max planck institut informatik

SIC Saarland Informatics Campus

High Level Computer Vision

Self-Supervised Learning (Part 2)

@ July 3, 2024

Bernt Schiele

cms.sic.saarland/hlcvss24/

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

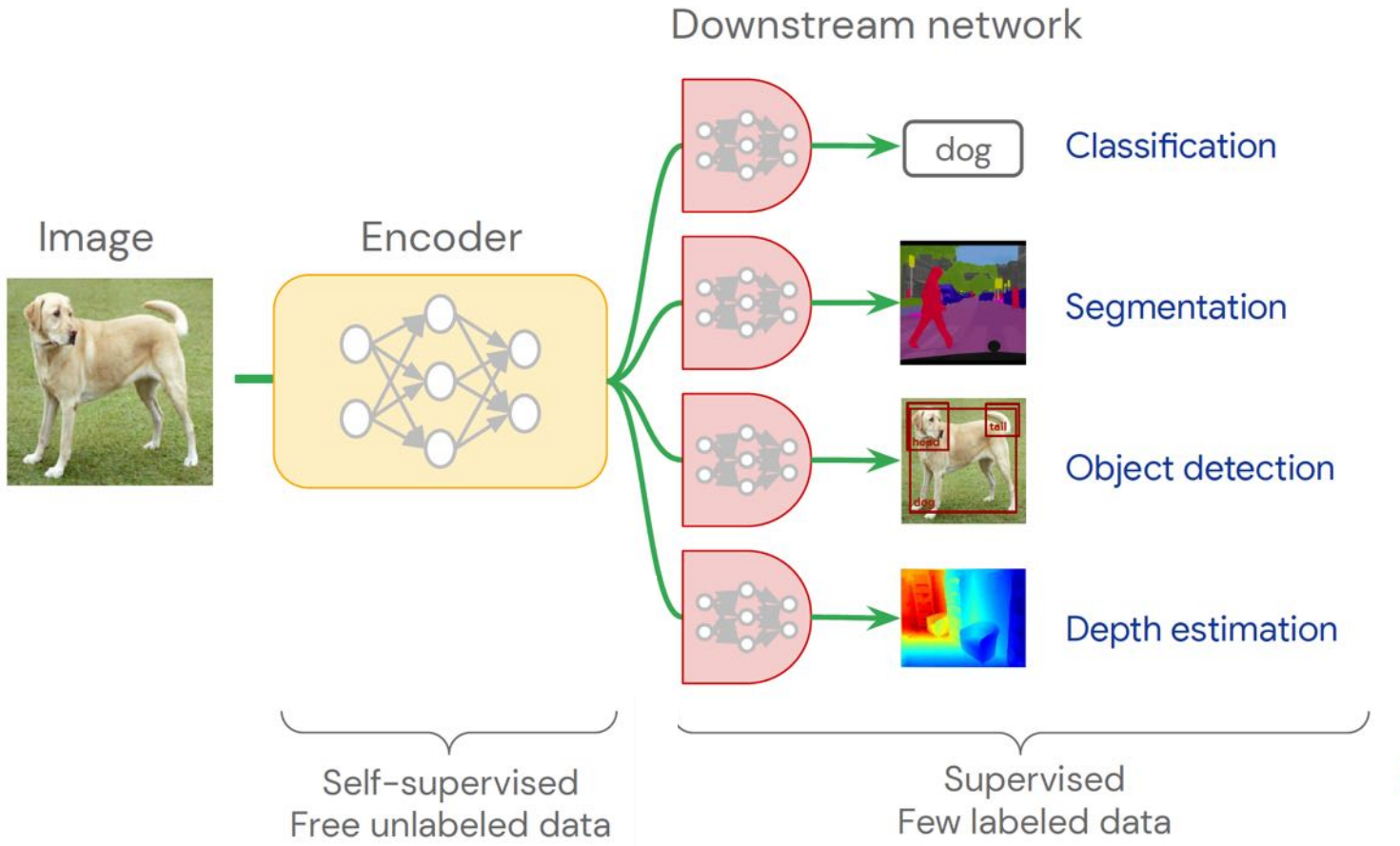
Overview of Today's Lecture

- Last Time: Self-Supervised Learning — Part 1:
  - Motivation of Self-Supervised Learning
  - Pretext tasks as image transformations (e.g. rotation, inpainting, coloring)
  - Contrastive representation learning (SimCLR, MoCo, CPC)

- Today: Self-Supervised Learning — Part 2:
  - Teacher-Student "feature reconstruction"
    - motivation, setting
    - methods: BYOL, DINO
  - Image Reconstruction
    - MAE - Masked Autoencoders

slide credit: Jean-Basien Grill, et al.

Idea of Self-Supervised Learning

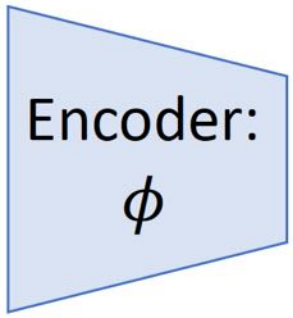


slide credit: Justin Johnson

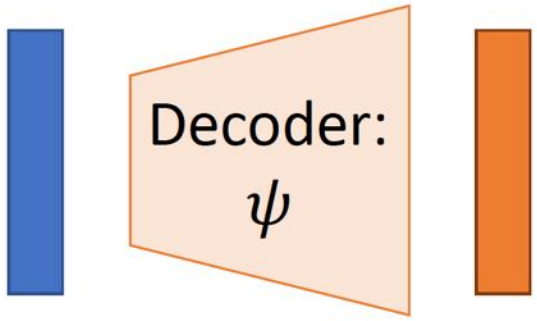
Self-Supervised Learning: Pretraining — then Transfer

Step 1: Pretrain a network on a pretext task that doesn't require supervision

Input Image: x



Features: $\phi(x)$

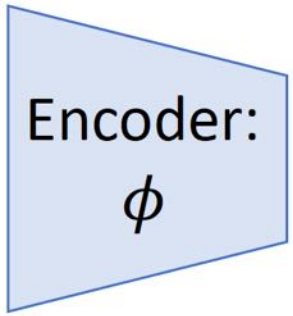


Prediction: $\hat{y}$

Loss: $L(\hat{y}, y)$

Step 2: Transfer encoder to downstream tasks via linear classifiers, KNN, finetuning

Input Image: x

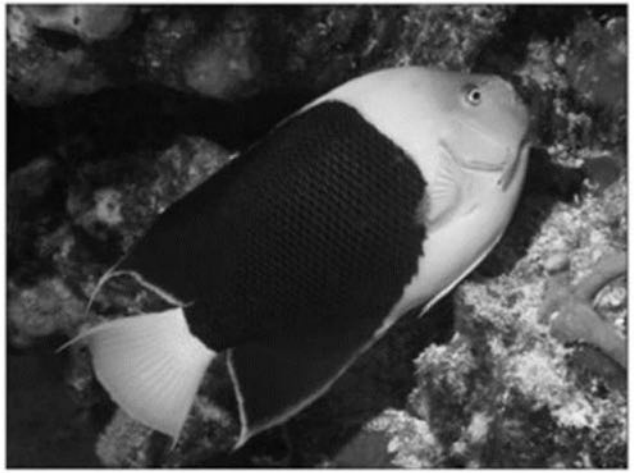


Features: $\phi(x)$

Downstream tasks: Image classification, object detection, semantic segmentation

slide credit: Spyros Gidaris, Andrei Bursuc

Colorization

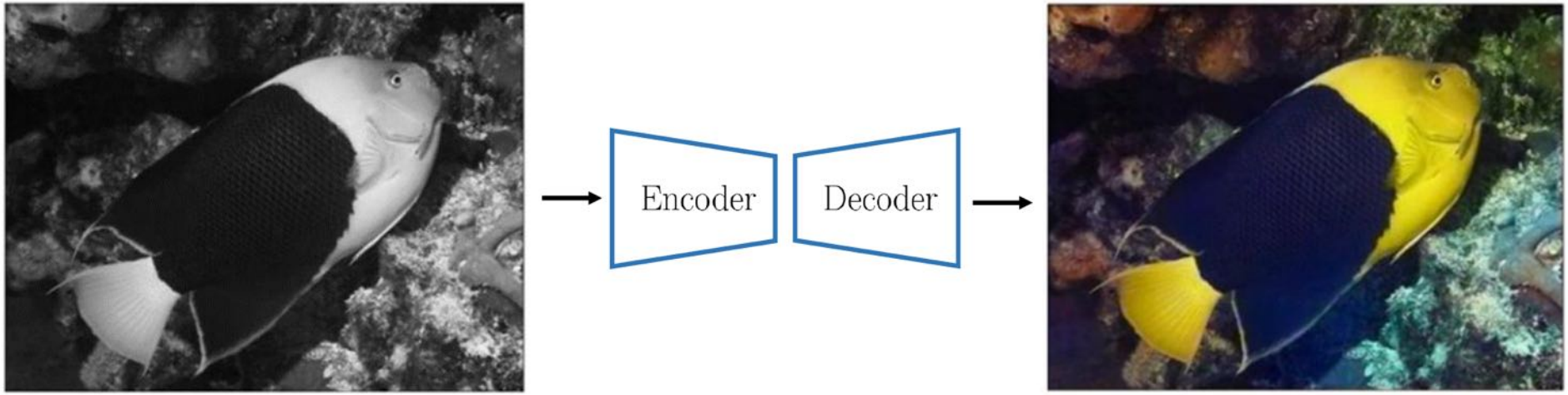


What is the colour of every pixel?
Hard if you don't recognize the objects!

Image colorization (Zhang et al. 2018)

slide credit: Spyros Gidaris, Andrei Bursuc

Colorization

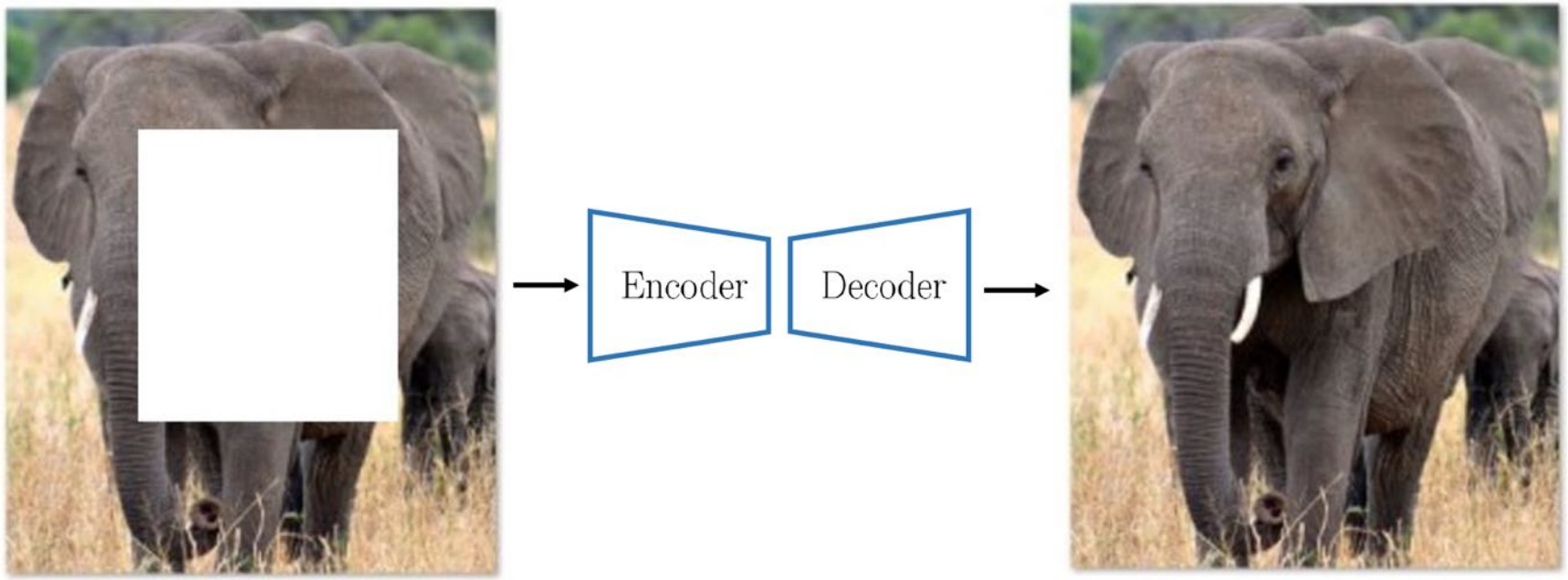


- Requires preservation of fine-grained information
- Input reconstruction is too hard and ambiguous
- Lots of effort spent on “useless” details: exact color, good boundary, etc.

Image colorization (Zhang et al. 2018)

slide credit: Spyros Gidaris, Andrei Bursuc

Context Encoders



- Requires preservation of fine-grained information and context-aware skills
- Input reconstruction is too hard and ambiguous
- Lots of effort spent on “useless” details: exact color, good boundary, etc.

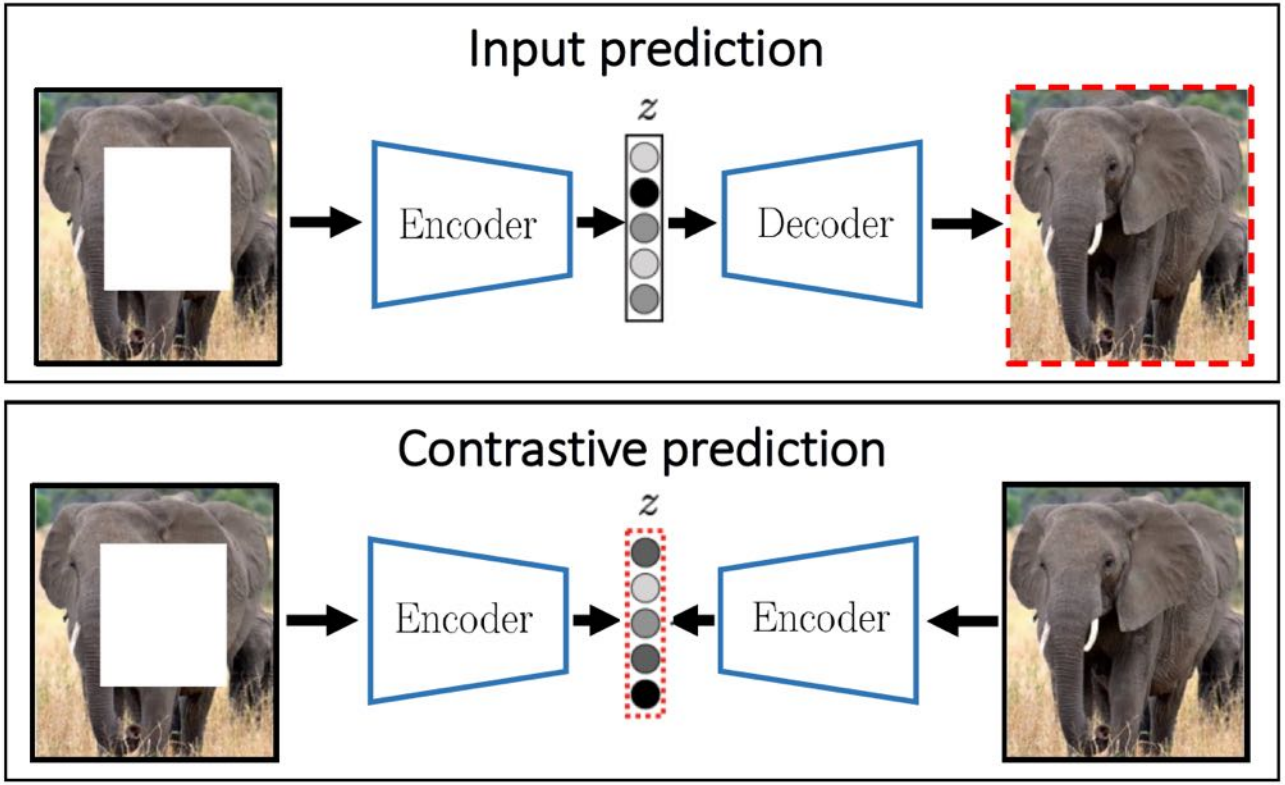
Context Encoders (Pathak et al. 2016)

slide credit: Spyros Gidaris, Andrei Bursuc

Contrastive Learning

Formulates self-supervised tasks in terms of learned representations:

- Recognize different views of the same image in the presence of distracting negative image views
- Requires many negative examples
- How to choose negatives?
- Impossible to know whether a sample is actually negative or actually (i.e., from the same object)



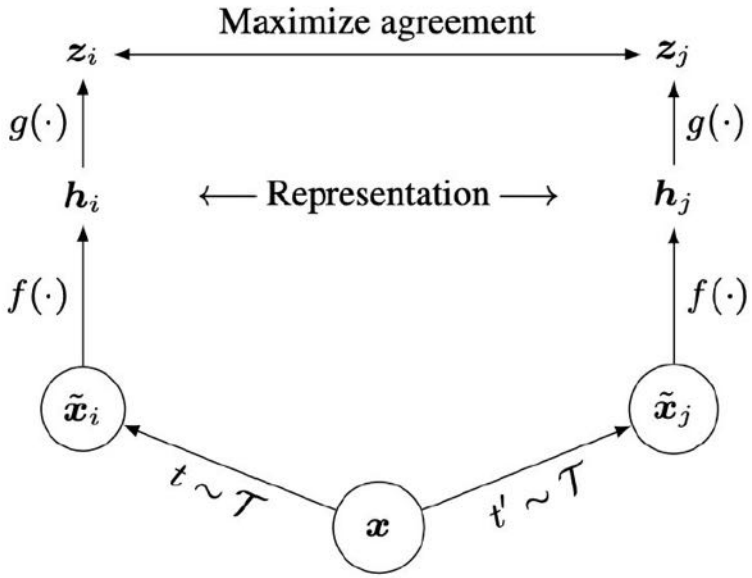
References:
"Representation learning with contrastive predictive coding", Oord et al, 2018
"Constraive multiview coding", Tian et al, 2020
"A simple framework for contrastive learning of visual representations", Chen et al, 2020
...

slide credit: Fei-Fei Li, Jiajun Wu, Ruohan Gao

Summary: Contrastive Representation Learning

SimCLR: a simple framework for contrastive representation learning

- **Key ideas**: non-linear projection head to allow flexible representation learning
- Simple to implement, effective in learning visual representation
- Requires large training batch size to be effective; large memory footprint

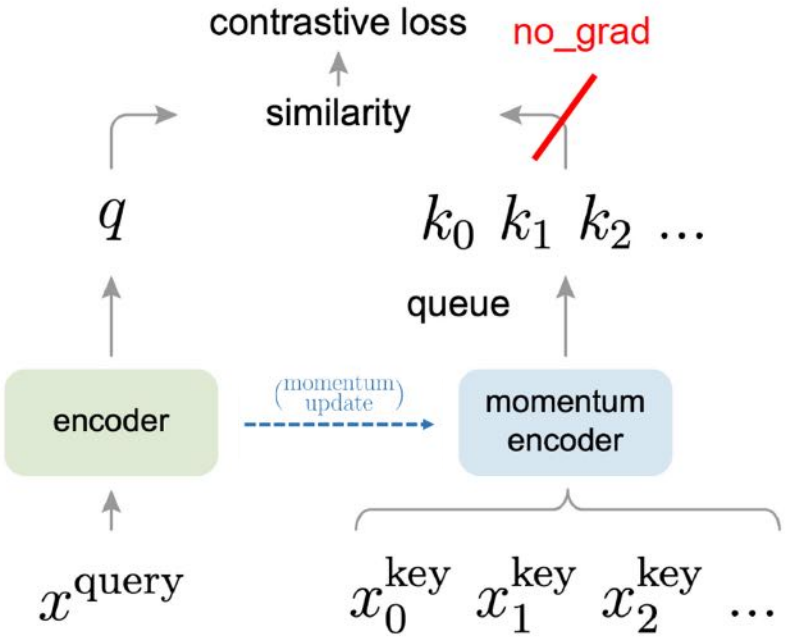


slide credit: Fei-Fei Li, Jiajun Wu, Ruohan Gao

Summary: Contrastive Representation Learning

MoCo (v1, v2): contrastive learning using momentum sample encoder

- Decouples negative sample size from minibatch size; allows large batch training without TPU
- MoCo-v2 combines the key ideas from SimCLR, i.e., nonlinear projection head, strong data augmentation, with momentum contrastive learning



