Vision as Bayesian Inference: The Importance of Generative Models.

Alan Yuille

Bloomberg Distinguished Professor

Depts. Cognitive Science and Computer Science

Johns Hopkins University

Vision: Interpreting & Understanding Images.

- **Vision is an extremely exciting, challenging, and interdisciplinary field of research. It has developed by borrowing and adapting techniques from many disciplines. Mathematics, Computer Science, Statistics, Engineering.**
- A Summer School in 2013 required 70 one hour lectures: *available online*
- http://www.ipam.ucla.edu/programs/summer-schools/graduate-summer-school-computer-vision/?tab=schedule
- *Luckily we invited Rob Fergus to lecture on Deep Networks, which was an obscure unfashionable topic at that time.*
- **Vision is extremely practical – automated cars, face recognition, medical image analysis.**
- Vision is starting to work. Big conferences used to be a couple of hundred people. Now some have tens of thousands. *Every student is a potential billionaire.*

What Techniques are involved?

- The summer school present core techniques in Computer Vision.
- *These techniques include filtering, geometry, differential equations, harmonic analysis, probabilistic methods, machine learning, and many more.*
- Machine learning is a vast field which includes Deep Networks, support vector machines, and many other topics. We also had lectures on sparsity, optimization theory, and much else.
- The school discusses *interactions with related disciplines such as image processing, machine learning, and biological vision.*

My Personal Motivations

- My motivations are to develop vision algorithms that work as well as the human visual system.
- **This is ridiculously ambitious.**
- ***Vision is human's underappreciated superpower. It enables us to get information about the universe from distances of a fraction of an inch to hundreds of light years.***
- *Humans are good at vision because our brains devote an enormous number of neurons to performing it. Roughly half the number of neurons in your cortex are doing vision.*

Why is Vision Difficult?

- **Vision is an inverse problem. Interpreting images requires inverting the process that generates the image.**
- In other words, inverse computer graphics.
- *Key Problem: the complexity and ambiguity of images.*
- ***The set of images is infinite. The set of possible scenes is combinatorially large.***

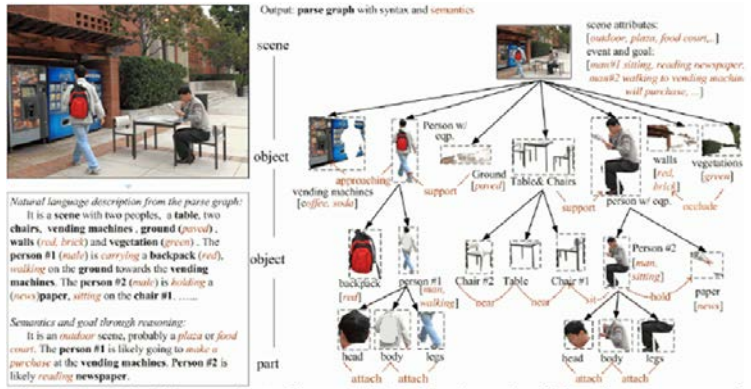
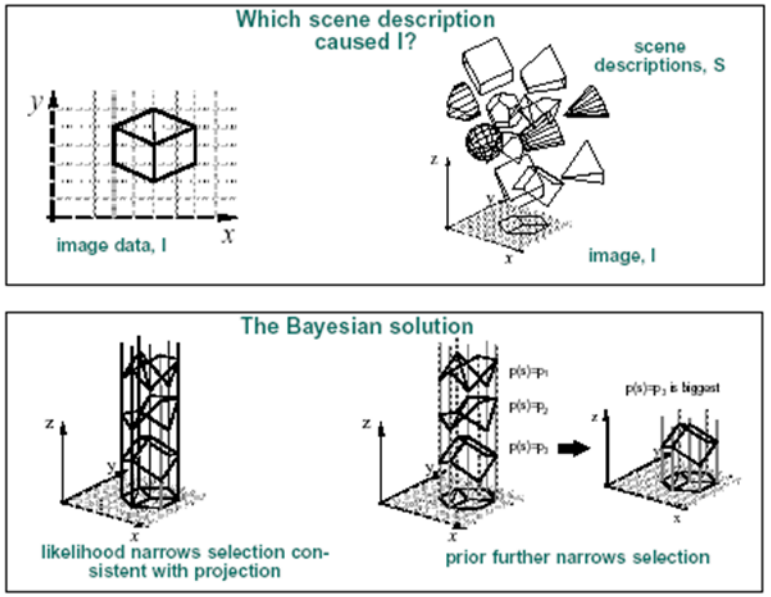


Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description in natural language (bottom-left).

Vision as Bayesian Inference?

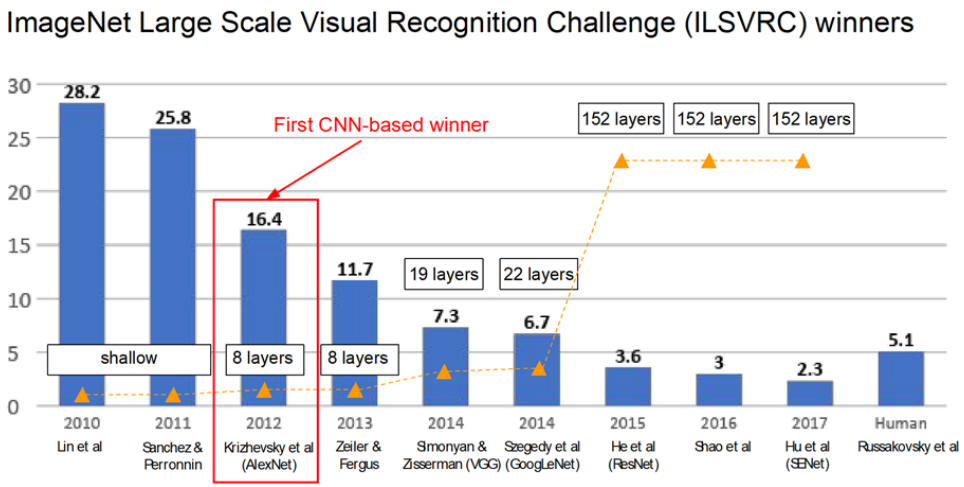
- Inverse Inference and Bayes Theorem:
- P(S|I) = P(I|S)P(S)/P(I).
- The likelihood function P(I|S) rules out most interpretations of the image.
- But there remain many possibilities.
- A prior P(S) is also required.



- Challenges. Can we extend this idea to real images? Can we define probability models that generate real images? *I've seen two genius level mathematicians have a heated arguments about whether this is practical.*

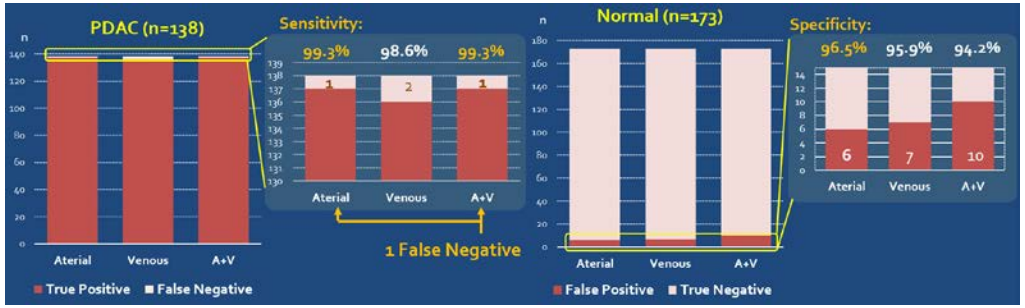
What About Deep Networks (DNs)?

- DNs learn distributions P(S|I) = F(S,I,W). Their weights W are learnt by gradient descent, because F(.,.,.) is a differentiable function of W.
- This is sophisticated regression.
- *Given a well-specified domain with annotated training data DNs perform incredibly well. Much better than anything else.*
- *Performance on the ImageNet challenge may beat humans!*



Early Detection of Pancreatic Tumors.

