Acknowledgements

We are grateful to Kaiming He for invaluable discussions and suggestions, and his advice around MAE pre-pretraining. We thank Mary Williamson for her help, guidance and support in project planning, execution, and managing various uncertainties throughout the research project. We are grateful to Vivek Pai for his help with the training infrastructure. We also thank Yanghao Li for his help with the detection evaluations, Dominik Kallusky for his help with the dataset pipeline, and Tsung-Yu Lin for his help with preparing the image-caption dataset. Lastly, we thank Stephen Roller and Naman Goyal for helpful discussions and feedback.

Appendix

A. Pretraining details

Model architectures. To ensure we were able to scale models out further than ViT-H, and that the models can be easily scaled beyond a few billion parameters, we decided to scale models along the same lines as GPT-3 [10], which has proven to be successful for NLP. We share the exact settings for all the models discussed in the paper in Table 8.

Arch.	Layers	Embed	MLP	Heads	Params
ViT-B	12	768	3072	12	86M
ViT-L	24	1024	4096	16	307M
ViT-H	32	1280	5120	16	632M
ViT-2B	24	2560	10240	32	1.89B
ViT-6.5B	32	4096	16384	32	6.44B

Table 8: Model architecture details.

MAE pretraining. We make no changes from He *et al*. [33] for MAE pretraining. We utilize the same decoder dimensions as well – 8 layers, 512 dimension, and 16 heads. Training hyperparameters are shared in Table 9.

WSP and MAE→WSP pretraining. We note that large scale WSP pretraining is quite robust to the choice of training hyperparameters, including the choice of (i) a single *vs*. two layer MLP head, (ii) softmax cross-entropy *vs*. binary cross-entropy training loss, (iii) additional training augmentations like mixup [86], CutMix [83], *etc*., (iv) learning rate over a $\sim 5\times$ range. Most of these choices seem to affect results at small scales, like 0.1 epochs over IG-3B (500 million samples seen), but the differences dissipate when training for a full epoch (5 billion samples seen). Our training hyperparameters are shared in Table 10, and we use these settings for both WSP and MAE→WSP. We follow Singh *et al*. [70] for the training setup and hyperparameters for consistency and easier comparisons. For a label vocabulary C, and an image img with labels $L_{img} \in \{0,1\}^{|C|}$, we utilize softmax cross-entropy, where our label is normalized to a probability distribution, $L_{im} / \sum_{c \in C} L^c_{im}$, and the model output is passed to a softmax followed by a cross-entropy loss [52, 70]. We attach a two layer MLP head to the trunk – `Linear(embed, embed), Tanh(), Linear(embed, classes)`.

Setting	Value
Epochs	1
Batch size	4096
Masking ratio	0.75
Optimizer	AdamW [51]
Learning rate:	
Schedule	Cosine
Peak	2.4e-3
Warmup Schedule	Linear
Warmup Fraction	5%
Weight decay	0.05
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$
Augmentations:	
`RandomResizedCrop`	
`size`	224px
`scale`	[0.2, 1.00]
`ratio`	[0.75, 1.33]
`interpolation`	Bicubic
`RandomHorizontalFlip`	$p = 0.5$
`Normalize`	

Table 9: MAE Pretraining hyperparameters. We follow the settings from [33] without any modifications.

Setting	Value
Epochs	1
Batch size	8192
Optimizer	AdamW [51]
Learning rate:	
Schedule	Linear
Peak	4e-4
Warmup Schedule	Linear
Warmup Fraction	5%
Weight decay	0.1
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
Augmentations:	
`RandomResizedCrop`	
`size`	224px
`scale`	[0.08, 1.00]
`ratio`	[0.75, 1.33]
`interpolation`	Bicubic
`RandomHorizontalFlip`	$p = 0.5$
`Normalize`	

Table 10: WSP Pretraining hyperparameters. We follow the settings from [70] without any modifications. For ViT-2B we were able to train with the same hyperparameters, but ultimately reduced the learning rate to 1e-4, enabled gradient clipping to 1.0 norm, and set AdamW's β_2 to 0.95 for improved training stability.

