Deep Learning

Self-supervised Learning

Technische Hochschule Rosenheim
Sommer 2023
Prof. Dr. Jochen Schmidt

Acknowledgements

Many of the slides presented here are based on the Deep Learning Slides Summer Semester 2020, courtesy of
A. Maier, V. Christlein, K. Breininger, F. Denzinger, F. Thamm,
Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nürnberg.
https://lme.tf.fau.de/

Supervised Learning

- We have seen impressive results achieved with…
  - large amounts of training data and
  - consistent, high-quality annotations.

Mask R-CNN image source [MAT19]

The Cost of Annotation

Image-level class labels: ~27 sec

Instance spotting: +14 sec

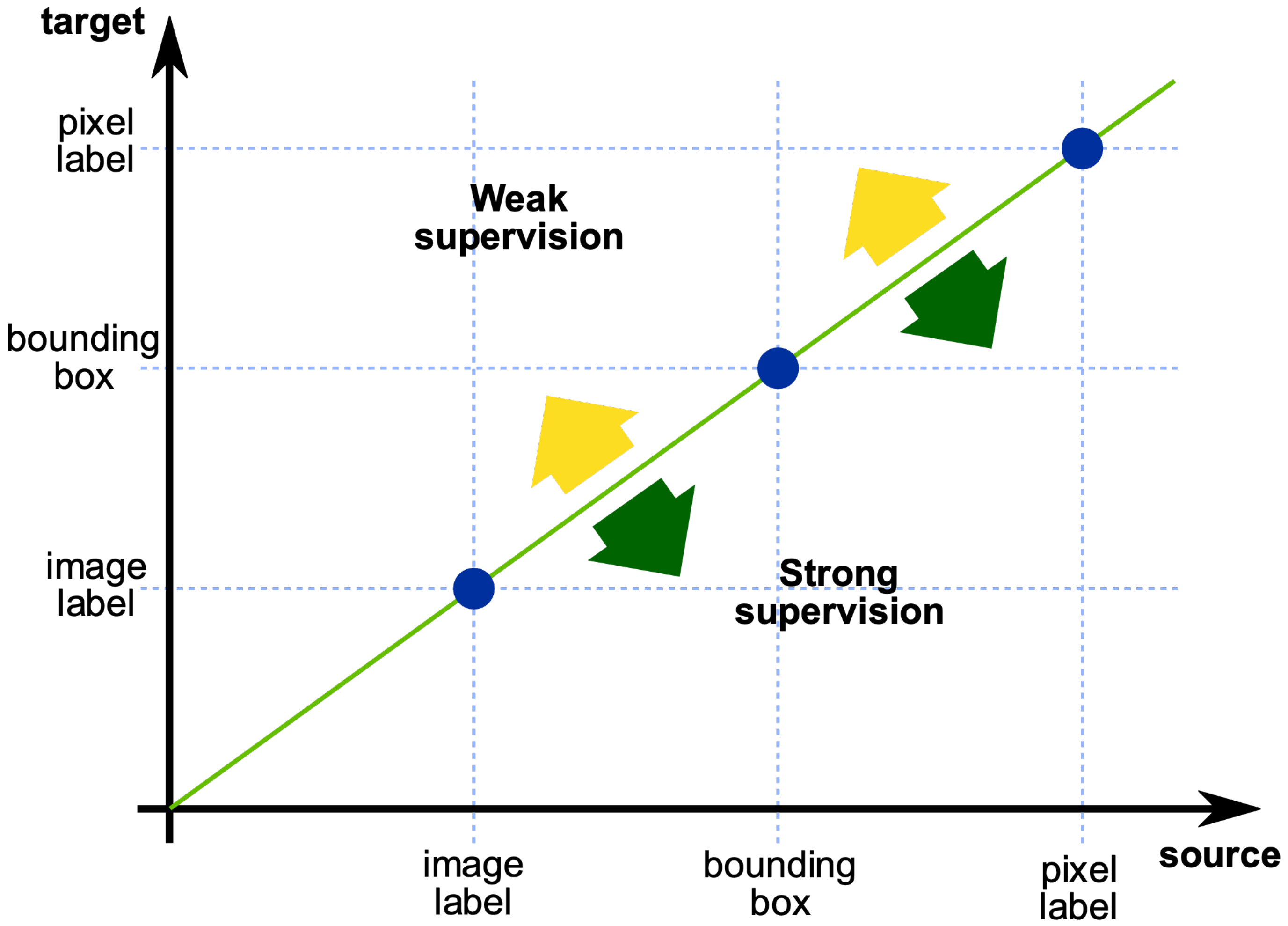
Instance Segmentation: +80 sec

Source: [Lin14]

Dense pixel-level annotations: 1.5h

Source: [Cor16]

