ICLR 22 Potpourri

Mert

Content

~Pre-training~

~Vision Transformers~

~Self-supervised Learning~

~Continual Learning~

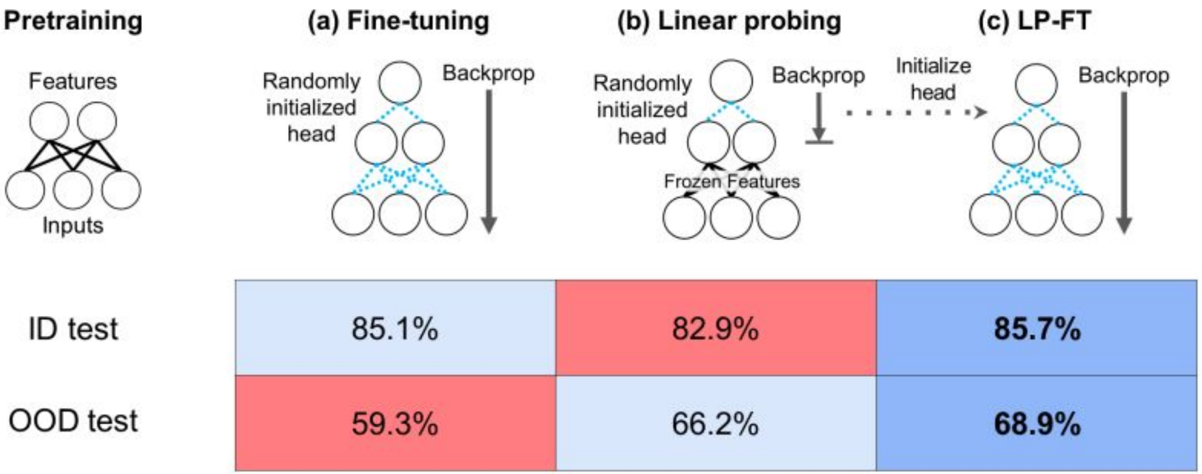
~Meta Learning~

~Novel Tasks & Paradigms~

~Pre-Training~

1

FINE-TUNING CAN DISTORT PRETRAINED FEATURES AND UNDERPERFORM OUT-OF-DISTRIBUTION



	(a) Fine-tuning	(b) Linear probing	(c) LP-FT
ID test	85.1%	82.9%	**85.7%**
OOD test	59.3%	66.2%	**68.9%**

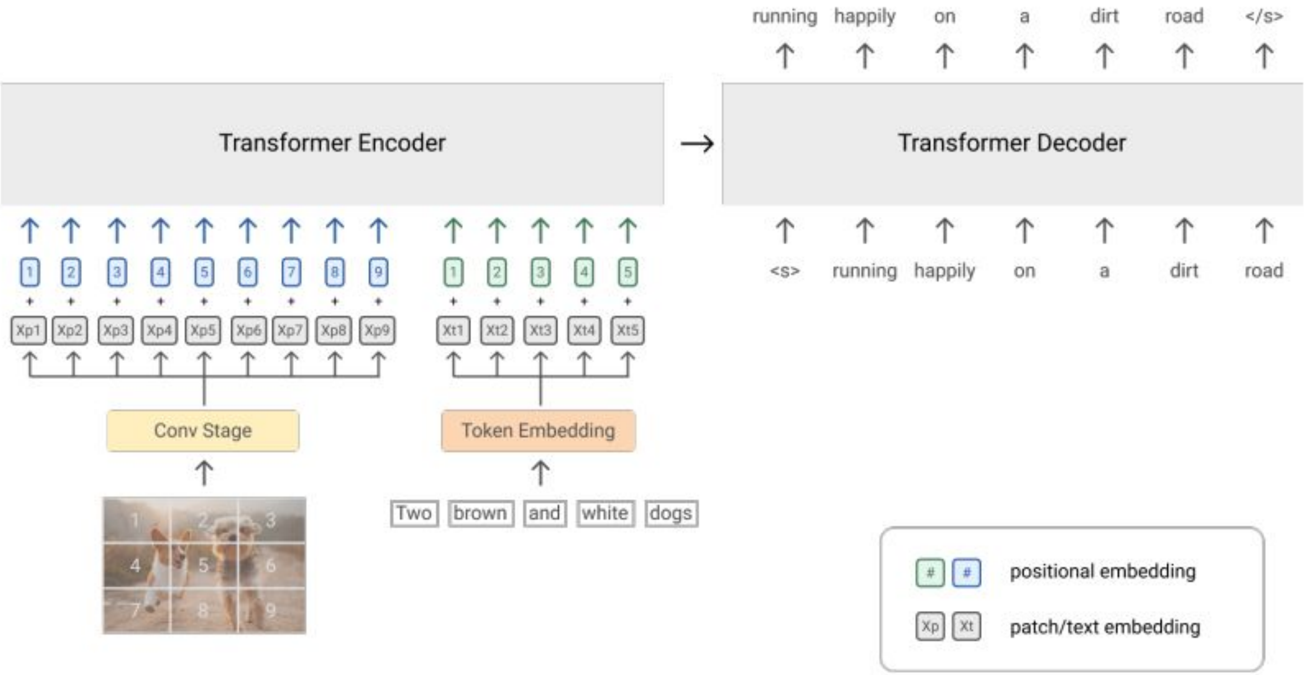
Average accuracies (10 distribution shifts)

Don’t insert a random classifier to a pre-trained network. First, linear probe, then fine-tune the rest.

