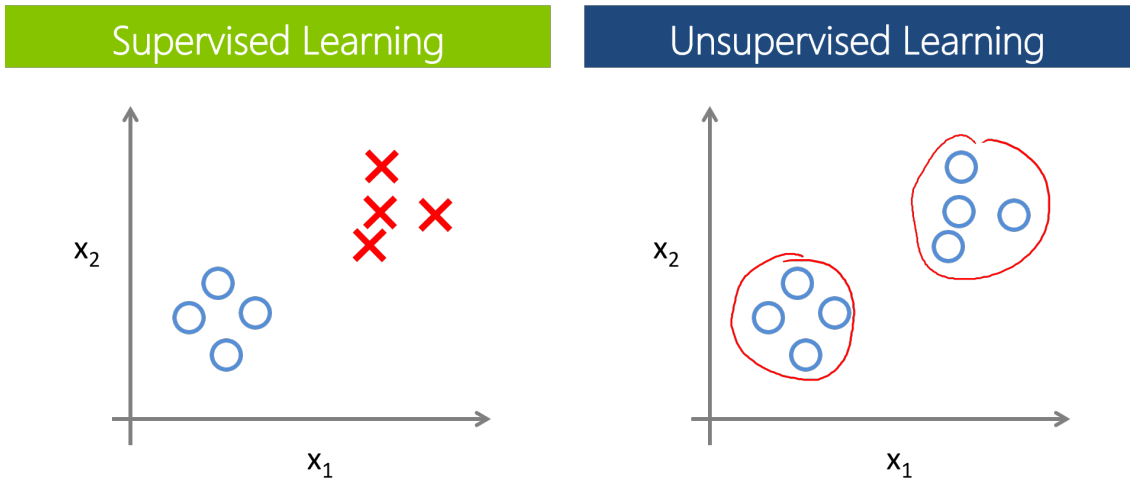
CAP 5516
Medical Image Computing
(Spring 2022)

Dr. Chen Chen
Center for Research in Computer Vision (CRCV)
University of Central Florida
Office: HEC 221
Address: 4328 Scorpius St., Orlando, FL 32816-2365
Email: chen.chen@crcv.ucf.edu
Web: https://www.crcv.ucf.edu/chenchen/

Lecture 11
Self-supervised Learning

Paradigm of Learning

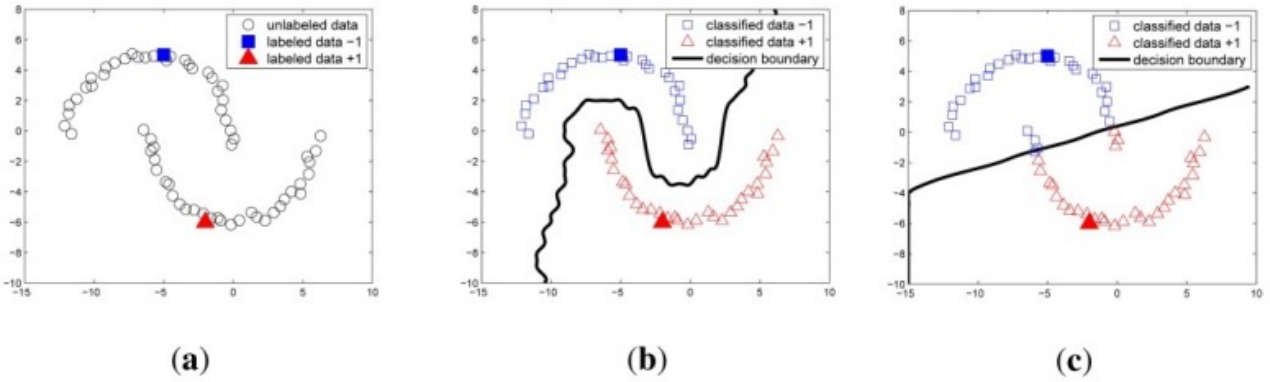
- Supervised Learning & Unsupervised Learning
  - Given desired output vs. No guidance at all



http://oliviaklose.azurewebsites.net/content/images/2015/02/2-supervised-vs-unsupervised-1.png

Paradigm of Learning

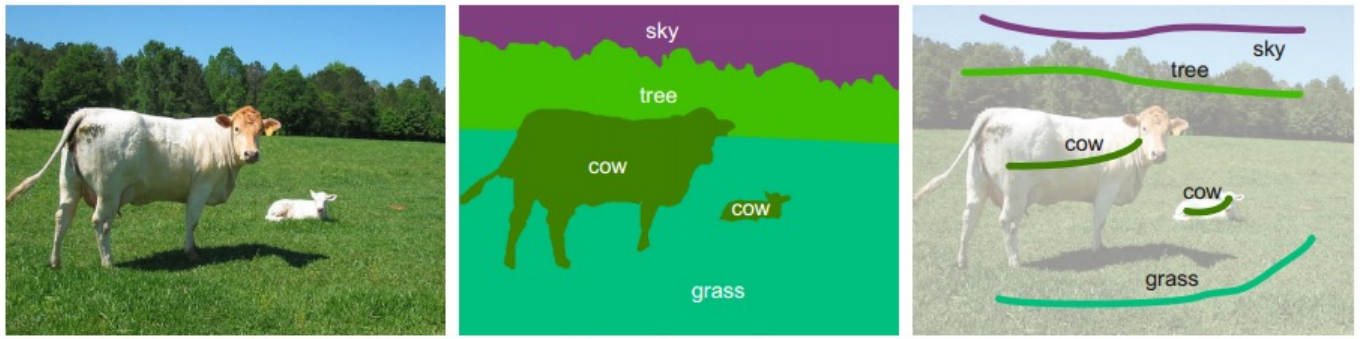
- In Between…
  - Semi-Supervised Learning
    - A small amount of labeled data in conjunction with a large amount of unlabeled data.



https://openi.nlm.nih.gov/imgs/512/371/4299091/PMC4299091_sensors-14-23871f4.png

Paradigm of Learning

- In Between…
  - Weakly-Supervised Learning
    - Use somewhat coarse or inaccurate supervision, e.g.
      - Given image level label, infer object level bounding box/ pixel level segmentation
      - Given video level label, infer image level label
      - Given scribble, infer the full pixel level segmentation
      - Given bounding box, infer the boundary of object



(a) image (b) mask annotation (c) scribble annotation

Lin, D., Dai, J., Jia, J., He, K., & Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR2016*.

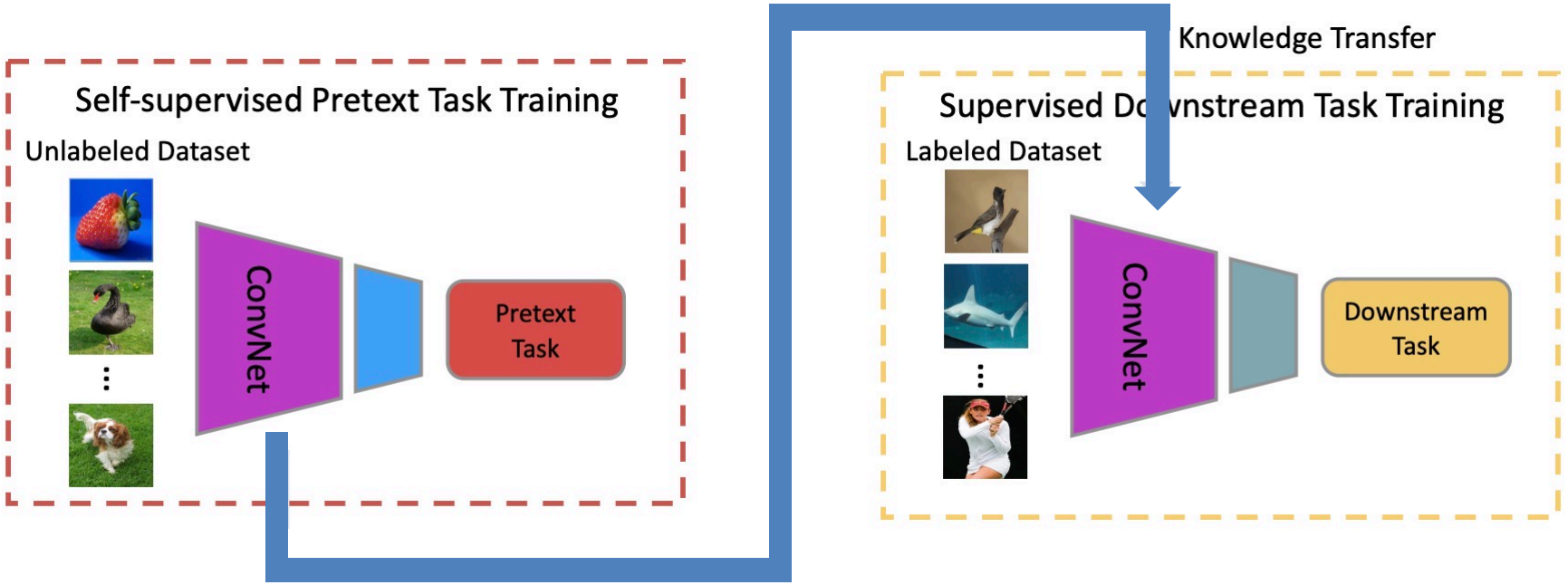
Self-supervised learning

- Self-supervised learning is a subset of unsupervised learning methods.

- Self-supervised learning refers to learning methods in which algorithms are explicitly trained with automatically generated labels.

  - **Human-annotated label:** Human-annotated labels refer to labels of data that are manually annotated by human workers.
  - **Pseudo label:** Pseudo labels are automatically generated labels based on data attributes for pretext tasks.

Self-supervised learning – general pipeline



Pretext Task: Pretext tasks are pre-designed tasks for networks to solve, and visual features are learned by learning objective functions of pretext tasks

Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." IEEE transactions on pattern analysis and machine intelligence 43.11 (2020): 4037-4058.

Self-Supervised Learning

- Why self-supervised learning?
  - Creating labeled datasets for each task is an expensive, time-consuming, tedious task
    - Requires hiring human labelers, preparing labeling manuals, creating GUIs, creating storage pipelines, etc.
    - High quality annotations in certain domains can be particularly expensive (e.g., medicine)
  - Self-supervised learning takes advantage of the vast amount of unlabeled data on the internet (images, videos, text)
    - Rich discriminative features can be obtained by training models without actual labels
  - Self-supervised learning can potentially generalize better because we learn more about the world
- ***Challenges*** for self-supervised learning
  - How to select a suitable pretext task for an application
  - There is no gold standard for comparison of learned feature representations

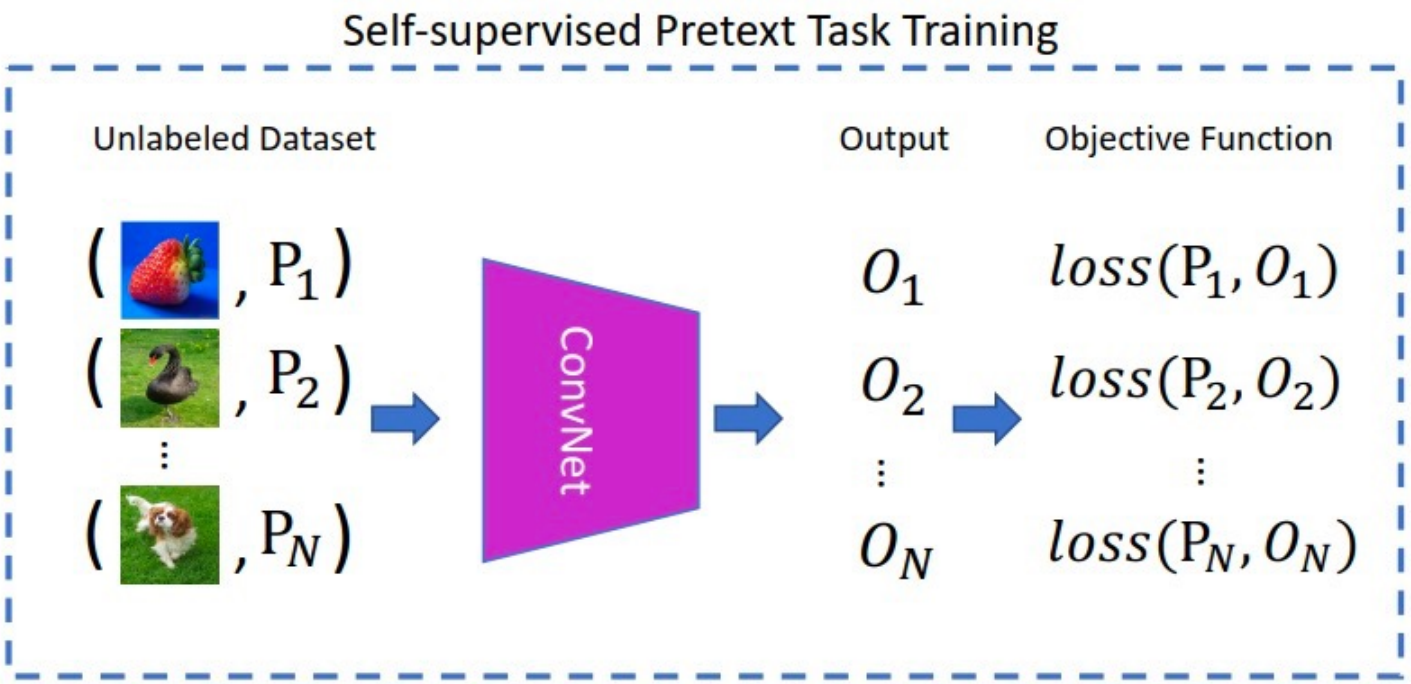
Slide credit: Dr. Alex Vakanski

Self-supervised learning – general pipeline

- Step 1: The visual feature is learned through the process of training neural networks (e.g., CNNs) to solve a pre-defined pretext task, e.g. image rotation prediction.