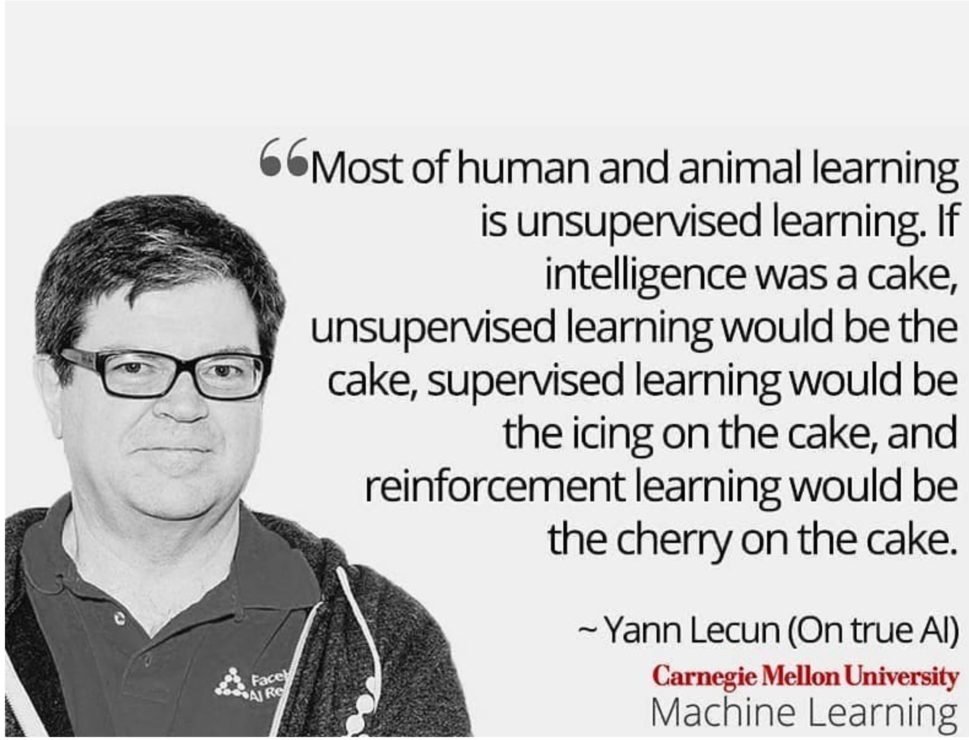
Strongly vs Weakly Supervised Learning



Reproduced from CVPR18 Tutorial: Weakly Supervised Learning for Computer Vision

Self-supervised Learning – Motivation

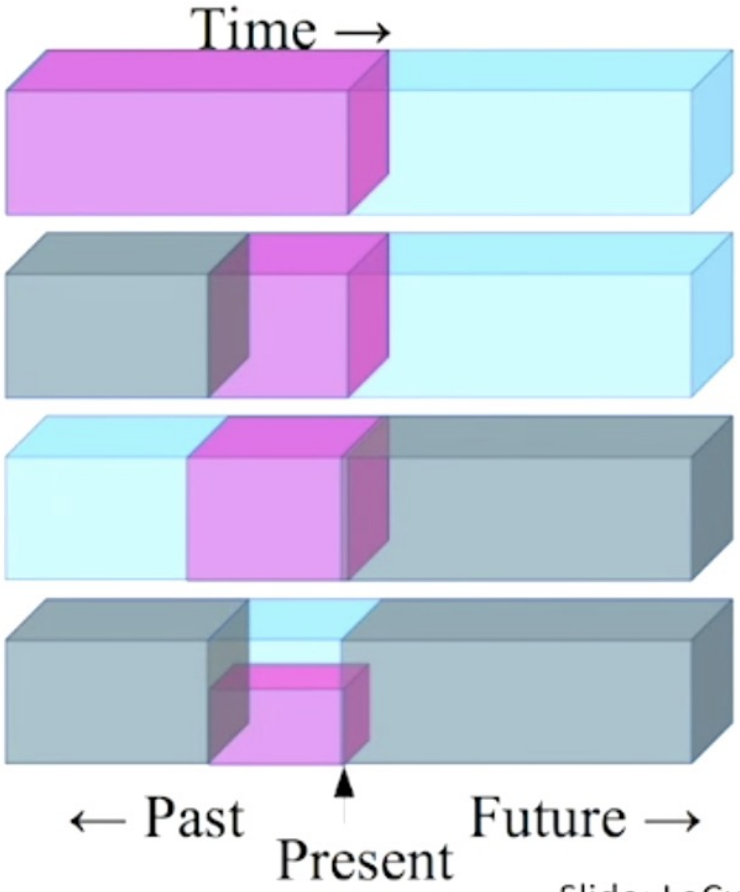
- Jitendra Malik: “Supervision is the opium of the AI researcher”
- Alyosha Efros: “The AI revolution will not be supervised”
- Yann LeCun:



Source: https://www.facebook.com/722677142/posts/10156036317282143/

Self-supervised Learning – Idea

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



Slide: LeCun

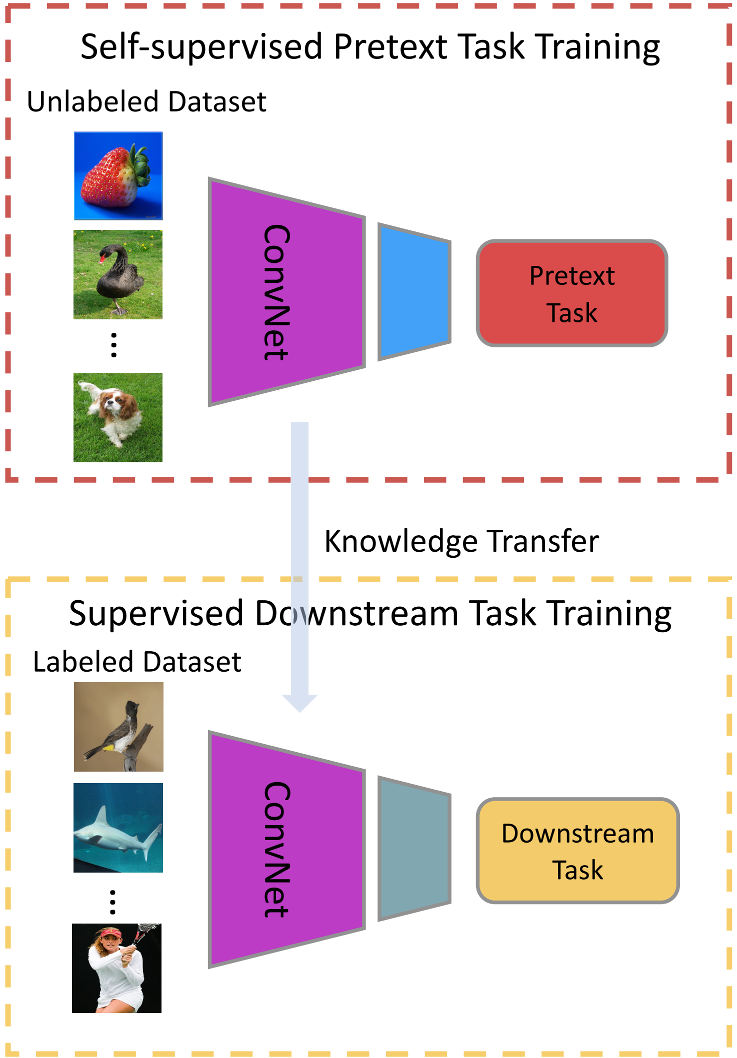
Source: https://www.youtube.com/watch?v=7I0Qt7GALVk