- **We can detect pancreatic tumors PDAC with very high sensitivity/specificity. Computer Tomography (CT) scans. Yellow is Pancreas, Red is a PDAC tumor.**



- *This is an ideal domain for Deep Nets, if you have annotated data from our colleagues in the JHU medical school.*
- Performance is very good. Approaching the level of an expert radiologist. Can be used to develop systems that can help practicing radiologists.
- ***Early detection of cancer from CT images can save a huge number of lives.***

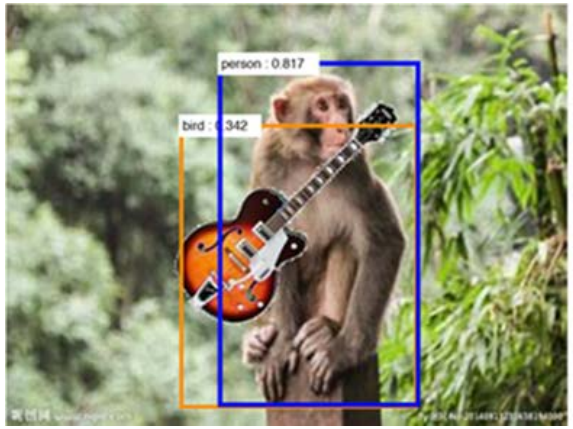
Deep Networks are Extremely Innovative

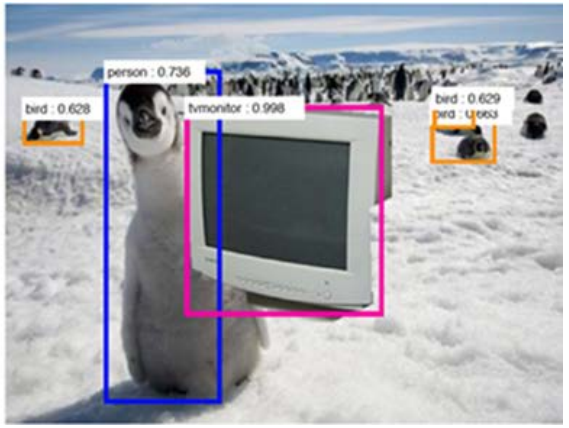
- *Recent inventions include:*
- 1. Unsupervised Learning.
- 2. Neural Architecture Search.
- 3. Adversarial Attacks and Robust Features.
- 4. Neuromodular networks.
- 5. Transformer networks.

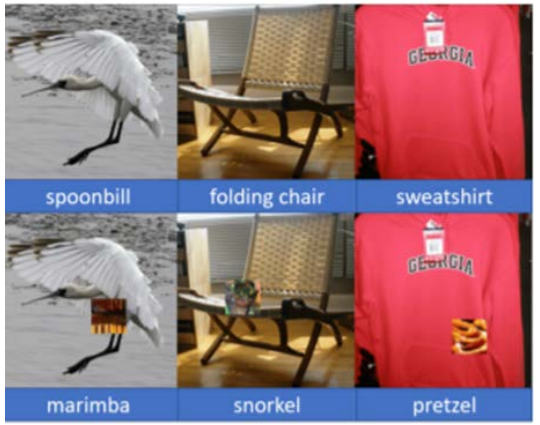
- A recent opinion paper gives entry points into the literature.
- A. Yuille & Chenxi Liu. Deep Nets: What have they ever done for Vision? IJCV. 2020/2021?. https://arxiv.org/abs/1805.04025

But current Deep Networks have limitations

- Deep Networks can be fooled by *occluders* (left), *context* (center) and *patch attacks* (right).







- More generally, current Deep Networks are not robust to changes in the images which would not fool a human.

Why Non-Robust? Complexity and Bias

- Current computer vision systems are trained and tested on finite-sized balanced annotated datasets.
- ***The problem is that the set of images is infinite and the set of scenes is combinatorial complex. Which means that finite-sized datasets are not representative, except in restricted domains like Computer Tomography (CT) Scans.***
- The datasets inevitably have biases which algorithms like Deep Networks can exploit.
- This means that our performance measures – on finite-sized datasets – can be misleading, and not may reflect how the algorithms will perform in real world conditions.

Deep Nets: Pros and Cons

- Deep Networks are amazing techniques that sometimes appear to be superhuman.
- But, in light the infinite set of images and combinatorial complexity, we need more challenging ways to evaluate algorithms like Deep Networks.
- *Perhaps Adversarial Examiners which probe for weaknesses of the algorithms by having a dynamic testing set, instead of standard fixed evaluation datasets.*
- ***This is not merely an academic problem.***

(1) Flawed performance measures can lead to unrealistic expectations.

(2) Results on biased datasets can have bad societal consequences.

- *AL Yuille & Chenxi Liu. Deep nets: What have they ever done for vision? IJCV. 2020/2021? https://arxiv.org/abs/1805.04025*

Two Robustness Challenges

- Classifying objects which are heavily occluded.
- Classifying objects despite patch attacks.
- *Both are situations where the statistics of the test set differs from the statistics of the training set.*
- Will standard Deep Networks succeed at these challenges?
- Can we develop alternative algorithms, perhaps variants of Deep Networks, that are more robust to these challenges?
- *Strategy: bring back generative models. But make them generative on deep network features (easier to design that generative models on image intensities, and can ignore unimportant details).*

Compositional Networks: A. Kortylewski et al. CVPR 2020.

- In natural images objects are surrounded and partially occluded by other objects
- Occluders are highly variable in terms of shape and texture -> **exponential complexity**
- Vision systems must generalize to exponentially complex domains

Deep Nets are not robust to severe occlusion

- DCNNs do not generalize when trained with non-occluded data



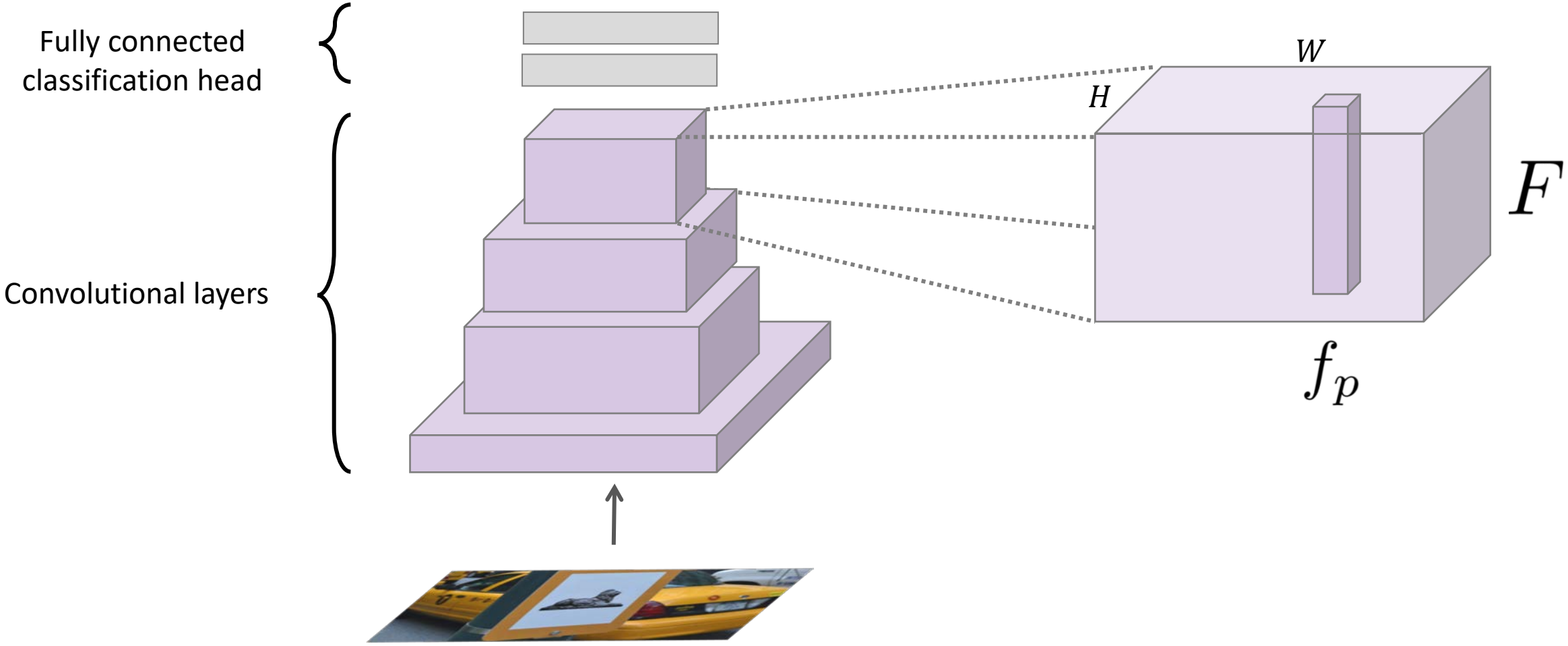

Occ. Area	0%	30%	50%	70%	Avg
VGG -16	99.1	88.7	78.8	63.0	82.4

- What if we train with lots of augmented data?

