max planck institut informatik

SIC Saarland Informatics Campus

High Level Computer Vision

Self-Supervised Learning (Part 2) & Vision-Language Models

@ June 28, 2023

Bernt Schiele

cms.sic.saarland/hlcvss23/

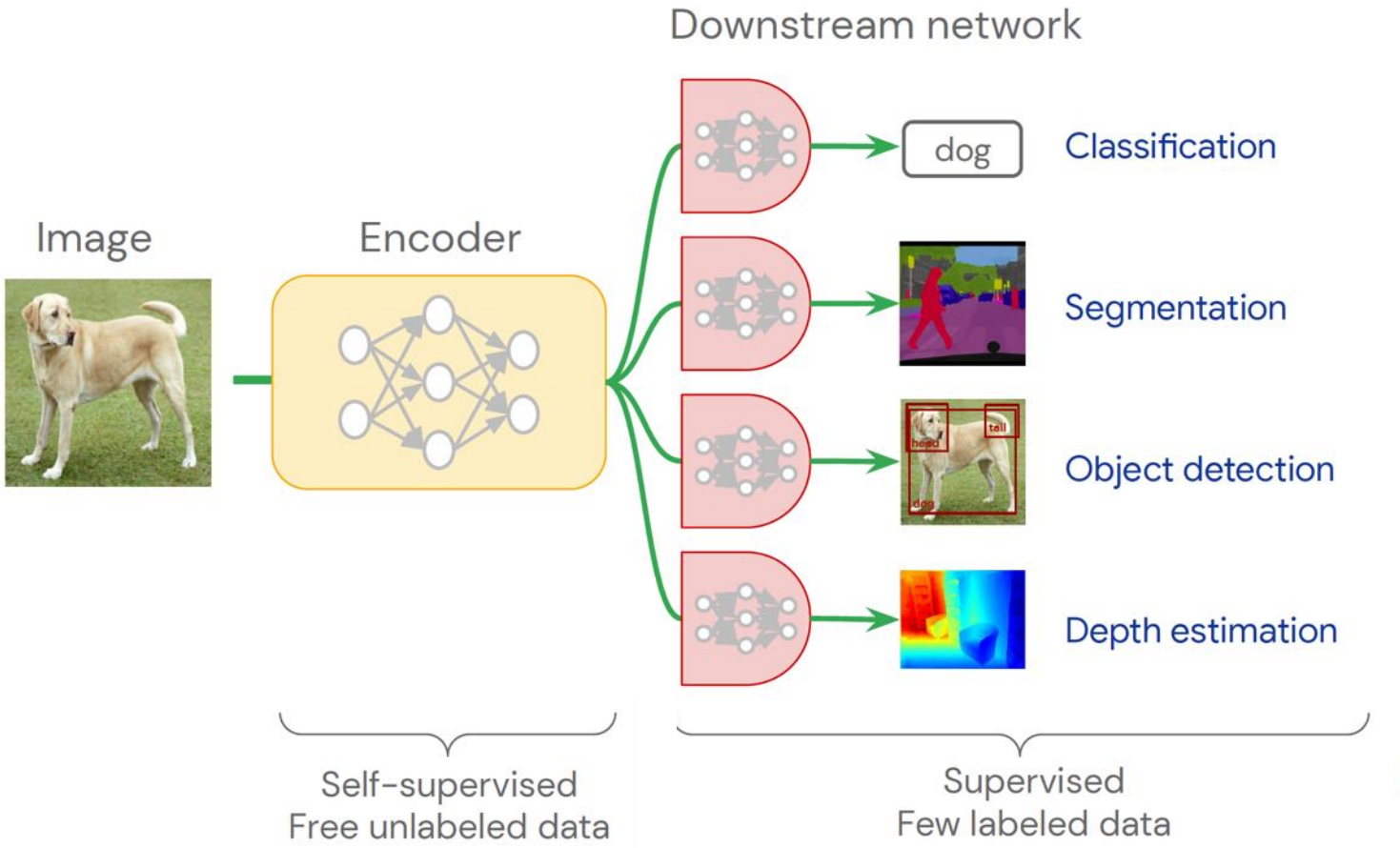
Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

Overview of Today's Lecture

- Continuation of Self-Supervised Learning
  - Teacher-Student "feature reconstruction"
    - motivation, setting
    - methods: BYOL, DINO
- Vision-Language Learning for Computer Vision
  - Motivation
  - CLIP, ALIGN
  - Some extensions

slide credit: Jean-Basien Grill, et al.

Idea of Self-Supervised Learning

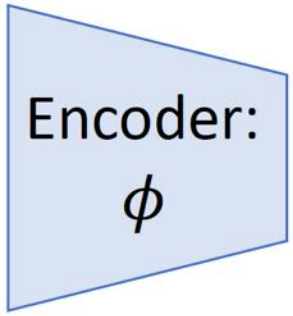


slide credit: Justin Johnson

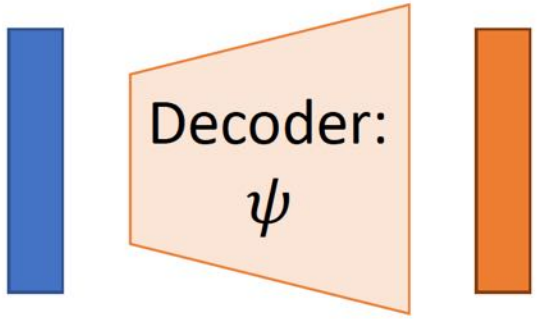
Self-Supervised Learning: Pretraining — then Transfer

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

Input Image: x



Features: $\phi(x)$

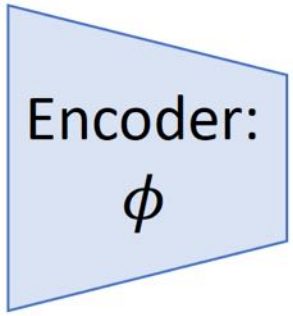


Prediction: $\hat{y}$

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

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

Input Image: x

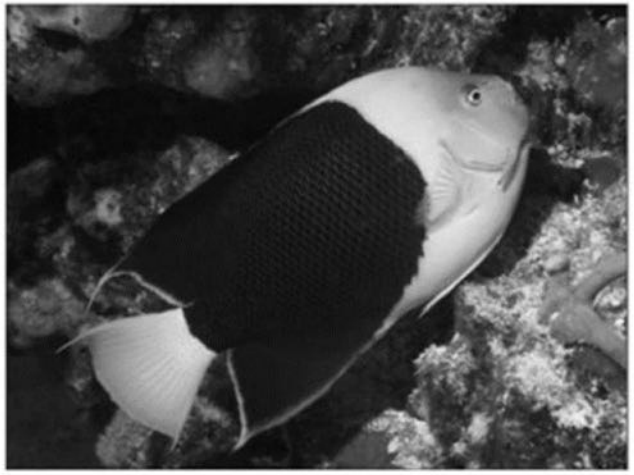


Features: $\phi(x)$

Downstream tasks: Image classification, object detection, semantic segmentation

slide credit: Spyros Gidaris, Andrei Bursuc

Colorization



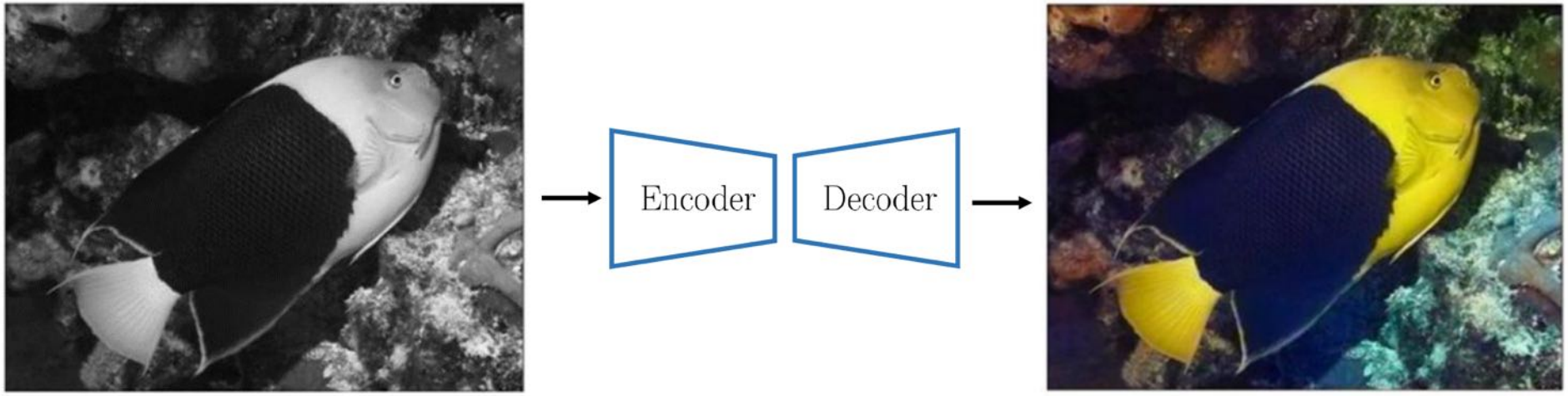
What is the colour of every pixel?

Hard if you don't recognize the objects!

Image colorization (Zhang et al. 2018)

slide credit: Spyros Gidaris, Andrei Bursuc

Colorization

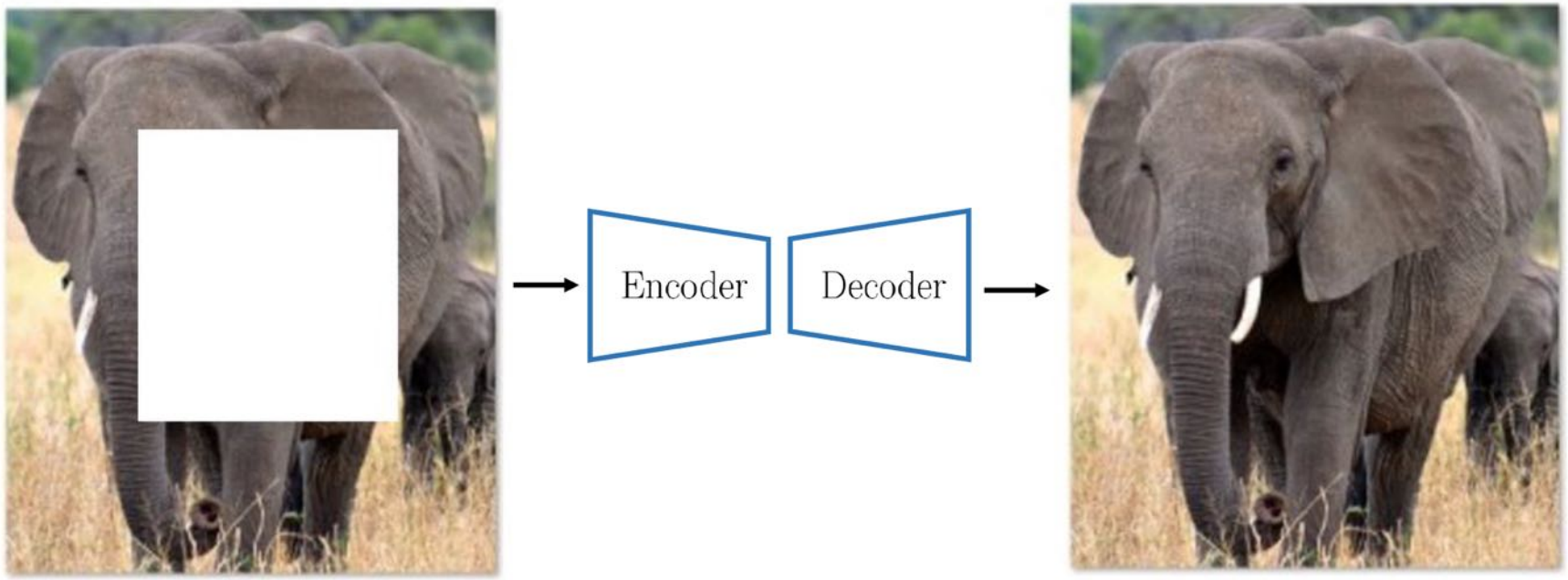


- Requires preservation of fine-grained information
- Input reconstruction is too hard and ambiguous
- Lots of effort spent on “useless” details: exact color, good boundary, etc.

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.

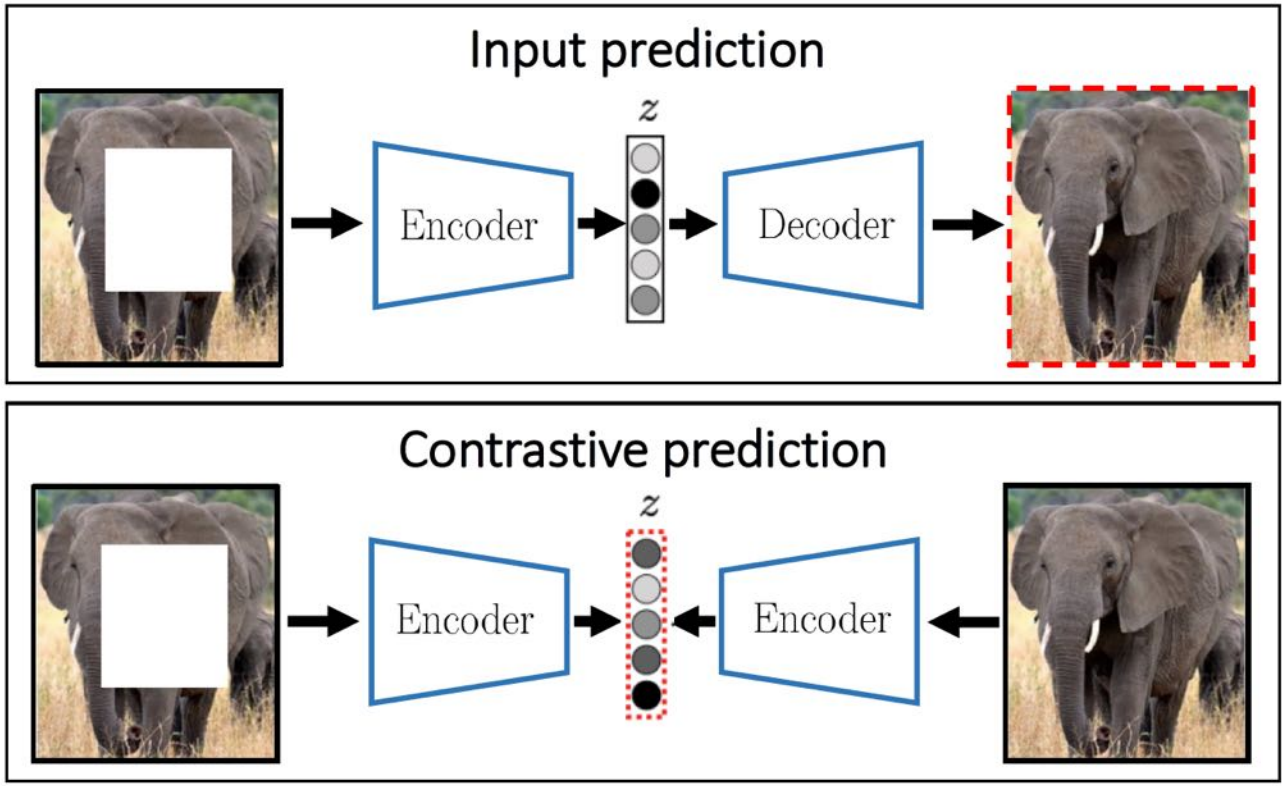
Context Encoders (Pathak et al. 2016)

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)



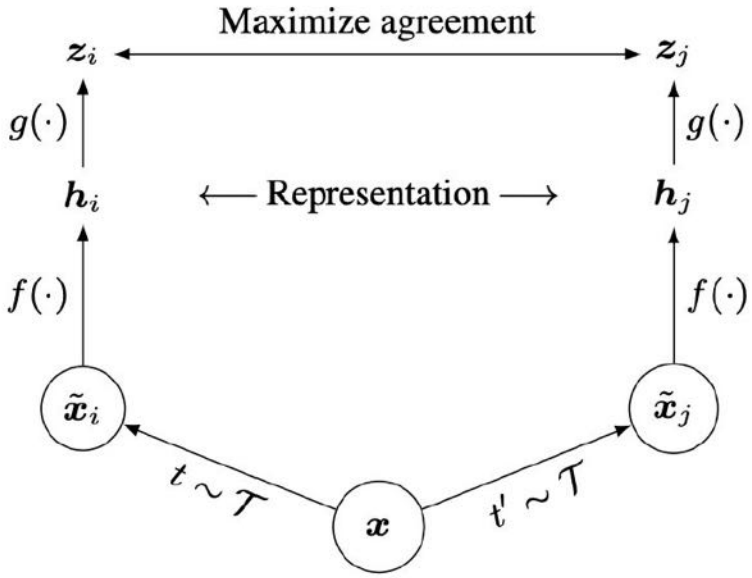
References:
"Representation learning with contrastive predictive coding", Oord et al, 2018
"Constraive multiview coding", Tian et al, 2020
"A simple framework for contrastive learning of visual representations", Chen et al, 2020
...