Self-supervised Learning – Definition

Yann LeCun
April 30, 2019 ·

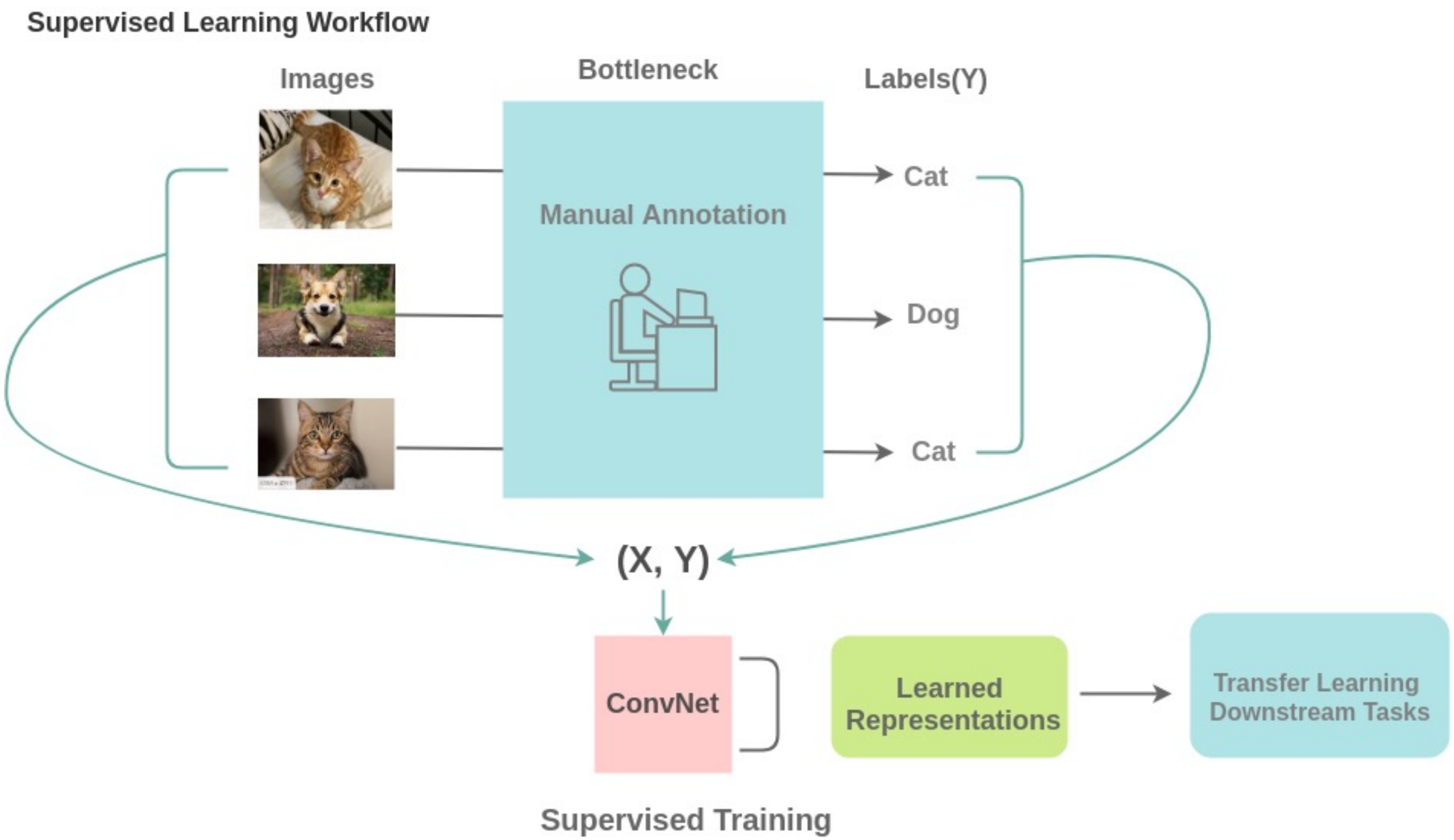
I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

- Subcategory of unsupervised learning
  - Use pretext/surrogate/pseudo tasks in a supervised fashion, i.e., we have
    - automatically generated labels
    - that can be used as a measure of correctness (for the loss function)
- Downstream task: retrieval, supervised or semi-supervised classification, etc.
- Note: Generative models (e.g., GANs) are also SSL methods



