Context Autoencoder for Scalable Self-Supervised Representation Pretraining

Jingdong Wang

Chief Scientist for Computer Vision

AI Group, Baidu
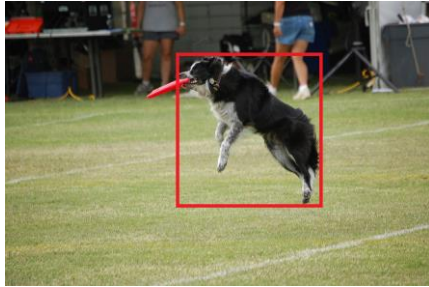
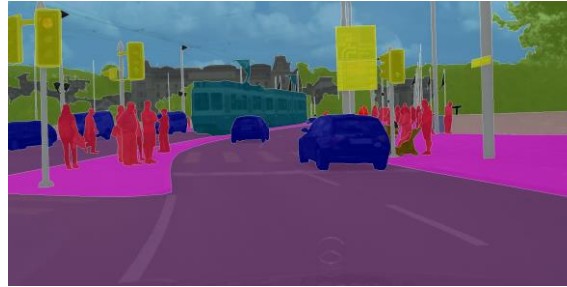https://jingdongwang2017.github.io/

June 2, 2022

@视觉模型专题论坛，北京智源大会

Computer Vision Tasks Need Representation Pretraining

Image classification

Object detection
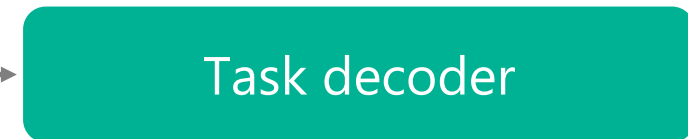
Semantic segmentation

Face alignment

Pose estimation

representation

Encoder

Task decoder

Representation Pretraining

Goal	Pretraining task	Downstream tasks
Learn an encoder mapping an image to a representation	Supervised classification, self-supervision tasks, etc.	Finetune the pretrained encoder on downstream tasks

Encoder → representation → Pretraining task decoder

Representation Pretraining

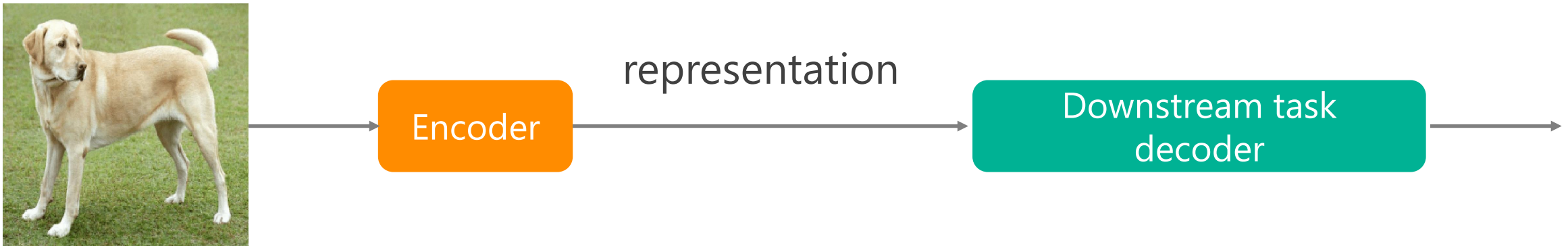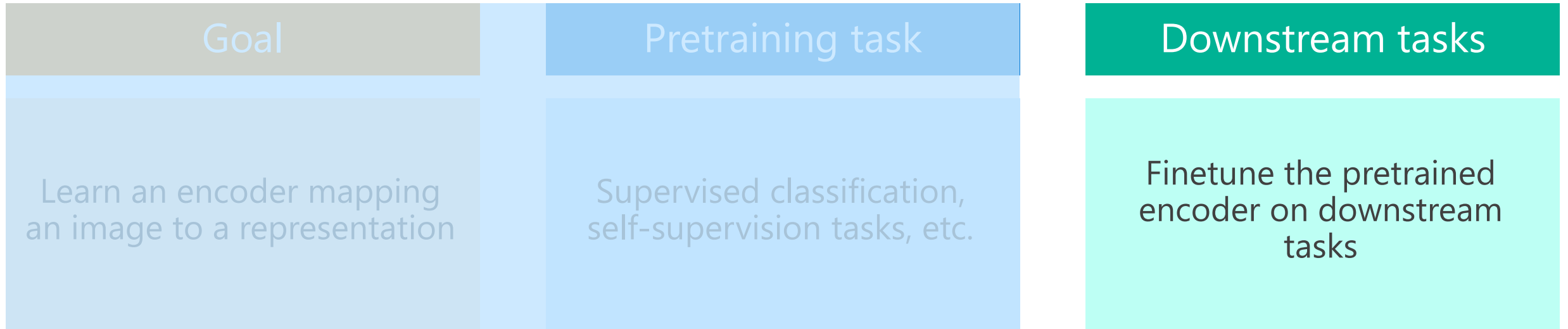| Goal	Pretraining task	Downstream tasks
Learn an encoder mapping an image to a representation	Supervised classification, self-supervision tasks, etc.	Finetune the pretrained encoder on downstream tasks

representation

Encoder → Downstream task decoder

Background

Representation Pretraining: Scaling Up

Sample scale	Concept scale
How many samples are used for pretraining	How many concepts are learned from pretraining

How Existing Pretraining Frameworks Scale Up

	Supervised	Semi-supervised	Vision-language	Self-supervised
Sample scale	🙁	🙂	🙂	🙂
Concept scale	🙁	🙁	🙂	🙂

Self-Supervised Representation Pretraining in Vision

Contrastive pretraining

- Maximize the similarity *between random crops from the same image*, and minimize the similarity *between random crops from different images*

- Representative methods: InstDis, SimCLR, MoCo, BYOL, Swav, DINO, etc.

Masked image modeling

- Divide the images into visible and masked patches

- Make predictions for the masked patches from the visible patches

- Representative methods: ViT, BEiT, MAE, SimMIM, SplitMask, etc.

Others

- Color shuffling: colorization, color shuffle, split-brain autoencoder

- Innate relationship prediction: predict relative positions between patches, Jigsaw puzzle, predict the rotation

Self-supervised pretraining is NOT unsupervised learning.
There is still supervision, that is from the images other than human annotation.