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

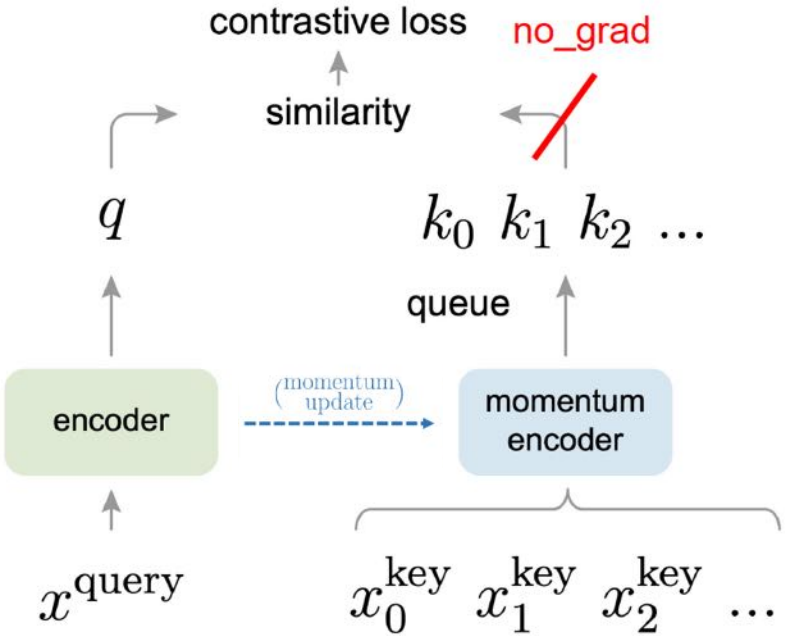


slide credit: Fei-Fei Li, Jiajun Wu, Ruohan Gao

Summary: Contrastive Representation Learning

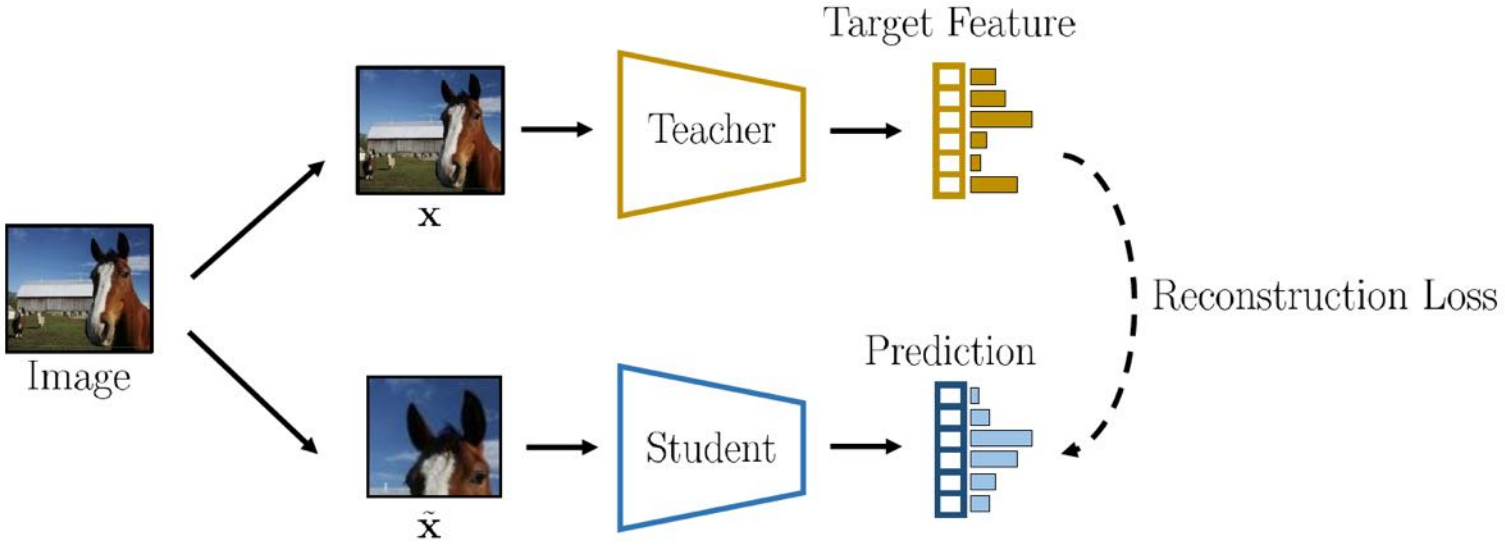
MoCo (v1, v2): contrastive learning using momentum sample encoder

- 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



slide credit: Spyros Gidaris, Andrei Bursuc

Teacher-Student Feature “Reconstruction”

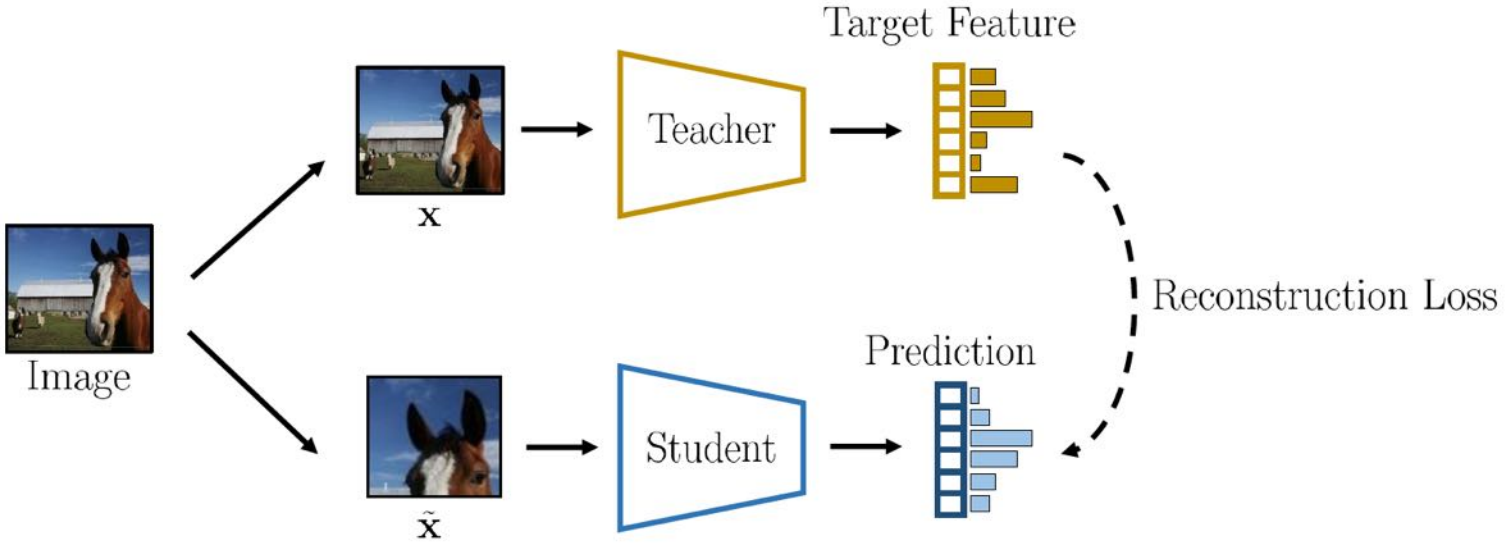


Teacher: generate a target feature vector from a given image

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

slide credit: Spyros Gidaris, Andrei Bursuc

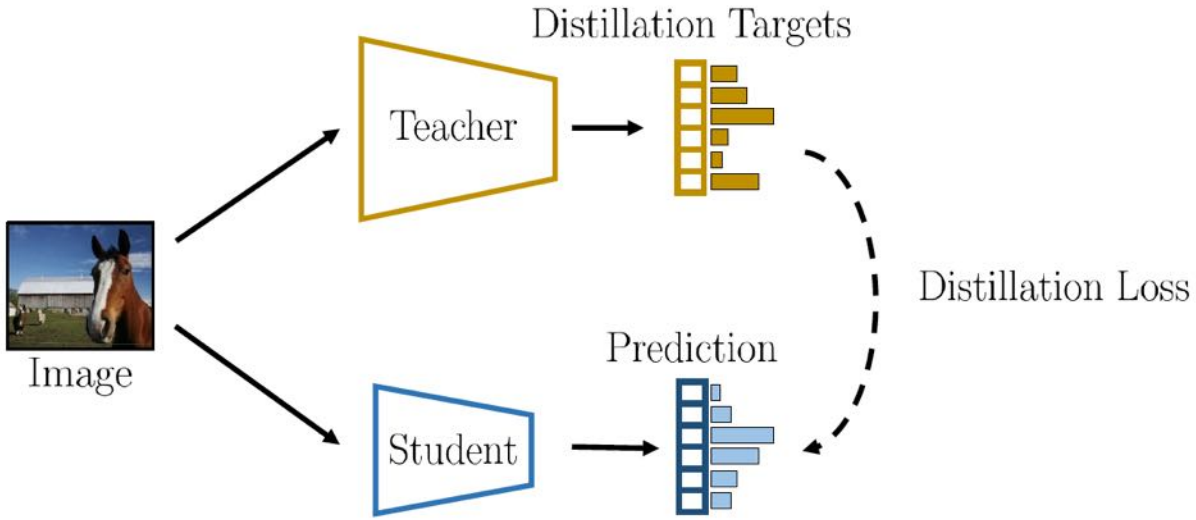
Teacher-Student Feature “Reconstruction”



- 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



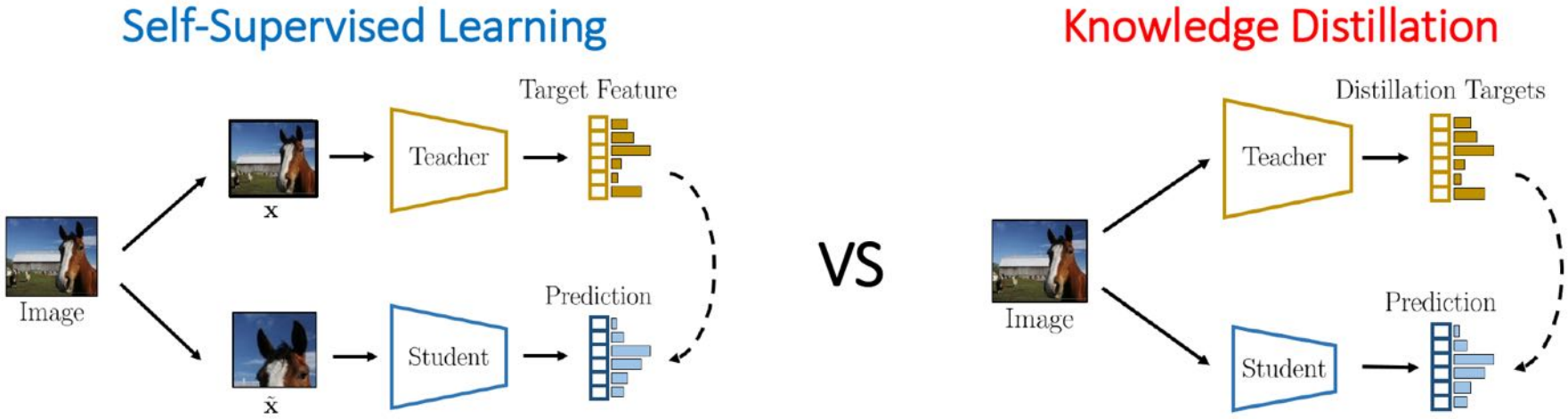
Goal: Distill the knowledge of a pre-trained teacher into a smaller student

- 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

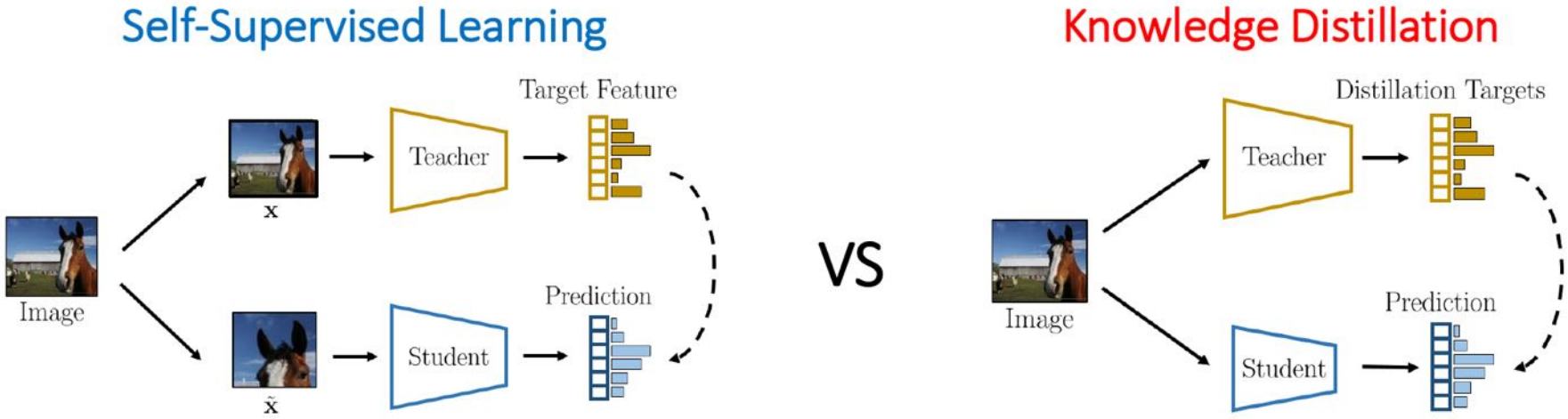
Teacher-Student Feature "Reconstruction" vs. Knowledge Distillation



- 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

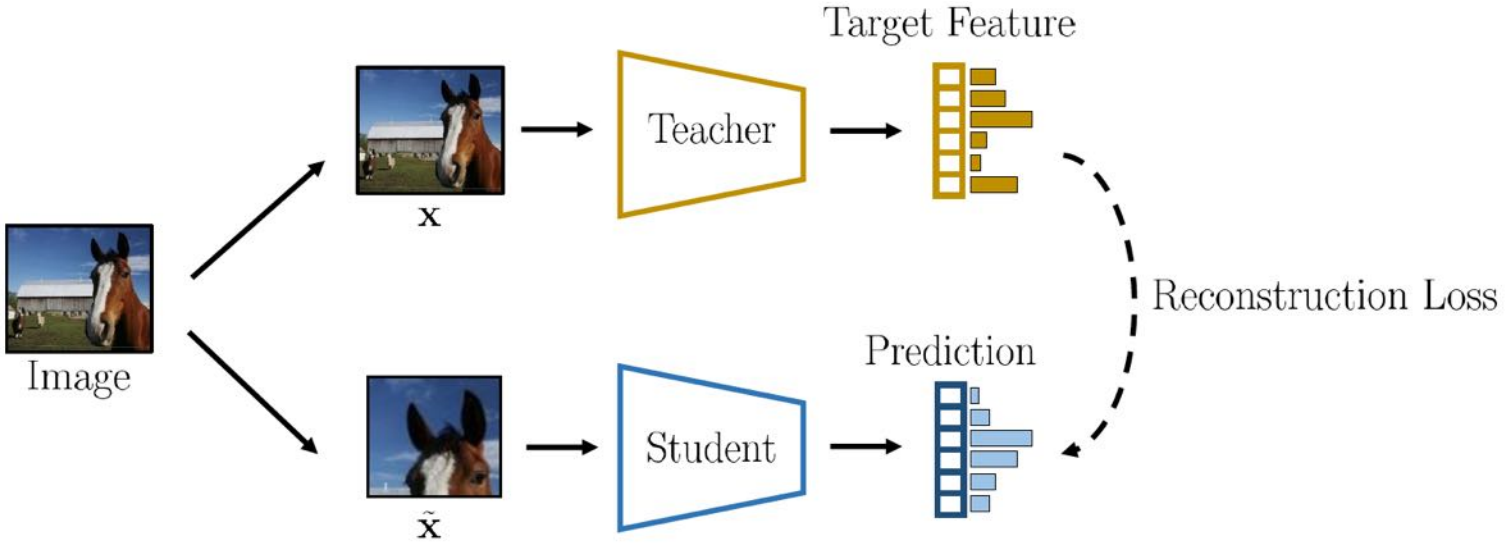
Teacher-Student Feature "Reconstruction" vs. Knowledge Distillation



- 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”



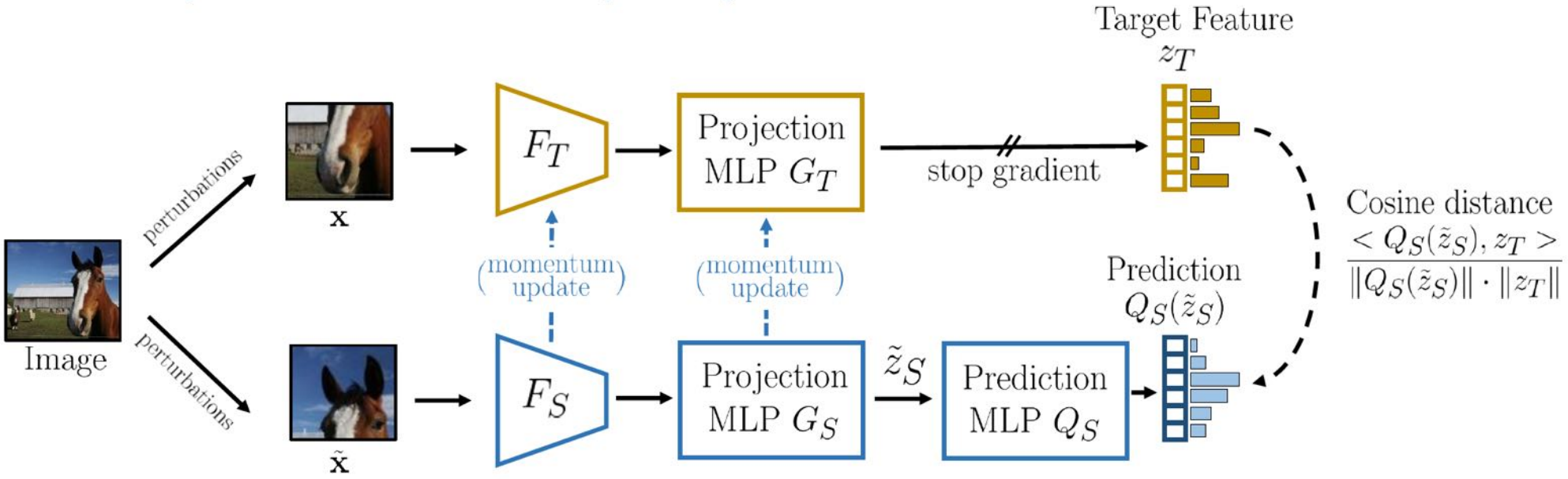
Key questions:

- 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)



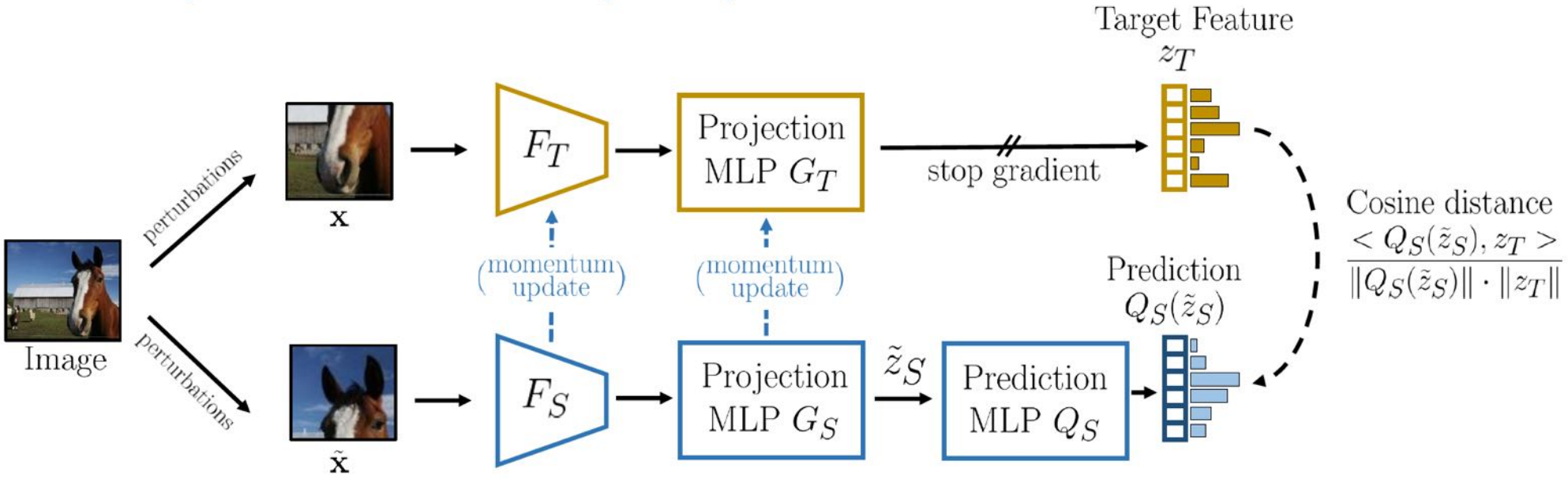
Feature reconstruction method:

- **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)