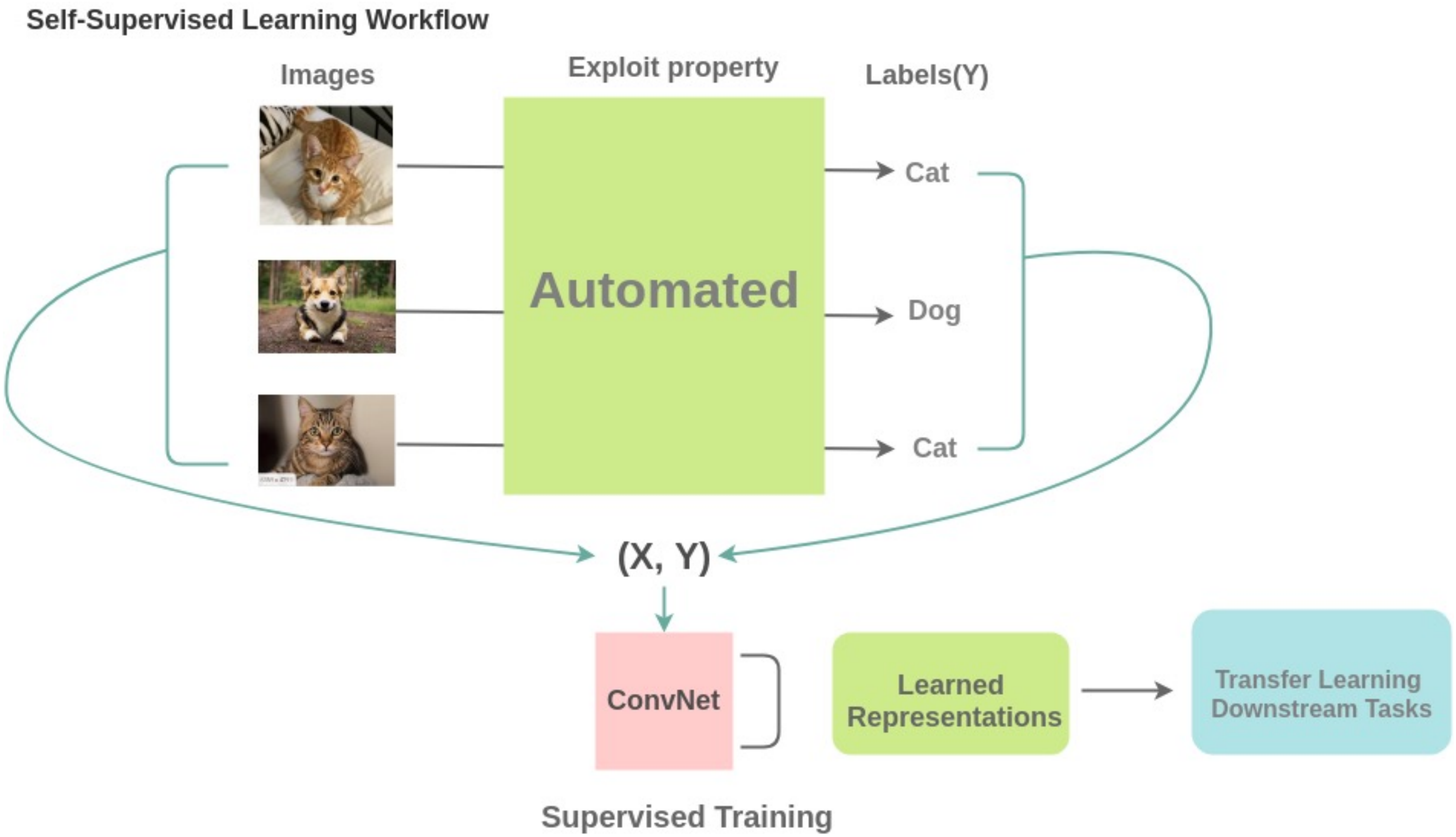
Source: https://www.facebook.com/722677142/posts/10155934004262143/

Advantages of Self-supervised Learning



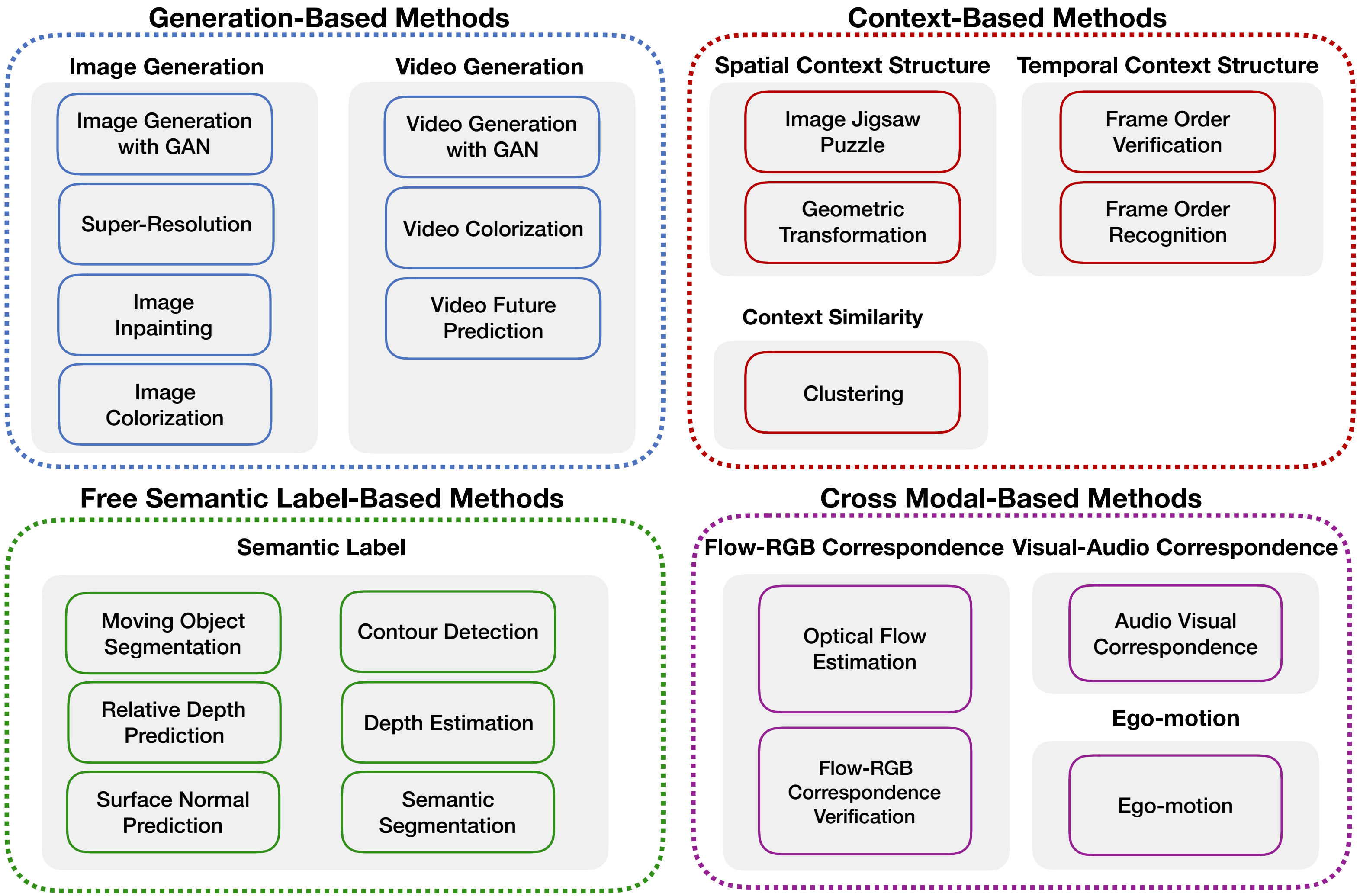
Source: https://amitness.com/2020/02/illustrated-self-supervised-learning/

