DEEP LEARNING

Lecture 12: Self-Supervised Learning

Dr. Yang Lu

Department of Computer Science and Technology

luyang@xmu.edu.cn

厦門大學信息学院（特色化示范性软件学院）
School of Informatics Xiamen University (National Characteristic Demonstration Software School)

厦门大学 计算机科学与技术系
Department of Computer Science and Technology, Xiamen University

Self-Supervised Learning

Lecun

Yann LeCun and Yoshua Bengio say at ICLR 2020:

Self-supervised learning could lead to the creation of artificial intelligence (AI) programs that are more humanlike in their reasoning.

Bengio

Image source: https://www.nytimes.com/2019/03/27/technology/turing-award-ai.html

Supervised and Unsupervised Learning

- Given a task and enough labels, supervised learning can solve it really well.
- However, good performance usually requires a decent amount of labels, but collecting manual labels is expensive (i.e. ImageNet) and hard to be scaled up.

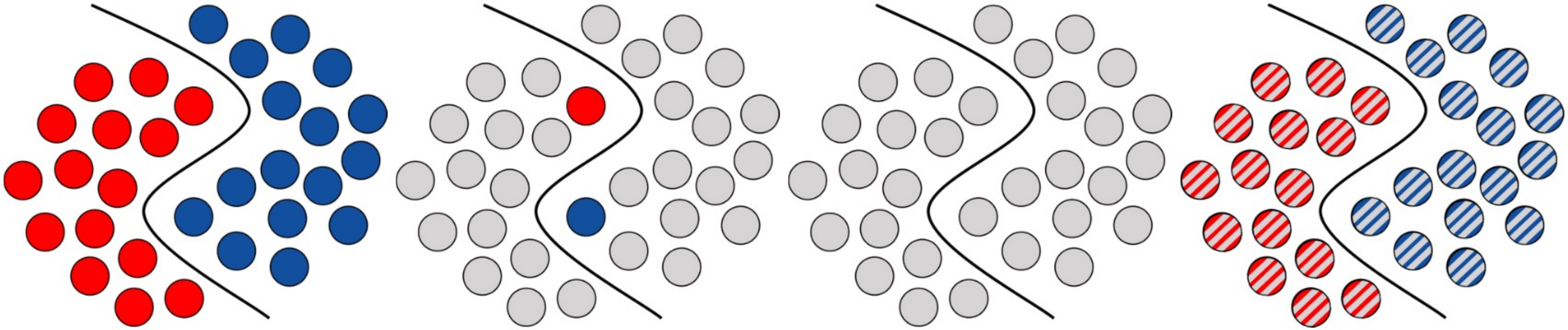
Supervised and Unsupervised Learning

- Unlabeled data (e.g. free text, all the images on the Internet) is substantially more than a limited number of human curated labelled datasets,
  - It is kind of wasteful not to use them.
- However, unsupervised learning is not easy and usually works much less efficiently than supervised learning.

Self-Supervised Learning

- What if we can automatically generate labels by some rules for unlabeled data and train unsupervised dataset in a supervised manner?
  - E.g. use a part of the data to predict the rest. The partition can be generated by rules, rather than human annotation.
- In this way, all the information needed, both inputs and labels, has been provided. This is known as self-supervised learning.
- The main purpose of self-supervised learning is to pre-train representations that can be transferred to downstream tasks by fine-tuning.

Self-Supervised Learning



(a) Supervised　(b) Semi-supervised　(c) Unsupervised　(d) Self-supervised

Image source:L Schmarje, Lars, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. "A survey on semi-, self-and unsupervised learning for image classification." arXiv preprint arXiv:2002.08721 2 (2020).

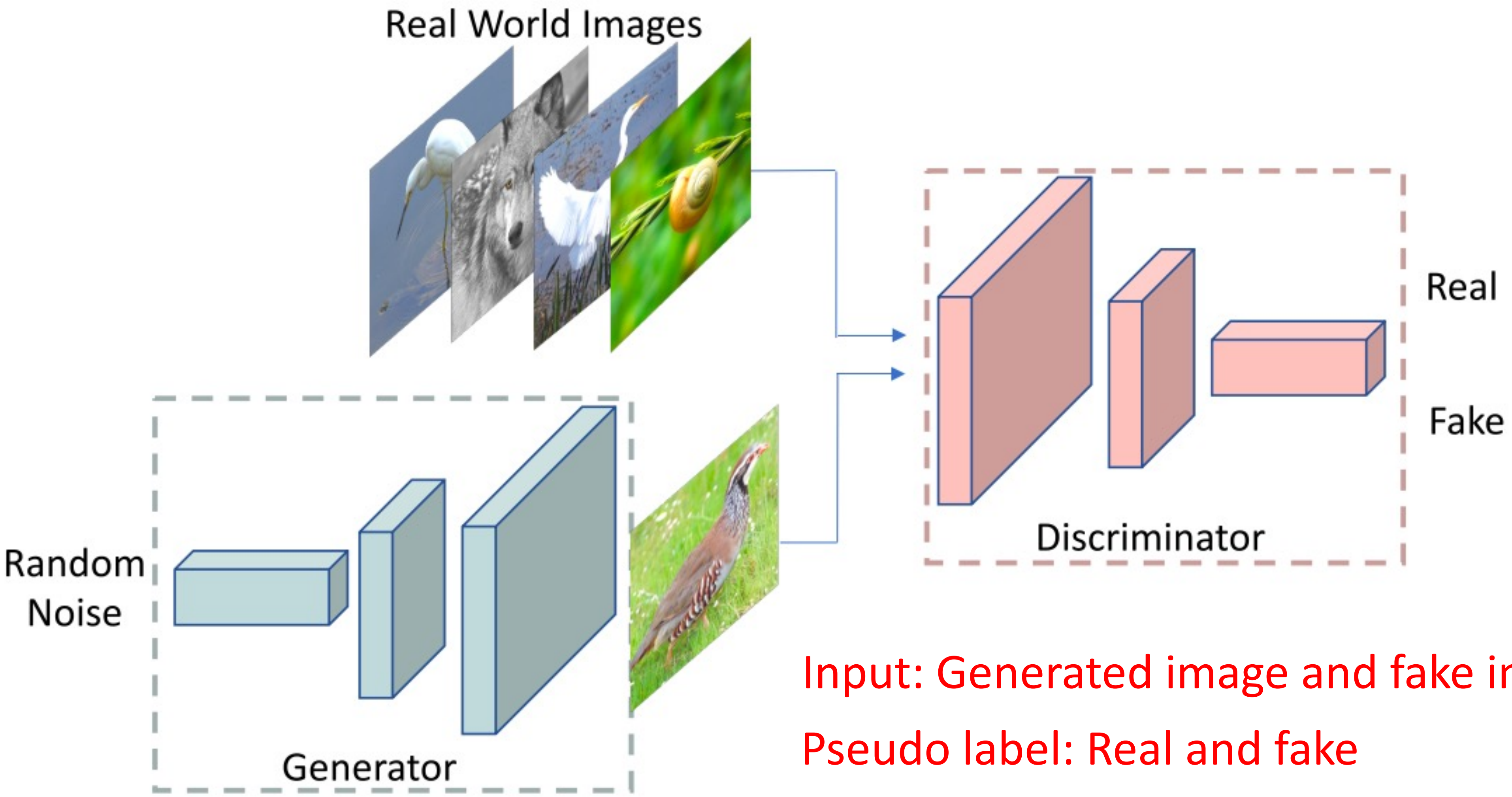
Self-Supervised Learning

We have seen examples of self-supervised learning.

- Word2vec uses center word to predict context words.
  - The label (context words) is generated by sliding window.
- BERT has two tasks:
  - Use mask token to predict the missing word.
  - Concat two sentences to predict their order.
- GAN uses real images and fake images as labels.
- Graph embedding uses neighbors as labels.

All the labels are automatically generated without human annotation for supervised learning task.

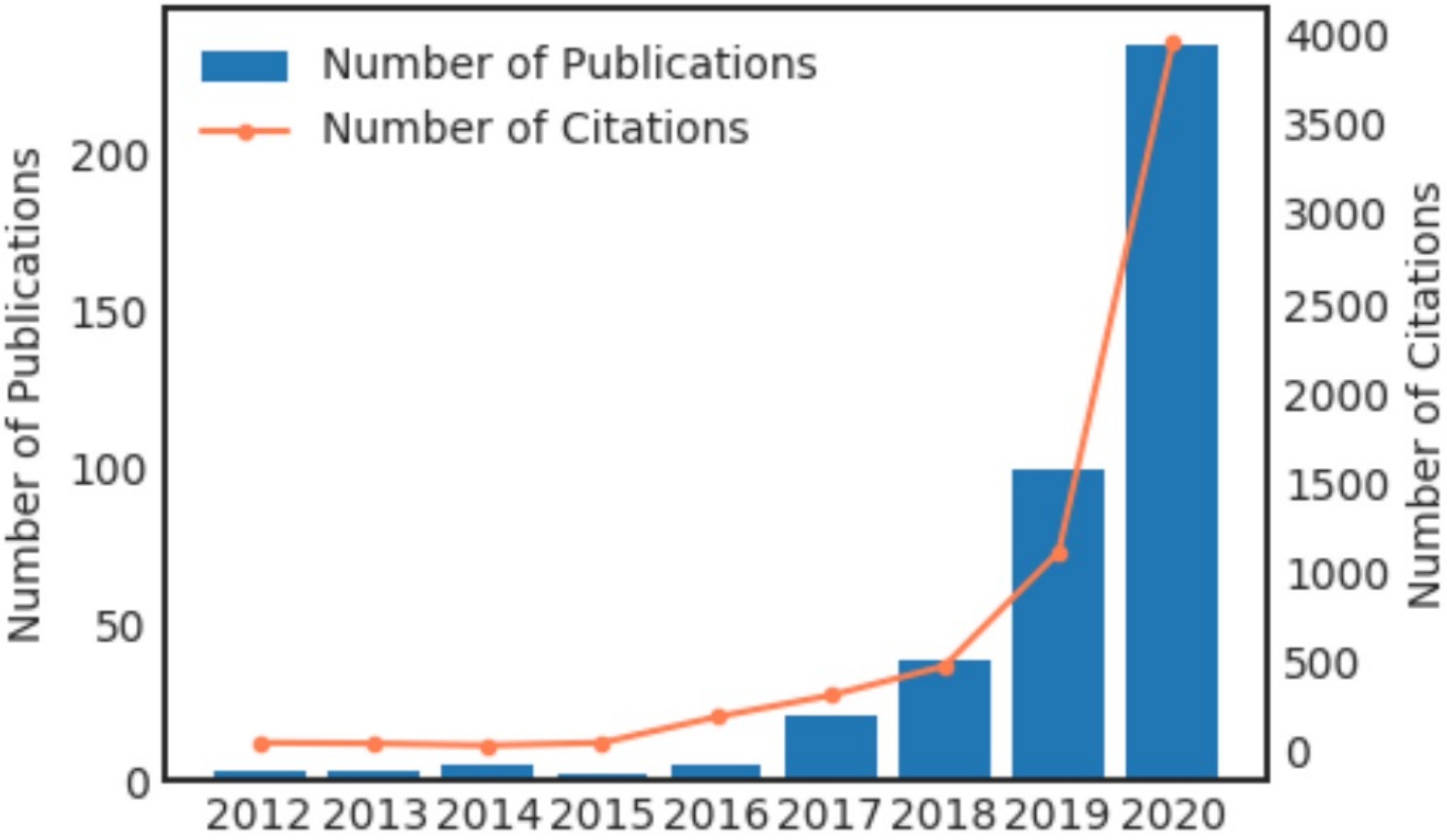
GAN



Input: Generated image and fake image

Pseudo label: Real and fake

Self-Supervised Learning



厦門大學信息学院（特色化示范性软件学院）
School of Informatics Xiamen University (National Characteristic Demonstration Software School)
厦门大学 计算机科学与技术系
Department of Computer Science and Technology, Xiamen University

Image source: Liu, Xiao, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. "Self-supervised learning: Generative or contrastive." arXiv preprint arXiv:2006.08218 1, no. 2 (2021).

Term Definition

- Pretext Task: Pre-designed tasks for networks to solve, in order to learn features as a pre-trained model.
- Downstream Task: Applications that are used to evaluate the quality of features learned by self-supervised learning.
- Human-annotated label: Labels of data that are manually annotated by human workers.
- Pseudo label: Automatically generated labels based on data attributes for pretext tasks.

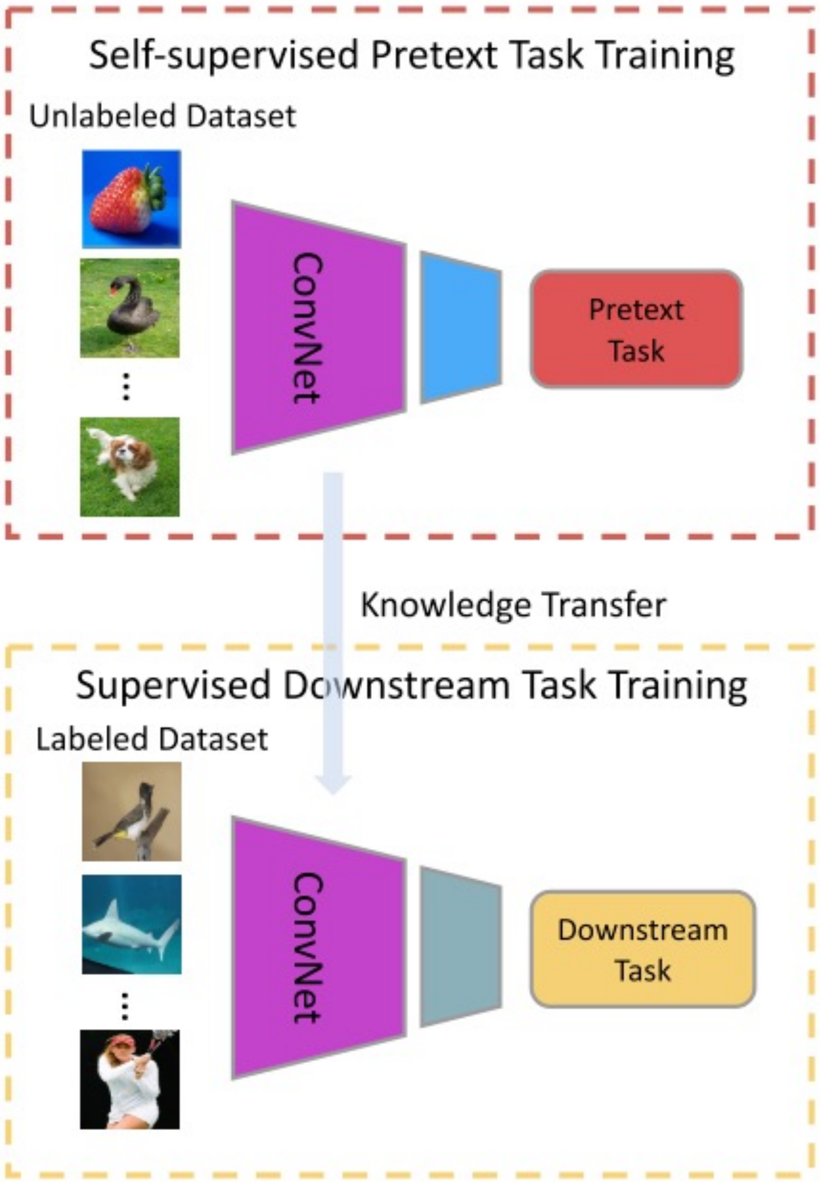


Image source: Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

Outlines

- Generation-Based Methods
- Context-Based Methods
- Free Semantic Label-Based Methods
- Cross Modal-Based Methods
- Contrastive Learning

GENERATION-BASED METHODS

Generation-Based Methods

- Idea: Use modified image to generate original image.
- The generator is able to learn image features by the loss between generated image and original image.
- The pseudo label is usually the original image.

Image Generation with Inpainting

Input:
Image with missing region

Pseudo label:
Original Image

Image source: Zeng, Yanhong, Jianlong Fu, Hongyang Chao, and Baining Guo. "Learning pyramid-context encoder network for high-quality image inpainting." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1486-1494. 2019.

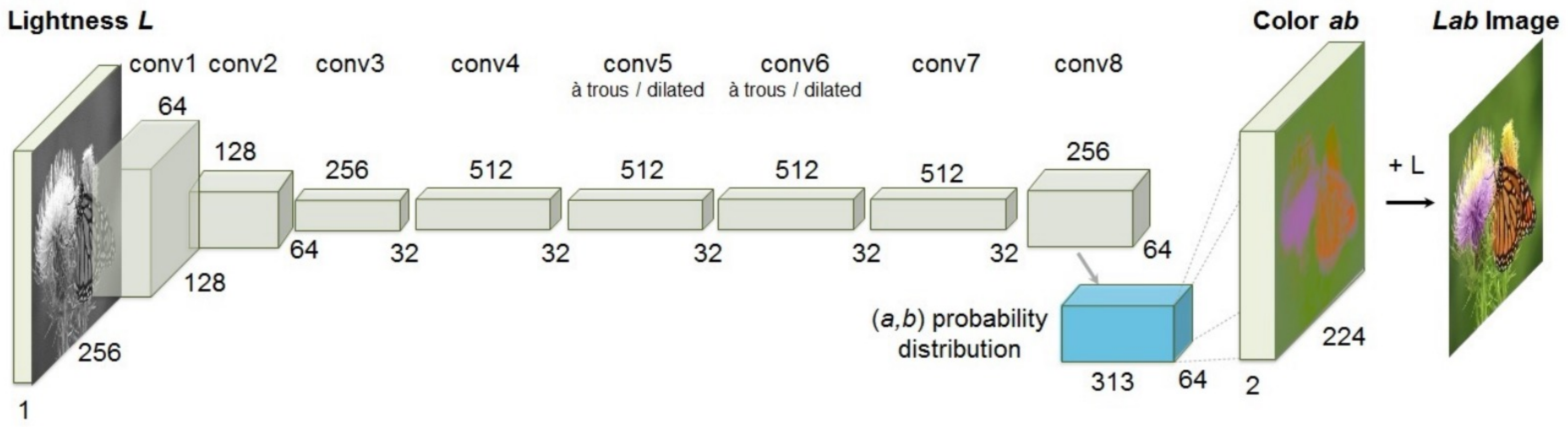
Image Generation with Super Resolution

Input:
Low resolution image

Pseudo label:
High resolution mage

Image source: Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken et al. "Photo-realistic single image super-resolution using a generative adversarial network." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681-4690. 2017.

Image Generation with Colorization



Input:
Transformed grey level image

Pseudo label:
Original colorful mage

Image source: Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." In European conference on computer vision, pp. 649-666. Springer, Cham, 2016.