LiT training. We follow LiT to train a text encoder on Instagram captions. We sanitize the captions and also remove the pound signs in front of hashtags. We use a context length (number of tokens) of 100 per caption. Following CLIP, we chose the embedding dimension for aligning the encoders to be 512, 768 and 1024 for ViT-B, ViT-L and ViT-H, respec-

Setting	Value
Epochs	1
Batch size	32768
Optimizer	AdamW [51]
Learning rate:	
Schedule	Cosine
Peak	2e-4
Warmup Schedule	Linear
Warmup Fraction	4%
Weight decay	0.1
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.98$
Loss:	
`CLIP` [59]	
`LabelSmoothing` [73]	0.1
Augmentations:	
`RandomResizedCrop`	
`size`	224px
`scale`	[0.9, 1.00]
`ratio`	[0.75, 1.33]
`interpolation`	Bicubic
`RandomHorizontalFlip`	$p = 0.5$
`Normalize`	

Table 11: LiT training hyperparameters for Table 5. For ablations with XLM-R Base we used a higher learning rate of 1e-3 with a dropout of 0.1.

Setting	Value
Epochs	90
Batch size	4096
Optimizer	AdamW [51]
Learning rate:	
Schedule	Cosine
Peak	1e-3
Warmup Schedule	Linear
Warmup Fraction	3.3%
Weight decay	0.1
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
Gradient Clipping	1.0
Augmentations:	
`RandomResizedCrop`	
`size`	224px
`scale`	[0.08, 1.00]
`ratio`	[0.75, 1.33]
`interpolation`	Bicubic
`RandomHorizontalFlip`	$p = 0.5$
`RandomAugment` [18]	
`num_layers`	2
`magnitude`	9
`Normalize`	
`mixup` [86]	0.5

Table 12: WSP Pretraining hyperparameters for IN21k.

tively, and set it to 2048 for ViT-2B. We use a pretrained XLM-R [17] text encoders, with the Base size (270M parameters) for ablations and Large (550M parameters) for the results in Table 5. Our LiT training hyperparameters are shared in Table 11.

ImageNet-21k pretraining. In IN21k each image is labeled with one class, but the classes are based on WordNet synsets [53] which are hierarchical in nature. Some images in this dataset are duplicated across more than one class – we deduplicate the images by hashing the image contents, and convert the dataset to a multi-label dataset. We disregard the class hierarchy amongst the labels and treat them independently.

For MAE pretraining we again follow [33] and use the hyperparameters in Table 9 and train for 160 epochs over the dataset. For WSP (and MAE→WSP) pretraining on ImageNet-21k, we use the hyperparameters from Steiner *et al*. [72], with a few minor differences. We train for 90 epochs over the dataset, and select the augmentation setting *medium2* of the paper. We train with a softmax cross-entropy loss similar to IG-3B pretraining, but utilize a head with just a single linear layer. Full hyperparameters are in Table 12.

For LiT finetuning of models pretrained on IN21k, we follow a similar setup and hyperparameters used for IG-3B (Table 11), and train for 20 epochs on PMD [69]. Unlike Instagram-3B dataset which has 5 billion image-text pairs, PMD has only about 70 million image-text pairs, so we train for multiple epochs on this dataset.

B. Transfer learning details

Setting	Value
Batch size	1024
Optimizer	AdamW [51]
Learning rate:	
Schedule	Constant
Peak	2e-3
Layerwise decay [5, 16]	0.75
Warmup Schedule	Linear
Warmup Fraction	5%
Weight decay	0.05
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
DropPath [39]	0.2
EMA [58]	1e-4
Augmentations:	
`RandomResizedCrop`	
`size`	518px
`scale`	[0.08, 1.00]
`ratio`	[0.75, 1.33]
`interpolation`	Bicubic
`RandomHorizontalFlip`	$p = 0.5$
`Normalize`	
`mixup` [86]	0.8
`CutMix` [83]	1.0
`LabelSmoothing` [73]	0.1

Table 13: MAE Image finetuning hyperparameters. We train for 50 epochs on IN1k and 100 epochs on iNat18.

