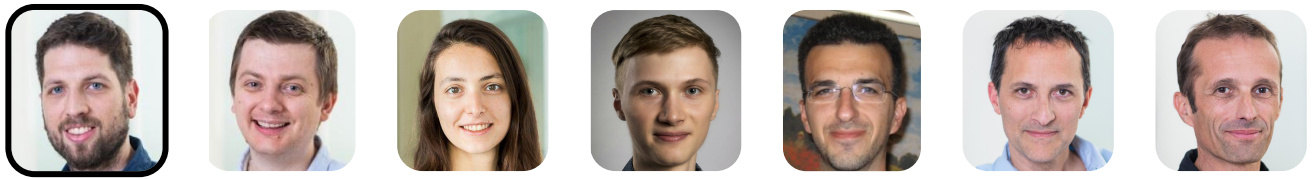
ICLR 2025

MOCA ☕: Self-Supervised Representation Learning by Predicting Masked Online Codebook Assignments

Spyros Gidaris, Andrei Bursuc, Oriane Siméoni, Antonin Vobecky

Nikos Komodakis, Matthieu Cord, Patrick Pérez

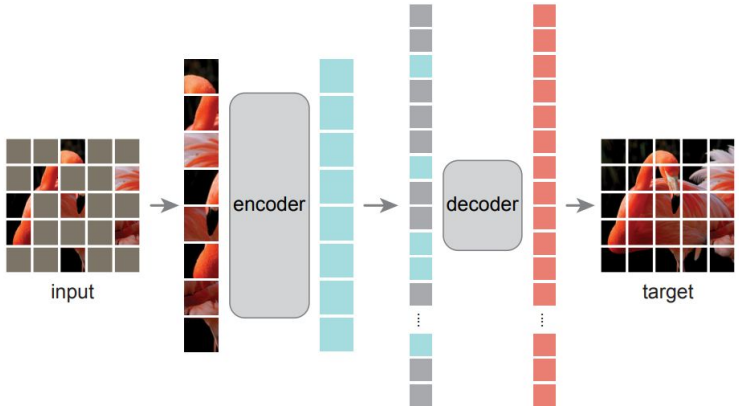


ICLR 2025 Poster — TMLR 2024 publication

Two main self-supervised learning paradigms for ViTs

Masked Image Modeling

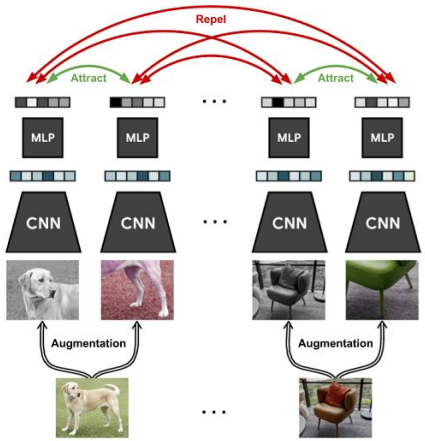
Example: mask image patches and task ViT to reconstruct their pixel values



Masked Autoencoders (He et al. 2022)