~Pre-Training~

2

SIMVLM: SIMPLE VISUAL LANGUAGE MODEL PRETRAINING WITH WEAK SUPERVISION

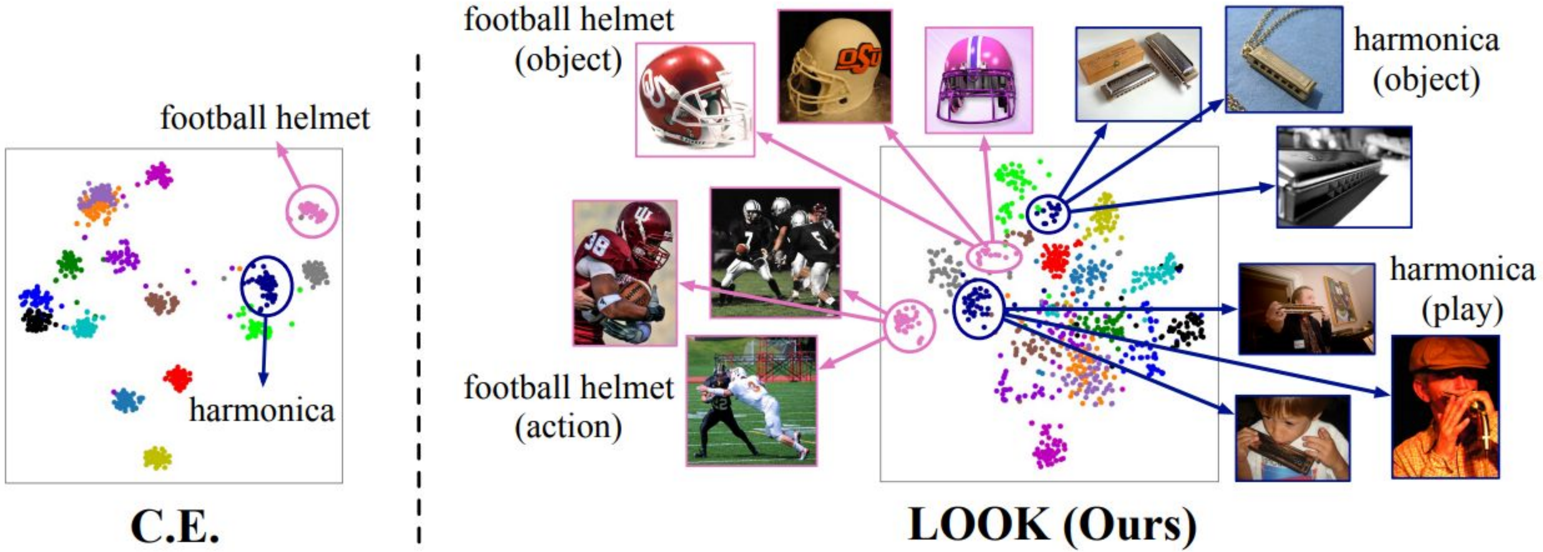


Simply predicting next tokens, given either i) previous tokens, ii) previous tokens + image works best.

~Pre-Training~

3

RETHINKING SUPERVISED PRE-TRAINING FOR BETTER DOWNSTREAM TRANSFERRING



Discover sub-clusters of the same object (i.e. [helmet-action] vs. [helmet-static]).

Exhibits even better transferability than self-supervision.

Content

~Pre-training~

~Vision Transformers~

~Self-supervised Learning~

~Continual Learning~

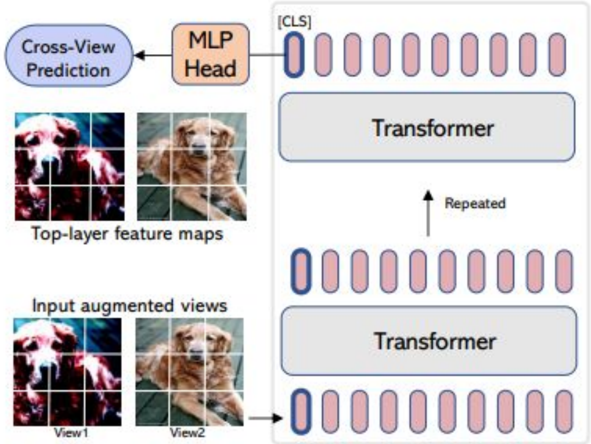
~Meta Learning~

~Novel Tasks & Paradigms~

~Vision Transformers~

1

Efficient Self-supervised Vision Transformers for Representation Learning



(a) Baseline monolithic architecture

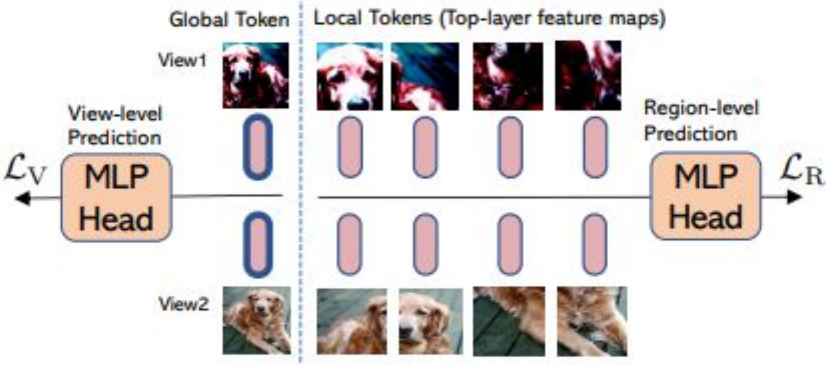


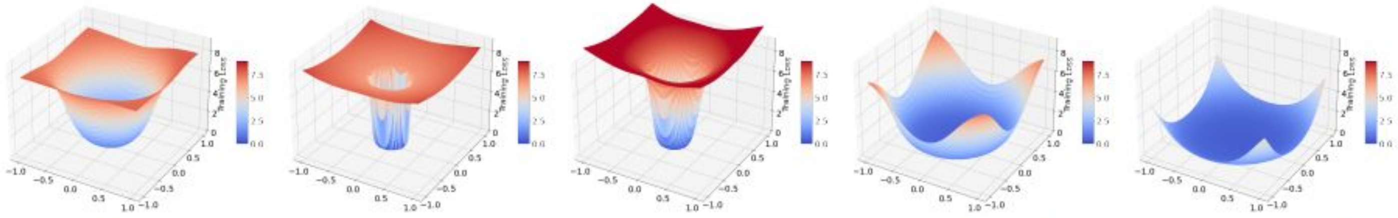
Figure 3: Pre-training objectives.

a) Feed different views of the same input to ViT, b) Match patch-level features across two views.

