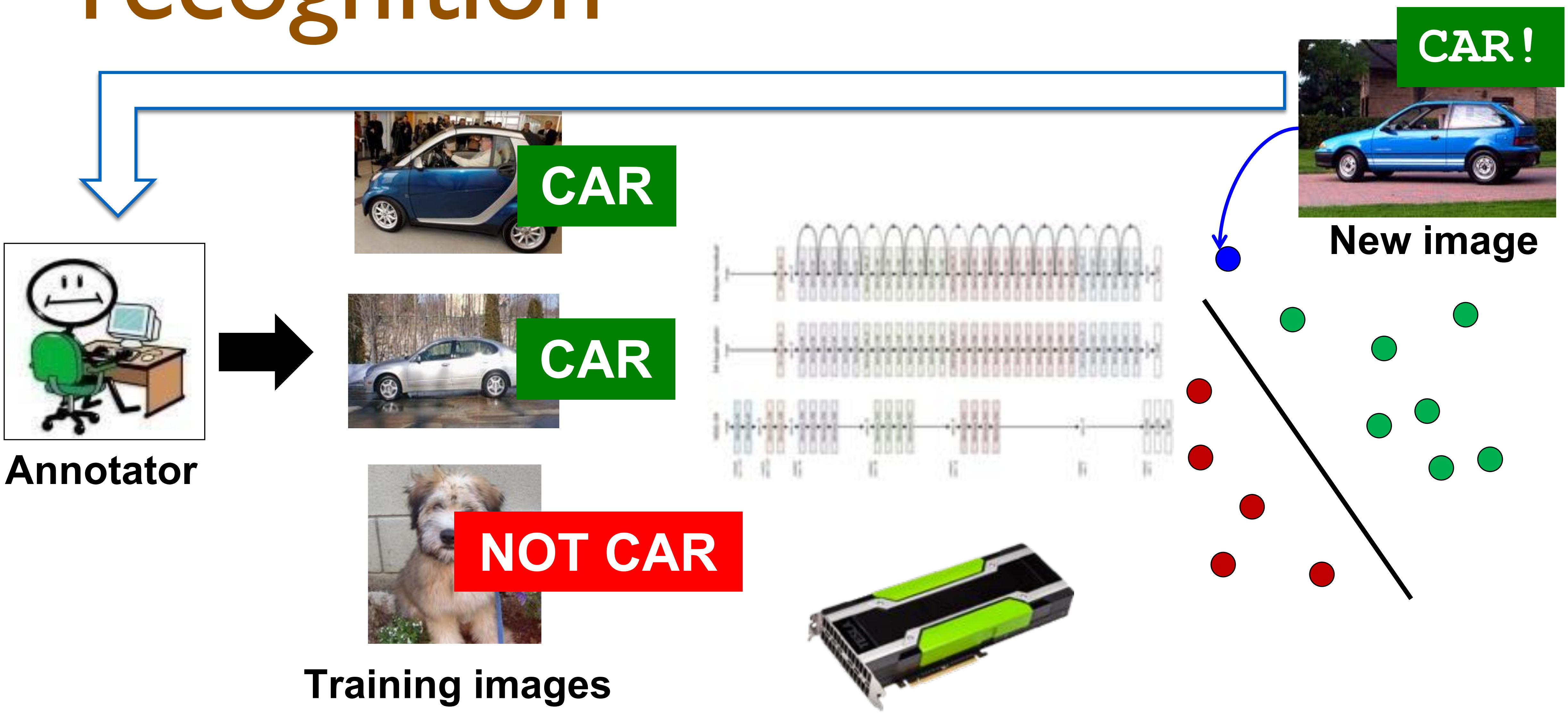
Learning and Adapting from the Web for Visual Recognition

Boqing Gong

Tencent AI Lab

Learning based visual recognition

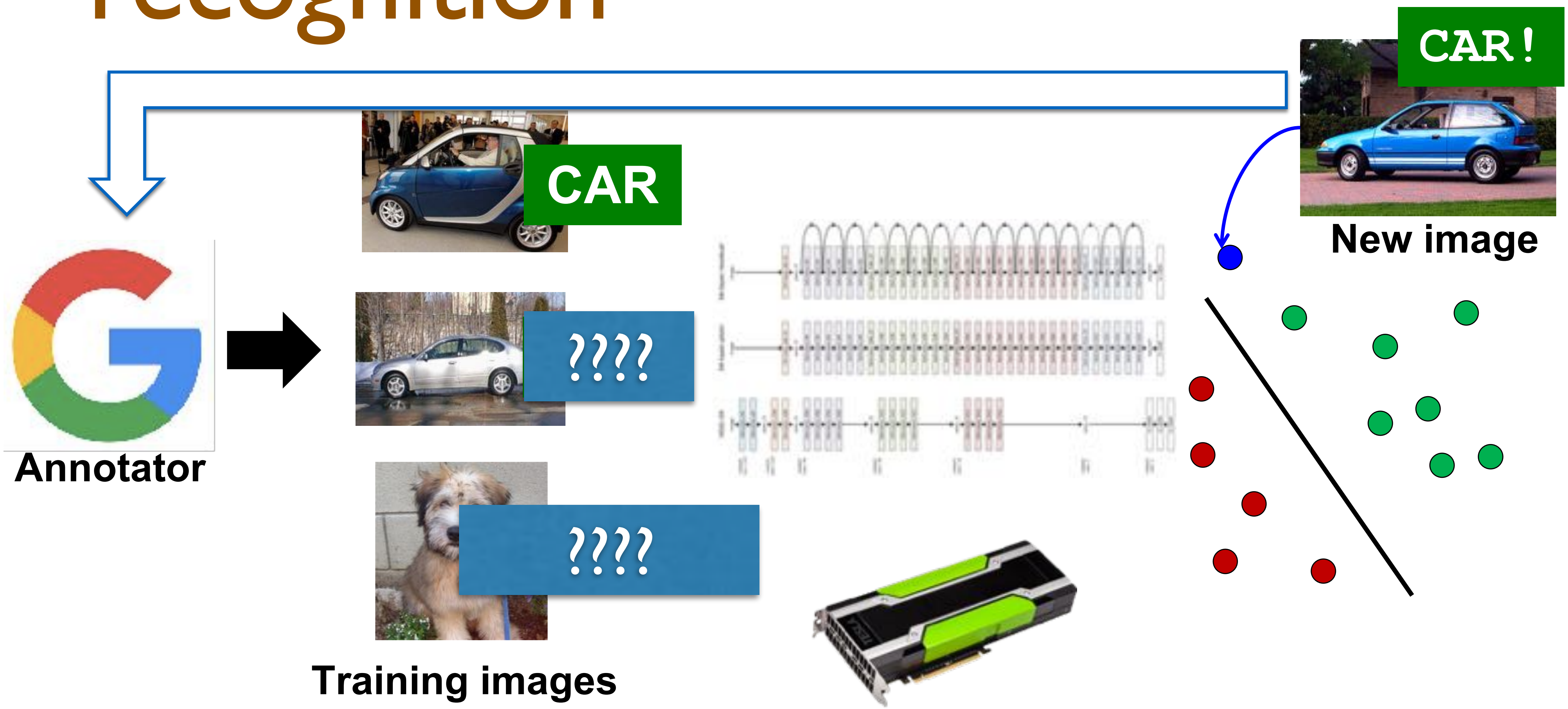


Courtesy K. Grauman

Domain adaptation: key to use simulation "for real"

Simulation to reality for segmentation, detection, Dynamics planning & control, etc.

Learning based visual recognition



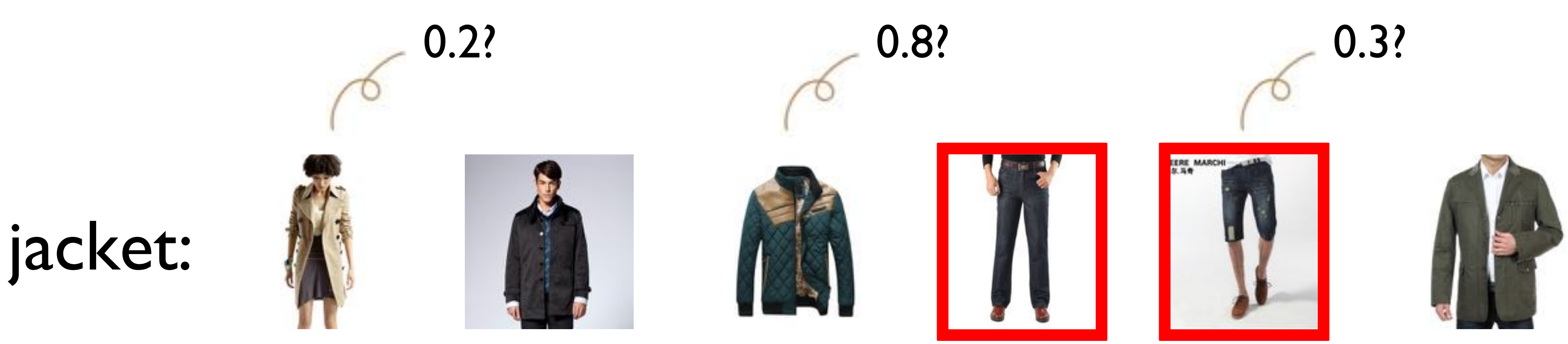
Web data with **noisy labels**
Need different training methods

Courtesy K. Grauman

Label correction & re-weighting

Correct wrong labels

Reweigh data/label terms



Label ~~correction & re-weighting~~ removal

Correct wrong labels

cat:

Hard to rectify wrong labels

Easier to simply remove them (but keep the images)

Semi-supervised learning?

Caveat: outlier images

Label ~~correction & re-weighting~~ removal

Correct wrong labels

cat:

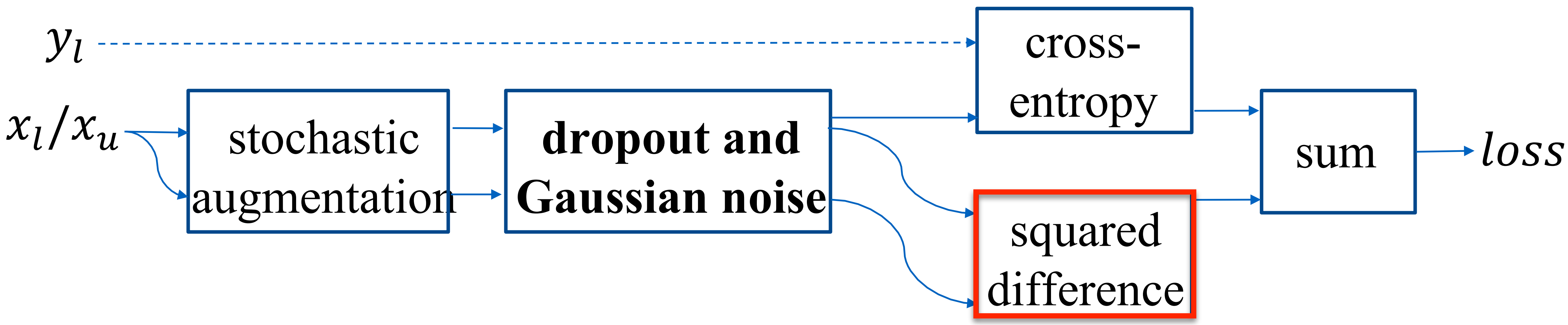
Hard to rectify wrong labels

Easier to simply remove them (but keep the images)

Semi-supervised learning?

Caveat: outlier images

A consistent term & its dual effect



Outlier images still help.

