Knowledge Transfer in Vision Tasks with Incomplete Data

Zhizhong Li

University of Illinois Urbana Champaign

Chair: Derek Hoiem

Committee: Derek Hoiem, Svetlana Lazebnik, Alexander Schwing, Linjie Luo

Knowledge in humans

- It comes quite naturally
  - Classify
  - Attribute
  - Applying to new circumstances
  - Inference between attributes
  - Know if answer is uncertain
  - Understand its behavior
  - Know how to interact with it
  - More…

Knowledge in ML models

- Harder for ML algorithms
  - Classify
  - Attribute
  - Applying to new circumstances
  - Inference between attributes
  - Know if answer is uncertain
  - Understand its behavior
  - Know how to interact with it
  - More…

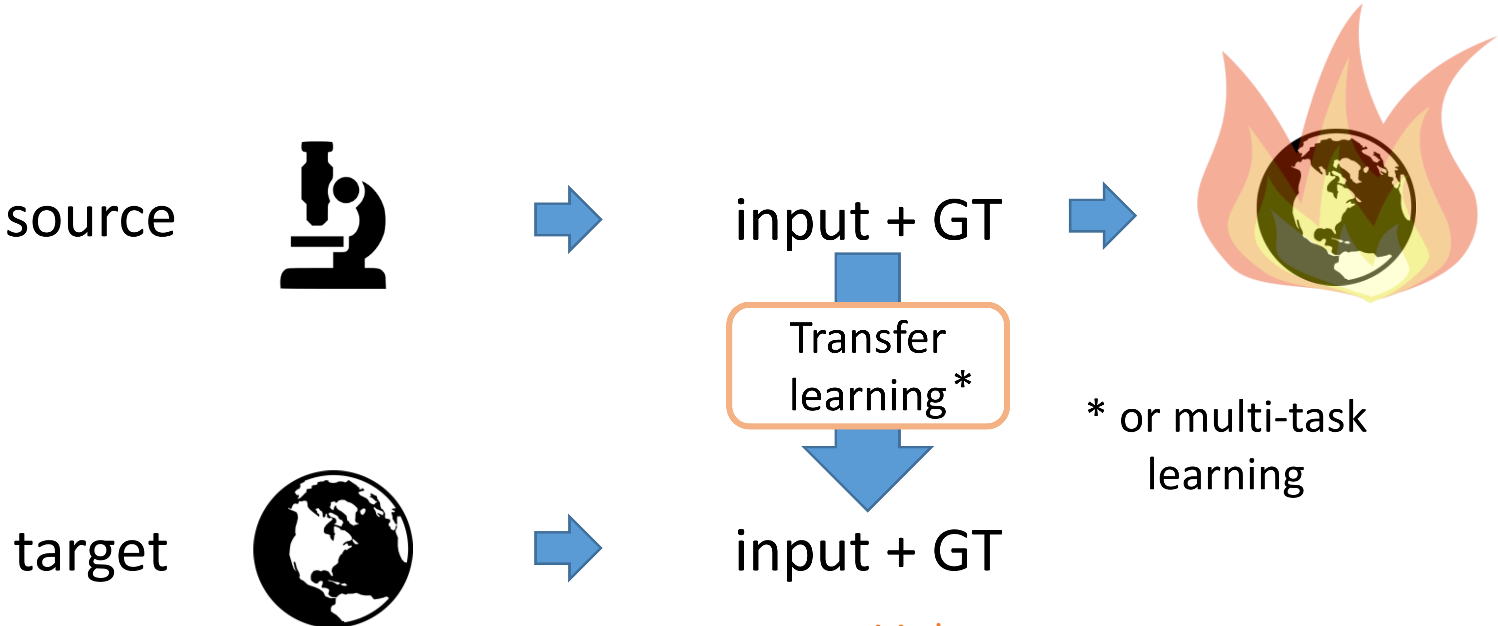
Easier for
models to learn

Knowledge transfer!

Trickier to learn
Help generalization



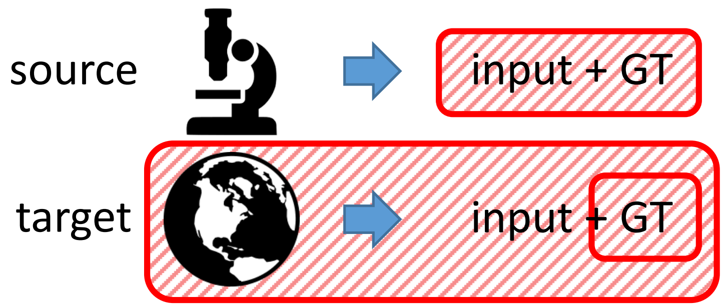

Transfer learning



non-i.i.d.
(different task or different distribution)

In practice:
what if we cannot obtain all aspects of necessary data?

Transfer learning in practice



① old data missing

Learning without Forgetting

② label missing

Task-assisted Domain Adaptation

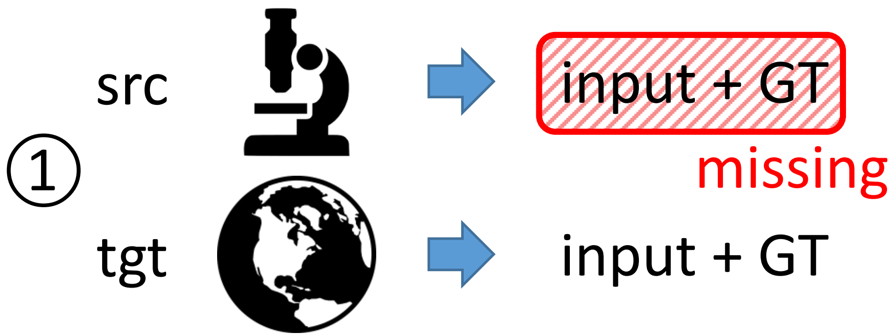
③ domain unknown

Improving Confidence Estimates
for Unfamiliar Examples

Learning without Forgetting

Zhizhong Li, Derek Hoiem

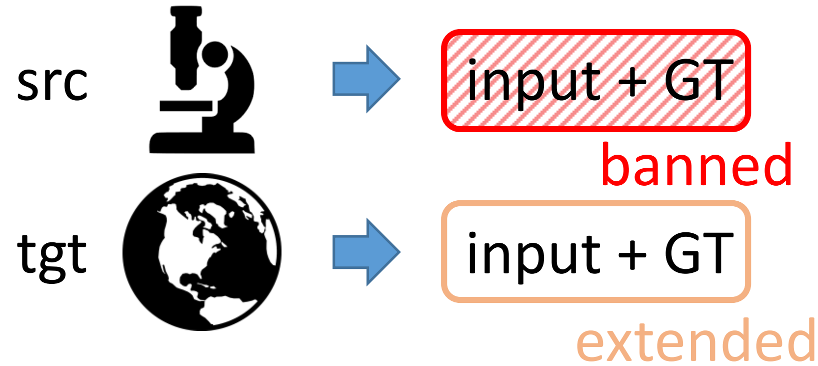
In ECCV 2016 (spotlight);
PAMI, 2018



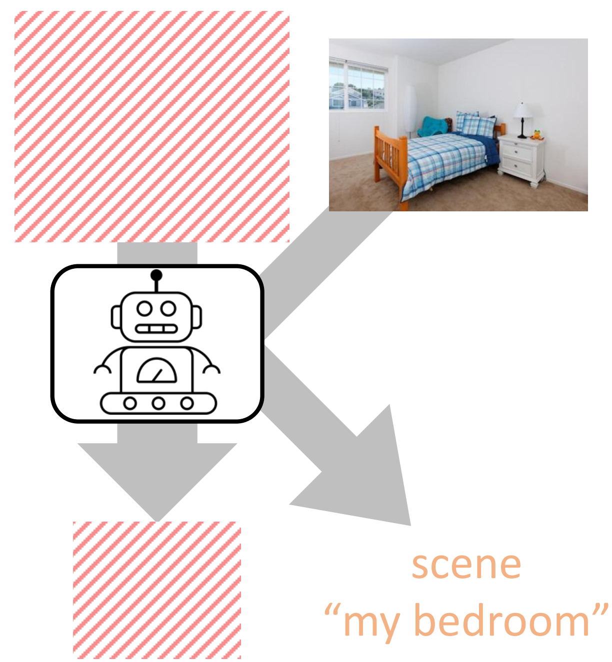
Motivation

- Task:
  extending capability
  (transfer to new task)

  * closer to multi-task learning

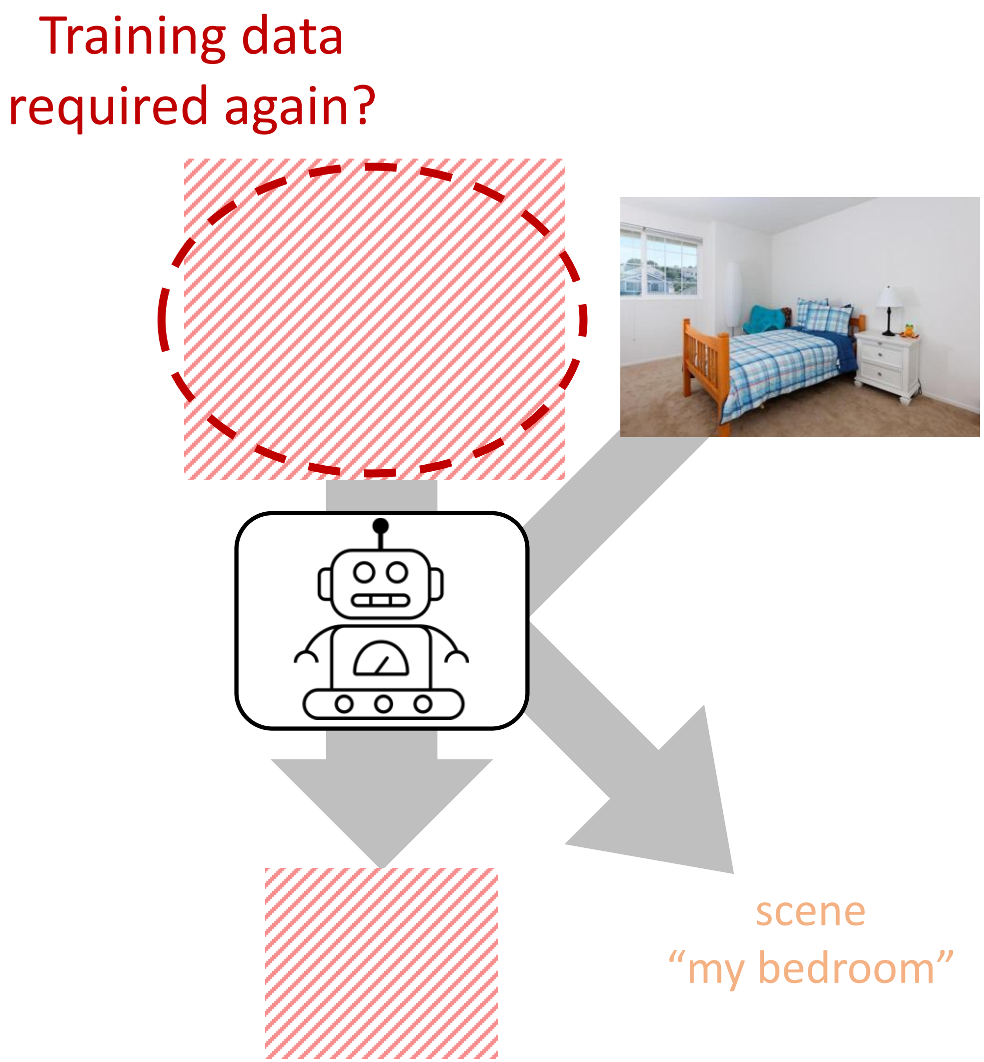


- Constraint:
  - Cannot access original dataset
  - Common in industry settings



- Challenge:
  - Catastrophic forgetting
  - ... but maintain old task performance

Baselines



- Fine-tuning?
- Feature extraction?
- Joint training?

Related work

- Fine-tuning, feature extracting, Multi-task learning

- Closely related:
  - Less Forgetting Learning [1]
  - A-LTM [2]

Compared in paper

- Other continual learning methods:
  - iCaRL [3]
  - EWC [4], SI [5]

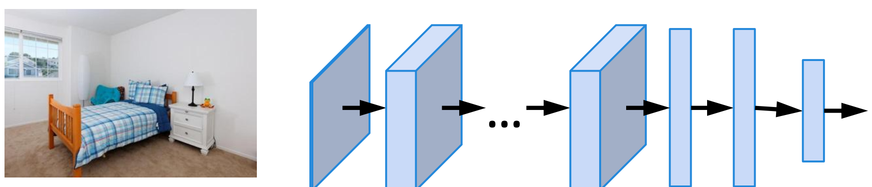
require past data

[1] Heechul Jung et al. “Less-forgetting Learning in Deep Neural Networks”
[2] T. Furlanello, J. Zhao, A. M. Saxe, L. Itti, and B. S. Tjan, “Activelong term memory networks”
[3] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning."
[4] Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks.“
[5] Zenke, Friedemann, Ben Poole, and Surya Ganguli. "Continual learning through synaptic intelligence."

Method

1. Obtain old task responses

“90% bed,
10% lamp” etc.



[new image]

[old task 1 response Y_o]

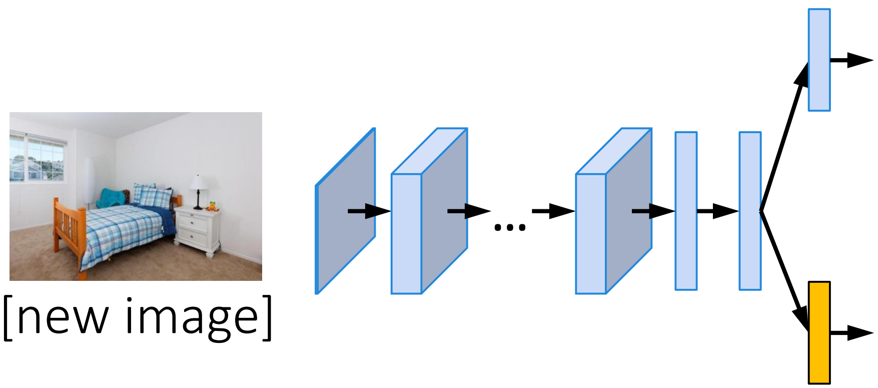
Serve as reminder
of old task

Method

2. Train on new images

Target: “90% bed, 10% lamp” etc.