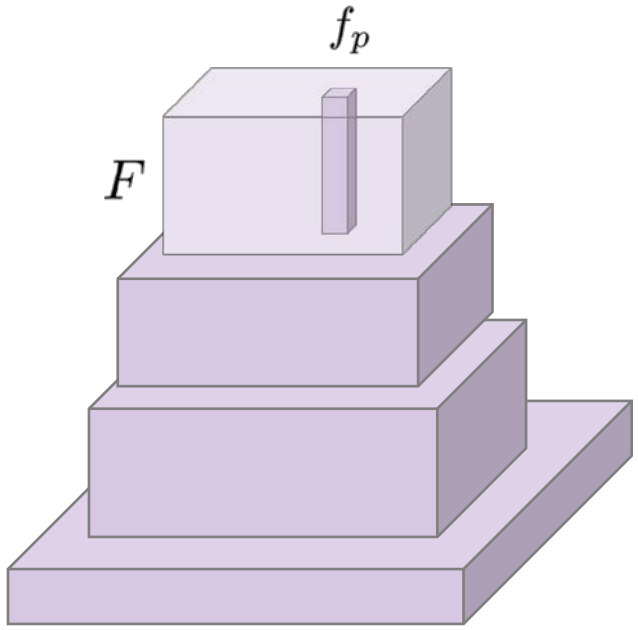


Occ. Area	0%	30%	50%	70%	Avg
VGG-16-Augmented	99.3	92.3	89.9	80.8	90.6

A Generative Model of Neural Feature Activations



A Generative Model of Neural Feature Activations

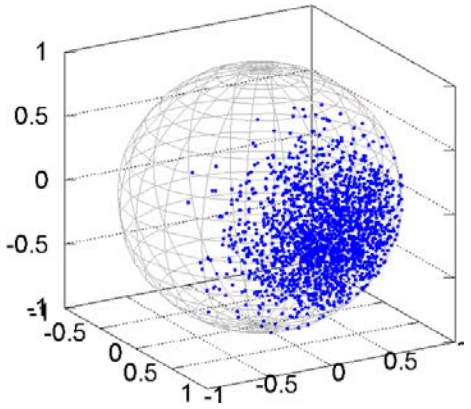


$$p(F|\Theta_y) = \sum_m \nu^m p(F|\theta_y^m)$$

$$p(F|\theta_y^m) = \prod_p p(f_p|\mathcal{A}_{p,y}^m, \Lambda)$$

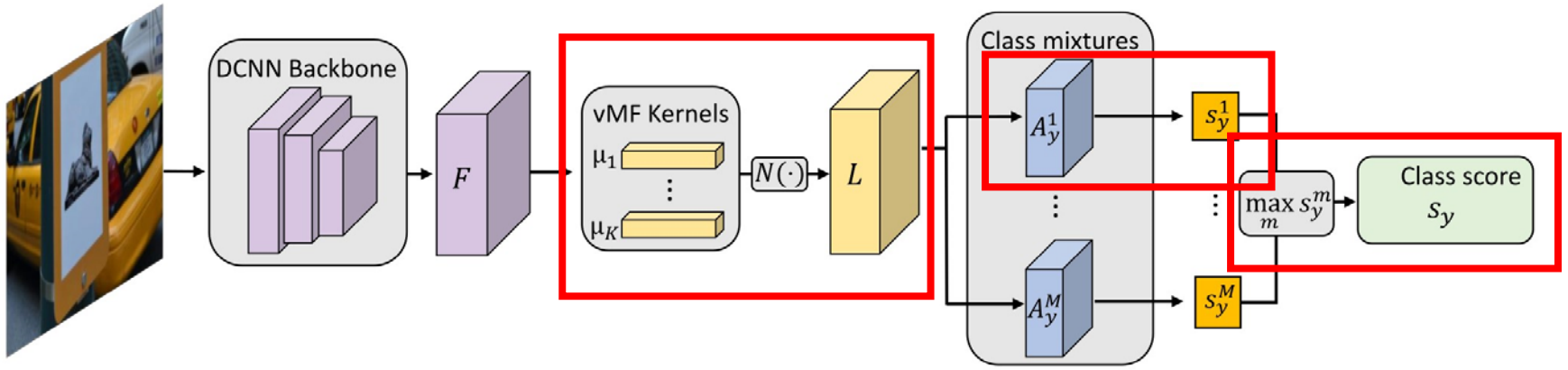
$$p(f_p|\mathcal{A}_{p,y}^m, \Lambda) = \sum_k \alpha_{p,k,y}^m p(f_p|\lambda_k), \quad \lambda_k = \{\mu_k, \sigma_k\}$$

$$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1$$



Y labels object class
P labels position in the image
fp are the feature vectors at p
m label the mixture (viewpoint)
alpha's,lambda's,mu's are parameters which are learnt.

