Universität St.Gallen

IGARSS

IGRARSS 2023 Tutorial

Self-supervised Learning & Contrastive Learning for Earth Observation

Joëlle Hanna, Michael Mommert, Linus Scheibenreif, Damian Borth

Universität St.Gallen

Machine Learning

Universität St.Gallen

Machine Learning

Supervised Learning

- Goal: Learn the mapping between data and output:

$$f\colon x_i \longrightarrow \hat{y}_i$$

- Data: (x_i, y_i) **labels are known**
- Objective: $\mathcal{L}(\hat{y}_i, y_i)$ based on label error/loss

Unsupervised Learning

- Goal: Learn unknown distributions, pattern and dependencies from data

$$f\colon x_i \longrightarrow z_i \quad g\colon z_i \longrightarrow \hat{\boldsymbol{x}}_i$$

- Data: (x_i) **labels are unknown**
- Objective: $\mathcal{L}(\hat{\boldsymbol{x}}_i, x_i)$ based on reconstruction / reassignment

Universität St.Gallen

Machine Learning

Supervised Learning

- Goal: Learn the mapping between data and output:

$$f: x_i \longrightarrow \hat{y}_i$$

- Data: (x_i, y_i) **labels are known**
- Objective: $\mathcal{L}(\hat{y}_i, y_i)$ based on label error/loss

Semi-supervised Learning

- Goal: Learn structure from data and supervise by using only few labels

Self-supervised Learning

- Goal: Learn representation of data by pre-text task and transfer to downstream tasks

Unsupervised Learning

- Goal: Learn unknown distributions, pattern and dependencies from data

$$f: x_i \longrightarrow z_i \quad g: z_i \longrightarrow \hat{\boldsymbol{x}}_i$$

- Data: (x_i) **labels are unknown**
- Objective: $\mathcal{L}(\hat{\boldsymbol{x}}_i, x_i)$ based on reconstruction / reassignment

Universität St.Gallen

Machine Learning

Supervised Learning

- Goal: Learn the mapping between data and output:

$$f: x_i \longrightarrow \hat{y}_i$$

- Data: (x_i, y_i) **labels are known**
- Objective: $\mathcal{L}(\hat{y}_i, y_i)$ based on label error/loss

Semi-supervised Learning

- Goal: Learn structure from data and supervise by using only few labels

Self-supervised Learning

- Goal: Learn representation of data by pre-text task and transfer to downstream tasks

Unsupervised Learning

- Goal: Learn unknown distributions, pattern and dependencies from data

$$f: x_i \longrightarrow z_i \quad g: z_i \longrightarrow \hat{\boldsymbol{x}}_i$$

- Data: (x_i) **labels are unknown**
- Objective: $\mathcal{L}(\hat{\boldsymbol{x}}_i, x_i)$ based on reconstruction / reassignment

Representation Learning / Transfer Learning

- Goal: Learn an internal representation which can be used as robust preprocessor of underlying data
- Goal: Use a previously learned representation to adapt towards a new task / domain

Universität St.Gallen

Machine Learning

Supervised Learning

- Goal: Learn the mapping between data and output:

$$f: x_i \longrightarrow \hat{y}_i$$

- Data: (x_i, y_i) **labels are known**
- Objective: $\mathcal{L}(\hat{y}_i, y_i)$ based on label error/loss

Semi-supervised Learning

- Goal: Learn structure from data and supervise by using only few labels

Self-supervised Learning

- Goal: Learn representation of data by pre-text task and transfer to downstream tasks

Unsupervised Learning

- Goal: Learn unknown distributions, pattern and dependencies from data

$$f: x_i \longrightarrow z_i \quad g: z_i \longrightarrow \hat{\boldsymbol{x}}_i$$

- Data: (x_i) **labels are unknown**
- Objective: $\mathcal{L}(\hat{\boldsymbol{x}}_i, x_i)$ based on reconstruction / reassignment

Representation Learning / Transfer Learning

- Goal: Learn an internal representation which can be used as robust preprocessor of underlying data
- Goal: Use a previously learned representation to adapt towards a new task / domain

Adversarial Learning

- Goal: Learn via a proxy loss given by a min-max optimization between a generator and discriminator

Reinforcement Learning

- Goal: Learn given an interaction of the agent and its environment maximize the notion of cumulative reward.

Universität St.Gallen

Machine Learning

Supervised Learning

- Goal: Learn the mapping between data and output:

$$f: x_i \longrightarrow \hat{y}_i$$

- Data: (x_i, y_i) **labels are known**
- Objective: $\mathcal{L}(\hat{y}_i, y_i)$ based on label error/loss

Semi-supervised Learning

- Goal: Learn structure from data and supervise by using only few labels

Self-supervised Learning

- Goal: Learn representation of data by pre-text task and transfer to downstream tasks

Unsupervised Learning

- Goal: Learn unknown distributions, pattern and dependencies from data

$$f: x_i \longrightarrow z_i \quad g: z_i \longrightarrow \hat{x}_i$$

- Data: (x_i) **labels are unknown**
- Objective: $\mathcal{L}(\hat{x}_i, x_i)$ based on reconstruction / reassignment

Representation Learning / Transfer Learning

- Goal: Learn an internal representation which can be used as robust preprocessor of underlying data
- Goal: Use a previously learned representation to adapt towards a new task / domain

Adversarial Learning

- Goal: Learn via a proxy loss given by a min-max optimization between a generator and discriminator

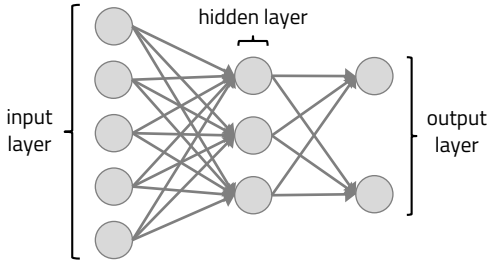
Reinforcement Learning

- Goal: Learn given an interaction of the agent and its environment maximize the notion of cumulative reward.

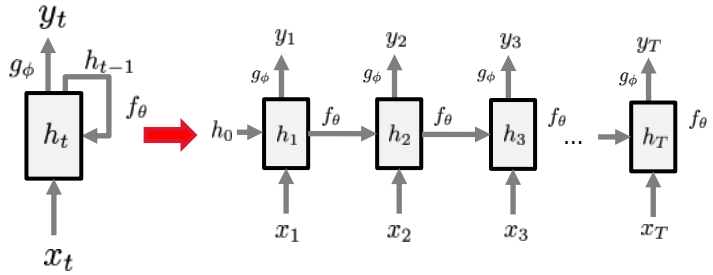
Representation Learning

Deep Neural Network Architectures

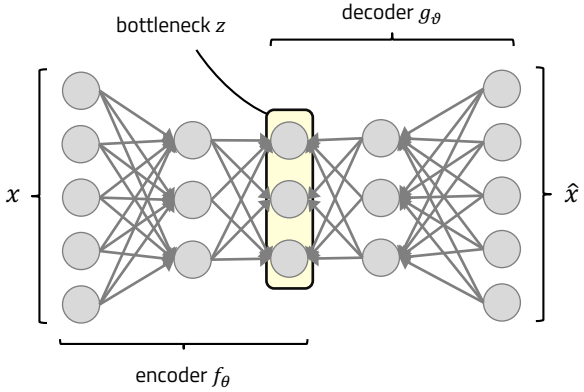
Multilayer Perceptrons (MLP)



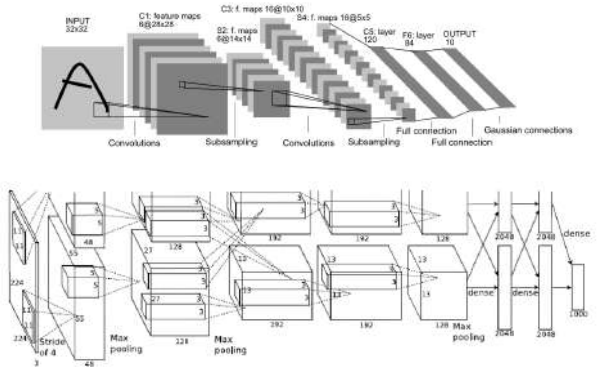
Recurrent Neural Networks (RNN) Long-Short Term Memory (LSTM)



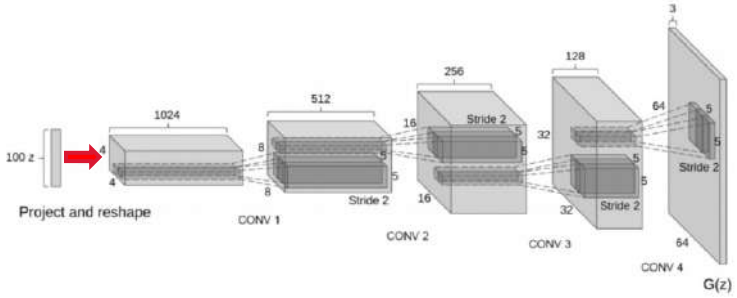
Deep Autoencoder



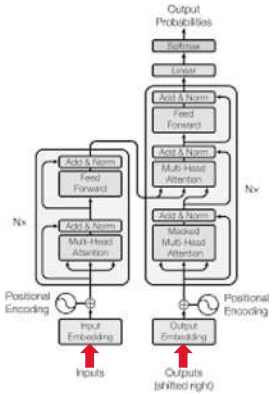
Convolutional Neural Networks (CNN)



Generative Adversarial Networks (GAN)

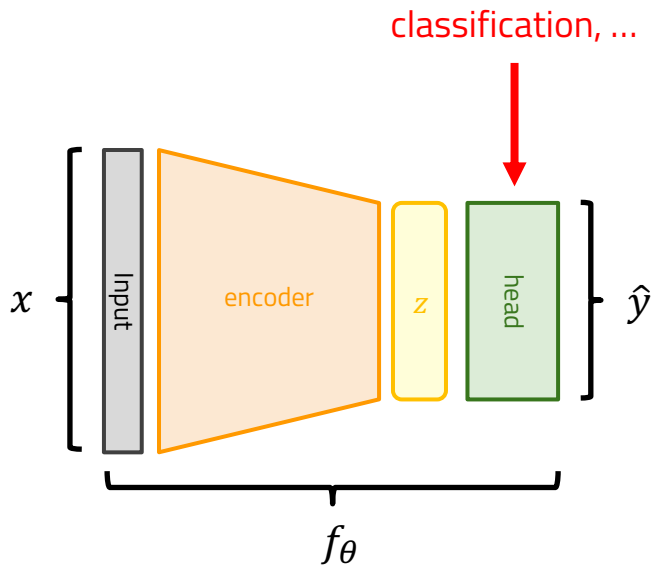


Attention & Transformer

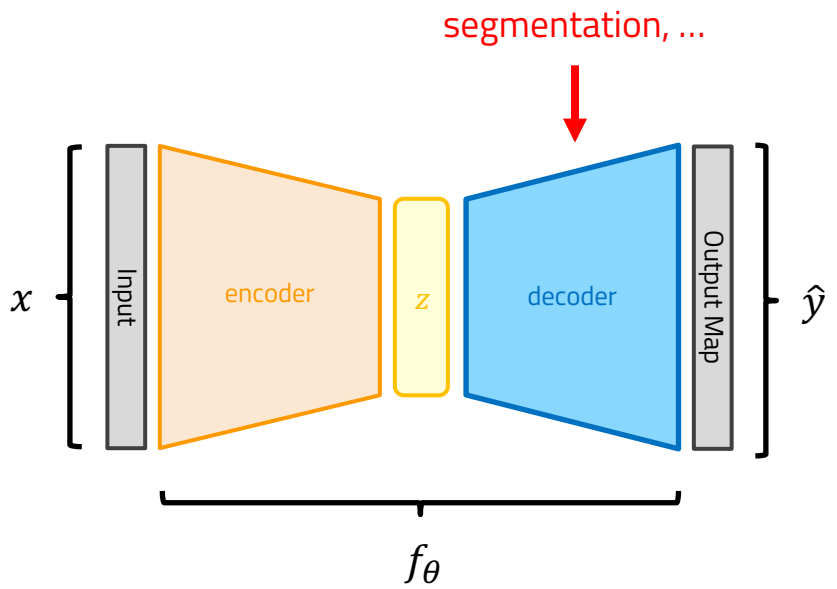


Deep Neural Networks = Representation Learning

Discriminative Tasks

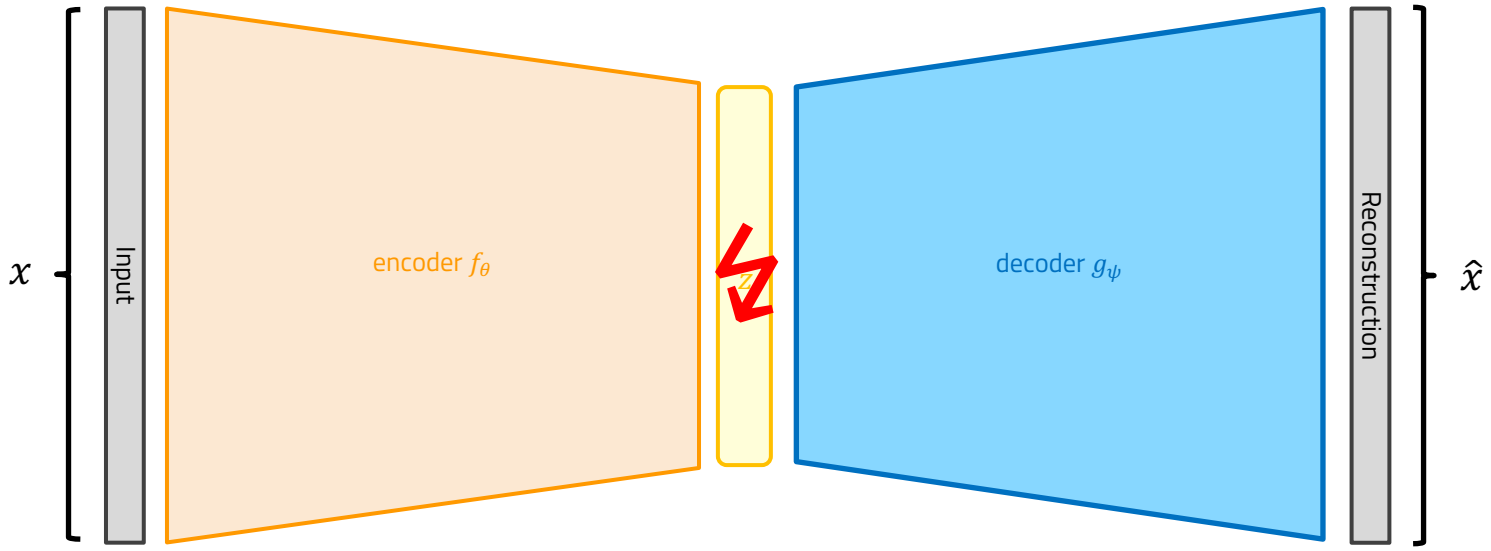


Generative Tasks



How can we be become more efficient in learning these representations?

Why not combining both?



Representation Learning

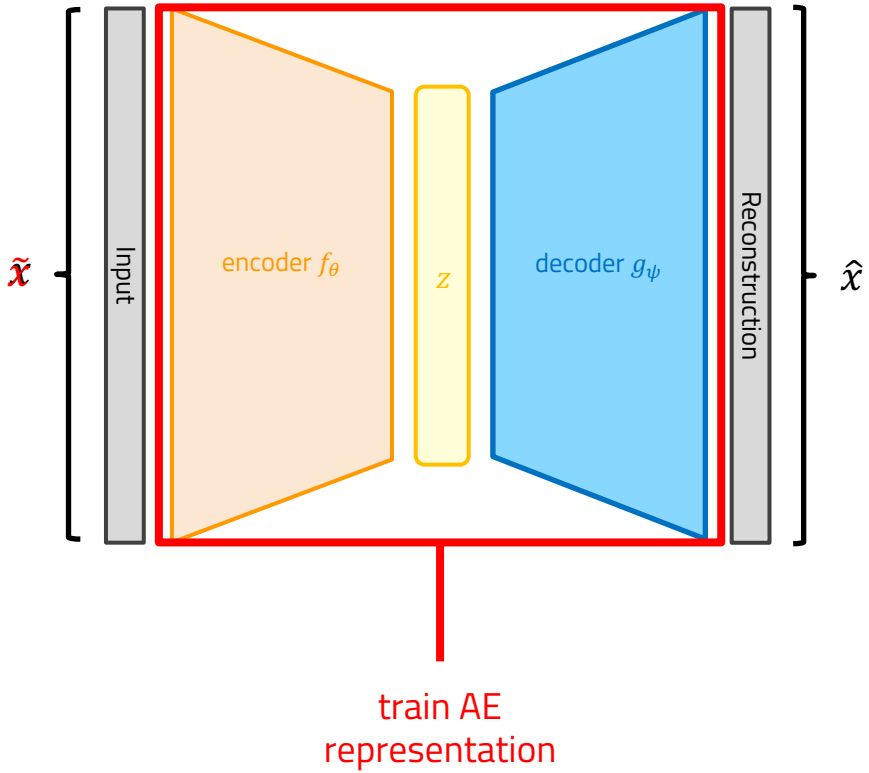
Challenge:
Autoencoder with depth > input dim runs the risk to learn the identify function i.e., to memorize samples.

Solution:
- Extend AE to VAE
- Stacked Denoising Autoencoder

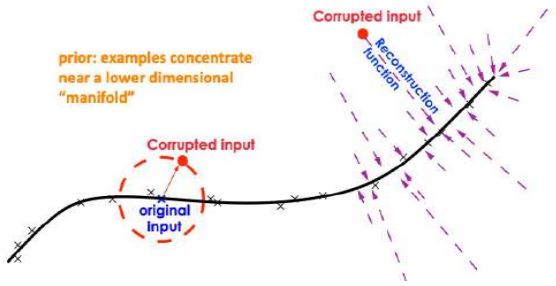
Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol.
Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
Journal of Machine Learning Research (JMLR), 11:3371–3408, 2010

Why not combining both?

$$\theta^* = \min_\theta \mathcal{J}(\theta) = \sum \mathcal{L}(\hat{x}_i, x_i)$$



$\mathcal{T}: x \rightarrow \tilde{x}$ $\hat{x}$



Representation Learning

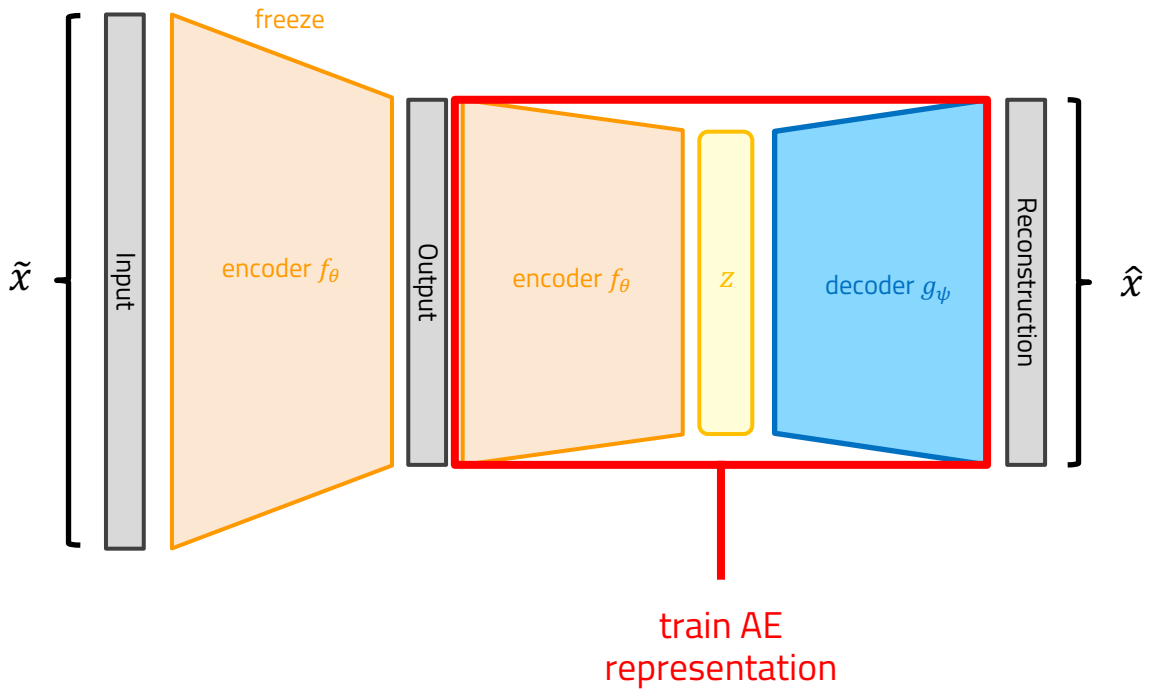
Challenge:
Autoencoder with depth > input dim runs the risk to learn the identify function i.e., to memorize samples.

Solution:
- Extend AE to VAE
- Stacked Denoising Autoencoder ←

Idea:
- unsupervised training of AE with randomly corrupted inputs (e.g., noise)

Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol.
Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
Journal of Machine Learning Research (JMLR), 11:3371–3408, 2010

Why not combining both?



Representation Learning

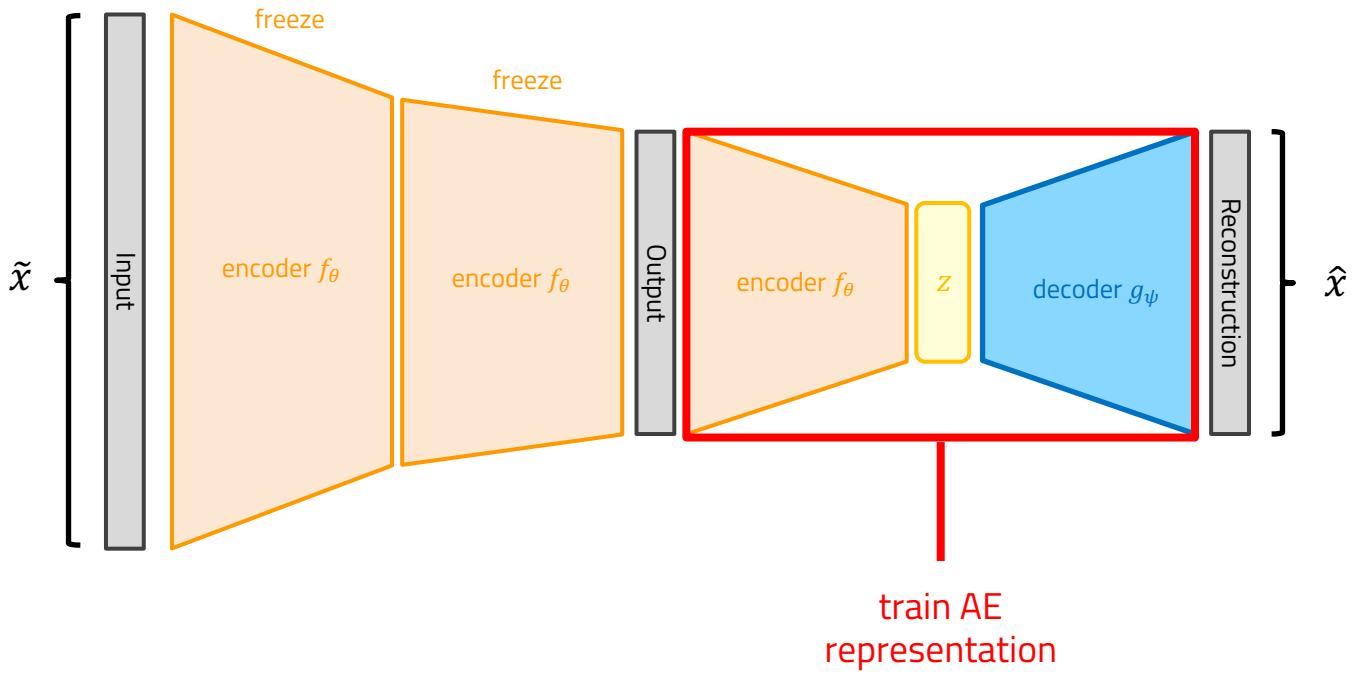
Challenge:
Autoencoder with depth > input dim runs the risk to learn the identify function i.e., to memorize samples.

Solution:
- Extend AE to VAE
- Stacked Denoising Autoencoder ←

Idea:
- unsupervised training of AE with randomly corrupted inputs (e.g., noise)

Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol.
Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
Journal of Machine Learning Research (JMLR), 11:3371–3408, 2010

Why not combining both?