Image classification details. We finetune models pretrained with MAE, WSP and MAE→WSP by attaching a single linear layer on IN1k and iNat18. We finetune models at either

Setting	Value
Batch size	2048
Optimizer	SGD
Learning rate:	
Schedule	Constant
Peak	2.4e-2
Warmup Schedule	Linear
Warmup Fraction	5%
Weight decay	0.0
Optimizer Momentum	0.9
Gradient Clipping	1.0
DropPath [39]	0.2
EMA [58]	1e-4
Augmentations:	
`RandomResizedCrop`	
`size`	518px
`scale`	[0.08, 1.00]
`ratio`	[0.75, 1.33]
`interpolation`	Bicubic
`RandomHorizontalFlip`	$p = 0.5$
`Normalize`	
`mixup [86]`	0.1

Table 14: WSP Image finetuning hyperparameters. We train for 50 epochs on IN1k and 150 epochs on iNat18.

224×224 resolution or 518×518 resolution (or 512×512 for models which use a patch size of 16) and use the same hyperparameters at all resolutions. The hyperparameters for high resolution finetuning are shared in Table 13 for MAE models and in Table 14 for models pretrained with WSP or MAE→WSP.

Video classification details. For video finetuning, we sample 32 frames out of 2.7 second clips for Kinetics-400 and 4 second clips for Something Something-v2, and train the models at 224×224 resolution. We convert the input into patches of size $2 \times 16 \times 16$, akin to MAE approaches applied to video models [25, 28, 75]. We initialize the video models with weights from our inflated [13] pretrained image models. Table 15 contains the finetuning hyperparameters for both K400 and SSv2.

Low-shot image classification details. We adapt all MAE→WSP and SWAG models with VPT [42] with the same hyperparameters – 8 tokens of size 192 per self attention layer, as this proved to be a reasonable default setting. The full settings for training our models with VPT are described in Table 16.

We also attempted to train CLIP and OpenCLIP models with Adapters, but however noticed that they do not work as great with VPT as they do with logistic regression, despite sweeping a range of different hyperparameters. We ultimately adopt the logistic regression protocol of MSN [3] for CLIP and OpenCLIP, and also for DINO and MSN. For MAE low-shot evaluations, we noticed that neither Adapters nor logistic regression led to great results, something already seen in MSN [3]. We therefore follow MSN and finetune

Setting	Value	
	K400	SSv2
Epochs	110	100
Batch size	256	512
Optimizer	AdamW [51]	
Learning rate:		
Schedule	Constant	
Peak	1e-4	2e-3
Layerwise decay [5, 16]	0.75	
Warmup Schedule	Linear	
Warmup Fraction	20%	
Weight decay	0.05	
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	
Gradient Clipping	1.0	
Dropout [71]	–	0.5
DropPath [39]	0.2	0.4
EMA [58]	1e-4	
Augmentations:		
`ShortSideScale`	256px	
`RandomResizedCrop`		
`size`	224px	
`scale`	[0.08, 1.0]	
`ratio`	[0.75, 1.33]	
`interpolation`	Bicubic	
`RandomAugment [18]`		
`magnitude`	7	
`num_layers`	5	
`RandomErasing [88]`	$p = 0.25$	
`Normalize`	Yes	
`mixup [86]`	0.8	
`CutMix [83]`	1.0	
`LabelSmoothing [73]`	0.1	

Table 15: Video finetuning hyperparameters. We used these settings for Table 3. For Figure 1 and Figure 6 we used shorter 50 epoch schedules for efficiency reasons, which has minimal impact on performance (< 0.5%).

MAE in low-shot settings, but improve upon the results published in [3] for MAE significantly. All these results are reported in Table 4.

Zero-shot transfer details. For evaluation of our LiT models, we follow the zero-shot evaluation strategy proposed in CLIP [59]. We leverage the prompt templates and class names introduced in CLIP for ImageNet-1k and Food-101. We compute the cosine similarity between the query image and all the generated prompts for each class. While doing so, we also take advantage of the class prompt ensembling approach introduced in CLIP by considering multiple prompt templates for each class.

Detection and segmentation details. We train all of our models with the Cascade Mask-RCNN [35] framework and use the ViTDet [47] architecture to leverage our ViT models within this framework.

