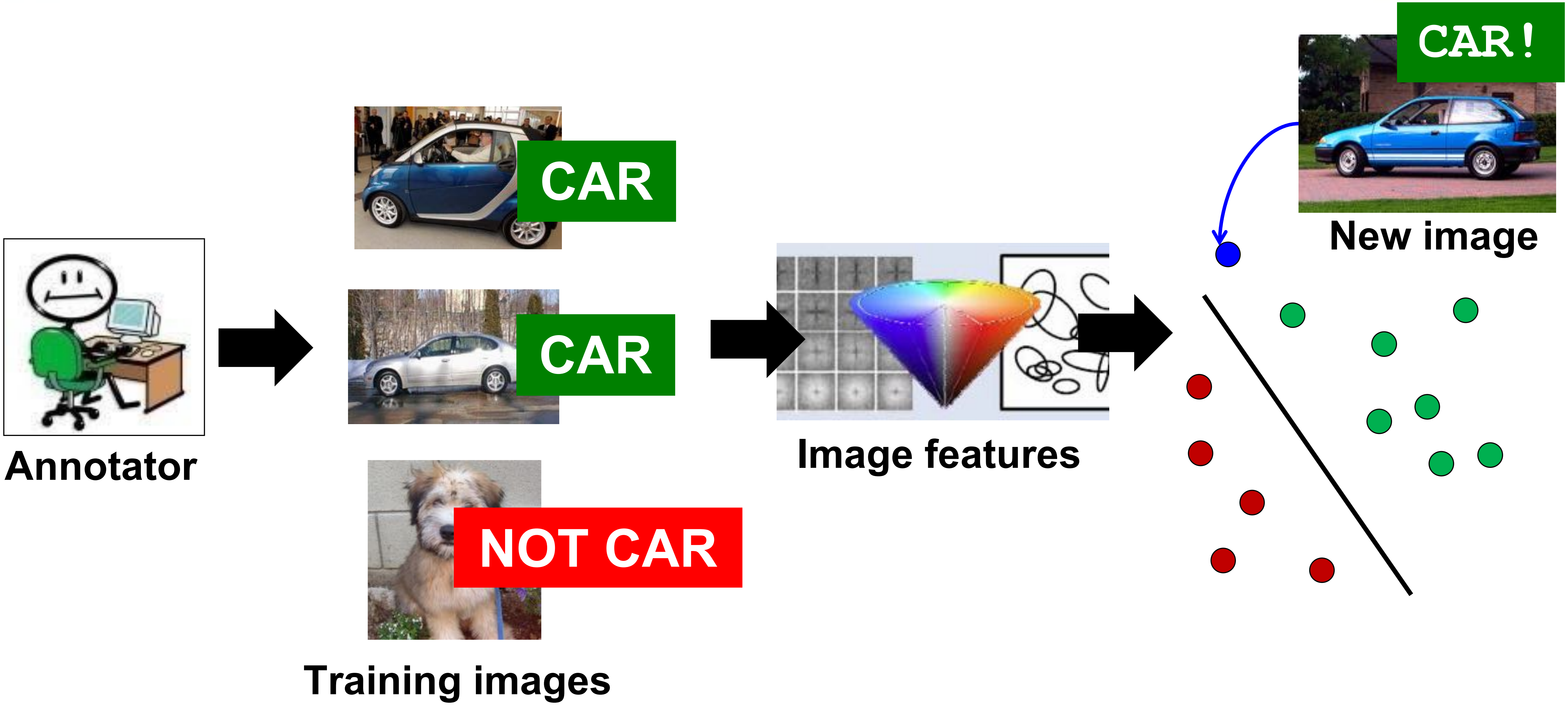
Learning from Web Data and Adapting beyond It

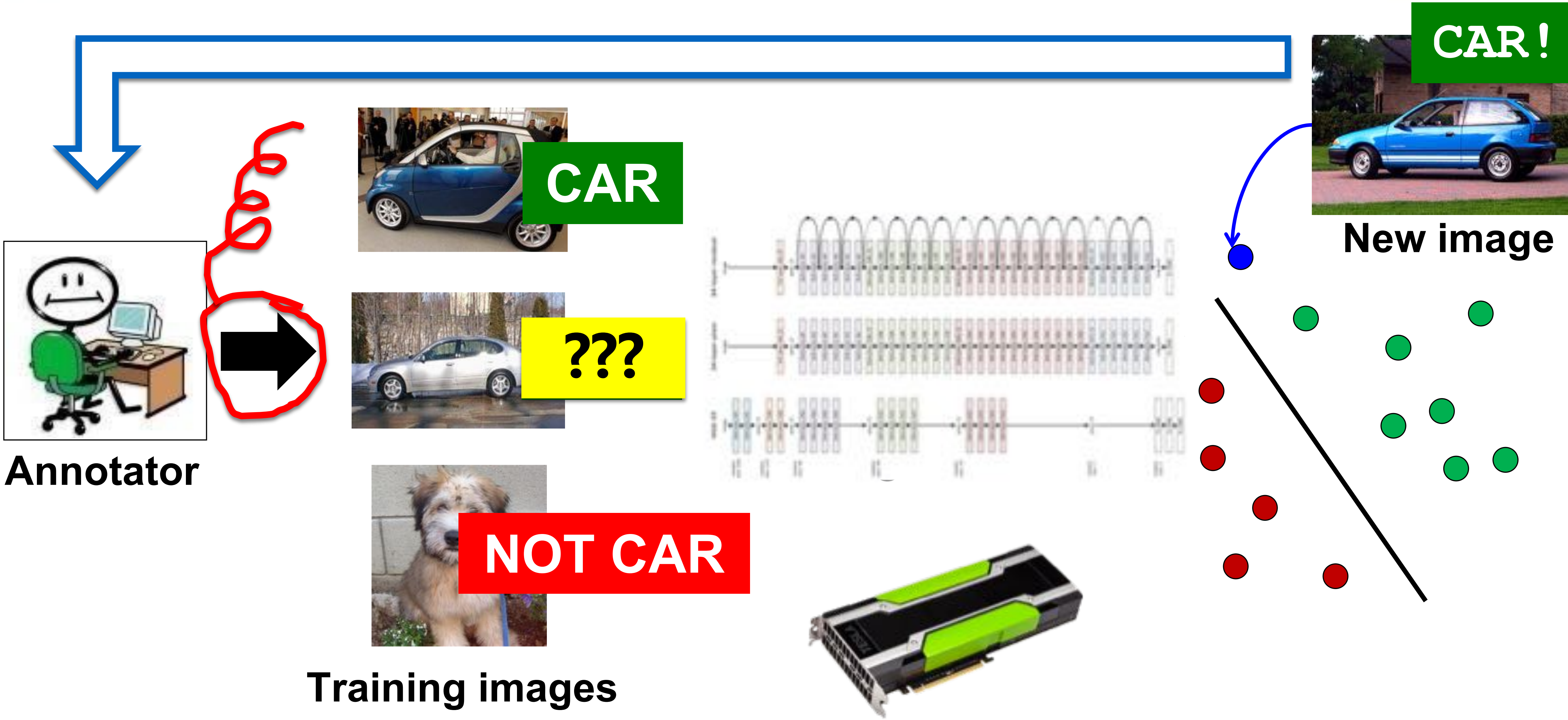
Boqing Gong
BoqingGo@outlook.com

Learning based visual recognition



Courtesy K. Grauman

Learning based visual recognition



1. Web data with noisy labels

➔ Need different training techniques

Courtesy K. Grauman

Label correction & re-weighting

Label Correction

cat:

Re-weigh labels/data terms

jacket:

Label ~~correction & re-weighting~~ removal

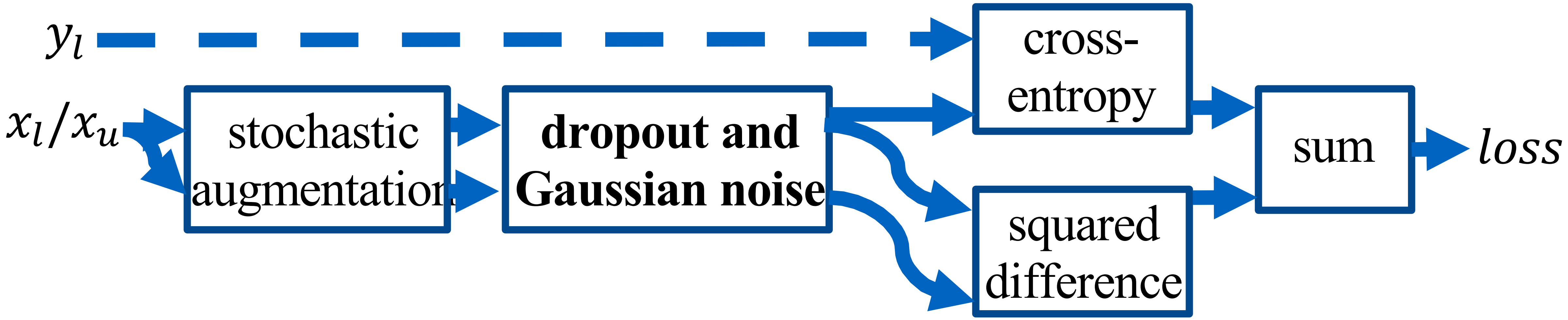
Label Correction

Hard to rectify wrong labels
Easier to just remove wrong labels

Semi-supervised learning?
Caveat: outlier images

A consistent term & its dual effect

[Laine & Aila, ICLR 2017]



Outlier still helps!

Noisy labels, no outlier

Result on CIFAR-10 and 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]

Noisy labels, & outlier images

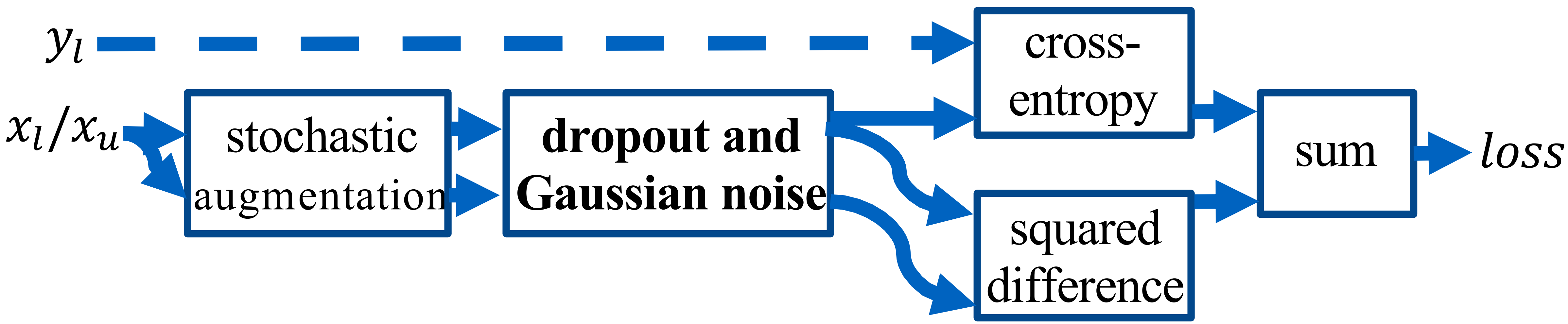
Results on Clothing1M

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]

Detour: a consistent term & its dual effect



$$\frac{d_Y(f(x_1), f(x_2))}{d_X(x_1, x_2)} \leq K.$$

Augment the same example twice

➔ Two data points around that example

➔ Lipschitz continuity in Wasserstein GAN

[Xiang*, Gong*, et al., ICLR 2018]