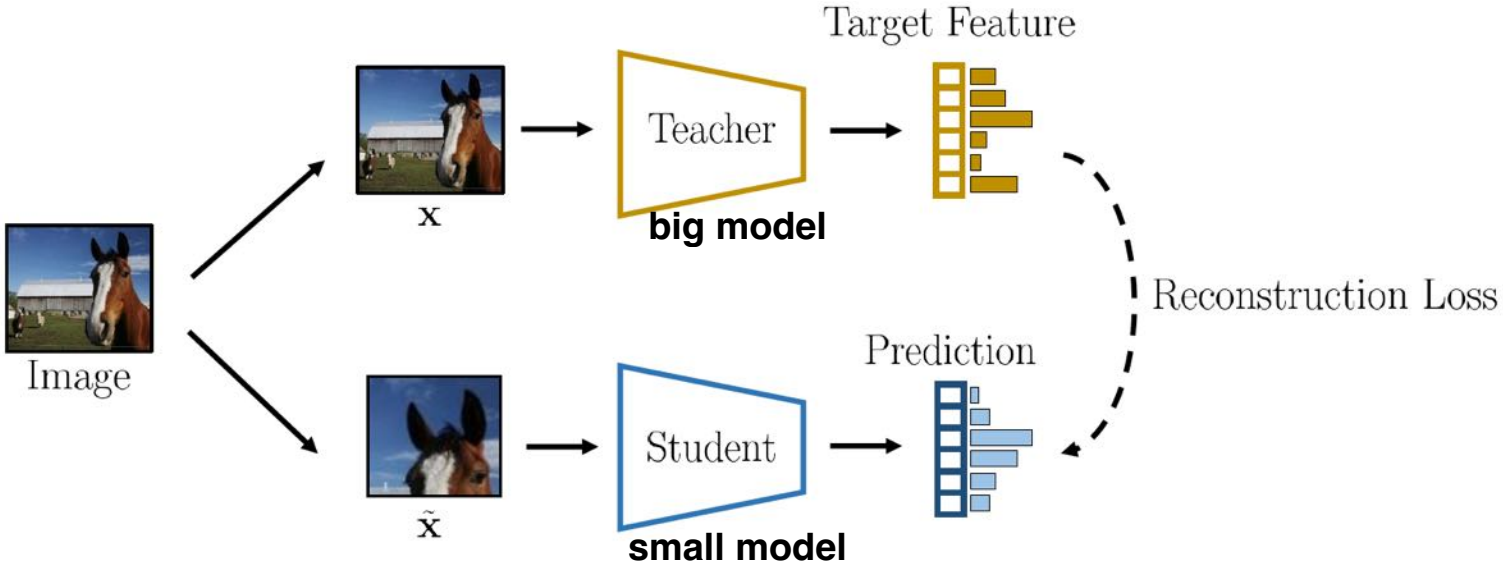
Overview of Today's Lecture

- Last Time: Self-Supervised Learning — Part 1:
  - Motivation of Self-Supervised Learning
  - Pretext tasks as image transformations (e.g. rotation, inpainting, coloring)
  - Contrastive representation learning (SimCLR, MoCo, CPC)

- Today: Self-Supervised Learning — Part 2:
  - Teacher-Student "feature reconstruction"
    - motivation, setting
    - methods: BYOL, DINO
  - Image Reconstruction
    - MAE - Masked Autoencoders

slide credit: Spyros Gidaris, Andrei Bursuc

Teacher-Student Feature "Reconstruction"

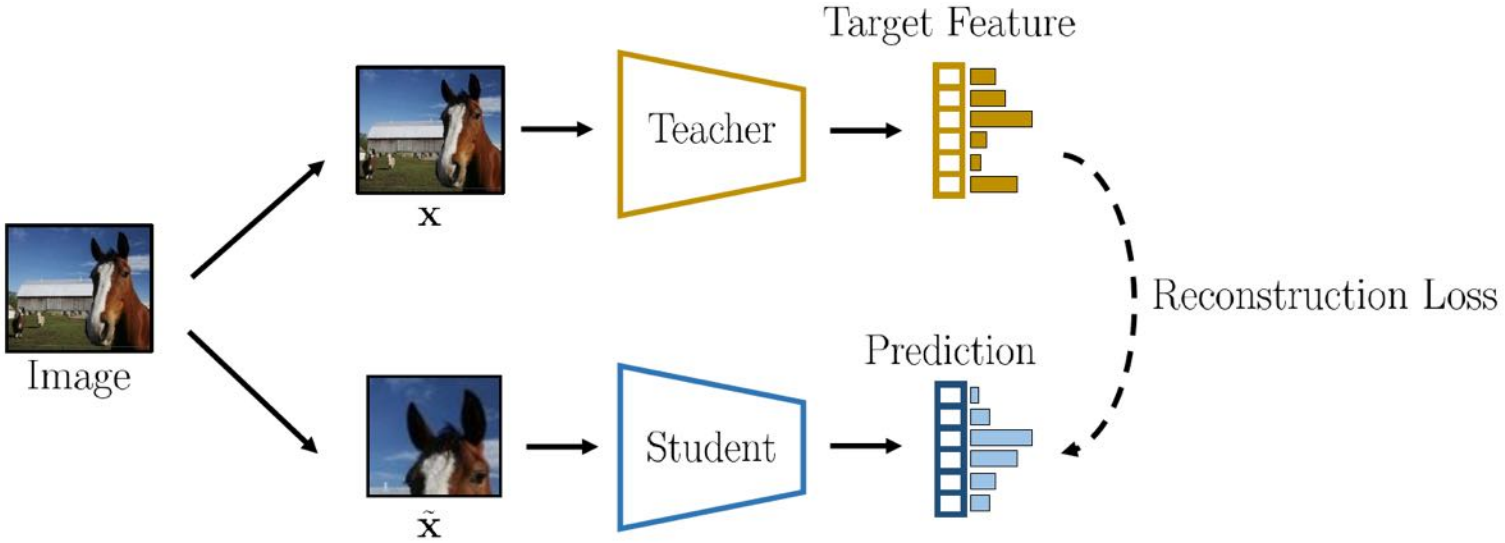


Teacher: generate a target feature vector from a given image

Student: predict this target, given as input a different random view of the same image

slide credit: Spyros Gidaris, Andrei Bursuc

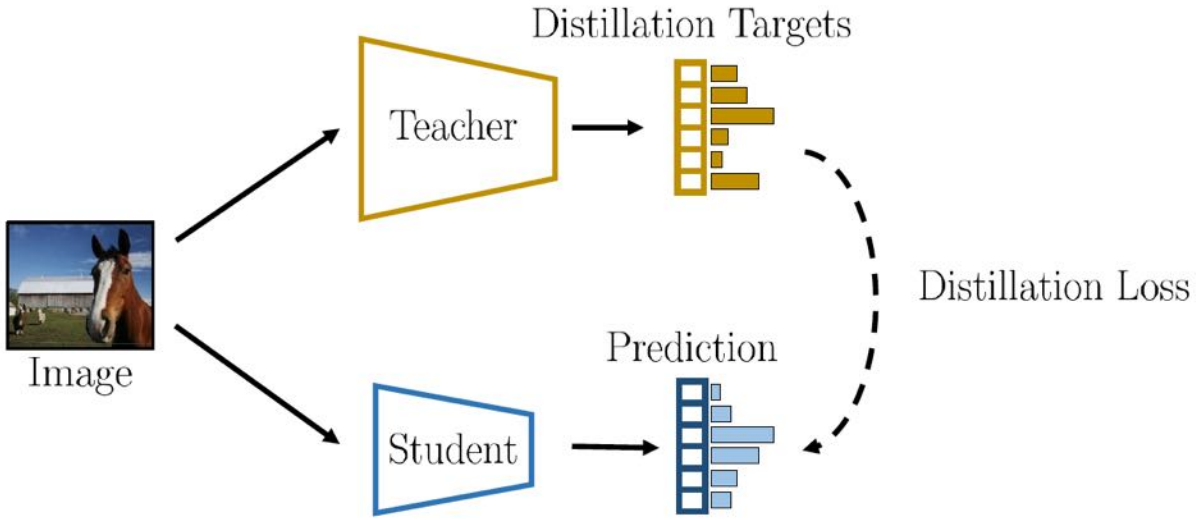
Teacher-Student Feature “Reconstruction”



- Goal: **focus on reconstructing high-level visual concepts** rid of “useless” image details
- Enforces **perturbation-invariant representations** without **requiring negative examples**

slide credit: Spyros Gidaris, Andrei Bursuc

Detour: Knowledge Distillation = Teacher-Student Approach for Model Compression



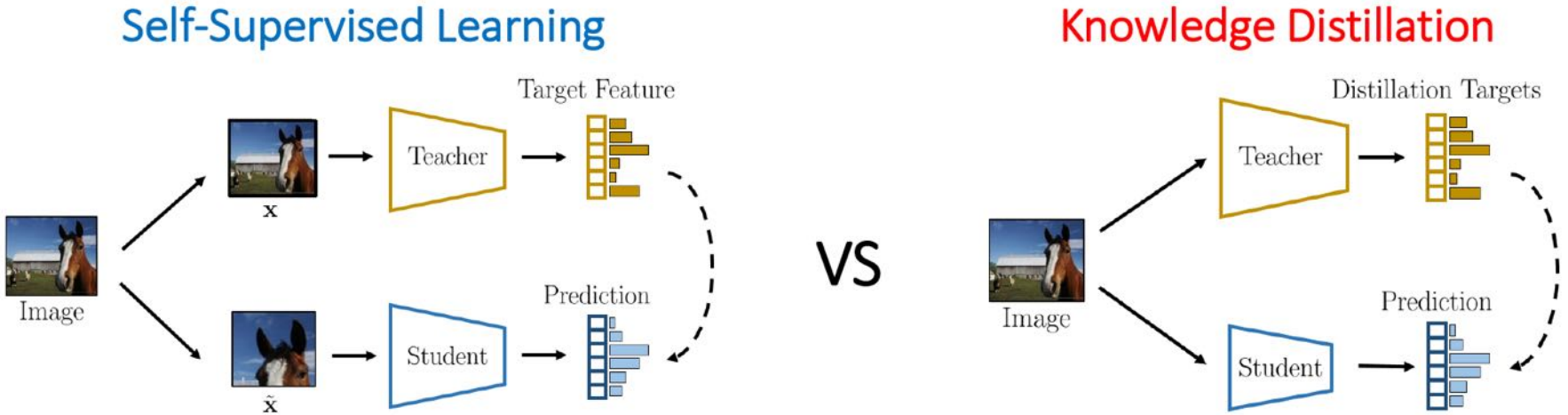
Goal: Distill the knowledge of a pre-trained teacher into a smaller student

- Commonly called **Knowledge Distillation**
- **Student:** trained to predict the teacher target when given the same input image
- Examples of targets: classification logits, intermediate features, attention maps, ...

"Distilling the knowledge in a neural network", Hinton et al, 2018

slide credit: Spyros Gidaris, Andrei Bursuc

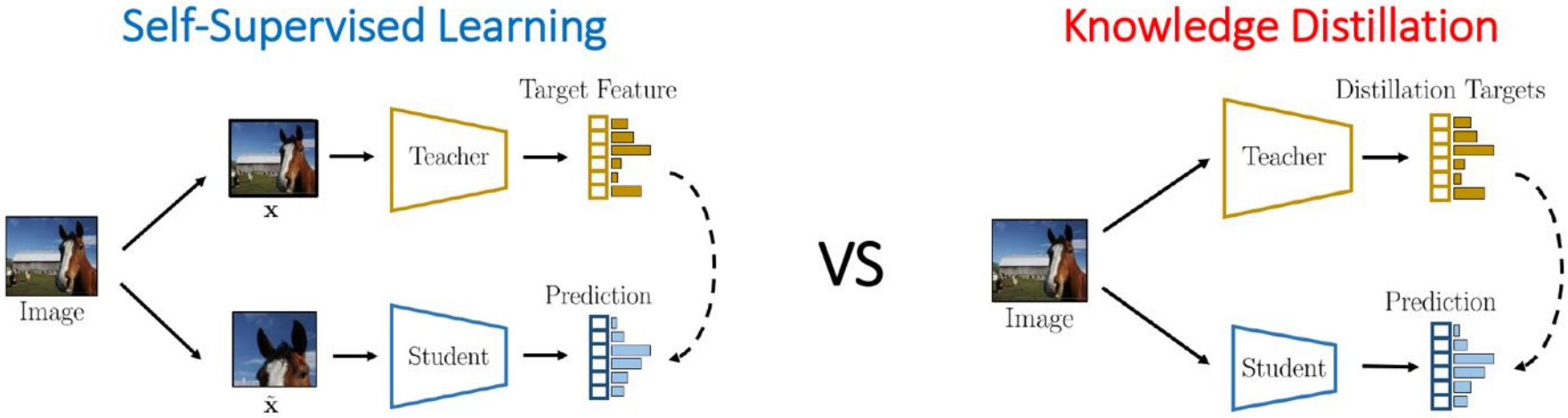
Teacher-Student Feature "Reconstruction" vs. Knowledge Distillation



- Access to a "good" teacher
- (Typically) For the same exactly input, the outputs should match.
- (Typically) Hopefully the student would reach the teacher
- (Typically) The student network is smaller

slide credit: Spyros Gidaris, Andrei Bursuc

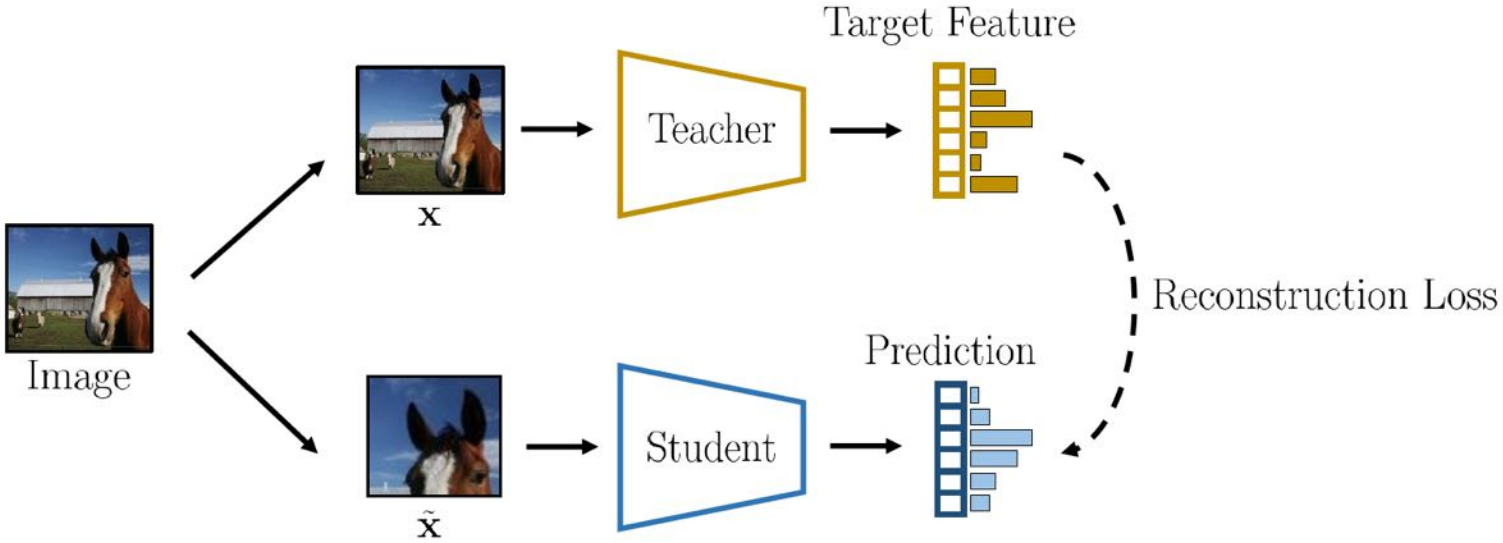
Teacher-Student Feature "Reconstruction" vs. Knowledge Distillation



- No access to a "good" teacher
- The student must predict the teacher output given a different version of the image
- The student MUST surpass the initial teacher
- Both networks are of the same size

slide credit: Spyros Gidaris, Andrei Bursuc

Teacher-Student Feature "Reconstruction"



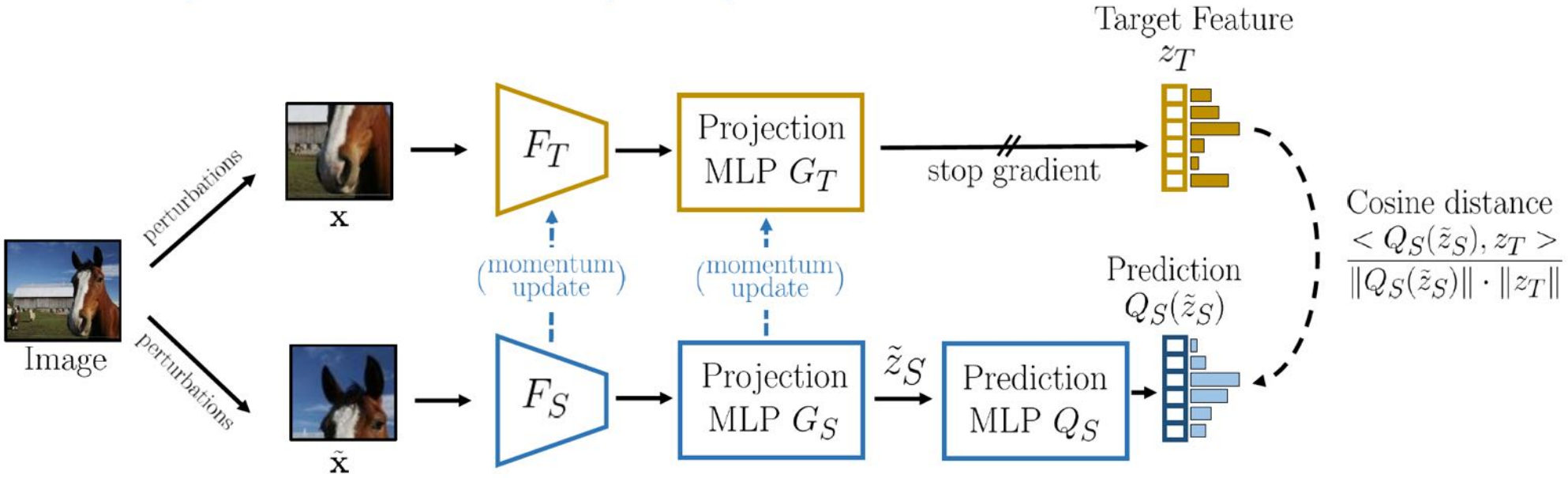
Key questions:

- What teacher to use?
- How to make the student surpass the teacher?
- What type of target features to use?

slide credit: Spyros Gidaris, Andrei Bursuc

Dynamic teacher-student feature “reconstruction” methods

Bootstrap Your Own Latent (BYOL)



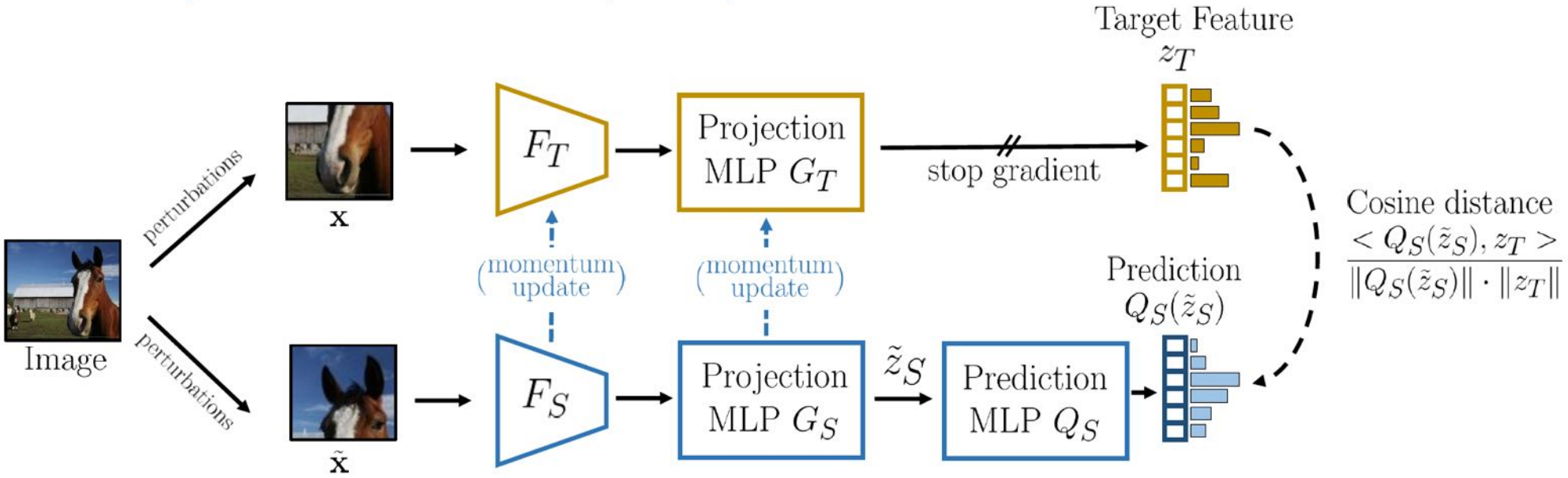
Feature reconstruction method:

- **Teacher:** extract a target feature vector from a random view of an image
- **Student:** predict this target, given as input a different random view of the same image
- **Symmetric loss:** from $\tilde{\mathbf{x}}$ predict the target of $\mathbf{x}$ and from $\mathbf{x}$ predict the target of $\tilde{\mathbf{x}}$

"Bootstrap Your Own Latent: a new approach to self-supervised learning", NeurIPs 2020

slide credit: Spyros Gidaris, Andrei Bursuc

Bootstrap Your Own Latent (BYOL)



