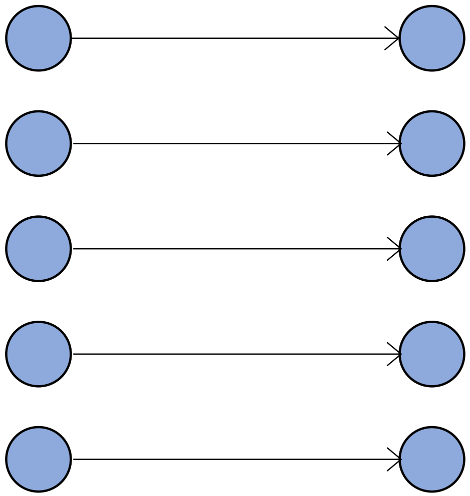
Application of Generative Models: X Learning

Hao Dong

Peking University

From Data Point of View

Data in both input x and output y
with known mappings
(Learn the mapping f)

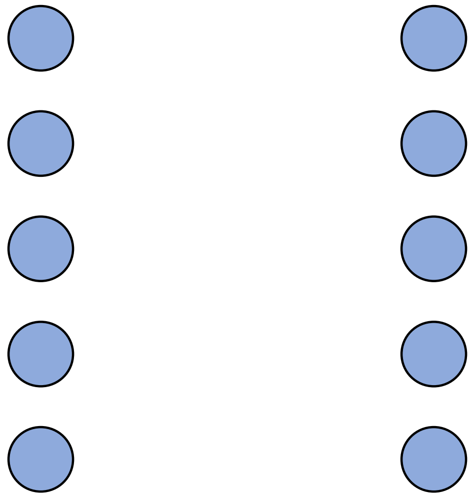


$$y = f(x)$$

Supervised Learning

- Image classification
- Object detection
- …

Data in both input x and output y
without known mappings
(Learn the mapping f)



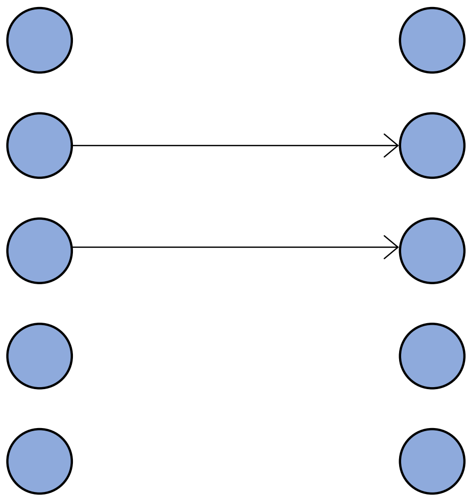
$$y = f(x)$$

Unsupervised Learning

- Autoencoder
(when output is features)
- GANs
- …

From Data Point of View

Data in both input x and output y
with known partial mappings
(Learn the mapping f)

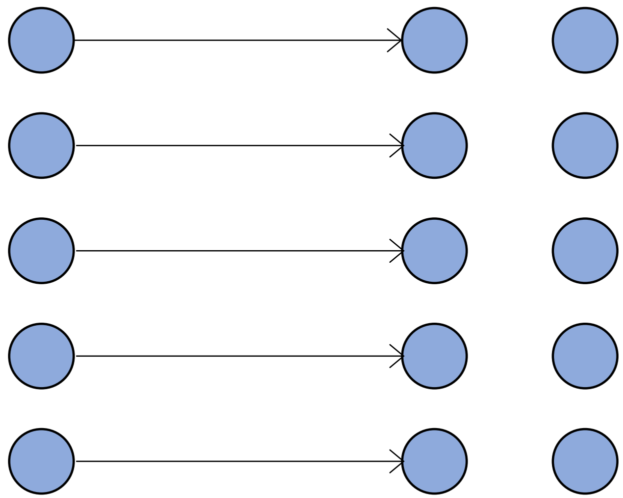


$$y = f(x)$$

Semi-supervised Learning

- …

Data in both input x and output y
with known mappings for y
(Learn the mapping f for another output y')

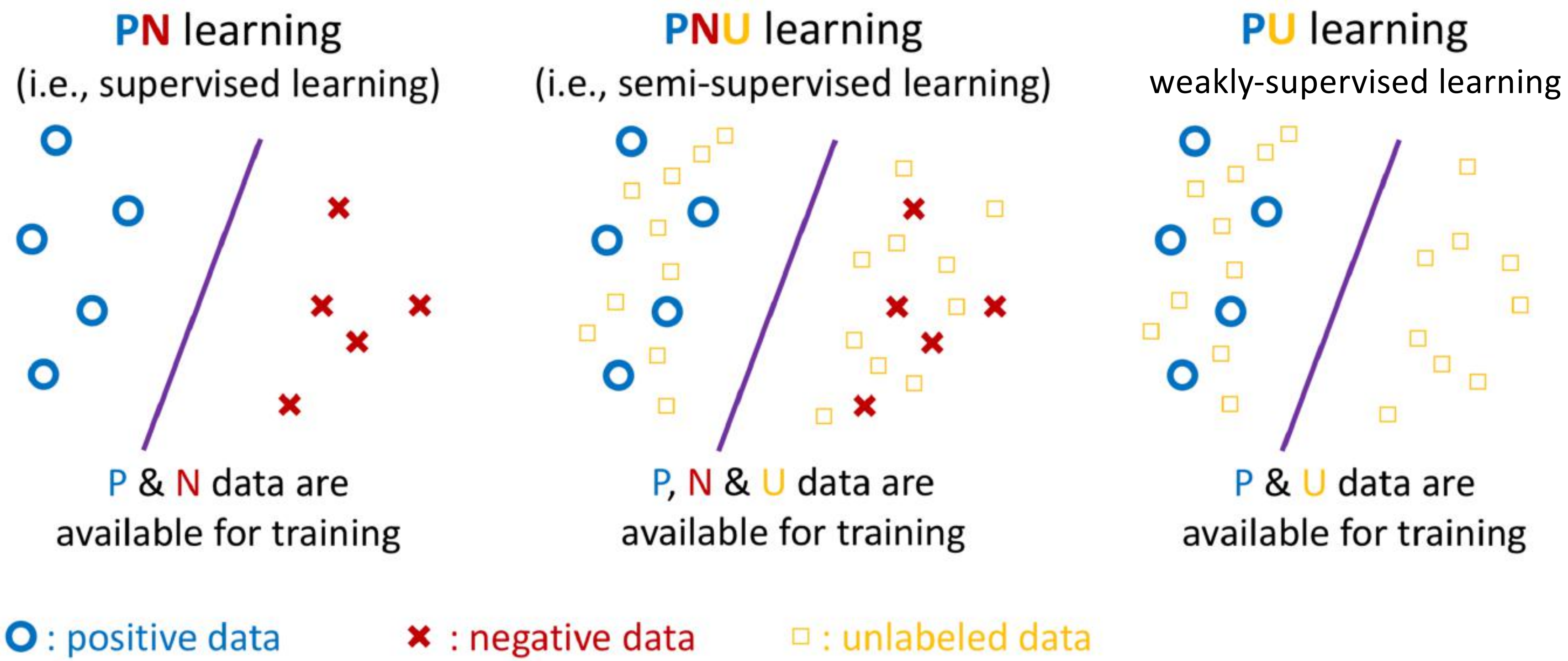


$$y' = f(x)$$

Weakly-supervised Learning

- Learn segmentation via classification
- …

From Data Point of View

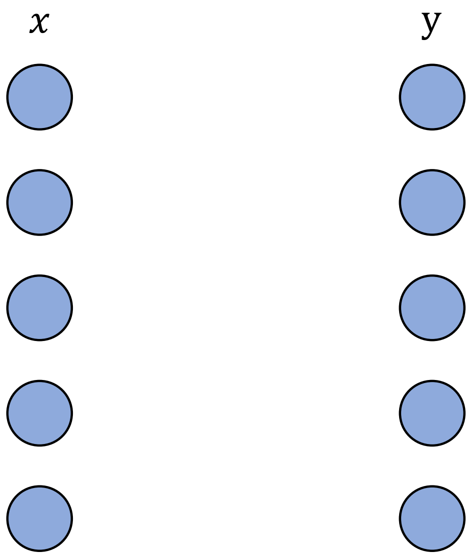


From https://niug1984.github.io/paper/niu_tdlw2018.pdf

From Mapping Point of View

PEKING UNIVERSITY

Data in both input and output
(Learn the mapping f, f')

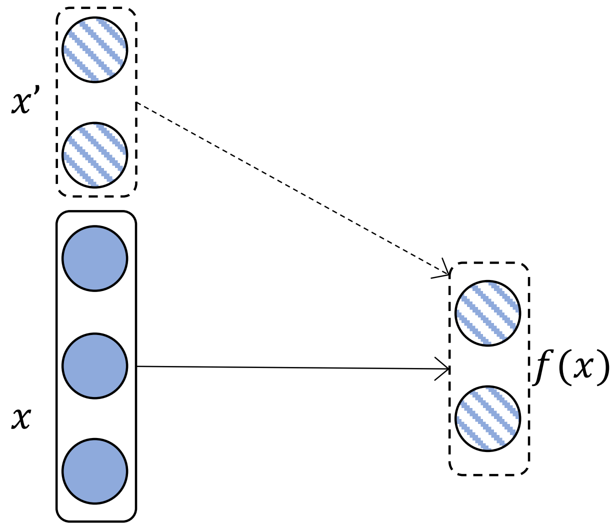


$y = f(x), x = f'(y)$

(Unsupervised) Dual Learning

- VAE
- CycleGAN
- …

Data in input x, x' only
with known mapping f'
(Learn the mapping f)

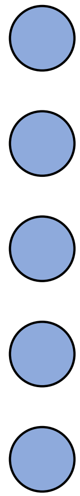


$x' = f(x)$

Self-supervised Learning

- Word2Vec
- Denoising Autoencoder
- …

Data in input only
with known inverse mapping f'
(Learn the mapping f and output y)



$y = f(x), x = f'(y)$

Self-augmented Learning

- ?

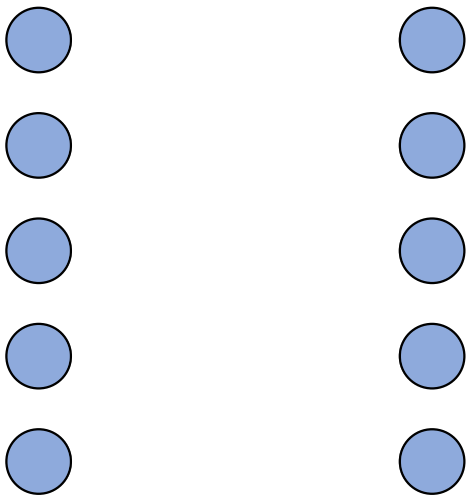
Application of Generative Models: Learning Methods

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

Unsupervised Learning

Data in both input x and output y
(Learn the mapping f)



$y = f(x)$

Unsupervised Learning

- In practice, it is difficult to obtain a large amount of labelled data, but it is easy to get a large amount of unlabeled data.
- Learn a good feature extractor using unlabelled data and then learn the classifier using labelled data can improve the performance.

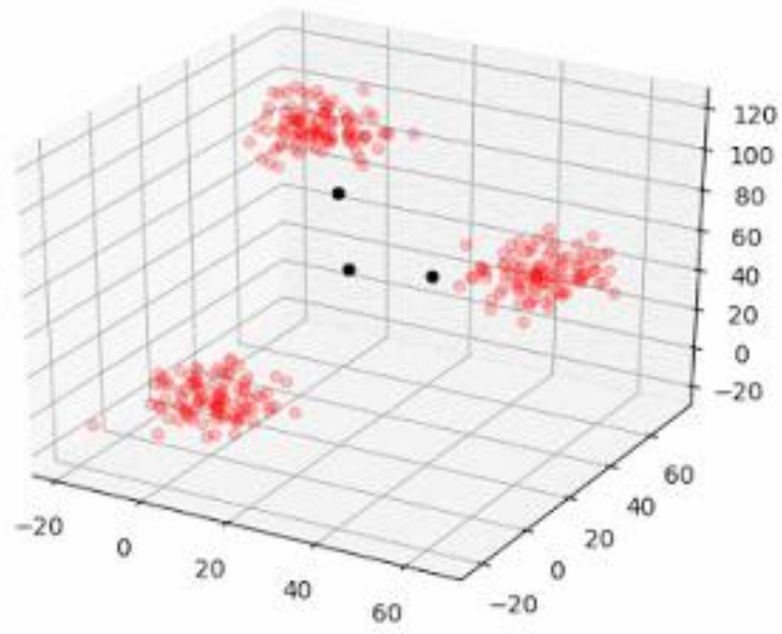
Unsupervised Learning

- Unsupervised learning is about problems where we don't have labelled answers, such as clustering, dimensionality reduction, and anomaly detection.

