**Equations – Self-supervised learning**

**Self-supervised learning I**

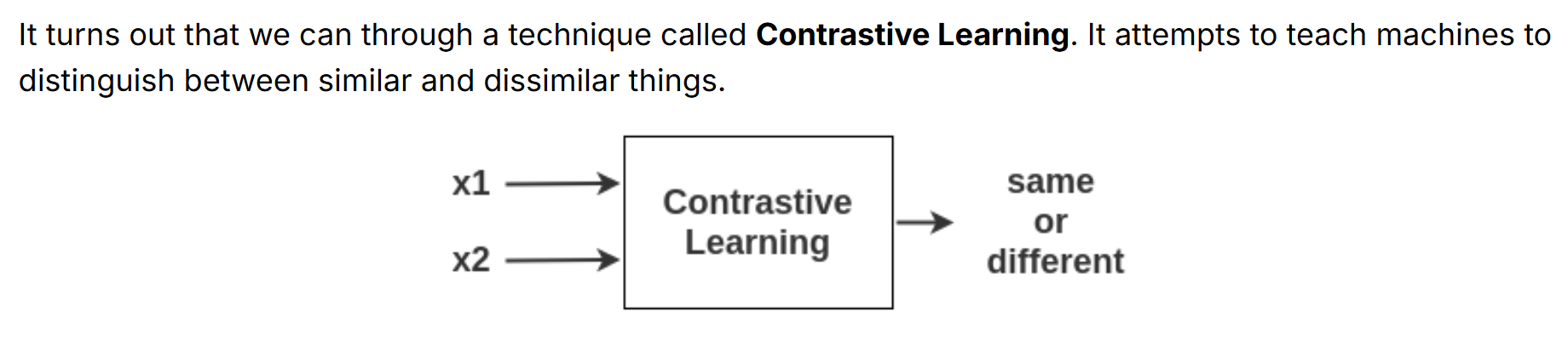
**Context-Based Methods**

* Image rotations
* Image colorization
* Jigsaw Puzzles
* Context Prediction

**Contrastive learning**

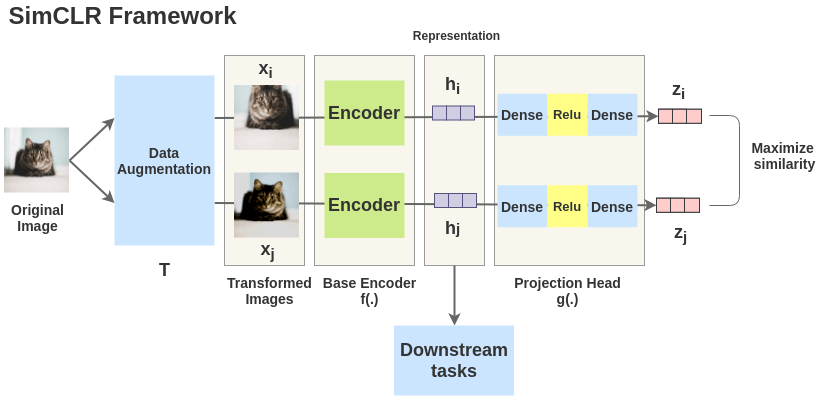
Definitions:

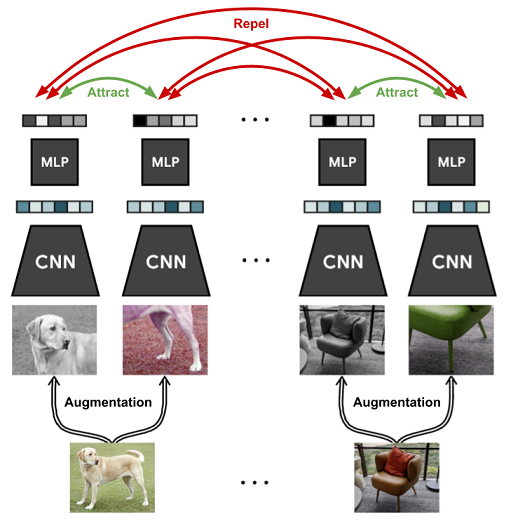
* “[…] trains models by maximizing the similarity between related data points while minimizing the similarity between unrelated ones.”
* “[…] to learn high-quality representations by ensuring that similar samples have closer representations while dissimilar samples are pushed apart in the feature space.”

****

**SimCLR (2020)**

* A simple framework for contrastive learning of visual representations.





Normalized Temperature-scaled Cross-Entropy (NT-Xent)

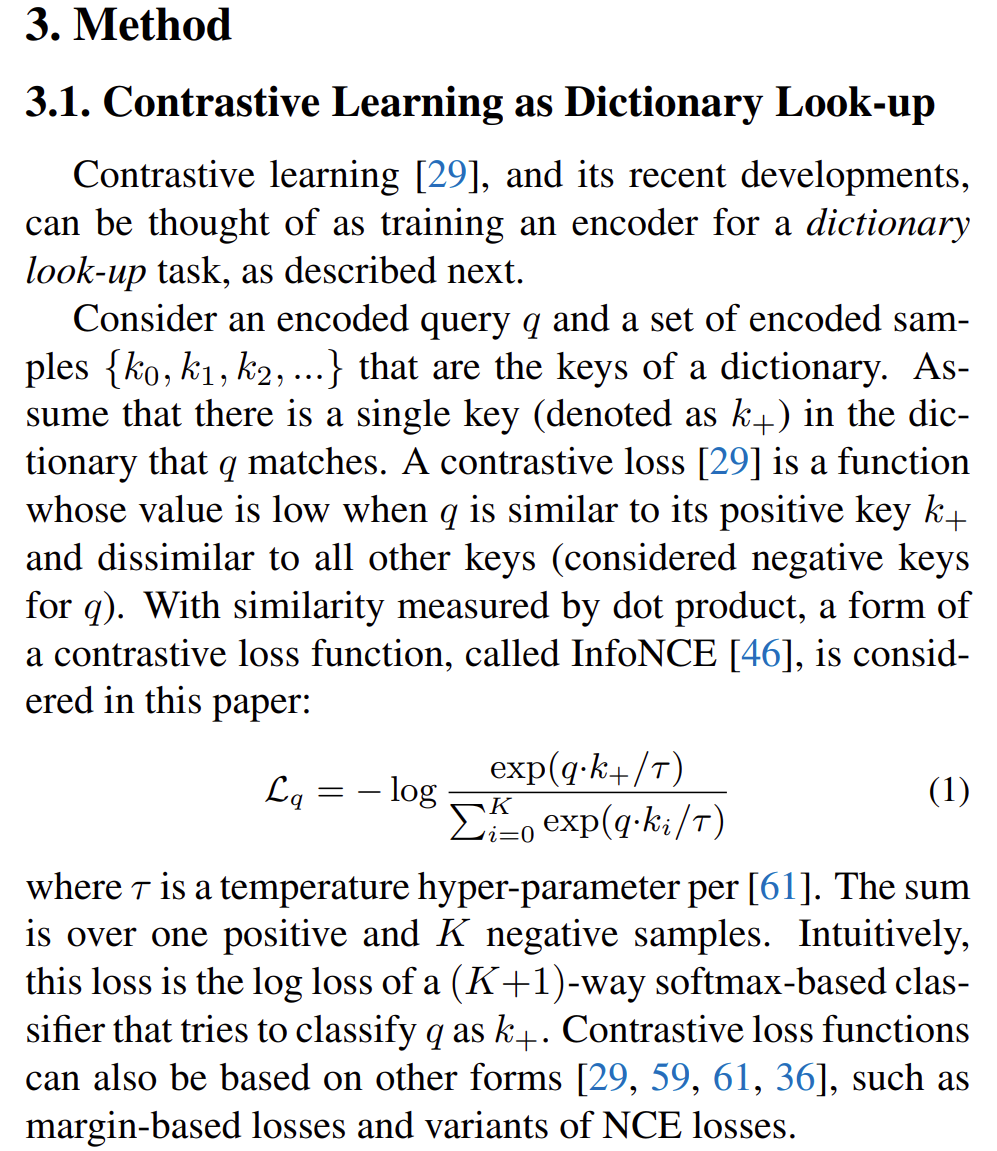
**MoCo (2020)**

Contrastive loss:

* “[…] function whose value is low when is similar to its positive key and dissimilar to all the other keys.”

Momentum update

matches



**CPC (2018)**

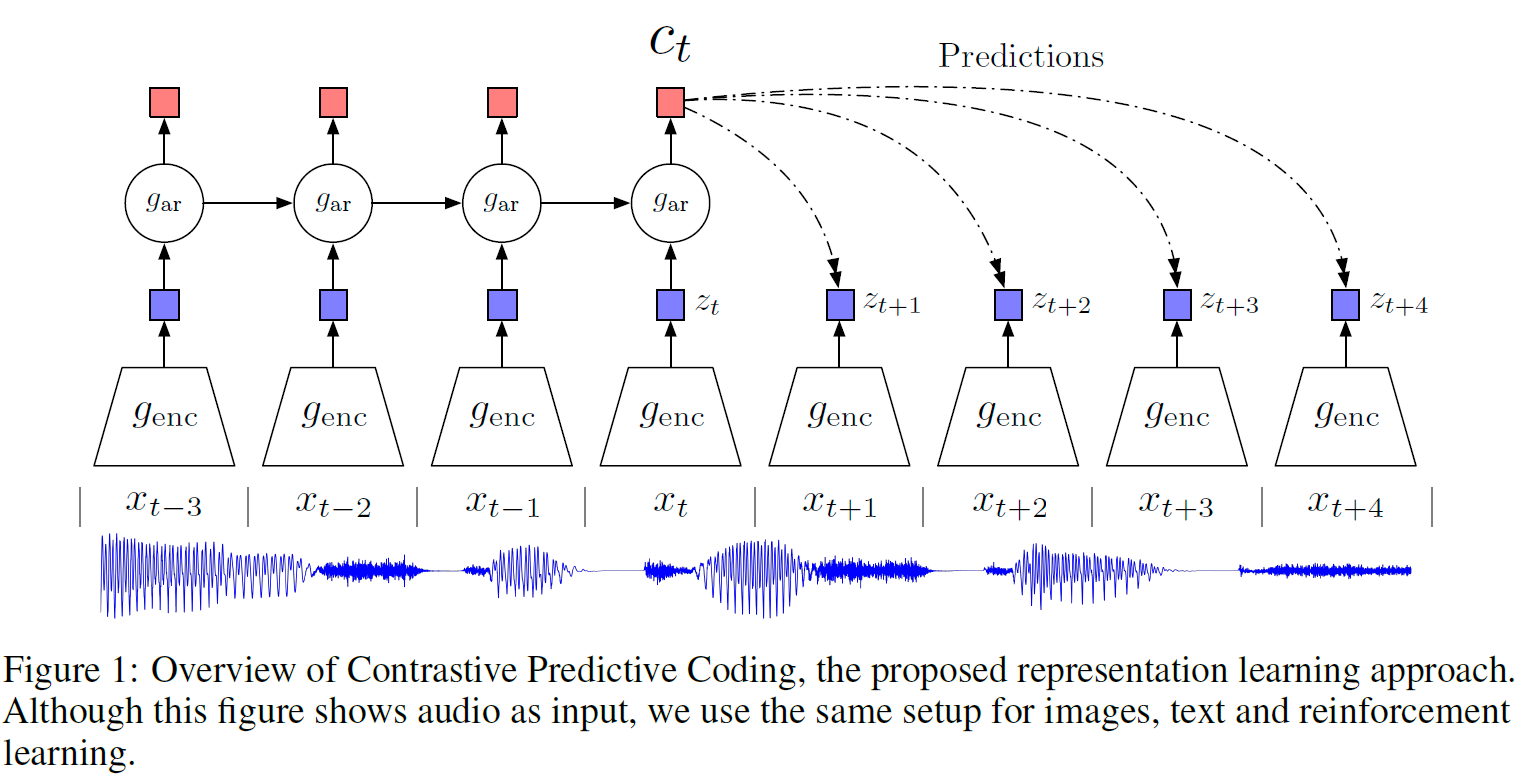
Representation Learning with Contrastive Predictive Coding