For our MAE models trained on IG data, we use the hyperparameters of MAE trained on IN1k from [47]. For the large ViT-2B model, we use settings similar to ViT-H and

Setting	Value	
	1-shot	{5, 10}-shot
Epochs	56	28
Batch size	128	
Optimizer	SGD	
Learning rate:		
Schedule	Cosine	
Peak	6e-3	
Weight decay	0.0	
Optimizer Momentum	0.9	
Augmentations:		
`ShortSideScale`	256px	
`RandomResizedCrop`		
`size`	224px	
`scale`	[0.08, 1.0]	
`ratio`	[0.75, 1.33]	
`interpolation`	Bicubic	
`RandomHorizontalFlip`	$p = 0.5$	
`Normalize`	Yes	
`mixup` [86]	0.1	
VPT:		
`NumTokens`	8	
`DimTokens`	192	
`TokensDropout`	0.0	

Table 16: Low-shot hyperparameters. Settings used for our adapting our models with VPT [42]. For the ViT-2B, we decrease the learning rate to 1e-3, double the training time and reduce the batch size to 64.

Setting	WSP			MAE
	ViT-L	ViT-H	ViT-2B	ViT-2B
Epochs	100	100	100	100
Image Size (px)	1024	1024	1024	1024
Learning rate				
Peak	1e-4	1e-4	1e-4	2e-4
Schedule	Step	Step	Step	Step
Layer Decay	0.8	0.85	0.8	0.8
Min Layer Decay	0.1	0.1	0.1	–
Optimizer				
AdamW β_1	0.9	0.9	0.9	0.9
AdamW β_2	0.999	0.998	0.999	0.999
Drop Path rate	0.4	0.4	0.4	0.5
Weight Decay	0.2	0.1	0.1	0.1

Table 17: Detection parameters used for LVIS. For MAE on ViTDet architectures smaller than ViT-H, we use the best set of hyper-parameters reported in the original ViTDet [47] work.

only change the layer decay to be the same as ViT-L, since both those models have 24 transformer layers.

For WSP and MAE→WSP pretraining, we adapt the parameters used for MAE pretraining. The most salient change is a modification of the layer decay scheme of MAE to cap it at a minimum value of 0.1, allowing WSP pretrained models to update the initial layers to align better for detection and segmentation tasks. This was particularly useful for instance segmentation.

The hyperparameters used for training our detection and instance segmentation models are available in Table 17 for LVIS and Table 18 for COCO.

Setting	WSP			MAE
	ViT-L	ViT-H	ViT-2B	ViT-2B
Epochs	100	75	75	75
Image Size (px)	1024	1024	1024	1024
Learning rate				
Peak	1e-4	1e-4	1e-4	1e-4
Layer Decay	0.8	0.85	0.8	0.8
Min Layer Decay	-	0.1	0.1	-
Optimizer				
AdamW β_1	0.9	0.9	0.9	0.9
AdamW β_2	0.999	0.999	0.999	0.999
Drop Path rate	0.4	0.4	0.5	0.5
Weight Decay	0.1	0.1	0.1	0.1

Table 18: Detection parameters used for COCO. For MAE on ViTDet architectures smaller than ViT-H, we use the best set of hyper-parameters reported in the original ViTDet [47] work.

C. Additional Results and Full Tables

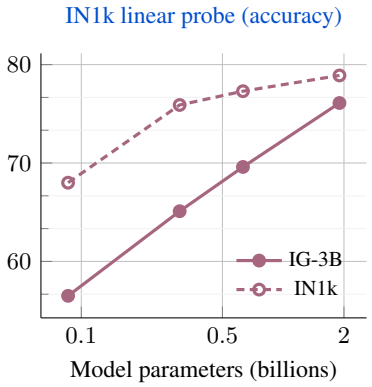

Figure 9: Scaling MAE with model and dataset size. MAE's linear performance on IN1k when pretrained on IN1k or IG, while scaling model size. MAE IN1k starts off much better, since the pretraining is performed on the same dataset used for linear probing. But the scaling behavior for MAE-IG-3B is exciting, showing linear performance improvements as we scale model sizes, with the potential to outperform IN1k pretraining if scaled out further.