[Laine & Aila, ICLR'17]

“Web data” with noisy labels, no outlier

Results on CIFAR10 & MNIST

Table 3. Comparison results on CIFAR-10 and MINIST

Methods	CIFAR-10 14-layer ResNet				MNIST fully connected			
	$p = 0$	sy.$p = 0.2$	asy.$p = 0.2$	asy.$p = 0.6$	$p = 0$	sy.$p = 0.2$	asy.$p = 0.2$	asy.$p = 0.6$
cross-entropy [37]	87.8	83.7	85.0	57.6	97.9± 0.0	96.9± 0.1	97.5± 0.0	53± 0.6
unhinged (BN) [57]	86.9	84.1	83.8	52.1	97.6± 0.0	96.9± 0.1	97.0± 0.1	71.2 ± 1.0
sigmoid (BN) [12]	76.0	66.6	71.8	57.0	97.2± 0.1	93.1± 0.1	96.7± 0.1	71.4± 1.3
savage [30]	80.1	77.4	76.0	50.5	97.3± 0.0	96.9± 0.0	97.0± 0.1	51.3± 0.4
bootstrap soft [40]	87.7	84.3	84.6	57.8	97.9± 0.0	96.9± 0.0	97.5± 0.0	53.0± 0.4
bootstrap hard [40]	87.3	83.6	84.7	58.3	97.9± 0.0	96.8± 0.0	97.4± 0.0	55.0± 1.3
backward [37]	87.7	80.4	83.8	66.7	97.9± 0.0	96.9± 0.0	96.7± 0.1	67.4± 1.5
forward [37]	87.4	83.4	**87.0**	74.8	97.9± 0.0	96.9± 0.0	97.7± 0.0	64.9± 4.4
cross-entropy	87.9	82.4	85.5	56.2	98.0± 0.1	97.1± 0.1	97.6± 0.2	52.9± 0.6
improved baseline	87.8	83.6	85.2	74.1	98.0± 0.1	97.1± 0.1	97.7± 0.1	**76.7± 1.6**
ours	**88.0**	**84.5**	85.6	**75.8**	**98.2± 0.1**	**97.7± 0.4**	**97.8± 0.1**	**83.4± 1.3**

[Ding et al., WACV’18]

"Web data" with noisy labels & outlier images

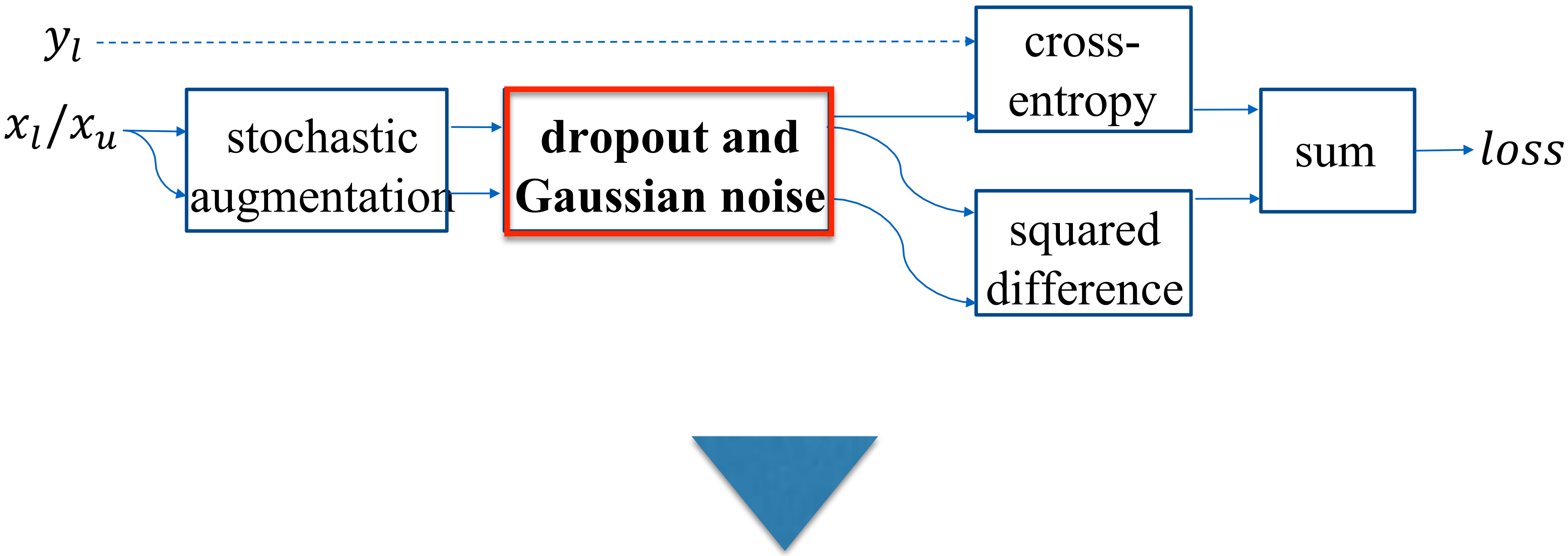
Results on Clothing1M

Table 4. Comparison results on the Clothing1M dataset [59].

#	model	loss / method	initialization	training set	accuracy (reported)	accuracy (our impl.)
1	AlexNet	pseudo-label [25]	#9	1M, 50K	73.04	–
2	AlexNet	bottom-up [47]	#9	1M, 50K	76.22	–
3	AlexNet	label noise model [59]	#9	1M, 50K	78.24	–
4	50-ResNet	cross-entropy	ImageNet	1M	68.94	69.03
5	50-ResNet	backward [37]	ImageNet	1M	69.13	–
6	50-ResNet	forward [37]	ImageNet	1M	69.84	–
7	50-ResNet	**ours**	ImageNet	1M	–	77.34
8	50-ResNet	**ours**	ImageNet	1M, 50K	–	**79.38**
9	AlexNet	cross-entropy	ImageNet	50K	72.63	–
10	50-ResNet	cross-entropy	ImageNet	50K	75.19	74.84
11	50-ResNet	cross-entropy	#6	50K	80.38	–
12	50-ResNet	cross-entropy	#7	50K	–	80.44
13	50-ResNet	cross-entropy	#8	50K	–	**80.53**

[Ding et al., WACV'18]

A consistent term & its dual effect



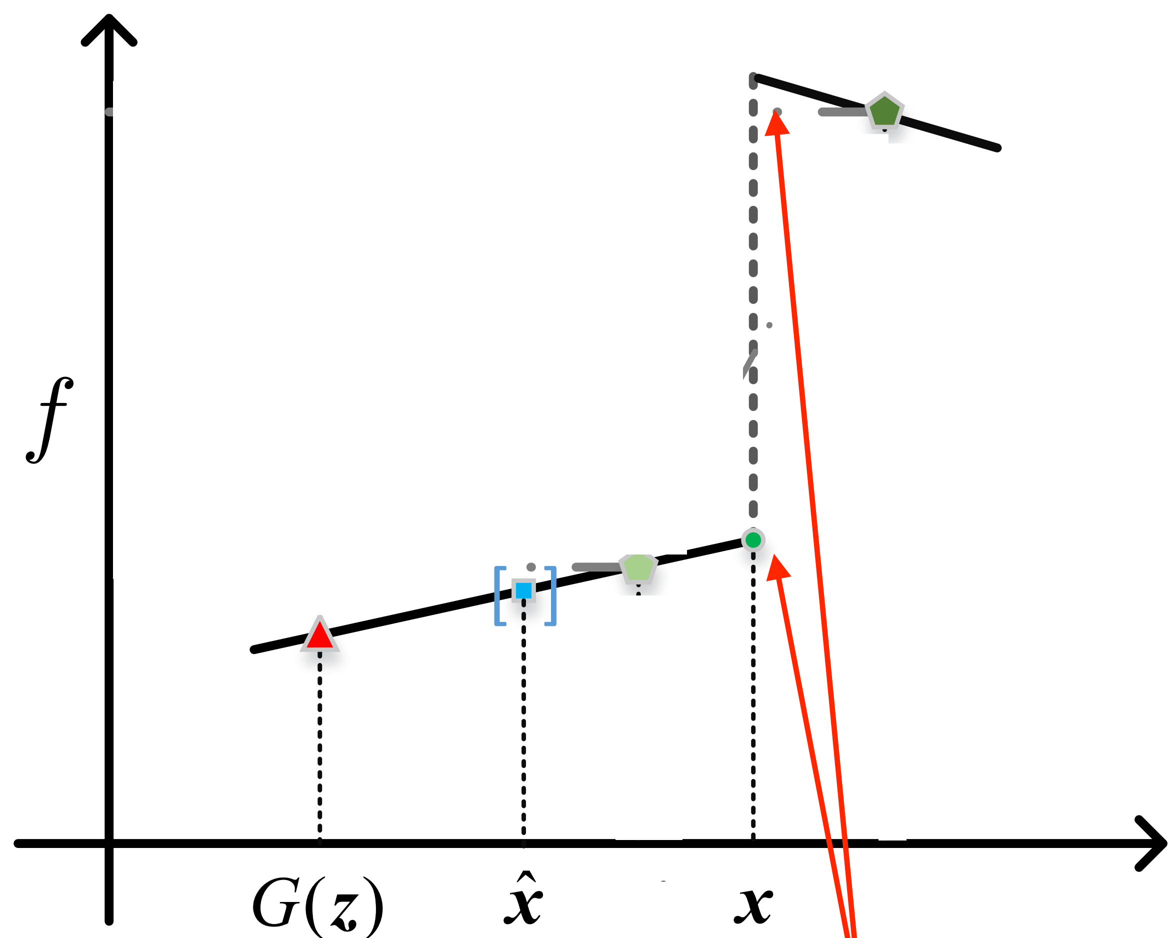
Lipschitz continuity in WGAN

WGAN

Discriminator/critic is Lipschitz continuous

$$\begin{aligned} \min_f \quad & \text{Loss} \\ \text{s.t.} \quad & \|f\|_L \leq 1 \end{aligned}$$