MAE

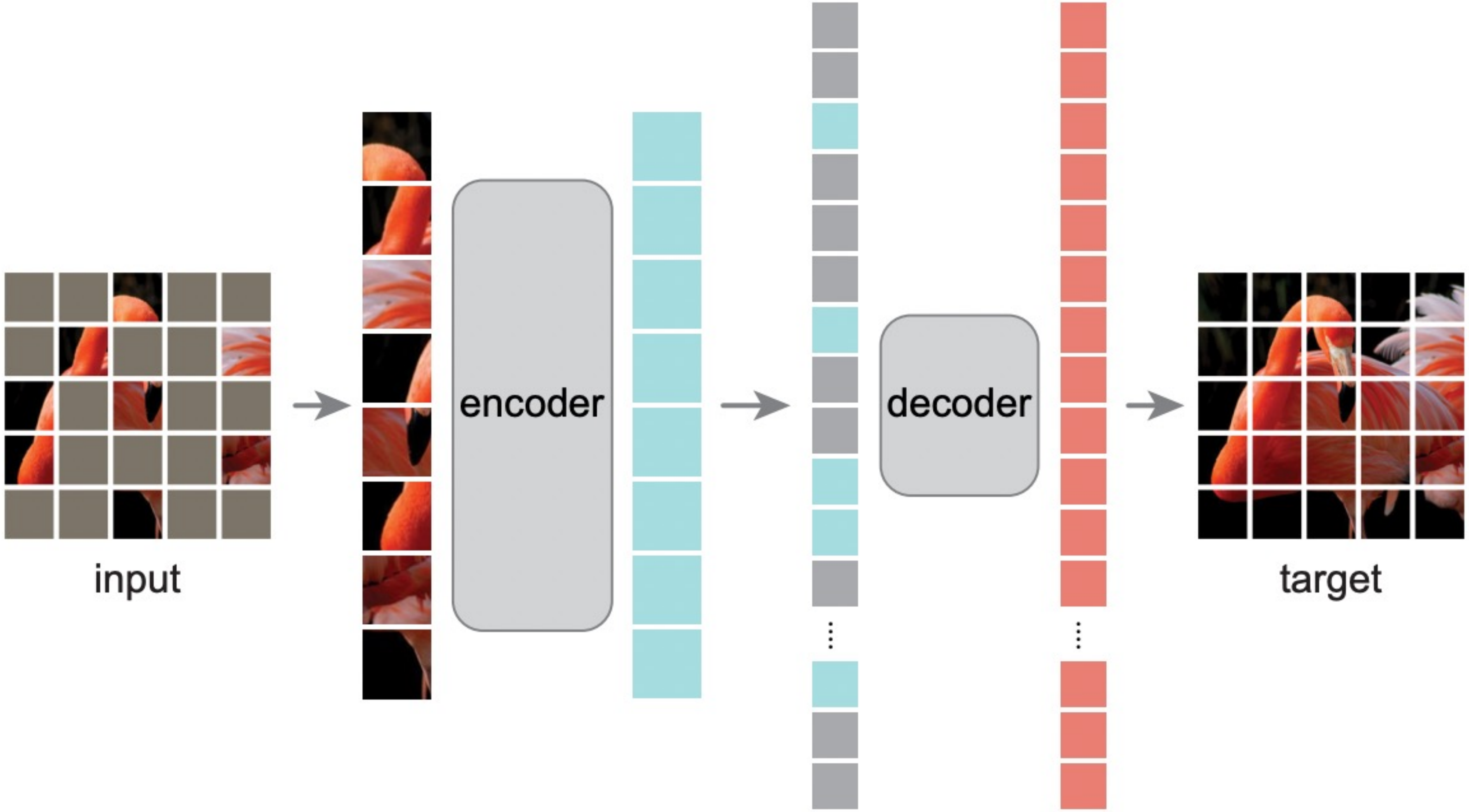


Image source: He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked autoencoders are scalable vision learners." arXiv preprint arXiv:2111.06377 (2021).

MAE

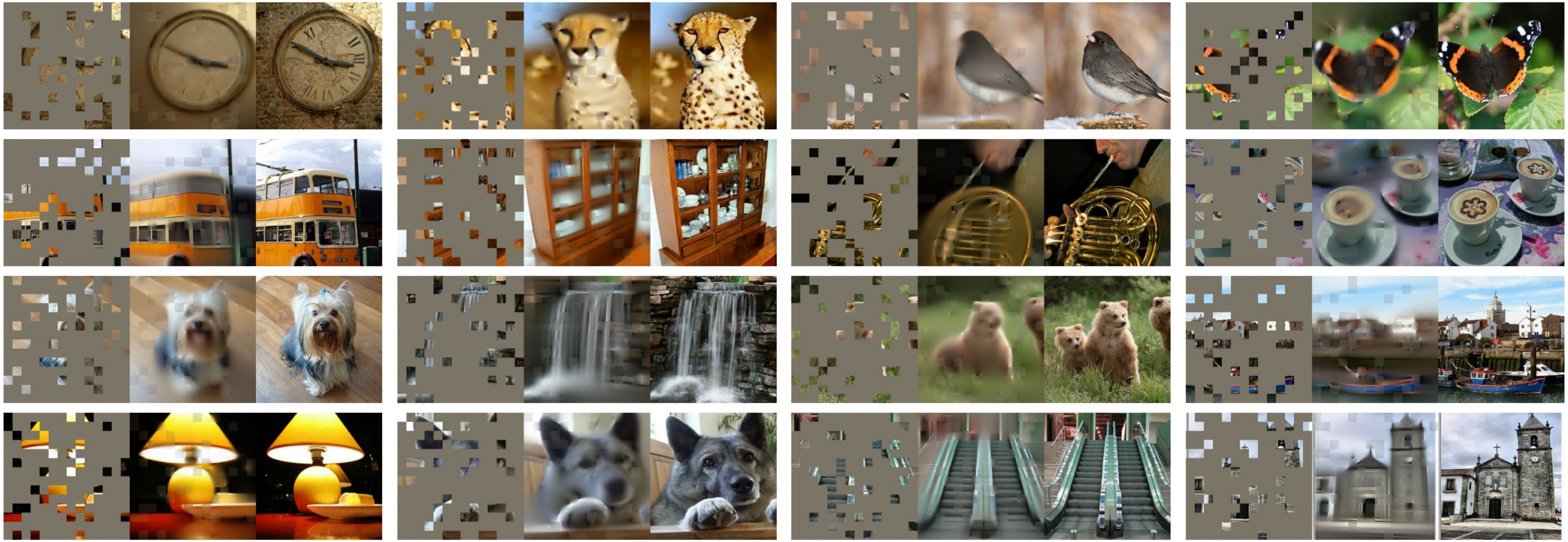


Image source: He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked autoencoders are scalable vision learners." arXiv preprint arXiv:2111.06377 (2021).

Video Generation with Colorization

Reference colored frame | Input video | Predicted colorized video

Video source: https://ai.googleblog.com/2018/06/self-supervised-tracking-via-video.html

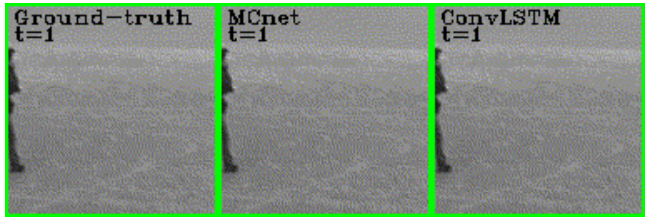
Video Generation with Colorization

After learning to colorize videos, a mechanism for tracking automatically emerges without supervision.

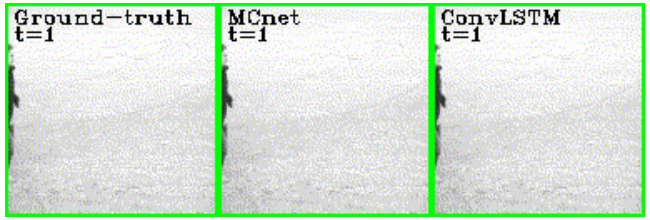
Video source: https://ai.googleblog.com/2018/06/self-supervised-tracking-via-video.html

Video Prediction

Running



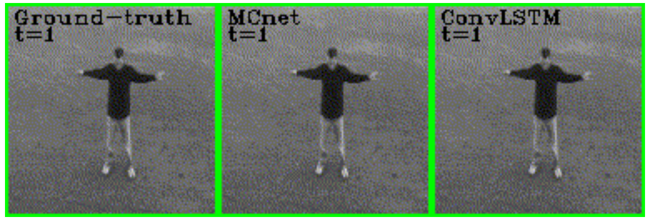
Jogging



Walking

Boxing

Handclapping



Handwaving

All models are trained to observe 10 frames (green) and predict 10 frames (red)

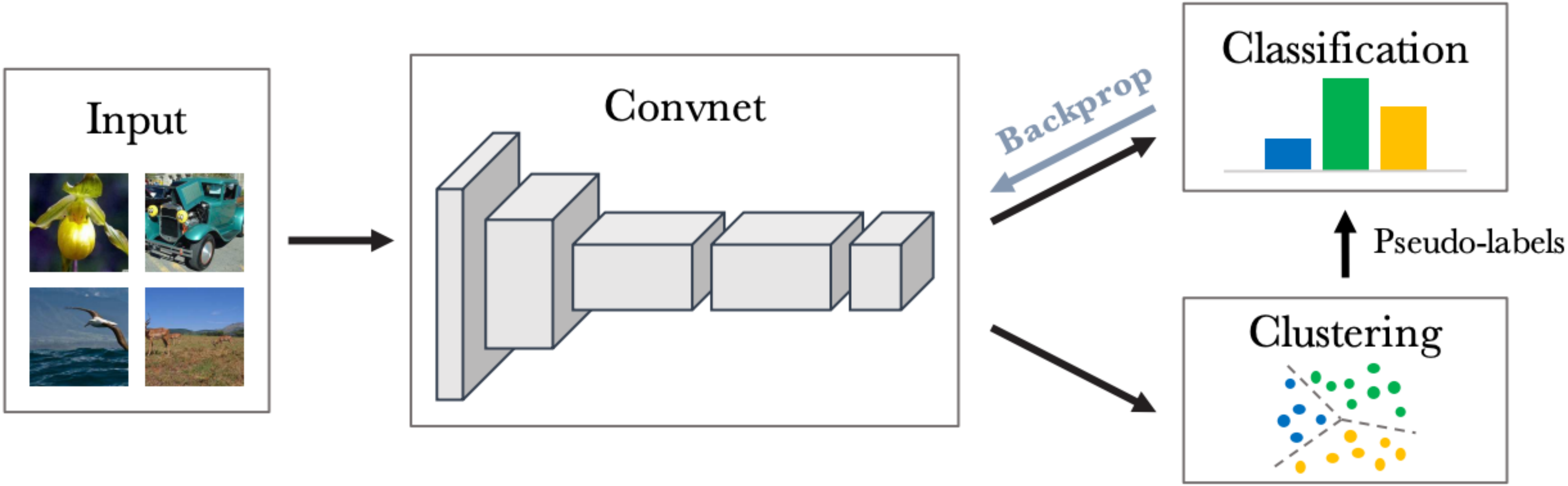
Video source: https://sites.google.com/a/umich.edu/rubenevillegas/iclr2017

CONTEXT-BASED METHODS

Context-Based Methods

- The context-based pretext tasks mainly employ the context features of images as the supervision signal, including
  - context similarity;
  - spatial structure;
  - temporal structure;
  - …

Context Similarity

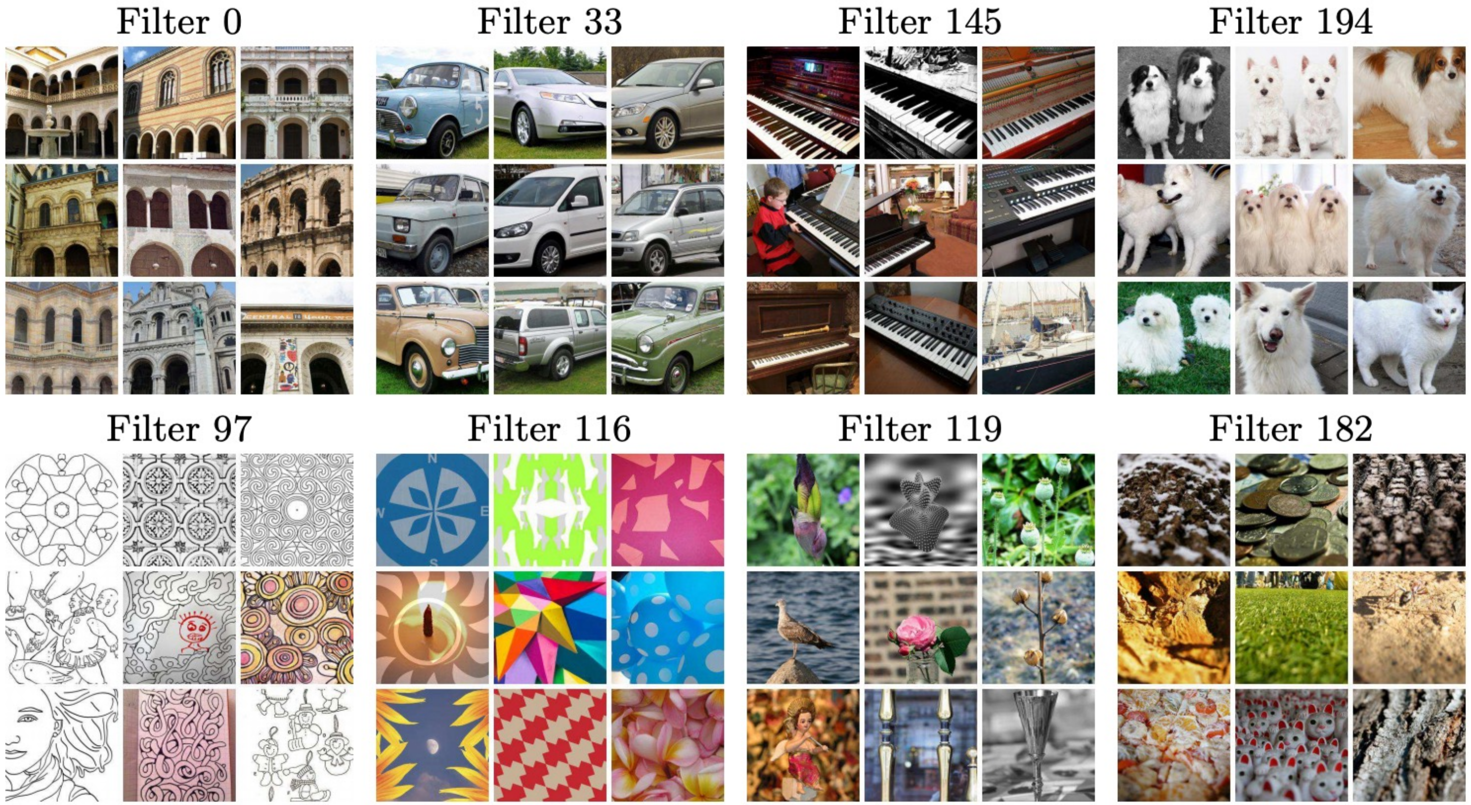


Iteratively cluster deep features and use the cluster assignments as pseudo-labels to learn the parameters of CNN

Image source: Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 132-149. 2018.

Context Similarity

- Clustering



Sensitive to specific objects

Sensitive to stylistic effect

Top 9 activated images from a random subset of 10 millions images from YFCC100M for target filters in the last convolutional layer.

Image source: Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 132-149. 2018.

Context Similarity

- Clustering

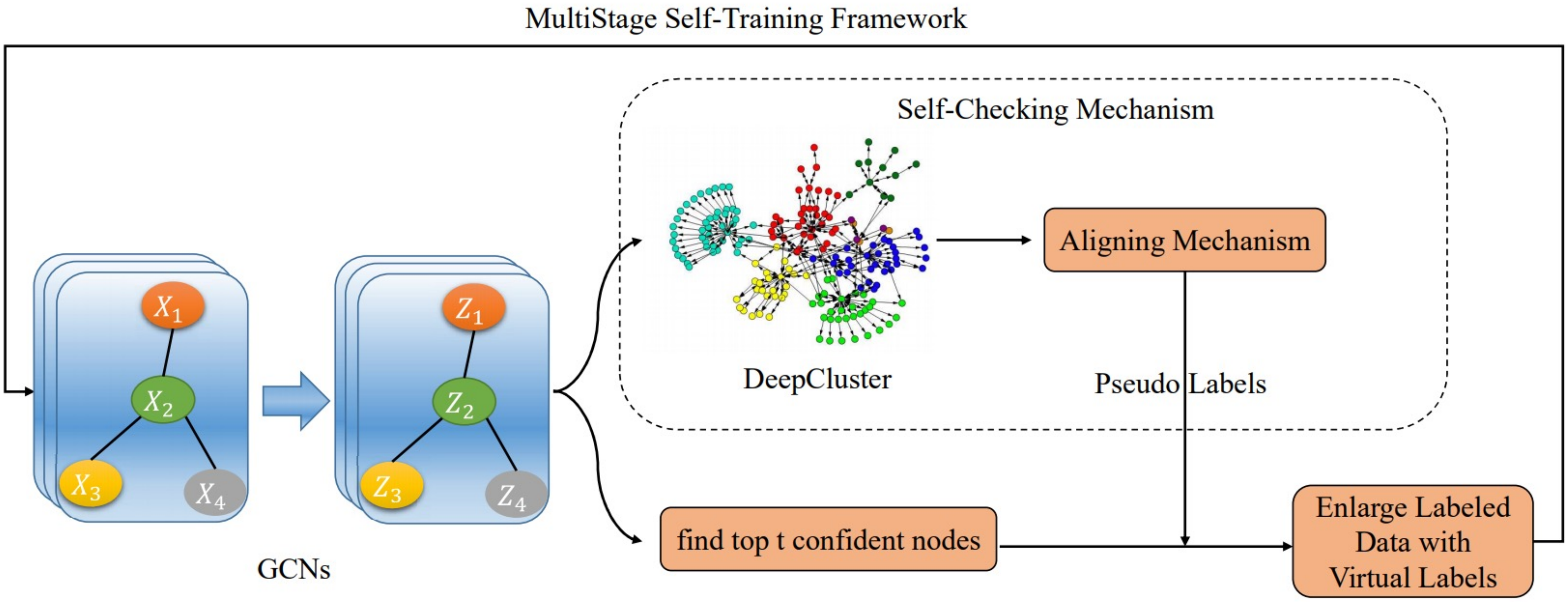


Image source: Sun, Ke, Zhanxing Zhu, and Zhouchen Lin. "Multi-stage self-supervised learning for graph convolutional networks." arXiv preprint arXiv:1902.11038 (2019).

