Self-Supervised Learning

Josef Kittler

Centre for Vision, Speech and Signal Processing, University of Surrey, UK.

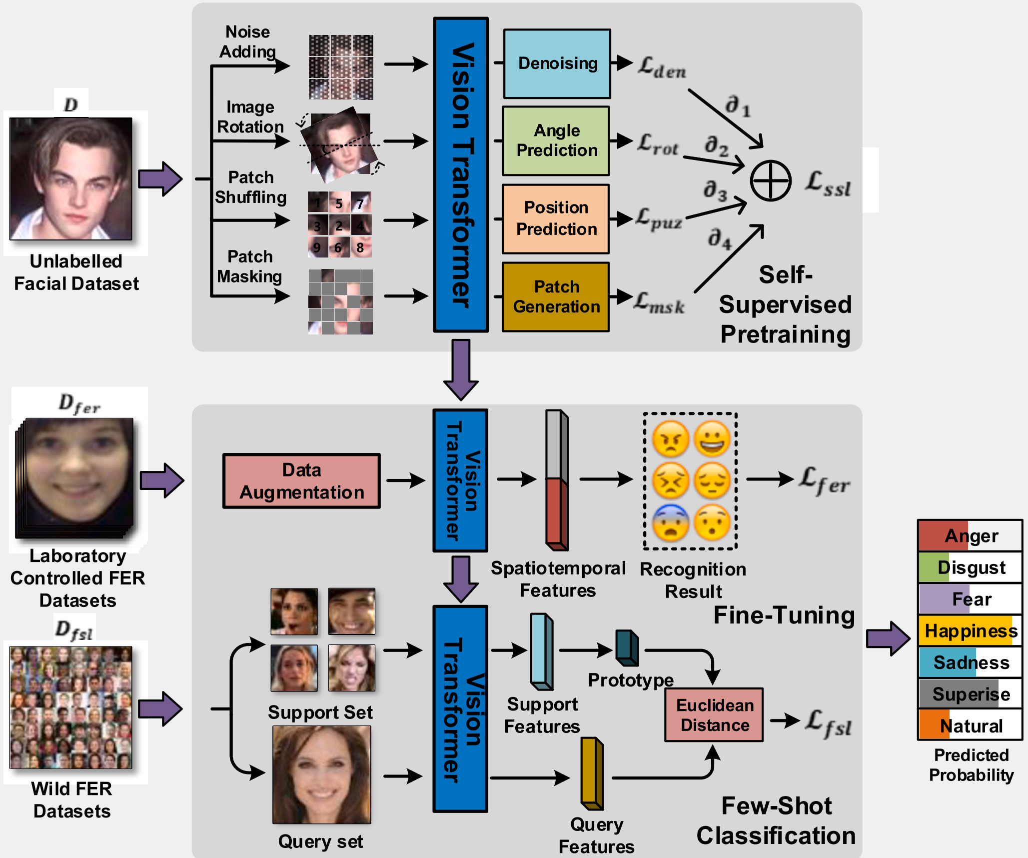
School of Artificial Intelligence and Computer Science, Jiangnan University, China.

J.Kittler@surrey.ac.uk

CVSSP Centre for Vision, Speech and Signal Processing

Motivation for SSL in biometrics

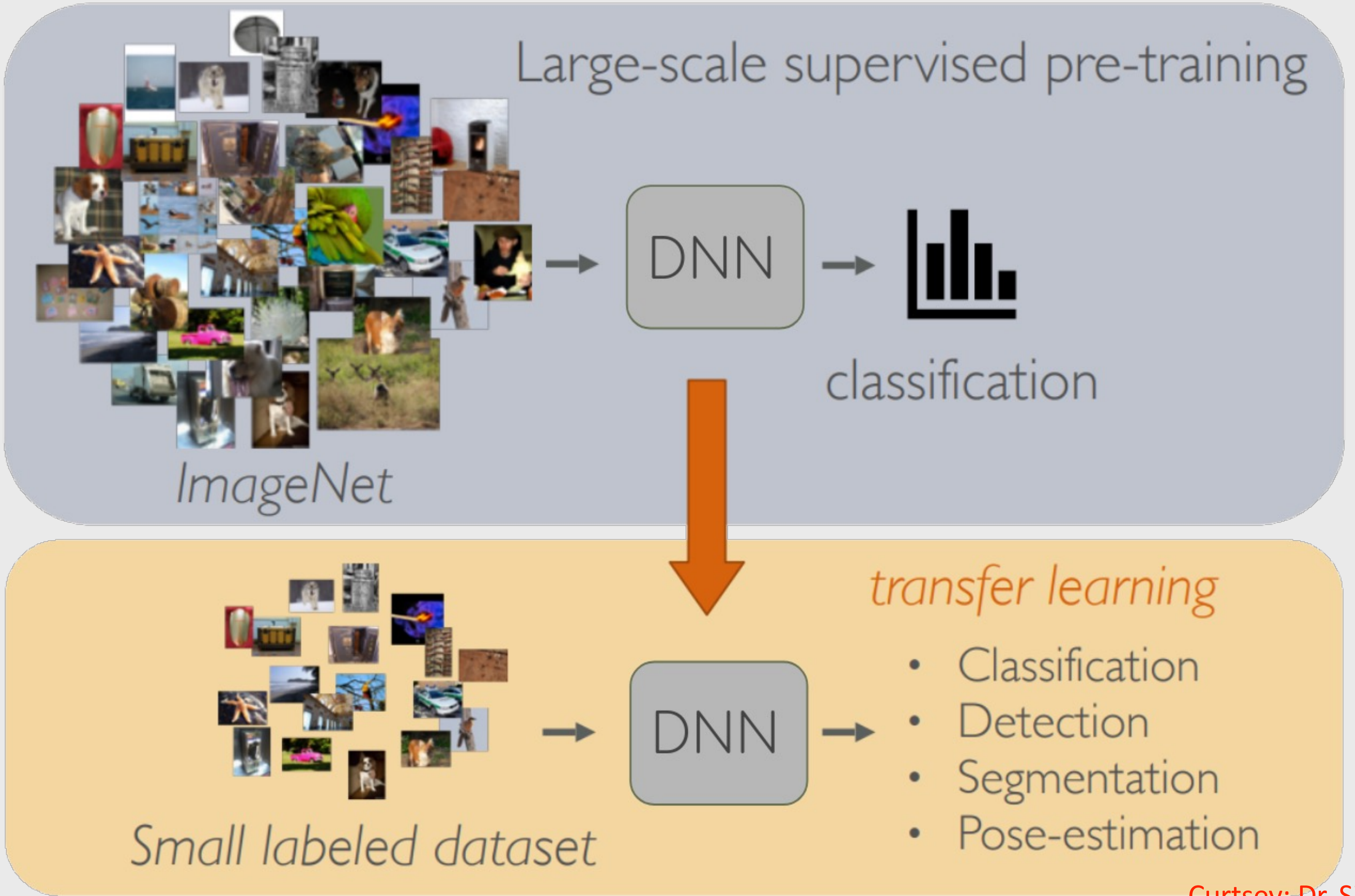
- To achieve successful designs even on small data set
  - new biometrics, such as breath
- Recent example: Chen et al, PR2023, Self-supervised vision transformer-based few-shot learning for facial expression recognition



Outline

- Introduction
- Foundation models: concept and challenges
- Self-supervised learning: the art
- Benefits of SSL
- Self-supervised learning: challenges
- Self-supervised learning: towards science
- Conclusions

Supervised Learning: A Success Story



Curtsey: Dr. Simon Jenni

Supervised Learning: The Limitations

- ImageNet model not panacea
- Expensive (cumbersome, domain experts)
- Not scalable
- Ambiguous
- Low information content
- Requires a lot of data to train
- Does not model directly image properties
- ……

"Foundation" Models

- The way forward – Self-supervised pretraining (SPP)

  - ❖ No annotation required

- Large-scale self-supervised pretrained (**SSP**) models are behind major transformation in how AI systems are built since their introduction in late 2018.

  - The foundation models emerged from natural language processing (NLP), by large pretrained models like BERT [b], GPT-3 [c] etc., based on transformer

Can the success be repeated elsewhere?

The early efforts largely unsuccessful (CNNs, Transformers)

[b] Devlin, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding reduction" ACL, 2019.
[c] Brown, et al. "Language Models are Few-Shot Learners" arXiv, 2020.

