Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv´e Jegou, Julien Mairal, Piotr Bojanowski, Armand Joulin,

Facebook AI Research, Inria, Sorbonne University

Arxiv 2021

Dense Contrastive Learning for Self-Supervised Visual Pre-Training

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, Lei Li

The University of Adelaide, Tongji University, ByteDance AI Lab

CVPR 2021 Oral

Self-Supervised Representation Learning

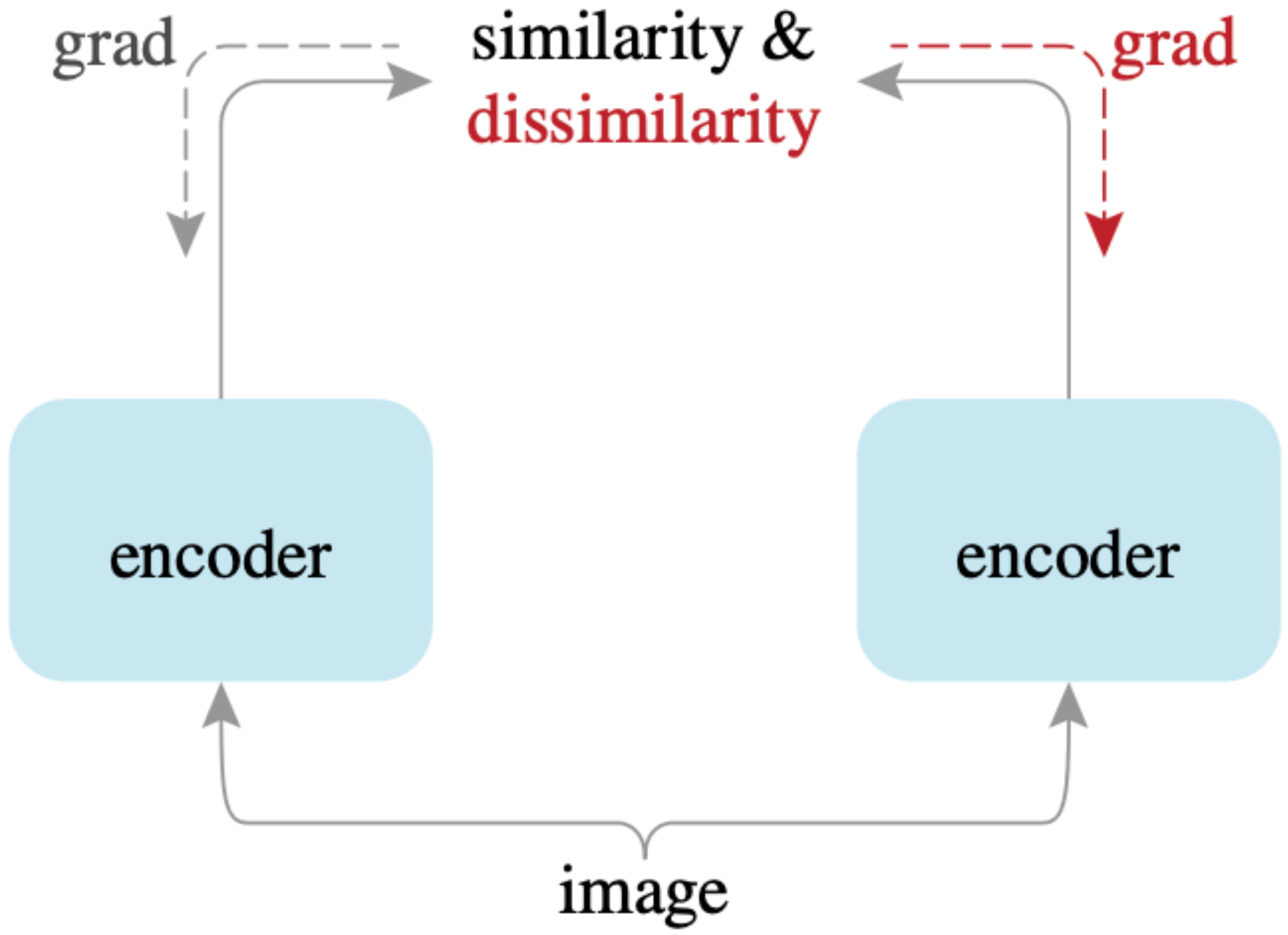
- Learn image features without human labels
- Map similar semantics closer
- Transferrable to downstream tasks

Common Idea:

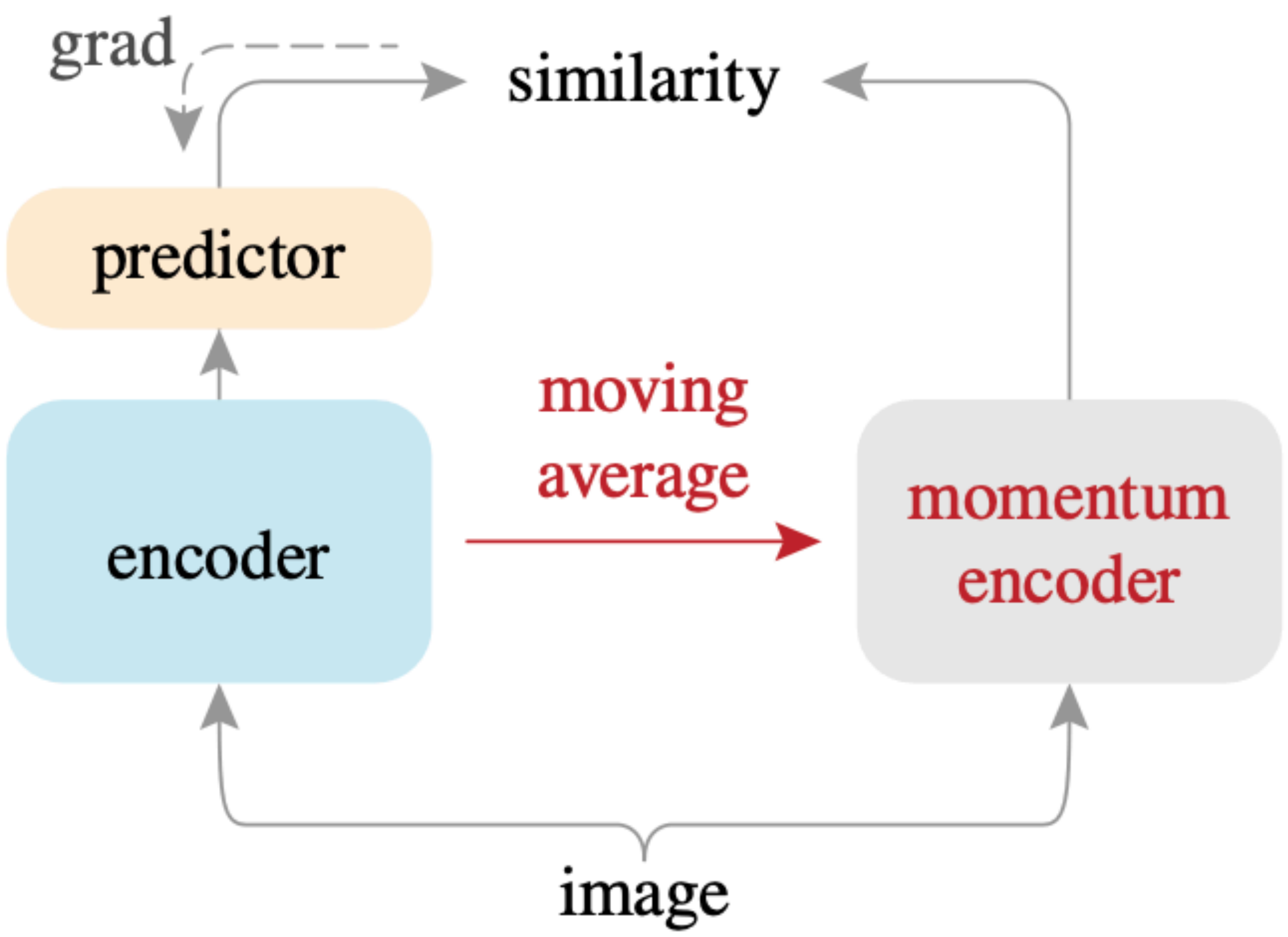
- Positive pair has similar features
- Negative pair has distinct features (Optional)