- Clustering: EM

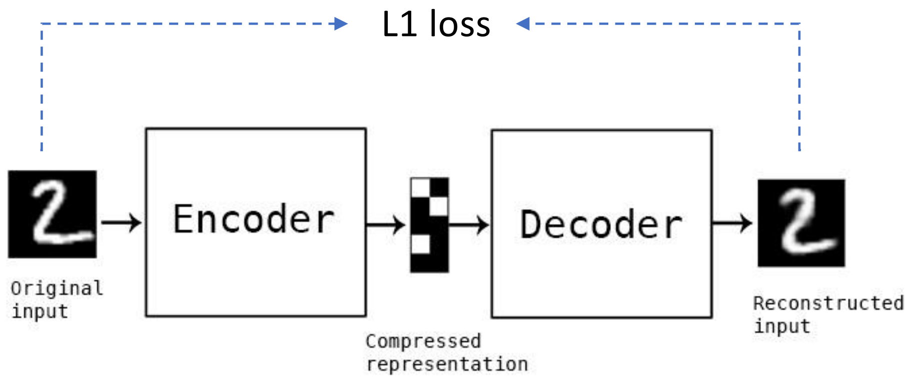
- Dimension Reduction: PCA

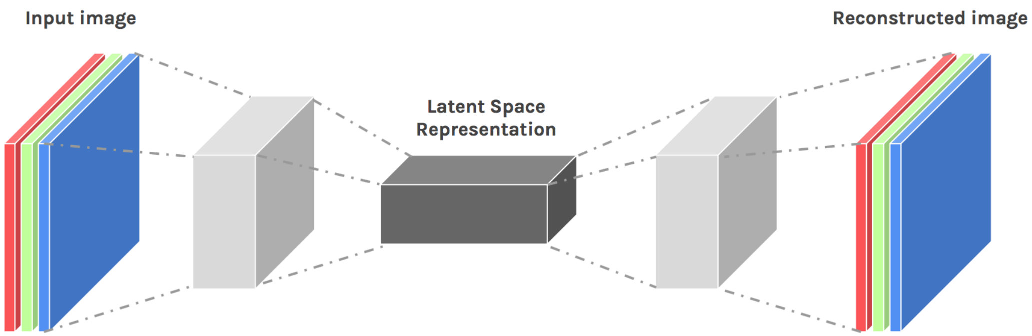
- ...



Unsupervised Learning

- **Autoencoder** (when the output is extracted features)

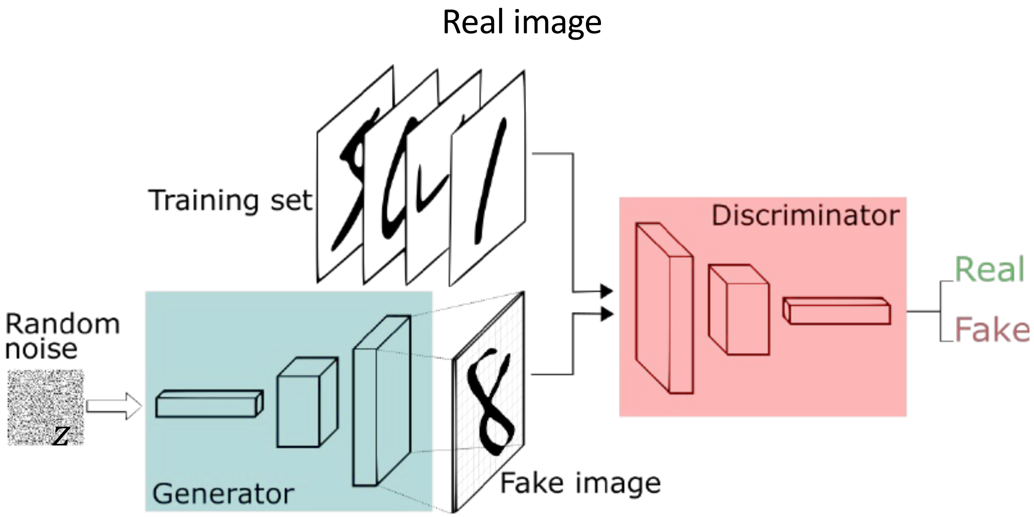


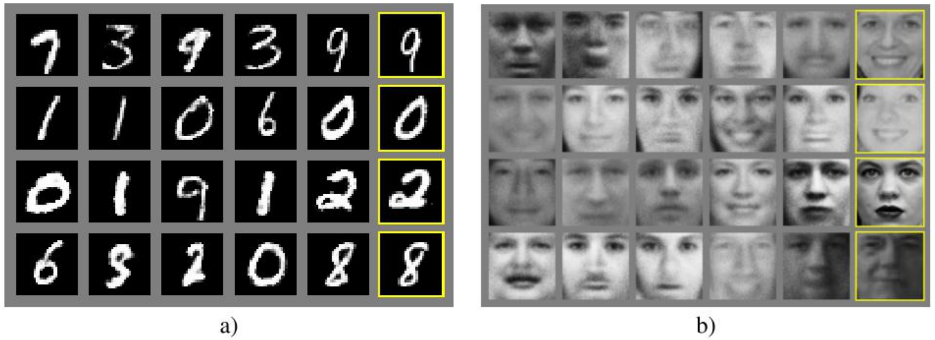


Autoencoder: Encode the input image x into a hidden state, then decode the latent space representation into a image $\bar{x}$. Then minimize the reconstruction loss between x and $\bar{x}$.

Unsupervised Learning

- **GANs**





Update the discriminator – ascending gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

Update the generator – descending gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

https://pathmind.com/wiki/generative-adversarial-network-gan

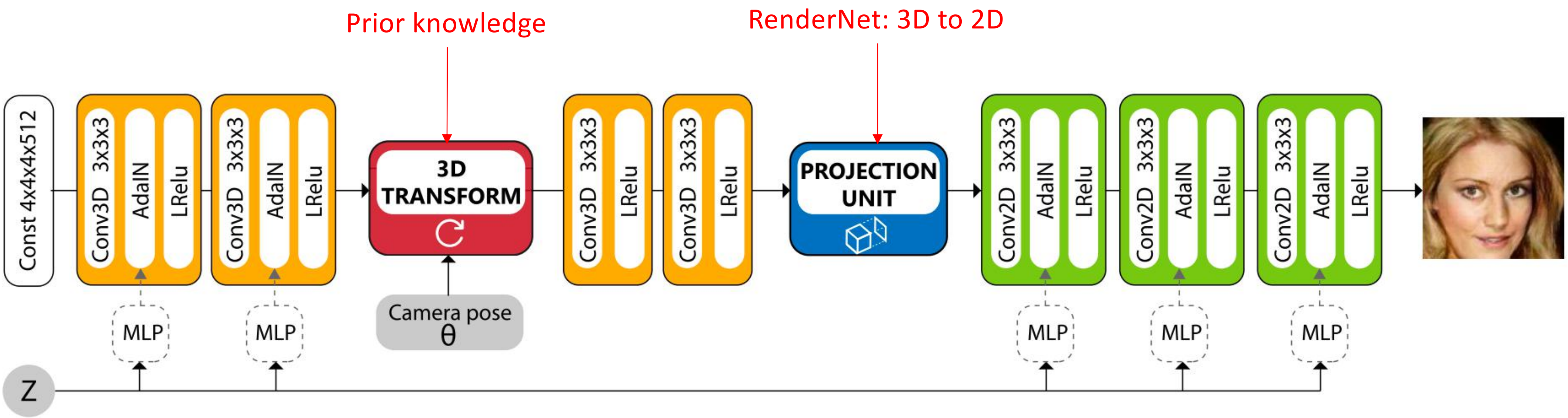
Unsupervised Learning

- **HoloGAN: learn the rotation concept**

HoloGAN: Unsupervised learning of 3D representations from natural images. NIPS 2019

Unsupervised Learning

- **HoloGAN: How it works**



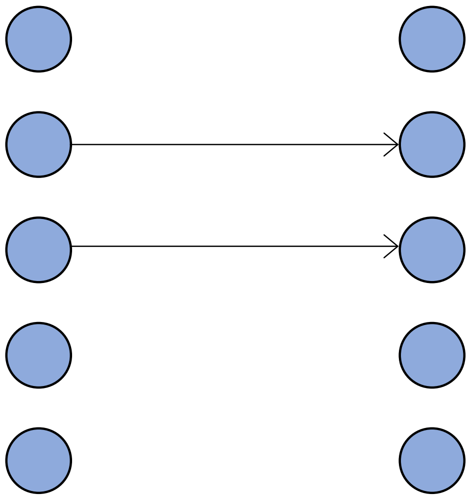
HoloGAN: Unsupervised learning of 3D representations from natural images. NIPS 2019

北京大學
PEKING UNIVERSITY

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

Semi-supervised Learning

Data in both input x and output y
with known partial mappings
(Learn the mapping f)



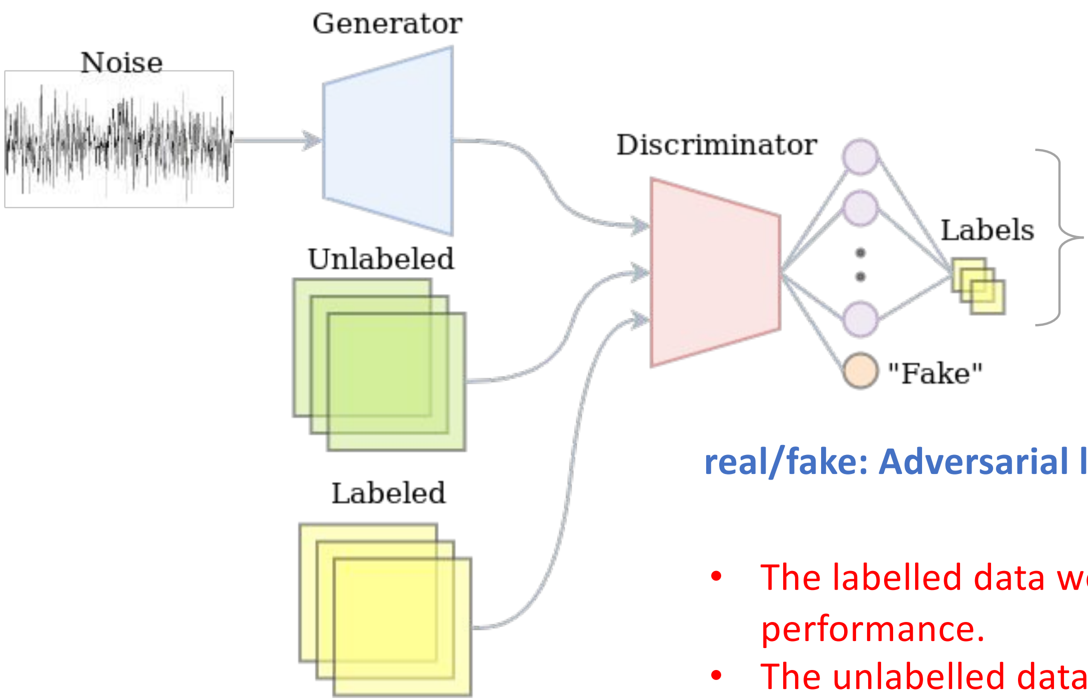
$y = f(x)$

Semi-supervised Learning

- **Motivation:**
  - Unlabelled data is easy to be obtained
  - Labelled data can be hard to get
- **Goal:**
  - Semi-supervised learning mixes labelled and labelled data to produce better models.
- **vs. Transductive Learning:**
  - Semi-supervised learning is eventually applied to the testing data
  - Transductive learning is only related to the unlabelled data

Semi-supervised Learning

- **Semi-supervised GAN**



Classify the real data **supervised learning**

(a Softmax output distribution for which we minimize the cross-entropy)

real/fake: Adversarial learning

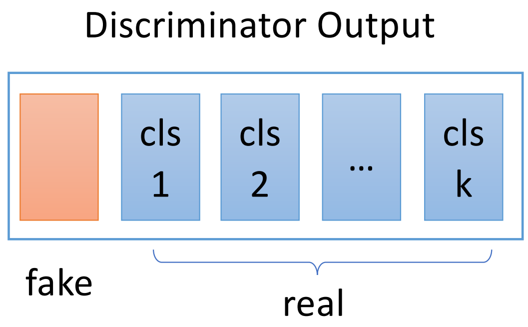
- The labelled data would be used to optimise the classification performance.
- The unlabelled data will be used to merely tell fake from real.

https://jostosh.github.io/ssl-gan/

PEKING UNIVERSITY 北京大學

Semi-supervised Learning

- **Semi-supervised GAN**
- Discriminator loss



the probability of it being real: $$p(x) = \frac{Z(x)}{Z(x) + \exp(l_{fake})} = \frac{Z(x)}{1 + Z(x)}$$

where $Z(x)$ is the sum of the unnormalised probabilities in the softmax operation.

$$\log\big(\mathrm{Z}(\mathrm{x})\big) = \mathrm{logsumexp}(l_1, \dots, l_k)$$

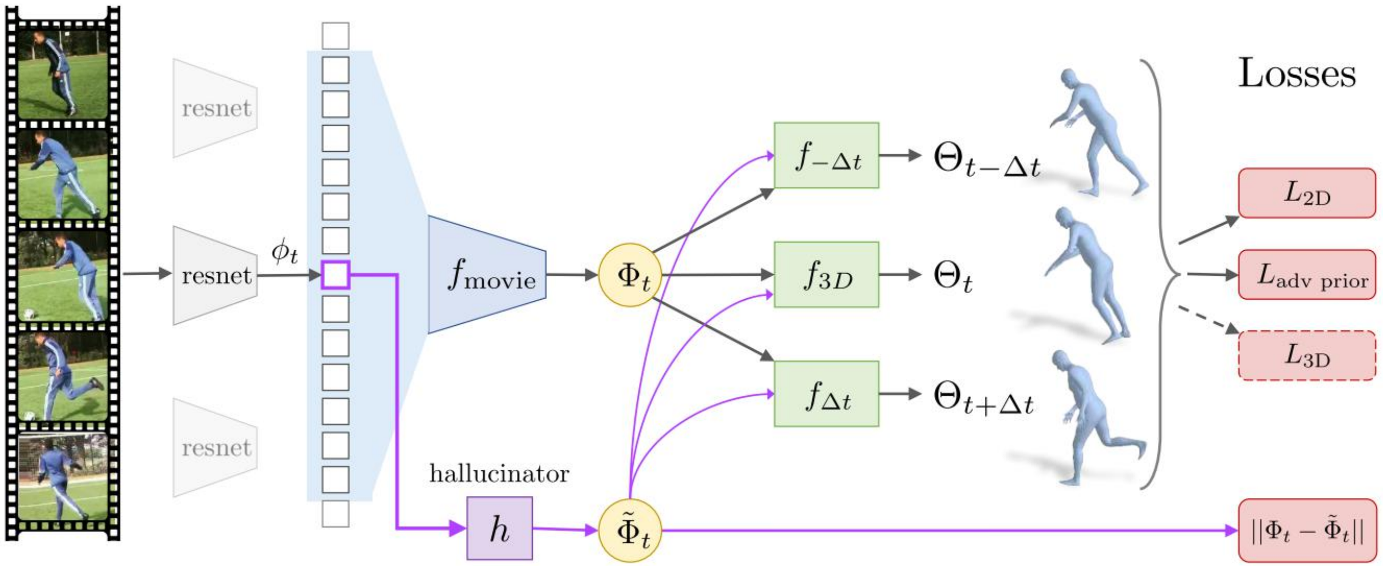
Gradient descent:
$$-\log(D(x)) - \log(1 - D(G(\boldsymbol{z})))$$
$$= -\log(\frac{Z(x)}{1 + Z(x)}) - \log(1 - \frac{Z(G(\boldsymbol{z}))}{1 + Z(G(\boldsymbol{z}))})$$

Semi-supervised Learning

- **Example: 2D Video to 3D shape**

The model can learn from videos with only 2D pose annotations in a semi-supervised manner.