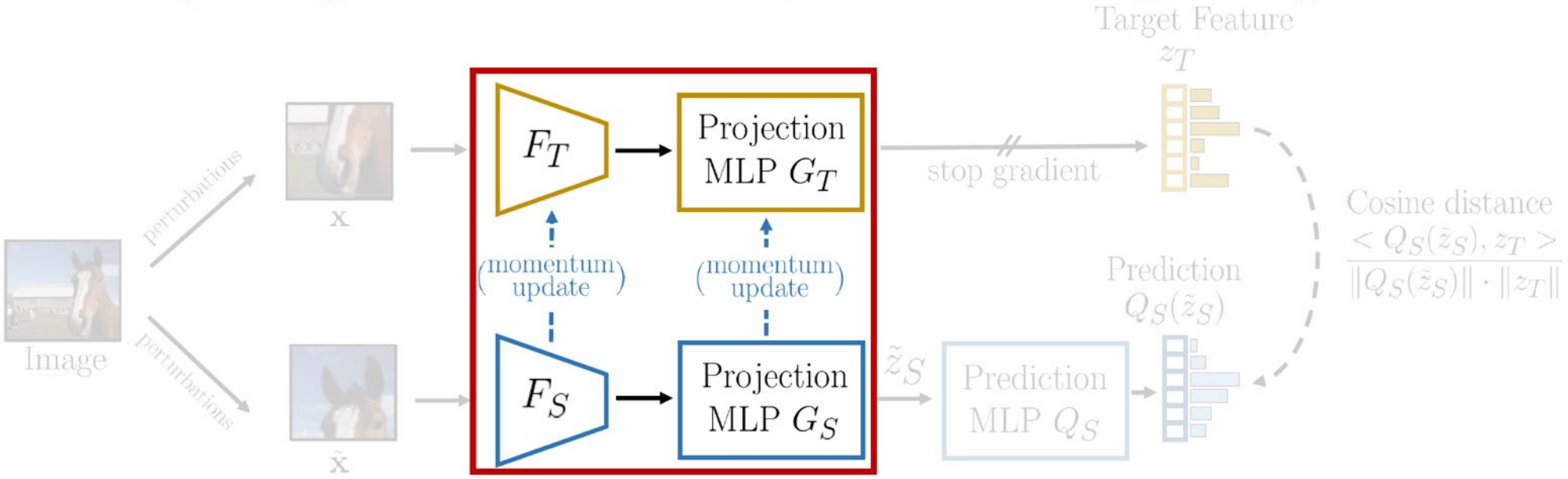
Bootstrap idea: builds a sequence of student representations of increasing quality

- 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:

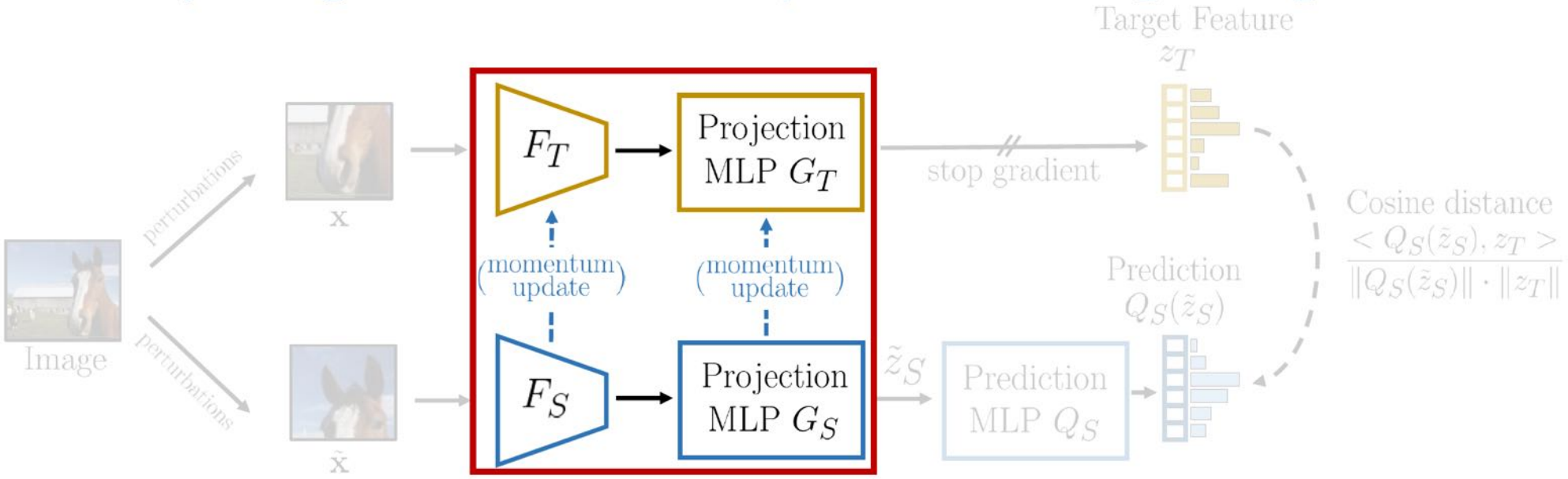
$$\theta_{\mathrm{T}}^{(t)} \leftarrow \alpha \cdot \theta_{\mathrm{T}}^{(t-1)} + (1-\alpha) \cdot \theta_{\mathrm{S}}^{(t)}$$

θ_{T} : teacher parameters

θ_{S} : student parameters

slide credit: Spyros Gidaris, Andrei Bursuc

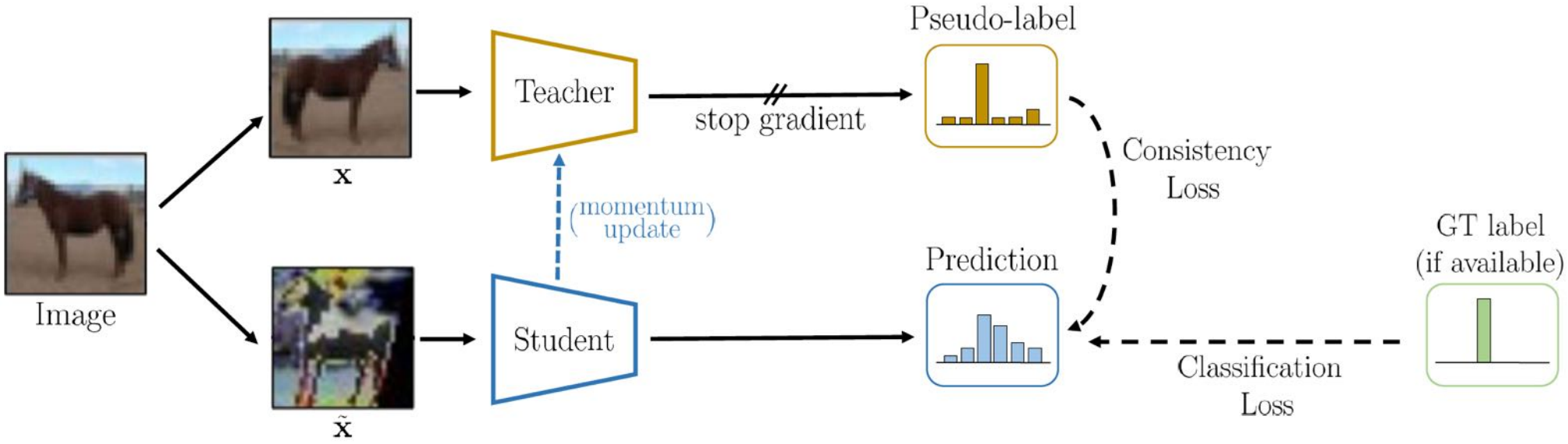
Online updating the teacher with exponential moving average



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

slide credit: Spyros Gidaris, Andrei Bursuc

Detour: mean / momentum teacher in semi-supervised learning



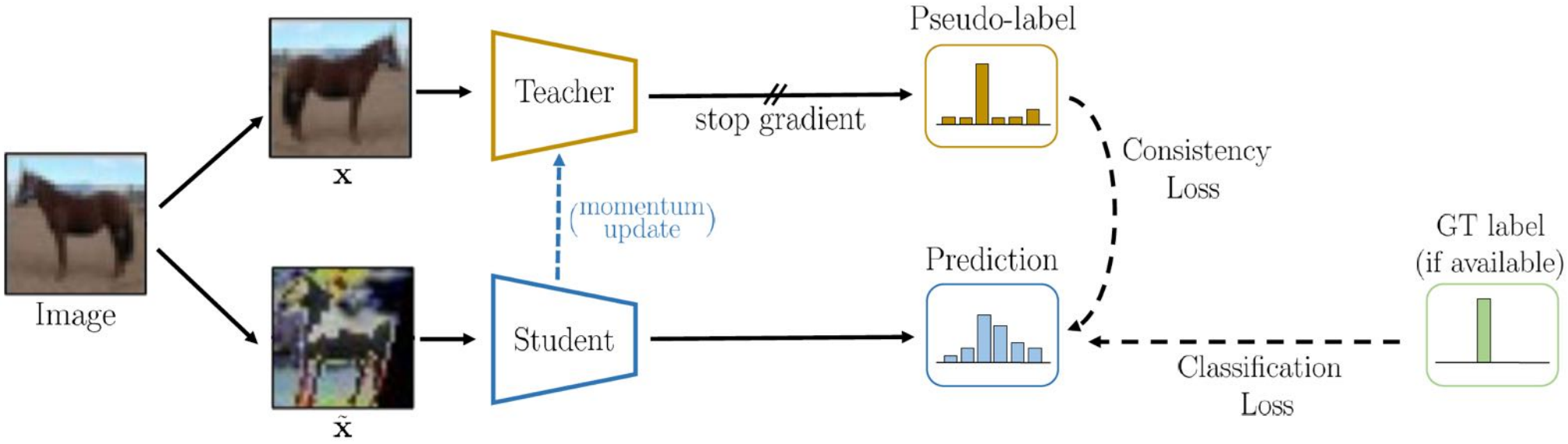
Teacher-student approaches are common in semi-supervised learning:

- **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



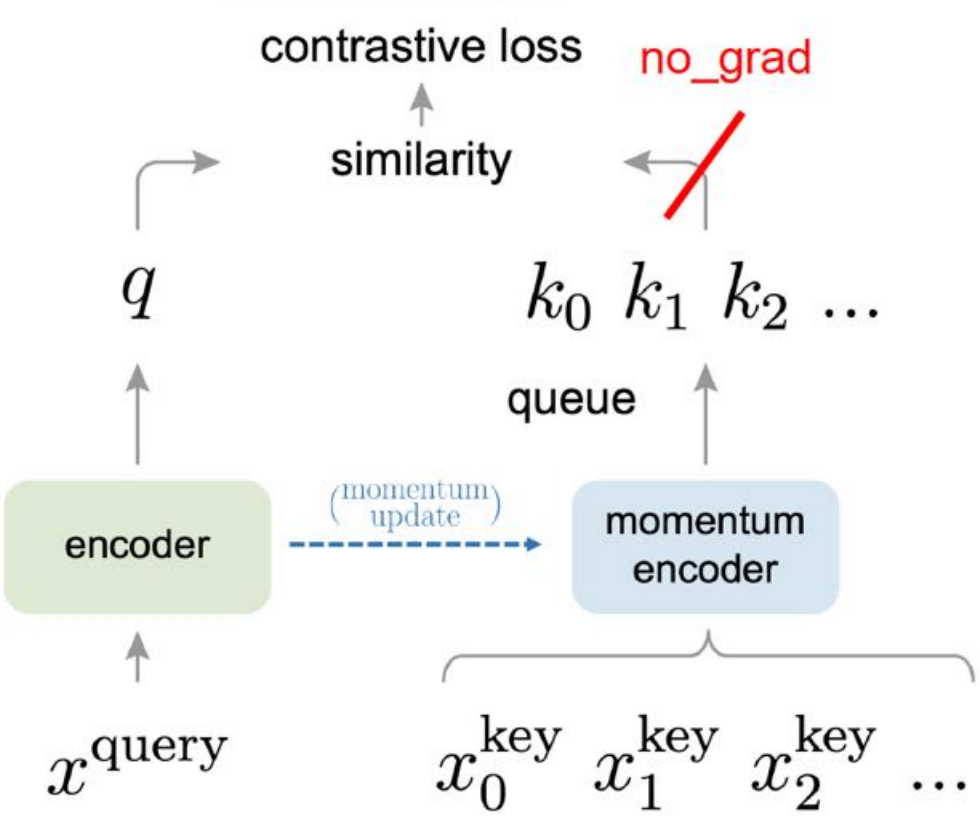
Mean teachers have been shown to improve the results:

- 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

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.

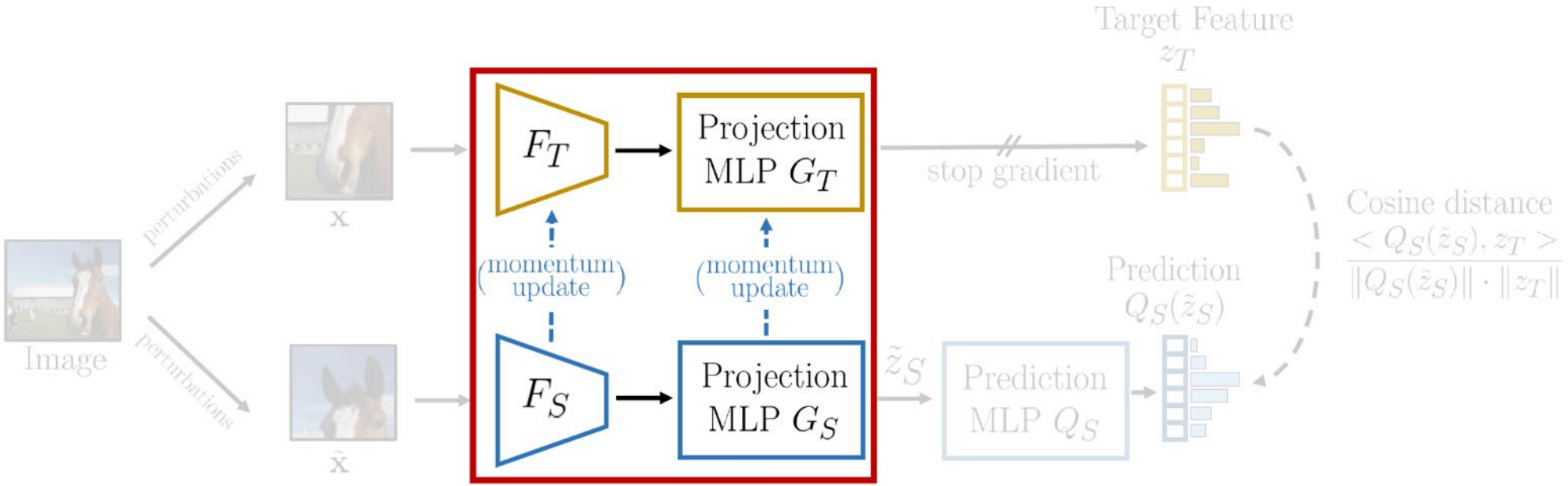
The key encoder is slowly progressing through the momentum update rules:

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

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

slide credit: Spyros Gidaris, Andrei Bursuc

Back to BYOL - Mean teacher for the feature reconstruction task

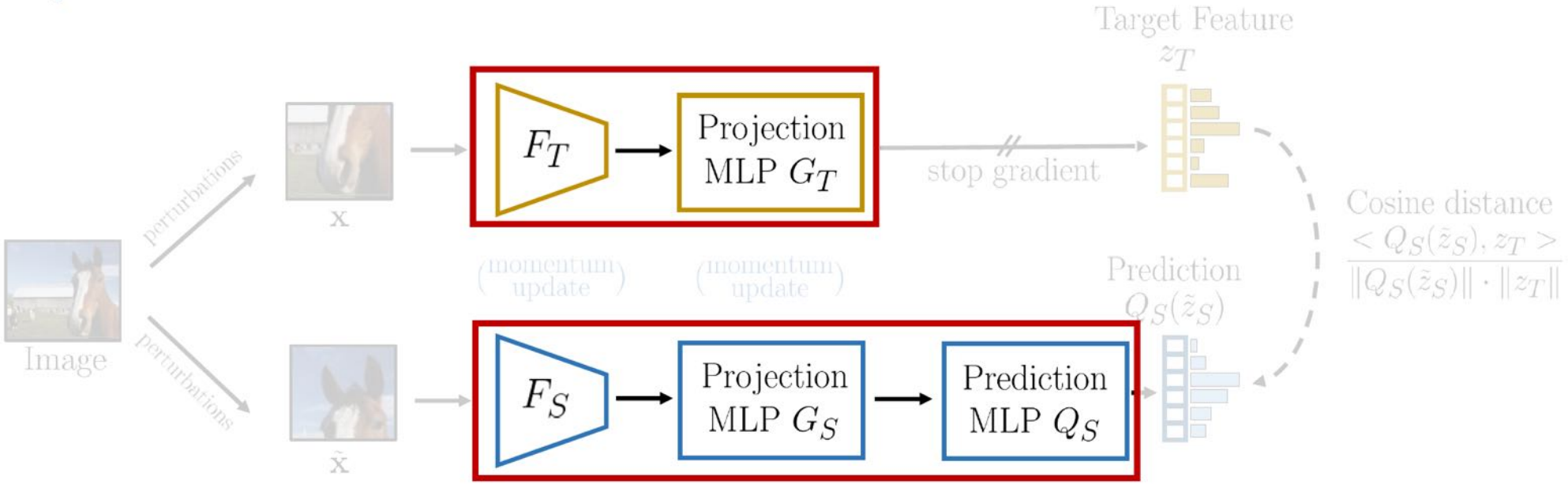


A mean teacher approach without any labels

- 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

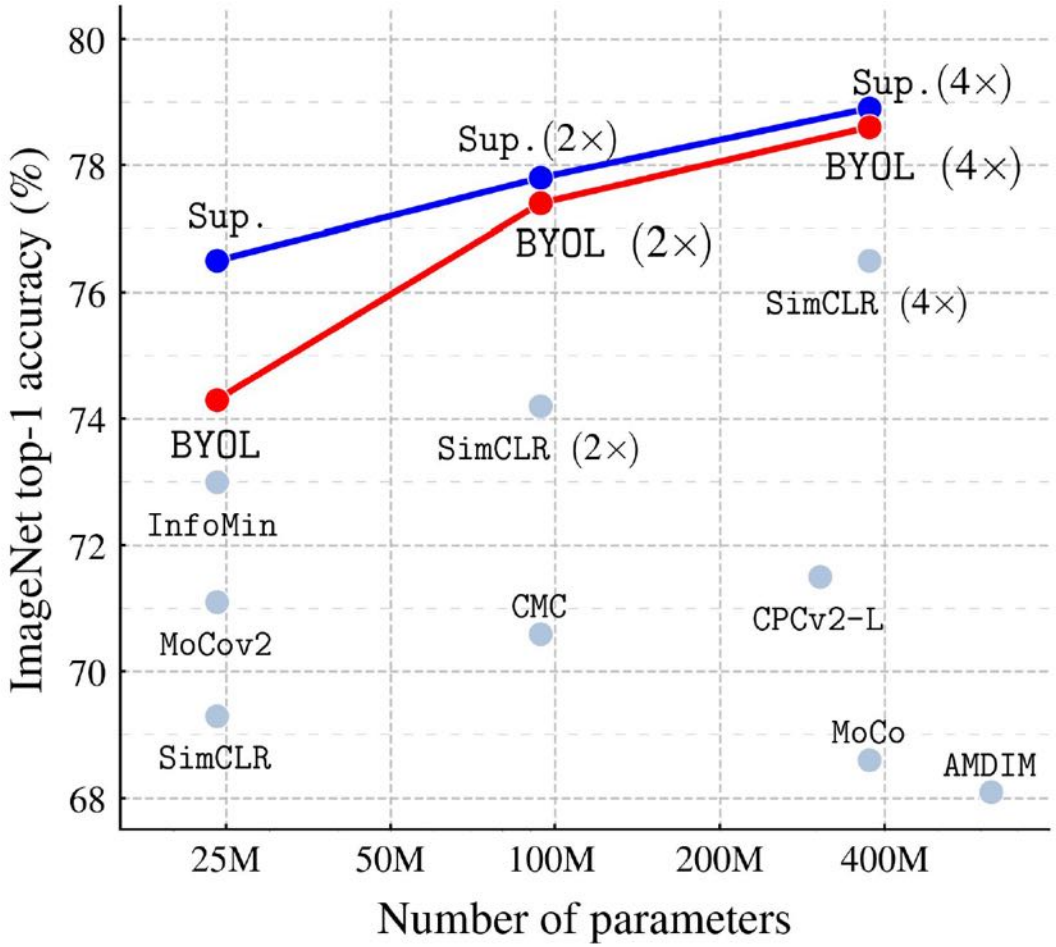
Asymmetric architecture



Asymmetric architecture: the student has an extra prediction MLP head

slide credit: Jean-Basien Grill, et al.