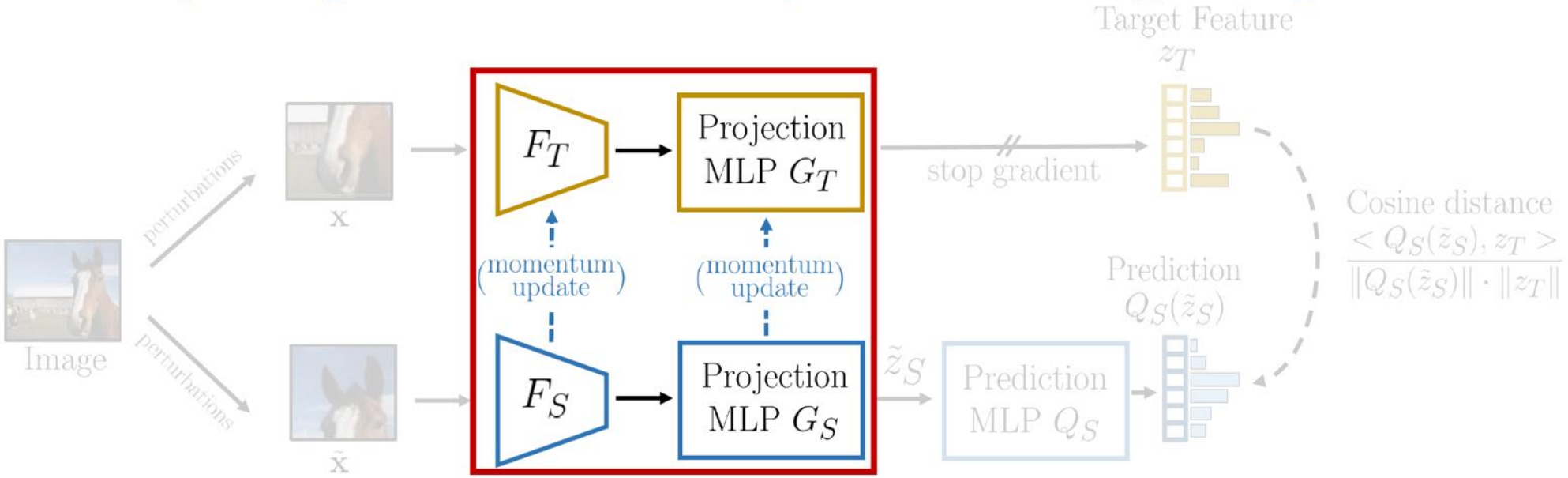
Bootstrap idea: builds a sequence of student representations of increasing quality

- Given a teacher, train a new enhanced student by predicting the teacher's features
- Iteratively apply this procedure by updating the teacher with the new student

"Bootstrap Your Own Latent: a new approach to self-supervised learning", NeurIPs 2020

slide credit: Spyros Gidaris, Andrei Bursuc

Online updating the teacher with exponential moving average



Use exponential moving average for online updating the teacher at each training step:

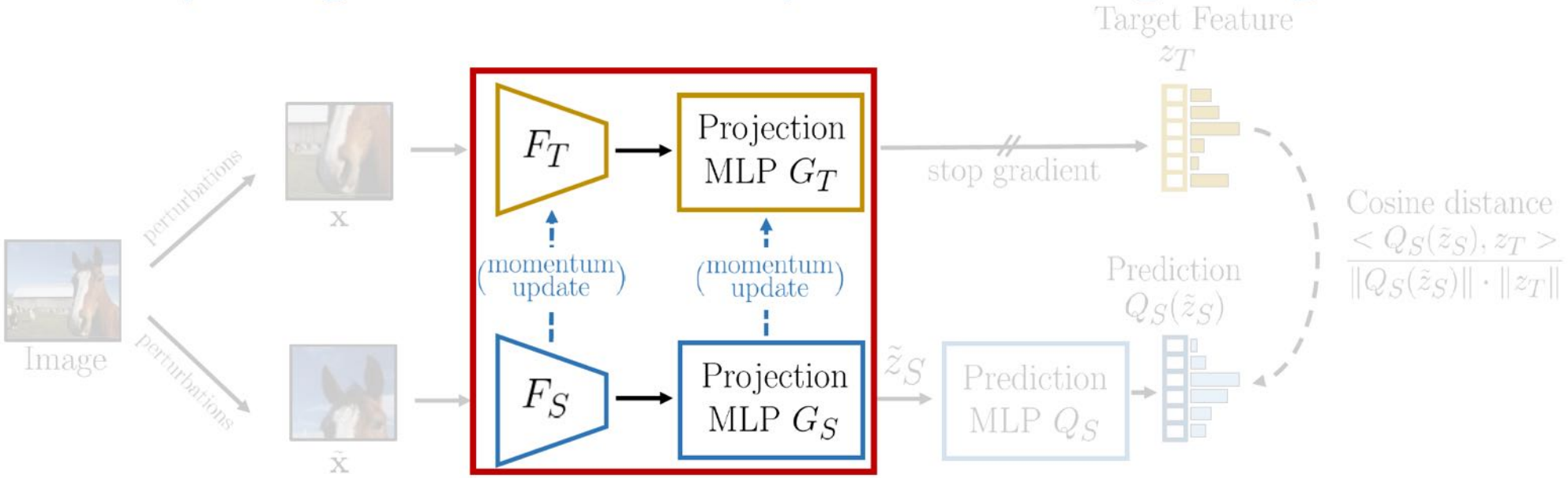
$$\theta_{\mathrm{T}}^{(t)} \leftarrow \alpha \cdot \theta_{\mathrm{T}}^{(t-1)} + (1-\alpha) \cdot \theta_{\mathrm{S}}^{(t)}$$

θ_{T} : teacher parameters

θ_{S} : student parameters

slide credit: Spyros Gidaris, Andrei Bursuc

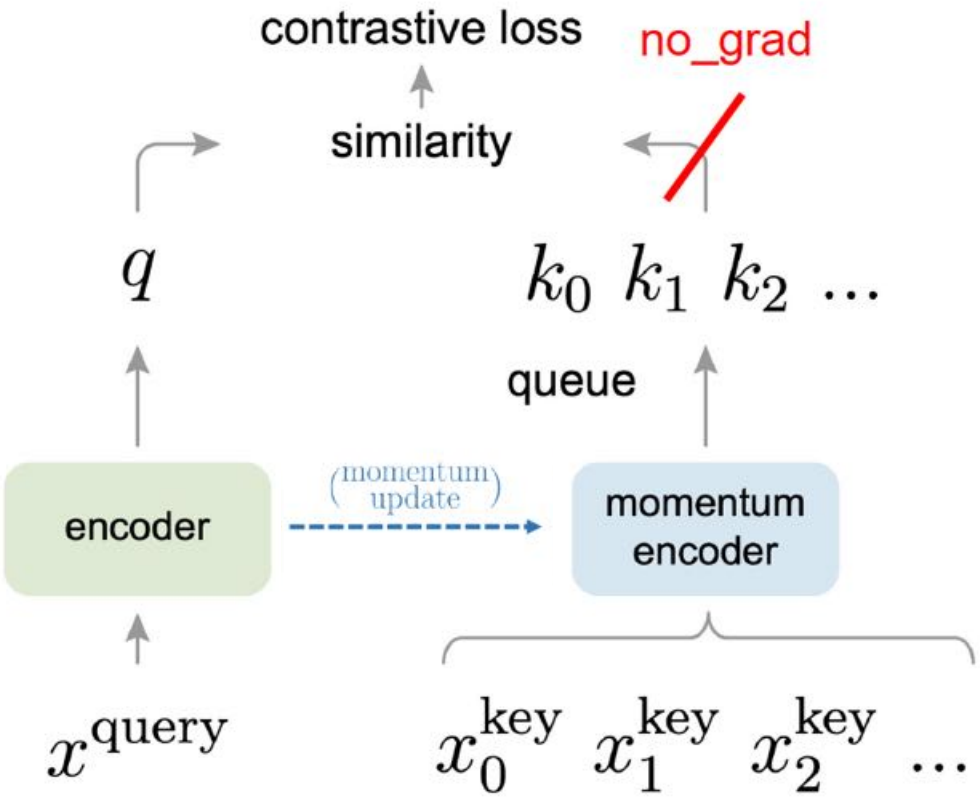
Online updating the teacher with exponential moving average



This type of teacher is typically called **momentum or mean teacher.**

slide credit: Spyros Gidaris, Andrei Bursuc

Detour: momentum / mean teacher in contrastive learning



MoCo exploits a momentum encoder network for maintaining a large and consistent dictionary of keys (positives + negatives examples) for contrastive learning.

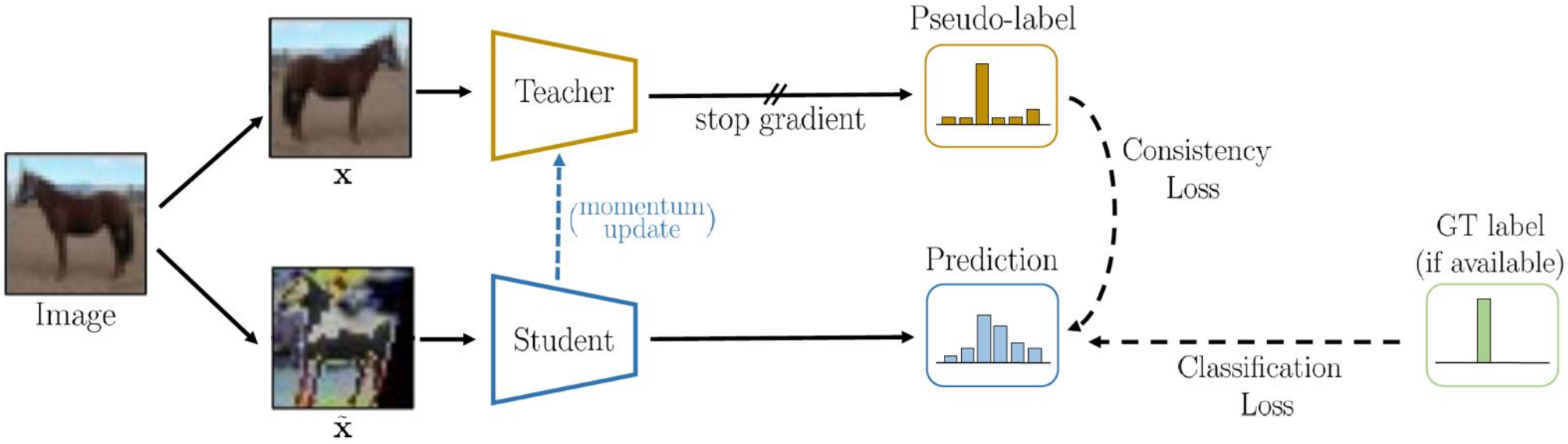
The key encoder is slowly progressing through the momentum update rules:

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}$$

"Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020

slide credit: Spyros Gidaris, Andrei Bursuc

Detour: mean / momentum teacher in semi-supervised learning



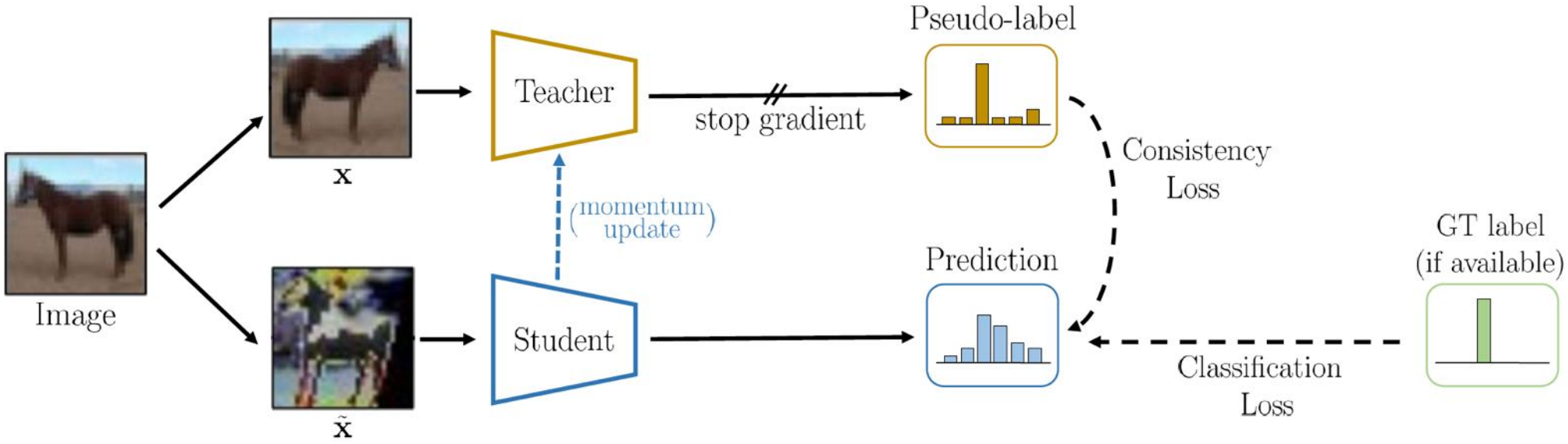
Teacher-student approaches are common in semi-supervised learning:

- **Teacher:** generate target classification predictions from an image
- **Student:** trained to predict this target given a different random view of the same image

"Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results", NeurIPs 2017

slide credit: Spyros Gidaris, Andrei Bursuc

Detour: mean / momentum teacher in semi-supervised learning



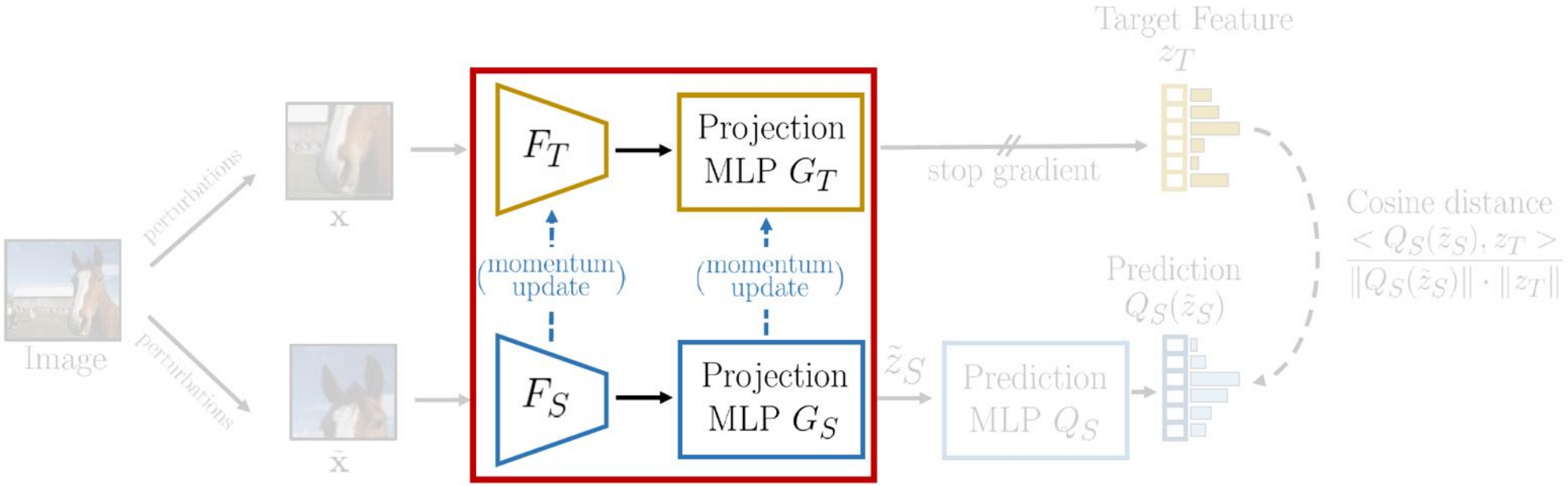
Mean teachers have been shown to improve the results:

- Similar to temporal ensembles of the student model but instead of averaging the predictions it averages the model weights
- More stable and accurate version of the student

"Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results", NeurIPs 2017

slide credit: Spyros Gidaris, Andrei Bursuc

Back to BYOL - Mean teacher for the feature reconstruction task

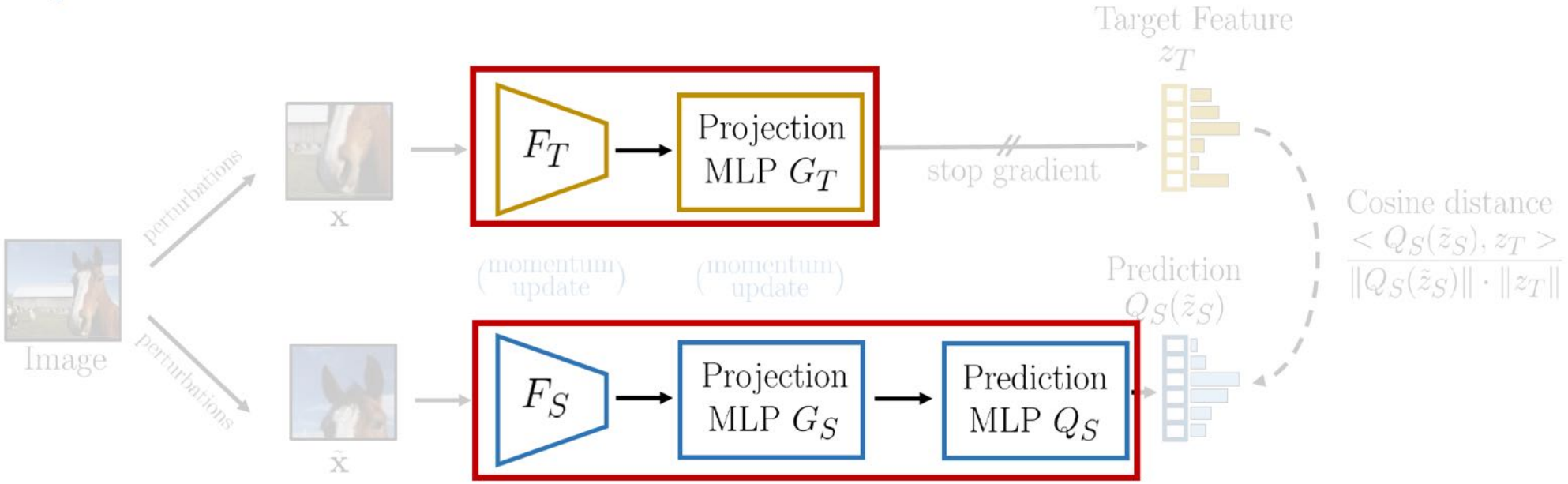


A mean teacher approach without any labels

- Offers stable but slowly evolving feature targets
- More efficient than using a fixed pre-training teacher that is updated only after the end of each training cycle

slide credit: Spyros Gidaris, Andrei Bursuc

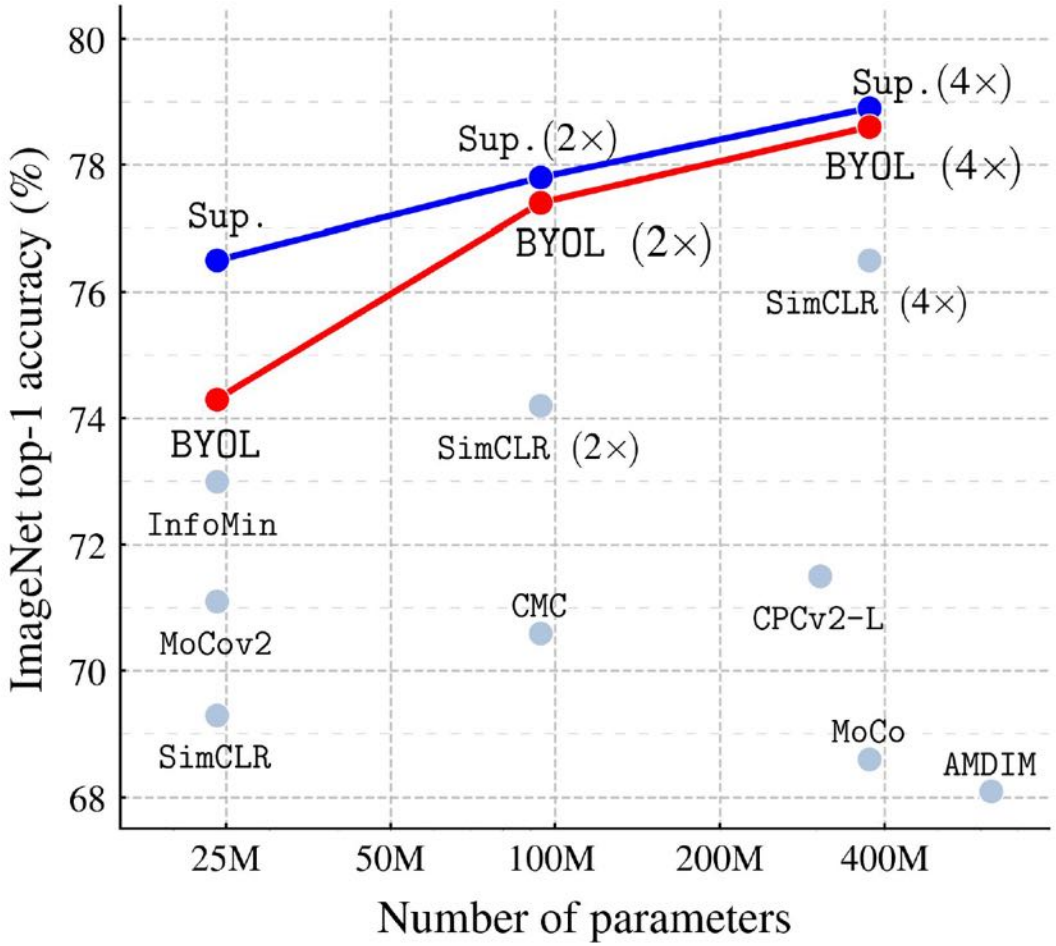
Asymmetric architecture



Asymmetric architecture: the student has an extra prediction MLP head

slide credit: Jean-Basien Grill, et al.