Inference as a Feed-Forward Neural Network

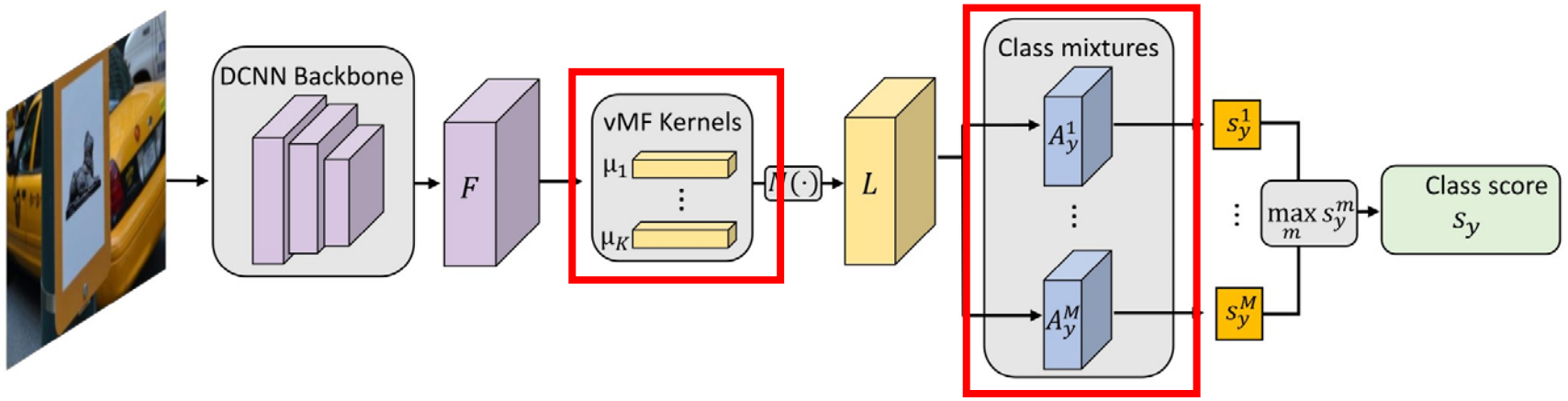


1. vMF likelihood: $$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1$$

2. Mixture likelihoods: $$p(F|\theta_y^m) = \prod_p \sum_k \alpha_{p,k,y}^m p(f_p|\lambda_k)$$

3. Class score: $$p(F|\Theta_y) = \sum_m \nu^m p(F|\theta_y^m), \quad \nu^m \in \{0,1\}, \sum_m \nu^m = 1$$

Learning the Model Parameters: Backpropagation & Clustering

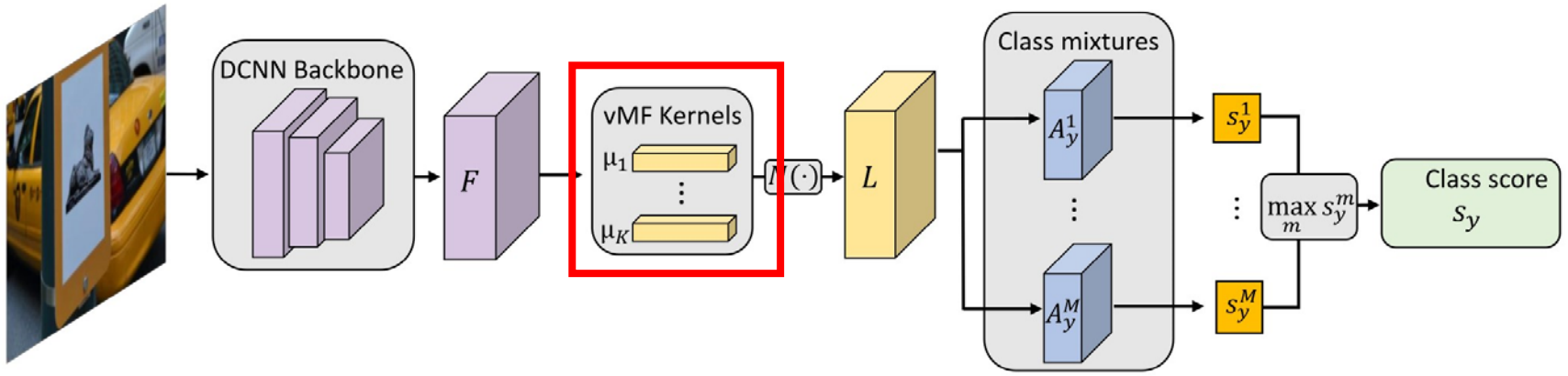


$$\mathcal{L} = \mathcal{L}_{class}(y, y') + \gamma_1 \mathcal{L}_{weight}(W) + \gamma_2 \mathcal{L}_{vmf}(F, \Lambda) + \gamma_3 \mathcal{L}_{mix}(F, \mathcal{A}_y)$$

$$\mathcal{L}_{vmf}(F, \Lambda) = -\sum_p \max_k \log p(f_p|\mu_k) = C \sum_p \min_k \mu_k^T f_p$$

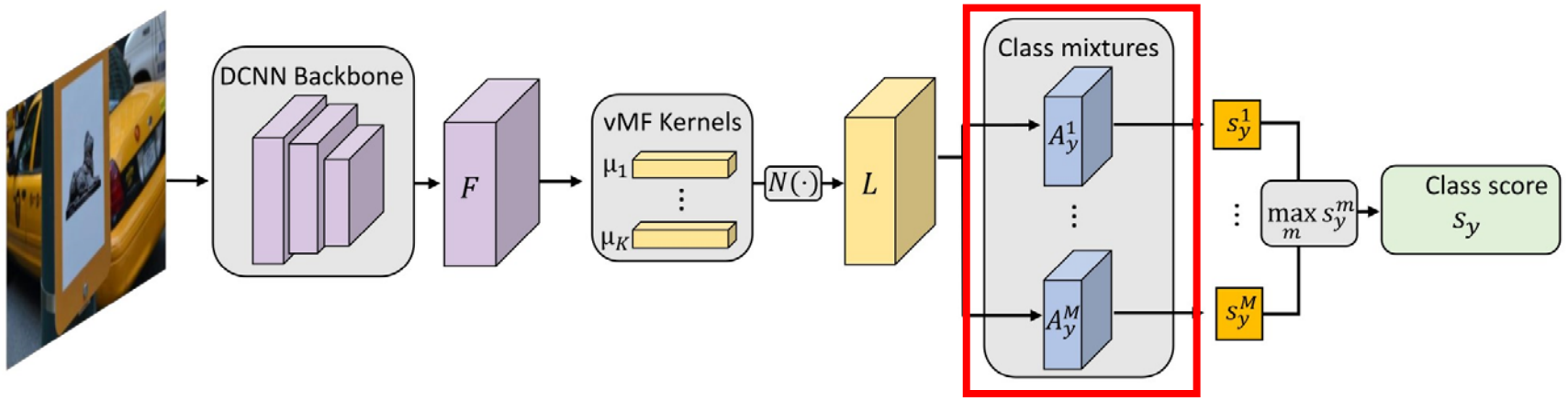
$$\mathcal{L}_{mix}(F, \mathcal{A}_y) = -\sum_p \log \left[\sum_k \alpha_{p,k,y}^m p(f_p|\mu_k) \right]$$

Explainability - vMF Kernels resemble „part detectors“



- Image patterns with highest likelihood:

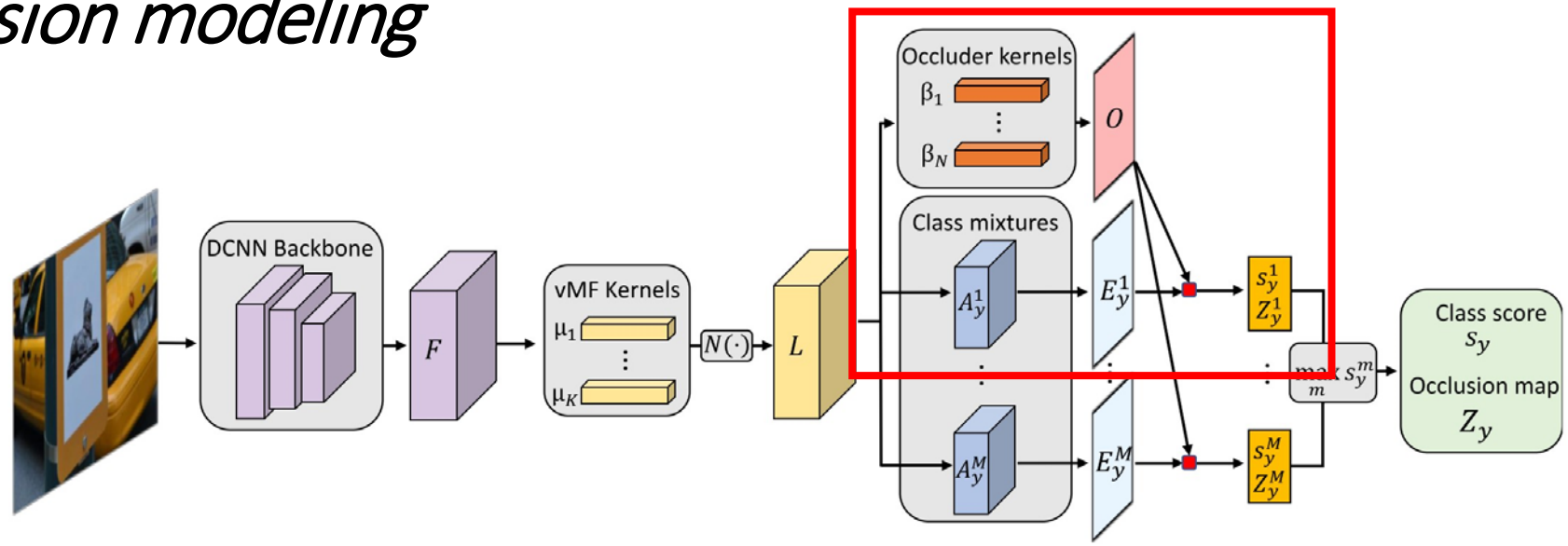
Explainability – Mixture components model object pose



- Images with highest likelihood for mixture components:




Occlusion modeling



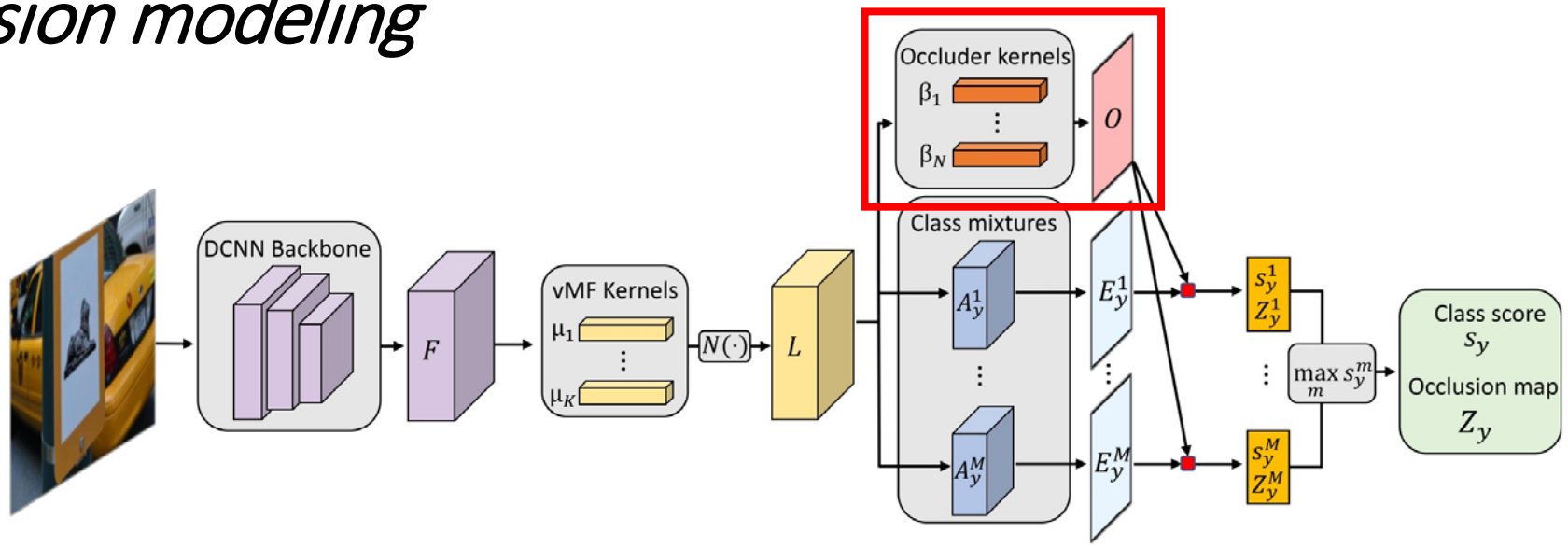
- *We introduce an outlier model:*

$$p(F|\theta_y^m,\beta)=\prod_p p(f_p,z_p^m=0)^{1-z_p^m}\,p(f_p,z_p^m=1)^{z_p^m},\quad \mathcal{Z}^m=\{z_p^m\in\{0,1\}|p\in\mathcal{P}\}$$

$$p(f_p,z_p^m=1)=p(f_p|\beta,\Lambda)\;p(z_p^m=1),$$

$$p(f_p,z_p^m=0)=p(f_p|\mathcal{A}_{p,y}^m,\Lambda)\;(1-p(z_p^m=1)).$$

Occlusion modeling

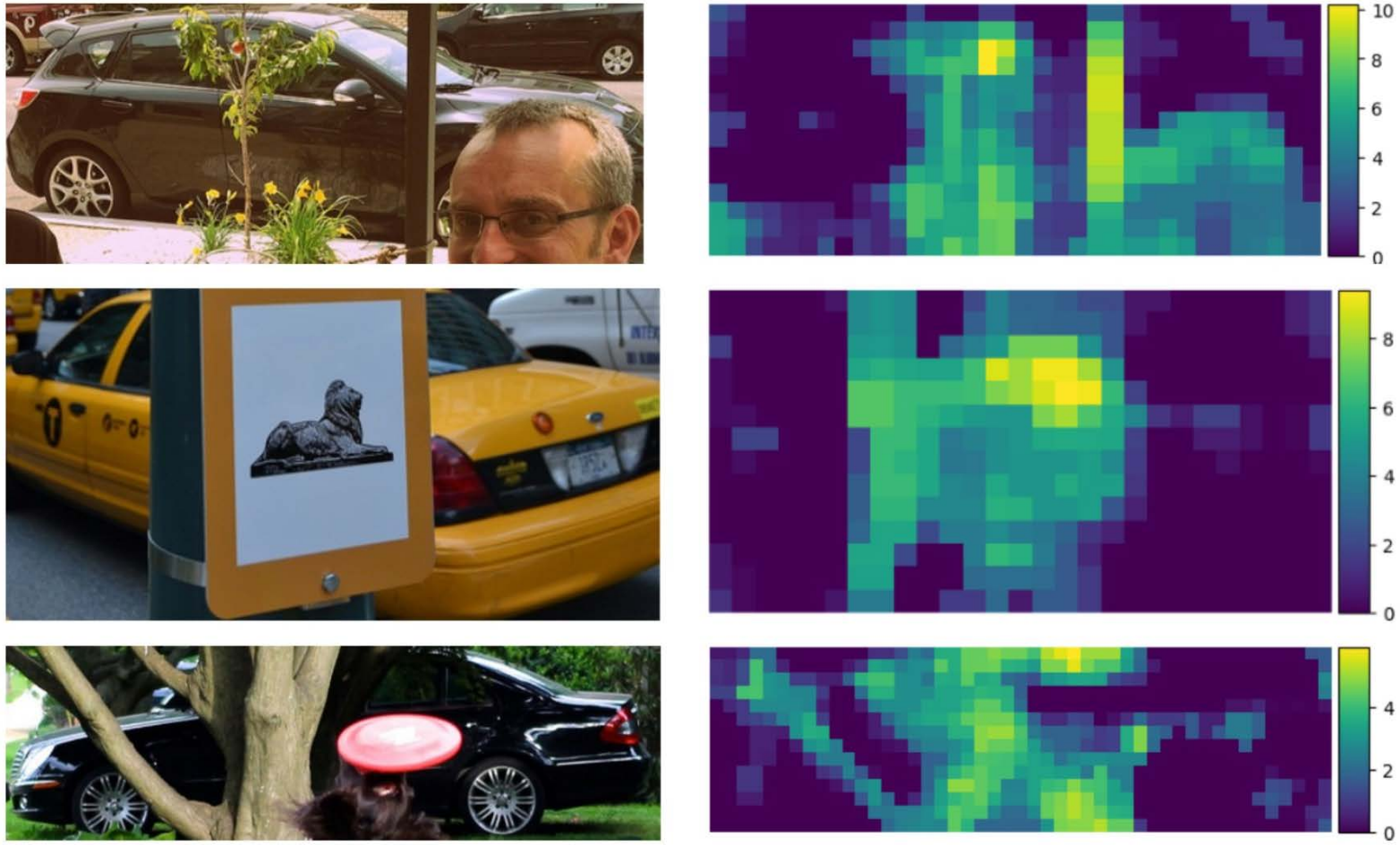


- *We introduce an outlier model:*

$$p(F|\theta_y^m, \beta) = \prod_p p(f_p, z_p^m = 0)^{1-z_p^m} p(f_p, z_p^m = 1)^{z_p^m}, \quad \mathcal{Z}^m = \{z_p^m \in \{0,1\} | p \in \mathcal{P}\}$$

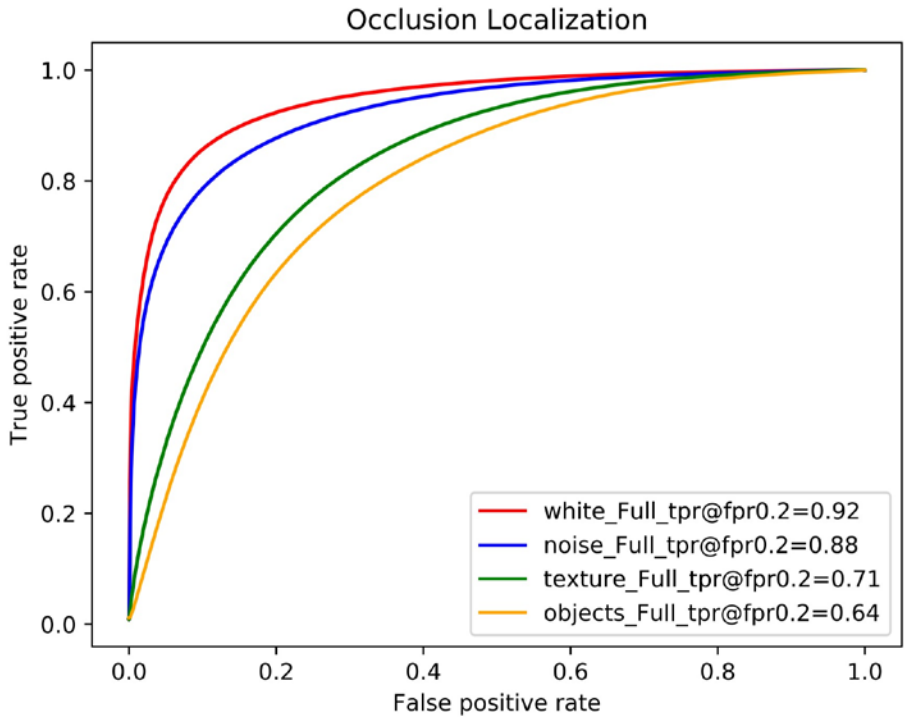
- A simple model of *what the object does not look like:*

Competition between object and outlier model



Quantitative Evaluation of Occluder Localization



CompNets can classify partially occluded vehicles robustly




Occ. Area	**L0**	**L1**	**L2**	**L3**	Avg
VGG	97.8	86.8	79.1	60.3	81.0
ResNet50	98.5	89.6	84.9	71.2	86.1
ResNext	98.7	90.7	85.9	75.3	87.7

Compositional Nets can also be applied to Object Detection

- This requires extending the occlusion model to deal with background context.