Contrastive Pretraining

Random crop

Encoder → Projector → Predictor

Random crop

Encoder → Projector

Alignment

9
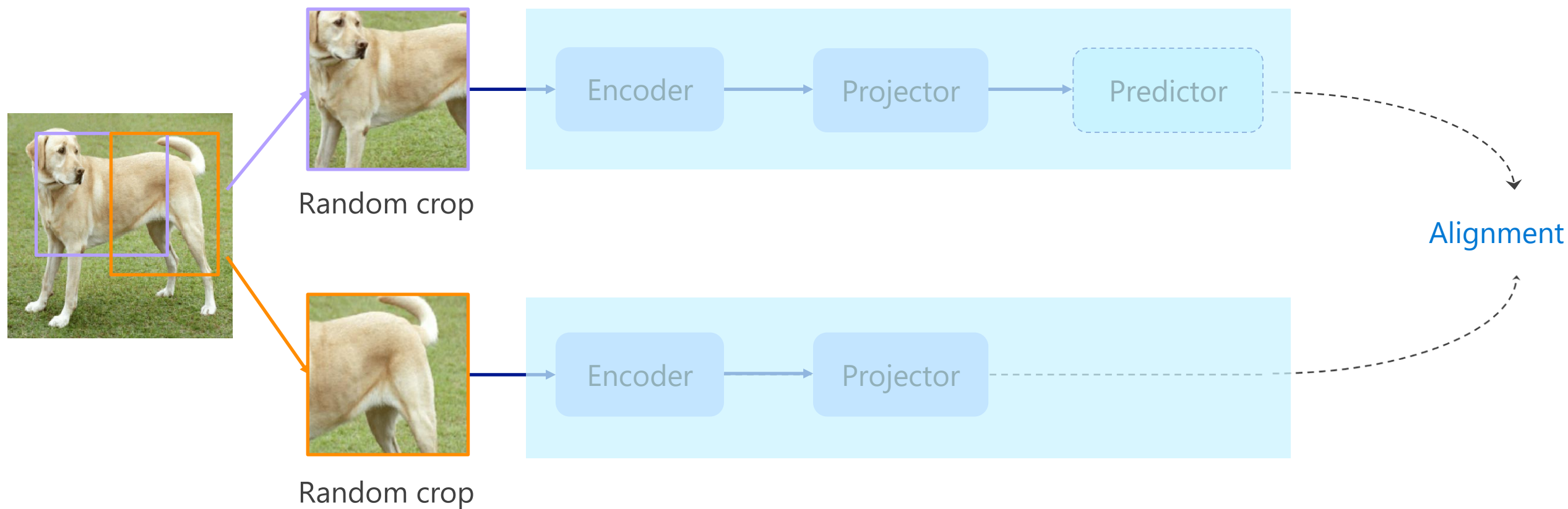
Masked Image Modeling
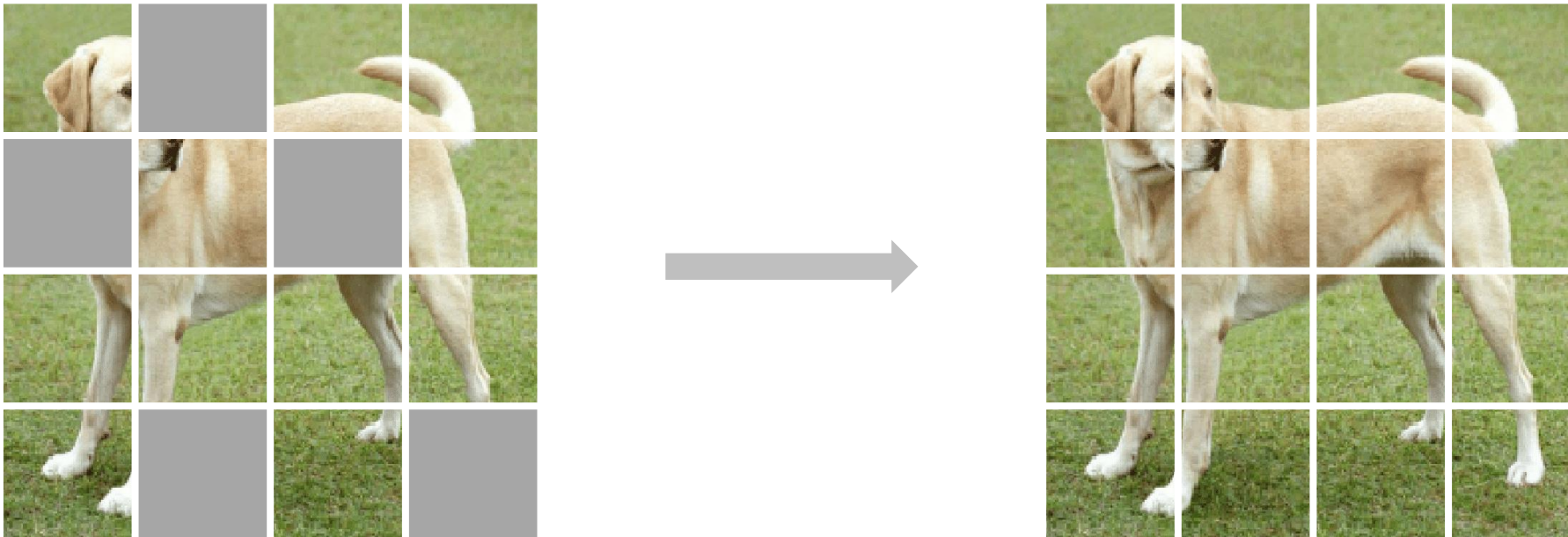
Predict the pixel RGB values or token IDs for masked patches from visible patches

Our Approach: Context Autoencoder (CAE)

CAE: Ideas

Goal in mind: Representation pretraining aims to learn an encoder, mapping an image to a representation that can be transferred to downstream tasks.

Core ideas:

❑ Make predictions for masked patches from visible patches in the encoded representation space for solving the masked image modeling task

❑ The encoder is dedicated for representation pretraining, and representation pretraining is only by the encoder

❑ The prediction part is separated from the encoder

One principle in deep learning: Operation in the latent representation space

CAE: Architecture

(a) Encoder (self-attention) - Extract the content representations for visible patches

1) Only the visible patches are fed into the encoder
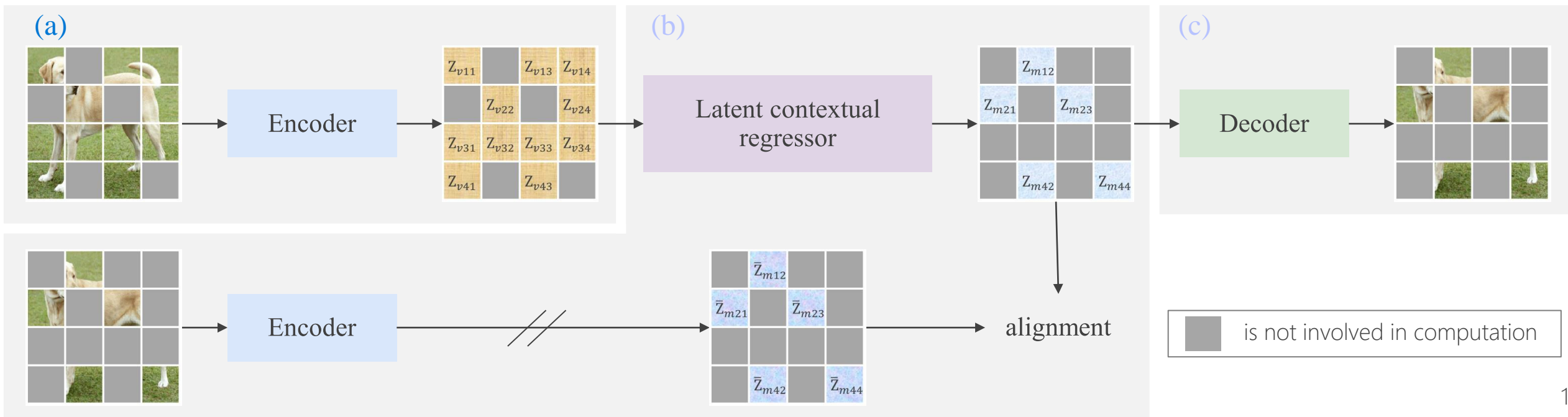
CAE: Architecture

(b) Making predictions in the encoded representation space

1) Latent contextual regressor (cross-attention) – Predict content representations for masked patches
    - ❑ The representations of the visible matches are not updated
    - ❑ The only part connecting visible patches and masked patches
2) Alignment – Make sure that the predictions are made in the encoded representation space
    - ❑ Expect that the predicted representations are close to the representations of masked patches computed from the encoder
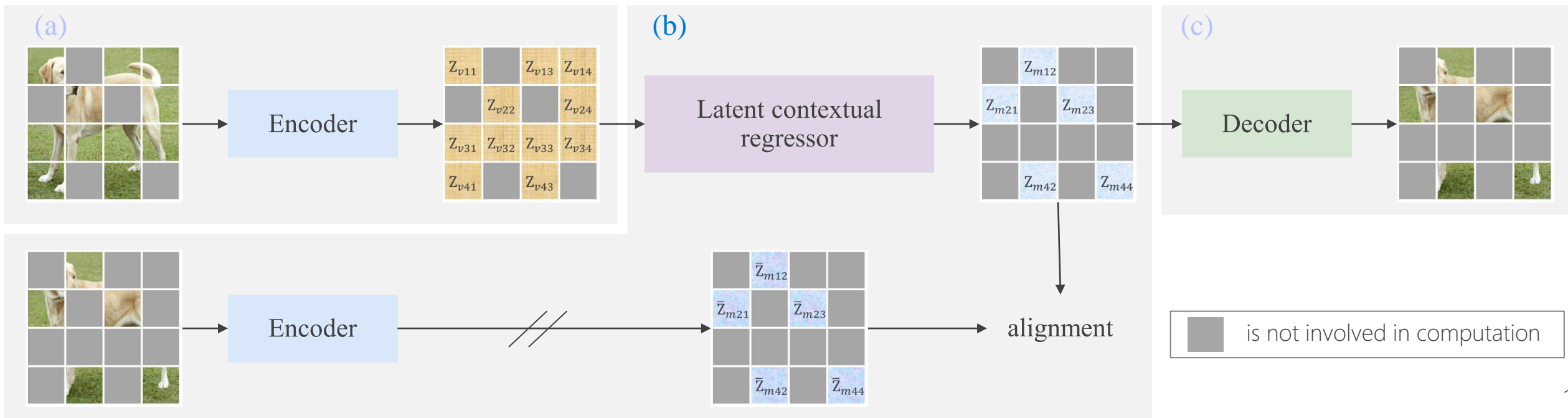
CAE: Architecture

