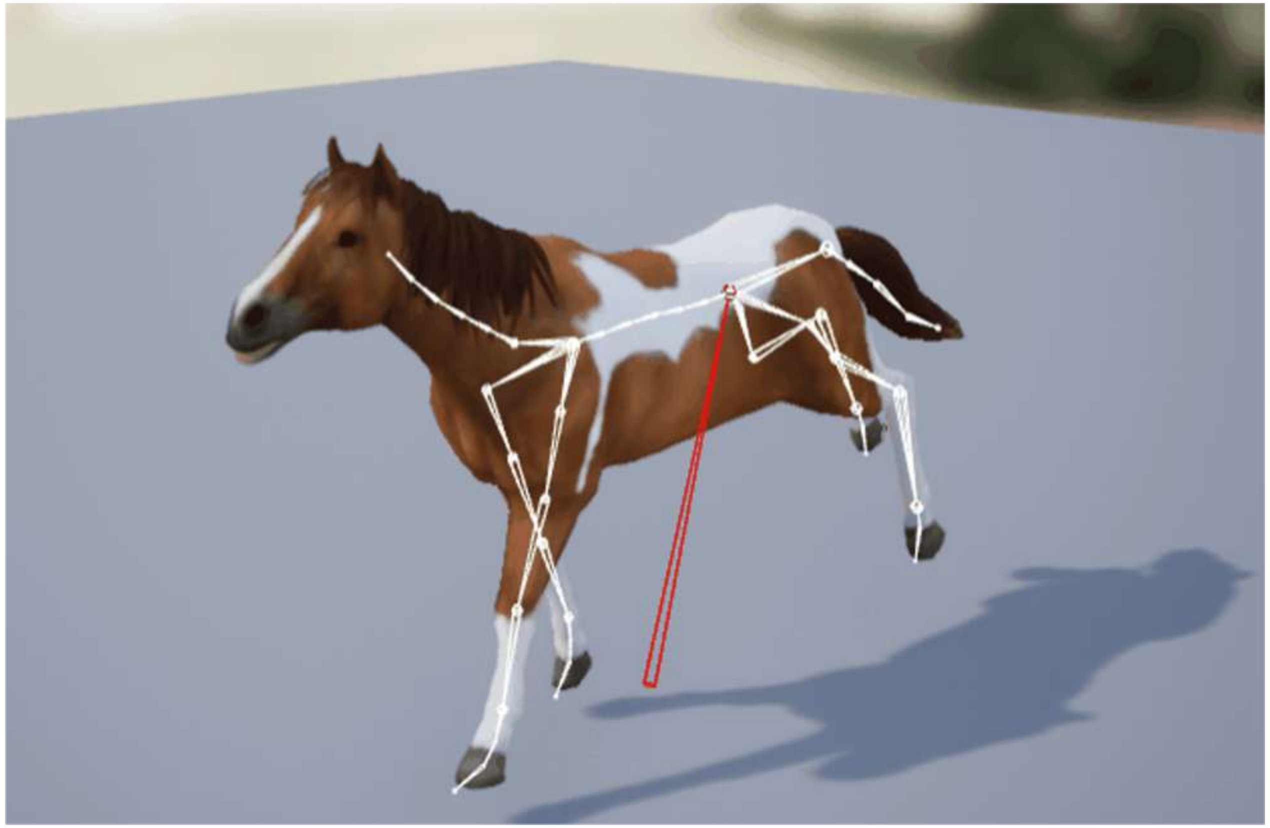
- Instead:
- *(I) Increase diversity by randomizing texture, lighting, background. (25.33 PCK@0.05)*
- *(II) Data augmentation – adding Gaussian noise, rotating the images. (60.85 PCK@0.05)*
- Recall Training with Real Data achieves 78.98 PCK@0.05.