Transfer Using Linear Probe on BYOL Features

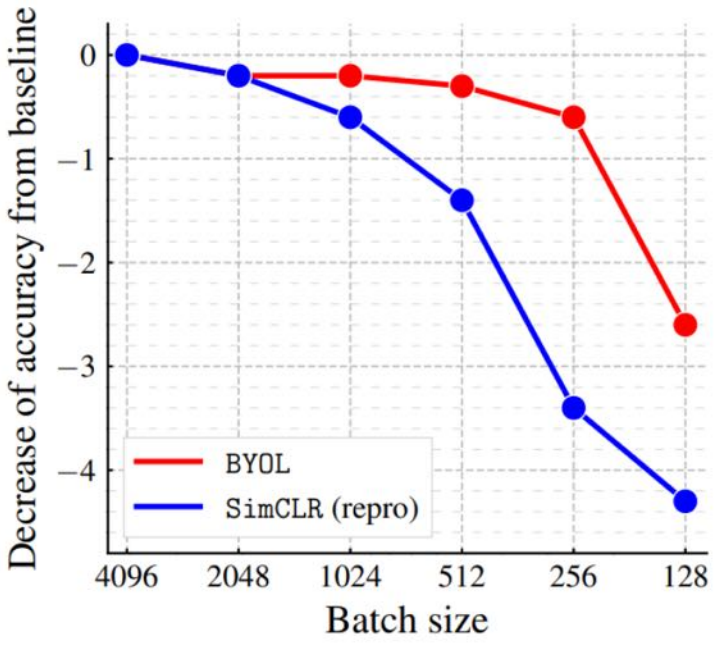


Note: these supervised baselines are from SimCLR (Chen & Hinton, ICML 2020)

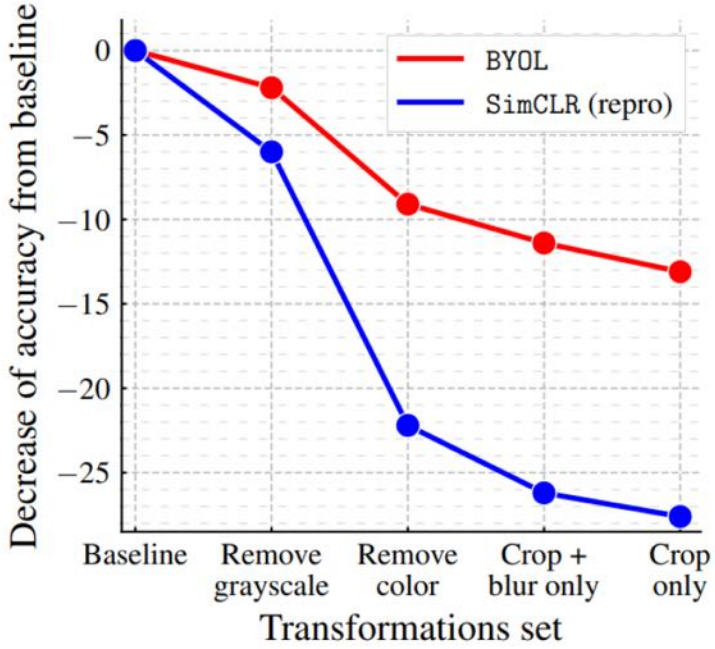
CPCv2: van den Oord et al., *Representation learning with contrastive predictive coding*. 2018
AMDIM: Bachman et al., *Learning representations by maximizing mutual information across views*. 2019
CMC: Tian et al.,*Contrastive multiview coding*. 2019.
MoCo: He et al., *Momentum contrast for unsupervised 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

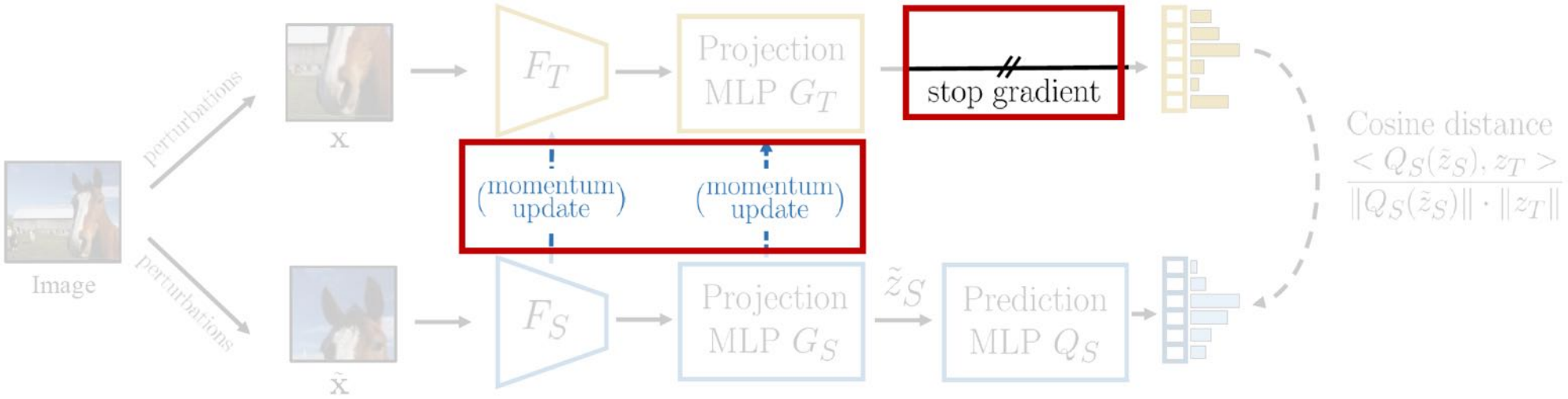
- BYOL **does not require negative examples** as the contrastive method SimCLR
- **More robust** to the choice of image augmentations and the batch-size
- Cropping is more important for BYOL and color jittering more important for SimCLR

slide credit: Spyros Gidaris, Andrei Bursuc

Key question: Why it avoids feature collapse?

slide credit: Spyros Gidaris, Andrei Bursuc

Why it avoids feature collapse?



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

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"**

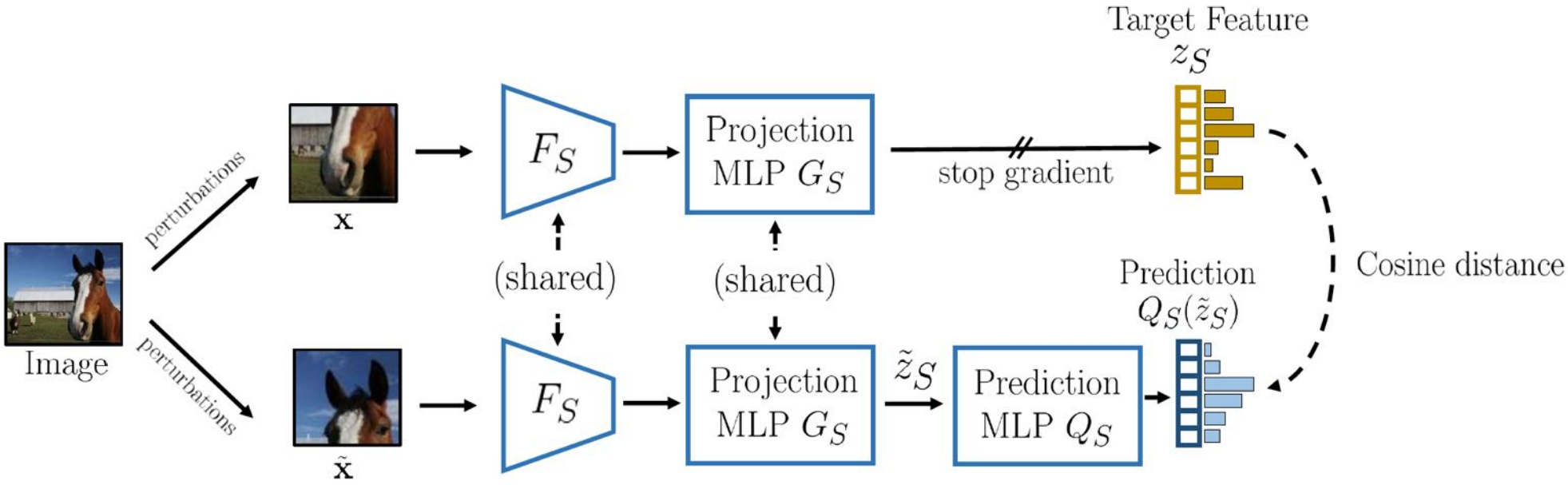
- Either by better tuning the network initialization
- Or replacing BN with Group Normalization and Weight Standardization (GN + WS)

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

SimSiam

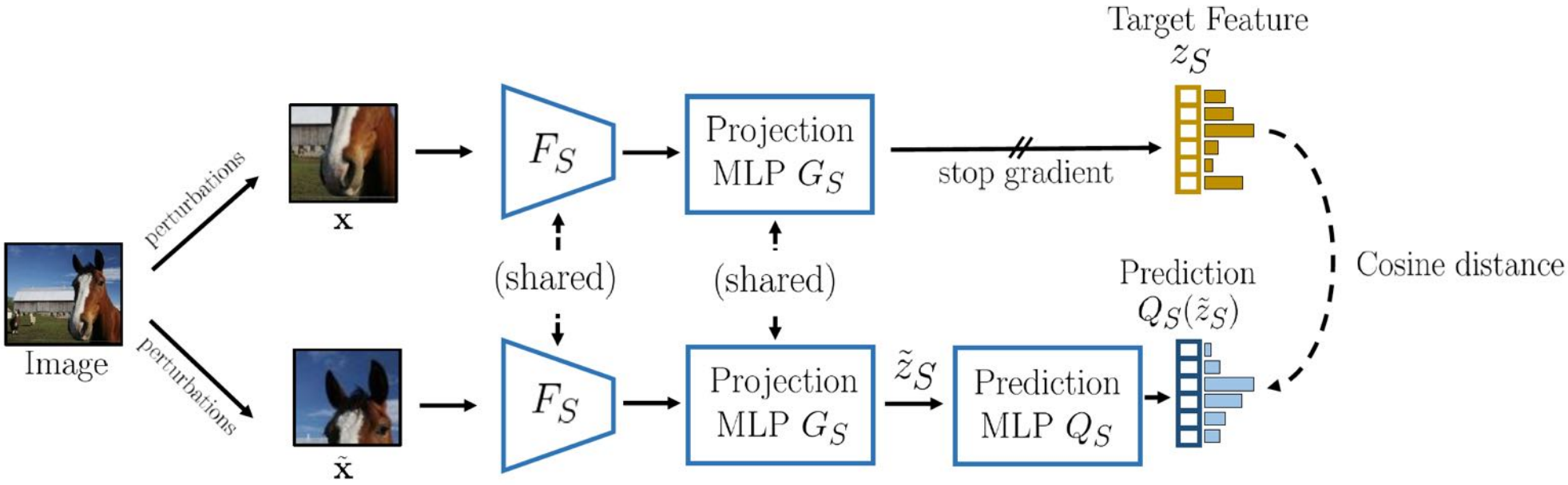


SimSiam: BYOL without the momentum teacher (the teacher is identical to the student)

"Exploring Simple Siamese Representation Learning", CVPR 2021

slide credit: Spyros Gidaris, Andrei Bursuc

SimSiam

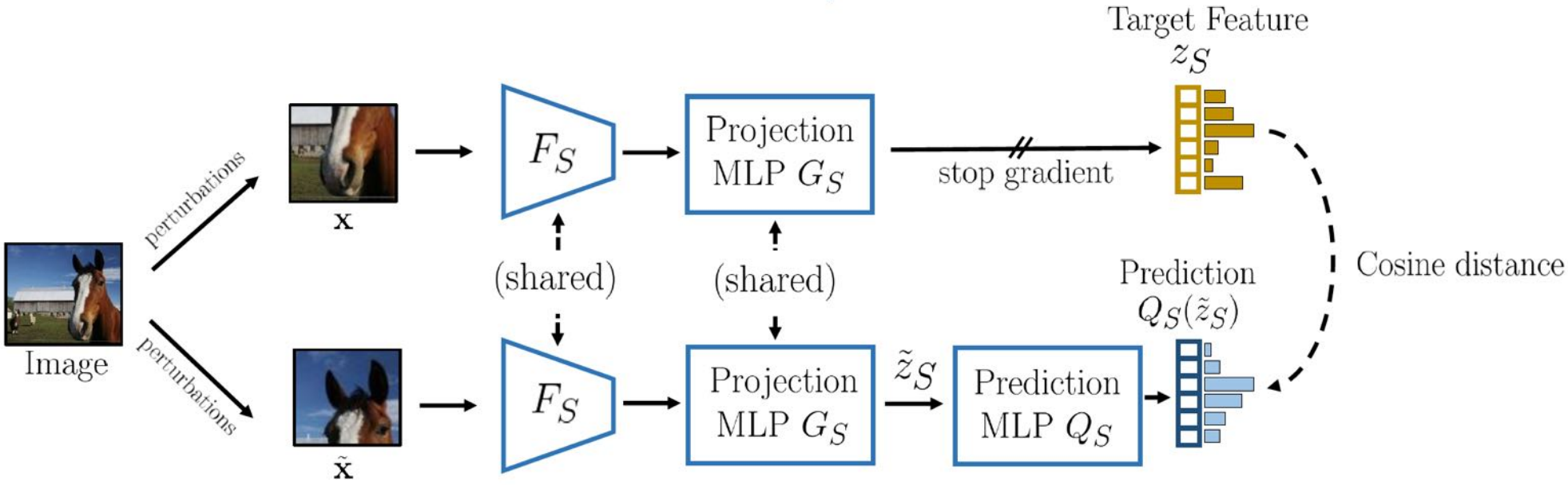


Momentum teacher: **improves performance but not necessary for avoiding feature collapse**

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

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

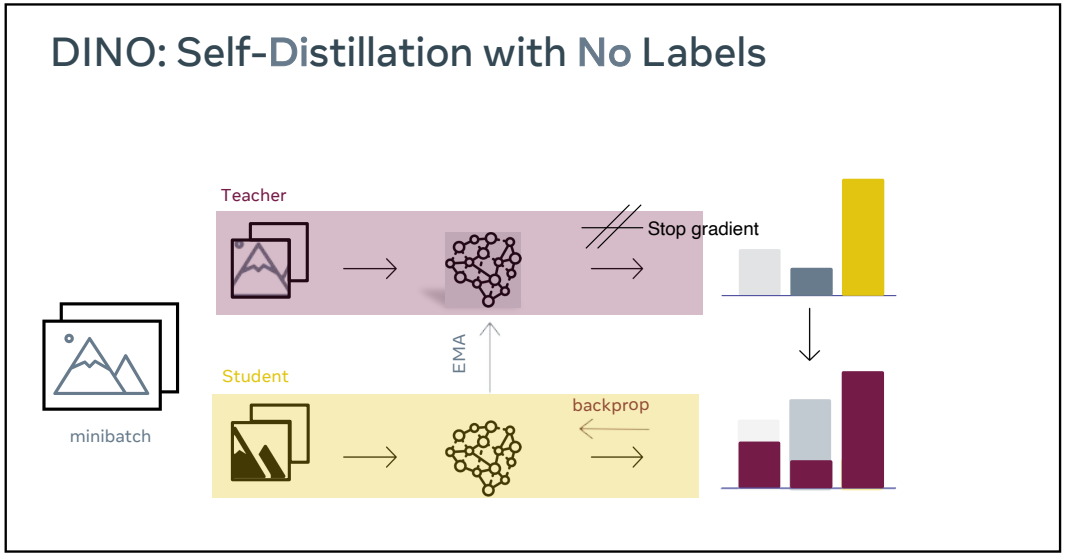
slide credit: Moritz Boehle

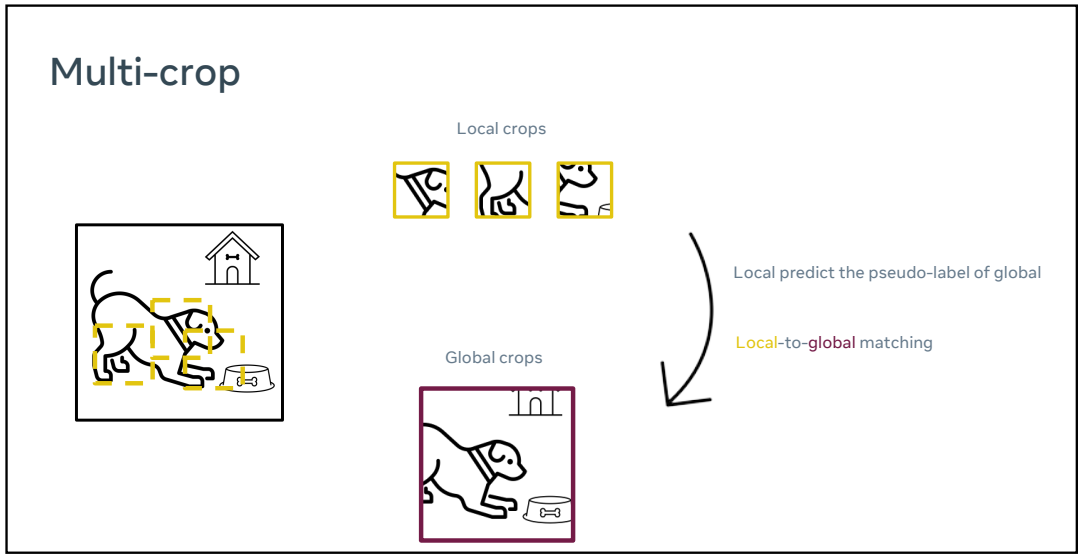
Source: https://github.com/facebookresearch/dino

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

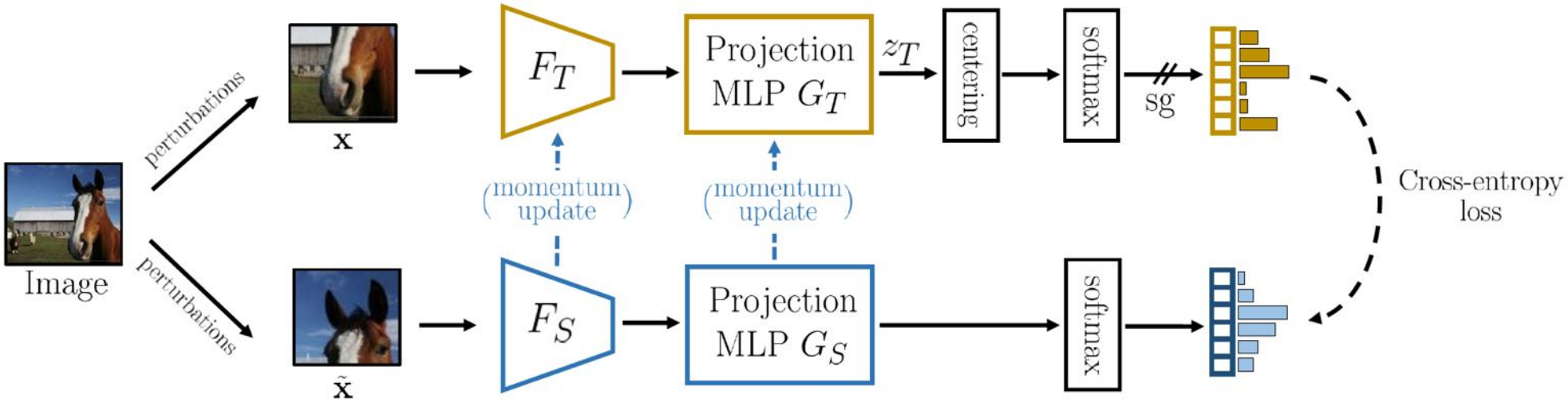
DINO-Pipeline: Summary

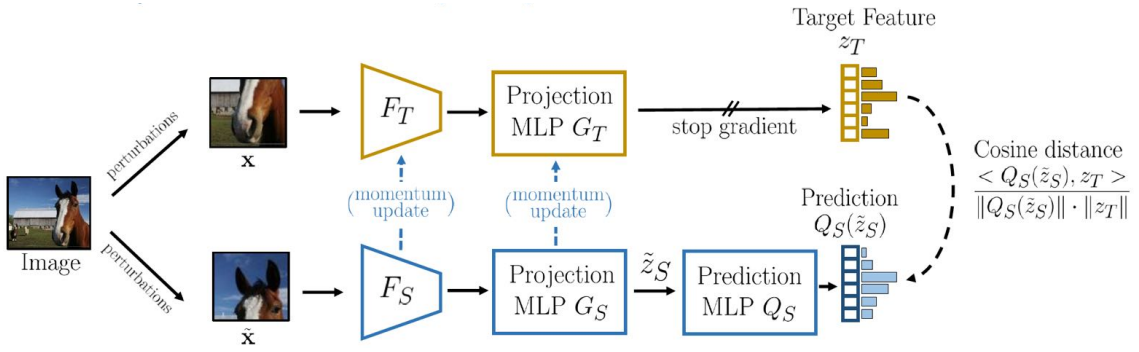
In short: Self-defining classification task, main ingredients: EMA and multi-crop, trained with CE-Loss