- Step 2: The learned parameters serve as a pre-trained model and are transferred to other downstream computer vision tasks by finetuning.

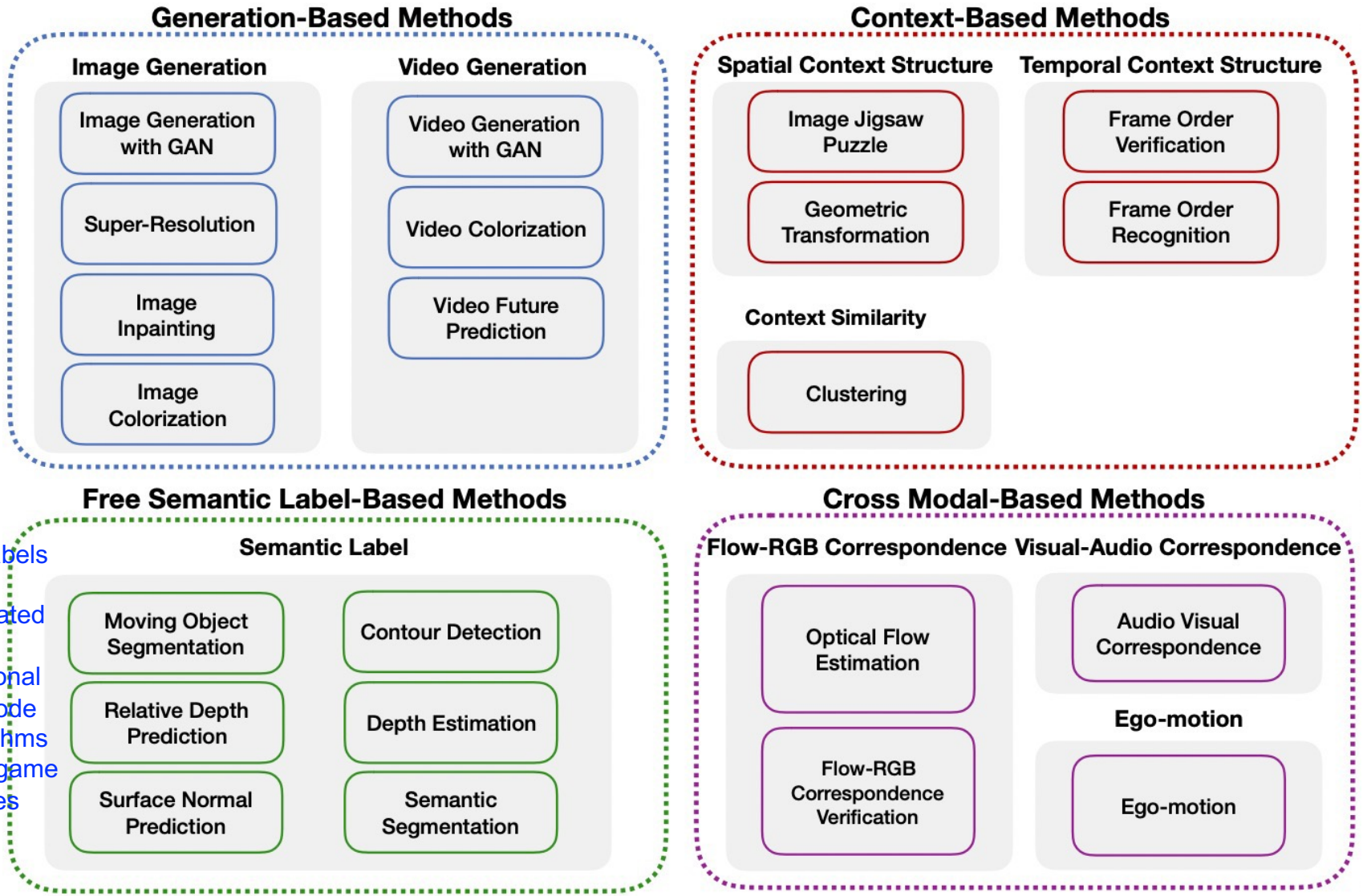
Pretext task training



The ConvNet is trained by minimizing errors between pseudo labels P and predictions O of the ConvNet. Since the pseudo labels are automatically generated, no human annotations are involved during the whole process.

Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." IEEE transactions on pattern analysis and machine intelligence 43.11 (2020): 4037-4058.

Categories of pretext task

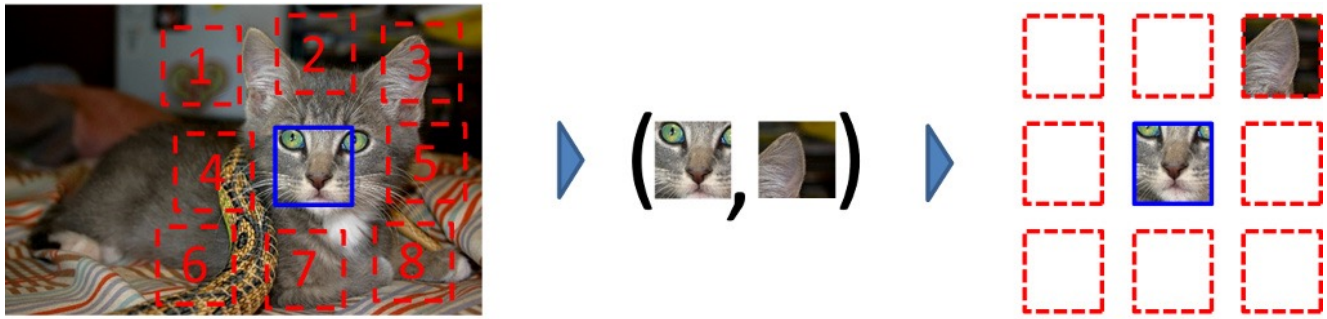


The labels are generated by traditional hardcode algorithms or by game engines

Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." IEEE transactions on pattern analysis and machine intelligence 43.11 (2020): 4037-4058.

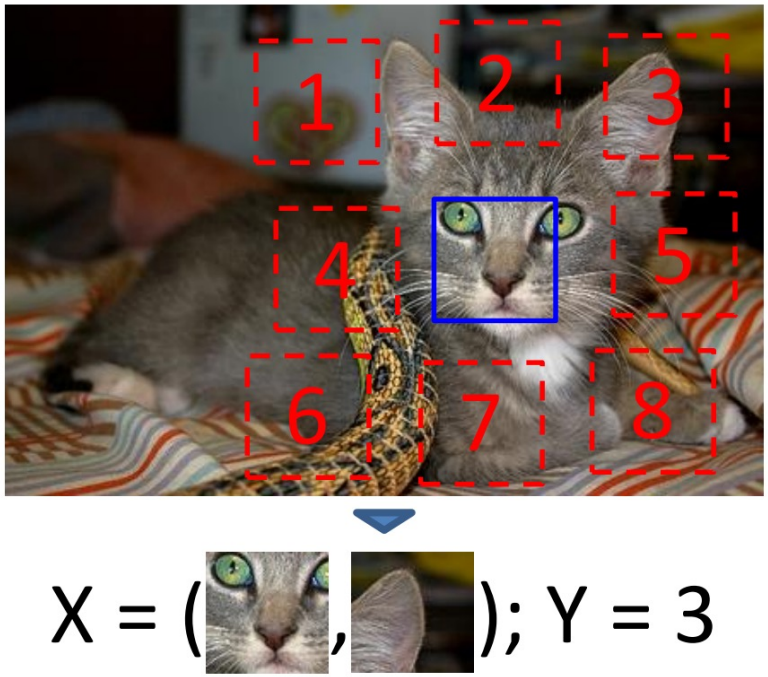
Context

- Solving the Jigsaw
  - Predict relative positions of patches
  - You have to understand the object to solve this problem!



Carl Doersch, Abhinav Gupta, and Alexei A. Efros. **Unsupervised Visual Representation Learning by Context Prediction.** In *ICCV 2015*

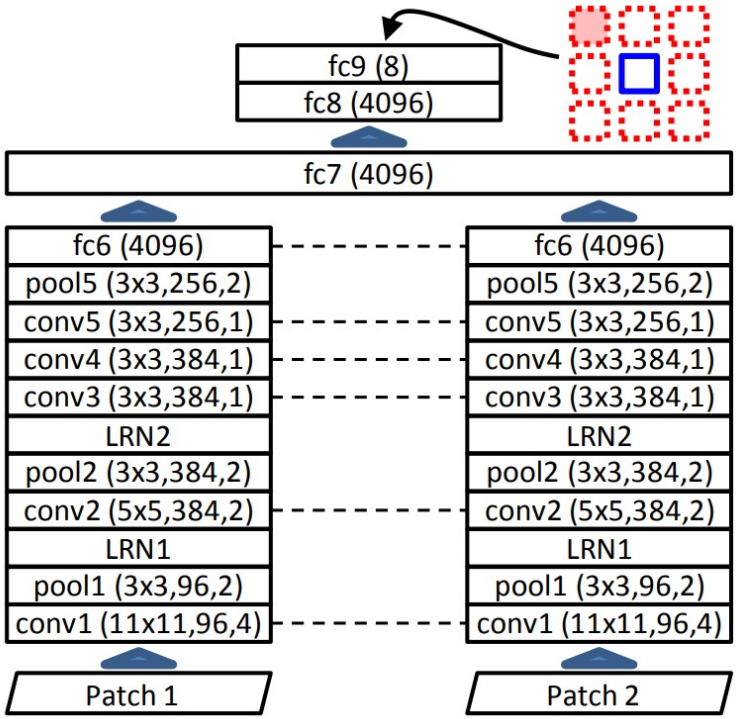
Spatial Context Prediction



The algorithm receives two patches in one of these eight possible spatial arrangements, without any context, and must then classify which configuration was sampled.