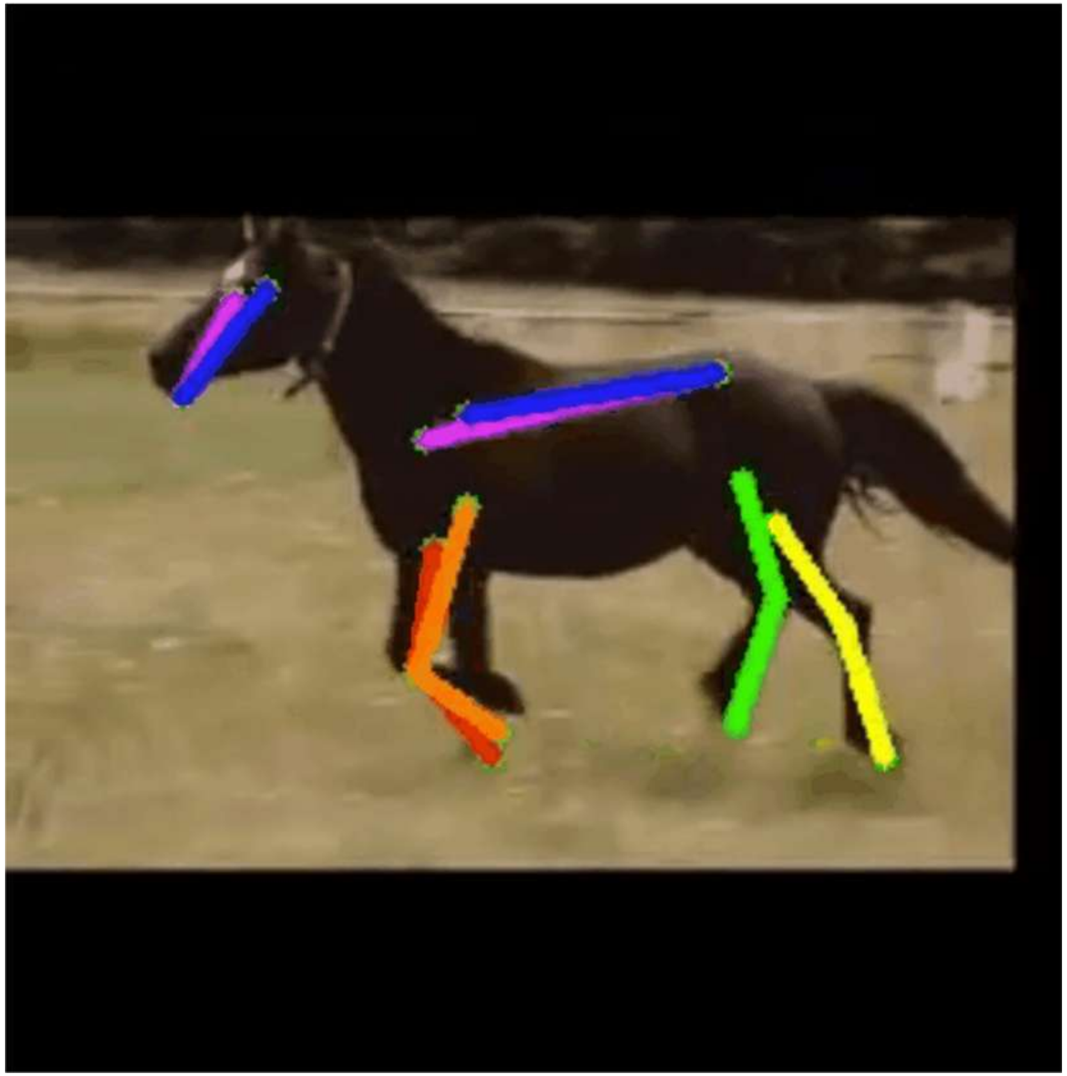
How to improve performance?