(c) Decoder (self-attention) – Map content representations for masked patches to the targets, e.g., patch token IDs

1) Only depend on the predicted representations of masked patches

(a)

(b)

(c)

Encoder

Z_{v11} Z_{v13} Z_{v14} Z_{v22} Z_{v24} Z_{v31} Z_{v32} Z_{v33} Z_{v34} Z_{v41} Z_{v43}

Latent contextual regressor

Z_{m12} Z_{m21} Z_{m23} Z_{m42} Z_{m44}

Decoder

Encoder

$\bar{Z}_{m12}$ $\bar{Z}_{m21}$ $\bar{Z}_{m23}$ $\bar{Z}_{m42}$ $\bar{Z}_{m44}$

alignment

is not involved in computation

15
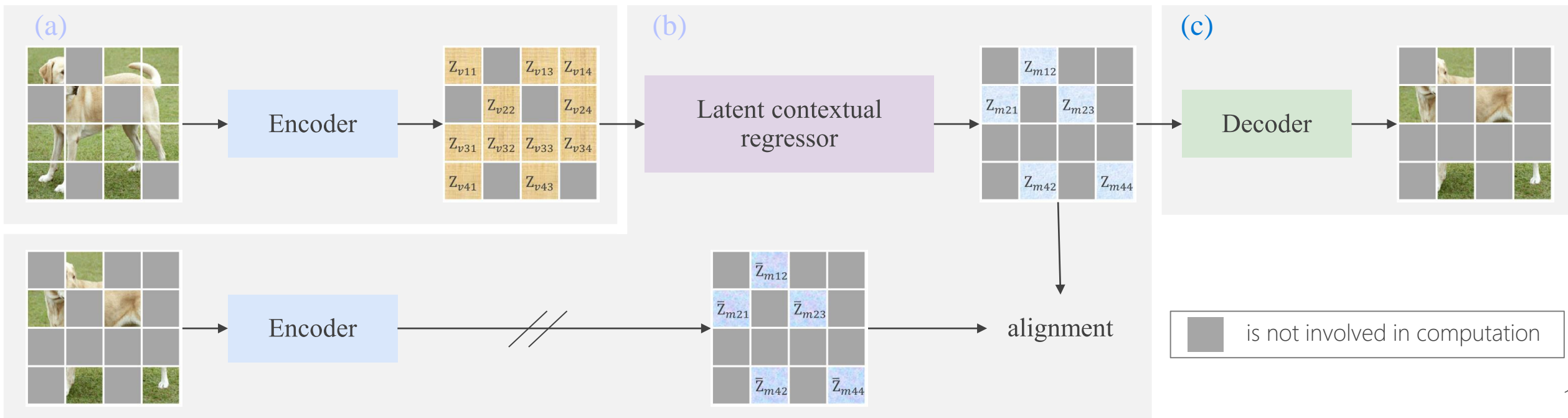
CAE: Key Points

❑ Latent contextual regressor + alignment (b) expect that the predictions are made in the latent representation space

❑ Separate the encoding (representation learning) role (a) from the MIM task completion role (b, c)

❑ Application to downstream tasks: simply replace the task completion part (b, c) with the downstream task part
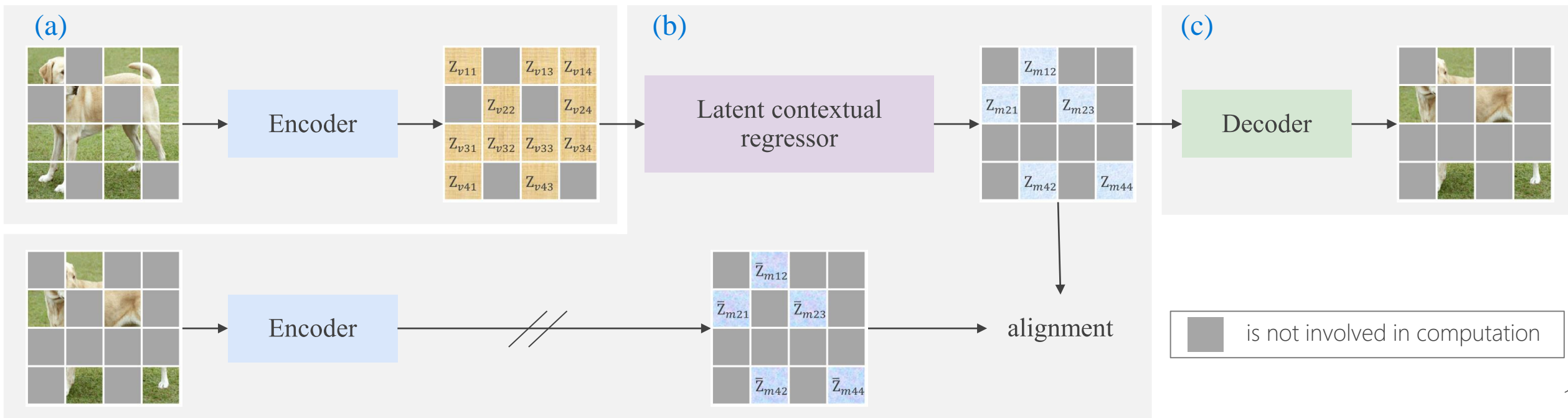
Analysis and Discussion