Representation Learning

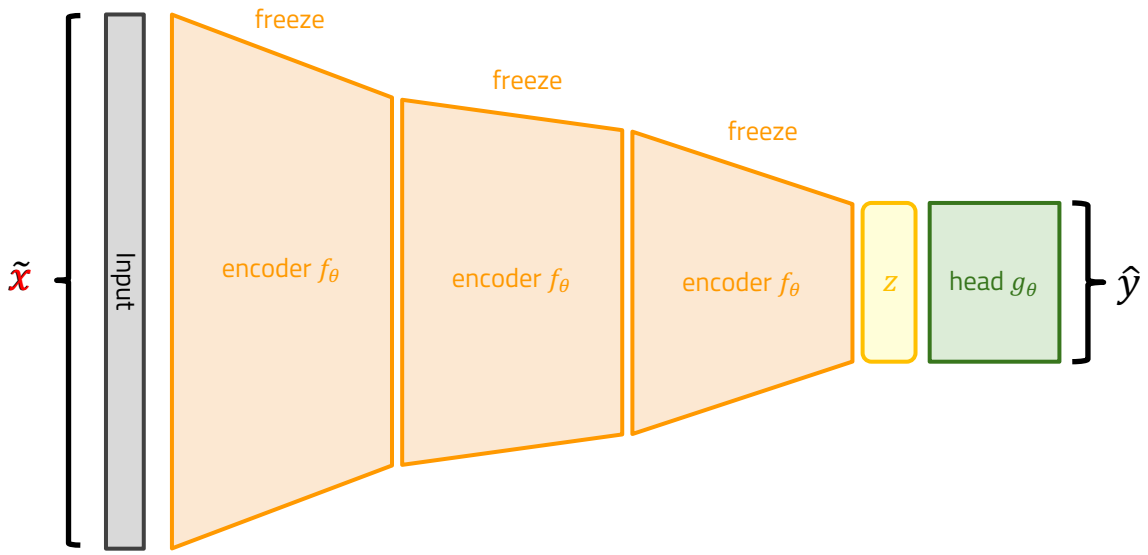
Challenge:
Autoencoder with depth > input dim runs the risk to learn the identify function i.e., to memorize samples.

Solution:
- Extend AE to VAE
- Stacked Denoising Autoencoder ←

Idea:
- unsupervised training of AE with randomly corrupted inputs (e.g., noise)

Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol.
Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
Journal of Machine Learning Research (JMLR), 11:3371–3408, 2010

Why not combining both?



Representation Learning

Challenge:
Autoencoder with depth > input dim runs the risk to learn the identify function i.e., to memorize samples.

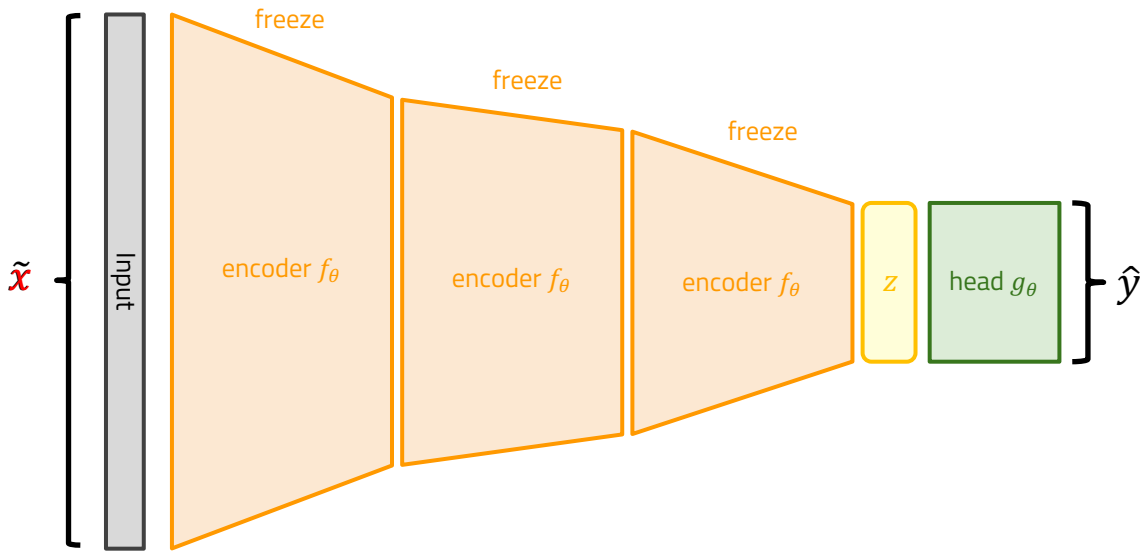
Solution:
- Extend AE to VAE
- Stacked Denoising Autoencoder

Idea:
- unsupervised training of AE with randomly corrupted inputs (e.g., noise)
- supervised training

Fundamental Concepts:
1. pre-training
2. pretext task

Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol.
Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
Journal of Machine Learning Research (JMLR), 11:3371–3408, 2010

Why not combining both?



Representation Learning

Challenge:
Autoencoder with depth > input dim runs the risk to learn the identify function i.e., to memorize samples.

Solution:
- Extend AE to VAE
- Stacked Denoising Autoencoder

Idea:
- unsupervised training of AE with randomly corrupted inputs (e.g., noise)
- supervised training

Fundamental Concepts:
1. pre-training -> transfer learning
2. pretext task -> self-supervised learning

Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol.
Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
Journal of Machine Learning Research (JMLR), 11:3371–3408, 2010

Self-supervised Learning

Universität St.Gallen

Machine Learning

Supervised Learning

- Goal: Learn the mapping between data and output:

$$f: x_i \longrightarrow \hat{y}_i$$

- Data: (x_i, y_i) **labels are known**
- Objective: $\mathcal{L}(\hat{y}_i, y_i)$ based on label error/loss

Semi-supervised Learning

- Goal: Learn structure from data and supervise by using only few labels

Self-supervised Learning

- Goal: Learn representation of data by pre-text task and transfer to downstream tasks

Unsupervised Learning

- Goal: Learn unknown distributions, pattern and dependencies from data

$$f: x_i \longrightarrow z_i \quad g: z_i \longrightarrow \hat{x}_i$$

- Data: (x_i) **labels are unknown**
- Objective: $\mathcal{L}(\hat{x}_i, x_i)$ based on reconstruction / reassignment

Representation Learning / Transfer Learning

- Goal: Learn an internal representation which can be used as robust preprocessor of underlying data
- Goal: Use a previously learned representation to adapt towards a new task / domain

Adversarial Learning

- Goal: Learn via a proxy loss given by a min-max optimization between a generator and discriminator

Reinforcement Learning

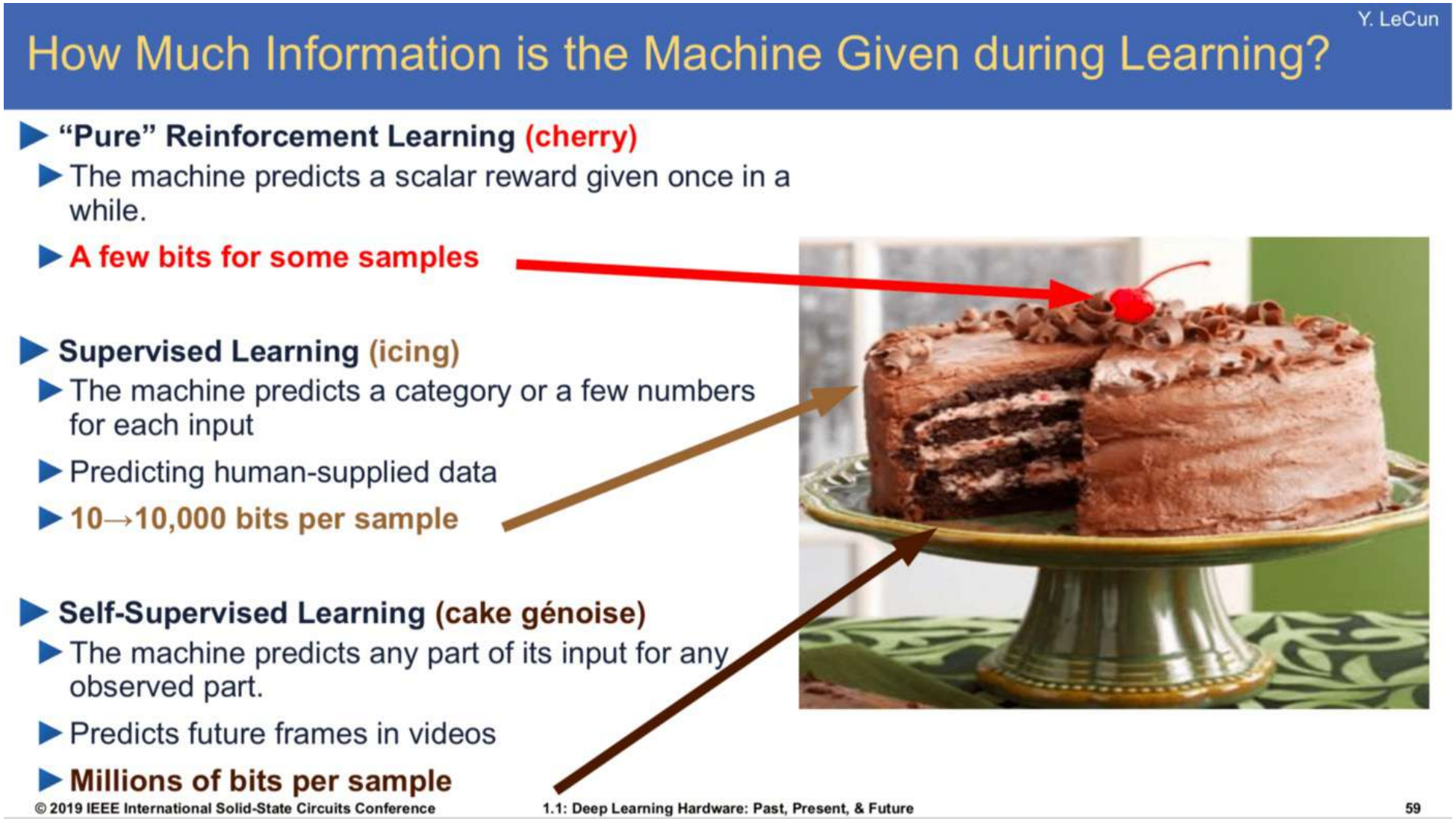
- Goal: Learn given an interaction of the agent and its environment maximize the notion of cumulative reward.

Yann LeCun's Cake (updated 2019)

ACM Turing Award 2018

Long-term DL researcher

CNN -> LeNet



SSL according to Yan LeCun

ACM Turing Award 2018

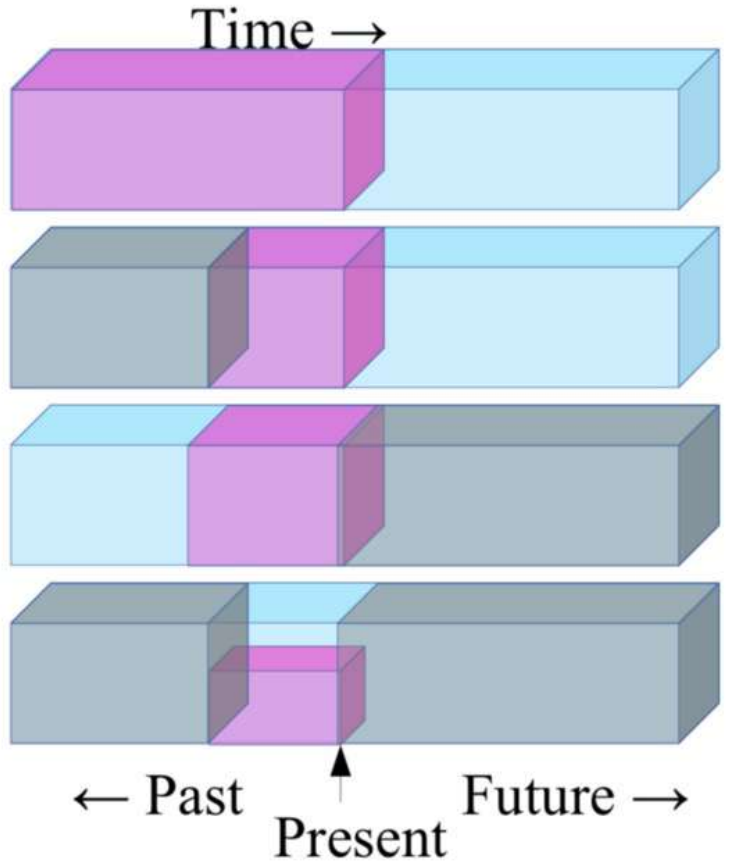
Long-term DL researcher

CNN -> LeNet

Self-Supervised Learning

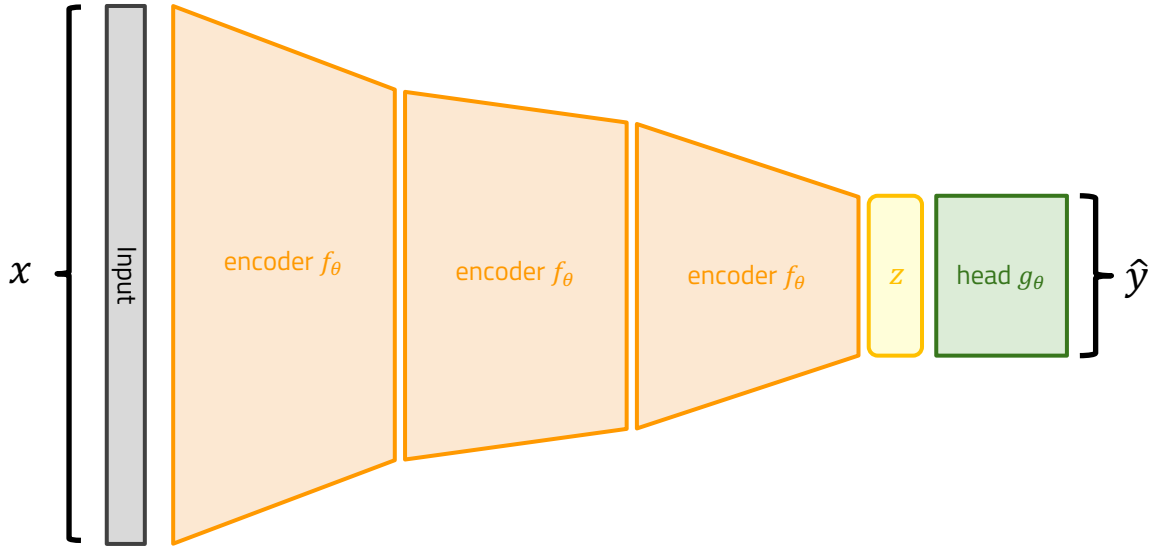
Y. LeCun

- 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.



Unsupervised = Self-supervised?

Stacked Denoising Autoencoder



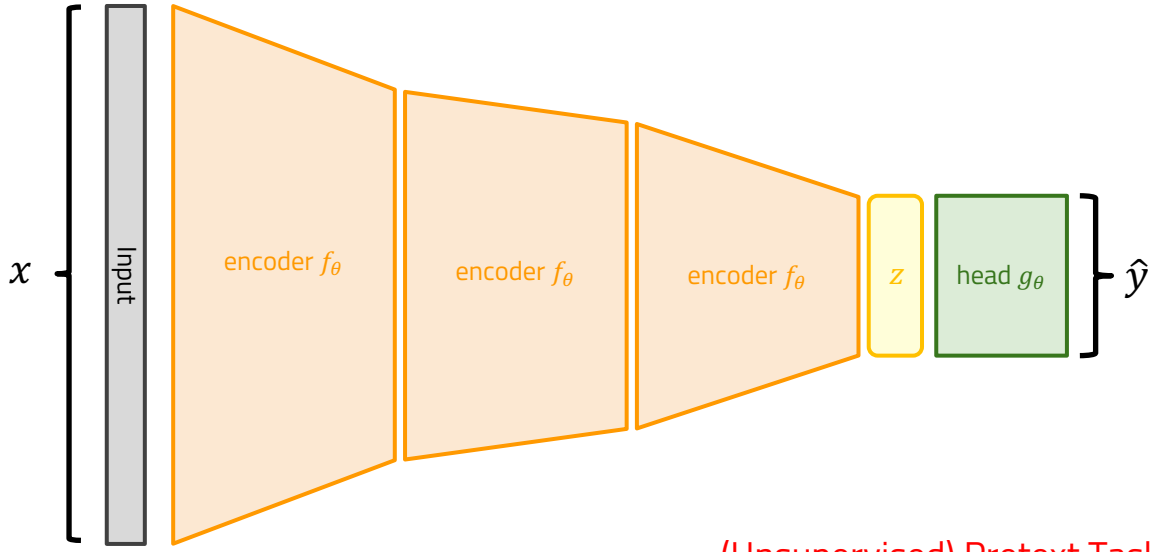
- Encoder projects x into latent space $z \Longrightarrow$ feature representation
- Decoder can reconstruct back to original space $\Longrightarrow$ generative model

Goal: learn a lower-dimensional embedding of data onto a manifold for linear separability of original non-linear separable data

- Pre-trained representation taken for fine-tuning $\Longrightarrow$ transfer learning

Unsupervised = Self-supervised?

Stacked Denoising Autoencoder



(Unsupervised) Pretext Task

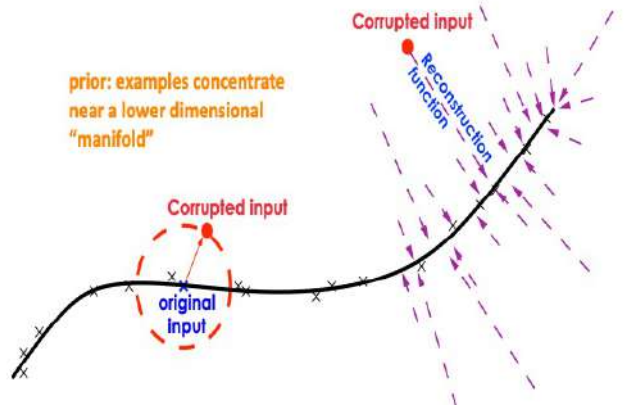
- Encoder projects x into latent space $z \Longrightarrow$ feature representation
- Decoder can reconstruct back to original space $\Longrightarrow$ generative model

Goal: learn a lower-dimensional embedding of data onto a manifold for linear separability of original non-linear separable data

- Pre-trained representation taken for fine-tuning $\Longrightarrow$ transfer learning

(Supervised) Downstream Task

Learn to reconstruct corrupted Images



Secret Ingredient:

- we made the task more difficult !
- we extended our learning task with data manipulation:

$$\mathcal{T}: x_i \rightarrow \widetilde{x_i}, \qquad \hat{x} = g_\psi\big(f_\theta(\widetilde{x_i})\big), \qquad \mathcal{L}(\hat{x_i}, x_i)$$

More general principle:

- pretend there is a part of the data you don't know and predict that
- following, the model has to learn a lot more about the data

Self-supervised Learning

Setup

Large amounts of unlabeled data

Self-supervised Learning

Learn a good feature representation from self-supervised pretext tasks, i.e., labels are generated automatically

f_θ

feature encoder

Supervised Learning (fine-tuning)

Attach a shallow network on the feature encoder; train the shallow network on the downstream task with small amount of labeled data

evaluate on the downstream task e.g. classification, regression, detection, ...

Self-supervised Learning

Setup

Large amounts of unlabeled data

Self-supervised Learning

Learn a good feature representation from self-supervised pretext tasks, i.e., labels are generated automatically

f_θ

feature encoder

Supervised Learning (fine-tuning)

Attach a shallow network on the feature encoder; train the shallow network on the downstream task with small amount of labeled data

evaluate on the downstream task e.g. classification, regression, detection, ...

Goal

SSL aims to learn rich task-agnostic representations from raw unlabeled data suitable for many different downstream tasks

We want to be able to train equally good models without too many labels

We potentially might be able to generalize better because we have to learn more about the world

Evaluation

Evaluation of SSL (usually) focuses on the downstream task and not so much on the self-supervised learning task

We don't care so much about the pretext task performance

We want rich (disentangled) representations -> evaluation on downstream task fine-tuning a linear model with labels

We want to know about generalizability of our rep-representation -> fine-tuning with non-linear model with labels

Discriminative Pretext tasks

Universität St.Gallen

Taxonomy of Pretext tasks

Generative

Pretext task aims for reconstruction of either the entire sample or parts of the sample to inform learning.

Reconstruction

- VAE
- GAN
- Autoregressive
- Flow-based
- Diffusion

Self-Prediction

- Masked Prediction
  - Denoising AE
  - Context AE
  - Image Completion
- Channel Shuffling
  - Colorization
  - Split-brain

Discriminative

Pretext task aims at discrimination either between innate relationship or contrast between sample pairs to learning.

Innate Relationship

- Patch Positioning
  - Relative Position
  - Jigsaw
- Image Rotation
- Feature Counting
- Contrastive Predictive Coding

Contrastive

- Instance Discrim. / Augmentation
  - Feature Discrimination
  - MoCo
  - SimCLR
- Clustering-based
- Positive-Contrast
  - BOYT

Source: OpenAI - SSL Tutorial, NeurIPS, 2021

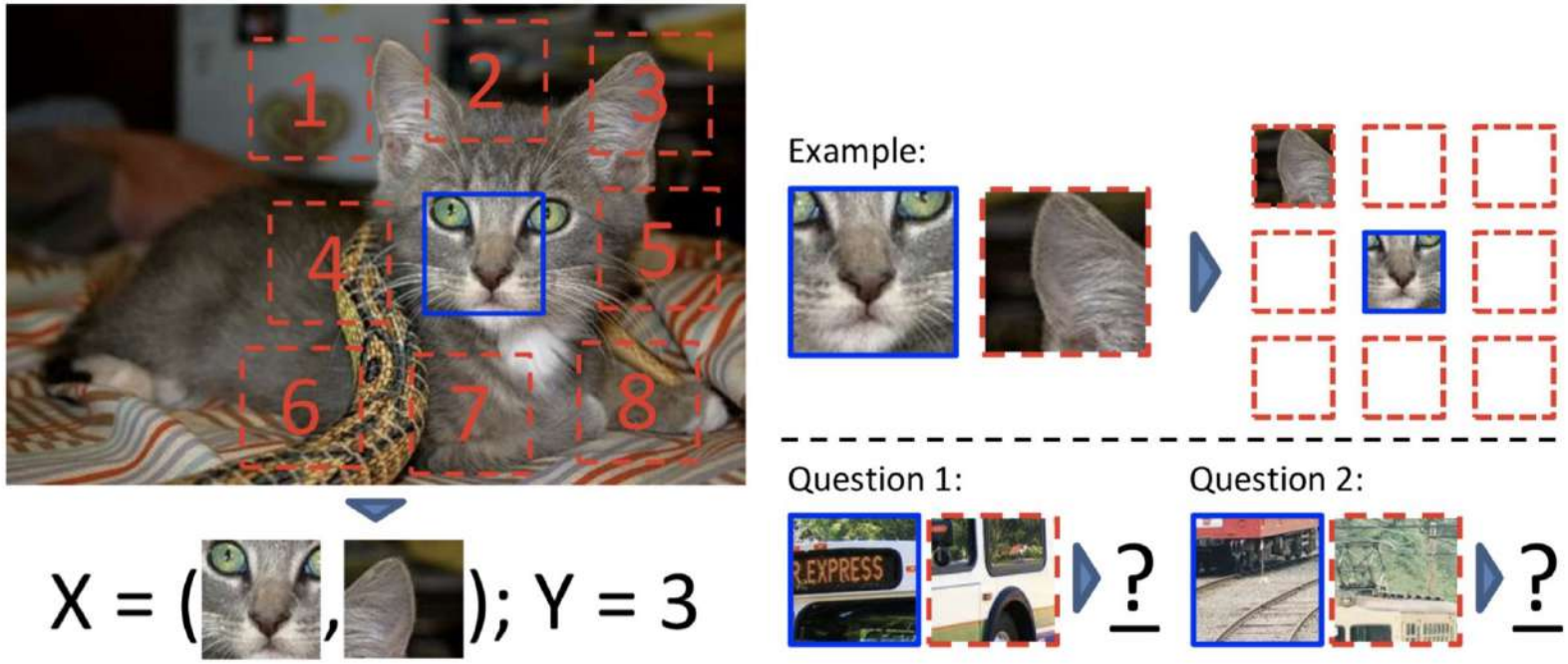
Discriminative Pretext Task

Predict relative patch locations

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

Discriminative Pretext Task

Predict relative patch locations

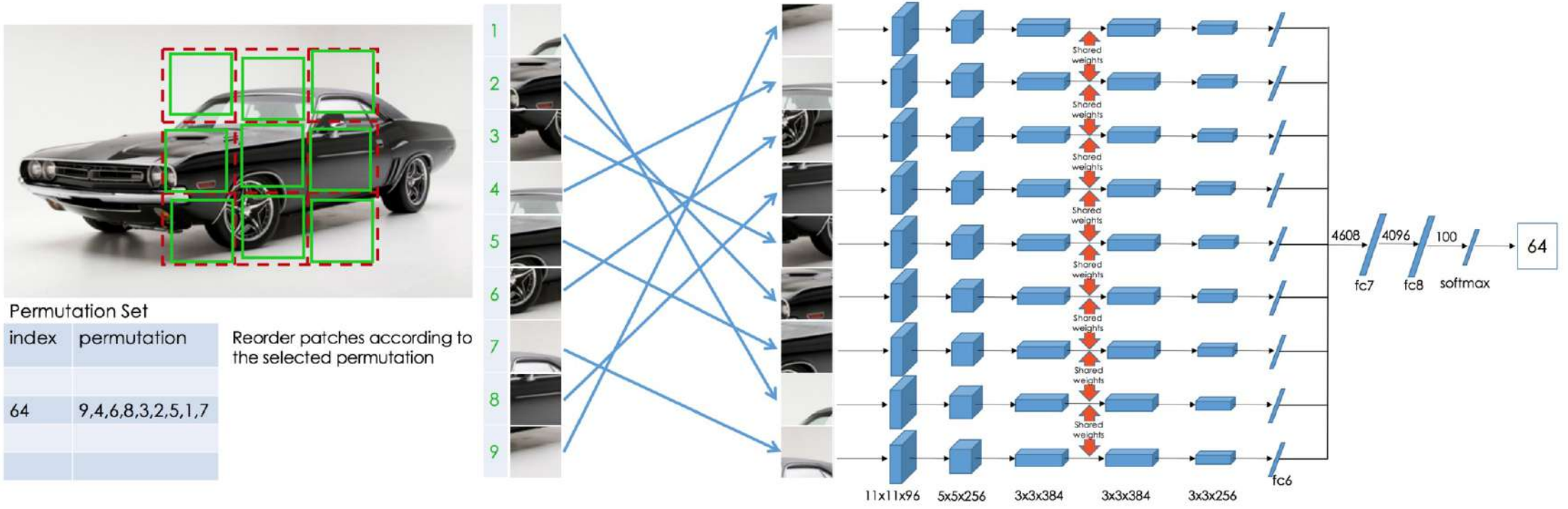


Idea: doing well on this task requires the model to learn to recognize objects and their parts. It learns feature representation using context, which indeed captures visual similarity across images.