[old task 1 response Y_o]



[new image]

new task ground truth Y_n

“bedroom”

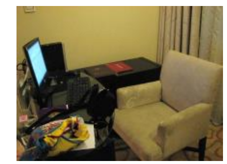
- Fine-tuning: no old task loss
- Feature extraction: freeze old layers
- Joint training (multi-task): use old task image + GT (oracle)

Experiments

- AlexNet


1 old task + 1 new task

ILSVRC 2012 Places365	+	PASCAL VOC 2012 Caltech-UCSD Birds MIT indoor scenes MNIST

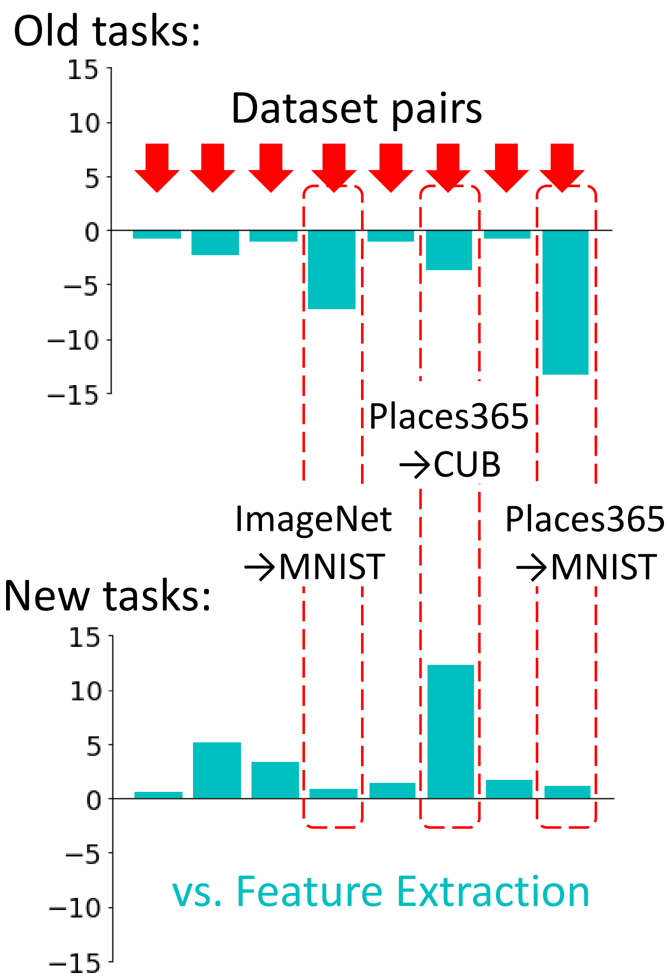



(8 combinations)

- Compared Methods:
  - Baselines
  - Less-forgetting Learning
  - Joint training (oracle)

Results: LwF vs. Feature Extraction

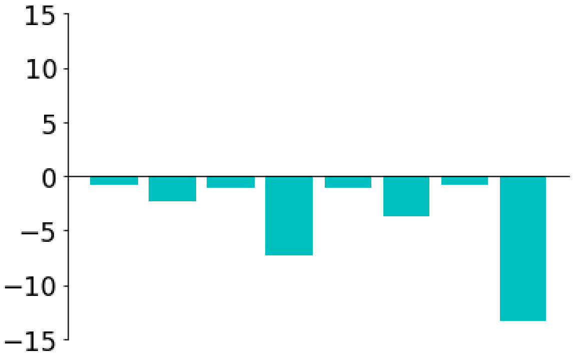
- Shown: accuracy (ours) relative to the baseline's on eight task pairs

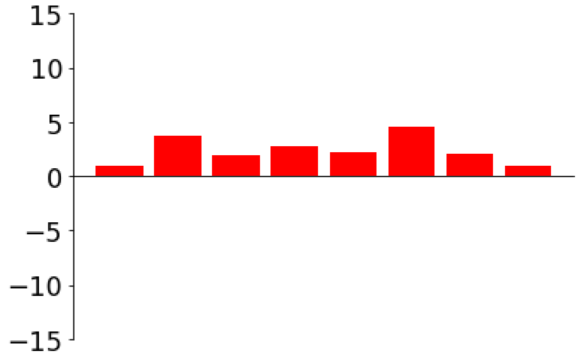


Results: LwF vs. Fine-tuning

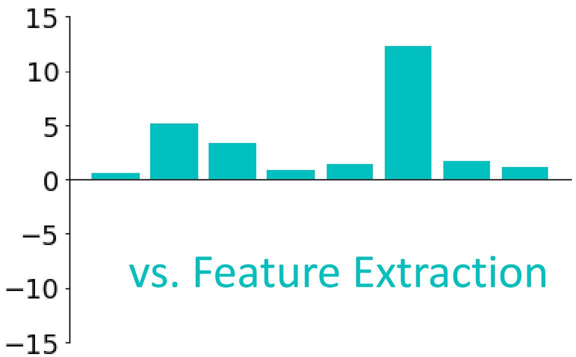
- Old task: actively preserves performance
- New task: mimics joint training

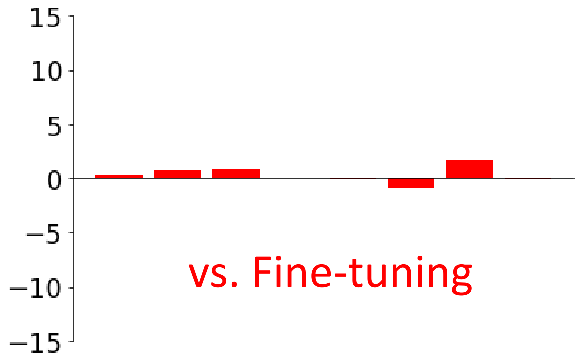
Old tasks:





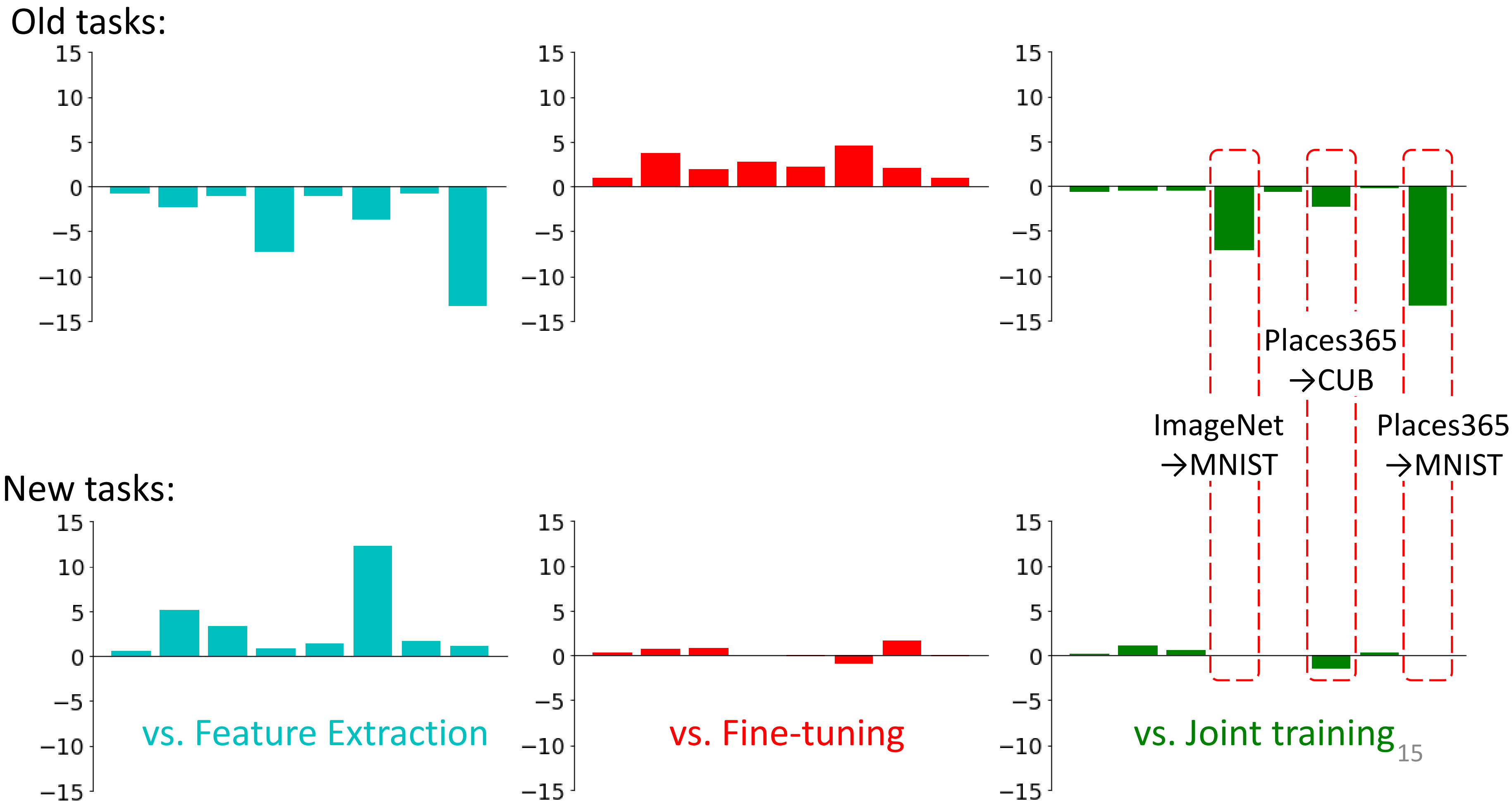
New tasks:





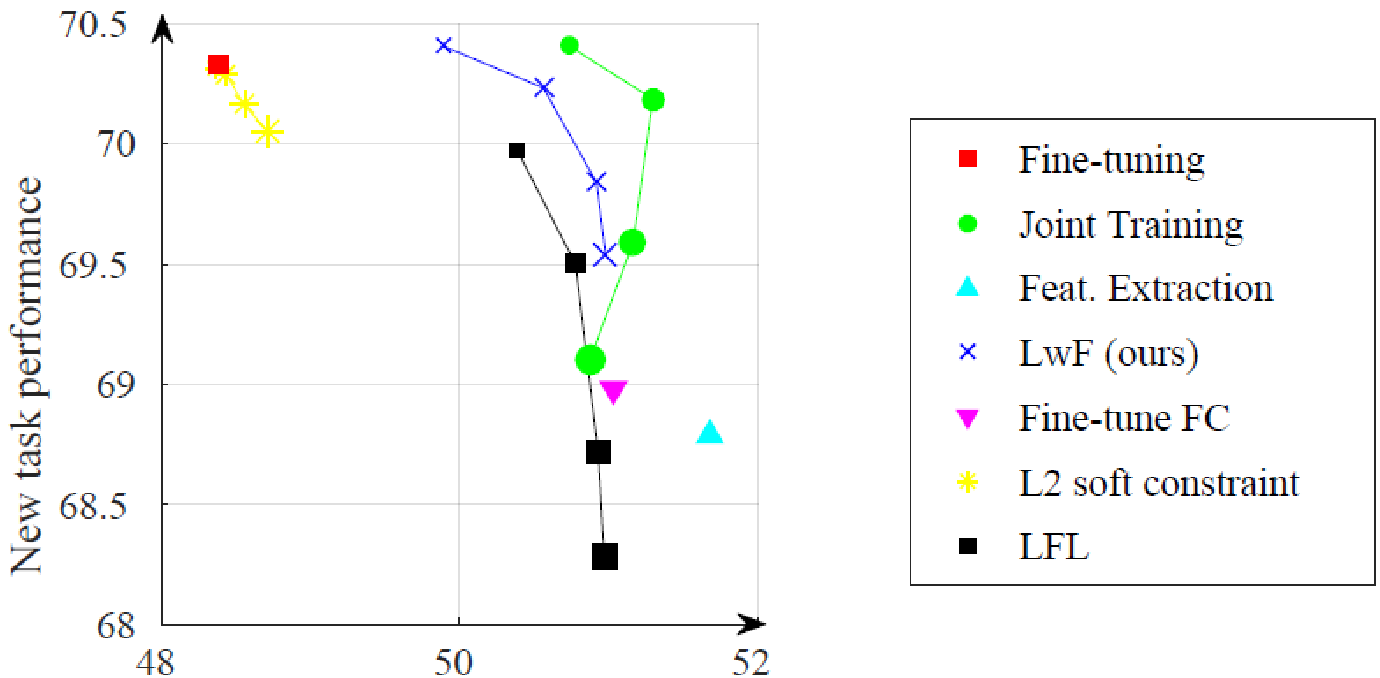
Results: LwF vs. oracle

- Joint training
- Similar performance



Results

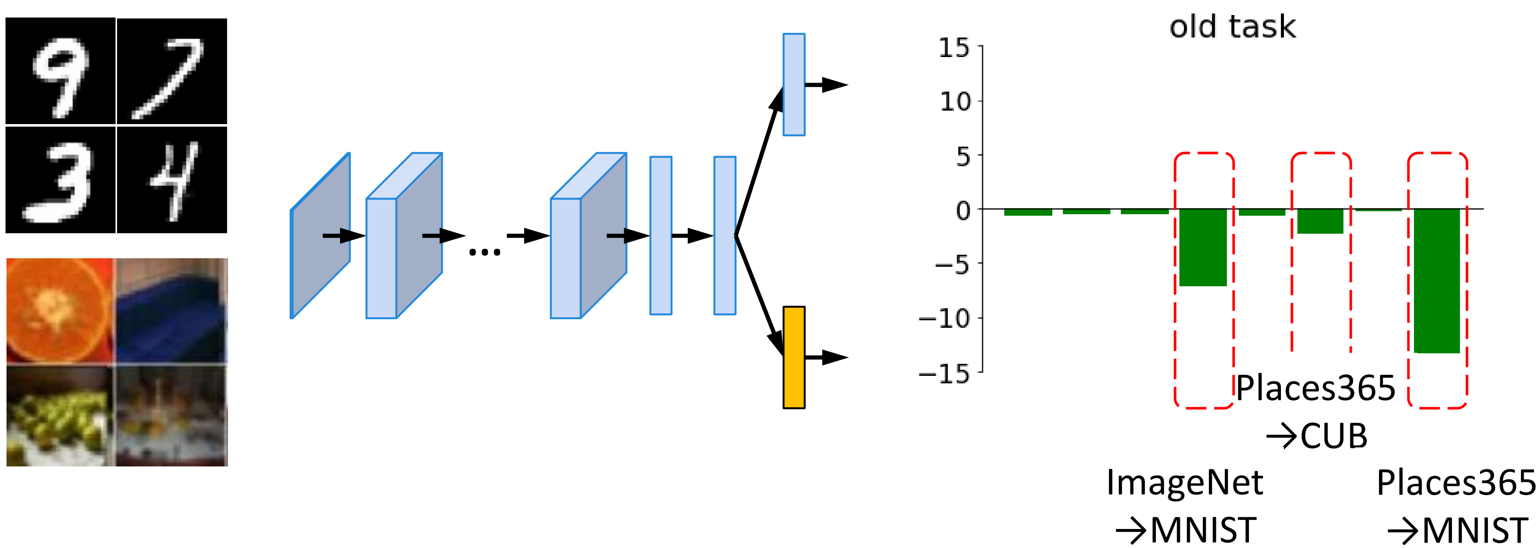
- Old-new trade-off (accuracy / VOC mAP)



