Investigating Data Augmentation Strategies for Advancing Deep Learning Training

Winston H. Hsu (徐宏民)
Professor, National Taiwan University
Director, NVIDIA AI Lab (NTU)

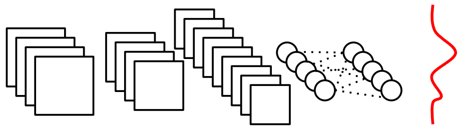
March 26, 2018

Office: R512, CSIE Building
Communication and Multimedia Lab (通訊與多媒體實驗室)
http://winstonhsu.info

Outline

- Why data augmentation in deep learning?
- Data augmentation strategies by
  - Data crawling
  - Weakly supervised learning (least effort for data)
  - Data transformation
  - Synthesizing
- Summary

Deep Learning – A Paradigm Shift in Machine Learning



- Competitive “**deep**” neural network
- Automatic feature learning (convolution)
- Huge improvements in image/video recognition tasks; so do in audio/speech applications; but marginally in text analytics

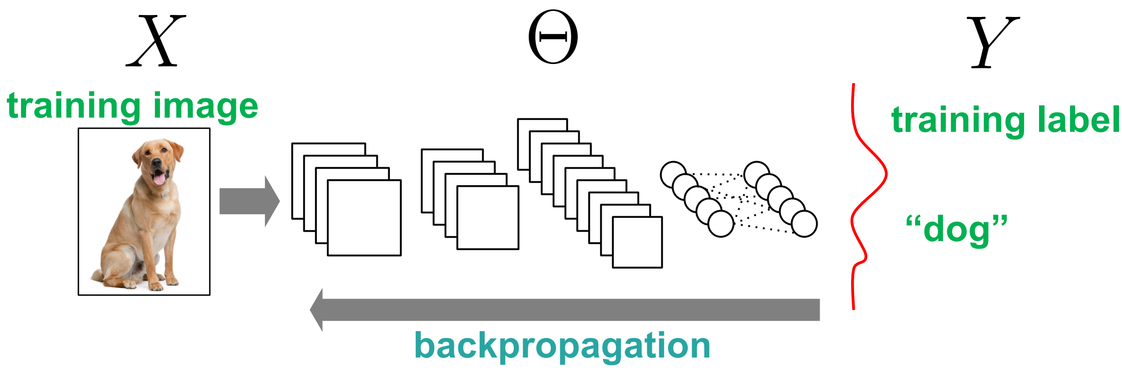
- For example, classification track in ILSVRC (Top 5 Error)

Year	Team	Result	Note
2012	SuperVision	0.15	1st Place (CNN)
2012	ISI	0.26	2nd Place (Conventional)
2015	MSRA	0.036	1st Place (CNN)

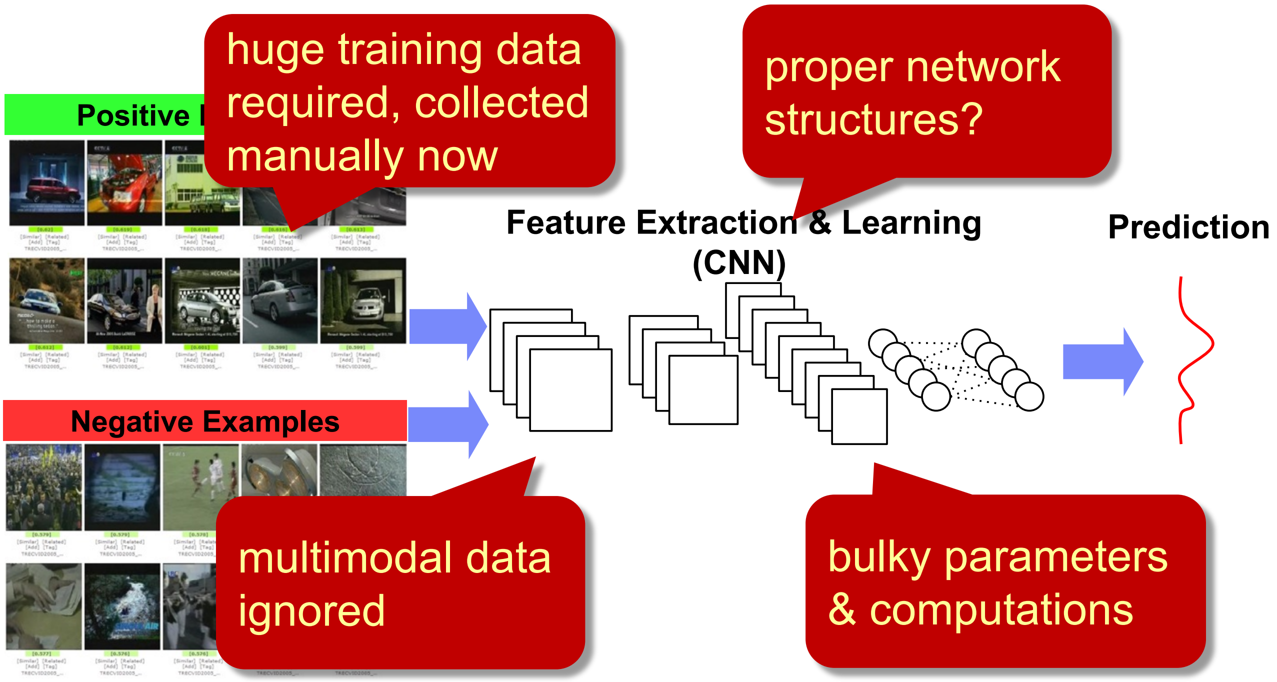
over 96% accuracy if 5 guesses are provided

Why Deep Neural Networks So Powerful?

- **“End-to-end training”** by
  - Huge training data, GPUs, advanced algorithms, etc.

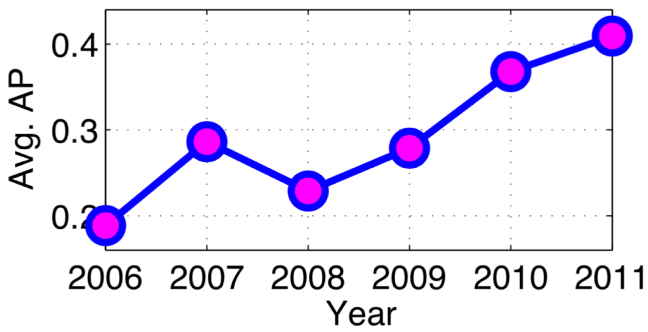


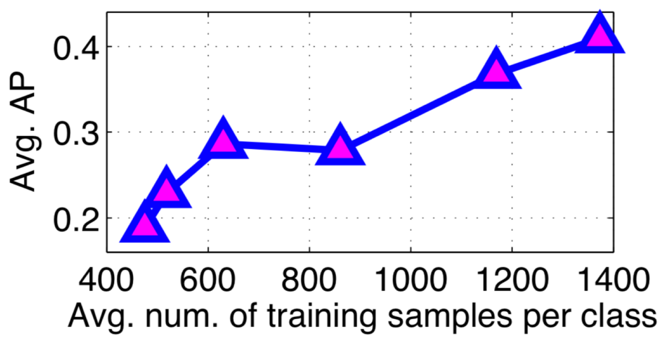
Deficiencies in Convolutional Neural Networks for Industry Products



Data is Vital across Learning Paradigms – Example via the (Old) Computer Vision Methods

- PASCAL VOC detection challenge provides realistic benchmark of object detection performance





Zhu et al. Do We Need More Training Data or Better Models for Object Detection? BMVC 2012

Data is Vital for Deep Learning

- AI algorithm is biased?
- Story covered in "*Facial Recognition Is Accurate, if You're a White Guy*," The New York Times, Feb. 9, 2018
- Actually, "gender classification" error caused by lacking quality data in certain categories
  - e.g., the darker the skin, the more errors arise
  - More specific training data will help

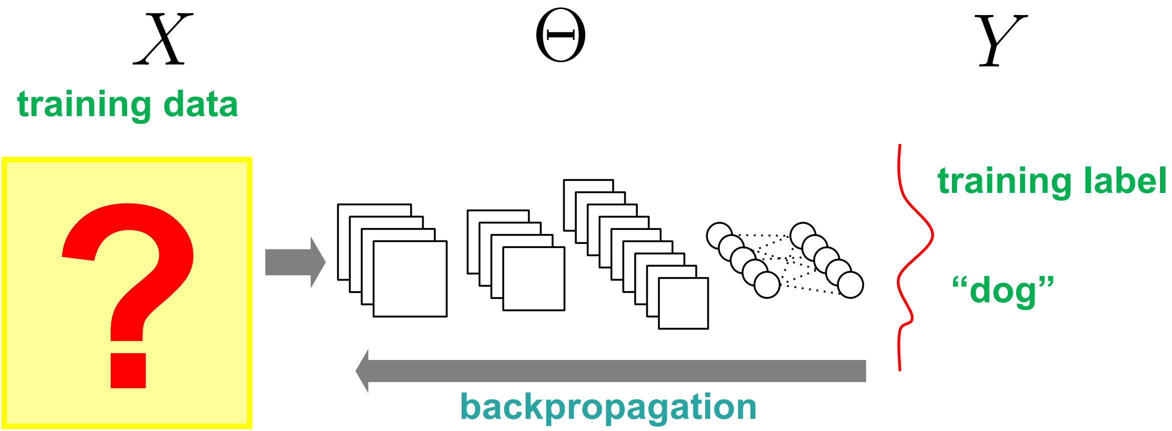
Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Where/How to Get Quality Training Data in an Efficient and Effective Way?