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

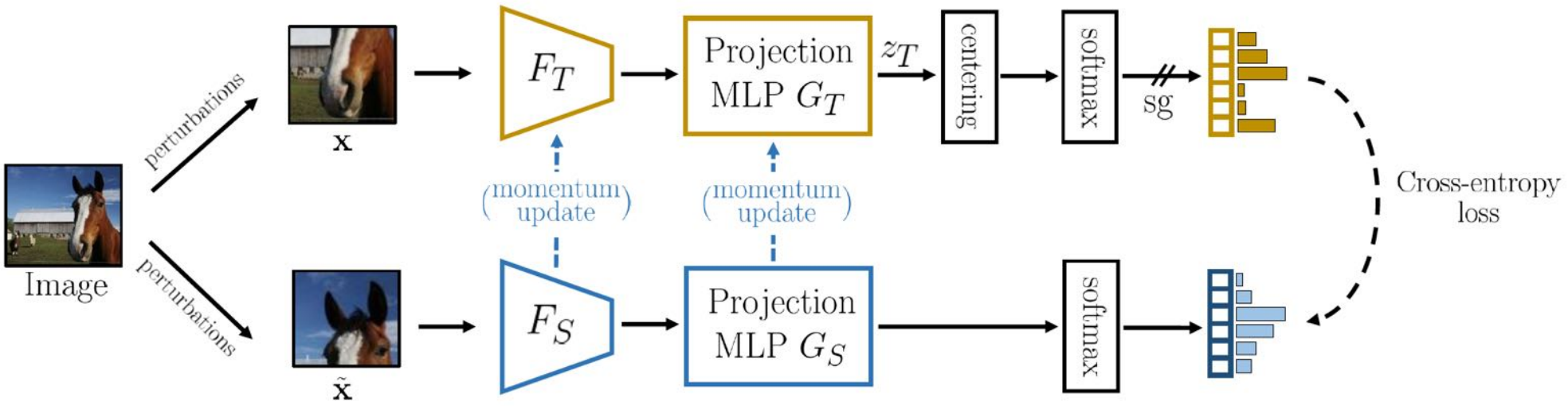
slide credit: Spyros Gidaris, Andrei Bursuc

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

- Training a linear classifier on top
- k-NN evaluation

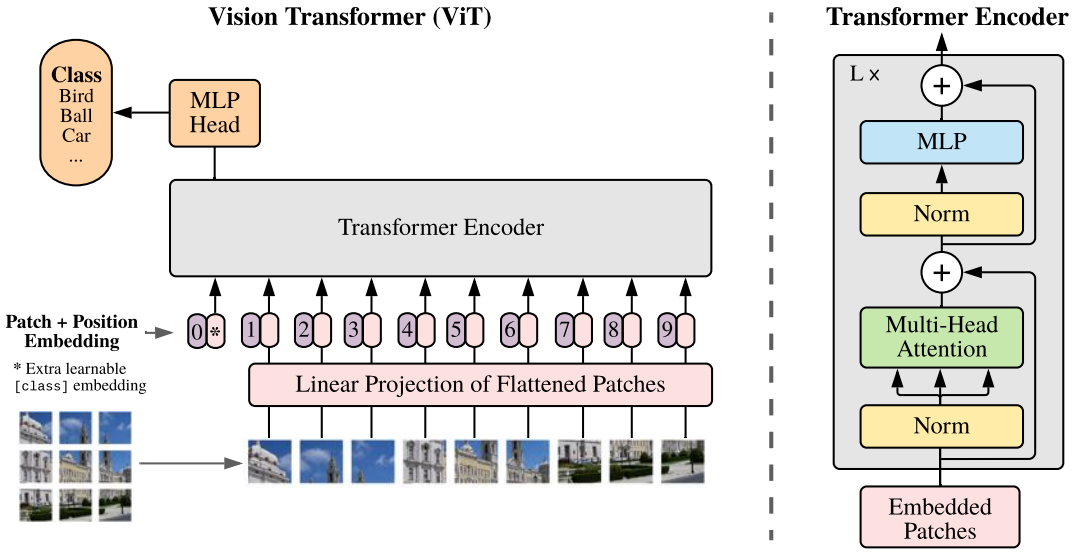
- DINO works well across architectures
  - But best with ViTs

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?

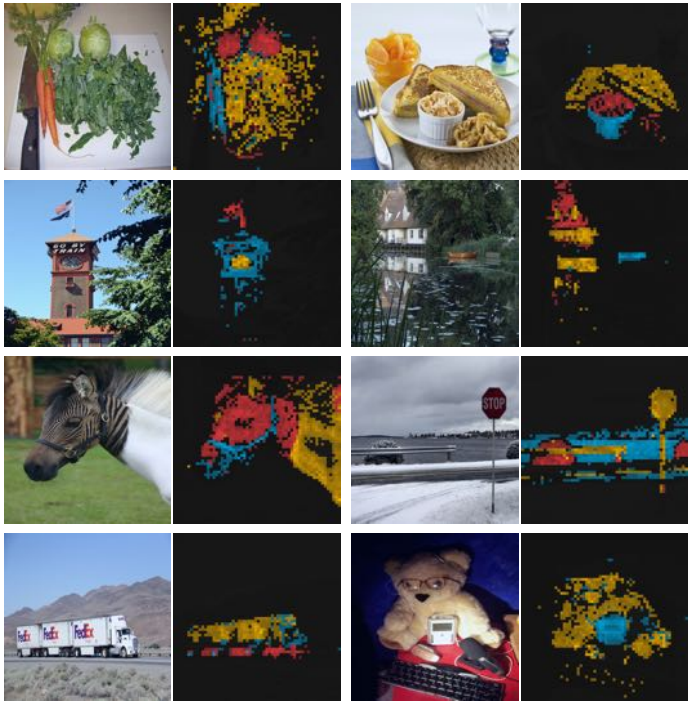
Supervised

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
  - …

Overview of Today's Lecture

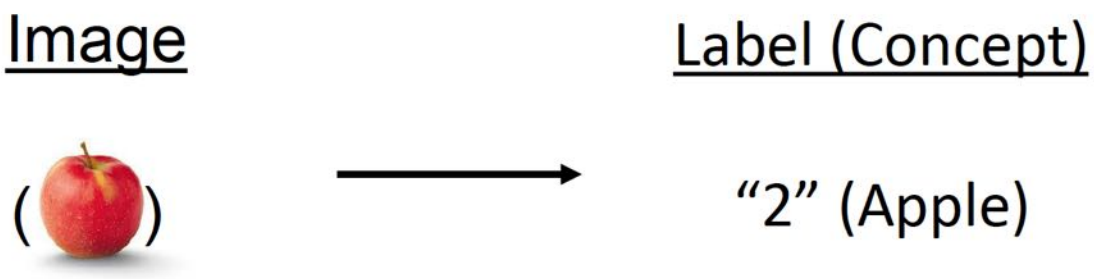
- Continuation of Self-Supervised Learning
  - Teacher-Student "feature reconstruction"
    - motivation, setting
    - methods: BYOL, DINO

- Vision-Language Learning for Computer Vision
  - Supervised learning vs. vision-language learning
  - CLIP, ALIGN
  - Some extensions

slide credit: Jianwei Yang

Supervised Learning

Map an image to a discrete label
which is associated a visual concept



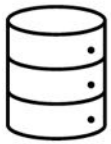
slide credit: Jianwei Yang

Supervised Learning

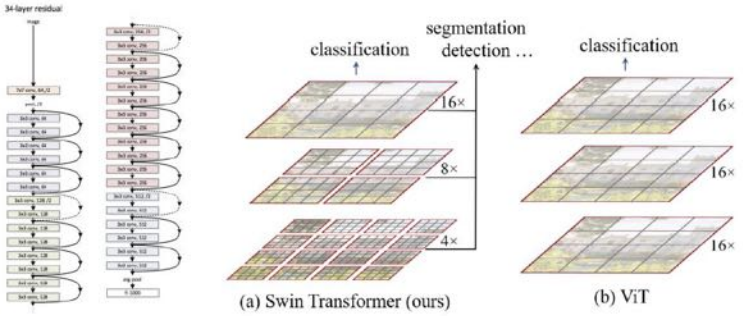
MNIST. LeCun *et al.*

CIFAR-10. Krizhevsky *et al.*

ImageNet. Deng *et al.*

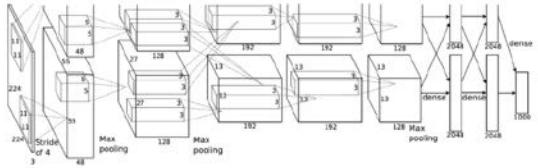


labels



ResNet. He *et al.* **Swin**. Liu *et al.*

images



AlexNet. Krizhevsky *et al.*

slide credit: Jianwei Yang

Supervised Learning

- Pros
  - Densely labeled samples for each category
- Cons
  - Requires a lot of human effort
  - Limited number of categories

slide credit: Jianwei Yang

Zero-Shot Learning (Canonical)

Map an image to description of a visual concept

Image		Descriptions (Concept)
	→	Fruit, Red, Sphere (Apple)
	→	Fruit, Yellow (Orange)

slide credit: Jianwei Yang

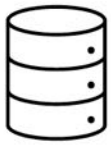
Zero-Shot Learning (Canonical)

CUB-200-2011. Wah *et al.*

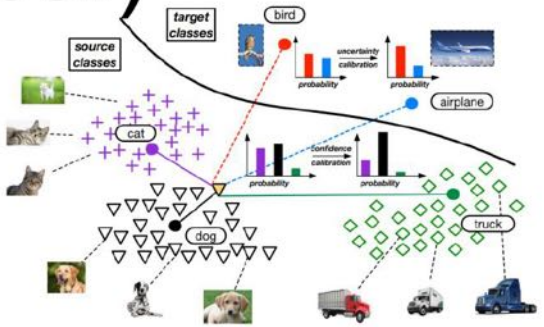
AwA2. Xian *et al.*

aPY. Farhadi *et al.*

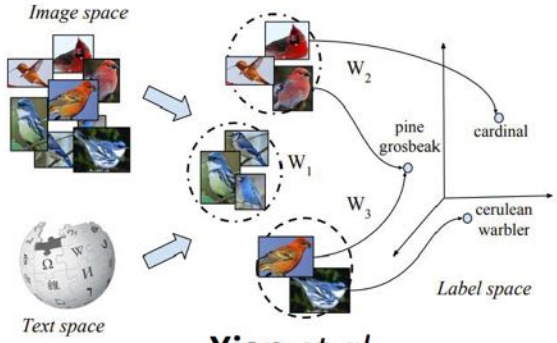
Label & descriptions



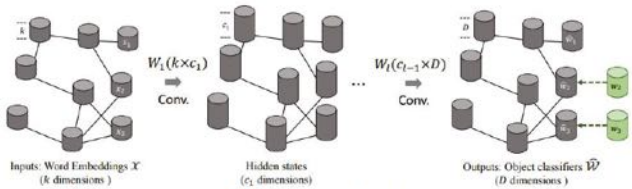
images



Liu *et al.*



Xian *et al.*



Wang *et al.*

slide credit: Jianwei Yang

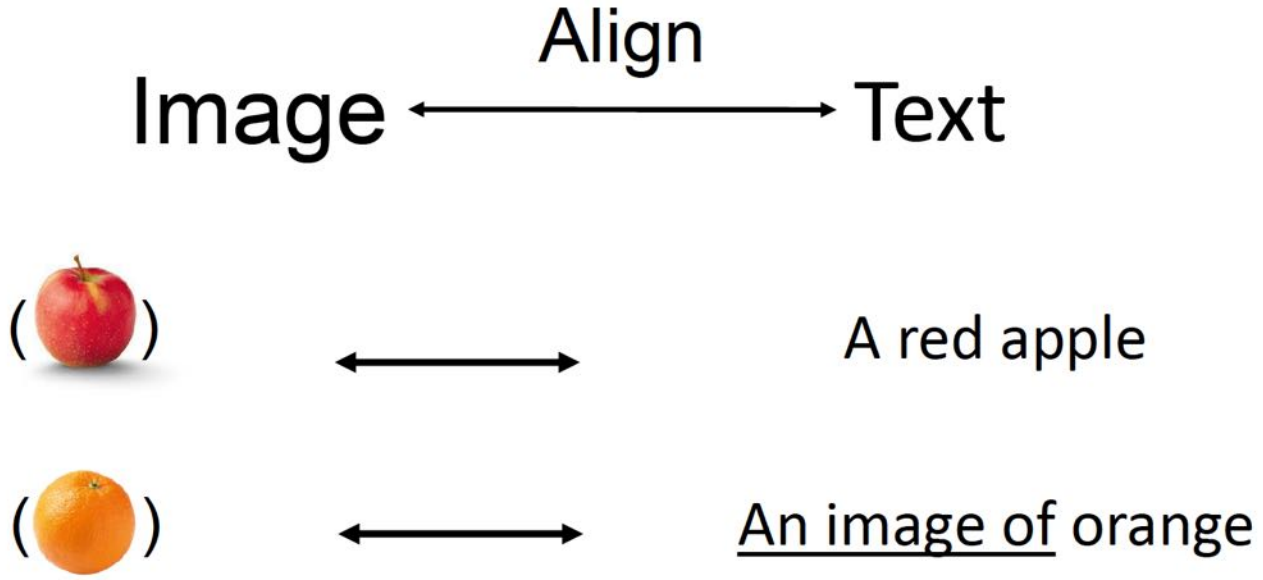
Zero-Shot Learning (Canonical)

- Pros
  - Directly learn the visual-semantic matching

- Cons
  - Small scale with limited vocabulary
  - Fixed visual and text encoder

slide credit: Jianwei Yang

Contrastive Vision-Language Learning



slide credit: Jianwei Yang

Contrastive Vision-Language Learning

Align

Image ⟷ Text

End-to-end learning on large-scale corpus

() ⟷ An image of orange

slide credit: Jianwei Yang

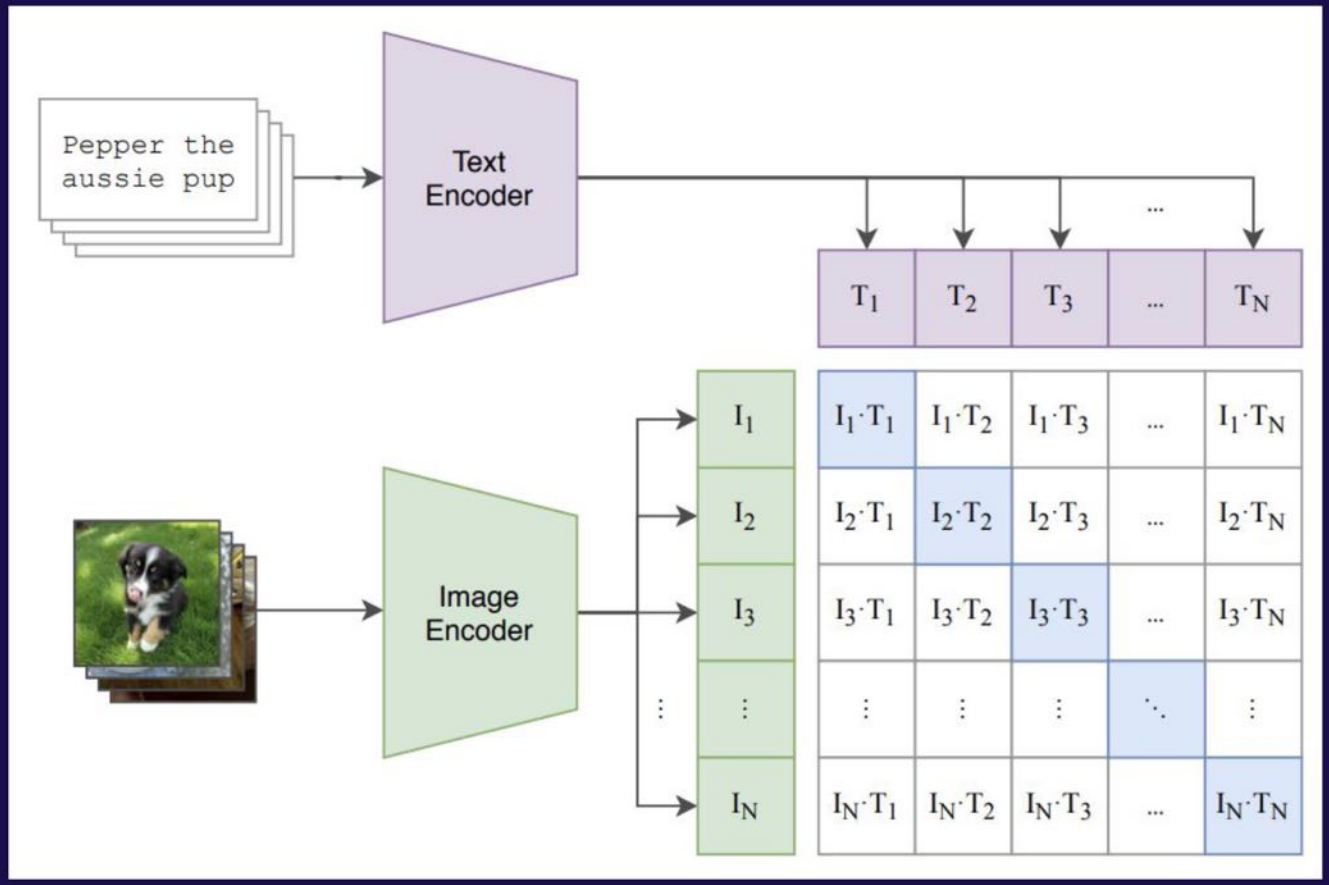
The most recent art

Contrastive Vision-Language Learning

A lot of research works come along the line of vision-language learning for vision

Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

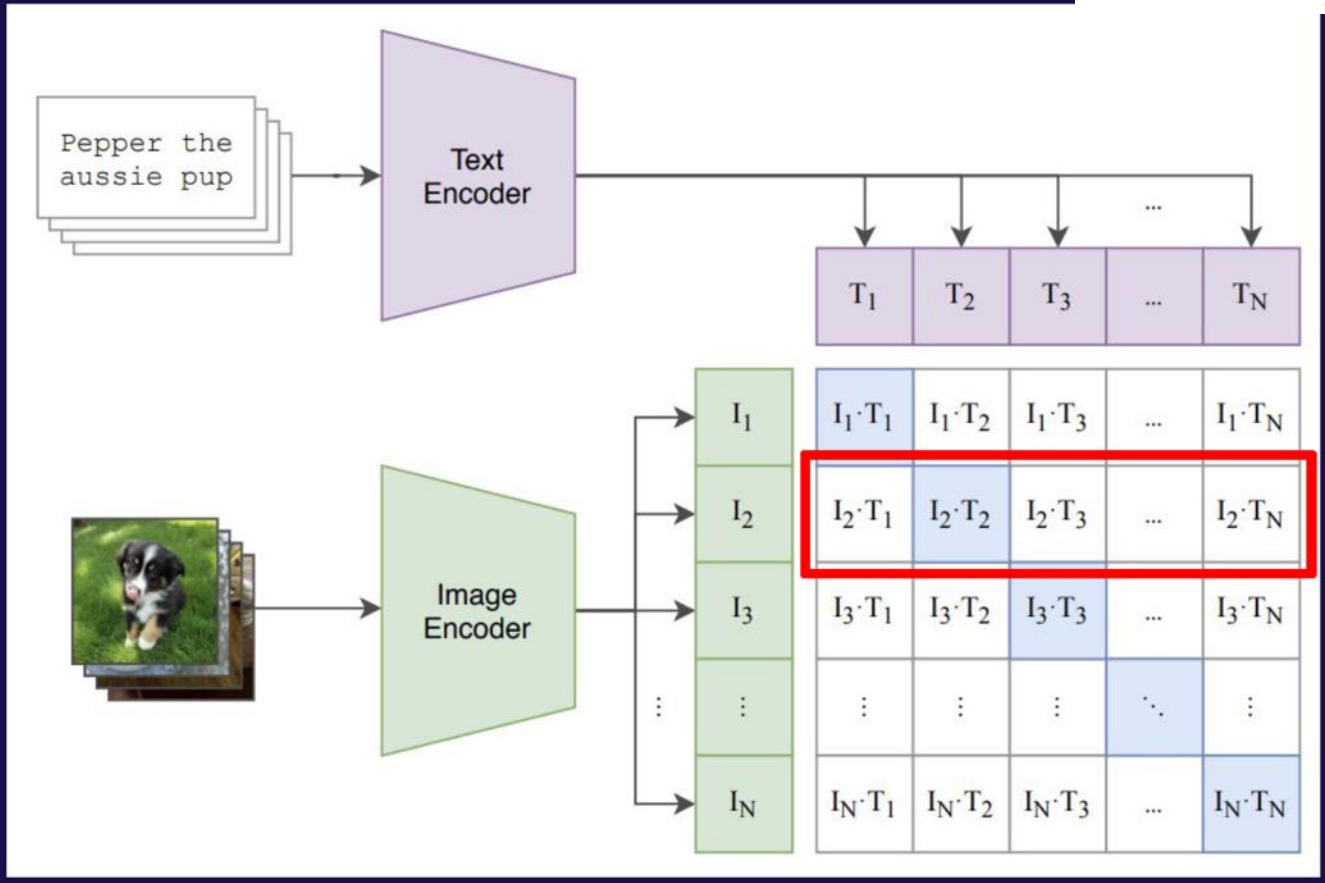
CLIP: Contrastive Language-Image Pre-training



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

CLIP: Contrastive Language-Image Pre-training

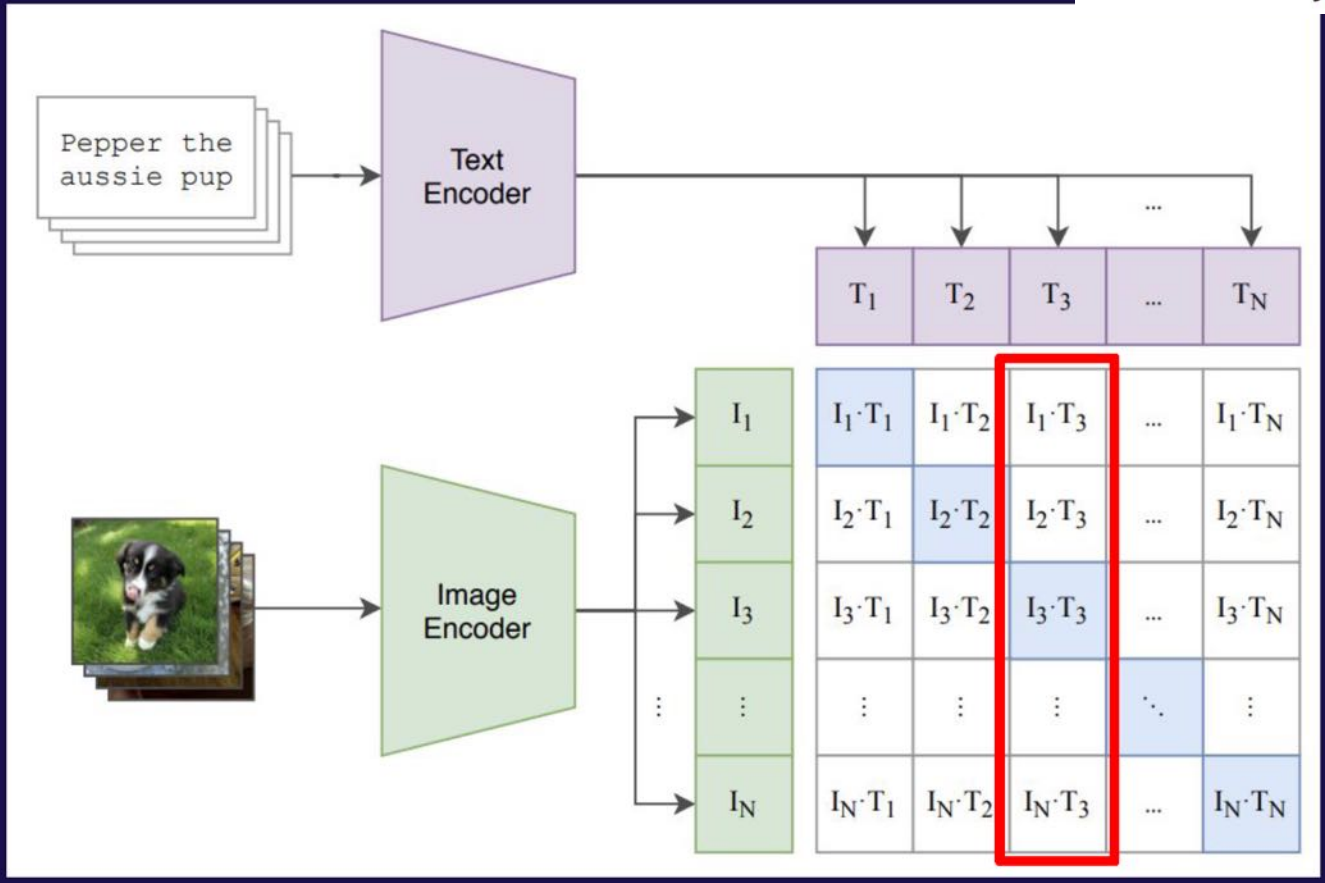
$$L_{i2t} = -\sum_{j} \log \frac{\exp I_j T_j^T}{\sum_k \exp I_j T_k^T}$$



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

CLIP: Contrastive Language-Image Pre-training

$$L_{t2i} = -\sum_j \log \frac{\exp I_j T_j^T}{\sum_k \exp I_k T_j^T}$$



Some CLIP details

Training

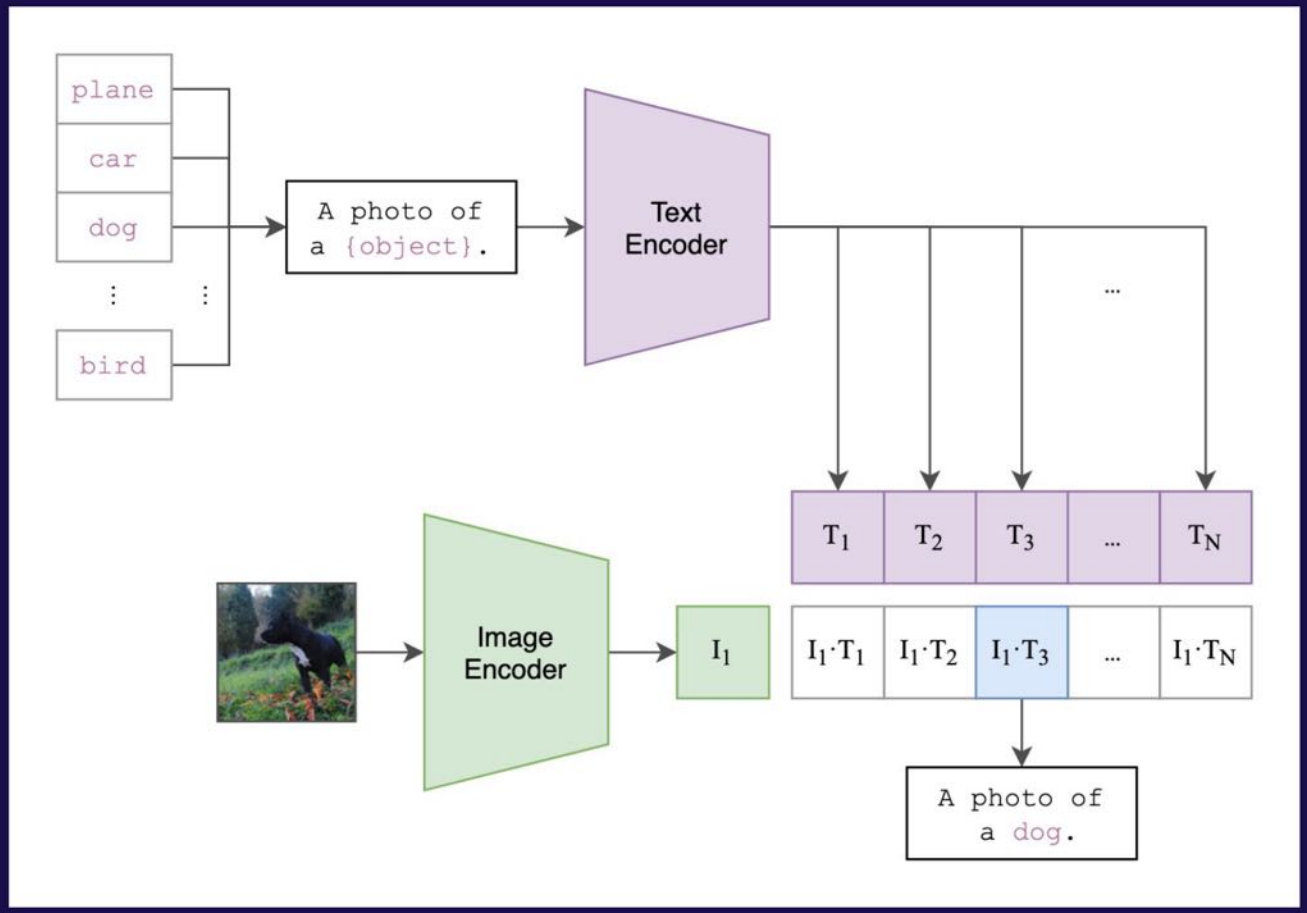
- Trained on 400M image-text pairs from the internet
- Batch size of 32,768
- 32 epochs over the dataset
- Cosine learning rate decay

Architecture

- ResNet-based or ViT-based image encoder
- Transformer-based text encoder

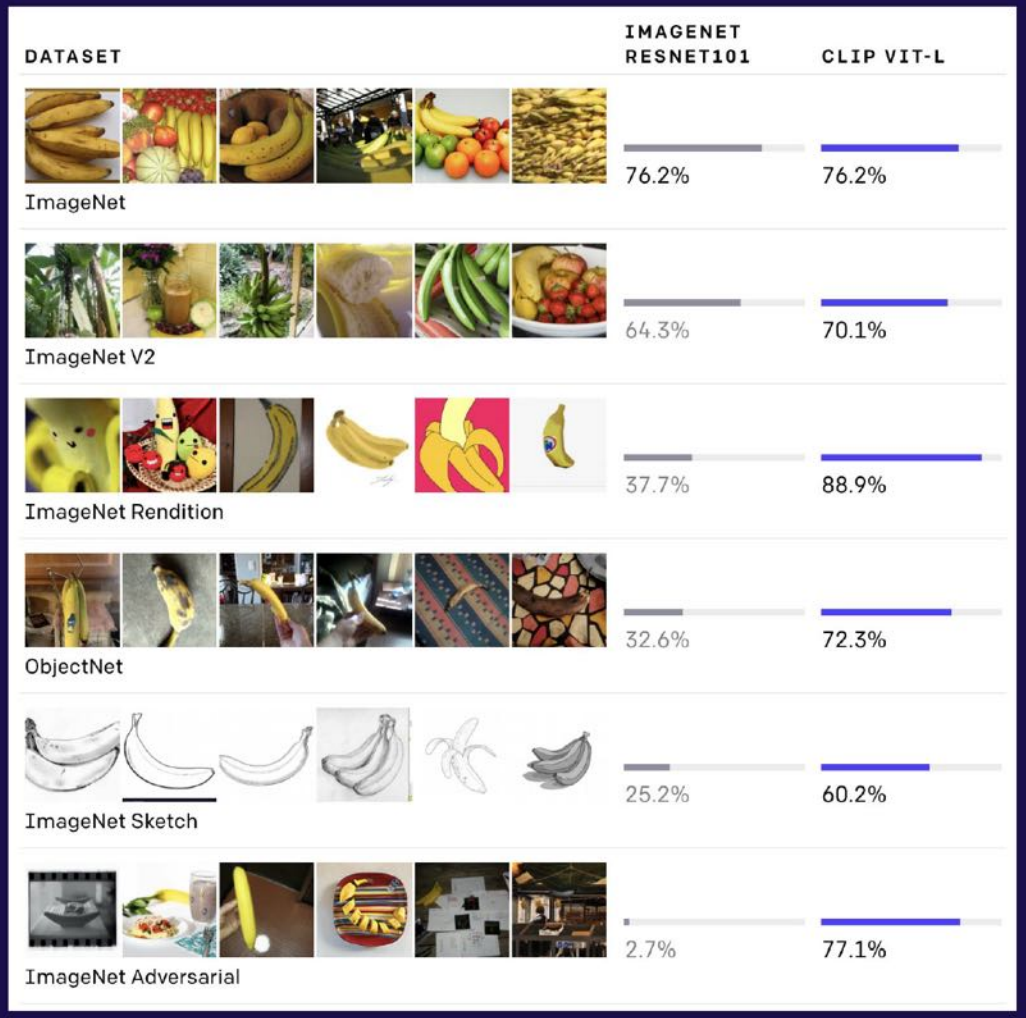
Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Zero-shot image classification



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Zero-shot CLIP is much more robust



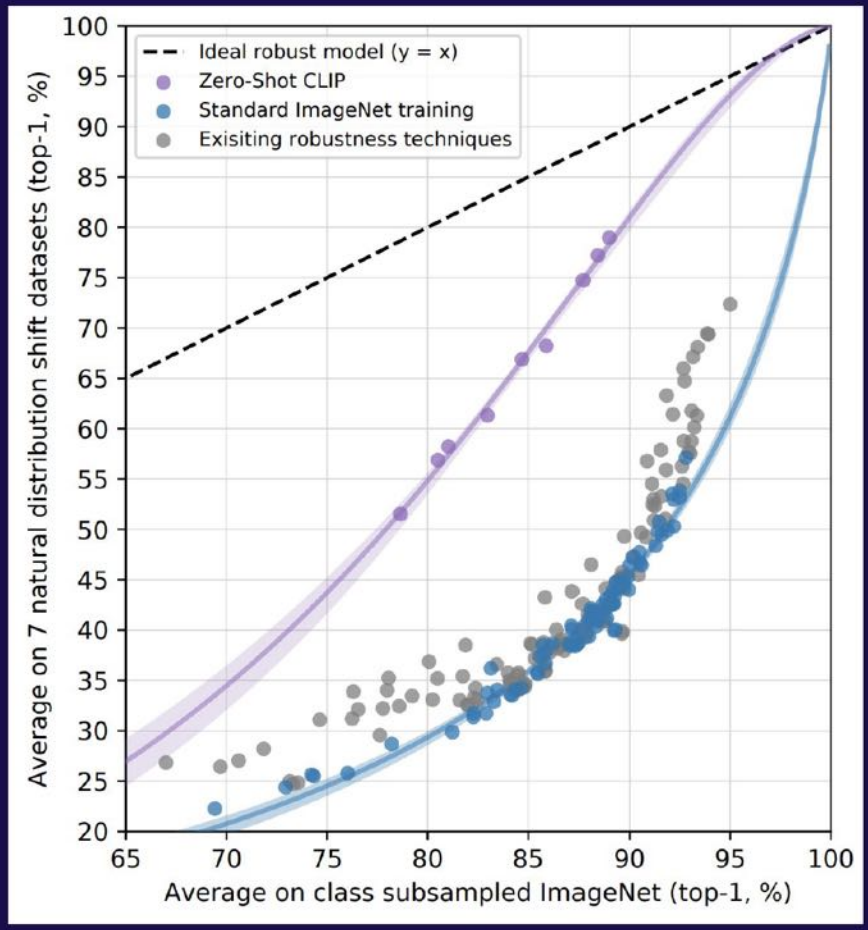
Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Robustness to natural distribution shift

CLIP is significantly more robust!

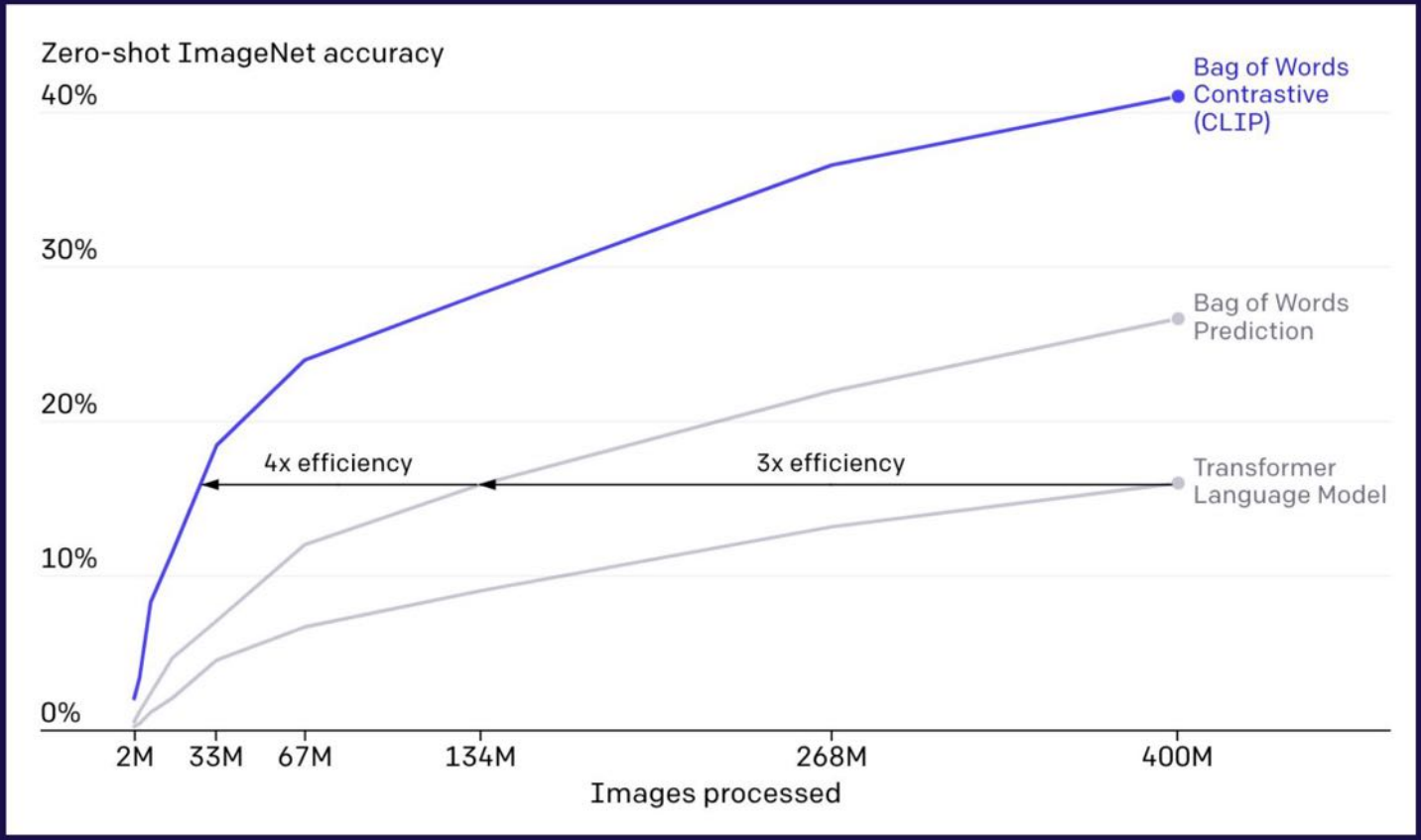
7 ImageNet-like Datasets (Taori et al.)