Previously on Self-Supervised Representation Learning

Positive pair
Negative pair
Symmetric training



SimCLR

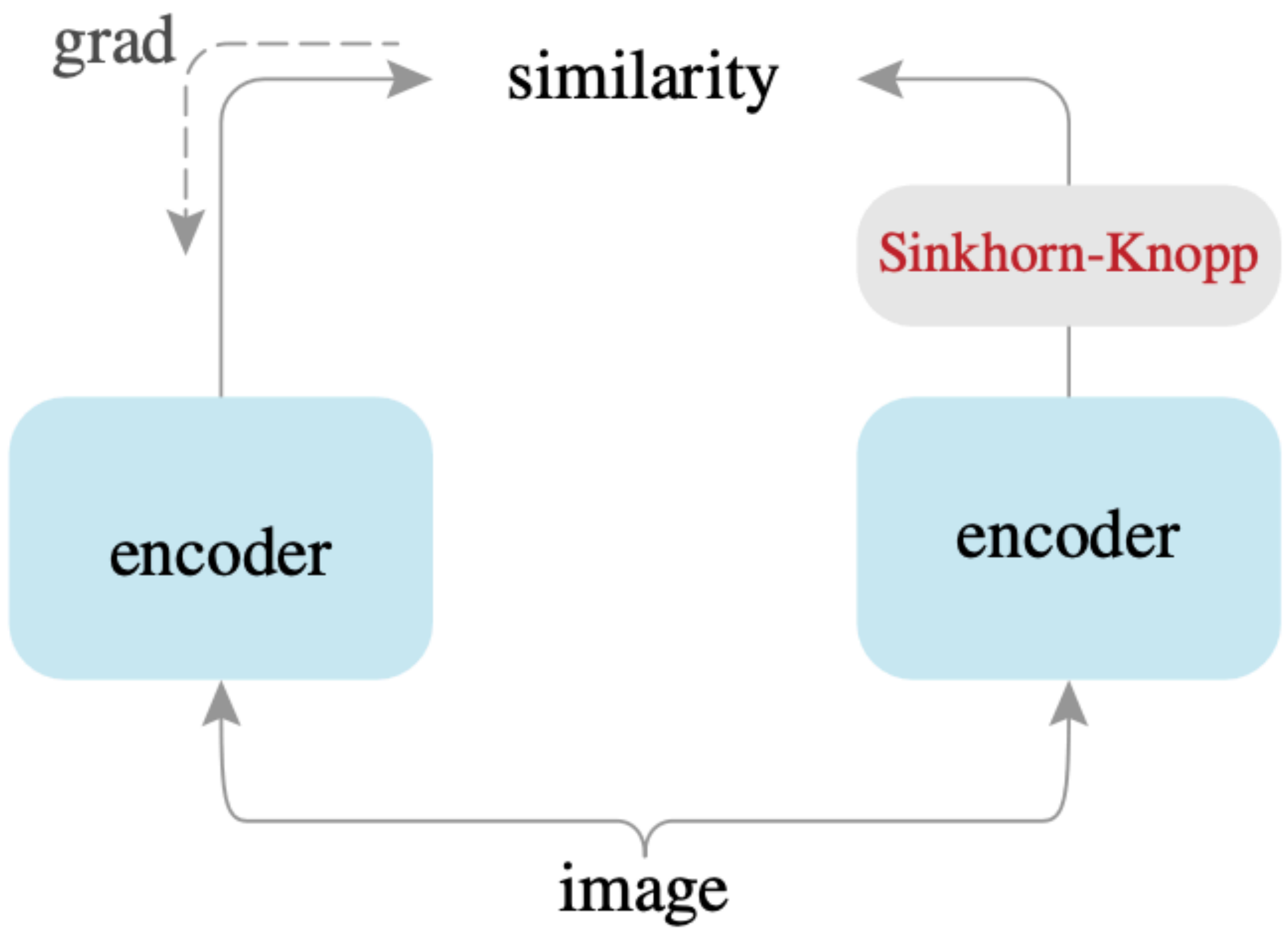


BYOL

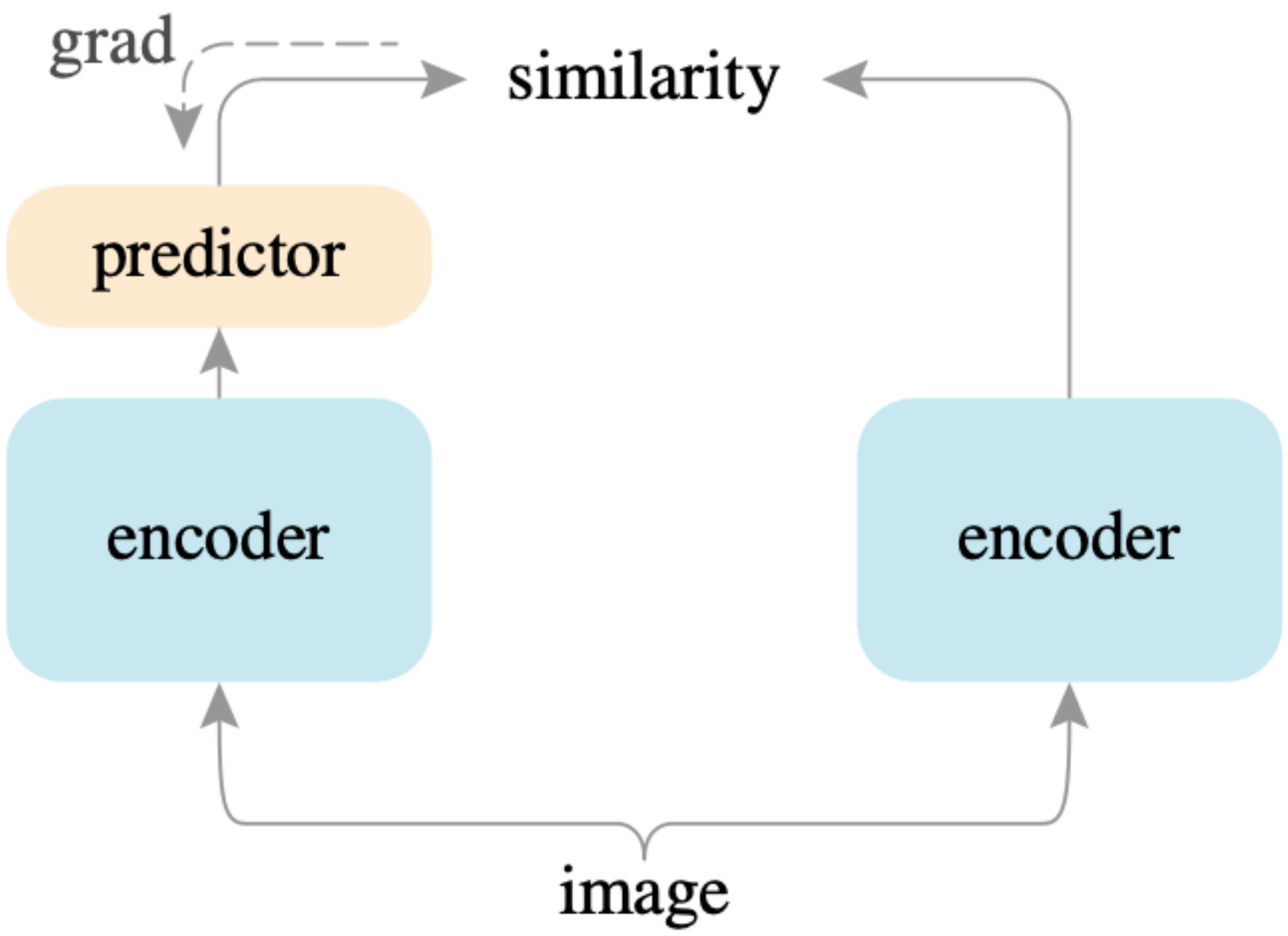
Positive pair
~~Negative pair~~
Mean teacher
Asymmetric training