(a) Places365→VOC

Limitation

- Worse when old/new images too different
  - How to add as new classes?
    - (a.k.a. class-incremental learning)



Follow-up: Dreaming to Distill

In collaboration with NVIDIA

Hongxu Yin, Pavlo Molchanov, **Zhizhong Li**, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, Jan Kautz

Accepted in CVPR 2020 as an oral presentation

A better old data proxy

- Network visualization methods
  - e.g. Deep dream, [Tensorflow lucid]

  - Generates images given only class ID or neuron ID
  - No data retention required!
- Too different from original data?

vs.

  - DeepInversion: use pretrained BatchNorm statistics

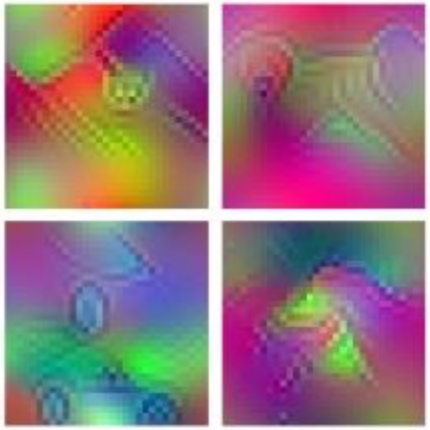
Image generation

- DeepDream $\min_{\hat{x}} \mathcal{L}(\hat{x}, y) + \mathcal{R}(\hat{x})$
- DeepInversion $\min_{\hat{x}} \mathcal{L}(\hat{x}, y) + \mathcal{R}(\hat{x}) + \boxed{\mathcal{R}_{\text{feature}}(\hat{x})}$

$$\mathcal{R}_{\text{feature}}(\hat{x}) = \sum_{l} \left|\left| \begin{array}{c}\hat{x}\text{'s} \\ \text{mean/var}\end{array} - \begin{array}{c}\text{BatchNorm} \\ \text{mean/var}\end{array} \right|\right|_2$$

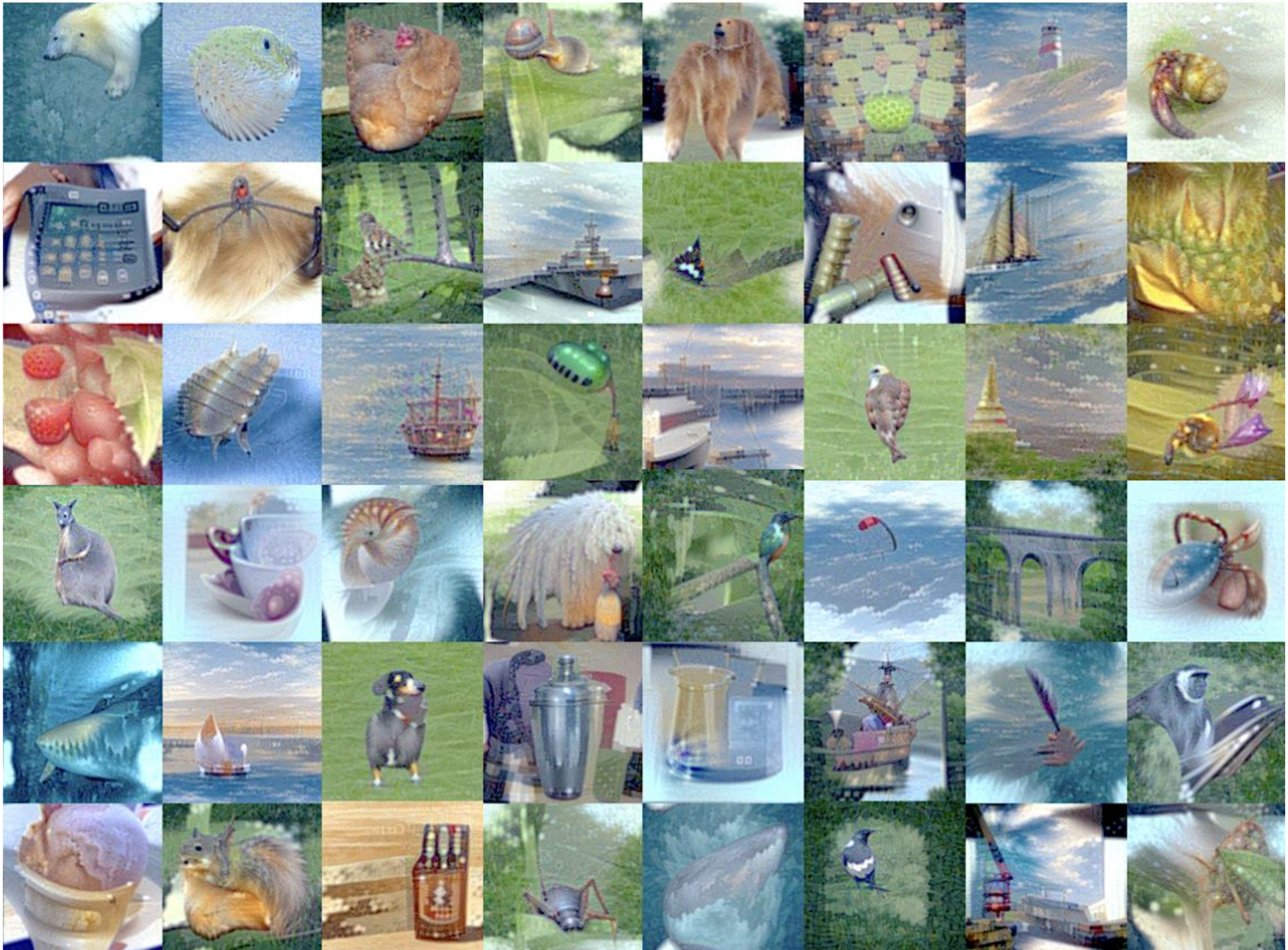
Makes feature distribution similar to training

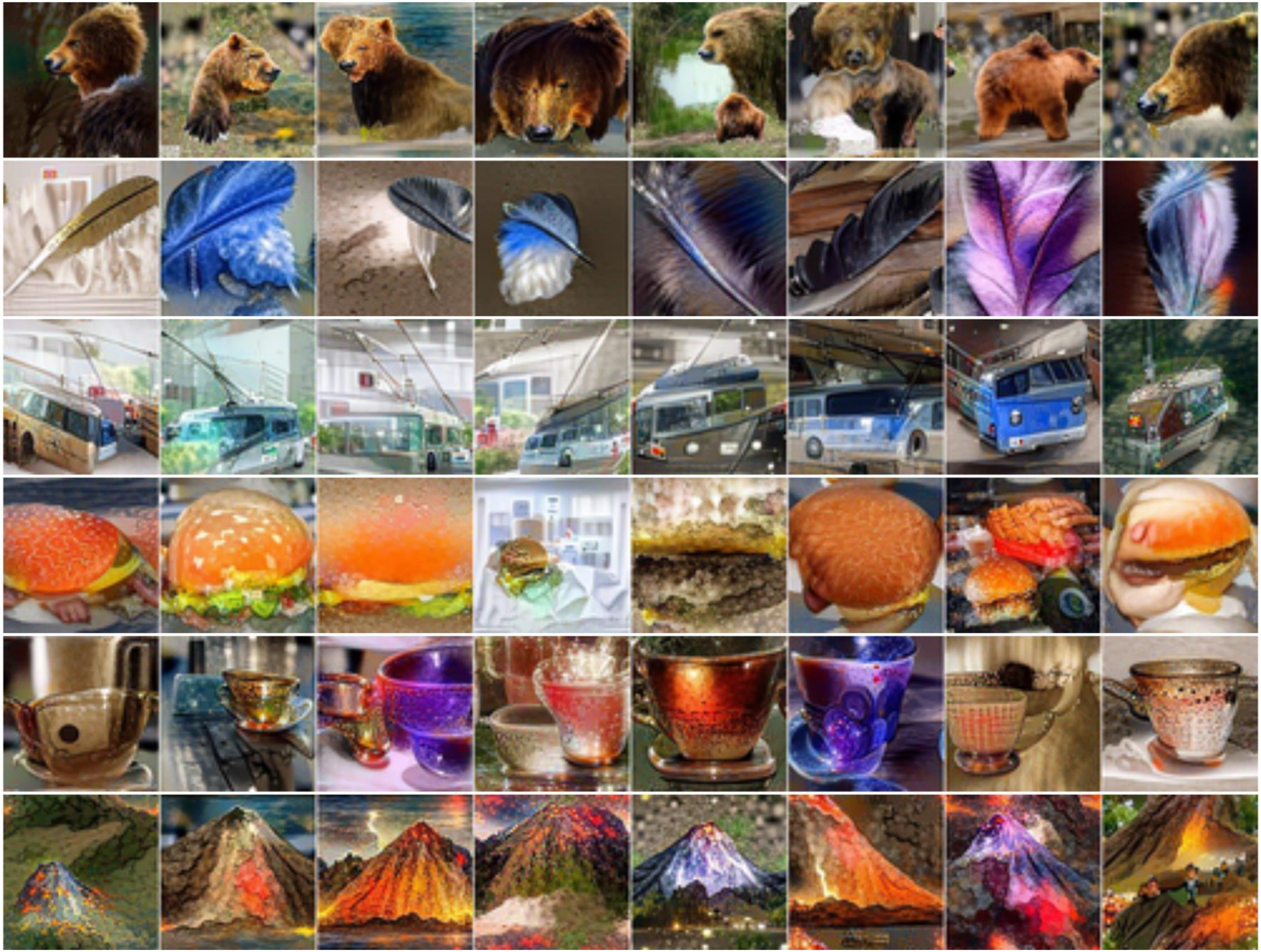
CIFAR10



DeepDream

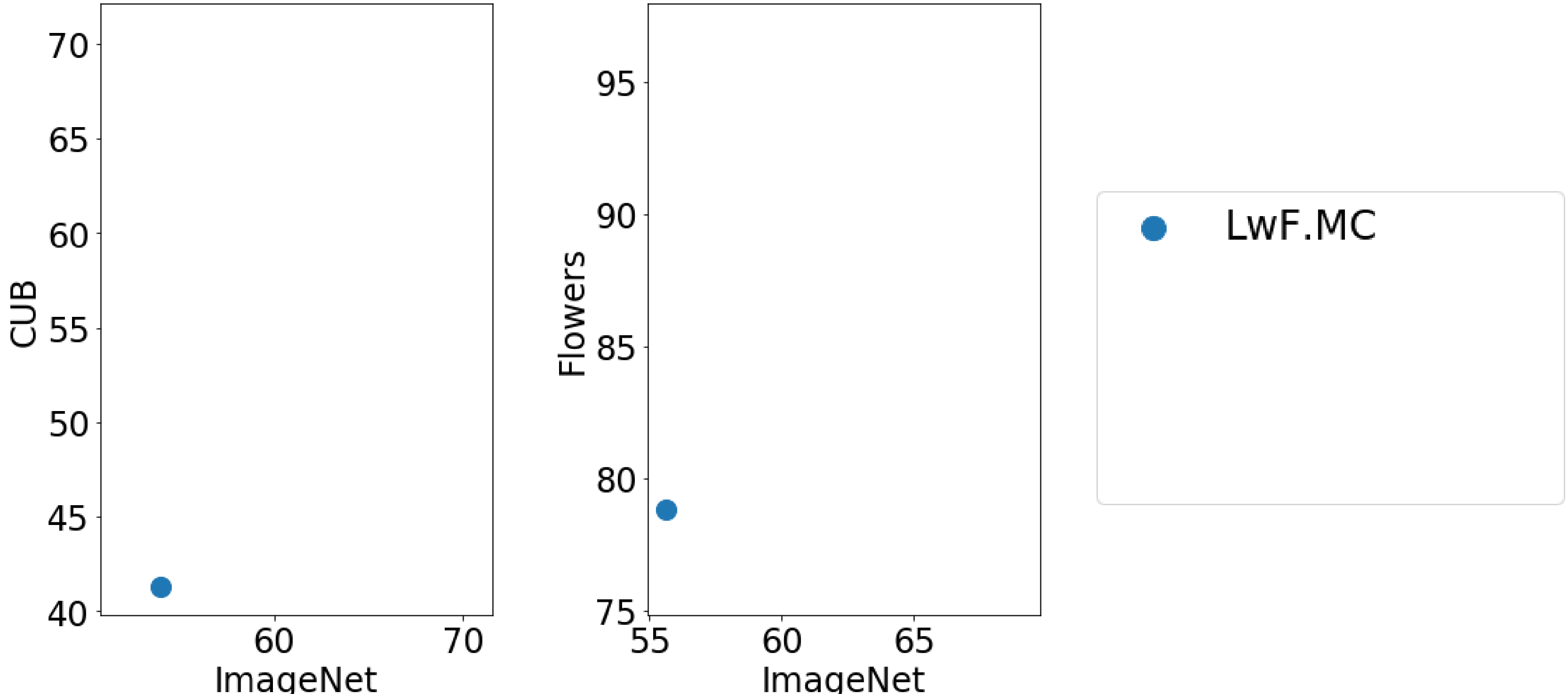
DeepInversion





Quantitative results

- ImageNet→CUB, ImageNet→Flowers
  - Allow confusion between old/new classes
    - (i.e. class-incremental instead of task-incremental)
  - Report accuracy on each dataset