- Training with Real Only (78.98)
- Better Data
  - More realistic model, realistic background (intuitive, but not work)
  - Texture Randomization (25.33)
  - Data Augmentation, rotation, gaussian noise (60.84)
- Better Training
  - 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.*

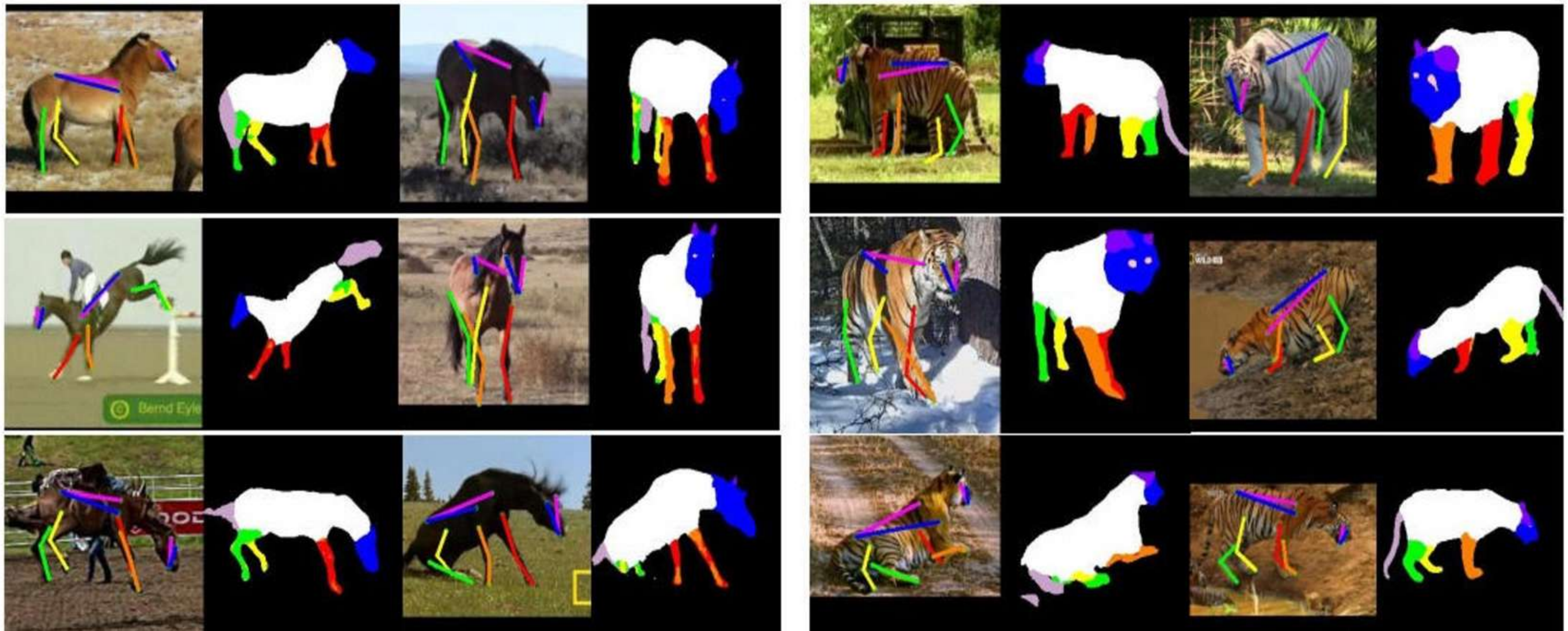
An animal keypoint video





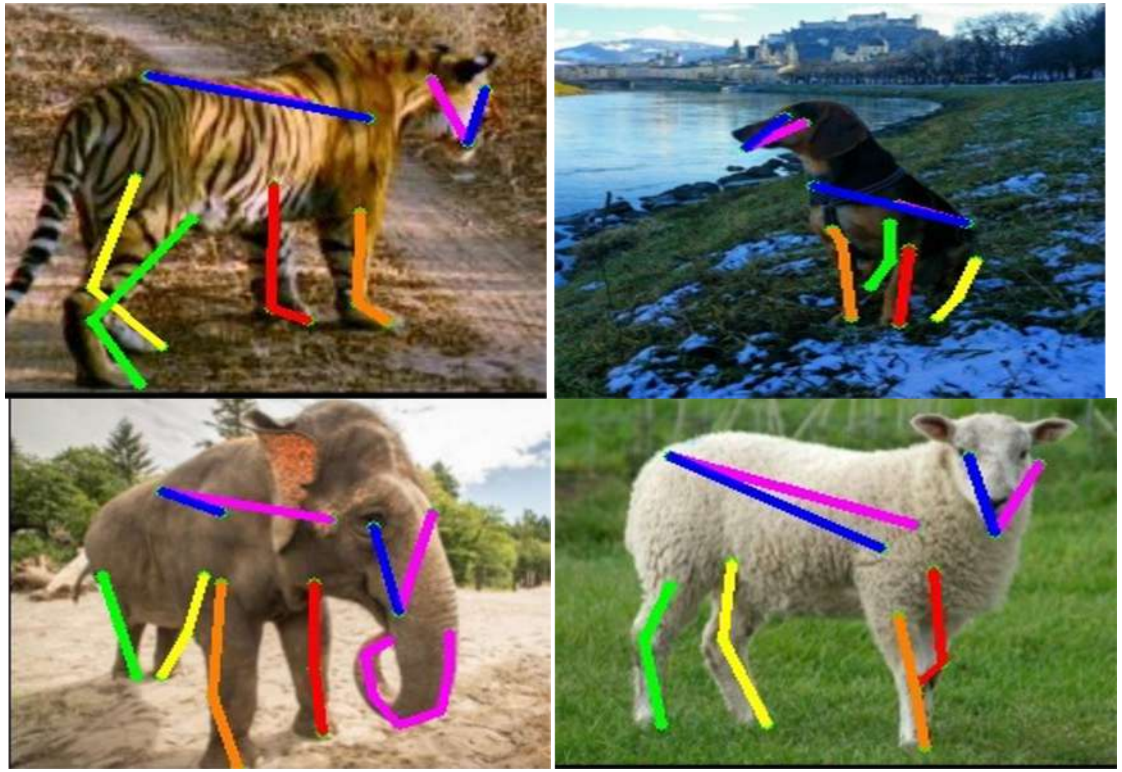
Stage 3: Scale Up –extend to new tasks.