Unsupervised Visual Representation Learning by Context Prediction.
Carl Doersch, Abhinav Gupta, Alexei A. Efros

Spatial Context Prediction

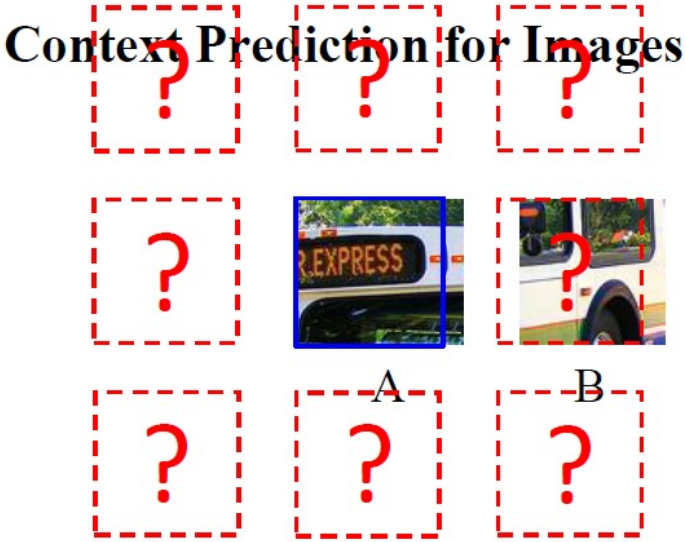


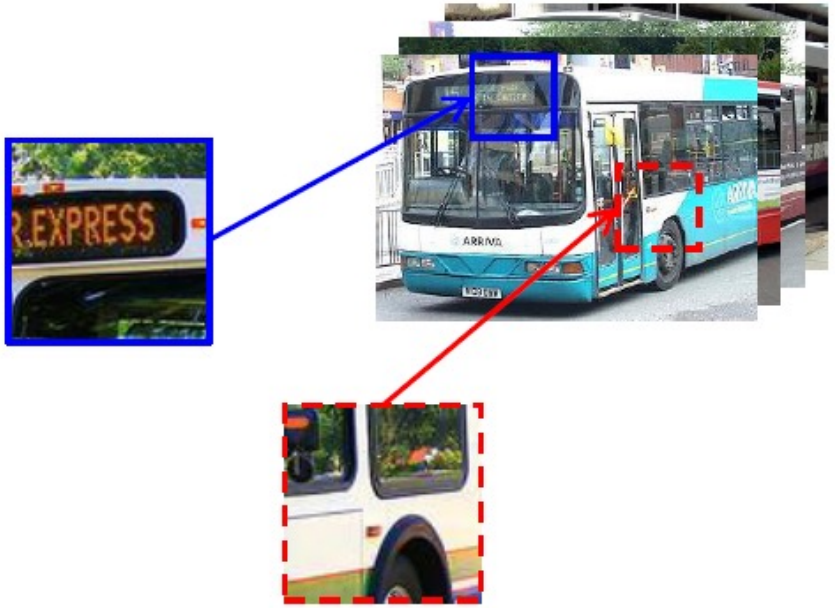
We ultimately wish to learn a feature embedding for individual patches, such that patches which are visually similar (across different images) would be close in the embedding space.

Unsupervised Visual Representation Learning by Context Prediction.
Carl Doersch, Abhinav Gupta, Alexei A. Efros

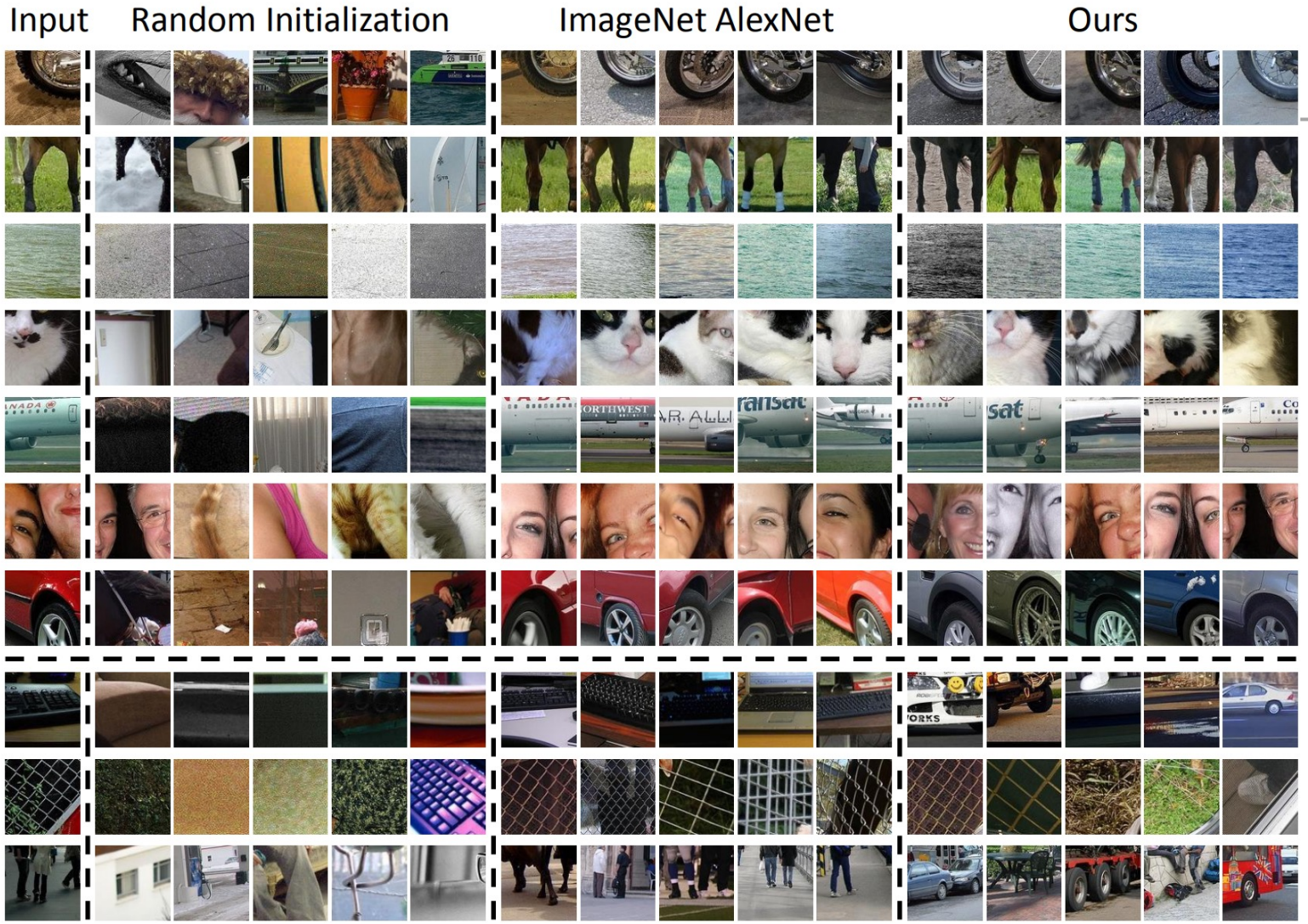
Spatial Context Prediction

- Example: predict the position of patch B with respect to patch A



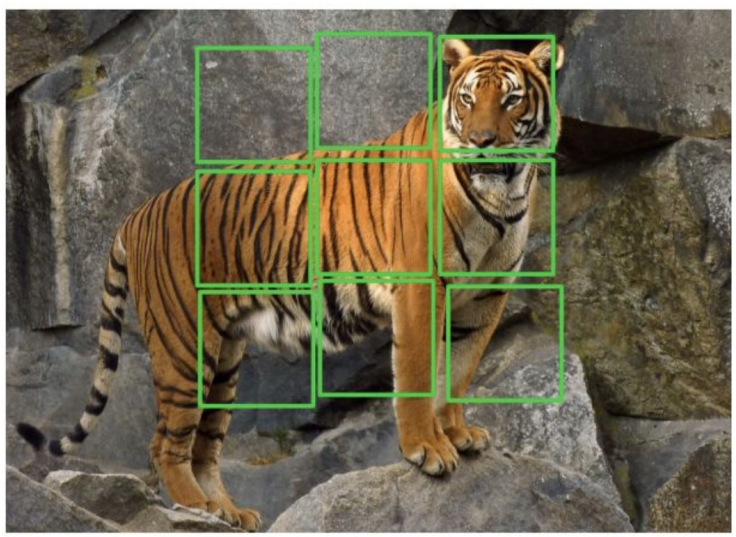


Slide credit: Dr. Alex Vakanski



The query patch is shown on the far left. Matches are for three different features: fc6 features from a random initialization of our architecture, AlexNet fc7 after training on labeled ImageNet, and the fc6 features learned from our method. Queries were chosen from 1000 randomly-sampled patches. The top group is examples where our algorithm performs well; for the bottom group, AlexNet outperforms our approach.

Solving Jigsaw Puzzles

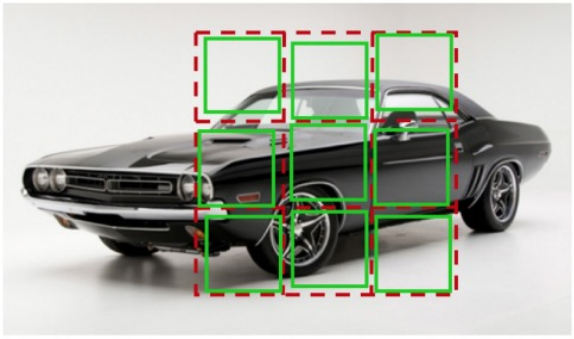


Training data: 9 patches extracted in images (similar to the previous approach)
Pretext task: predict the positions of all 9 patches

Instead of predicting the relative position of only 2 patches, this approach uses the grid of 3×3 patches and solves a jigsaw puzzle

Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." European conference on computer vision. Springer, Cham, 2016.