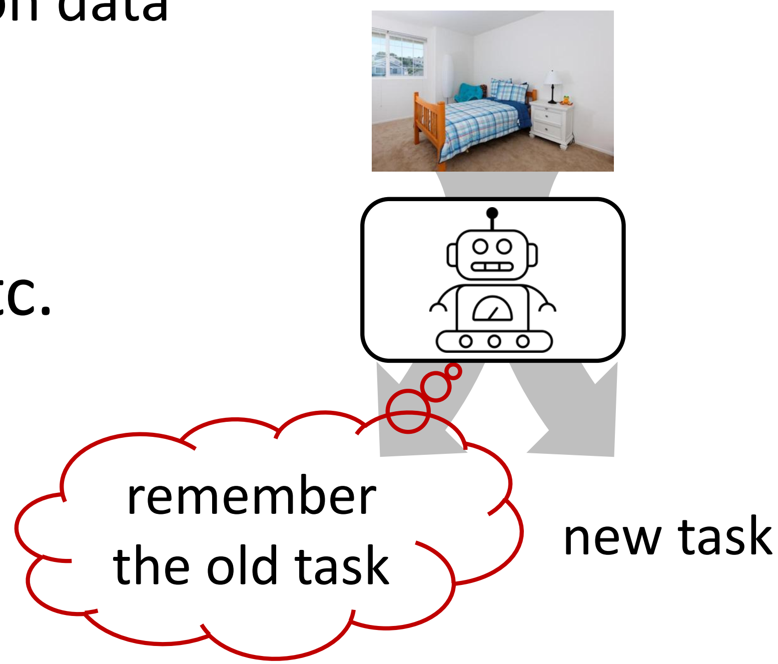


Take-away

- Data for existing knowledge can be missing

- Proxy for old task data
  - New task data / DeepInverson data
  - Original network responses

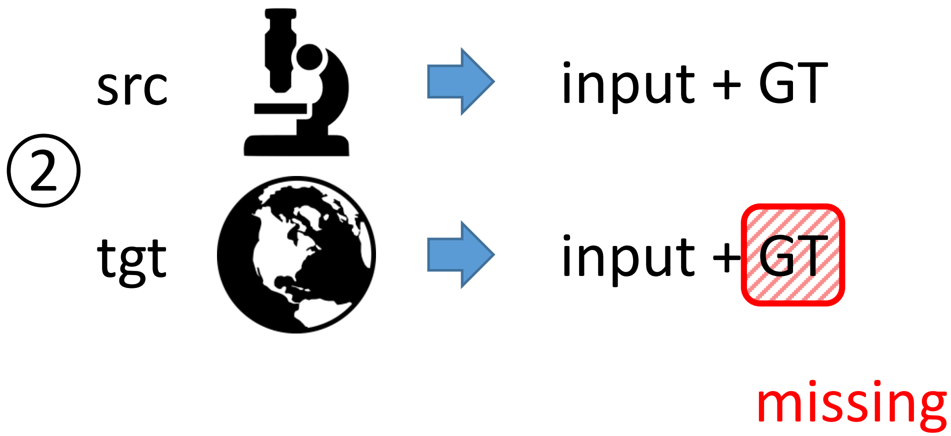
- Outperforms fine-tuning, etc.



Anchor Tasks for Domain Adaptation

Zhizhong Li, Linjie Luo, Sergey Tulyakov, Qieyun Dai, Derek Hoiem

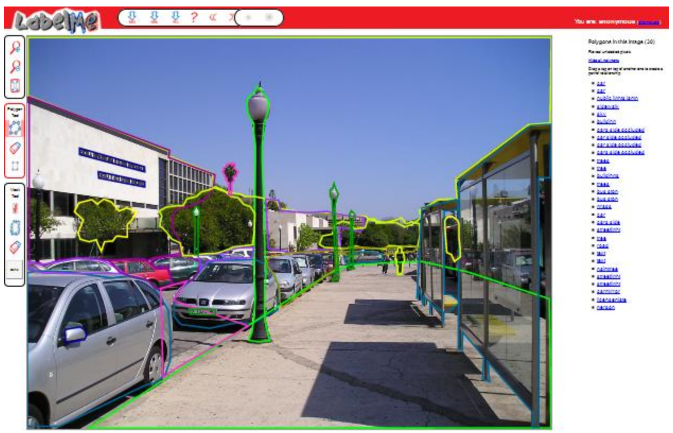
In collaboration with Snap Inc.



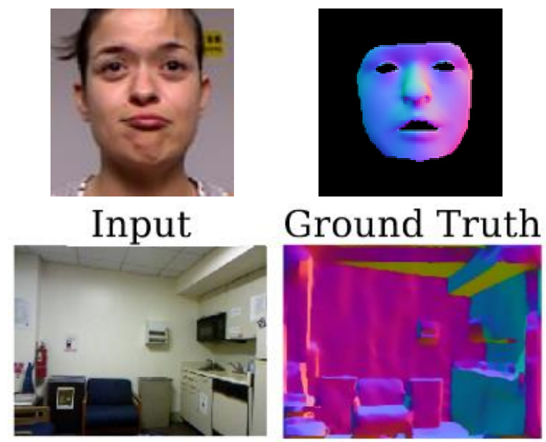
Spatial ground truth problems

- Hard to obtain

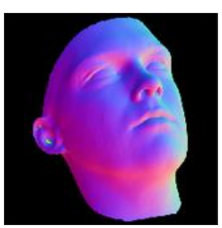
Time-consuming



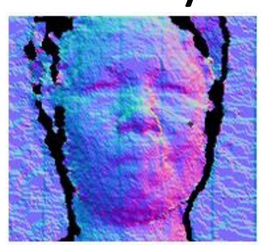
Estimations: unfaithful



Cannot manually annotate



Estimations: noisy



- Use domain adaptation (and synth data) to help!

Unsupervised domain adaptation

- Make distribution between domains match

  - Feature space [1,2]
  - Input space (Refiner [3], CyCADA [4])
  - Output space [5,6]

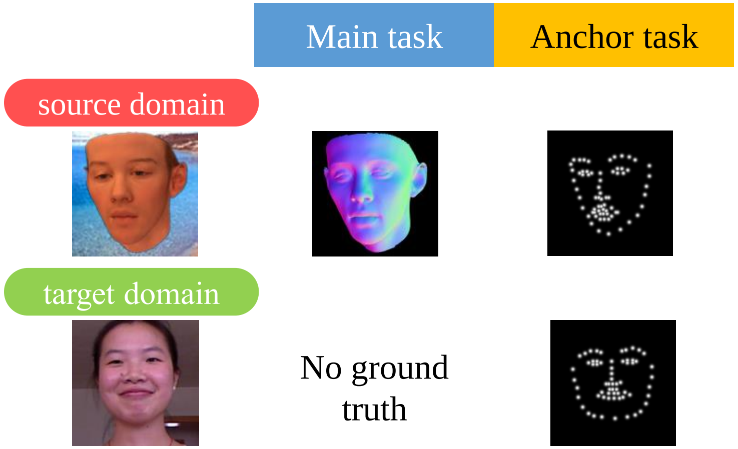
- Assume distributions *should* be made identical

Semantic and spatial info can help matching

[1] Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation."
[2] Mingsheng Long et al. "Learning transferable features with deep adaptation networks".
[3] Ashish Shrivastava et al. "Learning from Simulated and Unsupervised Images through Adversarial Training"
[4] Judy Hoffman et al. "CyCADA: Cycle Consistent Adversarial Domain Adaptation".
[5] Kuniaki Saito et al. "Maximum classifier discrepancy for unsupervised domain adaptation"
[6] Yi-Hsuan Tsai et al. "Learning to Adapt Structured Output Space for Semantic Segmentation".

Task-Assisted Domain Adaptation (TADA)

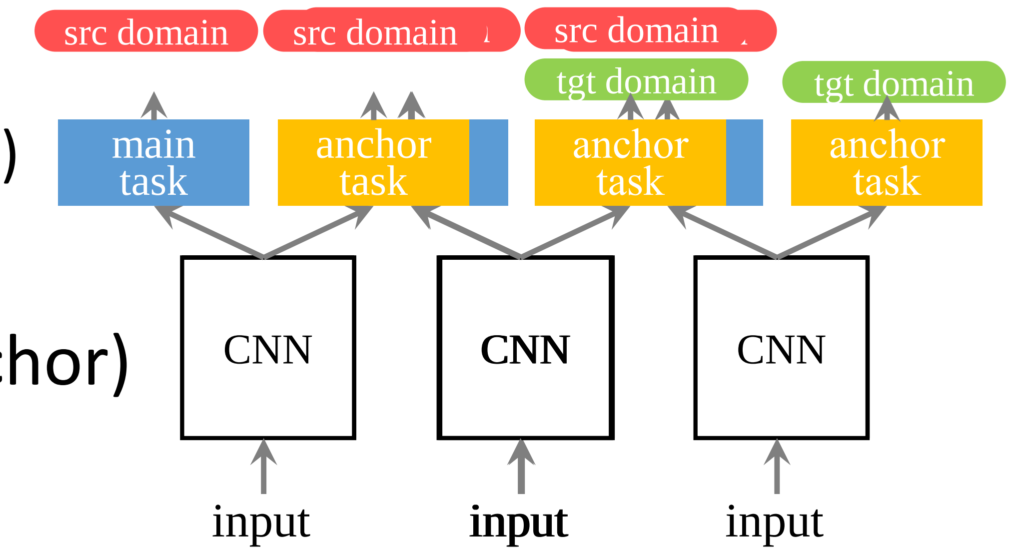
- How about an auxiliary supervised task?



- Pick "anchor task"
  - Easier to obtain
  - Guidance info (e.g. semantic / spatial)
  - On both domains
- No explicit task relationship needed!

Method

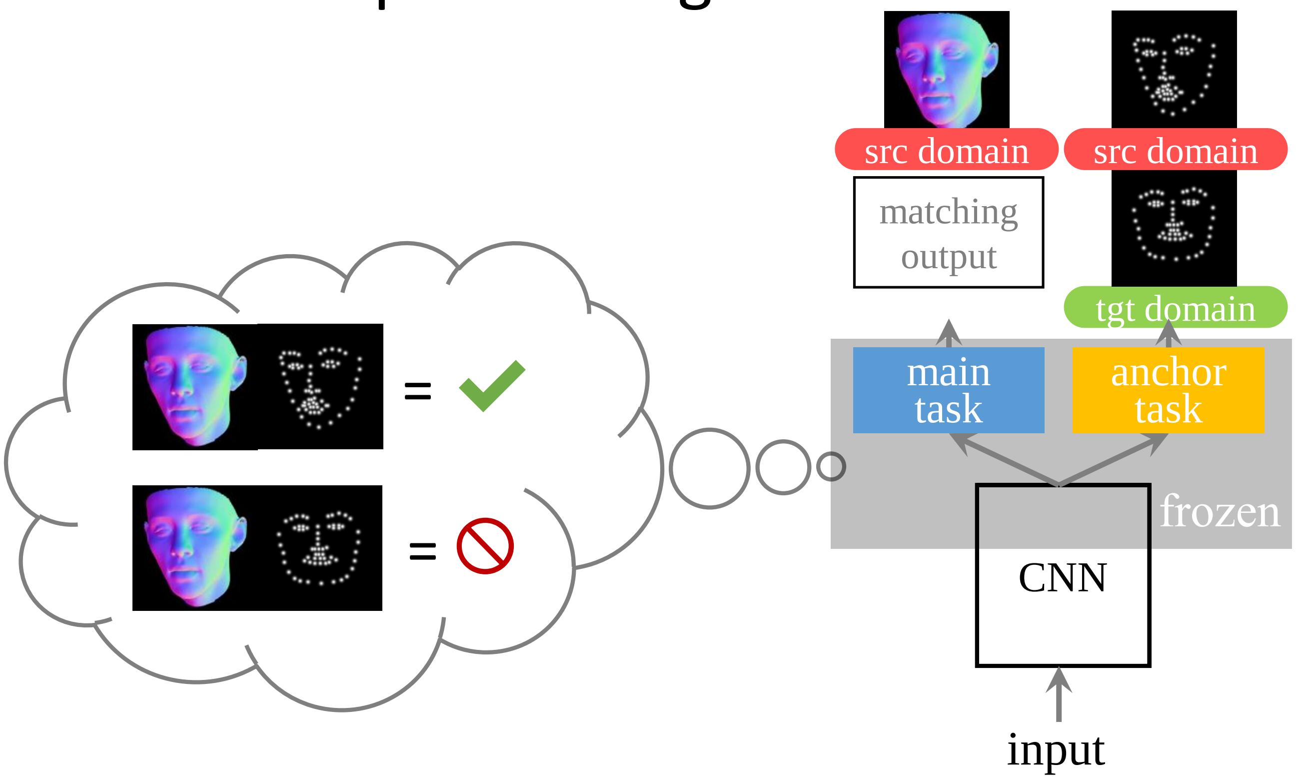
- Baselines
  - Single task
  - Multi-task (not shared)
- Multi-task (shared anchor)