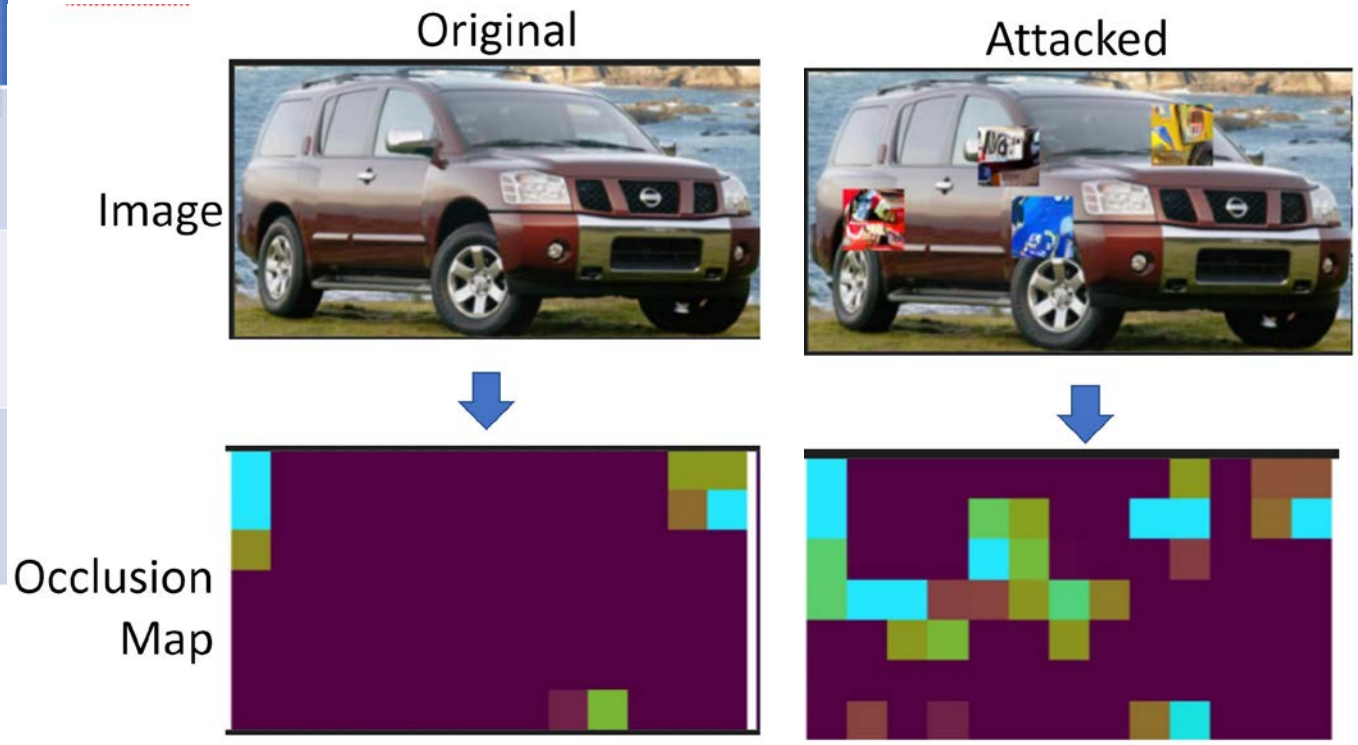
- A. Wang et al. CVPR. 2020.

CompNets vrs. Patch Attacks and Sparse-RS

- CompNets ***are*** robust against targeted patch attacks. (C. Cosgrove et al. 2020).
-

Model	Accuracy (%)	Attack success rate (%)	
		PatchAttack[1] (TPA) 4 patches	Sparse-RS[2] 1 patch
CompNet (vgg16 backbone)	98.5	12.6	0.9
vgg16	98.6	98.8	92.3

CompNets ***can*** detect patch attacks.

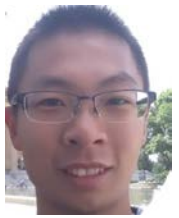


Learning to Parse Animals with Weak Prior Knowledge: "You Only Annotate Once".

- Motivated by infants playing 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?

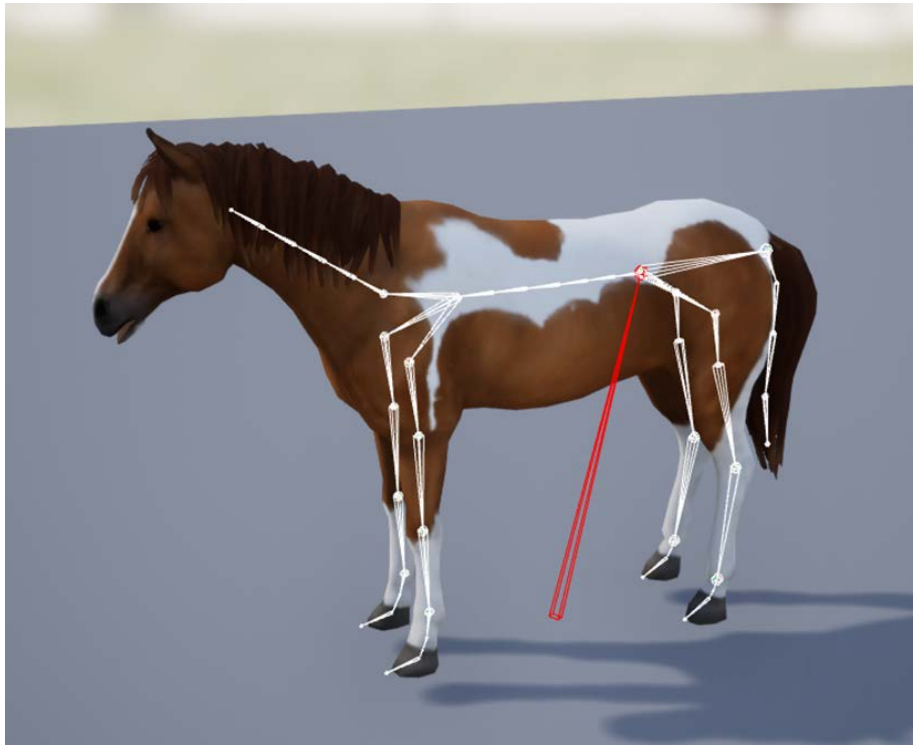
You Only Annotate Once

- 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.