What is self supervised learning?

- Aim of learning: find minimal sufficient representation
- Types of learning
  - supervised, all data annotated
  - weakly supervised, only some data annotated, exploit e.g. temporal contiguity in video
  - unsupervised, no annotation, discover structures in the data
  - self-supervised, relates to deep networks, no annotation, but data structure discovery is meaningless
- Aim of self-supervised (SSL) learning - pre-train a network so that it can subsequently be fine-tuned for a specific task on small quantity of data

What should SSL achieve?

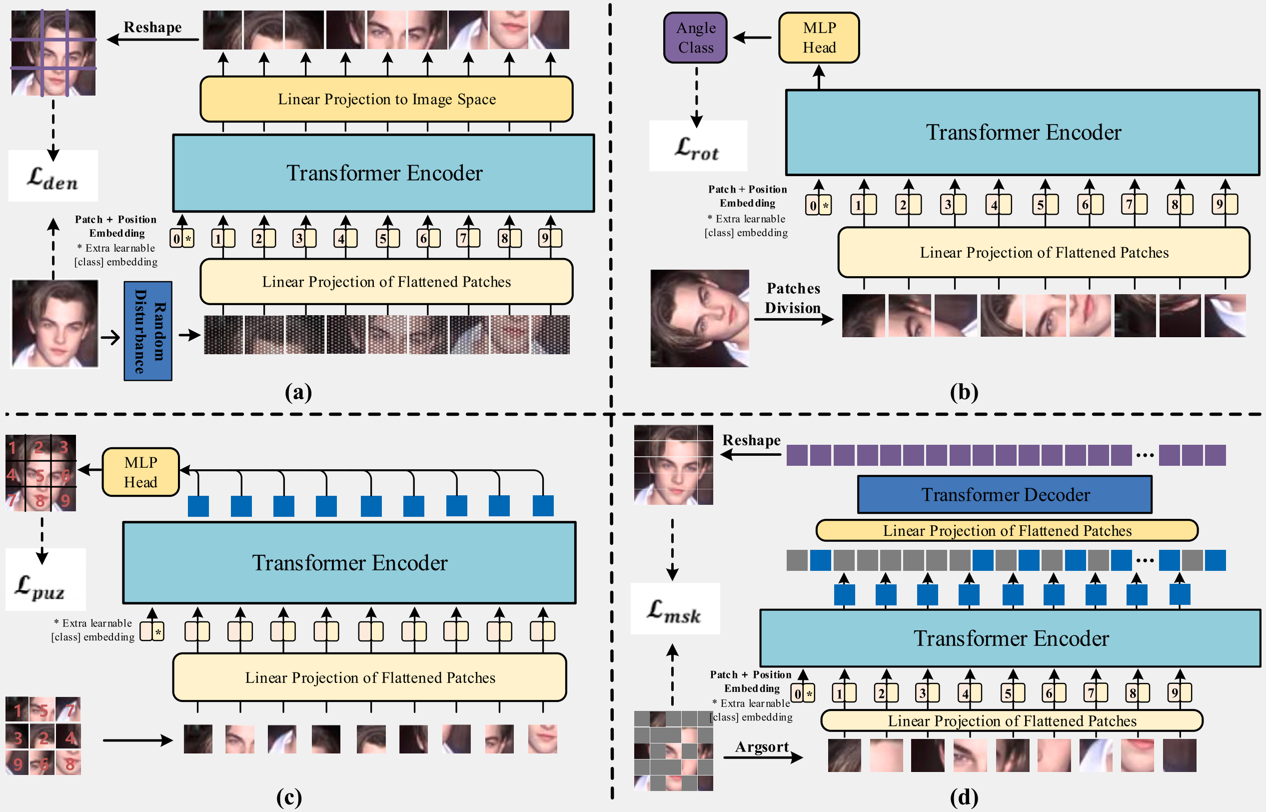
- Learn about
  - local image properties
  - notions of similarity and dissimilarity
  - the existence of different concepts
  - basic properties of concepts (text, shape, contiguity)
  - the diversity of concept manifestations
  - context
  - relationship of different modalities
- Robustness to variations and resilience to noise
- Eliminate redundancy

Key ingredients of SSL

- Pre-text (auxiliary) tasks
  - their accomplishment should endow the network with the ability to generalize to the target tasks
- Data augmentation
  - generation of training data in support of the pre-text tasks
- Auxiliary network architecture
  - e.g. siamese twin
- Loss functions
- Training strategy
- Optimisation procedures

Example pre-text tasks

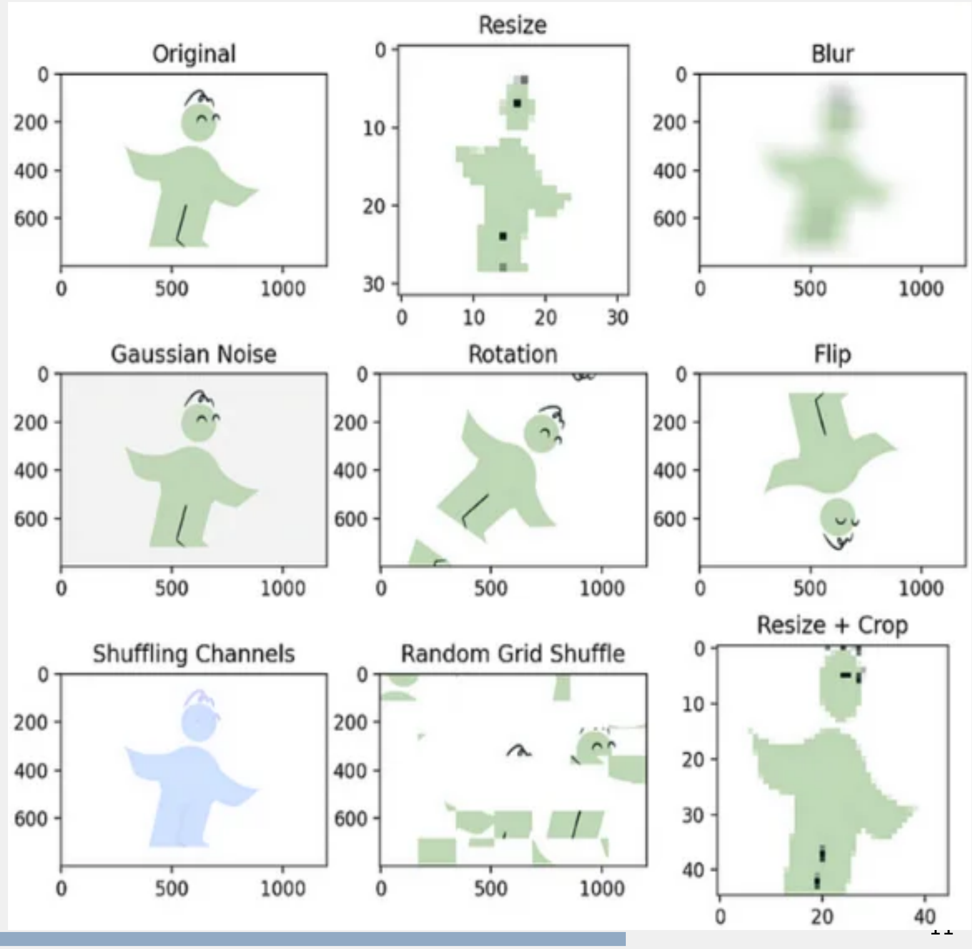
- Popular pre-text tasks
  - Reconstruction
  - Rotation classification
  - Juxtaposition
  - Masking



From Chen et al (PR2023)