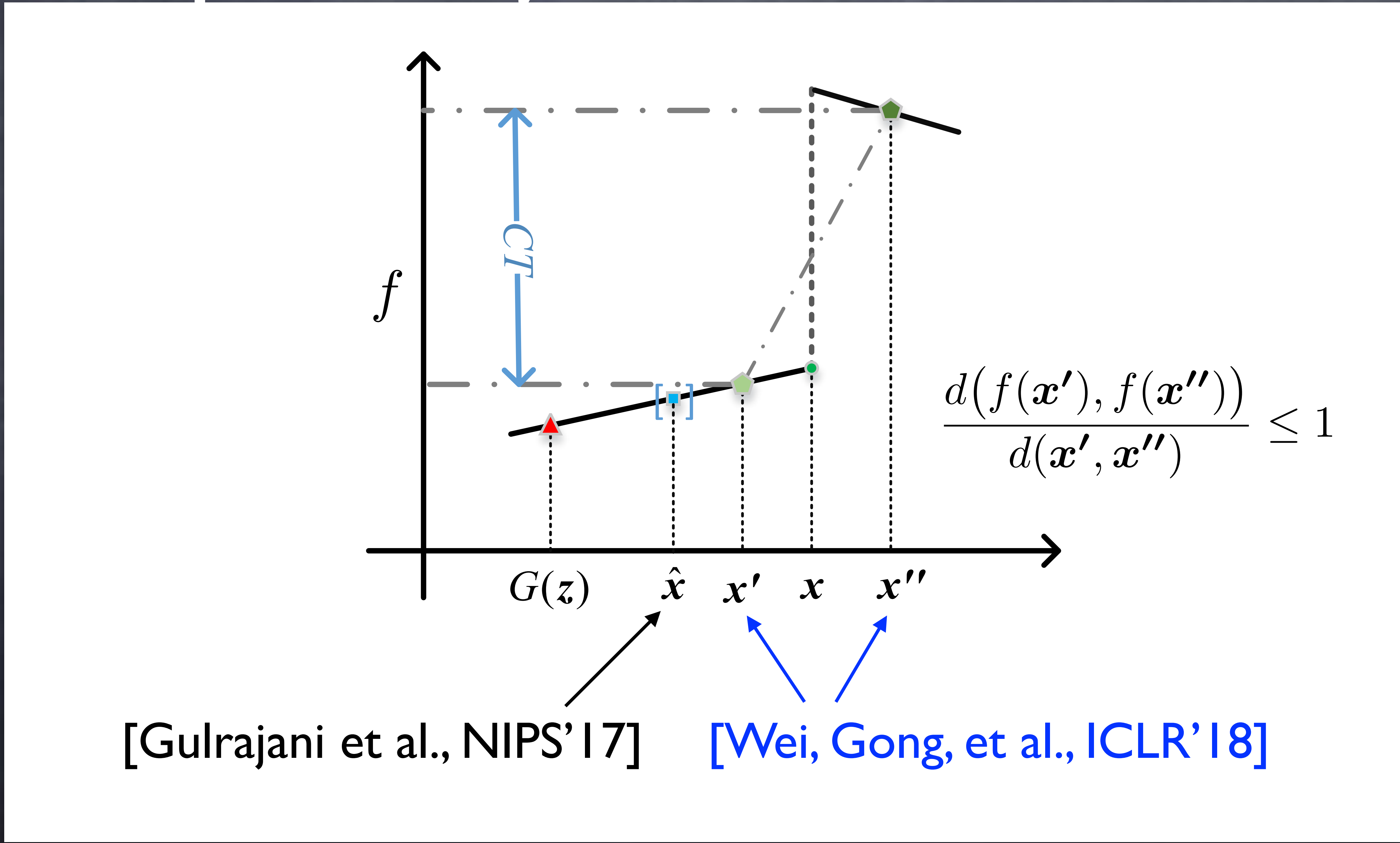
L-continuity by gradient penalty



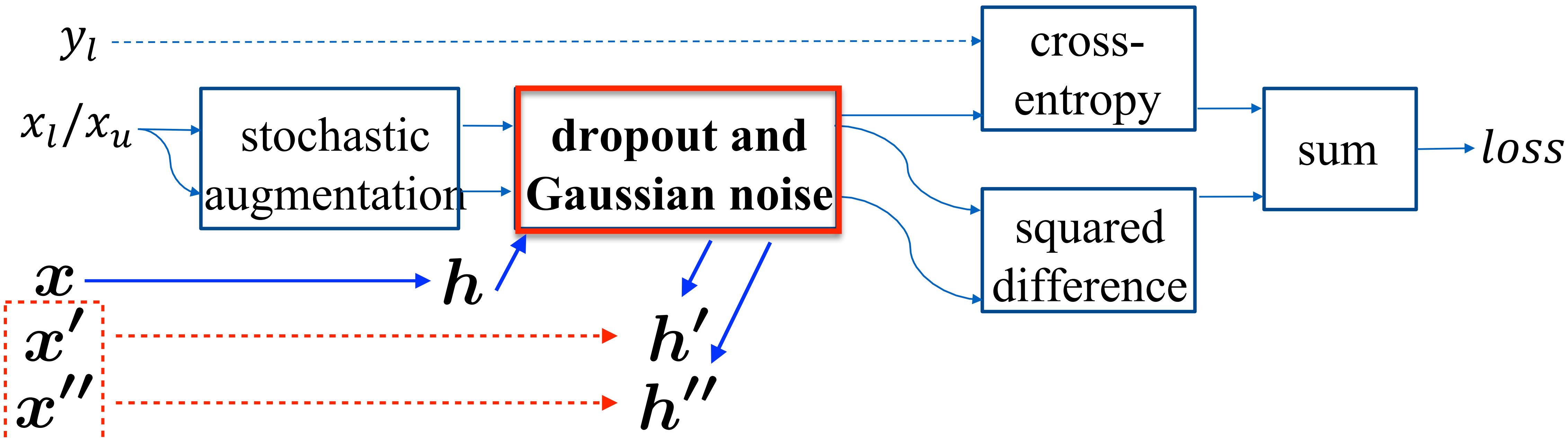
[Gulrajani et al., NIPS'17]

Fails to check the regions near real data

L-continuity by gradient penalty & definition



A consistent term & its dual effect



$$\min_f \quad \text{Loss} + \mathbb{E}_{x',x''} \max\left[0, \frac{d(f(x'),f(x''))}{d(x',x'')}\right]$$

A consistent term & its dual effect

Results on CIFAR10 (Semi-Sup.)

Method	**Test error (%)**
Ladder (Rasmus et al., 2015)	20.40 ± 0.47
VAT (Miyato et al., 2017)	10.55
TE (Laine & Aila, 2016)	12.16 ± 0.24
Teacher-Student (Tarvainen & Valpola, 2017)	12.31 ± 0.28
CatGANs (Springenberg, 2015)	19.58 ± 0.58
Improved GANs (Salimans et al., 2016)	18.63 ± 2.32
ALI (Dumoulin et al., 2016)	17.99 ± 1.62
CLS-GAN (Qi, 2017)	17.30 ± 0.50
Triple GAN (Li et al., 2017a)	16.99 ± 0.36
Improved semi-GAN (Kumar et al., 2017)	16.78 ± 1.80
Our CT-GAN	$\mathbf{9.98 \pm 0.21}$

Outline

Web data with noisy labels

- Hard to rectify wrong labels
- Easier to remove wrong labels

Semi-sup. Learning

WGAN

Web data with accurate labels

- 3D videos/movies

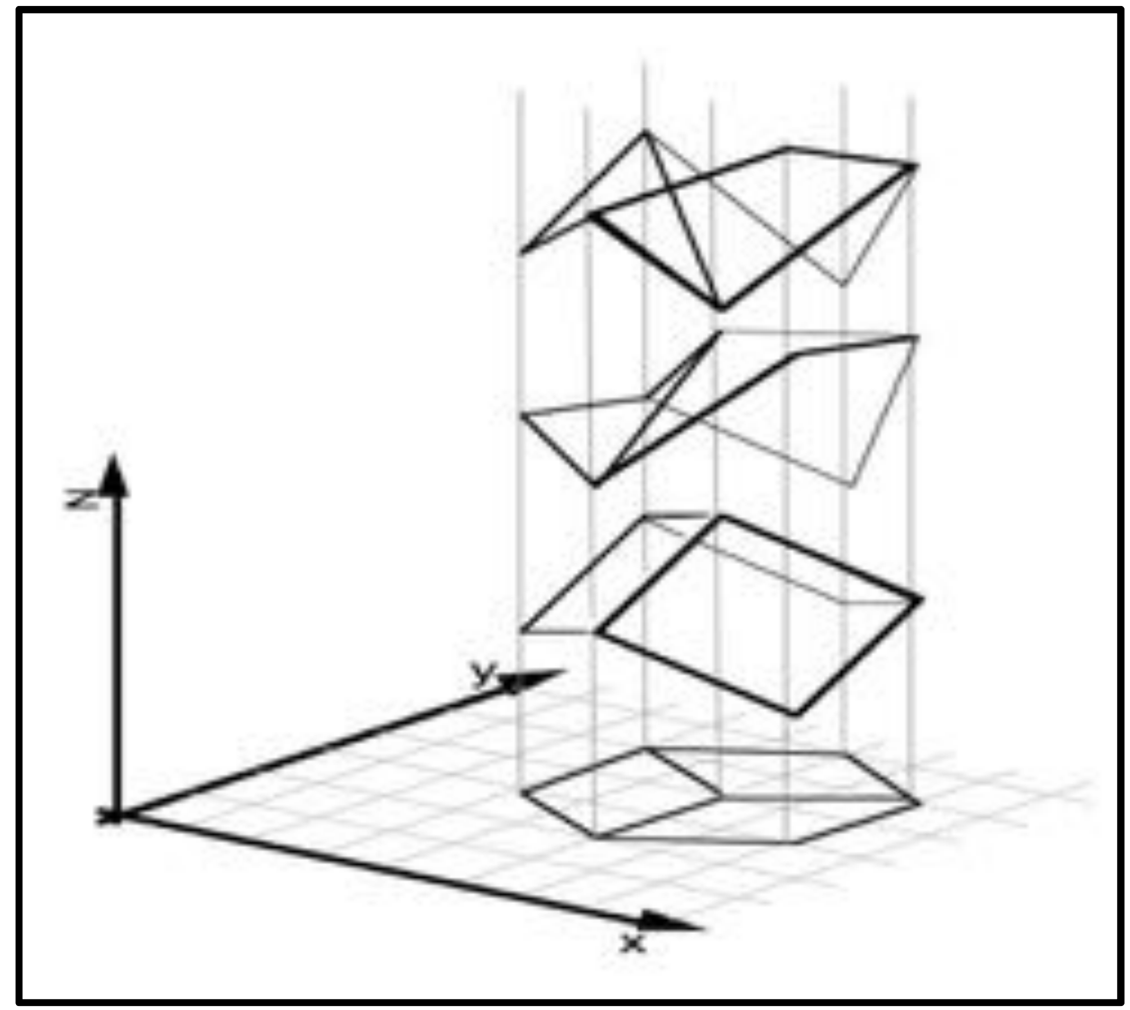
Web data of multi-modalities

- Web images vs. Web videos

3D videos/movies & geometry

Geometry & semantics

[Snavely et al, CVPR '06]



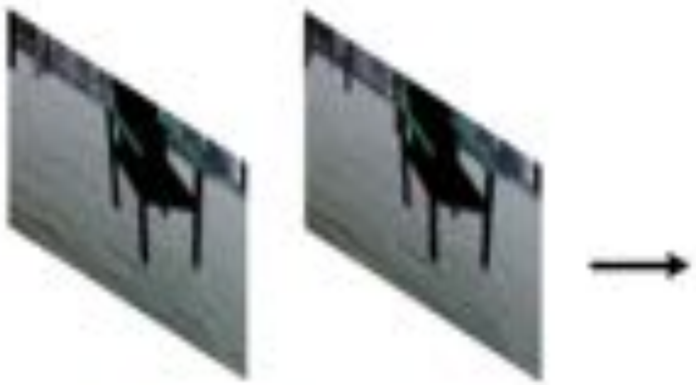
[Sinha et al, ICCV'93]