Data Crawling

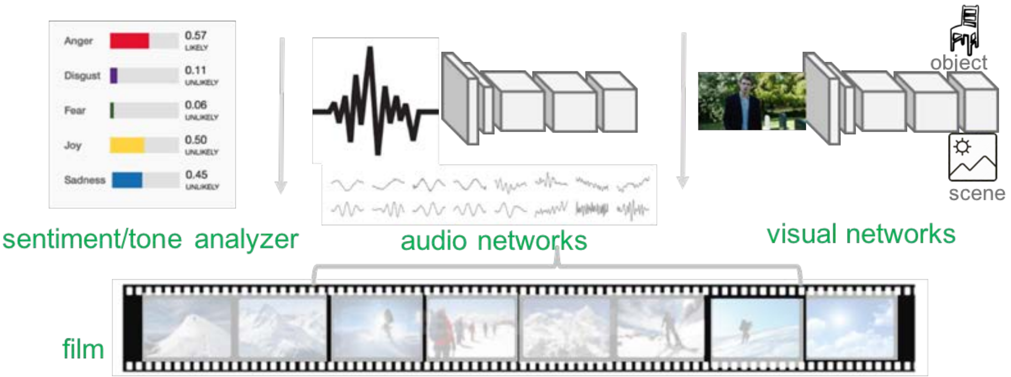
Rich Image/Videos, Comments, Metadata (GPS, Tags, Time, etc.) in Social Media

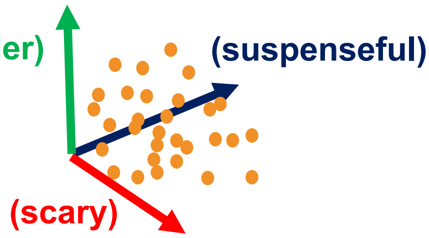
Why? Sharing for **organization** and **social communication**
[Ames, et al., CHI'07]

The First AI-Generated Movie Trailer – Learning from Hundreds of (Horror) Trailers

[Smith et al., ACMMM'17]

- Data availability dominates learning model(s)





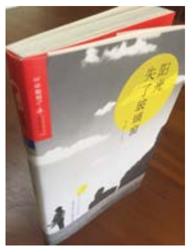
https://www.ibm.com/blogs/think/2016/08/cognitive-movie-trailer/

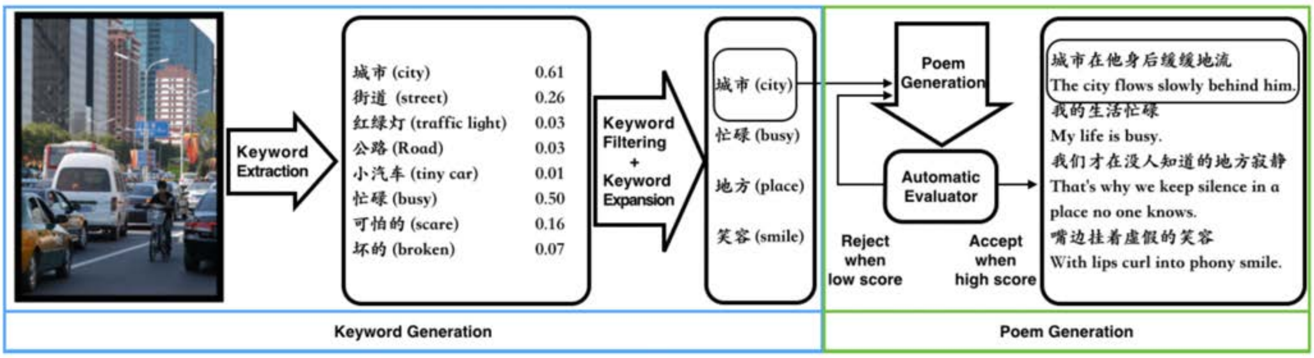
Image to Poetry by Cross-Modality Understanding

Microsoft Research

[Cheng et al., 2017]

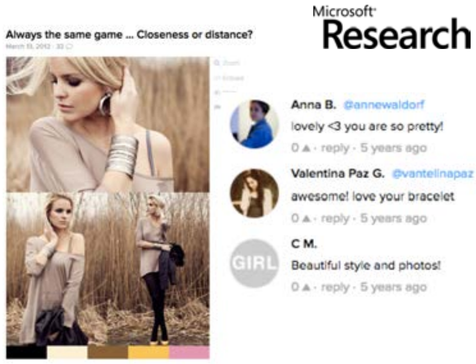
- Joint work with Microsoft Research Asia; deployed live in Microsoft chatbot XiaoIce (小冰)
- **Learning from the 519 poets** (1920~)
- Hierarchical LSTM-like models for ensuring the intra- and inter-sentence coherence





Netizen-Style Commenting by Learning from Fashion Communities – NetiLook (Public) Dataset

(a) **Human**: love ur ombre hair <3333 +1	
(b) **CaptionBot**: A group of women standing next to a woman.	
(c) **NC**: love the dress	
(d) **Attention**: love the shoes	
(e) **NSC**: I love the combinations :)) My heart for today goes to you! :)	



Dataset	Images	Sentences	Average Length	Unique Words
Flickr30k	30K	150K	13.39	23,461
MS COCO	200K	1M	10.46	54,231
NetiLook	350K	5M	3.75	597,629

- Contributing the first (large-scale) clothing dataset named **NetiLook** to discover netizen-style comments; **355,205** images from **11,034** users and **5 million** associated comments collected from Lookbook.
- Investigating commenting diversity by topic-parameterized neural networks (NSC)

Lin et al., Netizen-Style Commenting on Fashion Photos – Dataset and Diversity Measures, WWW 2018

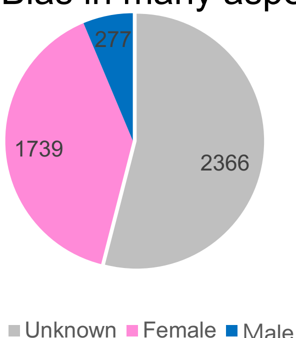
Social Media are Noisy and Biased

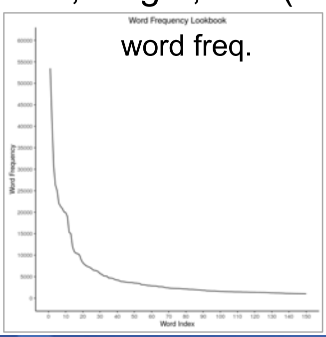
- Subjective and inaccurate for social tagging [Chang, 08]

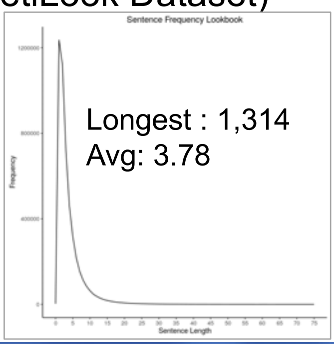
New York Landmark Labels (Flickr)

Locations	Precision
Brooklyn Br.	0.38
Chrysler Building	0.65
Columbia University	0.30
Empire State Building	0.18

- Bias in many aspects: gender, length, etc. (e.g., NetiLook Dataset)







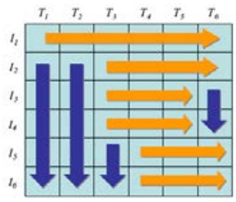
Data Annotation – Gaming with a Purpose

- **ESP Game:** labeling image as games [von Ahn, SIGCHI'04]
  – Two people see the same image, and type keywords until they match
- **Other variants:**
  – PeekABoom, Google Labeler, and more in www.gwap.com
  – Label Me

Data Annotation – Advanced Approaches

- Information beyond images / videos
  – speech, semantic network, location, hybrid tag/browse and … mind-reading

http://www.expertsystem.net



Hybrid Tag-Browse Labeler [Yan et al., 2008]

IBM Speech Recognition

Yahoo! ZoneTag

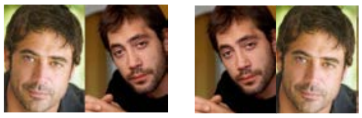
Brain-Computer Interface (Pic. From www.ice.hut.fi)

Data Annotation – Outsourcing Labeling Task

- Goal – outsourcing tasks to a distributed group of people
  - to <u>share</u> the annotation efforts
  - to reduce the <u>personal bias</u>
- Paid crowd-sourcing by *Amazon Mechanical Turk*

Worker   Requester

- **Mechanisms for ensuring quality**
  - Being completely answered in a HIT
  - Consistence for the "duplicated" questions in a HIT
  - Avoiding robots
  - …

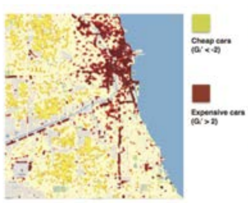


Question 12.
12 + 33 =? **Answer:**

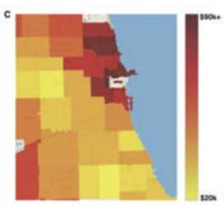
- Nice tutorial for annotations for numerous visual learning tasks in [Kovashka et al.]

Leveraging Product Images for Fine-Grained Vehicle Detection and Visual Census Estimation

- Goal – data-driven, machine learning-driven approaches are cheaper for collecting (predicting) <u>census data</u> (e.g., income, per capita carbon emission, crime rates, etc.) from Google Street View images
- Dataset – the largest fine-grained dataset reported to date consisting of over <u>2600 classes</u> of cars comprised of images from **Street View** and other **web sources**, classified by <u>car experts and AMT</u> (object)



Expensive/Cheap clusters of cars in Chicago



Zip code level median household income in Chicago.