L_{2D}, L_{3D} : supervision from ground-truth



$L_{adv\ prior}$: each prior discriminator judge a corresponding joint rotation of the body model

$$\sum_k (D_k(\mathbf{\Theta}) - 1)^2$$

make sure that the **hallucinator** can recover the current 3D mesh as well as its 3D past and future motion.

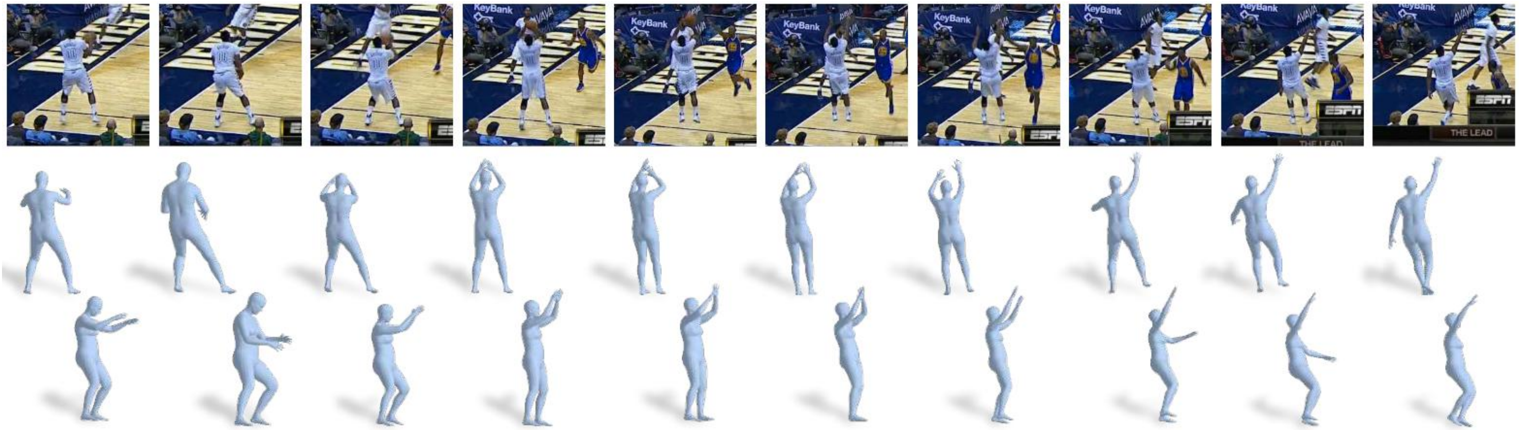
train a temporal encoder f_{movie} that learns a representation of 3D human dynamics Φ_t over the **temporal window centered at frame t**

Learning 3D Human Dynamics from Video. *A. Kanazawa, J. Zhang et al*. CVPR, 2019

Semi-supervised Learning

- **Example: 2D Video to 3D shape**

From a single image, the model can recover the current 3D mesh as well as its 3D past and future motion.



$$L_t = L_{\text{2D}} + L_{\text{3D}} + L_{\text{adv prior}} + L_{\beta \text{ prior}}$$

$$L_{\text{const shape}} = \sum_{t=1}^{T-1} ||\beta_t - \beta_{t+1}||. \qquad L_{\text{temporal}} = \sum_t L_t + \sum_{\Delta t} L_{t+\Delta t} + L_{\text{const shape}}.$$

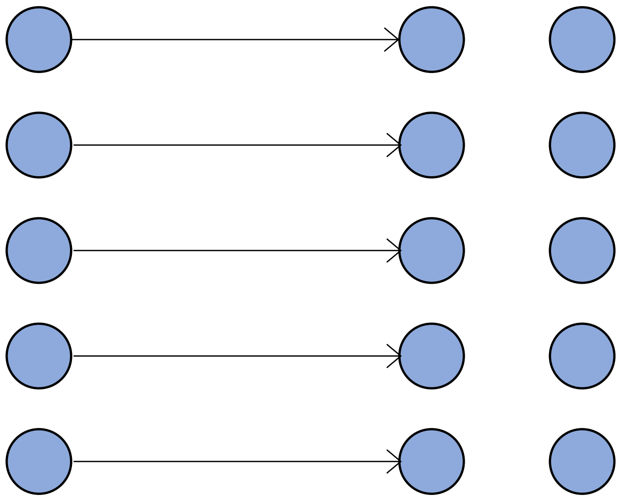
Learning 3D Human Dynamics from Video. *A. Kanazawa, J. Zhang et al*. CVPR, 2019

北京大學 PEKING UNIVERSITY

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

Weakly-supervised Learning

Data in both input x and output y
with known mapping for y
(Learn the mapping f for another output y')

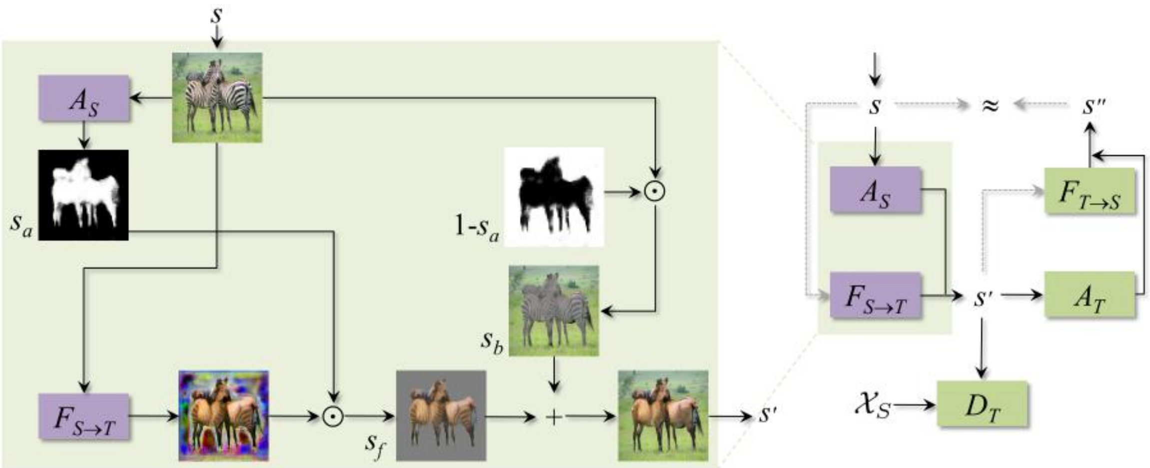


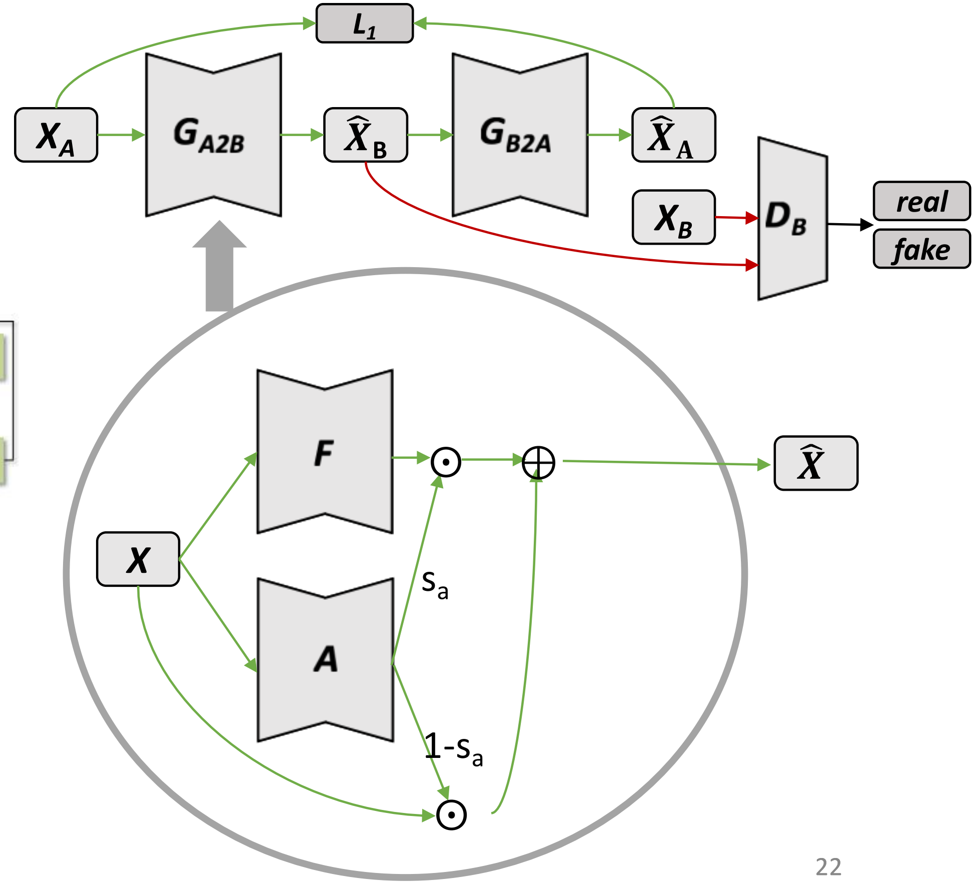
$y' = f(x)$
Weakly-supervised Learning

- Weakly supervised learning is a machine learning framework where the model is trained using examples that are only partially annotated or labeled.

Weakly-supervised Learning

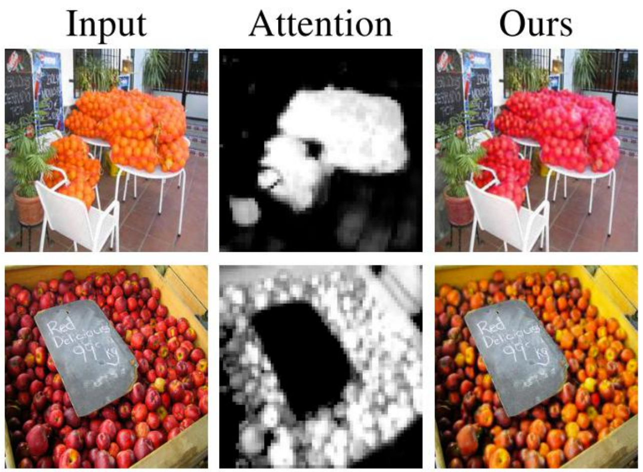
- **Attention CycleGAN**
- Learn the segmentation via synthesis