Illustration for Making Predictions in the Encoded Representation Space
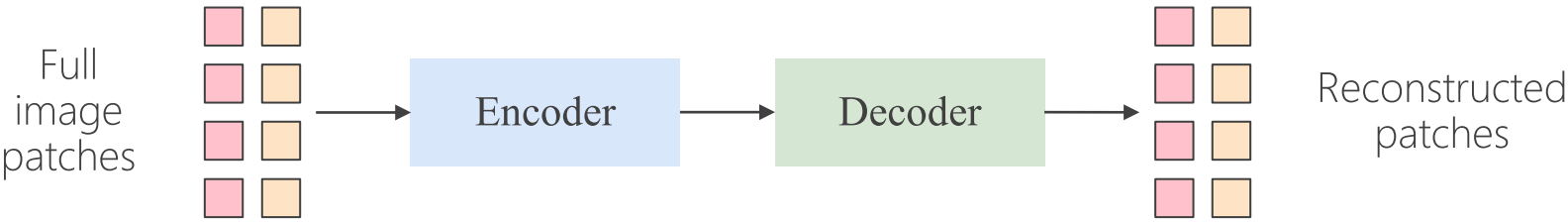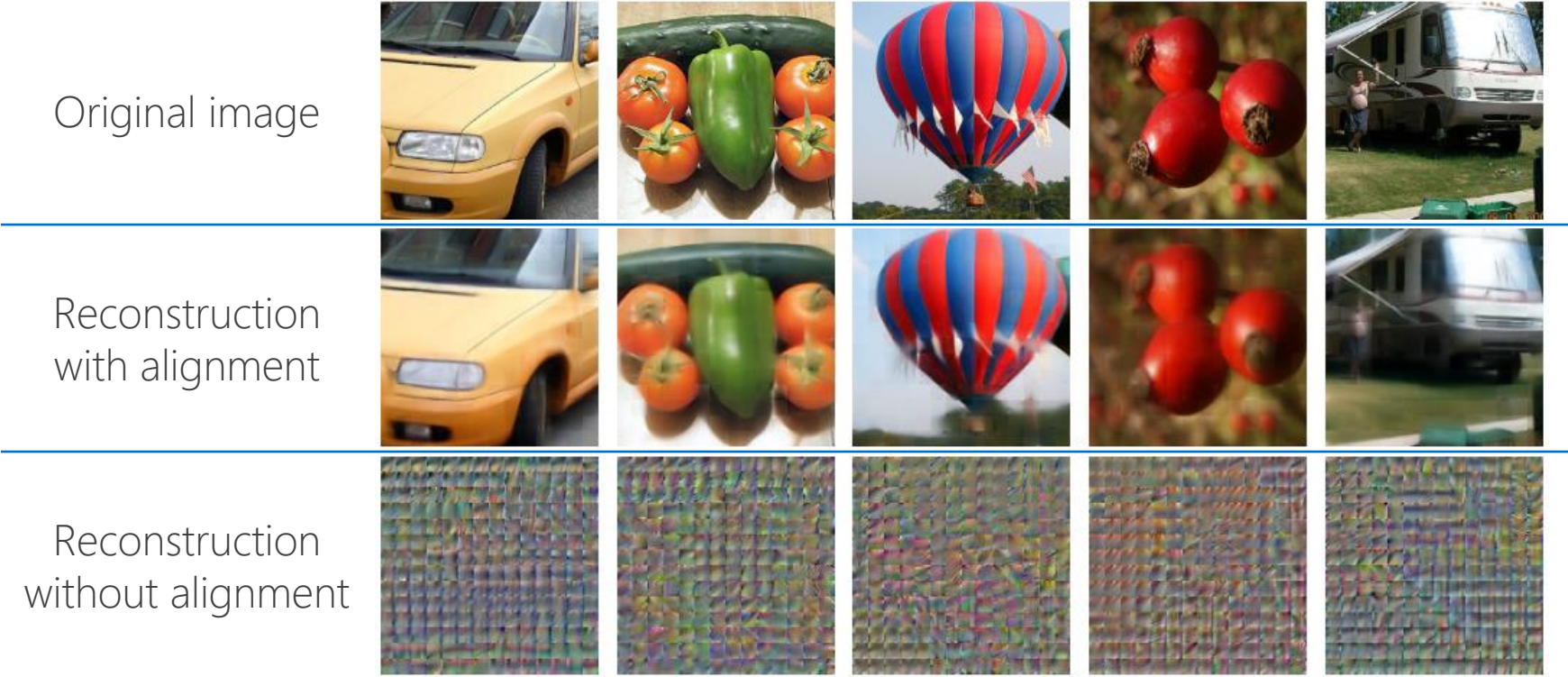
Representations are in the Semantic Space

The representations from CAE encoder are clustered for the same categories
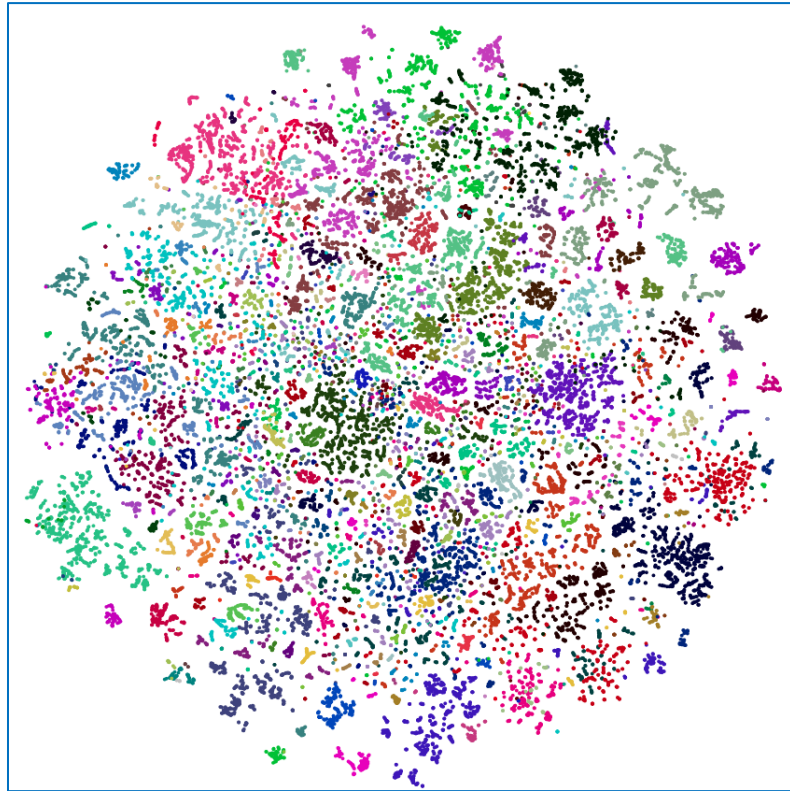
Representations from CAE encoder
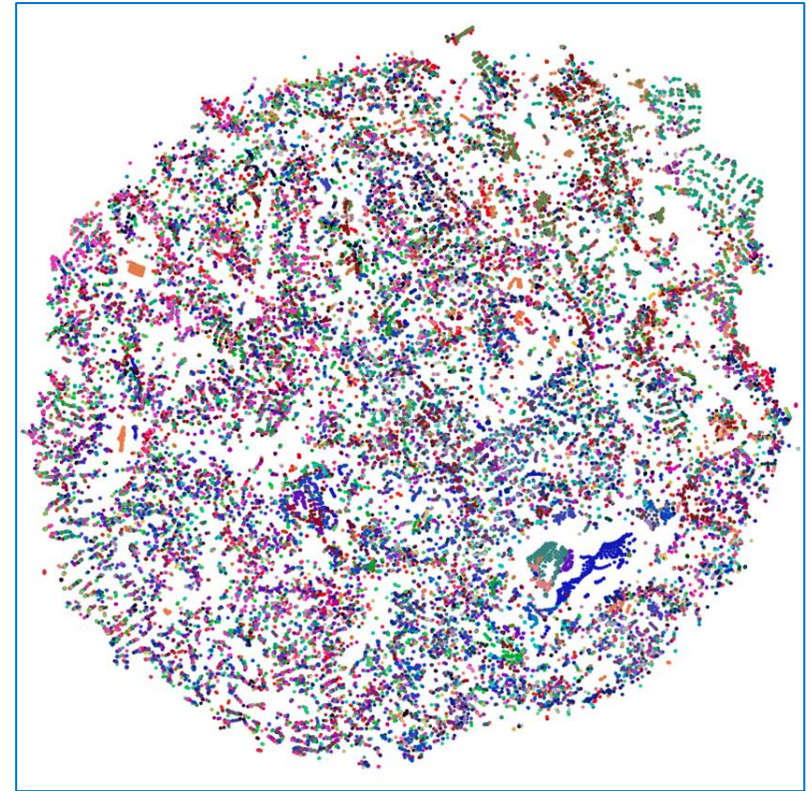
Representations from random encoder

t-SNE visualization of image patches from the ADE20K images. Different colors correspond to different categories

