Appendices

A Experimental details

A.1 Detailed ingredients and hyper-parameters

Table 8: Summary of our training procedures with ImageNet-1k and ImageNet-21k. We also provide DeiT [47], Wightman et al [56] and Steiner et al. [41] baselines for reference. Adapt. means the hparams is adapted to the size of the model. For finetuning to higher resolution with model pre-trained on ImageNet-1k only we use the finetuning procedure from DeiT see section A.2 for more details.

	Previous approaches				Ours		
Procedure →	ViT	Steiner	DeiT	Wightman	ImNet-1k	ImNet-21k	
Reference	[12]	et al. [41]	[47]	et al. [56]		Pretrain.	Finetune.
Batch size	4096	4096	1024	2048	2048	2048	2048
Optimizer	AdamW	AdamW	AdamW	LAMB	LAMB	LAMB	LAMB
LR	3.10^{-3}	3.10^{-3}	1.10^{-3}	5.10^{-3}	3.10^{-3}	3.10^{-3}	3.10^{-4}
LR decay	cosine	cosine	cosine	cosine	cosine	cosine	cosine
Weight decay	0.1	0.3	0.05	0.02	0.02	0.02	0.02
Warmup epochs	3.4	3.4	5	5	5	5	5
Label smoothing ε	0.1	0.1	0.1	✗	✗	0.1	0.1
Dropout	✓	✓	✗	✗	✗	✗	✗
Stoch. Depth	✗	✓	✓	✓	✓	✓	✓
Repeated Aug	✗	✗	✓	✓	✓	✗	✗
Gradient Clip.	1.0	1.0	✗	1.0	1.0	1.0	1.0
H. flip	✓	✓	✓	✓	✓	✓	✓
RRC	✓	✓	✓	✓	✓	✗	✗
Rand Augment	✗	Adapt.	9/0.5	7/0.5	✗	✗	✗
3 Augment (ours)	✗	✗	✗	✗	✓	✓	✓
LayerScale	✗	✗	✗	✗	✓	✓	✓
Mixup alpha	✗	Adapt.	0.8	0.2	0.8	✗	✗
Cutmix alpha	✗	✗	1.0	1.0	1.0	1.0	1.0
Erasing prob.	✗	✗	0.25	✗	✗	✗	✗
ColorJitter	✗	✗	✗	✗	0.3	0.3	0.3
Test crop ratio	0.875	0.875	0.875	0.95	1.0	1.0	1.0
Loss	CE	CE	CE	BCE	BCE	CE	CE

A.2 Baselines and default settings

The main task that we consider in this paper for the evaluation of our training procedure is image classification. We train on Imagenet1k-train and evaluate on Imagenet1k-val, with results on ImageNet-V2 to control overfitting. We also consider the case where we can pretrain on ImageNet-21k, Finally, we report transfer learning results on 6 different datasets/benchmarks.

Default setting. When training on ImageNet-1k only, by default we train during 400 epochs with a batch size 2048, following prior works [50,59]. Unless specified otherwise, both the training and evaluation are carried out at resolution 224×224 (even though we recommend to train at a lower resolution when targeting 224×224 at inference time).

When pre-training on ImageNet-21k, we pre-train by default during 90 epochs at resolution 224×224, followed by a finetuning of 50 epochs on ImageNet-1k. In this context, we consider two fine-tuning resolutions: 224×224 and 384×384.

Fine-tuning at higher resolution. When pre-training on ImageNet-1k at resolution 224×224 we fix the train-test resolution discrepancy by finetuning at a higher resolution [52]. Our finetuning procedure is inspired by DeiT, except that we adapt the stochastic depth rate according to the model size [50]. We fix the learning reate to $lr = 1 \times 10^{-5}$ with batch-size=512 during 20 epochs with a weight decay of 0.1 without repeated augmentation. Other hyper-parameters are similar to those employed in DeiT fine-tuning.

Stochastic depth. We adapt the stochastic depth drop rate according to the model size. We report stochastic depth drop rate values in Table 9.

Table 9: Stochastic depth drop-rate according to the model size. For 400 epochs training on ImageNet-1k and 90 epochs training on ImageNet-21k. See section B for further adaption with longer training.