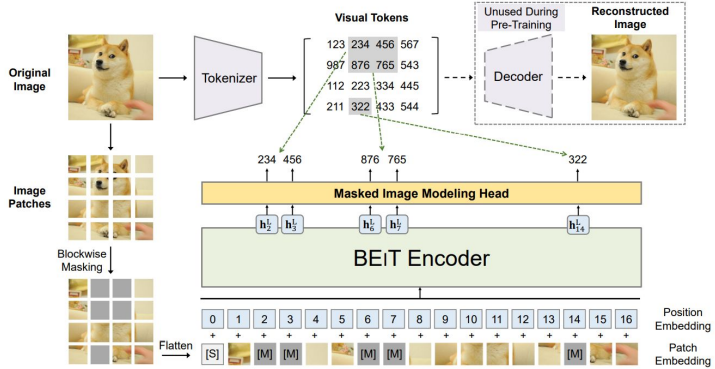
Discriminative approaches

Examples: contrastive-, teacher-student-, or clustering-style objectives

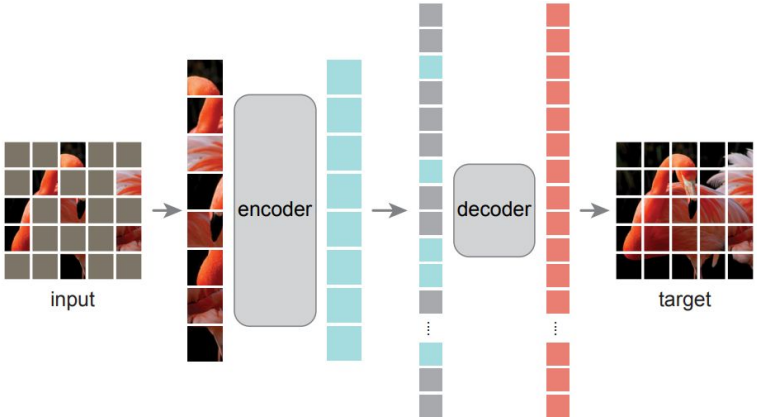


SimCLR (Chen et al. 2020)

Masked Image Modeling with Vision Transformers



BEiT (Bao et al. 2021)



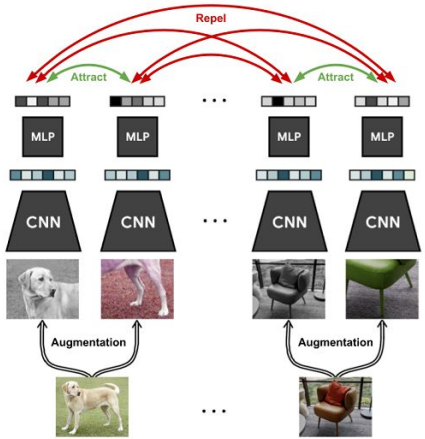
Masked Autoencoders (He et al. 2022)

Masked token prediction paradigm

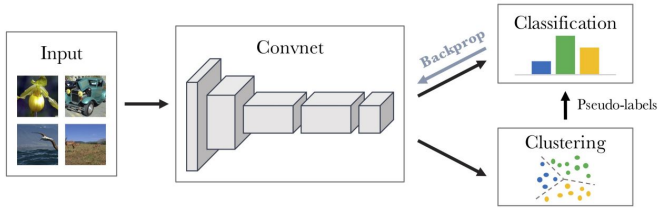
- ✓ Enforces the learning of detailed contextual and generative skills
  - ✓ Strong results when used as an initialization for downstream tasks
- ✘ Low-level reconstruction targets; Doesn't promote the learning of invariances
  - ✘ Does not provide "ready-to-use" / "out-of-the-box" representations

Discriminative self-supervised learning approaches

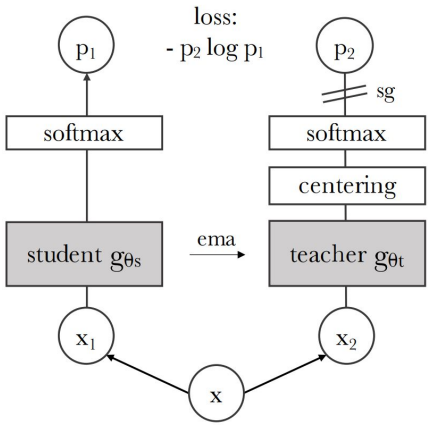
Contrastive-based
e.g., SimCLR (Chen et al. 2020)



Clustering-style objectives
e.g., DeepCluster (Caron et al. 2019)



Teacher-student based
e.g., DINO (Caron et al. 2021)