Outline

Web data with noisy labels

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

Semi-supervised learning

Web data with accurate labels

- 3D movies

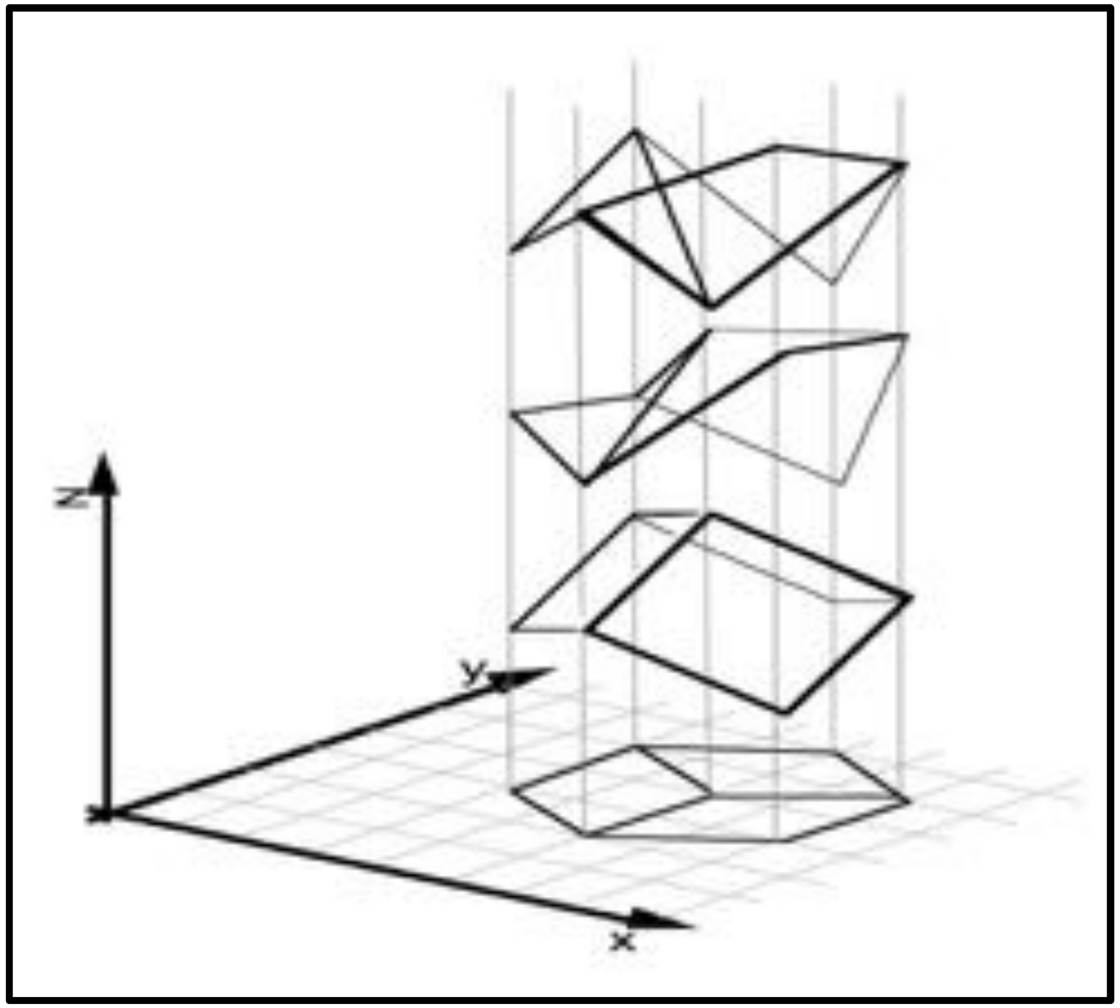
Web data of multi-modalities

- Web images vs. Web videos

3D movies

Geometry & semantics

[Snavely et al, CVPR '06]



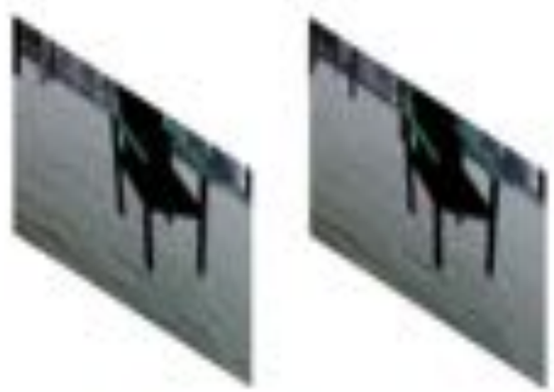
[Sinha et al, ICCV'93]

Shape from dense views
geometric problem

Shape from one view
semantic problem

Courtesy K. Grauman & D. Jayaraman

3D movies



Training on Flying chairs

Start with synthetic imagery and precise geometry cues

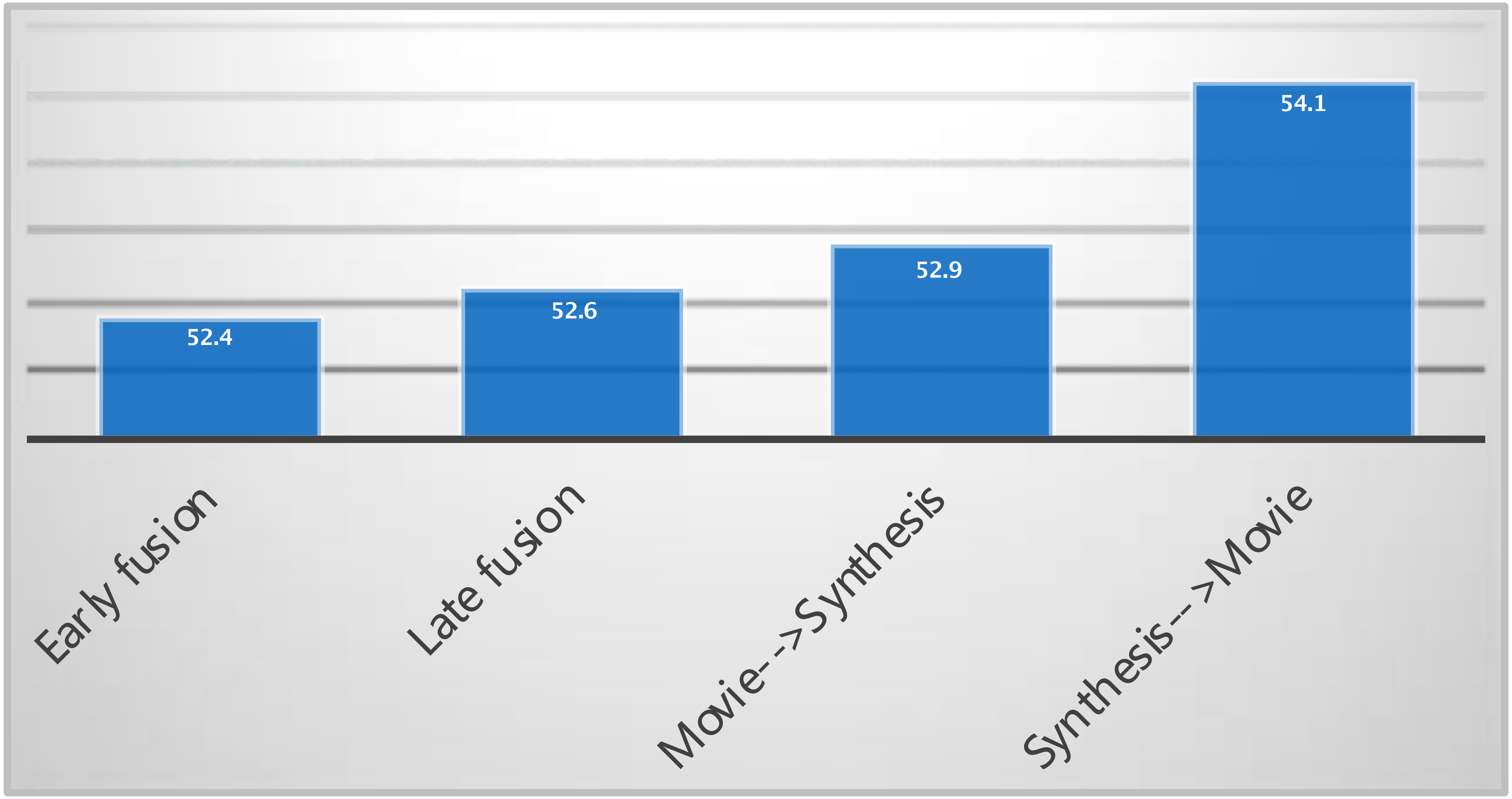
Training on 3D movies

Followed by 3D movies to Incorporate reality cues

Curriculum

Results on UCF101

It is important to follow the right curriculum!