Transfer Using Linear Probe on BYOL Features

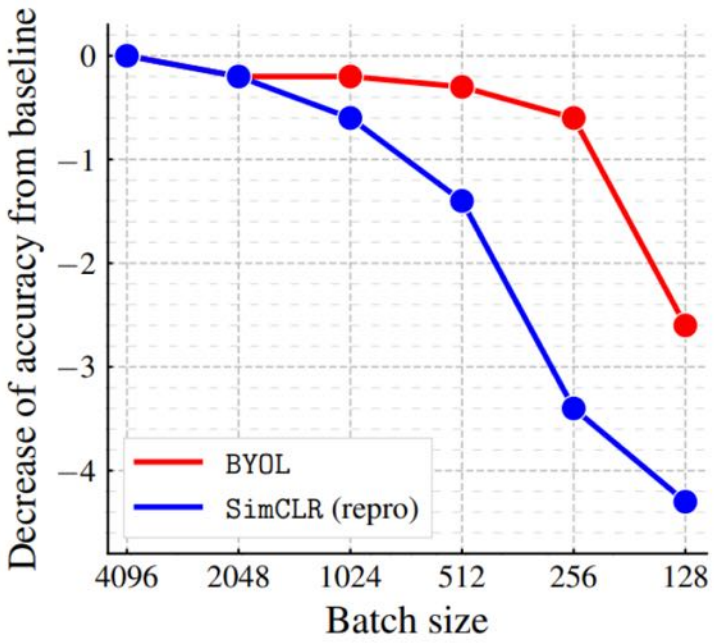


Note: these supervised baselines are from SimCLR (Chen & Hinton, ICML 2020)

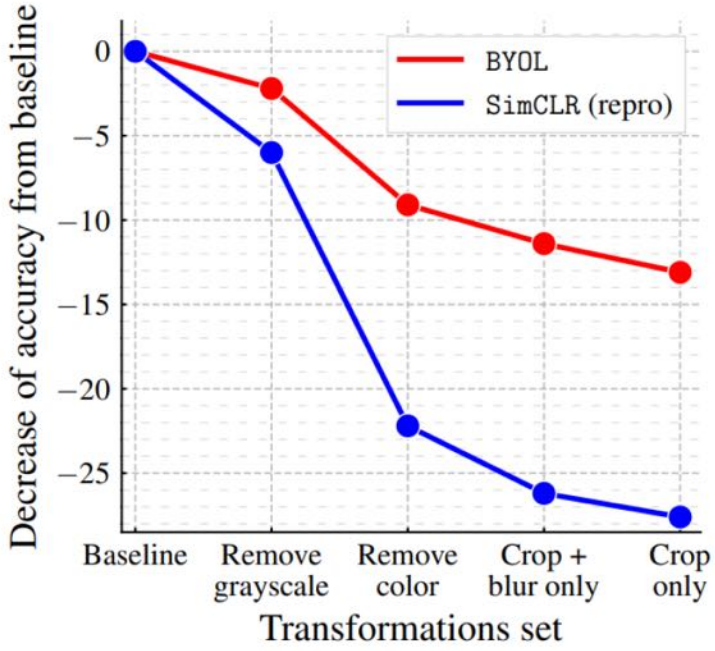
CPCv2: van den Oord et al., *Representation learning with contrastive predictive coding*. 2018
AMDIM: Bachman et al., *Learning representations by maximizing mutual information across views*. 2019
CMC: Tian et al.,*Contrastive multiview coding*. 2019.
MoCo: He et al., *Momentum contrast for unsupervised visual representation learning*. 2019
InfoMin: Tian et al., *What makes for good views for contrastive learning*. 2020
MoCov2: Jain et al., *Improved baselines with momentum contrastive learning*. 2020
SimCLR: Chen et al., *A simple framework for contrastive learning of visual representations*. 2020

slide credit: Spyros Gidaris, Andrei Bursuc

BYOL vs Contrastive methods (SimCLR)



(a) Impact of batch size



(b) Impact of progressively removing transformations

- BYOL **does not require negative examples** as the contrastive method SimCLR
- **More robust** to the choice of image augmentations and the batch-size
- Cropping is more important for BYOL and color jittering more important for SimCLR

slide credit: Spyros Gidaris, Andrei Bursuc

Key question: Why it avoids feature collapse?

slide credit: Spyros Gidaris, Andrei Bursuc

Why it avoids feature collapse?

Batch Normalization (BN) in BYOL implicitly causes a form contrastive learning: collapse is avoided because all samples in the mini-batch cannot take on the same value after BN

- suggested in "Understanding self-supervised and contrastive learning with BYOL", Fetterman et al).

However, according to BYOL authors **"BYOL works even without batch statistics"**

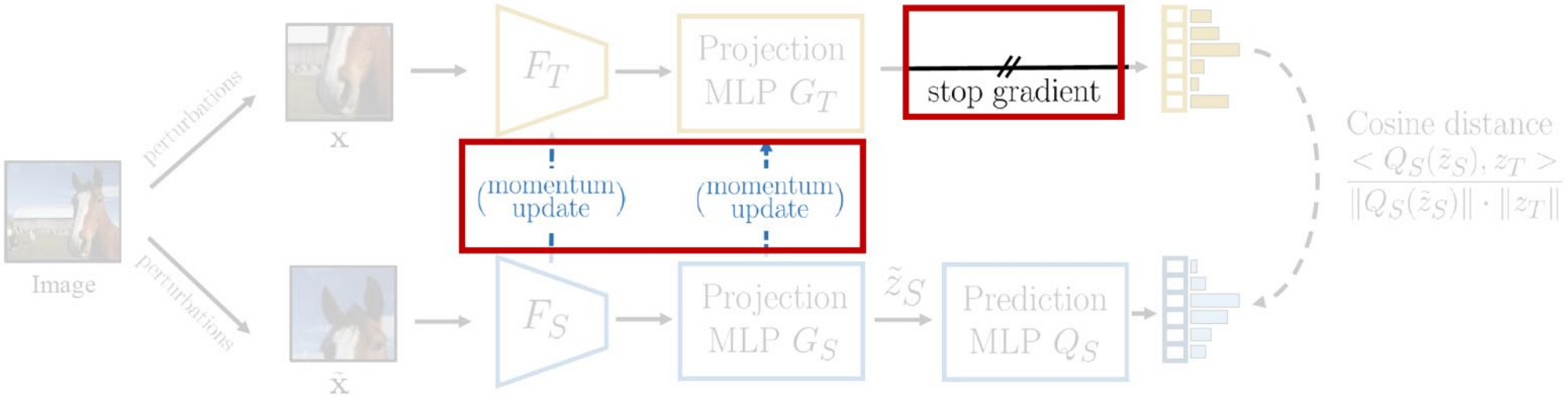
- Either by better tuning the network initialization
- Or replacing BN with Group Normalization and Weight Standardization (GN + WS)

Table 2: top-1 accuracy with linear evaluation on ImageNet

BYOL variant	Vanilla BN	No BN	Modified init.	GN + WS
Uses batch statistics	Yes	No	No	No
Top-1 accuracy (%)	74.3	0.1	65.7	73.9

slide credit: Spyros Gidaris, Andrei Bursuc

Why it avoids feature collapse?

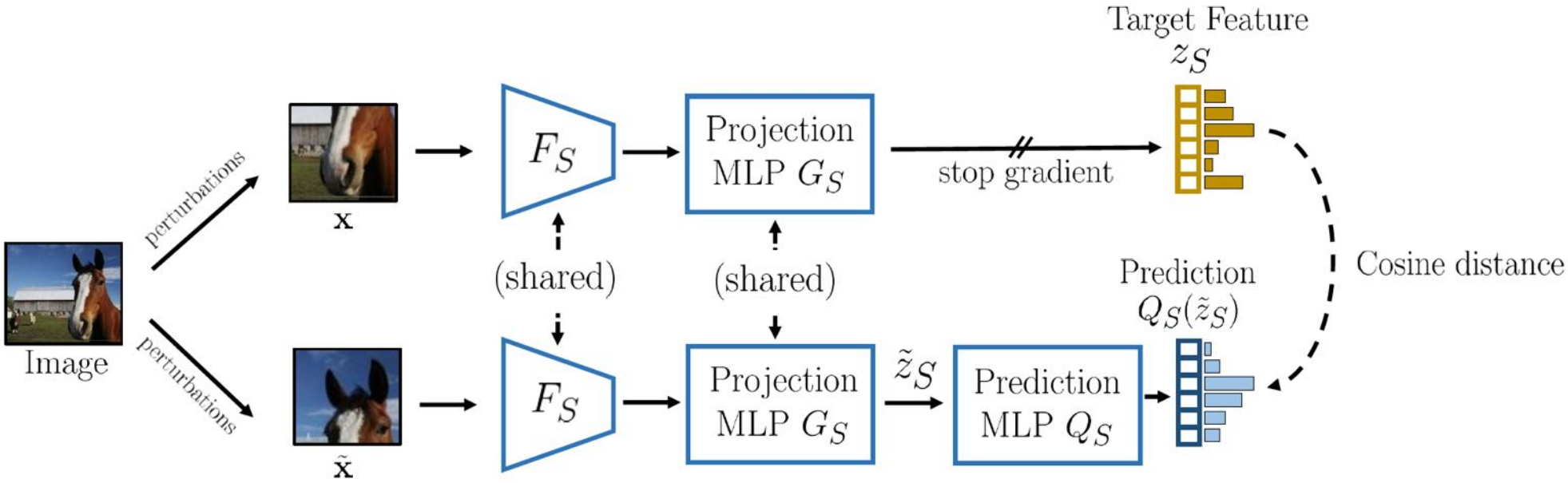


The teacher parameter updates ARE NOT NECASSARILY in the direction of minimizing the loss, i.e., BYOL does not explictly optimize the loss w.r.t. the teacher parameters.

"Bootstrap Your Own Latent: a new approach to self-supervised learning", NeurIPs 2020

slide credit: Spyros Gidaris, Andrei Bursuc

SimSiam

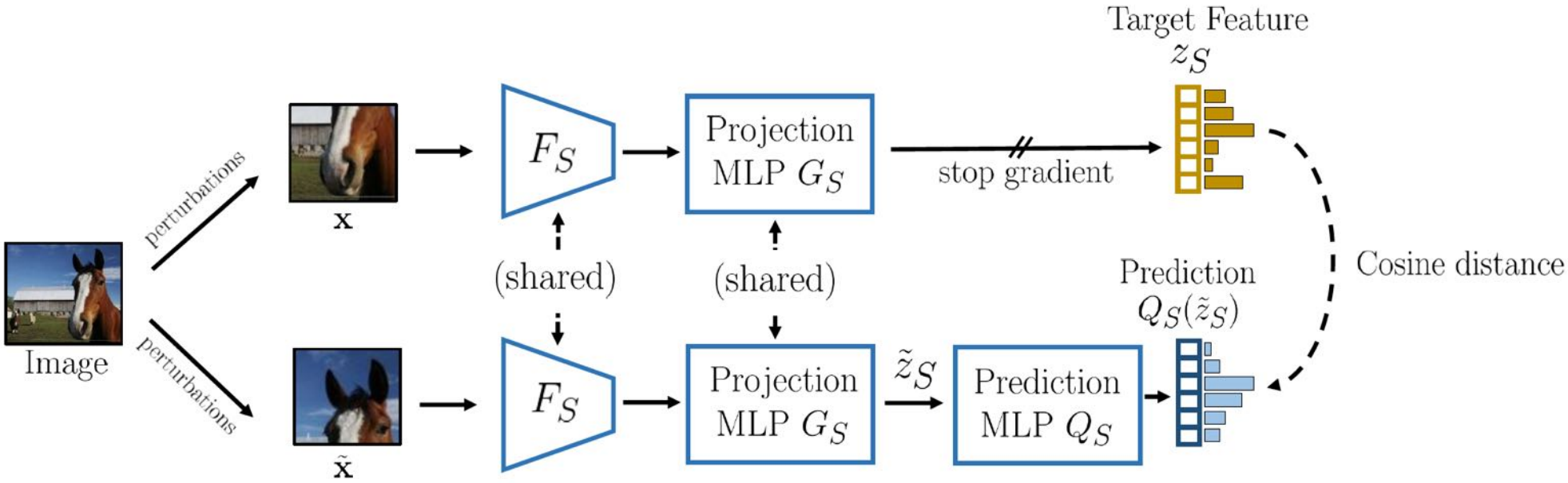


SimSiam: BYOL without the momentum teacher (the teacher is identical to the student)

"Exploring Simple Siamese Representation Learning", CVPR 2021

slide credit: Spyros Gidaris, Andrei Bursuc

SimSiam



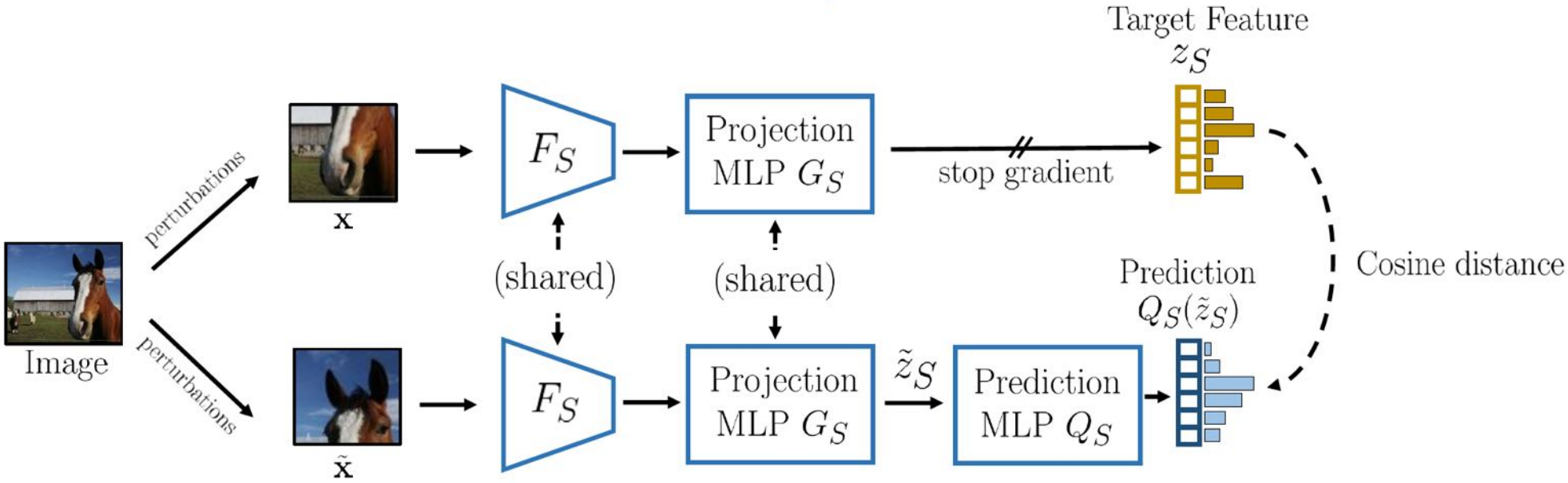
Momentum teacher: **improves performance but not necessary for avoiding feature collapse**

method	momentum encoder	100 ep	200 ep	400 ep	800 ep
BYOL	✓	66.5	**70.6**	**73.2**	**74.3**
SimSiam		**68.1**	70.0	70.8	71.3

epochs

slide credit: Spyros Gidaris, Andrei Bursuc

SimSiam: When it avoids feature collapse?



Without **stop-gradient** or the **predictor head** the network is trained to minimize the reconstruction loss for both image views at the same time, leading to constant features

	pred. MLP h	acc. (%)
baseline	lr with cosine decay	67.7
(a)	no pred. MLP	0.1

Table 1. **Effect of prediction MLP**

	acc. (%)
w/ stop-grad	67.7±0.1
w/o stop-grad	0.1

max planck institut informatik

SIC Saarland Informatics Campus

DINO: Self-Distillation with No Labels or Emerging Properties in Self-Supervised ViTs

Caron et al. ICCV'21

easy-to-read paper

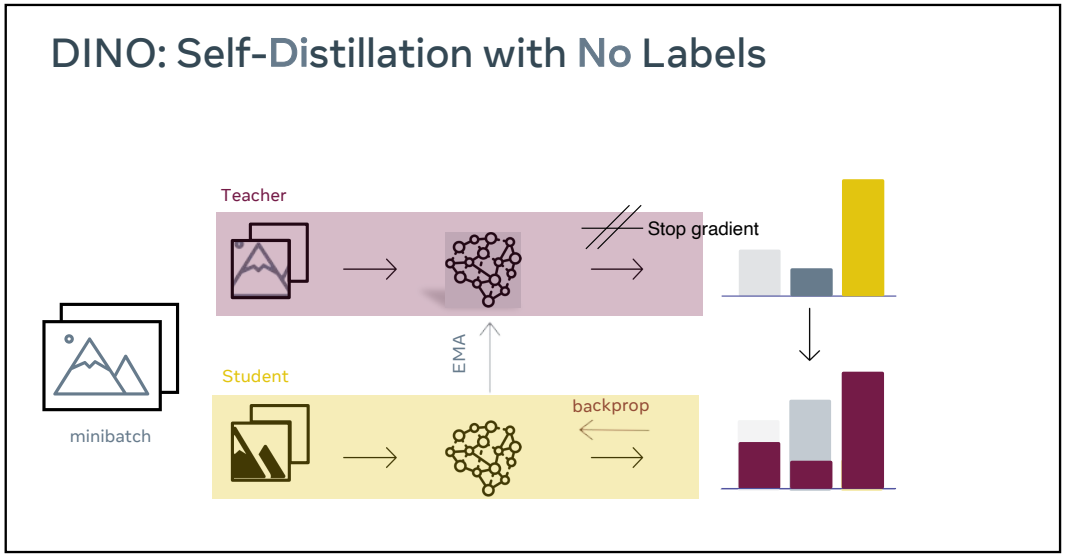
slide credit: Moritz Boehle

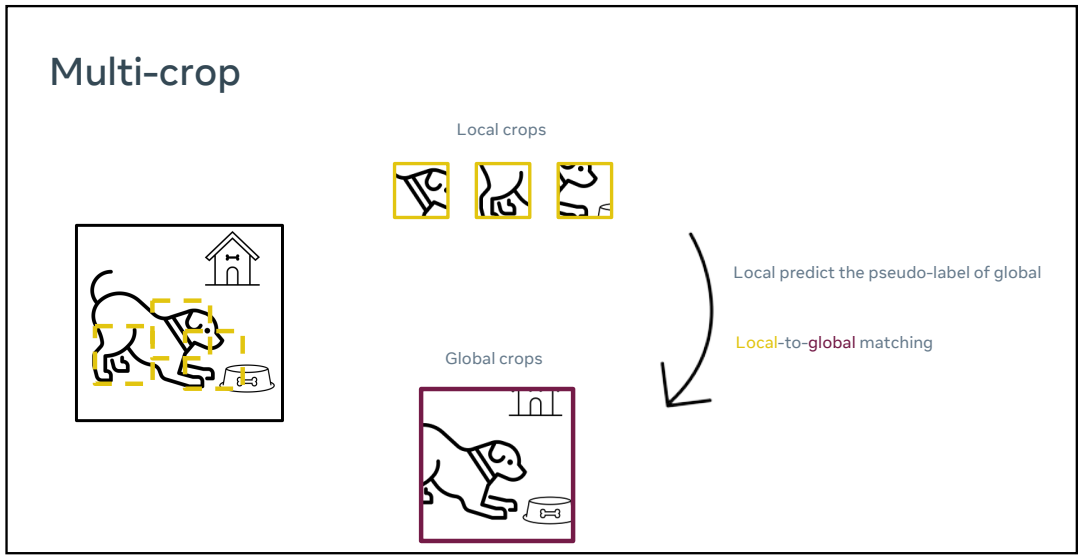
Source: https://github.com/facebookresearch/dino

slide credit: Moritz Boehle

DINO-Pipeline: Key Ingredients

- Student: predict teacher output





1. Extract small and large crops
2. Predict **teacher output on global crops** given **local crops**
3. **Updates:** Student—SGD I Trainer—EMA of student
   EMA = Exponential Moving Average