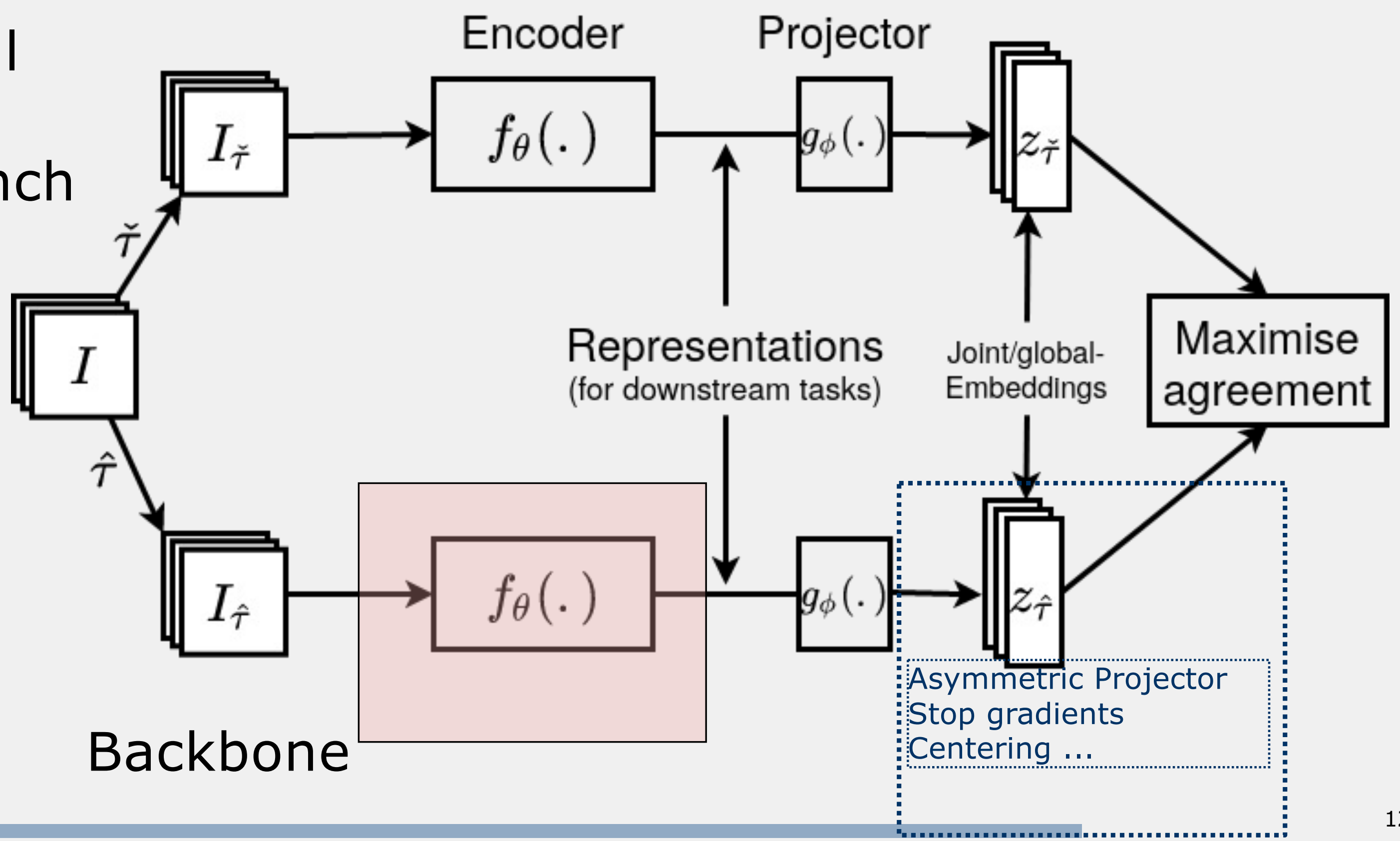
Data augmentation

- Geometric transformations
- Puzzle (random grid shuffle)
- Filtering (blurring)
- Compression
- Training set balancing
- Multiple views (global, local)
- Adding noise
- Physics based transformations (hazy)
- Rotation by a given angle
- Colour modification
- Transformation to a monochromatic image
- Masking and cut out



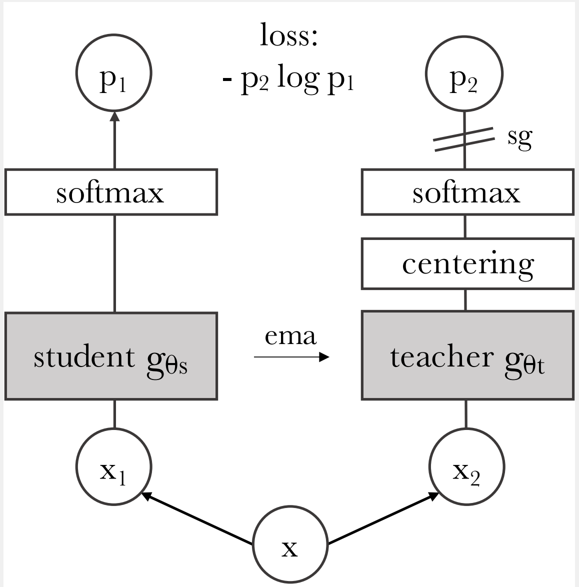
Architectures for SSL

- Backbone architecture
- Addition structural components, e.g. siamese twin branch

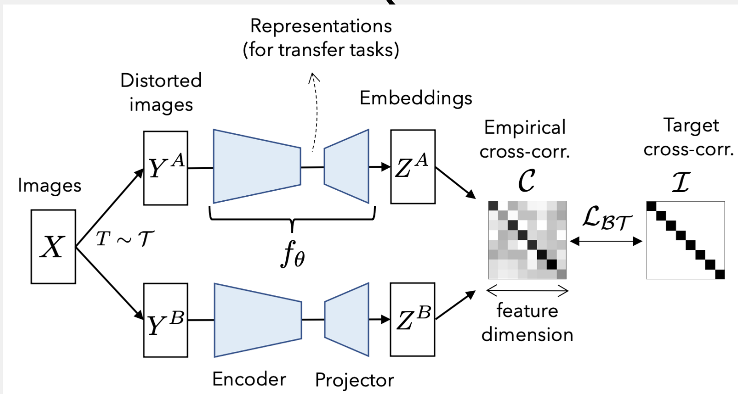


Architectures

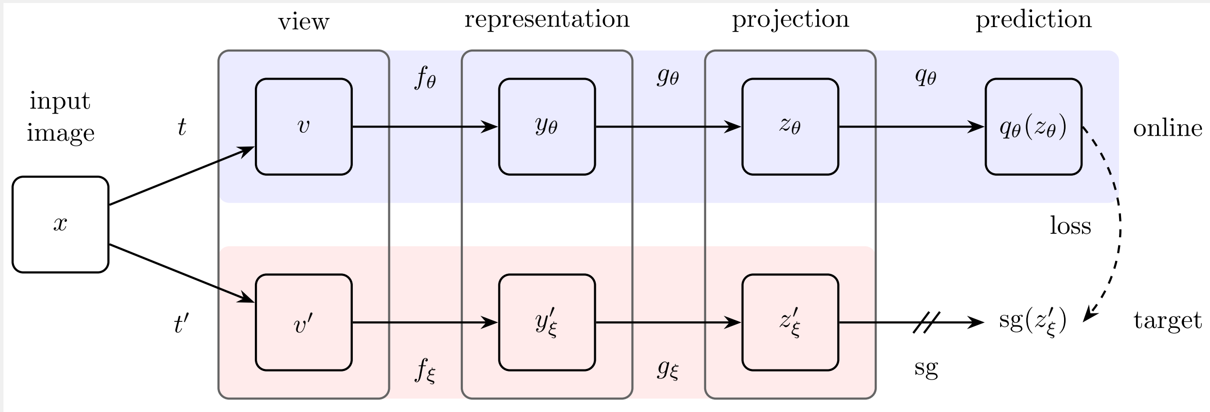
- Dino (Caron, 2021)



- Barlow twins (Zbontar 2021)



- Byol (Grill, 2020)



Sample loss functions

- Contrastive loss (SimCLR, van den Oord,2018)

$$\text{InfoNCE}_\theta^{\alpha,\beta} \triangleq \frac{2}{B}\sum_{i=1}^{B} S_\theta(v_i, v_i') - \frac{2\alpha\cdot\beta}{B}\sum_{i=1}^{B}\ln\left(\sum_{j\neq i}\exp\frac{S_\theta(v_i,v_j)}{\alpha} + \sum_j \exp\frac{S_\theta(v_i,v_j')}{\alpha}\right),$$

- where $S_\theta(v_i, v_i')$ represents similarity between two views (embedding, representation, or prediction)

$$S_\theta(u_1,u_2) \triangleq \frac{\langle\phi(u_1),\psi(u_2)\rangle}{\|\phi(u_1)\|_2\cdot\|\psi(u_2)\|_2}.$$