MoCo: +Negative

Positive pair
~~Negative pair~~
Online Clustering
Asymmetric training



SwAV



SimSiam

Positive pair
~~Negative pair~~
Asymmetric training

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv´e Jegou, Julien Mairal, Piotr Bojanowski, Armand Joulin,

Facebook AI Research, Inria, Sorbonne University

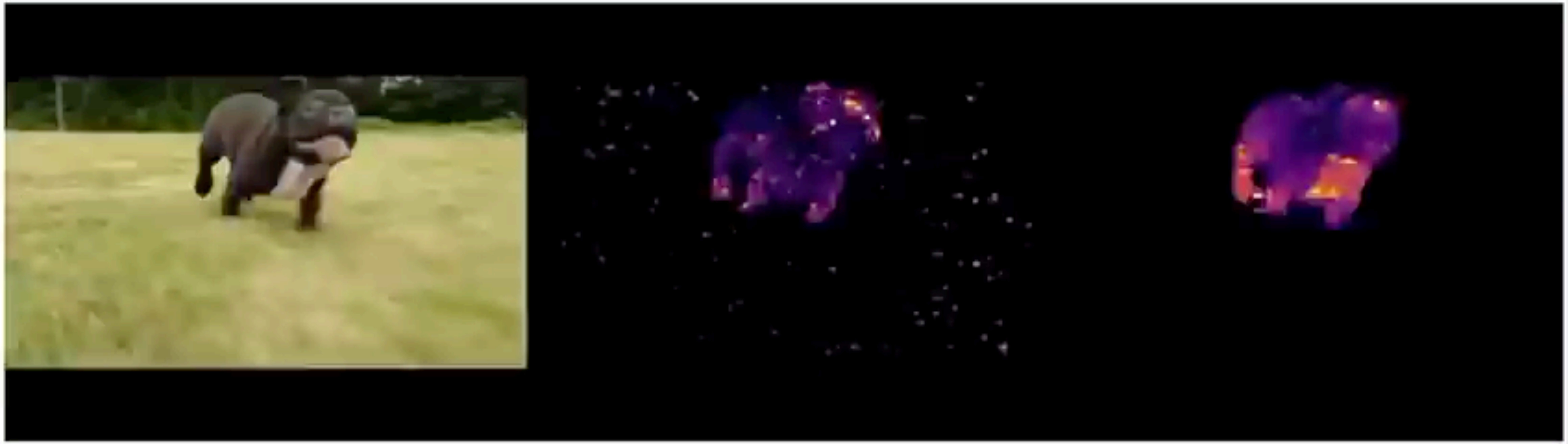
Arxiv 2021

Dense Contrastive Learning for Self-Supervised Visual Pre-Training

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, Lei Li

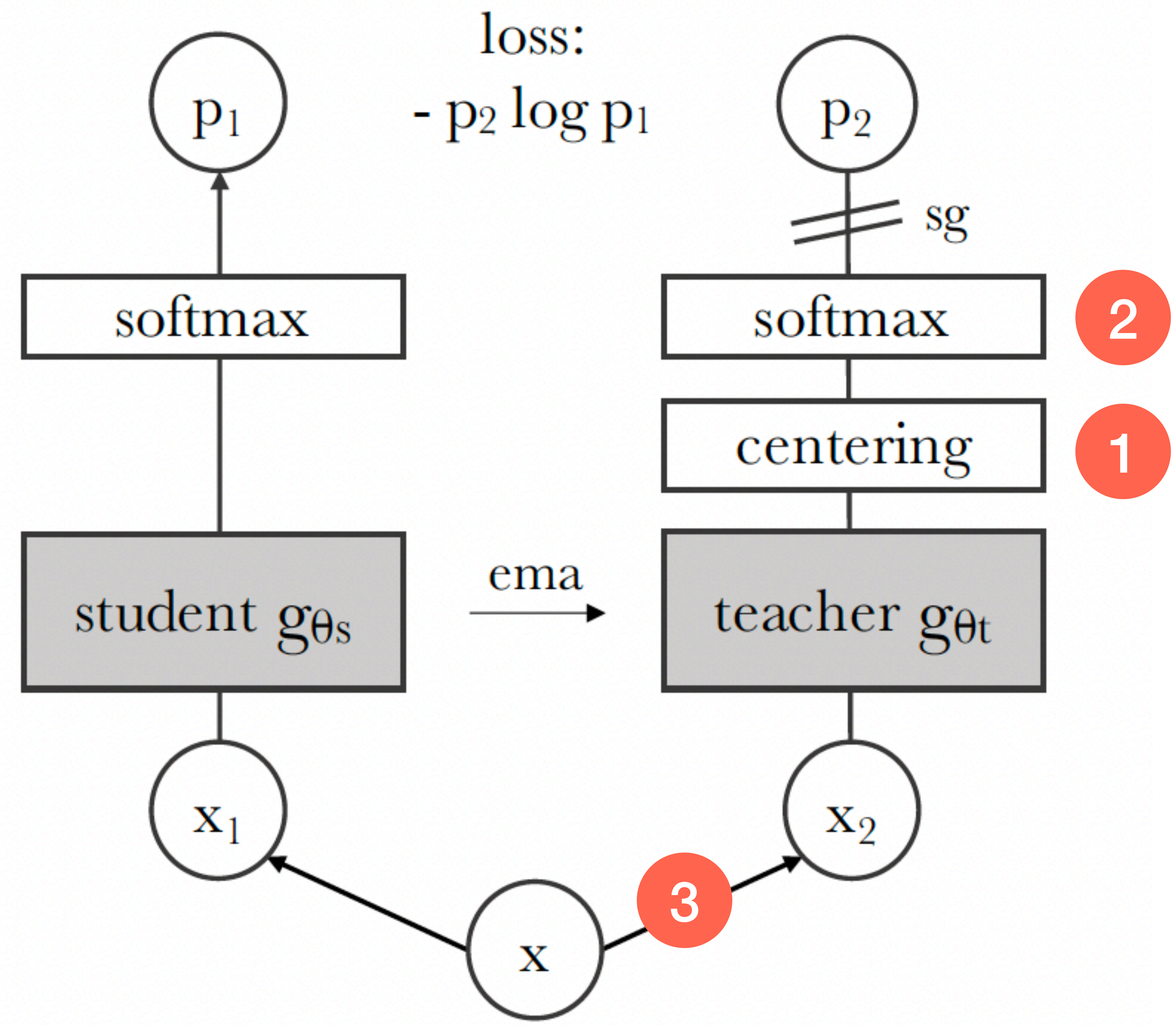
The University of Adelaide, Tongji University, ByteDance AI Lab