Model	# Params ($\times 10^6$)	FLOPs ($\times 10^9$)	Stochastic depth drop-rate	
			ImageNet-1k	ImageNet-21k
ViT-T	5.7	1.3	0.0	0.0
ViT-S	22.0	4.6	0.0	0.0
ViT-B	86.6	17.5	0.1	0.1
ViT-L	304.4	61.6	0.4	0.3
ViT-H	632.1	167.4	0.5	0.5

For transfer learning experiments we evaluate our models pre-trained at resolution 224×224 on ImageNet-1k only on 6 transfer learning datasets. We give the details of these datasets in Table 10 below.

B Additional details and Ablations

Number of training epochs In Table 11 we provide an ablation on the number of training epochs on ImageNet-1k. We do not observe a saturation when the increase of the number of training epochs, as observed with BerT like approaches [1,19]. For longer training we increase the weight decay from 0.02 to 0.05 and we increase the stochastic depth drop-rate by 0.05 every 200 epochs to prevent overfitting.

Table 10: Datasets used for our different transfer-learning tasks.

Dataset	Train size	Test size	#classes
iNaturalist 2018 [23]	437,513	24,426	8,142
iNaturalist 2019 [22]	265,240	3,003	1,010
Flowers-102 [35]	2,040	6,149	102
Stanford Cars [26]	8,144	8,041	196
CIFAR-100 [28]	50,000	10,000	100
CIFAR-10 [28]	50,000	10,000	10

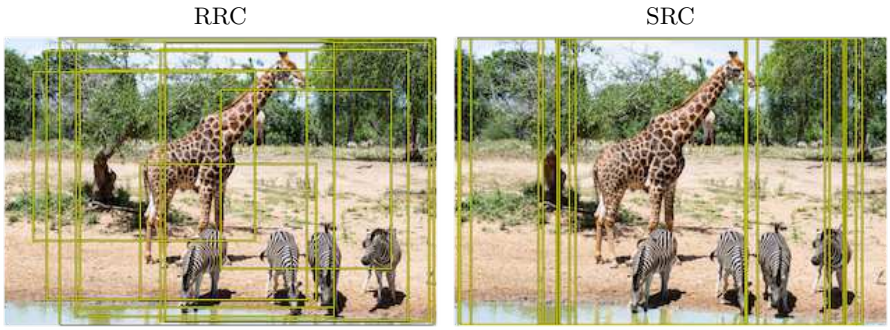

Fig. 7: Example of crops selected by Random Resized Crop and Simple Random Crop.

Impact of training resolution In Table 12 we report the evolution of the performance according to the training resolution. We observe that we benefit from the FixRes [52] effect. By training at resolution 192×192 (or 160×160) we get a better performance at 224 after a slight fine-tuning than when training from scratch at 224×224.

We observe that the resolution has a regularization effect. While it is known that it is best to use a smaller resolution at training time [52], we also observe in the training curves that this show reduces the overfitting of the larger models. This is also illustrated by our results Table 12 with ViT-H and ViT-L. This is especially important with longer training, where models overfit without a stronger regularisation. This smaller resolution implies that there are less patches to be processed, and therefore it reduces the training cost and increases the performance. In that respect it effect is comparable to that of MAE [19]. We also report results with ViT-H 52 layers and ViT-H 26 layers parallel [49] models with 1B parameters. With lower resolution training it is easier to train these models.

Detailed Tables for Image classification In Table 13 we compare ViT architectures trained with our training recipes on ImageNet-1k with other architectures. In Table 14 we compare ViT architecture pre-trained on ImageNet-21k with our training recipe then finetuned on ImageNet-1k.

Model	epochs	ImageNet top1 acc. val	real	v2
ViT-S	300	79.9	86.1	68.8
	400	80.4	86.1	69.7
	600	80.8	86.7	69.9
	800	81.4	87.0	70.5
ViT-B	300	82.8	87.6	72.1
	400	83.1	87.7	72.6
	600	83.2	87.8	73.3
	800	83.7	88.1	73.1
ViT-L	300	84.1	88.5	74.1
	400	84.2	88.6	74.3
	600	84.4	88.6	74.6
	800	84.5	88.8	75.0
ViT-H	300	84.6	89.0	74.9
	400	84.8	89.1	75.3

Table 11: Impact on the performance of the number of training epochs on ImageNet-1k.