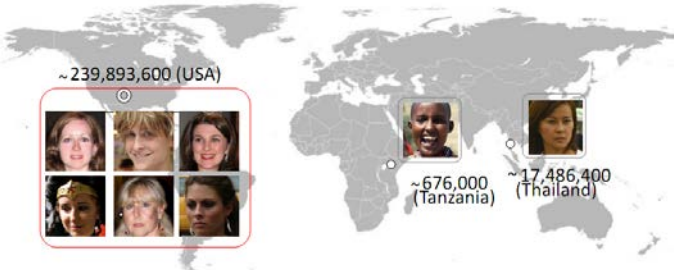
“Automatically” Acquiring Effective Training Images for Learning Facial Attributes

- Challenges:
  - Noise
  - Visual diversity
  - Geographical diversity

by textual relevance

by visual relevance

Geographically Uneven user-contributed photos



- Goals
  - Effectiveness for general facial attributes → data/feature selection
  - More diversity in training data → enhanced by contexts

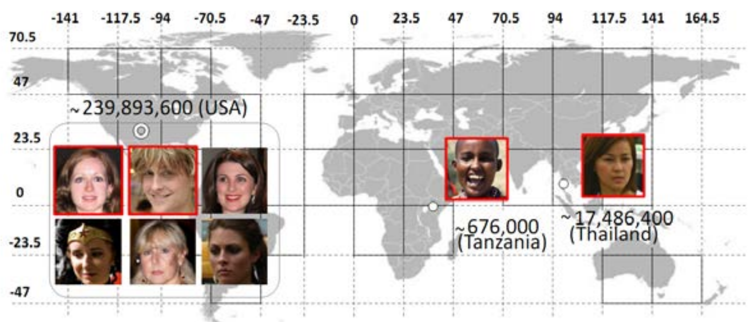
Chen et al. Automatic Training Image Acquisition and Effective Feature Selection From Community-Contributed Photos for Facial Attribute Detection. IEEE TMM 2013

Balancing Content and Context from Social Images

$$g_k = 1 - \frac{B_G(v_k)}{|G|}$$

v_k : votes from pseudo positives (negatives), weighted by x_m

g_k: relative v_k in a grid



Limit the number of images from a grid

$$\min_p \sum_k^K \boxed{(p_k - t_k)^2} - \boxed{\beta g_k p_k} + \boxed{\gamma \|p_k\|^2}$$

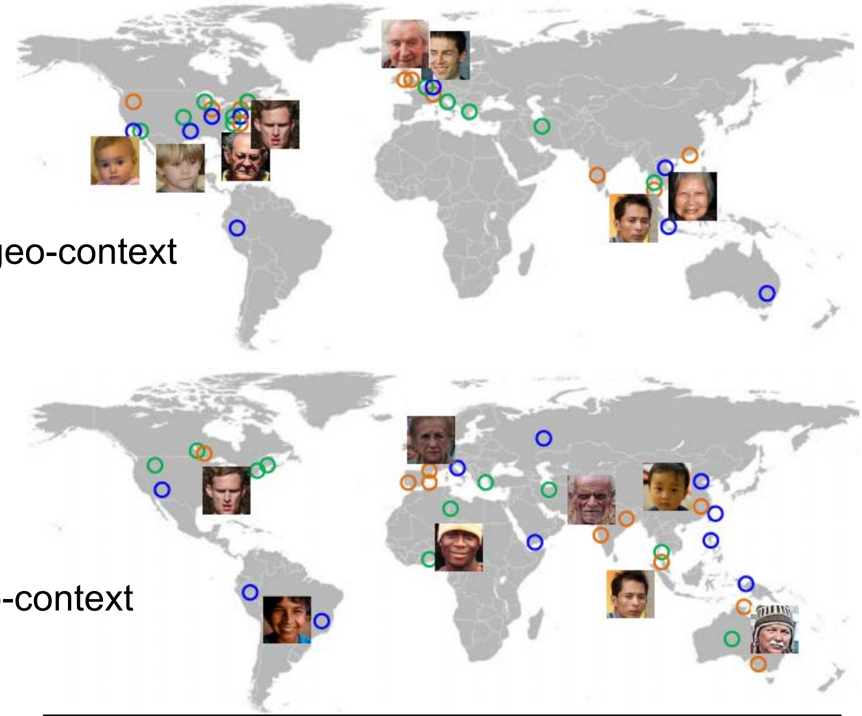
Textual Relevance — Visual Consistency — Regularization

p_k: annotation quality; selection indicator [0, 1]

Geographical Diversity for Training Facial Recognizers

without geo-context

with geo-context

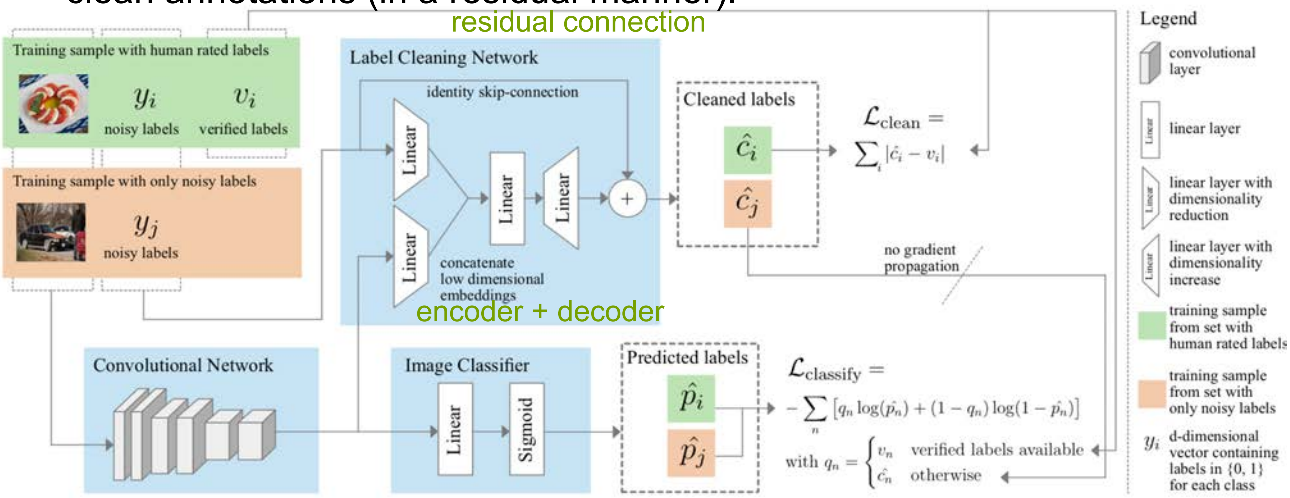


	elder	kid	male
Classification Error Rate (%)	26.33 (-0.00)	18.66 (-0.67)	24.50 (-3.00)

Least Effort for the Data

Learning from Noisy Labels – Annotation is Very Expensive (1/2)

- A multi-task network that jointly learns to clean noisy annotations and to accurately classify images
- Using the small clean dataset to learn a mapping between noisy and clean annotations (in a residual manner).



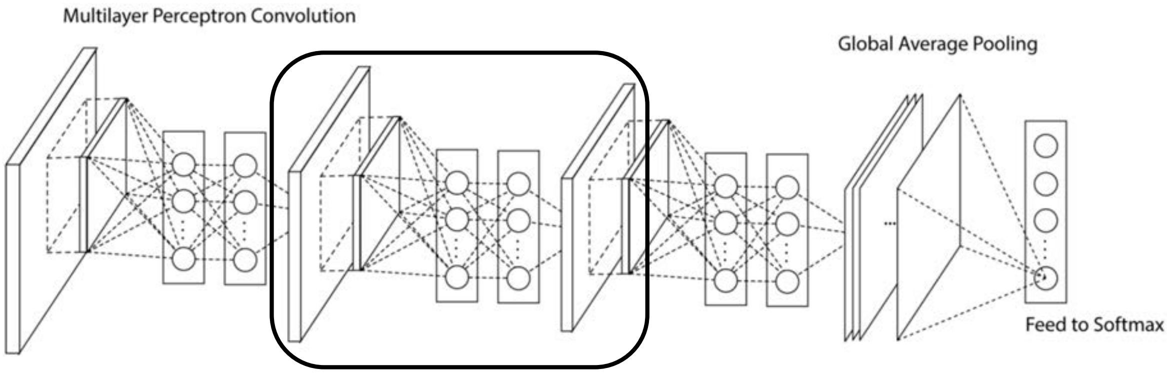
Veit at al. Learning From Noisy Large-Scale Datasets With Minimal Supervision. CVPR 2017

Learning from Noisy Labels – Annotation is Very Expensive (2/2)

- Using the clean labels to directly fine-tune a network trained on the noisy labels does not fully leverage the information
- Clean labels are used to reduce the noise in the large dataset **before** fine-tuning the network using both the clean labels and the full dataset with reduced noise.
- Experiments in Open Images dataset,
  - **Noisy set**: ~9 million images over 6000 unique classes
  - **Small clean set**: ~40k images.

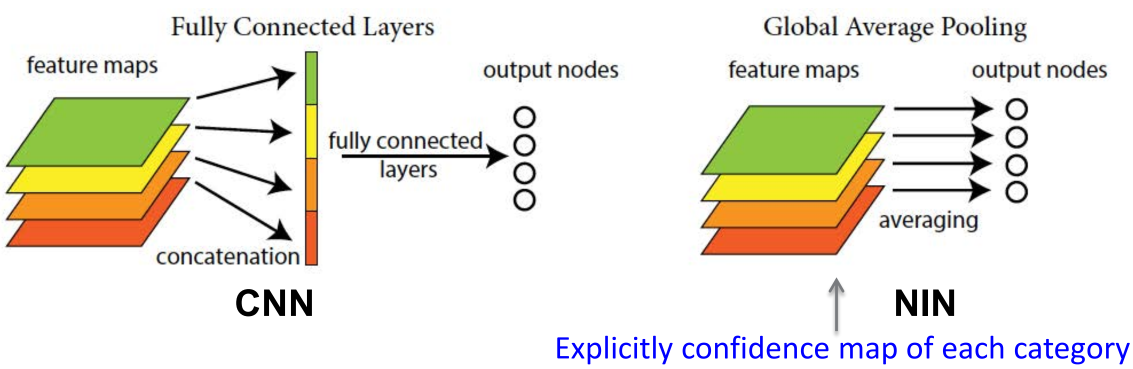
Model	AP_{all}	MAP
Baseline	83.82	61.82
Misra et al. [22] visual classifier	83.55	61.85
Misra et al. [22] relevance classifier	83.79	61.89
Fine-Tuning with mixed labels	84.80	61.90
Fine-Tuning with clean labels	85.88	61.53
Our Approach with pre-training	**87.68**	62.36
Our Approach trained jointly	87.67	**62.38**