CVPR 2021 Oral



Positive pair
~~Negative pair~~
Mean teacher
Asymmetric training

Centering
Softmax + Different temperatures
Augmentation: Local / Global Views

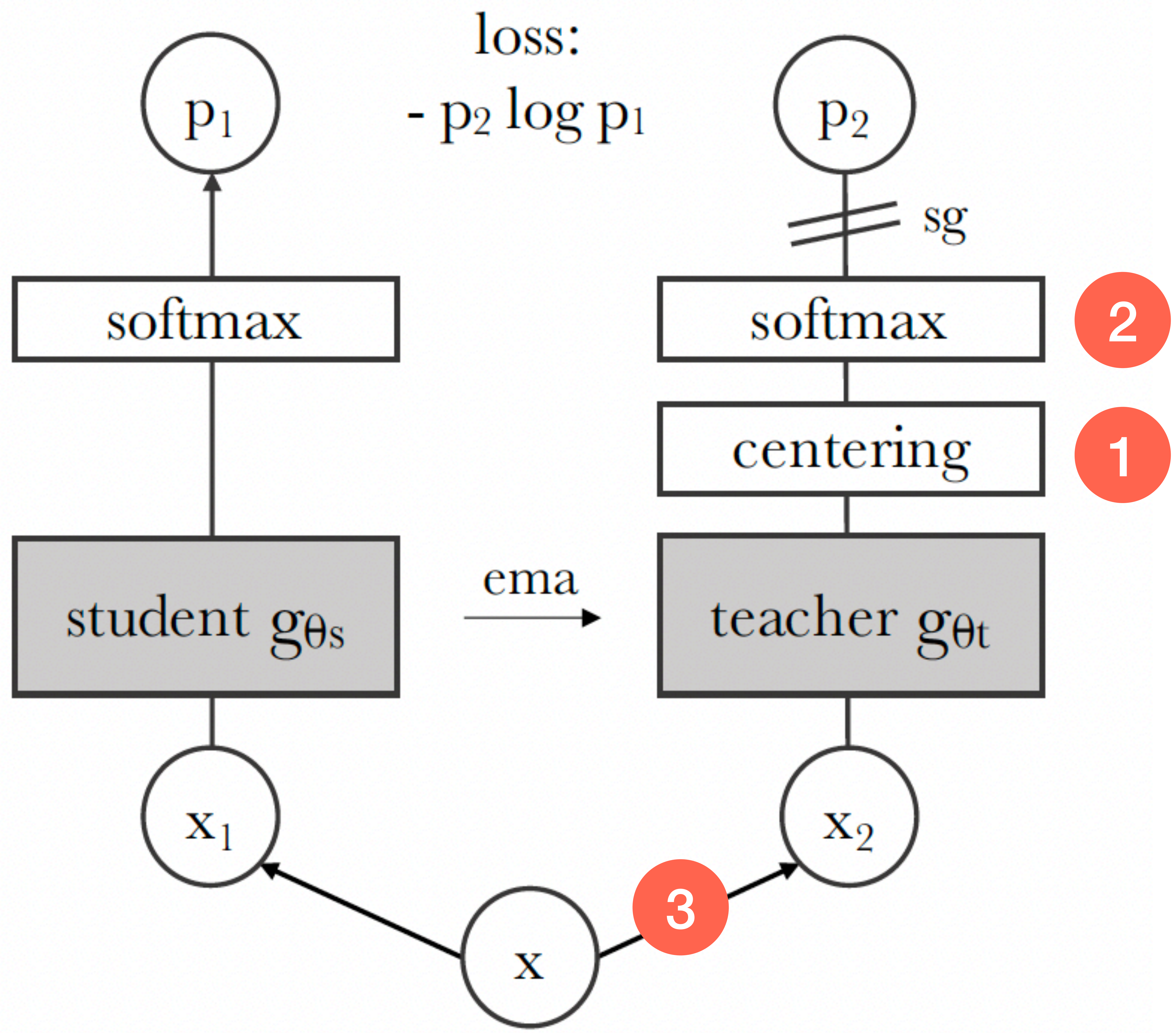


Interpreted as self-distillation

Positive pair
~~Negative pair~~
Mean teacher
Asymmetric training

Centering
Softmax + Different temperatures
Augmentation: Local / Global Views

1 2 Prevent Collapse



Interpreted as self-distillation

Positive pair

~~Negative pair~~

Mean teacher

Asymmetric training

Centering

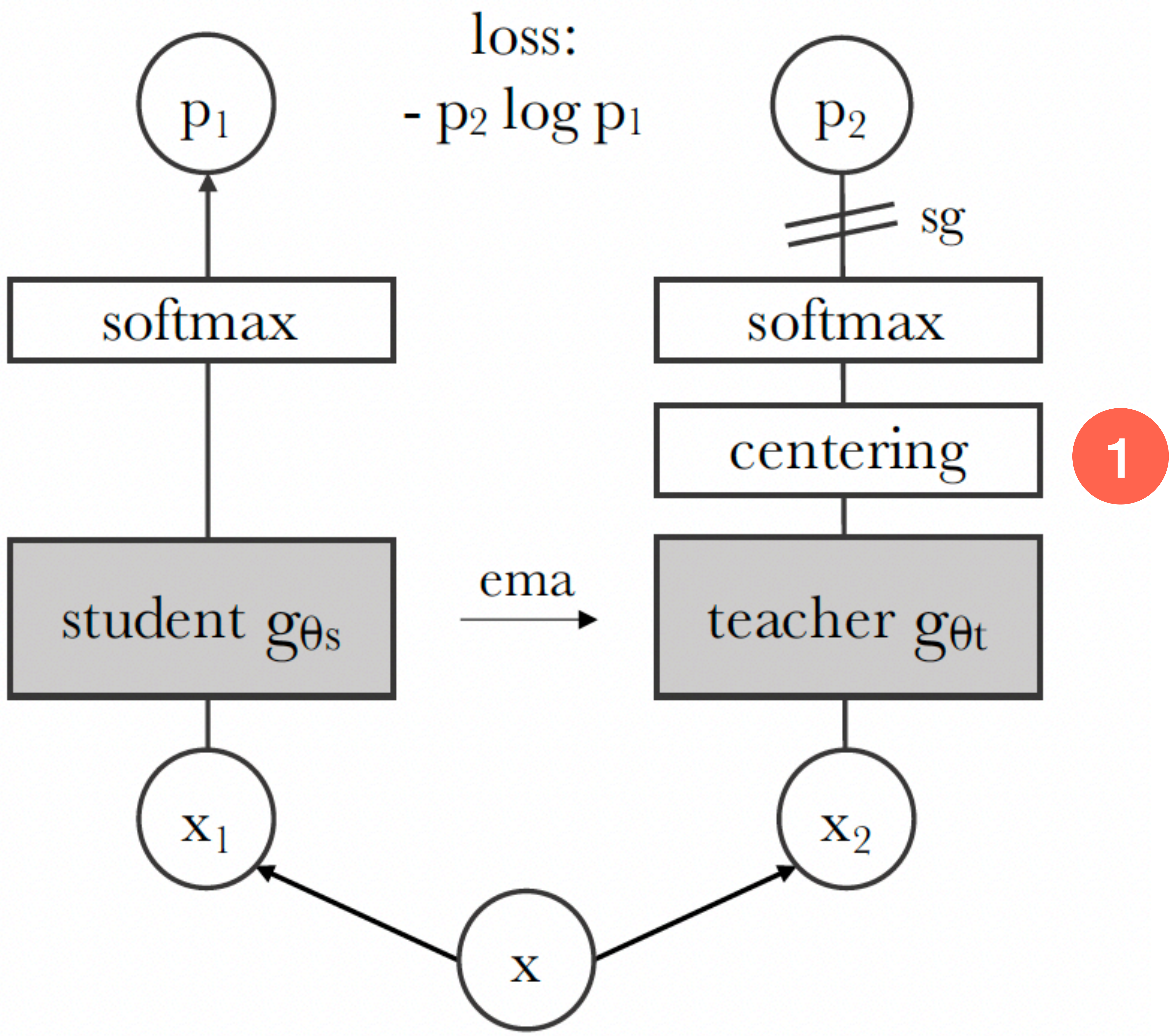
Softmax + Different temperatures

Augmentation: Local / Global Views

EMA center

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i),$$

Subtract the center from teacher features



(Moving first order BN)