- BYOL loss $\beta = 0$

- Reconstruction loss

$$reconstr = ||v_i - \tilde{v}_i(v_i')||_1$$

Current research in SSL

Heuristic notions in SSL

- Distillation
- Asymmetry
  - Architectural (projector, predictor)
  - Architecture parameters (momentum update)
  - Variance
  - Batch normalization
  - Temperature difference
- Principle of SSL (clustering, Barlow's redundancy reduction, info bottleneck, information theory)
- Architecture
- Batch size and normalization
- Augmentation methods/masking
- Loss functions – what, where, how
- Positive sample only versus positive/negative sample learning

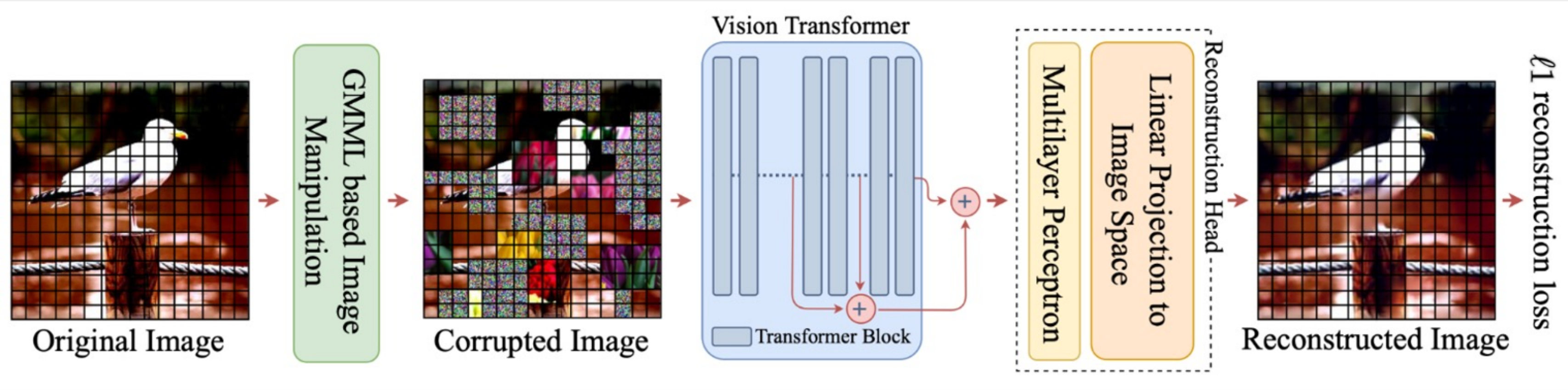
Selfsupervised learning

Previously reported SiT and GMML selfsupervised methods

- Image reconstruction
- Extensive masking (70%)
- Group masking
- Significant improvements on small data sets
- Better domain transfer

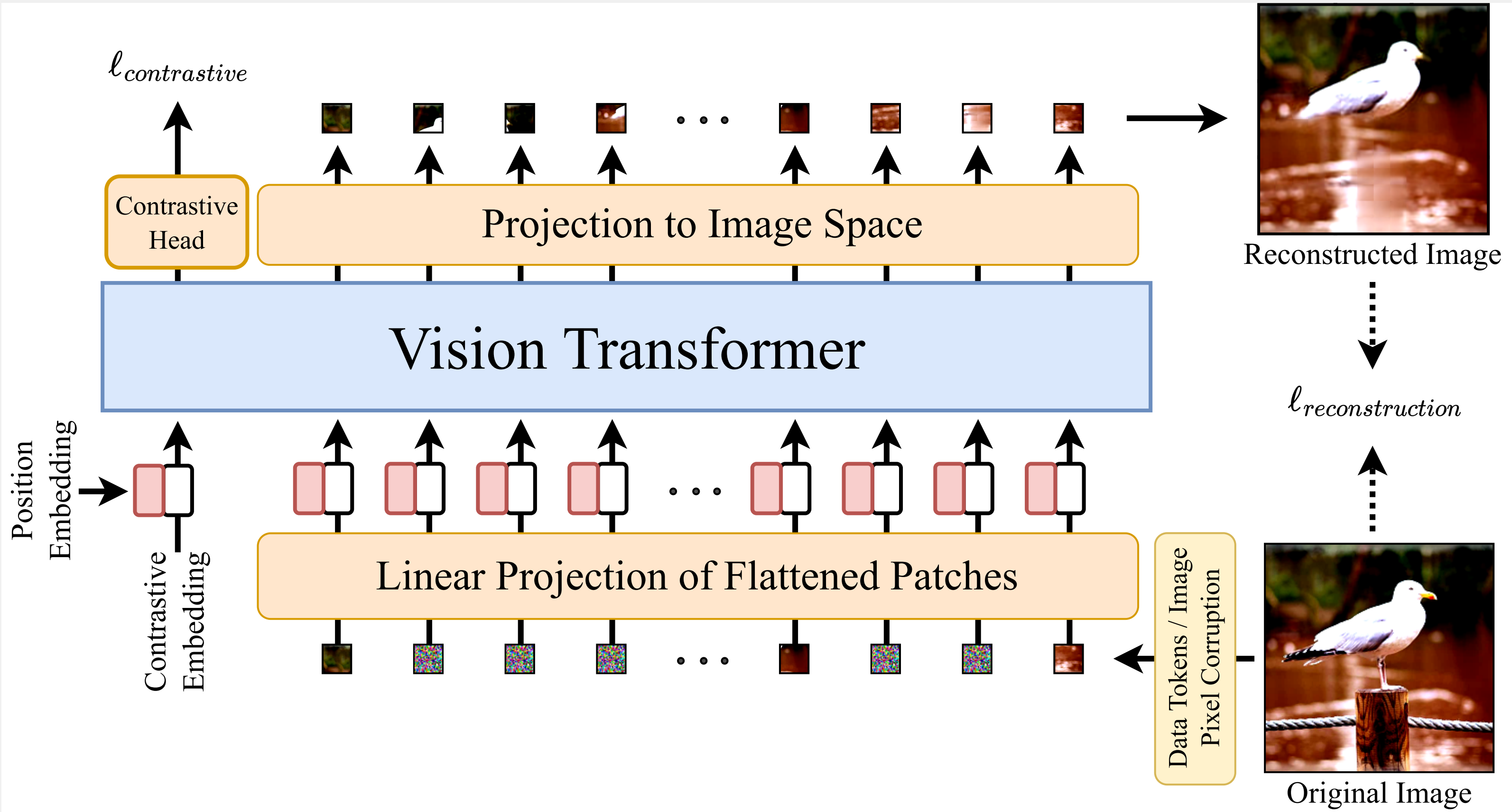
Applications

- Audio classification
- Knee x-ray classification



SiT: Selfsupervised image transformer

- Loss
  - Reconstruction
  - Contrastive



Magic of self-supervised pretraining

- Masking forces learning
  - image properties
  - image context
  - robustness to occlusion
- Augmentation
  - Robustness to transformations
  - Increases training data size
  - Enhances data diversity
  - Robustness to scale

SSL in Domain Transfer

TABLE 3: Domain Transfer. Fine-tuning self-supervised pretrained models on different datasets employing ViT-T variant of transformers.