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

Discriminative Pretext Task

Solve a Jigsaw puzzle



Idea: By training the model to solve Jigsaw puzzles, we can learn both a feature mapping of object parts as well as their correct spatial arrangement This way the model learns features able capture semantically relevant content.

Mehdi Noroozi and Paolo Favaro, **Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles**, ECCV 2016

Universität St.Gallen

Discriminative Pretext Task

Solve a Jigsaw puzzle

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%**

SSL Data:
Self-supervised learning on ImageNet (entire training set) with AlexNet.

Downstream Data:
Finetune on labelled data from Pascal VOC 2007

Downstream Tasks:
- classification
- detection
- segmentation

Mehdi Noroozi and Paolo Favaro, **Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles**, ECCV 2016

Universität St.Gallen

Discriminative Pretext Task

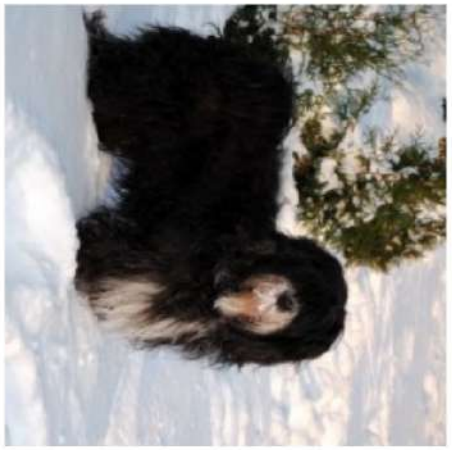
Rotation

Discriminative Pretext Task

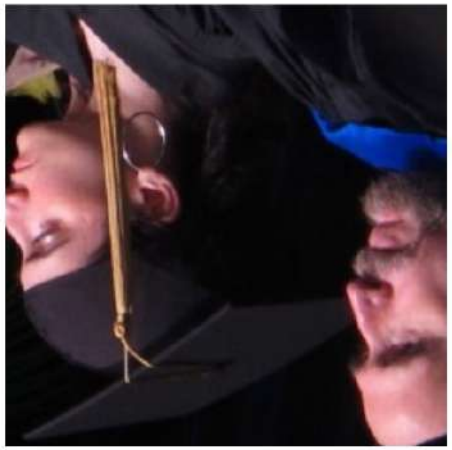
Rotation

x

90° rotation



270° rotation



180° rotation

0° rotation

270° rotation

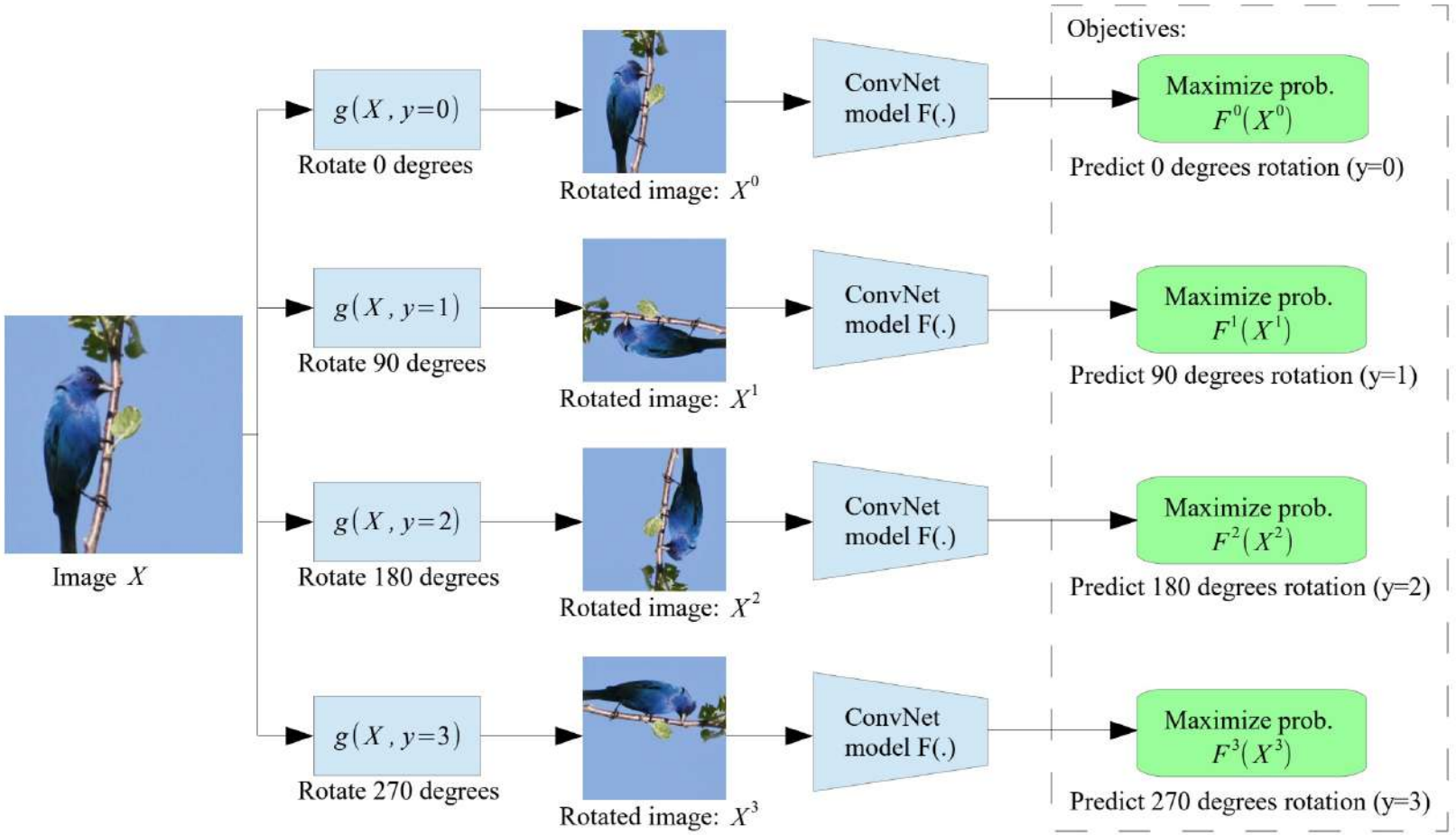
$\hat{y}$

Idea: a model could recognize the correct rotation of an object only if it has the "visual common sense" of what the object should look like unperturbed.

Gidaris et al., **Unsupervised Representation Learning by Predicting Image Rotations**, ICLR, 2018

Discriminative Pretext Task

Rotation



Idea:
Self-supervised learning by rotating the entire input images.

Label:
The model learns to predict which rotation is applied (4-way classification)

Gidaris et al., **Unsupervised Representation Learning by Predicting Image Rotations**, ICLR, 2018

Discriminative Pretext Task

Rotation

	Classification (%mAP)		Detection (%mAP)	Segmentation (%mIoU)
Trained layers	fc6-8	all	all	all
ImageNet labels	78.9	79.9	56.8	48.0
Random		53.3	43.4	19.8
Random rescaled Krähenbühl et al. (2015)	39.2	56.6	45.6	32.6
Egomotion (Agrawal et al., 2015)	31.0	54.2	43.9	
Context Encoders (Pathak et al., 2016b)	34.6	56.5	44.5	29.7
Tracking (Wang & Gupta, 2015)	55.6	63.1	47.4	
Context (Doersch et al., 2015)	55.1	65.3	51.1	
Colorization (Zhang et al., 2016a)	61.5	65.6	46.9	35.6
BIGAN (Donahue et al., 2016)	52.3	60.1	46.9	34.9
Jigsaw Puzzles (Noroozi & Favaro, 2016)	-	67.6	53.2	37.6
NAT (Bojanowski & Joulin, 2017)	56.7	65.3	49.4	
Split-Brain (Zhang et al., 2016b)	63.0	67.1	46.7	36.0
ColorProxy (Larsson et al., 2017)		65.9		38.4
Counting (Noroozi et al., 2017)	-	67.7	51.4	36.6
(Ours) RotNet	**70.87**	**72.97**	**54.4**	**39.1**

pre-trained on ImageNet + fine-tuning on PASCAL

no pretraining

proposed SSL

SSL Data:
Self-supervised learning on ImageNet (entire training set) with AlexNet.

Downstream Tasks:
- classification
- detection
- segmentation

Downstream Data:
Finetune on labelled data from Pascal VOC 2007.

Gidaris et al., **Unsupervised Representation Learning by Predicting Image Rotations**, ICLR, 2018

Generative Pretext Tasks

Taxonomy of Pretext tasks

Generative

Pretext task aims for reconstruction of either the entire sample or parts of the sample to inform learning.

Reconstruction

- VAE
- GAN
- Autoregressive
- Flow-based
- Diffusion

Self-Prediction

- Masked Prediction
  - Denoising AE
  - Context AE
  - Image Completion
- Channel Shuffling
  - Colorization
  - Split-brain

Discriminative

Pretext task aims at discrimination either between innate relationship or contrast between sample pairs to learning.

Innate Relationship

- Patch Positioning
  - Relative Position
  - Jigsaw
- Image Rotation
- Feature Counting
- Contrastive Predictive Coding

Contrastive

- Instance Discrim. / Augmentation
  - Feature Discrimination
  - MoCo
  - SimCLR
- Clustering-based
- Positive-Contrast
  - BOYT

Source: OpenAI - SSL Tutorial, NeurIPS, 2021

Generative Pretext Task

Predict missing pixels (inpainting)

Deepak Pathak Philipp Krahenbühl, Jeff Donahue Trevor Darrell Alexei A. Efros, **Context Encoders: Feature Learning by Inpainting**, CVPR, 2016

Generative Pretext Task

Predict missing pixels (inpainting)

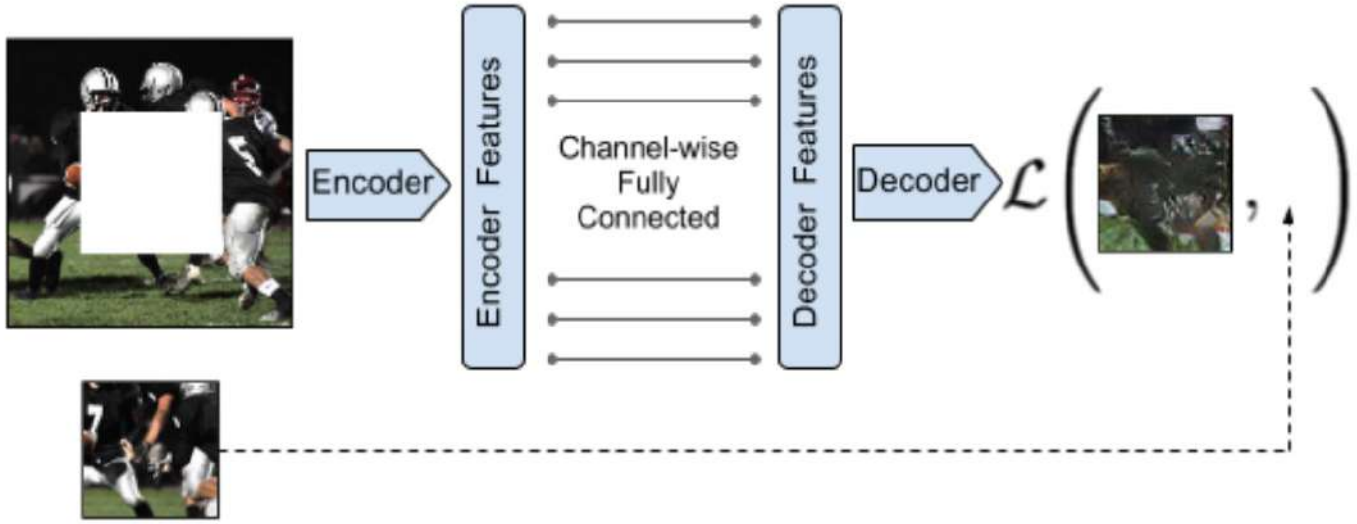
(a) Input context (b) Human artist (c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

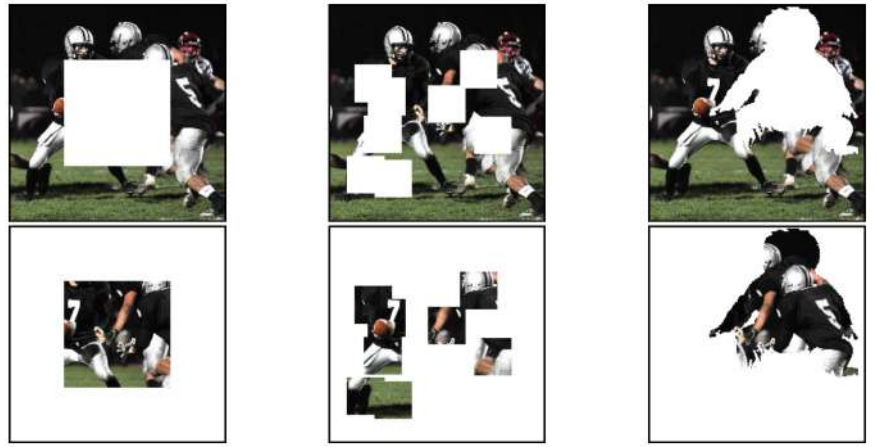
Idea: In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s).

Deepak Pathak Philipp Krahenbühl, Jeff Donahue Trevor Darrell Alexei A. Efros, **Context Encoders: Feature Learning by Inpainting**, CVPR, 2016

Generative Pretext Task

Predict missing pixels (inpainting)





(a) Central region (b) Random block (c) Random region

Architecture: Generative model based on an encoder-decoder architecture. The encoder is based on AlexNet Conv layers and the decoder is based on de-conv layers. No latent bottleneck, but rather a layer-wise fully connected bottleneck.

Deepak Pathak Philipp Krahenbühl, Jeff Donahue Trevor Darrell Alexei A. Efros, **Context Encoders: Feature Learning by Inpainting**, CVPR, 2016

Generative Pretext Task

Predict missing pixels (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%

SSL Data:
Self-supervised learning on ImageNet (entire training set) with AlexNet.

Downstream Data:
Finetune on labelled data from Pascal VOC 2007

Downstream Tasks:
- classification
- detection
- segmentation

Deepak Pathak Philipp Krahenbühl, Jeff Donahue Trevor Darrell Alexei A. Efros, **Context Encoders: Feature Learning by Inpainting**, CVPR, 2016

Generative Pretext Task

Image Coloring

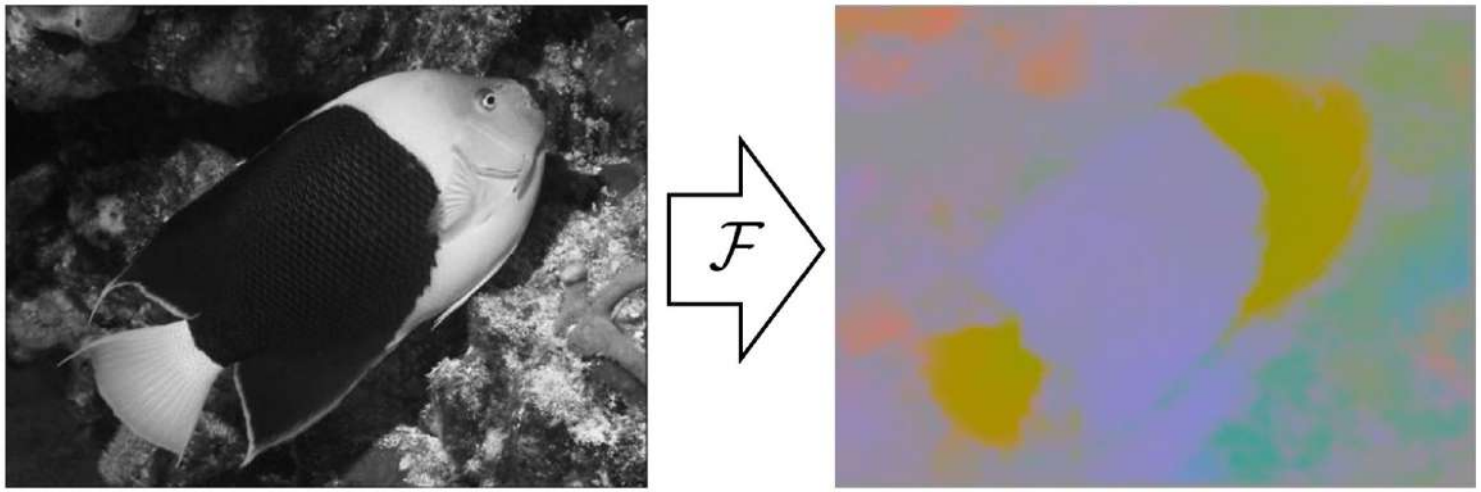
Richard Zhang, Phillip Isola, Alexei A. Efros, **Colorful Image Colorization**, ECCV 2016

Generative Pretext Task

Image Coloring

L*a*b* color space:

L* = perceptual lightness
(orthogonal to)
a*b* = color model



Idea: Colorization can be a powerful pretext task for self-supervised feature learning, acting as a cross-channel encoder.

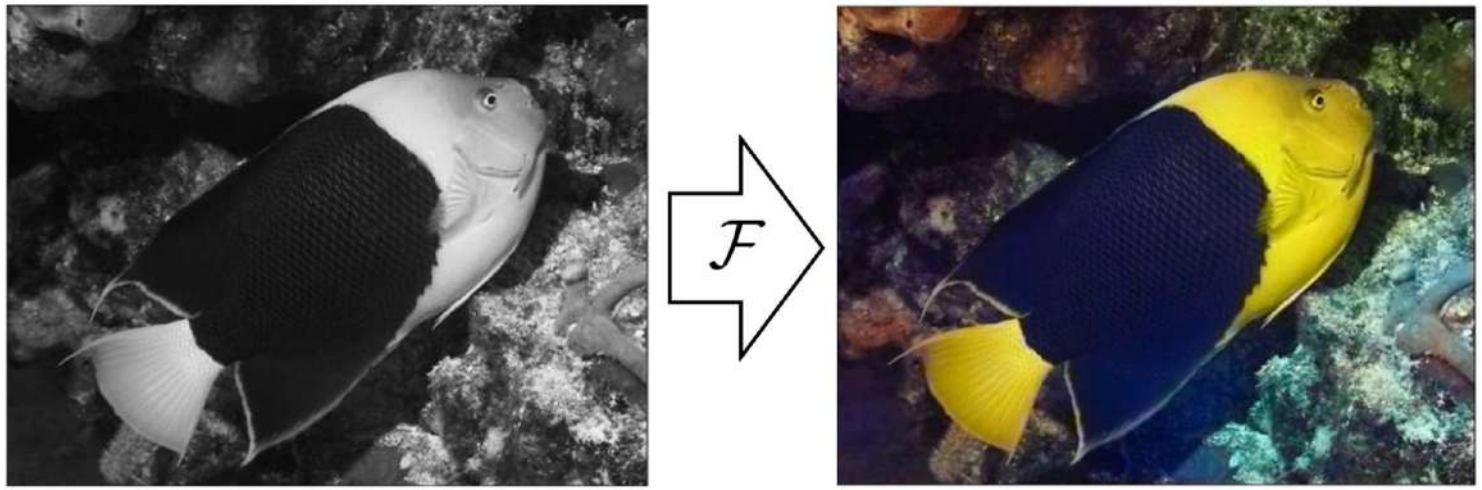
Richard Zhang, Phillip Isola, Alexei A. Efros, **Colorful Image Colorization**, ECCV 2016

Generative Pretext Task

Image Coloring

L*a*b* color space:

L* = perceptual lightness
(orthogonal to)
a*b* = color model



Grayscale image: L channel
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

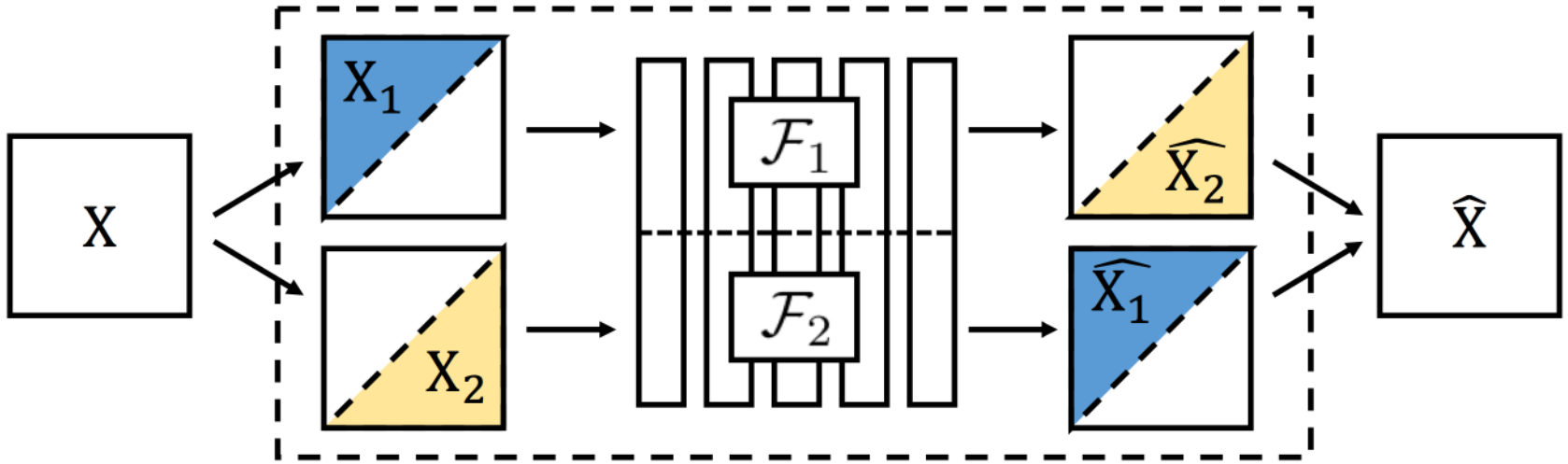
Concatenate (L,ab) channels
$(\mathbf{X}, \widehat{\mathbf{Y}})$

Idea: Colorization can be a powerful pretext task for self-supervised feature learning, acting as a cross-channel encoder.

Richard Zhang, Phillip Isola, Alexei A. Efros, **Colorful Image Colorization**, ECCV 2016

Generative Pretext Task

Image Coloring

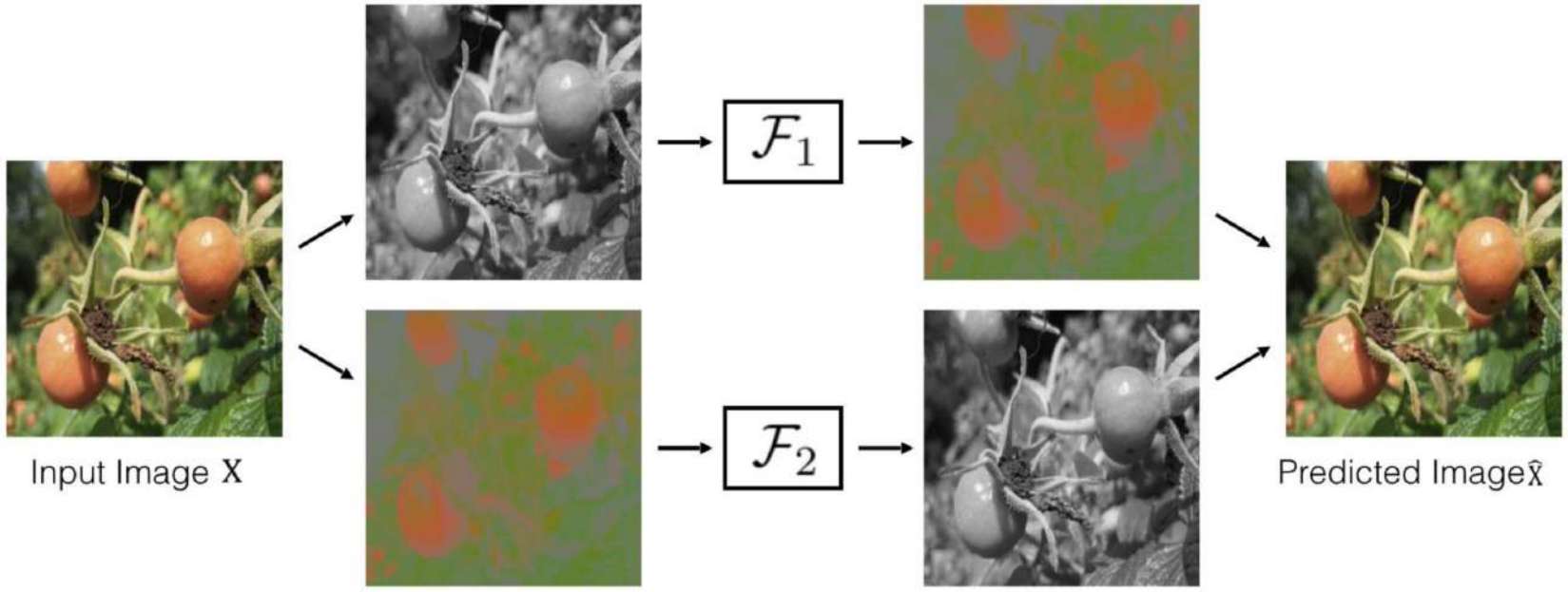


Idea: By forcing the network to solve cross-channel prediction tasks, we induce a representation within the network which transfers well to other, unseen tasks.