Advantages of Self-supervised Learning



Source: https://amitness.com/2020/02/illustrated-self-supervised-learning/

Pretext Tasks Overview

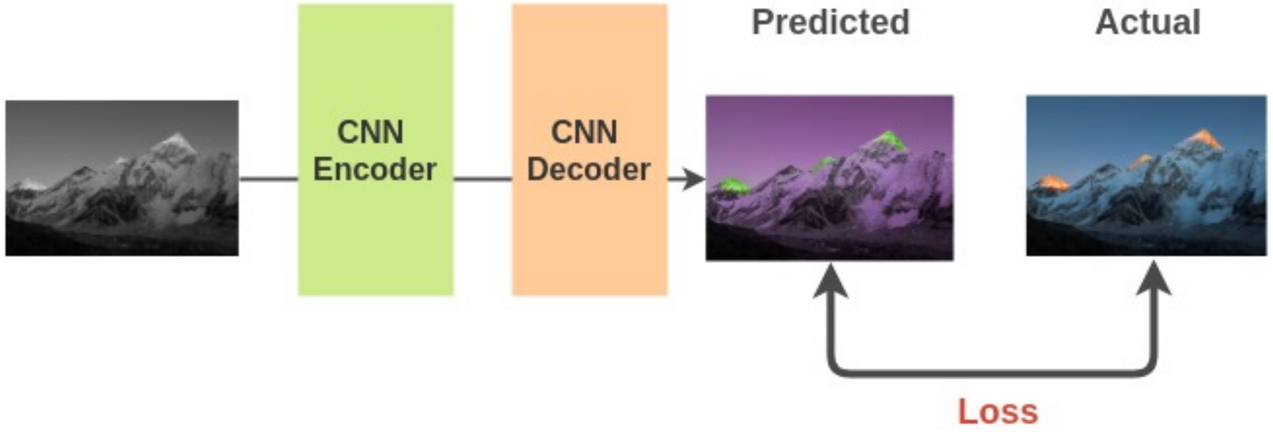


Source: [Jin19]

Image Colorization

Data generation

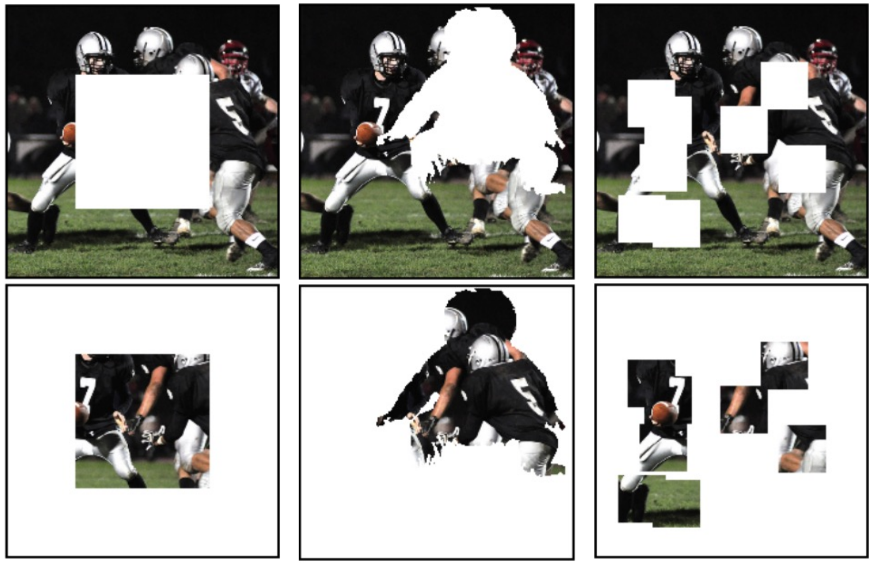
Pretext task: l_2 loss between gray and color version