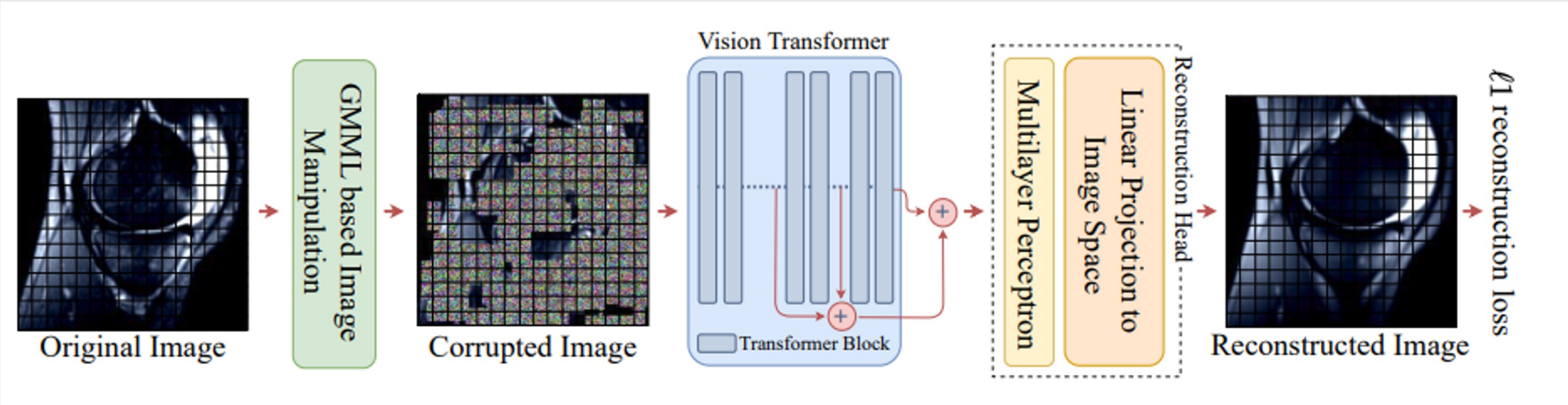
Pretraining	Fine-tuning					
	MNIST	Flowers	Pets	CUB	Aircraft	Cars
random init.	–	58.1	31.8	23.8	14.6	12.3
Transfer learning from toy dataset.						
MNIST	99.6	74.8	67.9	52.3	57.2	70.2
Transfer learning from small datasets.						
Flowers	99.6	90.6	78.7	61.8	67.4	80.2
Pets	99.5	88.8	86.0	61.7	69.1	82.7
CUB	99.5	89.1	84.8	71.2	77.79	88.7
Aircraft	99.5	89.2	84.4	68.7	85.1	89.7
Cars	99.6	89.2	85.7	69.4	81.1	92.7

TABLE 4: Domain Transfer. Fine-tuning self-supervised pretrained models on different datasets employing ViT-S.

Pretraining	Fine-tuning				
	Flowers	Pets	CUB	Aircraft	Cars
Transfer learning from toy dataset.					
MNIST	77.7	61.5	41.8	48.1	48.4
Transfer learning from small datasets.					
Flowers	94.7	84.4	67.7	74.9	89.3
Pets	92.5	88.1	70.9	78.0	89.7
CUB	92.2	84.4	73.4	78.9	90.7
Aircraft	90.5	82.5	69.8	85.1	90.9
Cars	92.6	86.9	71.1	83.7	93.3

Domain Transfer is so strong that we use transfer from toy MNIST dataset
Even MNIST pretraining outperformed supervised pretraining of ViTs with large margin

GMML in Medical Domain



Atito, Sara, Muhammad Awais, and Josef Kittler. "**Sit: Self-supervised vision transformer.**" *arXiv preprint arXiv:2104.03602* (2021).

Atito, Sara, Muhammad Awais, and Josef Kittler. "**GMML is All you Need.**" *arXiv preprint arXiv:2205.14986* (2022).

Sara Atito, Syed Muhammad Anwar, Muhammad Awais, and Josef Kittler. "SB-SSL: Slice-Based Self-Supervised Transformers for Knee Abnormality Classification from MRI", MICCAI MILLanD, 2022

GMML in Medical Domain

Table 1: Comparison with SOTA on ACL tears classification employing sagittal plane.

Method	Backbone	# params	ACL Tear (Sagittal plane) Accuracy (%)	ACL Tear (Sagittal plane) AUC
Training using only the given dataset				
Random Init	CNN	77M	71.67	0.754
Random Init	ViT-S	21M	70.00	0.721
[20]	CNN	77M	76.62	0.848
[20] + noise	CNN	77M	75.83	0.817
SB-SSL (Ours)	ViT-T	5M	85.83	0.952
SB-SSL (Ours)	ViT-S	21M	88.33	0.954
SB-SSL (Ours)	ViT-B	86M	89.17	0.954
Transfer learning from ImageNet-1K dataset				
MRNet [6]	AlexNet	61M	86.63	0.963

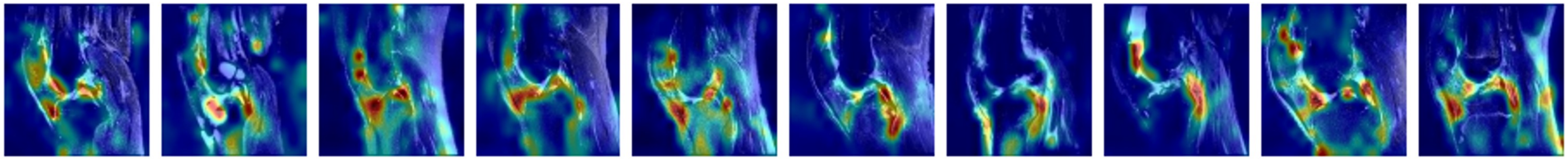
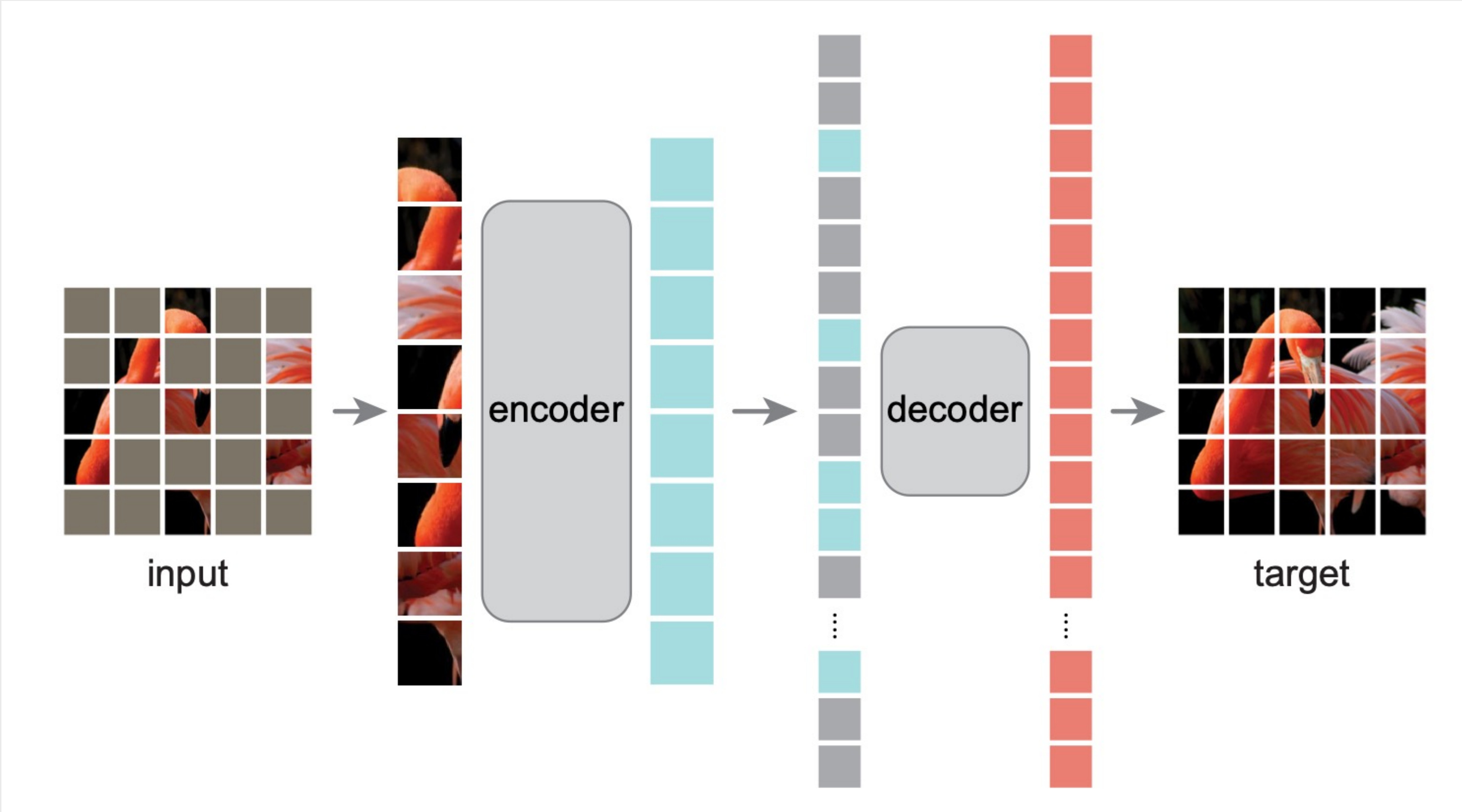


Fig. 5: Self-attention visualizations from the ViT-S model finetuned on the ACL tears task employing the sagittal plane.