Richard Zhang, Phillip Isola, Alexei A. Efros, **Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction**, CVPR 2017

Generative Pretext Task

Image Coloring

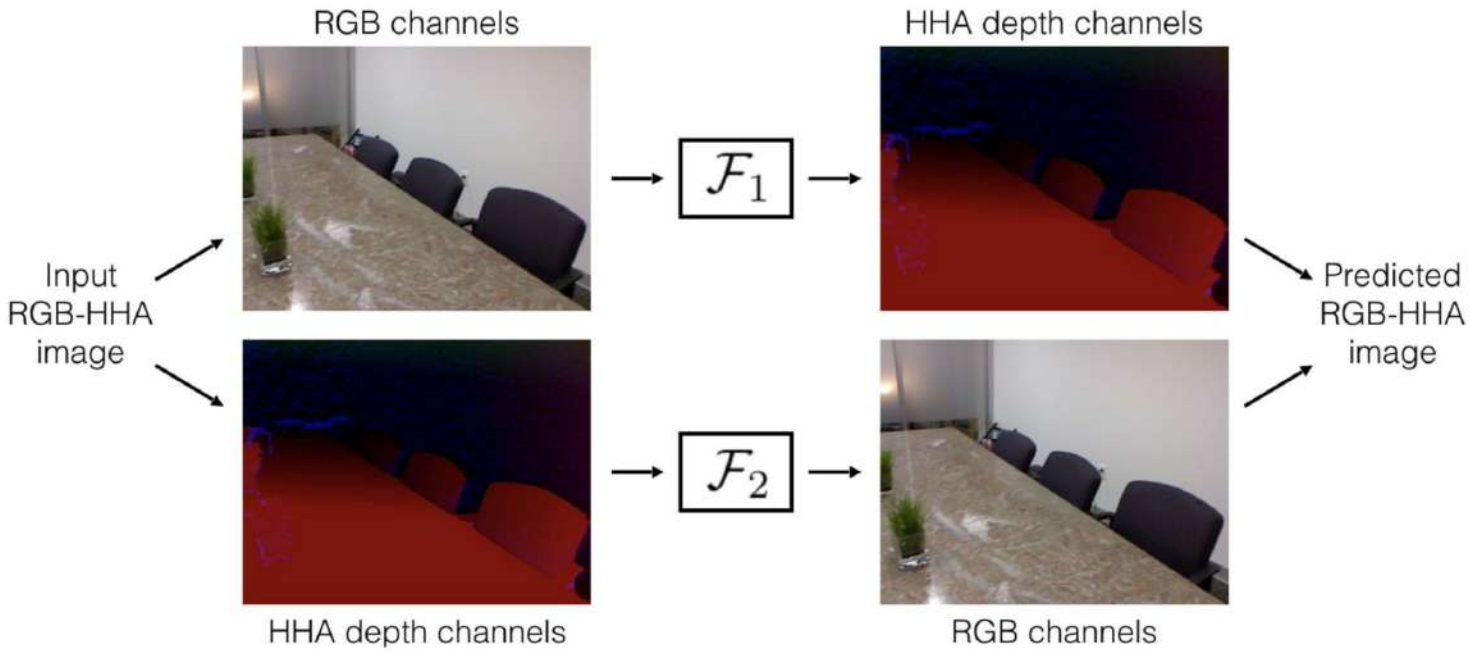


Idea: By forcing the network to solve cross-channel prediction tasks, we induce a representation within the network which transfers well to other, unseen tasks.

Richard Zhang, Phillip Isola, Alexei A. Efros, **Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction**, CVPR 2017

Generative Pretext Task

Image Coloring



Idea: By forcing the network to solve cross-channel prediction tasks, we induce a representation within the network which transfers well to other, unseen tasks.

Richard Zhang, Phillip Isola, Alexei A. Efros, **Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction**, CVPR 2017

Generative Pretext Task

Image Coloring

SSL Data:
Self-supervised learning on ImageNet (entire training set).

Tasks:
- classification
- detection
- segmentation

Downstream Data:
Concatenate features from F_1 & F_2

Finetune on labelled data from from Pascal VOC 2007

Task and Data Generalization on PASCAL VOC [12]

	Classification [25] (%mAP)			Detection [15] (%mAP)		Seg. [29] (%mIU)	
frozen layers		**conv5**	**none**		**none**		**none**
fine-tuned layers	**Ref**	**fc6-8**	**all**	**Ref**	**all**	**Ref**	**all**
ImageNet labels [26]	[49]	78.9	79.9	[25]	56.8	[29]	48.0
Gaussian	[35]	–	53.3	[35]	43.4	[35]	19.8
Autoencoder	[9]	16.0	53.8	[35]	41.9	[35]	25.2
Krähenbühl et al. [25]	[9]	39.2	56.6	[25]	45.6	[9]	32.6
Jayaraman & Grauman [23]	–	–	–	[23]	41.7	–	–
Agrawal et al. [1]	[25]	–	52.9	[25]	41.8	–	–
Agrawal et al. [1]†	[9]	31.0	54.2	[25]	43.9	–	–
Wang & Gupta [46]	[25]	–	62.8	[25]	47.4	–	–
Wang & Gupta [46]†	[25]	–	63.1	[25]	47.2	–	–
Doersch et al. [8]	[25]	–	55.3	[25]	46.6	–	–
Doersch et al. [8]†	[9]	55.1	65.3	[25]	51.1	–	–
Pathak et al. [35]	[35]	–	56.5	[35]	44.5	[35]	29.7
Donahue et al. [9]†	[9]	52.3	60.1	[9]	46.9	[9]	35.2
Misra et al. [30]	–	–	–	[30]	42.4	–	–
Owens et al. [33]	▷	54.6	54.4	[33]	44.0	–	–
Owens et al. [33]†	▷	52.3	61.3	–	–	–	–
Zhang et al. [49]†	[49]	61.5	65.9	[49]	46.9	[49]	35.6
Larsson et al. [28]◇	[28]	–	65.9	–	–	[28]	**38.4**
Pathak et al. [34]◇	[34]	–	61.0	[34]	**52.2**	–	–
Split-Brain Auto (cl,cl)†	▷	**63.0**	**67.1**	▷	46.7	▷	36.0

Richard Zhang, Phillip Isola, Alexei A. Efros, **Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction**, CVPR 2017

Generative Pretext Task

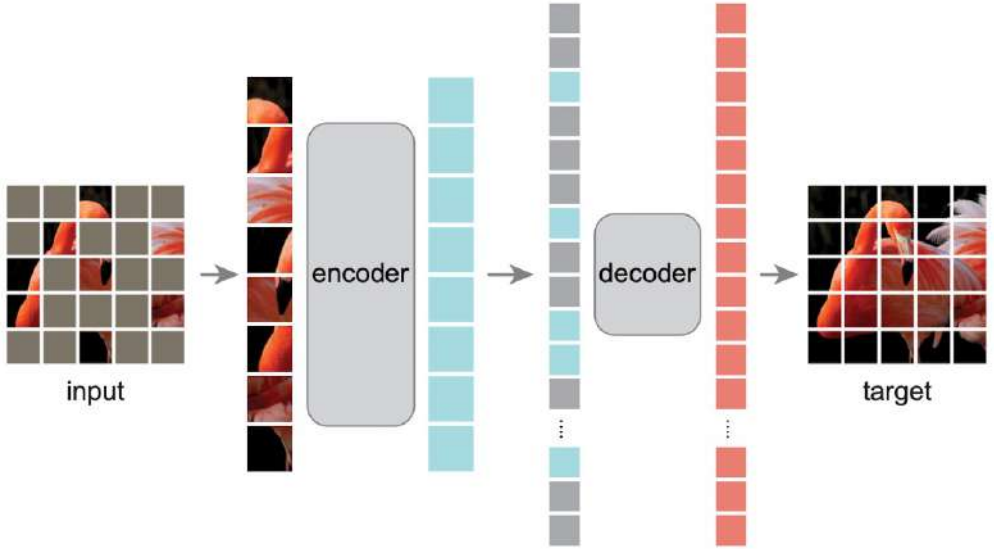
Image Coloring

Evaluation: Our goal is not necessarily to recover the actual ground truth color,
but rather to produce a plausible colorization that could potentially fool a human observer.

Richard Zhang, Phillip Isola, Alexei A. Efros, **Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction**, CVPR 2017

Generative Pretext Task

Masked Autoencoder (Masked Image Modelling)



Idea: masking random patches of the input image and reconstruct the missing pixels is much more scalable than inpainting -> An asymmetric encoder-decoder architecture i.e., an encoder that operates only on the visible subset of patches (without mask tokens), a lightweight decoder that reconstructs the original image from the latent representation and mask tokens.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick; **Masked Autoencoders Are Scalable Vision Learners**, CVPR, 2022

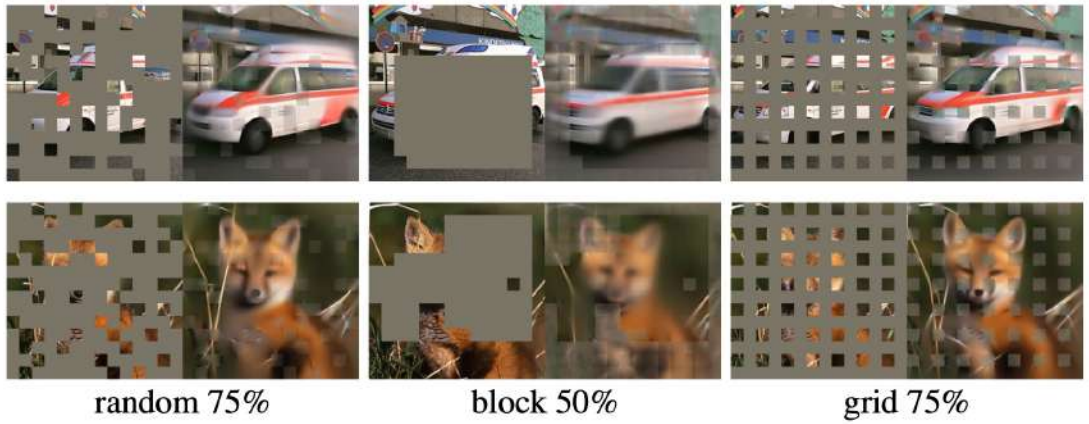
Generative Pretext Task

Masked Autoencoder (Masked Image Modelling)

SSL Data:
Self-supervised learning on ImageNet & COCO

Tasks:
- classification, detection, segmentation, …

Downstream Data:
- multiple evaluation protocols



method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

Table 3. **Comparisons with previous results on ImageNet-1K**. The pre-training data is the ImageNet-1K training set (except the tokenizer in BEiT was pre-trained on 250M DALLE data [50]). All self-supervised methods are evaluated by end-to-end fine-tuning. The ViT models are B/16, L/16, H/14 [16]. The best for each column is underlined. All results are on an image size of 224, except for ViT-H with an extra result on 448. Here our MAE reconstructs normalized pixels and is pre-trained for 1600 epochs.

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick; **Masked Autoencoders Are Scalable Vision Learners**, CVPR, 2022

Are the presented pretext task enough?

Challenge: (a) inventing individual pretext tasks for specific datasets is tedious
(b) run the risk that the learned representations may not be general.

image completion

rotation prediction

jigsaw puzzle

image coloring

Question:
Can we come up with a more general concept of a pretext task and exploit it for SSL?

Contrastive Learning

Taxonomy of Pretext tasks

Generative

Pretext task aims for reconstruction of either the entire sample or parts of the sample to inform learning.

Reconstruction

- VAE
- GAN
- Autoregressive
- Flow-based
- Diffusion

Self-Prediction

- Masked Prediction
  - Denoising AE
  - Context AE
  - Image Completion
- Channel Shuffling
  - Colorization
  - Split-brain

Discriminative

Pretext task aims at discrimination either between innate relationship or contrast between sample pairs to learning.

Innate Relationship

- Patch Positioning
  - Relative Position
  - Jigsaw
- Image Rotation
- Feature Counting
- Contrastive Predictive Coding

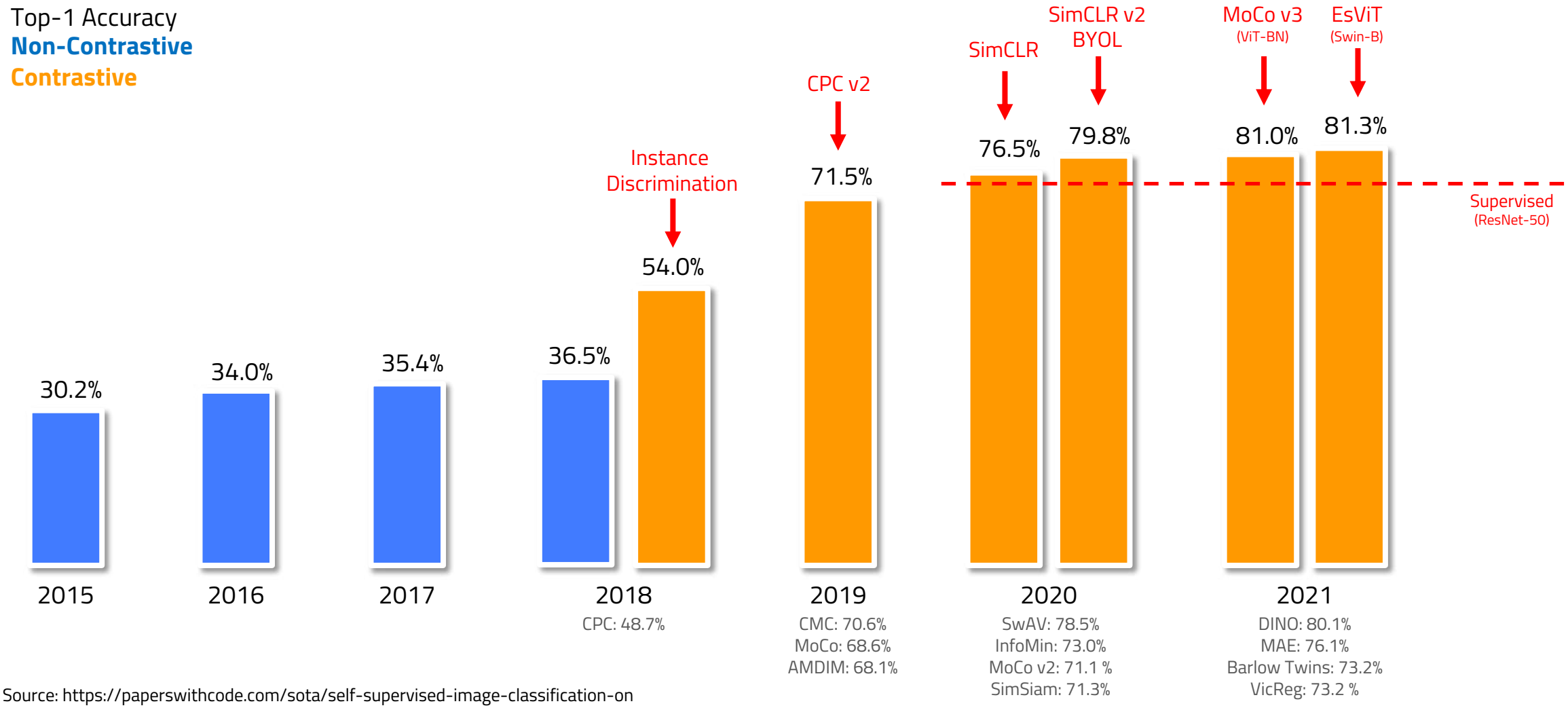
Contrastive

- Instance Discrim. / Augmentation
  - Feature Discrimination
  - MoCo
  - SimCLR
- Clustering-based
- Positive-Contrast
  - BOYT

Source: OpenAI - SSL Tutorial, NeurIPS, 2021

Motivation for Contrastive Learning

ImageNet Linear Benchmark



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

Idea

Latent Space

Representation Learning





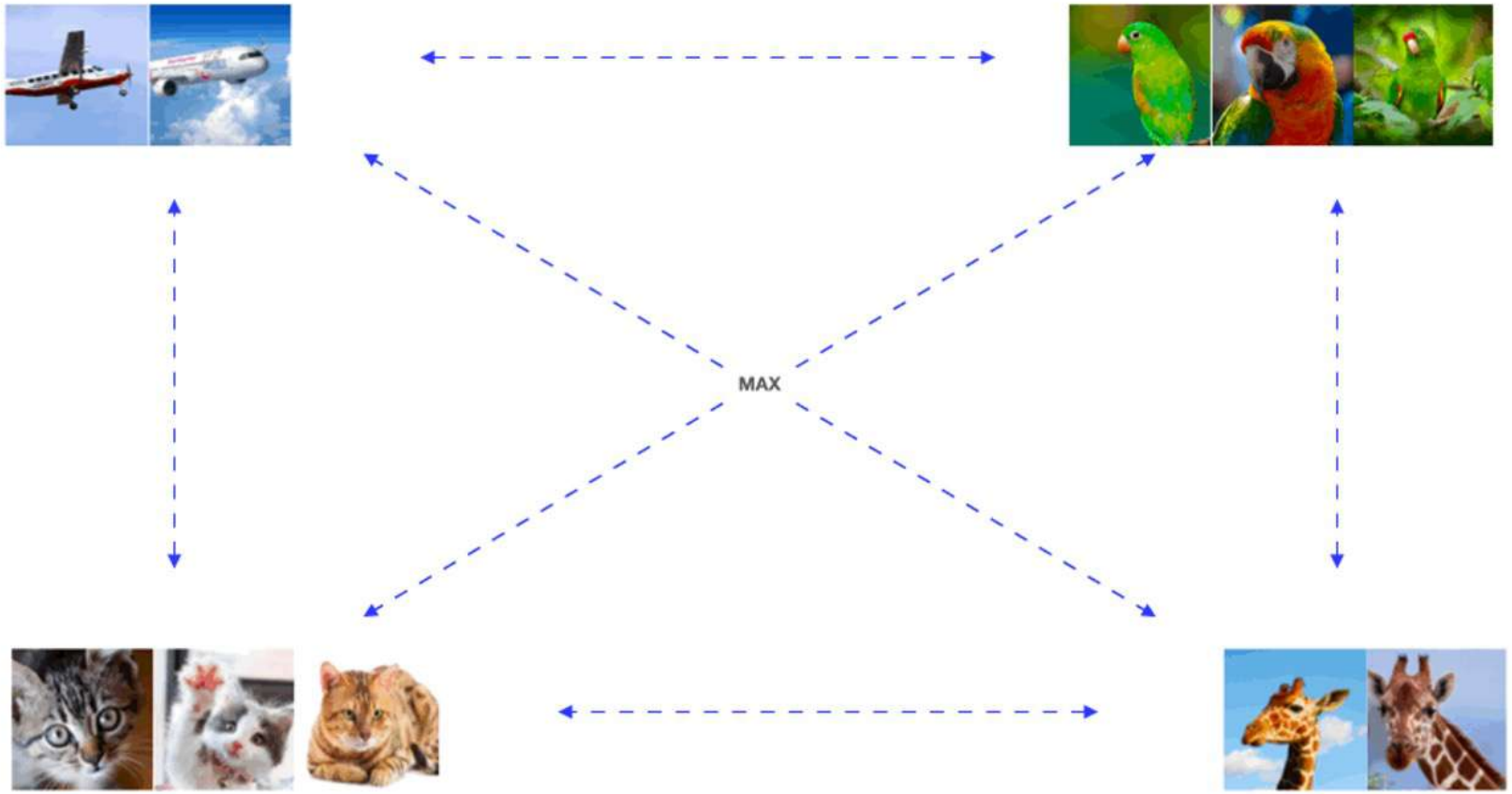





Images: https://towardsdatascience.com/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

Idea

Latent Space



Representation Learning

Ideally, we want to train a representation, that is able to discriminative between classes

Goal:

- min. distance between samples of the same class
- max. distance between samples of different classes

Until here this has nothing to do with Self-Supervise Learning

Question:
How can we do this without class labels?

Images: https://towardsdatascience.com/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

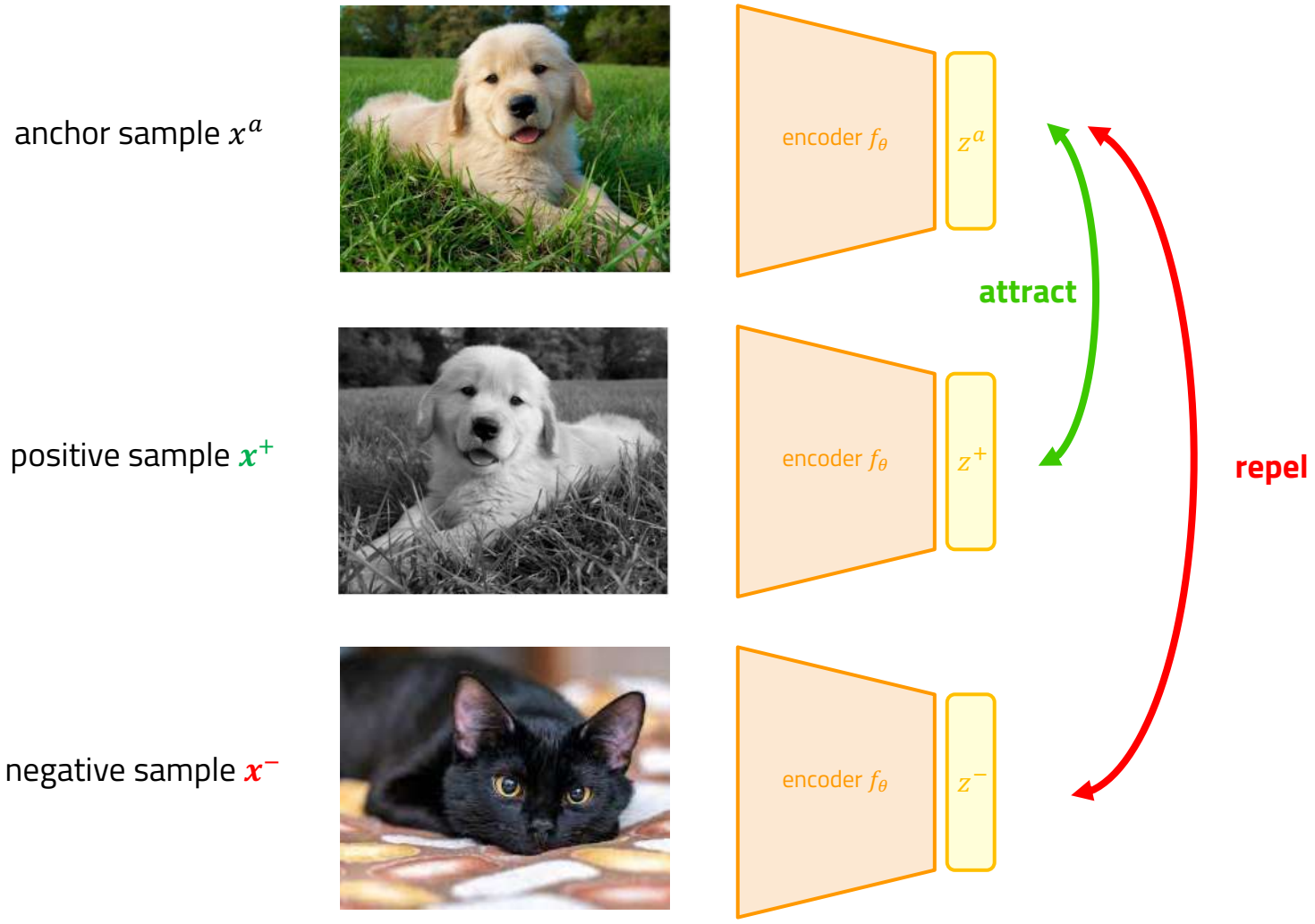
We use “contrast” to replace class labels

Question: What do you think is more difficult? To find good x^+ or x^-

anchor sample x^a | positive sample x^+ | negative sample x^- | negative sample x^-

Images: https://towardsdatascience.com/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

Basic Setup



Setup

Contrast is being defined in latent space i.e., the **embedding** vector of the image after a forward-pass through an (the same) **encoder** f_θ.

Since we have now vectors representing sample we have to quantify "attract" and "repel" and include this into a loss.