19

Representations are in the Semantic Space

Patch search result example: using the patch representation for similarity computation (the first patch as the query)
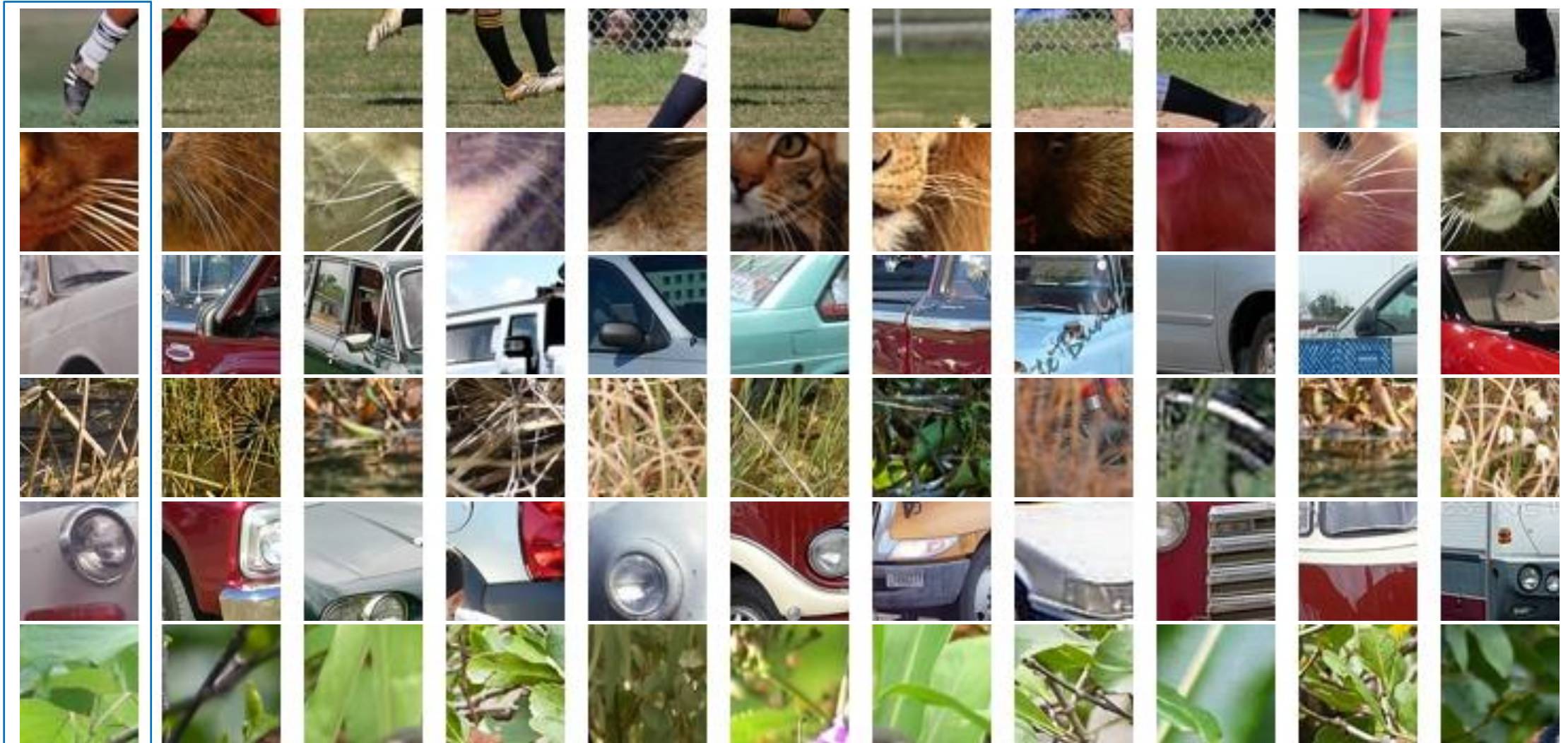
Intuitive Interpretation

Dog = a subspace of latent representations

i	Recognize the visible regions to be about a dog	Map the visible regions to latent representations
ii	Predict the regions where the other parts of the dog appear	Predict the latent representations for masked patches
iii	Guess what the other parts look like	Map the latent representations to targets, e.g. pixels
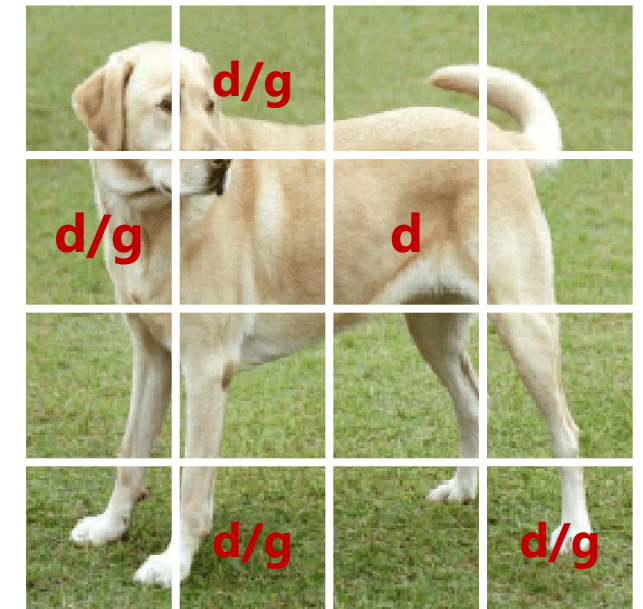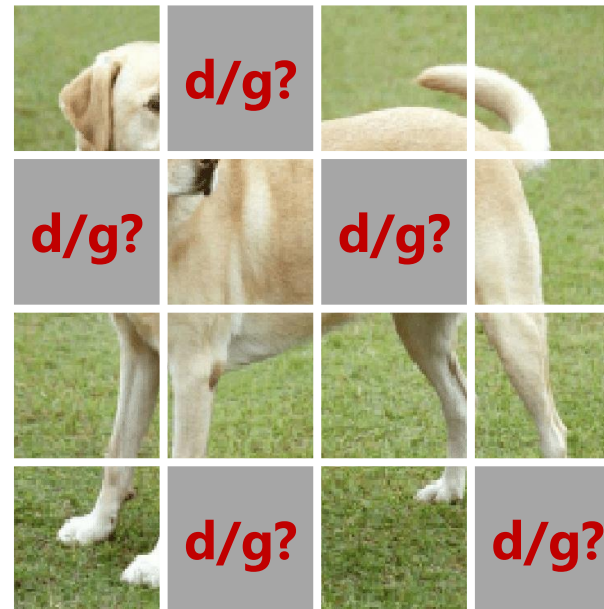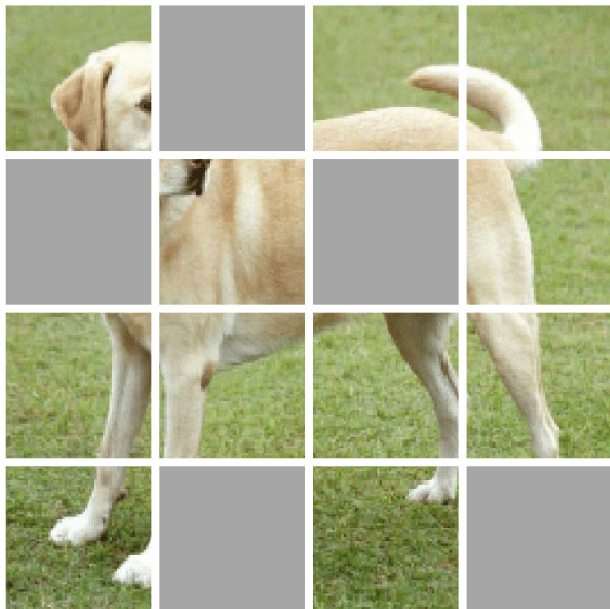
Comparison to BEiT and MAE

BEiT

(Bidirectional Encoder representation from Image Transformers)
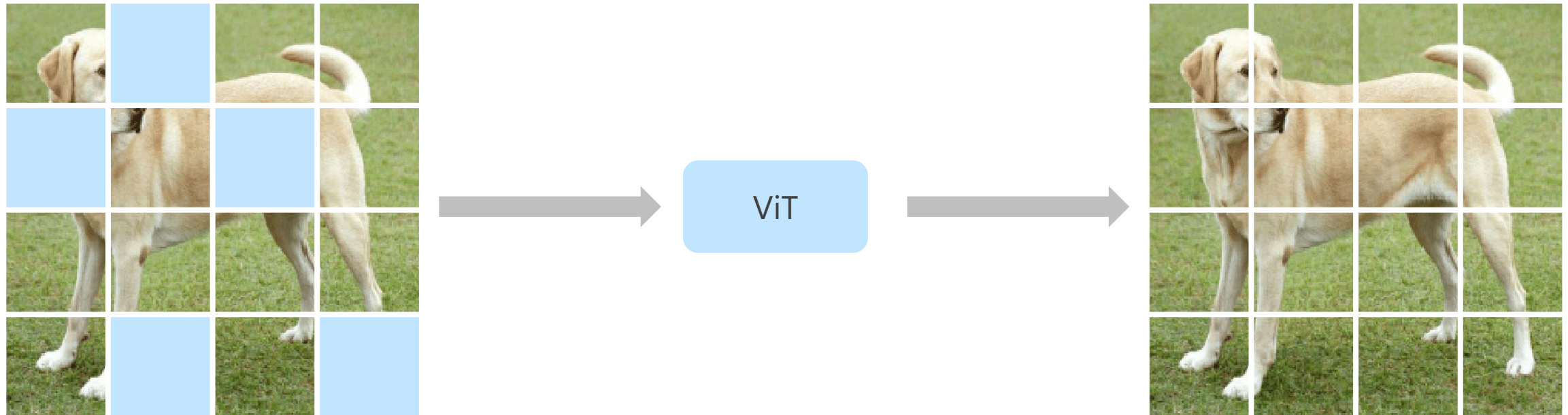
❑ All the tokens, including visible and mask tokens, are input to the ViT
❑ The ViT is responsible for both encoding and task completion roles