Shape from dense views
geometric problem

Shape from one view
semantic problem

Courtesy K. Grauman & D. Jayaraman

Geometry guided CNN for semantics tasks



Training on Flying chairs

Begin with synthetic imagery

and precise geometry cues

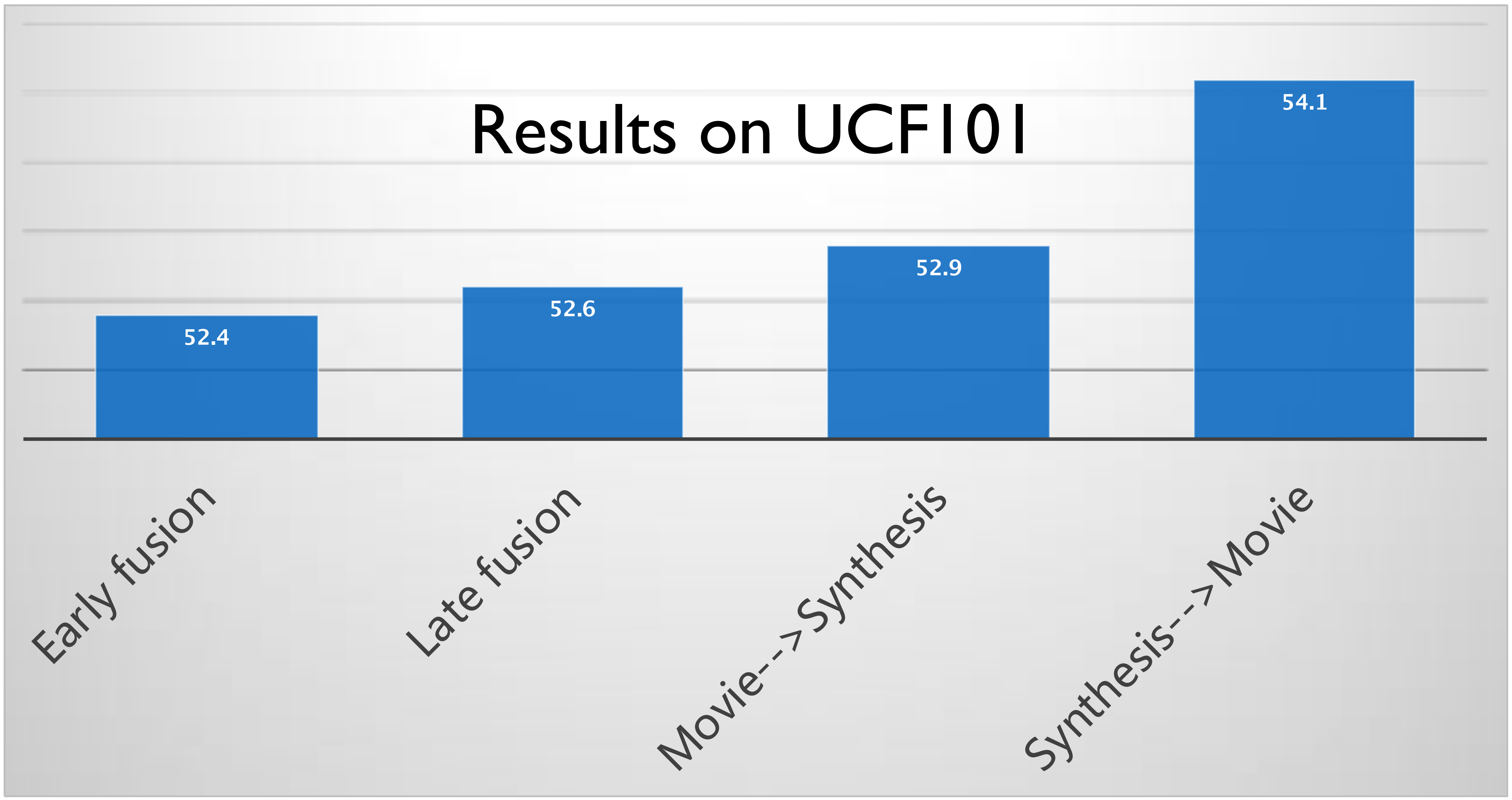
Training on 3D movies

Followed by 3D movies

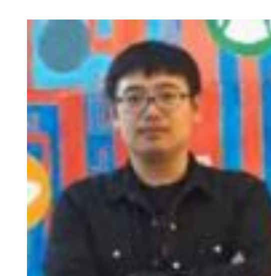
to incorporate reality

Curriculum

Key: to follow the right curriculum

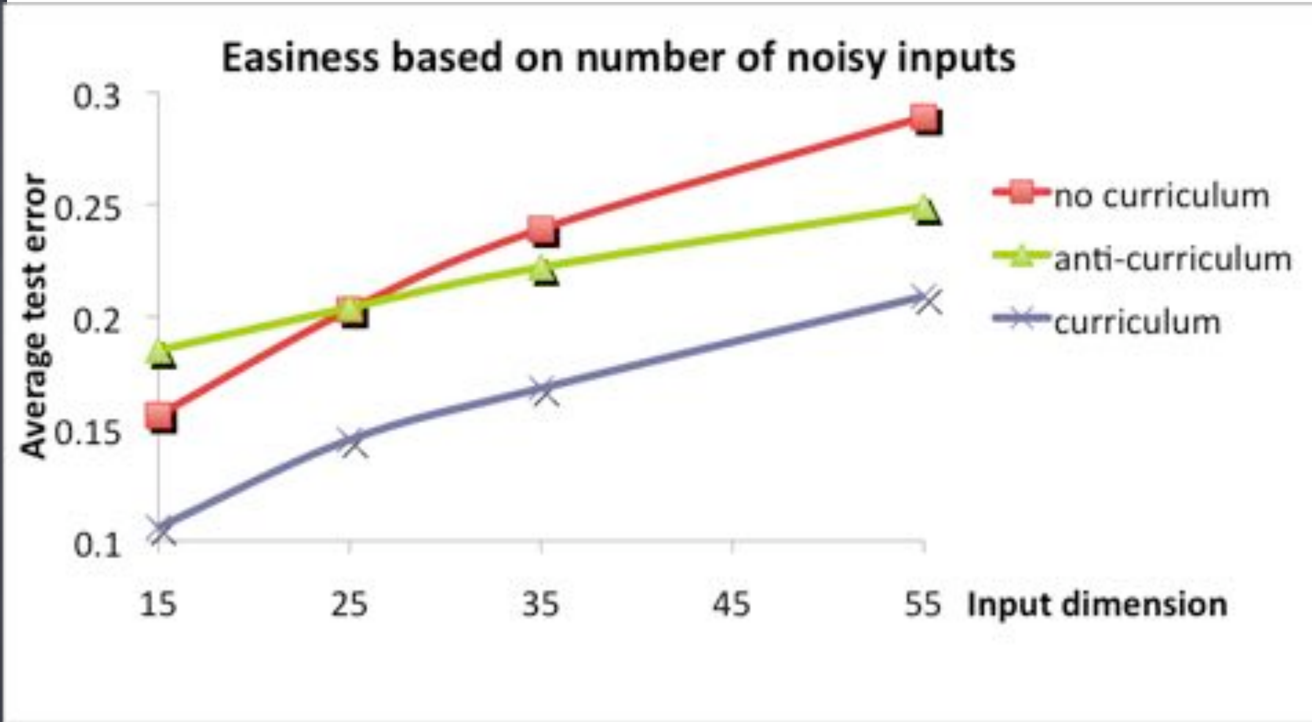


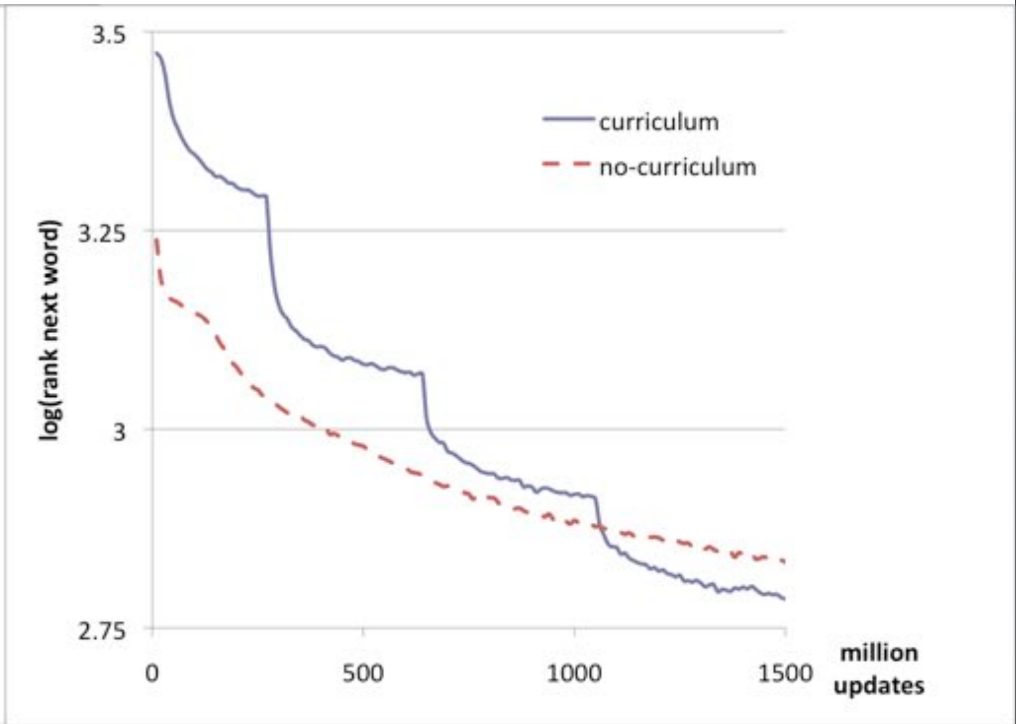
[Gan et al., CVPR'18]



Curriculum learning

Feed a learning system "easy" **examples** first
Gradually introduce more difficult ones





[Bengio et al., ICML'09]

Curriculum domain adaptation

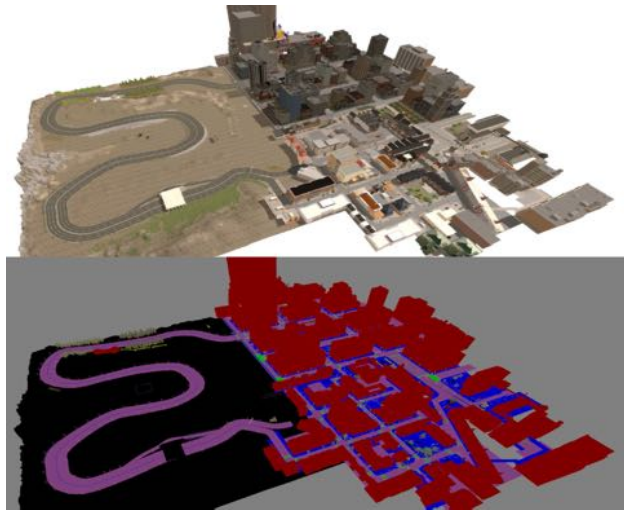
Feed a learning system "easy" **tasks** first

The solutions to them find good local optima, acting as an effective regularizer

Curriculum domain adaptation

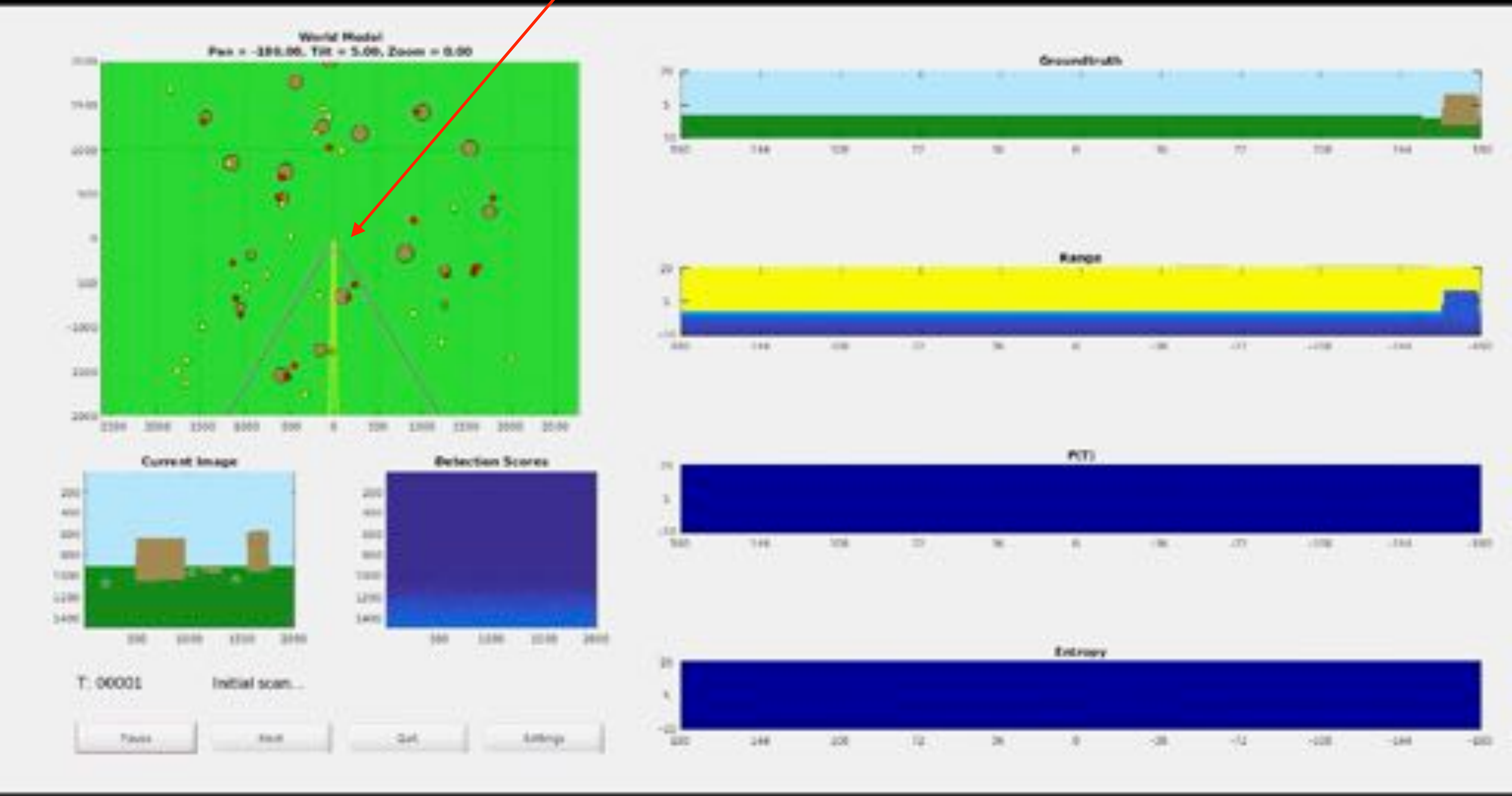
Feed a learning system "easy" **tasks** first

The solutions to them find good local optima, acting as an effective regularizer



Synthetic imagery → Real photos

An intelligent robot



Curriculum domain adaptation

About 1.5 hrs to label one such image!