Network in Network (NIN) – Compact Networks with Global Average Pooling



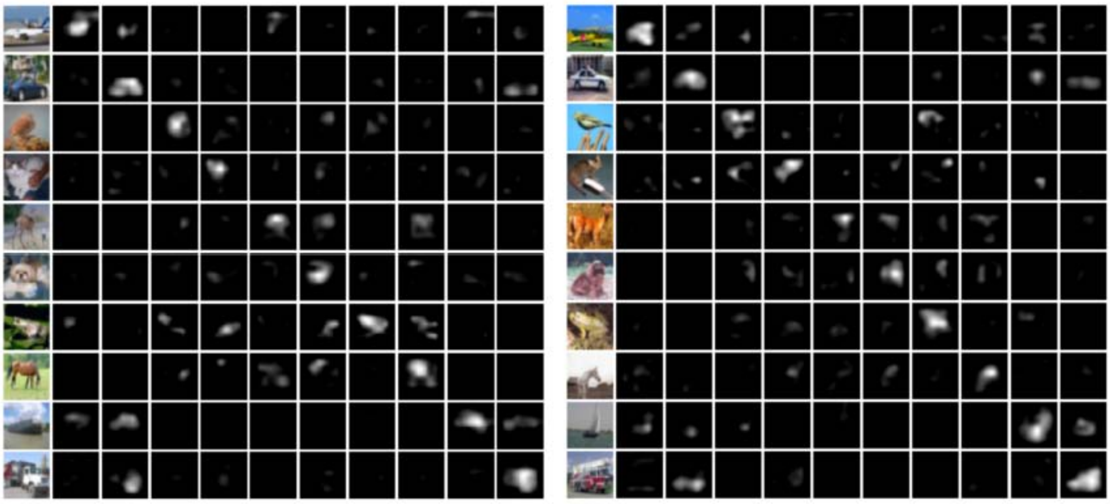
	Parameter Number	Performance	Time to train (GTX Titan)
AlexNet	60 Million (230 Megabytes)	40.7% (Top 1)	8 days
NIN	7.5 Million (**29 Megabytes**)	39.2% (Top 1)	4 days

Global Average Pooling – Huge Parameter Saving by Removing FC Layers



- Global average pooling layer produces spatial average of feature maps as confidence of categories
- Correspondence between feature maps and categories preserved; more meaningful and interpretable.
- No parameters (compared to fully connected layers) → prevent overfitting
- Robust to spatial translations of input

Side Product for Global Average Pooling – Visualizing Learned Spatial Correspondence



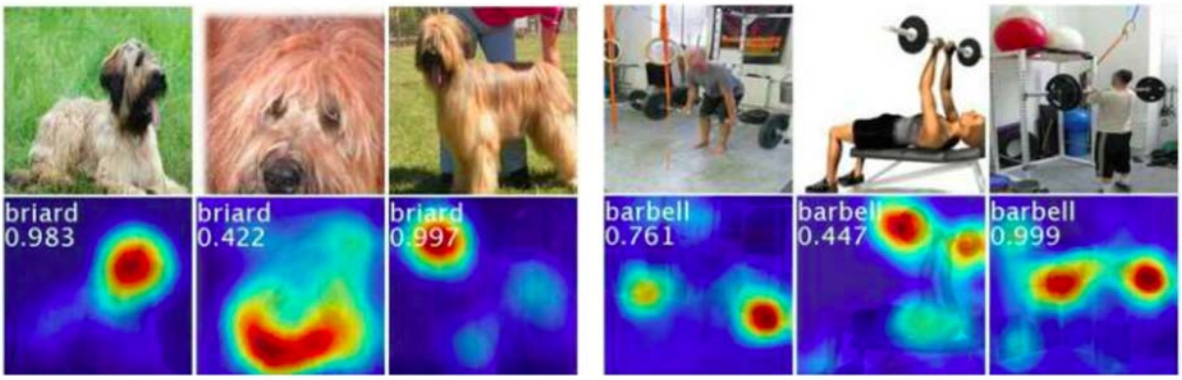
1. airplane, 2. automobile, 3. bird, 4. cat, 5. deer, 6. dog, 7. frog, 8. horse, 9. ship, 10. truck

Class Activation Map for Weakly Supervised Object Localization (1/3)

- Investigating what CNN is looking in image classification
  - Global Average Pooling (GAP): (1) Does not harm classification results, (2) Remarkable localization ability
  - Class Activation Map (CAM)

Class Activation Map for Weakly Supervised Object Localization (2/3)

- The CAMs of two classes from ILSVRC. The maps highlight the discriminative image regions used for image classification, the head of the animal for “briard” and the plates in “barbell”.



Class Activation Map for Weakly Supervised Object Localization (3/3)

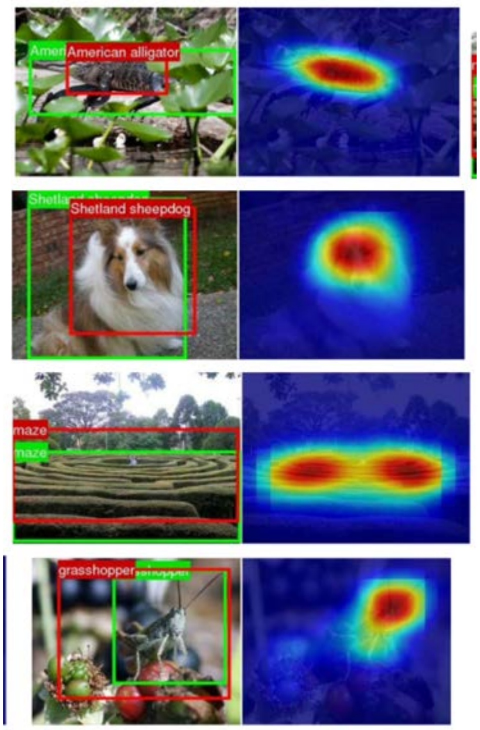


Table 2. Localization error on the ILSVRC validation set. *Backprop* refers to using [23] for localization instead of CAM.

Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

Method	supervision	top-5 test error
GoogLeNet-GAP (heuristics)	weakly	**37.1**
GoogLeNet-GAP	weakly	42.9
Backprop [23]	weakly	46.4
GoogLeNet [25]	full	26.7
OverFeat [22]	full	29.9
AlexNet [25]	full	34.2

* Green: Groundtruth
* Red: Predict

Weakly Supervised Object Detection

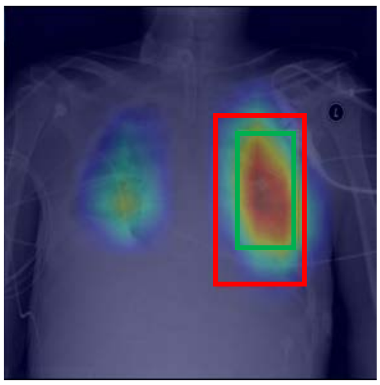
- Usual object detector is trained by dataset annotated with bounding boxes
  - Collecting those labels can be very costly and labor intensive.
  - **For fields like medical imaging, the labels are even more expensive.**
  - Image-level annotation is much easier to get
- **Weakly Supervised object detection**
  - Aim to train the model to localize the object <u>with only image level supervision</u> (only class label, no bounding boxes)

CAT CAT

Weakly Supervised Learning for Localizing Thoracic Diseases – Problem Definition

- Bounding box labels for medical images require professionals to generate the training data. It's rather time-consuming and expensive.
- Goal – train the network to <u>automatically localize the lesions</u> with only image level supervision. (**no bounding box info)**

Infiltration



Red: Ground truth bbox
Green: Predicted bbox

Wang et al., ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. CVPR 2017

NIH Chest X-Ray 8 Dataset

- Training: 108,948, 8 frontal view X-ray images of 32,717 unique patients with the recording containing disease image class labels **(noisy)**
- 985 human annotated bounding boxes on 880 images by 8 chest pathologies

ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases

Xiaosong Wang[1], Yifan Peng[2], Le Lu[1], Zhiyong Lu[2], Mohammadhadi Bagheri[1], Ronald M. Summers[1]
[1]Department of Radiology and Imaging Sciences, Clinical Center,
[2] National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20892
{xiaosong.wang,yifan.peng,le.lu,luzh,mohammad.bagheri,rms}@nih.gov

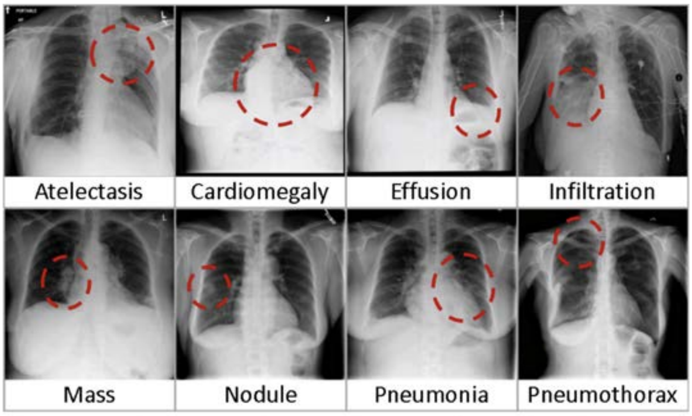
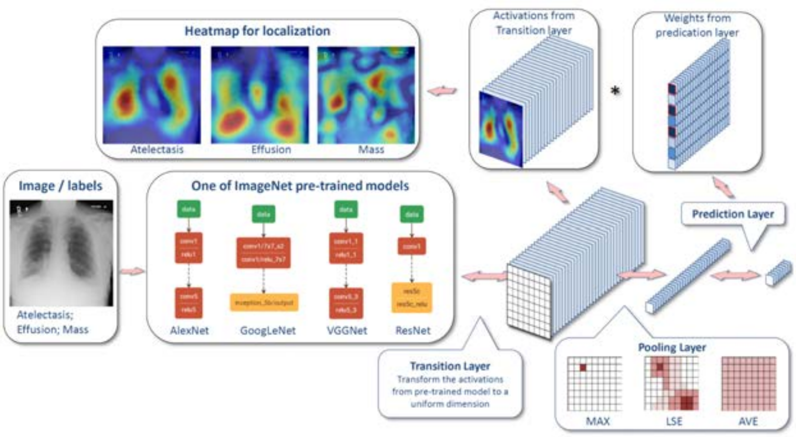
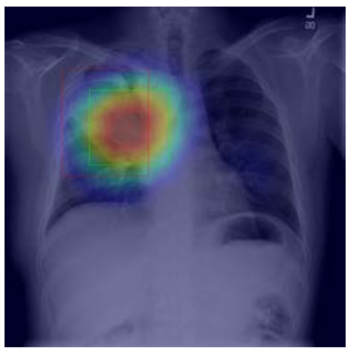