~Vision Transformers~

2

WHEN VISION TRANSFORMERS OUTPERFORM RESNETS WITHOUT PRE-TRAINING OR STRONG DATA AUGMENTATIONS



(a) ResNet (b) ViT (c) Mixer (d) ViT-SAM (e) Mixer-SAM

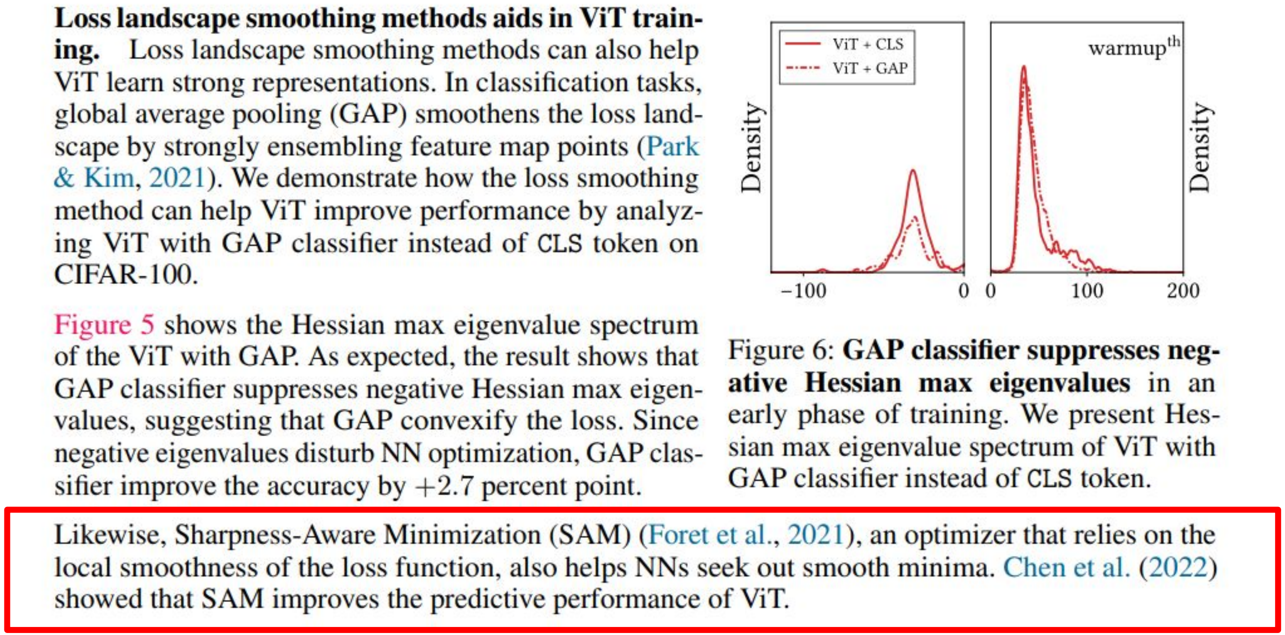
Figure 1: Cross-entropy loss landscapes of ResNet-152, ViT-B/16, and Mixer-B/16. ViT and MLP-Mixer converge to sharper regions than ResNet when trained on ImageNet with the basic Inception-style preprocessing. SAM, a sharpness-aware optimizer, significantly smooths the landscapes.

If you simply smooth loss landscape of ViTs, then you don't need **JFT-300M** or heavier augmentations.

~Vision Transformers~

3

HOW DO VISION TRANSFORMERS WORK?



Loss landscape smoothing methods aids in ViT training. Loss landscape smoothing methods can also help ViT learn strong representations. In classification tasks, global average pooling (GAP) smoothens the loss landscape by strongly ensembling feature map points (Park & Kim, 2021). We demonstrate how the loss smoothing method can help ViT improve performance by analyzing ViT with GAP classifier instead of CLS token on CIFAR-100.

Figure 5 shows the Hessian max eigenvalue spectrum of the ViT with GAP. As expected, the result shows that GAP classifier suppresses negative Hessian max eigenvalues, suggesting that GAP convexify the loss. Since negative eigenvalues disturb NN optimization, GAP classifier improve the accuracy by +2.7 percent point.

Figure 6: **GAP classifier suppresses negative Hessian max eigenvalues** in an early phase of training. We present Hessian max eigenvalue spectrum of ViT with GAP classifier instead of CLS token.

Likewise, Sharpness-Aware Minimization (SAM) (Foret et al., 2021), an optimizer that relies on the local smoothness of the loss function, also helps NNs seek out smooth minima. Chen et al. (2022) showed that SAM improves the predictive performance of ViT.

Same conclusion! If you smoothify ViT loss landscape with SAM, it improves.

~Vision Transformers~

4 COSFORMER : RETHINKING SOFTMAX IN ATTENTION

5 VITGAN: TRAINING GANS WITH VISION TRANSFORMERS

6 AUTO-SCALING VISION TRANSFORMERS WITHOUT TRAINING

7 SPARSE DETR: EFFICIENT END-TO-END OBJECT DETECTION WITH LEARNABLE SPARSITY

Content

~Pre-training~

~Vision Transformers~

~Self-supervised Learning~

~Continual Learning~

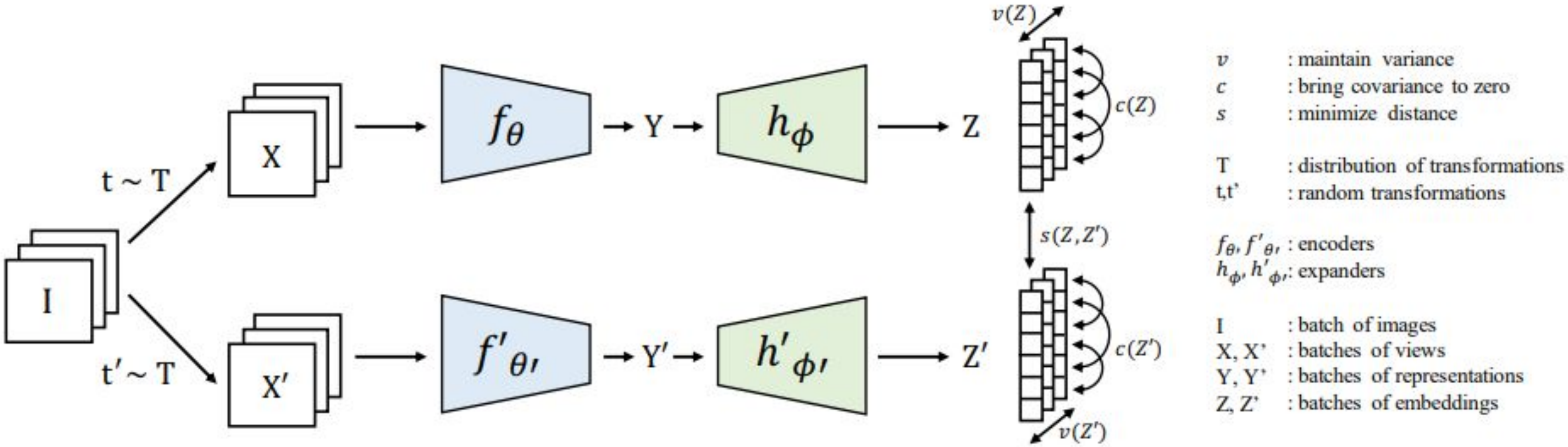
~Meta Learning~

~Novel Tasks & Paradigms~

~Self-supervised Learning~

1

VICREG: VARIANCE-INVARIANCE-COVARIANCE REGULARIZATION FOR SELF-SUPERVISED LEARNING (Bardes, Ponce, LeCun)



I
t ~ T
t' ~ T
X
X'
f_θ
$f'_{\theta'}$
Y
Y'
h_ϕ
$h'_{\phi'}$
Z
Z'
$v(Z)$
$c(Z)$
$s(Z,Z')$
$c(Z')$
$v(Z')$
v : maintain variance
c : bring covariance to zero
s : minimize distance
T : distribution of transformations
t,t' : random transformations
$f_\theta, f'_{\theta'}$: encoders
$h_\phi, h'_{\phi'}$: expanders
I : batch of images
X, X' : batches of views
Y, Y' : batches of representations
Z, Z' : batches of embeddings

Variance: Force feature dimensions to maintain a certain variability.