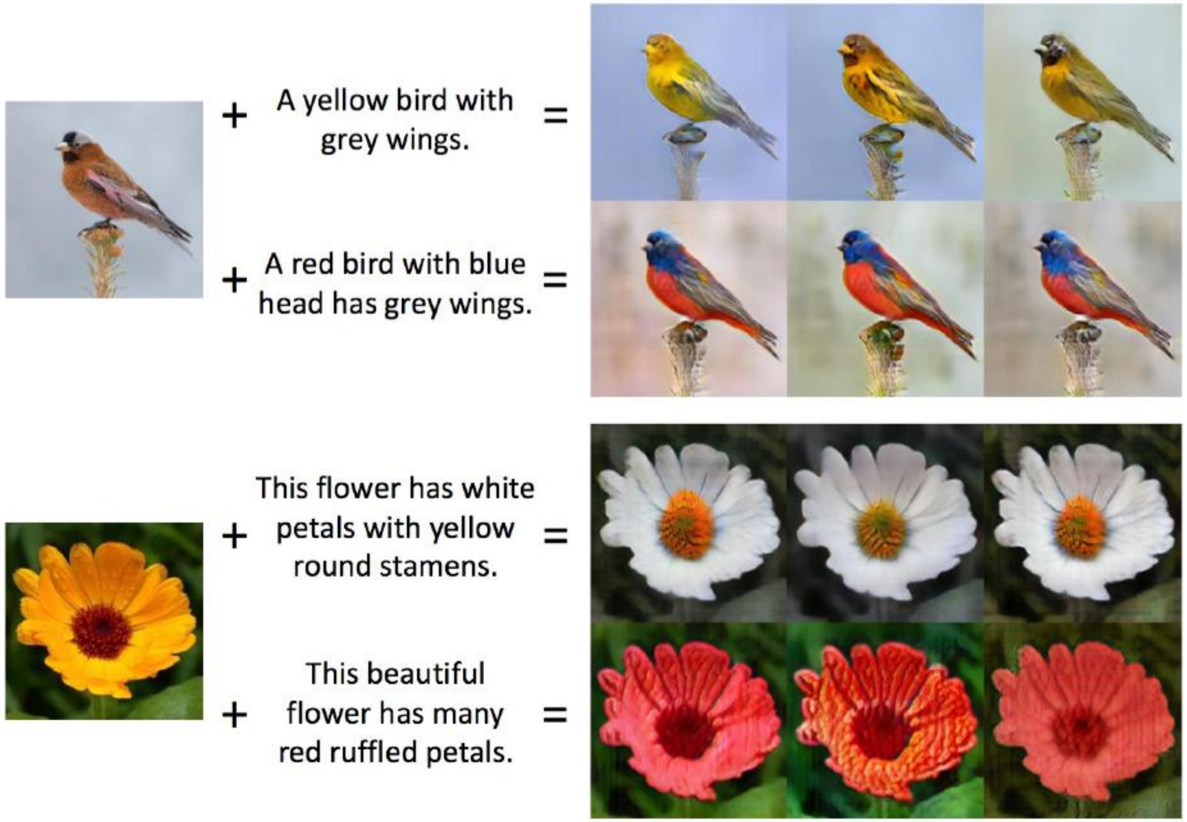
Weakly-supervised Learning

- **Attention CycleGAN**
  - Learn the segmentation without segmentation masks



Weakly-supervised Learning

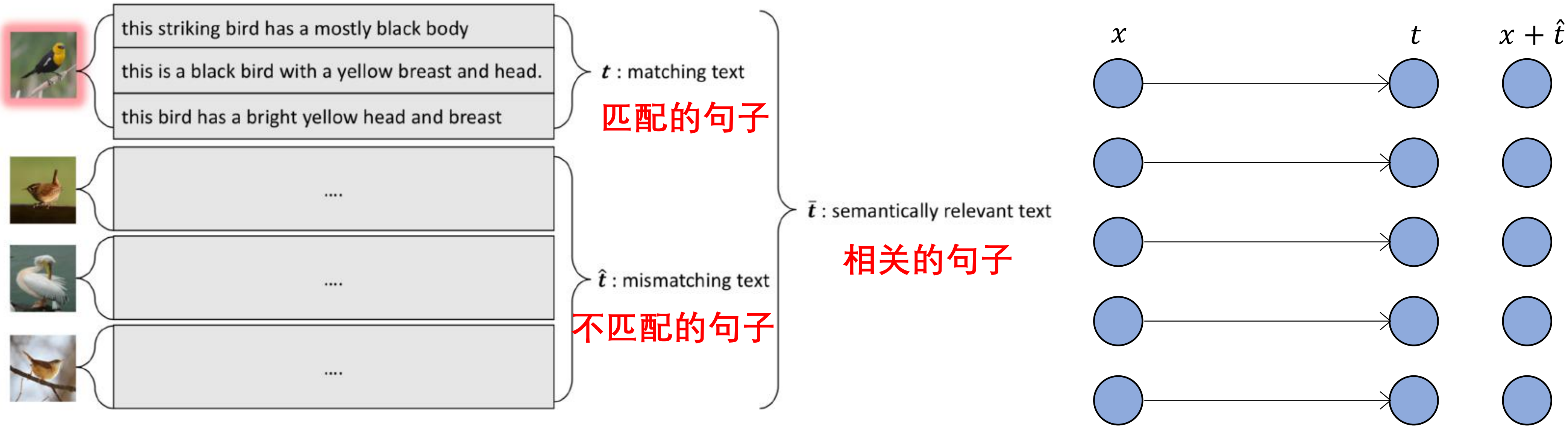
- **Semantic Image Synthesis: Language Image Manipulation**



Semantic Image Synthesis via Adversarial Learning. *H. Dong, S. Yu et al. ICCV 2017.*

Weakly-supervised Learning

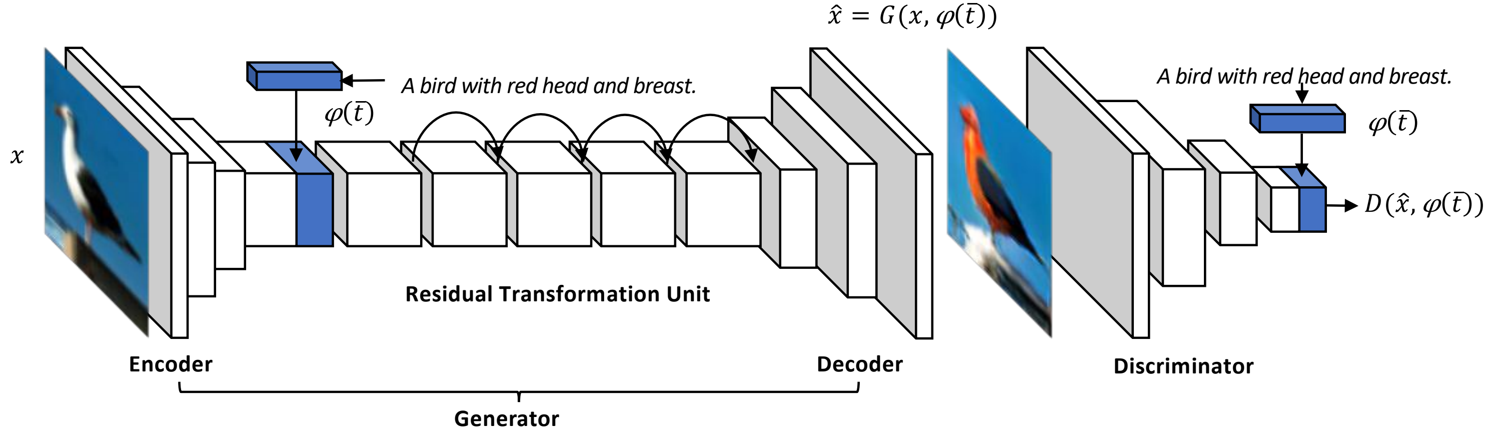
- **Semantic Image Synthesis: Language Image Manipulation**

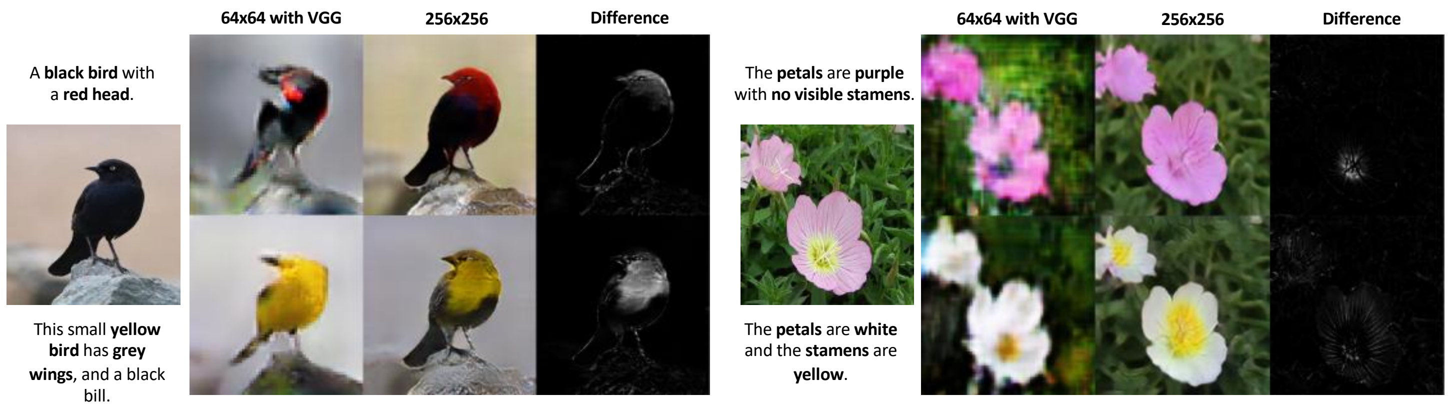


Semantic Image Synthesis via Adversarial Learning. *H. Dong, S. Yu et al. ICCV 2017.*

Weakly-supervised Learning

- **Semantic Image Synthesis: Learn the segmentation via synthesis**



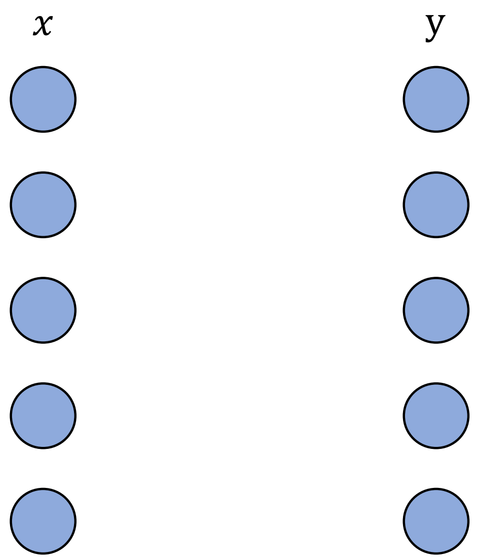


Semantic Image Synthesis via Adversarial Learning. *Dong, H., Yu, S., Wu, C., Guo, Y*. 2017. ICCV

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

Dual Learning

Data in both input and output
(Learn the mapping f, f')



$y = f(x), x = f'(y)$

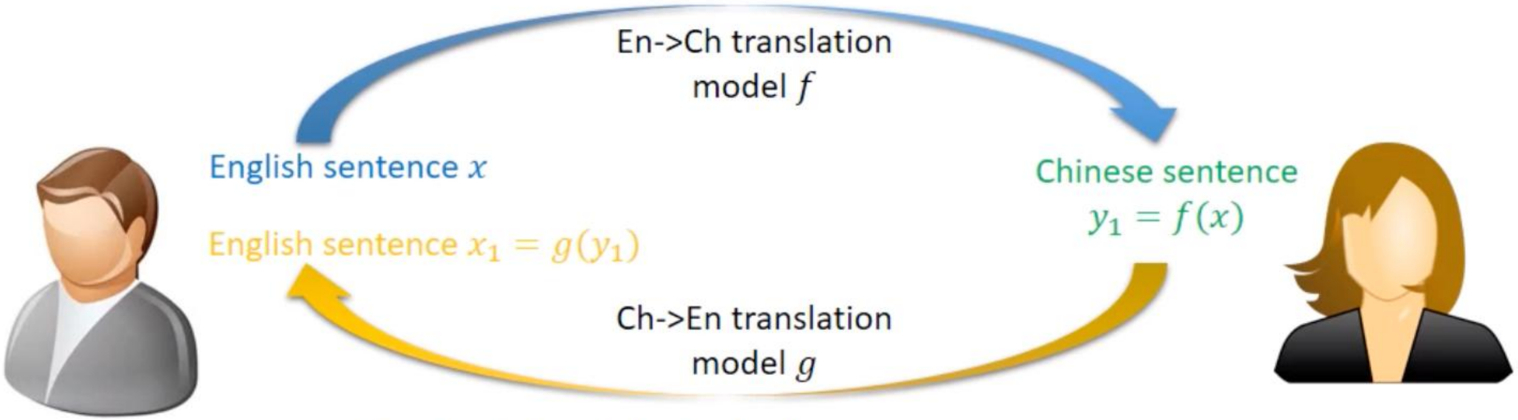
(Unsupervised) Dual Learning

- Motivation
  - Human label is expensive
  - No feedback if using unlabeled data

Application	Primal Task	Dual (Inverse) Task
Machine translation	Translate language from A to B	Translate language from B to A
Speed processing	Speech to text (STT)	Text to speech (TTS)
Image understanding	Image captioning	Image generation
Conversation engine	Question	Answer
Search engine	Search	Query

Dual Learning

- **Language Translation**



Feedback signals during the loop:

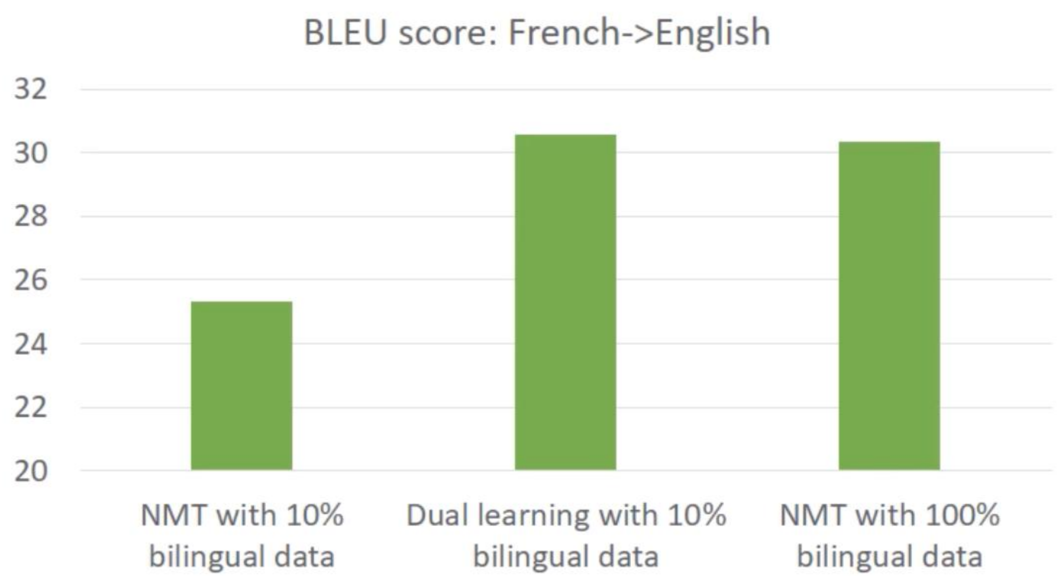
- $s(x, x_1)$: BLEU score of x_1 given x
- $L(y)$ and $L(x_1)$: Likelihood and language model of y_1 and x_1

Reinforcement learning is used to improve the translation models from these feedback signals

Dual Learning for Machine Translation.
Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, Wei-Ying Ma. NIPS, 2016

Dual Learning

- **Language Translation**

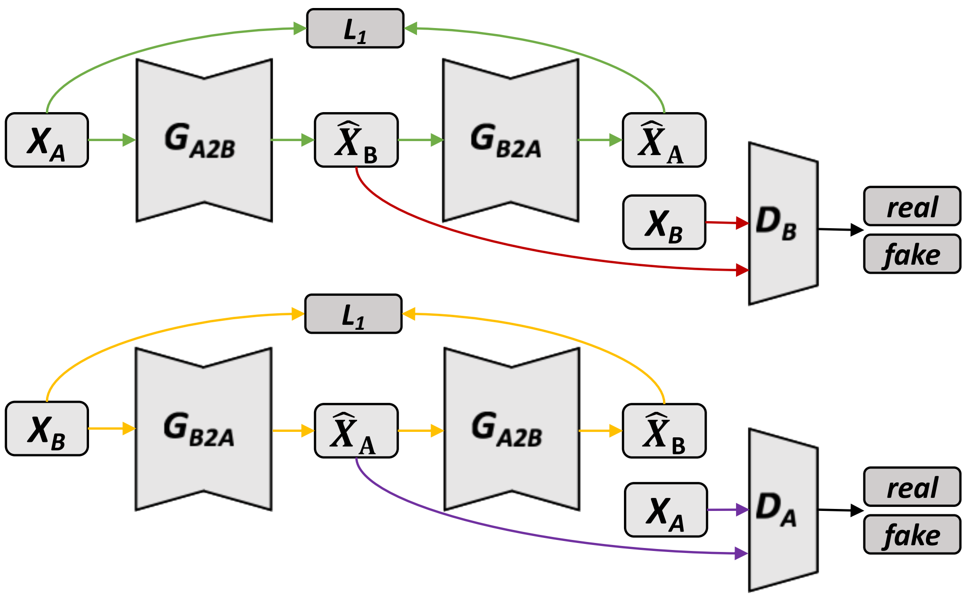


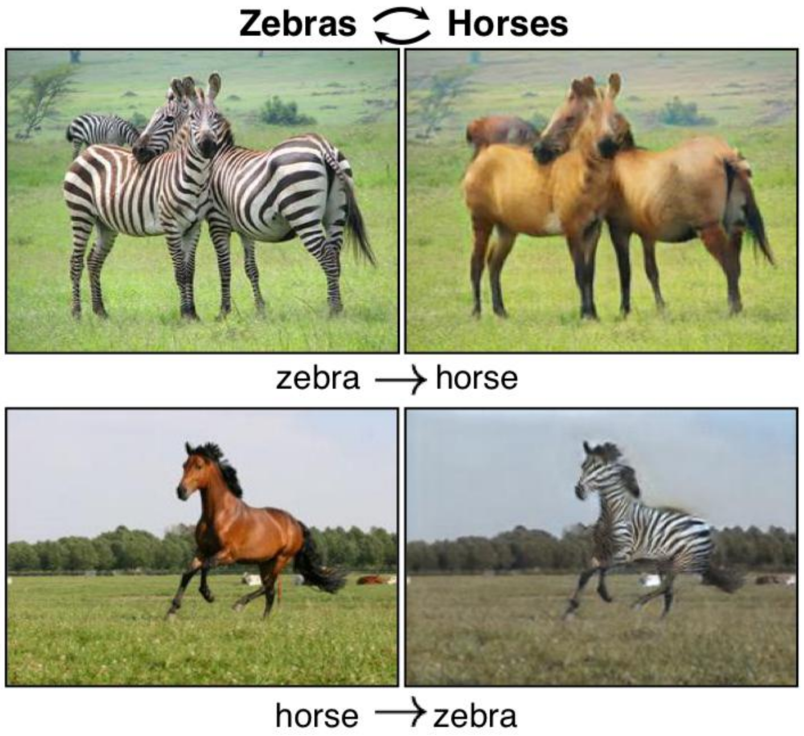
Starting from initial models obtained from only 10% bilingual data, dual learning can achieve similar accuracy as the NMT model learned from 100% bilingual data!

Dual Learning for Machine Translation.
Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, Wei-Ying Ma. NIPS, 2016

Dual Learning

- **Unpaired Image-to-Image Translation**



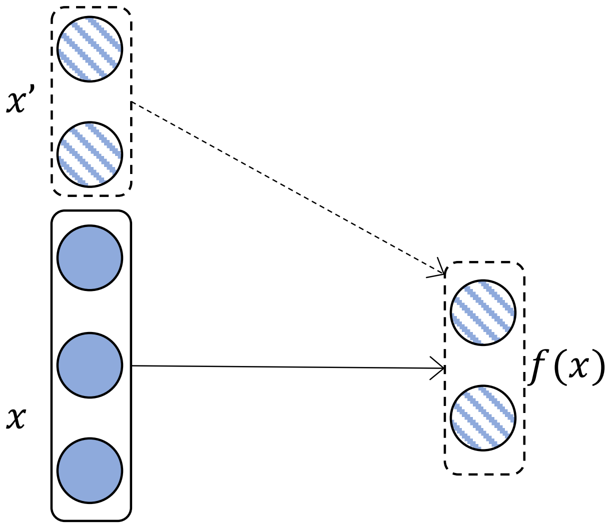


Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *J. Zhu, T. Park et al. ICCV 2017.*

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

Self-supervised Learning

Data in input x, x' only
with known mapping f'
(Learn the mapping f)



$x' = f(x)$
Self-supervised Learning

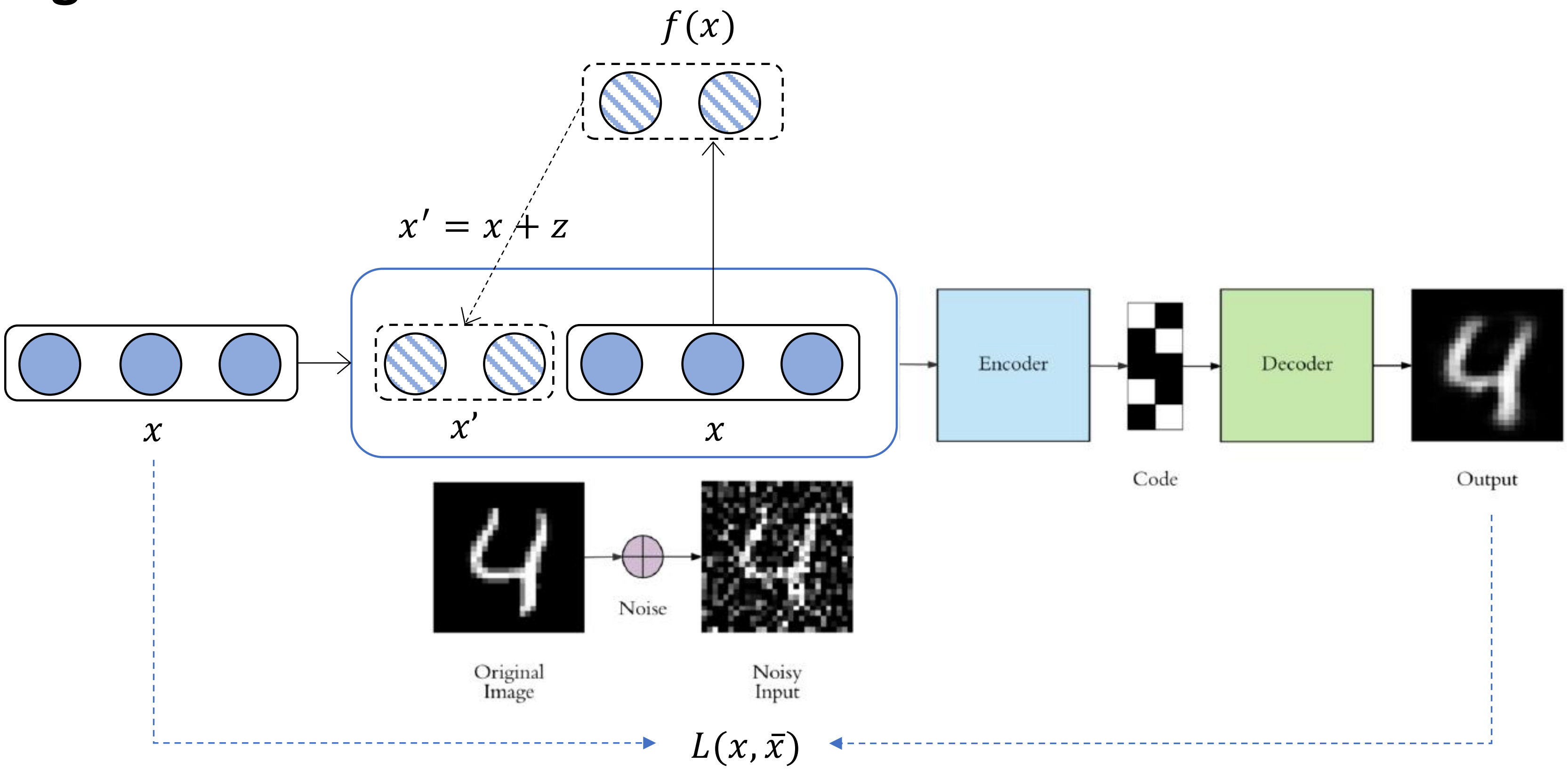
- Self-supervised learning is autonomous supervised learning, it learns to predict part of its input from other parts of its input.

- Examples: Word2Vec, Denoising Autoencoder

- Self-supervised vs. unsupervised learning: Self-supervised learning is like unsupervised Learning because the system learns without using explicitly-provided labels. It is different from unsupervised learning because we are not learning the inherent structure of data. Self-supervised learning, unlike unsupervised learning, is not centered around clustering and grouping, dimensionality reduction, recommendation engines, density estimation, or anomaly detection.

Self-supervised Learning

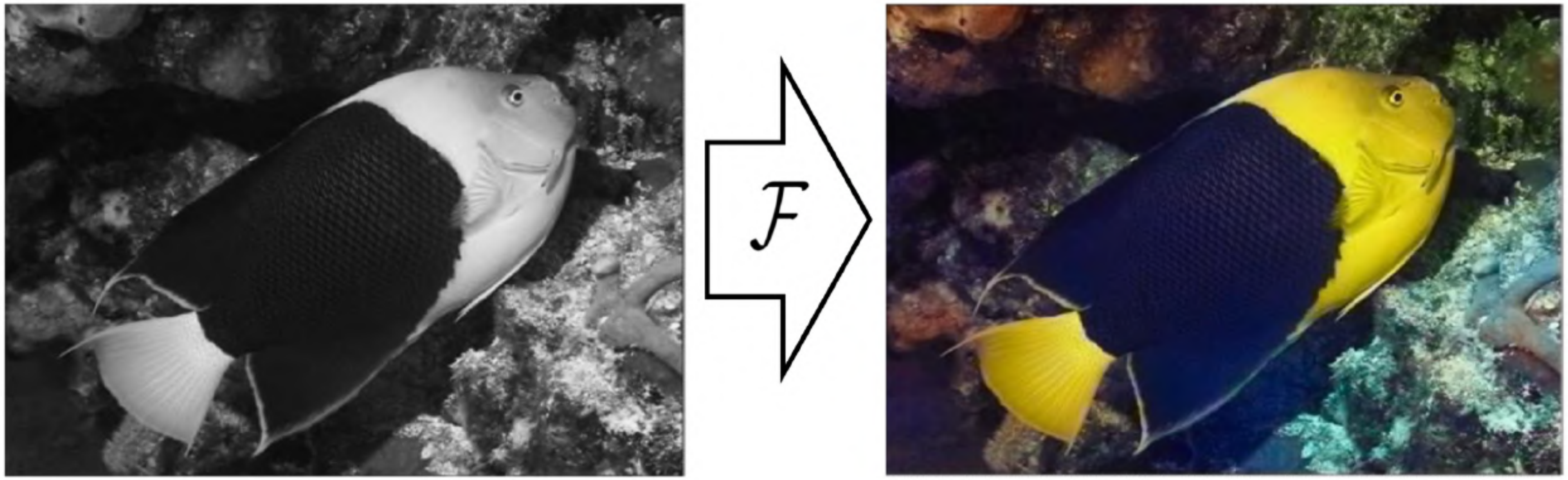
- **Denoising Autoencoder**



Extracting and composing robust features with denoising autoencoders, Pascal Vincent etc, 2008