BEiT

(Bidirectional Encoder representation from Image Transformers)
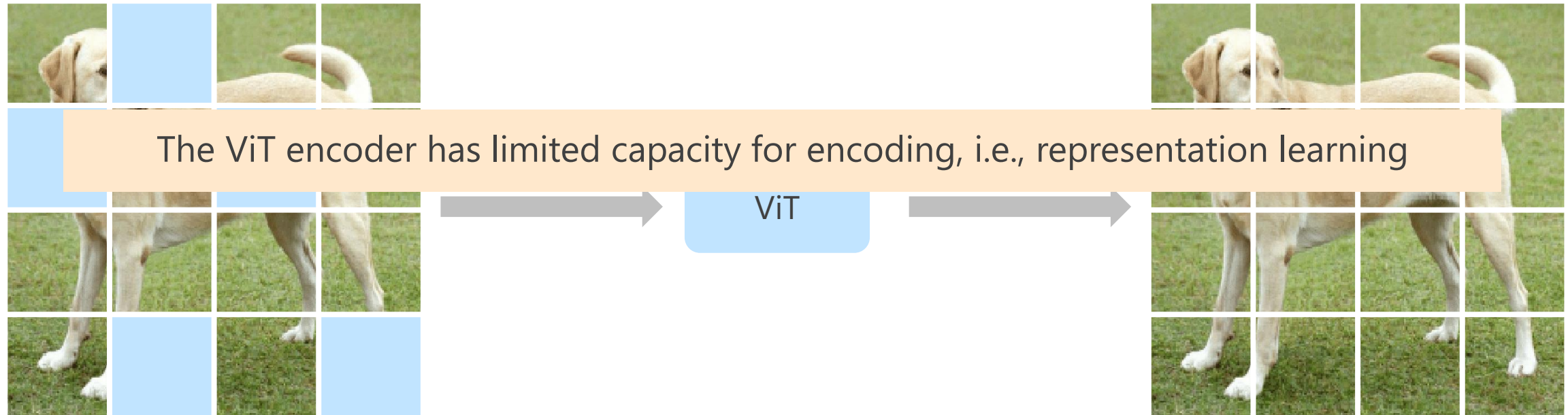
The ViT encoder has limited capacity for encoding, i.e., representation learning

ViT

❑ All the tokens, including visible and mask tokens, are input to the ViT
❑ The ViT is responsible for both encoding and task completion roles

MAE
Masked Autoencoder
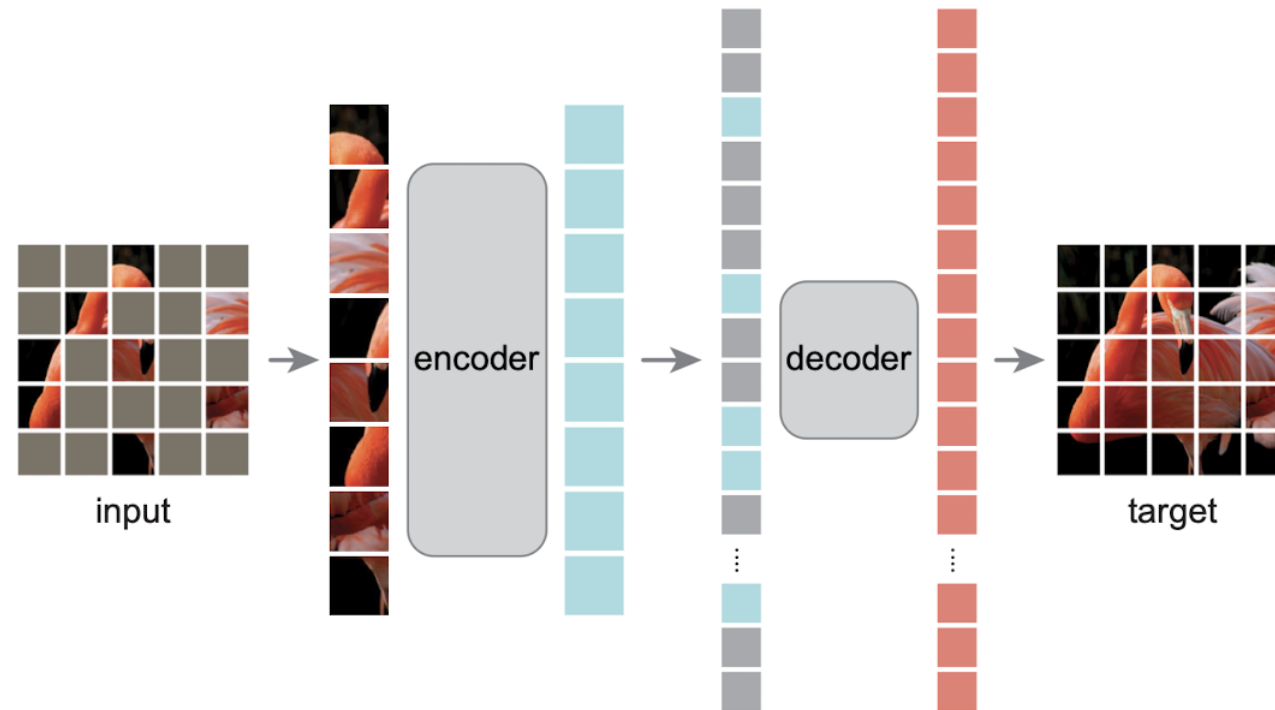
input — encoder — decoder — target

- ❑ Encoder: transformer with self-attention on visible tokens
- ❑ Decoder: transformer with self-attention on both visible and mask tokens
- ❑ The decoder takes some responsibility for encoding, i.e., representation learning

MAE
Masked Autoencoder

input

target

The encoder does not take full responsibility for representation learning

❑ Encoder: transformer with self-attention on visible tokens
❑ Decoder: transformer with self-attention on both visible and mask tokens
❑ The decoder takes some responsibility for encoding, i.e., representation learning

CAE
Context Autoencoder
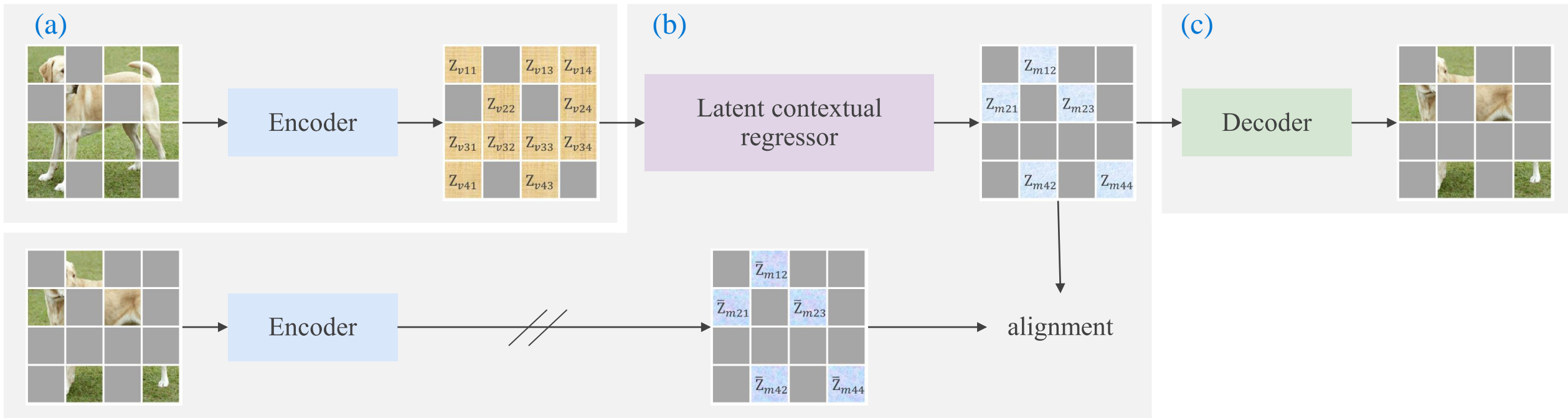
□ Latent contextual regressor + alignment (b) expect that the predictions are made in the latent representation space