Design Decisions:

1. Select encoder
2. Select similarity / distance (metric)
3. Define a proper loss function

Please note:
There is a difference between "similarity", "distance" or "metric"

Transformations to generate x^+

Augmentation

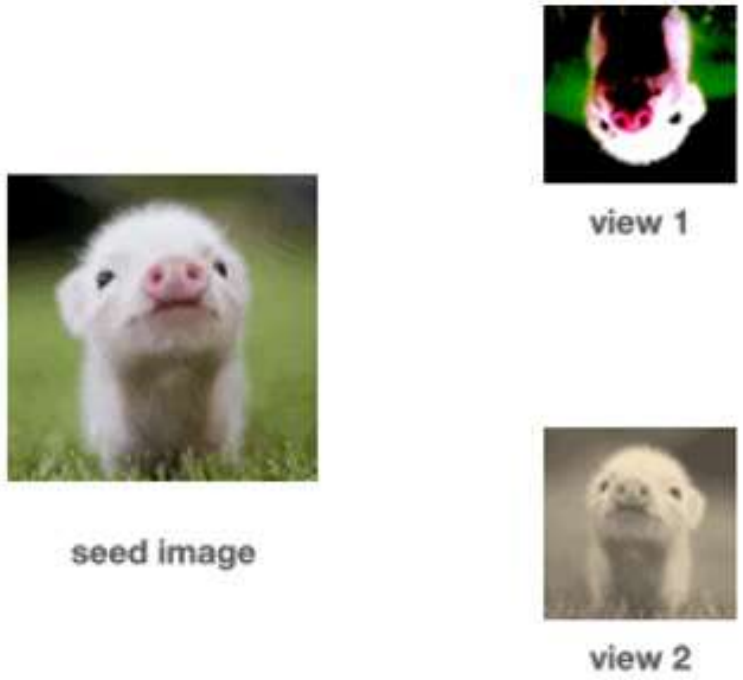
seed image

- same class + similar appearance
- randomly select augmentation
- **used in:** AMDIM, SimCLR, MOCO

Images: https://towardsdatascience.com/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

Transformations to generate x^+

Augmentation



- same class + similar appearance
- randomly select augmentation
- **used in:** AMDIM, SimCLR, MOCO

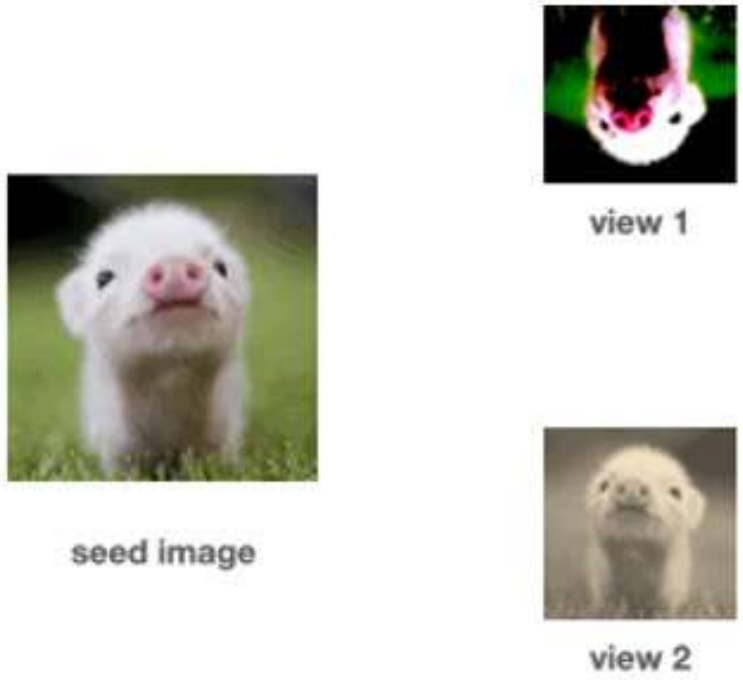
Overlaps

- same class + patches are partially similar
- sliding window in reading order
- **used in:** CPC

Images: https://towardsdatascience.com/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

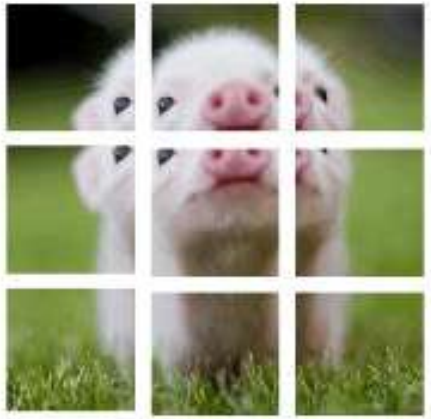
Transformations to generate x^+

Augmentation



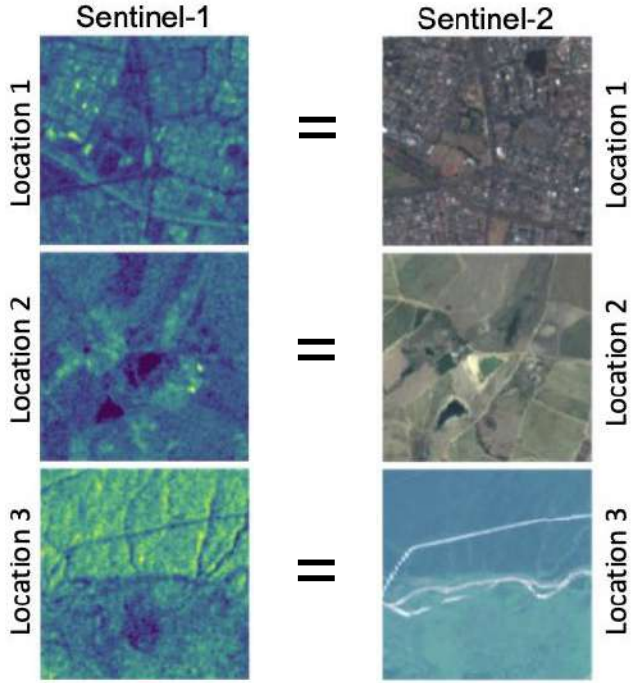
- same class + similar appearance
- randomly select augmentation
- **used in:** AMDIM, SimCLR, MOCO

Overlaps



- same class + patches are partially similar
- sliding window in reading order
- **used in:** CPC

Multi-view (Multi-modal)

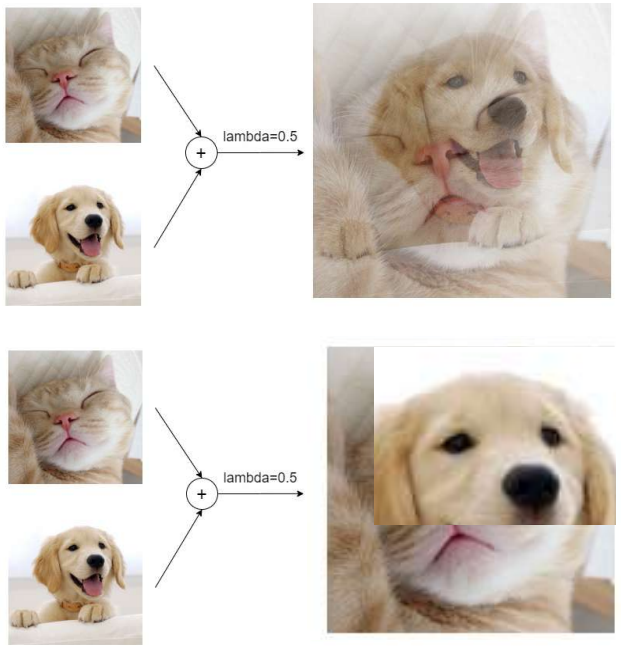


- same class = same location + diff sensor
- multi-modal fusion
- **used in:** CVPR Earth Vision: [our work]

Images: https://towardsdatascience.com/a-framework-for-contrastive-self-supervised-learning-and-designing-a-new-approach-3caab5d29619

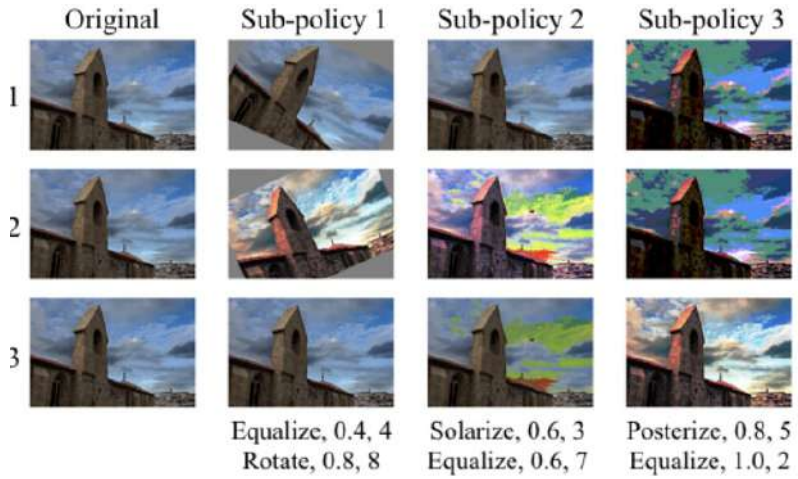
Transformations to generate x^+

Image Mixture



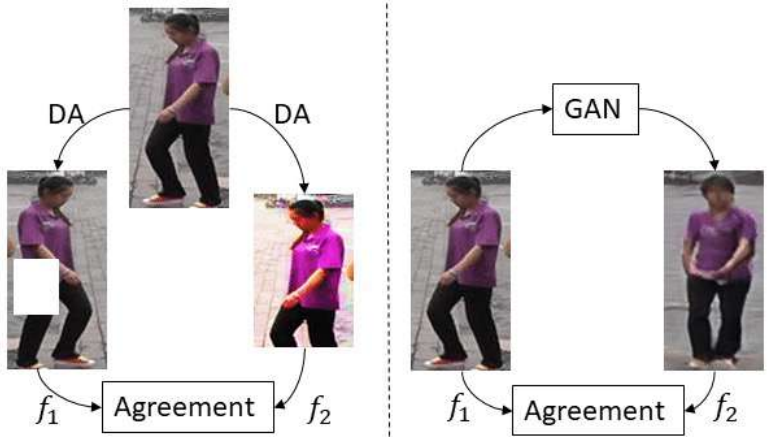
- same class + distortion with other class
- pixel-wise linear overlay of two images
- **used in:** Mixup, Cutmix,

Data-adapted



- same class + optimized similar apparency
- learning augmentation policy from data
- **used in:** AutoAugment, SelfAugment, UDA

Generative



- same class + generated transformation
- conditioned GAN given anchor image
- **used in:** GCL

(Hard) Negatives x^-

Hard Negative = different semantic label but close to embedding

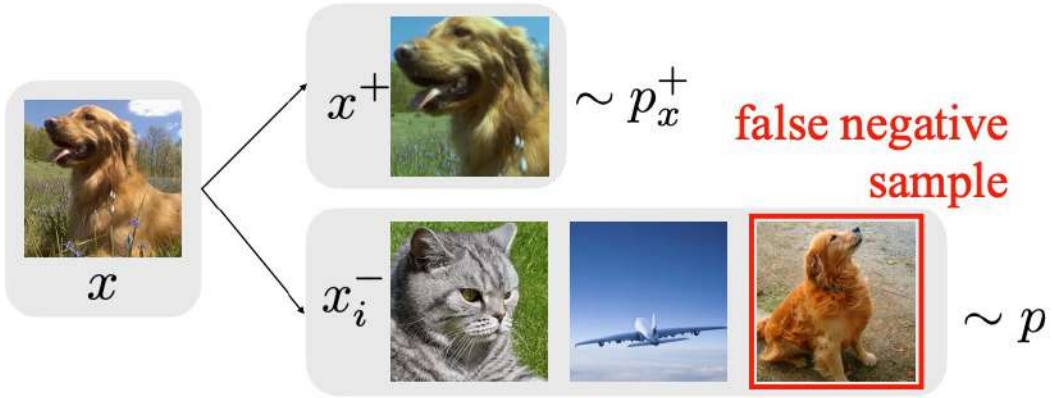


Figure 1: **"Sampling bias":** The common practice of drawing negative examples x_i^- from the data distribution $p(x)$ may result in x_i^- that are actually similar to x.

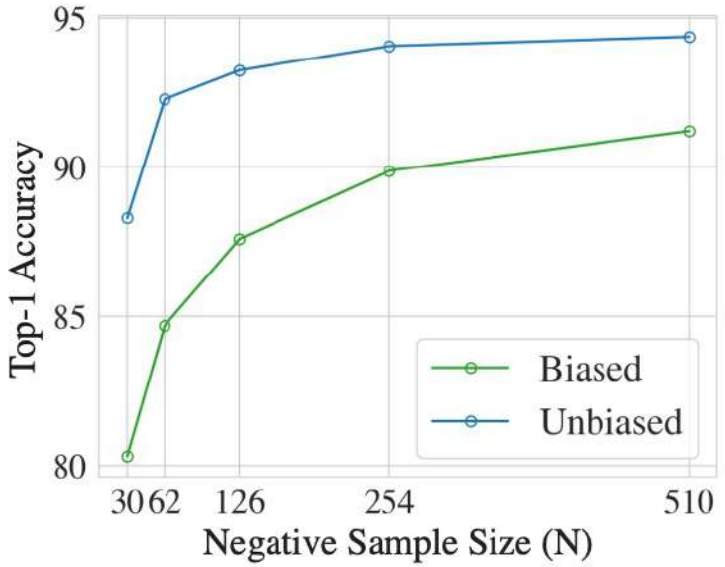


Figure 2: **Sampling bias leads to performance drop:** Results on CIFAR-10 for drawing x_i^- from $p(x)$ (biased) and from data with different labels, i.e., truly semantically different data (unbiased).

Source: A Debiased Contrastive Learning, NeurIPS, 2020

Loss function for our Objective

Overview

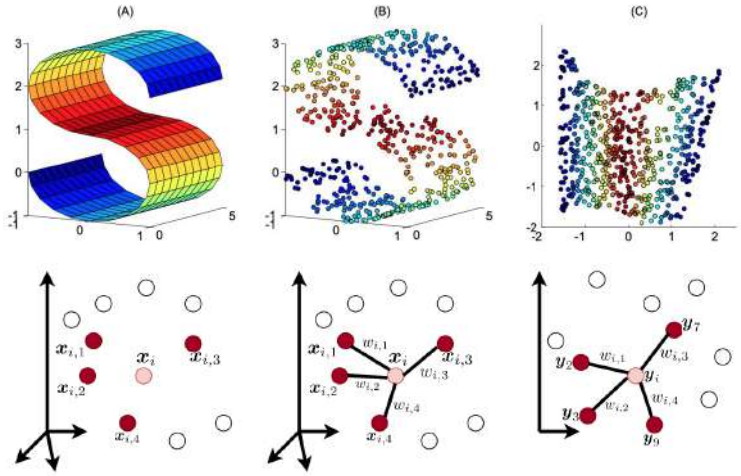
Historical

Precursor of this type of learning objective comes from two disciplines:

- **Multiple Instance Learning**
- **Metric Learning**

with ideas inspired by:

- Multidimensional scaling (MDS)
  [MDS; Cox et al. 1994]
- Locally linear embedding (LLE)
  [LLE; Roweis et al. 2000]



General

Contrastive learning loss functions can be applied to both setups:

- supervised
- unsupervised

Vector similarity is at the core of contrastive learning loss functions.

Early loss functions have only considered one positive and one negative sample to compute loss.

Recently, proposed training objectives include multiple positive and negative pairs in one batch

Loss functions

- Contrastive loss
  [Chopra et al. 2005]
- Triplet loss
  [Schroff et al. 2015; FaceNet]
- Lifted structured loss
  [Song et al. 2015]
- N-pair loss
  [Sohn 2016]
- InfoNCE loss
  [van den Oord, et al. 2018]
- NT-Xent loss
  [Chen et al., 2020]

Loss calculation is done within the mini batch
i.e., batch size is a limiting factor for sample size
as it related directly to the GPU or TPU memory!

Contrastive Loss

Notation

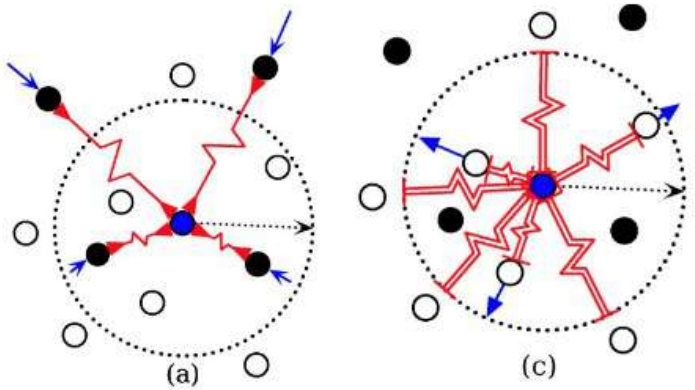
Given two labelled data pairs $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$:

max margin defined as penalty if diff. label samples are to close

$$\mathcal{L}_{\text{cont}}(\mathbf{x}_i, \mathbf{x}_j, \theta) = \mathbb{1}[y_i = y_j] \underbrace{\|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2^2}_{\text{attract}} + \mathbb{1}[y_i \neq y_j] \max(0, \epsilon - \underbrace{\|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2}_{\text{repel}})^2$$

- **Contrastive loss** [Chopra et al. 2005]
- Requires labelled data -> ground truth labels
- Use encoder f_θ to extract embedding z
- Encodes data into an embedding vector such that examples from the same class have similar embeddings and samples from different classes have different ones.

Idea



Triplet Loss

Notation

Given a triplet input $(\mathbf{x}^a, \mathbf{x}^+, \mathbf{x}^-)$:

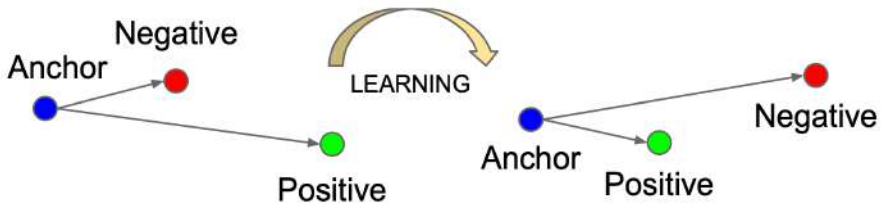
min margin defined between positive and negative samples

$$\mathcal{L}_{\text{triplet}}(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = \sum_{\mathbf{x} \in \mathcal{X}} \max\left(0, \|f(\mathbf{x}) - f(\mathbf{x}^+)\|_2^2 - \|f(\mathbf{x}) - f(\mathbf{x}^-)\|_2^2 + \epsilon\right)$$

attract — repel

- **Triplet Loss** [Schroff et al. 2015]
- Augmentation defines triplet with $\mathbf{x}^+$
- Use encoder f_θ to extract embedding z
- Learns to minimize the distance between the anchor $\mathbf{x}^a$ and positive $\mathbf{x}^+$ and maximize the distance between the anchor x and negative $\mathbf{x}^-$ at the same time..

Idea



N-pair Loss

Notation

Given an anchor, one positive and N-1 negative samples $\{\mathbf{x}^{a}, \mathbf{x}^{+}, \mathbf{x}_1^{-}, \dots, \mathbf{x}_{N-1}^{-}\}$:

$$\mathcal{L}_{\text{N-pair}}(\mathbf{x}, \mathbf{x}^{+}, \{\mathbf{x}_i^{-}\}_{i=1}^{N-1}) = \log\left(1 + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^{\top} f(\mathbf{x}_i^{-}) - f(\mathbf{x})^{\top} f(\mathbf{x}^{+}))\right)$$

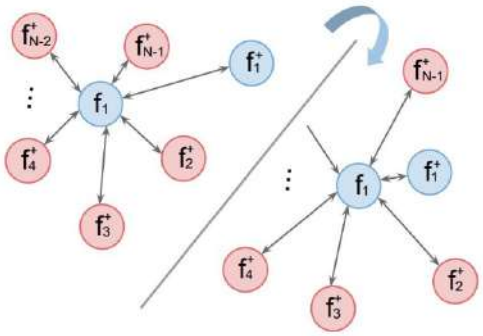
$$= -\log \frac{\overbrace{\exp(f(\mathbf{x})^{\top} f(\mathbf{x}^{+}))}^{\text{attract}}}{\underbrace{\exp(f(\mathbf{x})^{\top} f(\mathbf{x}^{+}))}_{\text{attract}} + \underbrace{\sum_{i=1}^{N-1} \exp(f(\mathbf{x})^{\top} f(\mathbf{x}_i^{-}))}_{\text{repel}}}$$

could be interpreted as the probability of the augmented sample being most similar to the anchor samples as compared to all negative samples

- **N-pair loss** [Sohn 2016]
- Use encoder f_θ to extract embedding z

- Generalizes triplet loss to include multiple negative samples
- reminds us of a softmax alike loss
- uses dot product for similarity computation not a distance

Illustration



Norm. Temperature-Scaled Cross-Entropy Loss

Notation

Given an anchor, one positive and N-1 negative samples $\{\mathbf{x}^{a}, \mathbf{x}^{+}, \mathbf{x}_1^{-}, \dots, \mathbf{x}_{N-1}^{-}\}$:

attract

$$\mathcal{L}^{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

temperature term

repel

- **NT-Xent Loss /** often called **InfoNCE** for simplicity
- **InfoNCE** [van den Oord, 2018] & **Soft-nearest neighbors Loss** [Salakhutdinov & Hinton 2007, Frosst et al. 2019]
- used in SimCLR
- Cross entropy loss for a N-way softmax classifier, i.e., learn to find the positive sample from the N samples

Specifics

NT-Xent Loss combines previous ideas:

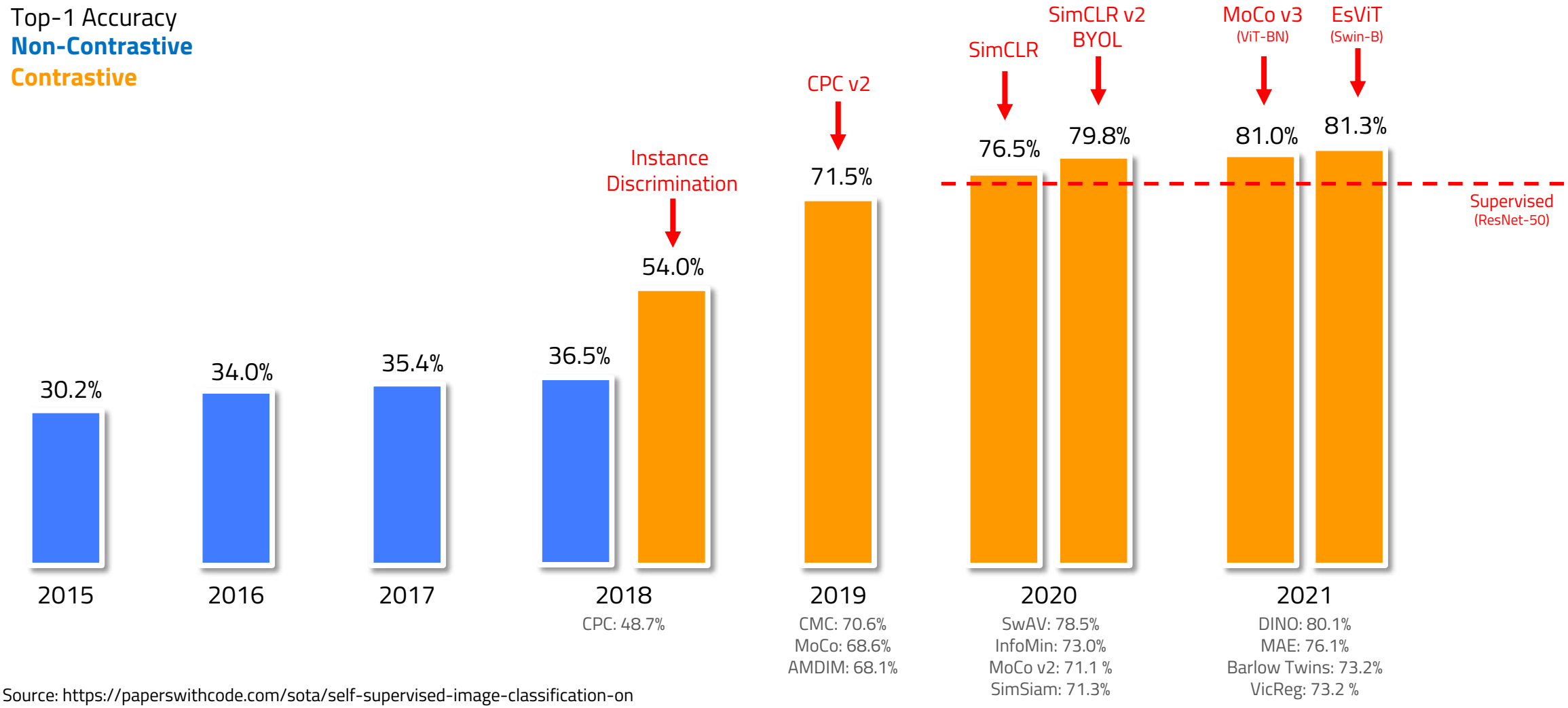
- uses a projection head g_θ to extract z
- uses cosine similarity sim()
- uses a temperature term τ

Temperature defines how concentrated the features are in representation space. When low, the loss is dominated by the small distances and widely separated representations cannot contribute much and become irrelevant.