JOHNS HOPKINS UNIVERSITY

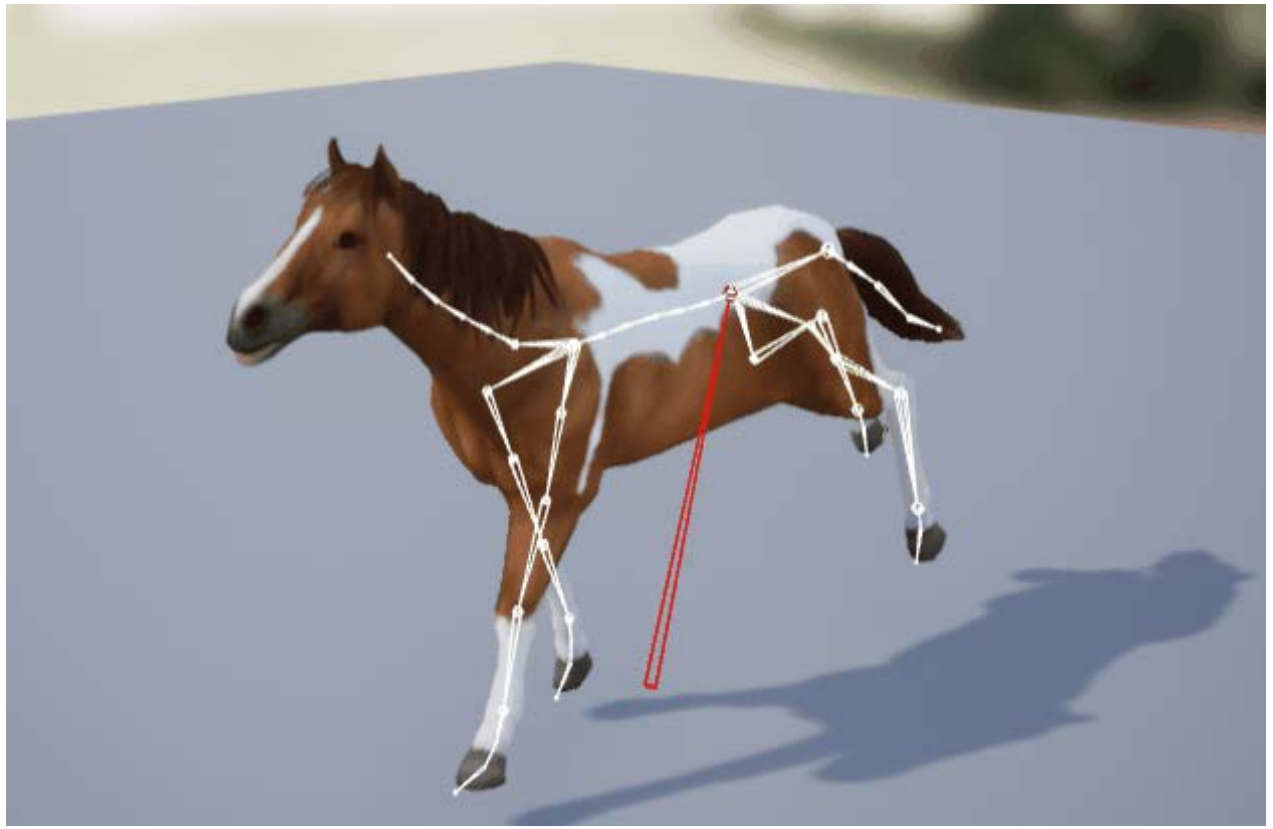
Train with Synthetic Data (one horse, one tiger)

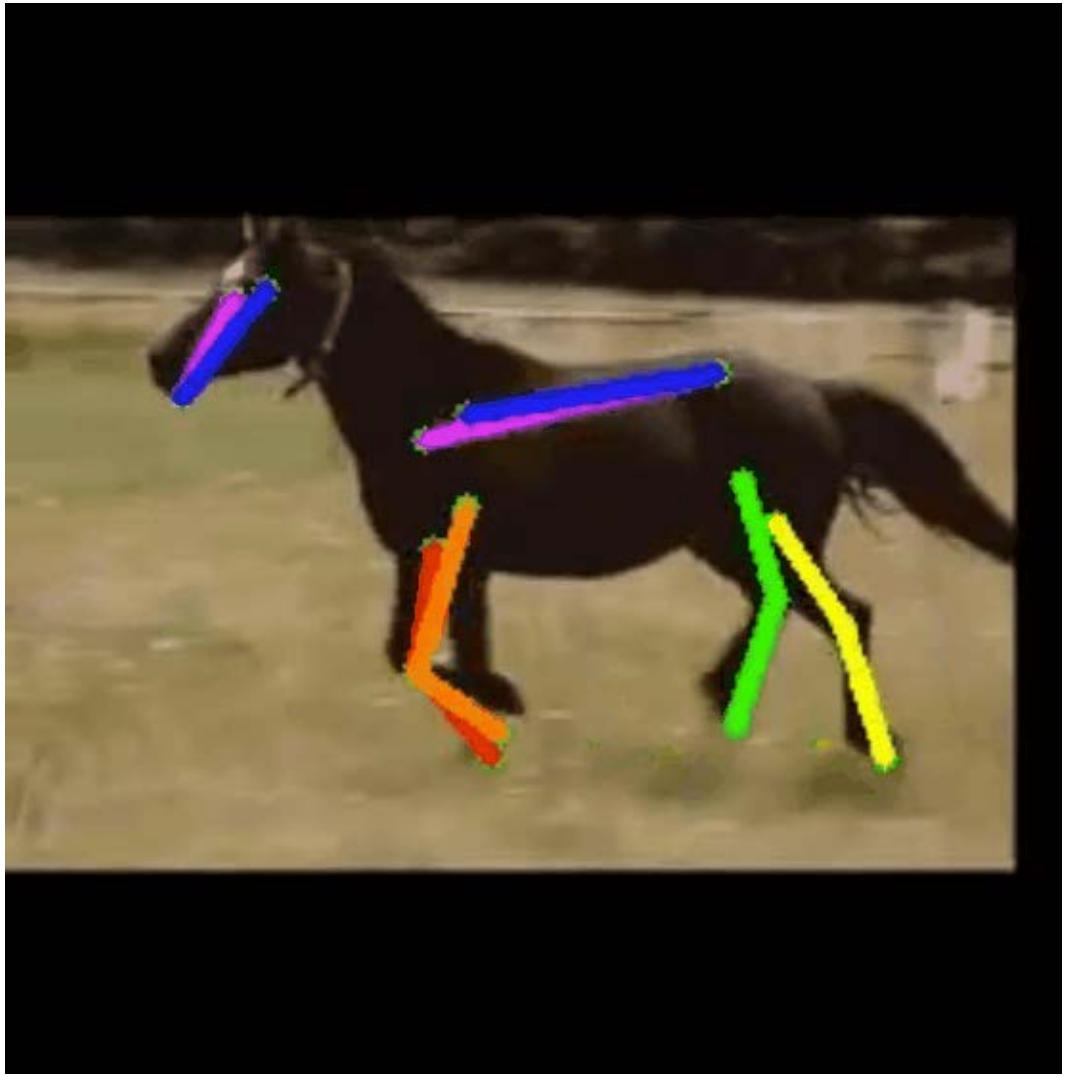
- Training with Real Data with Annotations (78.98)
- Better Data
  - More realistic model, realistic background (intuitive, but not work)
  - Texture Randomization (25.33)
  - Data Augmentation, rotation, gaussian noise (60.84)
- Better Training
  - Domain adaptation
    - *synthetic +unlabeled real 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.*

JOHNS HOPKINS UNIVERSITY

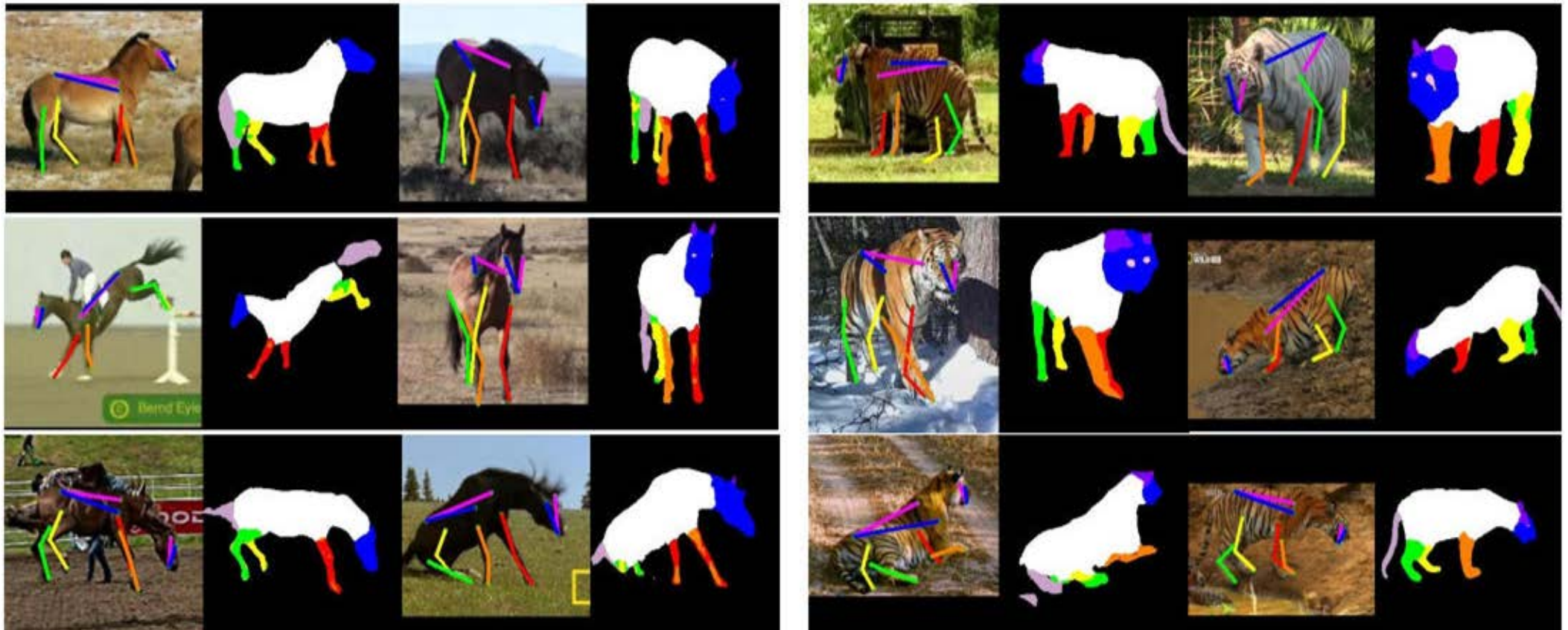
An animal keypoint video





Scale Up –extend to new tasks.