Source: https://amitness.com/2020/02/illustrated-self-supervised-learning/

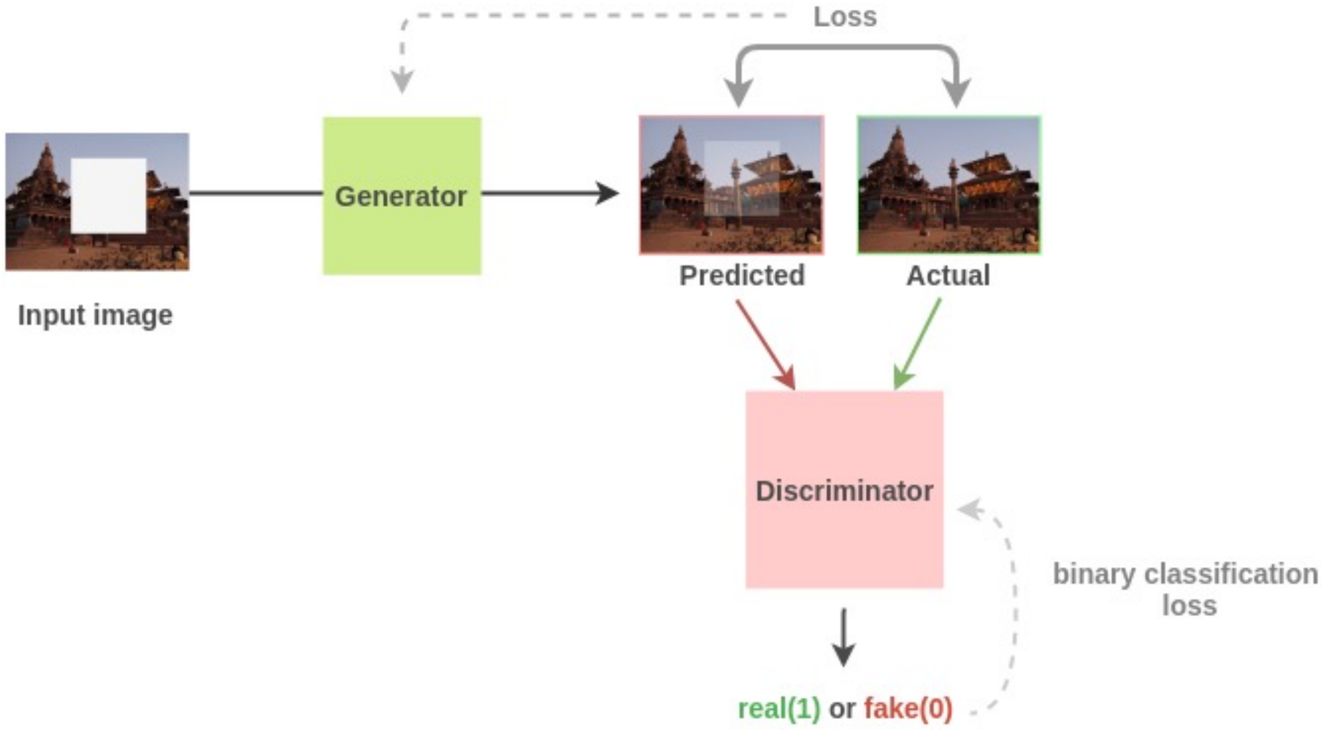
Image Inpainting

Data generation



Source: [Pat16]

Pretext task

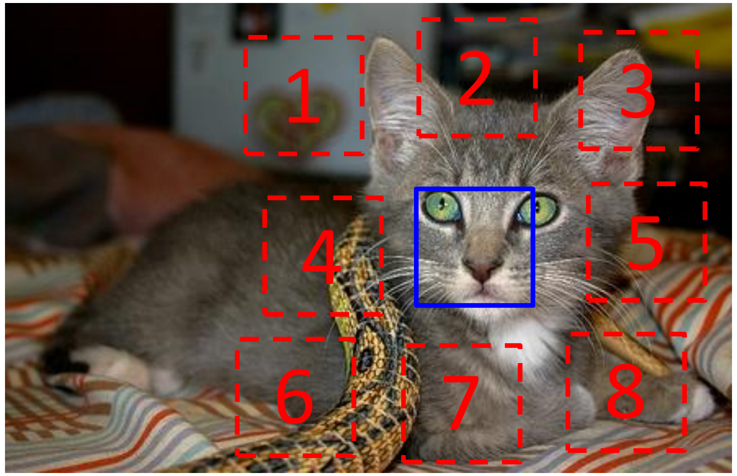


Source: https://amitness.com/2020/02/illustrated-self-supervised-learning/

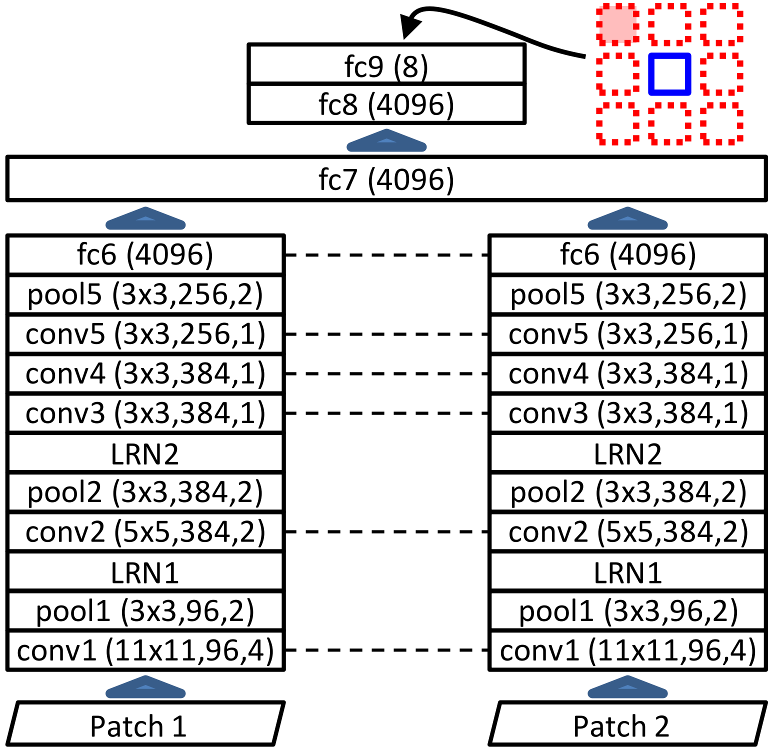
Solve Jigsaw Puzzle

Technische Hochschule Rosenheim

Predict spatial configuration of a patch to selected reference patch (blue)



X = ( , ); Y = 3



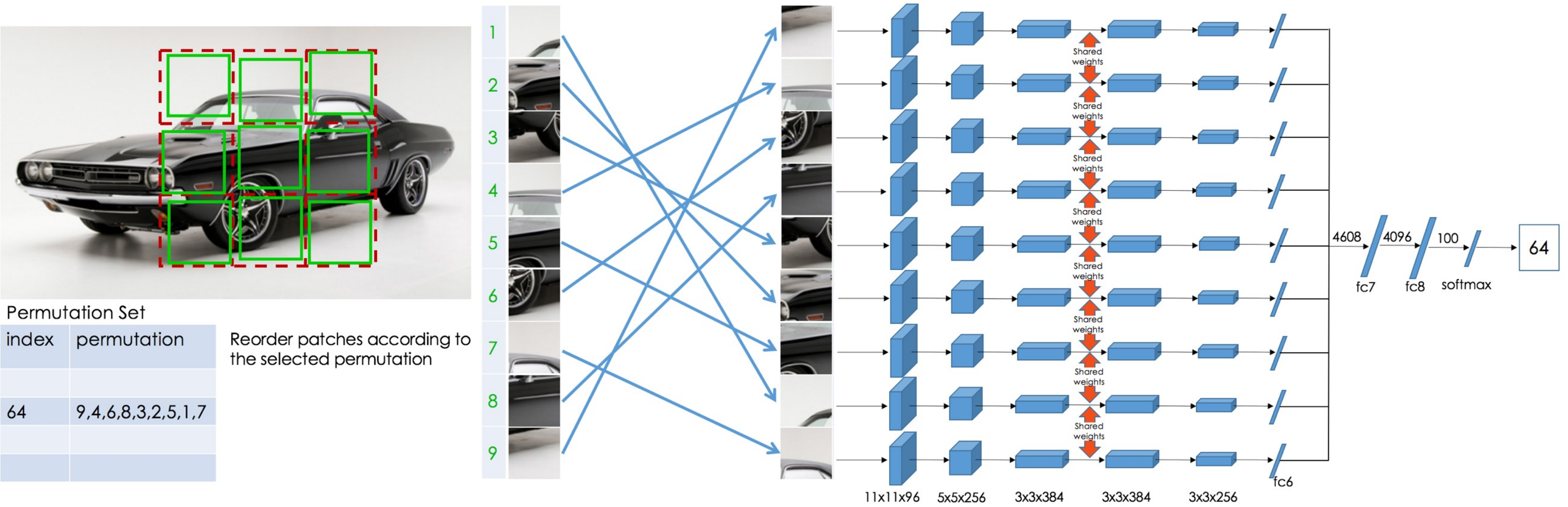
Attention: Avoid trivial shortcuts that the network might learn to use