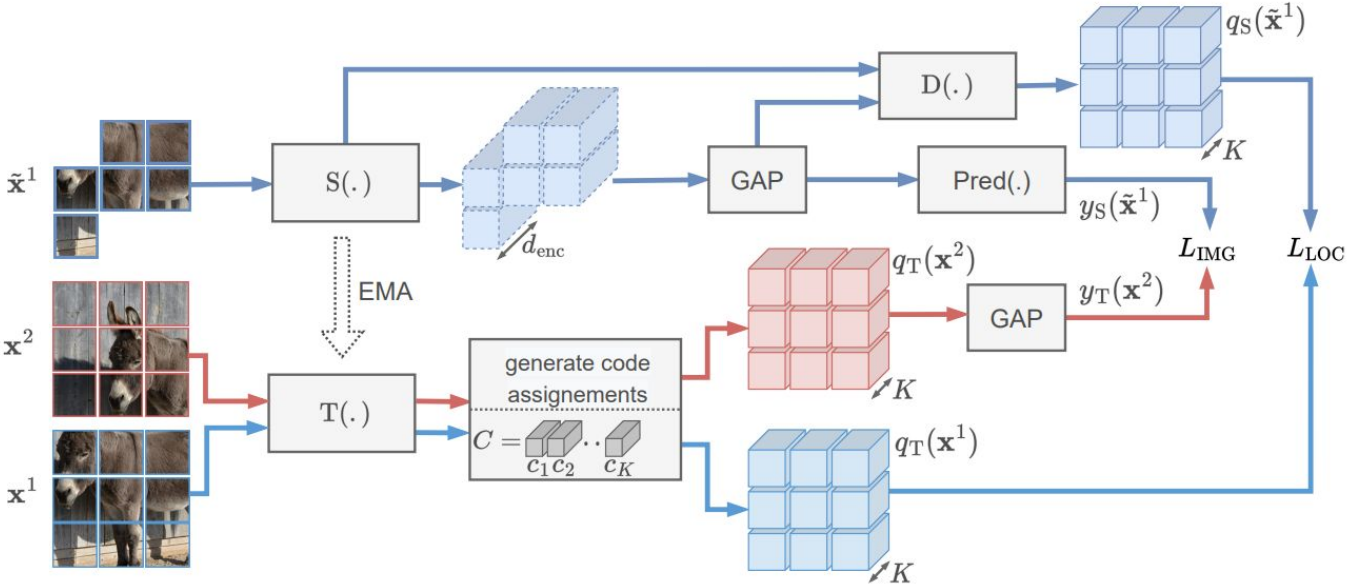
✓ Focuses on predicting high-level local visual concepts rid of “useless” image details

✘ Relies on an image-wise loss ⇒ does not promote detailed feature generation skills

MOCA: Unifies two different self-supervised learning paradigms

S(.) : Student encoder
T(.) : Teacher encoder
D(.) : Decoder

original image



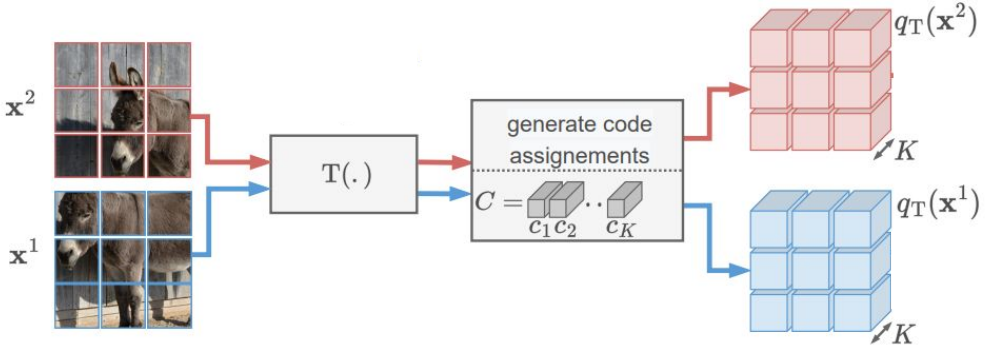
MOCA, a masked-based teacher-student method enforcing good reconstruction of patch- wise codebook assignments, which encode high-level & perturbation invariant features