Spatial Context Structure

- Rotation prediction

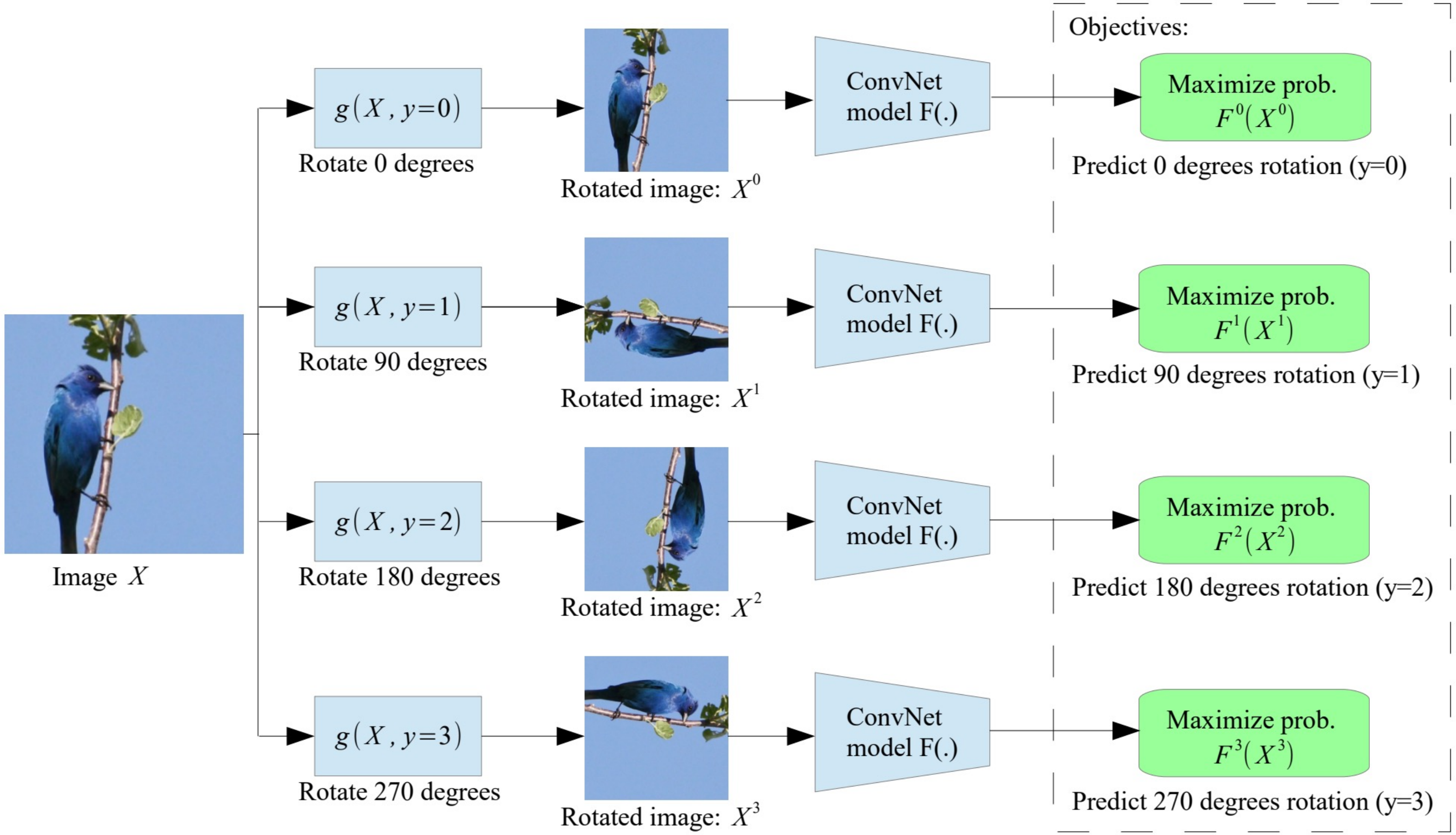
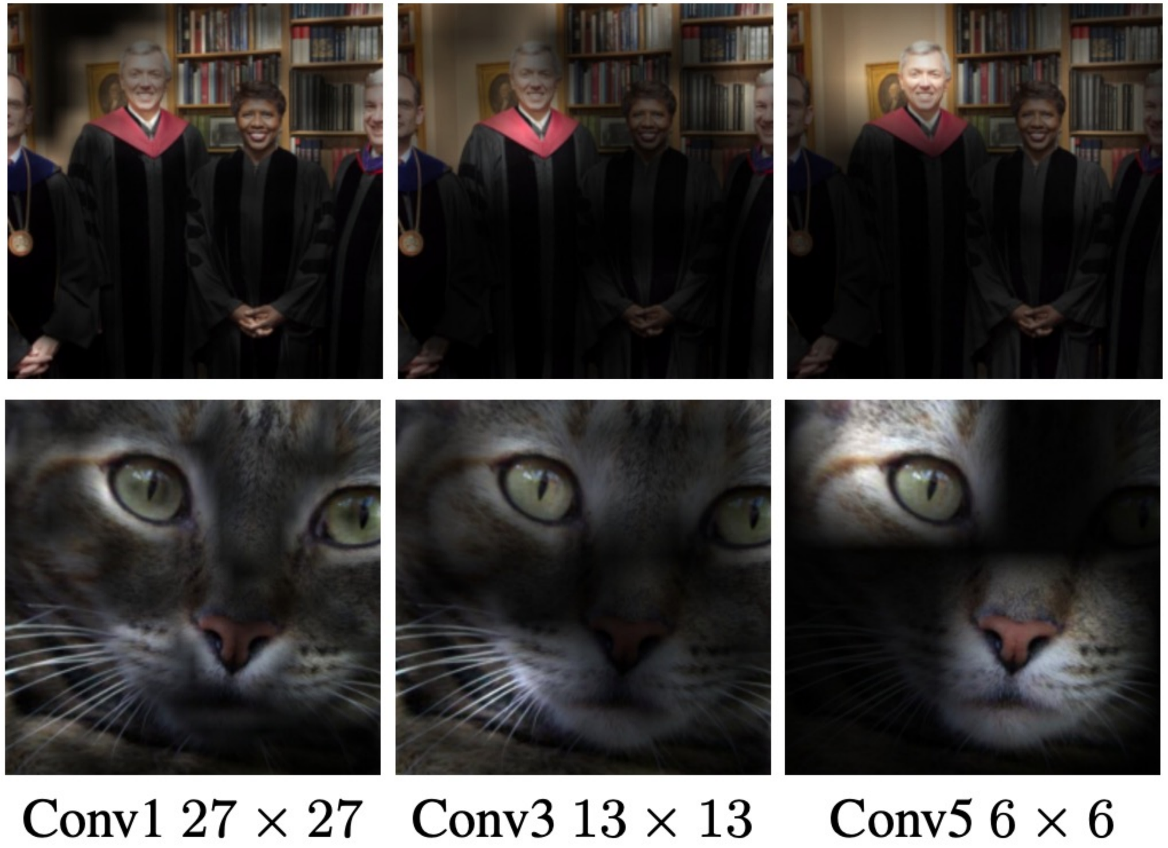


Image source: Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).

Spatial Context Structure

- Rotation prediction



Attention maps of self-supervised model

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

Spatial Context Structure

- Relative position prediction

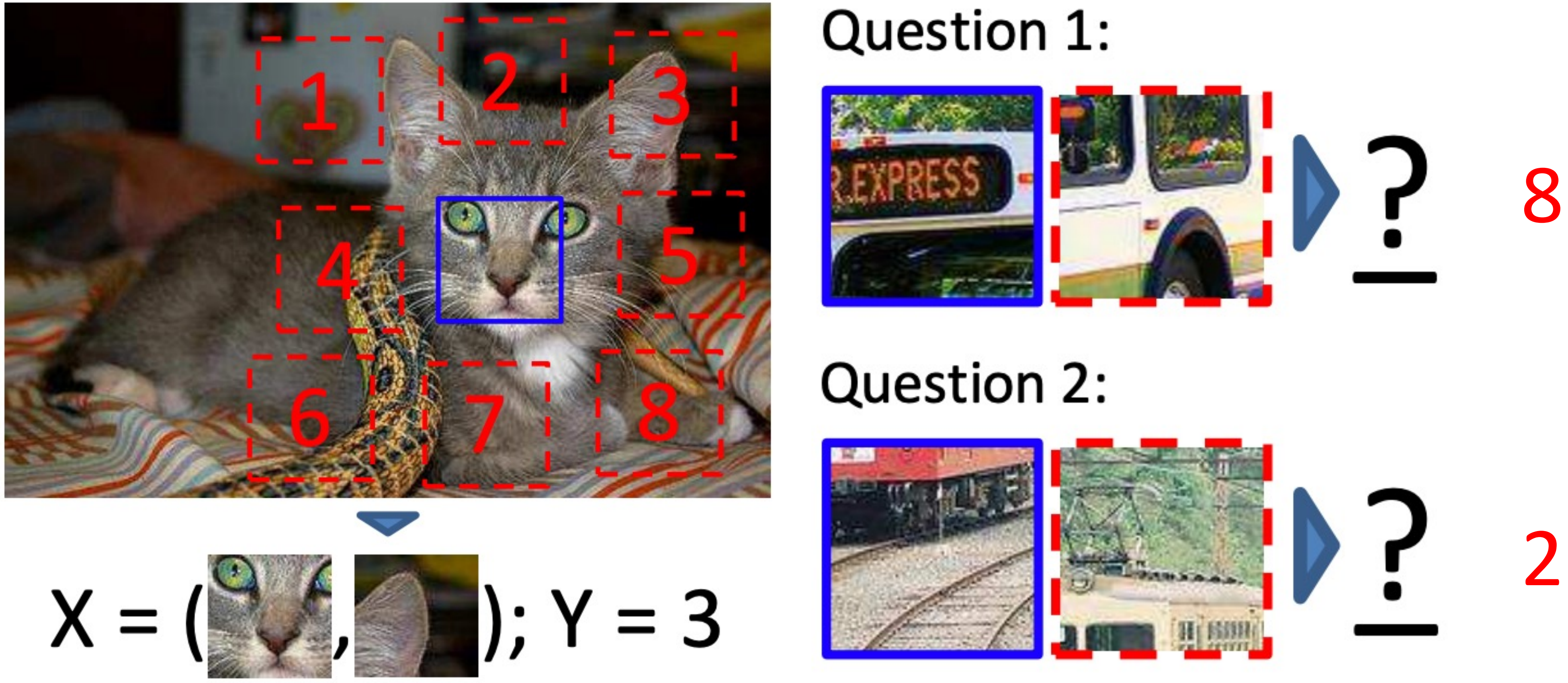


Image source: Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." In Proceedings of the IEEE international conference on computer vision, pp. 1422-1430. 2015.

Spatial Context Structure

- Jigsaw puzzle

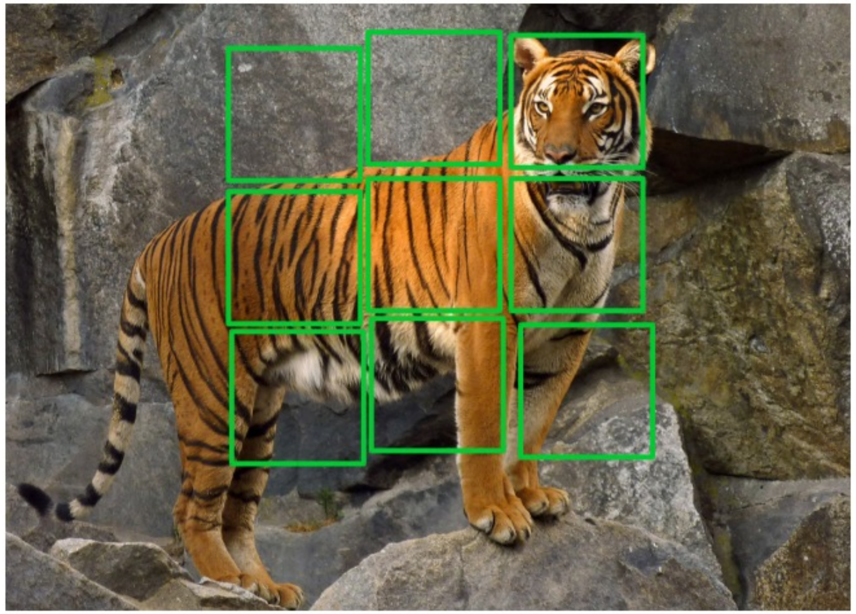


Image with 9 sampled image patches

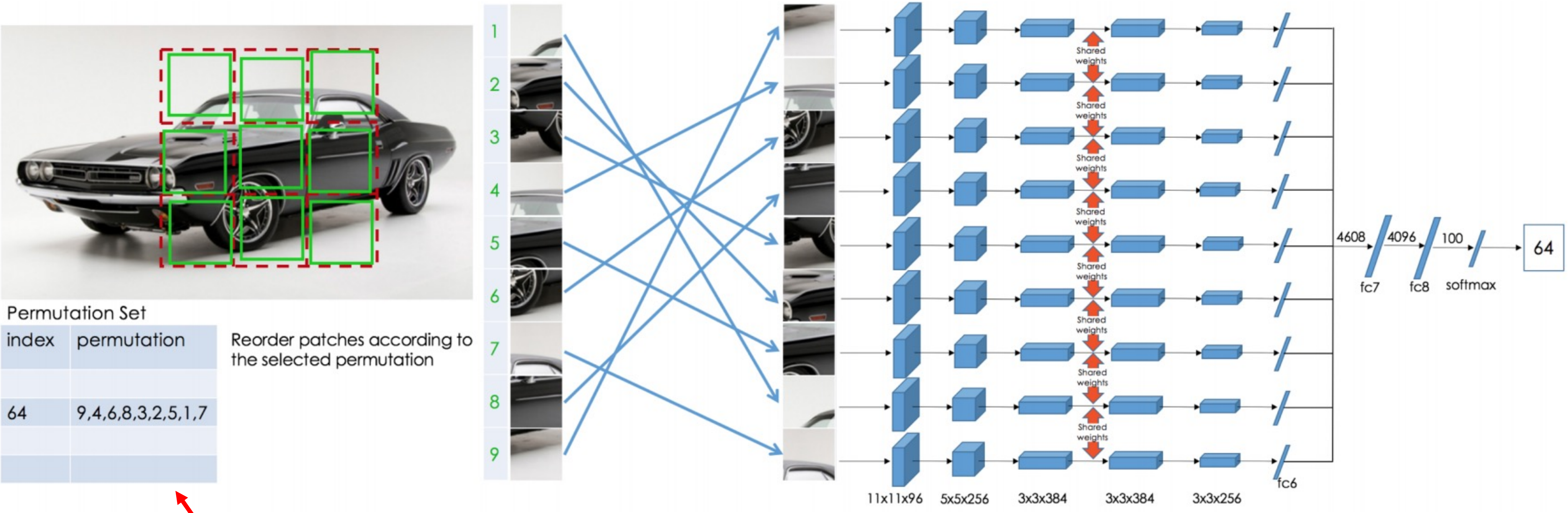
Shuffled image patches

Correct order of the sampled 9 patches

Image source: Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." In European Conference on Computer Vision, pp. 69-84. Springer, Cham, 2016.

Spatial Context Structure

- Jigsaw puzzle



How many classes? 9!=362,880.

Image source: Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." In European Conference on Computer Vision, pp. 69-84. Springer, Cham, 2016.

Spatial Context Structure

- Impossible to iterate over all possible permutation.
  - Similar permutation is somehow redundant.
- Choose permutations with max Hamming distance.

Number of permutations	Average hamming distance	Minimum hamming distance	Jigsaw task accuracy	Detection performance
1000	8.00	2	71	**53.2**
1000	6.35	2	62	51.3
1000	3.99	2	54	50.2

Image source: Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." In European Conference on Computer Vision, pp. 69-84. Springer, Cham, 2016.

Temporal Context Structure

- Temporal order verification

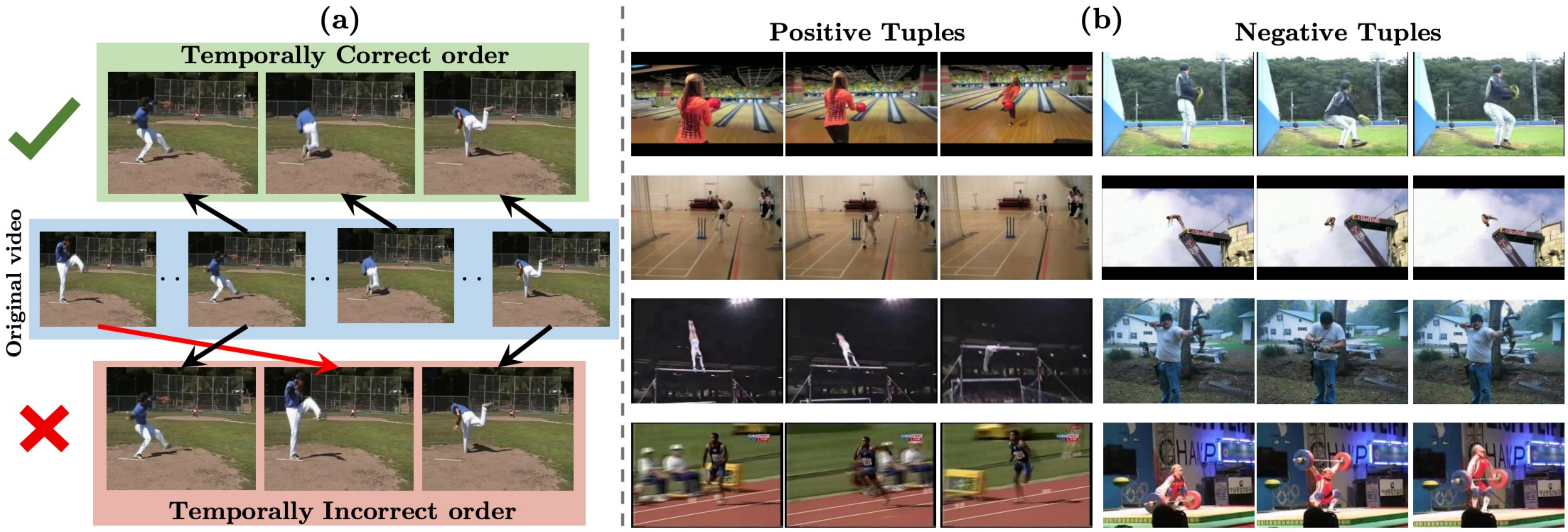


Image source: Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification." In European Conference on Computer Vision, pp. 527-544. Springer, Cham, 2016.

Temporal Context Structure

- Temporal order prediction

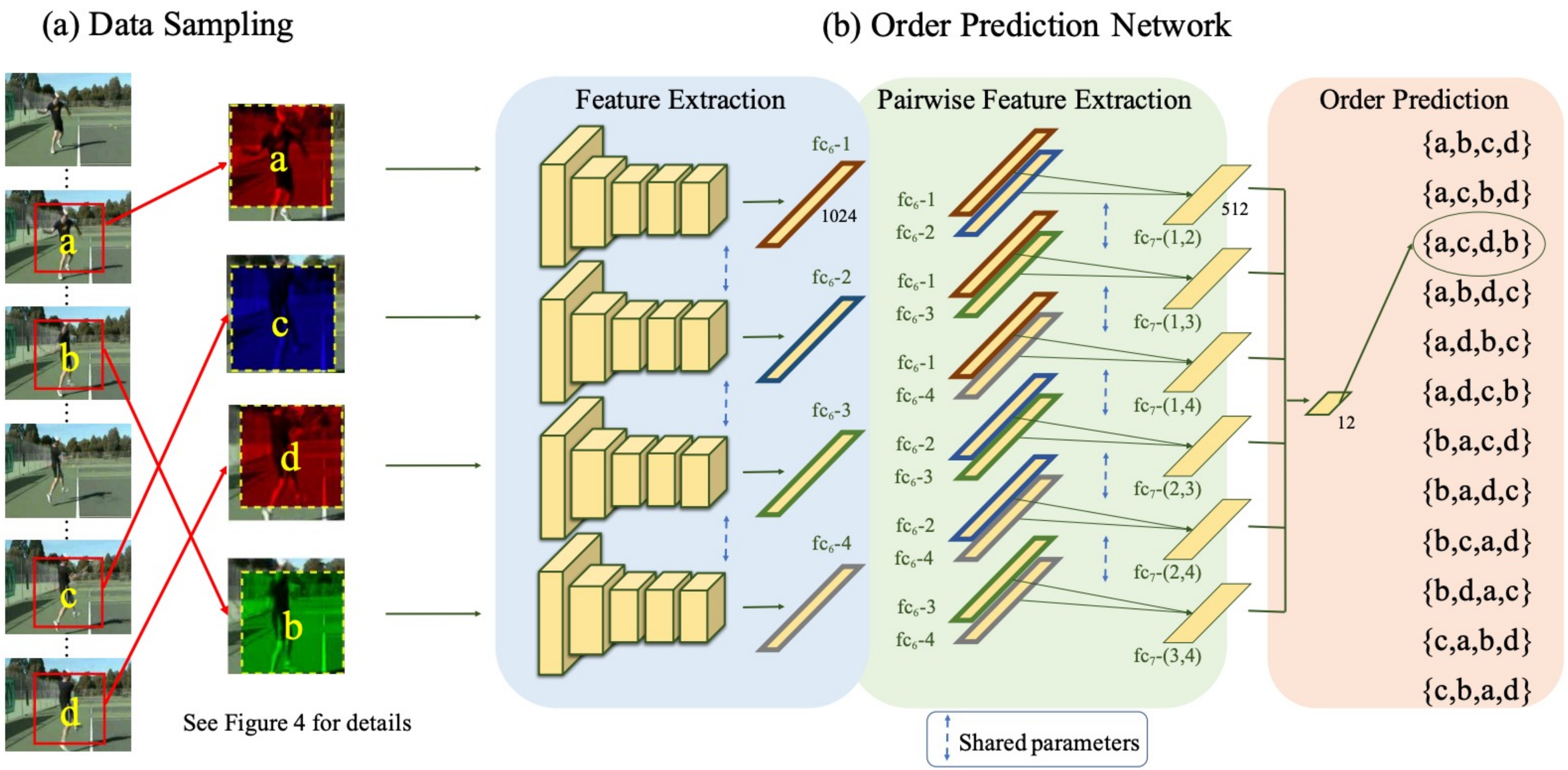


Image source: Lee, Hsin-Ying, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. "Unsupervised representation learning by sorting sequences." In Proceedings of the IEEE International Conference on Computer Vision, pp. 667-676. 2017.

Sentence Context Structure

- Emoji prediction

Image source: https://amitness.com/2020/05/self-supervised-learning-nlp/

Sentence Context Structure

- Sentence permutation and rotation

I did X. Then I did Y. Finally I did Z.

I am going outside. I will be back in the evening.

original text

Image source: https://amitness.com/2020/05/self-supervised-learning-nlp/

Sentence Context Structure

- Gap sentence generation

Image source: https://amitness.com/2020/05/self-supervised-learning-nlp/

FREE SEMANTIC LABEL-BASED METHODS

Free Semantic Label-Based Methods

- Self-supervised learning requires no human annotations.
- Alternatively, we may obtain some semantic information as labels by.
  - Game engines: generate realistic images with accurate pixel-level labels with very low cost.
  - Auxiliary automatic annotators: generate salience, foreground masks, contours, depth for images and videos.