Self-supervised Learning

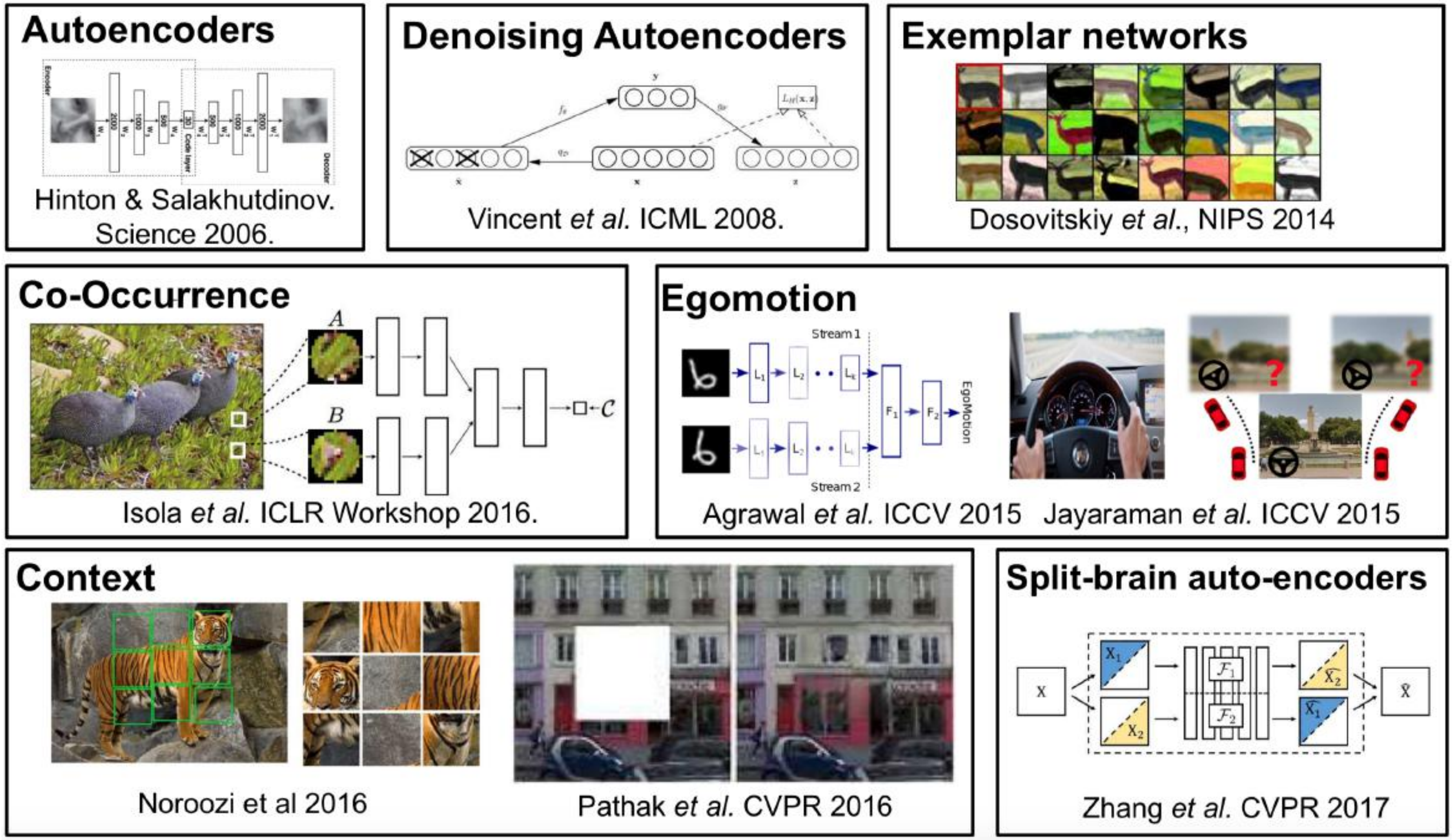
- **Image Example: Colorisation**



Colorful Image Colorization. *Zhang et al.,* ECCV 2016

Self-supervised Learning

- Image Examples



Self-supervised Learning

- **Video Example**

- Videos contain
  - Colour, Temporal info
- Possible proxy tasks
  - Temporal order of the frames
  - Optical flow: Motion of objects
  - …

Self-supervised Learning

- Video Example: Shuffle and Learn

Given a start and an end, can this point lie in between?

Unsupervised Learning using Temporal Order Verification. *Ishan Misra, C. Lawrence Zitnick and Martial Hebert.* ECCV, 2016.

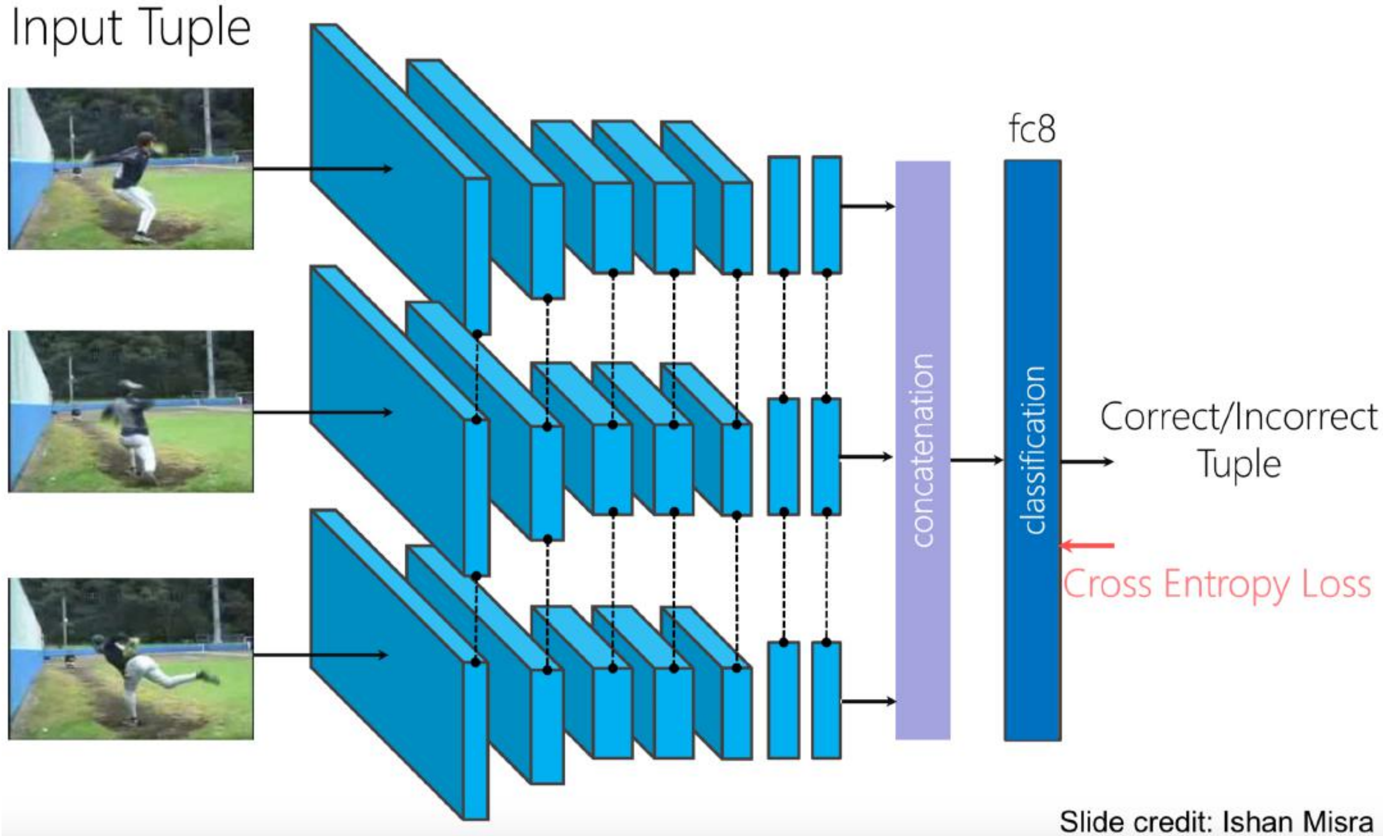
Self-supervised Learning

- Video Example: Shuffle and Learn

Unsupervised Learning using Temporal Order Verification. *Ishan Misra, C. Lawrence Zitnick and Martial Hebert.* ECCV, 2016.

Self-supervised Learning

- Video Example: Shuffle and Learn

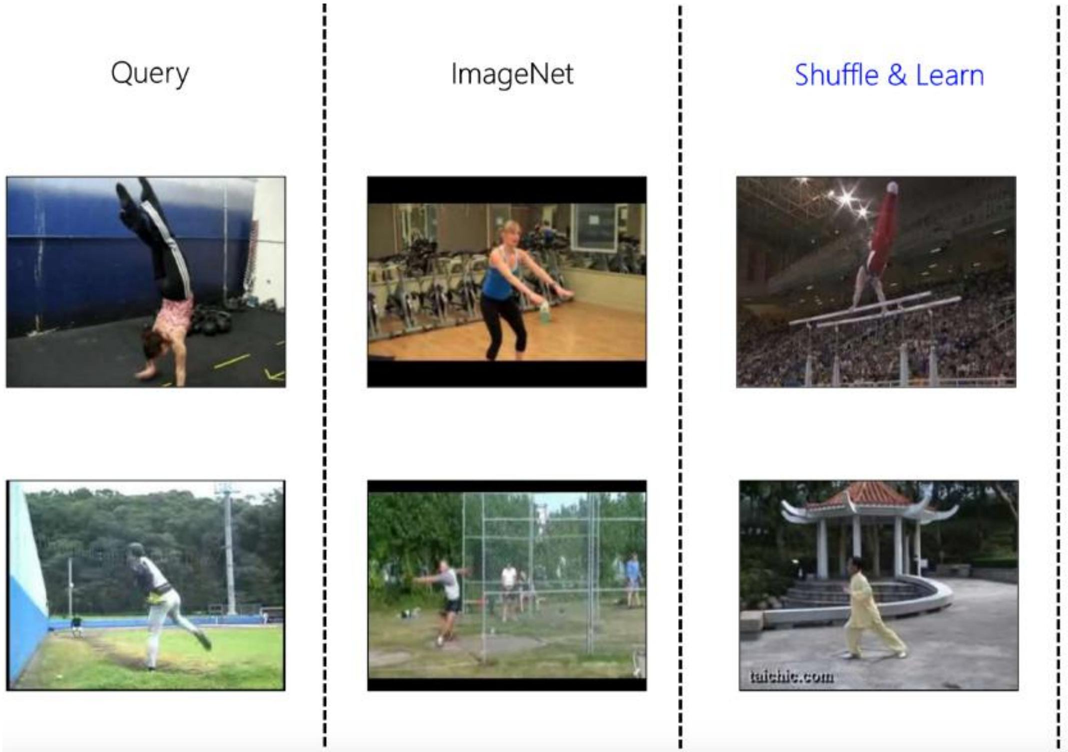


Unsupervised Learning using Temporal Order Verification. *Ishan Misra, C. Lawrence Zitnick and Martial Hebert.* ECCV, 2016.

Self-supervised Learning

- Video Example: Shuffle and Learn

Image Retrieval: Nearest Neighbors of Query Frame (FC5 outputs)



Unsupervised Learning using Temporal Order Verification. *Ishan Misra, C. Lawrence Zitnick and Martial Hebert.* ECCV, 2016.

Self-supervised Learning

- Video Example: Shuffle and Learn

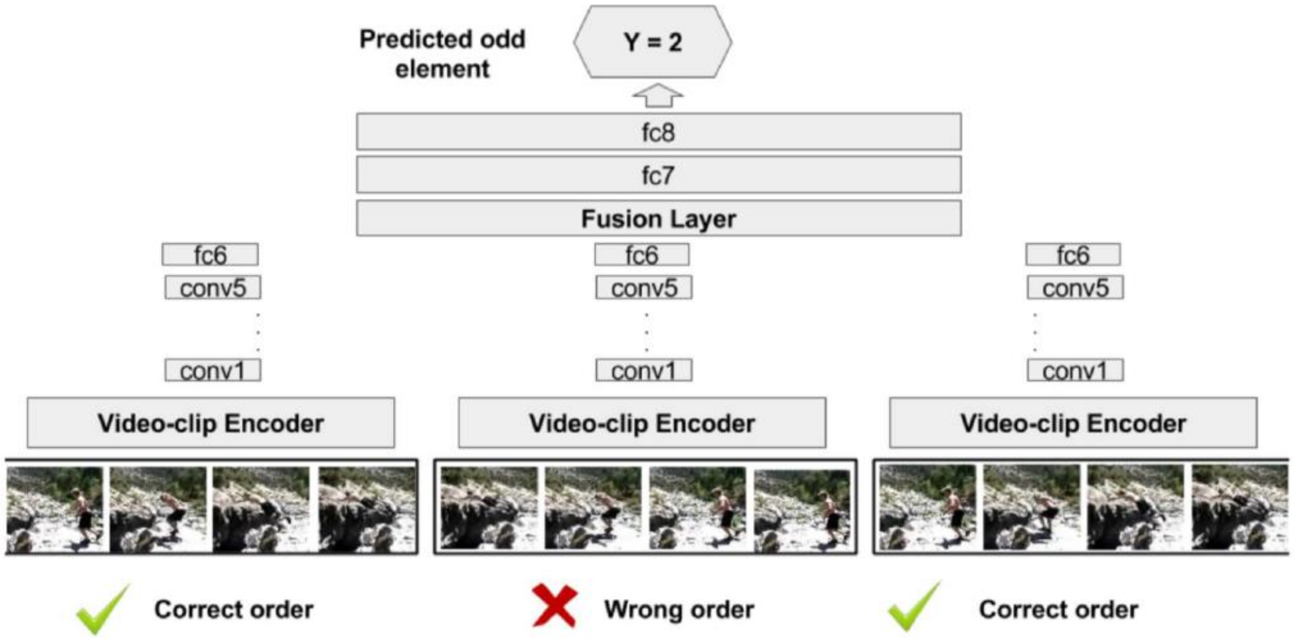
Action Recognition: Self-supervised pre-train + Finetune

Dataset	Initialization	Mean Classification Accuracy
UCF101	Random	38.6
	Shuffle & Learn	50.2
	ImageNet pre-trained	**67.1**

Unsupervised Learning using Temporal Order Verification. *Ishan Misra, C. Lawrence Zitnick and Martial Hebert.* ECCV, 2016.