✓ Unifies both discriminative and masked image modeling paradigms

MOCA: Unifies two different self-supervised learning paradigms

S(.) : Student encoder
T(.) : Teacher encoder
D(.) : Decoder

original image

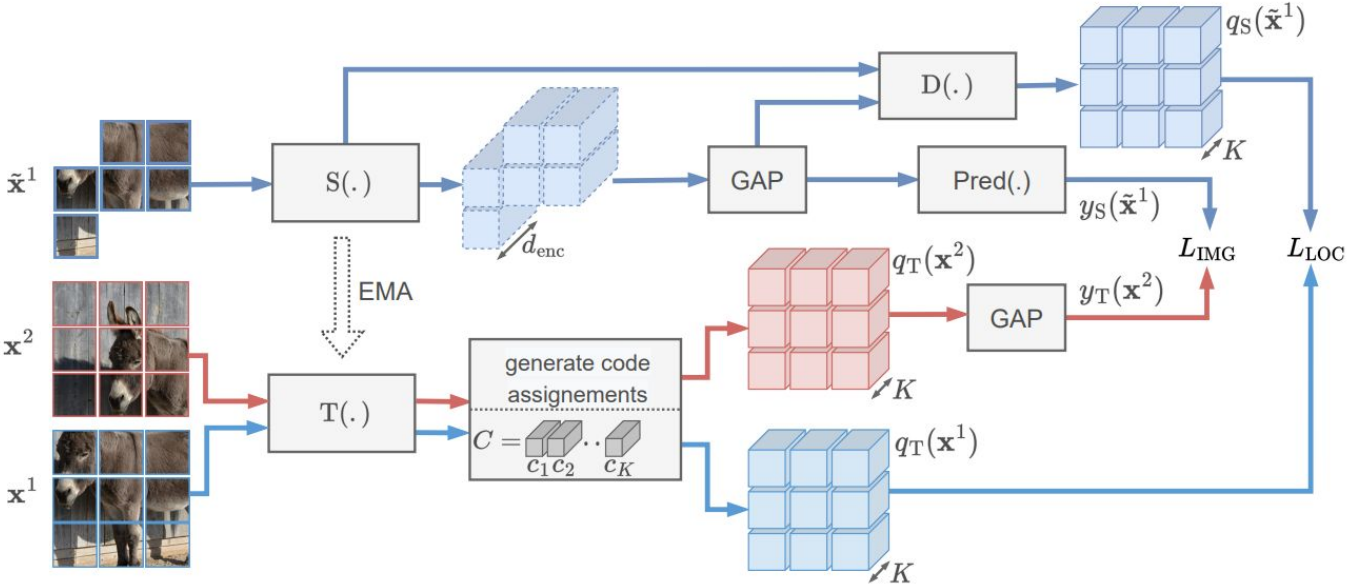


Teacher (EMA) network: takes 2 unmasked random views of same image and generates dense token-wise code assignments for them, i.e, soft-assigns codebook items to the patch tokens

MOCA: Unifies two different self-supervised learning paradigms

S(.) : Student encoder
T(.) : Teacher encoder
D(.) : Decoder

original image



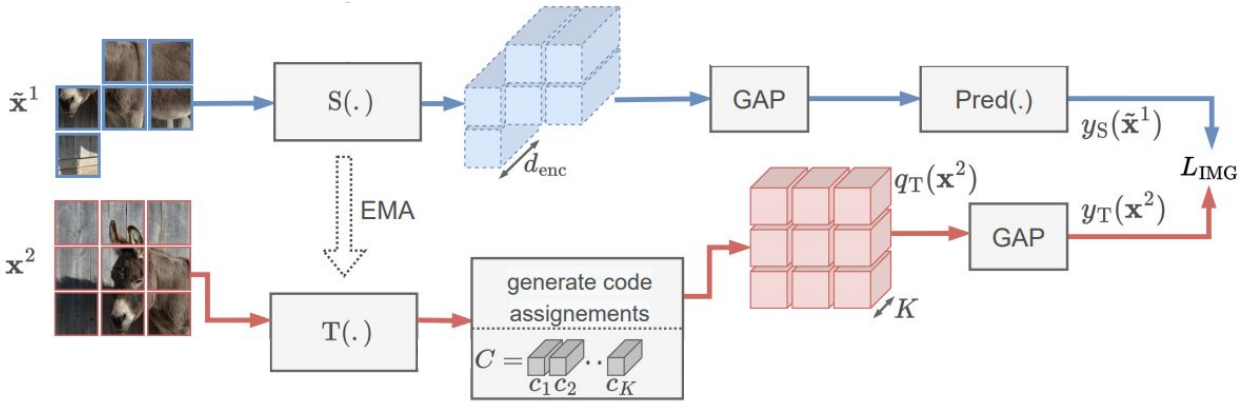
Student network is trained to minimize two types of self-supervised losses:

1. *Image-wise loss*: masked cross-view average assignment prediction
2. *Dense patch-wise loss*: masked same-view token assignment prediction