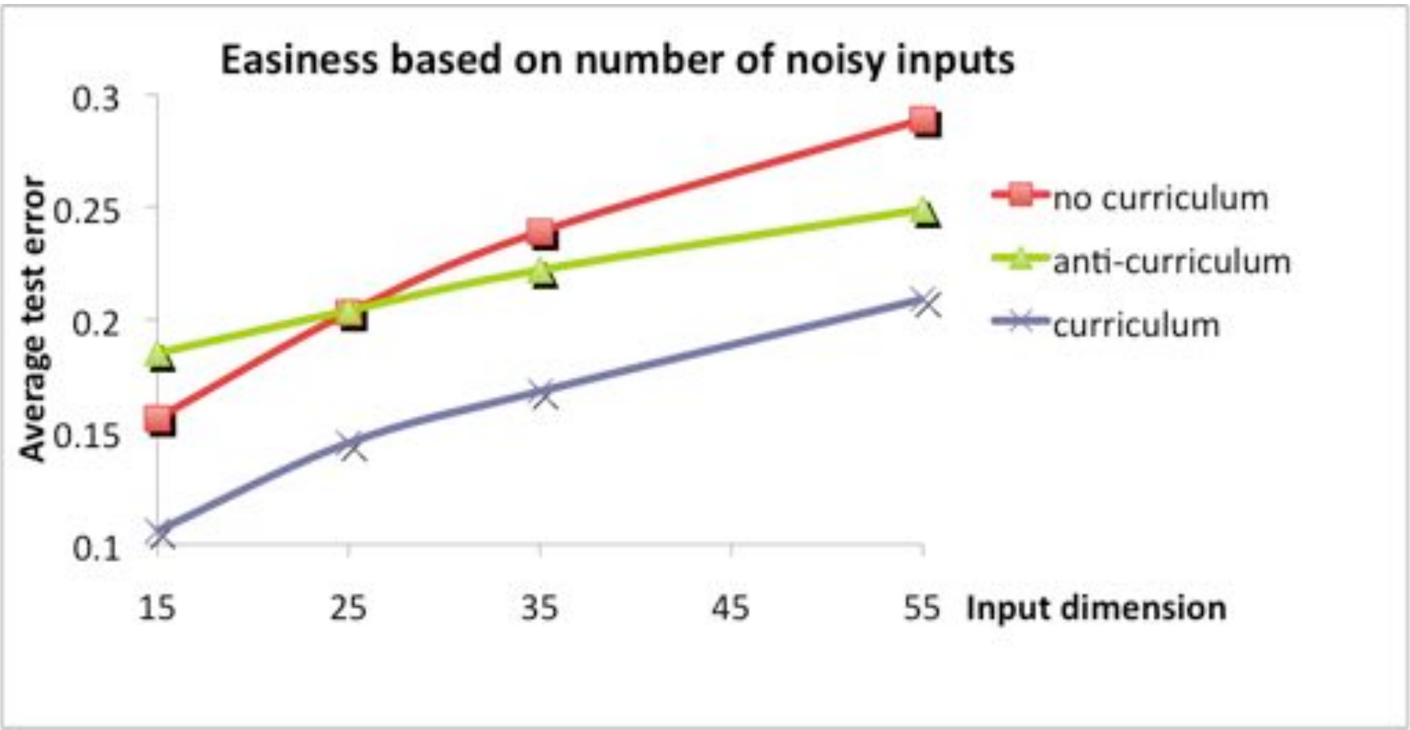


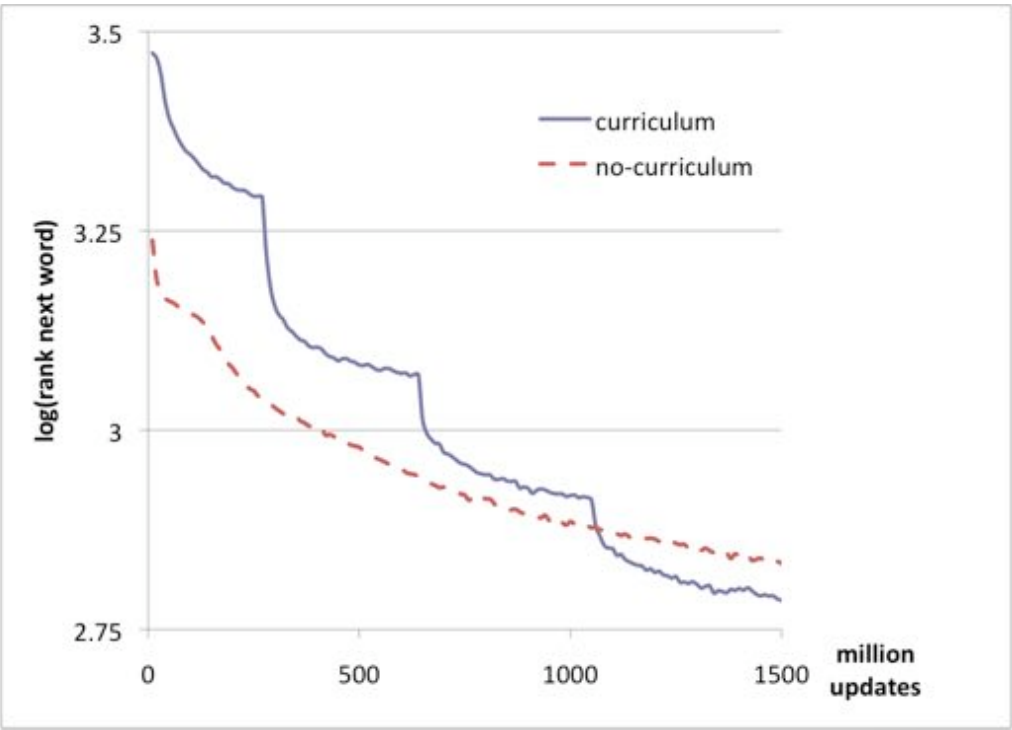
[Gan et al., CVPR’18]

Detour: curriculum learning

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

[Bengio et al., ICML'09]