- boundary patterns, continuing textures $\longrightarrow$ use large enough gaps
- prevent network from only learning about color information: introduce chromatic aberration
  - pre-process images by shifting green and magenta toward gray or
  - randomly drop 2 color channels

Source: [Doe15]

Solve Jigsaw Puzzle++

Predict position of each of the tiles



9 tiles ⟶ 9! = 362 880 possible permutations

After training of Jigsaw problem: Transfer Learning

- Transfer weights of conv layers to AlexNet
- init fully connected AlexNet layers randomly
- use Fast R-CNN architecture for detection

Source: [Nor16]

Solve Jigsaw Puzzle++

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
100	8.08	2	88	52.6
95	8.08	3	90	52.4
85	8.07	4	91	52.7
71	8.07	5	92	52.8
35	8.13	6	94	52.6
10	8.57	7	97	49.2
7	8.95	8	98	49.6
6	9	9	99	49.7

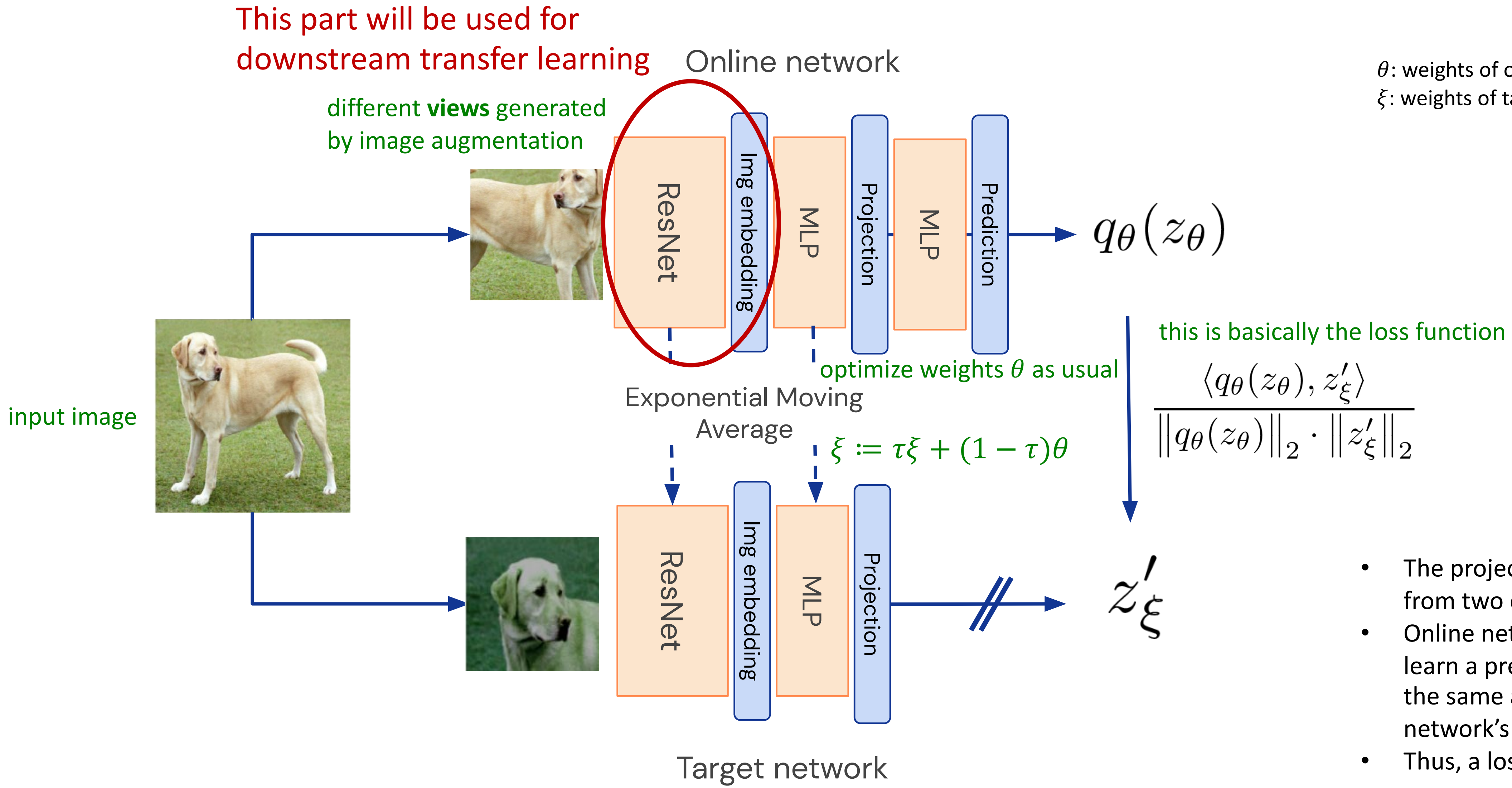
Detection performance: Evaluated after transfer learning on Pascal VOC dataset