Self-supervised Learning

- Video Example: Odd-One-Out



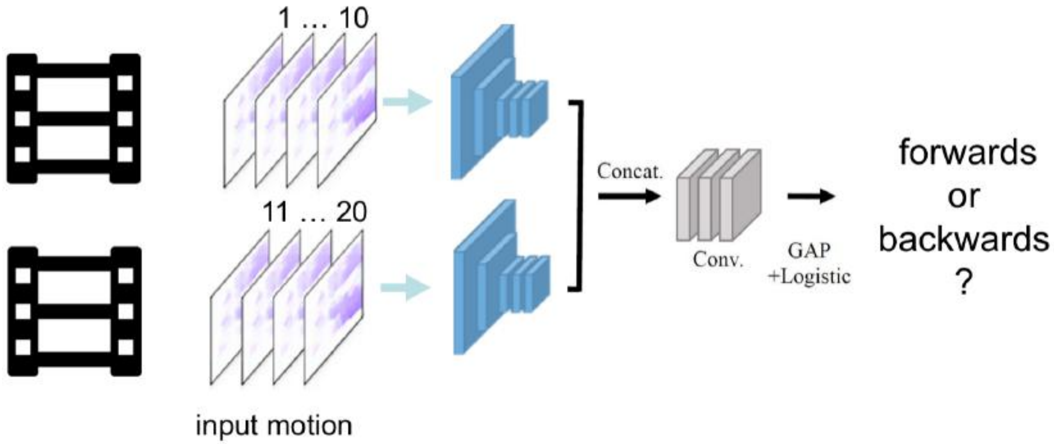
Initialization	Mean Classification Accuracy
Random	38.6
Shuffle and Learn	50.2
Odd-One-Out	60.3
ImageNet pre-trained	**67.1**

Self-Supervised Video Representation Learning With Odd-One-Out Networks. *Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould*, ICCV 2017

Self-supervised Learning

- Video Example: Learning the Arrow of Time

Forward or backward plays?



- Depending on the video, solving the task may require
(a) low-level understanding (e.g. physics)
(b) high-level reasoning (e.g. semantics)
(c) familiarity with very subtle effects
(d) camera conventions

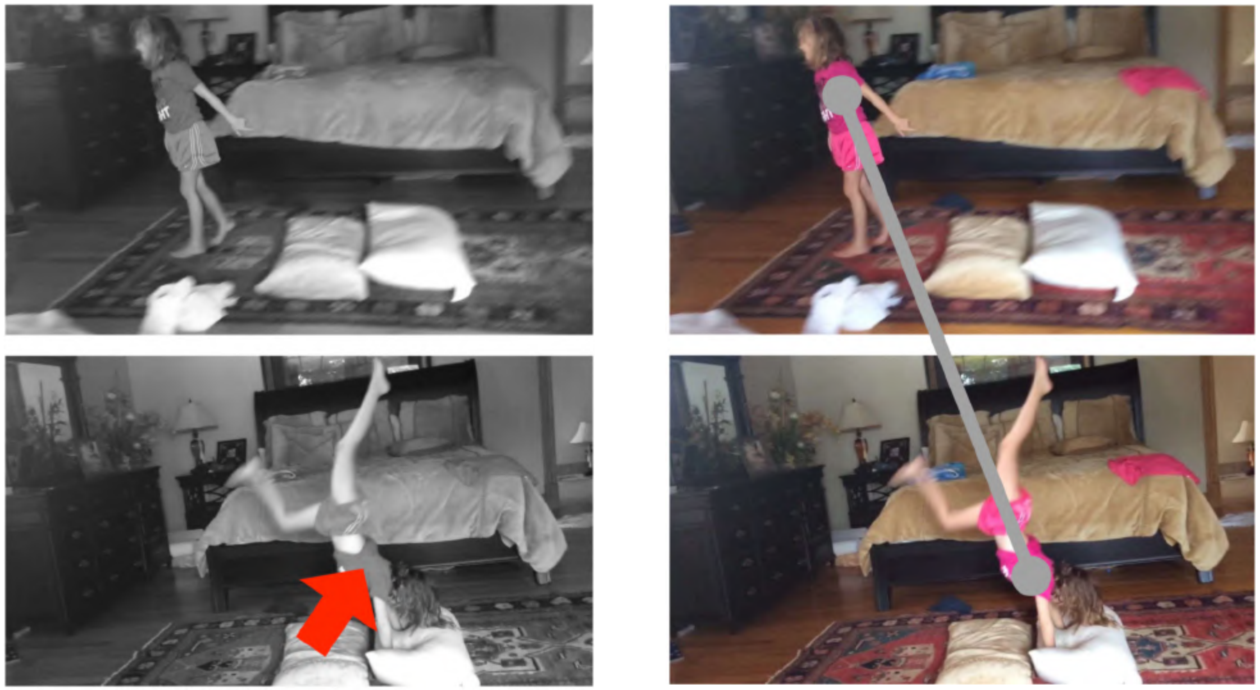
- Input: optical flow in two chunks
- Final layer: global average pooling to allow class activation map (CAM)

Learning and Using the Arrow of Time. *Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman.* CVPR 2018

Self-supervised Learning

- Video Example: Temporal Coherence of Color

Colorize all frames of a grey scale version using a reference frame



Reference Frame

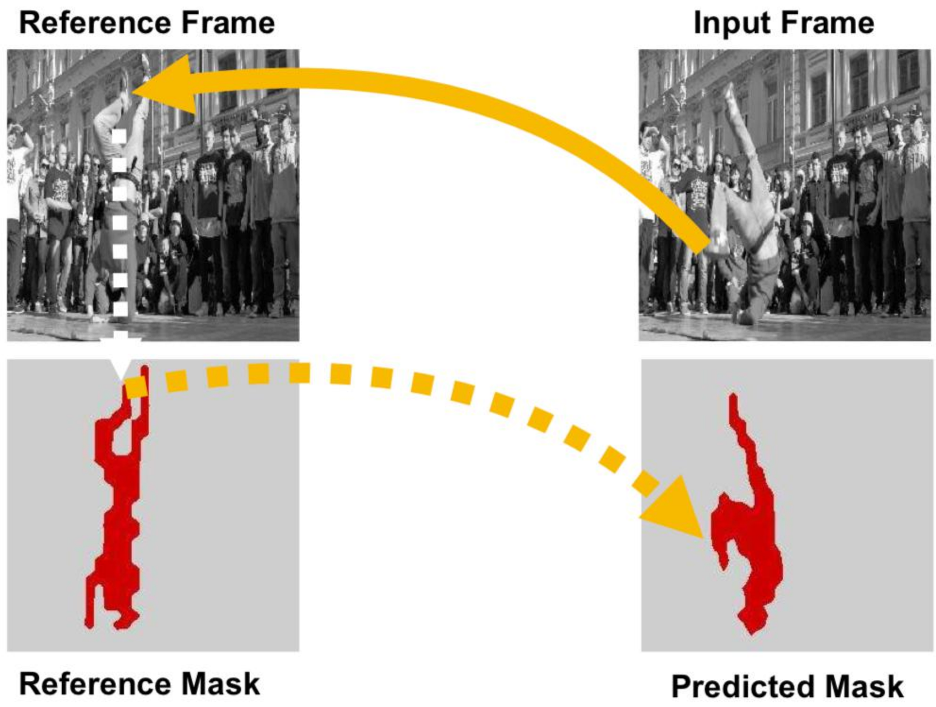
What color is that?

Tracking Emerges by Colorizing Videos
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy, ECCV 2018

Self-supervised Learning

- Video Example: Temporal Coherence of Color

Tracking Emerges: Only the first frame is given, colors indicate different instances



Tracking Emerges by Colorizing Videos
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy, ECCV 2018

Self-supervised Learning

- Video Example: Temporal Coherence of Color

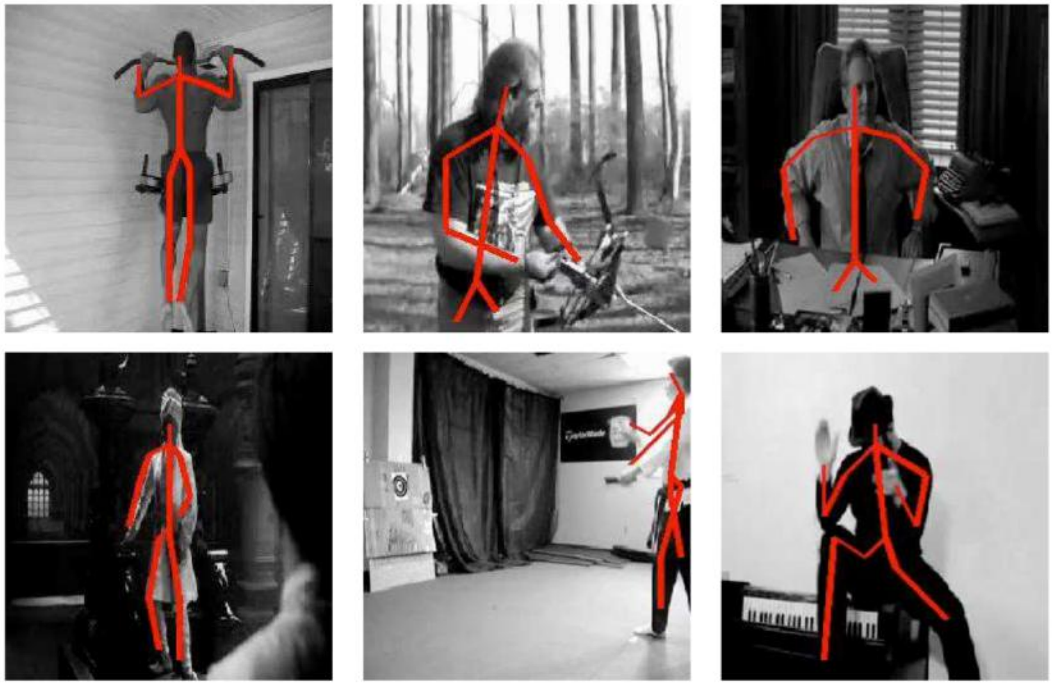
Segment Tracking: Only the first frame is given, colors indicate different instances

Tracking Emerges by Colorizing Videos
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy, ECCV 2018

Self-supervised Learning

- Video Example: Temporal Coherence of Color

Pose Tracking: Only the skeleton in the first frame is given

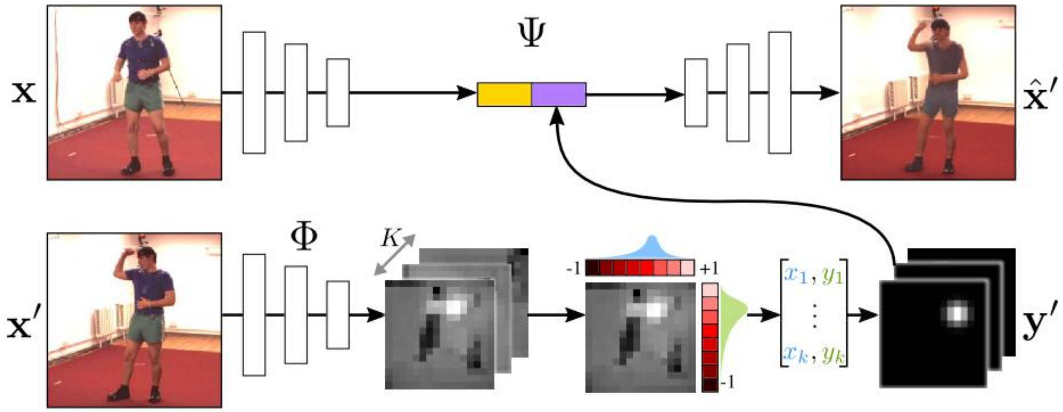


Tracking Emerges by Colorizing Videos
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy, ECCV 2018

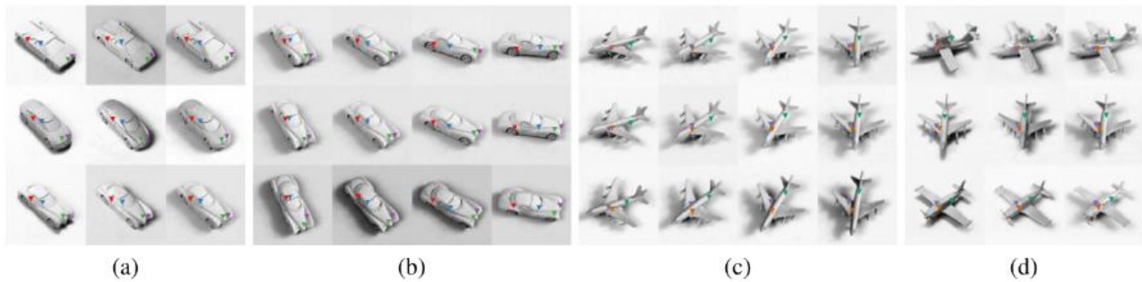
Self-supervised Learning

- Video Example: Temporal Coherence of Color

Unsupervised Key-point Detection: Only paired images of the same object is given



- Achieve retargeting
- Disentangling Style and Geometry
- Invariant Localization



Unsupervised Learning of Object Landmarks through Conditional Image Generation
Tomas Jakab, Ankush Gupta et al. NIPS, 2018.