Learns useful representations from high-dimensional data. Learns by predicting the future in latent space by using autoregressive models.

Shared information

Mutual information:

Maximize mutual information.

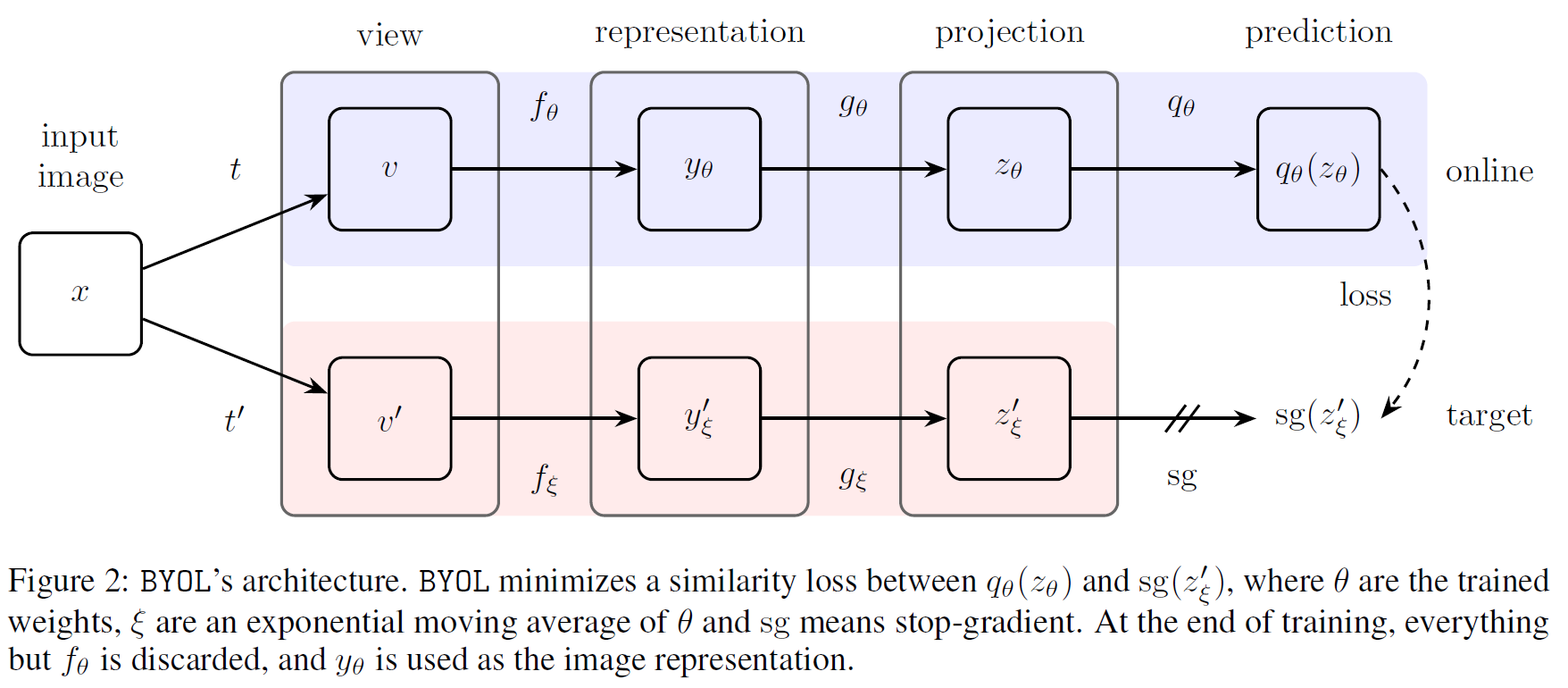


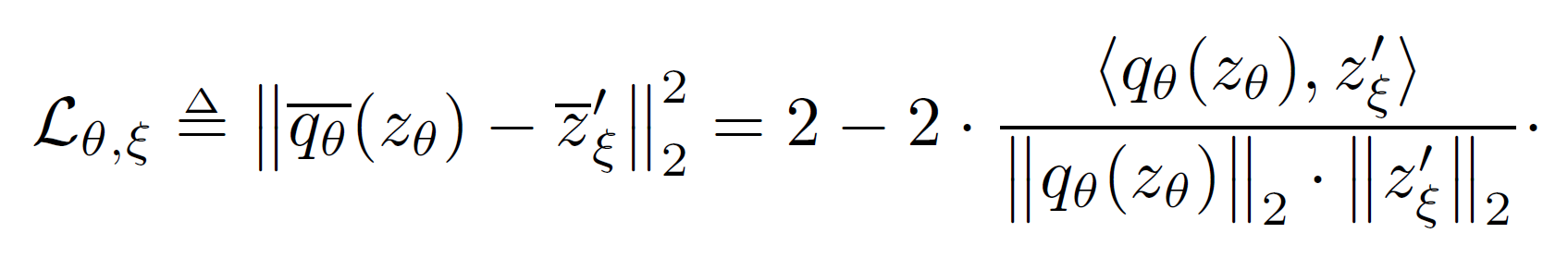
**BYOL (2020)**

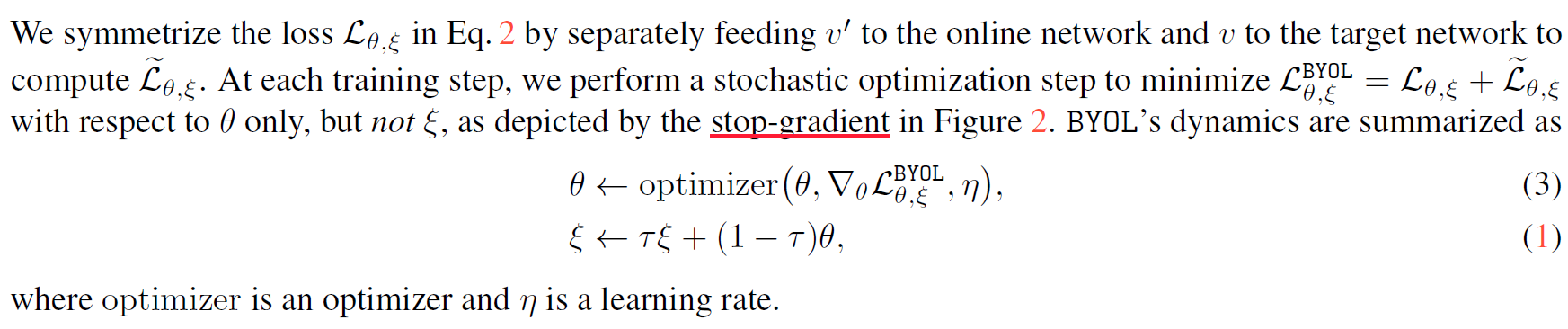
After training, BYOL produces good representations for downstream tasks.