MOCA: Unifies two different self-supervised learning paradigms

S(.) : Student encoder
T(.) : Teacher encoder
D(.) : Decoder

original image

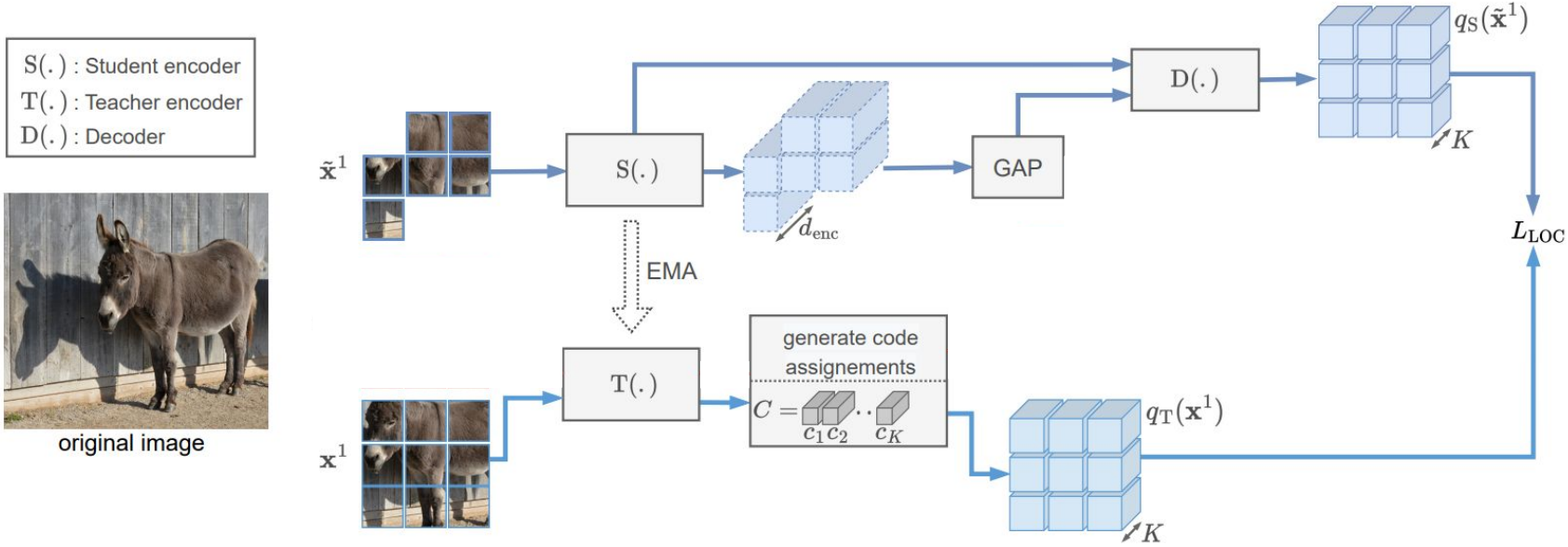


this is essentially a teacher-student image–wise loss

Student network is trained to minimize two types of self-supervised losses:

1. *Image-wise loss*: masked cross-view average assignment prediction
   ✓ Promotes invariance to perturbations

MOCA: Unifies two different self-supervised learning paradigms



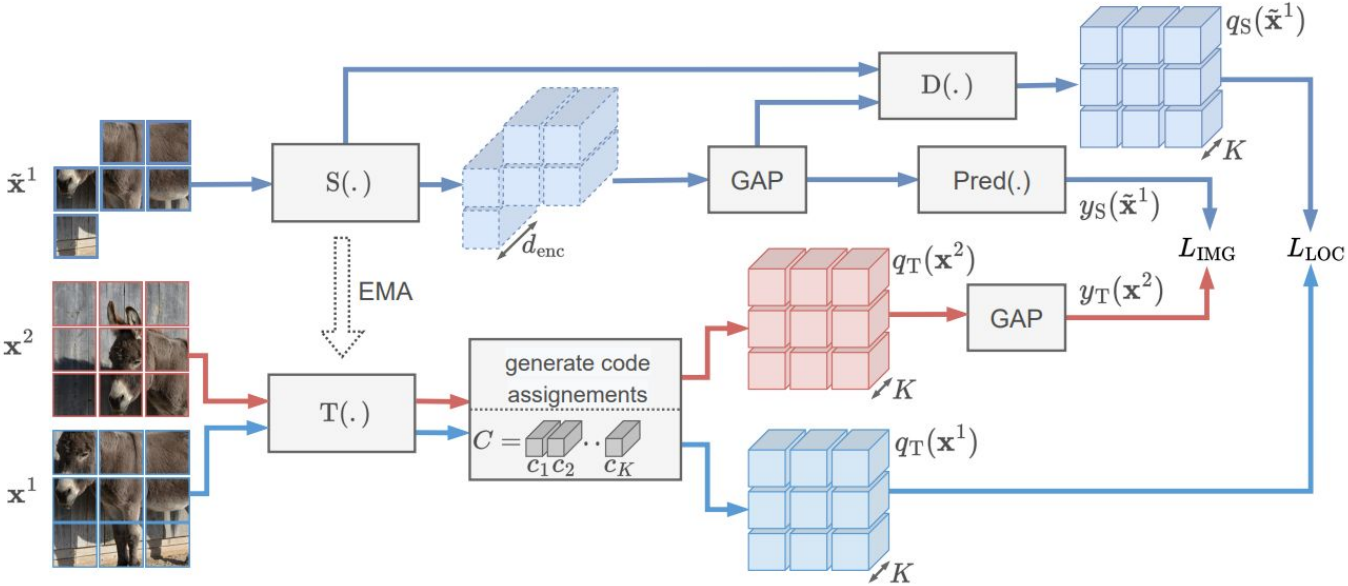
Student network is trained to minimize two types of self-supervised losses:

2. *Dense patch-wise loss*: masked same-view token assignment prediction
   - ✓ A masked image modeling loss: encourages detailed feature generation

MOCA: Unifies two different self-supervised learning paradigms

S(.) : Student encoder
T(.) : Teacher encoder
D(.) : Decoder

original image



1. Image-wise loss
2. Dense patch-wise loss

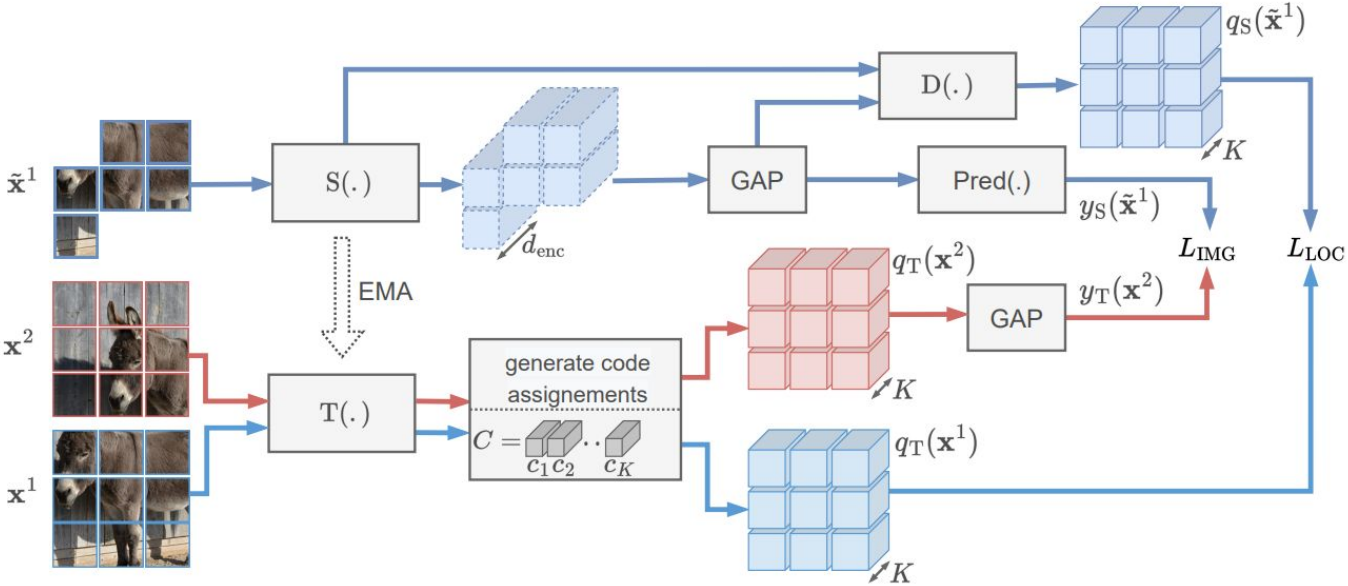
$$L = \lambda L_{\text{IMG}} + (1 - \lambda) L_{\text{LOC}}$$