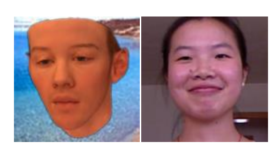
(b) Multitask learning (source only) (c) Multitask learning with anchor tasks (d) Multitask learning (one task per domain)

Method

- Relationship modeling



(e) FREEZE (ours)



Experiments

- Two datasets

normal | keypoints

Main task | Anchor task

source domain

SfSNet [1]

target domain

Face-Warehouse

No ground truth

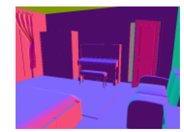
state-of-the-art keypoint estimator [2]

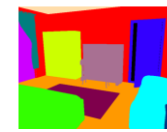
normal | semantic

Main task | Anchor task

source domain

SUNCG





renders

target domain

NYUdv2

No ground truth

estimations (from depth)

labels

- Compared methods
  - Baseline, oracle
  - SfSNet [1]

[1] Soumyadip Sengupta et al. “SfSNet: Learning Shape, Reflectance and Illuminance of Facesin the Wild’”

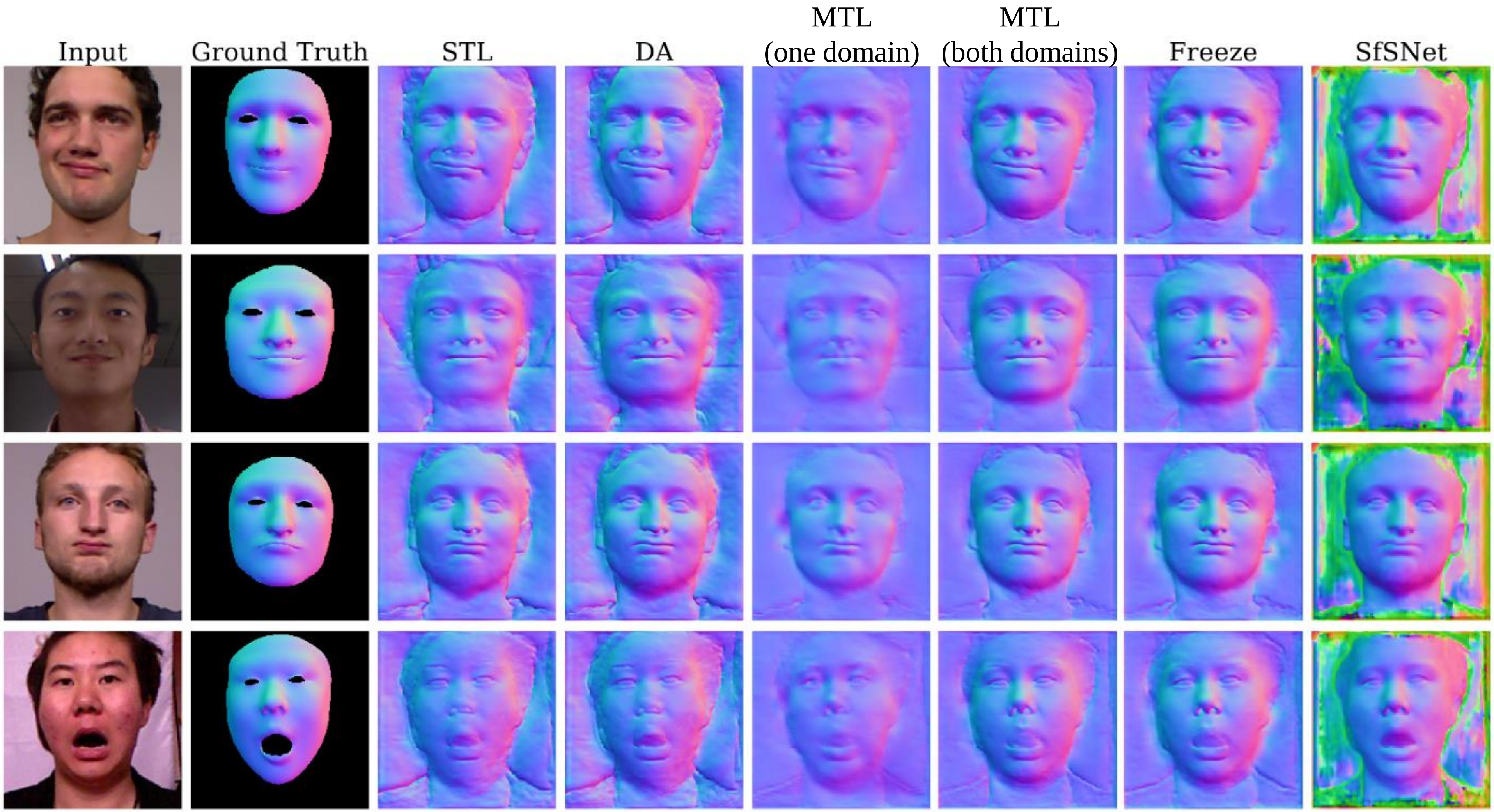
[2] Adrian Bulat and Georgios Tzimiropoulos. “How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)”.

Results

	src main	src anch	tgt main	tgt anch	Faces SfSsyn→FaceWH $< 11.25°$	$< 30°$	RMSE	Mean	Median
STL	✓				0.424	0.929	17.8	14.8	12.8
DA [1]	✓				0.456	0.937	17.2	14.2	12.1
MTL	✓	✓			0.409	0.935	17.7	14.9	13.1
MTL	✓			✓	0.162	0.791	24.3	21.8	20.4
MTL	✓	✓		✓	0.492	**0.953**	16.0	13.3	11.4
FREEZE (ours)	✓	✓		✓	**0.519**	**0.954**	**15.8**	**12.9**	**10.9**

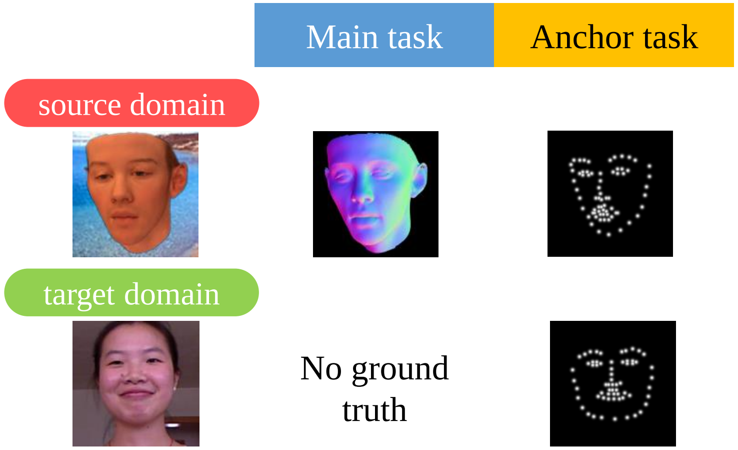
[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domainadaptation by backpropagation. InInternational Conferenceon Machine Learning, pages 1180–1189, 2015

Qualitative results



Take-away

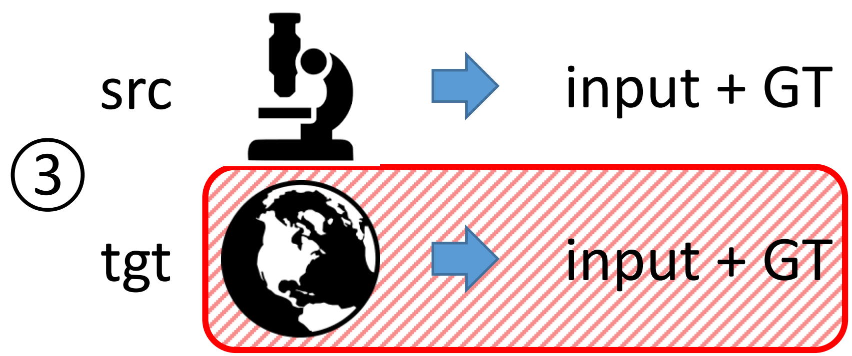
- Matching distributions are not enough for unsupervised domain adaptation
- Easy-to-obtain labels for another task can help
- Modeling task relationship can help



Study: Improving Confidence Estimates for Unfamiliar Examples

Zhizhong Li, Derek Hoiem

Accepted in CVPR 2020
as an oral presentation



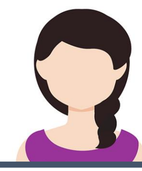
[redacted] driver dies in first fatal crash while using autopilot mode

The autopilot sensors on [redacted] failed to distinguish a white tractor-trailer crossing the highway against a bright sky

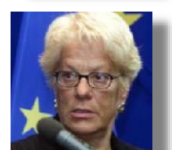
(biased) dataset collection

97% accuracy

user input



P(female) = 99.9%



P(male) = 99.3%

CNN model

??

99%+ confidence = <1% error rate?

- 0.5% w/ familiar
- 6.0% w/ unfamiliar
  - 12x errors!

Problems:

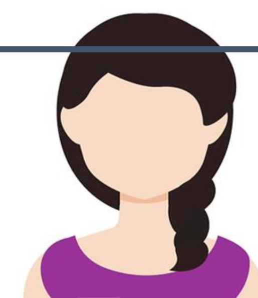
- Test data different in unexpected ways
- Underrepresented data get confidently misclassified

Prior work

- Domain adaptation, Domain generalization
  - Needs knowing variations of future domains

I'm adapted to in-the-wild images!

I've got you some new stuff



Adapted models

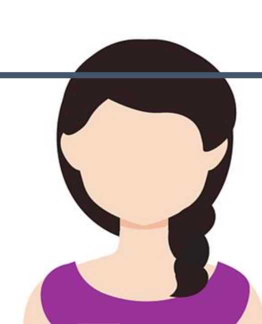
Prior work

Novelty detection

These are not faces I have seen.

Yeah, but aren't you going to classify them?

Novelty detector

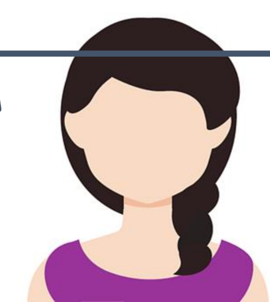


Prior work

Modeling epistemic uncertainty

- (i.e. uncertainty due to lack of knowledge)

I am not familiar with these, so I make predictions with adjusted confidence



Desired model

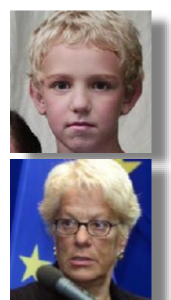
Goal

- Comparative study
  - Which prior work has the most well-behaved confidence on unseen data?
- How to evaluate?

P(female) = 99.9%

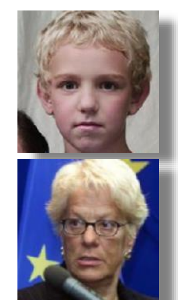
P(male) = 99.3%

CNN model



These are not faces I have seen.

Novelty detector

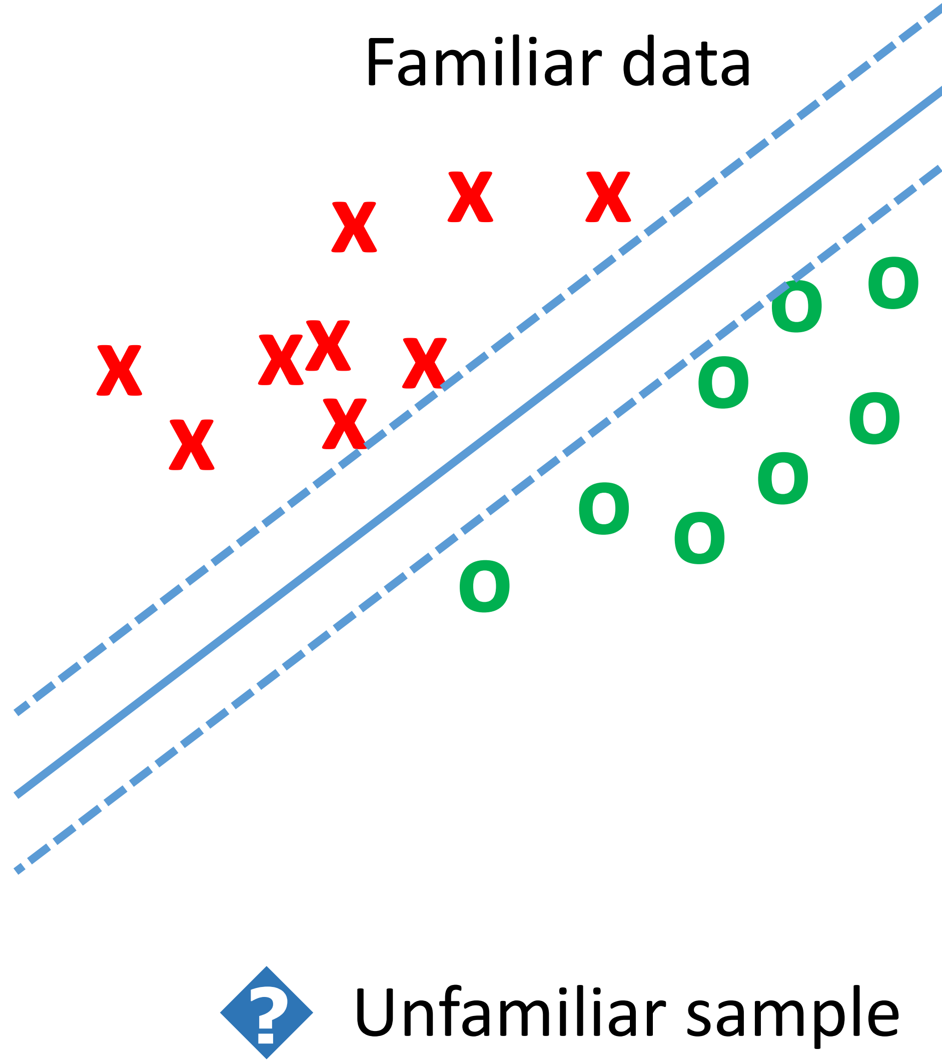


I am not familiar with these, so I make predictions with adjusted confidence

Uncertainty model

Comparative study

- List of compared works
  - Regularly-trained model (baseline)
  - Modeling uncertainty [2]
  - Calibration with temperature-scaling [1]
  - Ensemble
    - Calibrated ensemble
    - Distilling [3]
    - Distilling [3] (modified)
  - Novelty detection [4] (modified)