Source: https://gidariss.github.io/self-supervised-learning-cvpr2021

slide credit: Moritz Boehle

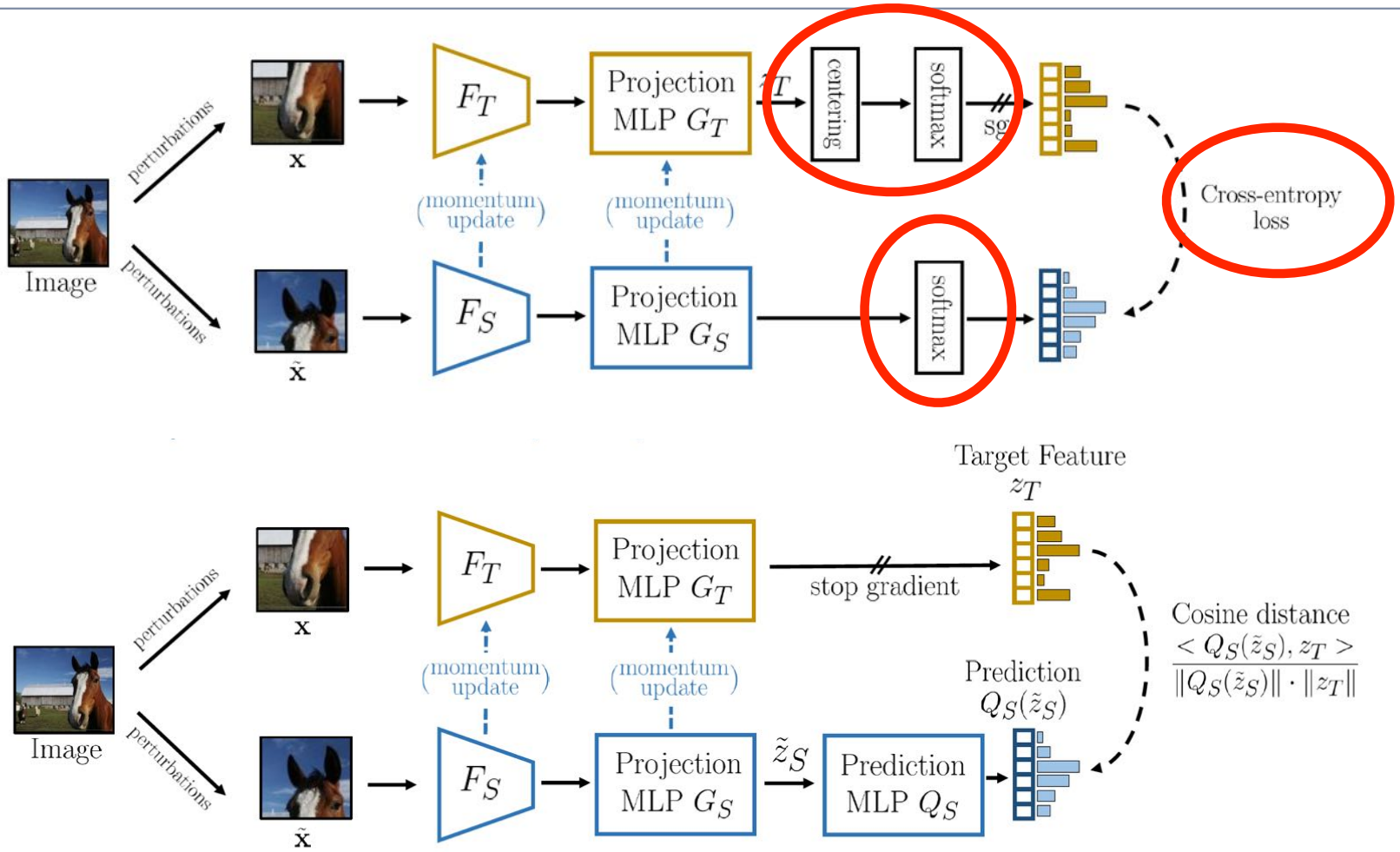
DINO-Pipeline: Summary

In short: Self-defining classification task, main ingredients: EMA and multi-crop, trained with CE-Loss

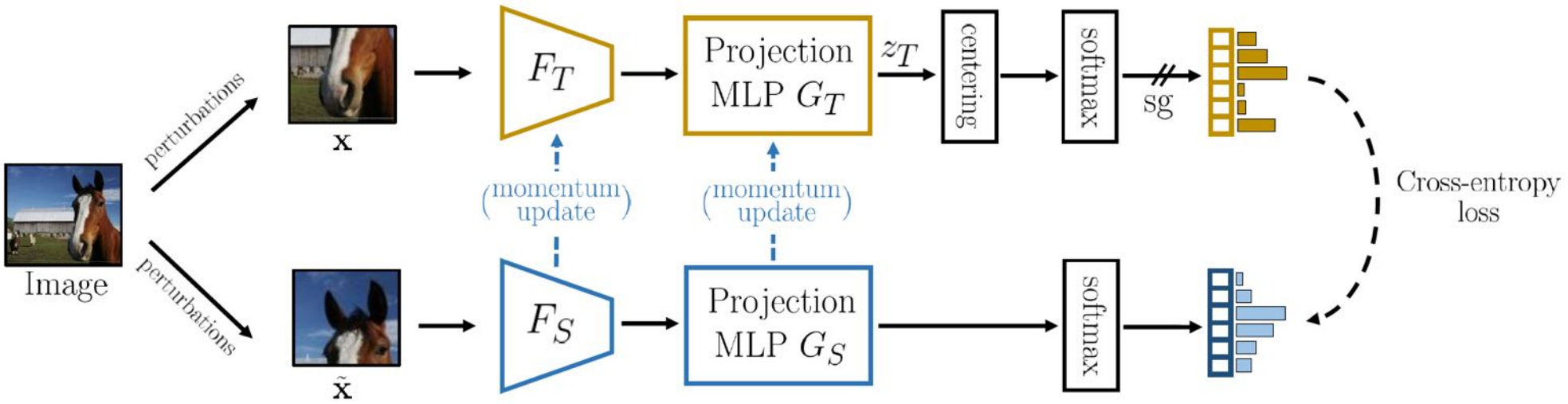
Source: https://github.com/facebookresearch/dino

slide credit: Spyros Gidaris, Andrei Bursuc

DINO (top) vs. BYOL (bottom)



DINO



No prediction head - post-processing of teacher outputs to avoid feature collapse:

- **Centering by subtracting the mean feature:** prevents collapsing to constant 1-hot targets
- **Sharpening by using low softmax temperature:** prevents collapsing to a uniform target vector

slide credit: Moritz Boehle

Results: Testing the Representations

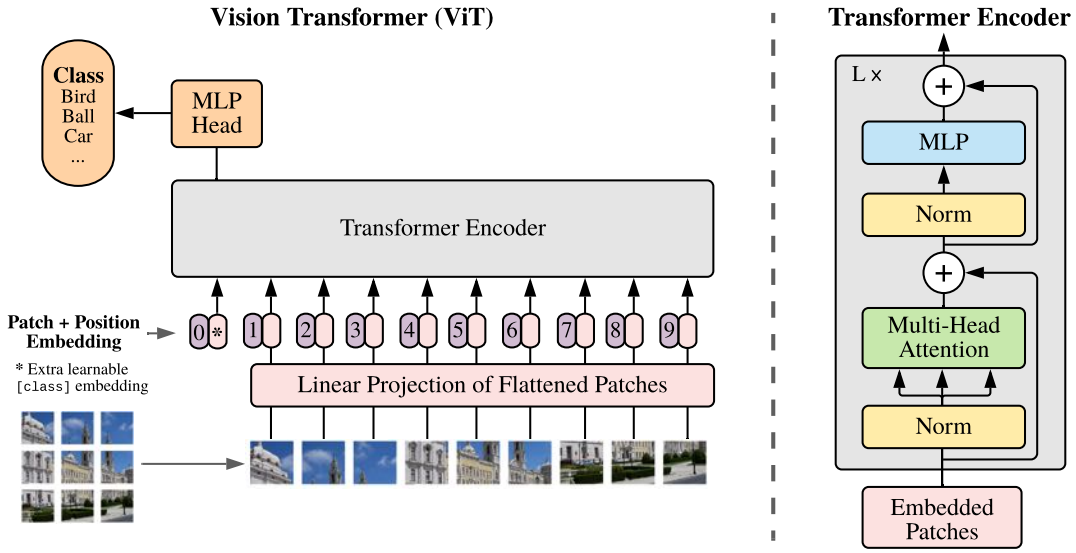
- Evaluation
  - Training a linear probe / classifier only
  - k-NN evaluation
- DINO works well across architectures
  - But best with ViTs

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**

slide credit: Moritz Boehle

DINO + Vision Transformer

- DINO is independent of architecture
  - However, developed to improve SSL performance of ViTs
- Advantage: visualise attention
- "Emerging properties":
  - attention segments image well

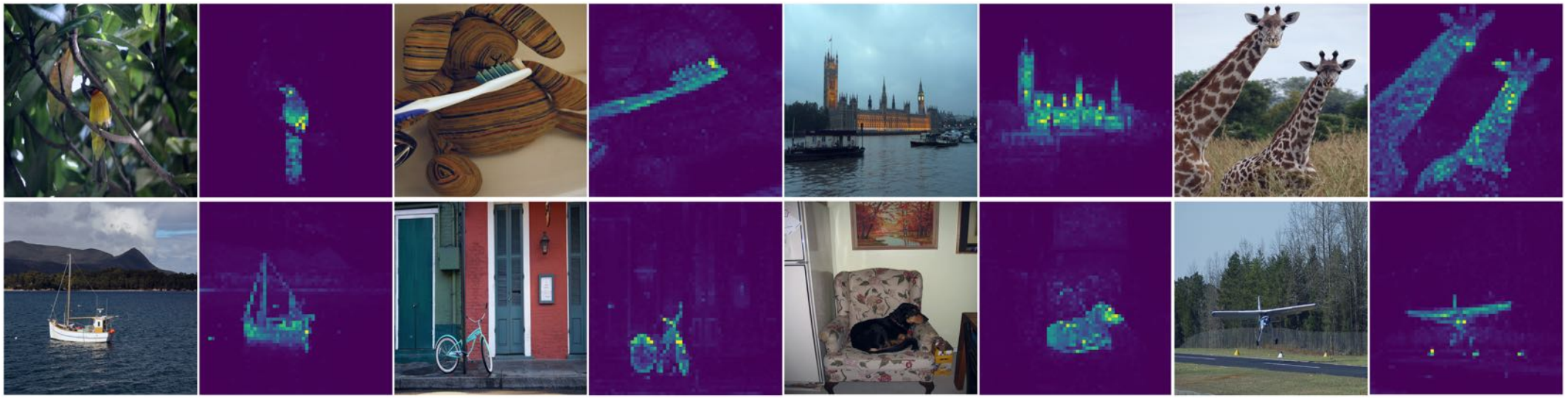


Source: An Image ist worth 16x16 words (Dosovitskiy et al., 2021)

slide credit: Moritz Boehle

Testing the Attention Maps

- Despite no supervision, attention segments objects well
  - showing the attention of a single attention head at the end of the network



slide credit: Moritz Boehle

Testing the Attention Maps

- Despite no supervision, attention segments objects well
  - showing the attention of a single attention head at the end of the network
- *Because* of no (classification) supervision?

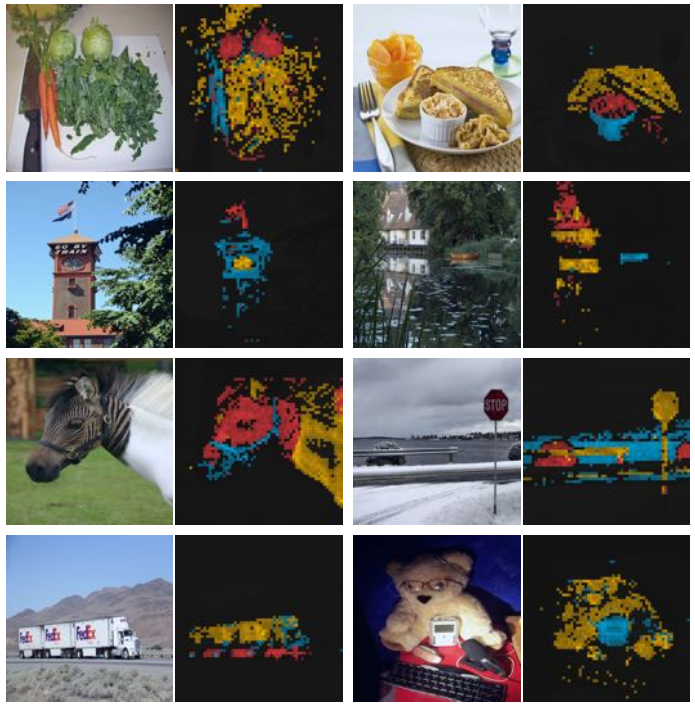
Supervised

DINO

slide credit: Moritz Boehle

Testing the Attention Maps

- Despite no supervision, attention segments objects well
  - showing the attention of a single attention head at the end of the network
- *Because* of no (classification) supervision?
- Different heads might also collect information from different parts
- Works well for object tracking in videos



slide credit: Spyros Gidaris, Andrei Bursuc

DINO

Method	Mom.	Loss	Pred.	k-NN	Lin.
DINO	✓	CE	✗	72.8	76.1
	✗	CE	✗	0.1	0.1
	✓	MSE	✗	52.6	62.4
	✓	CE	✓	71.8	75.6
BYOL	✓	MSE	✓	66.6	71.4

- **Loss:** Cross-Entropy (CE) instead of Mean-Squared Error (MSE)
- **Momentum teacher:** avoid collapsing
- **Better without predictor**

slide credit: Spyros Gidaris, Andrei Bursuc

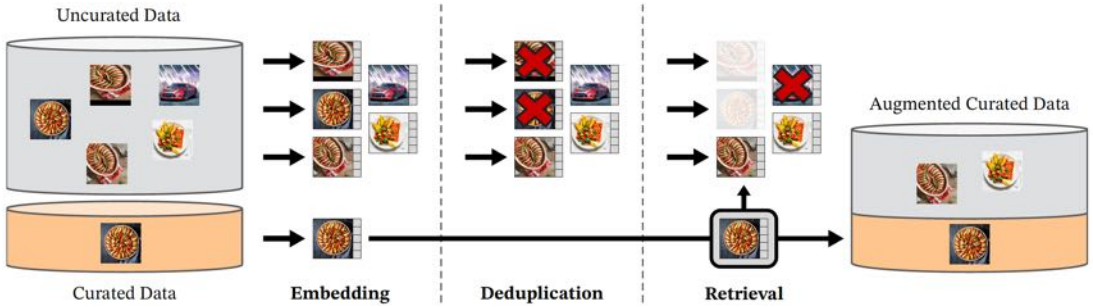
Conclusions

- Feature "reconstruction" self-supervised methods are gaining increased attention
- Manage to learn SOTA self-supervised representations without requiring negatives
  - Surpassing even supervised representations
- However, it's not entirely clear why they avoid feature collapse
- Recent trends: mid-way between contrastive and feature reconstruction
  - "Whitening for self-supervised representation learning", arXiv 2020
  - "Barlow Twins: self-supervised learning via redundancy reduction", ICML 2021
  - "VICReg: Variance-Invariance-Covariance Regularization for self-supervised learning", arXiv 2021
  - …

https://arxiv.org/pdf/2304.07193

DINOv2 — state of the art features today :)

- DINOv2 — vs. DINO — major changes (high level ;-)
  - lots of “engineering” to obtain better “foundational features”
  - building large, curated, and diverse dataset to train models



  - algorithmic / technical improvements
    - additional regularization terms that help training (otherwise unstable training)
  - strong, lightweight models obtained with model distillation
    - first train larger models (also requiring more powerful hardware)
    - then use knowledge distillation to compress the larger models into performant smaller ones

Overview of Today's Lecture

- Last Time: Self-Supervised Learning — Part 1:
  - Motivation of Self-Supervised Learning
  - Pretext tasks from image transformations (e.g. rotation, inpainting, coloring)
  - Contrastive representation learning (SimCLR, MoCo, CPC)

- Today: Self-Supervised Learning — Part 2:
  - Teacher-Student "feature reconstruction"
    - motivation, setting
    - methods: BYOL, DINO
  - Image Reconstruction
    - MAE - Masked Autoencoders

Masked Auto-Encoders as Scalable Vision Learners

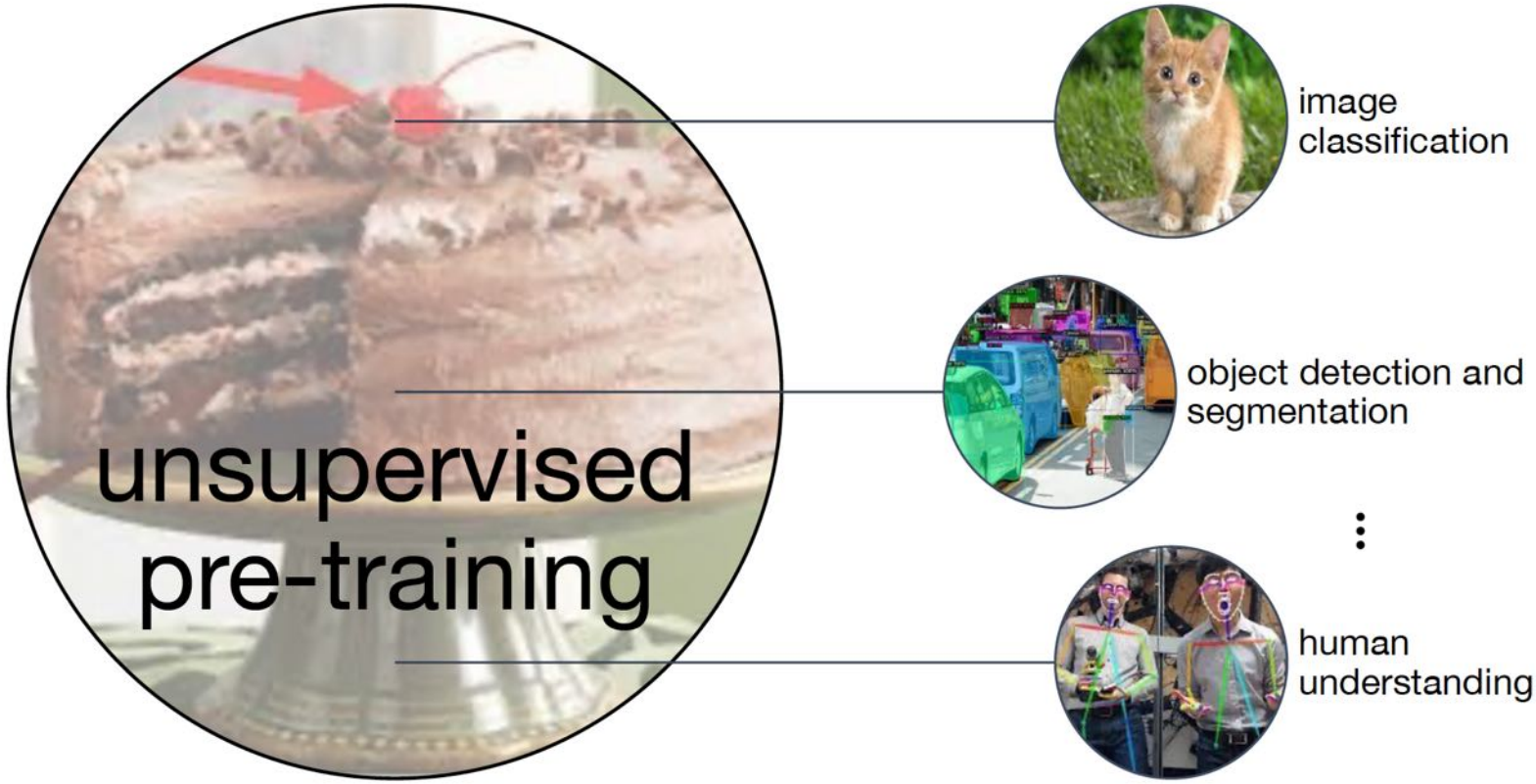
Xinlei Chen

ECCV 2022 tutorial on self-supervised representation learning in computer vision

facebook
Artificial Intelligence Research

Self-Supervised Learning

- Pre-train representations without labels for downstream tasks



slide credit: Xinlei Chen

[Devlin et al, NAACL 2019] [He et al, CVPR 2022]

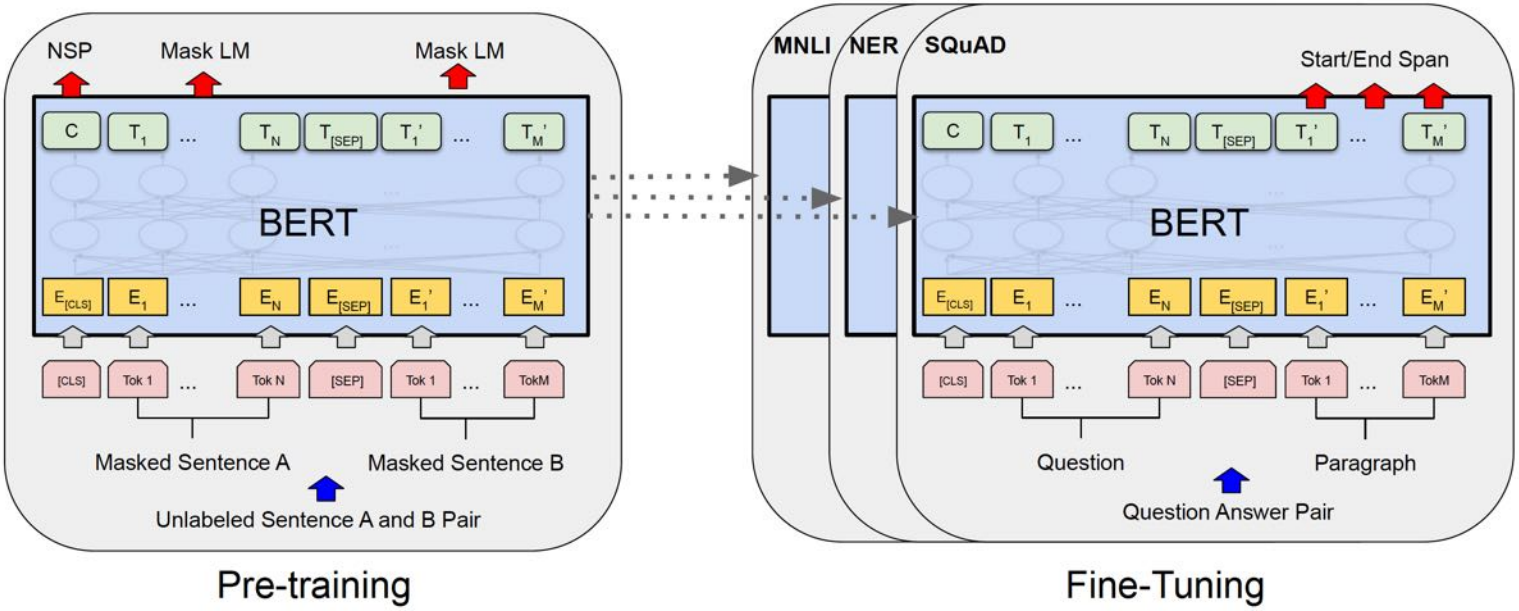
What is MAE?