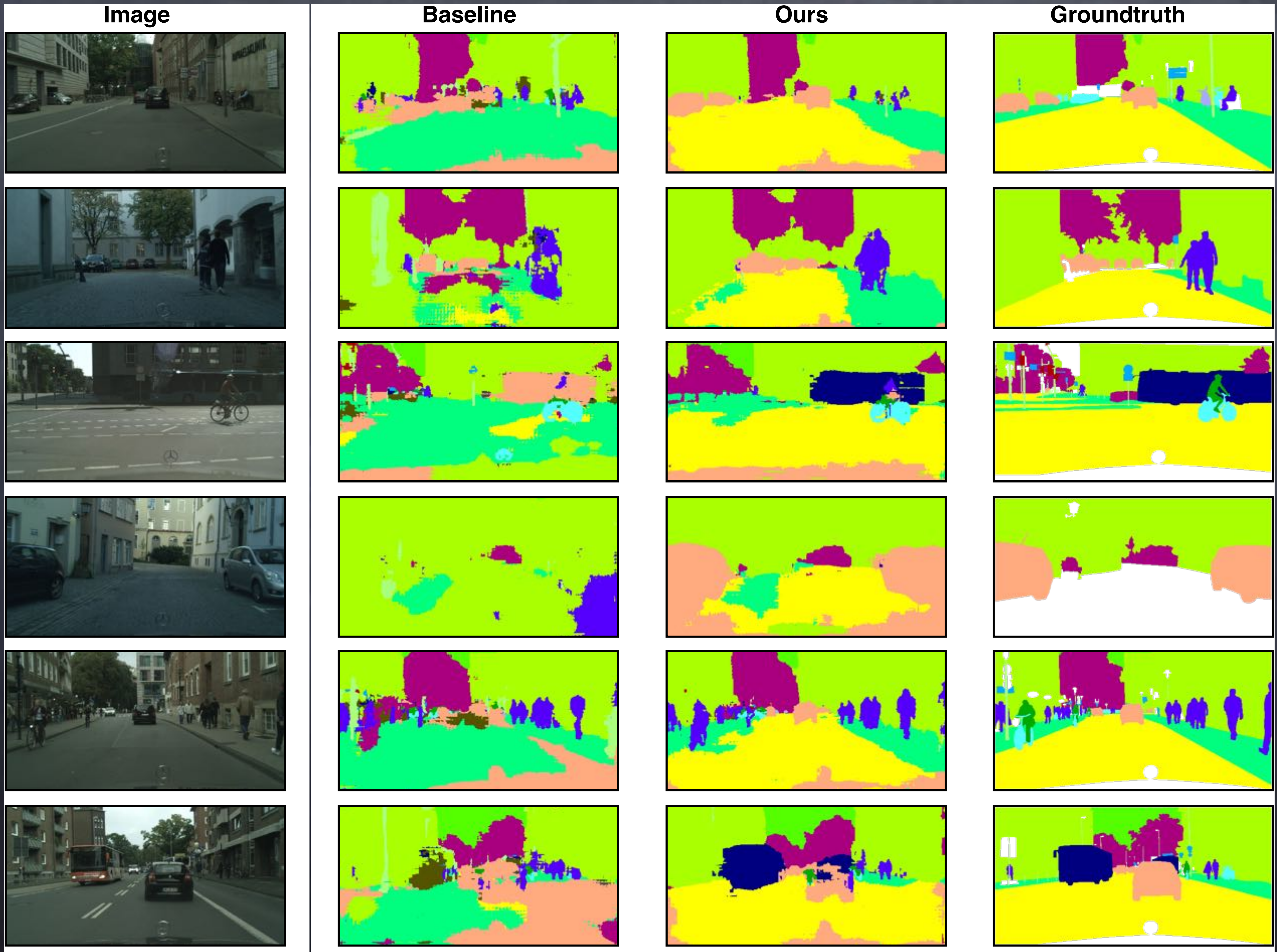
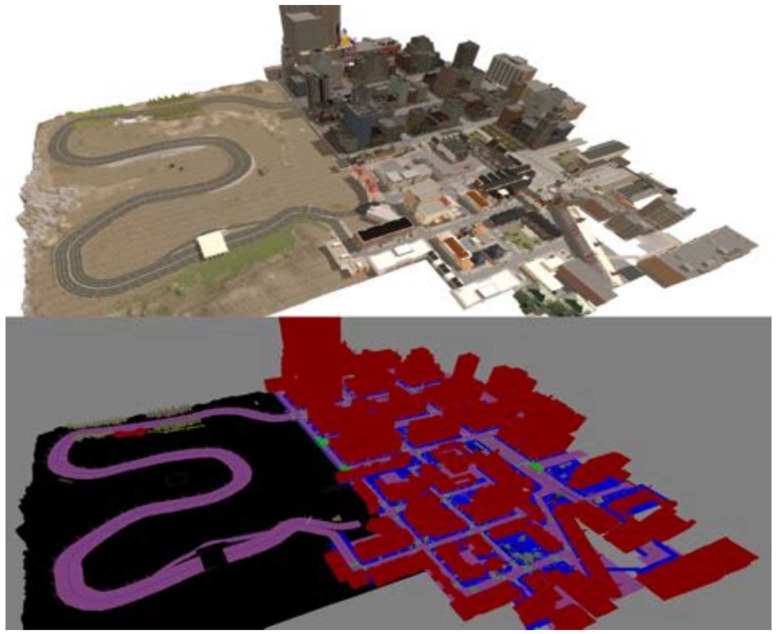
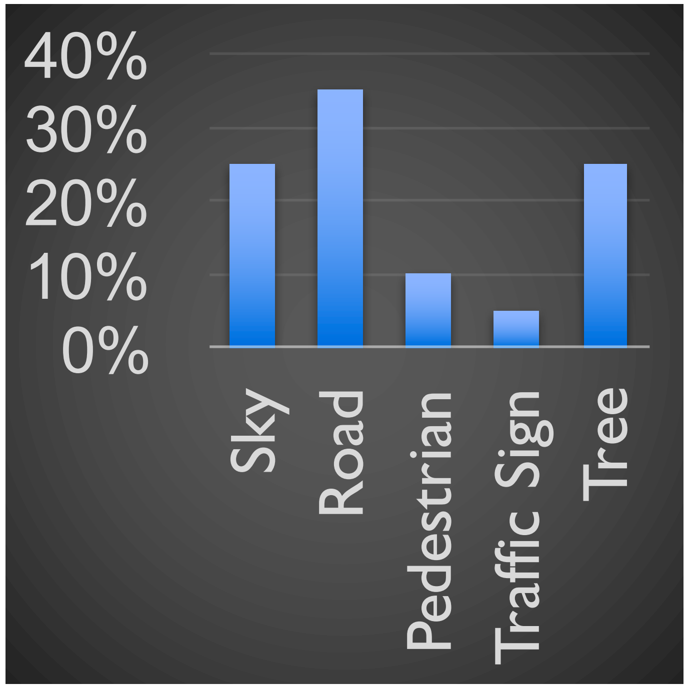


Image
Baseline
Ours
Groundtruth

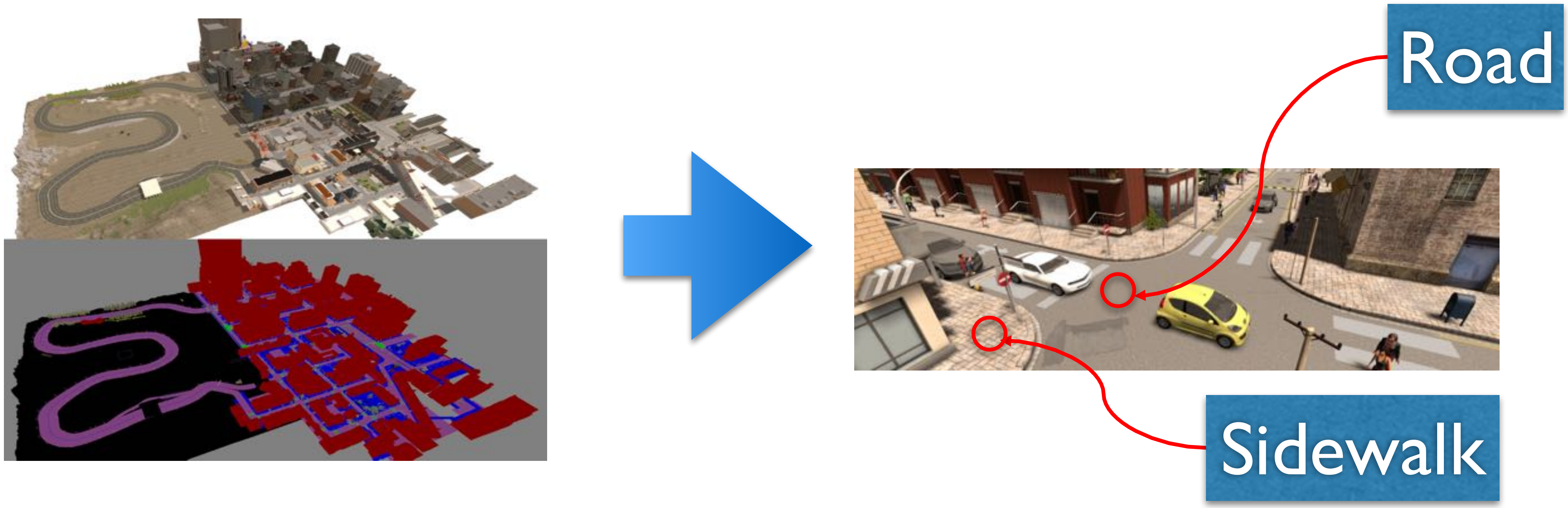
Easy task 1: predict label distributions





Input: An urban scene image
Algorithm: Logistic regression
Output: Label distributions

Easy task 2: Label landmark superpixels

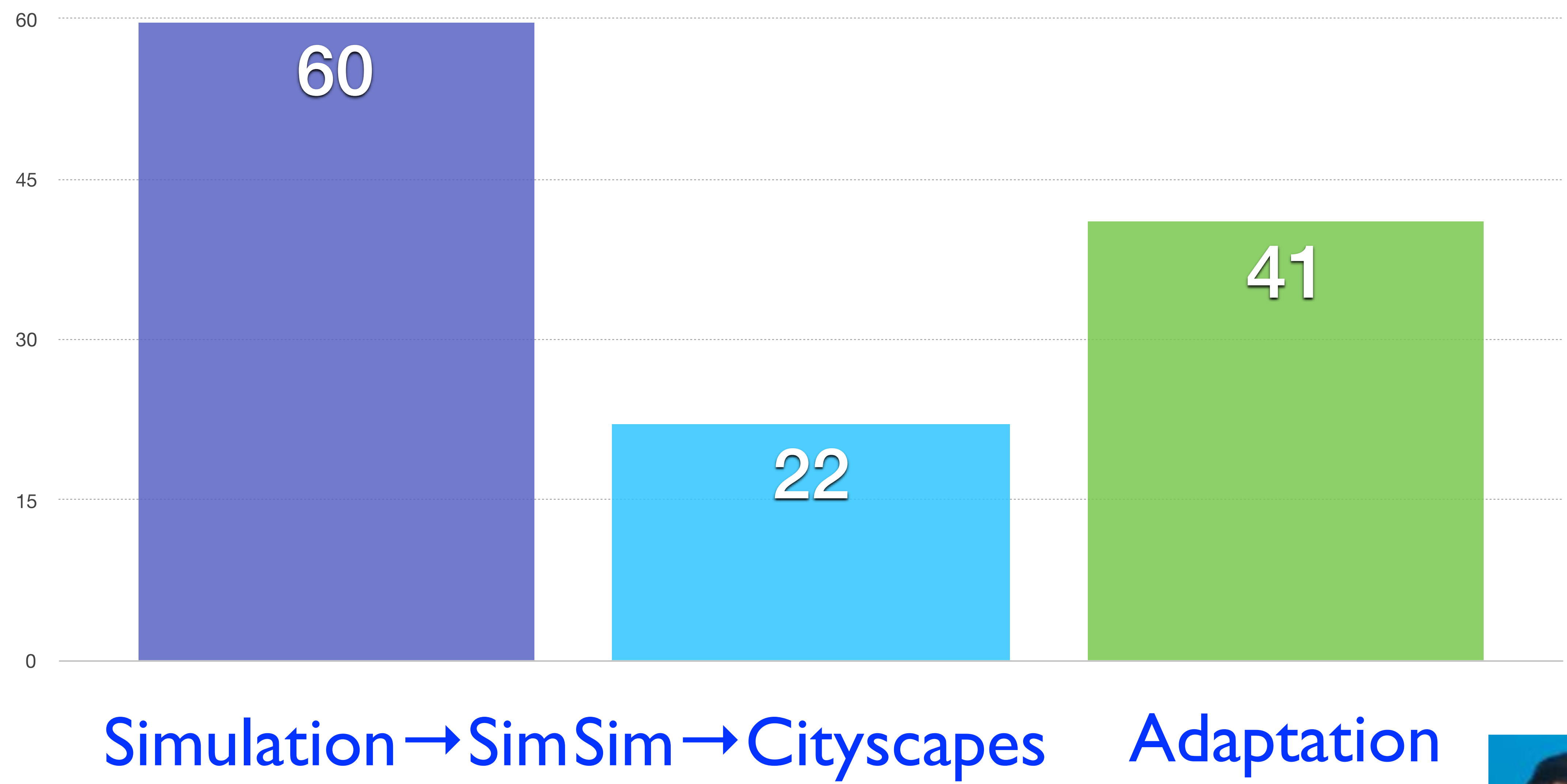


Input: An urban scene image

Algorithm: Superpixel + Logistic regression

Output: Labels of some super-pixels

Simulation → real world:
~~catastrophic~~ performance drop



[Zhang et al., ICCV'17]