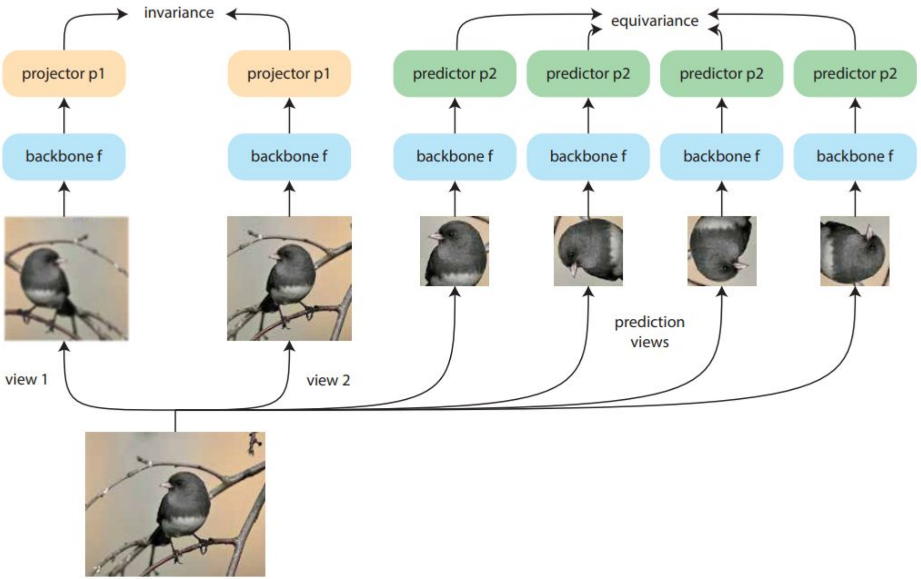
In-Variance: Force features of different view to be as similar as possible.

Co-Variance: Force de-correlated features across batch.

~Self-supervised Learning~

2

EQUIVARIANT CONTRASTIVE LEARNING



Forcing the network to retain only rotation variation (while ignoring others) helps significantly.

~Self-supervised Learning~

3

CHAOS IS A LADDER: A NEW THEORETICAL UNDERSTANDING OF CONTRASTIVE LEARNING VIA AUGMENTATION OVERLAP

Theoretical work that blew my mind.

4

DISENTANGLING PROPERTIES OF CONTRASTIVE METHODS

Contrastive learning factorizes different input variations (light, color, rotation) into separate channels.

Content

~Pre-training~

~Vision Transformers~

~Self-supervised Learning~

~Continual Learning~

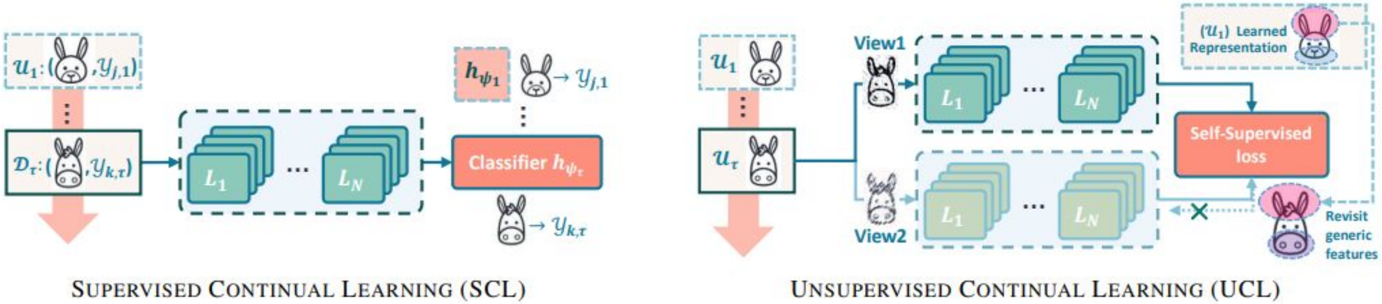
~Meta Learning~

~Novel Tasks & Paradigms~

~Continual Learning~

1

REPRESENTATIONAL CONTINUITY FOR UNSUPERVISED CONTINUAL LEARNING



SUPERVISED CONTINUAL LEARNING (SCL)

UNSUPERVISED CONTINUAL LEARNING (UCL)