- Very simple method, but highly effective
- BERT-like algorithm, but with crucial design changes for vision
- Intriguing properties – better scalability and more from analysis

slide credit: Xinlei Chen

BERT: Pretraining of Bidirectioanal Transformers for Language Understanding @NAACL 2019 — https://arxiv.org/abs/1810.04805

- Pretraining & Finetuning
- Two Pretraining Losses:
  - Mask LM = Masked Language Model: prediction of words masked out (without direction !)
  - NSP = Next Sentence Prediction (does sentence B follow sentence A)

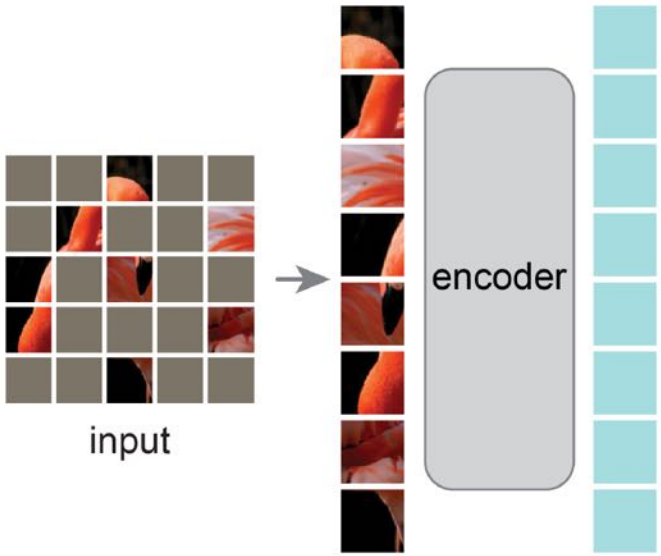


How MAE Works?

Random masking

slide credit: Xinlei Chen

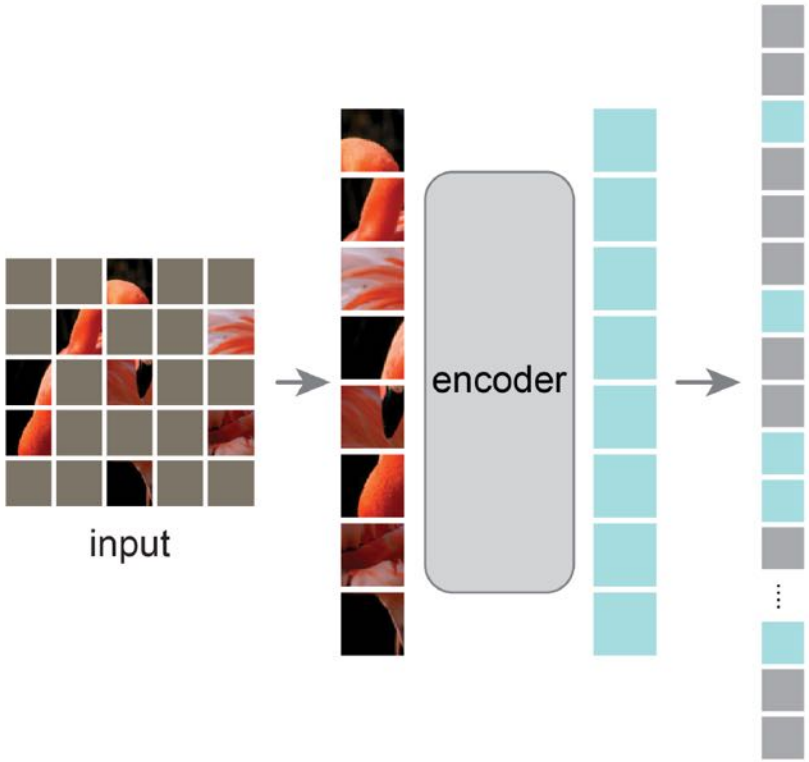
How MAE Works?



Encode visible patches

slide credit: Xinlei Chen

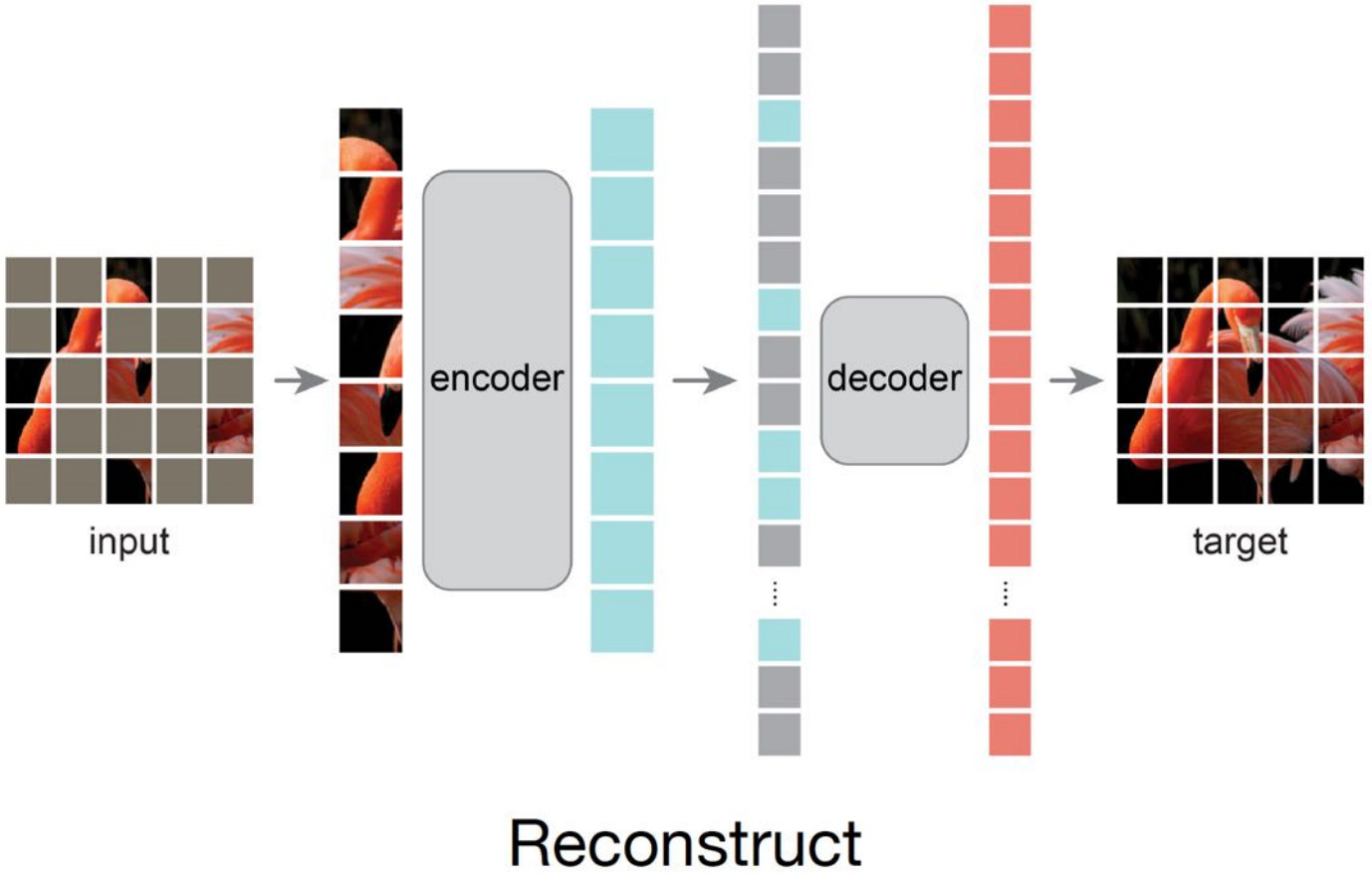
How MAE Works?



Add mask tokens

slide credit: Xinlei Chen

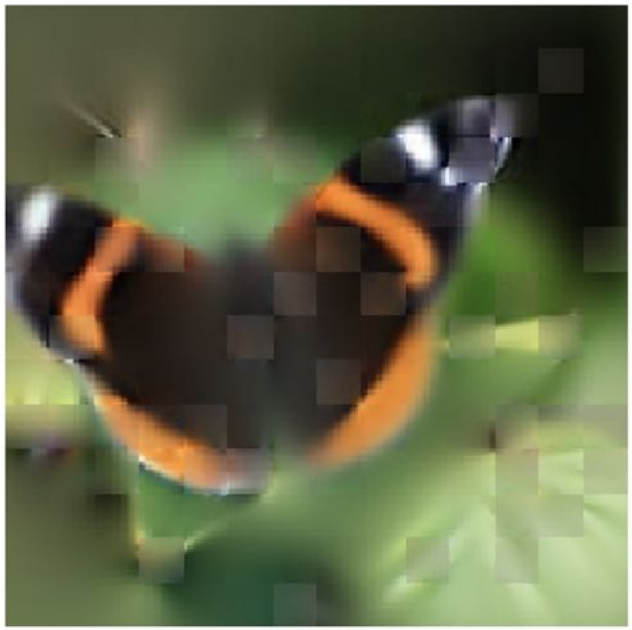
How MAE Works?