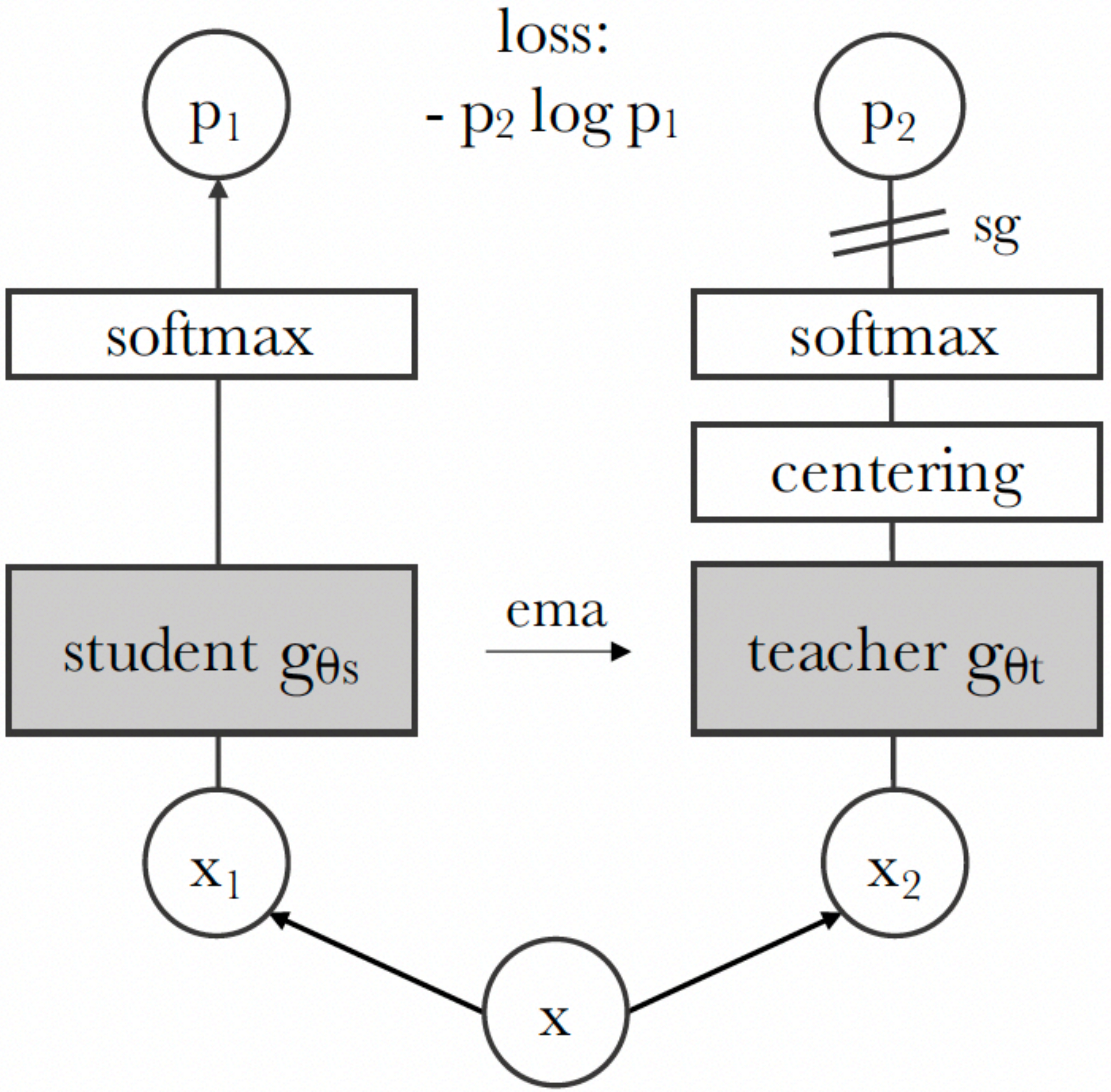
Positive pair
~~Negative pair~~
Mean teacher
Asymmetric training

Centering
Softmax + Different temperatures
Augmentation: Local / Global Views

Softmax: Fake classification

Different temperatures

Student Teacher



2

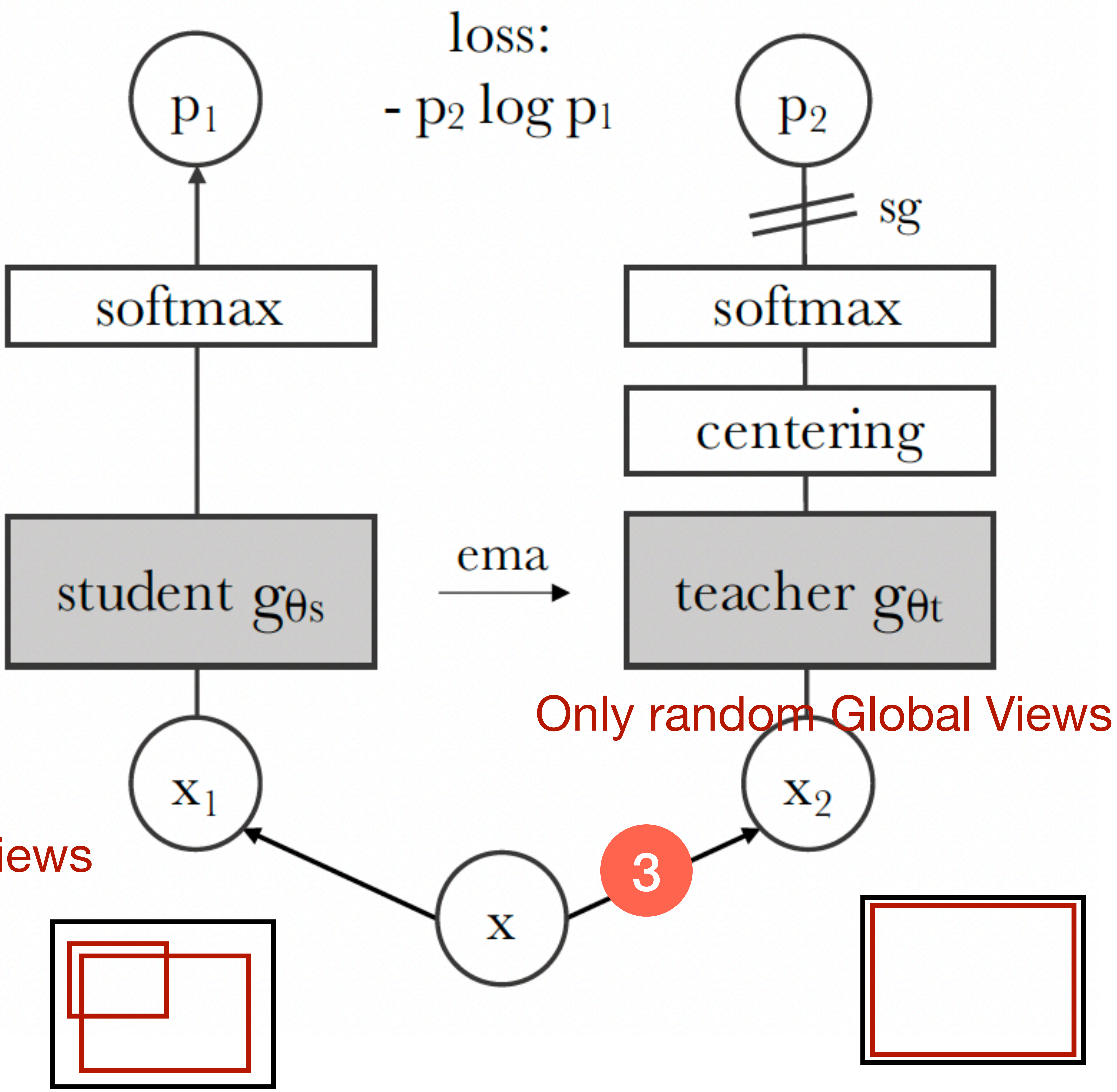
Making the student to be more confident gradually