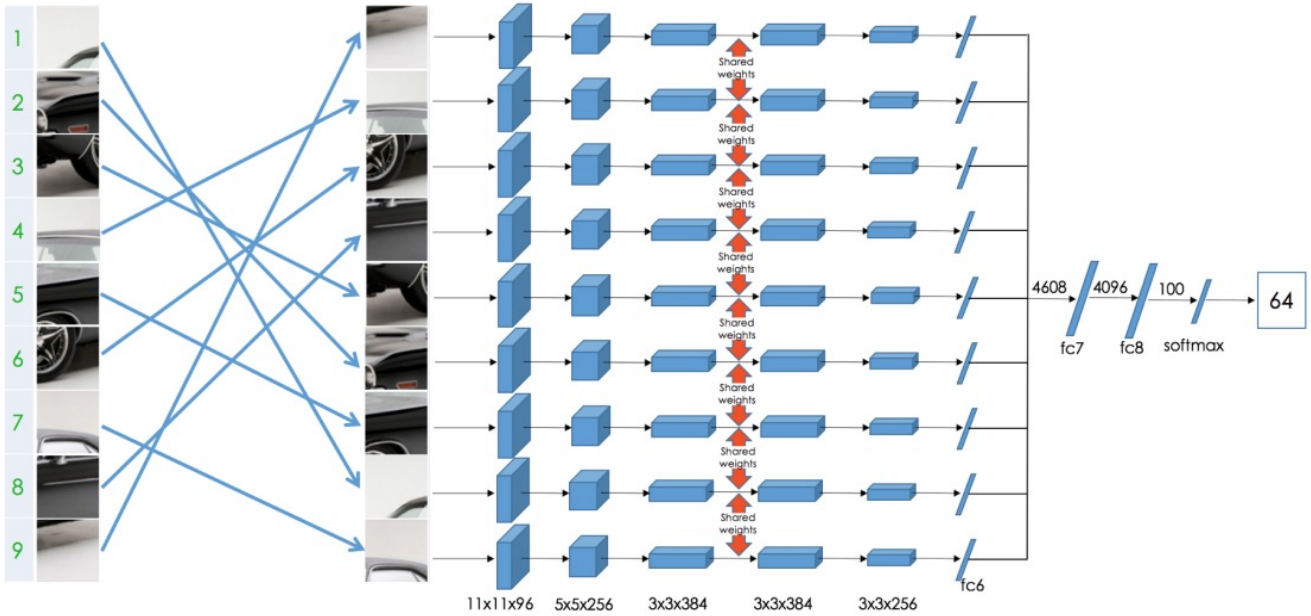
Solving Jigsaw Puzzles



Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation



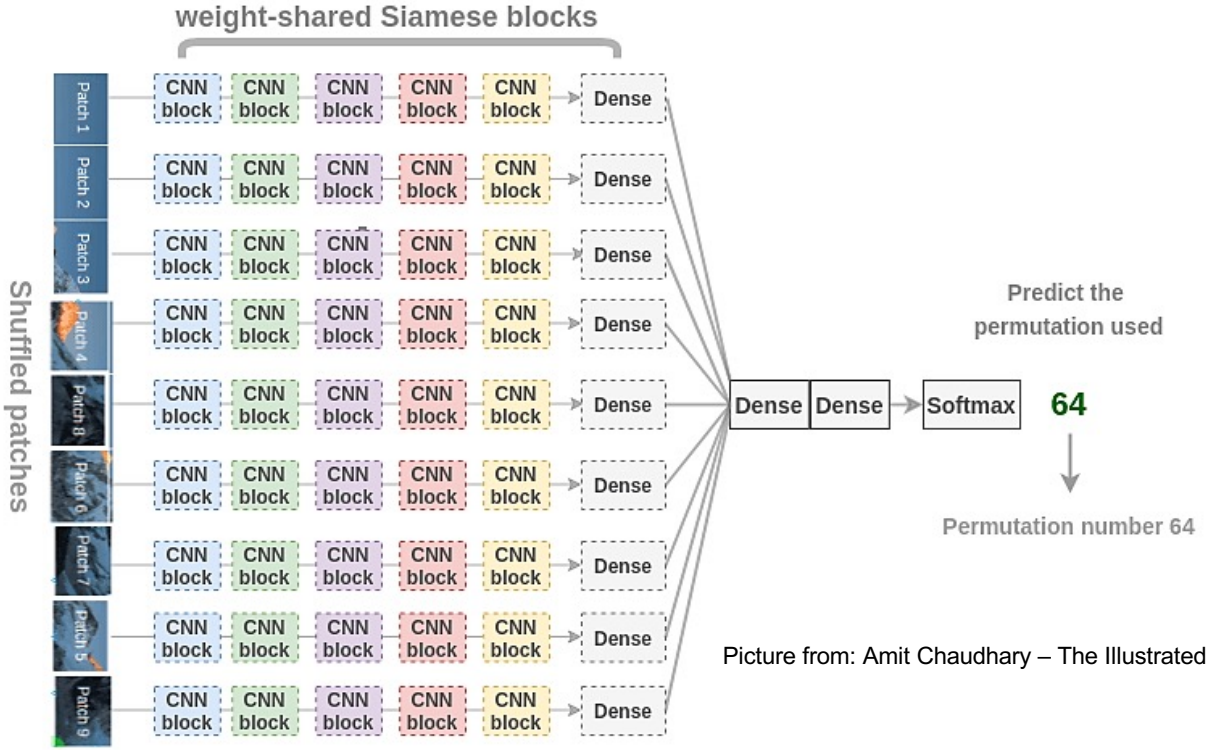
Context Free Network (CFN)

- A ConvNet model passes the individual patches through the same Conv layers that have shared weights
  - The features are combined and passed through fully-connected layers
  - Output is the positions of the patches (i.e., the shuffling permutation of the patches)
  - The patches are shuffled according to a set of 64 predefined permutations
    - Namely, for 9 patches, in total there are 362,880 possible puzzles
    - The authors used a small set of 64 shuffling permutations

Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." European conference on computer vision. Springer, Cham, 2016.

Solving Jigsaw Puzzles

- The model needs to learn to identify how parts are assembled in an object, relative positions of different parts of objects, and shape of objects
  - The learned representations are useful for downstream tasks in classification and object detection



Picture from: Amit Chaudhary – The Illustrated Self-Supervised Learning

Slide credit: Dr. Alex Vakanski

Solving Jigsaw Puzzles

- Transfer learning
  - Train the CFN on the self-supervised learning task.
  - Use the CFN weights to initialize all the conv layers of a standard AlexNet network (freeze weights).
  - Retrain the rest of the network from scratch (Gaussian noise as initial weights) for object classification on ImageNet dataset.

Solving Jigsaw Puzzles

Table 1: Results on PASCAL VOC 2007 Detection and Classification. The results of the other methods are taken from Pathak *et al.* [30].

Method	Pretraining time	Supervision	Classification	Detection	Segmentation
Krizhevsky*et al.* [25]	3 days	1000 class labels	**78.2%**	**56.8%**	**48.0%**
Wang and Gupta[39]	1 week	motion	58.4%	44.0%	-
Doersch *et al.* [10]	4 weeks	context	55.3%	46.6%	-
Pathak *et al.* [30]	14 hours	context	56.5%	44.5%	29.7%
Ours	2.5 days	context	**67.6%**	**53.2%**	**37.6%**

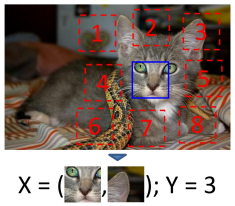
Krizhevsky*et al.* [25] AlexNet (supervised learning)

Wang and Gupta[39]

Doersch *et al.* [10] → Spatial context prediction

Pathak *et al.* [30]

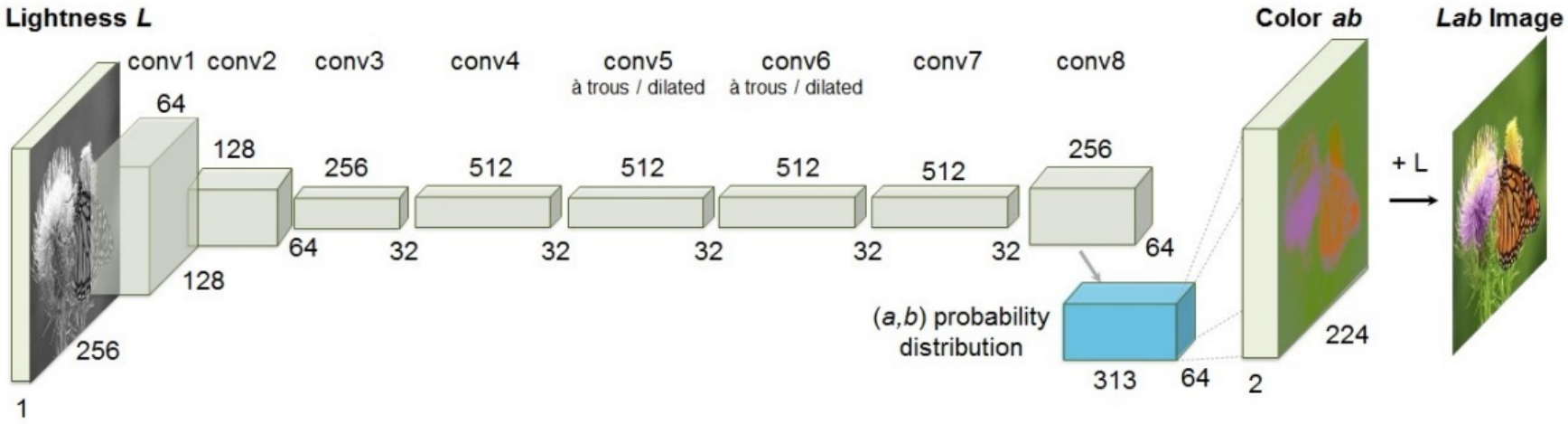
Ours



Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." European conference on computer vision. Springer, Cham, 2016.

Context

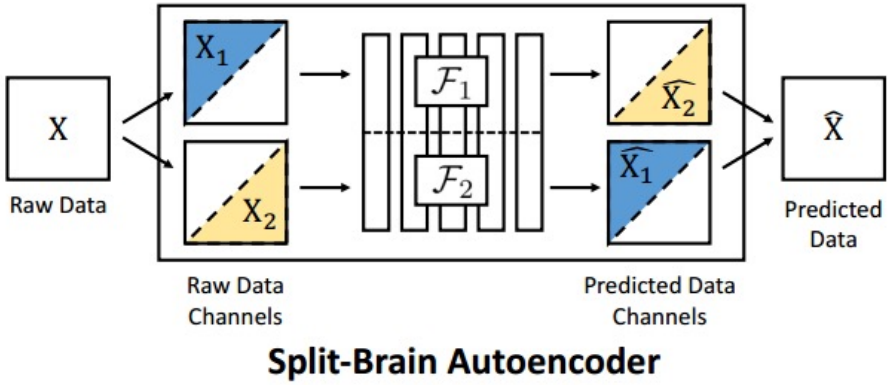
- Colorization
  - You have to know what the object is before you predict its color
  - E.g. Apple is red/green, sky is blue, etc.

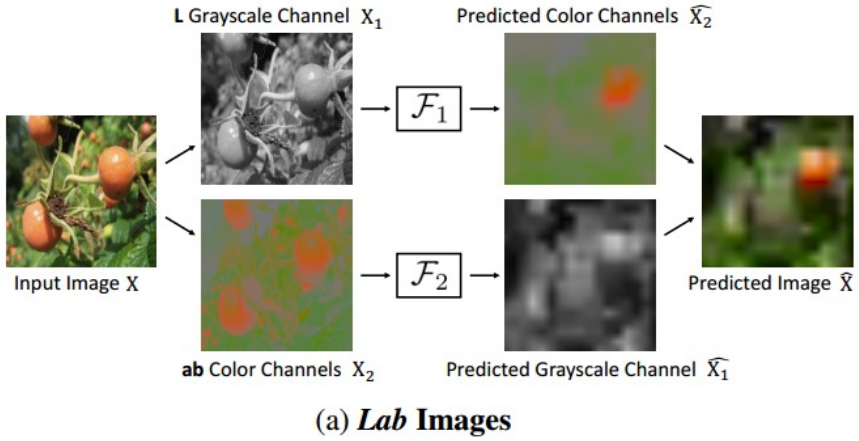


Zhang, R., Isola, P., & Efros, A. A. Colorful image colorization. In *ECCV 2016*

Context

- Colorization
  - Stronger supervision, cross-supervision of different parts of data



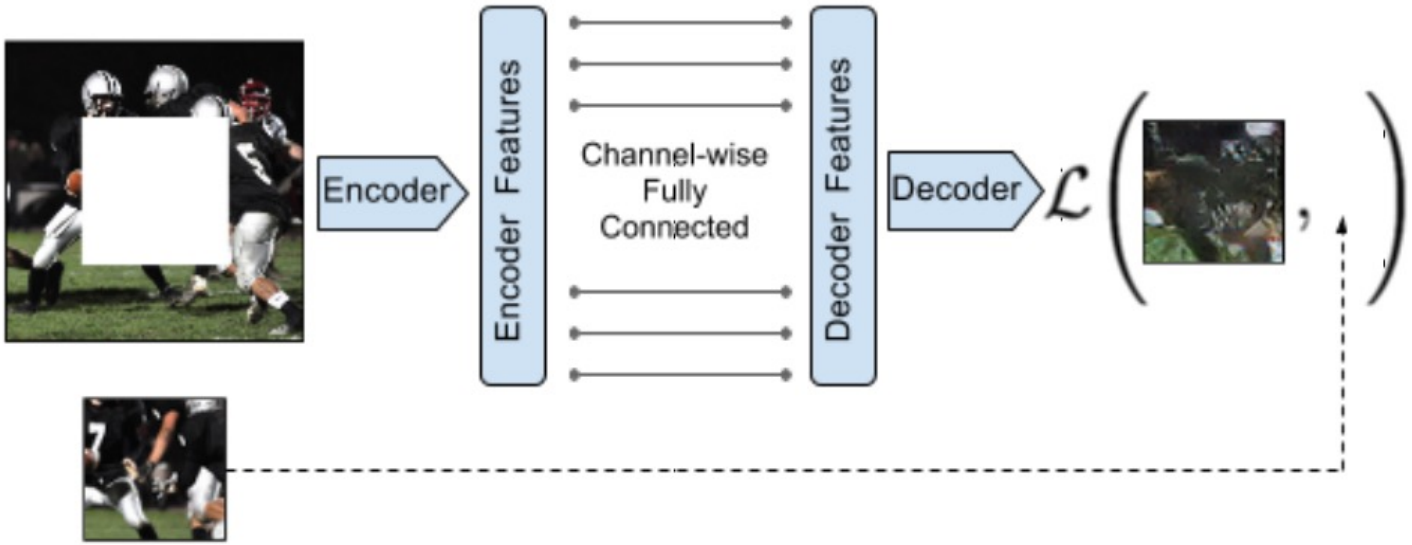


Zhang, R., Isola, P., & Efros, A. A. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. *In CVPR 2017*