Contrastive Learning Frameworks

Motivation for Contrastive Learning

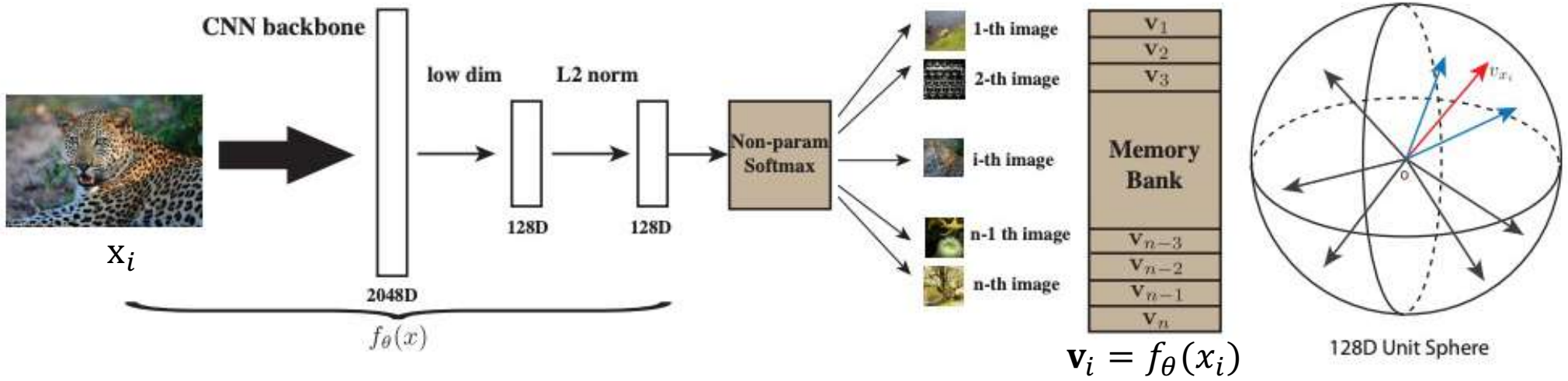
ImageNet Linear Benchmark



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

Instance Discrimination

Approach



- **Idea:** each image belongs to a unique class
- Each class has only one instance ⇒ $\mathbf{v}_i$ is the class prototype
- Computing $P(i|\mathbf{v})$ is **inefficient** ⇒ it requires all $\mathbf{v}_i = f_\theta(x_i)$ and $\mathbf{v}_j^T\mathbf{v}$
- **Solution:** use memory bank
  ⇒ store all $\mathbf{v}_i$ and update them for mini-batch
  ⇒ to stabilize training use momentum-update for memory bank

Non-parametric softmax classifier + NCE loss

$$P(i|\mathbf{v}) = \frac{\exp(\mathbf{v}_i^\top \mathbf{v}/\tau)}{\sum_{j=1}^{n} \exp(\mathbf{v}_j^\top \mathbf{v}/\tau)}$$

$$P(D=1|i,\mathbf{v}) = P(i|\mathbf{v}) = \frac{\exp(\mathbf{v}_i^\top \mathbf{v})}{\exp(\mathbf{v}_i^\top \mathbf{v}) + \sum_{k=1}^{m} \exp(\mathbf{v}_{j_k}^\top \mathbf{v})}$$

$$\mathcal{L}_{\text{NCE}} = -\underbrace{\mathbb{E}_{P_d}[\log P(D=1|i,\mathbf{v})]}_{\text{data distribution}} - \underbrace{m\mathbb{E}_{P_n}[\log P(D=0|i,\mathbf{v}')]}_{\text{noise distribution = uniform}}$$

Source: Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination, CVPR 2018

Instance Discrimination

Results (linear)

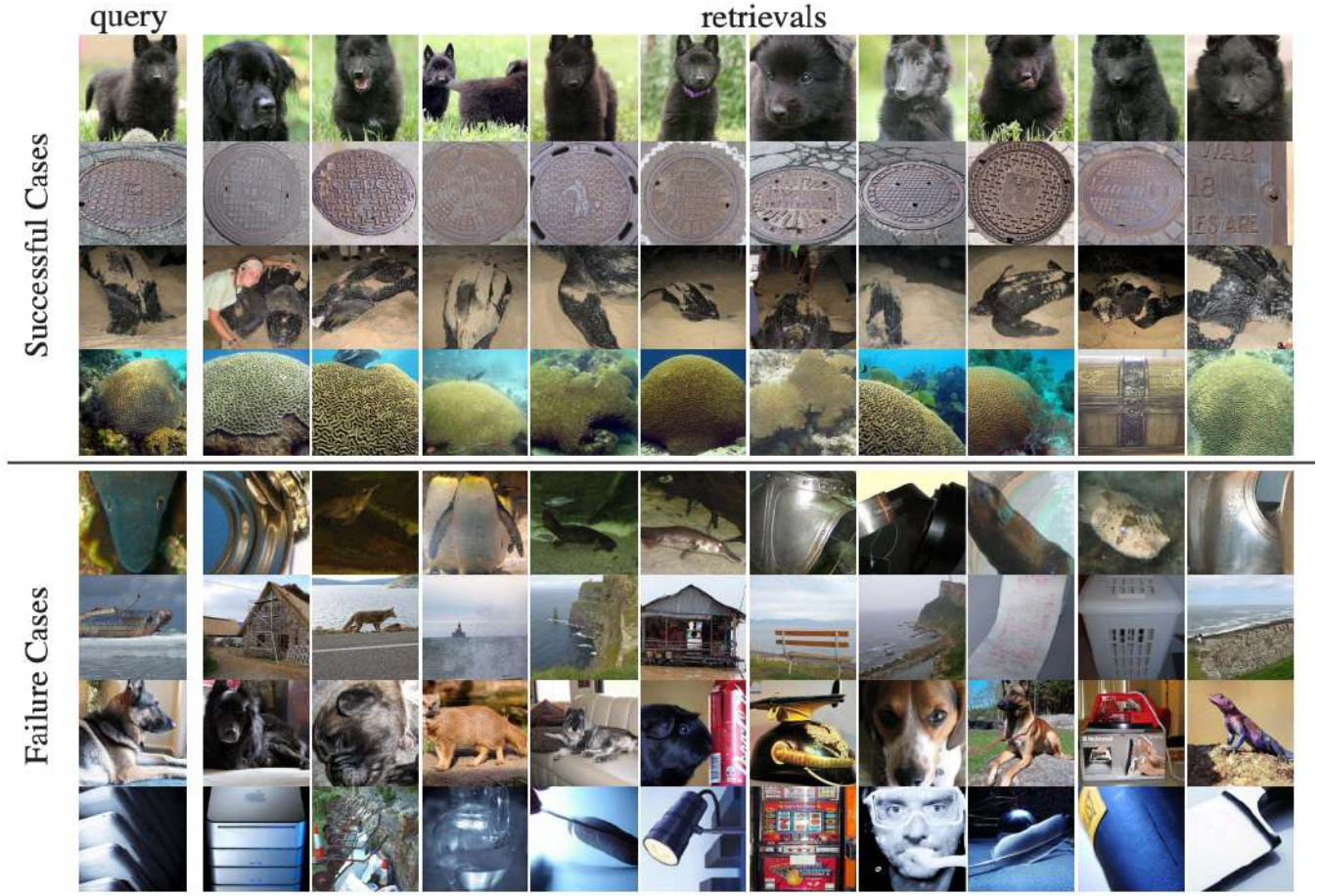
Image Classification Accuracy on ImageNet							
method	conv1	conv2	conv3	conv4	conv5	kNN	#dim
Random	11.6	17.1	16.9	16.3	14.1	3.5	10K
Data-Init [16]	17.5	23.0	24.5	23.2	20.6	-	10K
Context [2]	16.2	23.3	30.2	31.7	29.6	-	10K
Adversarial [4]	17.7	24.5	31.0	29.9	28.0	-	10K
Color [47]	13.1	24.8	31.0	32.6	31.8	-	10K
Jigsaw [27]	19.2	30.1	34.7	33.9	28.3	-	10K
Count [28]	18.0	30.6	34.3	32.5	25.7	-	10K
SplitBrain [48]	17.7	29.3	35.4	35.2	32.8	11.8	10K
Exemplar[3]			31.5			-	4.5K
Ours Alexnet	16.8	26.5	31.8	34.1	**35.6**	31.3	128
Ours VGG16	16.5	21.4	27.6	35.1	**39.2**	33.9	128
Ours Resnet18	16.0	19.9	29.8	39.0	**44.5**	**41.0**	128
Ours Resnet50	15.3	18.8	24.9	40.6	**54.0**	**46.5**	128

backbone makes a difference

Table 2: Top-1 classification accuracy on ImageNet.

embedding size	32	64	128	256
top-1 accuracy	34.0	38.8	41.0	40.1

saturation of embedding size



Source: Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination, CVPR 2018

Simple Framework for Contrastive Learning
(SimCLR)

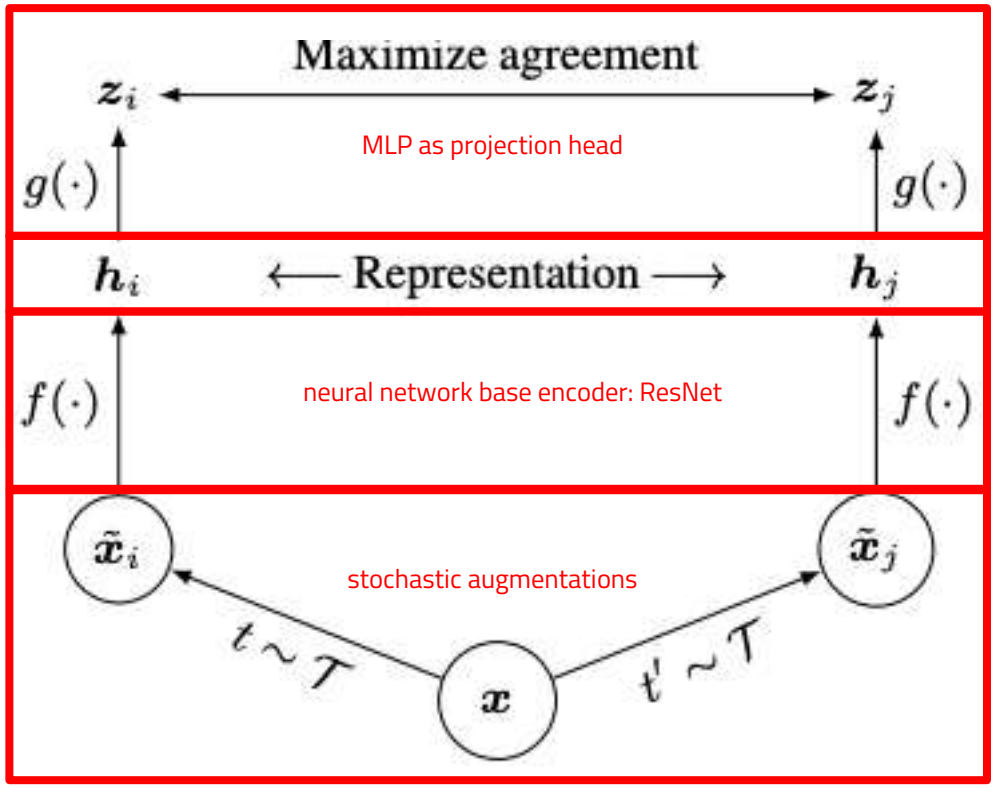
Approach

- **Idea:** decoupling representation space for downstream task and contrastive space, where loss is applied
- **Simple**: No memory bank, no momentum
- **Loss:** NT-Xent loss as objective function
- First contrastive learning approach comparable to supervised learning performance on the ImageNet linear classification protocol

- **Specifics**:
  ⇒ non-linear projection head:
  instead of computing loss on $\boldsymbol{z}_i = f_\theta(\tilde{x}_i)$, it uses $\boldsymbol{z}_i = g_\varphi(f_\theta(\tilde{x}_i))$

  ⇒ strong augmentations:
  multiple random augmentations used together

  ⇒ large batch size and long training:
  incredibly large batch-size of 4096 or 8192 trained over 1000 epoches
  (128 TPU v3 cores, it takes ~1.5 hours to train ResNet-50 with a batch size of 4096 for 100 epochs.)



Source: A Simple Framework for Contrastive Learning of Visual Representations, ICML, 2020

Simple Framework for Contrastive Learning
(SimCLR)

Augmentations

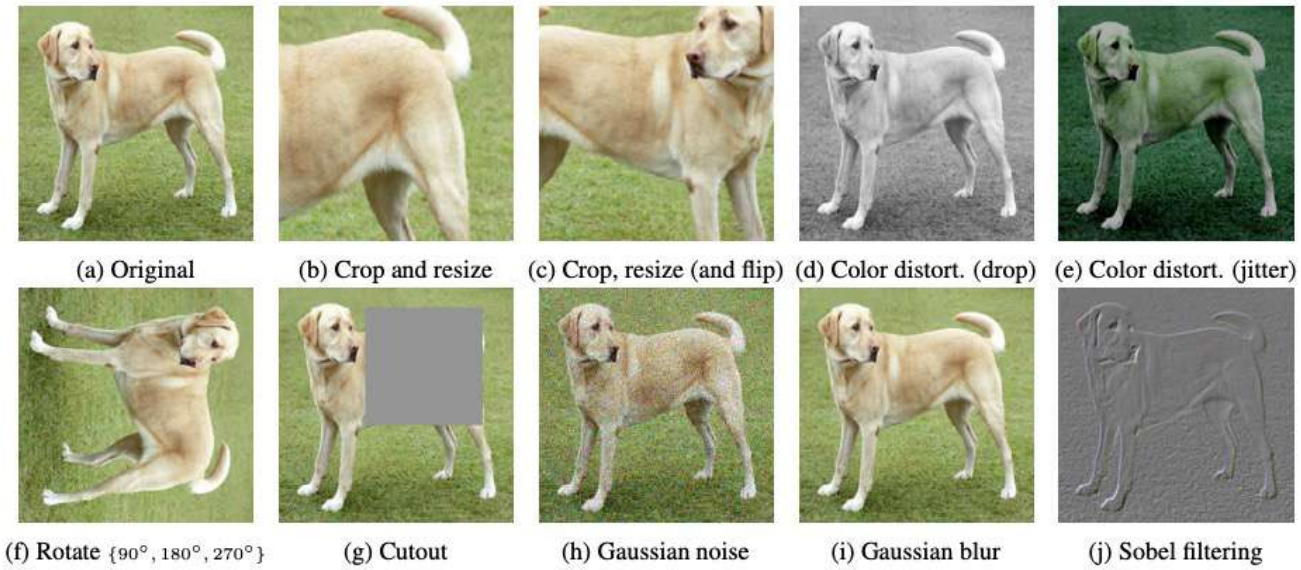


Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

Source: A Simple Framework for Contrastive Learning of Visual Representations, NeurIPS, 2020

Simple Framework for Contrastive Learning
(SimCLR)

Results (linear)

- $g_\varphi(\cdot)$ can make contrastive learning invariant to some information (e.g., color)
- this improves contrastive learning but might be lost later in the representation $\mathbf{h}$
- using nonlinear $g_\varphi(\cdot)$ allows representation $\mathbf{h}$ to preserve such information

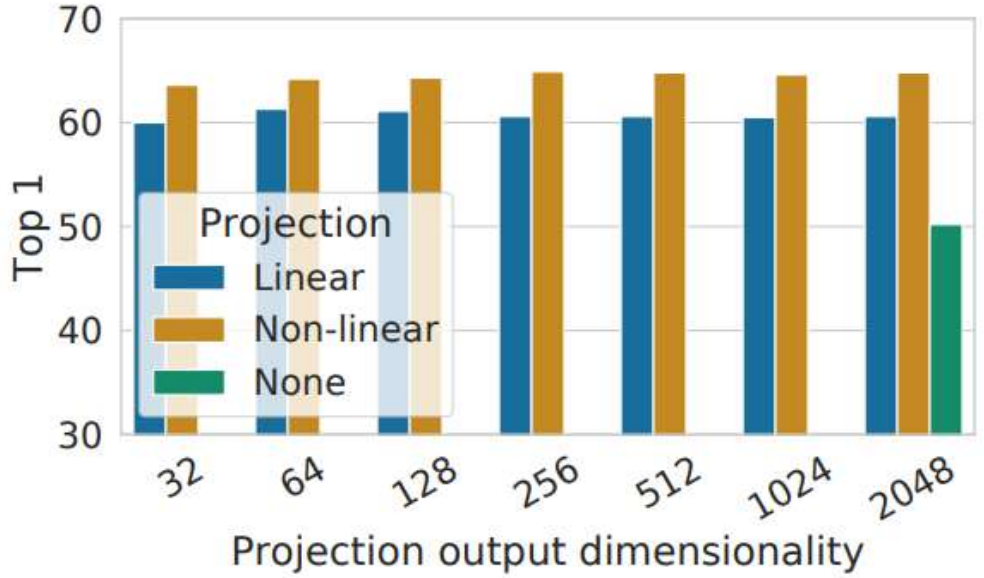


Figure 8. Linear evaluation of representations with different projection heads $g(\cdot)$ and various dimensions of $\boldsymbol{z} = g(\boldsymbol{h})$. The representation $\boldsymbol{h}$ (before projection) is 2048-dimensional here.

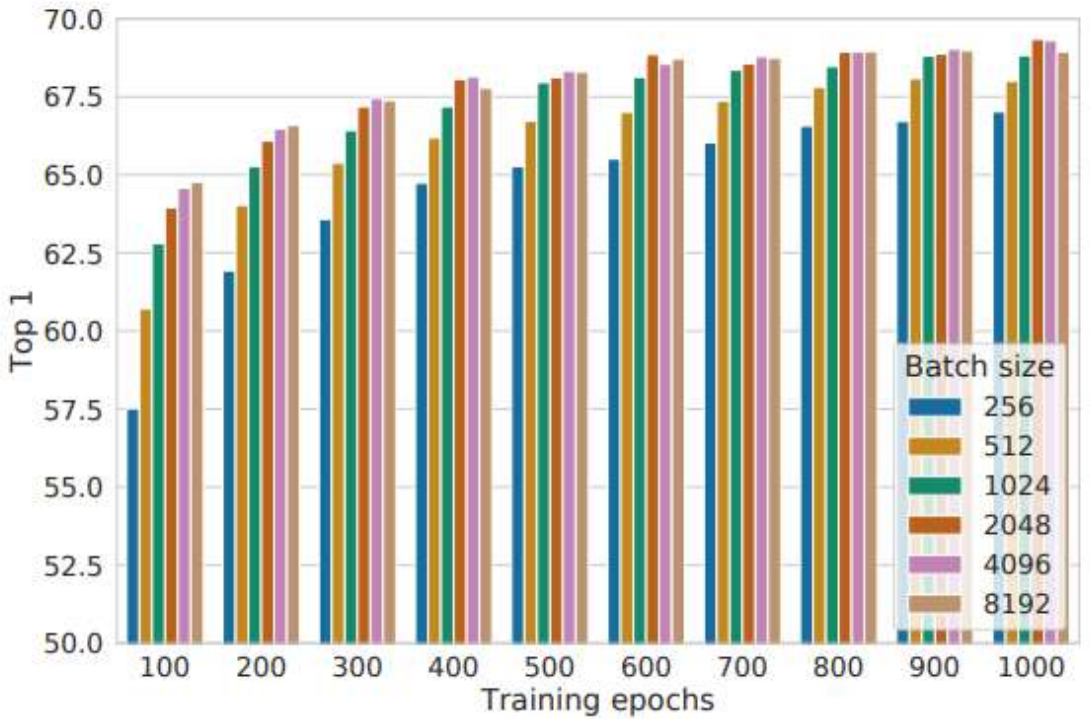
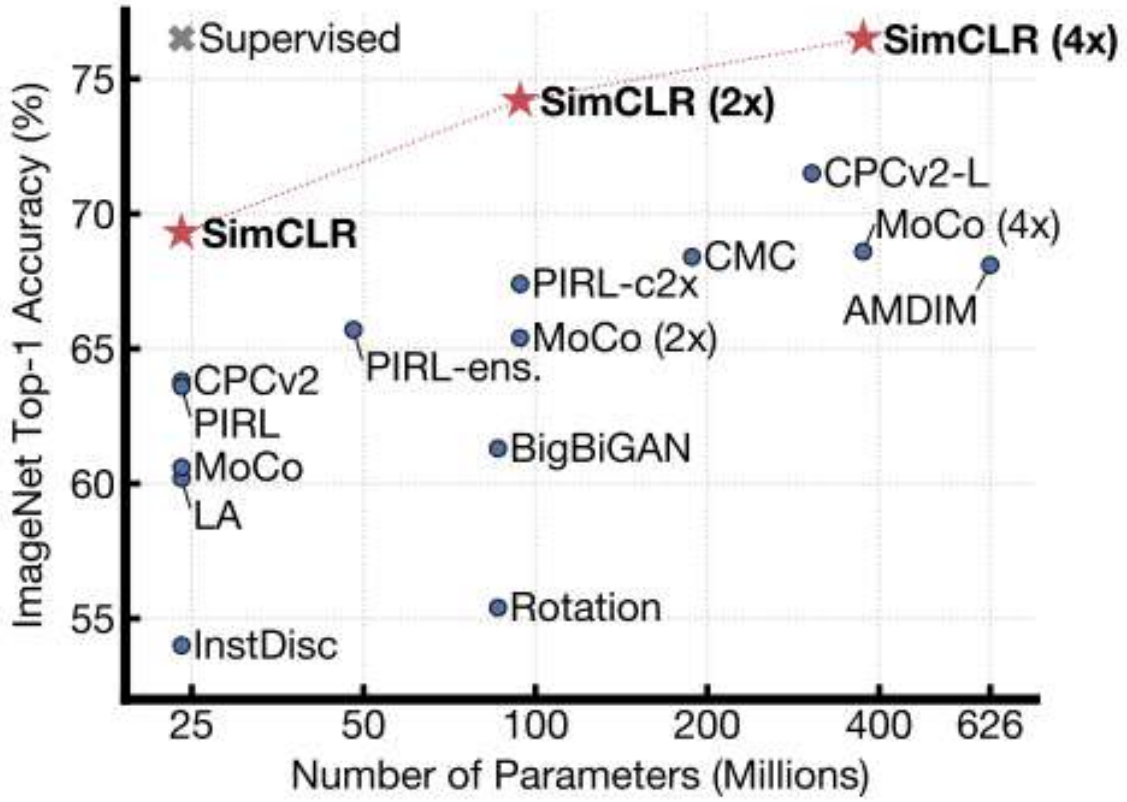


Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.[10]

Simple Framework for Contrastive Learning (SimCLR)

Results (linear)



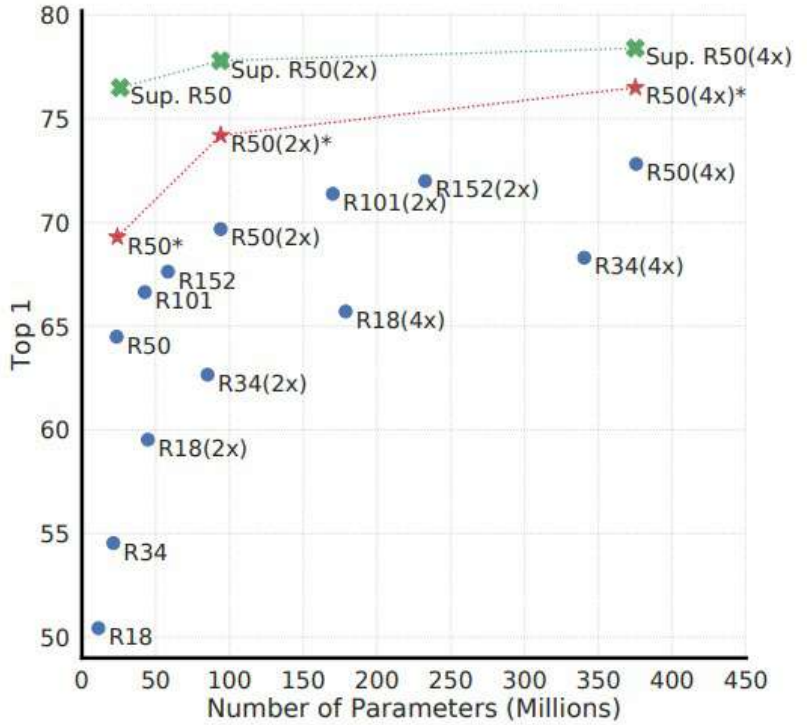


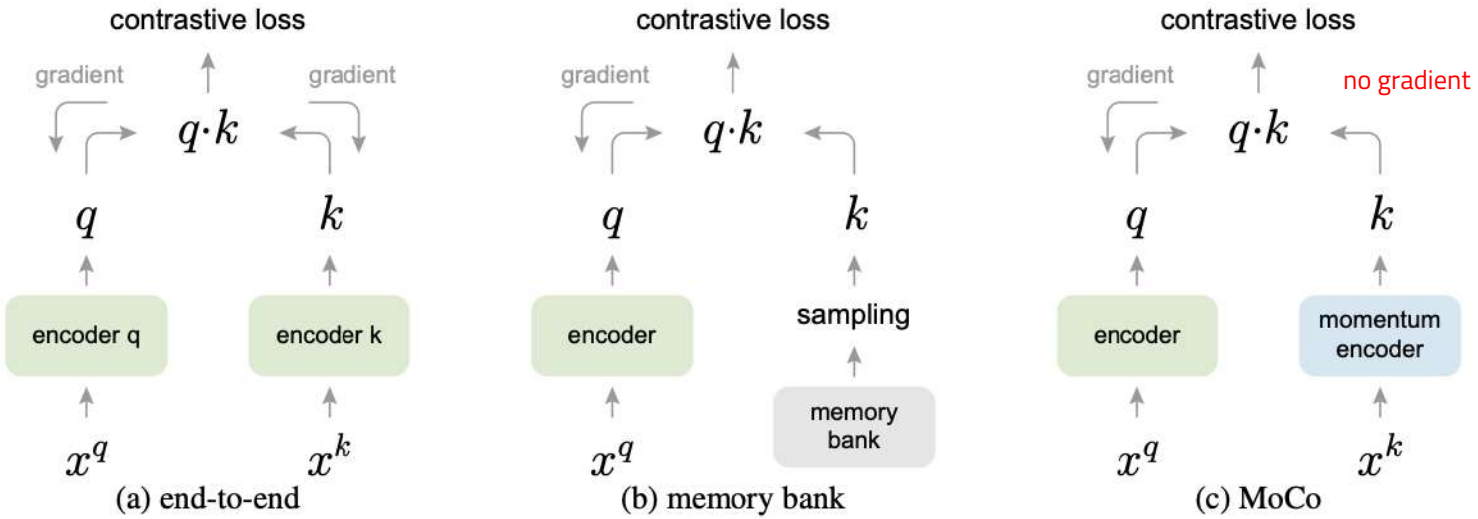
Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs[7] (He et al., 2016).