Positive pair
~~Negative pair~~
Mean teacher
Asymmetric training

Centering
Softmax + Different temperatures
Augmentation: Local / Global Views

Only random Global Views

Random Local or Global Views



ImageNet is object-centric

Guess the object with partial view

Positive pair

~~Negative pair~~

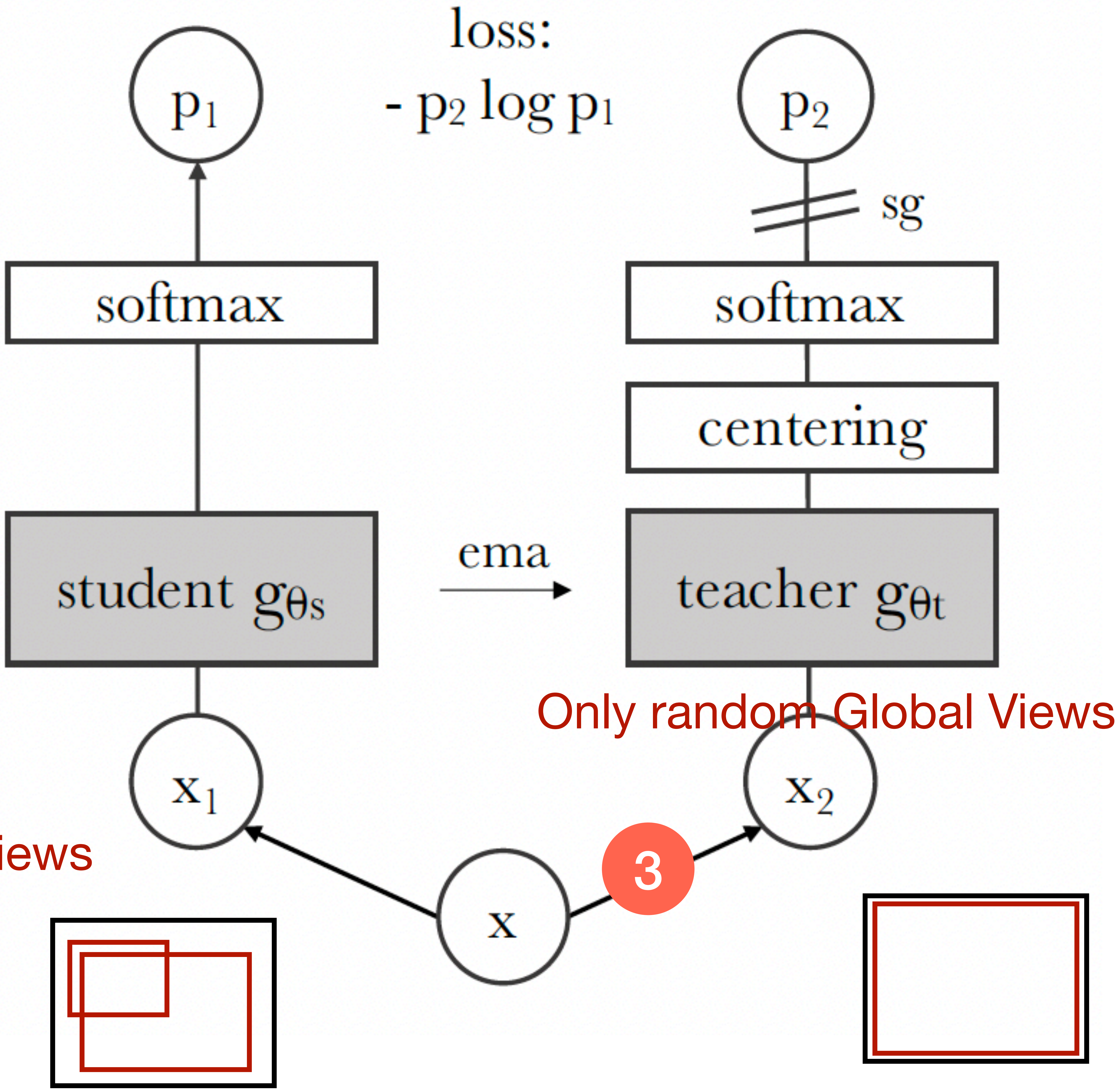
Mean teacher

Asymmetric training

Centering

Softmax + Different temperatures

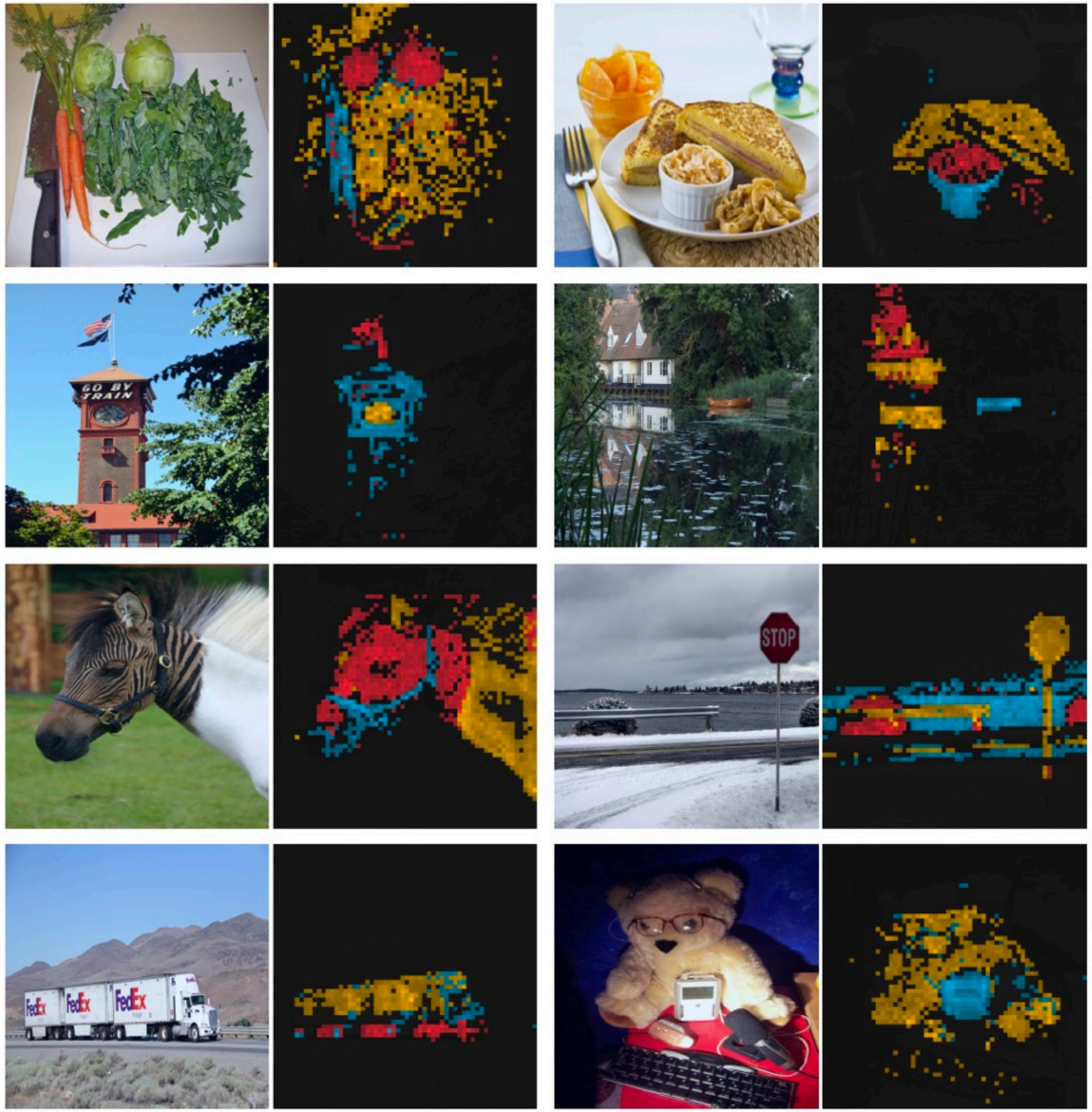
Augmentation: Local / Global Views



Linear / k-NN probe on ImageNet

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**
Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