Outline

Web data with noisy labels

- Hard to rectify wrong labels
- Easier to remove wrong labels

Semi-sup. Learning

WGAN

Web data with accurate labels

- 3D videos/movies

Curriculum learning / domain adaptation

Web data of multi-modalities

- Web images vs. Web videos

The perils of mismatched domains

Cause: standard assumption in machine learning

Same underlying distribution for training and testing

Consequence:

Poor cross-domain generalization

Brittle systems in dynamic and changing environment

A realistic obstacle for autonomous systems

Systems often deployed to new environments, not re-producible in house

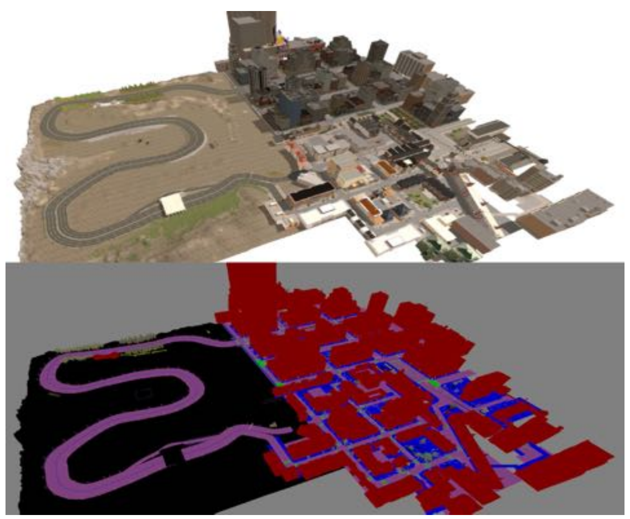
Expensive to collect training data to cover some target environments

Systems degrade over time

Environments change over time

Etc.

The perils of mismatched domains



Synthetic imagery → Real photos

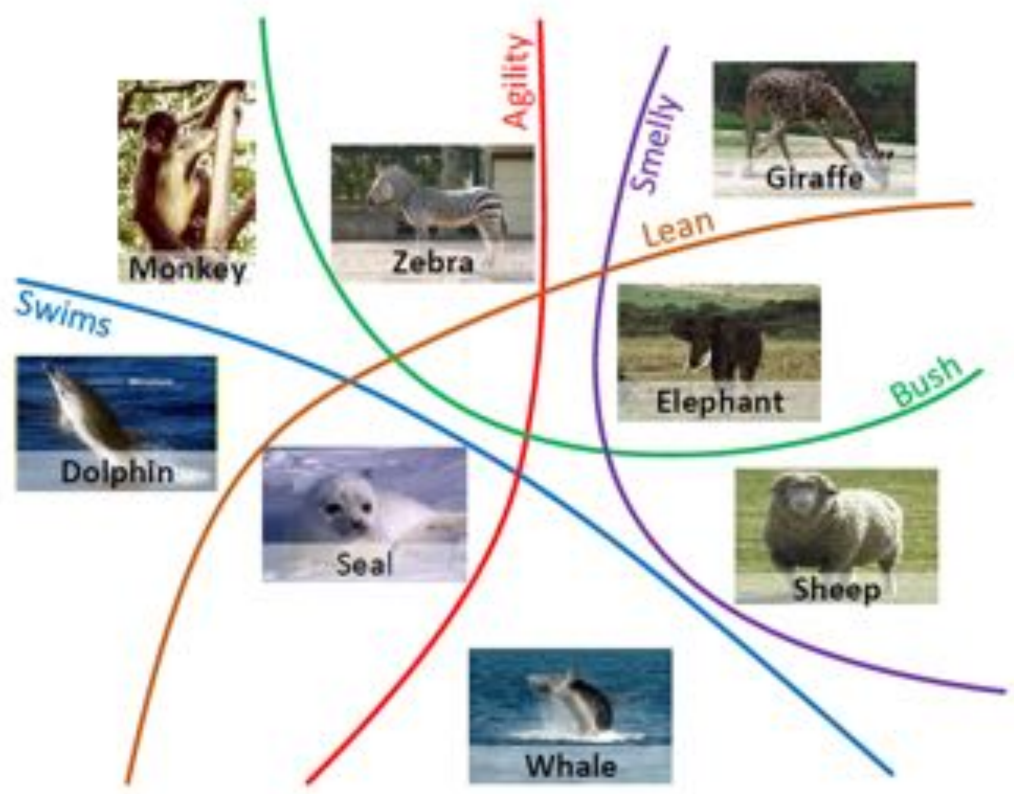
[Zhang et al., ICCV'17]

The perils of mismatched domains

Adapting face detector to a user's album

[Jamal et al., CVPR'18]

The perils of mismatched domains



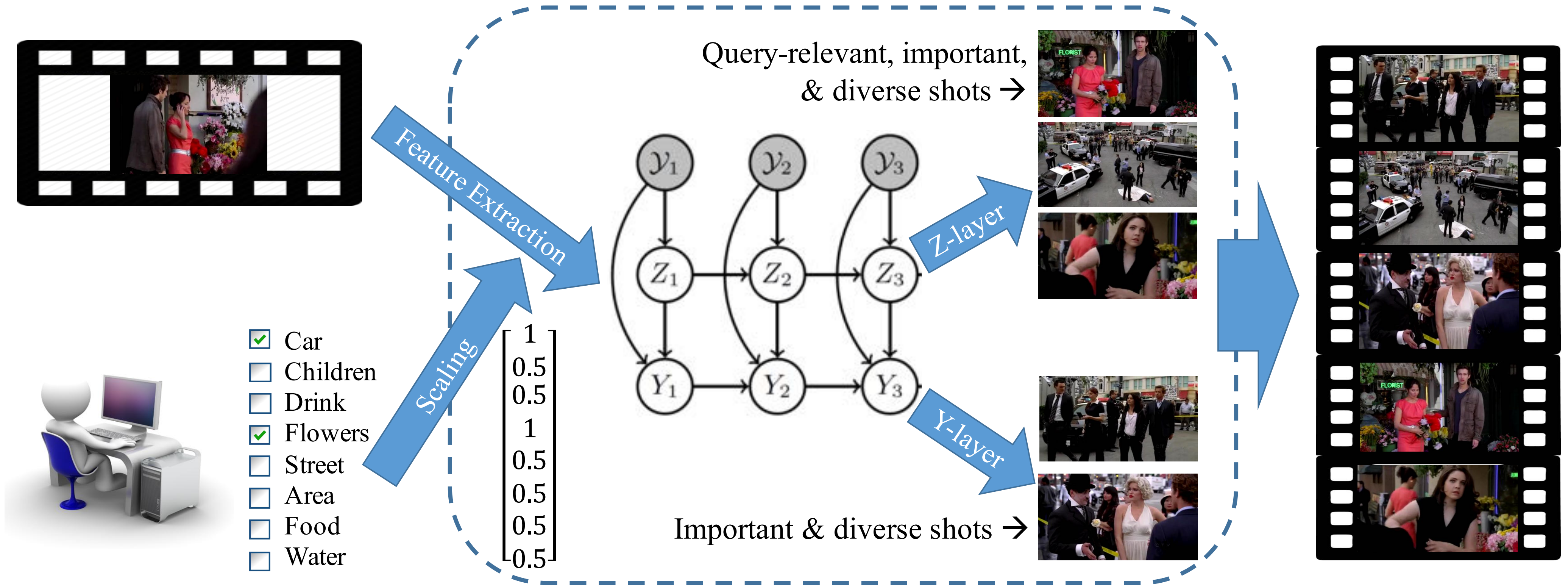
Middle-level concepts describing objects, faces, etc.

Shared by different categories

Attribute detection

[Gan et al., CVPR'17]

The perils of mismatched domains

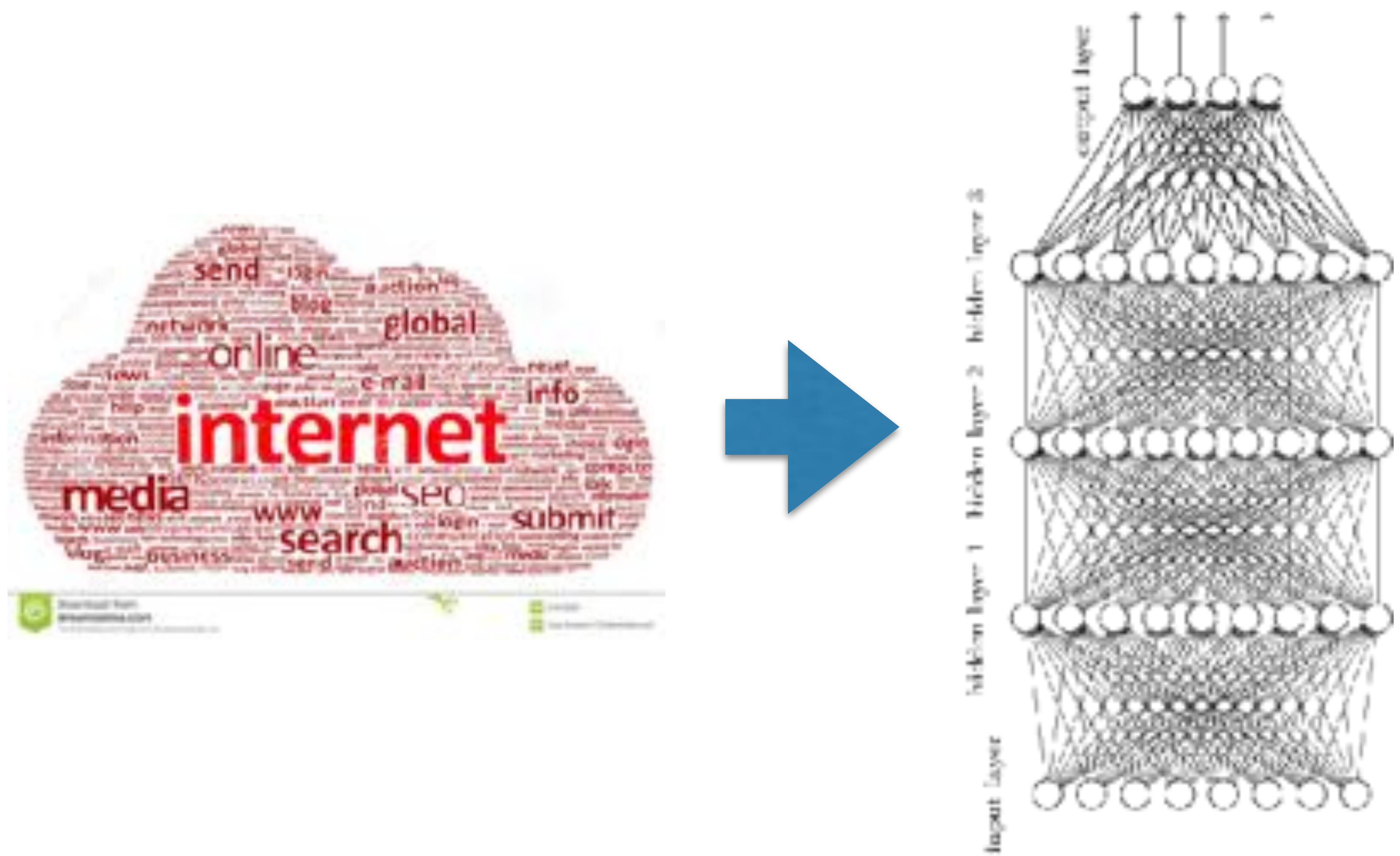


(a) Input: Video & Query **(b) Algorithm**: Sequential & Hierarchical Determinantal Point Process (SH-DPP) **(c) Output**: Summary