- ImageNetV2
- ImageNet-A
- ImageNet-R
- ImageNet Sketch
- ObjectNet
- ImageNet Vid
- Youtube-BB



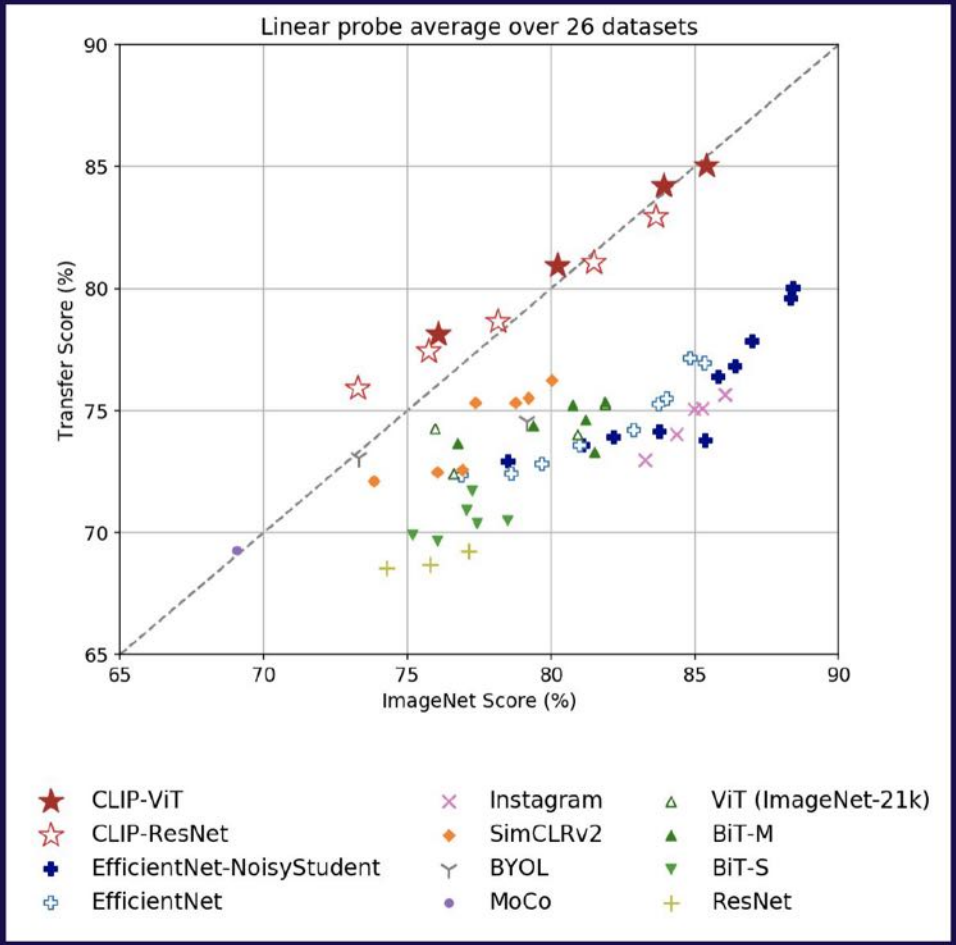
Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Why contrastive



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

vs ImageNet score

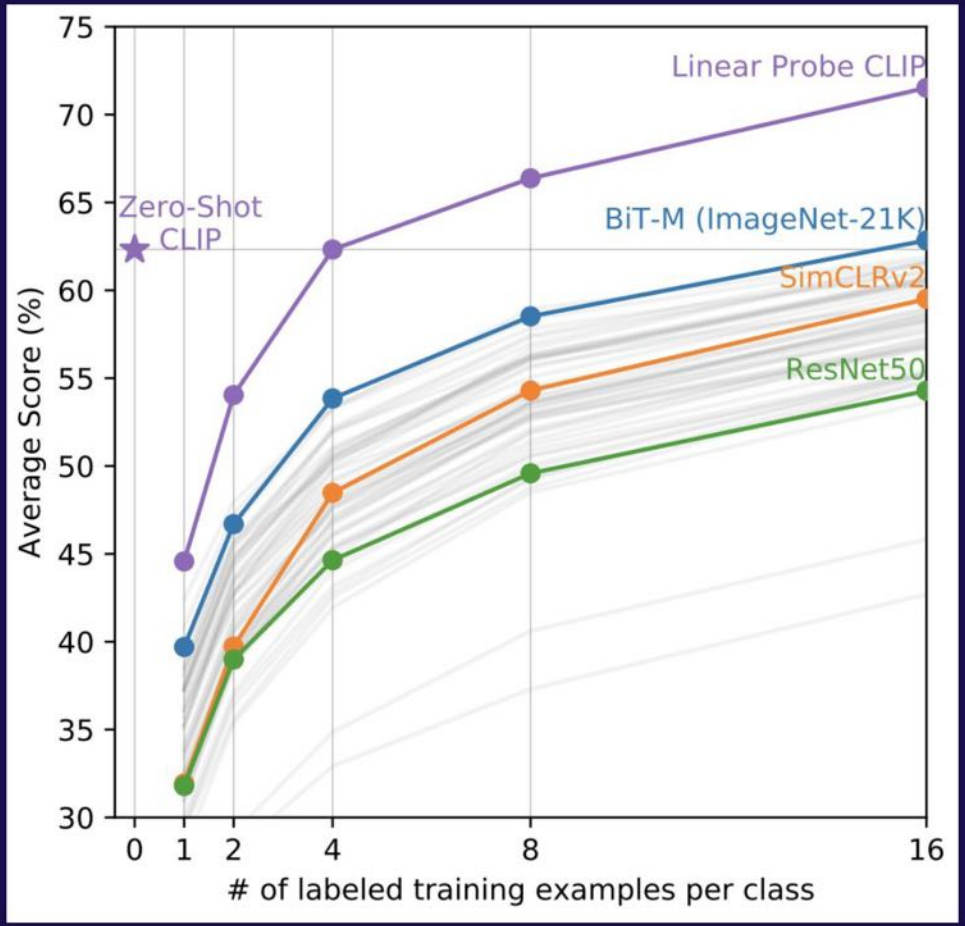


Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Zero-shot CLIP vs Few-shot linear probes

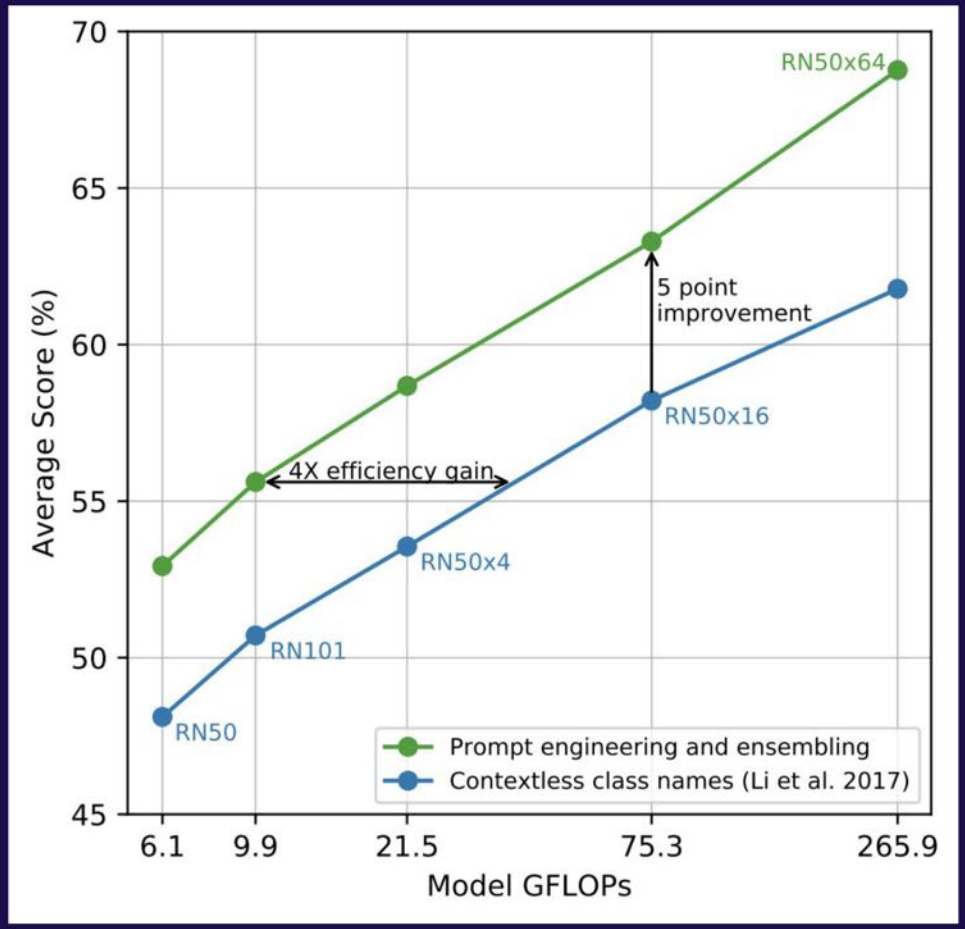
Zero-shot CLIP is as good as

- 4-shot linear-probe CLIP
- 16-shot BiT-M



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Prompt engineering



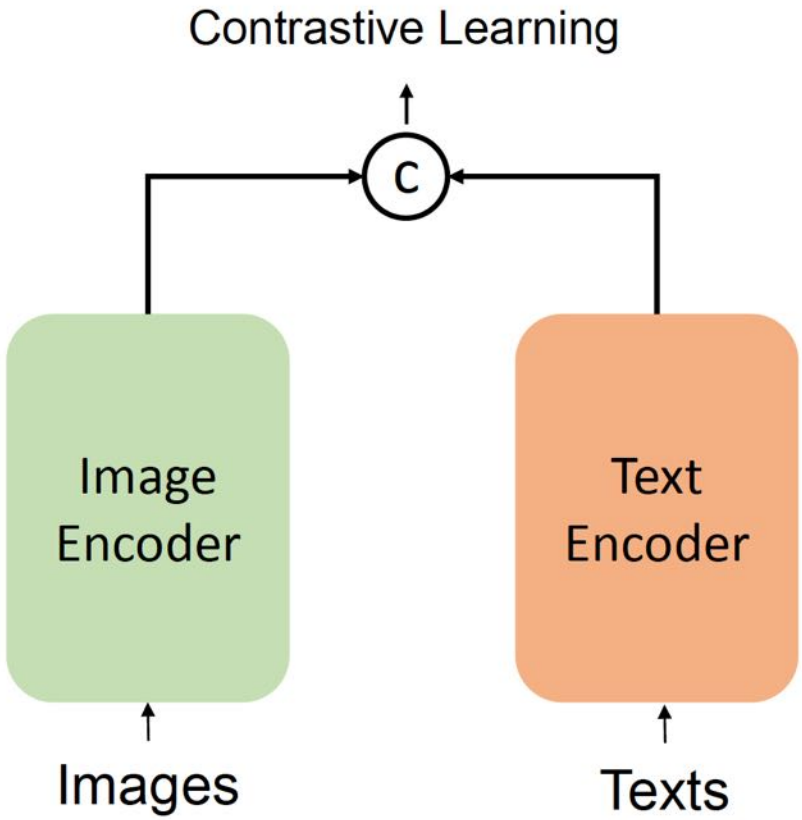
slide credit: Jianwei Yang

The Lesson from CLIP

- Image recognition can be formulated as an **image-text matching problem** instead of image-label mapping problem
- Image recognition does not require human-annotated image-label data but **huge amount of (noisy) image-text pairs**
- **Contrastive learning is a good learning objective** for multi-modal learning strategy compared with generative learning
- **Two-tower model without fusion** is sufficient to learn good and generic visual and language representations

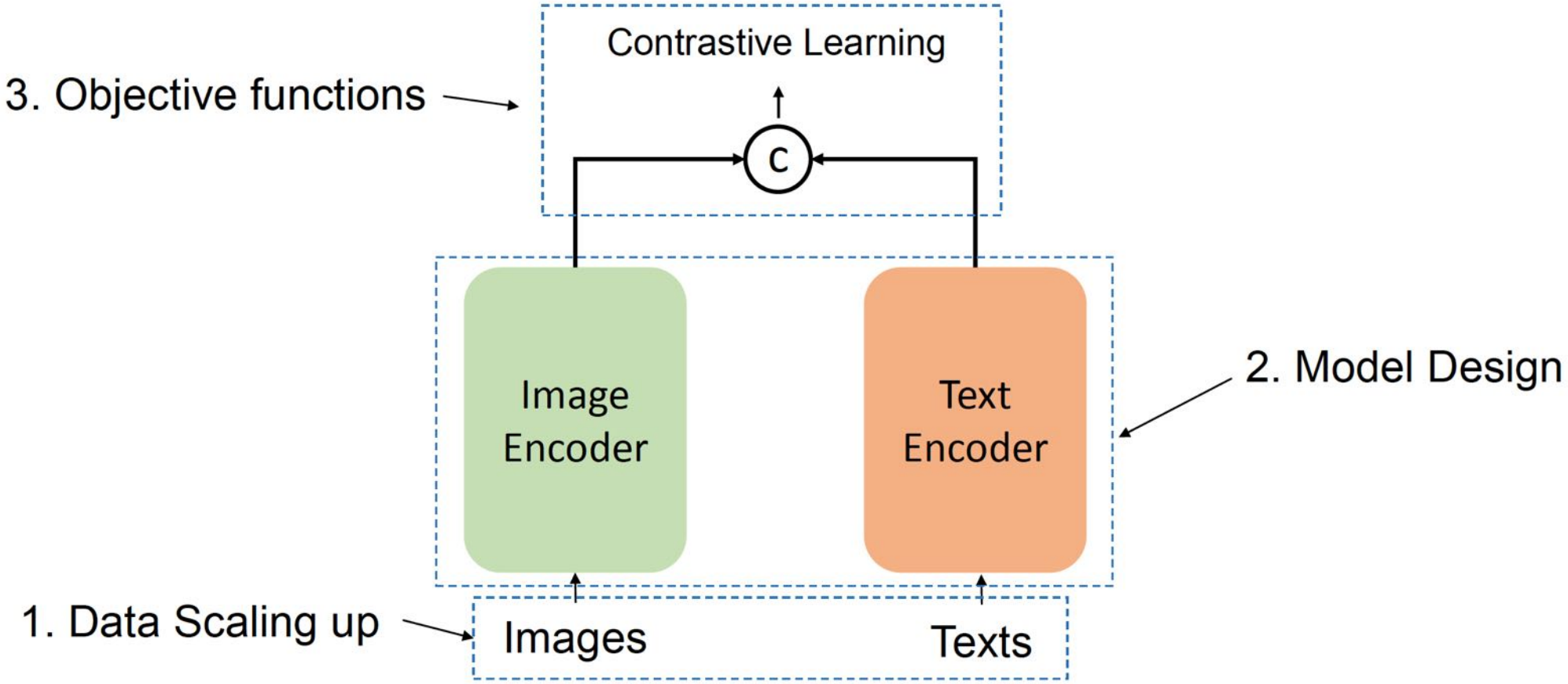
slide credit: Jianwei Yang

The most recent art



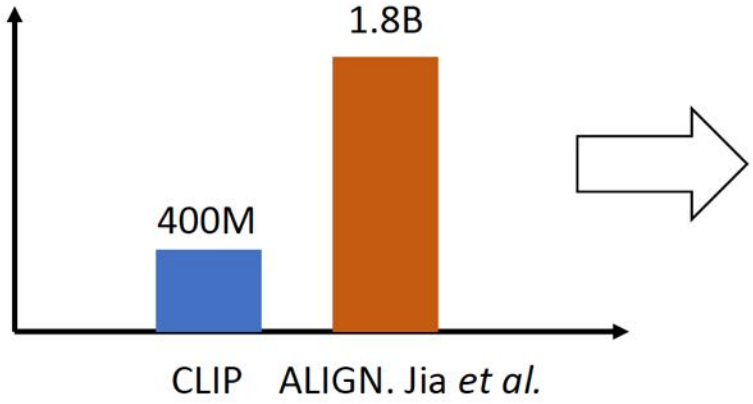
slide credit: Jianwei Yang

The most recent art



ALIGN = Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision - Jia et al@ICML'21 slide credit: Jianwei Yang

Data Scaling-Up



Larger scale but more noisy data


"motorcycle front wheel"

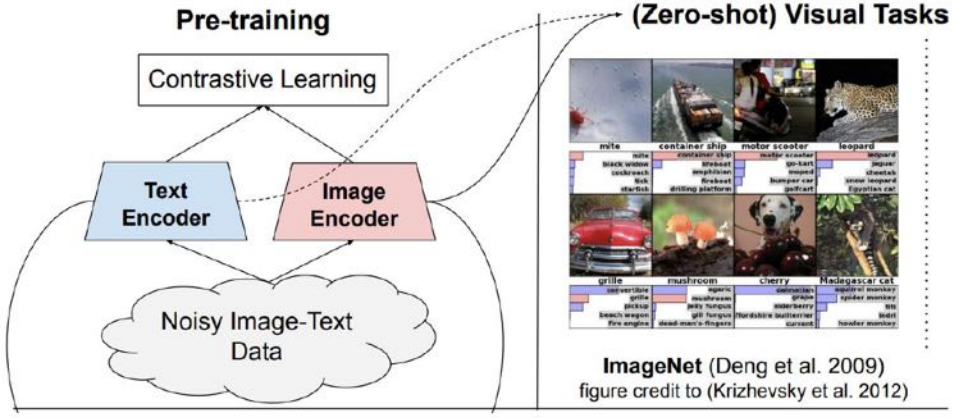

"*thumbnail for version as of 21 57 29 june 2010*"


"file frankfurt airport skyline 2017 05 jpg"


"file london barge race 2 jpg"


"moustache seamless wallpaper design"


"st oswalds way and shops"



Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	**77.2**	**70.1**
ALIGN	**76.4**	**92.2**	75.8	**70.1**