slide credit: Xinlei Chen

MAE Reconstruction Example

Masked input: 80%

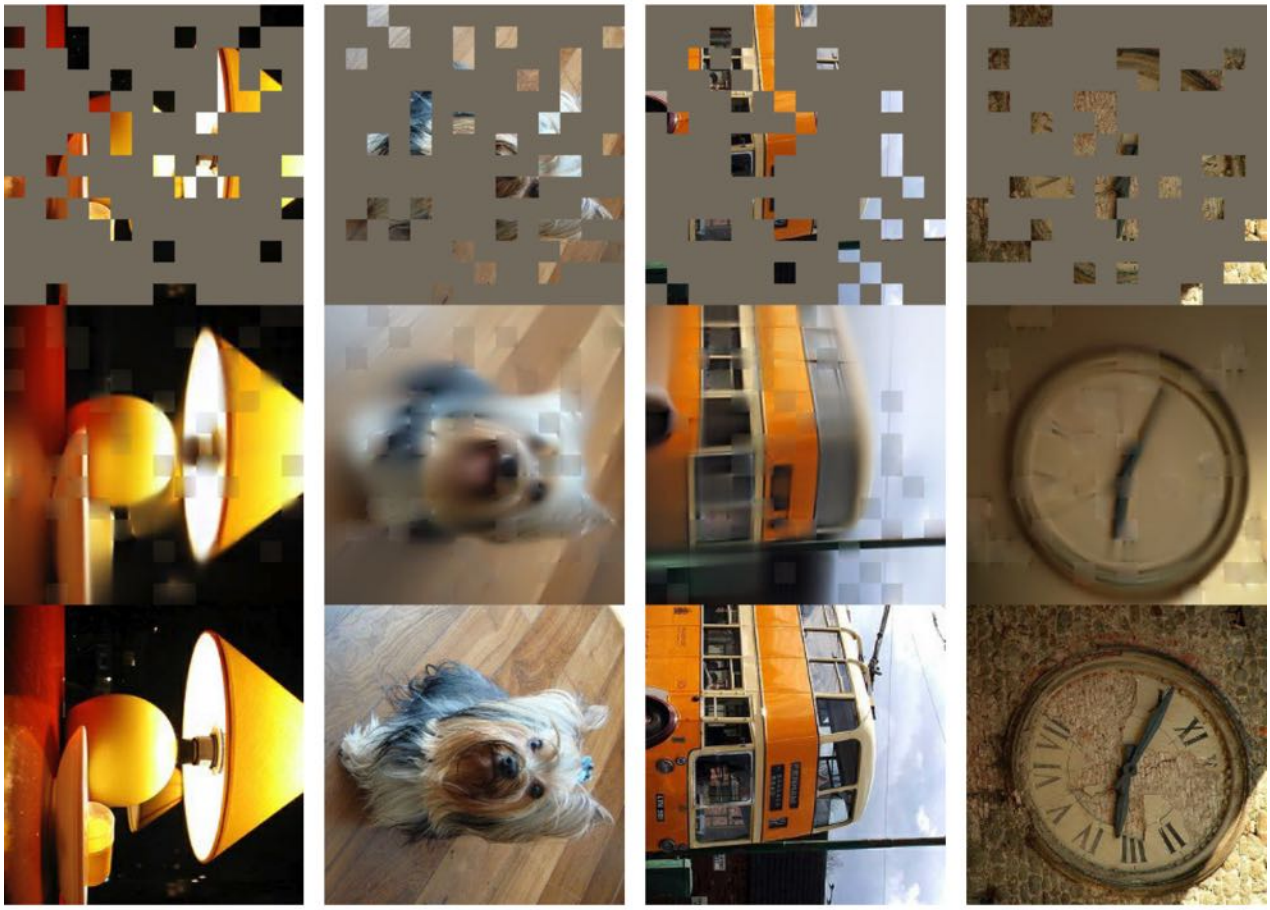


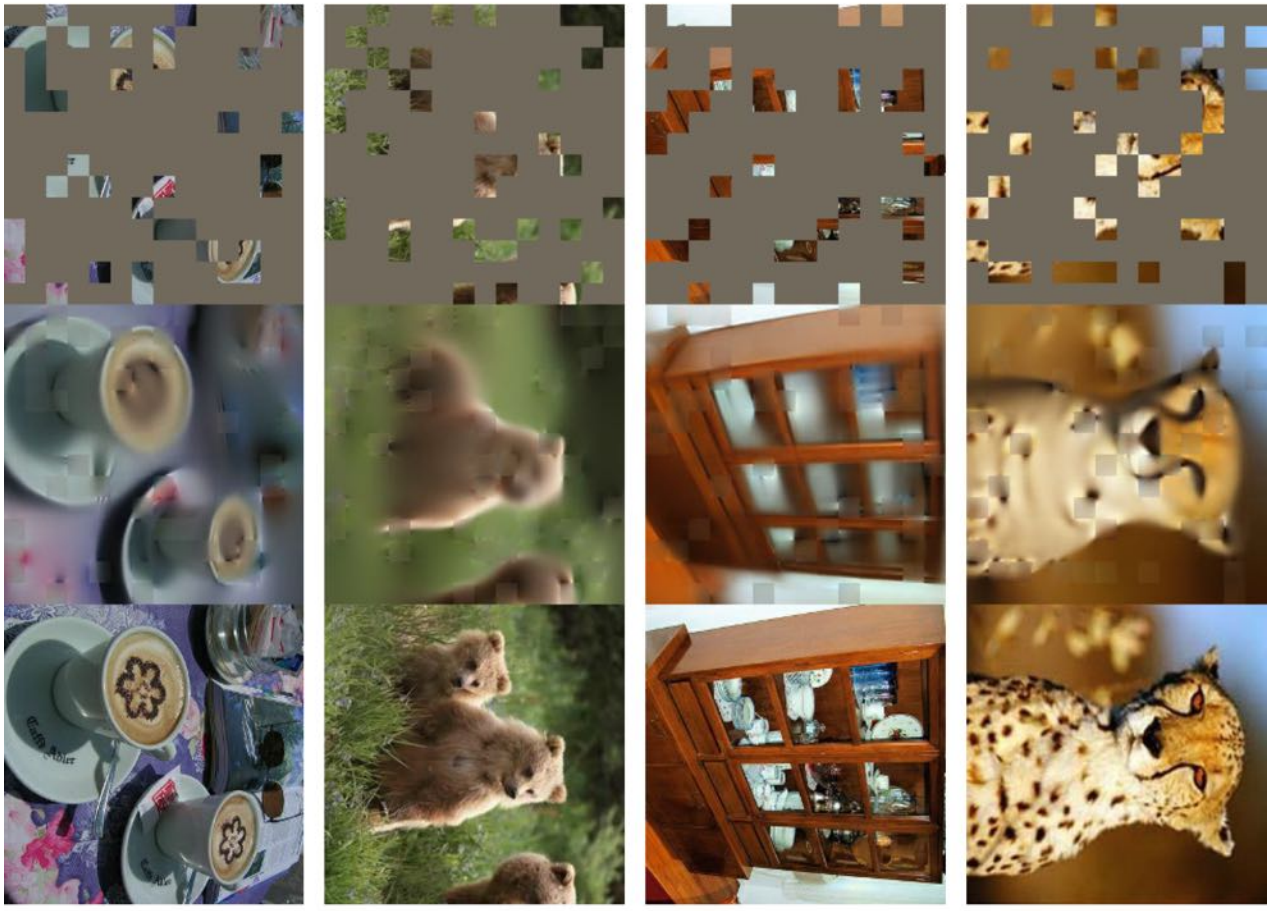
MAE's guess

Ground truth

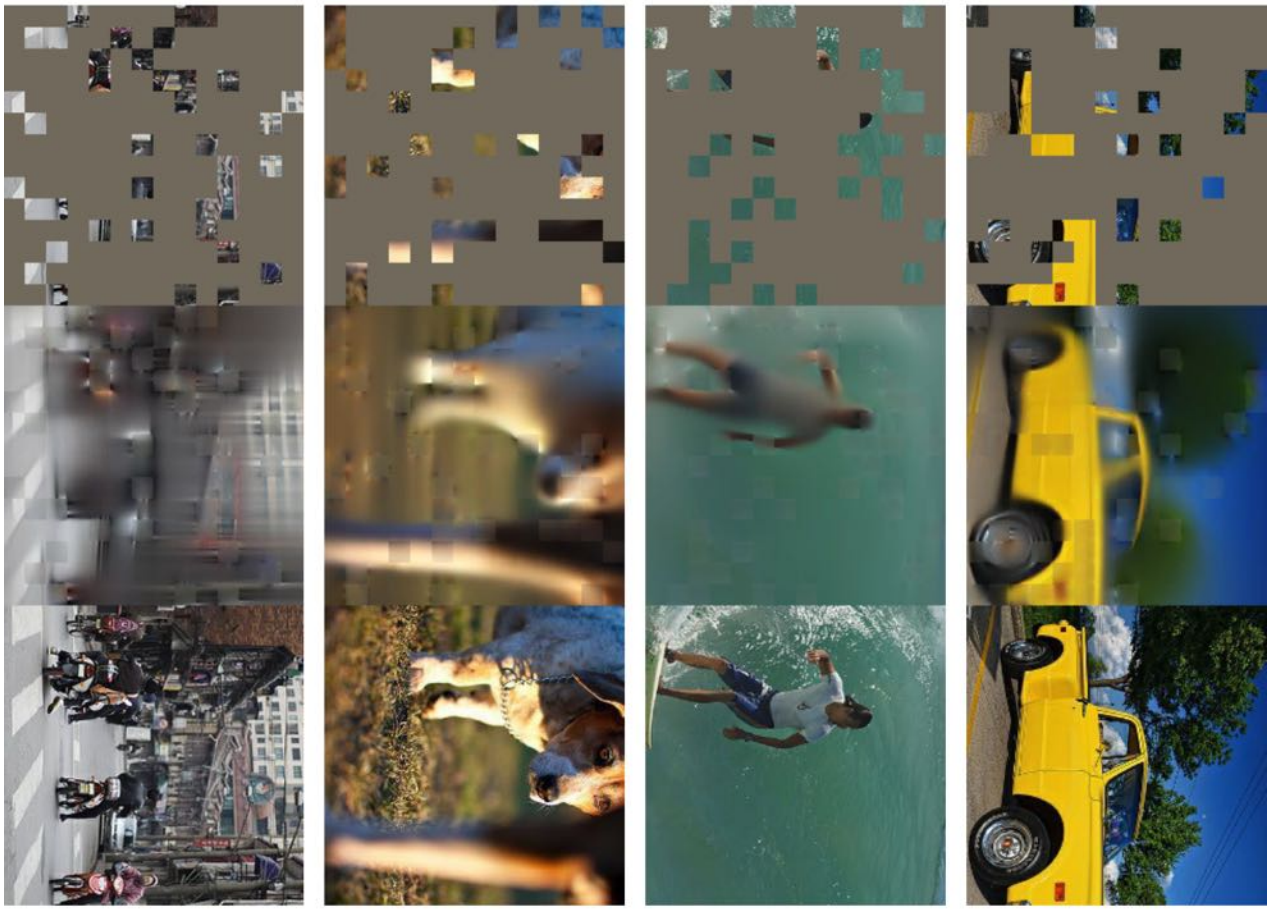
slide credit: Xinlei Chen

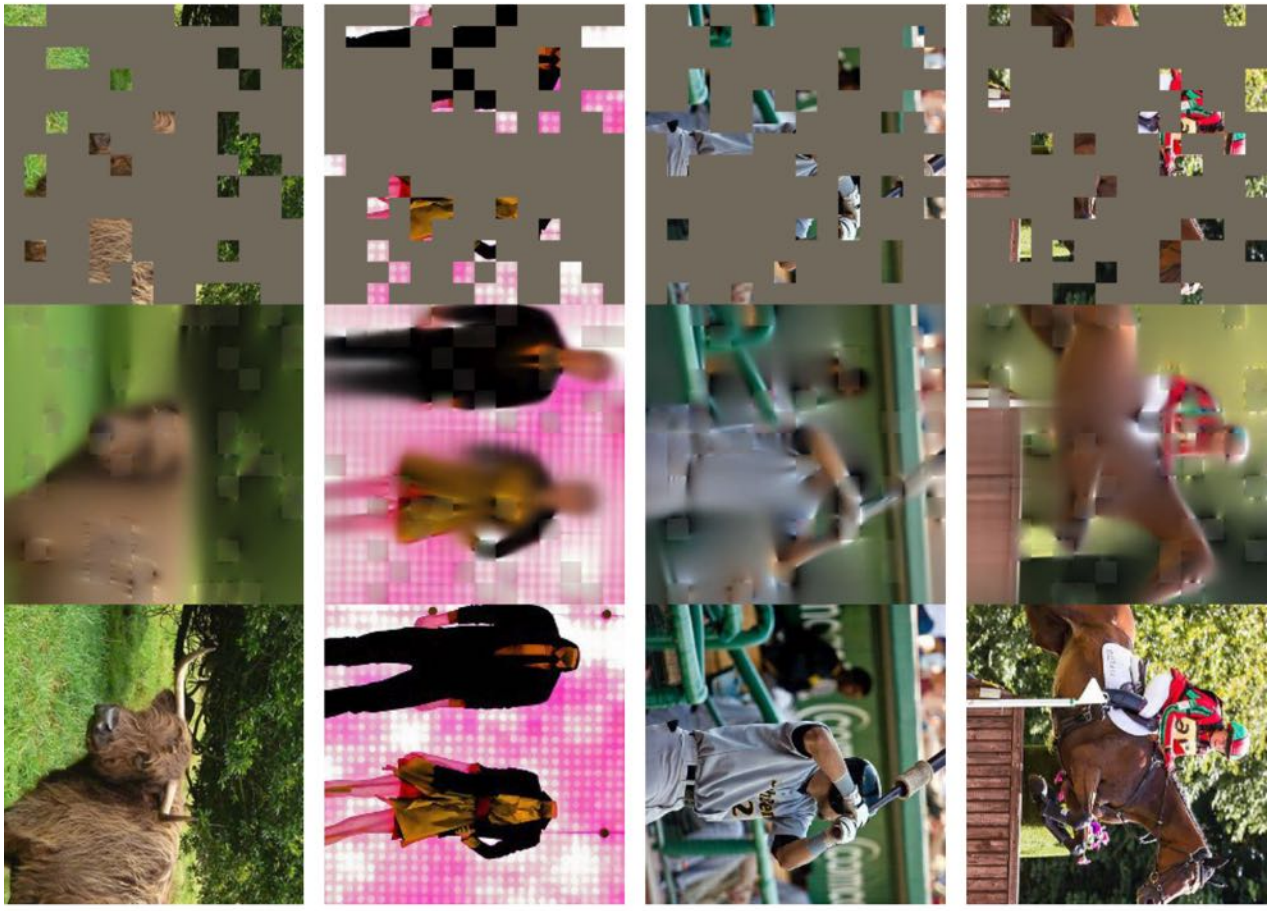
ImageNet val set (unseen)





COCO val set (unseen)

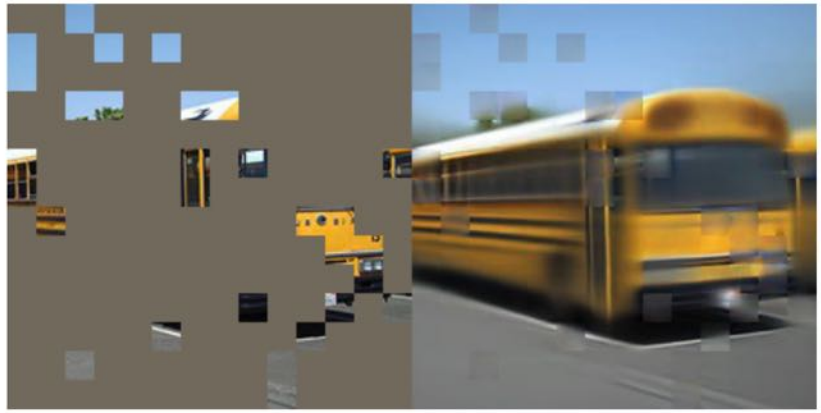




slide credit: Xinlei Chen

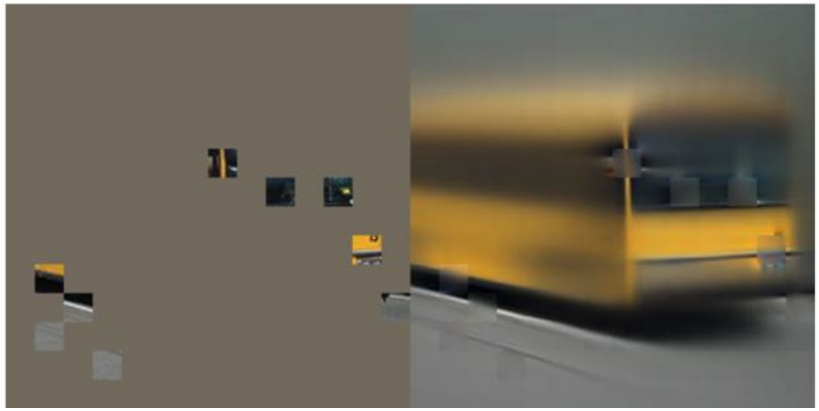
original

75% mask



85% mask

95% mask



MAE Can Generalize

slide credit: Xinlei Chen

original

75% mask

85% mask

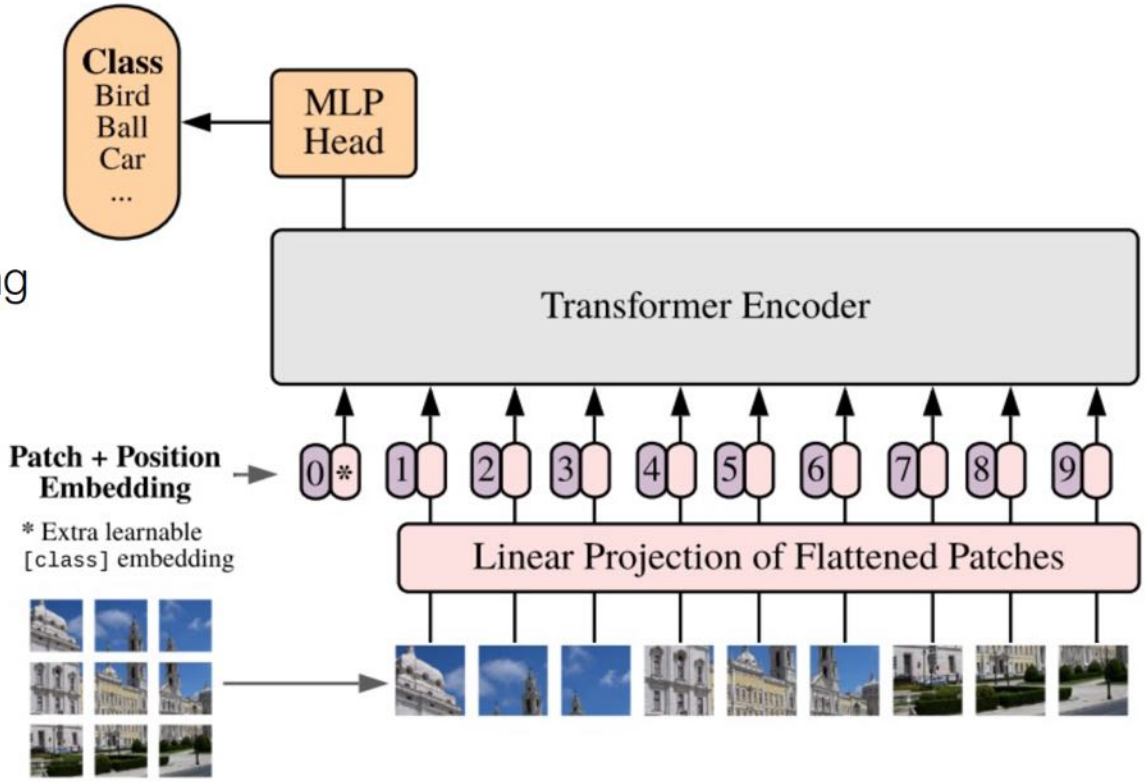
95% mask

MAE Can Generalize

[Dosovitskiy et al, ICLR 2021]

BERT-like: Transformers

- Vision Transformer (ViT)
  - Less inductive bias
  - Non-overlapping tokenization
    - Easier for masked auto-encoding
- *Scalable*
  - with larger models
  - on larger datasets

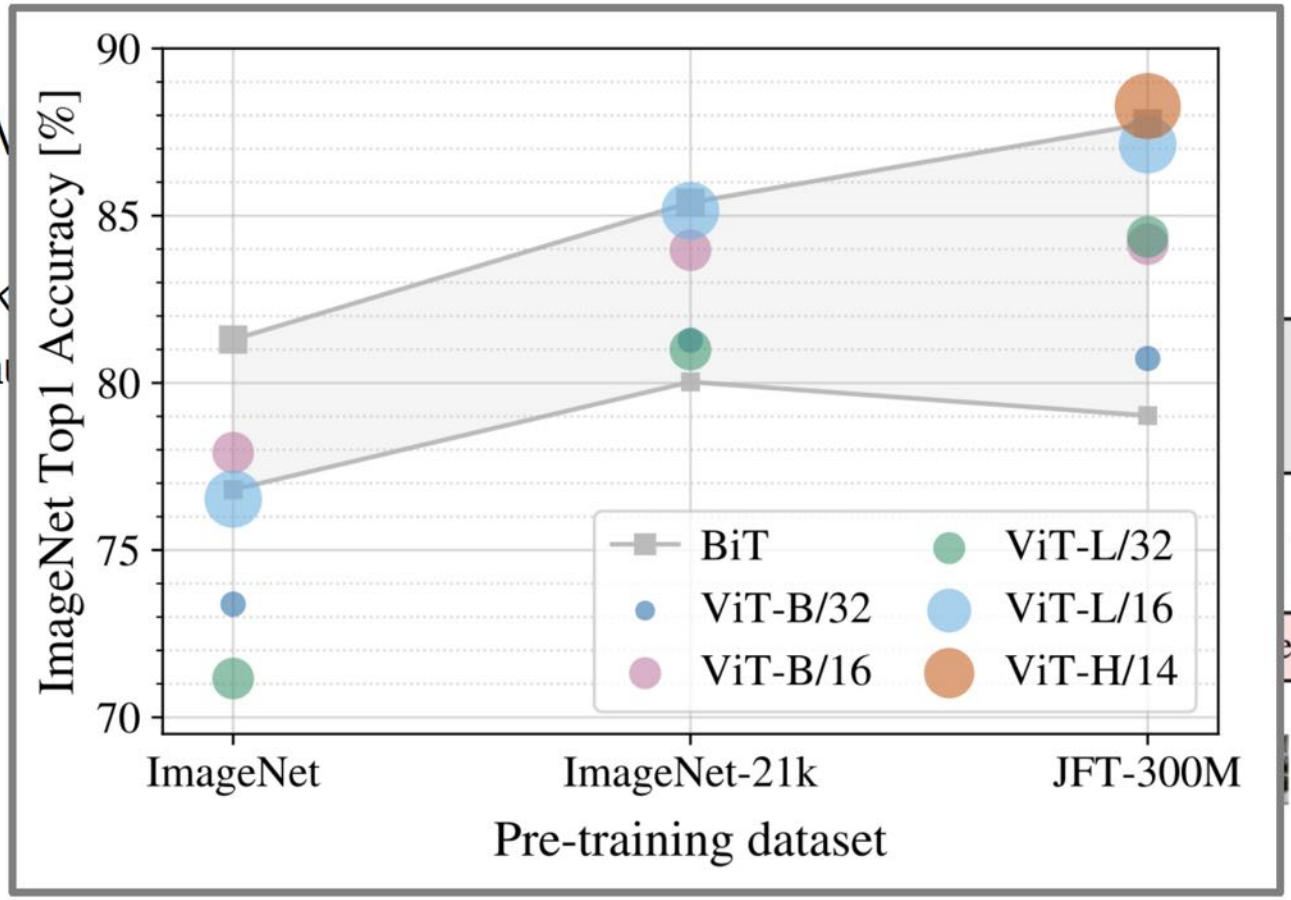


slide credit: Xinlei Chen

[Dosovitskiy et al, ICLR 2021]

BERT-like: Transformers

- Vision Transformer (V
  - Less inductive bias
  - Non-overlapping tok
    - Easier for masked a
- *Scalable*
  - with larger models
  - on larger datasets

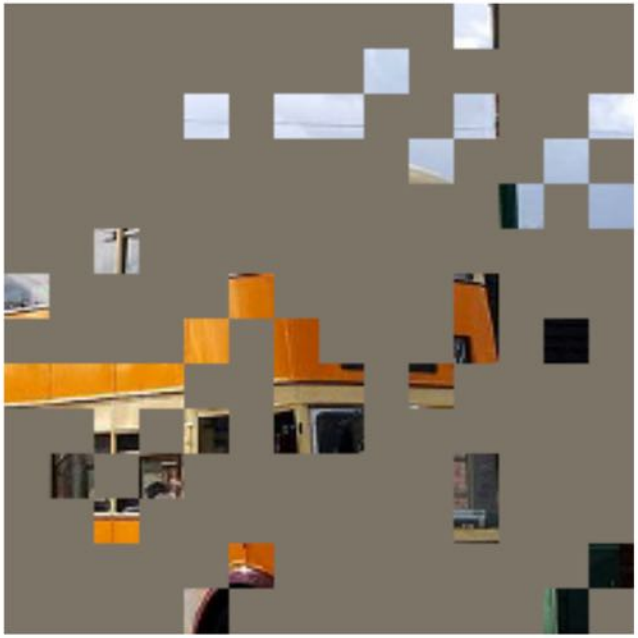


slide credit: Xinlei Chen

[Devlin et al, NAACL 2019]

BERT-*unlike*: Mask Ratio

- BERT: 15% is enough to create a challenging task
- MAE: a high ratio of 75% - 80% is about optimal

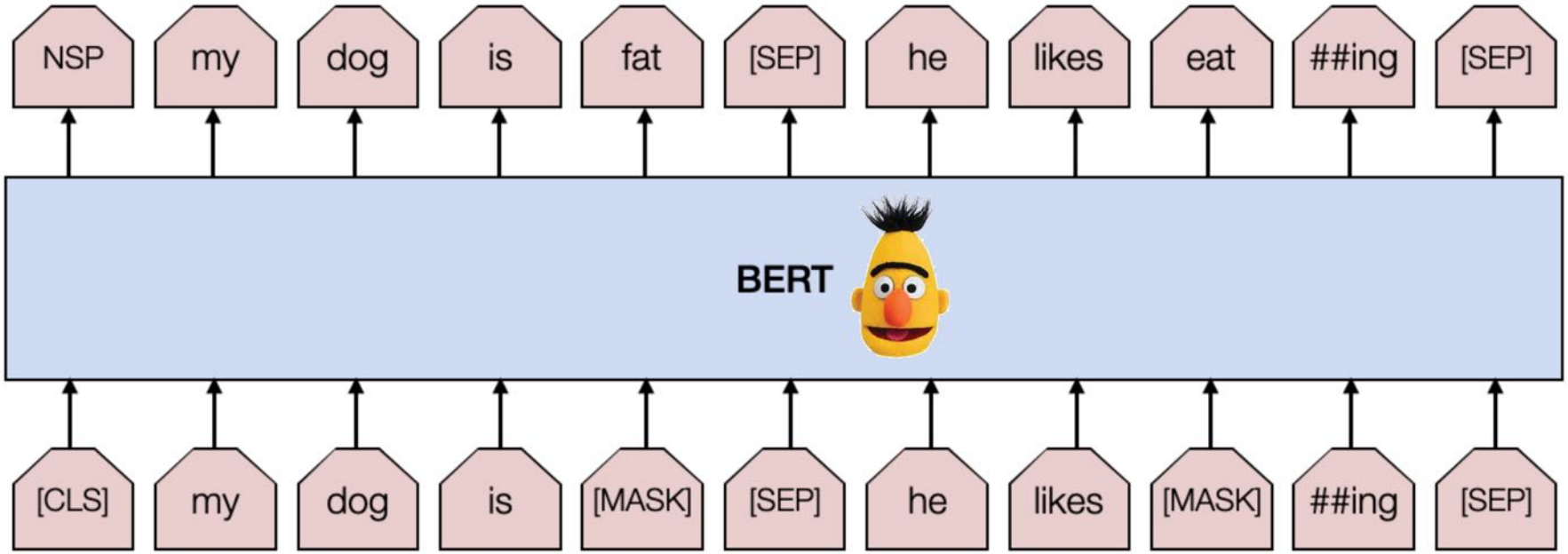


slide credit: Xinlei Chen

[Devlin et al, NAACL 2019]

BERT-*unlike*: Encoder-Decoder

- BERT: encoder-*only* pre-training



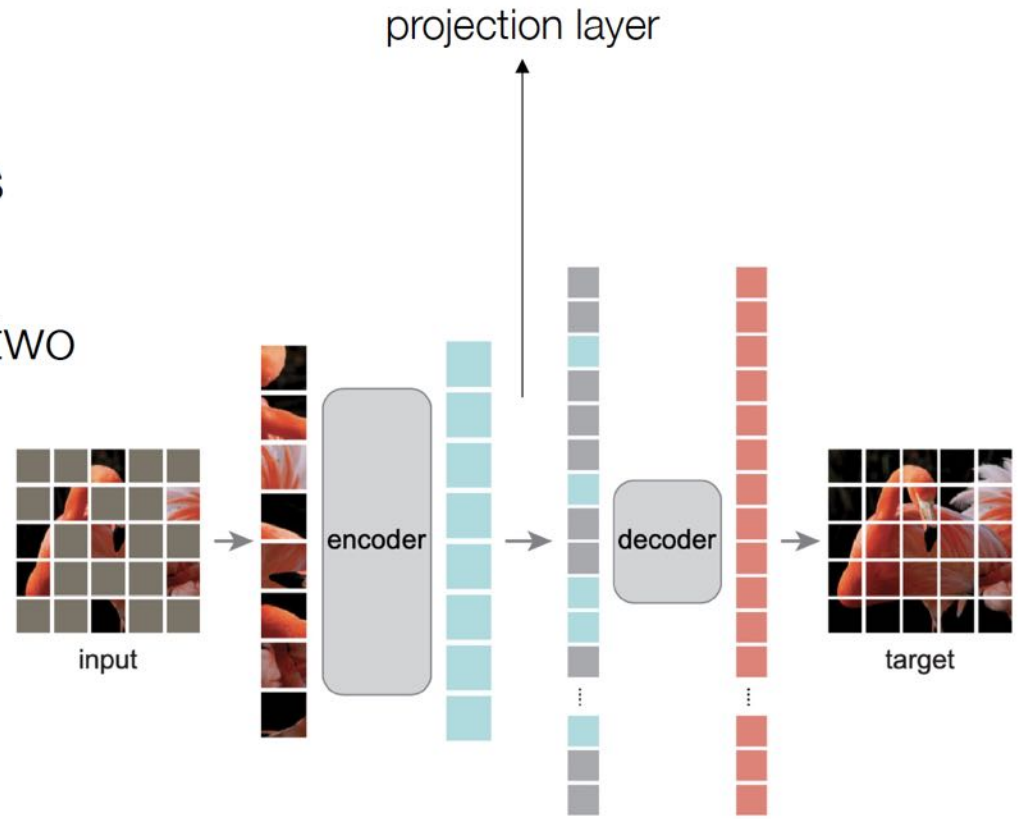
slide credit: Xinlei Chen

[Devlin et al, NAACL 2019]

BERT-*unlike*: Encoder-Decoder

- MAE:
  - *Large* encoder on *visible* tokens
  - Small decoder on *all* tokens
  - *Projection* layer to connect the two