Masked Autoencoder



- MAE

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF Conference on CVPR. 2022.

MAE on Small Dataset

Pre-training and finetuning MAE on small datasets

- Employing the official publicly available code of MAE
- Model: ViT-Small
- GMML is trained for 3000 epochs
- MAE is trained for 6000 epochs

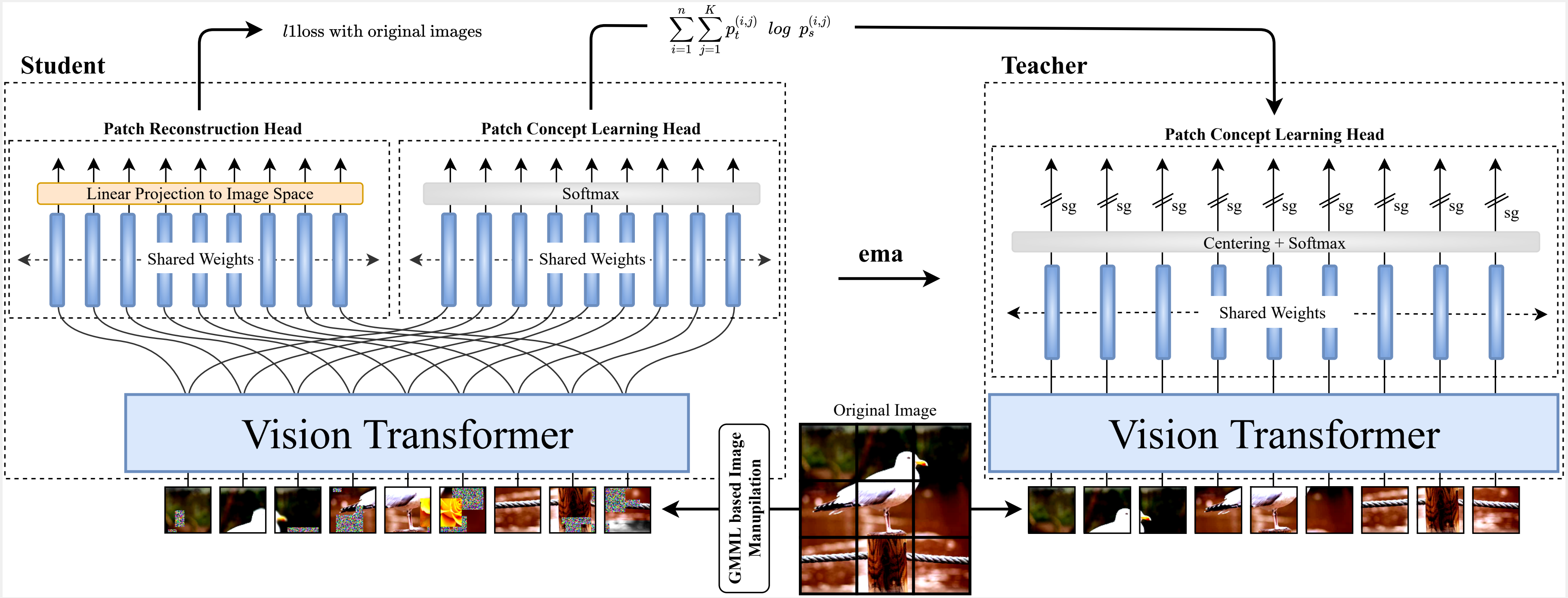
Method	Flowers	Pets	CUB	Aircraft	Cars
MAE	86.87	73.01	59.35	69.03	91.03
GMML	94.52 (↑ 7.65)	88.09 (↑ 15.08)	77.44 (↑ 18.09)	84.52 (↑ 15.49)	93.10 (↑ 2.07)

Multi-concept self-supervised learning

- Images typically contain multiple objects
- Yet, supervised/self-supervised methods assume there is a single dominant class
- This inconsistency makes the learning problem challenging
- We extended our pioneering selfsupervised learning methods SiT and GMML to the multi-concept SSL case

Architecture for multi-concept SSL

- Masked image reconstruction
- Clustering of patch embeddings

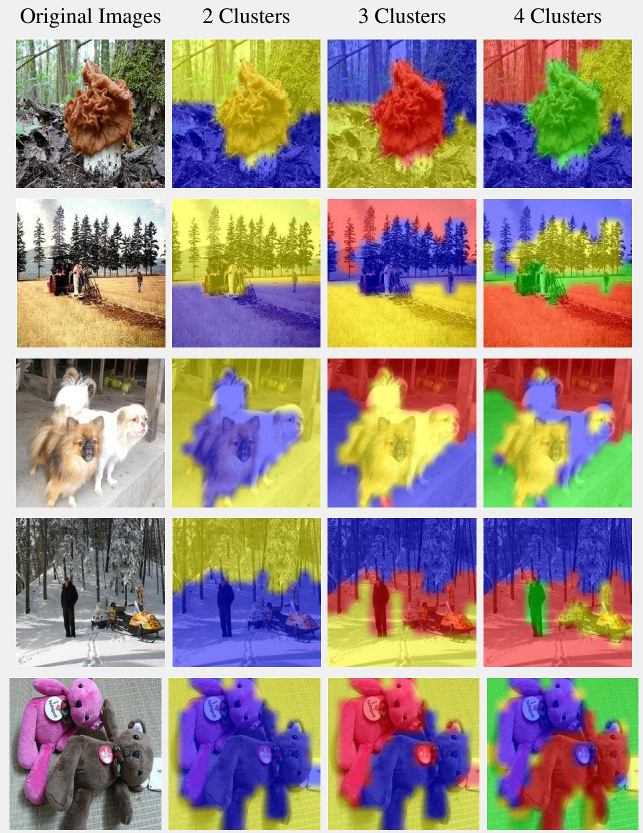


Atito, Sara, et al. **MC-SSL0. 0: Towards Multi-Concept Self-Supervised Learning**. *arXiv:2111.15340*.