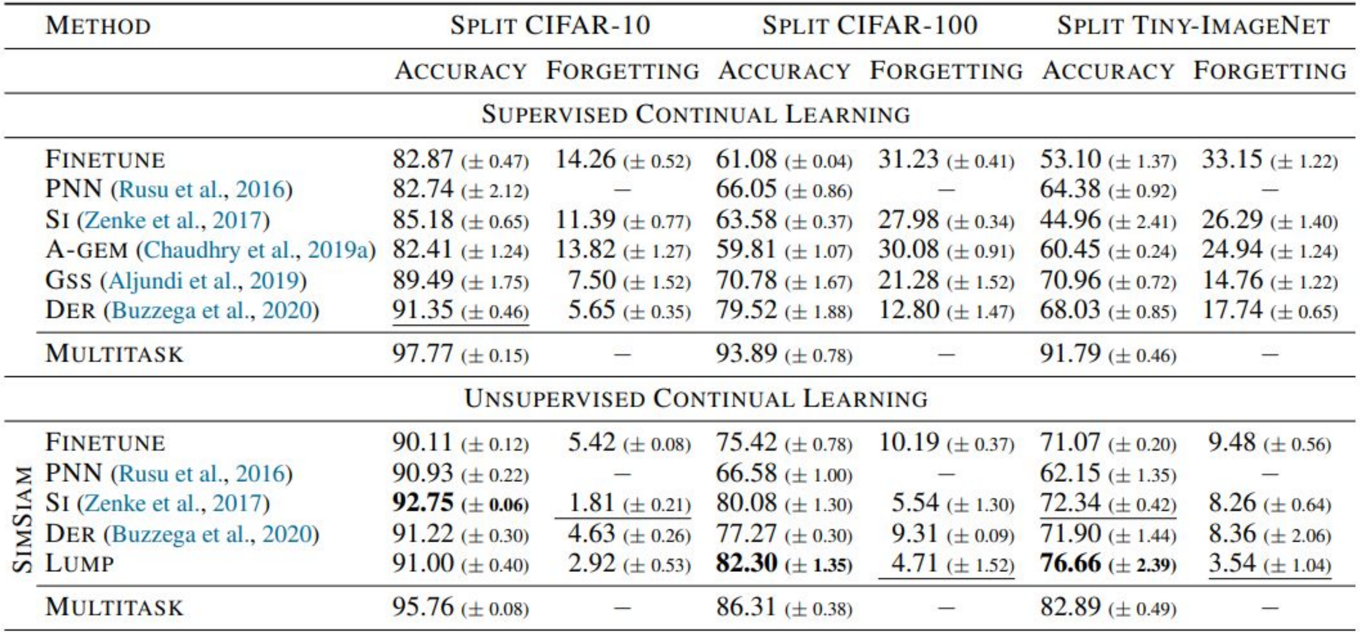
Figure 1: **Illustration of supervised and unsupervised continual learning.** The objective of SCL is to learn the ability to classify labeled images in the current task while preserving the past tasks' knowledge, where the tasks are non-iid to each other. On the other hand, UCL aims to learn the representation of images without the presence of labels and the model learns general-purpose representations during sequential training.

Self-supervised backbones are much better continual learners (they forget less).

~Continual Learning~

1

REPRESENTATIONAL CONTINUITY FOR UNSUPERVISED CONTINUAL LEARNING



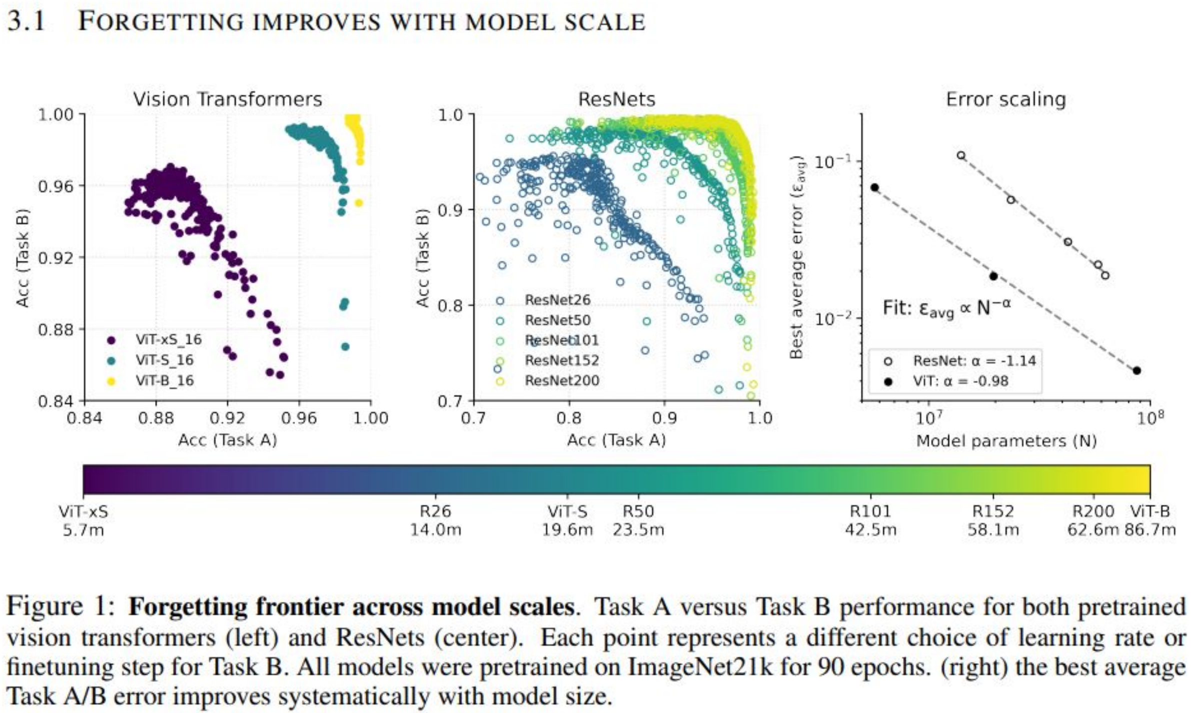
	Method	Split CIFAR-10		Split CIFAR-100		Split Tiny-ImageNet	
		Accuracy	Forgetting	Accuracy	Forgetting	Accuracy	Forgetting
		Supervised Continual Learning					
	Finetune	82.87 (± 0.47)	14.26 (± 0.52)	61.08 (± 0.04)	31.23 (± 0.41)	53.10 (± 1.37)	33.15 (± 1.22)
	PNN (Rusu et al., 2016)	82.74 (± 2.12)	—	66.05 (± 0.86)	—	64.38 (± 0.92)	—
	SI (Zenke et al., 2017)	85.18 (± 0.65)	11.39 (± 0.77)	63.58 (± 0.37)	27.98 (± 0.34)	44.96 (± 2.41)	26.29 (± 1.40)
	A-GEM (Chaudhry et al., 2019a)	82.41 (± 1.24)	13.82 (± 1.27)	59.81 (± 1.07)	30.08 (± 0.91)	60.45 (± 0.24)	24.94 (± 1.24)
	GSS (Aljundi et al., 2019)	89.49 (± 1.75)	7.50 (± 1.52)	70.78 (± 1.67)	21.28 (± 1.52)	70.96 (± 0.72)	14.76 (± 1.22)
	DER (Buzzega et al., 2020)	91.35 (± 0.46)	5.65 (± 0.35)	79.52 (± 1.88)	12.80 (± 1.47)	68.03 (± 0.85)	17.74 (± 0.65)
	Multitask	97.77 (± 0.15)	—	93.89 (± 0.78)	—	91.79 (± 0.46)	—
		Unsupervised Continual Learning					
SimSiam	Finetune	90.11 (± 0.12)	5.42 (± 0.08)	75.42 (± 0.78)	10.19 (± 0.37)	71.07 (± 0.20)	9.48 (± 0.56)
	PNN (Rusu et al., 2016)	90.93 (± 0.22)	—	66.58 (± 1.00)	—	62.15 (± 1.35)	—
	SI (Zenke et al., 2017)	**92.75 (± 0.06)**	1.81 (± 0.21)	80.08 (± 1.30)	5.54 (± 1.30)	72.34 (± 0.42)	8.26 (± 0.64)
	DER (Buzzega et al., 2020)	91.22 (± 0.30)	4.63 (± 0.26)	77.27 (± 0.30)	9.31 (± 0.09)	71.90 (± 1.44)	8.36 (± 2.06)
	LUMP	91.00 (± 0.40)	2.92 (± 0.53)	**82.30 (± 1.35)**	4.71 (± 1.52)	**76.66 (± 2.39)**	3.54 (± 1.04)
	Multitask	95.76 (± 0.08)	—	86.31 (± 0.38)	—	82.89 (± 0.49)	—