Game Engines

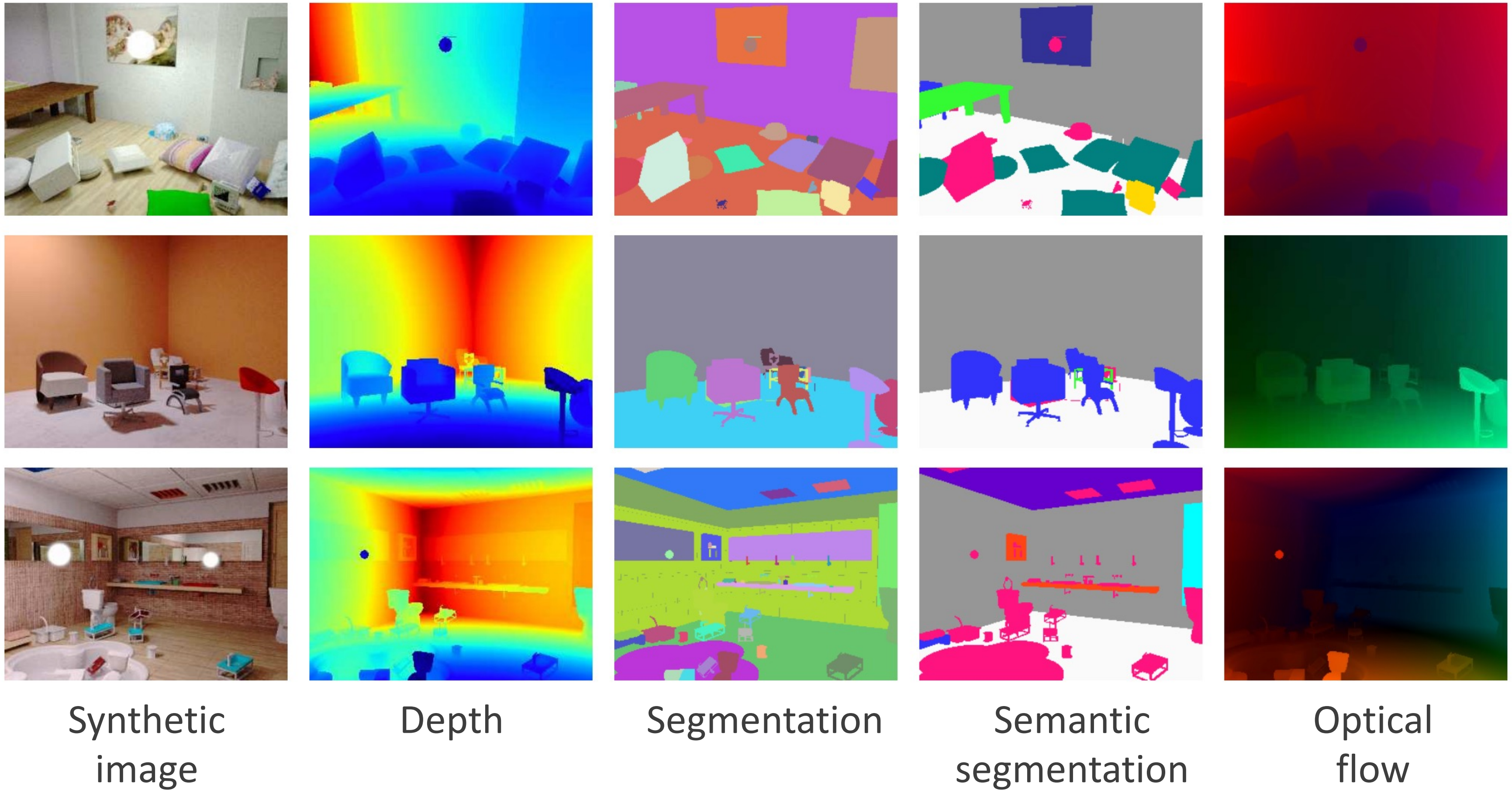
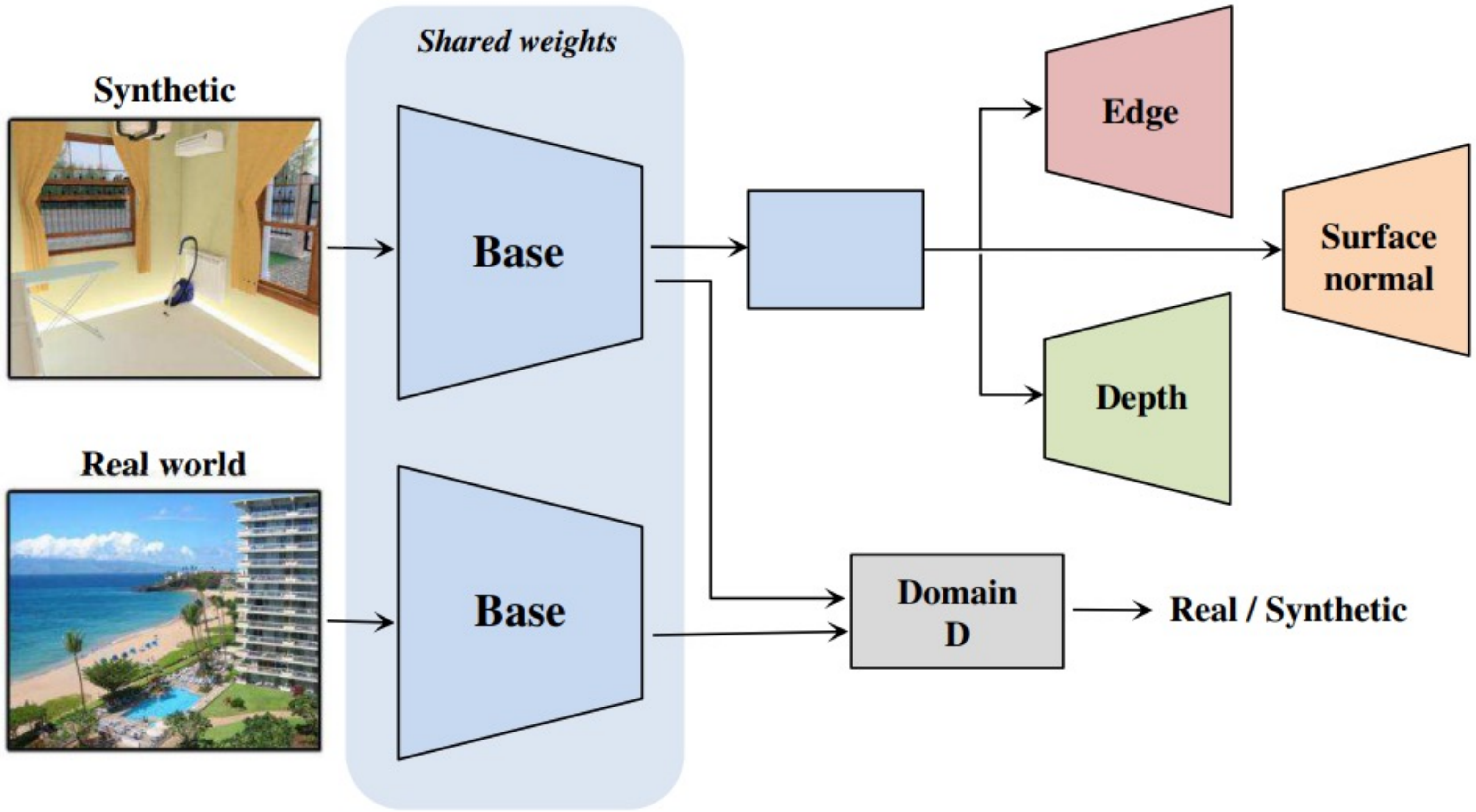


Image source: McCormac, John, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. "SceneNet RGB-D: Can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation?." In Proceedings of the IEEE International Conference on Computer Vision, pp. 2678-2687. 2017.

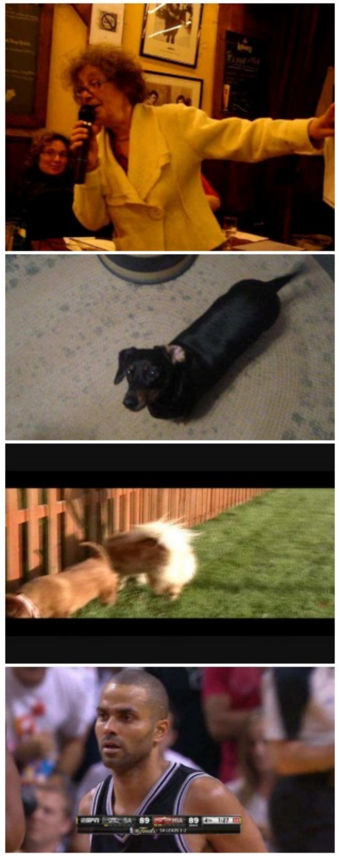
Game Engines



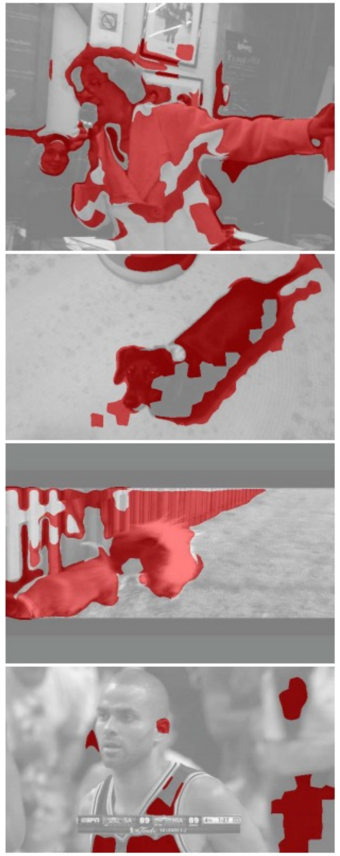
A discriminator network D is employed to minimize the difference of feature space domains between real-world and synthetic data

Image source: Ren, Zhongzheng, and Yong Jae Lee. "Cross-domain self-supervised multi-task feature learning using synthetic imagery." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 762-771. 2018.

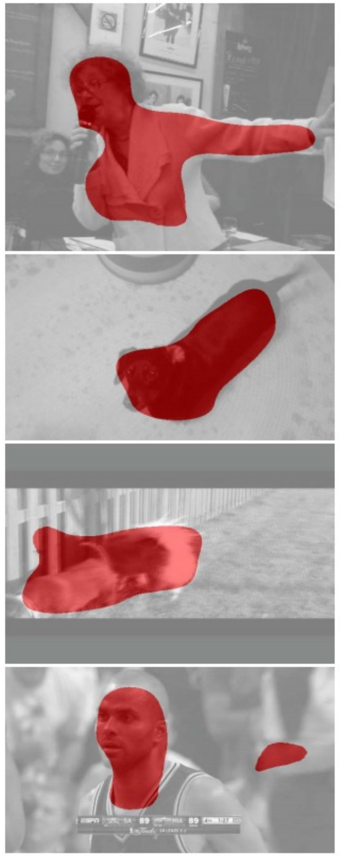
Auxiliary Automatic Annotators



A video frame



Auxiliary motion detector



Trained detector

Image source: Pathak, Deepak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. "Learning features by watching objects move." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2701-2710. 2017.

Auxiliary Automatic Annotators

- Top: input image.
- Middle: relative depth image computed using a formula.
- Bottom: Predicted depth maps using our trained model.

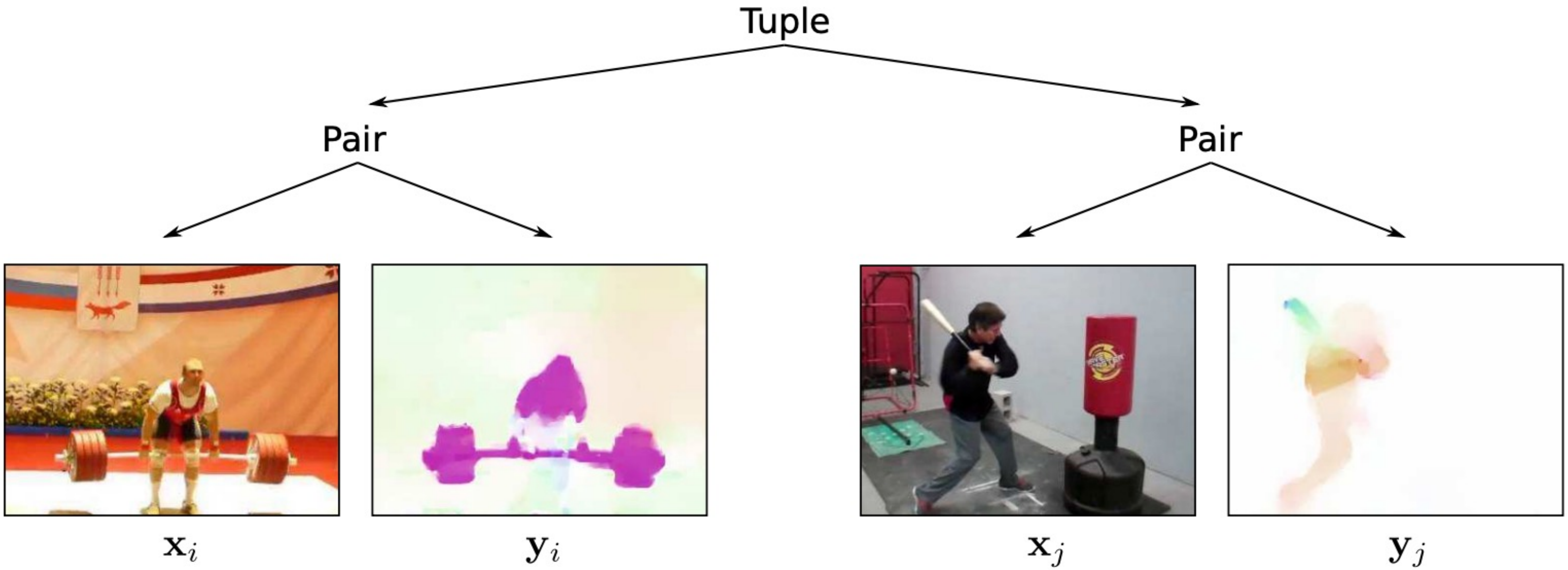
Image source: Jiang, Huaizu, Gustav Larsson, Michael Maire Greg Shakhnarovich, and Erik Learned-Miller. "Self-supervised relative depth learning for urban scene understanding." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 19-35. 2018.

CROSS MODAL-BASED METHODS

Cross Modal-based Learning

- Use different modal as pseudo label.
  - Optical flow;
  - Audio;
  - Text;
  - Camera poses…

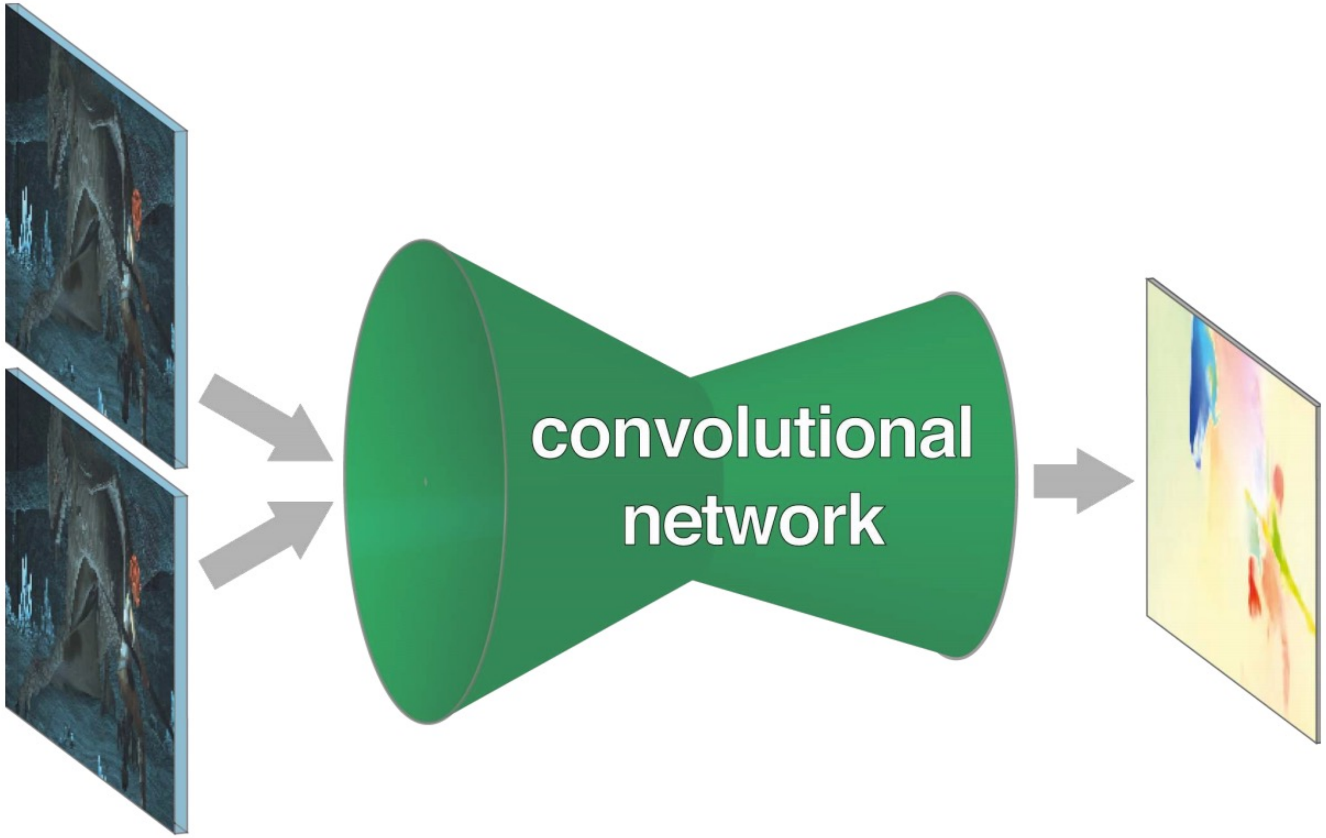
RGB-Flow Correspondence



Optical flow is another modal that can be used as pseudo label.

Image source: Sayed, Nawid, Biagio Brattoli, and Björn Ommer. "Cross and learn: Cross-modal self-supervision." In German Conference on Pattern Recognition, pp. 228-243. Springer, Cham, 2018.

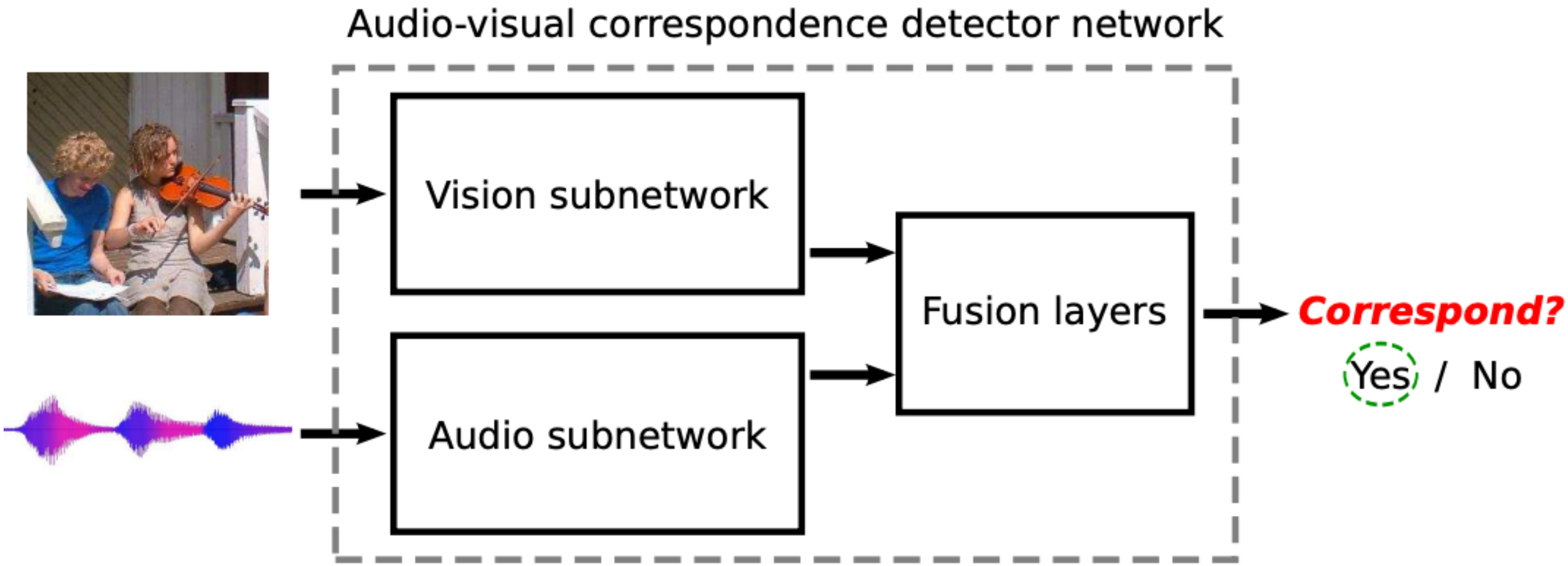
RGB-Flow Correspondence



Large scale optical flow modal is also hard to obtain. It can also be generated by some auxiliary algorithm.