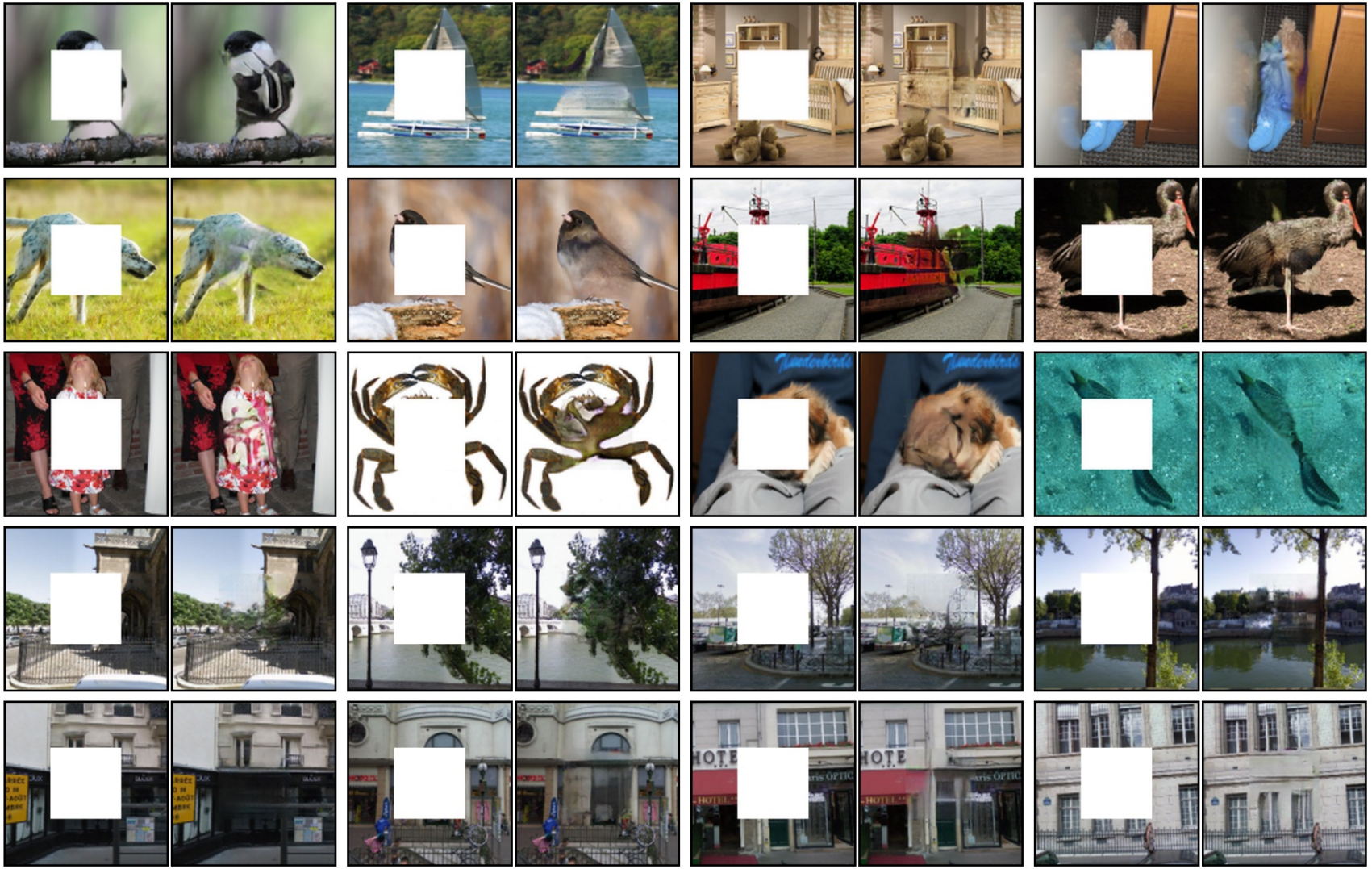
Image Inpainting

Decoder loss

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}.$$



Context Encoders: Feature Learning by Inpainting
Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros



Context Encoders: Feature Learning by Inpainting
Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

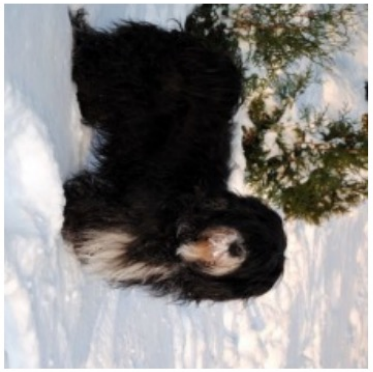
Image Inpainting

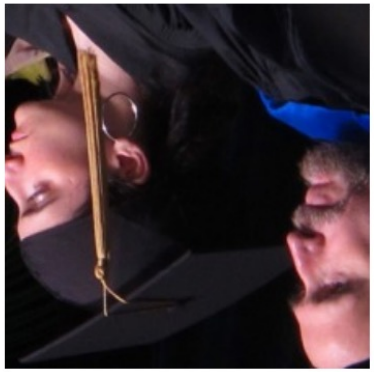
Pretraining Method	Supervision	Pretraining time	Classification	Detection	Segmentation
ImageNet [26]	1000 class labels	3 days	78.2%	56.8%	48.0%
Random Gaussian	initialization	$<$ 1 minute	53.3%	43.4%	19.8%
Autoencoder	-	14 hours	53.8%	41.9%	25.2%
Agrawal *et al.* [1]	egomotion	10 hours	52.9%	41.8%	-
Wang *et al.* [39]	motion	1 week	58.7%	47.4%	-
Doersch *et al.* [7]	relative context	4 weeks	55.3%	46.6%	-
Ours	context	14 hours	56.5%	44.5%	30.0%

Context Encoders: Feature Learning by Inpainting
Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

Learning by Rotating


90° rotation


270° rotation


180° rotation

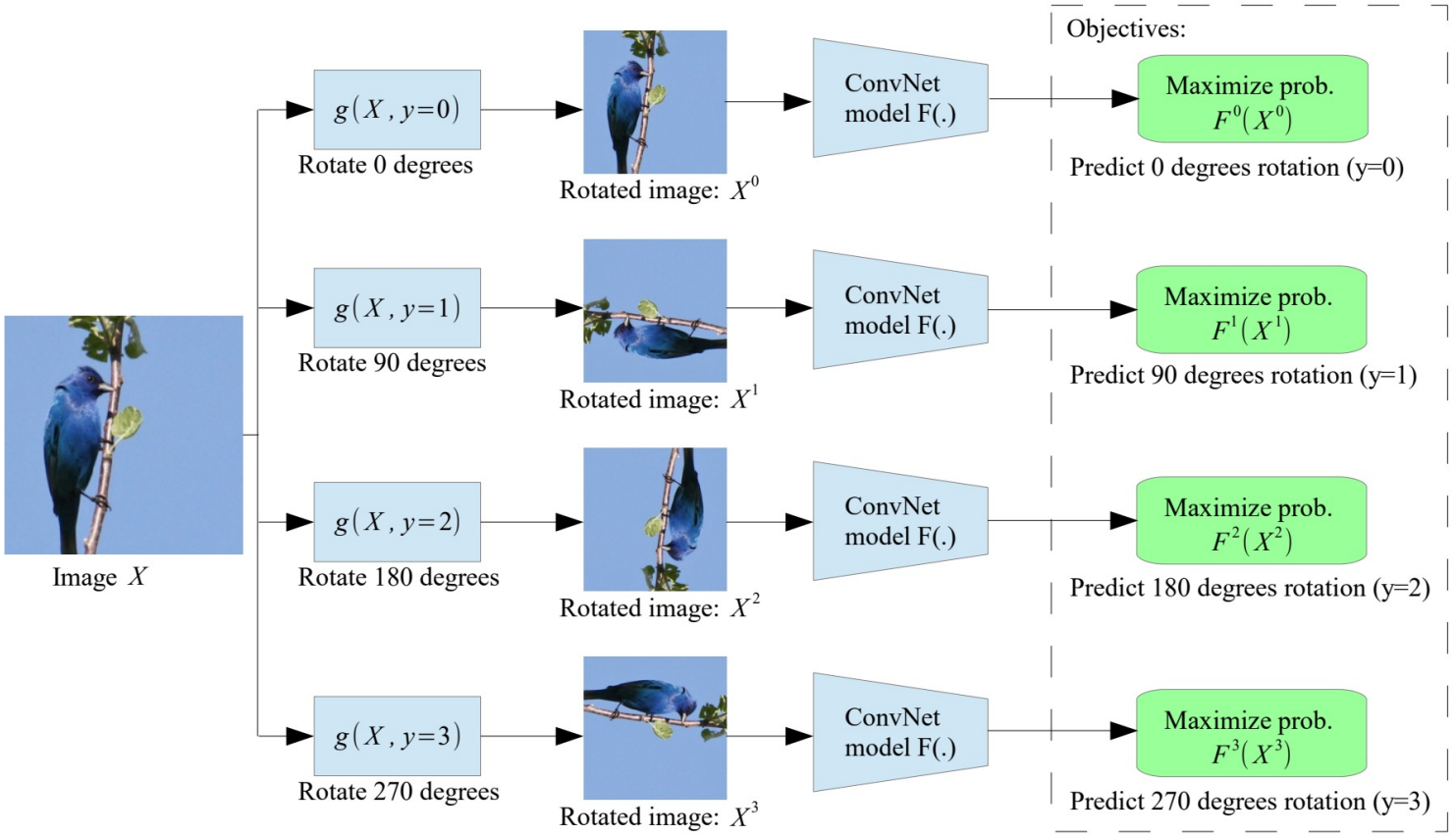

0° rotation


270° rotation

Motivation: if someone is not aware of the concepts of the objects depicted in the images, he/she cannot recognize the rotation that was applied to them.

Unsupervised Representation Learning by Predicting Image Rotations
Spyros Gidaris, Praveer Singh, Nikos Komodakis

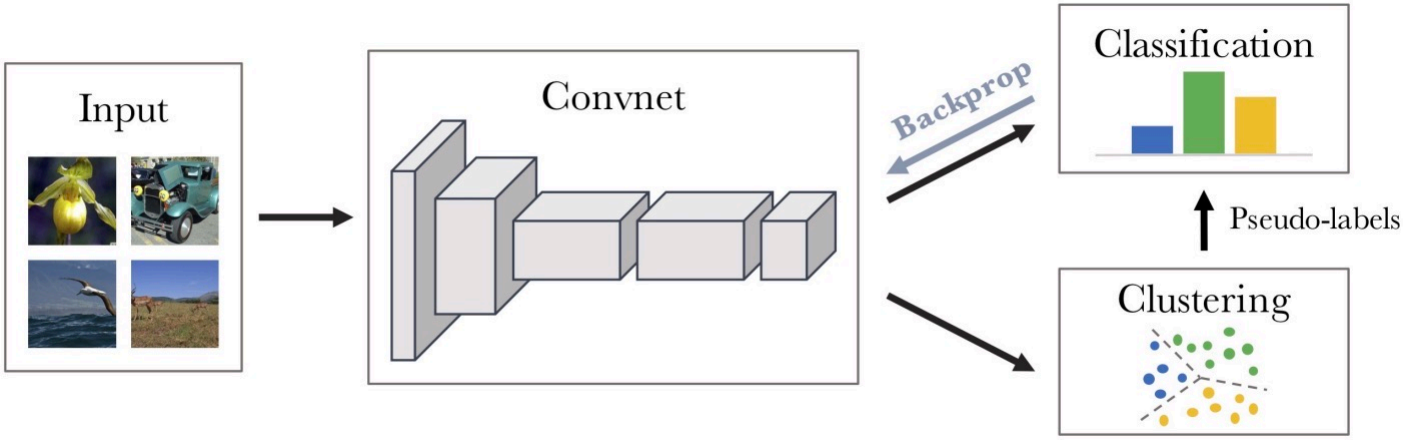
Learning by Rotating



Given four possible geometric transformations, the 0, 90, 180, and 270 degrees rotations, we train a ConvNet model F(.) to recognize the rotation that is applied to the image that it gets as input.

Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).

Learning by clustering



- Iteratively cluster deep features
- Use the cluster assignments as pseudo-labels to learn the parameters of the convnet.

- The authors used *k*-means for clustering the extracted feature maps.
- The model needs to learn the content in the images in order to assign them to the corresponding cluster.

Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

Performance Comparison

TABLE 4
Linear classification on ImageNet and Places datasets using activations from the convolutional layers of an AlexNet as features. "Convn" means the linear classifier is trained based on the n-th convolution layer of AlexNet. "Places Labels" and "ImageNet Labels" indicate using supervised model trained with human-annotated labels as the pre-trained model.

		ImageNet					Places				
Method	Pretext Tasks	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels [8]	—	—	—	—	—	—	22.1	35.1	40.2	43.3	44.6
ImageNet labels [8]	—	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random(Scratch) [8]	—	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
ColorfulColorization [18]	Generation	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
BiGAN [122]	Generation	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
SplitBrain [42]	Generation	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
ContextEncoder [19]	Context	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
ContextPrediction [41]	Context	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Jigsaw [20]	Context	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Learning2Count [130]	Context	18.0	30.6	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
DeepClustering [44]	**Context**	13.4	**32.3**	**41.0**	**39.6**	**38.2**	19.6	33.2	**39.2**	**39.8**	**34.7**

Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." IEEE transactions on pattern analysis and machine intelligence 43.11 (2020): 4037-4058.

Performance Comparison

TABLE 5
Comparison of the self-supervised image feature learning methods on classification, detection, and segmentation on PASCAL VOC dataset. "ImageNet Labels" indicates using supervised model trained with human-annotated labels as the pre-trained model.

Method	Pretext Tasks	Classification	Detection	Segmentation
ImageNet Labels [8]	—	79.9	56.8	48.0
Random(Scratch) [8]	—	57.0	44.5	30.1
ContextEncoder [19]	Generation	56.5	44.5	29.7
BiGAN [122]	Generation	60.1	46.9	35.2
ColorfulColorization [18]	Generation	65.9	46.9	35.6
SplitBrain [42]	Generation	67.1	46.7	36.0
RankVideo [38]	Context	63.1	47.2	35.4†
PredictNoise [46]	Context	65.3	49.4	37.1†
JigsawPuzzle [20]	Context	67.6	53.2	37.6
ContextPrediction [41]	Context	65.3	51.1	—
Learning2Count [130]	Context	67.7	51.4	36.6
DeepClustering [44]	**Context**	**73.7**	**55.4**	**45.1**
WatchingVideo [81]	Free Semantic Label	61.0	52.2	—
CrossDomain [30]	Free Semantic Label	68.0	52.6	—
AmbientSound [154]	Cross Modal	61.3	—	—
TiedToEgoMotion [95]	Cross Modal	—	41.7	—
EgoMotion [94]	Cross Modal	54.2	43.9	—

Predict rotation		72.97	54.4	39.1

Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." IEEE transactions on pattern analysis and machine intelligence 43.11 (2020): 4037-4058.

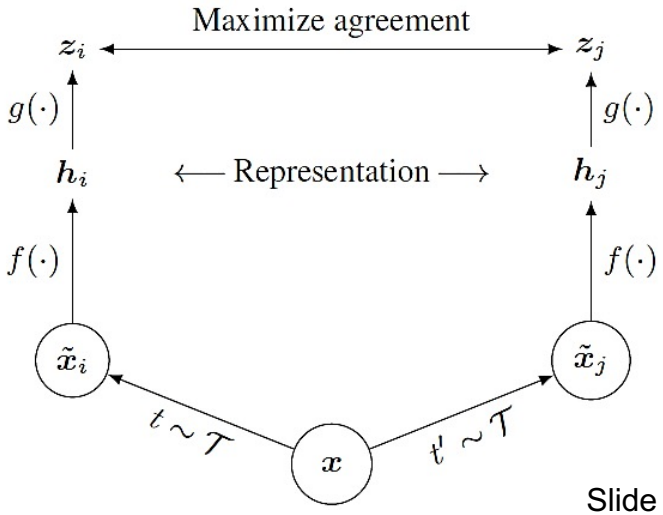
Contrastive learning (SimCLR)

Chen, Ting, et al. “A simple framework for contrastive learning of visual representations.” ICML 2020.

https://github.com/google-research/simclr

Contrastive learning (SimCLR)

- Randomly sample a mini-batch of n inputs $\boldsymbol{x}$, and apply two different data augmentation operations t and t', resulting in $2n$ samples $\tilde{\boldsymbol{x}}_i = t(\boldsymbol{x})$ and $\tilde{\boldsymbol{x}}_j = t'(\boldsymbol{x})$
  - Data augmentation includes random crop, resize with random flip, color distortions, and Gaussian blur (data augmentation is crucial for contrastive learning)
- Apply a base encoder $f(\cdot)$ to $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ to obtain the code representations $\boldsymbol{h}_i = f(\tilde{\boldsymbol{x}}_i)$ and $\boldsymbol{h}_j = f(\tilde{\boldsymbol{x}}_j)$
- Apply another prediction head encoder $g(\cdot)$ (one fully-connected layer) to $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ to obtain the code representations $\boldsymbol{z}_i = g(\boldsymbol{h}_i)$ and $\boldsymbol{z}_j = g(\boldsymbol{h}_j)$



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." ICML 2020.

https://github.com/google-research/simclr

Slide credit: Dr. Alex Vakanski

SimCLR

- For one positive pair of samples $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ and for the remaining $2(n-1)$ samples treated as negative, a cosine similarity is calculated as

$$\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{z}_i^T \boldsymbol{z}_j}{\|\boldsymbol{z}_i\| \|\boldsymbol{z}_i\|}$$

- The contrastive prediction task aims for a given sample $\tilde{\boldsymbol{x}}_i$ to identify a positive pairing sample $\tilde{\boldsymbol{x}}_j$
- The NT-Xent (a.k.a. the normalized temperature-scaled cross-entropy loss) loss for the instances $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ (a positive pair) is calculated as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)/\tau\right)}{\sum_{k=1}^{2n} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

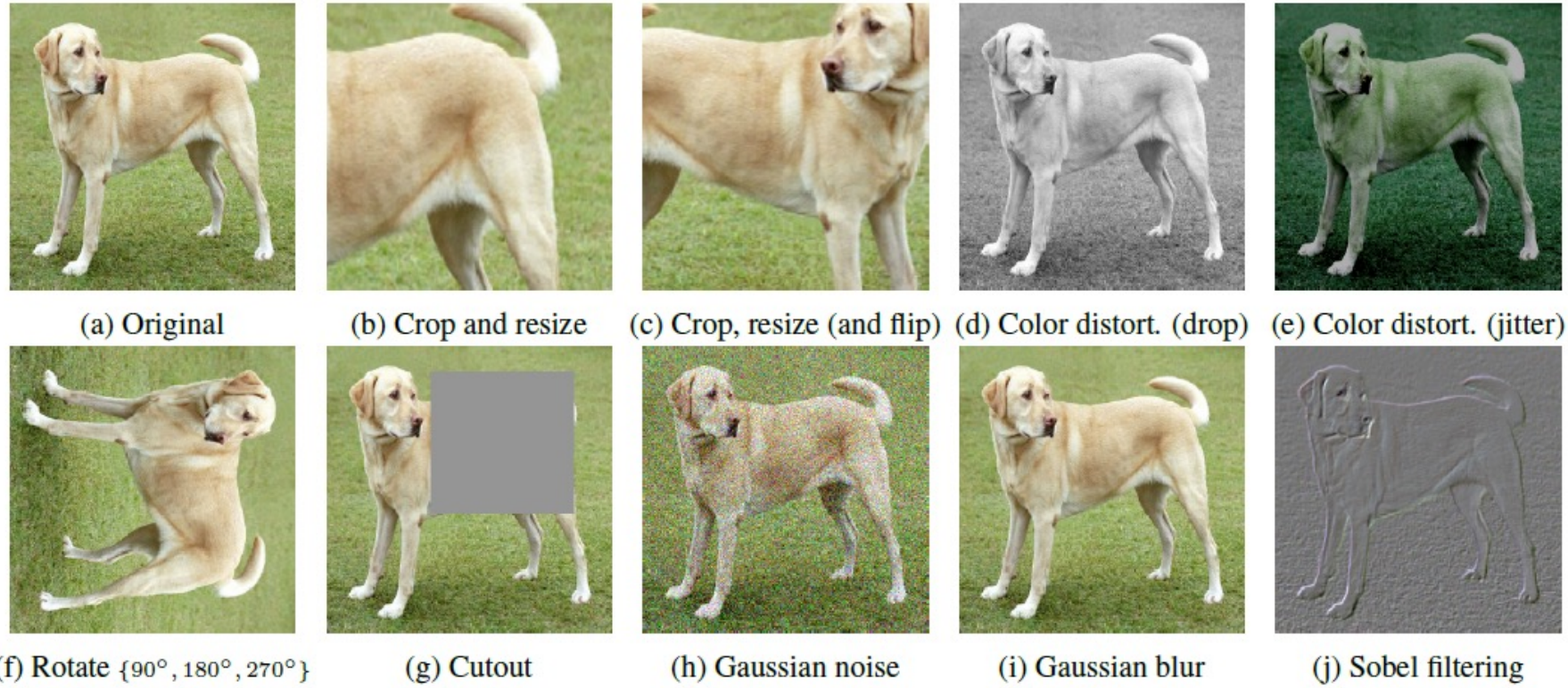
  - $\mathbf{1}_{[k \neq i]}$ has a value of 1 if $k \neq i$ and 0 otherwise, τ is a temperature hyperparameter
- The overall loss $\sum_{i,j} \mathcal{L}_{i,j}$ is calculated across all positive pairs $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ in a mini-batch
- For downstream tasks, the head $g(\cdot)$ is discarded and only the representation $\boldsymbol{h}_i$ is used

Walk through with an example:
https://amitness.com/2020/03/illustrated-simclr/

Slide credit: Dr. Alex Vakanski

SimCLR

- Data augmentation



Slide credit: Dr. Alex Vakanski

SimCLR

- Experimental results on 10 image datasets
  - SimCLR outperformed supervised models on most datasets

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 (4×) models pretrained on ImageNet.

Slide credit: Dr. Alex Vakanski

Other Contrastive SSL Approaches

- Other recent self-supervised approaches based on contrastive learning include:
  - ***Augmented Multiscale Deep InfoMax*** or ***AMDIM***
    - Bachman (2019) Learning Representations by Maximizing Mutual Information Across Views
  - ***Momentum Contrast*** or ***MoCo***
    - He (2019) Momentum Contrast for Unsupervised Visual Representation Learning
  - ***Bootstrap Your Own Latent*** or ***BYOL***
    - Grill (2020) Bootstrap your own latent: A new approach to self-supervised Learning
  - ***Swapping Assignments between multiple Views of the same image*** or ***SwAV***
    - Caron (2020) Unsupervised Learning of Visual Features by Contrasting Cluster Assignments
  - ***Yet Another DIM*** or ***YADIM***
    - Falcon (2020) A Framework for Contrastive Self-Supervised Learning and Designing a New Approach
  - SimSiam – no negative sample pairs
    - Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Slide credit: Dr. Alex Vakanski

Discussion

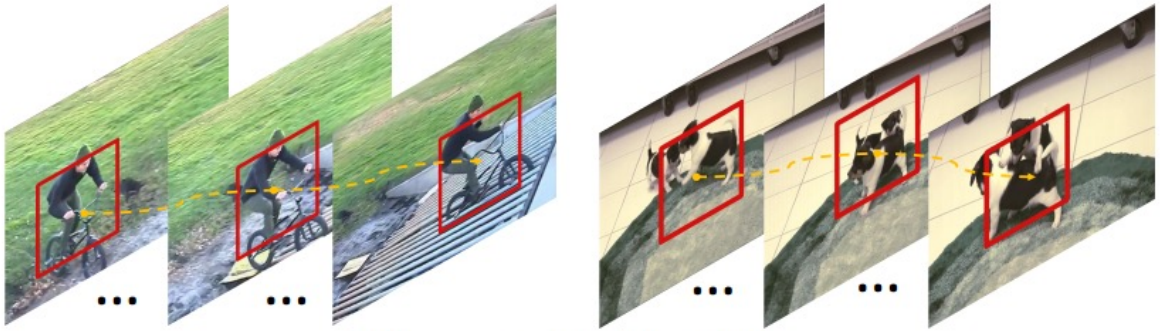
- What other self-supervision signals/schemes are useful for visual representation learning?