Significant drop in *forgetting* in comparison to supervised (standard) continual learning.

~Continual Learning~

2

EFFECT OF MODEL AND PRETRAINING SCALE ON CATASTROPHIC FORGETTING IN NEURAL NETWORKS



3.1 FORGETTING IMPROVES WITH MODEL SCALE

Figure 1: **Forgetting frontier across model scales**. Task A versus Task B performance for both pretrained vision transformers (left) and ResNets (center). Each point represents a different choice of learning rate or finetuning step for Task B. All models were pretrained on ImageNet21k for 90 epochs. (right) the best average Task A/B error improves systematically with model size.

Bigger backbones forget less, with the help of: 1) Supervised, 2) Self-supervised pre-training

~Continual Learning~

3 Task-Induced Representation Learning

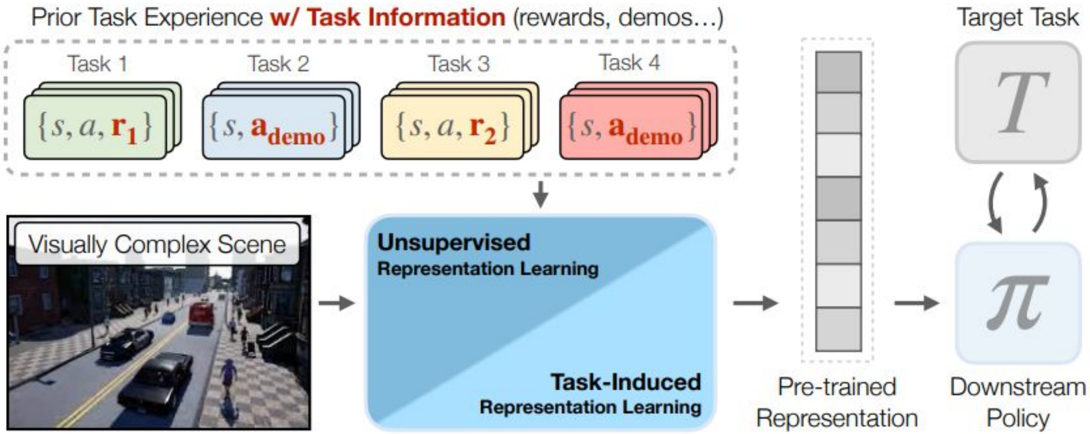


Figure 1: Overview of our pipeline for evaluating representation learning in visually complex scenes: given a multi-task dataset of prior experience we pre-train representations using unsupervised objectives, such as prediction and contrastive learning, or task-induced approaches, which leverage task-information from prior tasks to learn to focus on task-relevant aspects of the scene. We then evaluate the efficiency of the pre-trained representations for learning unseen tasks.

Label + Self-supervision drastically improves visual reinforcement learning.

Content

~Pre-training~

~Vision Transformers~

~Self-supervised Learning~

~Continual Learning~

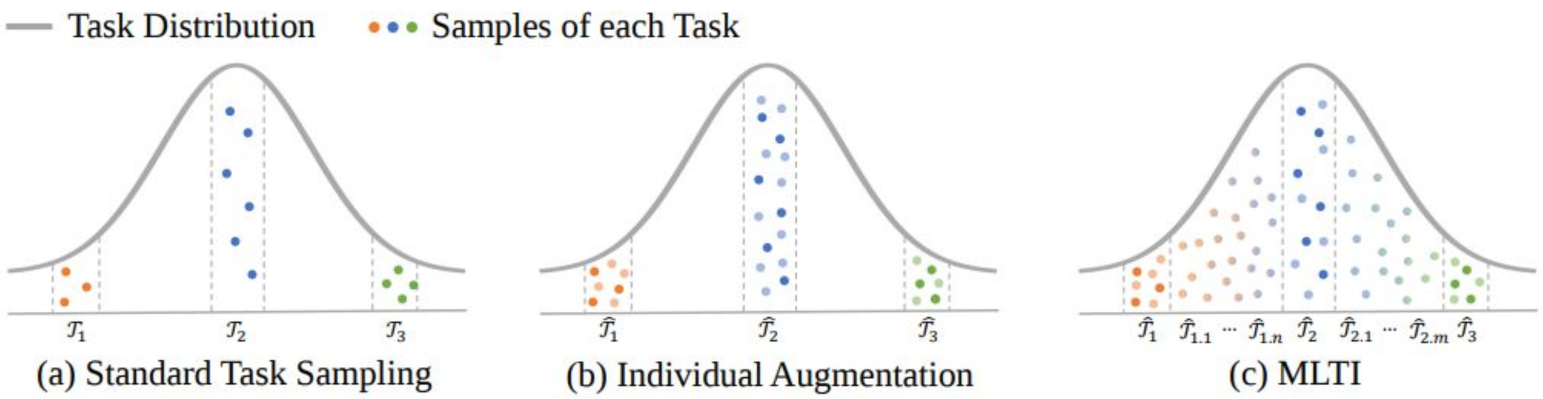
~Meta Learning~

~Novel Tasks & Paradigms~

~Meta Learning~

1

META-LEARNING WITH FEWER TASKS THROUGH TASK INTERPOLATION



(a) Standard Task Sampling

(b) Individual Augmentation

(c) MLTI

Figure 1: Motivations behind MLTI. (a) three tasks are sampled from the task distribution; (b) individual augmentation methods (e.g., (Ni et al., 2021; Yao et al., 2021) augment each task within its own distribution); (c) MLTI densifies the task-level distribution by performing cross-task interpolation.