Architecture

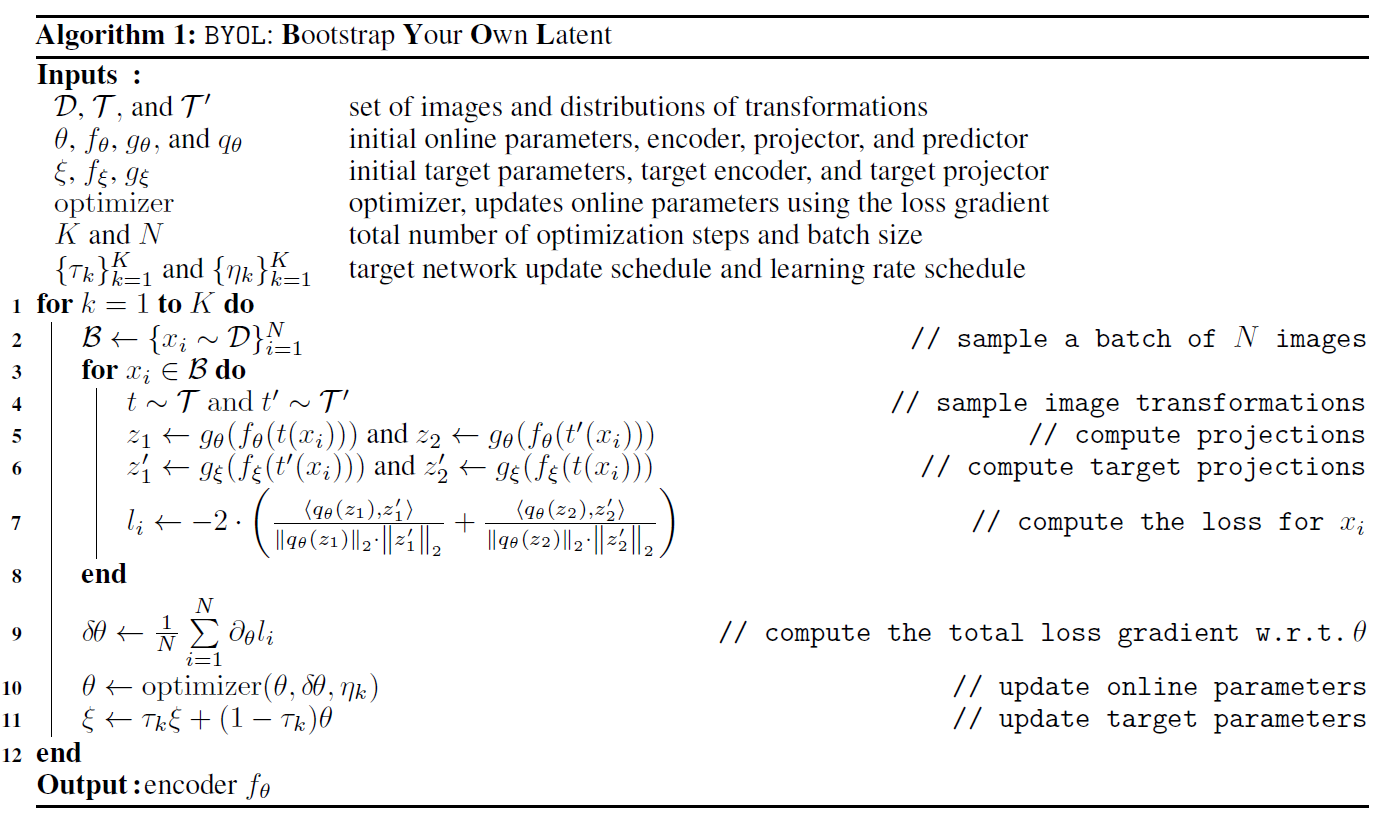
* Set of images Image sampled uniformly from .
* distributions of image augmentations. image augmentations.
* augmented views.
* **Online network** outputs representation and a projection
* **Target network** outputs representation and the target projection .
* **Online network** outputs a prediction of and normalize both.







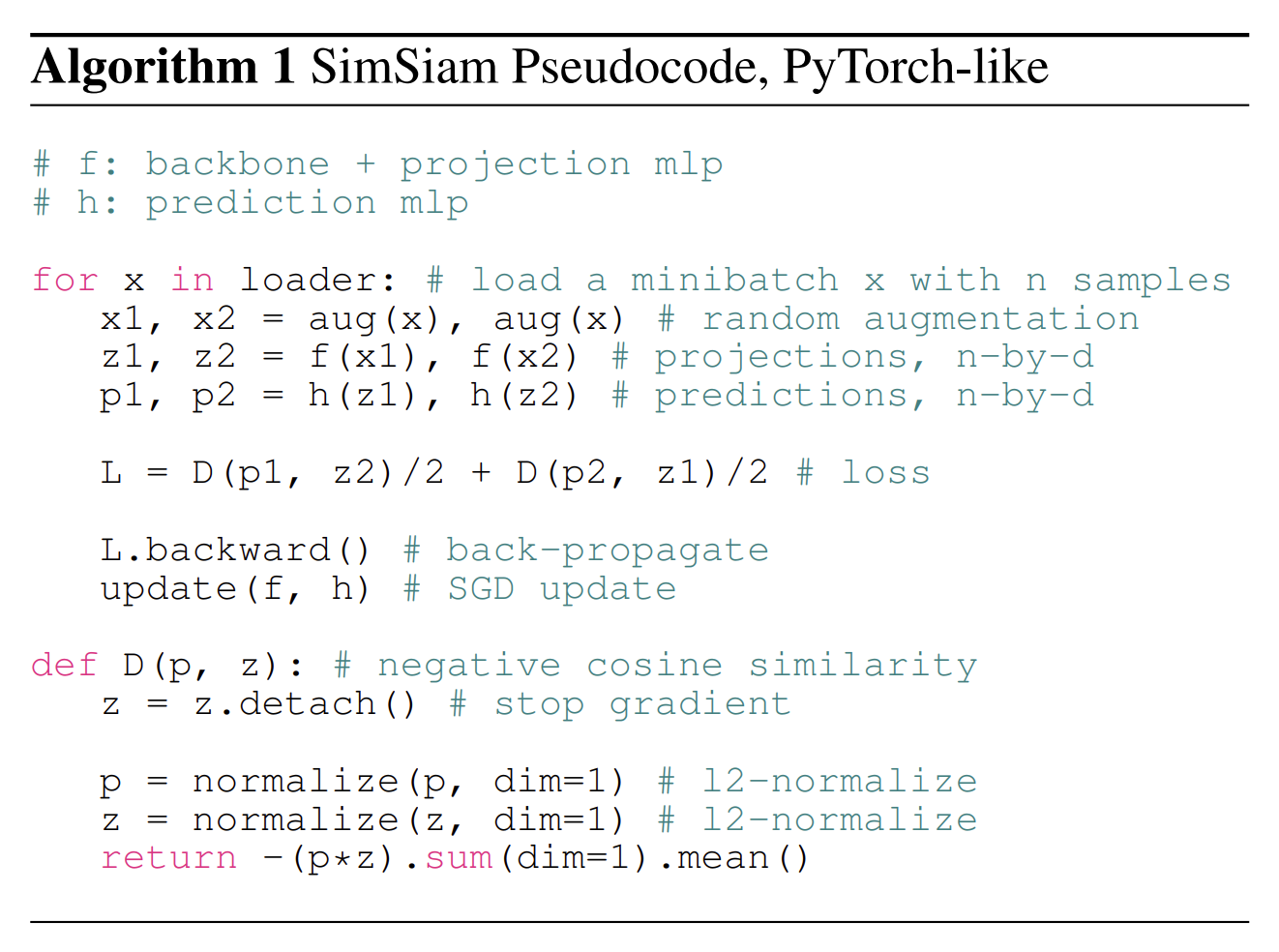
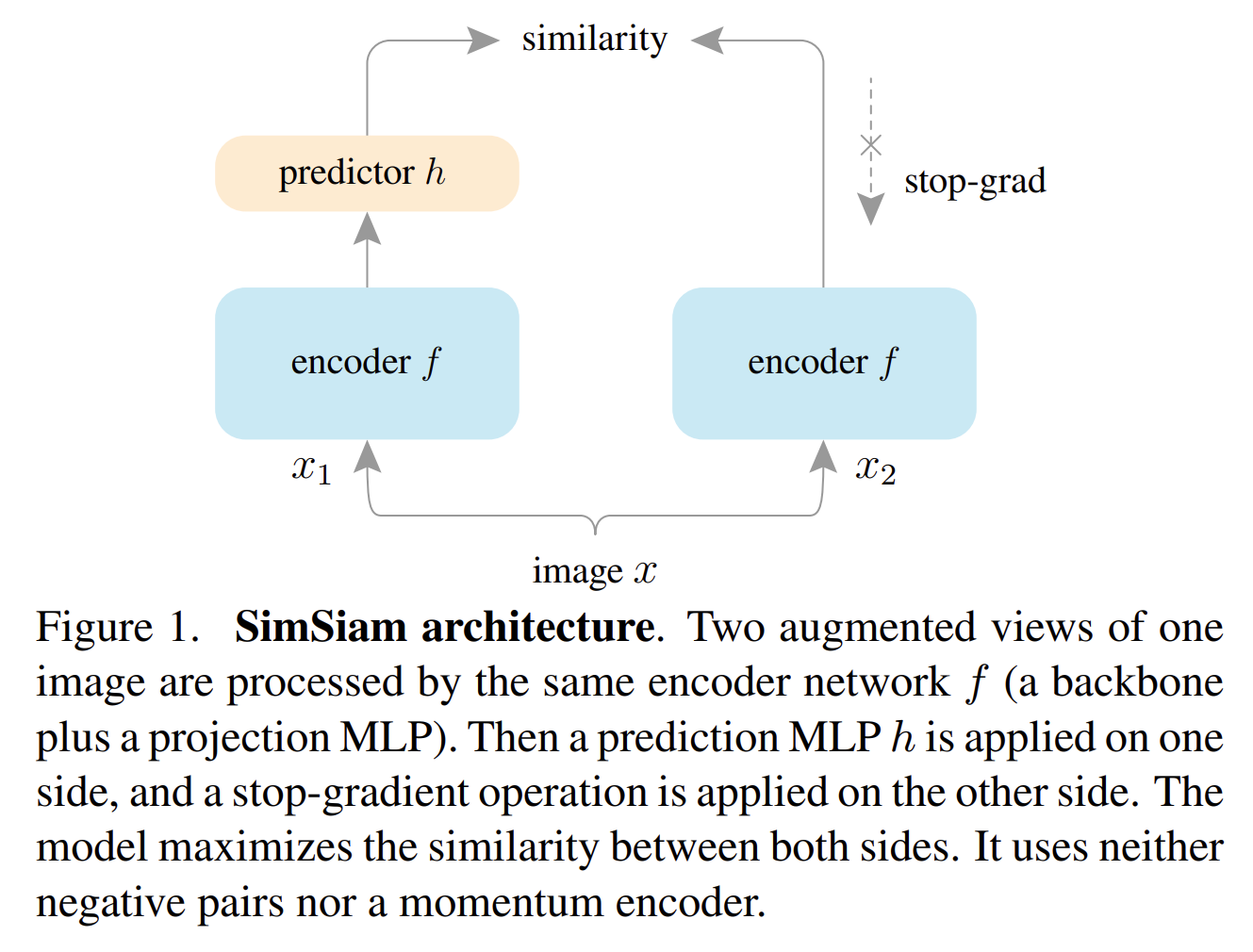
Update target network parameters using EMA