Personalization of video summarizers

[Sharghi et al., ECCV'16, CVPR'17, ECCV'18]

The perils of mismatched domains



Webly supervised learning

Abstract form: *unsupervised* domain adaptation (DA)

Source

$$D_{\mathcal{S}} = \{(x_m, y_m)\}_{m=1}^{\mathsf{M}} \sim \boxed{P_{\mathcal{S}}(X, Y)}$$

Target

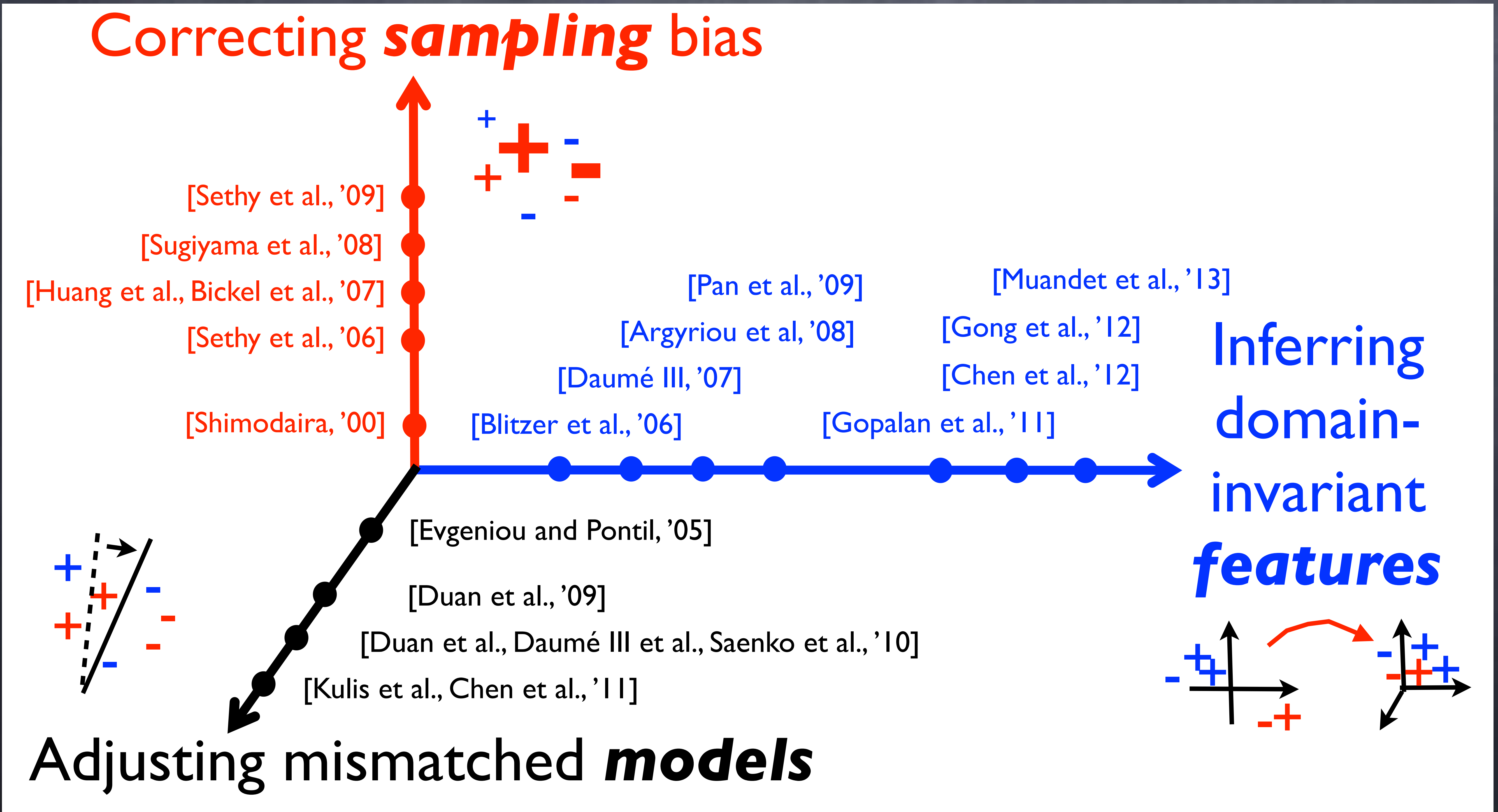
$$D_{\mathcal{T}} = \{(x_n, \ ? \ \}_{n=1}^{\mathsf{N}} \sim \boxed{P_{\mathcal{T}}(X, Y)}$$

Different distributions

Objective

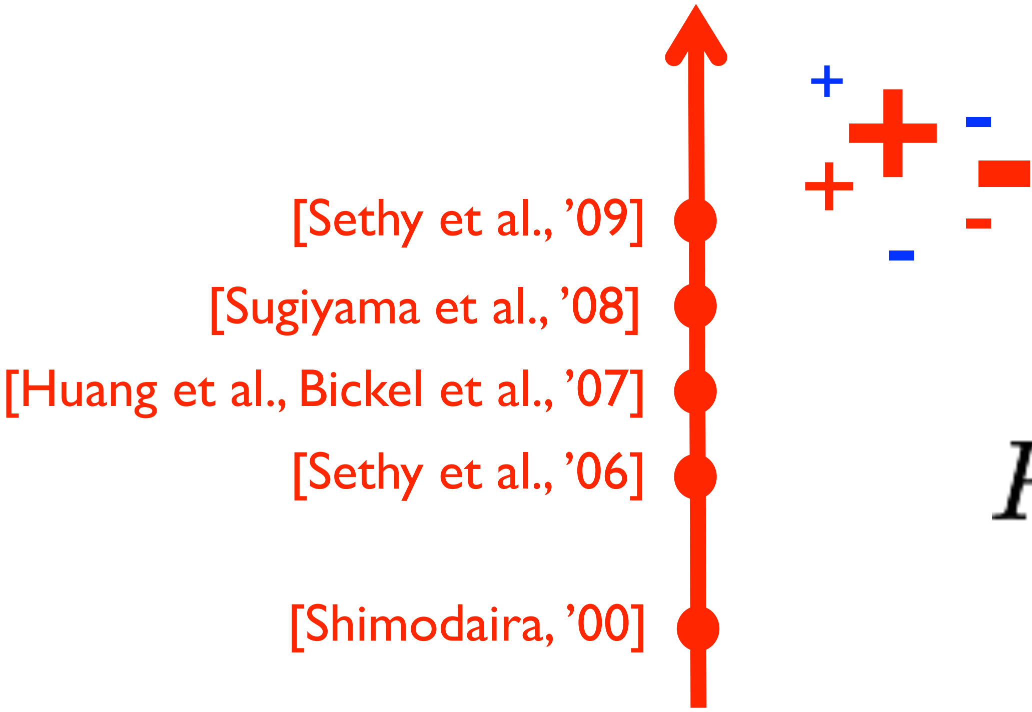
Learn models to work well on target

Popular methods



Data selection for DA

Correcting ***sampling*** bias



$$P_{\mathcal{L}}(\text{landmarks}) \approx P_{\mathcal{T}}(\text{target})$$

$$\min_{\text{landmarks}} \quad d(P_{\mathcal{L}}, P_{\mathcal{T}})$$

Data selection for DA

Landmarks are labeled source instances distributed similarly to the target domain.

Source

Target

[Gong et al., ICML'13]

Data selection for DA

Landmarks are labeled source instances distributed similarly to the target domain.

Source

Identifying landmarks:

$$P_{\mathcal{L}}(\text{landmarks}) \approx P_{\mathcal{T}}(\text{target})$$

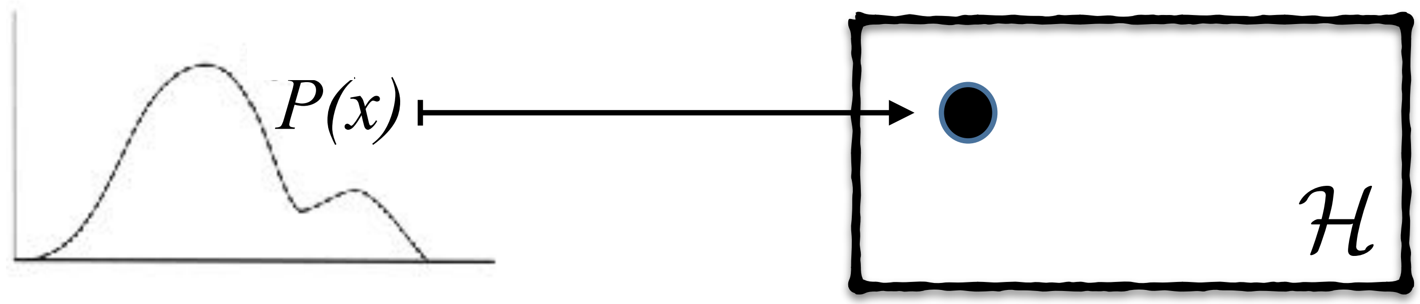
$$\min_{\text{landmarks}} \quad d(P_{\mathcal{L}}, P_{\mathcal{T}})$$

[Gong et al., ICML'13]

Target

Kernel mean embedding of distributions

$$\mu[P] \triangleq \mathbb{E}_x[\phi(x)]$$



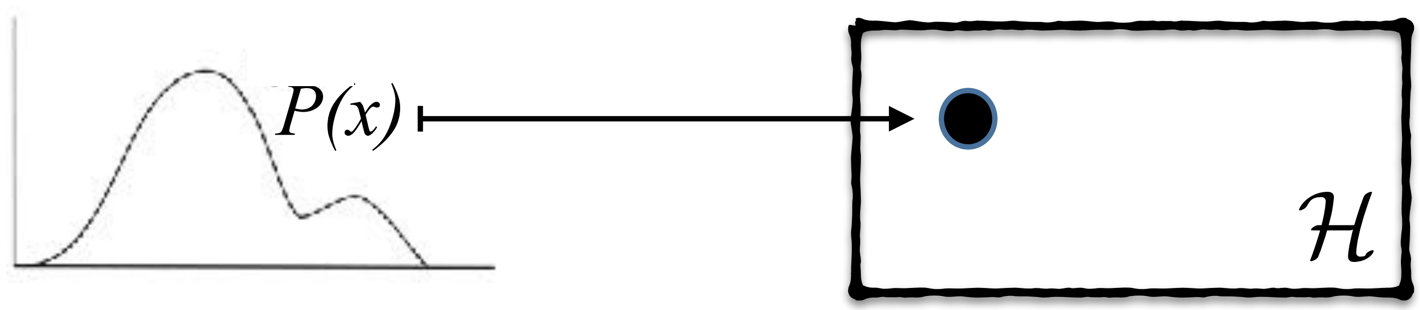
μ maps distribution P to Reproducing Kernel Hilbert Space

μ is injective if $\phi(\cdot)$ is characteristic

[Müller'97,Gretton et al.'07,Sriperumbudur et al.'10]

Kernel mean embedding of distributions

$$\mu[P] \triangleq \mathbb{E}_x[\phi(x)]$$



Empirical kernel embedding:

$$\hat{\mu}[P] = \frac{1}{\mathsf{n}} \sum_{i=1}^{\mathsf{n}} \phi(x_i), \quad x_i \sim P$$

[Müller'97,Gretton et al.'07,Sriperumbudur et al.'10]

Identifying landmarks by matching kernel means

Integer programming

$$\min_{\{\alpha_m\}} \left\| \frac{1}{\sum_i \alpha_i} \sum_{m=1}^{\mathsf{M}} \alpha_m \phi(x_m) - \frac{1}{\mathsf{N}} \sum_{n=1}^{\mathsf{N}} \phi(x_n) \right\|_{\mathcal{H}}^2$$

where

$$\alpha_m = \begin{cases} 1 & \text{if } x_m \text{ is a landmark wrt target} \\ 0 & \text{else} \end{cases}$$

$$m = 1, 2, \cdots, \mathsf{M}$$

Other details

Convex relaxation

Recovering $\alpha_m^\star$ from $\beta_m^\star (= \frac{\alpha_m}{\sum_i \alpha_i})$

Multi-scale analysis

Class balance constraint

How landmarks look like?