Experimental results

MS-COCO dataset

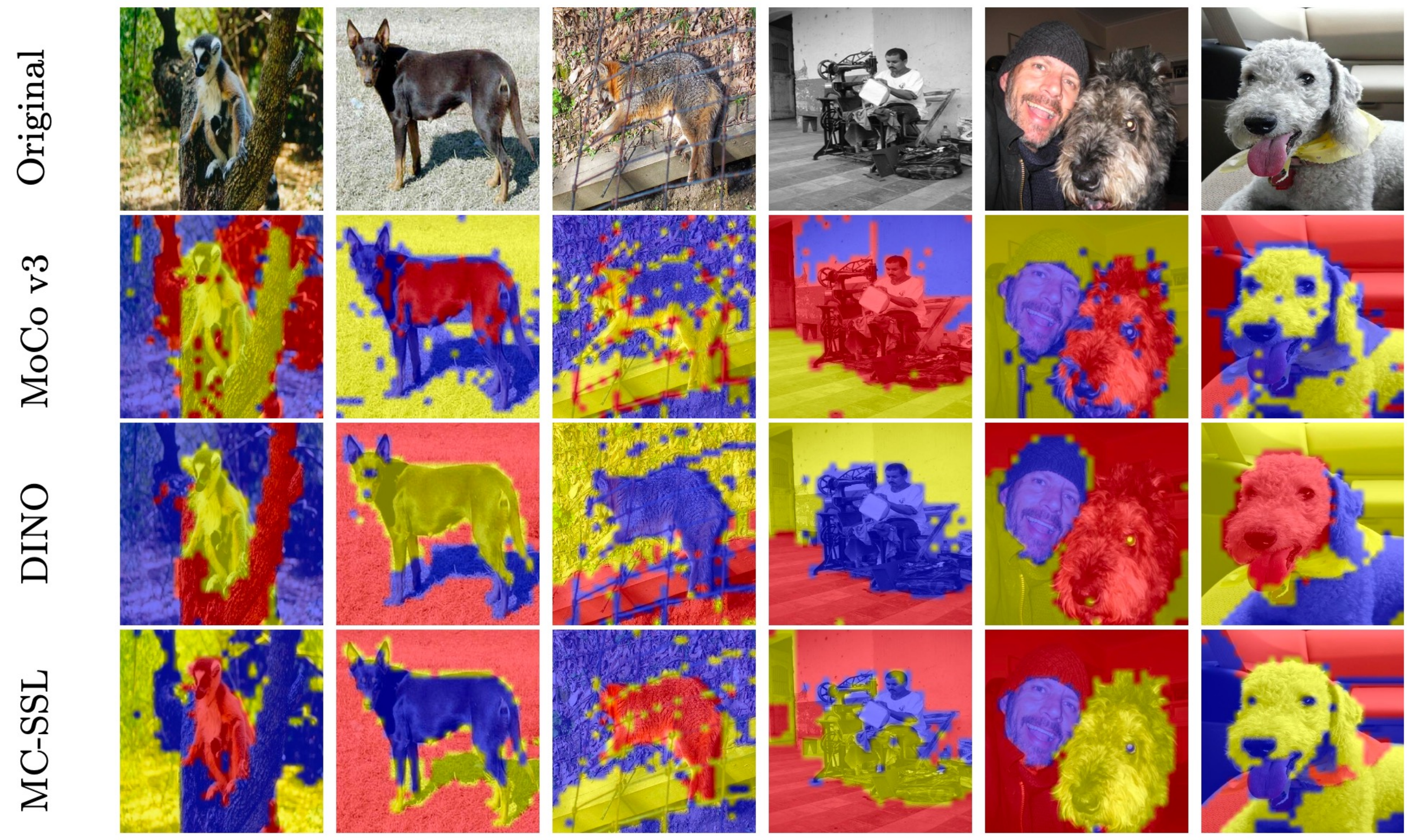
From scratch (i.e., random initialization)							
ViT-S/16*	44.7	32.7	58.7	42.0	37.1	67.9	48.0
Selfsupervised pretraining on MS-COCO							
MC-SSL0.0	73.1	56.2	75.2	64.3	58.6	80.1	67.7
Selfsupervised pretraining on 10% of ImageNet-1K							
Dino*	63.4	50.8	66.5	57.6	54.0	73.1	62.1
MC-SSL0.0	70.5	54.8	74.0	63.0	56.3	79.1	65.8
Selfsupervised pretraining on 10% ImageNet-1K with multi-crop							
Dino‡	69.0	56.0	70.1	62.2	59.4	75.4	66.5
MC-SSL0.0‡	72.7	56.8	74.1	64.3	59.6	79.0	67.9



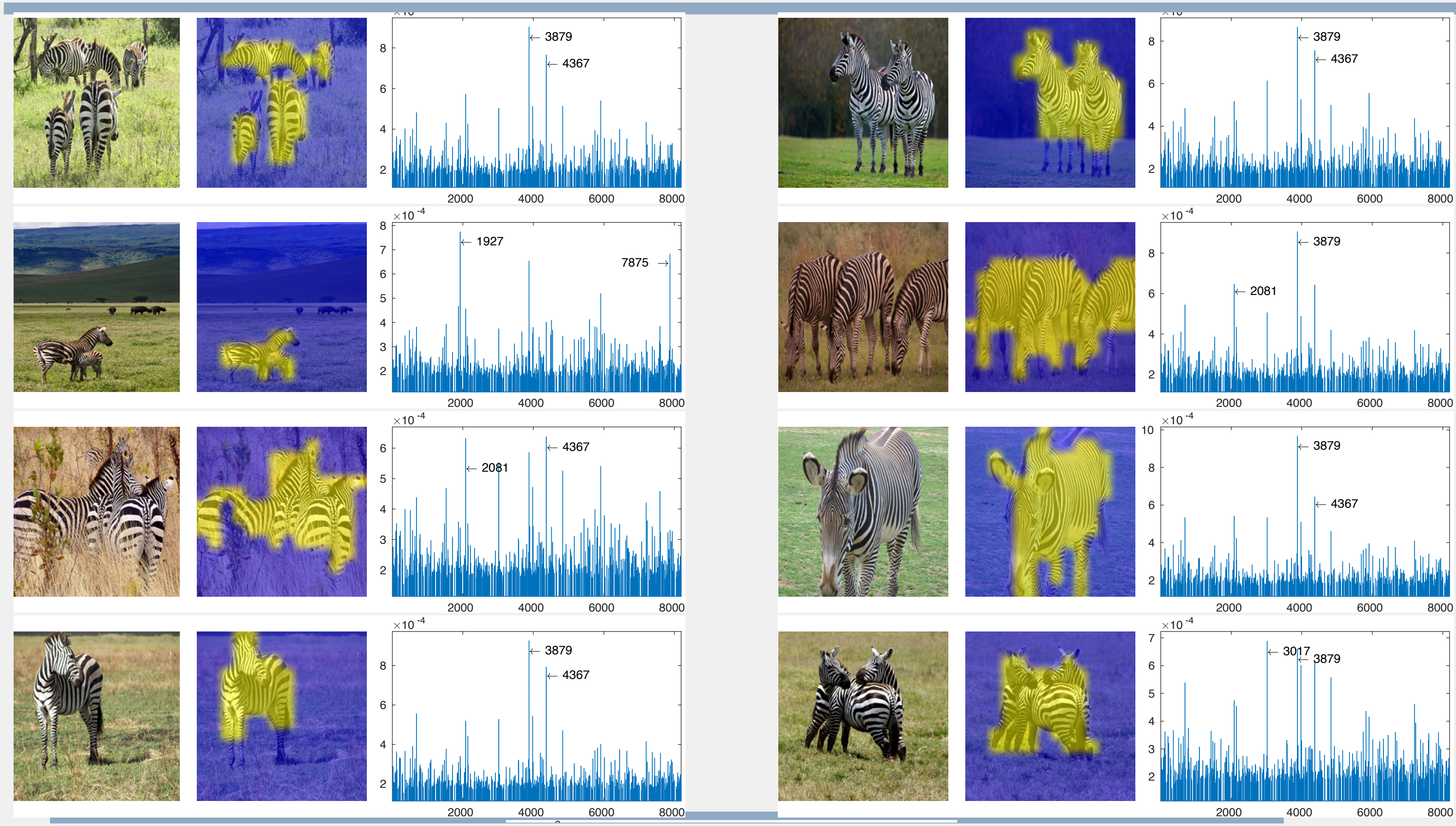
	CIFAR10	CIFAR100	Cars	Flowers
Random Init.	*91.42*	*70.14*	*10.67*	*54.04*
	w/o multi-crop			
MC-SSL0.0 [PR]	97.19	81.98	76.78	88.21
MC-SSL0.0 [PC]	97.77	84.25	83.93	94.89
MC-SSL0.0 [PC + PR]	**97.82**	**84.98**	**86.15**	**95.56**