Sampling in-between exemplars during meta-training (slightly) mitigates over-fitting (MAML and ProtoNet)

~Meta Learning~

2

HOW TO TRAIN YOUR MAML TO EXCEL IN FEW-SHOT CLASSIFICATION

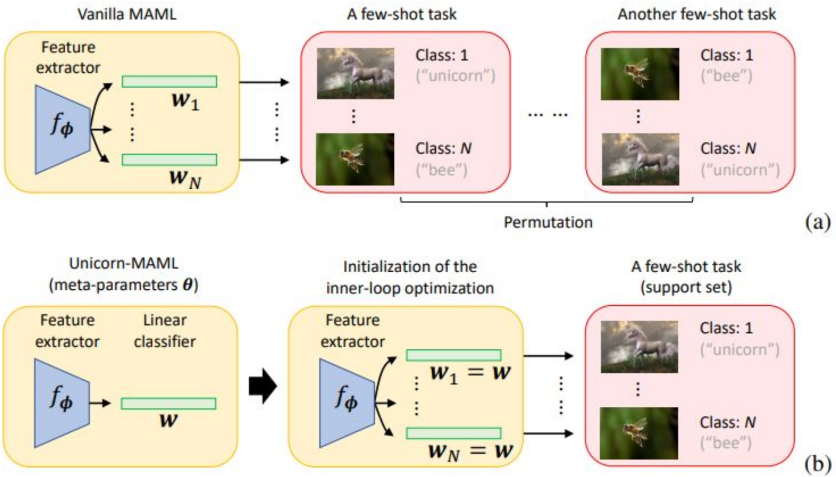


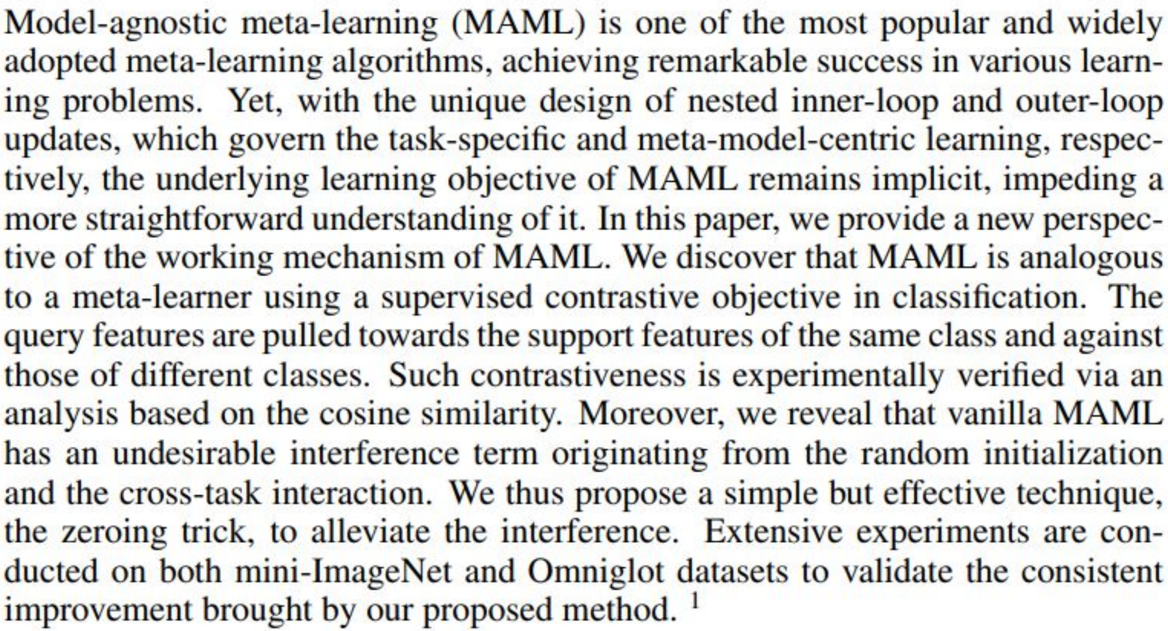
Figure 1: **The problem of permutations in label assignments, and the illustration of UNICORN-MAML.** (a) A vanilla MAML learns the initialization of ϕ and the N weight vectors $\{\boldsymbol{w}_c\}_{c=1}^N$. Each of $\{\boldsymbol{w}_c\}_{c=1}^N$ is paired with the corresponding class label $c \in [N]$ of a few-shot task. A few-shot task, however, may consist of the same set of semantic classes but in different permutations of class label assignments, leading to a larger variance in meta-testing accuracy. (b) In contrast, our **UNICORN-MAML**, besides learning ϕ, learns only a single weight vector $\boldsymbol{w}$ and uses it to initialize all the N weight vectors $\{\boldsymbol{w}_c\}_{c=1}^N$ at the beginning of the inner loop. That is, UNICORN-MAML directly forces the learned model initialization to be permutation-invariant.

Initialize the inner loop (class-specific params) with all the same parameter (not separately).

~Meta Learning~

3

MAML IS A NOISY CONTRASTIVE LEARNER IN CLASSIFICATION



Model-agnostic meta-learning (MAML) is one of the most popular and widely adopted meta-learning algorithms, achieving remarkable success in various learning problems. Yet, with the unique design of nested inner-loop and outer-loop updates, which govern the task-specific and meta-model-centric learning, respectively, the underlying learning objective of MAML remains implicit, impeding a more straightforward understanding of it. In this paper, we provide a new perspective of the working mechanism of MAML. We discover that MAML is analogous to a meta-learner using a supervised contrastive objective in classification. The query features are pulled towards the support features of the same class and against those of different classes. Such contrastiveness is experimentally verified via an analysis based on the cosine similarity. Moreover, we reveal that vanilla MAML has an undesirable interference term originating from the random initialization and the cross-task interaction. We thus propose a simple but effective technique, the zeroing trick, to alleviate the interference. Extensive experiments are conducted on both mini-ImageNet and Omniglot datasets to validate the consistent improvement brought by our proposed method. [1]

Training MAML is equivalent of supervised contrastive learning (bridging gap to metric learning).