Image source: Dosovitskiy, Alexey, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. "Flownet: Learning optical flow with convolutional networks." In Proceedings of the IEEE international conference on computer vision, pp. 2758-2766. 2015.

Visual-Audio Correspondence



Learn to determine whether a pair of video and audio clip correspond to each other or not

Image source: Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." In Proceedings of the IEEE International Conference on Computer Vision, pp. 609-617. 2017.

Visual-Audio Correspondence

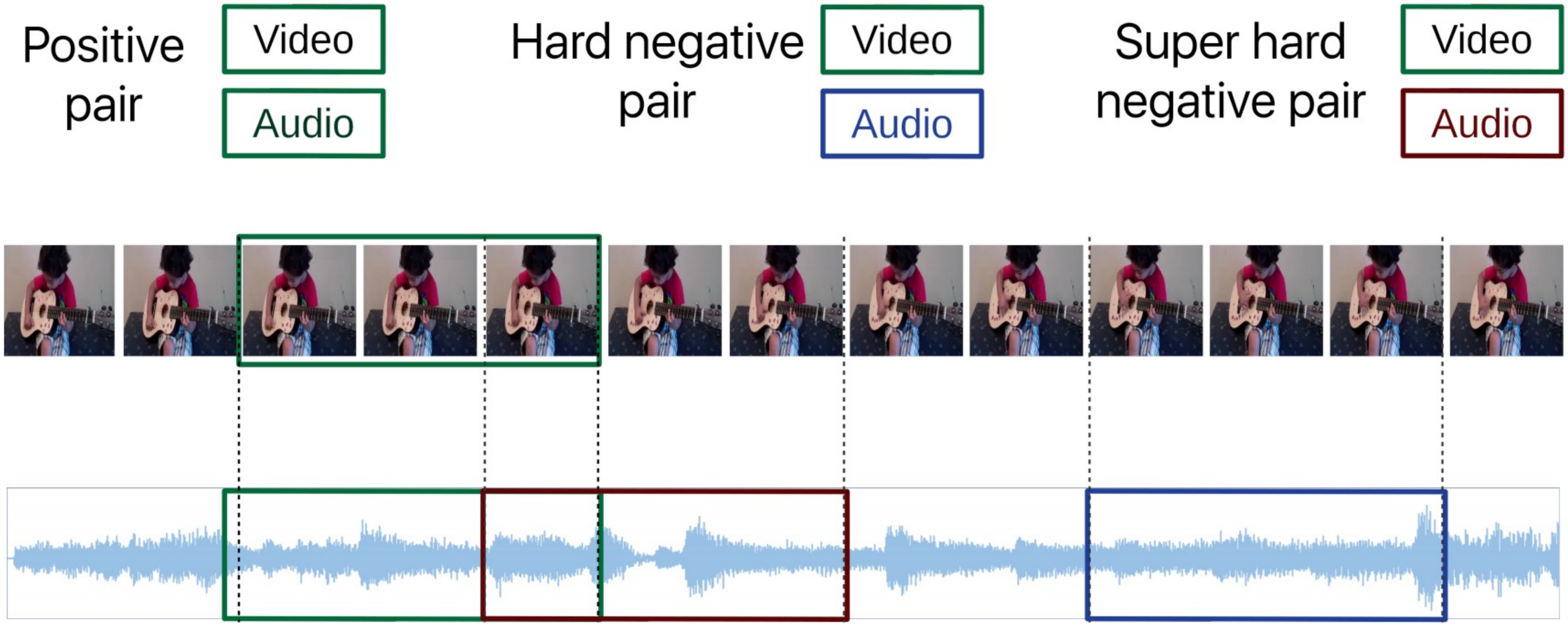
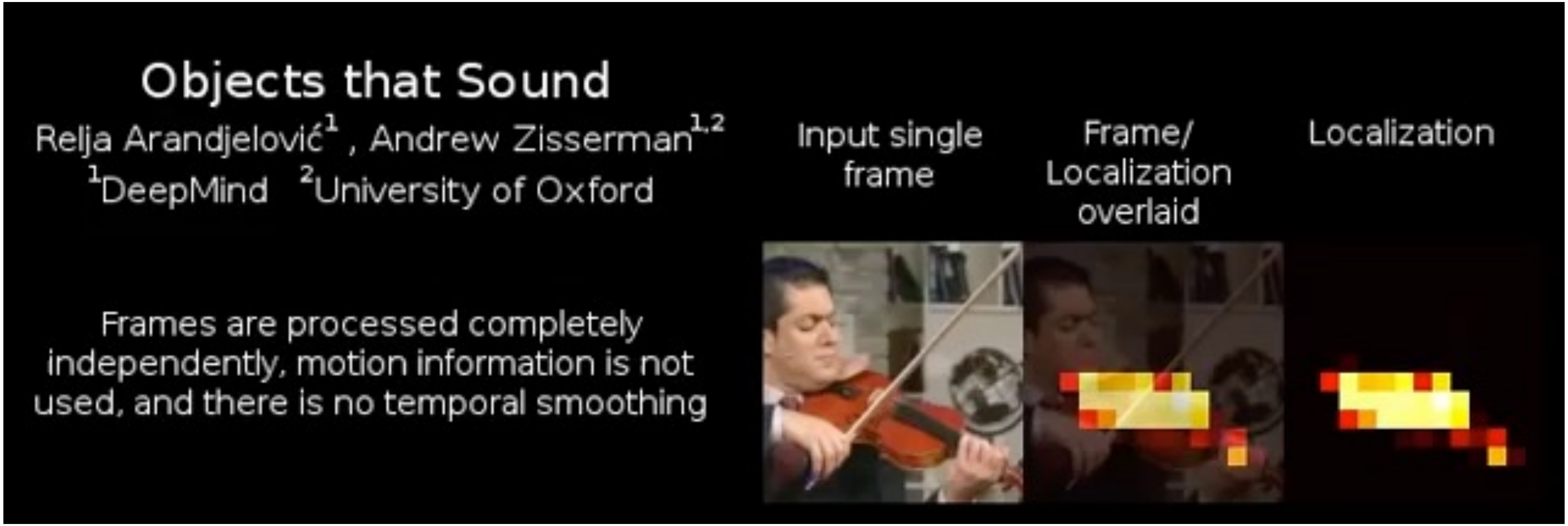


Image source: Korbar, Bruno, Du Tran, and Lorenzo Torresani. "Cooperative learning of audio and video models from self-supervised synchronization." Advances in Neural Information Processing Systems 31 (2018): 7763-7774.

Visual-Audio Correspondence

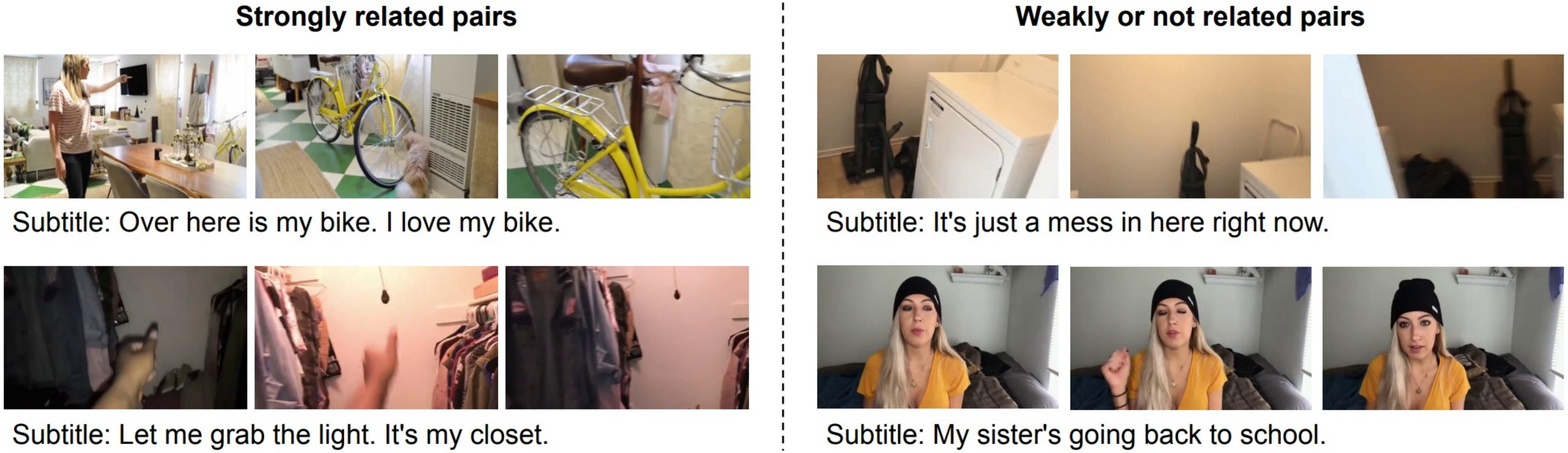


Localizing objects that sound

Video source: https://www.youtube.com/watch?v=TFyohksFd48

Reference: Arandjelovic, Relja, and Andrew Zisserman. "Objects that sound." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 435-451. 2018.

Visual-Text Correspondence



Use subtitle as supervision. But subtitles usually contain large number of noises.

Image source: Zhong, Yujie, Linhai Xie, Sen Wang, Lucia Specia, and Yishu Miao. "Watch and Learn: Mapping Language and Noisy Real-world Videos with Self-supervision." arXiv preprint arXiv:2011.09634 (2020).

CONTRASTIVE LEARNING

Contrastive Learning

Generate ⇄ Distinguish

?

Do we have to able to draw cash, in order to distinguish cash?

Image source: https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html

Generative, Predictive and Contrastive Methods

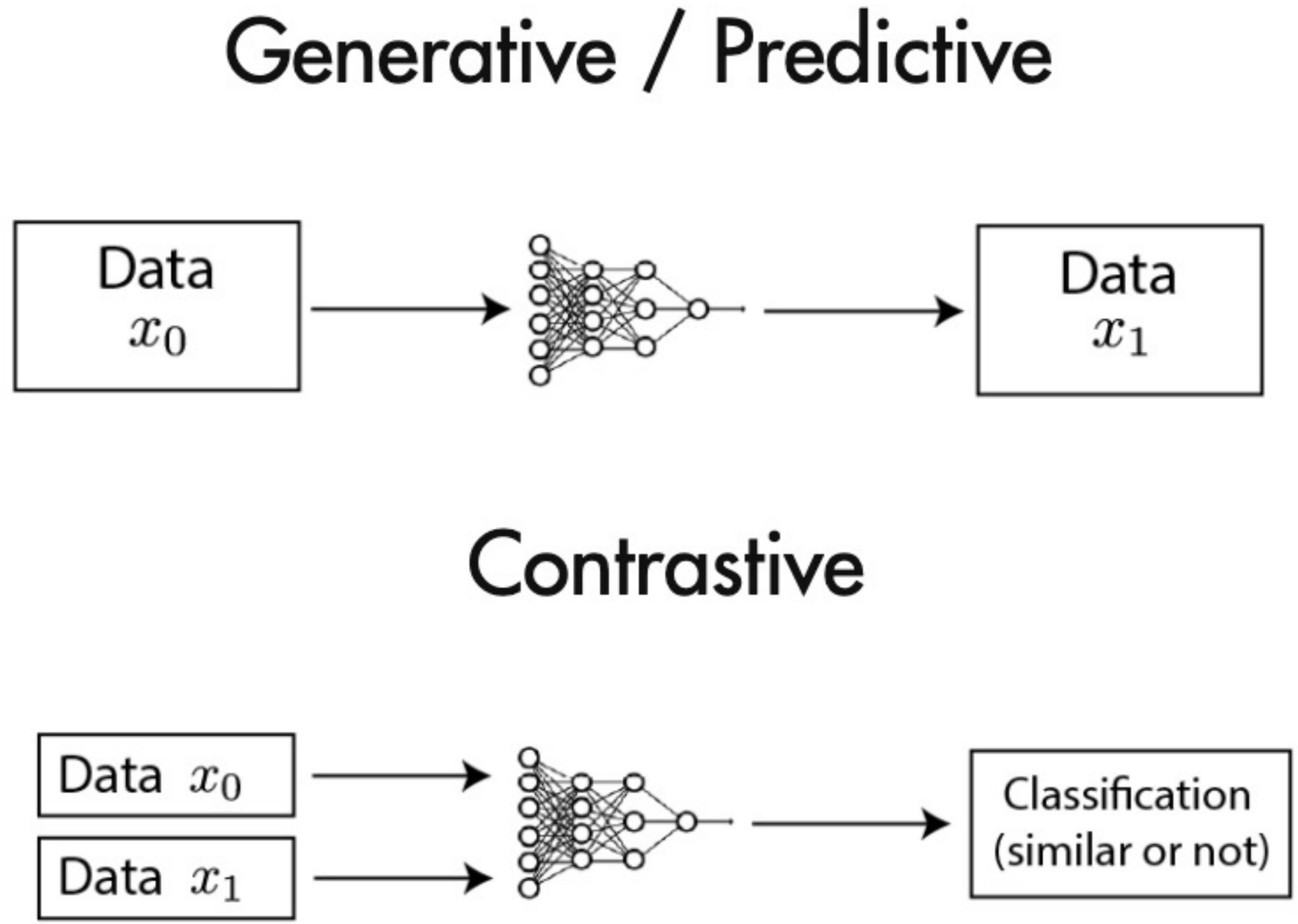


Image source: https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html

Contrastive Learning

- For any data point $\boldsymbol{x}$, which is is commonly referred to as an "anchor" data point, contrastive methods aim to learn a feature mapping f such that:

$$\text{score}\big(f(\boldsymbol{x}), f(\boldsymbol{x}^{+})\big) \gg \text{score}\big(f(\boldsymbol{x}), f(\boldsymbol{x}^{-})\big).$$

  - $\boldsymbol{x}^{+}$ is a data point similar to $\boldsymbol{x}$, referred to as a positive sample.
  - $\boldsymbol{x}^{-}$ is a data point dissimilar to $\boldsymbol{x}$, referred to as a negative sample.
  - the score function is a metric that measures the similarity between two features.

Contrastive Learning

- To optimize for this property, we can construct a softmax classifier that classifies positive and negative samples correctly:

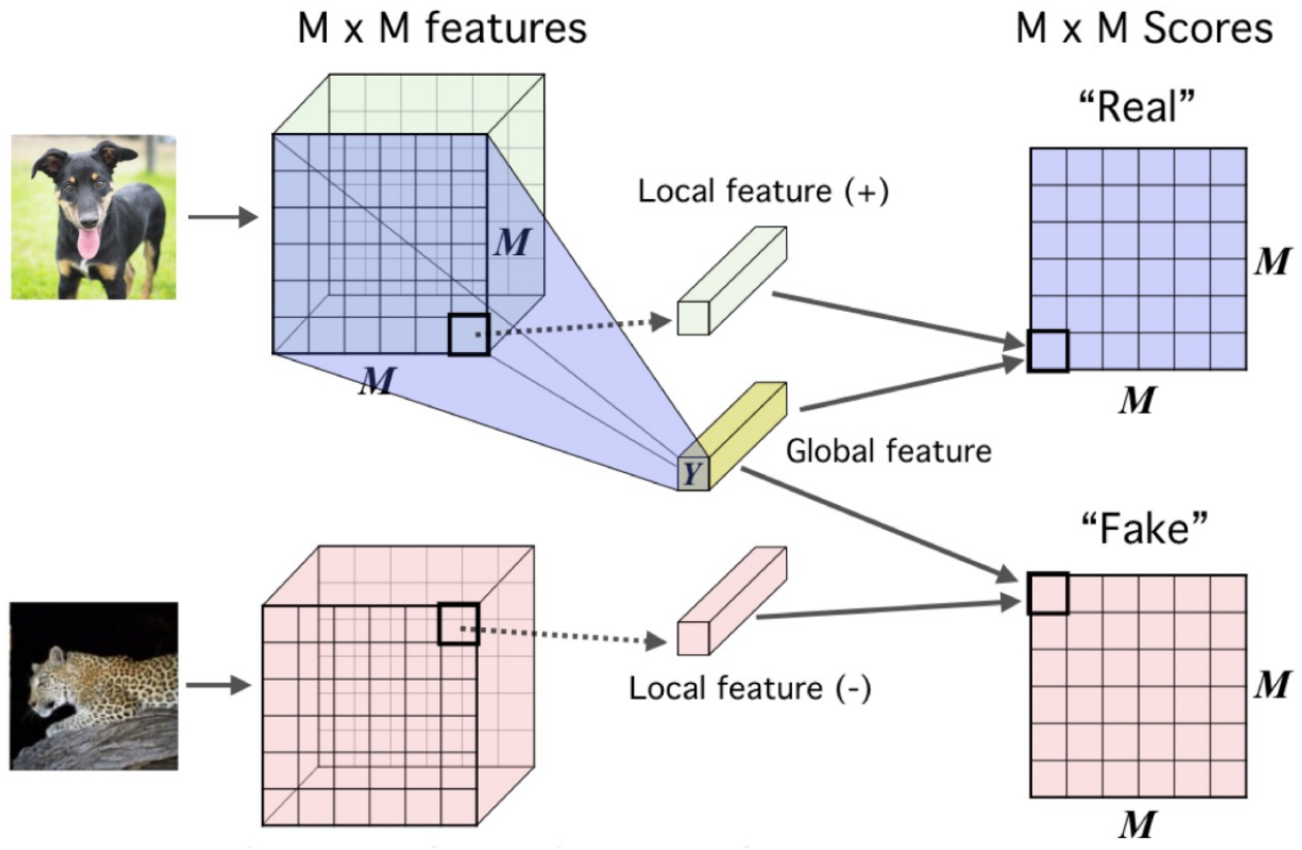
$$\mathcal{L} = -\mathbb{E}_X\left[\log \frac{\exp\left(f(\boldsymbol{x})^T f(\boldsymbol{x}^+)\right)}{\exp\left(f(\boldsymbol{x})^T f(\boldsymbol{x}^+)\right) + \sum_{j=1}^{N-1} \exp\left(f(\boldsymbol{x})^T f(\boldsymbol{x}_j)\right)}\right]$$

- It is commonly called the InfoNCE loss in the contrastive learning literature.
- But the key problem is:

How do we know data similarity?

Deep Infomax

Learning deep representations by mutual information estimation and maximization
RD Hjelm, A Fedorov, S Lavoie-Marchildon… - arXiv preprint arXiv …, 2018 - arxiv.org
In this work, we perform unsupervised learning of representations by maximizing mutual information between an input and the output of a deep neural network encoder. Importantly, we show that structure matters: incorporating knowledge about locality of the input to the objective can greatly influence a representation's suitability for downstream tasks. We further control characteristics of the representation by matching to a prior distribution adversarially. Our method, which we call Deep InfoMax (DIM), outperforms a number of popular …
Cited by 483 Related articles All 4 versions



Classify whether a pair of global features and local features are from the same image or not.

Image source: Hjelm, R. Devon, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. "Learning deep representations by mutual information estimation and maximization." arXiv preprint arXiv:1808.06670 (2018).

Deep Infomax

- Global features $E_\psi(X)$ are the final output of a convolutional encoder.
- Local features $C_\psi^{(i)}(X)$ are the output of an intermediate layer in the encoder (an $M \times M$ feature map).
  - Each local feature map has a limited receptive field.
- We want to maximize the mutual information between local and global features of the same image:

$$\underset{\omega,\psi}{\operatorname{argmax}} \frac{1}{M^2} \sum_{i=1}^{M^2} I_{\omega,\psi}\left(C_\psi^{(i)}(X); E_\psi(X)\right)$$

and minimize it for different image.