Figure 1. *Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully-automated diagnosis.*

Baseline Proposal and Results

- A multi-label classifier with pooling layer (LSE) to increase the localize capability of the network (mixture of GAP and GMP).
- Multiply the weights from prediction layers with the conv feature map to generate the activation heatmap of a specific class, similar to CAM



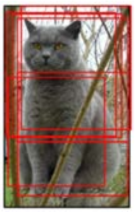


Mass

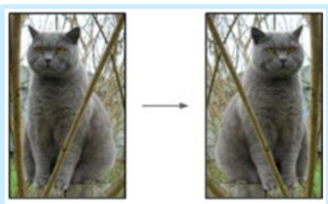
Data Transformation

Recent Data Augmentation Methods

- Summarized by Thoma in arXiv'17



crops



horizontal flip

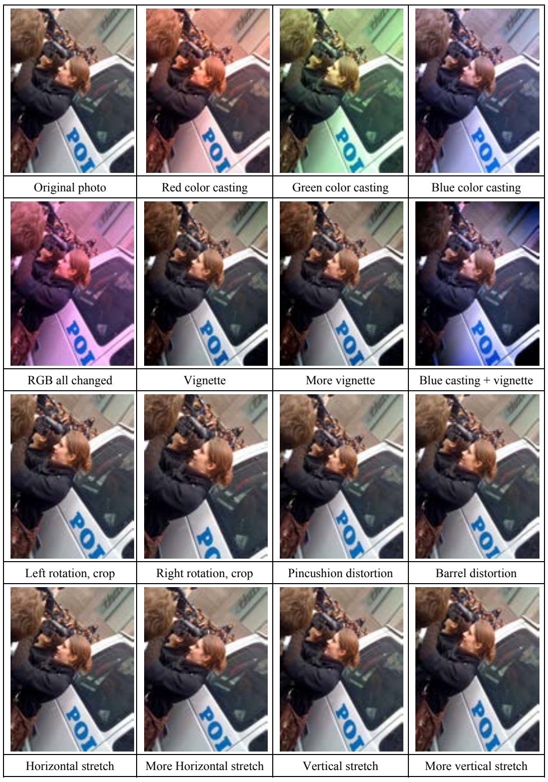
- Further operations
  - Adding noise
  - Elastic deformations
  - Color casting
  - Vignetting
  - Lens distortion

Name	Augmentation Factor
Horizontal flip	2
Vertical flip	2
Rotation	$\sim 40\ (\delta = 20)$
Scaling	$\sim 14\ (\delta \in [0.7, 1.4])$
Crops	$32^2 = 1024$
Shearing	
GANs	
Brightness	$\sim 20\ (\delta \in [0.5, 1.5])$
Hue	$51\ (\delta = 0.1)$
Saturation	$\sim 20\ (\delta = 0.5)$
Contrast	$\sim 20\ (\delta \in [0.5, 1.5])$
Channel shift	

Martin Thoma. Analysis and Optimization of Convolutional Neural Network Architectures. arXiv 2017.

Deep Image: Scaling up Image Recognition –
Wu et al., arxiv, 2015 (Baidu)

- **Data augmentation**
  - Cropping, shifting, color casting, lens distortion, vignetting, etc.
- Training on multi-scale images, including high-resolution ones
  - 512x512 vs. 224x224
- Hardware/software co-design for parallel computation
  - The number of weights is 212.7M
  - Estimated with 1GB for parameters



Deep Image: Scaling up Image Recognition –
Wu et al., arxiv, 2015 (Baidu)

- Configurations
  - 36 server nodes; each with 4 nvidia Tesla K40; FDR InfiniBand (56Gb/s)
  - Data parallelism in convolutional layers and model parallelism in FC layers
  - SGD synchronization: asynchronous updates
- Impacts

Team	Year	Place	Top-5 error
SuperVision	2012	1	16.42%
ISI	2012	2	26.17%
VGG	2012	3	26.98%
Clarifai	2013	1	11.74%
NUS	2013	2	12.95%
ZF	2013	3	13.51%
GoogLeNet	2014	1	6.66%
VGG	2014	2	7.32%
MSRA	2014	3	8.06%
Andrew Howard	2014	4	8.11%
DeeperVision	2014	5	9.51%
Deep Image	-	-	**5.98%**

Table 4: Single model comparison.

Team	Top-1 val. error	Top-5 val. error
GoogLeNet [21]	-	7.89%
VGG [20]	25.9%	8.0%
Deep Image	**24.88%**	**7.42%**

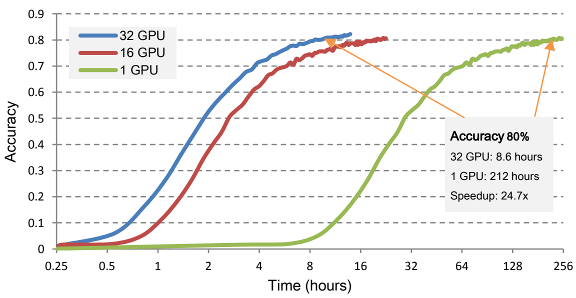


Figure 4: Validation set accuracy for different numbers of GPUs.

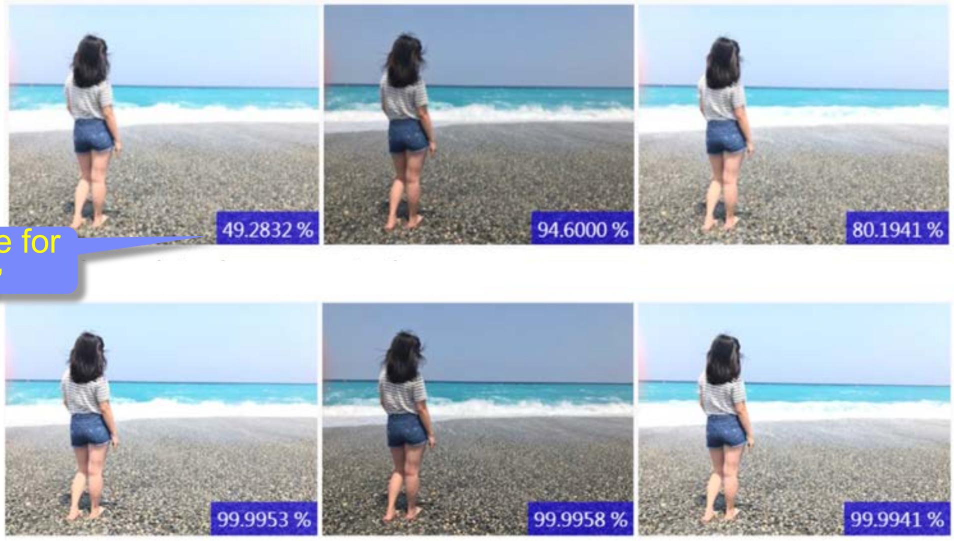
Industry Example – Image Recognition for Consumer Photo Management by Synology Inc. (1/2)

- Securing **robustness** in recognizing consumer photos, often suffering from varying lighting conditions

w/o augmentation

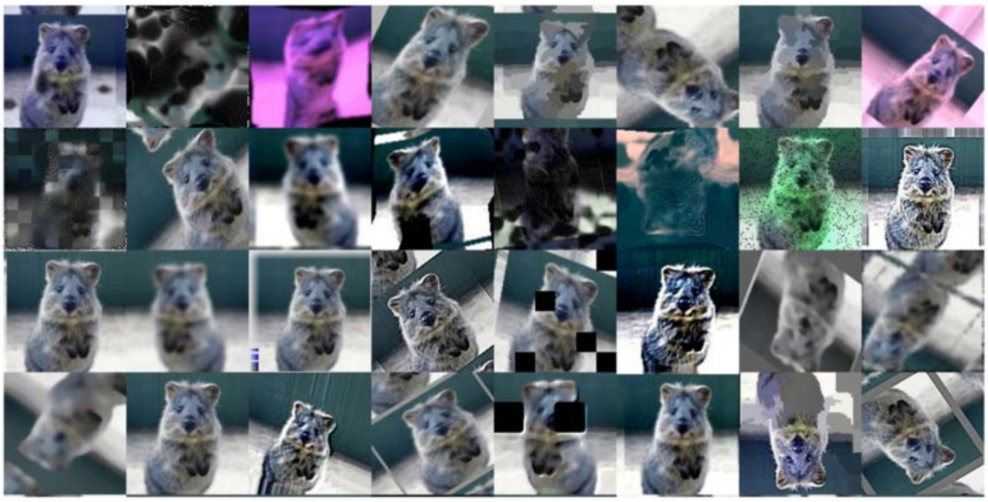
Confidence score for label “beach”

w/ augmentation



Industry Example – Image Recognition for Consumer Photo Management by Synology Inc. (2/2)

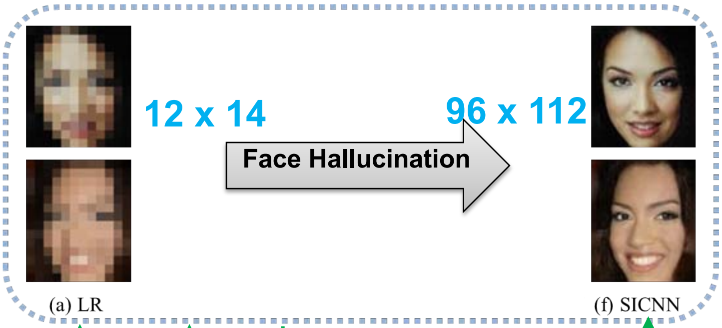
- Random flip, random crop
- Python open source image augmentation library: imgaug
  - Using blur, gaussian noise, brightness, hue, contrast and gray scale