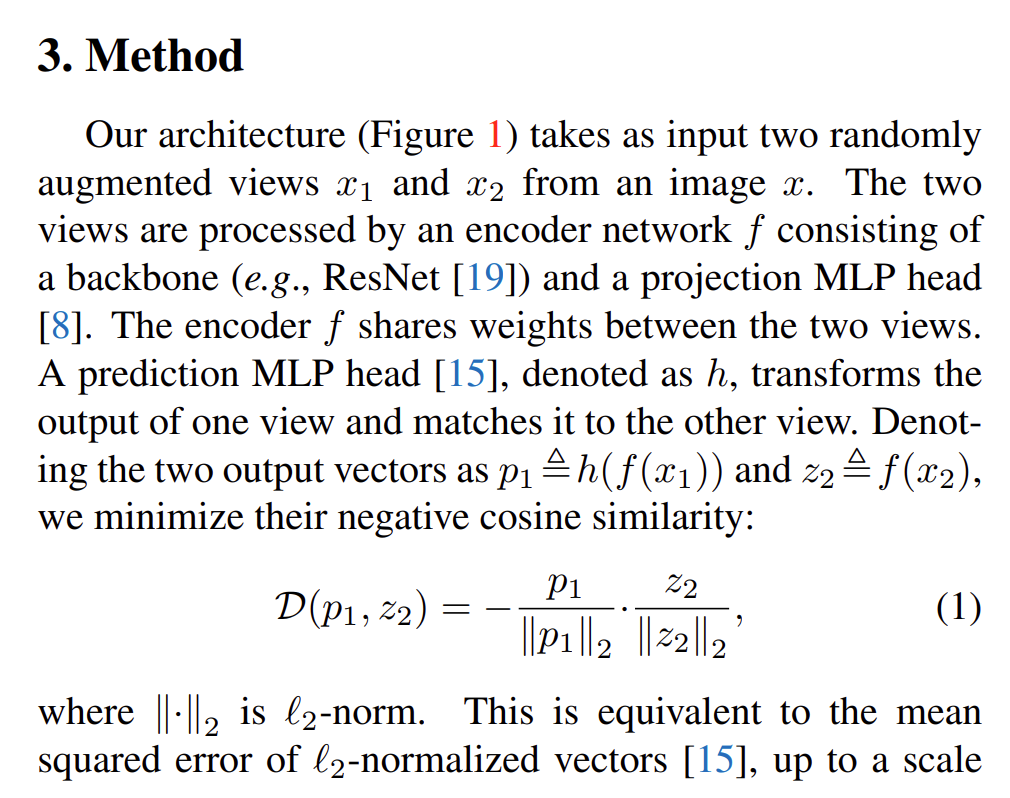
**SimSiam (2021)**

Framework

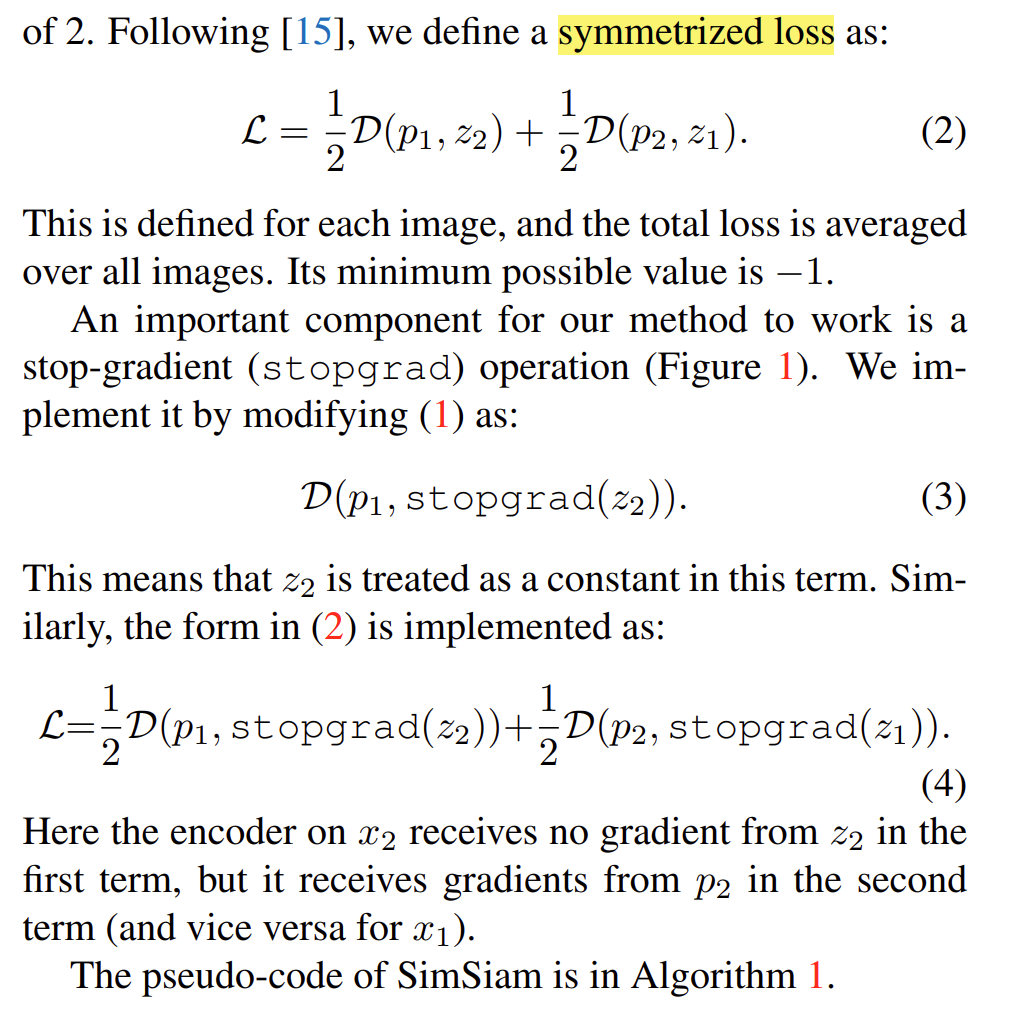
* Let be augmented views from an image .
* Encoder and a projection MLP head