Contrastive Predictive Coding

Representation learning with **contrastive** predictive coding
A Oord, Y Li, O Vinyals - arXiv preprint arXiv:1807.03748, 2018 - arxiv.org
… **learning** approach to extract useful **representations** from high-dimensional data, which we call **Contrastive** … The key insight of our model is to **learn** such **representations** by predicting the …
☆ Save 99 Cite Cited by 6707 Related articles All 4 versions

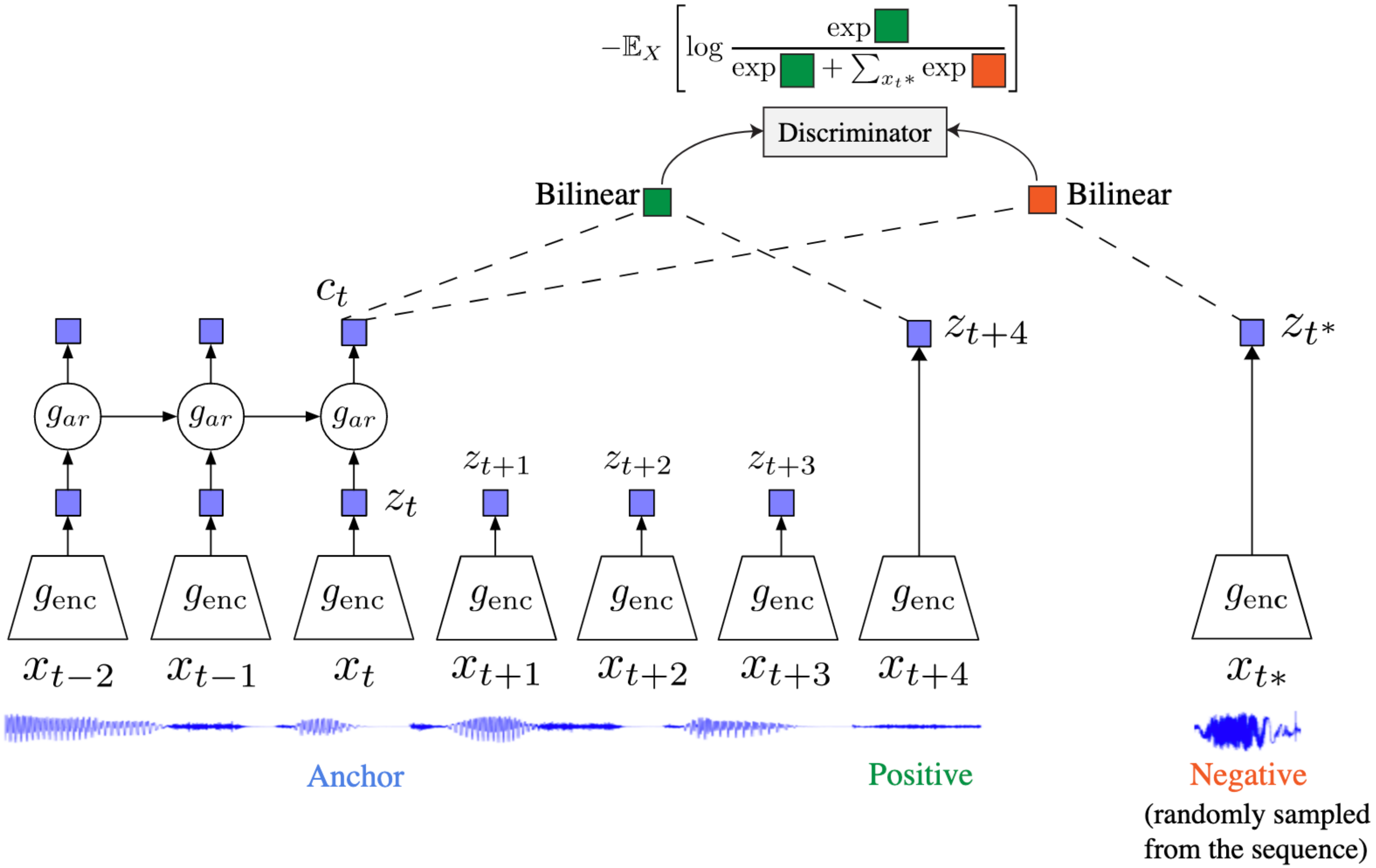
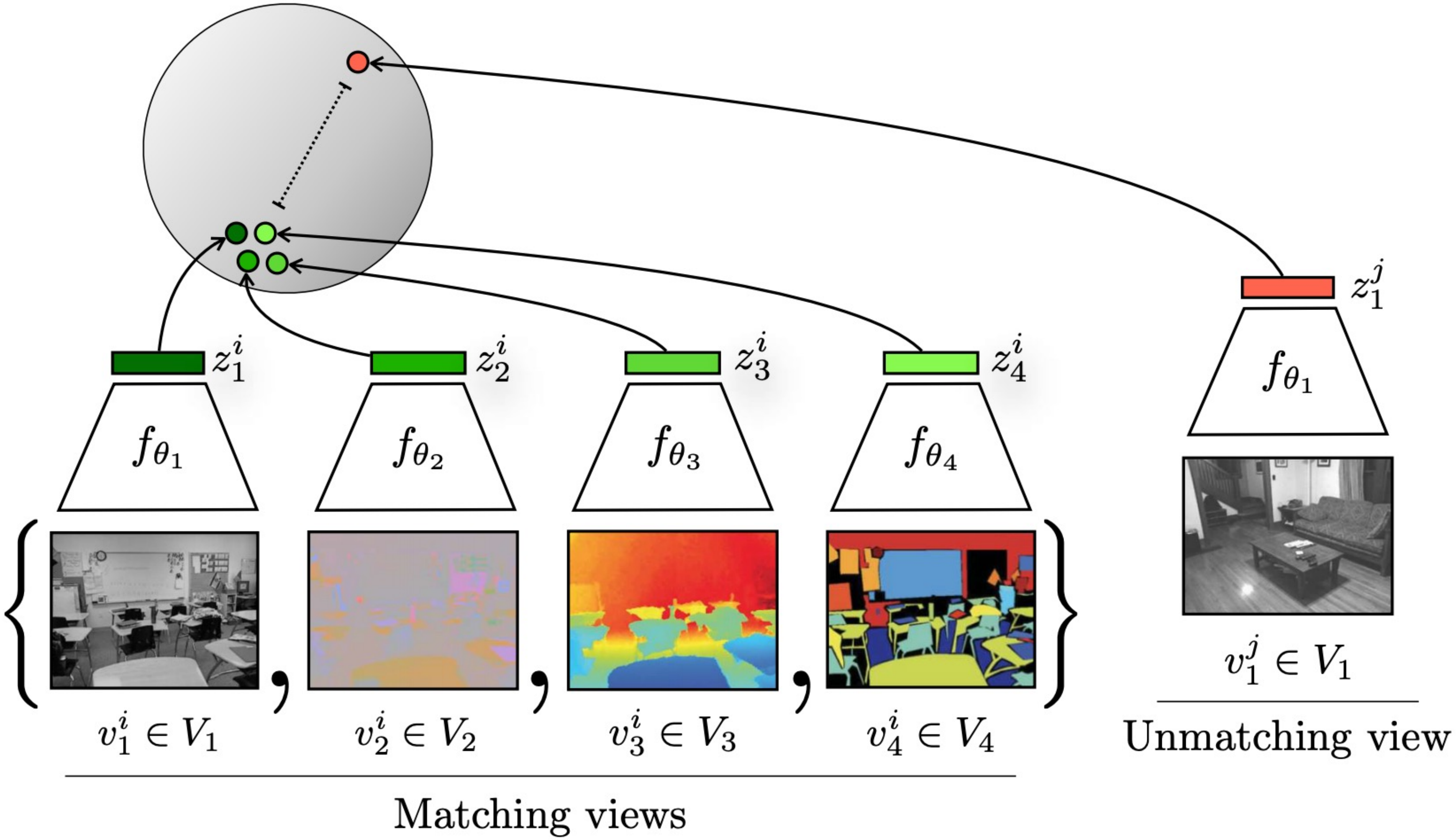


Image source: https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html

Contrastive Multiview Coding

Contrastive multiview coding
Y Tian, D Krishnan, P Isola - Computer Vision–ECCV 2020: 16th European …, 2020 - Springer
… Finally, we compare the **contrastive** formulation of **multiview** learning to the recently … **contrastive** approach learns stronger representations. The core ideas that we build on: **contrastive** …



Graph Contrastive Learning

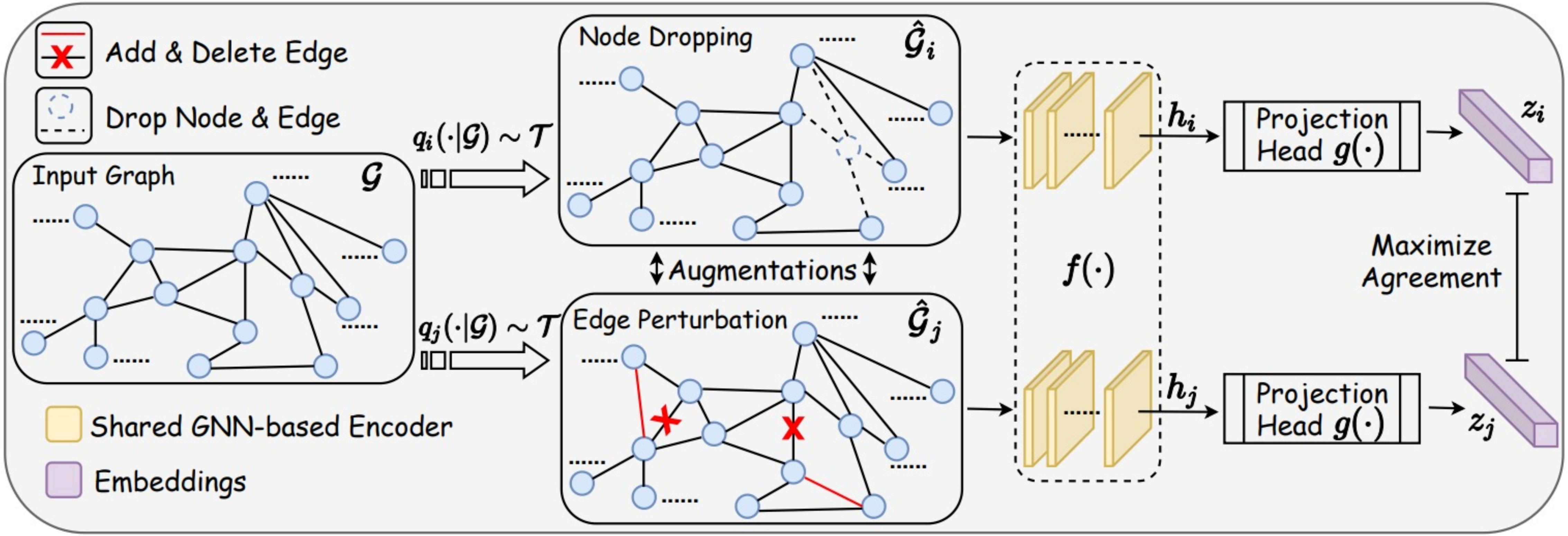
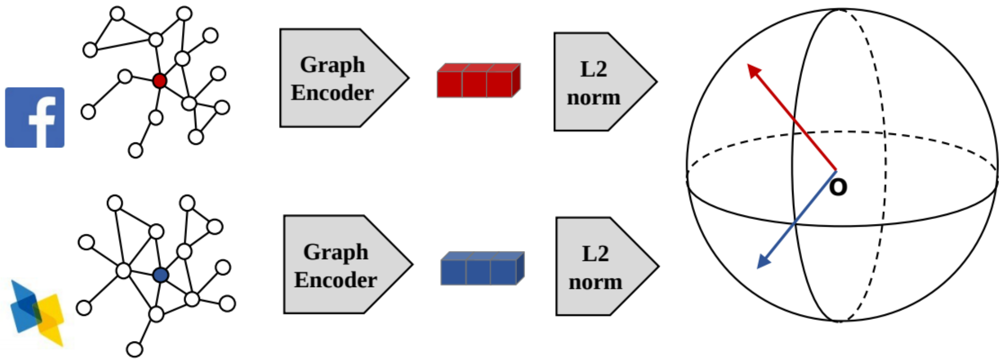


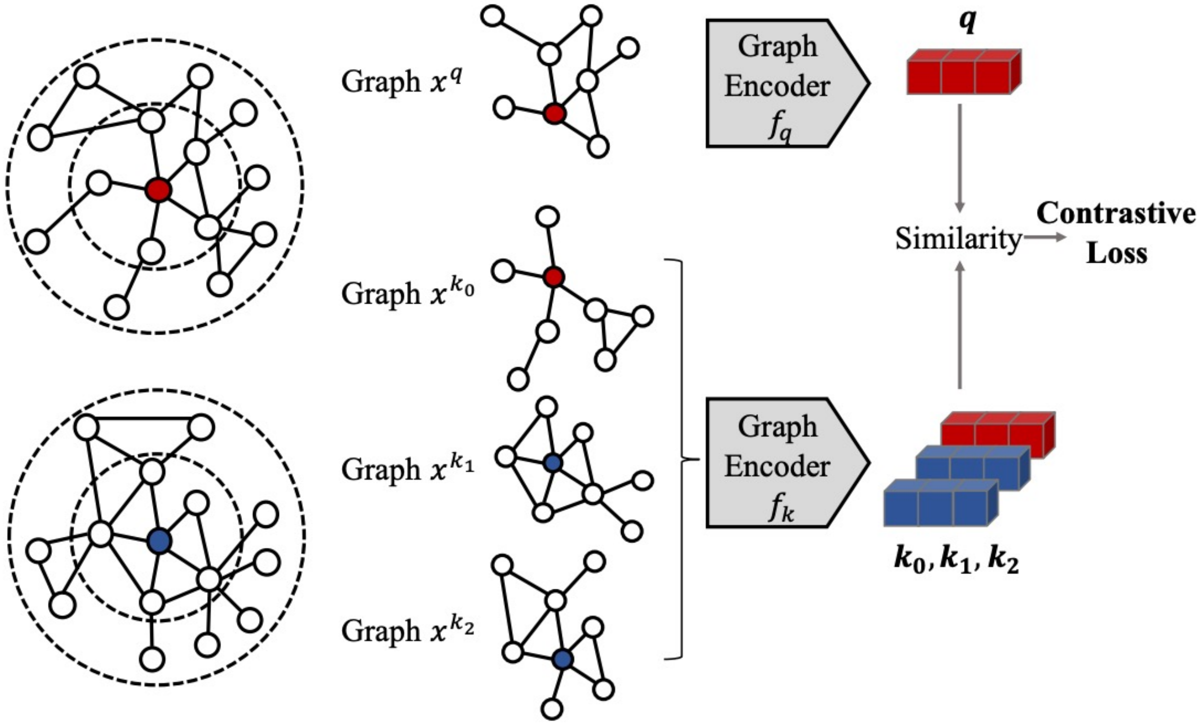
Image source: You, Yuning, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. "Graph Contrastive Learning with Augmentations." Advances in Neural Information Processing Systems 33 (2020).

Graph Contrastive Coding

Facebook social graph



DBLP co-authorship graph



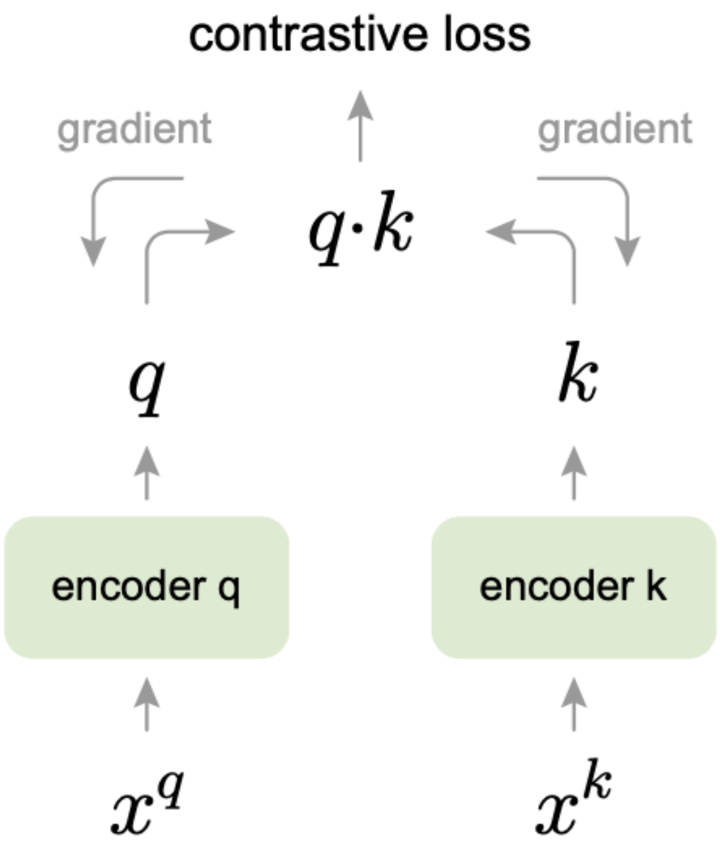
Capture the universal network topological properties across multiple networks

Image source: Qiu, Jiezhong, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. "Gcc: Graph contrastive coding for graph neural network pre-training." In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1150-1160. 2020.

MoCo

Momentum contrast for unsupervised visual representation **learning**
K He, H Fan, Y Wu, S Xie… - Proceedings of the IEEE …, 2020 - openaccess.thecvf.com
… From a perspective on **contrastive learning** [29] as dictionary look-up, we build a dynamic … on-the-fly that facilitates **contrastive** unsupervised **learning**. MoCo provides competitive results …
☆ Save 99 Cite Cited by 9482 Related articles All 19 versions

- A general framework for contrastive learning optimization.
- Motivation: number of negative samples should be large to make contrast.
  - However, the size is usually limited by batch size and GPU memory size.
- Idea: Reuse the representations of negative samples.



(a) end-to-end

Image source: He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729-9738. 2020.

MoCo

- Contrastive learning can be thought of as training an encoder for a dictionary look-up task.
- Query is the anchor sample. N keys contains 1 positive sample and $N-1$ negative samples.
- q and $k_0, k_1, \ldots$ are encoded samples. InfoNCE is calculated on them.

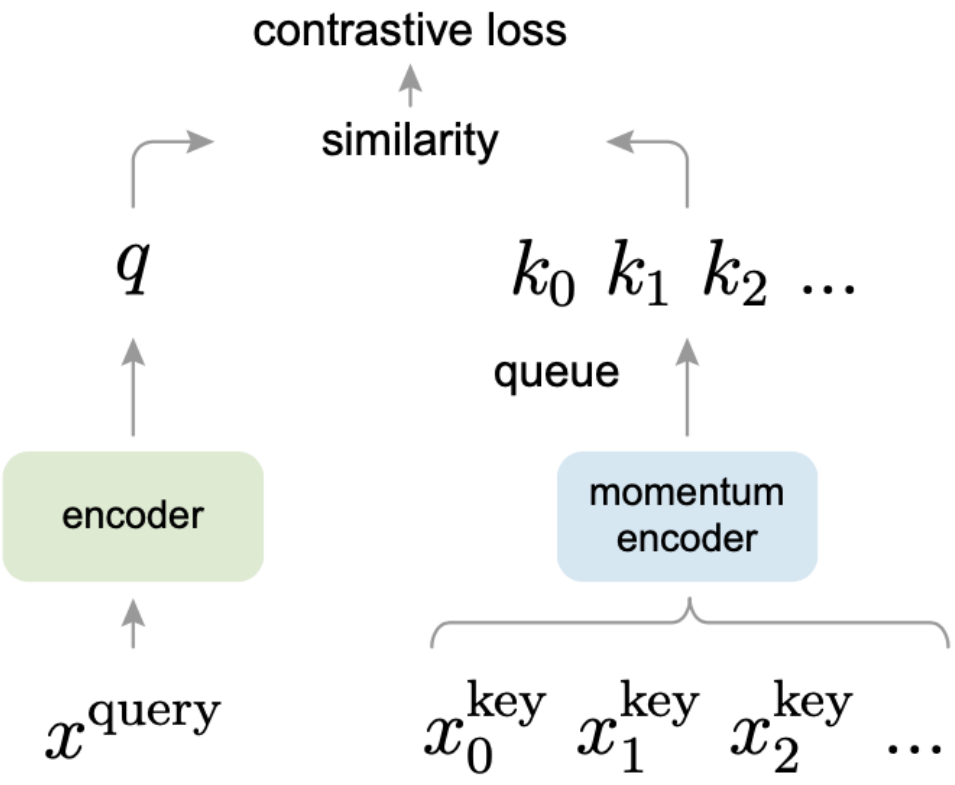


Image source: He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729-9738. 2020.