Source: A Simple Framework for Contrastive Learning of Visual Representations, NeurIPS, 2020

Momentum Contrast (MoCo)

Approach

(a) SimCLR style
(b) Instance Discrimination
(c) their approach



- **Idea:** decouple mini-batch size by a multiple keys in a queue
- number of negative samples is very crucial (see SimCLR)
- **Solution:** use momentum-updated encoder and maintain a queue
  ⇒ momentum = increases the key representation consistency
  ⇒ queue = allows us to use recent and many negative samples
- very similar to previous work ⇒ they called it memory bank

Update Procedure

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)} \qquad \theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

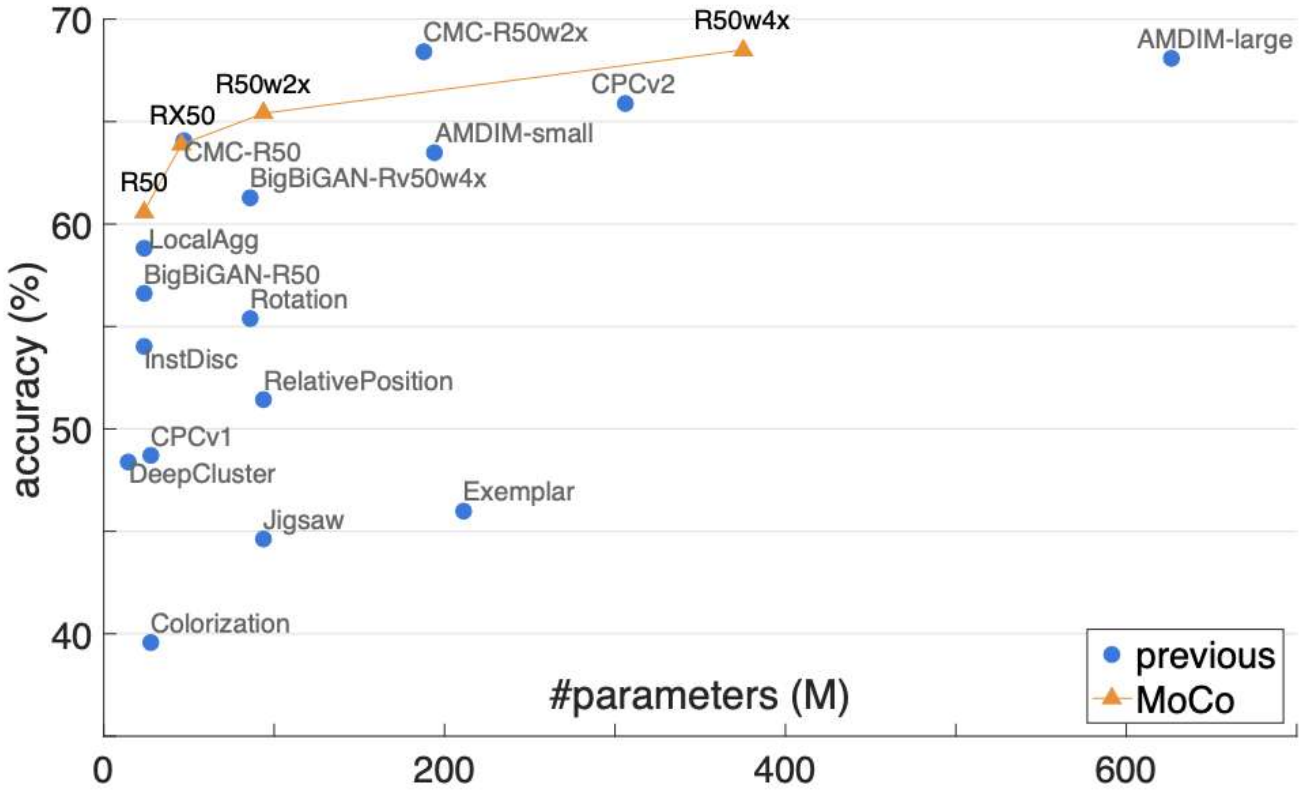
$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)}$$

first encoder is updated | **then** momentum is updated, add k^+

Source: Momentum Contrast for Unsupervised Visual Representation Learning, CVPR, 2020

Momentum Contrast (MoCo)

Results (linear)



method	architecture	#params (M)	accuracy (%)
Exemplar [17]	R50w3×	211	46.0 [38]
RelativePosition [13]	R50w2×	94	51.4 [38]
Jigsaw [45]	R50w2×	94	44.6 [38]
Rotation [19]	Rv50w4×	86	55.4 [38]
Colorization [64]	R101*	28	39.6 [14]
DeepCluster [3]	VGG [53]	15	48.4 [4]
BigBiGAN [16]	R50	24	56.6
	Rv50w4×	86	61.3
methods based on contrastive learning follow:			
InstDisc [61]	R50	24	54.0
LocalAgg [66]	R50	24	58.8
CPC v1 [46]	R101*	28	48.7
CPC v2 [35]	$R170^{*}_{wider}$	303	65.9
CMC [56]	$R50_{L+ab}$	47	64.1†
	$R50w2\times_{L+ab}$	188	68.4†
AMDIM [2]	$AMDIM_{small}$	194	63.5†
	$AMDIM_{large}$	626	68.1†
MoCo	R50	24	60.6
	RX50	46	63.9
	R50w2×	94	65.4
	R50w4×	375	**68.6**

Source: Momentum Contrast for Unsupervised Visual Representation Learning, CVPR, 2020

MoCo-v2 vs. SimCLR

MoCo v2 / v3

Improved Baselines with Momentum Contrastive Learning

Xinlei Chen Haoqi Fan Ross Girshick Kaiming He
Facebook AI Research (FAIR)

- **Specifics:**
  ⇒ From SimCLR: non-linear projection head
  ⇒ From SimCLR: strong augmentations
  ⇒ From MoCo: momentum-update with large number of negatives does well with smaller batches
- **Findings**:
  ⇒ Confirmation of projection head and strong augment.
  ⇒ MoCo v2 can run with batch-sizes of 256 vs. 8192
  ⇒ much smaller memory footprint
- **MoCo v3:**
  ⇒ using a ViT as encoder

Results

case	unsup. pre-train MLP	aug+	cos	epochs	batch	ImageNet acc.
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
*results of **longer** unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

Table 2. **MoCo *vs.* SimCLR**: ImageNet linear classifier accuracy (**ResNet-50, 1-crop 224×224**), trained on features from unsupervised pre-training. "aug+" in SimCLR includes blur and stronger color distortion. SimCLR ablations are from Fig. 9 in [2] (we thank the authors for providing the numerical results).

mechanism	batch	memory / GPU	time / 200-ep.
MoCo	256	**5.0G**	**53 hrs**
end-to-end	256	7.4G	65 hrs
end-to-end	4096	93.0G†	n/a

Table 3. **Memory and time cost** in 8 V100 16G GPUs, implemented in PyTorch. †: based on our estimation.

Source: Improved Baselines with Momentum Contrastive Learning, arxiv, March 2020

MoCo-v2 vs. SimCLR-v2

SimCLR v2

Big Self-Supervised Models are Strong Semi-Supervised Learners

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey Hinton
Google Research, Brain Team

- **Specifics:**
  ⇒ Larger encoder i.e., ResNet-152 and 3x wider channels
  ⇒ Increased capacity of projection head
  ⇒ Extension to memory mechanism with momentum

- **Findings**:
  ⇒ focus on finetuning and semi-supervised training towards ImageNet performs really well with larger models
  ⇒ distillation (teacher/student) with unlabeled samples can be used for refinement and knowledge transfer

Source: Big Self-Supervised Models are Strong Semi-Supervised Learner, NeurIPS 2020

Results

Depth	Width	Use SK [28]	Param (M)	Fine-tuned on 1%	Fine-tuned on 10%	Fine-tuned on 100%	Linear eval	Supervised
50	1×	False	**24**	**57.9**	**68.4**	**76.3**	**71.7**	**76.6**
		True	35	64.5	72.1	78.7	74.6	78.5
	2×	False	94	66.3	73.9	79.1	75.6	77.8
		True	140	70.6	77.0	81.3	77.7	79.3
101	1×	False	43	62.1	71.4	78.2	73.6	78.0
		True	65	68.3	75.1	80.6	76.3	79.6
	2×	False	170	69.1	75.8	80.7	77.0	78.9
		True	257	73.2	78.8	82.4	79.0	80.1
152	1×	False	58	64.0	73.0	79.3	74.5	78.3
		True	89	70.0	76.5	81.3	77.2	79.9
	2×	False	233	70.2	76.6	81.1	77.4	79.1
		True	354	74.2	79.4	82.9	79.4	80.4
152	3×	True	**795**	**74.9**	**80.1**	**83.1**	**79.8**	**80.5**

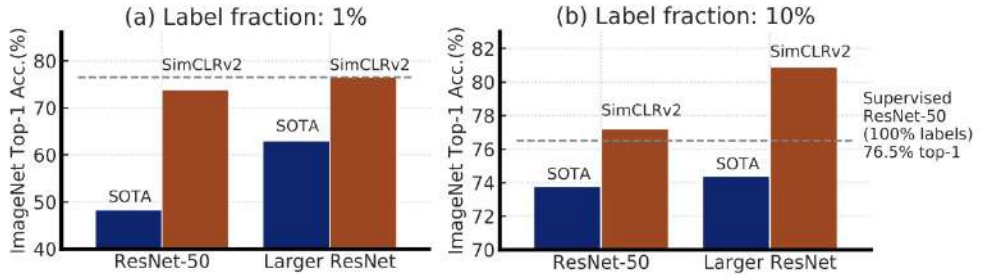


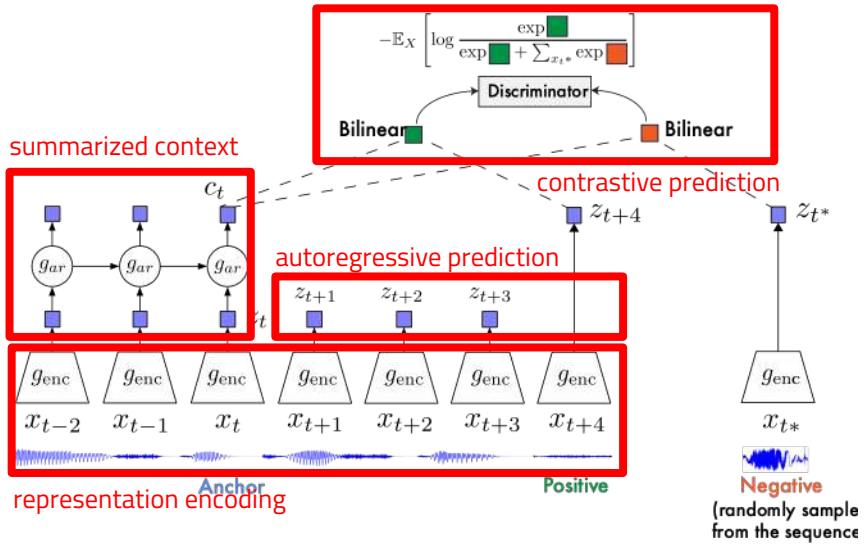
Figure 2: Top-1 accuracy of previous state-of-the-art (SOTA) methods [1, 2] and our method (SimCLRv2) on ImageNet using only 1% or 10% of the labels. Dashed line denotes fully supervised ResNet-50 trained with 100% of labels. Full comparisons in Table 3.

Universität St.Gallen

And many other approaches

Contrastive Predictive Coding (CPC)

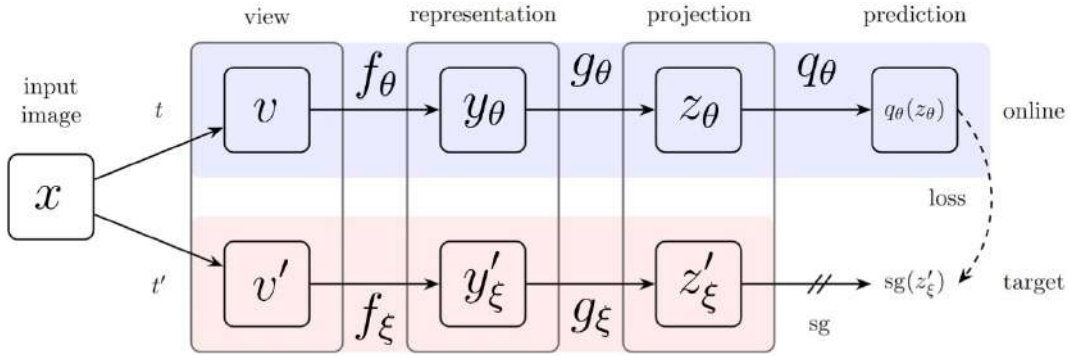
Contrast over a sequence such as timesteps for audio



Source: Representation Learning with Contrastive Predictive Coding Arxiv https://arxiv.org/abs/1807.03748, 2018

Bootstrap your own latent (BYOL)

Only positive samples are used, and contrast is calculated against target (momentum) network



Source: Bootstrap your own latent: A new approach to self-supervised Learning, NeurIPS, 2020

And many other approaches

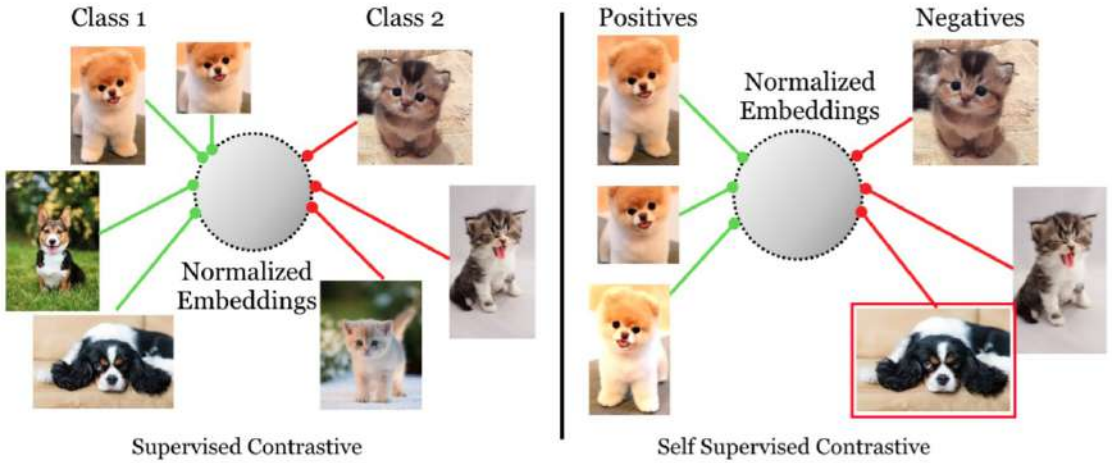
SwAV

Cluster embeddings to prototypes and swap.

Source: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, NeurIPS 2020

Supervised Contrast

Use supervision to prevent false negatives



Source: Supervised Contrastive Learning, NeurIPS 2020

And many other approaches

Multi-view Coding

Matching representation of multiple views of an image

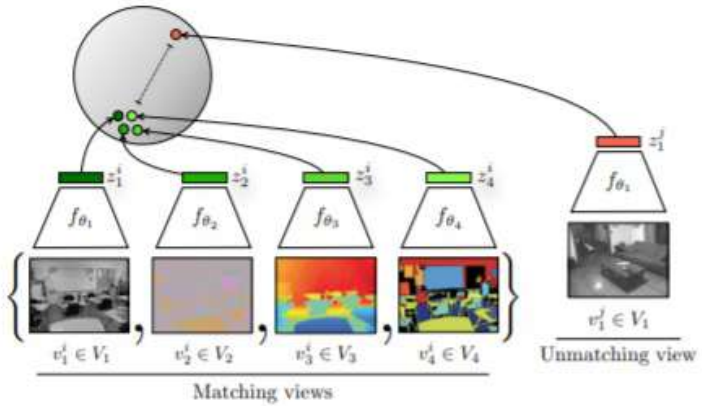
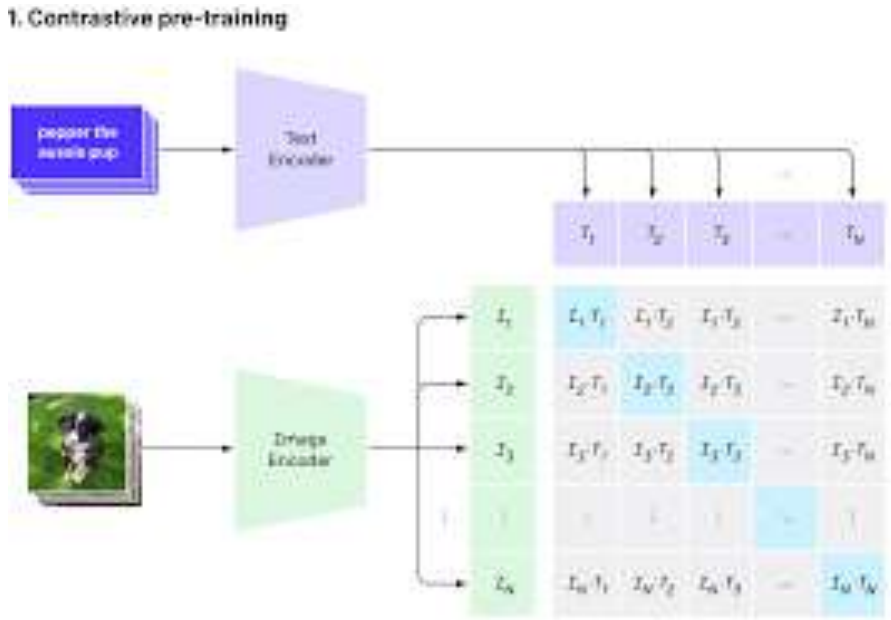


Figure 1: Given a set of sensory views, a deep representation is learnt by bringing views of the *same* scene together in embedding space, while pushing views of *different* scenes apart. Here we show and example of a 4-view dataset (NYU RGBD [53]) and its learned representation. The encodings for each view may be concatenated to form the full representation of a scene.

Source: SSL Tutorial, NeurIPS 2021

Clip – Multi-modal Alignment

"Views" can be paired input from two or more modalities



Source: OpenAI - https://openai.com/blog/clip/

Contrastive Learning for Remote Sensing

Manas, Lacoste, Giro-i-Nieto, Vazquez, Rodriguez, Seasonal contrast:
Unsupervised pre-training from uncurated remote sensing data,
ICCV, 2021

Seasonal Contrast

Setup

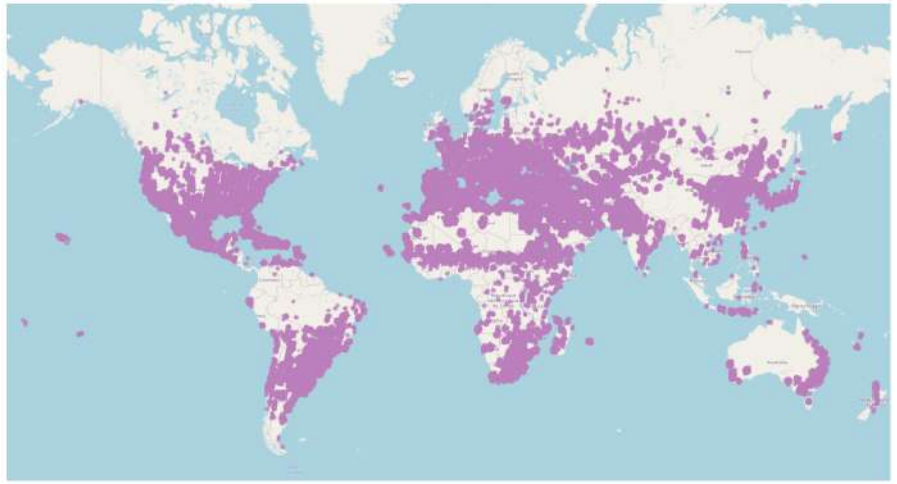
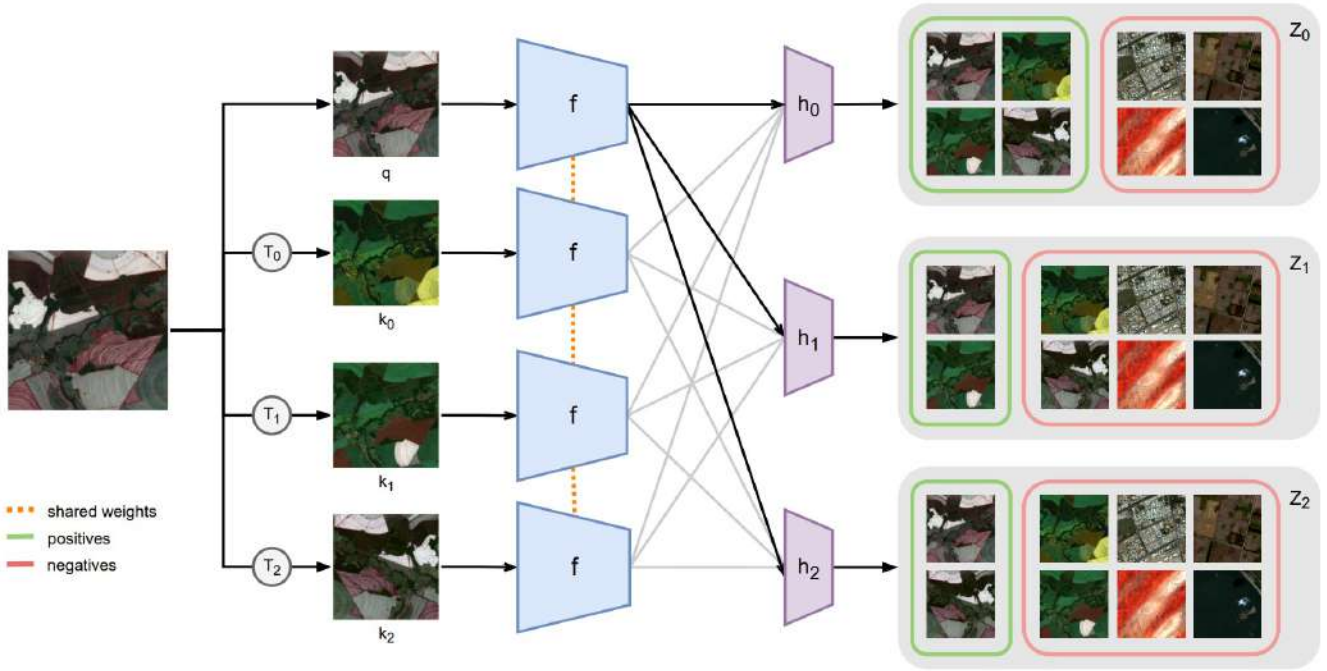


Figure 1. **Distribution of the Seasonal Contrast (SeCo) dataset.** Each point represents a sampled location. Images are collected around human settlements to avoid monotonous areas such as oceans and deserts.



Idea: using multiple views: including image augmentations, different seasons of the same location and combine them into multiple embeddings sub-spaces (being invariant to different combinations) to calculate contrast for pre-training.

Manas, Lacoste, Giro-i-Nieto, Vazquez, Rodriguez, Seasonal contrast: **Unsupervised pre-training from uncurated remote sensing data**, ICCV, 2021

Seasonal Contrast

Results (linear probing & finetuning)

Pre-training	Backbone	100k images				1M images			
		Linear probing		Fine-tuning		Linear probing		Fine-tuning	
		10%	100%	10%	100%	10%	100%	10%	100%
Random init.	ResNet-18	43.05	45.95	68.11	79.80	43.05	45.95	68.11	79.80
ImageNet (sup.)		65.69	66.40	78.76	85.90	65.69	66.40	78.76	85.90
MoCo-v2	ResNet-18	69.70	70.90	78.76	85.17	69.28	70.79	78.33	85.23
MoCo-v2+TP		70.20	71.08	79.80	85.71	72.58	73.60	80.68	86.59
SeCo (ours)		**74.67**	**75.52**	**81.49**	**87.04**	**76.05**	**77.00**	**81.86**	**87.27**
Random init.	ResNet-50	43.95	46.92	69.49	78.98	43.95	46.92	69.49	78.98
ImageNet (sup.)		70.46	71.82	80.04	86.74	70.46	71.82	80.04	86.74
MoCo-v2	ResNet-50	71.85	73.27	79.23	85.79	73.71	75.65	80.08	86.05
MoCo-v2+TP		72.61	73.91	79.04	85.35	74.50	76.32	80.20	86.11
SeCo (ours)		**77.49**	**79.13**	**81.72**	**87.12**	**78.56**	**80.35**	**82.62**	**87.81**

Table 1. Mean average precision on the BigEarthNet land-cover classification task. Results cover different pre-training approaches and different ResNet backbones. We also explore the effect of the unlabeled pre-training set size between 100k and 1M images, and the size of the BigEarthNet training set between 10% and 100%.