Especially good for transformers



Attention maps from multiple heads

ViT [CLS]

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv´e Jegou, Julien Mairal, Piotr Bojanowski, Armand Joulin,

Facebook AI Research, Inria, Sorbonne University

Arxiv 2021

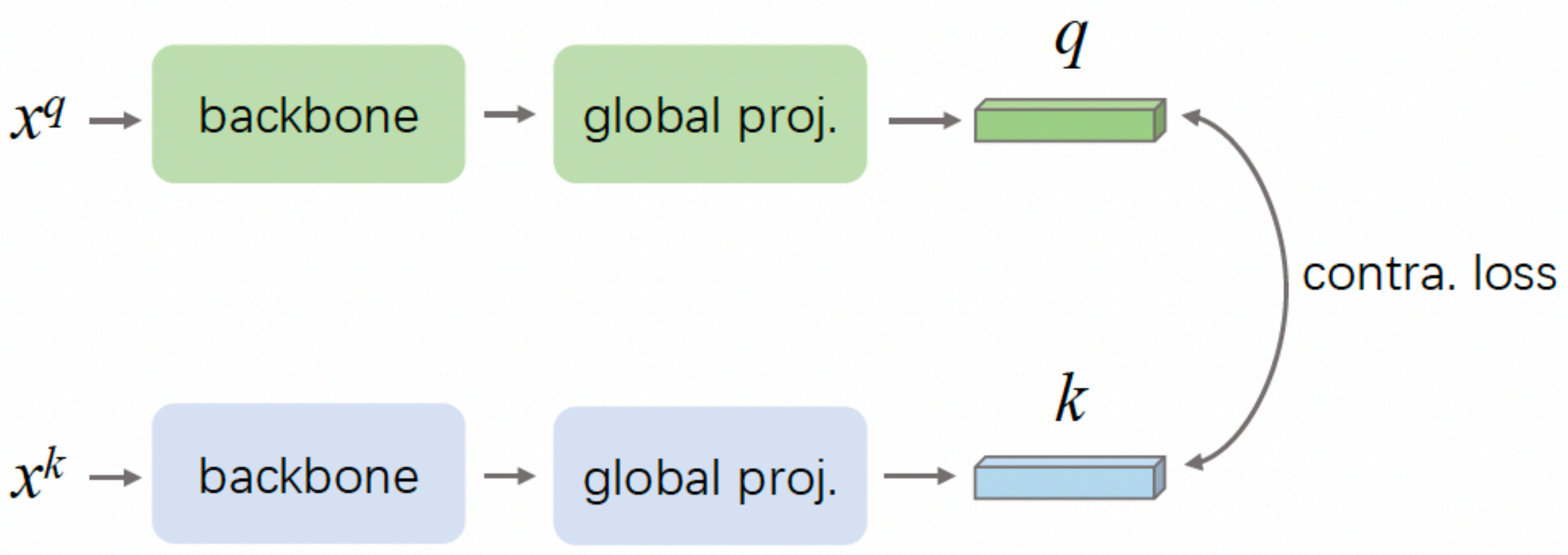
Dense Contrastive Learning for Self-Supervised Visual Pre-Training

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, Lei Li

The University of Adelaide, Tongji University, ByteDance AI Lab

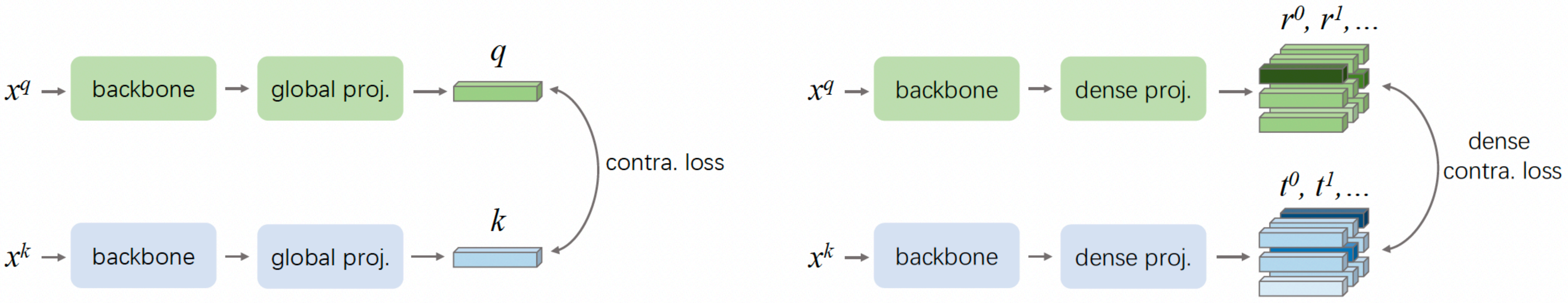
CVPR 2021 Oral

Self-Supervised Pre-Training for Dense Prediction Downstream Tasks



(a) Global Contrastive Learning

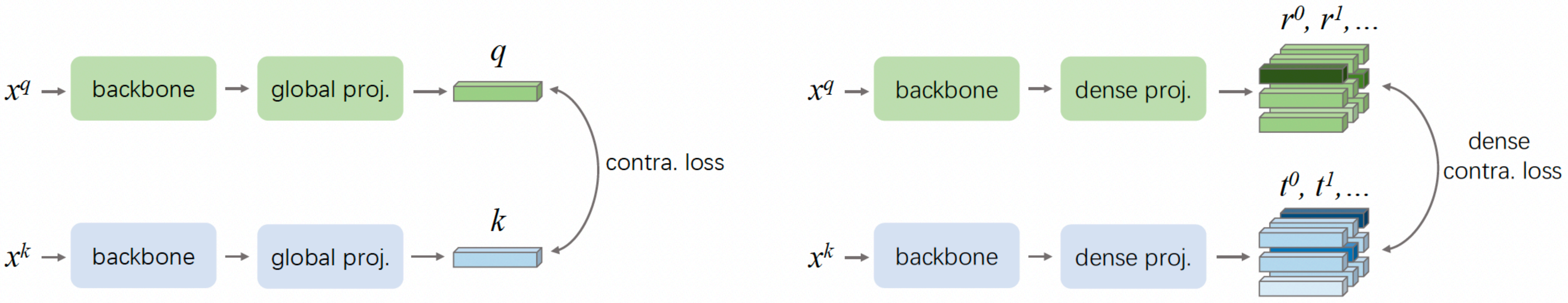
Self-Supervised Pre-Training for Dense Prediction Downstream Tasks



(a) Global Contrastive Learning

(b) Dense Contrastive Learning

Self-Supervised Pre-Training for Dense Prediction Downstream Tasks



(a) Global Contrastive Learning

(b) Dense Contrastive Learning

No pooling

Positive pair:

Feature vector at a location in a image

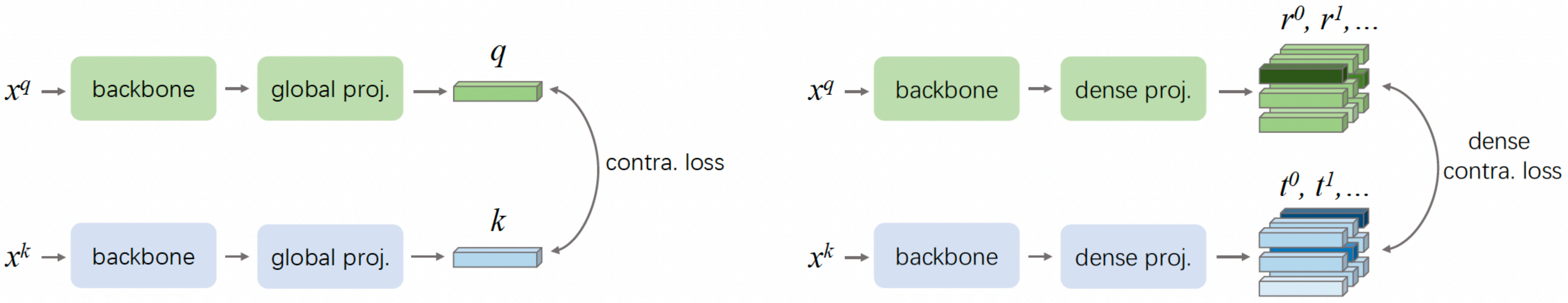
Feature vector at corresponding location in the augmented view of the same image

Negative pair:

Feature vector at a location in a image

Average of feature vectors of all locations in a different image

Self-Supervised Pre-Training for Dense Prediction Downstream Tasks



(a) Global Contrastive Learning

(b) Dense Contrastive Learning

No pooling

Positive pair:

Feature vector at a location in a image

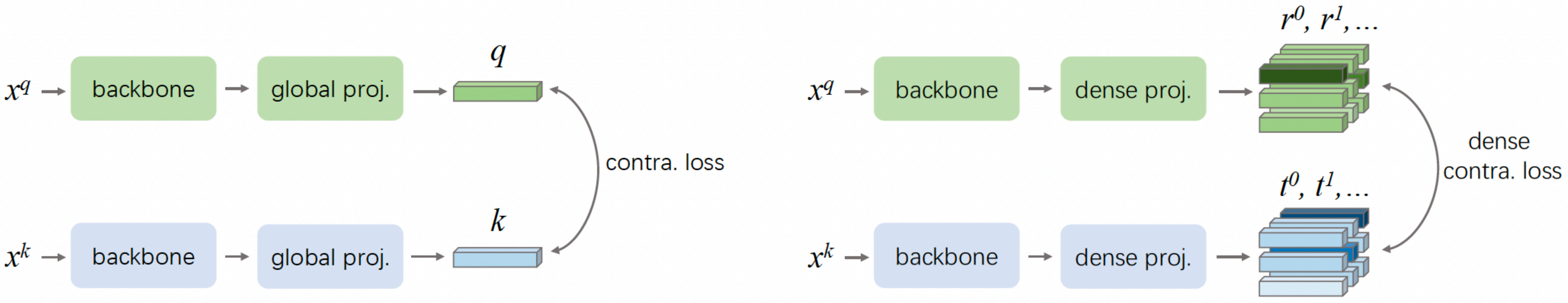
Feature vector at corresponding location in the augmented view of the same image

Negative pair:

Feature vector at a location in a image

Average of feature vectors of all locations in a different image

Self-Supervised Pre-Training for Dense Prediction Downstream Tasks



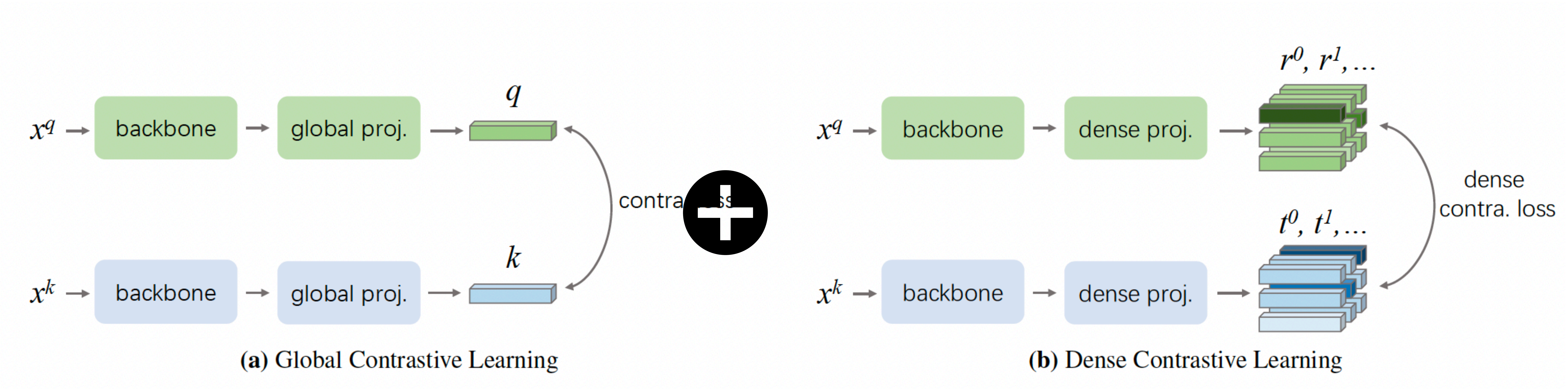
(a) Global Contrastive Learning **(b)** Dense Contrastive Learning

Correspondence:

Matching of feature maps

Computed from geometrics

Self-Supervised Pre-Training for Dense Prediction Downstream Tasks



Correspondence:

Matching of feature maps

Computed from geometrics

pre-train	AP	AP_{50}	AP_{75}
random init.	32.8	59.0	31.6
super. IN	54.2	81.6	59.8
MoCo-v2 CC	54.7	81.0	60.6
DenseCL CC	56.7	81.7	63.0
SimCLR IN [2]	51.5	79.4	55.6
BYOL IN [14]	51.9	81.0	56.5
MoCo IN [17]	55.9	81.5	62.6
MoCo-v2 IN [3]	57.0	82.4	63.6
MoCo-v2 IN*	57.0	82.2	63.4
DenseCL IN	58.7	82.8	65.2

Table 1 – Object detection fine-tuned on PASCAL VOC. 'CC' and 'IN' indicate the pre-training models trained on COCO and ImageNet respectively. The models pre-trained on

pre-train	mIoU
random init.	40.7
super. IN	67.7
MoCo-v2 CC	64.5
DenseCL CC	67.5
SimCLR IN	64.3
BYOL IN	63.3
MoCo-v2 IN	67.5
DenseCL IN	69.4

(a) PASCAL VOC

pre-train	mIoU
random init.	63.5
super. IN	73.7
MoCo-v2 CC	73.8
DenseCL CC	75.6
SimCLR IN	73.1
BYOL IN	71.6
MoCo-v2 IN	74.5
DenseCL IN	75.7

(b) Cityscapes

Table 4 – Semantic segmentation on PASCAL VOC and Cityscapes. 'CC' and 'IN' indicate the pre-training models trained on COCO and ImageNet respectively. The metric is the

strategy	Detection AP	Detection AP_{50}	Detection AP_{75}	Classification mAP
random	56.0	81.3	62.0	81.7
max-sim $\mathbf{\Theta}$	56.0	81.5	62.1	81.8
max-sim $\mathbf{F}$	56.7	81.7	63.0	82.9

Table 6 – Ablation study of matching strategy. To extract the dense correspondence according to the backbone features $\mathbf{F}_1$ and $\mathbf{F}_2$ shows the best results.

Conclusion

- DINO: Self-supervised learning + ViT
- DenseCL: Self-supervised learning for dense prediction

Comments

- Datasets and augmentations matter
- Downstream tasks matter