"Shrinking Image" for Learning Super-Resolution (or Face Hallucination)

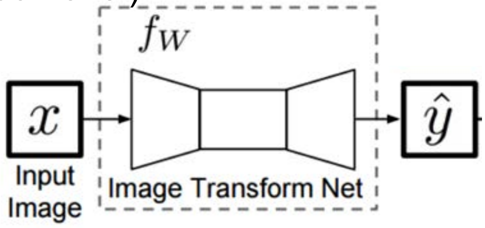
[Zhang et al., 2018]

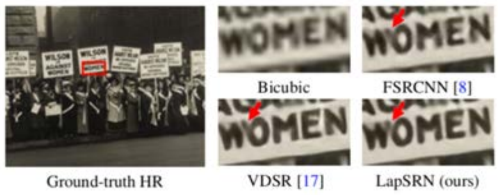
- 1/16 or 1/64 size of the original (high quality) one for measuring the reconstruction quality (pixel level or perceptual level)
- Mostly with encoder-decoder structure



(a) LR (f) SICNN

Low Resolution
ground truth
Generated by our method

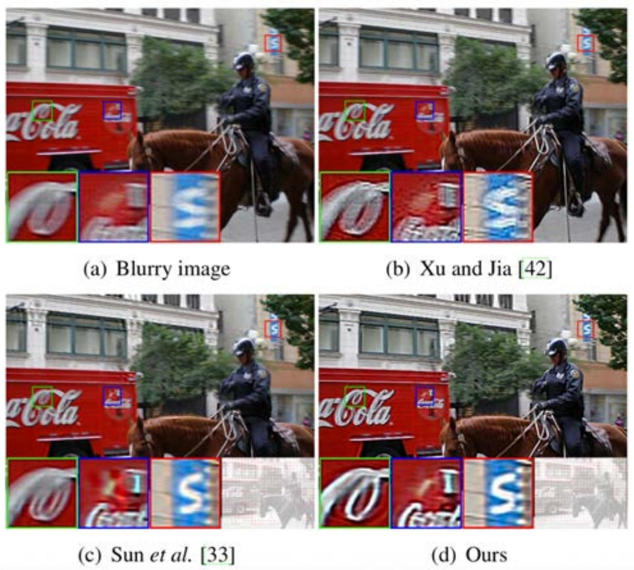


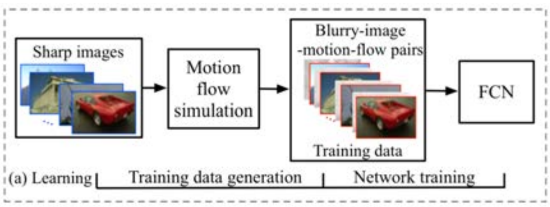

Ground-truth HR Bicubic FSRCNN [8] VDSR [17] LapSRN (ours)

- Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. CVPR 2017 '
- Johnson et al. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. ECCV 2016
- Dong et al. Accelerating the Super-Resolution Convolutional Neural Network. ECCV 2016

"Simulated Blurred Images" for Learning De-blurring

- Simulated blurred data from the original (high quality) one


(a) Blurry image (b) Xu and Jia [42]
(c) Sun *et al.* [33] (d) Ours

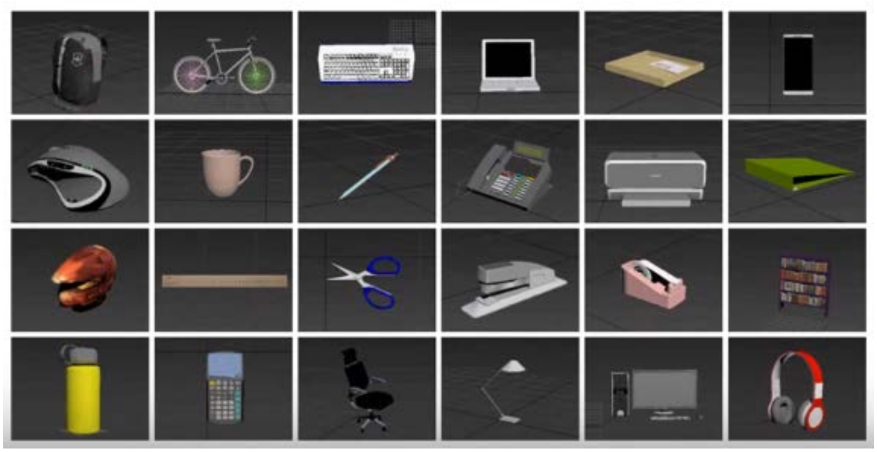


- Gong et al., From Motion Blur to Motion Flow: a Deep Learning Solution for Removing Heterogeneous Motion Blur. CVPR 2017
- Liang et al., Dual Motion GAN for Future-Flow Embedded Video Prediction. ICCV 2017

Synthesizing Data

Learning Object Detector from 3D Models (1/5)

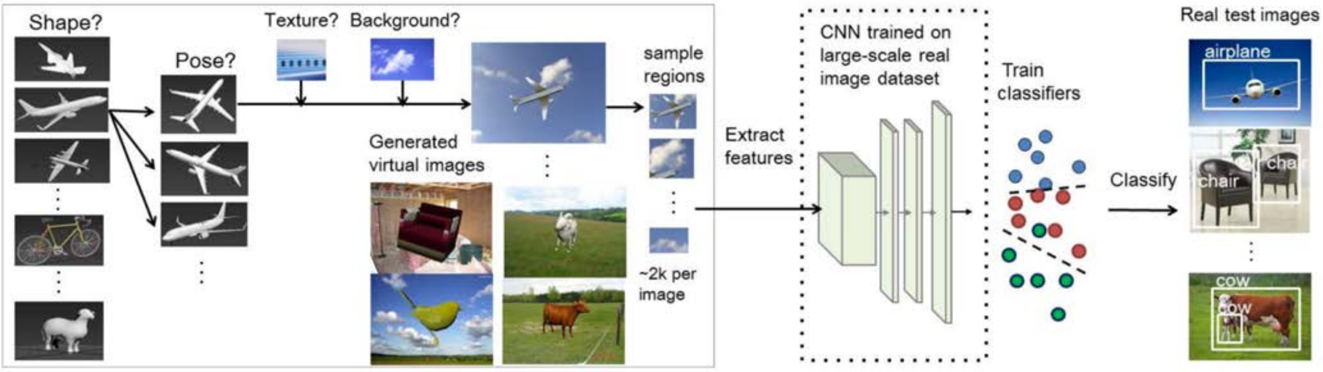
- Motivations – labelling images for detection is time-consuming.
  - Every object must be marked with a bounding box.
- Augmenting the training data with synthetic images rendered from 3D CAD models (e.g., 3dwarehouse)



Peng et al. Learning Deep Object Detectors from 3D Models. ICCV 2015.

Learning Object Detector from 3D Models (2/5)

- How variations in low-level cues affect the features by CNN on the object detection (e.g., PASCAL VOC2007 dataset).
  - Object color, texture and context
  - Synthetic image pose
  - 3D Shape



Learning Object Detector from 3D Models (3/5) – Object color, Texture and Context

the network has learned to be invariant to the color and texture of the object and its background.

	RR-RR	W-RR	W-UG	RR-UG	RG-UG	RG-RR
BG	Real RGB	White	White	Real RGB	Real Gray	Real Gray
TX	Real RGB	Real RGB	Unif. Gray	Unif. Gray	Unif. Gray	Real RGB

PASC-FT	aero	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
RR-RR	50.9	57.5	28.3	20.3	17.8	50.1	37.7	26.1	11.5	27.1	2.4	25.3	40.2	52.2	14.3	11.9	40.4	16.3	15.2	32.2	28.9
W-RR	46.5	55.8	28.6	21.7	21.3	50.6	46.6	28.9	14.9	38.1	0.7	27.3	42.5	53.0	17.4	22.8	30.4	16.4	16.7	43.5	31.2
W-UG	54.4	49.6	31.5	24.8	27.0	42.3	62.9	6.6	21.2	34.6	0.3	18.2	35.4	51.3	33.9	15.0	8.3	33.9	2.6	49.0	30.1
RR-UG	55.2	57.8	24.8	17.1	11.5	29.9	39.3	16.9	9.9	35.1	4.7	30.1	37.5	53.1	18.1	9.5	12.4	18.2	2.1	21.1	25.2
RG-UG	49.8	56.9	20.9	15.6	10.8	25.6	42.1	14.7	4.1	32.4	9.3	20.4	28.0	51.2	14.7	10.3	12.6	14.2	9.5	28.0	23.6
RG-RR	46.5	55.8	28.6	21.7	21.3	50.6	46.6	28.9	14.9	38.1	0.7	27.3	42.5	53.0	17.4	22.8	30.4	16.4	16.7	43.5	31.2
IMGNET	aero	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
RR-RR	34.3	34.6	19.9	17.1	10.8	30.0	33.0	18.4	9.7	13.7	1.4	17.6	17.7	34.7	13.9	11.8	15.2	12.7	6.3	26.0	18.9
W-RR	35.9	23.3	16.9	15.0	11.8	24.9	35.2	20.9	11.2	15.5	0.1	15.9	15.6	28.7	13.4	8.9	3.7	10.3	0.6	28.8	16.8
W-UG	38.6	32.5	18.7	14.1	9.7	21.2	36.0	9.9	11.3	13.6	0.9	15.7	15.5	32.3	15.9	9.9	9.7	19.9	0.1	17.4	17.1
RR-UG	26.4	36.3	9.5	9.6	9.4	5.8	24.9	0.4	1.2	12.8	4.7	14.4	9.2	28.8	11.7	9.6	0.7	4.9	0.1	12.2	11.6
RG-UG	32.7	34.5	20.2	14.6	9.4	7.5	30.1	12.1	2.3	14.6	9.3	15.2	11.2	30.2	12.3	11.4	2.2	9.9	0.5	13.1	14.7
RG-RR	26.4	38.2	21.0	15.4	12.1	26.7	34.5	18.0	8.8	16.4	0.4	17.0	20.9	32.1	11.0	14.7	18.4	14.8	6.7	32.0	19.3