Summary

Landmarks

[Gong *et al.*, ICML'13]

- *Labeled source instances, distributed similarly to target*
- *Better approximation of discriminative loss of target*
- *Automatically identifying landmarks*
- *Benefiting other adaptation methods*

Outline

Web data with noisy labels

- Hard to rectify wrong labels
- Easier to remove wrong labels

Semi-sup. Learning

WGAN

Web data with accurate labels

- 3D videos/movies

Curriculum learning / domain adaptation

Web data of multi-modalities

- Web images vs. Web videos

Web videos are often redundant, *sometimes misleading*

Bench Press

Pizza Tossing

Web images are informative for activity detection, *and noisy*

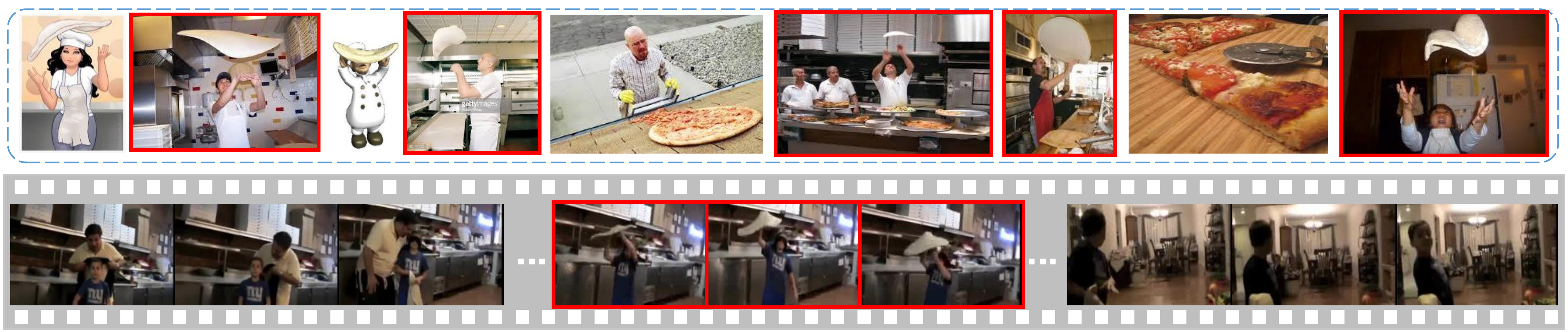
Bench Press

Pizza Tossing

Pruning by mutually voting

Query-relevant Web images and video frames *are alike;*

An irrelevant Web image or video frame *is irrelevant in its own way***.**



(c) Pizza Tossing

Pruning by mutually voting

Query-relevant Web images and video frames *are alike*;

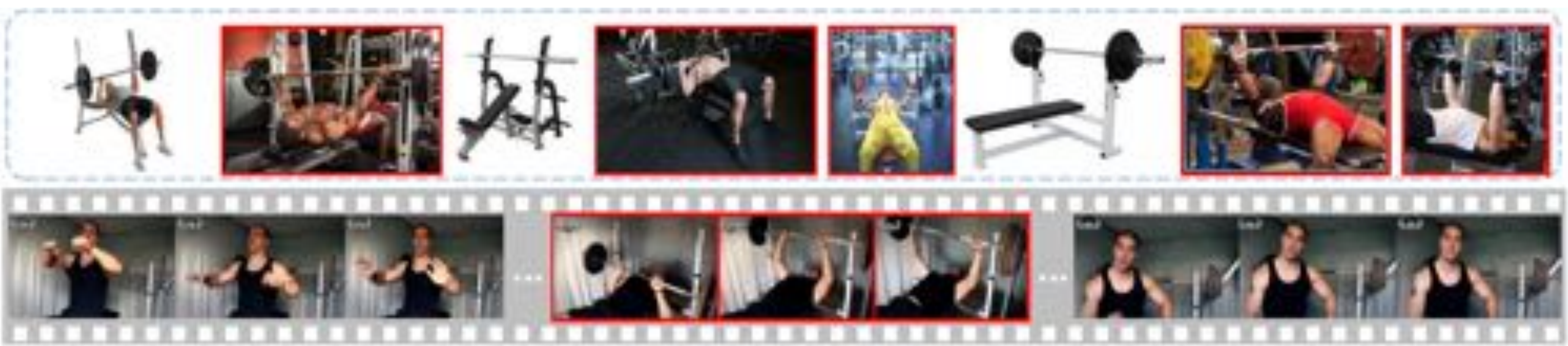
An irrelevant Web image or video frame *is irrelevant in its own way*.

(a) Basketbal Dunk

Pruning by mutually voting

Query-relevant Web images and video frames *are alike;*

An irrelevant Web image or video frame *is irrelevant in its own way*.



(b) Bench Press

Mutually vote by matching kernel means

$$\min_{\alpha,\beta\in\{0,1\}} \left\| \frac{1}{\sum_m \alpha_m} \sum_{m'} \underbrace{\boxed{\alpha_{m'}\phi(I_m)}}_{\text{Landmark images}} - \frac{1}{\sum_n \beta_n} \sum_{m'} \overbrace{\boxed{\beta_{m'}\phi(F_m)}}^{\text{Landmark video frames}} \right\| + \mathcal{R}(\beta)$$

$$\alpha_m = \begin{cases} 1 & \text{if } I_m \text{ is similar to selected video frames} \\ 0 & \text{else} \end{cases}$$

$$\mathcal{R}(\beta) = \text{Reconstruct video from the selected video frames}$$

Experimental results on UCF101

Table 1: Comparison results on UCF101.

Method	Accuracy (%)
Karpathy et al. [20]	65.4
LRCN [7]	71.1
Spatial stream net. [29]	73.0
Ours	69.3

Sophisticated models learned from ***manually pruned and labeled*** training videos.

SVM trained from ***mutually pruned Google labeled*** Web images & Web videos.

Experimental results on UCF101

Table 1: Comparison results on UCF101.

Method	Accuracy (%)
Karpathy et al. [20]	65.4
LRCN [7]	71.1
Spatial stream net. [29]	73.0
LSTM composite [34]	75.8
C3D [40]	82.3
IDT + FV [41]	87.9
Ours	69.3

Sophisticated model learned from ***manually*** *pruned and labeled* training videos.

Motion, or temporal features

SVM trained from ***mutually pruned*** *Google labeled* Web images & Web videos.

Web for visual recognition

Web data with noisy labels

- Hard to rectify wrong labels
- Easier to remove wrong labels

Semi-sup. Learning

WGAN

Web data with accurate labels

- 3D videos/movies

Curriculum learning / domain adaptation

Web data of multi-modalities

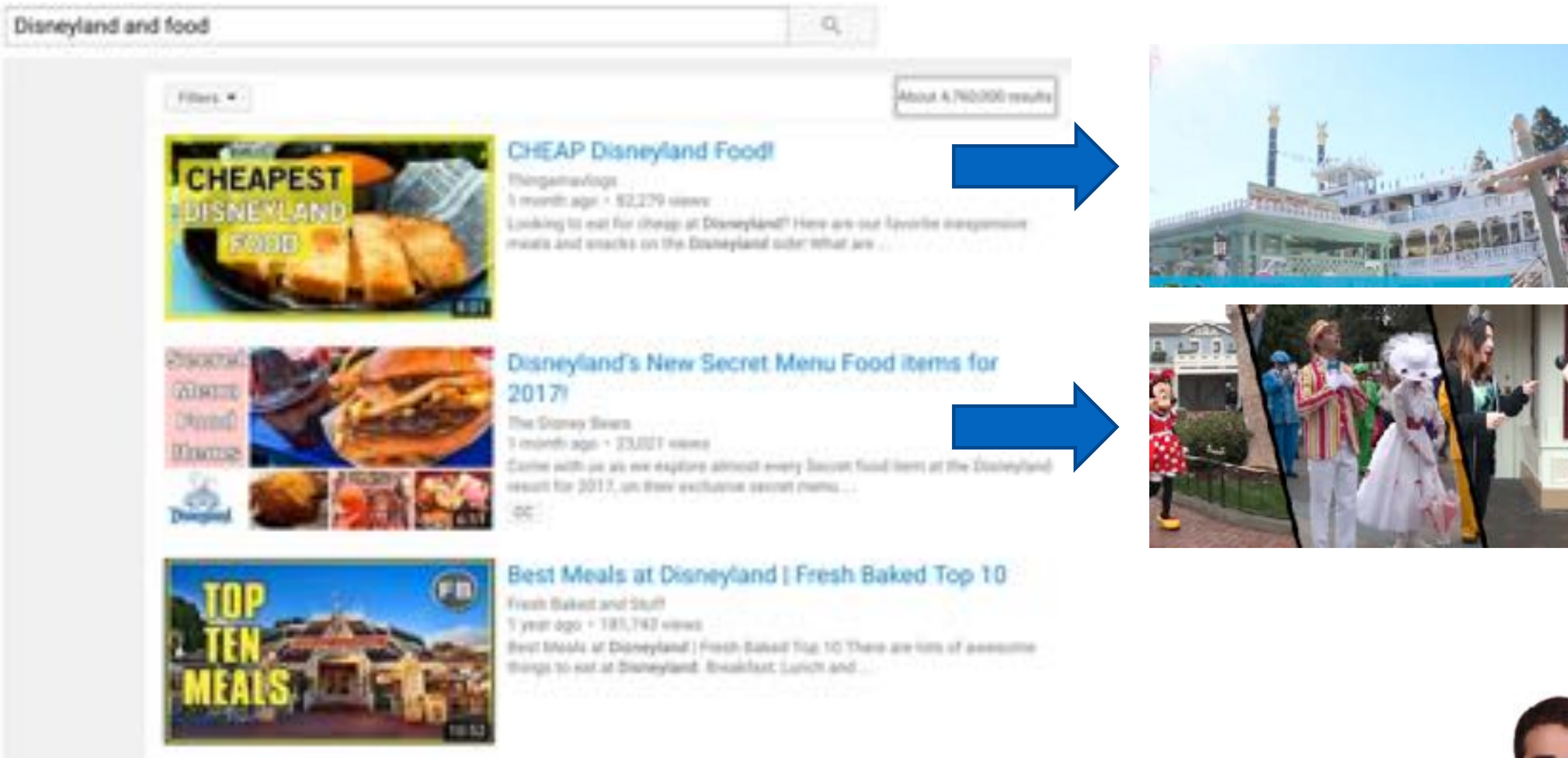
- Web images vs. Web videos

Kernel mean matching

Web for visual recognition

Web for supervised video summarization

Query-focused supervised video summarization



[Sharghi et al., ECCV'16, CVPR'17, ECCV'18]

Web for visual recognition

Web for supervised video summarization

1. $\nabla \cdot \mathbf{D} = \rho_V$
2. $\nabla \cdot \mathbf{B} = 0$
3. $\nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}$
4. $\nabla \times \mathbf{H} = \dfrac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$

Web for X (VQA, 3D reconstruction, etc.)

[1] A Semi-Supervised Two-Stage Approach to Learning from Noisy Labels. Y Ding, L Wang, D Fan, & **B Gong**. WACV 2018.

[2] Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect. X Wei*, **B Gong***, Z Liu, & L Wang. ICLR 2018.

[3] Geometry-Guided CNN for Self-Supervised Video Representation Learning. C Gan, **B Gong**, K Liu, H Su, & L Guibas. CVPR 2018.

[4] Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. Y Zhang, P David, & **B Gong**. ICCV 2017.

[5] Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation. **B Gong**, K Grauman, & F Sha. ICML 2013.

[6] Webly-supervised Video Recognition by Mutually Voting for Relevant Web Images and Web Video Frames. C Gan, C Sun, L Duan, & **B Gong**. ECCV 2016.

[7] Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach. A. Sharghi, J Laurel, & **B Gong**. CVPR 2017.