Minimize negative cosine similarity

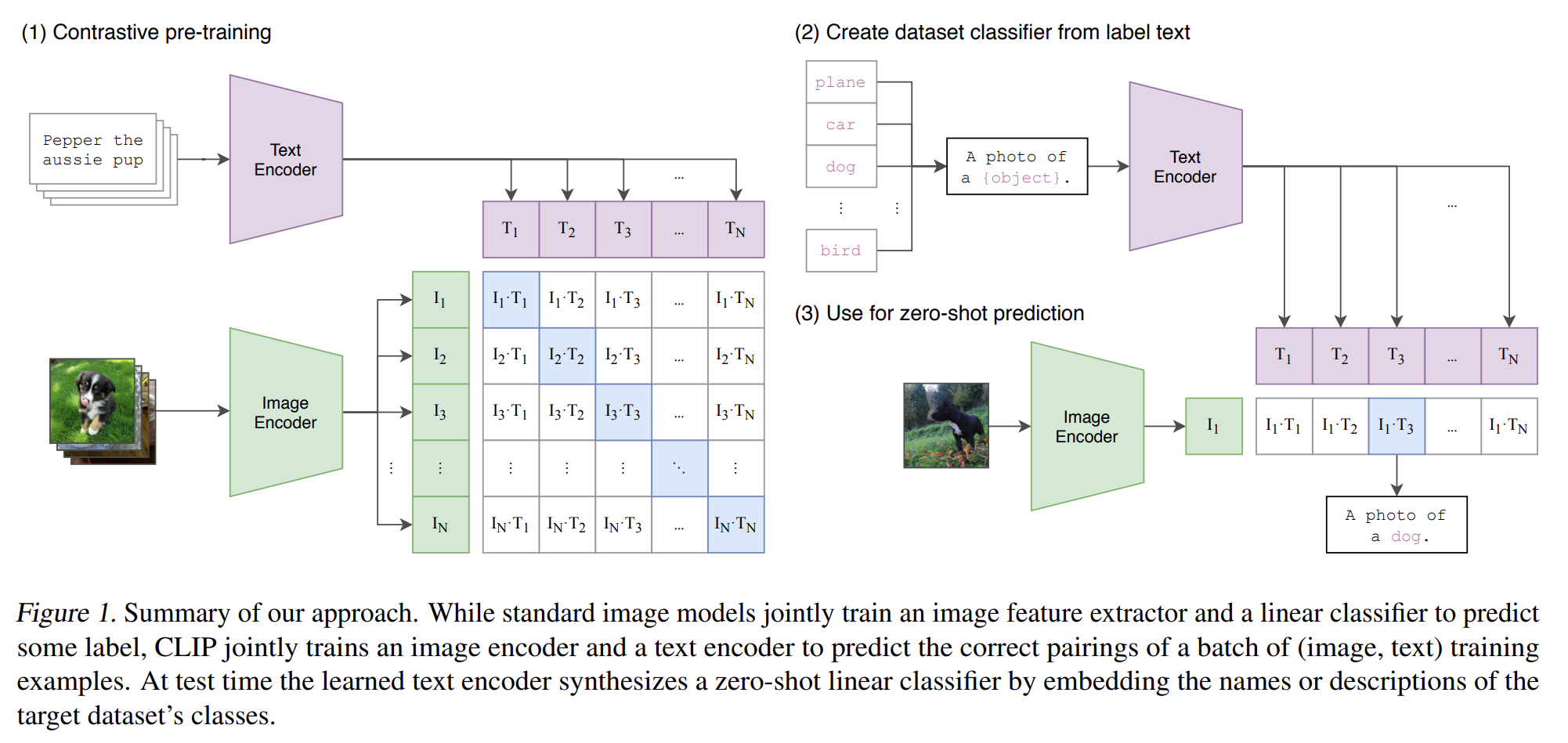


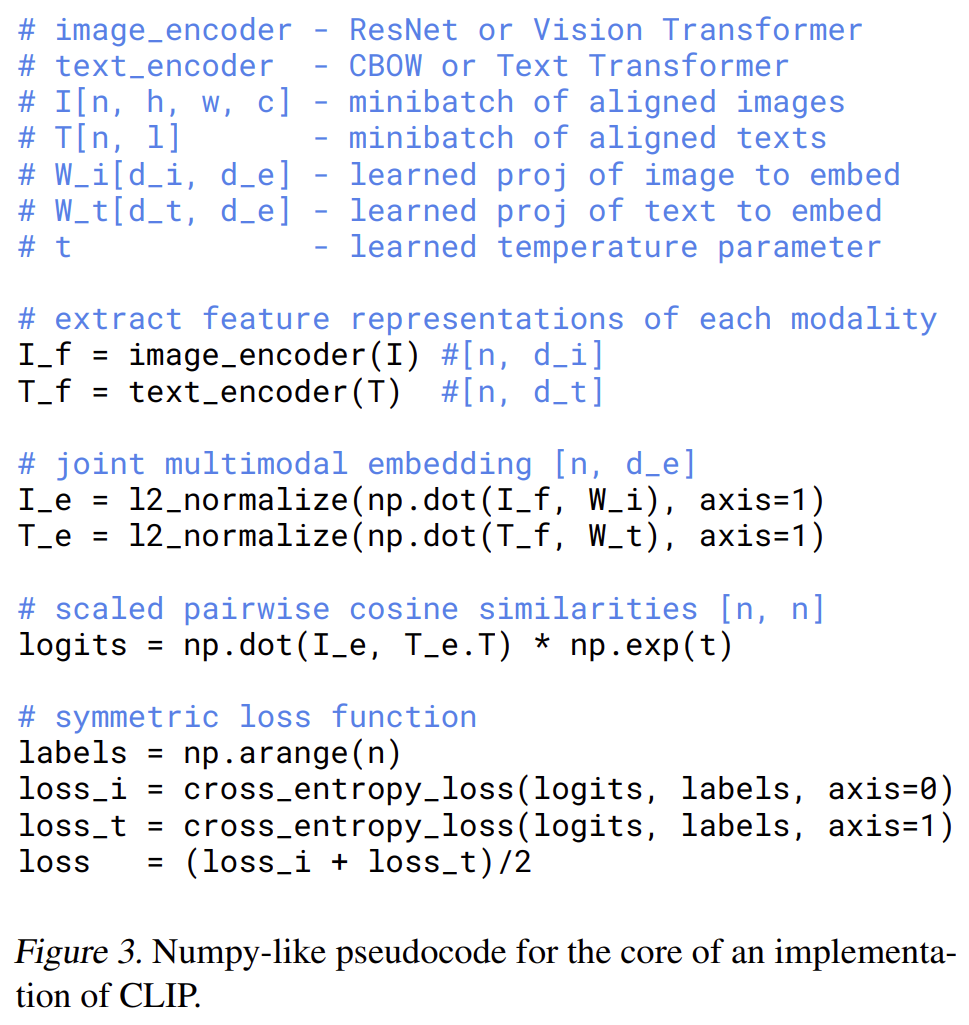
Symmetrized loss



**CLIP (2021)**

* The model learns to associate images with textual descriptions.
* Uses a text encoder and an image encoder



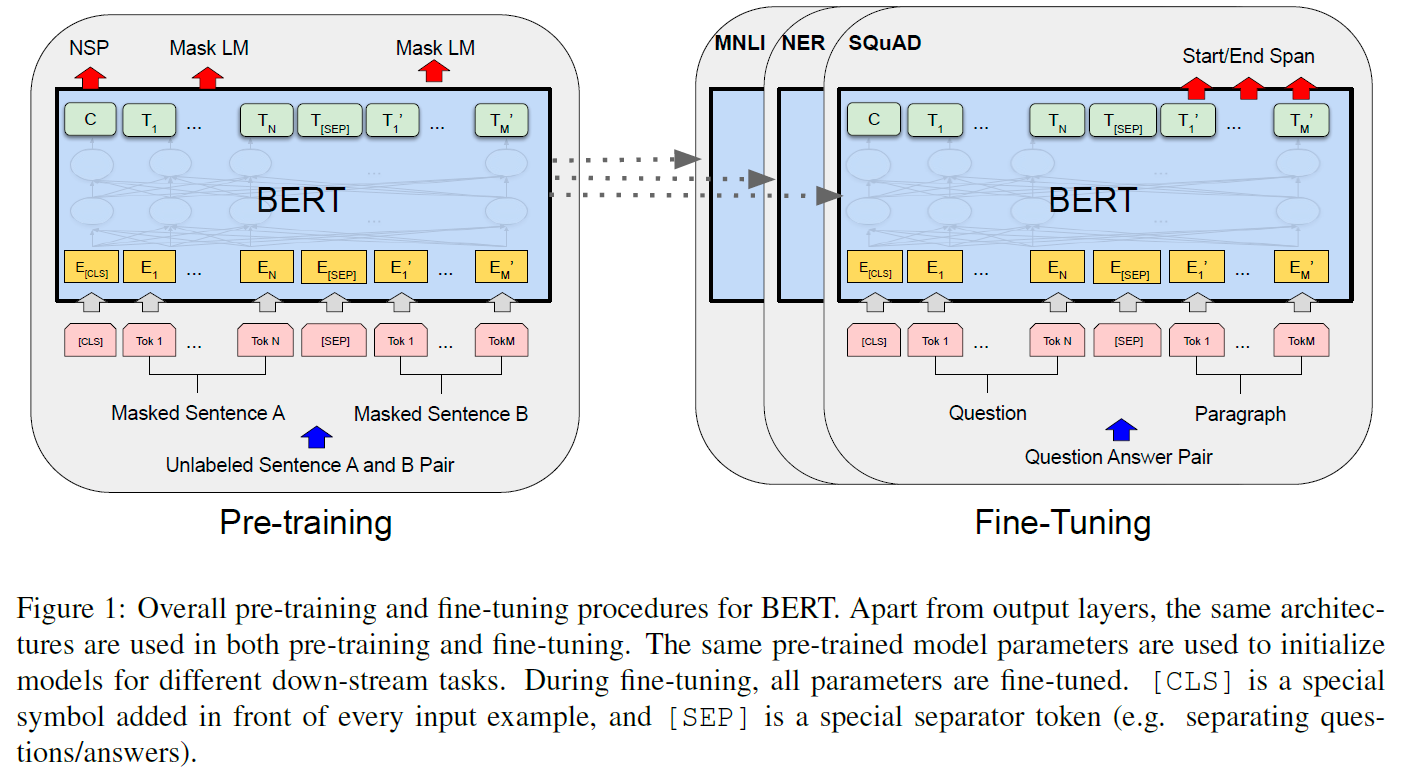


**Self-supervised learning II**

**BERT (2018)**

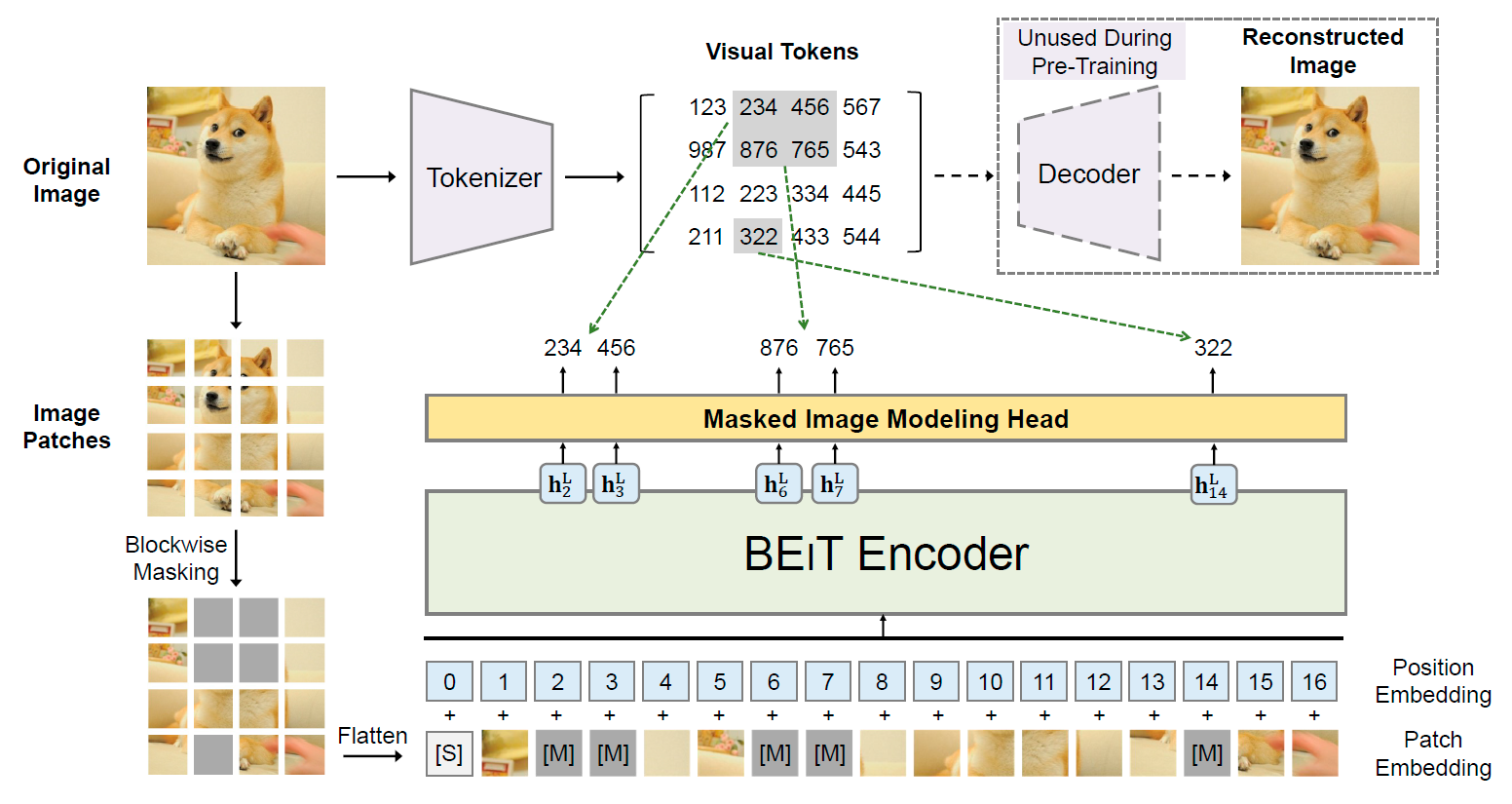
Framework steps:

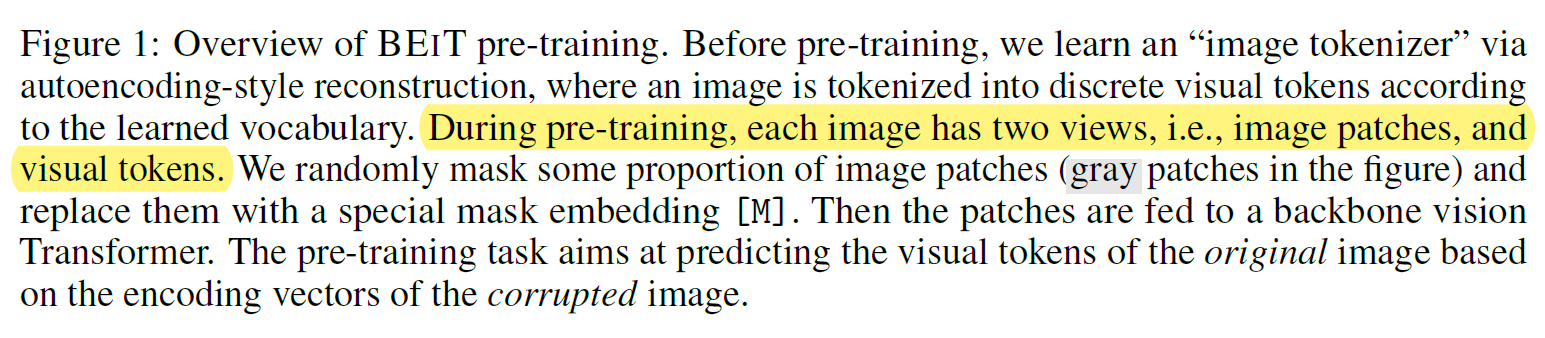
1. Pre-training: unlabeled data, pre-training tasks (Masked LM, Next Sentence Prediction).
2. Fine-tuning: labeled data, downstream tasks, initialized with the pre-trained parameters.

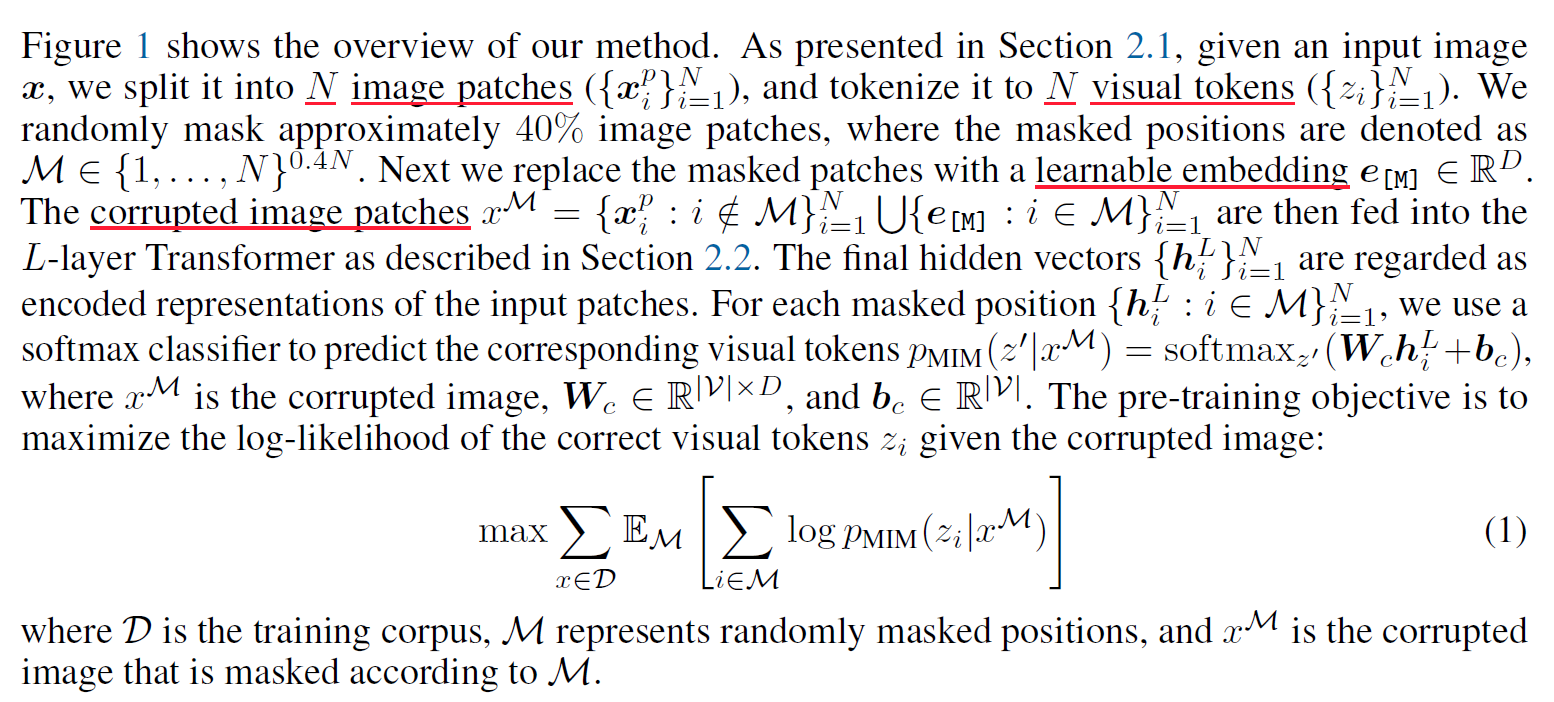
****

**BEiT (2021)**

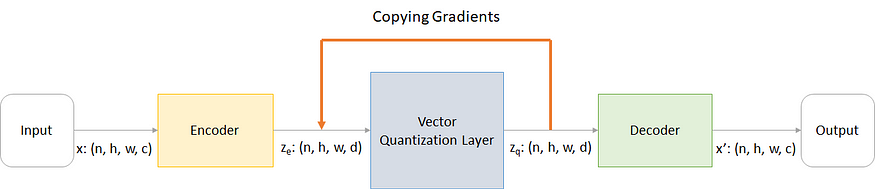
* It adapts BERT's masked prediction approach to images.
* Pre training task: masked image modeling (MIM)
* Problem handled: no pre-exist vocabulary for vision Transformer’s input unit (image patches).
* Predicts **visual tokens**.
* From VAEs import VQ-VAE

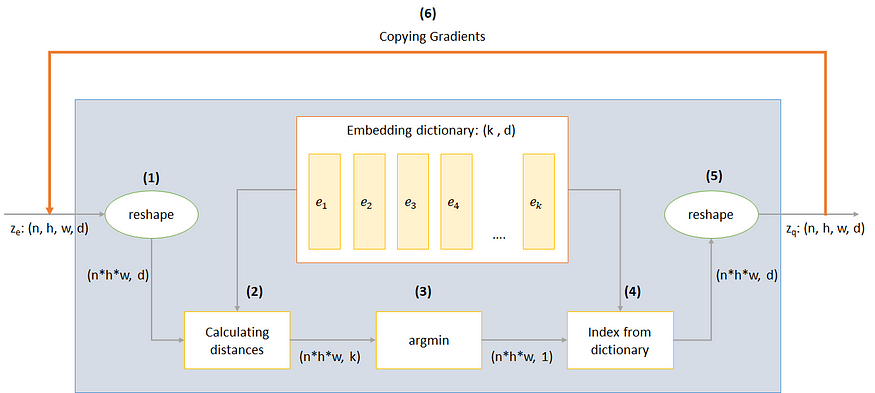






VQ-VAE (Vector Quantized-Variational Autoencoder)





**iBOT (2022)**

Preliminaries:

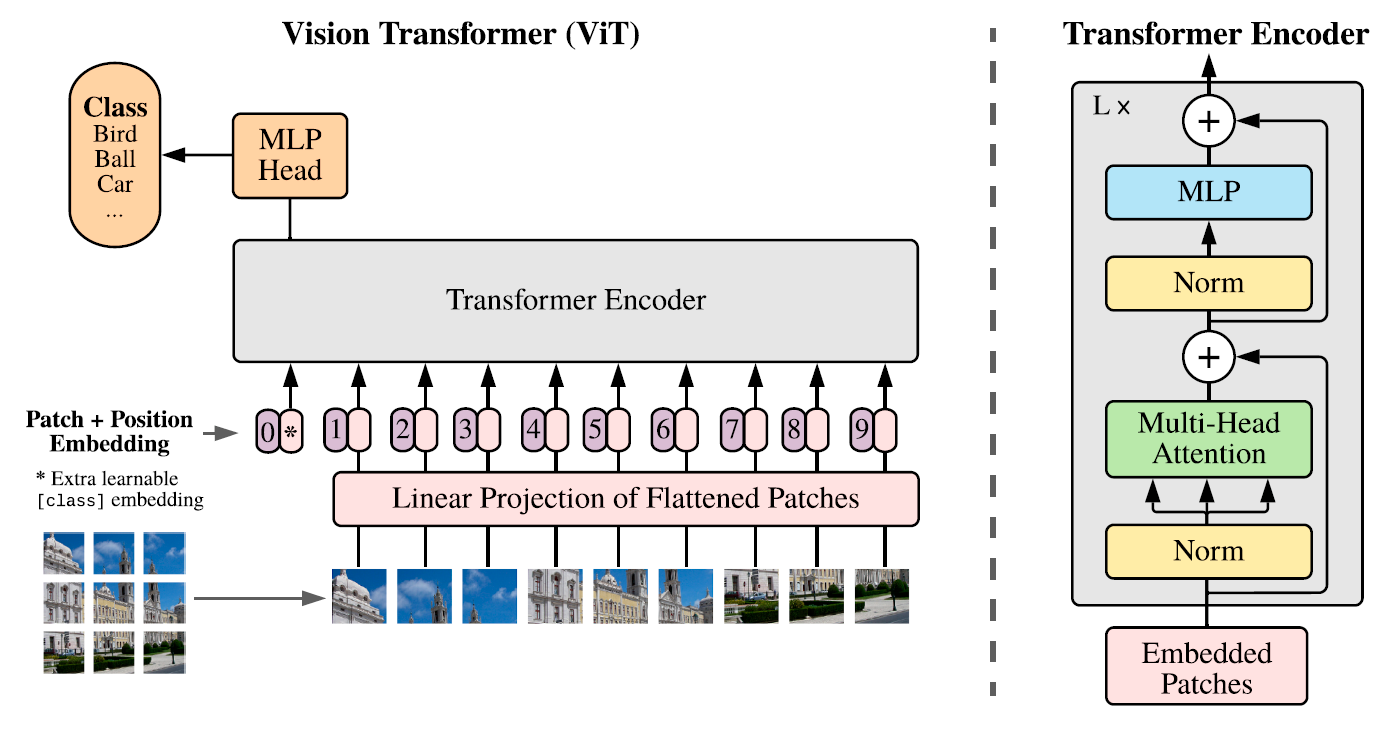
Masked image modeling (MIM)