Self-supervised Learning Using Video

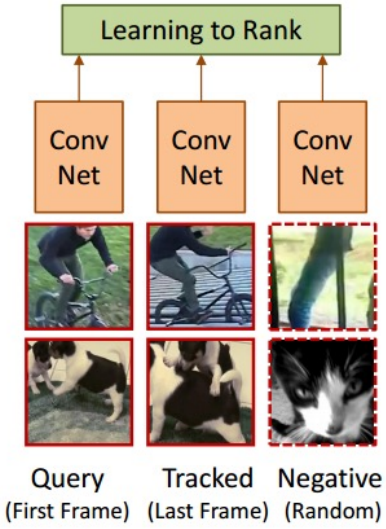
- Video can provide rich information
  - Temporal continuity
  - Motion consistency
  - Action order

Tracking Moving Objects

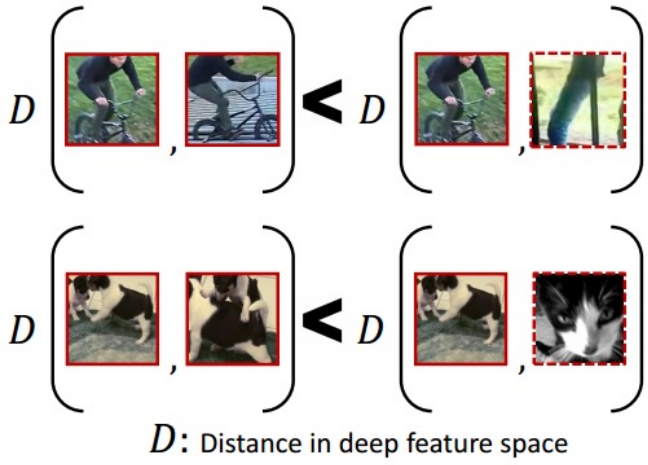
- Find corresponding pairs using visual tracking



(a) Unsupervised Tracking in Videos



(b) Siamese-triplet Network



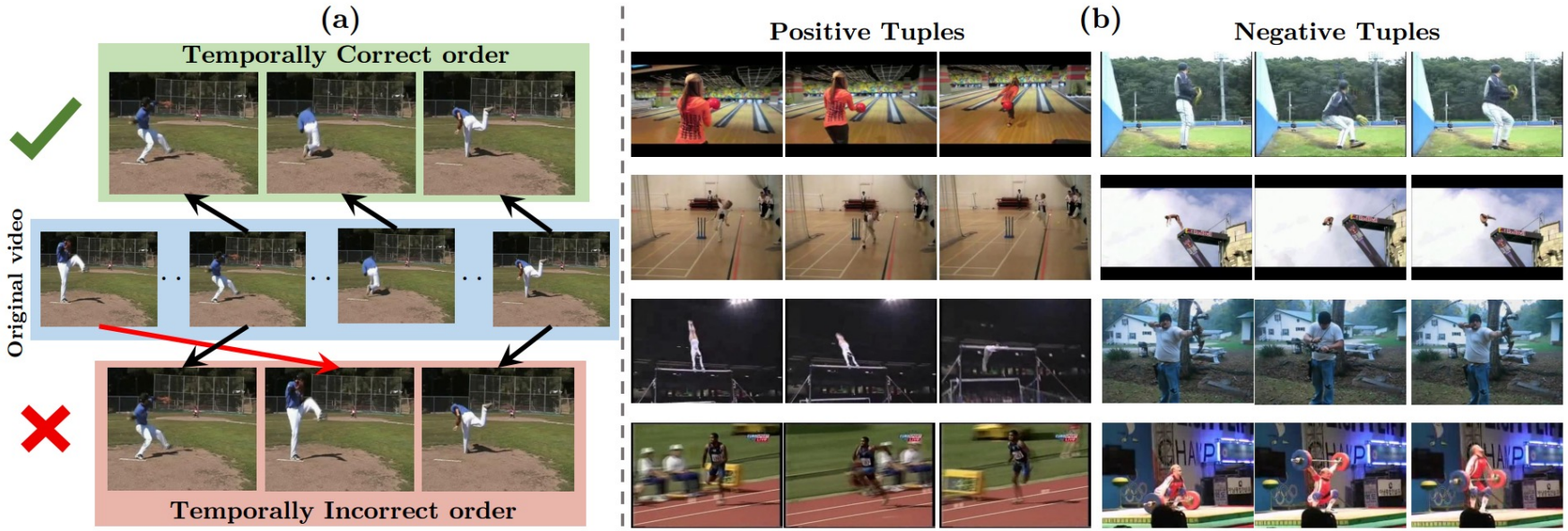
D: Distance in deep feature space

(c) Ranking Objective

Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *ICCV2015*

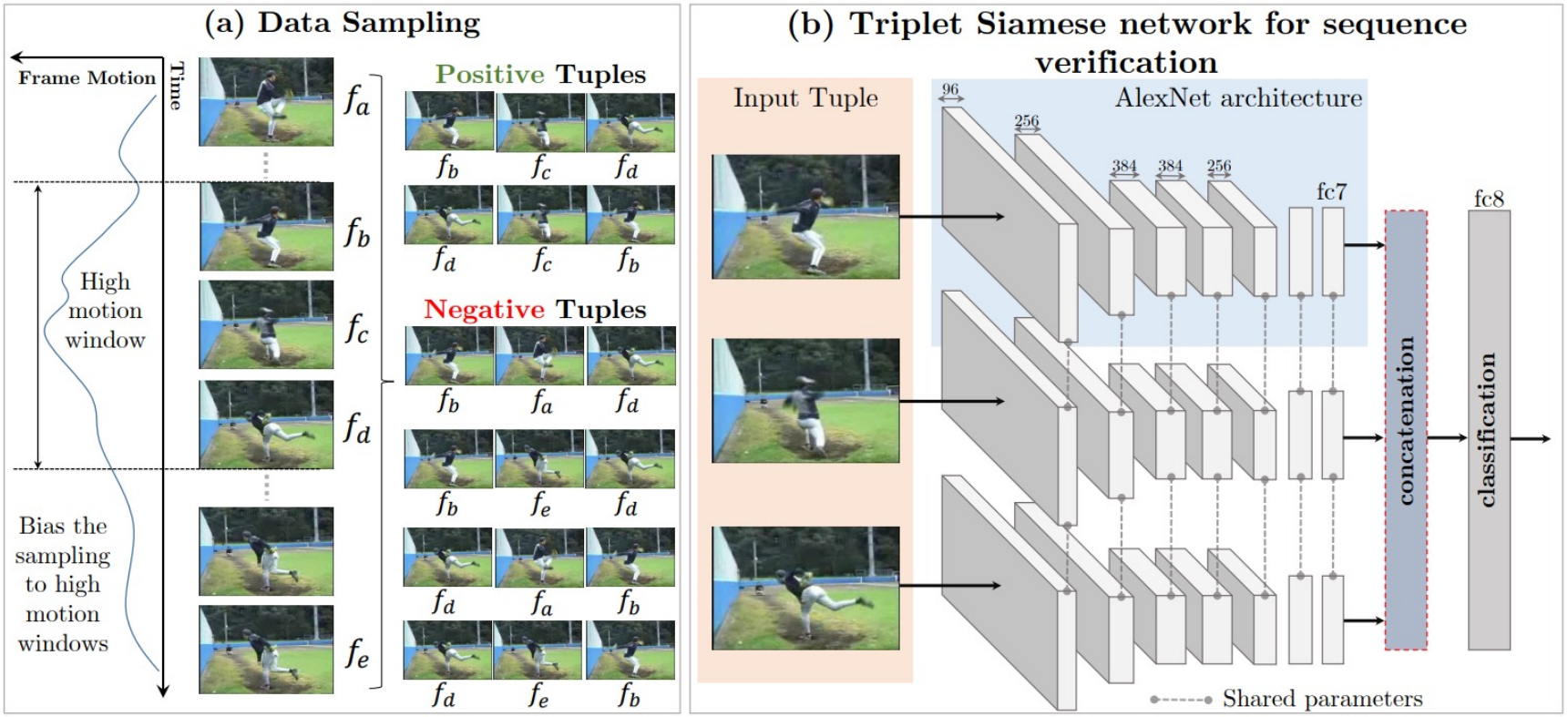
Learning from Temporal Ordering

- Is the temporal order of a video correct?
  - Encode the cause and effect of action



Shuffle and Learn: Unsupervised Learning using Temporal Order Verification
Ishan Misra, C. Lawrence Zitnick, Martial Hebert

Learning from Temporal Ordering



Shuffle and Learn: Unsupervised Learning using Temporal Order Verification
Ishan Misra, C. Lawrence Zitnick, Martial Hebert

Learning from Temporal Ordering

Table 2: Mean classification accuracies over the 3 splits of UCF101 and HMDB51 datasets. We compare different initializations and finetune them for action recognition.

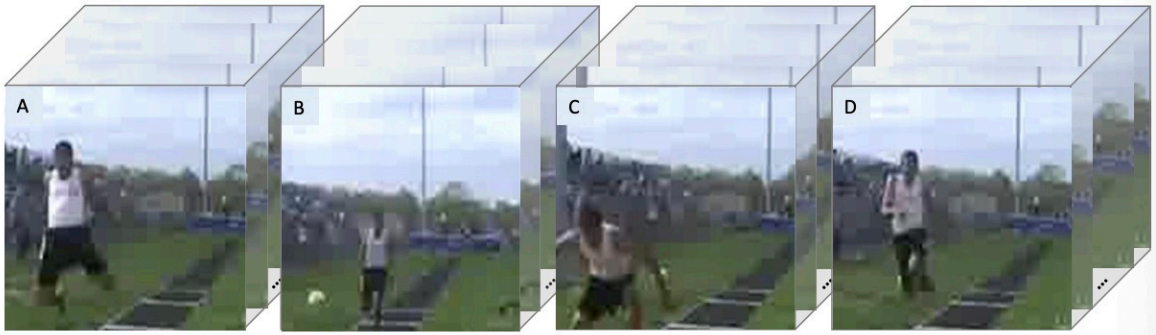
Dataset	Initialization	Mean Accuracy
UCF101	Random	38.6
	(Ours) Tuple verification	**50.2**
HMDB51	Random	13.3
	UCF Supervised	15.2
	(Ours) Tuple verification	**18.1**

Shuffle and Learn: Unsupervised Learning using Temporal Order Verification
Ishan Misra, C. Lawrence Zitnick, Martial Hebert

Video - Space-Time Cubic Puzzles

(Spatial)

Q. Can you arrange these?



A. Spatial: A-D-B-C

A. Temporal: B-D-A-C

(Temporal)

Kim, Dahun, Donghyeon Cho, and In So Kweon. "Self-supervised video representation learning with space-time cubic puzzles." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

Cross-Modality

- In some applications, it is easy to collect and align the data from several modalities
  - Lidar & GPS/IMU & Camera
  - RGB & D
  - Image & Text
  - Video & audio