Training with others architectures In Table 15 we measure the top-1 accuracy on ImageNet-val, ImageNet-real and ImageNet-v2 with different architecture train with our training procedure at resolution 224×224 on ImageNet-1k only. We can observe that for some architectures like PiT or CaiT our training method will improve the performance. For some others like TNT our approach is neutral and for architectures like Swin it decreases the performance. This is consistent with the findings of Wightman et al. [56] and illustrates the need to improve the training procedure in conjunction to the architecture to obtain robust conclusions. Indeed, adjusting these architectures while keeping the training procedure fixed can probably have the same effect as keeping the architecture fixed and adjusting the training procedure. That means that with a fixed training procedure we can have an overfitting of an architecture for a given training procedure. In order to take overfitting into account we perform our measurements on the ImageNet val and ImageNet-v2 to quantify the amount of overfitting.

Semantic segmentation details The ADE20k dataset [62] consists of 20k training and 5k validation images with labels over 150 categories. For the training, we adopt the same schedule as in Swin: 160k iterations with UperNet [58]. Our UperNet implementation is based on the XCiT [15] repository. By default the UperNet head uses an embedding dimension of 512. In order to save compute, for small and tiny models we set it to the size of their working dimension, i.e. 384 for small and 192 for tiny. We keep the 512 by default as it is done in XCiT for other models.

Model	epochs		Resolution		ImageNet top-1 acc		
	Train.	FT	Train.	FT	val	real	v2
ViT-B	400	20	128 × 128	224 × 224	83.2	**88.1**	73.2
			160 × 160		83.3	88.0	**73.4**
			192 × 192		**83.5**	88.0	72.8
		-	224 × 224		83.1	87.7	72.6
	800	20	128 × 128	224 × 224	83.5	**88.3**	73.4
			160 × 160		83.6	88.2	73.5
			192 × 192		**83.8**	88.2	**73.6**
		-	224 × 224		83.7	88.1	73.1
ViT-L	400	20	128 × 128	224 × 224	83.9	**88.8**	74.3
			160 × 160		84.4	**88.8**	74.3
			192 × 192		**84.5**	**88.8**	**75.1**
		-	224 × 224		84.2	88.6	74.3
	800	20	128 × 128	224 × 224	84.5	**88.9**	74.7
			160 × 160		84.7	**88.9**	**75.2**
			192 × 192		**84.9**	88.7	75.1
		-	224 × 224		84.5	88.8	75.0
ViT-H	400	20	126 × 126	224 × 224	84.7	89.2	75.2
			154 × 154		**85.1**	**89.3**	75.3
			182 × 182		**85.1**	89.2	**75.4**
		-	224 × 224		84.8	89.1	75.3
	800	20	126 × 126	224 × 224	85.1	**89.2**	75.6
			154 × 154		**85.2**	**89.2**	**75.9**
			182 × 182		85.1	88.9	**75.9**
		-	224 × 224		84.9	89.1	75.6
ViT-H-52	400	20	126 × 126	224 × 224	84.9	89.2	75.6
ViT-H-26×2	400	20	126 × 126	224 × 224	84.9	89.1	75.3

Table 12: We compare ViT architectures pre-trained on ImageNet-1k only with different training resolution followed by a fine-tuning at resolution 224 × 224. We benefit from the FixRes effect [52] and get better performance with a lower training resolution (e.g resolution 160 × 160 with patch size 16 represent 100 tokens vs 196 for 224 × 224. This represents a reduction of 50% of the number of tokens).

Table 13: **Classification with ImageNet-1k training.** We compare architectures with comparable FLOPs and number of parameters. All models are trained on ImageNet-1k only without distillation nor self-supervised pre-training. We report Top-1 accuracy on the validation set of ImageNet1k and ImageNet-V2 with different measure of complexity: throughput, FLOPs, number of parameters and peak memory usage. The throughput and peak memory are measured on a single V100-32GB GPU with batch size fixed to 256 and mixed precision. For ResNet [20] and RegNet [37] we report the improved results from Wightman et al. [56]. Note that different models may have received a different optimization effort. ↑R indicates that the model is fine-tuned at the resolution R and -R indicates that the model is trained at resolution R.