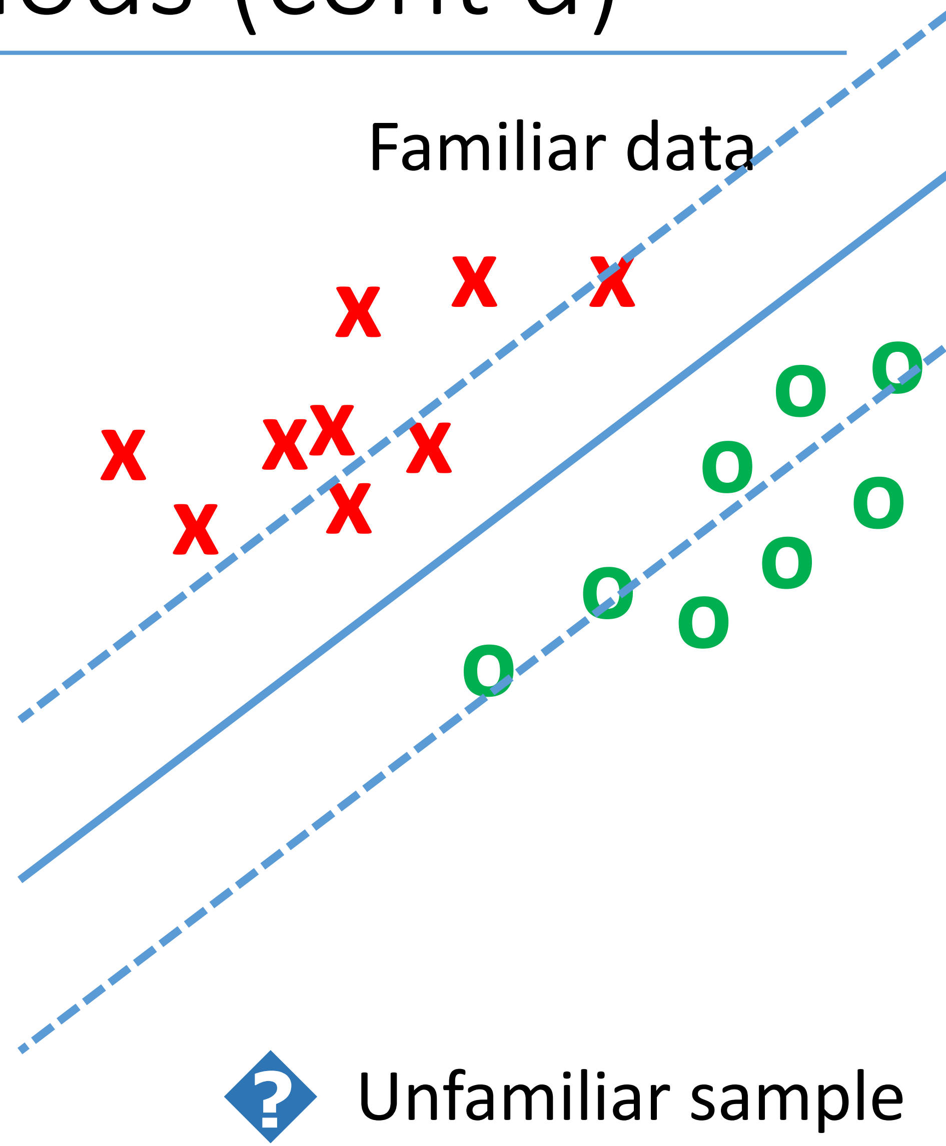
[1] Guo, Chuan, et al. "On Calibration of Modern Neural Networks."
[2] Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?."
[3] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network."
[4] Liang, Shiyu, Yixuan Li, and R. Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks."

Compared methods (cont'd)

- Calibration
  - Temperature scaling [1]

$$\mathbf{p}(x) = \mathrm{softmax}\big(\mathbf{f}(x)\big)$$
$$\mathbf{p}'(x) = \mathrm{softmax}(\mathbf{f}(x)/T)$$

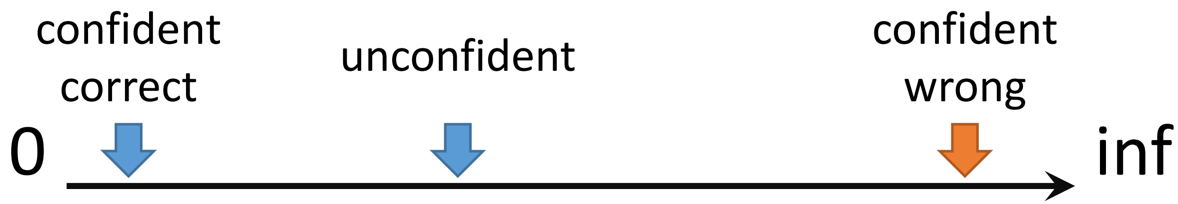
  - Use a higher temperature in the prediction
  - Calibrate the temperature in a validation set



[1] Guo, Chuan, et al. "On Calibration of Modern Neural Networks."

Experimental setup

- Evaluate confidence: Negative log-likelihood (NLL)



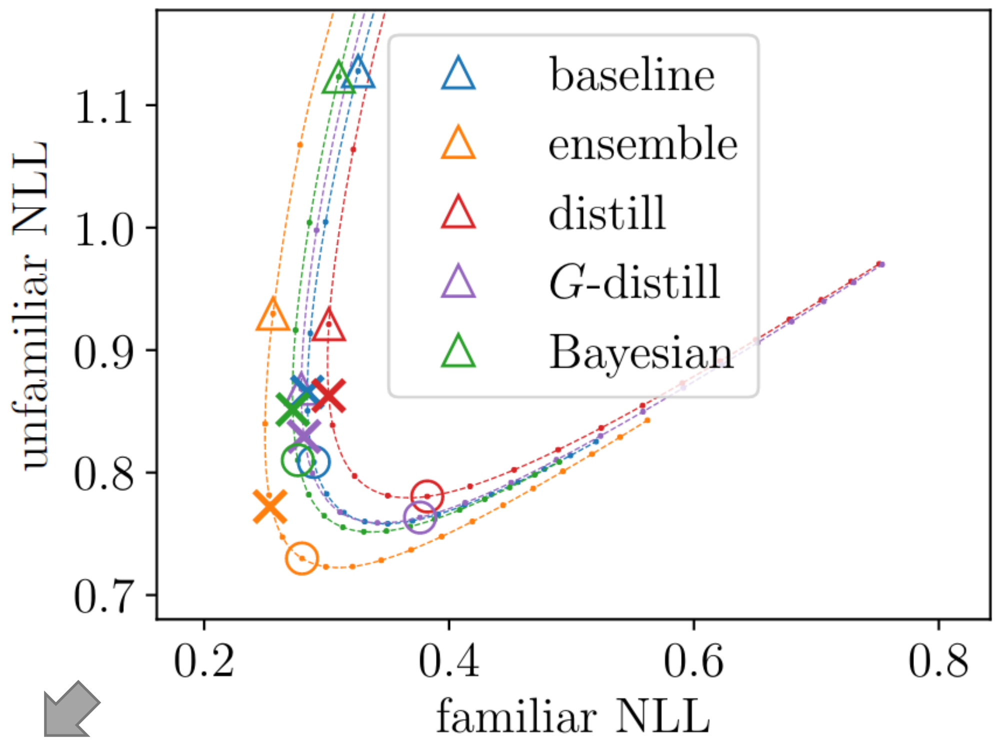
- Get underrepresented data: split by subcategories

Dataset	Familiar	Unfamiliar
LFW+ (face gender)	Ages 18-59	Ages 0-17, 60+
ImageNet superclass*	Some species	Other species
Pets (cat v. dog)	Some breeds	Other breeds
VOC-COCO, 20 classes	PASCAL VOC, whole dataset	MSCOCO, ignoring non-VOC classes

* mammals vs. herptiles vs. birds vs. fishes

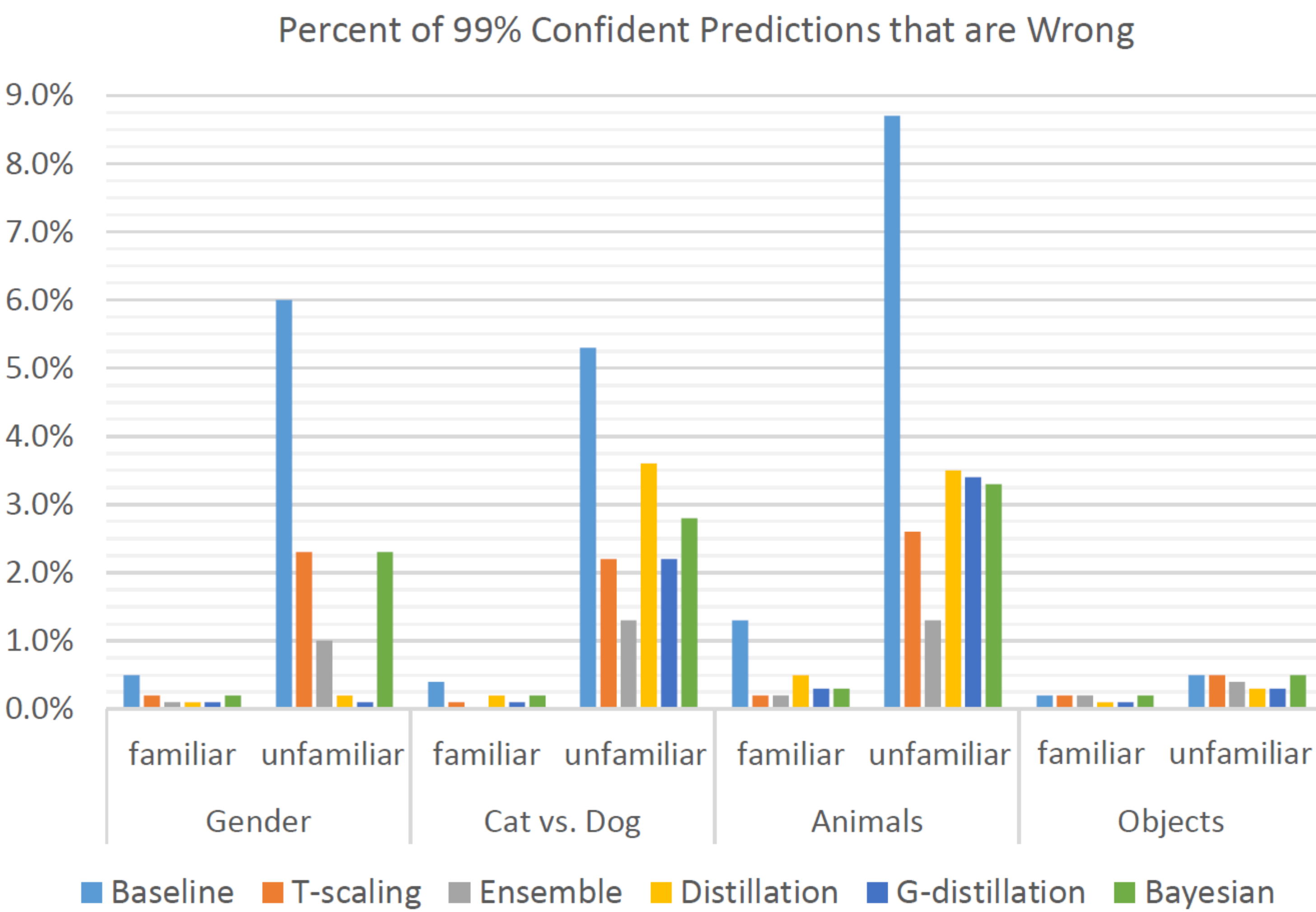
Results: Negative log-likelihood

- Animal classification (ImageNet subset)
  - Smoothing effect: trade-off familiar / unfamiliar



bottom-left better

Results: errors among 99% confident



Take-away

Issue highlight: data underrepresented in training can get confidently misclassified