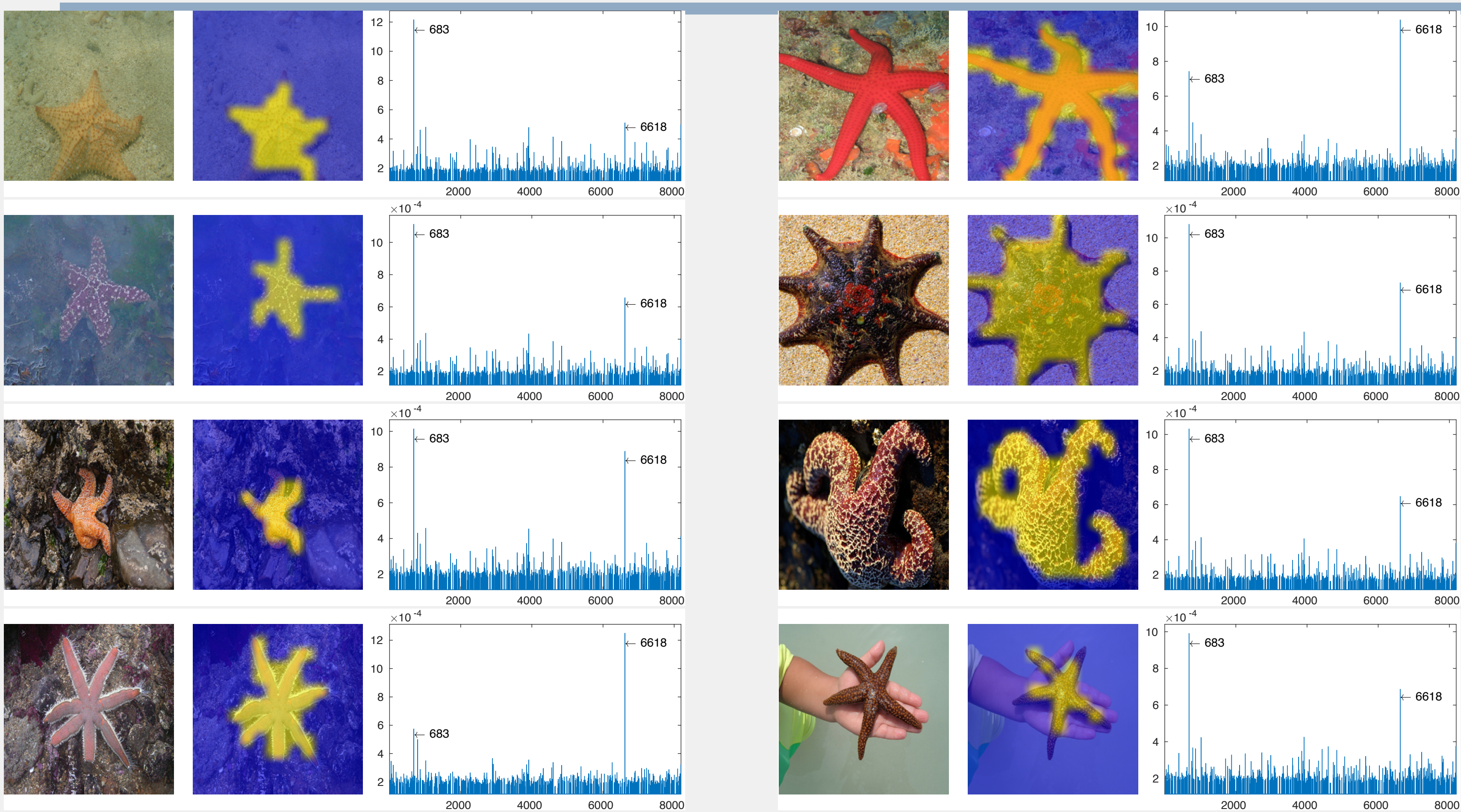
	CIFAR10	CIFAR100	Cars	Flowers
From scratch (i.e., random initialization)				
ViT-S/16	91.42	70.14	10.67	54.04
Self-supervised pre-training on the given dataset				
MC-SSL0.0‡	98.00	85.38	89.20	87.30
Selfsupervised pretraining on 10% of ImageNet-1K				
	w/o multi-crop			
Dino	97.27	81.77	82.08	92.68
MC-SSL0.0	**97.82**	**84.98**	**86.15**	**95.56**
	with multi-crop			
Dino‡	97.90	84.61	88.21	95.46
MC-SSL0.0‡	**98.08**	**85.82**	**90.44**	**96.31**

Comparison of MC-SSL with baselines



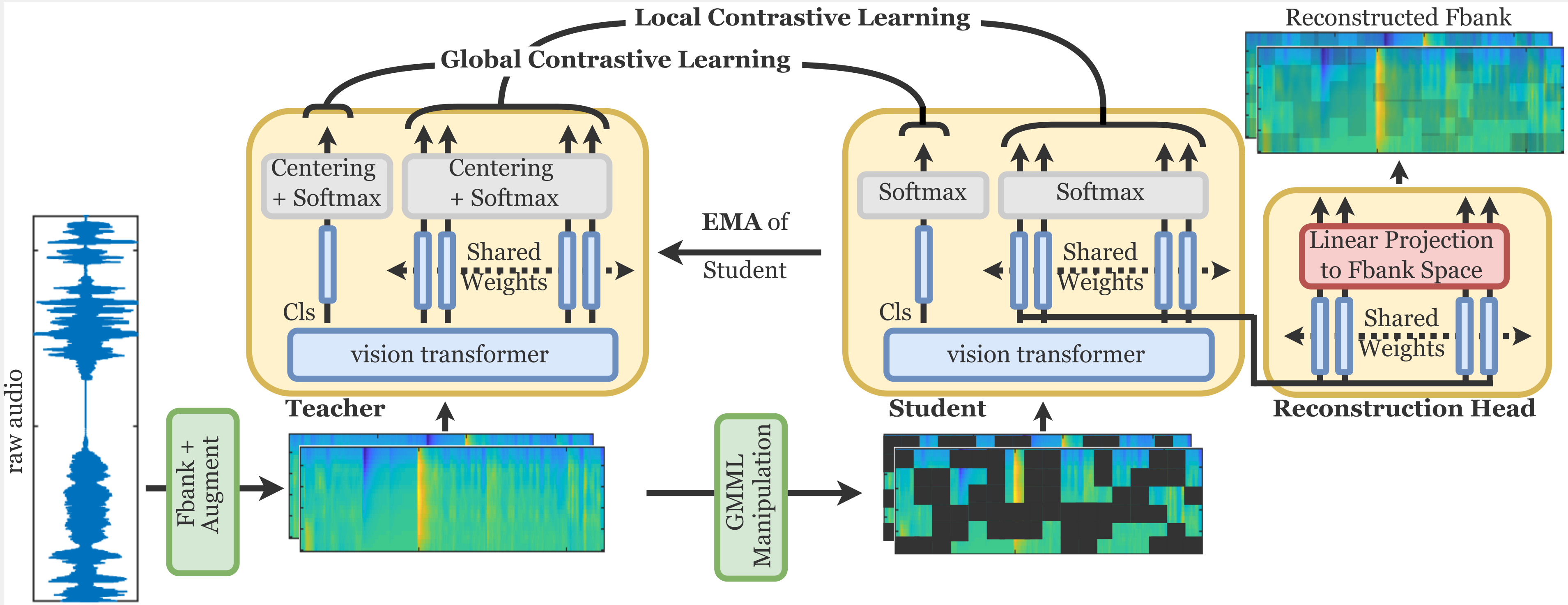
mm


3879
4367
1927
7875
3879
2081
2081
4367
3879
4367
3879
4367
3017
3879
×10^-4
2000
4000
6000
8000


683
6618
×10^-4
2000
4000
6000
8000

Audio classification

- System architecture



Audio classification evaluation

- Results

Table 1. Comparison with state-of-the-art works on audio and speech classification tasks. Evaluation metrics are mean average precision (mAP) for AS-2K and accuracy (%) for ESC-5, SC-V1, SC-V2, and SID. ↑ shows the improvement over best SOTA.

Method	Backbone	Pretraining Data	Transfer Learning				
			AS-20K	ESC-50	SC-V2	SC-V1	SID
Supervised-learning-based methods							
PANNs [44]	CNN	–	27.8	83.3	–	61.8	–
AST [5]	ViT-B	AS-2M	28.6	86.8	96.2	91.6	35.2
Self-supervised-learning-based methods							
COLA [21]	CNN	AS-2M	–	–	98.1	95.5	37.7
SSAST [6]	ViT-B	AS-2M	29.0	84.7	97.8	94.8	57.1
MaskSpec [8]	ViT-B	AS-2M	32.3	89.6	97.7	–	–
ASiT (ours)	ViT-B	AS-2M	**35.2** (↑ 2.9)	**92.0** (↑ 2.4)	**98.8** (↑ 0.7)	**98.1** (↑ 2.6)	**63.1** (↑ 6.0)
SSL based methods for reference not comparison as they are pretrained on additional speech dataset LS [45]							
SSAST [6]	ViT-B	AS-2M + LS	31.0	88.8	98.0	96.0	64.3
MAE-AST [7]	ViT-B	AS-2M + LS	30.6	90.0	97.9	95.8	63.3

Conclusions

- SSL provides a much better prospect for building foundation models in AI
- Its main benefits
  - no need for data annotation
  - does not propagate supervised learning biases
  - enables solving downstream tasks using small datasets
- Recent significant advances in SSL owe to masked image modelling
- Many challenges still outstanding
  - no theoretical underpinning

Thanks