MoCo

- Use a queue to store encoded negative samples for reuse.
- The queue is dynamic updated during training.
- Momentum update is adopted to slow down the frequency of key encoder:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$$

  - m is usually set very close to 1 (e.g. 0.999).

MoCo

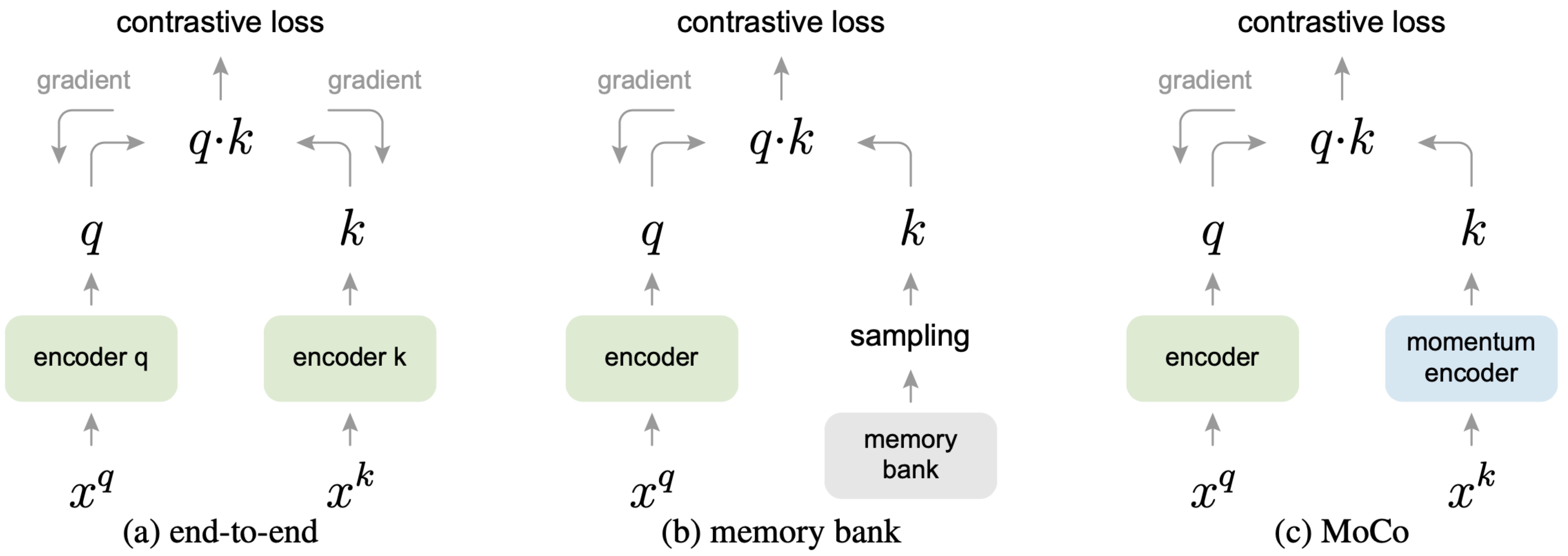


Image source: He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729-9738. 2020.

Algorithm 1 Pseudocode of MoCo in a PyTorch-like style.

```
# f_q, f_k: encoder networks for query and key
# queue: dictionary as a queue of K keys (CxK)
# m: momentum
# t: temperature

f_k.params = f_q.params # initialize
for x in loader: # load a minibatch x with N samples
    x_q = aug(x) # a randomly augmented version
    x_k = aug(x) # another randomly augmented version

    q = f_q.forward(x_q) # queries: NxC
    k = f_k.forward(x_k) # keys: NxC
    k = k.detach() # no gradient to keys

    # positive logits: Nx1
    l_pos = bmm(q.view(N,1,C), k.view(N,C,1))

    # negative logits: NxK
    l_neg = mm(q.view(N,C), queue.view(C,K))

    # logits: Nx(1+K)
    logits = cat([l_pos, l_neg], dim=1)

    # contrastive loss, Eqn.(1)
    labels = zeros(N) # positives are the 0-th
    loss = CrossEntropyLoss(logits/t, labels)

    # SGD update: query network
    loss.backward()
    update(f_q.params)

    # momentum update: key network
    f_k.params = m*f_k.params+(1-m)*f_q.params

    # update dictionary
    enqueue(queue, k) # enqueue the current minibatch
    dequeue(queue) # dequeue the earliest minibatch
```

No parameter update for keys

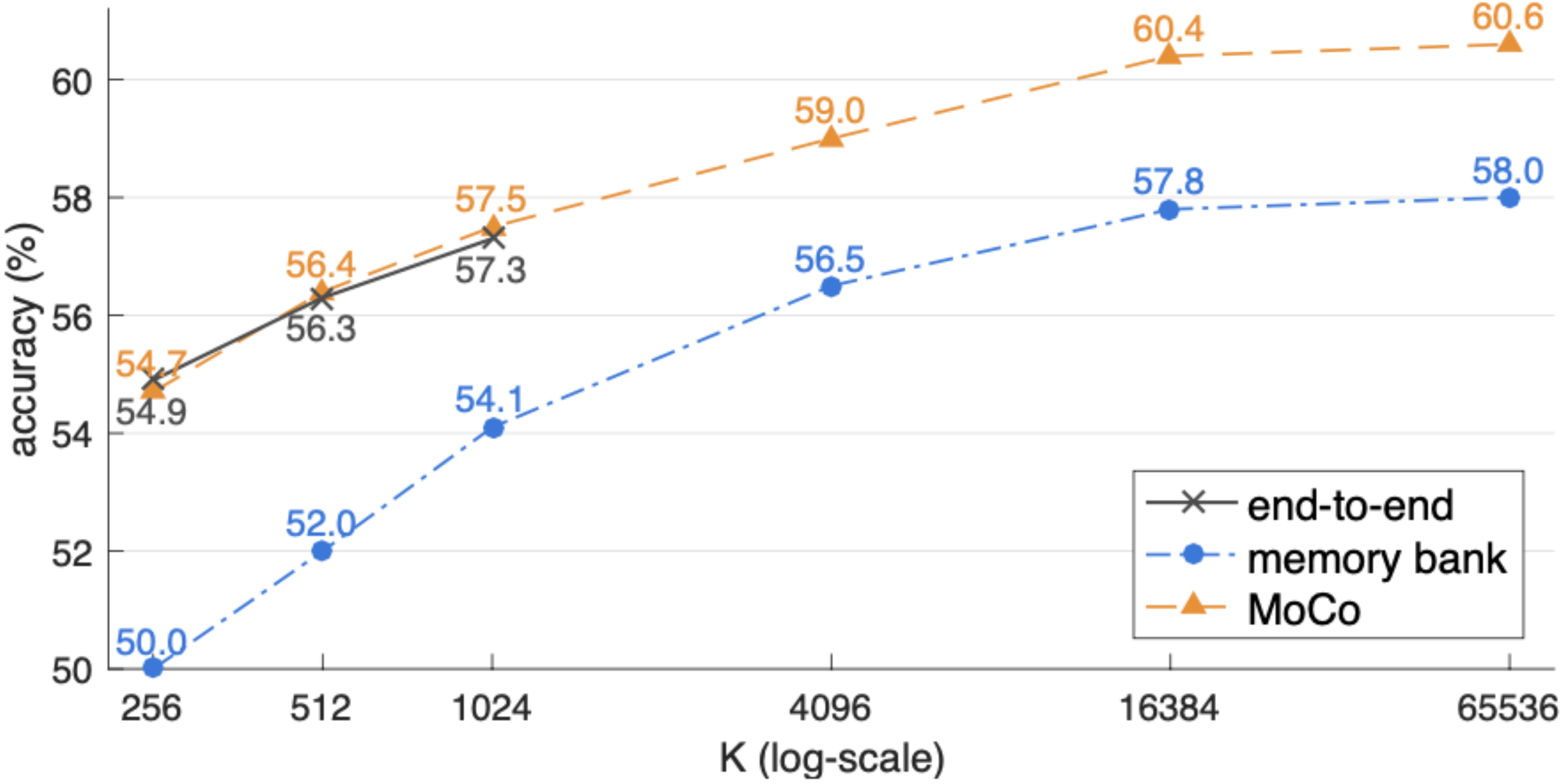
Use all negative samples in the queue

Momentum update

Enqueue negative samples

Image source: He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729-9738. 2020.

MoCo



Comparison of three contrastive loss mechanisms under the ImageNet linear classification protocol

Image source: He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729-9738. 2020.

SimCLR

A simple framework for contrastive learning of visual representations
T Chen, S Kornblith, M Norouzi… - … on machine learning, 2020 - proceedings.mlr.press
… proposed **contrastive** selfsupervised **learning** algorithms … the **contrastive** prediction tasks to **learn** useful representations, we systematically **study** the major components of our **framework**. …
☆ Save 99 Cite Cited by 13611 Related articles All 21 versions

Contributions:

- Use data augmentations.
- Introduce a learnable nonlinear transformation between the representation and the contrastive loss.
- Contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning.
  - Batch size 8192 with 128 TPU v3 cores...

SimCLR

SimCLR

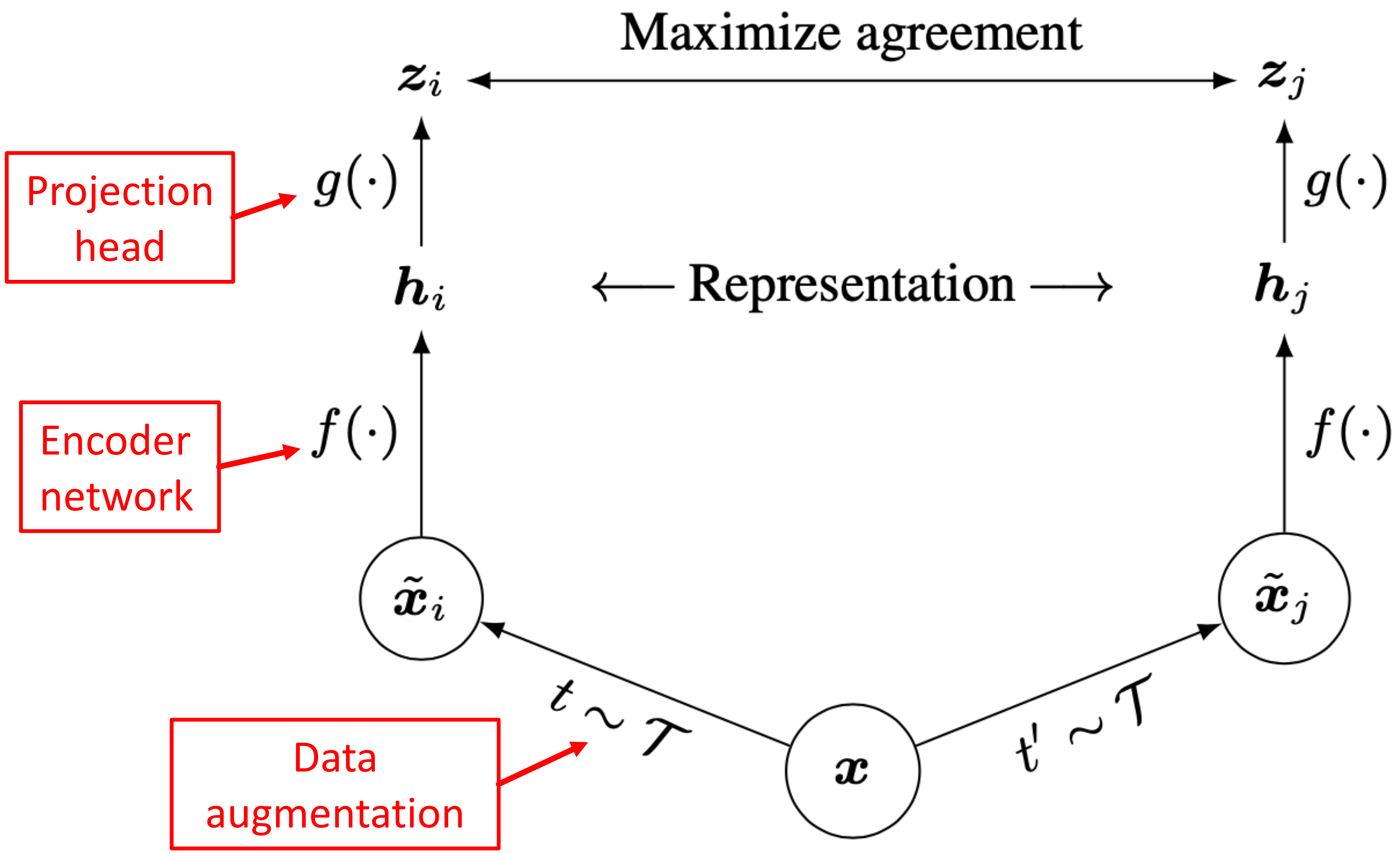


Image source: Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." arXiv preprint arXiv:2002.05709 (2020).

SimCLR

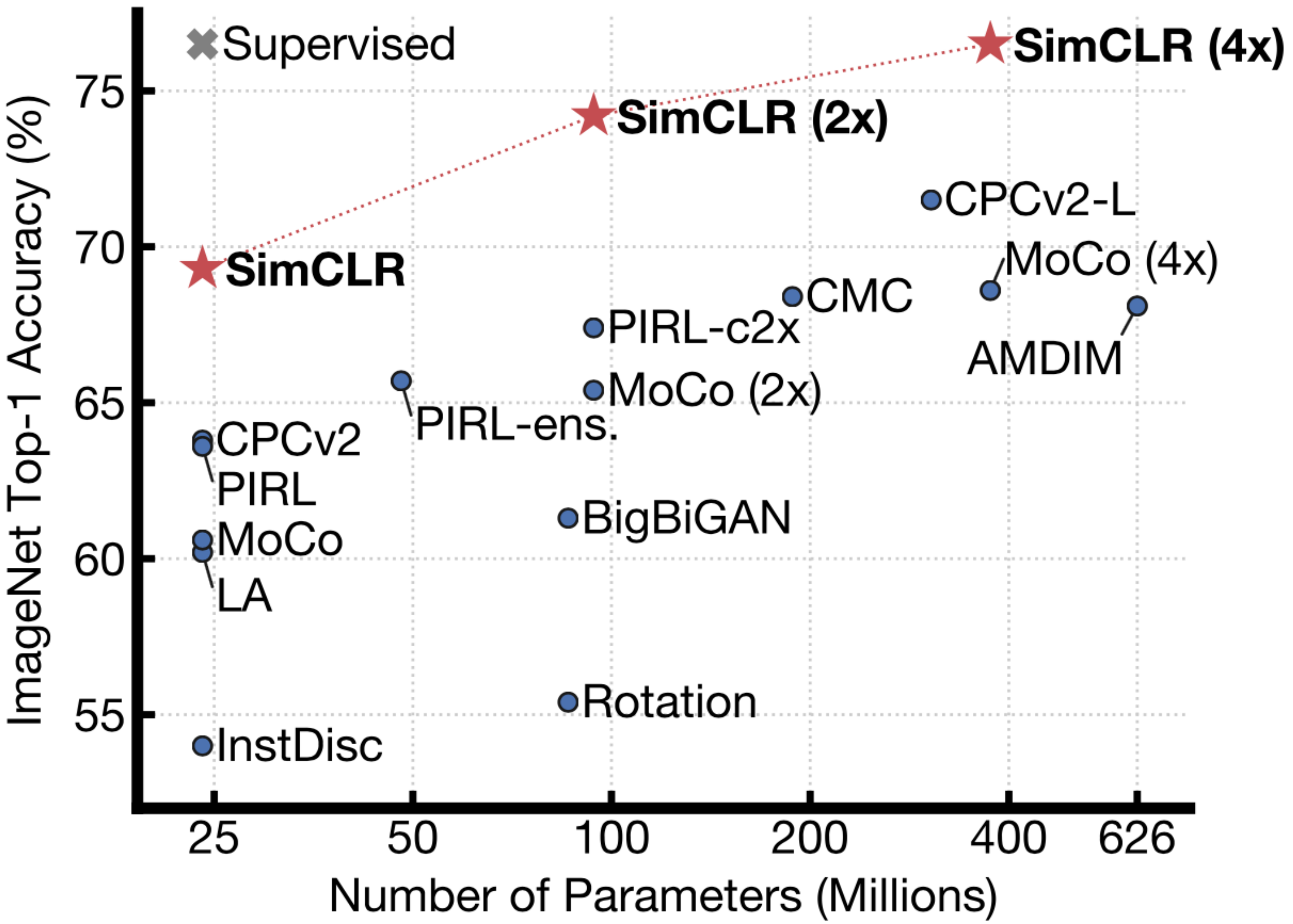


Image source: Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." arXiv preprint arXiv:2002.05709 (2020).

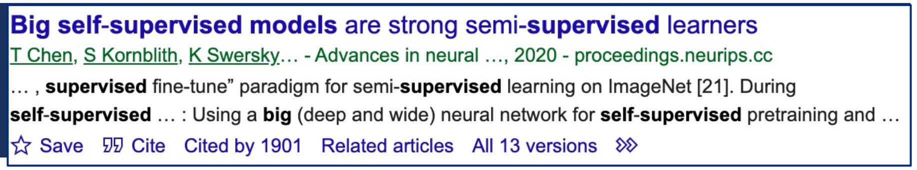
MoCo v2

Improved baselines with **momentum** contrastive learning
X Chen, H Fan, R Girshick, K He - arXiv preprint arXiv:2003.04297, 2020 - arxiv.org
… We report that two design **improvements** used in SimCLR, … " **baselines** can run on a typical 8-GPU machine and achieve better results than SimCLR. We hope these **improved baselines** …

This 2-page short paper declares:

- Two design improvements used in SimCLR, namely, an MLP projection head and stronger data augmentation, are orthogonal to the frameworks of MoCo and SimCLR, and when used with MoCo they lead to better image classification and object detection transfer learning results.
- In contrast to SimCLR's large 4k~8k batches, which require TPU support, our "MoCo v2" baselines can run on a typical 8-GPU machine and achieve better results than SimCLR.