~~= 99.9% female~~

= 84.5% female

~~= 99.3% male~~

= 53.7% male

Model reducing confident errors

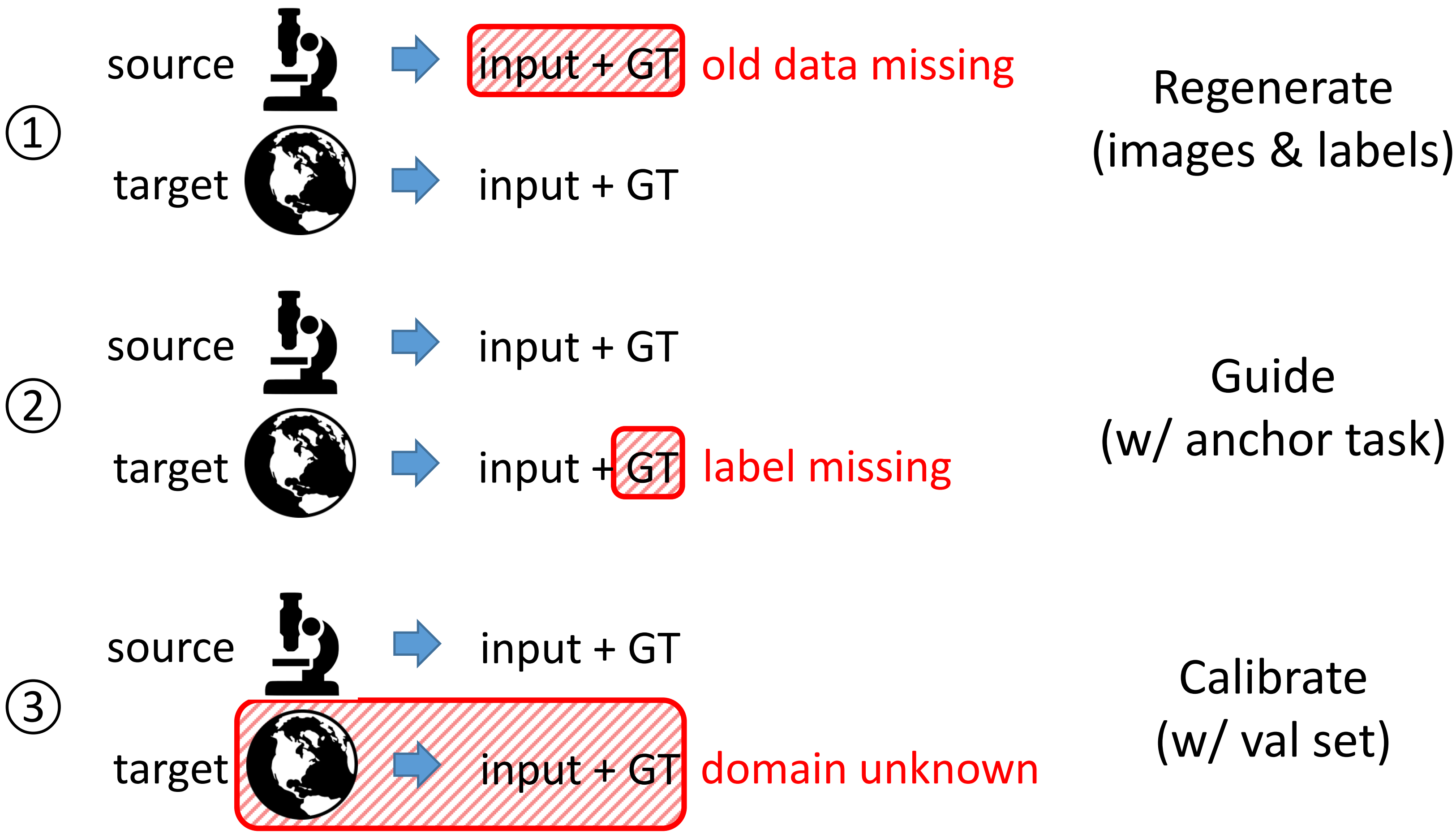
Best-performing methods

- Calibrated ensembles
  - -32% unfamiliar NLL
- Calibration (T-scaling)
  - -23% unfamiliar NLL

Experimental method

- Split familiar / unfamiliar by subcategories

Story so far


①
source
input + GT
old data missing
target
input + GT
Regenerate
(images & labels)
②
source
input + GT
target
input + GT
label missing
Guide
(w/ anchor task)
③
source
input + GT
target
input + GT
domain unknown
Calibrate
(w/ val set)

Future work

Open-ended question:

- How to improve knowledge transfer?
- How to circumvent data constraints in industry settings?

What IS knowledge?

① (x_i, y_i)? $\hat{y} = f(x)$? $\mathbb{P}(x)$?

② Interaction between two tasks' y_1, y_2?

③ Generalization where $\mathbb{P}(x) \approx 0$?

More?

Future work

Leverage other knowledge in humans

- How different things behave
- Why things behave this way
- Does this new thing behave the same way
- How does this knowledge affect my decisions
- Etc.

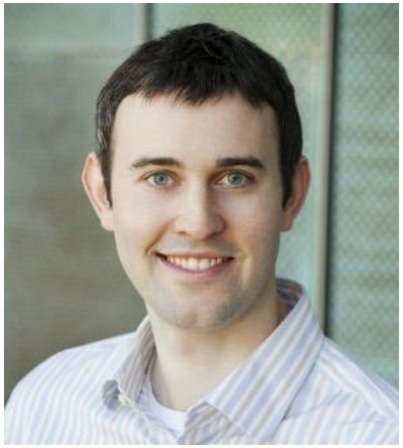
Can we extract these from models?

Can these be represented without using data?

Can we use these to improve knowledge transfer?

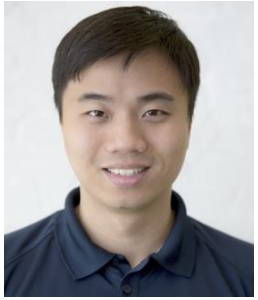
Acknowledgements

- My advisor



Derek Hoiem

- Collaborators



Linjie Luo

Sergey Tulyakov

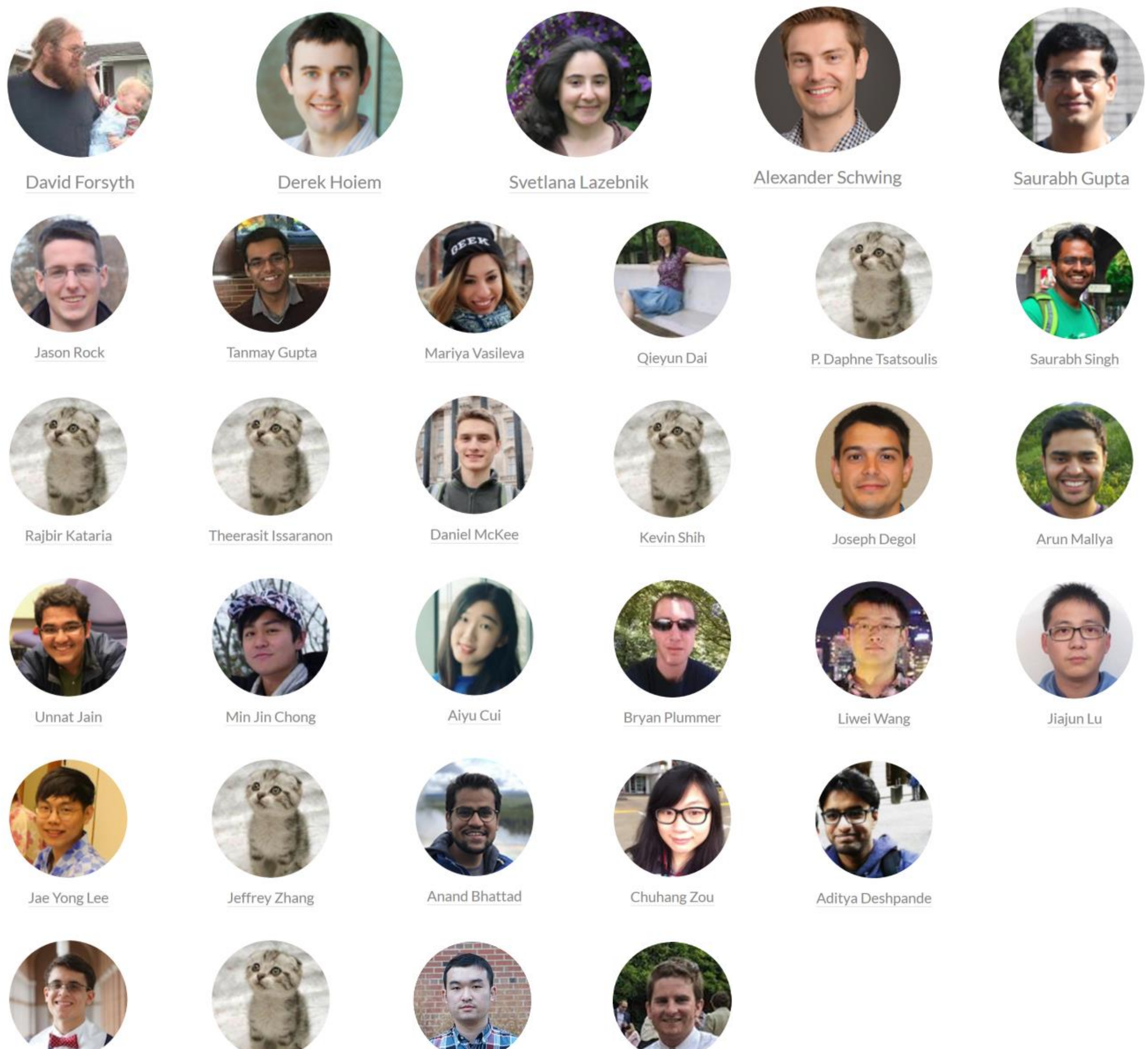
Qieyun Dai

Arun Mallya

Hongxu Yin

Pavlo Molchanov

Thanks everyone!



David Forsyth, Derek Hoiem, Svetlana Lazebnik, Alexander Schwing, Saurabh Gupta

Jason Rock, Tanmay Gupta, Mariya Vasileva, Qieyun Dai, P. Daphne Tsatsoulis, Saurabh Singh

Rajbir Kataria, Theerasit Issaranon, Daniel McKee, Kevin Shih, Joseph Degol, Arun Mallya

Unnat Jain, Min Jin Chong, Aiyu Cui, Bryan Plummer, Liwei Wang, Jiajun Lu

Jae Yong Lee, Jeffrey Zhang, Anand Bhattad, Chuhang Zou, Aditya Deshpande

Victor Gonzalez, Mantas Mazeika, Shuai Tang, Dominic Roberts

Questions?

Network pruning results

- Resnet 50; Compared to methods that use real data

Image Source	Top-1 acc. (%) -50% filters -71% FLOPs	Top-1 acc. (%) -20% filters -37% FLOPs
No finetune	1.9	16.6
Partial ImageNet		
0.1M images / 0 label	69.8	74.9
Proxy datasets		
MS COCO	66.0	73.8
PASCAL VOC	54.4	70.8
GAN		
Generator, BigGAN	63.0	73.7
Noise (Ours)		
DeepInversion (DI)	55.9	72.0
Adaptive DeepInversion (ADI)	60.7	73.3

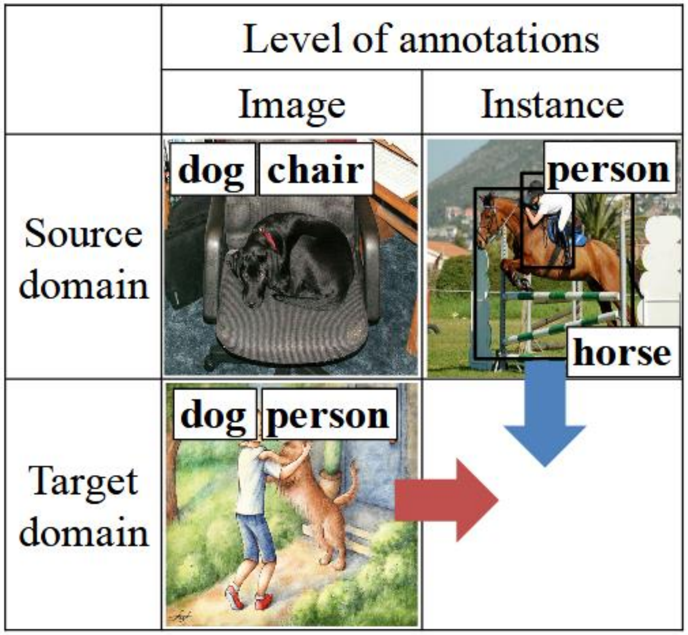
Knowledge transfer results

- Resnet 50 v1.5; from scratch

Image source	Data	Top-1 acc.
Base Model	1.3M, Real	77.26%
DeepInversion	140K, Dream	73.8%

Existing work with TADA structure

- Focus on known, explicit main-auxiliary label relationships [1,2,3,4]



[1] Kuan Fang et al. “Multi-Task Domain Adaptation for Deep Learning of Instance Grasping from Simulation”
[2] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. “Fine-Grained Recognition in the Wild: A Multi-task Domain Adaptation Approach”
[3] Naoto Inoue et al. “Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation”.
[4] Wei Yang et al. “3D Human Pose Estimation in the Wild by Adversarial Learning”.