□ Separate the encoding role (a) from the task completion role (b, c)

Comparing CAE, MAE and BEiT

	BEiT	MAE	CAE
Encoder and decoder separation	✖ ☹ Coupled	– 😐 Encoder only for representation learning, Decoder also partially for representation learning	✔ ☺ Designed for separation
Make predictions in the encoded representation space?	✖ ☹	– 😐	✔ ☺
Representation semantics	✖ ☹	– 😐	✔ ☺

Computational Graph Comparison
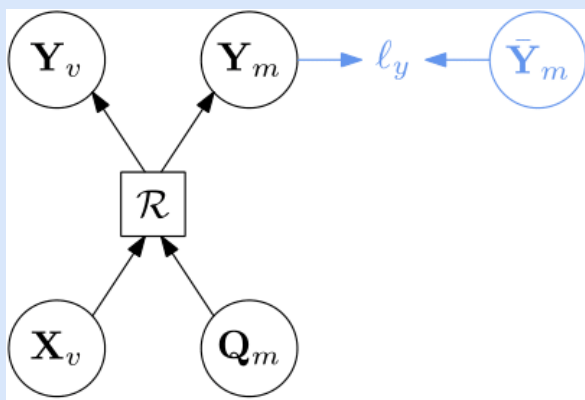
BEiT
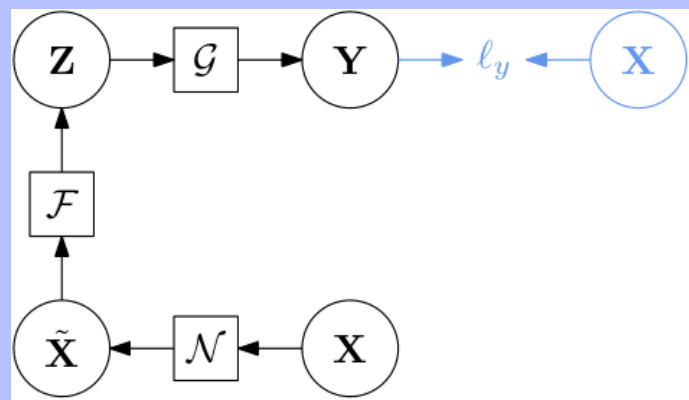
Denoising autoencoder
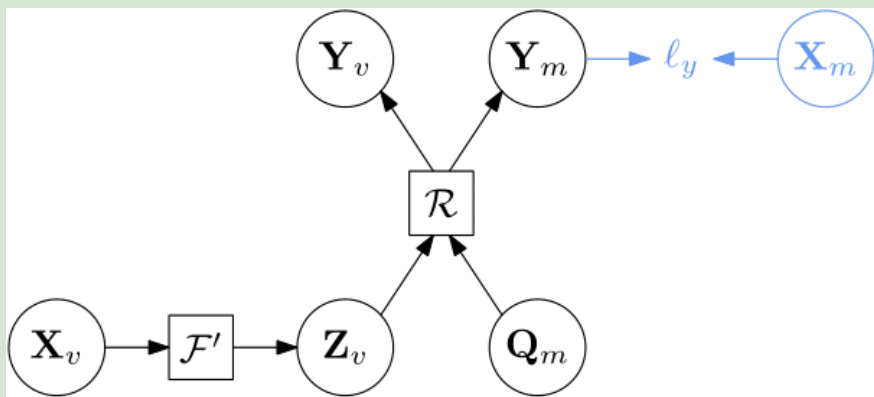
Masked autoencoder
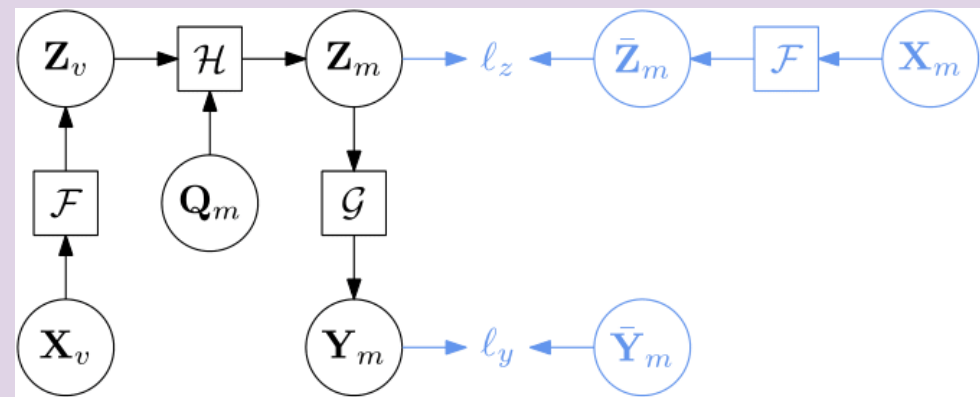
Context autoencoder

Comparison to Contrastive Pretraining

How Contrastive Pretraining Works

- ❑ **Typical contrastive pretraining methods (e.g., MoCo)** study the similarity between the representations of random crops
- ❑ **Question:** What makes the representations of random crops from the same original image be similar?
- ❑ **Observation:** The common among random crops lie in the center of the original image; The object in ImageNet lies in the center
- ❑ **Speculation:** Contrastive pretraining extracts the representation of the part of the object (encoder), then projects it to the representation of the object (projector)
- ❑ **Conclusion:** Contrastive pretraining mainly learns the semantics of the center region, i.e., 1000 categories in ImageNet-1K
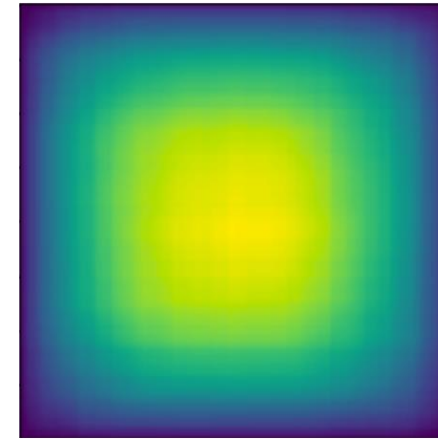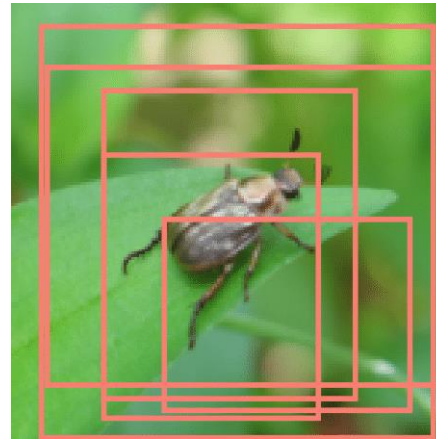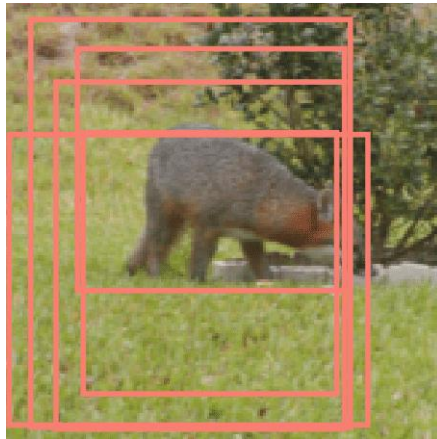
Center pixels have large probability to appear in random crops