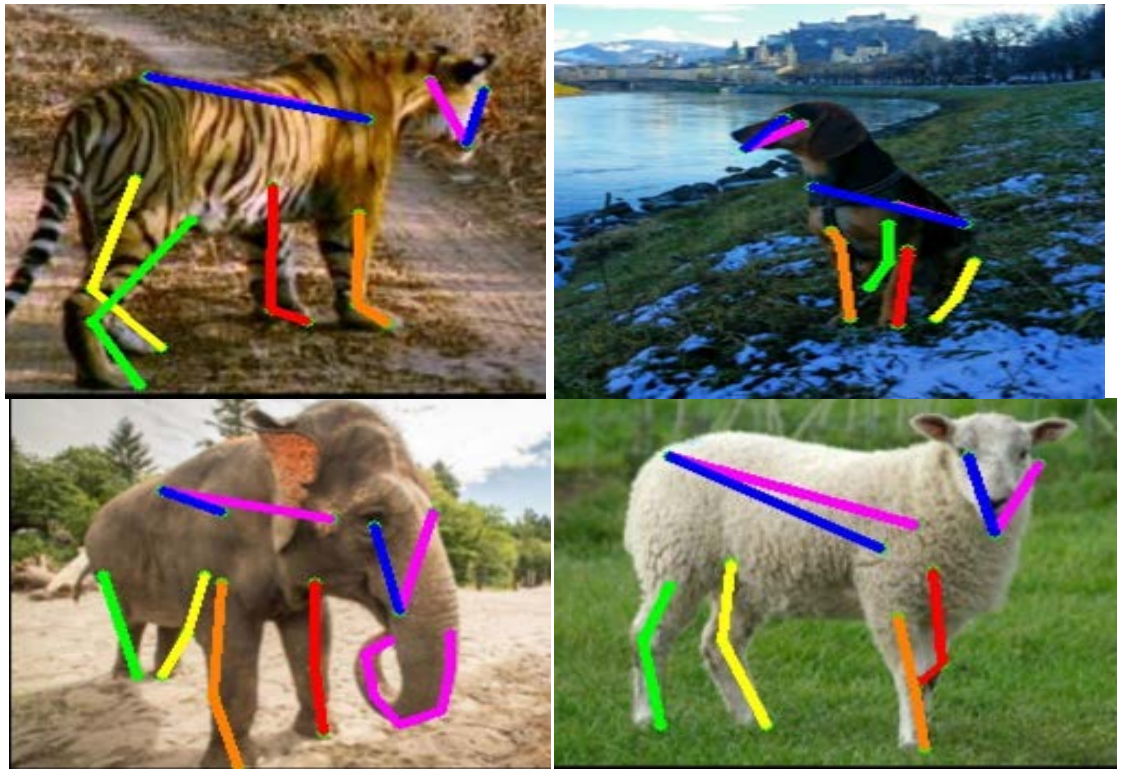
New Visual Task: Part Segmentation: Identify head, torso, legs, tails.
Same diversity plus learning strategy.



JOHNS HOPKINS UNIVERSITY

Scale Up -- extend to more categories

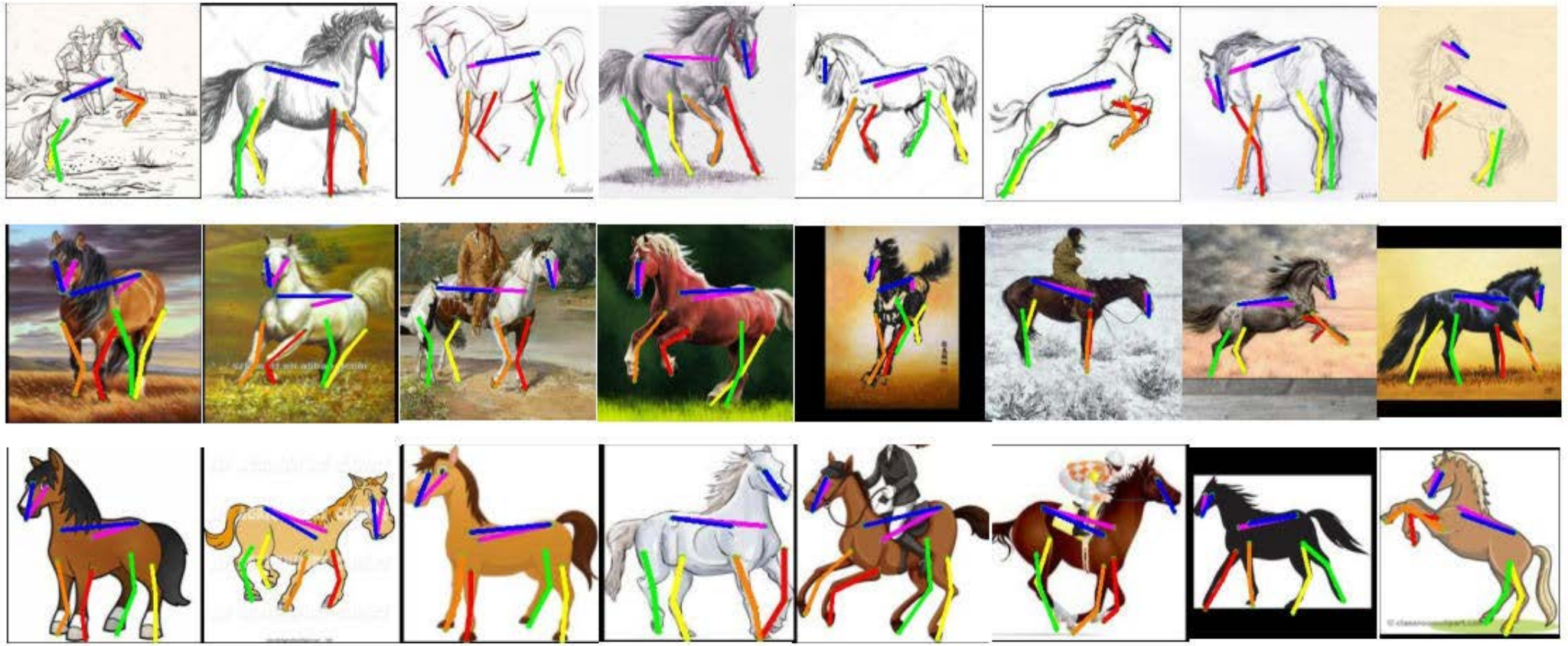
You only annotate once (for each object category) but same diversity and learning strategies still apply.



JOHNS HOPKINS UNIVERSITY

Scale Up: extend to domain generalization

Better Domain Generalization



Summary: Part 1.

Vision is an extremely interesting, challenging, and rewarding research field.

It involves an enormous range of techniques – summer school.

Deep Networks can give wonderful performance – and are extremely innovative and constantly evolving.

But remember they are only one of many techniques. Do not forget geometry, probabilistic generative models, and many more.

Finite-sized balanced annotated datasets are problematic in face of the combinatorial complexity of the real world. We need tougher challenges and performance measures, like adversarial examiners.

Summary Part 2:

- Two Challenges: Generative models can be more robust to occluders and patch attacks, because they can exploit outlier processes.
- Computer Graphics is very useful. Perhaps it is a surrogate for how human infants learn.

- Further Study: The IPAM 2013 summer school has 70 one hour lectures: http://www.ipam.ucla.edu/programs/summer-schools/graduate-summer-school-computer-vision/?tab=schedule
- Opinion paper: AL Yuille & Chenxi Liu. Deep nets: What have they ever done for vision? IJCV. 2020/2021? https://arxiv.org/abs/1805.04025