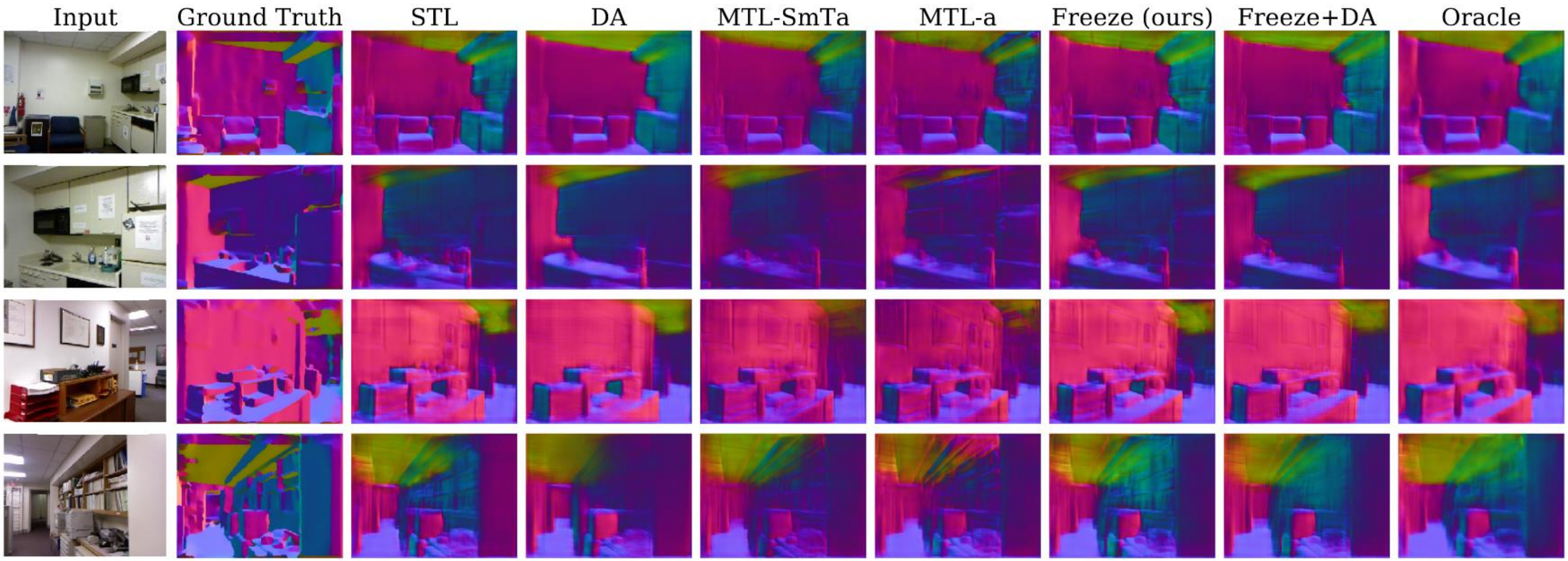
Results

- + Domain adaptation

	src main	src anch	tgt main	tgt anch	Faces SfSsyn→FaceWH $< 11.25°$	$< 30°$	RMSE	Mean	Median
STL	✓				0.424	0.929	17.8	14.8	12.8
MTL-src	✓	✓			0.409	0.935	17.7	14.9	13.1
MTL-SmTa	✓			✓	0.162	0.791	24.3	21.8	20.4
MTL-a	✓	✓		✓	0.492	**0.953**	16.0	13.3	11.4
FREEZE (ours)	✓	✓		✓	**0.519**	**0.954**	**15.8**	**12.9**	**10.9**
with unsupervised domain adaptation: [1]									
DA	✓				**0.456**	0.937	17.2	14.2	**12.1**
MTL-src	✓	✓			0.402	0.932	18.0	15.1	13.3
MTL-SmTa	✓			✓	0.216	0.854	22.0	19.5	18.1
MTL-a	✓	✓		✓	**0.455**	**0.946**	**16.7**	**13.9**	**12.1**
FREEZE (ours)	✓	✓		✓	**0.455**	0.935	17.2	14.2	**12.1**

[1] Yi-Hsuan Tsai et al. "Learning to Adapt Structured Output Space for Semantic Segmentation".

Qualitative results



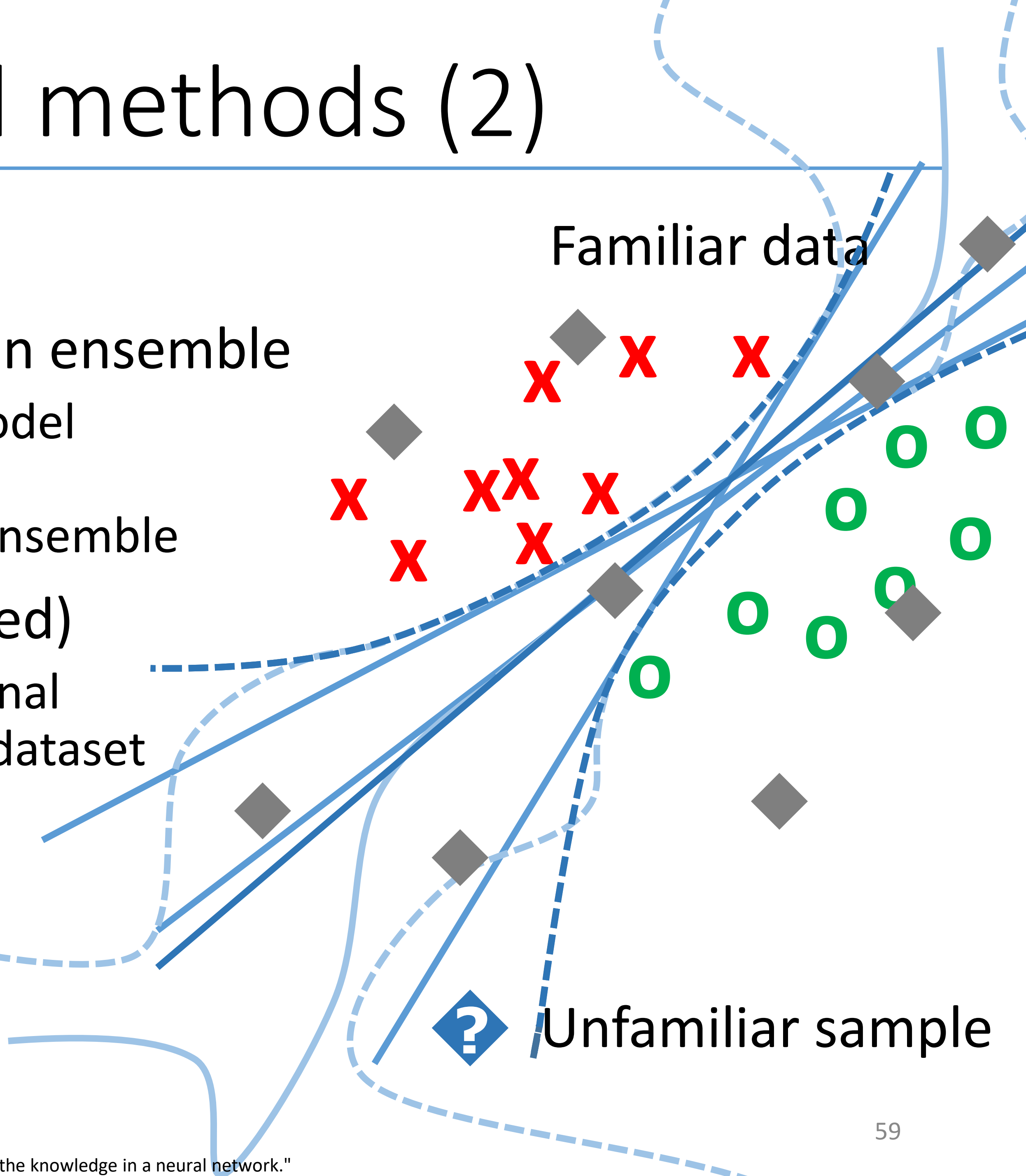
Better ceiling / wall

(STL works pretty well already)

Compared methods (2)

- Ensemble
- “Distilling” [1] an ensemble
  - Train single model on soft labels to mimic the ensemble
- G-distill (modified)
  - Use an additional unsupervised dataset

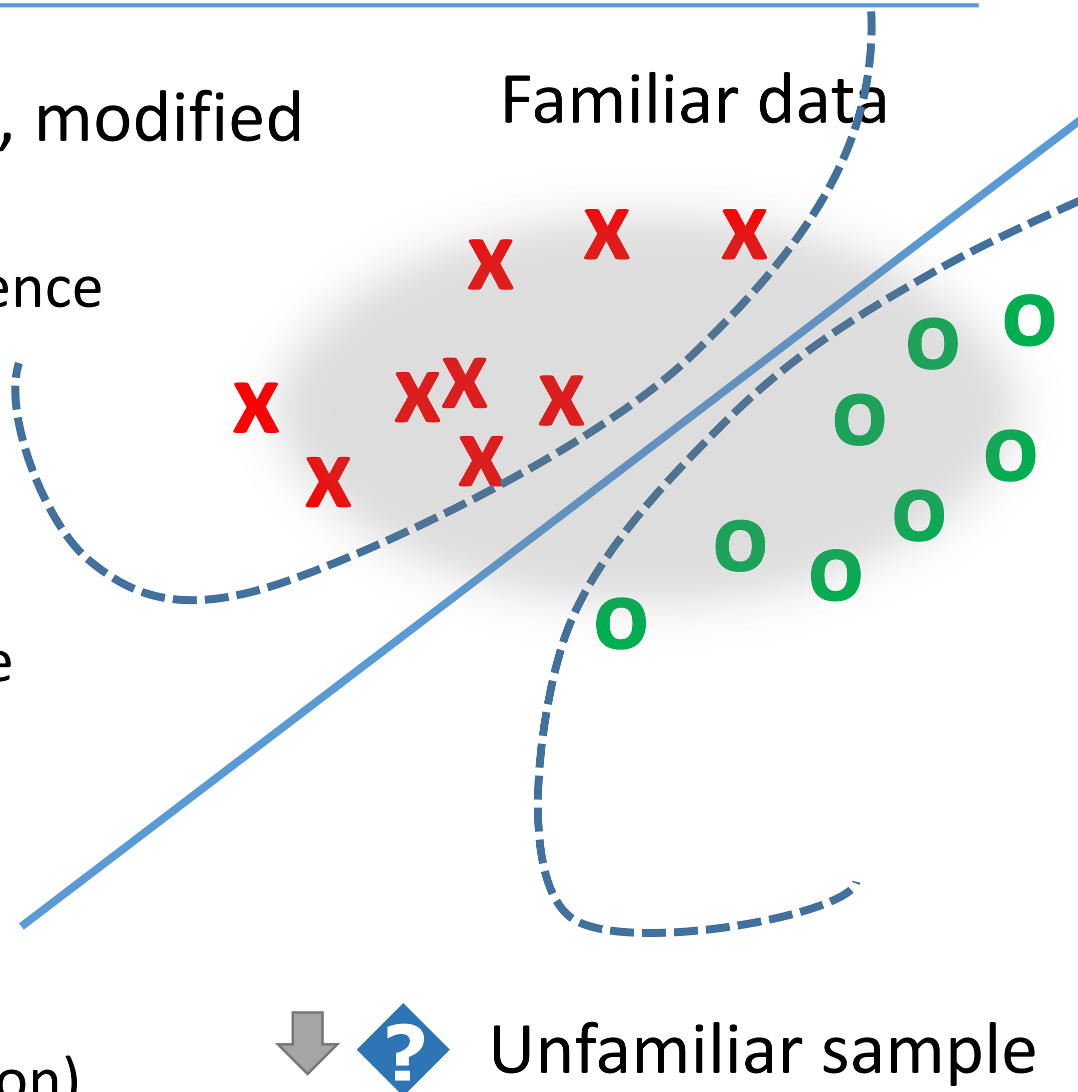
e.g. Internet pictures



[1] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network."

Compared methods (3)

- Novelty detection [1], modified (cannot use directly)
  1. Get original confidence
  2. Run novelty detection procedure
  3. Higher outlier score ▼ more reduction in confidence

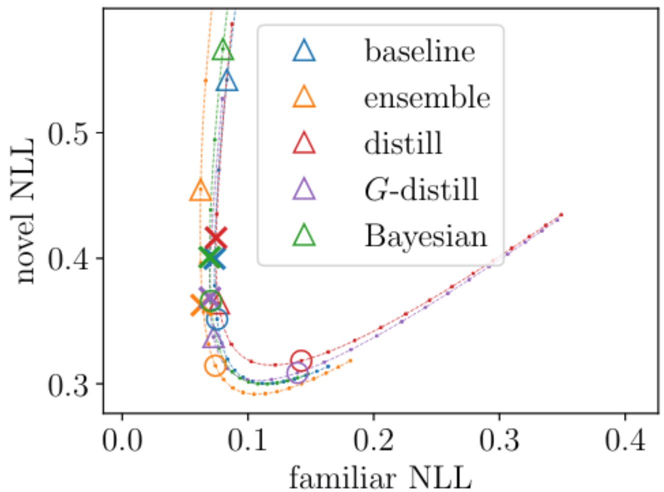


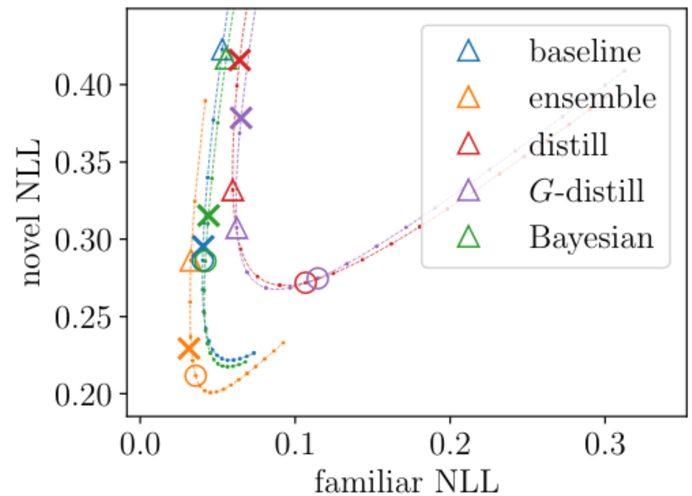
"NCR"
(Novel Confidence Reduction)

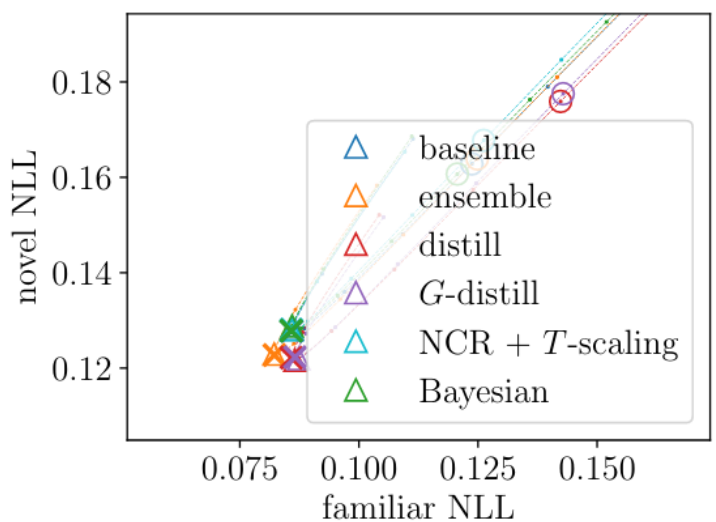
[1] Liang, Shiyu, Yixuan Li, and R. Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks."

Results: Negative log-likelihood

- Face gender, Pets cat vs. dogs, VOC-COCO







VOC too similar to COCO

Results: errors among 99% confident