Zero-shot image classification on ImageNet

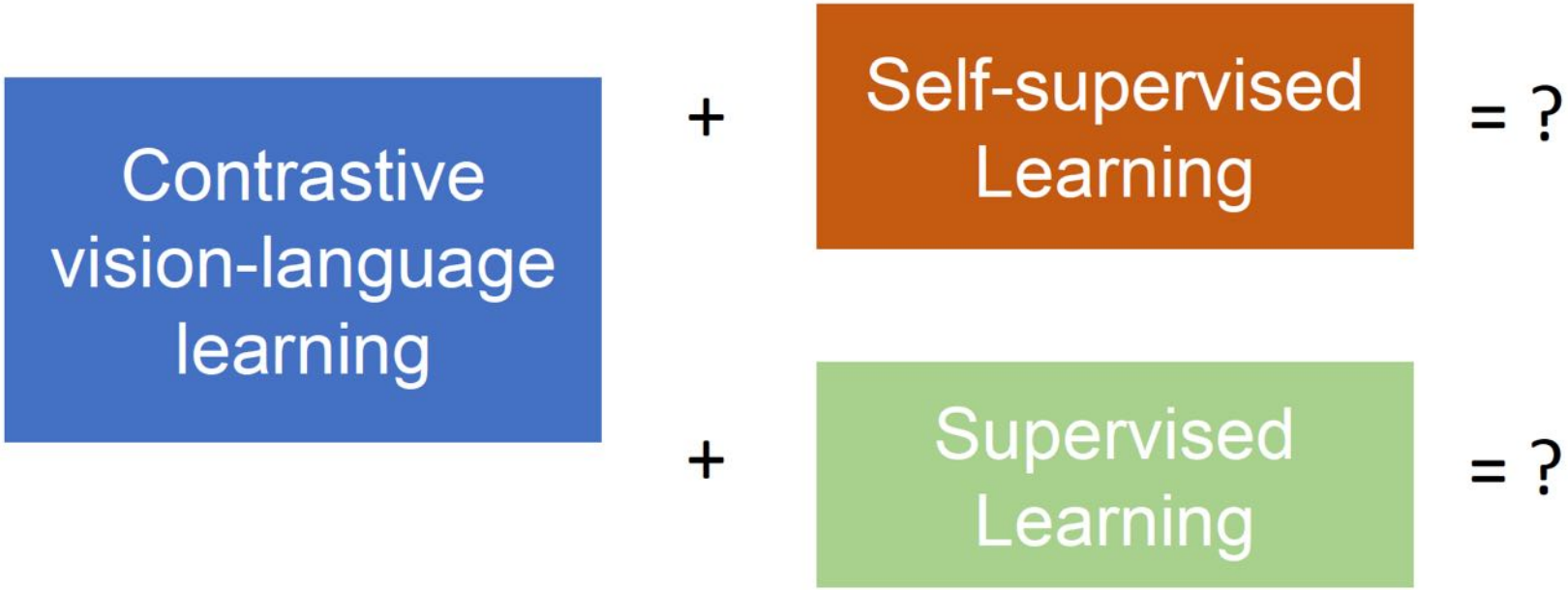
Model (backbone)	Acc@1 w/ frozen features	Acc@1	Acc@5
WSL (ResNeXt-101 32x48d)	83.6	85.4	97.6
CLIP (ViT-L/14)	85.4	-	-
BiT (ResNet152 x 4)	-	87.54	98.46
NoisyStudent (EfficientNet-L2)	-	88.4	98.7
ViT (ViT-H/14)	-	88.55	-
Meta-Pseudo-Labels (EfficientNet-L2)	-	**90.2**	**98.8**
ALIGN (EfficientNet-L2)	**85.5**	88.64	98.67

Image classification finetuning

slide credit: Jianwei Yang

Learning Objectives

Combining contrastive learning with other learning objectives

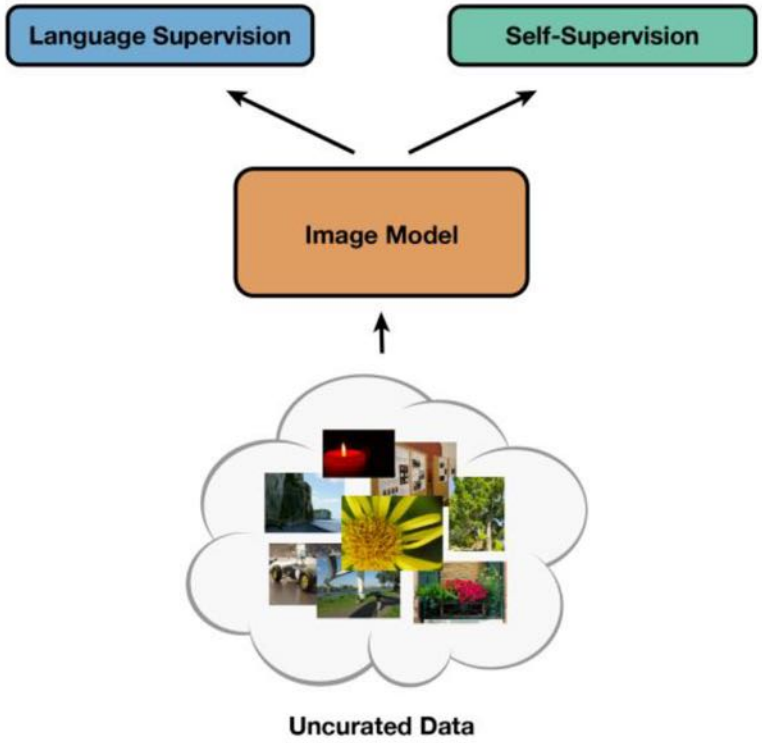


slide credit: Jianwei Yang

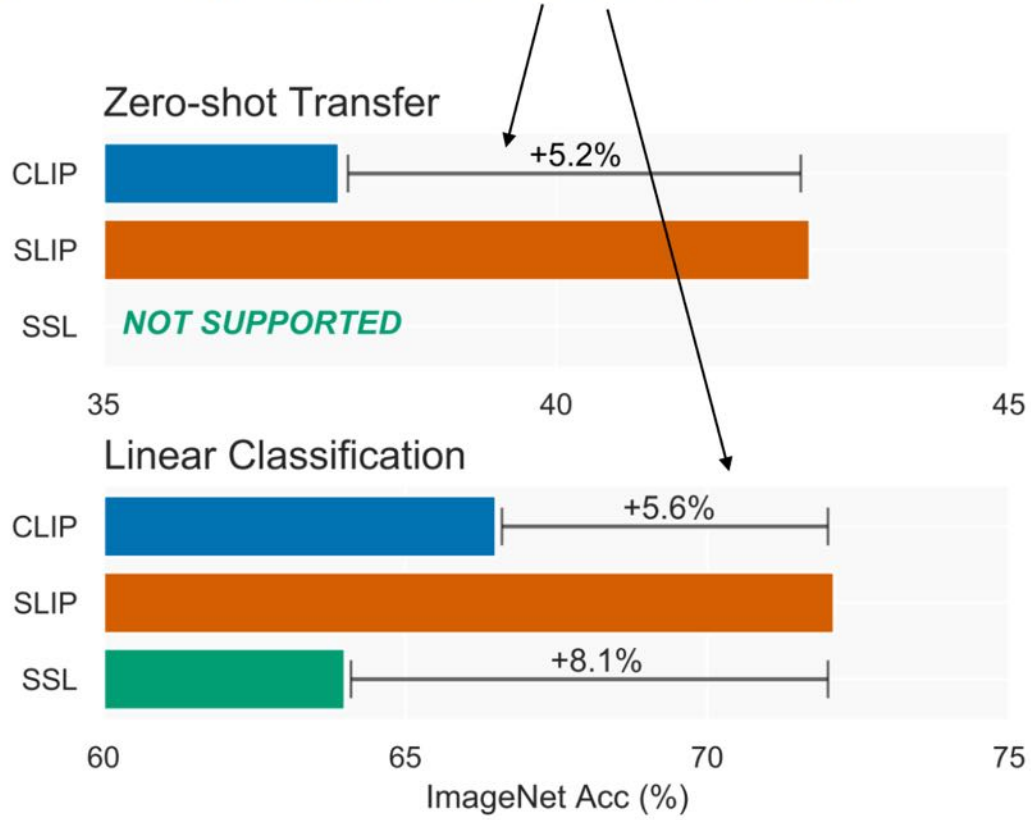
SLIP: Self-supervision meets Language-Image Pre-training: Mu et al.@arxiv'21

Learning Objectives

Simply combining contrastive language-image pretraining with self-supervised learning



SLIP outperforms CLIP on both zero-shot transfer and linear classification



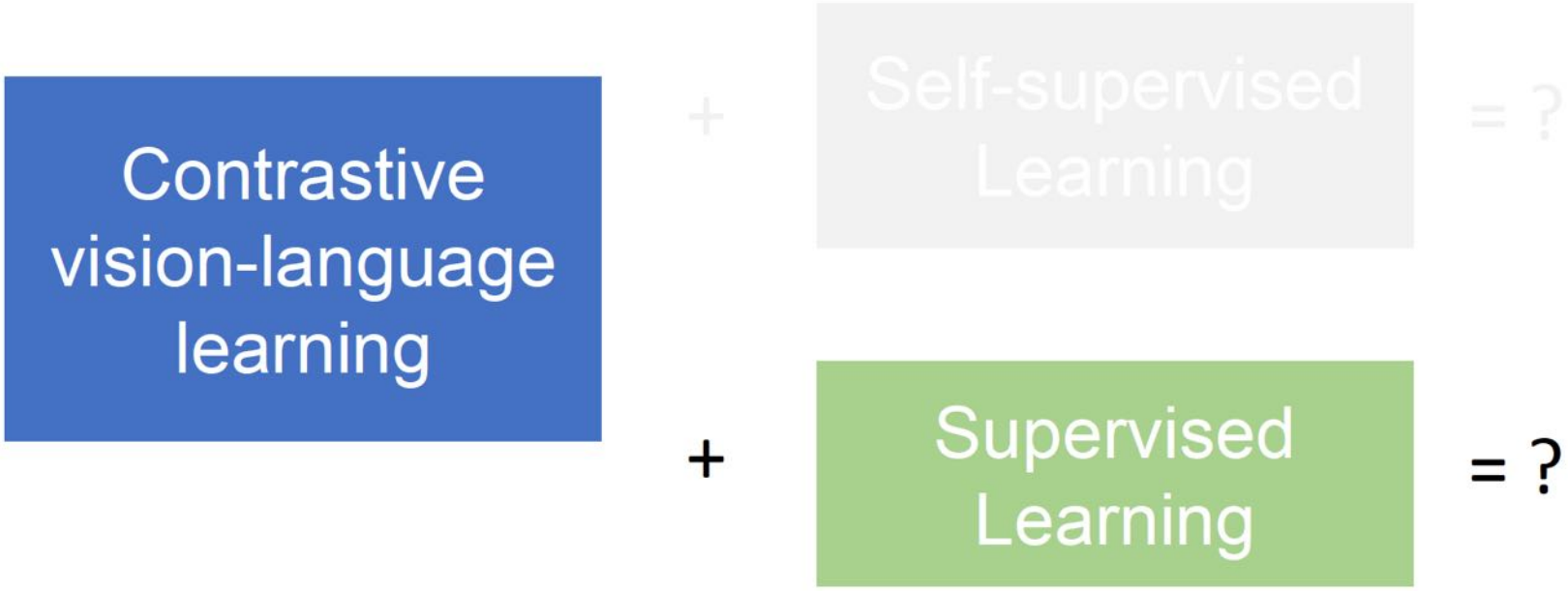
SLIP: Mu et al@

SLIP. Mu **et al.**

slide credit: Jianwei Yang

Learning Objectives

Combining contrastive learning with other learning objectives



slide credit: Jianwei Yang

Unified Contrastive Learning in Image Text-Label Space: Yang et al @CVPR'22

Image-Text-Label Space

Label

Image

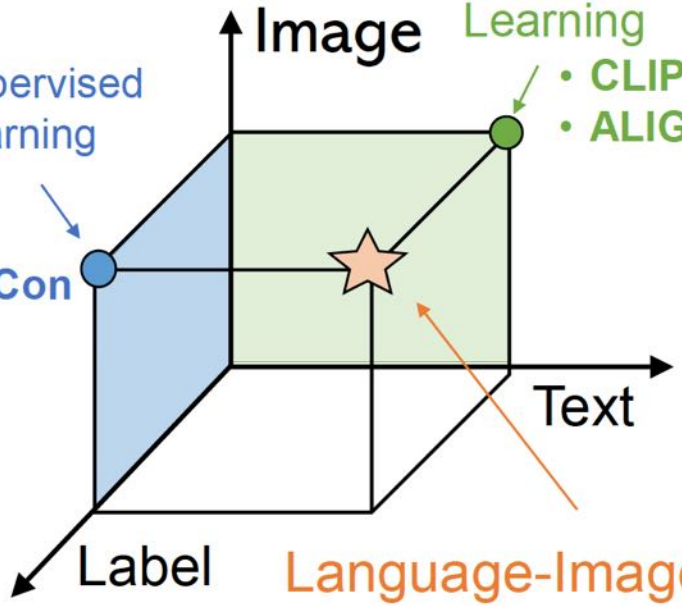
1 1 2 2

Learning a mapping function from image to label

Supervised Learning

- **CE**
- **SupCon**

Pros: Discriminative visual representations
Cons: Limited visual concepts



Language-Image Learning

- **CLIP**
- **ALIGN**


A white dog sitting on a couch


A dog looking bashfully to the side


A black cat is sleeping on a computer


Two parrots perching on a tree branch

Image

Text

Pros: Broad semantic coverage
Cons: Representations are not discriminative enough

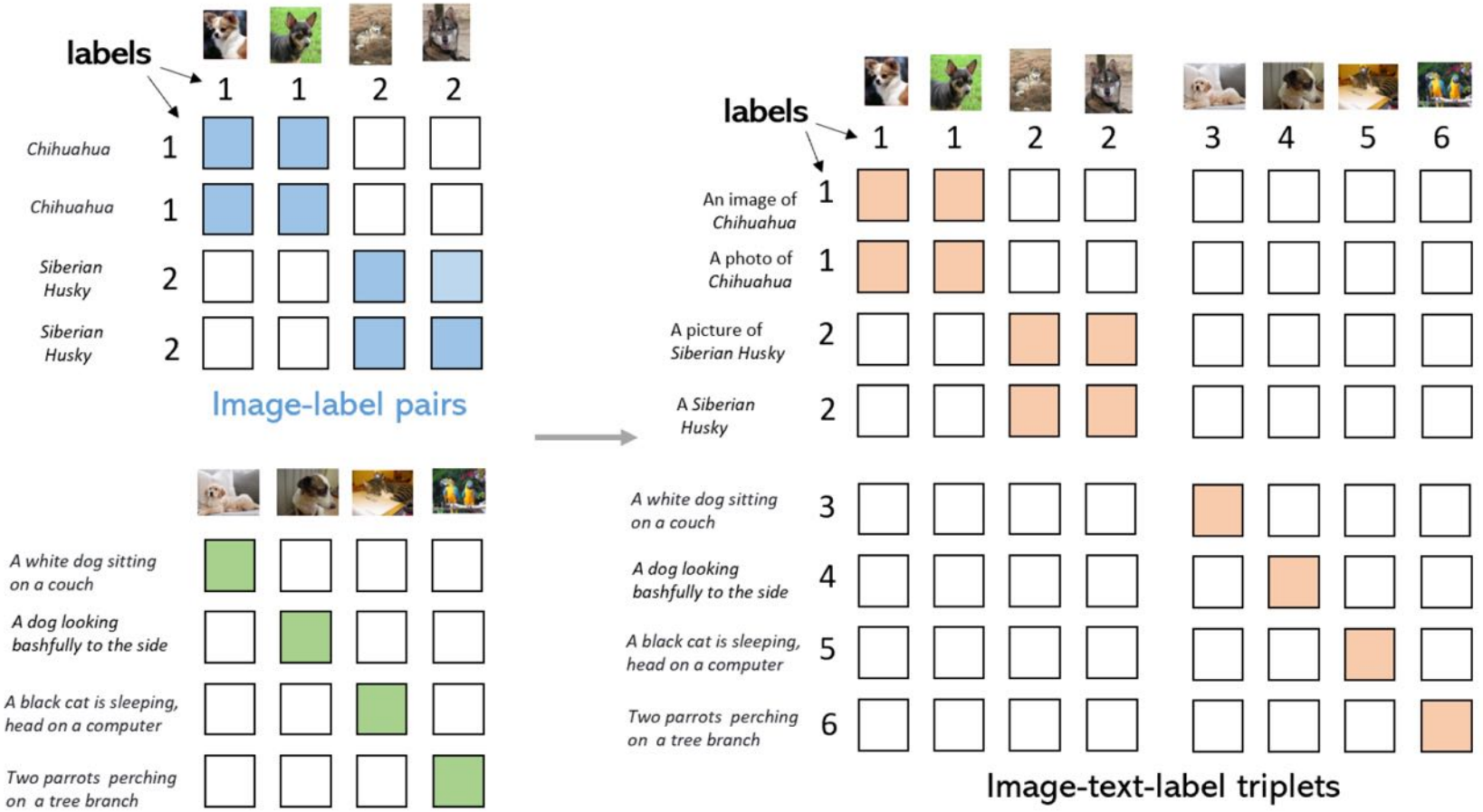
Language-Image-Label Learning

Best of both world:
Discriminative visual representations and
Broad semantic coverage

slide credit: Jianwei Yang

Unified Contrastive Learning in Image Text-Label Space: Yang et al @CVPR'22

Learning Objectives



UniCL. Yang *et al.*

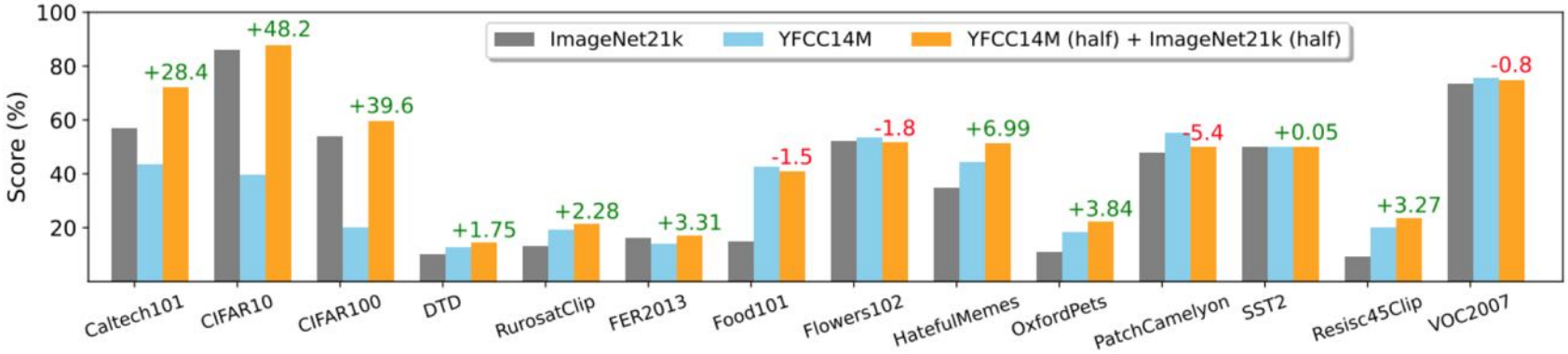
slide credit: Jianwei Yang

Unified Contrastive Learning in Image Text-Label Space: Yang et al @CVPR'22

Learning Objectives

How image-label data benefits image-text pairs on low-shot recognition?

Training Data	Method	Metric			
		Zero-shot		ImageNet-1K Finetuning	Linear Probing 18 datasets
		ImageNet-1K	14 datasets		
YFCC-14M	CLIP	30.1	36.3	77.5	72.7
ImageNet-21K	UniCL	28.5	37.8	78.8	80.5
YFCC-14M(half) + ImageNet-21K(half)	UniCL	36.4	45.5	79.0	80.0
YFCC-14M + ImageNet-21K	UniCL	**40.5**	**49.1**	**80.2**	**81.6**



Adding image-label data to image-text pairs can significantly improve the zero-shot, few-shot recognition