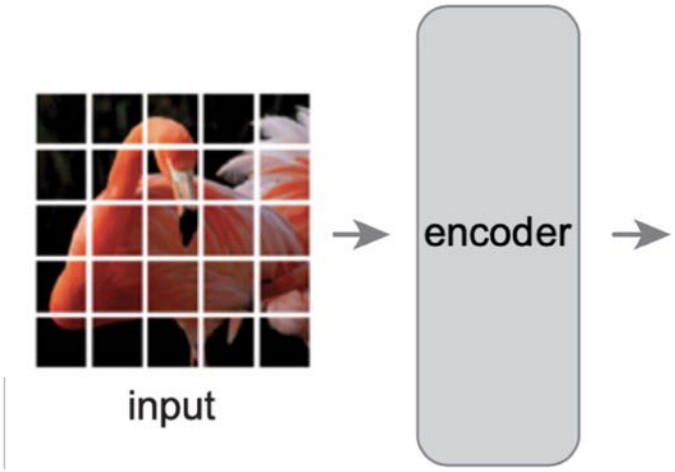
- Very efficient when coupled with high mask ratio (75%)



slide credit: Xinlei Chen

MAE for Downstream Tasks: *Encoder Only*

- After MAE pre-training, just *throw away* the decoder
- Encoder is used for representations with *full-sequence* input



slide credit: Xinlei Chen

Experimental Protocols

- Pre-training dataset: ImageNet-1K
- Architecture: ViT-*Large* encoder, 512-dim decoder
- Transfer task: ImageNet-1K classification
  - "*ft*": end-to-end tuning with MAE as an initialization
  - "*lin*": linear probing, a single classifier on top of frozen encoder features

slide credit: Xinlei Chen

Analysis: Decoder Size

- Encoder has 24-blocks, 1024-dimensional

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

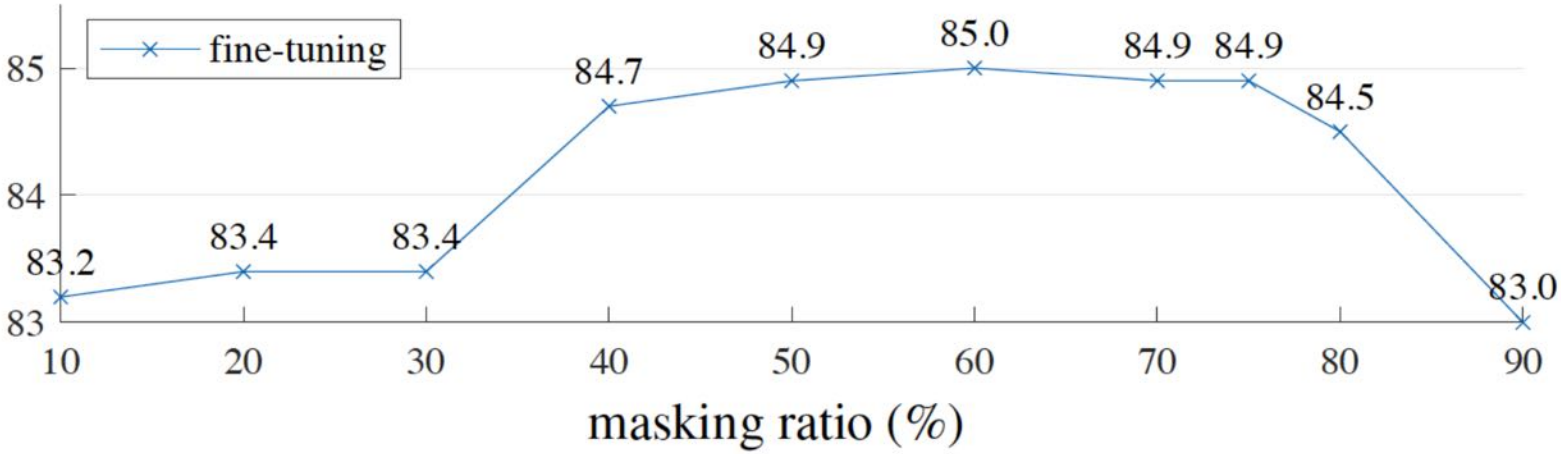
Decoder depth

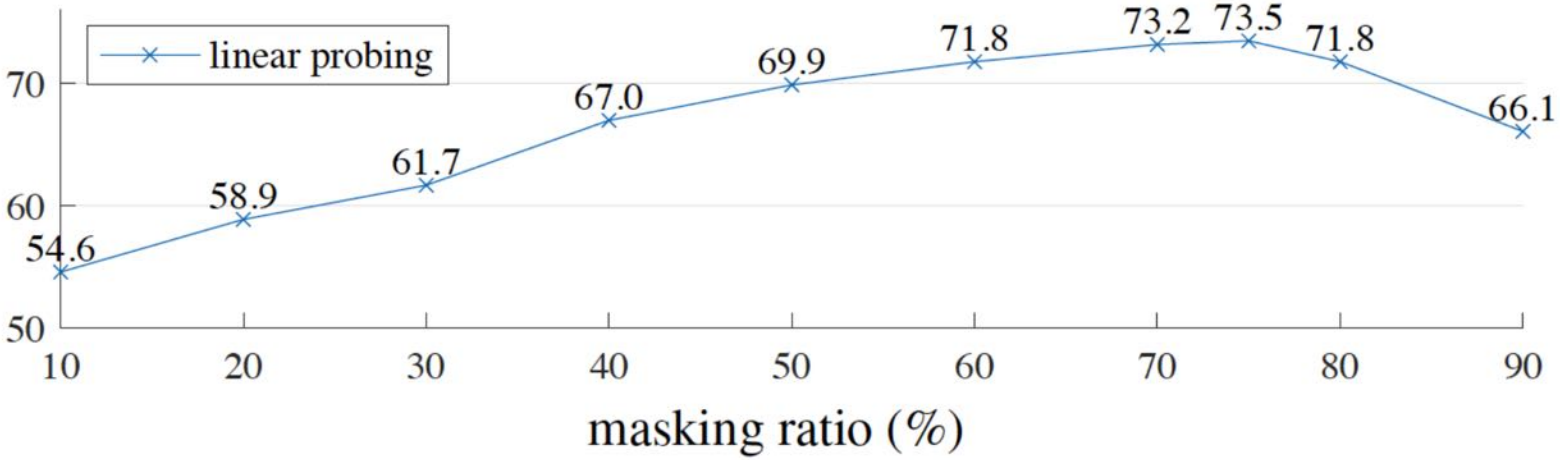
dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

Decoder width

slide credit: Xinlei Chen

Analysis: Mask Ratio





slide credit: Xinlei Chen

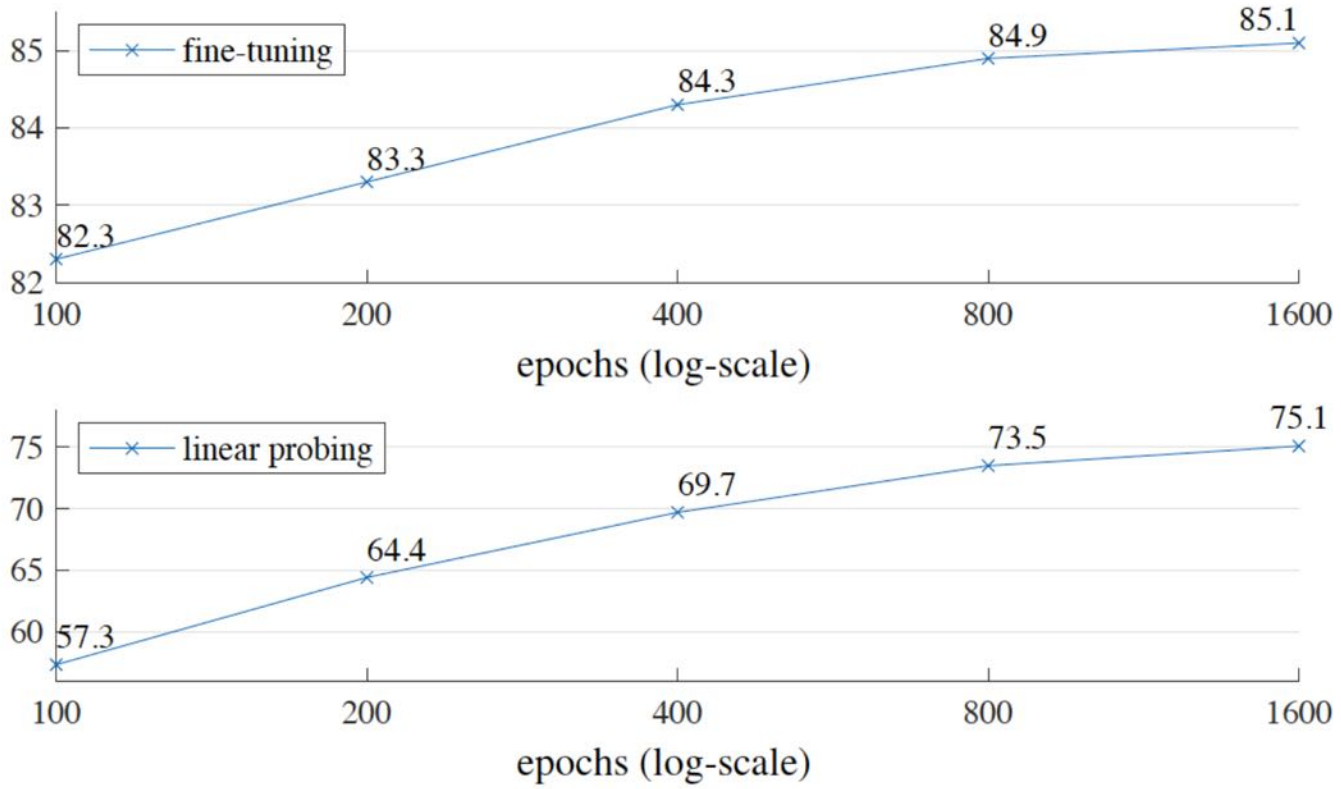
Analysis: Augmentations

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

- MAE can work with minimal data augmentation

slide credit: Xinlei Chen

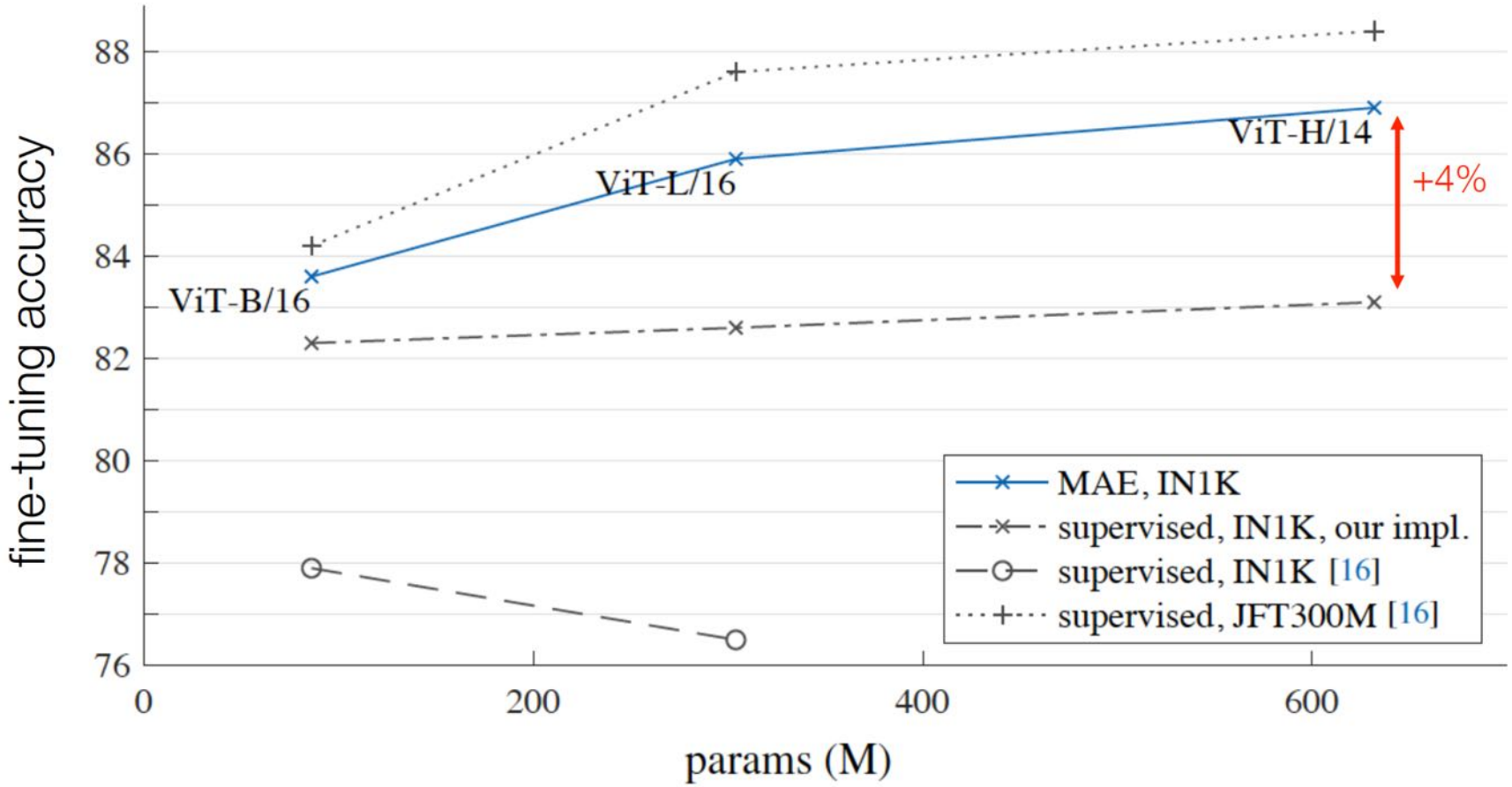
Scalability: Longer Training



Wall-clock speed still efficient thanks to MAE design

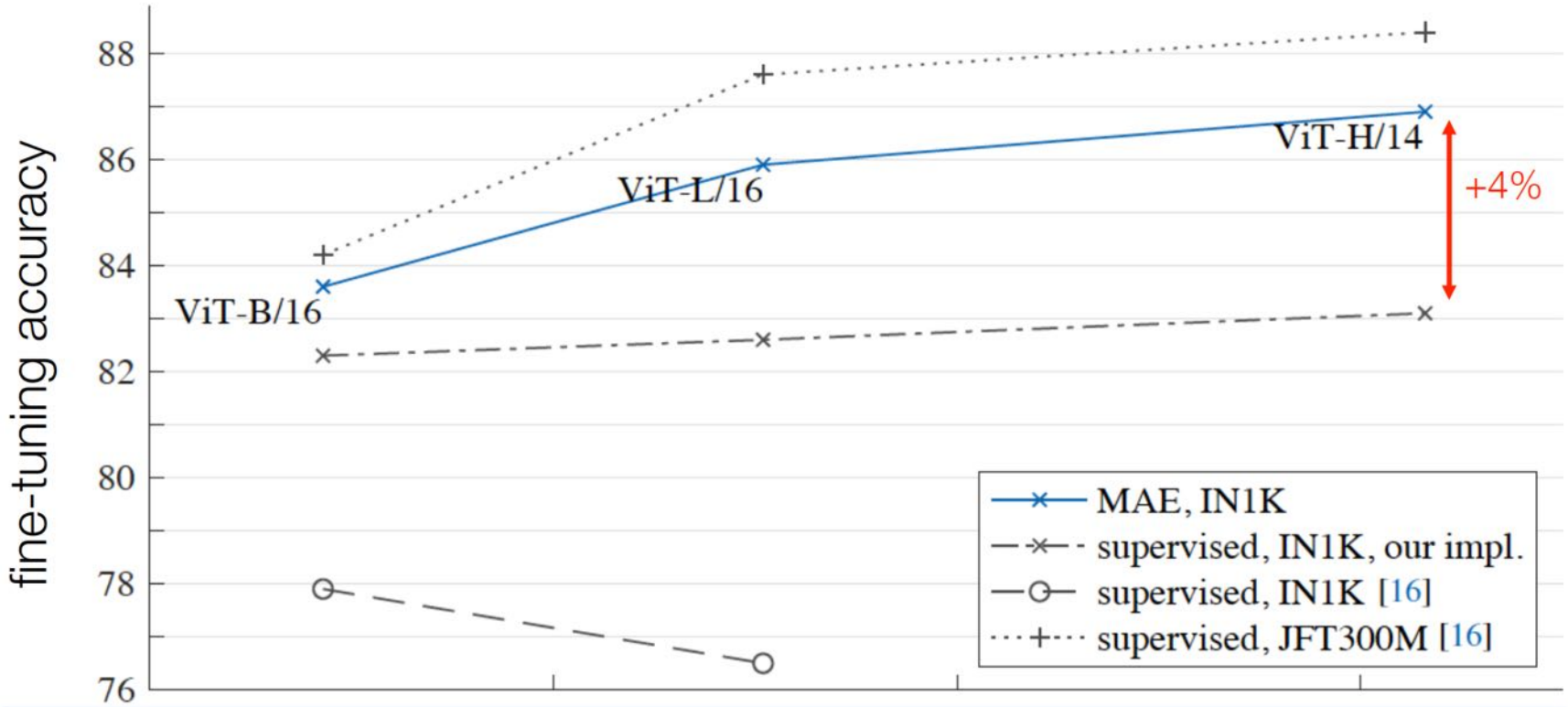
slide credit: Xinlei Chen

Scalability: Larger Models



slide credit: Xinlei Chen

Scalability: Larger Models



new SOTA on ImageNet-1K (no extra data): **87.8%**

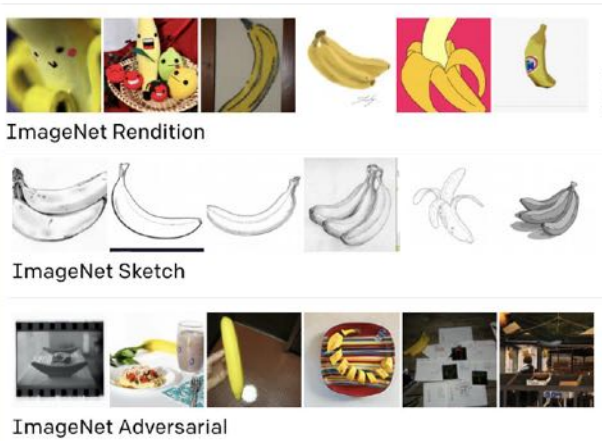
slide credit: Xinlei Chen

slide credit: Xinlei Chen

Scalability: Larger Models

dataset	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [50]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [49]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [49]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [36]‡

new SOTA on 5 large-scale classification datasets



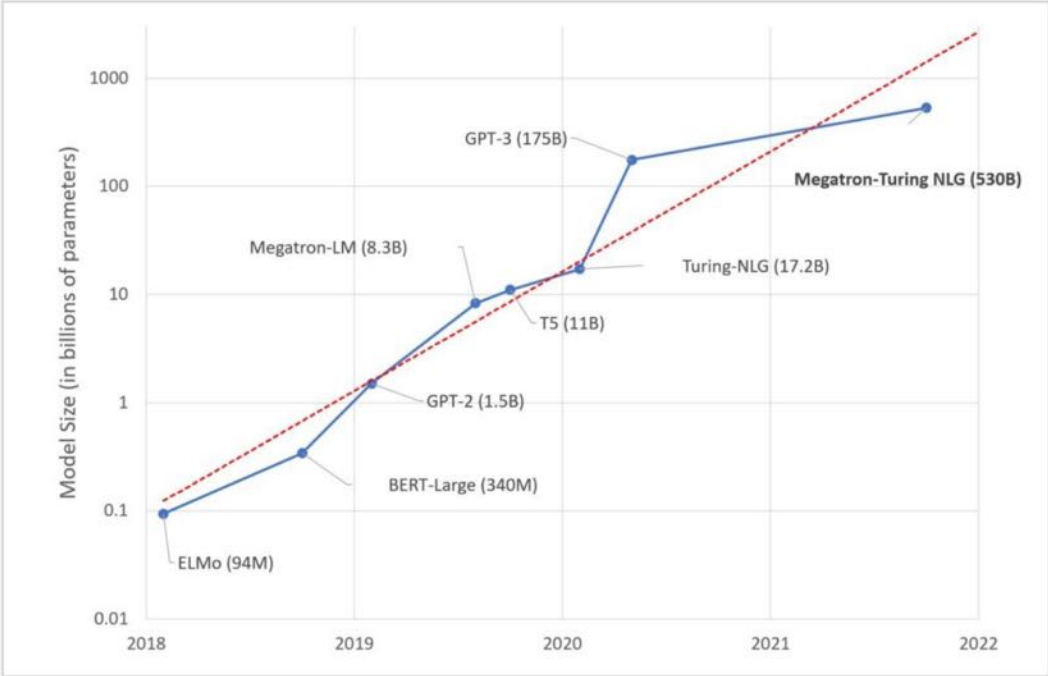
dataset	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$	prev best
IN-Corruption ↓ [27]	51.7	41.8	**33.8**	36.8	42.5 [32]
IN-Adversarial [28]	35.9	57.1	68.2	**76.7**	35.8 [41]
IN-Rendition [26]	48.3	59.9	64.4	**66.5**	48.7 [41]
IN-Sketch [60]	34.5	45.3	49.6	**50.9**	36.0 [41]

new SOTA on 4 ImageNet robust evaluations

Is the Journey 99% Done?

- NLP has witnessed amazing progress in scaling since BERT

- It's just *starting* in vision:
  - Temporal data
  - Architectures – ConvNets?
  - Other modalities? 3D?
  - Other downstream tasks?
  - Other axes to scale?
  - [`Your exploration`] here!



slide credit: Xinlei Chen

code (GPU): https://github.com/facebookresearch/mae
code (TPU): https://github.com/facebookresearch/long_seq_mae

Take-aways

- Self-supervised learning aims at *scalable* representation learning
- Masked auto-encoders can serve as scalable vision learners
- Exciting years ahead in this direction!

slide credit: Xinlei Chen

Overview of Today's Lecture

- Last Time: Self-Supervised Learning Part 1:
  - Motivation of Self-Supervised Learning
  - Pretext tasks from image transformations (e.g. rotation, inpainting, coloring)
  - Contrastive representation learning (SimCLR, MoCo, CPC)

- Today: Self-Supervised Learning Part 2
  - Teacher-Student "feature reconstruction"
    - motivation, setting
    - methods: BYOL, DINO
  - Image Reconstruction
    - MAE - Masked Autoencoders