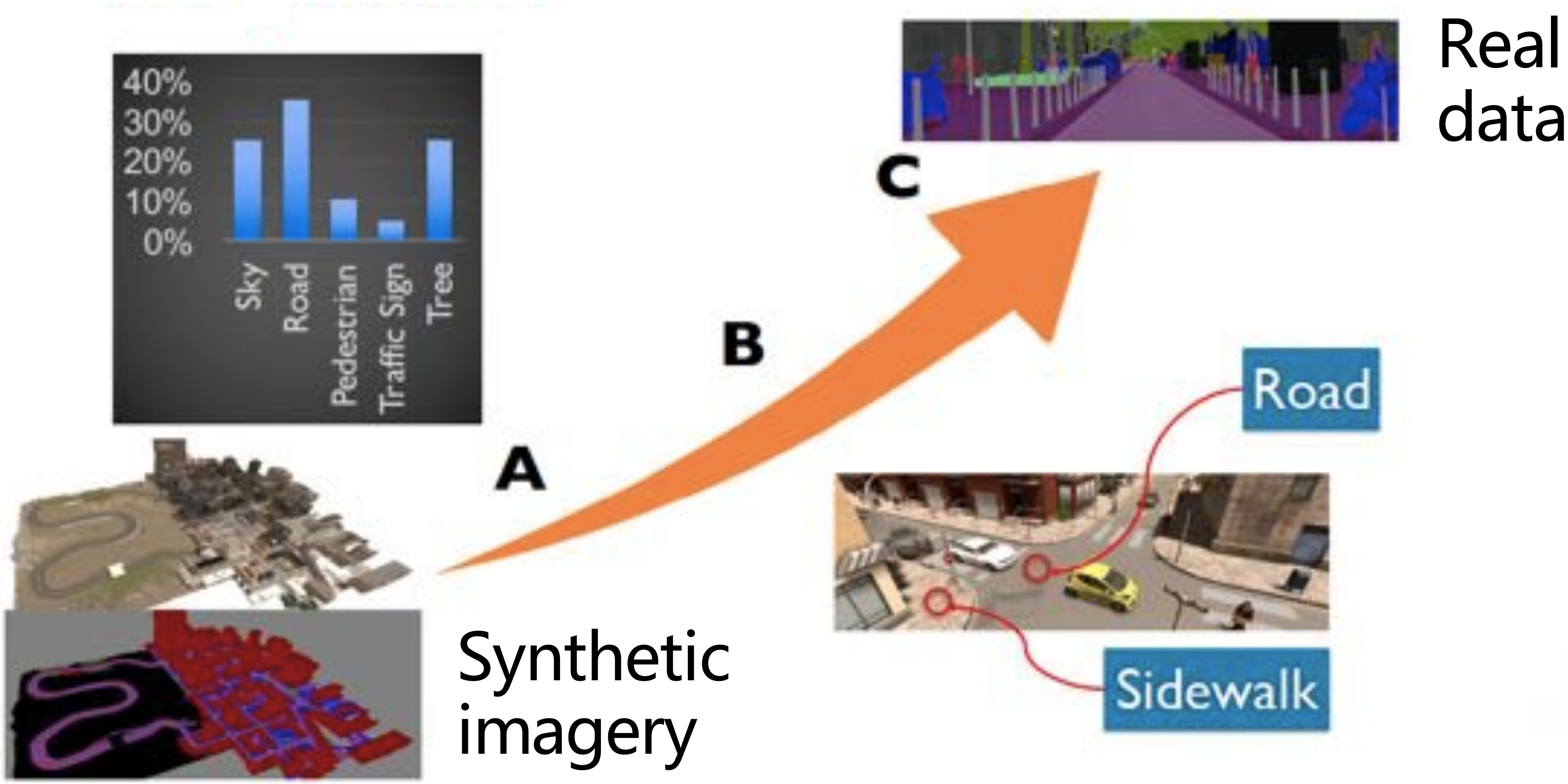




Detour: curriculum domain adaptation

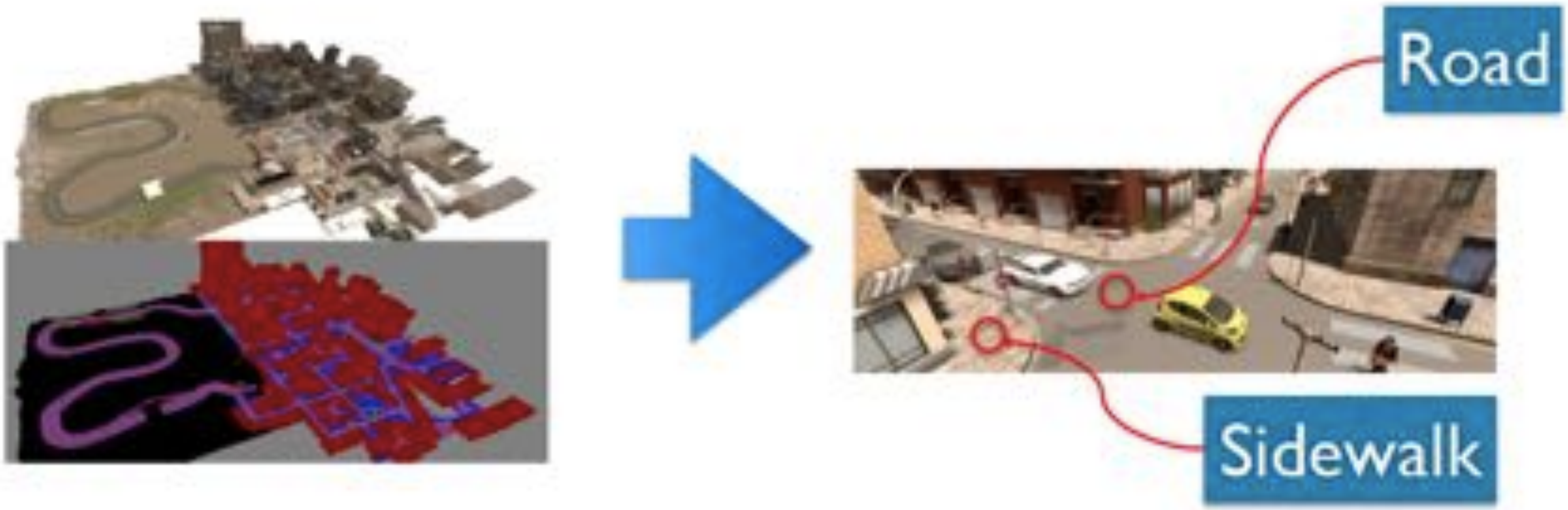
Feed a learning system "easy" tasks first

Their solutions find better local optima,
and act as a regularizer, i.e., focusing on the test



Detour: curriculum domain adaptation

Feed a learning system "easy" tasks first
Their solutions find better local optima,
and act as a regularizer, i.e., focusing on the test

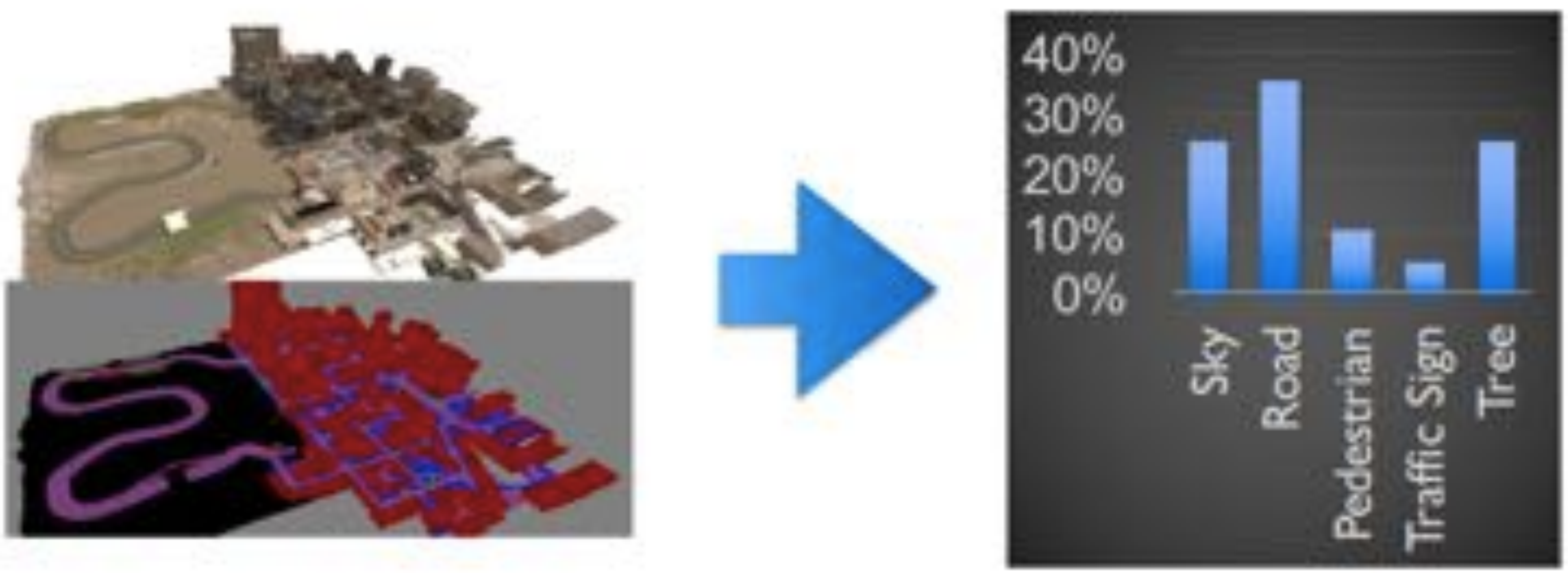


Input: An urban scene image
Algorithm: Super-pixel + Logistic regression
Output: Labels of **some** super-pixels