Source: [Nor16]

BYOL – Bootstrap Your Own Latent

- New approach to self-supervised learning [Gri20]
- Uses two neural networks that learn from each other
- Transfer weights of one network to downstream tasks
  - 74.3% top-1 classification accuracy on ImageNet (ResNet-50), 91.6% top-5
  - 79.6% top-1 classification accuracy on ImageNet (ResNet-200), 94.8% top-5

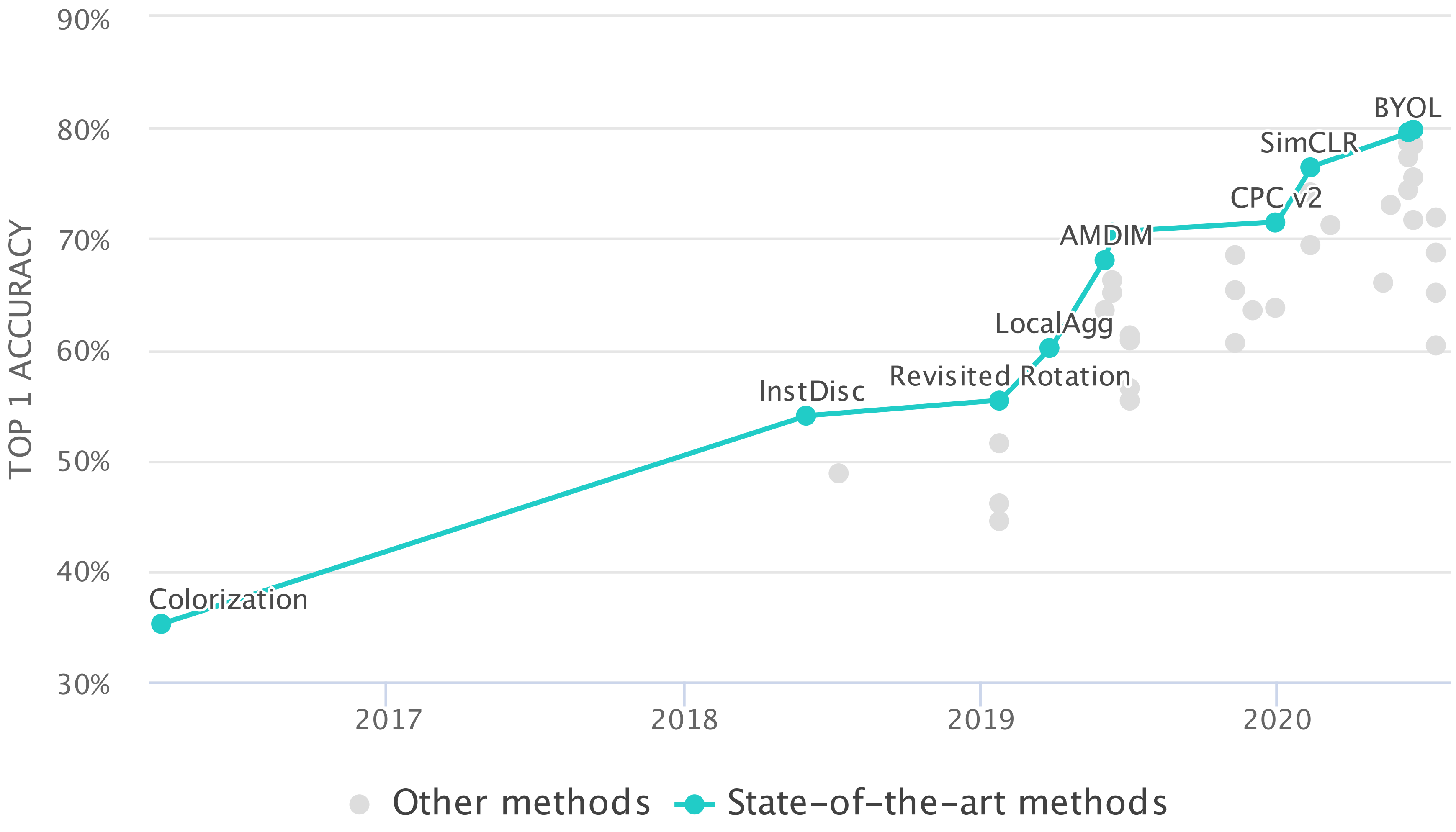
BYOL – Architecture



θ: weights of online network
ξ: weights of target network

- The projections originate from two different views
- Online network tries to learn a prediction that is the same as the target network's projection.
- Thus, a loss can be defined.

SSL – Classification Performance on ImageNet



Source: https://paperswithcode.com/sota/self-supervised-image-classification-on

References

[Cor16] Marius Cordts, Mohamed Omran, Sebastian Ramos, et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: CoRR abs/1604.01685 (2016). arXiv: 1604.01685.

[Doe15] C. Doersch, A. Gupta, and A. A. Efros. “Unsupervised Visual Representation Learning by Context Prediction”. In: 2015 IEEE International Conference on Computer Vision (ICCV). Dec. 2015, pp. 1422–1430.

[Gri20] Jean-Bastien Grill, Florian Strub, Florent Altché, et al. “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning”. In: arXiv e-prints, arXiv:2006.07733 (June 2020)

[Jin19] Longlong Jing and Yingli Tian. “Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey”. In: arXiv e-prints, arXiv:1902.06162

[Lin14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, et al. “Microsoft COCO: Common Objects in Context”. In: CoRR abs/1405.0312 (2014). arXiv: 1405.0312.

[Mat19] Matterport, Inc. Mask R-CNN for Object Detection and Segmentation. https://github.com/matterport/Mask_RCNN

[Nor16] Mehdi Noroozi and Paolo Favaro. “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles”. In: Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016, pp. 69–84.

[Pat16] D. Pathak, P. Krähenbühl, J. Donahue, et al. “Context Encoders: Feature Learning by Inpainting”. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, pp. 2536–2544.