SimCLR v2



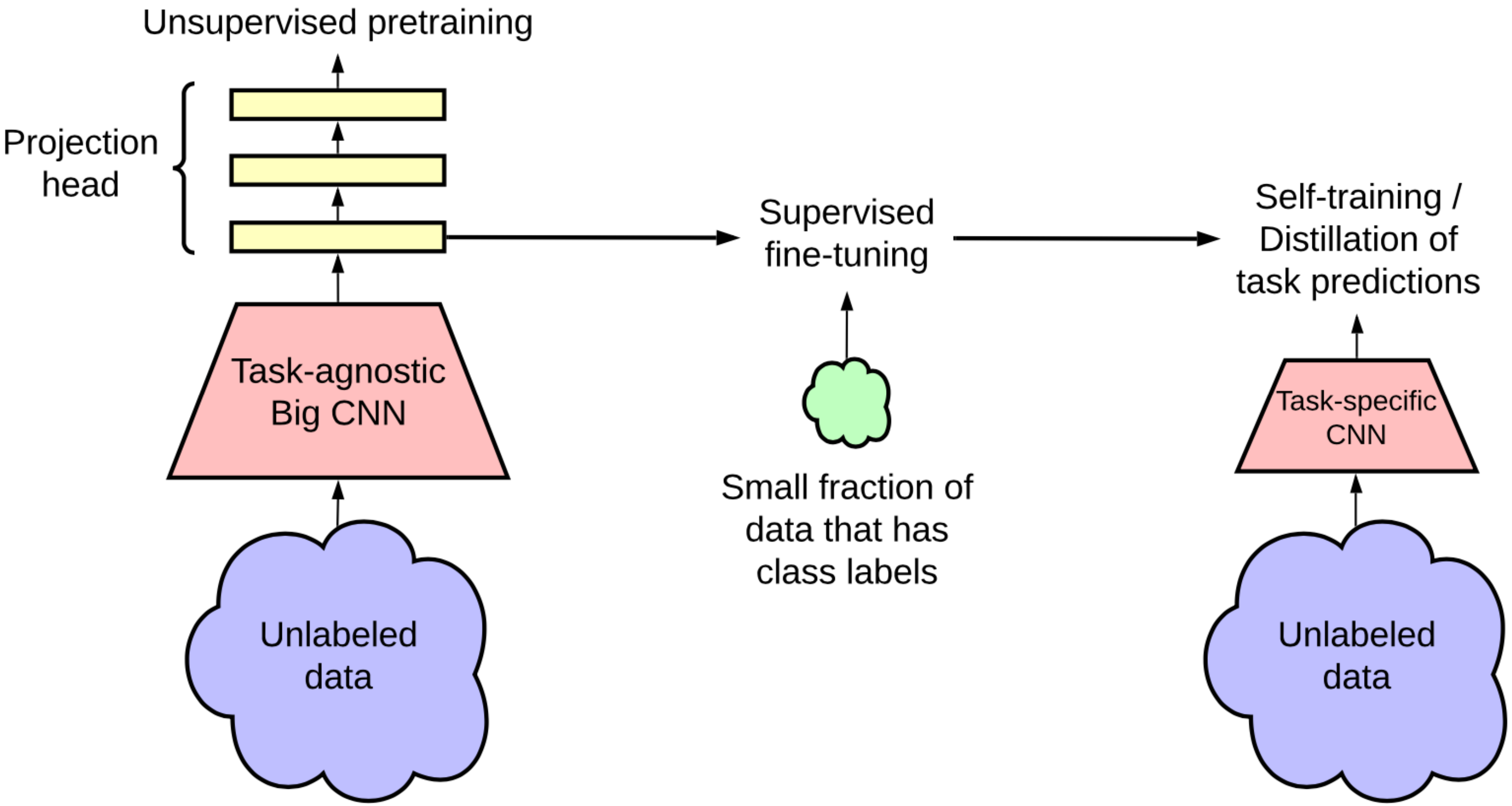
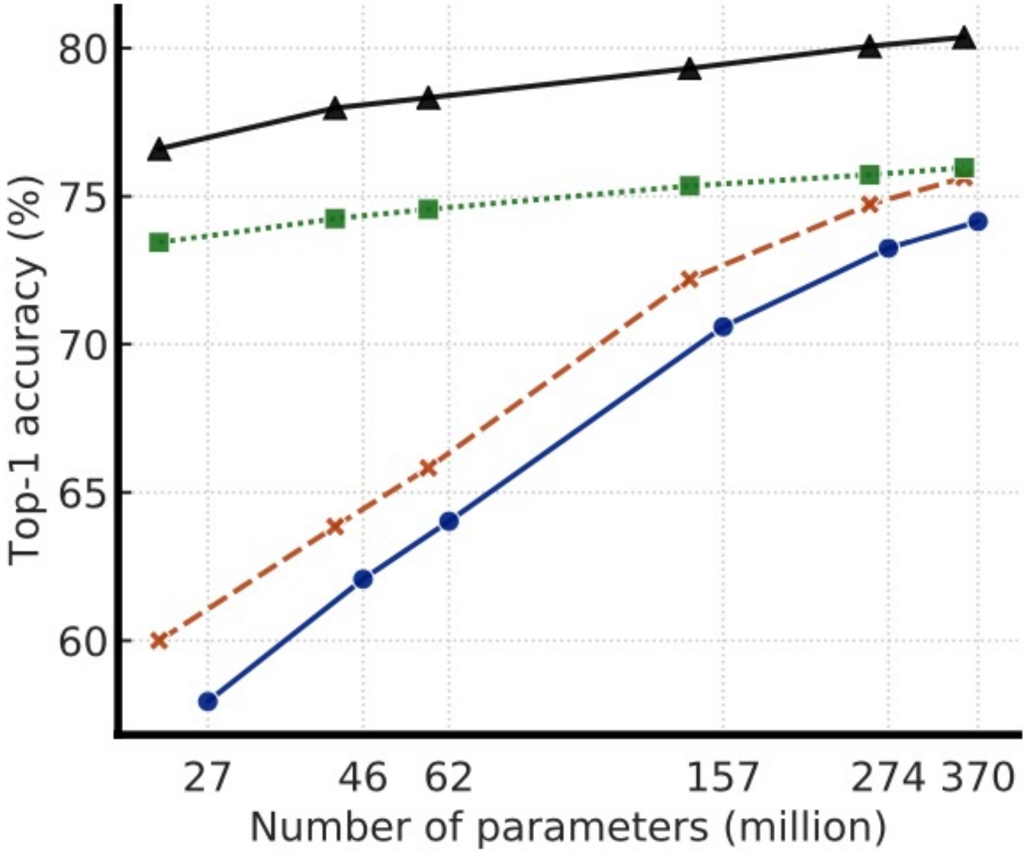
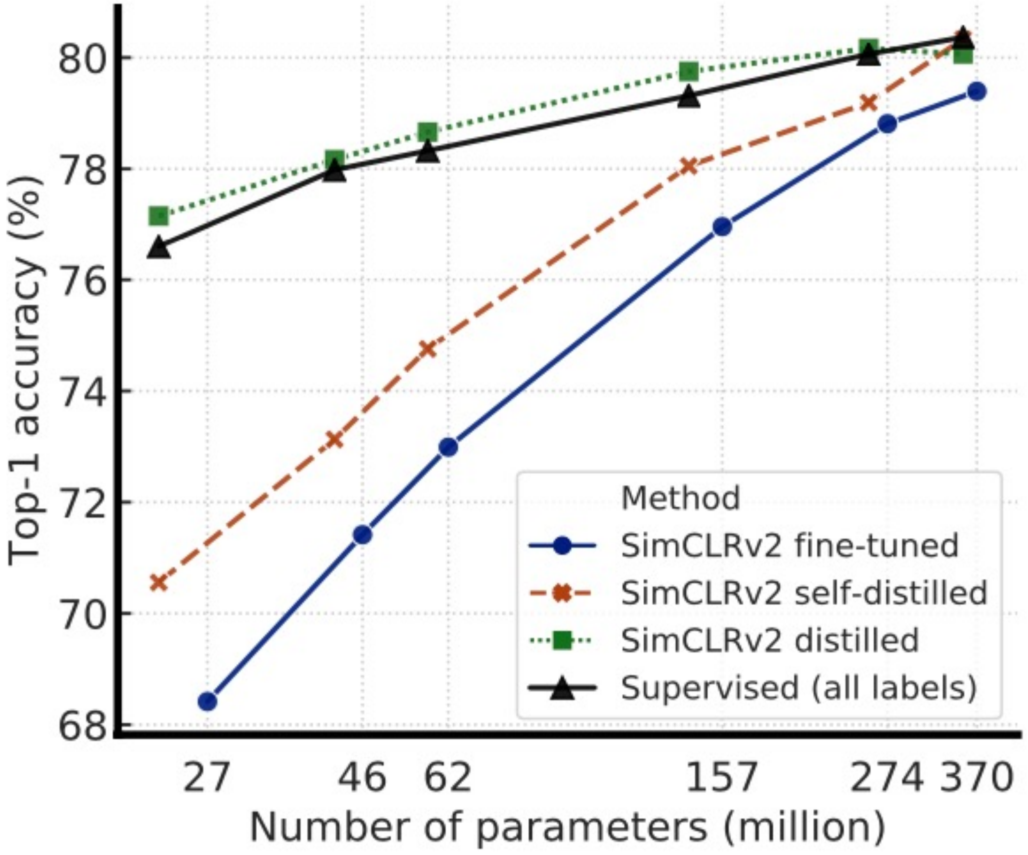


Image source: Chen, Ting, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. "Big self-supervised models are strong semi-supervised learners." Advances in Neural Information Processing Systems 33 (2020).

SimCLR v2



(a) Label fraction 1%

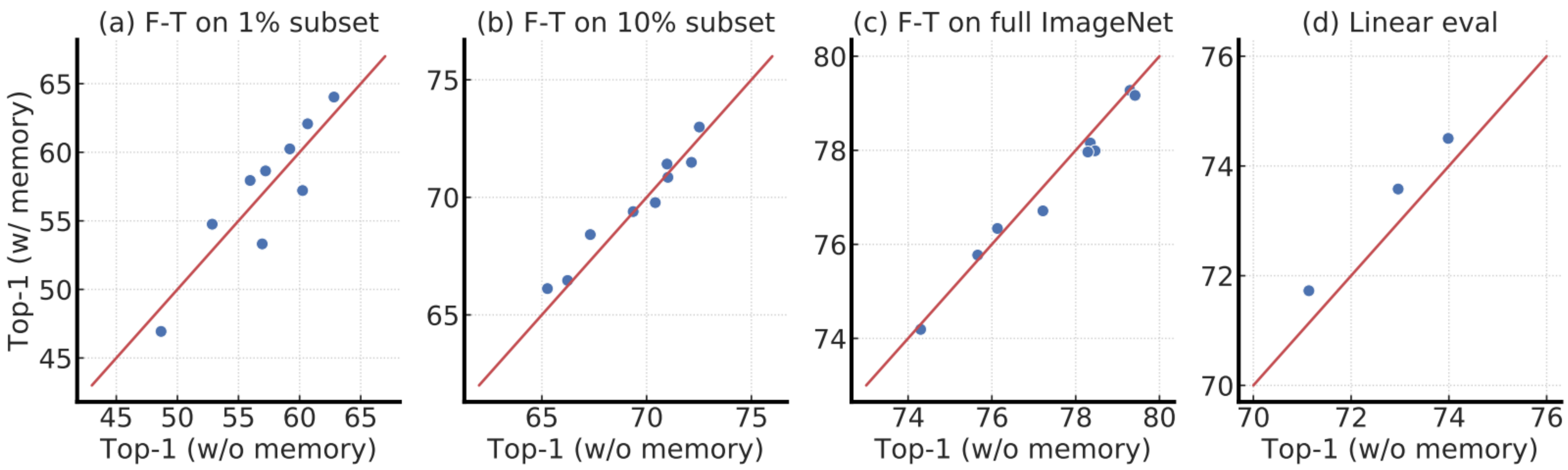


(b) Label fraction 10%

Image source: Chen, Ting, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. "Big self-supervised models are strong semi-supervised learners." Advances in Neural Information Processing Systems 33 (2020).

SimCLR v2

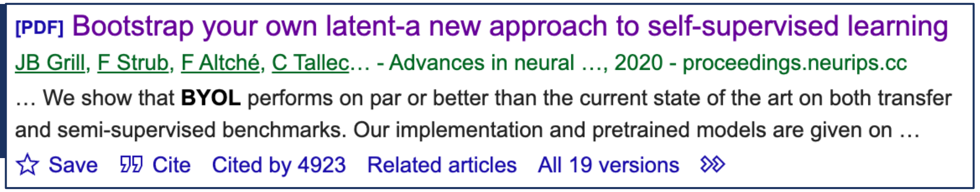
- “Memory provides modest advantages in terms of linear evaluation and fine-tuning with 1% of the labels; the improvement is around 1%.”
- “We believe the reason that memory only provides marginal improvement is that we already use a big batch size (i.e. 4096).”



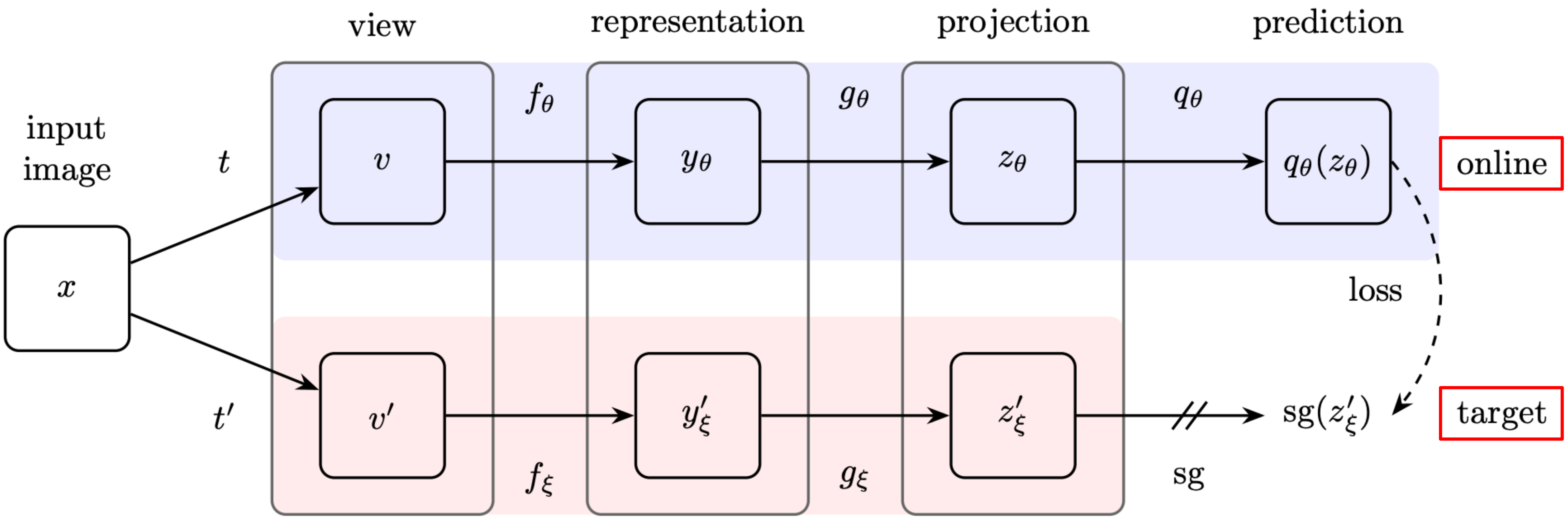
Top-1 results of ResNet-50, ResNet-101, and ResNet-152 trained with or without memory.

Image source: Chen, Ting, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. "Big self-supervised models are strong semi-supervised learners." Advances in Neural Information Processing Systems 33 (2020).

BYOL



- Are negative samples necessary for contrastive learning?



SimSiam

Exploring simple siamese representation learning
X Chen, K He - Proceedings of the IEEE/CVF conference on …, 2021 - openaccess.thecvf.com
… In this paper, we report surprising empirical results that **simple** Siamese networks can learn meaningful … We hope this **simple** baseline will motivate people to rethink the roles of Siamese …
Save Cite Cited by 2986 Related articles All 6 versions

- Surprising empirical results that simple Siamese networks can learn meaningful representations, even using none of the following:
  - negative sample pairs,
  - large batches,
  - momentum encoders.

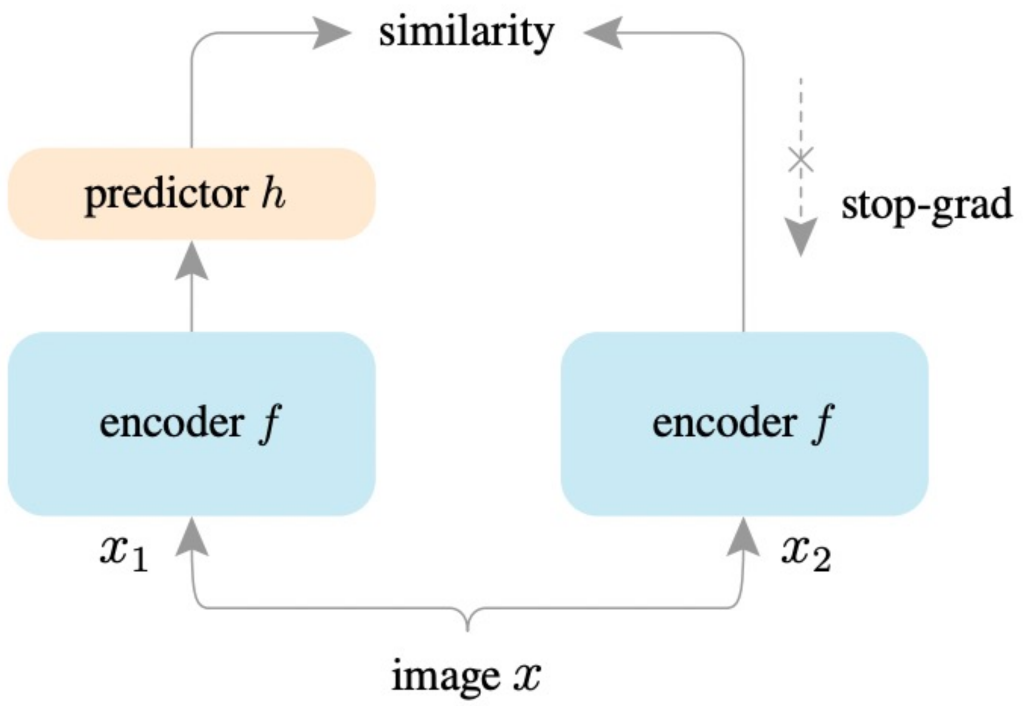


Image source: Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750-15758. 2021.

Collapse

- Collapse: All representations tend to be the same.

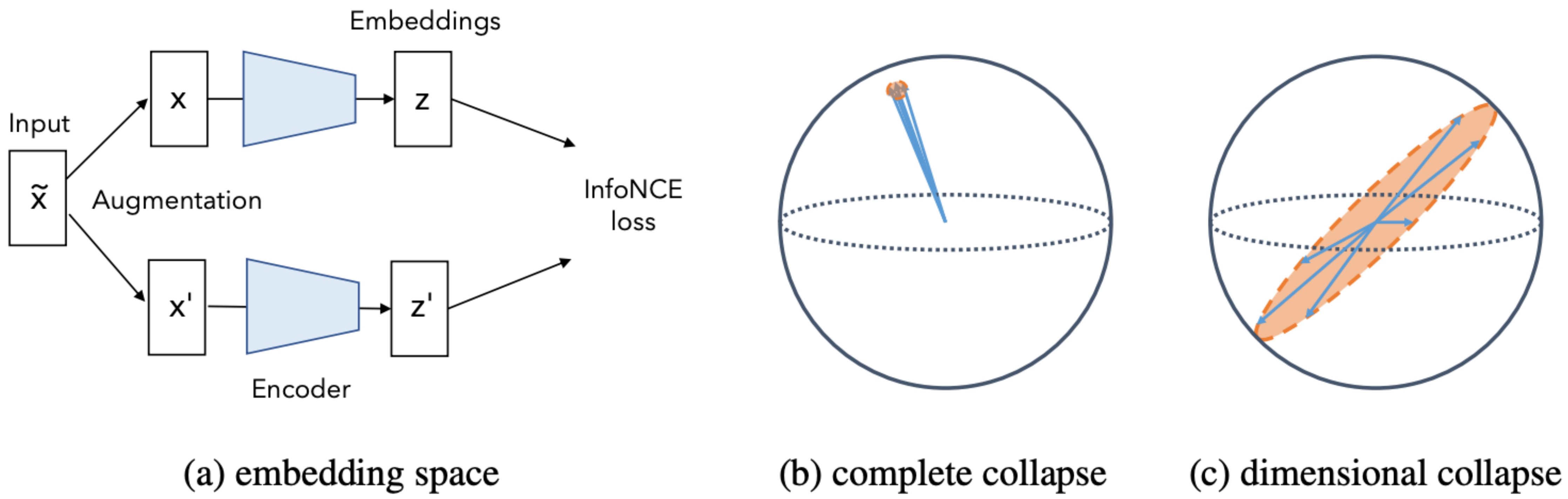


Image source: Jing, Li, Pascal Vincent, Yann LeCun, and Yuandong Tian. "Understanding dimensional collapse in contrastive self-supervised learning." arXiv preprint arXiv:2110.09348 (2021).

Downstream Tasks for Evaluation

- To compare different self-supervised learning methods, there are some commonly used downstream tasks for evaluation.

CV:

- Semantic segmentation
- Object detection
- Image classification
- Huma action recognition
- ...

NLP:

- Question answering
- Named entity recognition
- Sentiment classification
- Natural language inference
- ...

Conclusion

After this lecture, you should know:

- What is the difference between supervised and self-supervised learning.
- What is pretext task and pseudo label?
- How can we generate pseudo label?
- What is contrastive learning?

Suggested Reading

- Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey
- Self-supervised Learning: Generative or Contrastive
- Awesome Self-Supervised Learning
- Contrastive Self-Supervised Learning
- 对比学习(Contrastive Learning)相关进展梳理

Thank you!

- Any question?
- Don’t hesitate to send email to me for asking questions and discussion. ☺