Detour: curriculum domain adaptation

Feed a learning system "easy" tasks first

Their solutions find better local optima,
and act as a regularizer, i.e., focusing on the test



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

Detour: curriculum domain adaptation

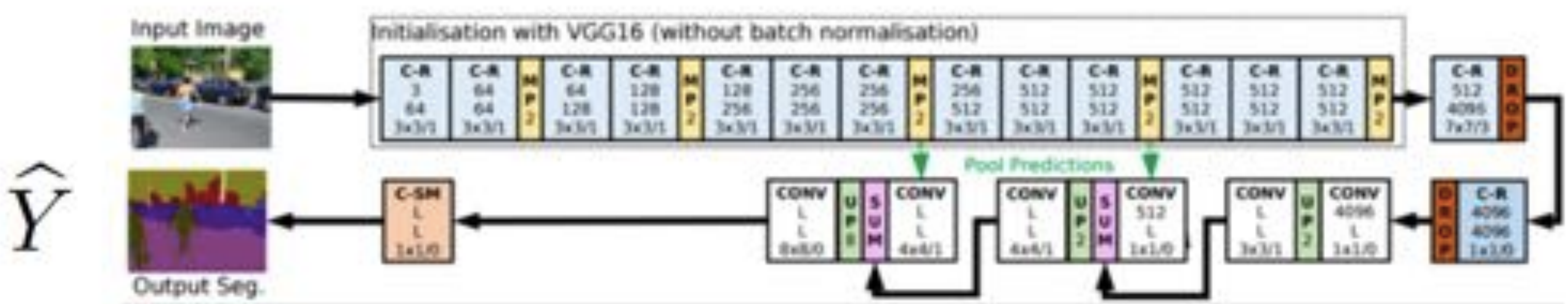
Feed a learning system "easy" tasks first

Their solutions find better local optima,
and act as a regularizer, i.e., focusing on the test

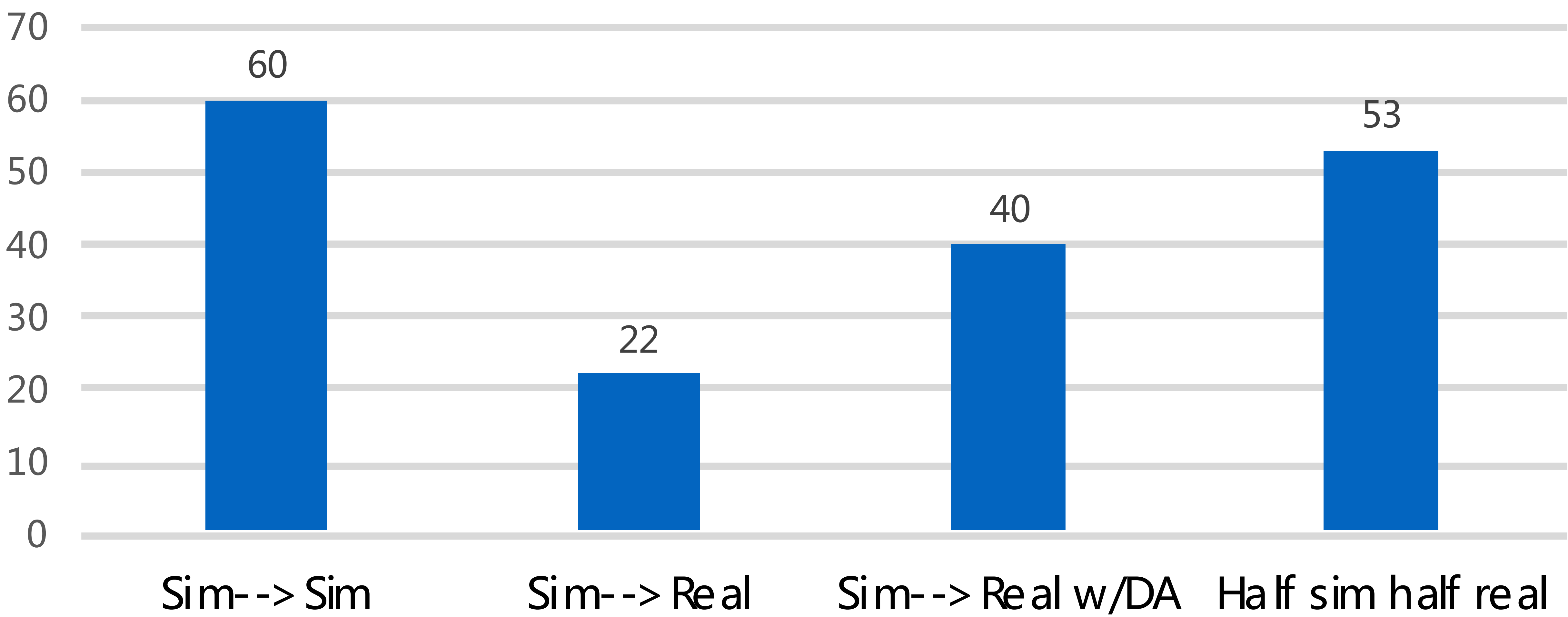
$$\min_{\Theta} \mathcal{L}(Y_s, \widehat{Y}_s) + d(p_t, p_t(\widehat{Y}_t))$$

s : Source, t : Target

p_t : Perturbation function



Detour: curriculum domain adaptation



[Yang et al., ICCV'17]

Outline

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Semi-supervised learning

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Curriculum learning & curriculum adaptation

Web data of multi-modalities
- Web images vs. Web videos

A comment on self-supervised learning

Geometry Guided Convolutional Neural Networks for ~~Self-Supervised~~ Video Representation ~~Learning~~

Self-supervised learning??

Supervised learning from self-labeled data

[Gan et al., CVPR'18]