Contrastive Pretraining Learns the Semantics of the Center Object

Contrastive Pretraining Learns the Semantics of the Center Object
The CAE (MIM) Learns the Semantics for All the Regions

Experimental Results

Representation Pretraining Evaluation

	Fine-tuning on ImageNet	Linear probing	Attentive probing
Operations	Retraining from pretrained model weights	Average-pool the patch representations or use the class-token embedding in the encoder, followed by a learned linear classifier with pretrained model weights fixed	Aggregate the patch representations by cross attention with an extra class token as the query, followed by a learned linear classifier with pretrained model weights fixed
Characteristics	No knowledge added. Pretrained weights may help optimization	Include all the patch representations	Select the proper patch representations for ImageNet classification
Pretraining evaluation?	Not suitable for pretraining evaluation	Contrastive pretraining: OK Masked image modeling: Not suitable	Contrastive pretraining: OK Masked image modeling: OK

Representation Pretraining Evaluation: Attentive Probing

Original image

Attention map
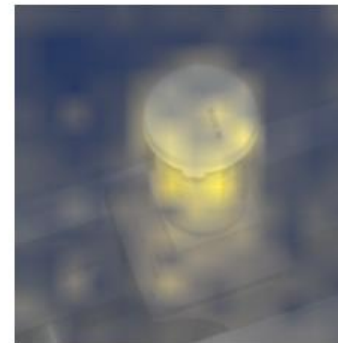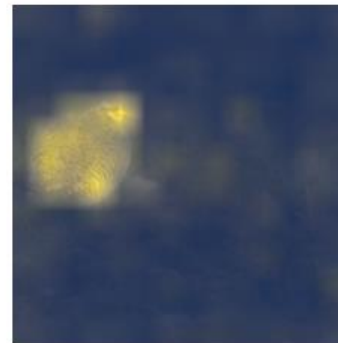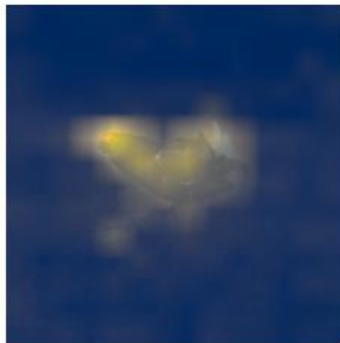
Representation Pretraining Evaluation: Results

Method	#Epochs	#Crops	Finetune	Linear prob.	Attentive prob.
Deit	300				81.8
MoCo v3	300	2	83.0	76.2	77.0
DINO	400	12	83.3	77.3	77.8
BEiT	300	1	83.0	37.6	49.4
MAE	300	1	82.9	61.5	71.1
CAE	300	1	83.6	64.1	73.8
MAE	1600	1	83.6	67.8	74.2
CAE	1600	1	83.9	70.4	77.1

❑ Contrastive pretraining: similar scores for linear and attentive probing
❑ Masked image modeling (MIM): larger score for attentive probing than linear probing
❑ MIM consider all image patches, and MIM is about the center region and thus 1K classes in ImageNet

CAE: Ablation Study

Decoder	Alignment	Attentive prob.	ADE Segmentation	COCO Detection
×	×	71.2	47.0	46.9
√	×	72.7	47.1	47.2
√	√	73.8	48.3	48.4

❏ Alignment is critical to make sure that predictions are made in the latent representation space.
❏ Latent representation is discriminative

Downstream Task: Semantic Segmentation

Method	#Epochs	Supervised	Self-Supervised	mIoU
Deit	300	√		47.0
MoCo v3	300		√	47.2
DINO	400		√	47.2
BEiT	300		√	45.5
BEiT	800		√	46.5
MAE	300		√	45.8
CAE	300		√	48.3
MAE	1600		√	48.1
CAE	1600		√	50.2
MAE (large)	1600		√	53.6
CAE (large)	1600		√	54.7

Downstream Task: Object Detection and Instance Segmentation

Method	#Epochs	Supervised	Self-Supervised	Detection	Segmentation
Deit	300	√		46.9	41.5
MoCo v3	300		√	45.5	40.5
DINO	400		√	46.8	41.5
BEiT	300		√	39.5	35.9
BEiT	800		√	42.1	37.8
MAE	300		√	45.4	40.6
CAE	300		√	48.4	42.6
MAE	1600		√	48.4	42.6
CAE	1600		√	50.0	44.0
MAE (large)	1600		√	54.0	47.1
CAE (large)	1600		√	54.5	47.6

Application to OCR Text Recognition (BCTR)

Method	Pretraining	Scene	Web	Document	Handwriting	Avg.
CRNN		53.4	54.5	97.5	46.4	67.0
ASTER		54.5	52.3	93.1	38.9	64.7
MORAN		51.8	49.9	95.8	39.7	64.3
SAR		62.5	54.3	93.8	31.4	67.3
SRN		60.1	52.3	96.7	18.0	65.0
SEED		49.6	46.3	93.7	32.1	61.2
TransOCR		63.3	62.3	96.9	53.4	72.8
Baseline (base)		68.8	70.7	98.6	49.4	75.8
+ CAE	√	72.3	73.7	99.2	62.5	79.8
MaskOCR (small)	√	71.4	72.5	98.8	55.6	78.1
MaskOCR (base)	√	73.9	74.8	99.3	63.7	80.8
MaskOCR (large)	√	76.2	76.8	99.4	67.9	82.6

Our MaskOCR: pretrain both the encoder and the decoder

Application to OCR Text Recognition (English scene texts)

Method	Pretraining	IC13	SVT	IIIT5K	IC15	SVTP	CUTE	Avg.	#Params
ASTER		91.8	89.5	93.4	76.1	78.5	79.5	86.7	-
SEED		92.8	89.6	93.8	80.0	81.4	83.6	88.3	-
PIMNet		95.2	91.2	95.2	83.5	84.3	84.4	90.5	-
SRN		95.5	91.5	94.8	82.7	85.1	87.8	90.4	55M
VisionLan		95.7	91.7	95.8	83.7	86.0	88.5	91.2	33M
ABINet	√	97.4	93.5	96.2	86.0	89.3	89.2	92.7	37M
ConCLR	√	97.7	94.3	96.5	85.4	89.3	91.3	92.8	37M
PerSec	√	97.2	94.6	96.3	84.4	89.5	90.2	92.4	-
MaskOCR (small)	√	97.1	95.7	95.5	87.0	90.1	90.3	92.9	31M
MaskOCR (base)	√	98.1	94.7	95.8	87.3	89.9	89.2	93.1	97M
MaskOCR (large)	√	97.8	94.1	96.5	88.7	90.2	92.7	93.8	315M

Our MaskOCR: pretrain both the encoder and the decoder

Summary

The core of CAE: Make predictions in the encoded representation space; Separate the encoding role and the task completion role

Masked image modeling learns all the region information, and discriminative and generative representations

Contrastive pretraining learns the center information, and is inferior to masked image modeling

Attentive probing is more suitable for pretraining evaluation. Fine-tuning on ImageNet is not suitable

Context Autoencoder for Self-Supervised Representation Learning: https://arxiv.org/abs/2202.03026
MaskOCR: Text Recognition with Masked Encoder-Decoder Pretraining: https://arxiv.org/abs/2206.00311
Code link: https://github.com/PaddlePaddle/VIMER

文心·CV大模型

https://github.com/PaddlePaddle/VIMER

多任务视觉表征学习 **VIMER-UFO** 多任务（20+） 大模型（170亿参数） 易部署（同时小模化）	**自监督视觉表征学习** **VIMER-CAE** 从无标注数据学习语义表征， 支持通用图像，文档图像等， 已落地实际应用	**文档图像表征学习** **VIMER-StrucTexT** 1.0：图像和文字多模态建模 2.0：图像建模 已落地实际应用
商品图文搜索表征学习 **VIMER-UMS** 融合学习单图，图+场景文字，图+文字描述等表征 已落地实际应用	**图文表征学习** Available soon	**视文表征学习** Available soon

行业大模型：视觉基础大模型落地行业应用

Thanks!
Q&A