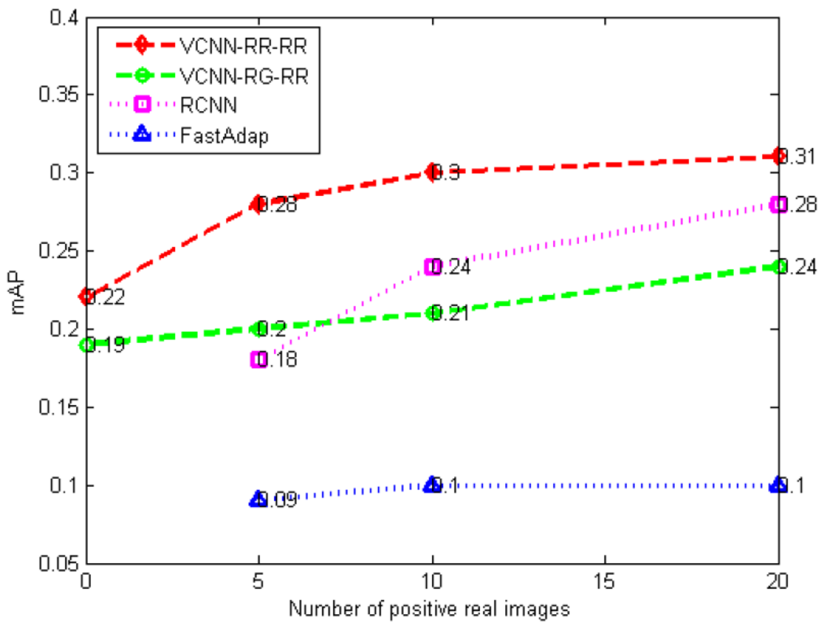
Learning Object Detector from 3D Models (4/5) – Experiments in Synthetic Pose

- Adding side view to front view gives a boost.
- Less invariance.

IMGNET	areo	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
front	24.9	38.7	12.5	9.3	9.4	18.8	33.6	13.8	9.7	12.5	2.1	18.0	19.6	27.8	13.3	7.5	10.2	9.6	13.8	28.8	16.7
front,side	24.3	36.8	19.0	17.7	11.9	26.6	36.0	10.8	9.7	15.5	0.9	21.6	21.1	32.8	14.2	12.0	14.3	12.7	10.1	32.6	19.0
front,side,intra	33.1	40.2	19.4	19.6	12.4	29.8	35.3	16.1	5.2	16.5	0.9	19.7	19.0	34.9	15.8	11.8	19.7	16.6	14.3	29.8	20.5
PASC-FT	aero	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
front	41.8	53.7	14.5	19.1	11.6	42.5	40.4	25.5	9.9	24.5	0.2	29.4	37.4	47.1	14.0	11.9	18.9	12.7	22.6	38.8	25.8
front,side	45.6	50.2	24.4	28.8	17.4	51.9	41.8	24.5	7.2	27.9	9.2	23.1	37.0	51.3	17.8	13.2	28.6	18.9	9.3	37.8	28.3
front,side,intra	54.2	55.5	22.7	27.0	20.5	52.6	40.1	26.8	8.1	27.3	2.3	30.6	36.6	53.3	17.8	14.2	34.1	26.4	19.3	37.5	30.3

Learning Object Detector from 3D Models (5/5) – Training Object Detection with Limited Real Images

- When the number of real training images is limited, 3D models performs better than traditional RCNN over limited training data.



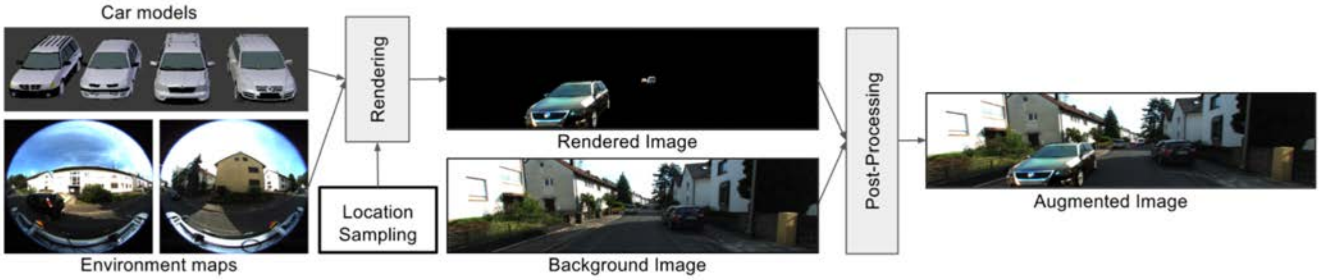
Augmented Reality for Data Generation (1/3) – Motivations

- Creating realistic 3D content is challenging and labor-intensive.
- Real-world images at large scale is easy and directly provides real background appearances without complex 3D models
- **Augmented imagery** generalizes better than **synthetic 3D data** or **limited real data**

Alhaija et al., Augmented Reality Meets Deep Learning for Car Instance Segmentation in Urban Scenes. BMVC 2017

Augmented Reality for Data Generation (2/3) – Augmentation Pipeline

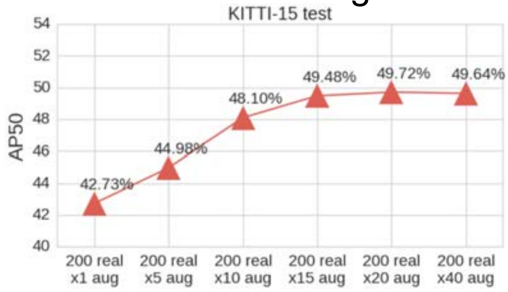
- Given a set of 3D car models, locations (**manual** or **automatic**) and environment maps
- Rendering high quality cars and overlay them on top of real images.
- The final post-processing step ensures better visual matching between the rendered and real parts of the resulting image.

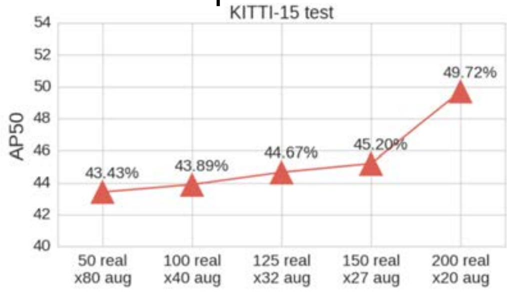


Alhaija et al., Augmented Reality Meets Deep Learning for Car Instance Segmentation in Urban Scenes. BMVC 2017

Augmented Reality for Data Generation (3/3) – Impacts Measured by Instance (Car) Segmentation

- More data – real or augmented – are both helpful

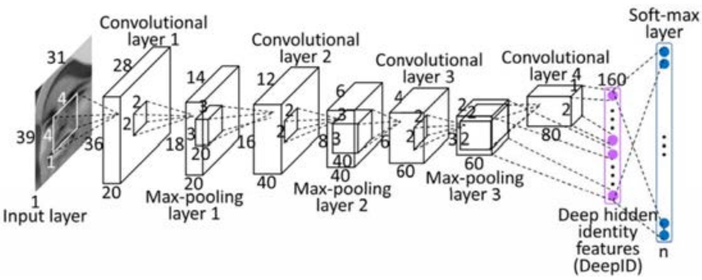


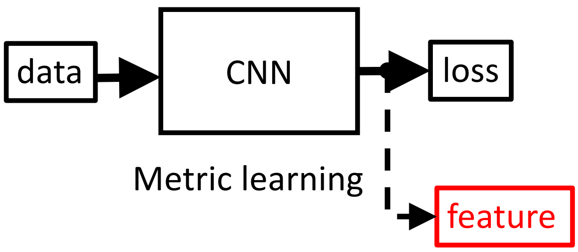


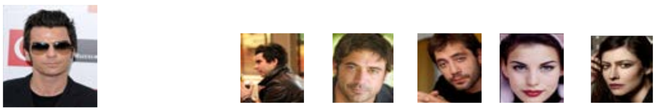
- Augmented foreground cars with real backgrounds are effective

(a) Black BG AP50 = 21.5% (b) Flickr BG AP50 = 40.3% (c) Virtual KITTI BG AP50 = 47.7.3% (d) Real BG AP50 = 49.7%

Face Recognition (Verification or Identification) by Varying (Multi-Tasking) Loss and Datasets







Probe set Gallery set

1 : N

Face Identification

Face verification

1 : 1

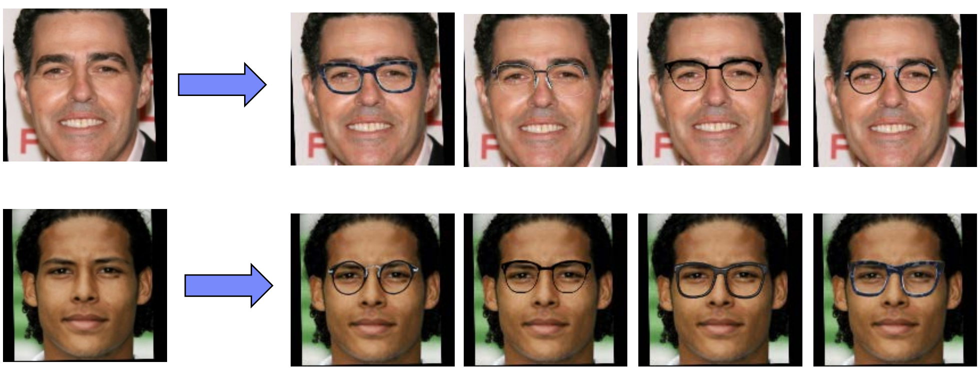
Dataset	#Images	#Persons	#Images per person
CASIA	0.49M	10K	50
VGGFace	2M	2K	1000
Megaface2	4.8M	672K	7
MS-celeb	10M	100K	100
UMDFace	0.37M	8.5K	40