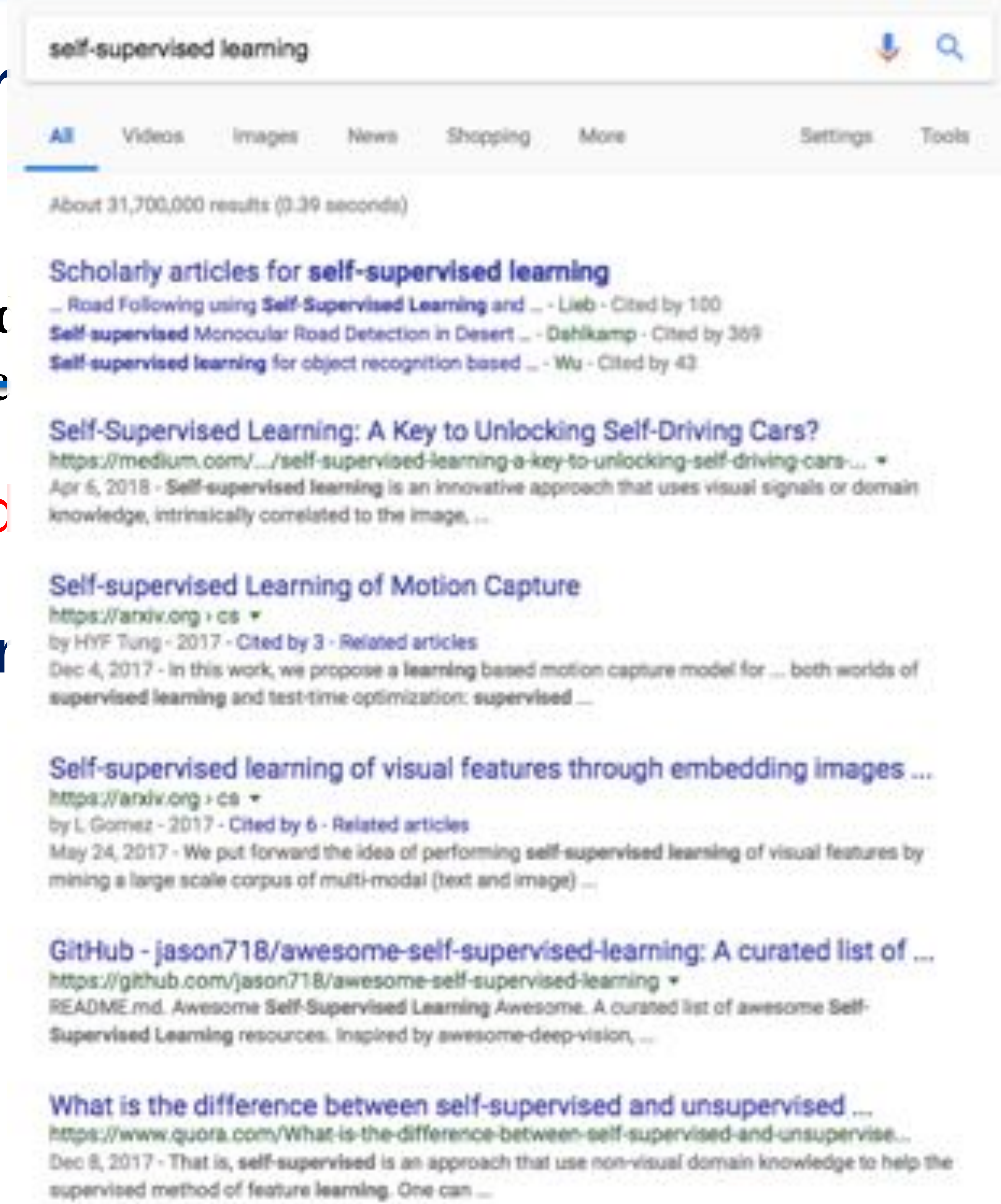
A comment or

Geometry Guided

~~**Self-Supervise**~~

Self-supervised

Supervised lear



Outline

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Semi-supervised learning

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Curriculum learning & curriculum adaptation

Web data of multi-modalities
- Web images vs. Web videos

Web images vs. Web videos

Given a query,

- Relevant Web images & video frames are alike
- An irrelevant Web image or video frame is irrelevant in its own way

(a) Basketbal Dunk

Web images vs. Web videos

Given a query,

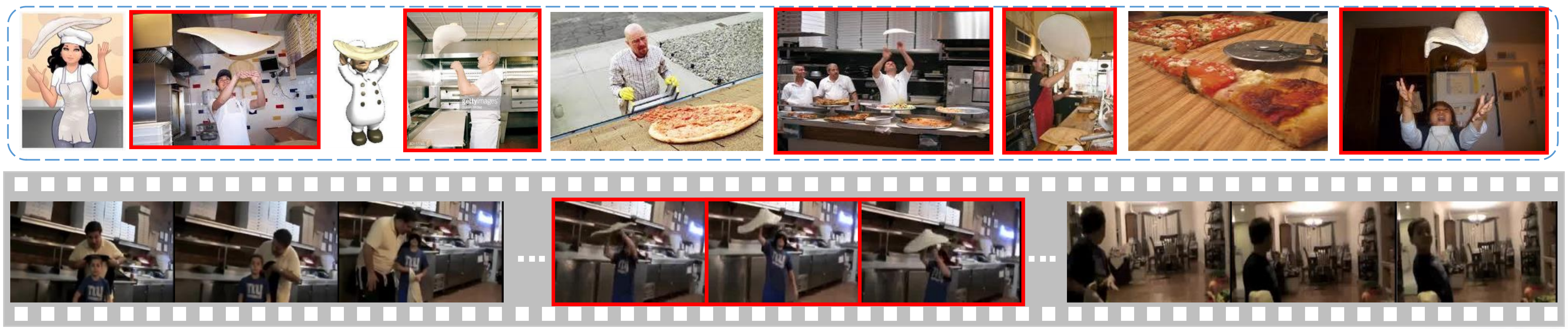
- Relevant Web images & video frames are alike
- An irrelevant Web image or video frame is irrelevant in its own way

(b) Bench Press

Web images vs. Web videos

Given a query,

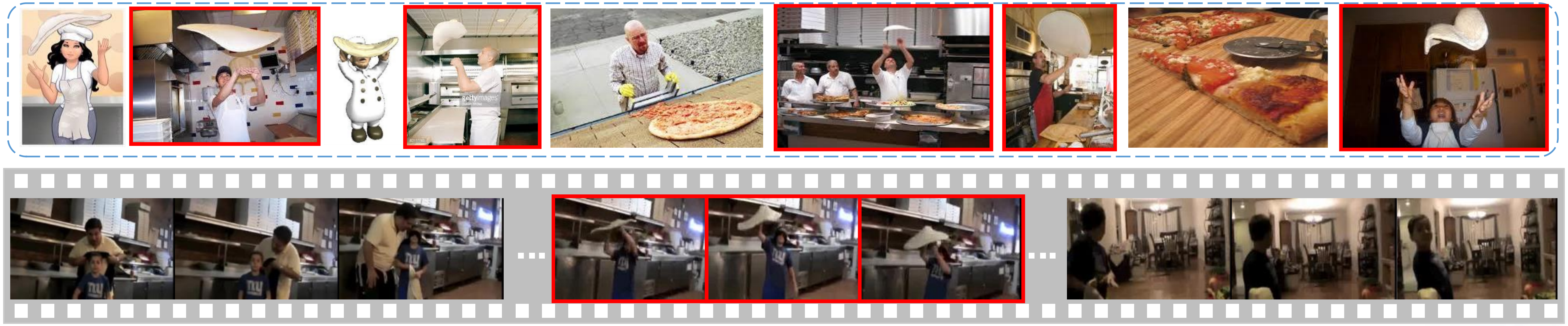
- Relevant Web images & video frames are alike
- An irrelevant Web image or video frame is irrelevant in its own way



(c) Pizza Tossing

Web images vs. Web videos

Mutually vote for commonness
to select training examples



(c) Pizza Tossing

Kernel mean embedding

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

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]