~Meta Learning~

4 CURRICULUM LEARNING AS A TOOL TO UNCOVER LEARNING PRINCIPLES IN THE BRAIN

5 LEARNING META-FEATURES FOR AUTOML

6 TASK RELATEDNESS-BASED GENERALIZATION BOUNDS FOR META LEARNING

Content

~Pre-training~

~Vision Transformers~

~Self-supervised Learning~

~Continual Learning~

~Meta Learning~

~Novel Tasks & Paradigms~

~Novel Tasks~

1

OPEN-WORLD SEMI-SUPERVISED LEARNING

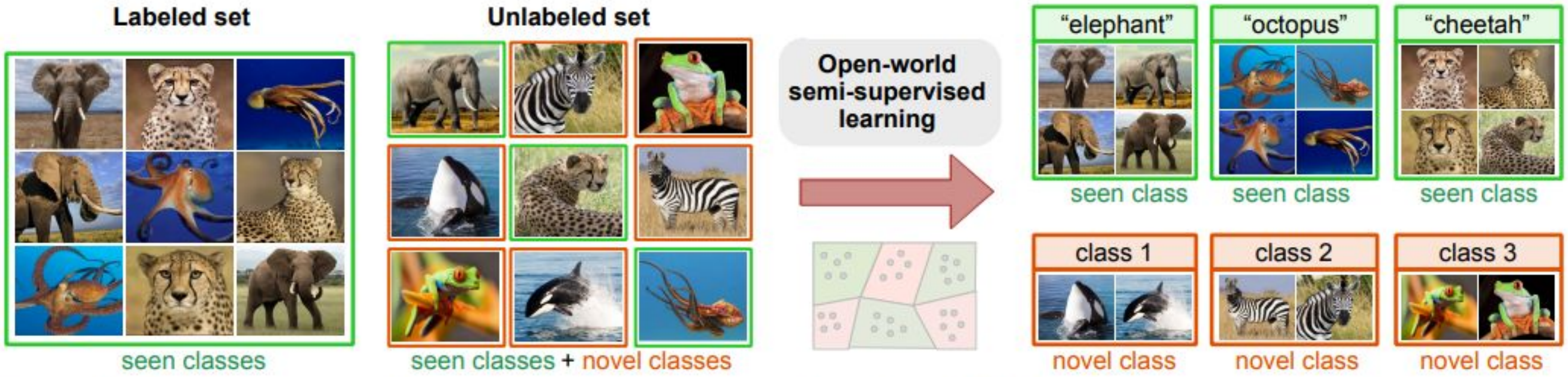


Figure 1: In the open-world semi-supervised learning, the unlabeled dataset may contain classes that have never been encountered in the labeled set. The model needs to be able to classify samples into previously seen classes, but also distinguish between unseen classes.

A learner observes: 1) Unlabeled examples of seen classes, 2) Unlabeled examples of unseen classes.

~Novel Tasks~

2 VISUAL REPRESENTATION LEARNING OVER LATENT DOMAINS

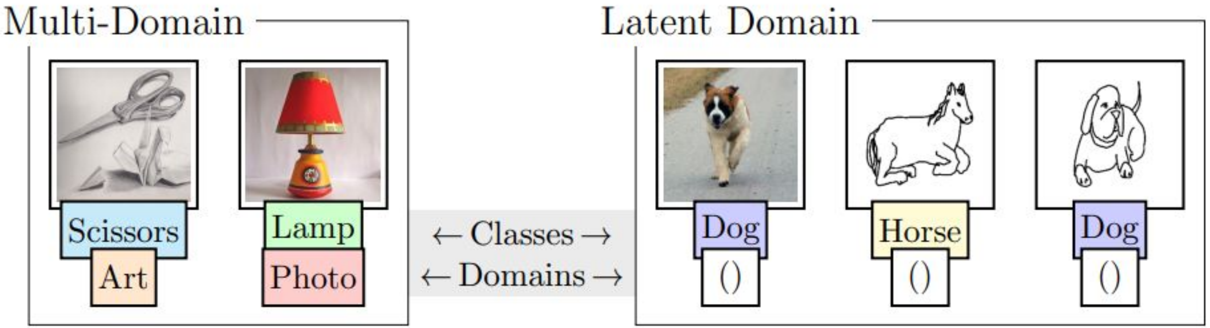


Figure 1: In multi-domain learning every sample has a domain label. Latent domain learning studies how models may best be learned without this information.

The learner observes images from multiple domains, although domains are not known (latent).

~Novel Tasks~

3

An Image Classifier Can Suffice For Video Understanding

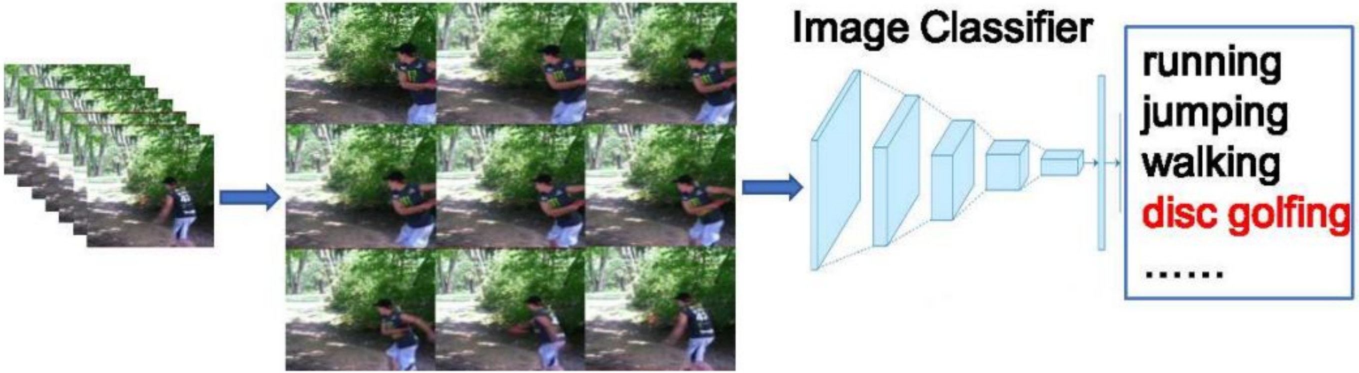


Figure 2: **Overview of SIFAR**. A sequence of input frames from a video are first made into a super image, which is then fed into a conventional image classifier for action recognition.

Treat a video as a canvas of N frames observed at the same time (i.e. convert time to spatial domain).

Summary

~To improve Pre-training~

Retain visual variation of subtypes of objects (i.e. object states).

~ To improve Vision Transformers Efficiency~

Retain the smoothness of loss landscape -> No need for heavy pre-training & augmentation.

~To improve Self-supervised Learning~

Retain variation of visual features: Variance/Rotation.

~To improve Continual Learning~

Retain self-supervised representation learning.

Discussion

Is **full supervision** dead yet?

Is **image-to-human label** paradigm causing lots of information to be lost (therefore less transferable)?

How to determine **which variation** to store (and neglect)?