Industry Example – Augmented Glasses for Face Recognition (1/2) by Video Surveillance SBU, LiteOn

- Problem – Each pair is the same person but predicted to different people (red pairs) by the deep methods because of thick glasses
- Why? Few faces with glasses in most face datasets but common in Asian

Industry Example – Augmented Glasses for Face Recognition (2/2) by Video Surveillance SBU, LiteOn

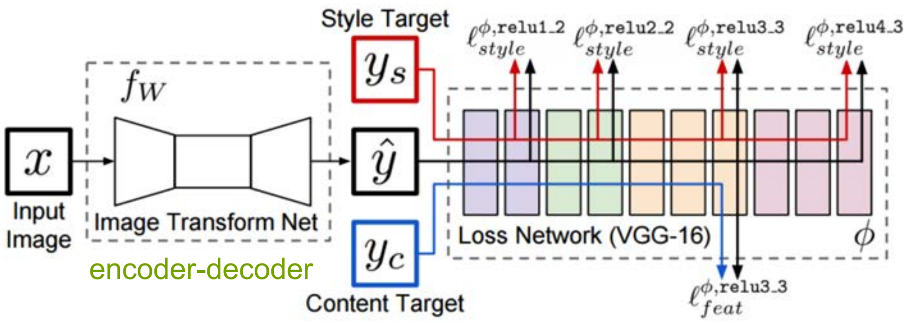
- Glass invariance augmentation by overlapping variant glass models



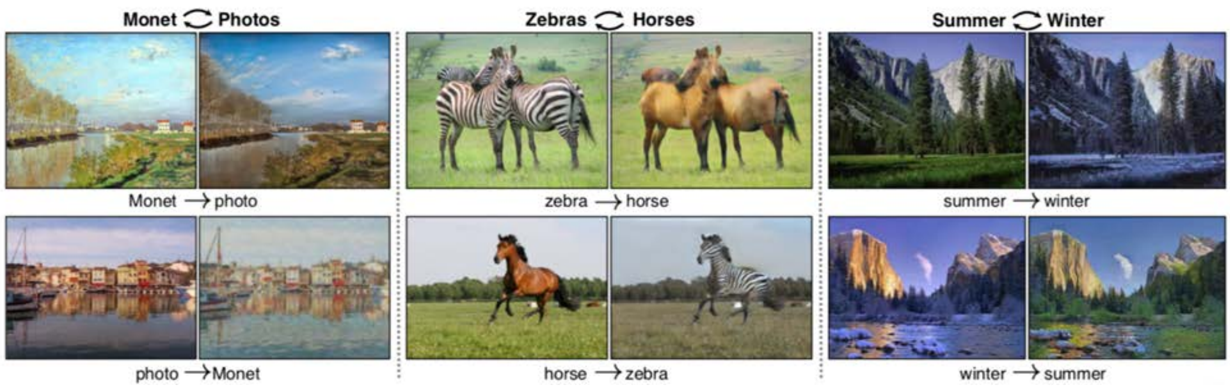
- Impacts on proprietary benchmarks include 21,570 face (glass) pairs

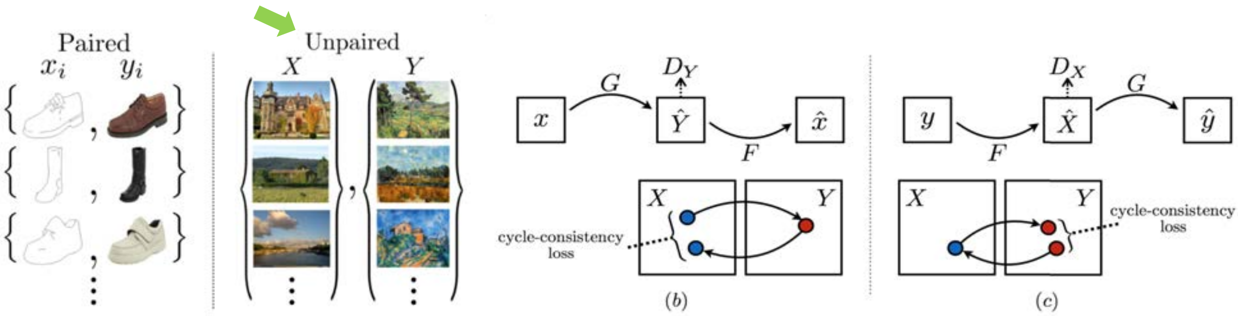
	w/o glasses aug.	w/ glasses aug.
Accuracy	99.207%	99.420%
Error case on glasses	109	40

(Real-Time) Style Transfer by Perceptual Losses



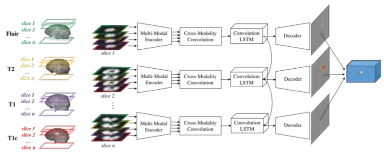
CycleGAN for Synthesizing “Realistic” Training Data





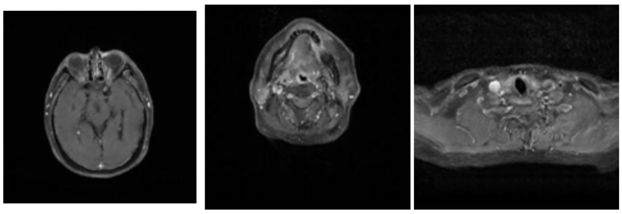
- Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks ICCV 2017
- Zhang et al., Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network. CVPR 2018

CycleGAN for Synthesizing “Realistic” Training Data

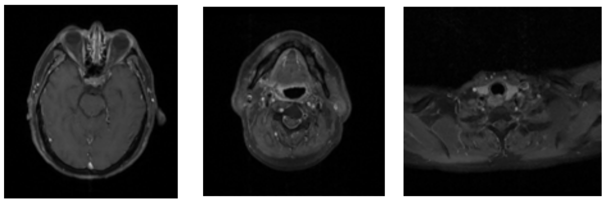


- Purposes for synthesized medical images
  - as an intermedium in cross-modality image registration or learning
  - as supplementary training samples to boost the generalization capability
- Our modified CybleGAN for multimodal recognition for utilizing MRI and CT (here CT → MRI) in Nasopharyngeal Carcinoma (NPC)

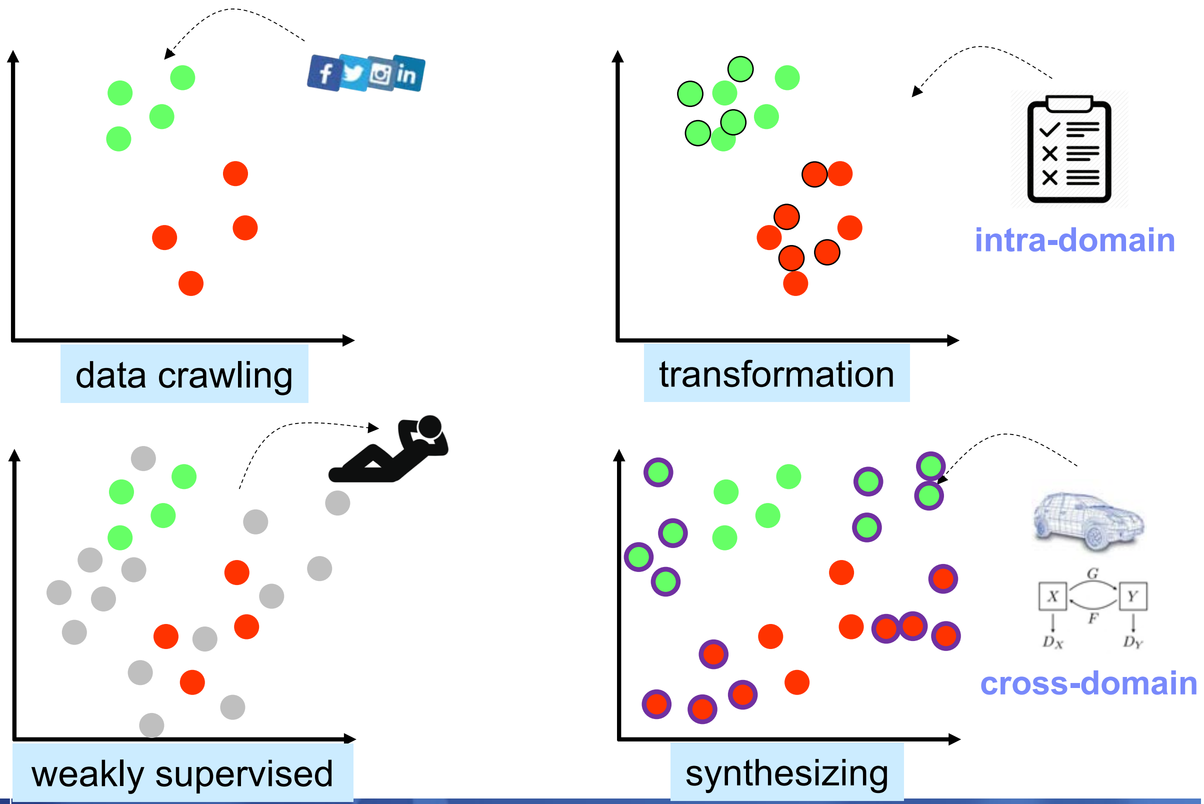
synthesized or real?



synthesized or real?



Summary – Utilizing Data Efficiently and Effectively



Take Home Messages

Take Home Messages

- Data are vital for learning paradigms but very costly
- Collecting more training data from public datasets
  - Used for multi-task learning or pre-training
- Augment data with
  - Social media
  - Synthesized data: transformed data, 3D, AR, GAN,
  - Work with the noisy data
  - Weakly supervised methods for minimizing human costs
- Data augmentation is vital for industry applications and will emerge as an important technical component
- Privacy! Privacy! Privacy!

Facebook, LinkedIn: “Winston Hsu”