Empirical kernel mean estimation

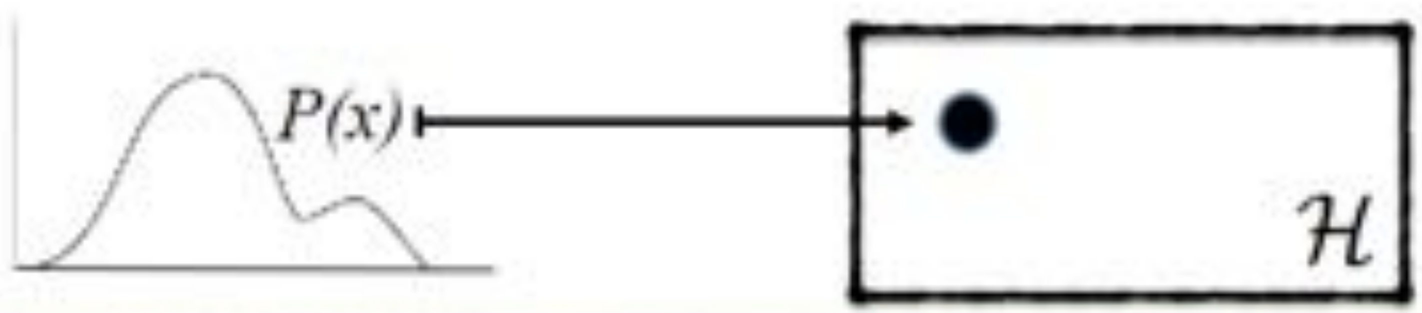
$$\mu[P] \triangleq \mathbb{E}_x[\phi(x)]$$

Empirical kernel embedding:

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

Mutually vote by matching kernel means

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



Empirical kernel embedding:

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

Mutually vote by matching kernel means

$$\min_{\alpha,\beta\in\{0,1\}} \left\| \frac{1}{\sum_m \alpha_m} \sum_{m'} \alpha_{m'} \phi(I_m) - \frac{1}{\sum_n \beta_n} \sum_{m'} \beta_{m'} \phi(F_m) \right\| + \mathcal{R}(\beta)$$

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

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

Mutually vote by matching kernel means

Table 6. Comparisons with state of the arts results using fully labeled data on UCF101.

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

[Gan et al., ECCV'16]

Outline

Web data with noisy labels

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

Semi-supervised learning

Web data with accurate labels

- Geometry from 3D movies
- Geometry encodes semantics

Curriculum learning & curriculum adaptation

Web data of multi-modalities

- Web images vs. Web videos

Mutually vote by kernel means

Future work: The Web is rich & inspiring

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Semi-supervised learning

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Curriculum learning & curriculum adaptation

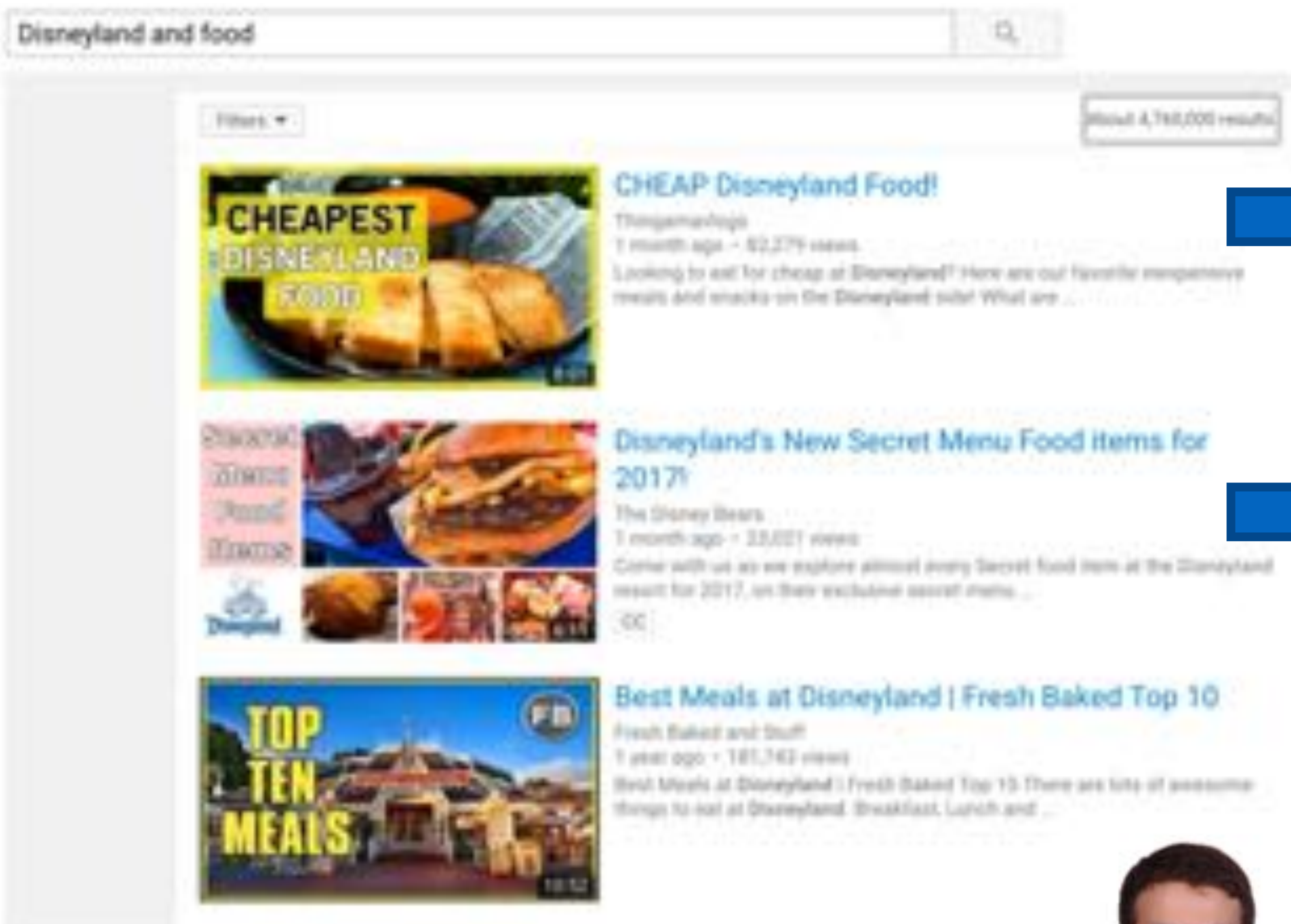
Web data of multi-modalities
- Web images vs. Web videos

Mutually vote by kernel means

Query, tags, news, audio, etc.

Future work: The Web is rich & inspiring

Query-focused video summarization



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

Future work: The Web is rich & inspiring

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Semi-supervised learning

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Curriculum learning & curriculum adaptation

Web data of multi-modalities
- Web images vs. Web videos

Mutually vote by kernel means

Query, tags, news, audio, etc.

Future work: The Web is rich & inspiring

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Semi-supervised learning

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Curriculum learning & curriculum adaptation

Web data of multi-modalities
- Web images vs. Web videos

Mutually vote by kernel means

Query, tags, news, audio, etc.

Multi-modal methods
Domain adaptation

References

[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] 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.

[6] Query-Focused Extractive Video Summarization. A. Sharghi, B Gong, & M Shah. ECCV 2016.