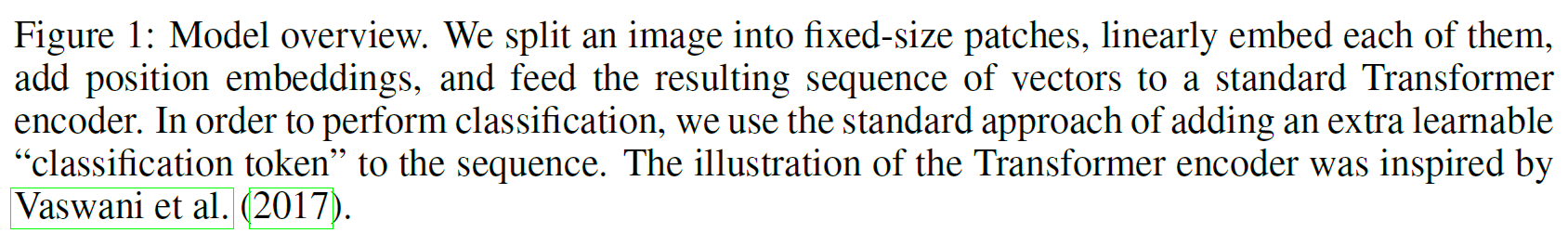
Self-distillation (proposed in DINO).

Let be distorted views of . are then put through a teacher-student framework to get the predictive categorical distributions from the [CLS] token. The knowledge is distilled from teacher to student by minimizing their cross-entropy:

iBOT uses ViTs as backbones.

ViT overview (CLS token, extra learnable “classification token”)

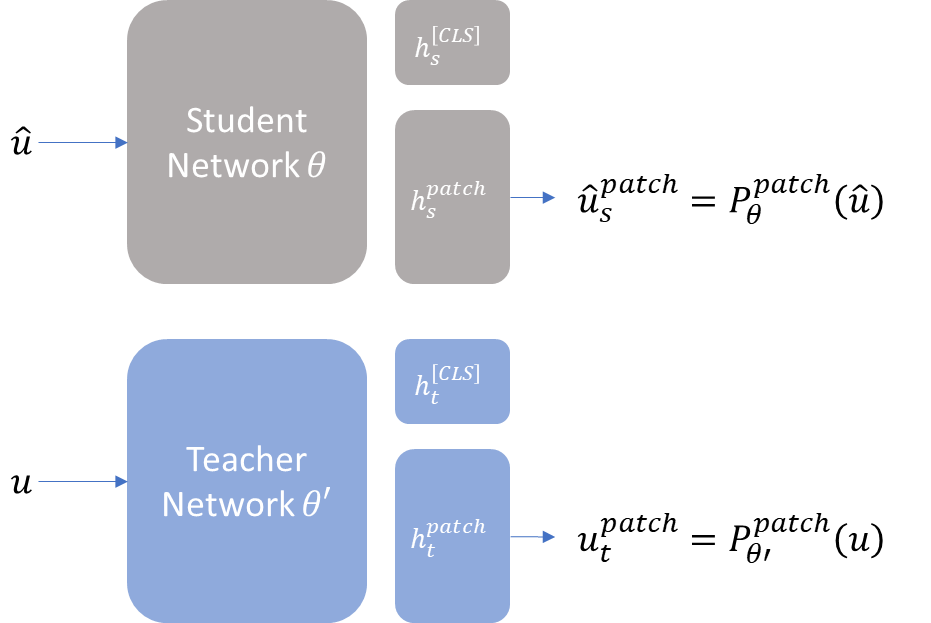




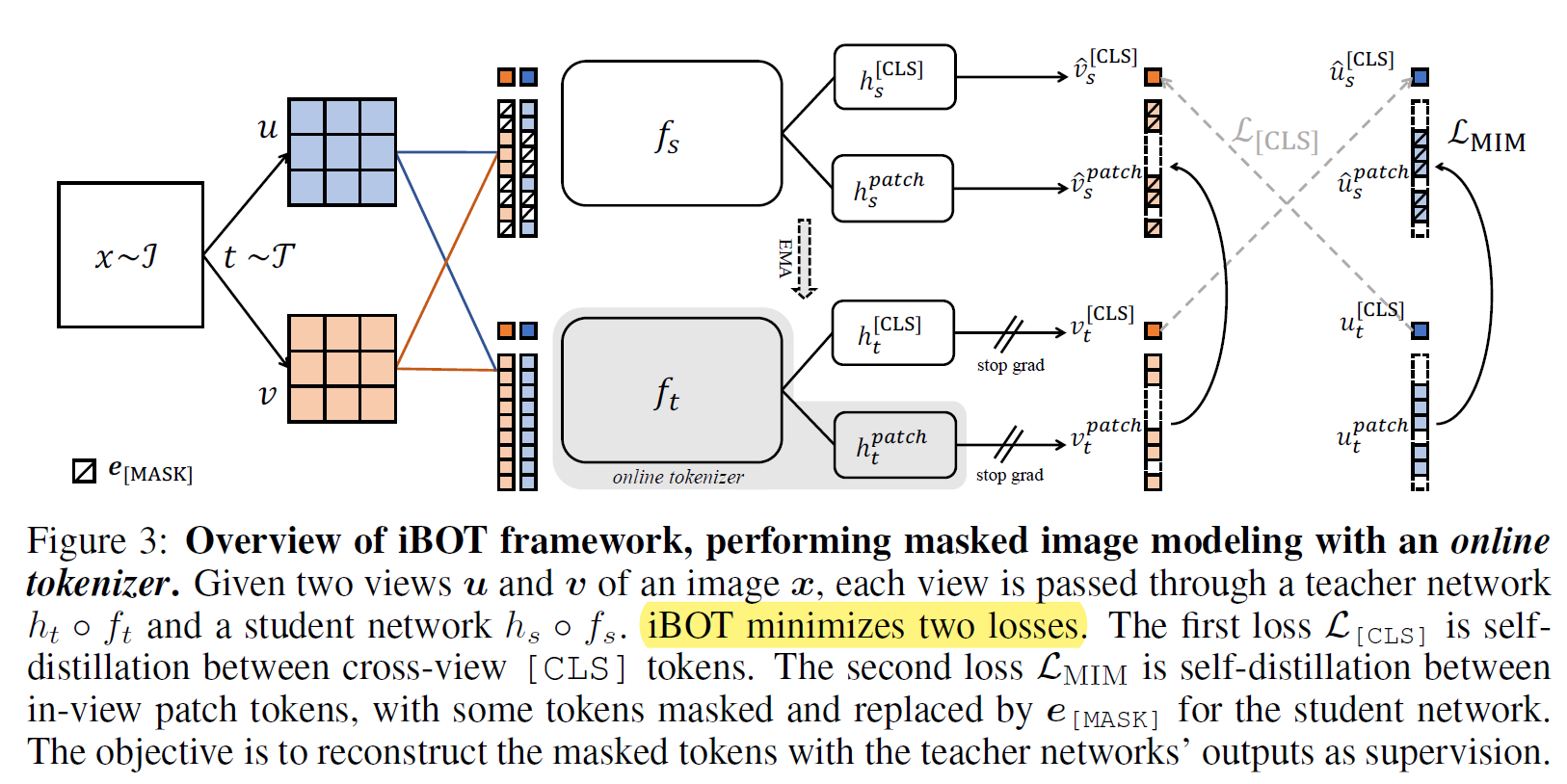
Framework

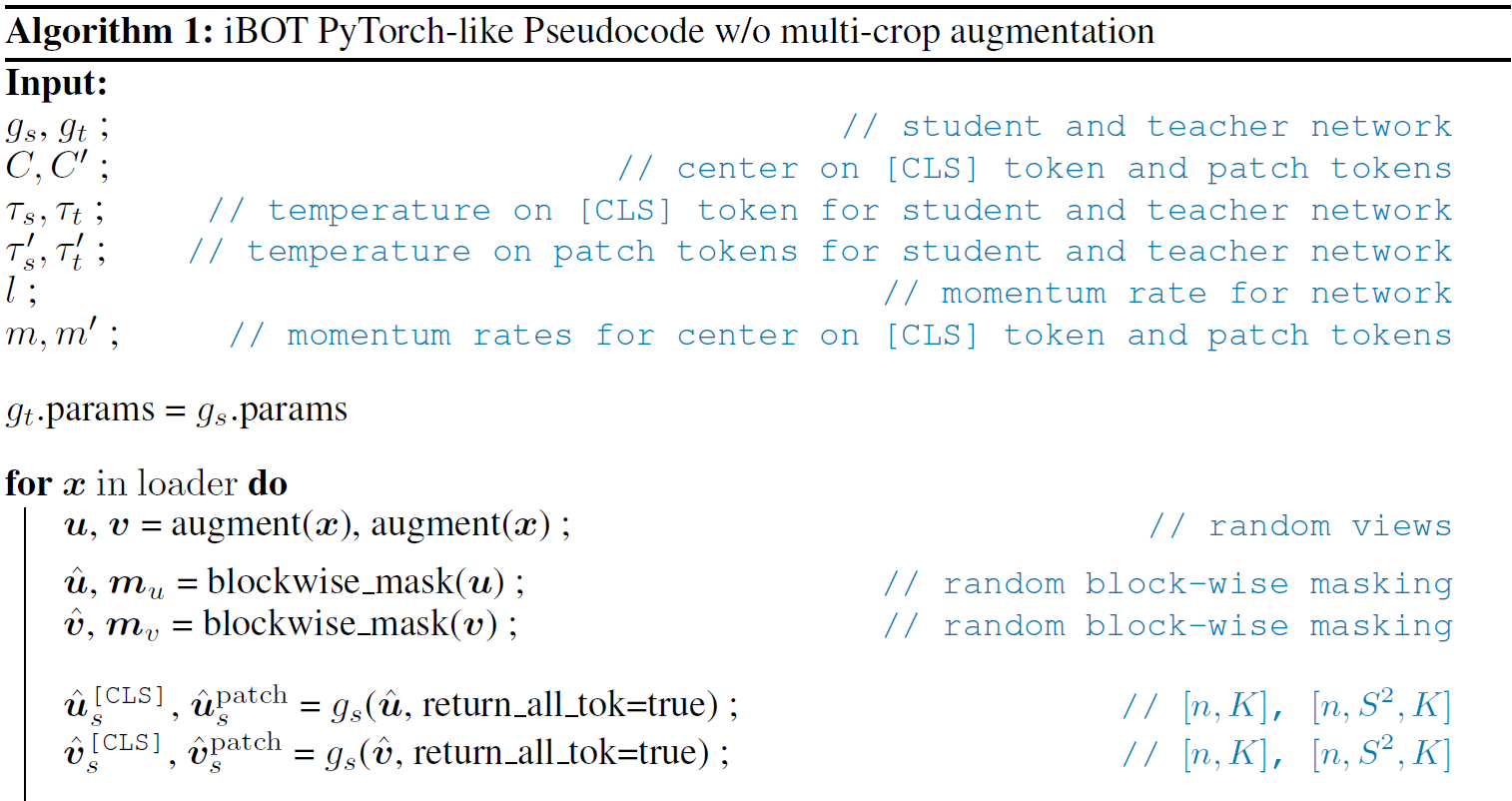
* Let be masked views.
* Student network outputs
* Teacher network outputs
* Training objective of MIM in iBOT:

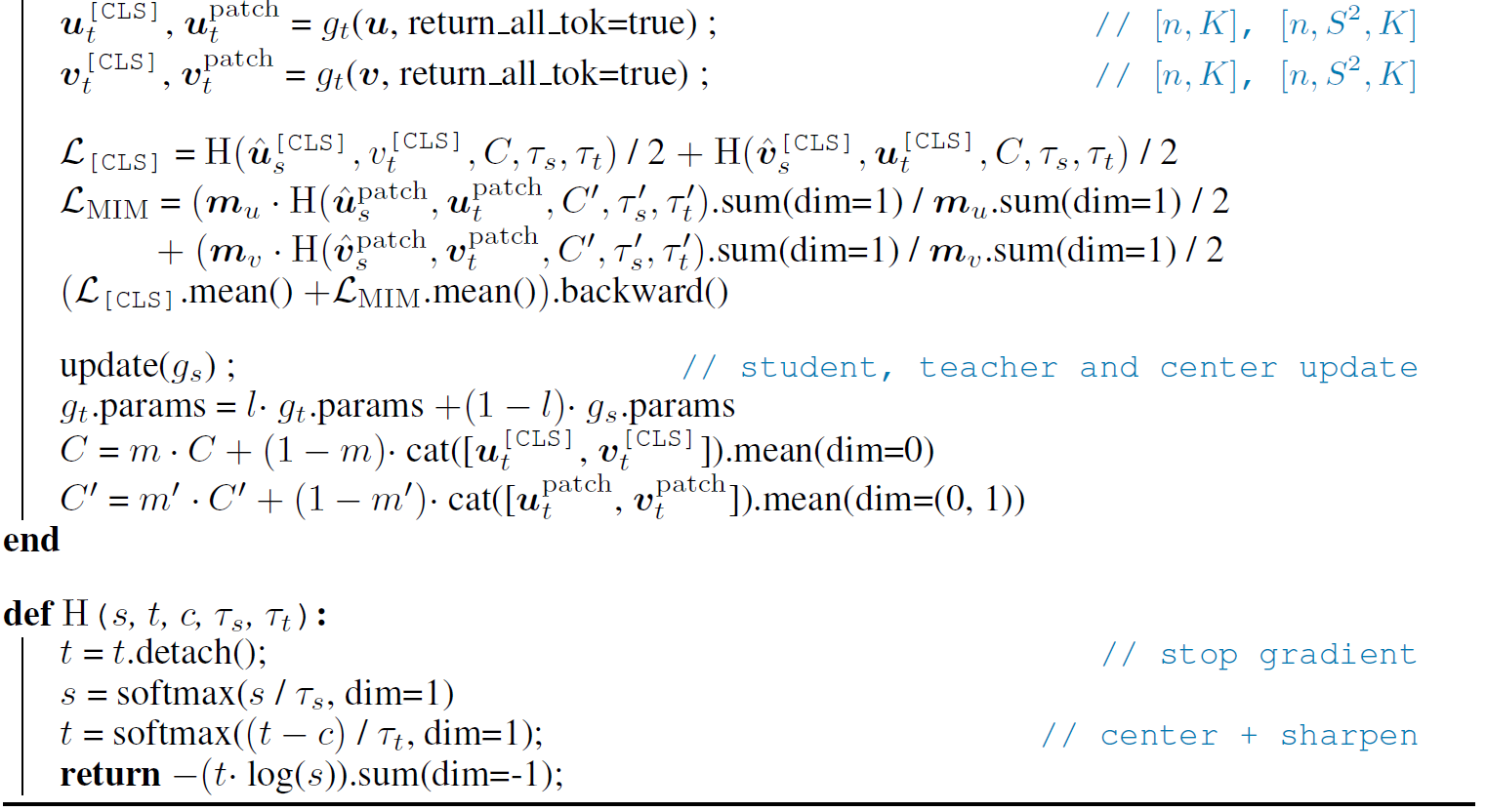
The loss symmetrized by averaging with another CE term between and .



Complete pipeline







**SimMIM (2022)**

Simple Masked Image Modeling

**DINO (2021)**

Both networks produce a probability distribution and over classes using the SoftMax function.

In Knowledge distillation, the student network learns using .

The problem is adapted to self-supervised learning using:

Where: is a set of local and global views, are global views.

“All crops are passed through the student while only the global views are passed through the teacher […]”

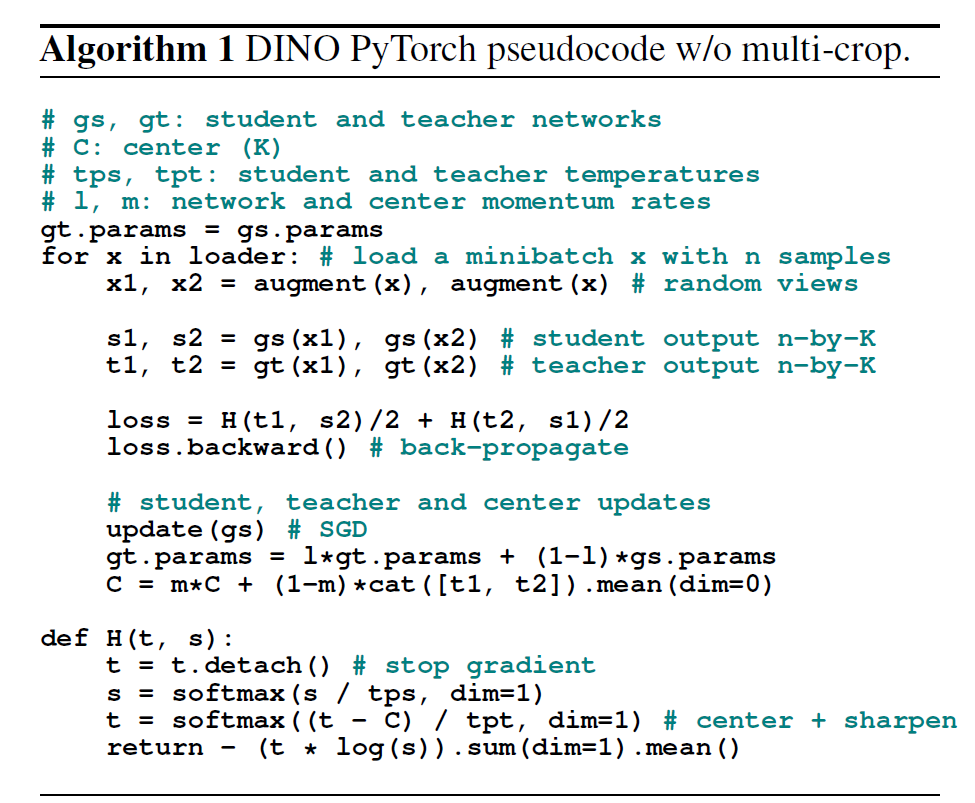
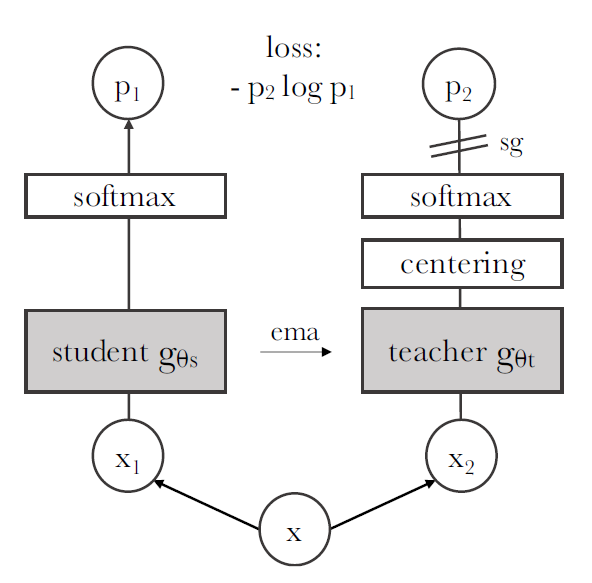
The teacher is built from past iterations of the student (Self-distillation).

Update rule:

follows a cosine schedule from 0.996 to 1 during training.

The framework prevents **collapse** by centering and sharpening the momentum teacher's outputs.

The center is updated with EMA.



**DEiT (2021)**

Training data-efficient image transformers & distillation through attention

Vision transformer (attention layers)

**Distillation token:** on output of the network its objective is to reproduce the (hard) label predicted by the teacher, instead of true label.

**Soft distillation:** minimizes .

**Hard distillation:** takes the hard decision of the teacher as a true label.