λ	1.0	0.75	0.5	0.25	0.00
k-NN	66.8	70.2	**71.8**	71.5	13.1

MOCA: Unifies two different self-supervised learning paradigms

S(.) : Student encoder
T(.) : Teacher encoder
D(.) : Decoder

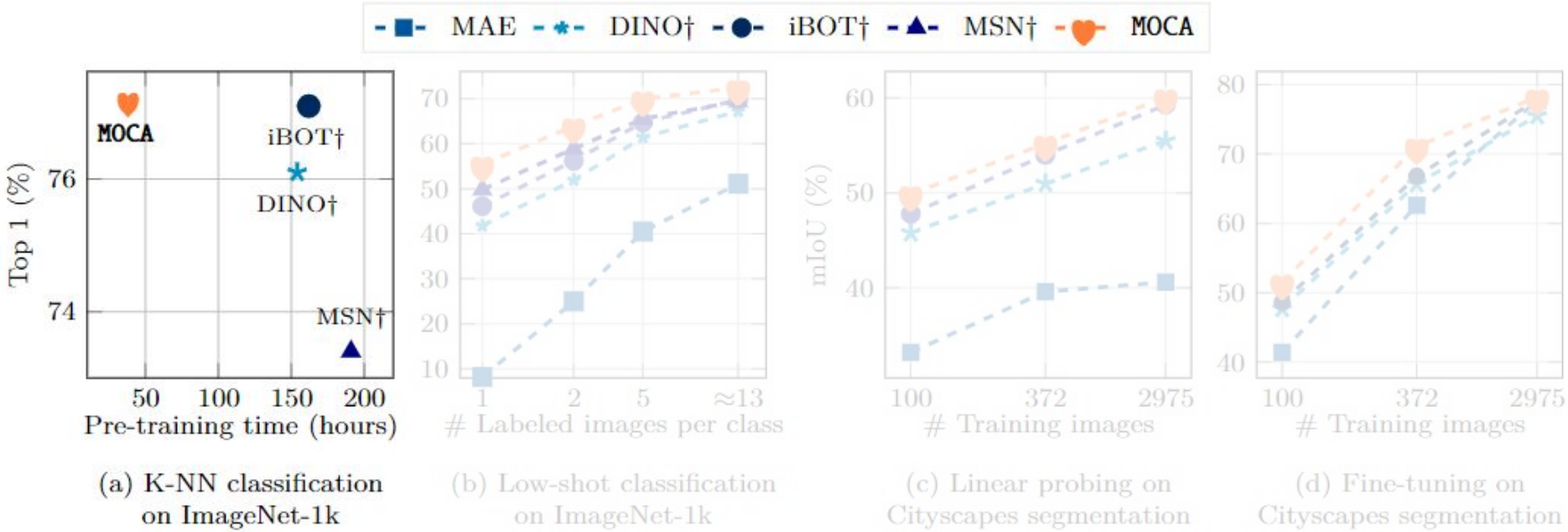
original image



1. Image-wise loss
2. Dense patch-wise loss

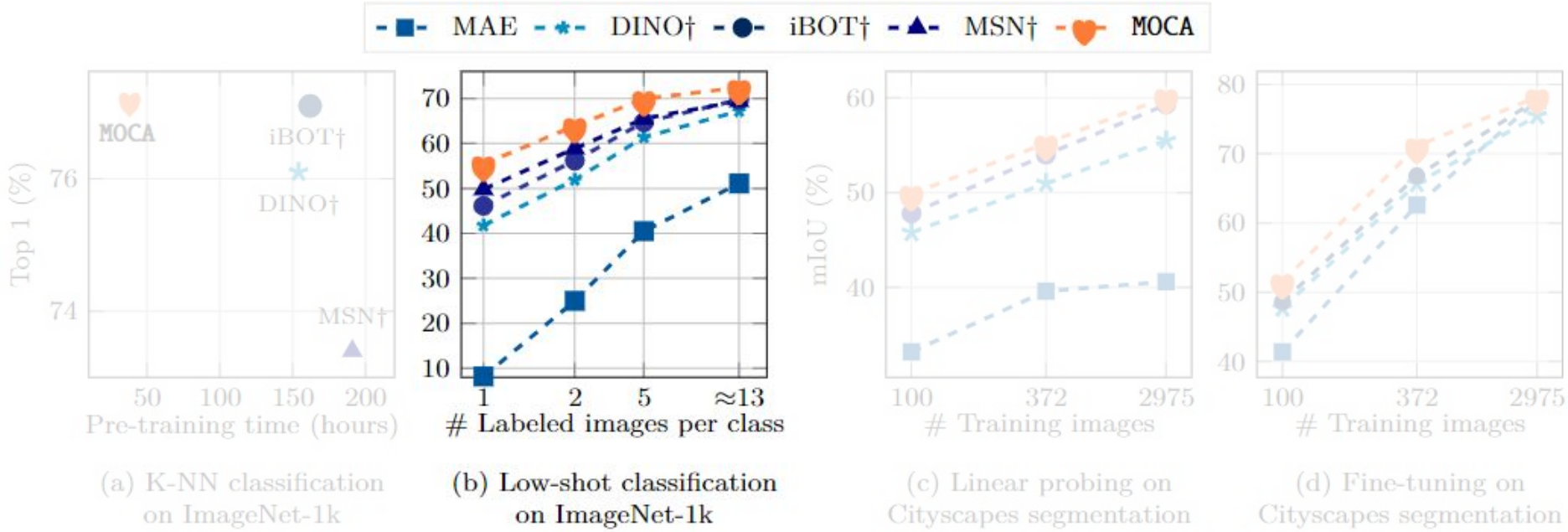
Both tasks are defined in the same space of high-level features