Architecture	nb params ($\times 10^6$)	throughput (im/s)	FLOPs ($\times 10^9$)	Peak Mem (MB)	Top-1 Acc.	V2 Acc.
"Traditional" ConvNets						
ResNet-50 [20,56]	25.6	2587	4.1	2182	80.4	68.7
ResNet-101 [20,56]	44.5	1586	7.9	2269	81.5	70.3
ResNet-152 [20,56]	60.2	1122	11.6	2359	82.0	70.6
RegNetY-4GF [37,56]	20.6	1779	4.0	3041	81.5	70.7
RegNetY-8GF [37,56]	39.2	1158	8.0	3939	82.2	71.1
RegNetY-16GF [37,47]	83.6	714	16.0	5204	82.9	72.4
EfficientNet-B4 [43]	19.0	573	4.2	10006	82.9	72.3
EfficientNet-B5 [43]	30.0	268	9.9	11046	83.6	73.6
EfficientNetV2-S [44]	21.5	874	8.5	4515	83.9	74.0
EfficientNetV2-M [44]	54.1	312	25.0	7127	85.1	75.5
EfficientNetV2-L [44]	118.5	179	53.0	9540	85.7	76.3
Vision Transformers derivative						
PiT-S-224 [21]	23.5	1809	2.9	3293	80.9	_
PiT-B-224 [21]	73.8	615	12.5	7564	82.0	_
Swin-T-224 [31]	28.3	1109	4.5	3345	81.3	69.5
Swin-S-224 [31]	49.6	718	8.7	3470	83.0	71.8
Swin-B-224 [31]	87.8	532	15.4	4695	83.5	_
Swin-B-384 [31]	87.9	160	47.2	19385	84.5	_
Vision MLP & Patch-based ConvNets						
Mixer-B/16 [45]	59.9	993	12.6	1448	76.4	63.2
ResMLP-B24 [46]	116.0	1120	23.0	930	81.0	69.0
PatchConvNet-S60-224 [48]	25.2	1125	4.0	1321	82.1	71.0
PatchConvNet-B60-224 [48]	99.4	541	15.8	2790	83.5	72.6
PatchConvNet-B120-224 [48]	188.6	280	29.9	3314	84.1	73.9
ConvNeXt-B-224 [32]	88.6	563	15.4	3029	83.8	73.4
ConvNeXt-B-384 [32]	88.6	190	45.0	7851	85.1	74.7
ConvNeXt-L-224 [32]	197.8	344	34.4	4865	84.3	74.0
ConvNeXt-L-384 [32]	197.8	115	101.0	11938	85.5	75.3
Our Vanilla Vision Transformers						
ViT-S	22.0	1891	4.6	987	81.4	70.5
ViT-S↑384	22.0	424	15.5	4569	83.4	73.1
ViT-B	86.6	831	17.5	2078	83.8	73.6
ViT-B↑384	86.9	190	55.5	8956	85.0	74.8
ViT-L	304.4	277	61.6	3789	84.9	75.1
ViT-L↑384	304.8	67	191.2	12866	85.8	76.7
ViT-H	632.1	112	167.4	6984	85.2	75.9

Table 14: **Classification with Imagenet-21k training.** We compare architectures with comparable FLOPs and number of parameters. All models are trained on ImageNet-21k without distillation nor self-supervised pre-training. We report Top-1 accuracy on the validation set of ImageNet-1k and ImageNet-V2 with different measure of complexity: throughput, FLOPs, number of parameters and peak memory usage. The throughput and peak memory are measured on a single V100-32GB GPU with batch size fixed to 256 and mixed precision. For Swin-L we decrease the batch size to 128 in order to avoid out of memory error and re-estimate the memory consumption. ↑R indicates that the model is fine-tuned at the resolution R.

Architecture	nb params ($\times 10^6$)	throughput (im/s)	FLOPs ($\times 10^9$)	Peak Mem (MB)	Top-1 Acc.	V2 Acc.
"Traditional" ConvNets						
R-101x3↑384 [25]	388	_	204.6	_	84.4	_
R-152x4↑480 [25]	937	_	840.5	_	85.4	_
EfficientNetV2-S↑384 [44]	21.5	874	8.5	4515	84.9	74.5
EfficientNetV2-M↑480 [44]	54.1	312	25.0	7127	86.2	75.9
EfficientNetV2-L↑480 [44]	118.5	179	53.0	9540	86.8	76.9
EfficientNetV2-XL↑512 [44]	208.1	_	94.0	_	87.3	77.0
Patch-based ConvNets						
ConvNeXt-B [32]	88.6	563	15.4	3029	85.8	75.6
ConvNeXt-B↑384 [32]	88.6	190	45.1	7851	86.8	76.6
ConvNeXt-L [32]	197.8	344	34.4	4865	86.6	76.6
ConvNeXt-L↑384 [32]	197.8	115	101	11938	87.5	77.7
ConvNeXt-XL [32]	350.2	241	60.9	6951	87.0	77.0
ConvNeXt-XL↑384 [32]	350.2	80	179.0	16260	87.8	77.7
Vision Transformers derivative						
Swin-B [31]	87.8	532	15.4	4695	85.2	74.6
Swin-B↑384 [31]	87.9	160	47.0	19385	86.4	76.3
Swin-L [31]	196.5	337	34.5	7350	86.3	76.3
Swin-L↑384 [31]	196.7	100	103.9	33456	87.3	77.0
Vanilla Vision Transformers						
ViT-B/16 [41]	86.6	831	17.6	2078	84.0	_
ViT-B/16↑384 [41]	86.7	190	55.5	8956	85.5	_
ViT-L/16 [41]	304.4	277	61.6	3789	84.0	_
ViT-L/16↑384 [41]	304.8	67	191.1	12866	85.5	_
Our Vanilla Vision Transformers						
ViT-S	22.0	1891	4.6	987	83.1	73.8
ViT-B	86.6	831	17.6	2078	85.7	76.5
ViT-B↑384	86.9	190	55.5	8956	86.7	77.9
ViT-L	304.4	277	61.6	3789	87.0	78.6
ViT-L↑384	304.8	67	191.2	12866	87.7	79.1
ViT-H	632.1	112	167.4	6984	87.2	79.2