Pre-training	Accuracy
Random init.	63.21
Imagenet (sup.)	86.44
MoCo-v2	83.72
MoCo-v2+TP	89.51
SeCo (ours)	**93.14**

Table 3. Fine-tuning accuracy on the EuroSAT land-cover classification task. We use a ResNet-18 backbone pre-trained on 1M images.

Manas, Lacoste, Giro-i-Nieto, Vazquez, Rodriguez, Seasonal contrast: **Unsupervised pre-training from uncurated remote sensing data**, ICCV, 2021

Universität St.Gallen

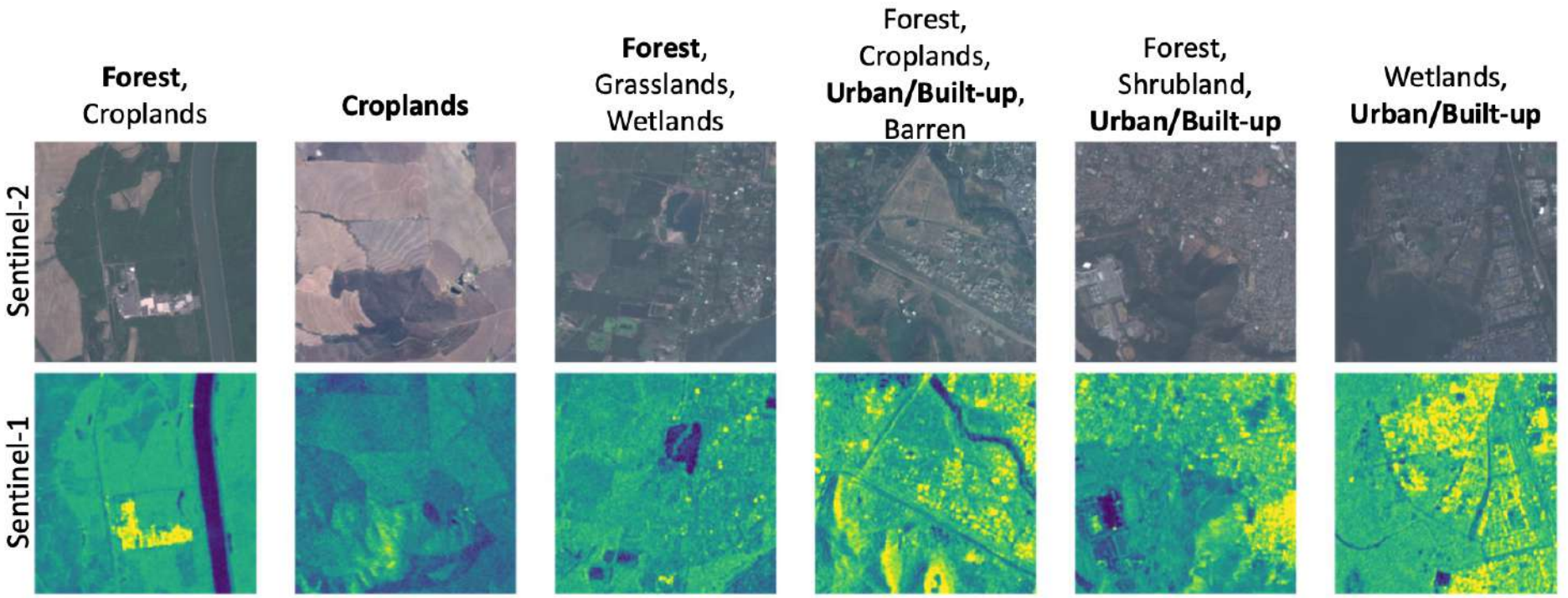
Multimodal Contrastive Learning for Remote Sensing

L Scheibenreif, J Hanna, M Mommert, D Borth
Self-supervised Vision Transformer for Land-cover Segmentation and Classification
CVPR Earth Vision Workshop, 2022

LScheibenreif, M Mommert, D Borth
Contrastive Self-supervised Data Fusion for Satellite Imagery
Int. Society for Photogrammetry and Remote Sensing (ISPRS), 2022

Multimodal Contrastive Learning for Remote Sensing

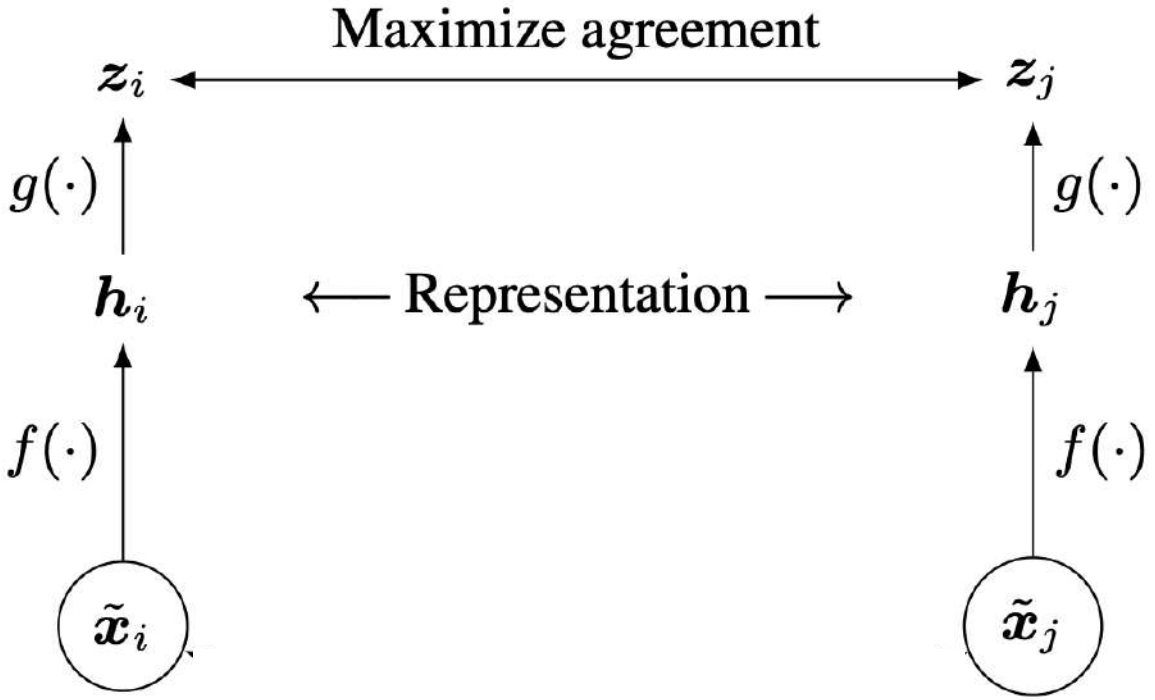
In satellite imagery, there are multiple views of the same location



Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X.
SEN12MS--A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion.
arXiv preprint arXiv:1906.07789, 2019

Multimodal Contrastive Learning for Remote Sensing

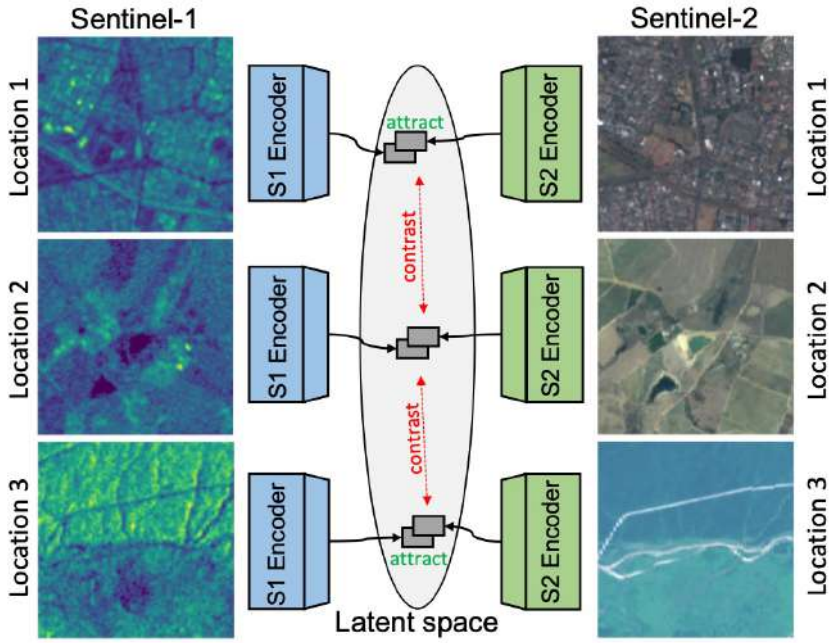
- Contrastive SSL yields great performance on natural images (e.g., SimCLR)
- Based on multiple views of same instance
- In natural images, multiple views are generated with **random augmentations**
- In remote sensing, unlabeled data is abundant, but less labeled data
- What could multiple views be in remote sensing and earth observation?



Data & Setup

Self-supervised pre-training

- Co-located Sentinel-1/2 image pairs
- SEN12MS dataset [Schmitt 2019]
- Low-resolution land cover labels are ignored

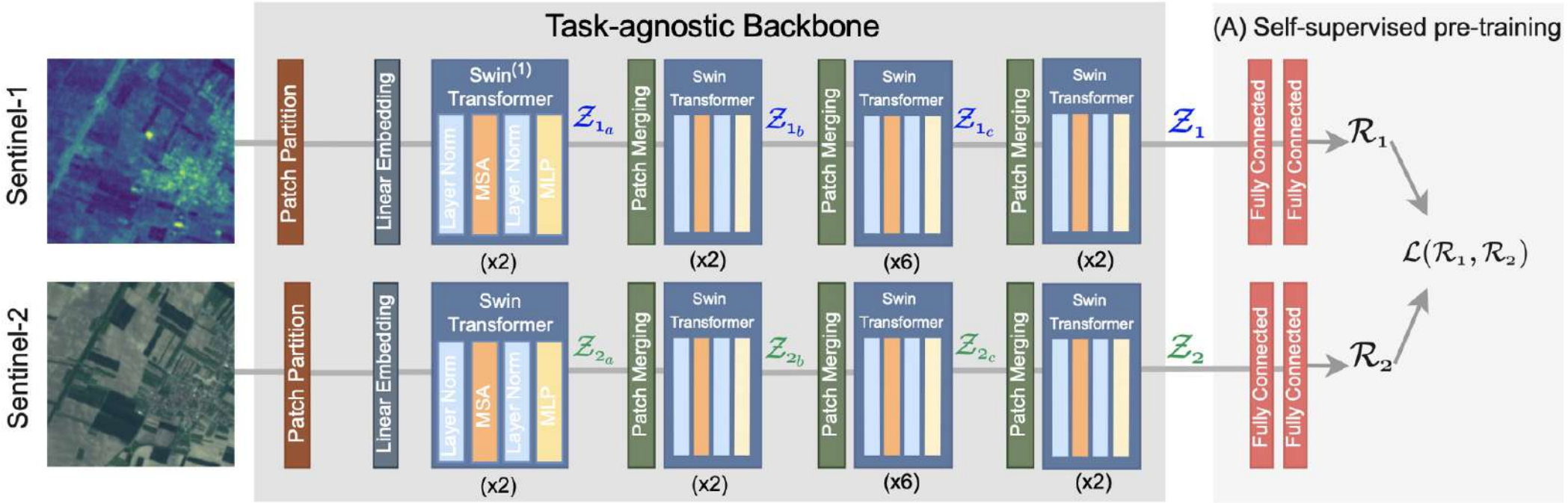


Land-cover classification downstream tasks

- Dataset from Data Fusion Contest (DFC2020) [Yokoya 2020]
- **Task 1:** Single- and multilabel classification
- **Task 2:** Segmentation

Universität St.Gallen

Self-supervised Pre-training

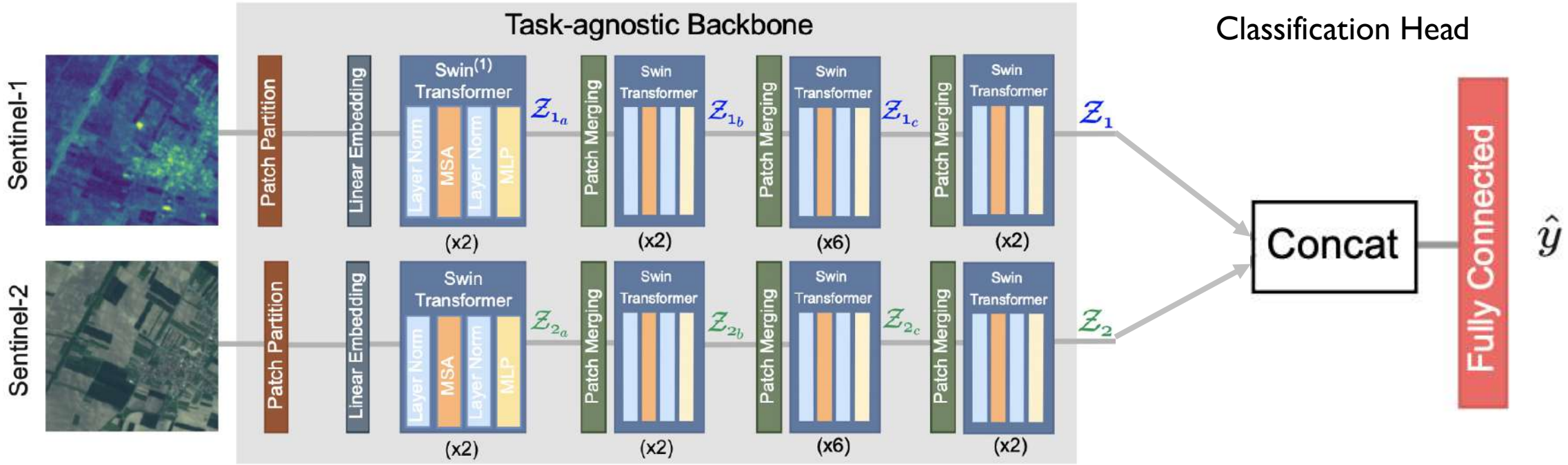


1. Encode Sentinel-1/2 images with distinct encoders
2. Compute contrastive loss on projected representations

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_k)/\tau)}$$

Swin Transformer: [Liu 2021]

Classification

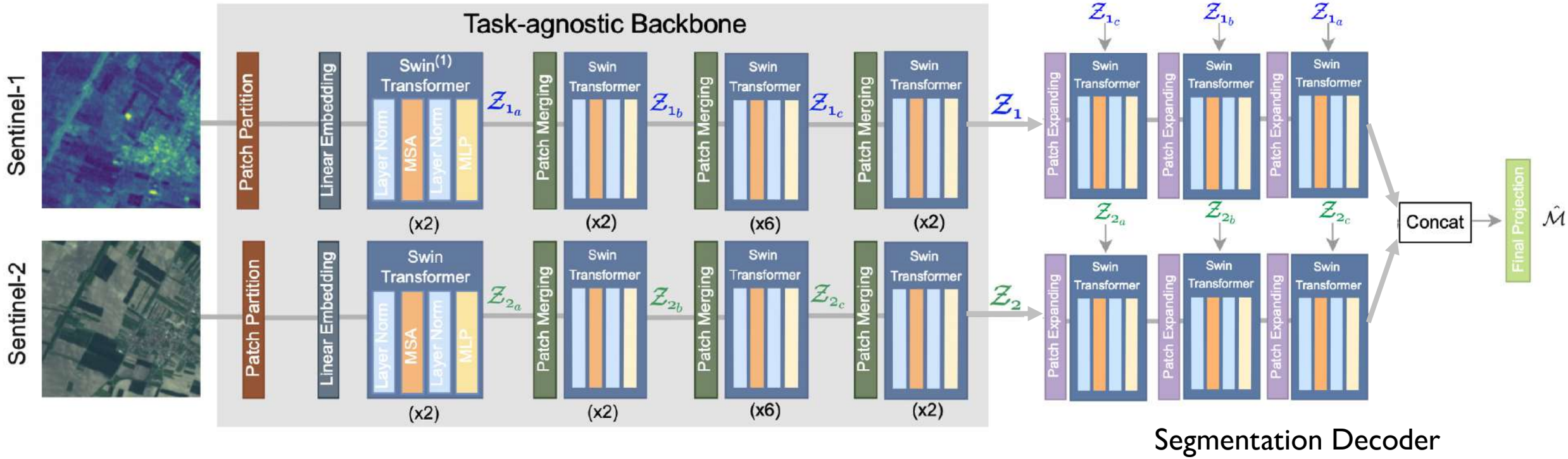


1. Encode Sentinel-1/2 images with distinct encoders
2. Compute contrastive loss on projected representations
3. Replace projection head by downstream task specific head

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_k)/\tau)}$$

Swin Transformer: [Liu 2021]

Segmentation



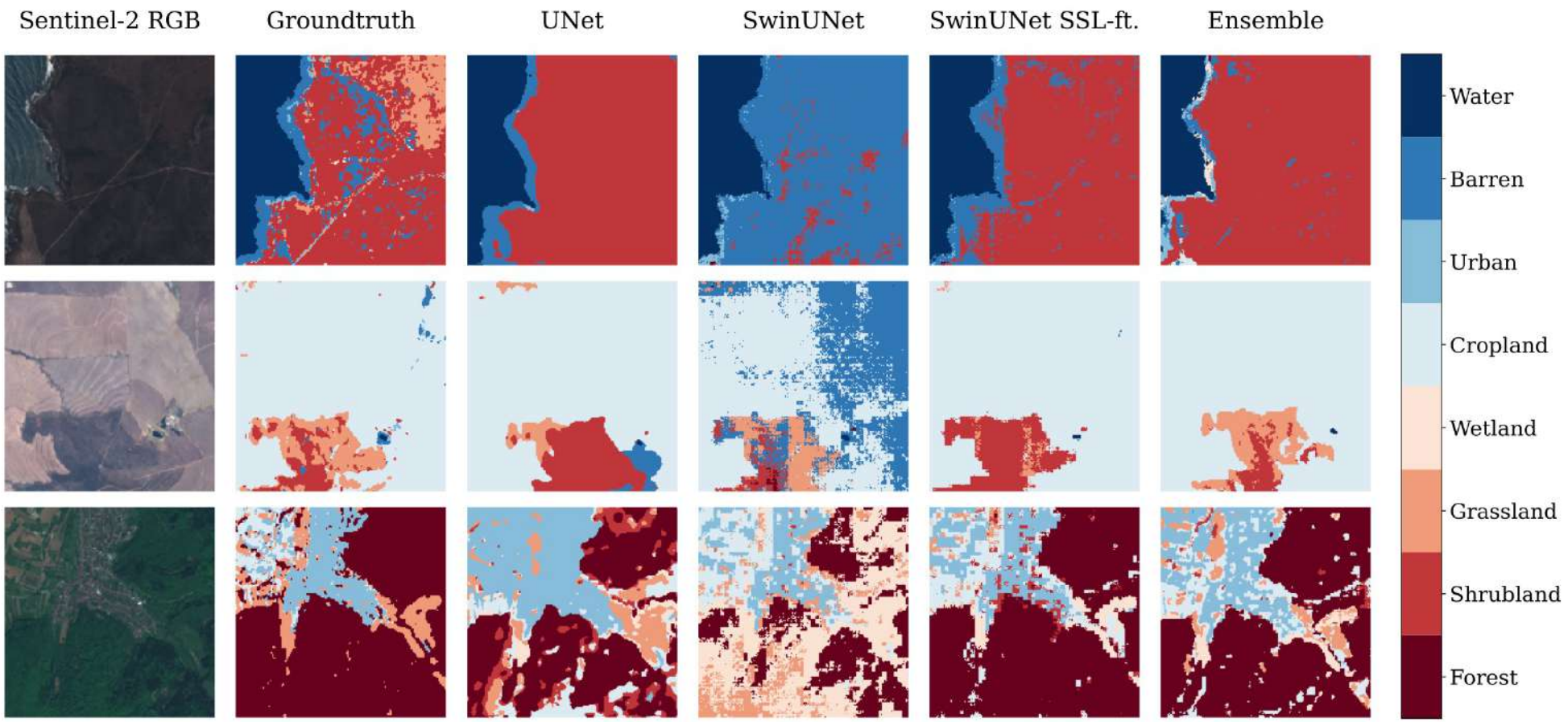
1. Encode Sentinel-1/2 images with distinct encoders
2. Compute contrastive loss on projected representations
3. Replace projection head by downstream task specific head

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathcal{R}_i, \mathcal{R}_k)/\tau)}$$

Swin Transformer: [Liu 2021]

Universität St.Gallen

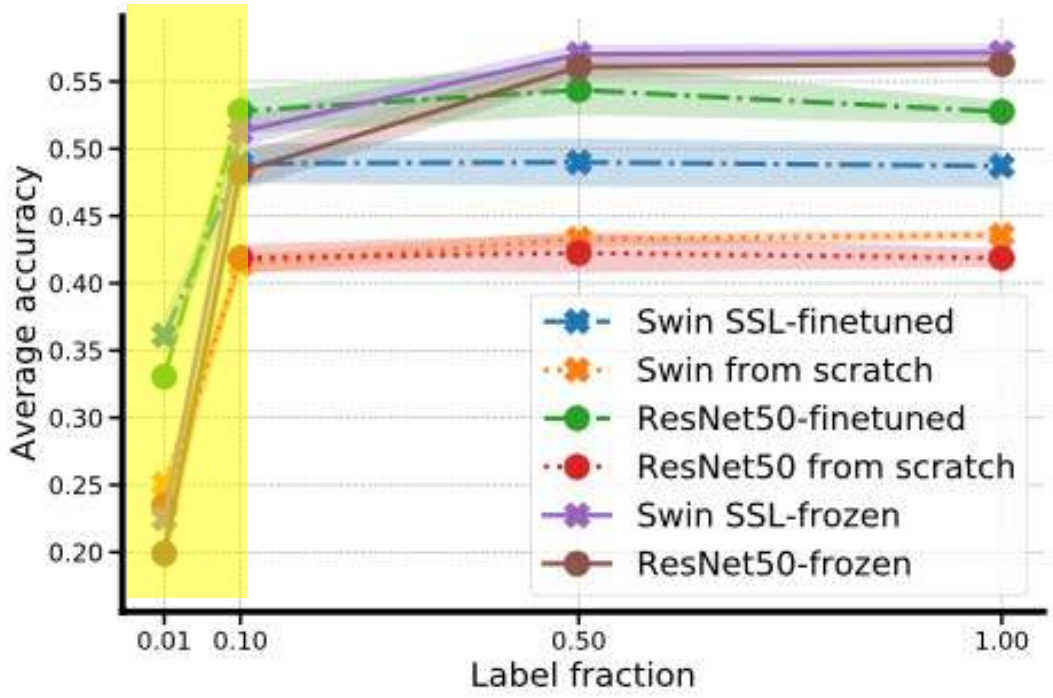
Results



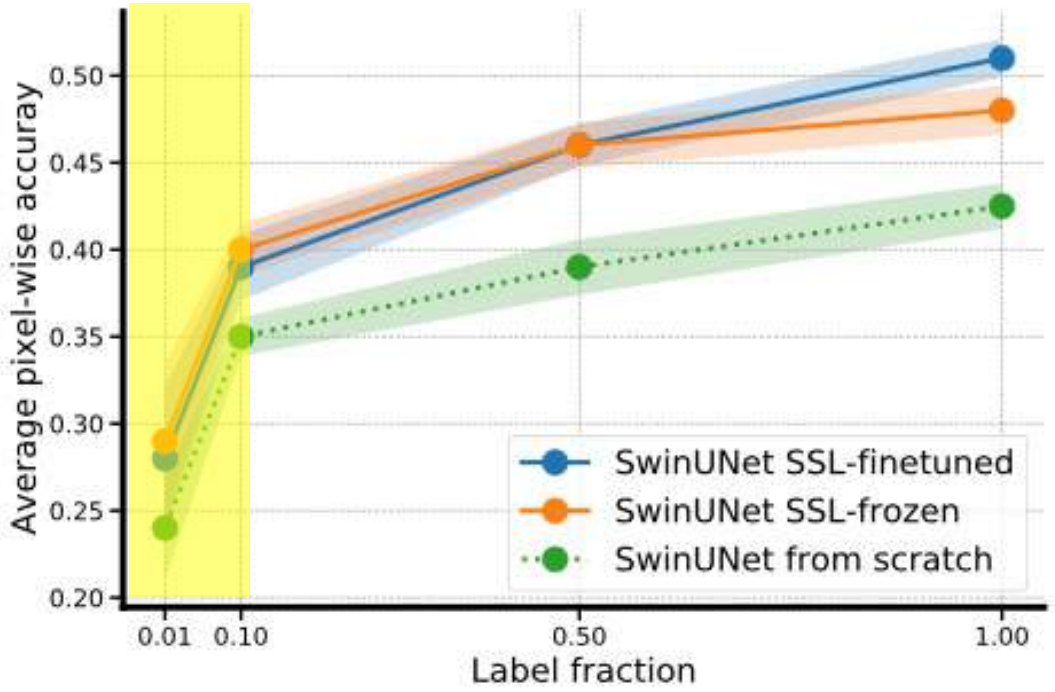
Universität St.Gallen

Results

Classification



Segmentation



SSL pre-training and 10-20% of labeled data outperform fully supervised training

Universität St.Gallen

Summary

Universität St.Gallen

This lecture

- (Brief) Recap SSL

- Contrastive Learning
  - Goal and Idea
  - Basic Notation
  - Learning Objective
  - Frameworks
- SSL & Contrastive Learning for Remote Sensing

Universität St.Gallen

Questions?