Evaluating self-supervised pre-trained ViT-B/16



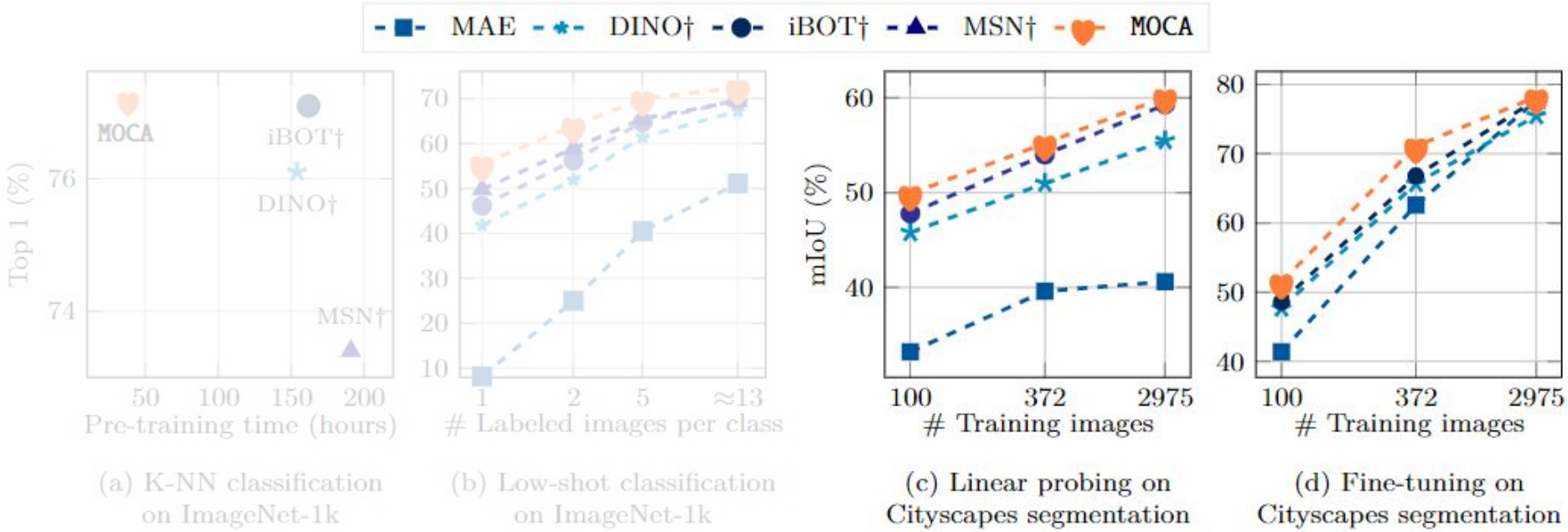
MOCA delivers **superior performance with 3x faster training** than prior methods

Evaluating self-supervised pre-trained ViT-B/16



MOCA **outperforms prior methods in low-shot ImageNet classification** achieving strong gains with just 1, 2, 5 or 13 examples per class

Evaluating self-supervised pre-trained ViT-B/16



(a) K-NN classification on ImageNet-1k

(b) Low-shot classification on ImageNet-1k

(c) Linear probing on Cityscapes segmentation

(d) Fine-tuning on Cityscapes segmentation

MOCA achieves **better results on Cityscapes semantic segmentations** across:

- Linear probing & full fine-tuning
- Full-shot & low-shot training settings

MOCA: Self-Supervised Representation Learning by Predicting Masked Online Codebook Assignments

- Unifies perturbation invariance (**discriminative**) + dense contextual reasoning (**masked image modeling**) objectives in a single framework
- A single **end-to-end teacher-student training** stage
- **3x more training efficient** than competing methods
- **Strong performance**, especially in low-shot settings

code: https://github.com/valeoai/MOCA