Model	Params ($\times 10^6$)	Flops ($\times 10^9$)	ImageNet-1k orig.	val	real	v2
ViT-S [47]	22.0	4.6	79.8	80.4	86.1	69.7
ViT-B [12,47]	86.6	17.6	81.8	83.1	87.7	72.6
PiT-S [21]	23.5	2.9	80.9	80.4	86.1	69.2
PiT-B [21]	73.8	12.5	82.0	82.4	86.8	72.0
TNT-S [18]	23.8	5.2	81.5	81.4	87.2	70.6
TNT-B [18]	65.6	14.1	82.9	82.9	87.6	72.2
ConViT-S [7]	27.8	5.8	81.3	81.3	87.0	70.3
ConViT-B [7]	86.5	17.5	82.4	82.0	86.7	71.3
Swin-S [31]	49.6	8.7	83.0	82.1	86.9	70.7
Swin-B [31]	87.8	15.4	83.5	82.2	86.7	70.7
CaiT-B12 [50]	100.0	18.2	-	83.3	87.7	73.3

Table 15: We report the performance reached with our training recipe with 400 epochs at resolution 224×224 for other transformers architectures. We have not performed an extensive grid search to adapt the hyper-parameters to each architecture. Our results are overall similar to the ones achieved in the papers where these architectures were originally published (reported in column 'orig.'), except for Swin Transformers, for which we observe a drop on ImageNet-val.

Crop.	LS	Mixup	Aug. policy	#Imnet21k epochs	finetuning resolution	Imagenet-1k val top-1 ViT-S	ViT-B	ViT-L	Imagenet-1k v2 top-1 ViT-S	ViT-B	ViT-L
RRC	✗	0.8	RA	90	224^2	81.6	84.6	86.0	70.7	74.7	76.4
SRC	✗	0.8	RA	90	224^2	82.1	84.8	86.3	71.8	75.0	76.7
SRC	✓	0.8	RA	90	224^2	82.4	85.0	86.4	72.4	75.7	77.4
SRC	✓	✗	RA	90	224^2	82.3	85.1	86.5	72.4	75.6	77.2
SRC	✓	✗	3A	90	224^2	82.6	85.2	86.8	72.6	76.1	78.3
SRC	✓	✗	3A	240	224^2	**83.1**	**85.7**	**87.0**	**73.8**	**76.5**	**78.6**
SRC	✓	✗	3A	240	384^2	*84.8*	*86.7*	*87.7*	75.1	77.9	79.1

Table 16: Ablation path: **augmentation and regularization** with ImageNet-21k pre-training (at resolution 224×224) and ImageNet-1k fine-tuning. We measure the impact of changing Random Resize Crop (RRC) to Simple Random Crop (SRC), adding LayerScale (LS), removing Mixup, replacing RandAugment (RA) by 3-Augment (3A), and finally employing a longer number of epochs during the pre-training phase on ImageNet-21k. All experiments are done with Seed 0 with fixed hparams except the drop-path rate of stochastic depth, which depends on the model and is increased by 0.05 for the longer pre-training. We report 2 digits top-1 accuracy but note that the standard standard deviation is around 0.1 on our ViT-B baseline. Note that all these changes are neutral w.r.t. complexity except in the last row, where the fine-tuning at resolution 384×384 significantly increases the complexity.