Self-supervised Learning

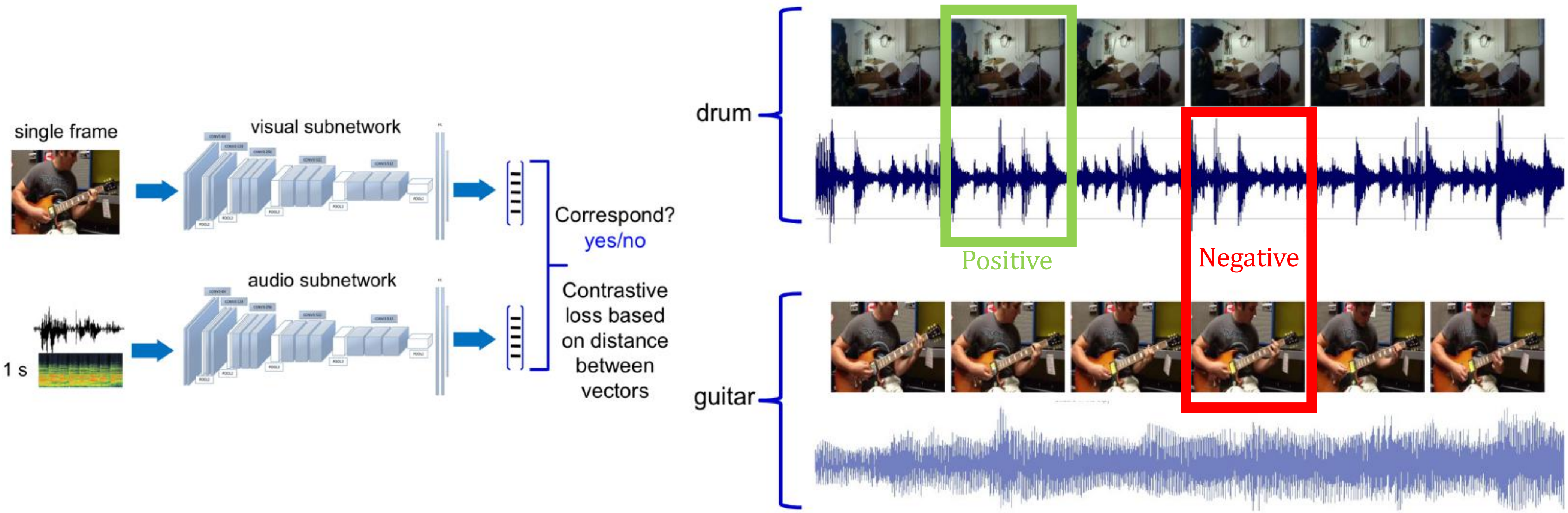
- **Video + Sound Example**

- Sound and frames are:
  - Semantically consistent
  - Synchronized
- Two types of proxy task:
  - Predict audio-visual correspondence
  - Predict audio-visual synchronization

Self-supervised Learning

- Video + Sound Example: Audio-Visual Co-supervision

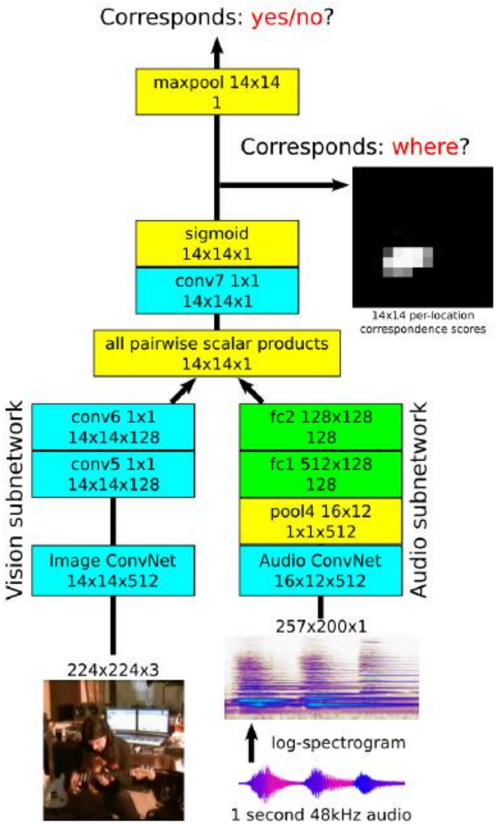
Train a network to predict if image and audio clip correspond



Objects that Sound. *Arandjelović and Zisserman*, ICCV 2017 & ECCV 2018

Self-supervised Learning

- Video + Sound Example: Audio-Visual Co-supervision



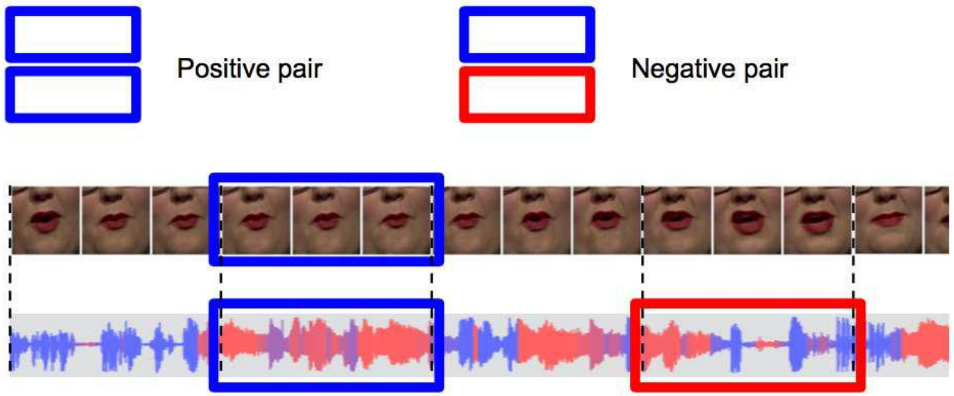
- Learn good visual features
- Learn good audio features
- Learn aligned audio-visual embeddings
- Learn to localize objects that sound

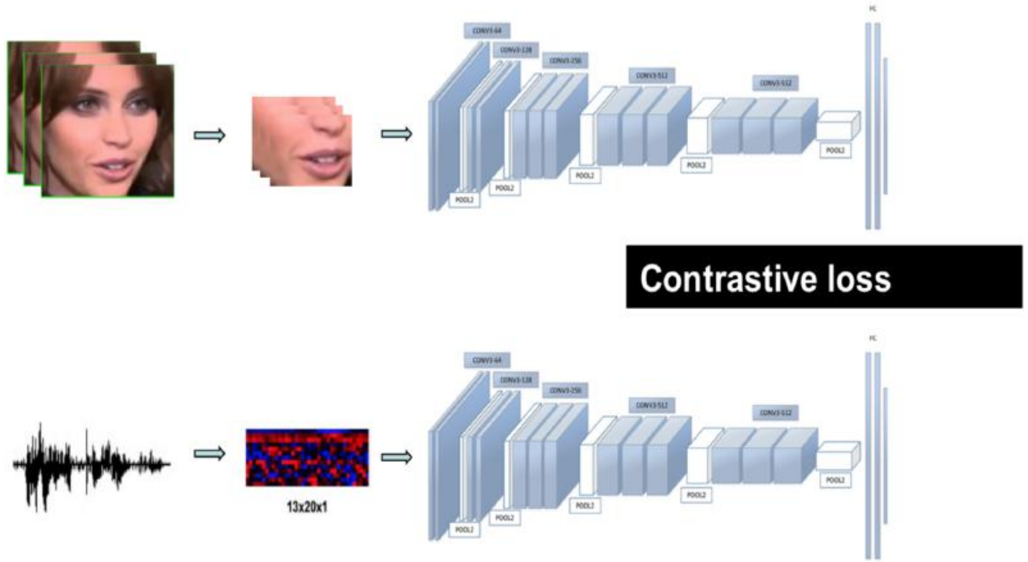
- Using learned features
  - Sound classification
  - Query on image to retrieve audio
  - Localizing objects with sound

Objects that Sound. *Arandjelović and Zisserman (DeepMind, Ox)*, ICCV 2017 & ECCV 2018

Self-supervised Learning

- Video + Sound Example: Audio-Visual Co-supervision





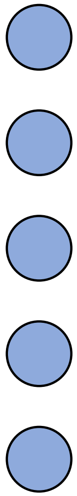
- Applications
  - Active speaker detection
  - Audio-to-video synchronization
  - Voice-over rejection
  - Visual features for lip reading

Out of time: Automatic lip sync in the wild. *Chung, Zisserman,* 2016

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

Self-augmented Learning

Data in input only
with known inverse mapping f'
(Learn the mapping f and output y)

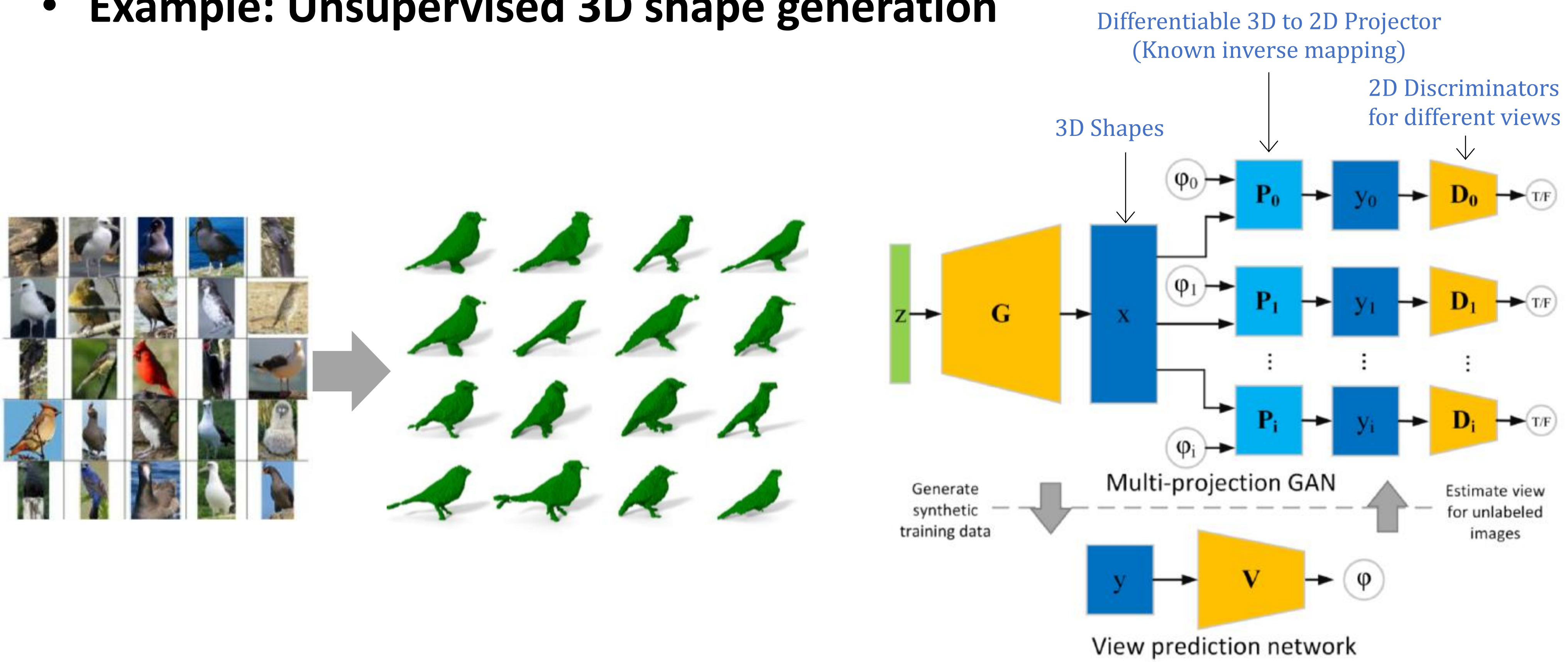


$y = f(x), x = f'(y)$

Self-augmented Learning

Self-augmented Learning

- **Example: Unsupervised 3D shape generation**



Synthesizing 3D Shapes from Unannotated Image Collections using Multi-projection Generative Adversarial Networks.
Xiao Li, Yue Dong, Pieter Peers, Xin Tong. CVPR, 2019

Summary

- Unsupervised Learning
- Semi-supervised Learning
- Weakly-supervised Learning
- Dual Learning
- Self-supervised Learning
- Self-augmented Learning

Thanks