Cross-Modality

- Audio and Visual Correspondence

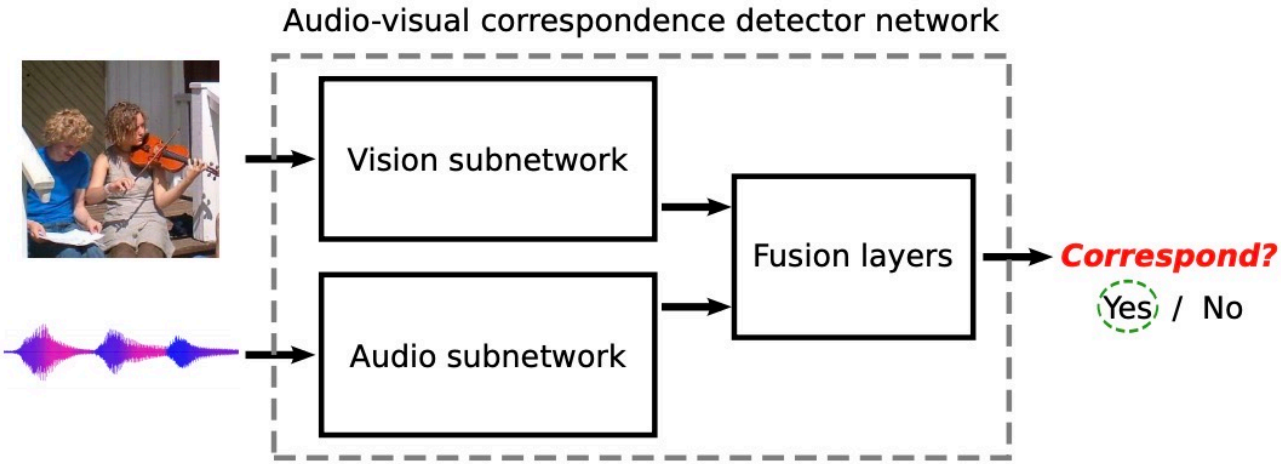
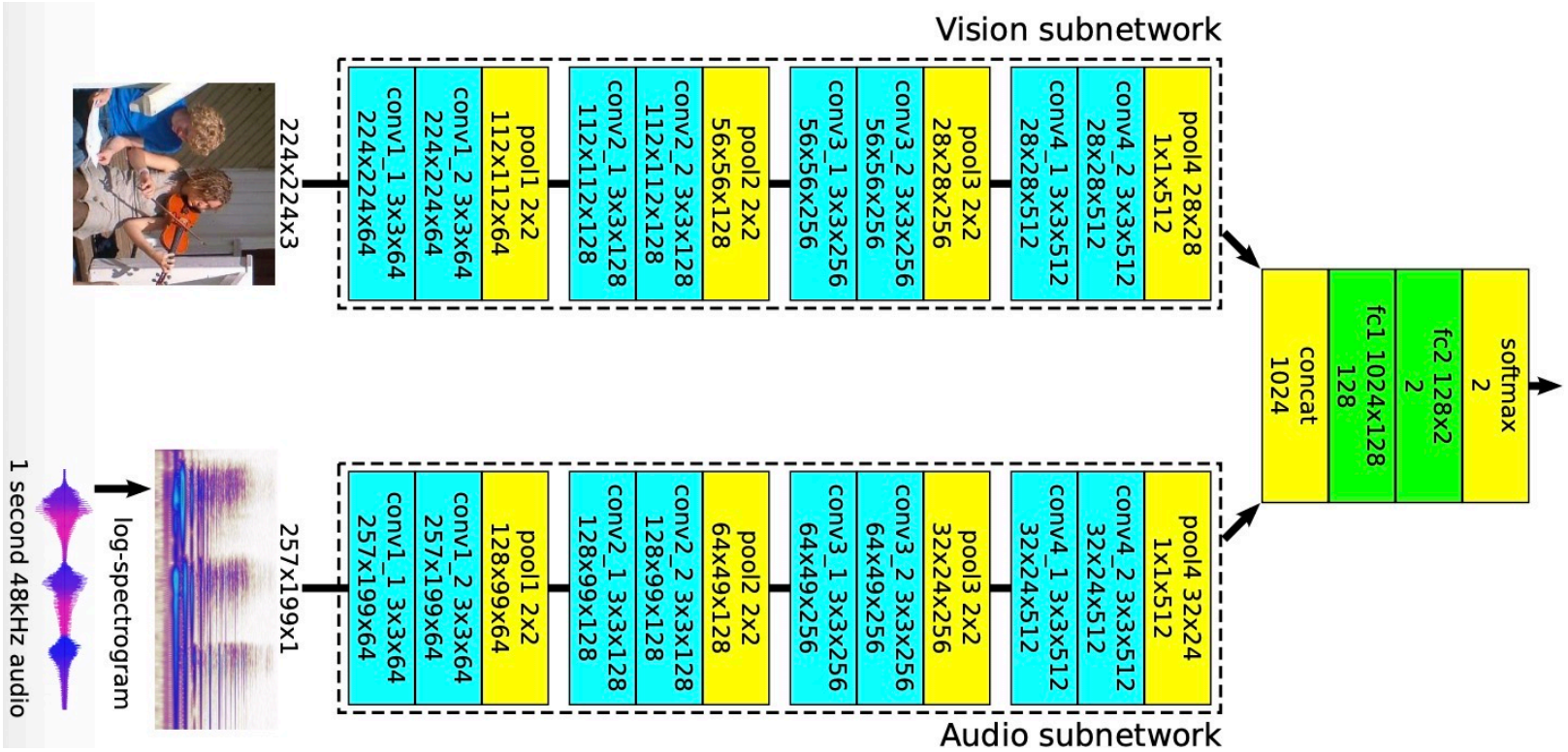


Figure 1. **Audio-visual correspondence task (AVC).** A network should learn to determine whether a pair of (video frame, short audio clip) correspond to each other or not. Positives are (frame, audio) extracted from the same time of one video, while negatives are a frame and audio extracted from different videos.

Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." Proceedings of the IEEE International Conference on Computer Vision. 2017.

Cross-Modality

- Audio and Visual Correspondence



Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." Proceedings of the IEEE International Conference on Computer Vision. 2017.

Self-Supervised Learning for Medical Image Computing

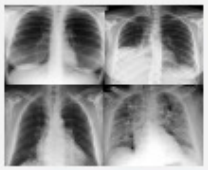
- High-quality labeled data is often scarce due to the time-consuming effort needed to annotate medical images.
- Self-supervised learning on unlabeled data can be an effective pre-training strategy within the domain of medical image computing.

Self-Supervised Models for Medical Image Classification

(1) Self-supervised learning on **unlabeled** natural images

(2) Self-supervised learning on **unlabeled** medical images and ***Multi-Instance Contrastive Learning (MICLe)*** if multiple images of each medical condition are available

Unlabeled dermatology images



Unlabeled chest x-rays

(3) Supervised fine-tuning on **labeled** medical images

Labeled dermatology images

Labeled chest x-rays

Azizi, Shekoofeh, et al. "Big self-supervised models advance medical image classification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Self-Supervised Models for Medical Image Classification

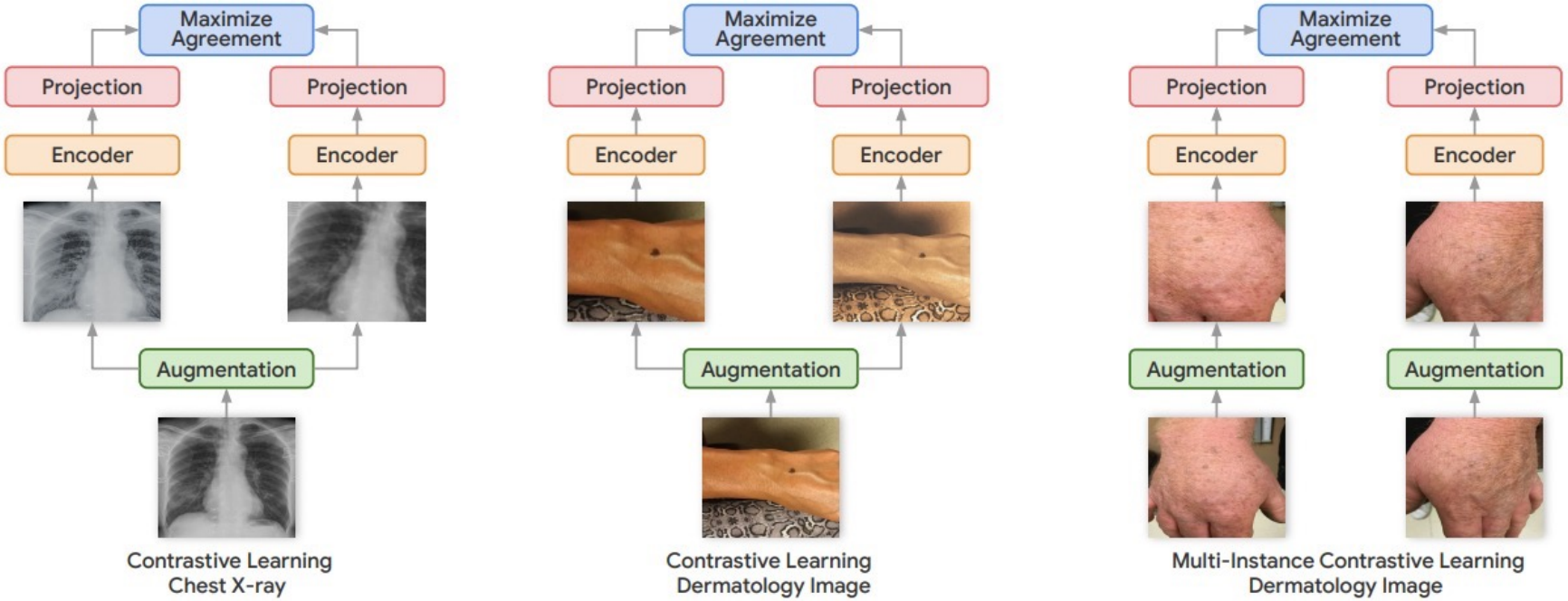


Figure 3: An illustrations of our self-supervised pretraining for medical image analysis. When a single image of a medical condition is available, we use standard data augmentation to generate two augmented views of the same image. When multiple images are available, we use two distinct images to directly create a positive pair of examples. We call the latter approach Multi-Instance Contrastive Learning (MICLe).

Azizi, Shekoofeh, et al. "Big self-supervised models advance medical image classification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Self-Supervised Models for Medical Image Classification

Table 3: Comparison of best self-supervised models *vs.* supervised pretraining baselines on dermatology classification.

Architecture	Method	Pretraining Dataset	Top-1 Accuracy
ResNet-152 (2×)	Supervised	ImageNet	63.36 ± 0.12
ResNet-101 (3×)	BiT [24]	ImageNet-21k	68.45 ± 0.29
ResNet-152 (2×)	SimCLR	ImageNet	66.38 ± 0.03
ResNet-152 (2×)	SimCLR	ImageNet → Derm	69.43 ± 0.43
ResNet-152 (2×)	MICLe	ImageNet → Derm	**70.02** ± 0.22

Table 4: Comparison of best self-supervised models *vs.* supervised pretraining baselines on chest X-ray classification.

Architecture	Method	Pretraining Dataset	Mean AUC
ResNet-152 (2×)	Supervised	ImageNet	0.7625 ± 0.001
ResNet-101 (3×)	BiT [24]	ImageNet-21k	0.7720 ± 0.002
ResNet-152 (2×)	SimCLR	ImageNet	0.7671 ± 0.008
ResNet-152 (2×)	SimCLR	CheXpert	0.7702 ± 0.001
ResNet-152 (2×)	SimCLR	ImageNet → CheXpert	**0.7729** ± 0.001

Azizi, Shekoofeh, et al. "Big self-supervised models advance medical image classification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Additional References

1. Lilian Weng – Self-Supervised Representation Learning, link: Lil'Log
2. Pieter Abbeel, UC Berkley, CS294-158 Deep Unsupervised Learning, Lecture 7 – Self-Supervised Learning
3. Amit Chaudhary – The Illustrated Self-Supervised Learning, link
4. Jing and Tian (2019) Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey
5. William Falcon – A Framework for Contrastive Self-Supervised Learning and Designing a New Approach, link
6. Andrew Zisserman – Self-Supervised Learning, slides from: Carl Doersch, Ishan Misra, Andrew Owens, Carl Vondrick, Richard Zhang: https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf
7. Amit Chaudhary –Self-Supervised Representation Learning in NLP, link
8. Awesome Self-Supervised Learning: https://github.com/jason718/awesome-self-supervised-learning

Slide credit: Dr. Alex Vakanski

Thank you!

Question?