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



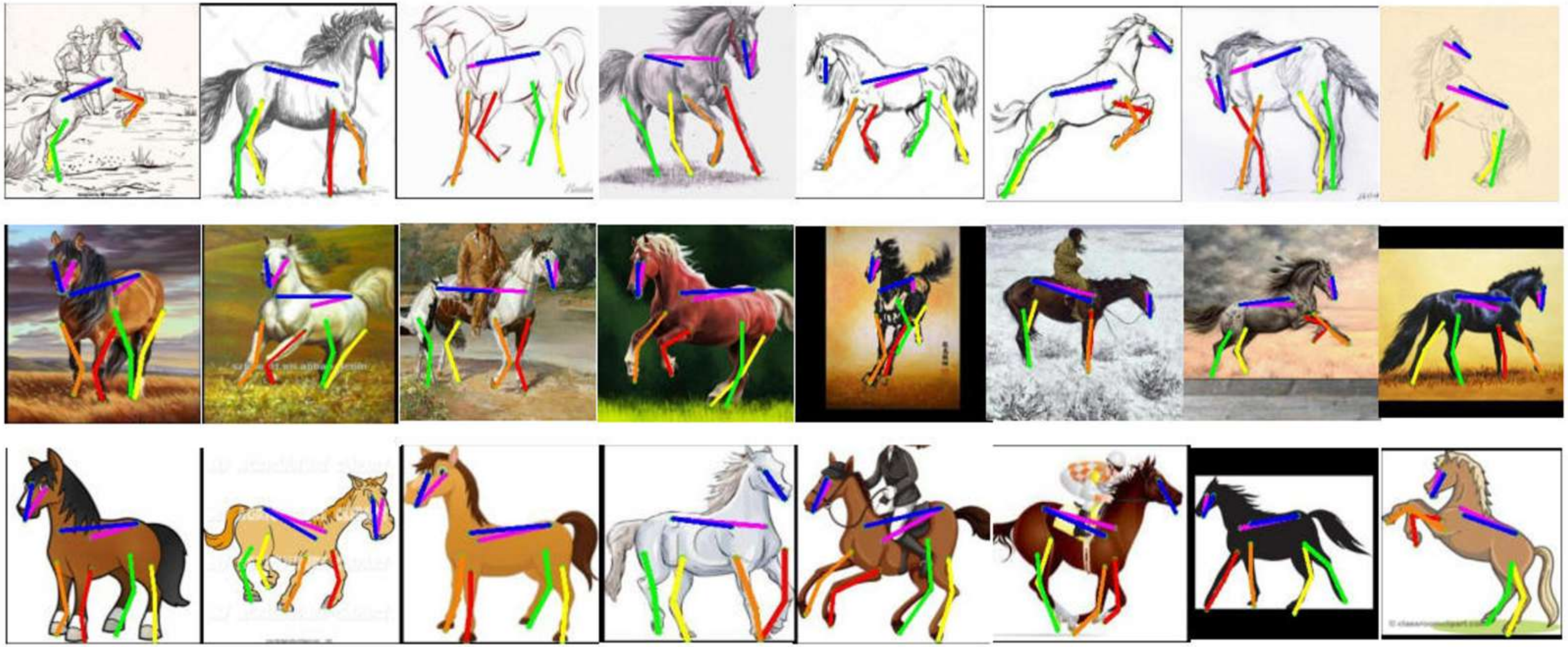
Scale Up -- extend to more categories

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



Scale Up: extend to domain generalization

Better Domain Generalization



Conclusion

- *Learning with weak supervision is not only very important. It is also possible and practical.*
- Three examples: (I) Learning Geometry. (II) Learning Features and Neural Architectures. (III) Learning to Parse Animals with Weak Prior Knowledge – You Only Annotate Once.
- *To approach human level performance, the computer vision community needs to move to a paradigm where we use limited annotations to train but are tested for our worst case performance on an infinite set (by our worst enemy).*
- Human infant learning, and human visual abilities, are great motivations for the next wave of computer vision!

Brief References (1)

- ***Human Infant Learning:***
- M.E. Arterberry & P.J. Kellman. “Development of Perception in Infancy: The Cradle of Knowledge Revisited”, Oxford University Press, 2016.
- A. Gopnik, A. N. Meltzoff, P. K. Kuhl. “The Scientist in the Crib: What Early Learning Tells Us About the Mind”. William Morrow Paperbacks, 2000.
- ***Learning Correspondence and Geometry:***
- ***S***.M. Smirnakis & A.L. Yuille. “Neural Implementation of Bayesian Vision Theories by Unsupervised Learning.” The Neurobiology of Computation, eds J. M. Bower, Kluwer Academic Publishers 1995; p: 427-432.
- Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, Hongyuan Zha: Unsupervised Deep Learning for Optical Flow Estimation. AAAI 2017: 1495-1501
- Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, Alan Yuille. “Every Pixel Counts ++: Joint Learning of Geometry and Motion with 3D Holistic Understanding”. TPAMI 2019.

- *No space of exhaustive references – sorry. At best, these references offer access to the literature.*

Brief References (2)

Unsupervised Learning of Features and Neural Architectures:

Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." In ICLR. 2018.

Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." In ECCV. 2016.

Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." In ECCV. 2016.

No space for exhaustive references on unsupervised feature learning (sorry).

C. Liu, P. Dollar, K. He, R. Girshick, A. Yuille, S. Xie. Unsupervised Neural Architecture Search. Arxiv. 2020.

We believe this is the first work on unsupervised NAS.

Brief References (3)

- ***Learning by Prior Models. You Only Annotate Once:***
- Jiteng Mu, Weichao Qiu, Gregory Hager, Alan Yuille. “Learning from Synthetic Animals”. CVPR (oral). 2020.
- *See this paper for related references.*

- ***Need for New Testing Paradigm:***
- Shu, Michelle, Chenxi Liu*, Weichao Qiu, and Alan Yuille. "Identifying Model Weakness with Adversarial Examiner." In AAAI. 2020.
- Yuille, Alan L., and Chenxi Liu. "Deep Nets: What have they ever done for Vision?." arXiv preprint arXiv:1805.04025 (2018).
- *Very little literature on these topics.*