Arch.	MAE	WSP	MAE→WSP
ViT-B	56.5	82.1	82.8
ViT-L	65.1	84.9	86.0
ViT-H	69.6	86.1	87.0
ViT-2B	76.1	86.9	88.1
ViT-6.5B	–	87.8	–

Table 22: Linear classifier results on IN1k at 224px resolution. Even though MAE's linear performance is not remarkable, MAE→WSP benefits from MAE pre-pretraining and outperforms WSP by a wide margin.

Arch.	IN1k Top-1	iNat18 Top-1	LVIS AP^{box}	LVIS AP^{mask}	COCO AP^{box}	COCO AP^{mask}
IN1k pretraining						
ViT-B	83.5	75.0	43.0	38.9	54.0	46.7
ViT-L	86.0	80.2	49.2	44.5	57.6	50.0
ViT-H	86.9	82.8	**51.5**	**46.6**	**58.7**	**51.0**
ViT-2B	**87.4**	**84.5**	–†	–†	–†	–†
IG-3B pretraining						
ViT-B	83.5	74.7	42.9	38.8	53.8	46.5
ViT-L	86.1	80.7	49.0	44.5	58.0	50.3
ViT-H	87.4	84.0	52.7	47.5	59.1	51.2
ViT-2B	**87.8**	**85.6**	**53.6**	**48.6**	**59.9**	**52.0**

Table 19: Scaling MAE with model and dataset size. We show the performance in tabular form of MAE pretraining on ImageNet-1k and Instagram-3B, earlier shown in Figure 2. MAE scales with both data and model size. IN1k and iNat18 finetuning results are at 224px resolution. †Training was unstable.

Method	Image Classification						Video Cls.		Detection	
	IN1k	iNat18	IN1k 5-shot	iNat18 5-shot	IN1k Zero-shot	IN1k Linear	K400	SSv2	LVIS	COCO
IG-3B Pretraining										
MAE	87.2	86.5	37.2	43.8	46.3	65.1	82.7	73.8	**58.0**	49.0
WSP	87.6	86.6	75.7	62.7	77.2	84.9	83.7	69.8	54.9	47.1
MAE→WSP	**88.8**	**89.0**	**77.9**	**65.2**	**78.3**	**86.0**	**85.6**	**74.1**	57.3	**49.6**
IN21k Pretraining										
MAE	86.9	86.1	37.1	43.4	41.7	68.8	80.5	**72.6**	**57.2**	**47.7**
WSP	84.9	85.4	76.3	58.9	69.8	81.3	80.2	65.2	53.3	43.9
MAE→WSP	**87.1**	**87.6**	**78.7**	**63.9**	**72.7**	**84.7**	**83.9**	71.5	56.3	47.5

Table 20: MAE pre-pretraining improves performance. We show in tabular form the transfer performance of a ViT-L pretrained with MAE, WSP, and MAE→WSP using IG-3B (earlier shown in Figure 1) and IN21k (earlier shown in Figure 6). We **bolden** the best results and underline the second best results for easier comparison. Regardless of whether WSP or MAE performs better, MAE→WSP can match or even outperform either approaches. IN1k and iNat18 finetuning results are at 512px resolution.

MAE Dataset	WSP Dataset	Arch.	IN1k	iNat18
IG-3B	IG-3B	ViT-H	89.3	90.5
IN1k	IG-3B	ViT-H	89.4	90.5

Table 21: Different datasets for pre-pretraining. MAE pre-pretraining is just as effective when trained with the small IN1k dataset with a different size and distribution than the pretraining dataset (IG-3B).

Method	Arch.	AP	Cls	Loc	Both	Dupl	Bkgd	Miss
Results with IoU threshold of 0.5								
WSP	ViT-L	56.4	11.46	6.09	1.5	0.16	11.64	12.34
MAE	ViT-L	57.4	13.87	5.29	1.21	0.13	11.37	10.75
MAE→WSP	ViT-L	58.4	11.73	5.64	1.69	0.17	10.82	11.54
Results with IoU threshold of 0.75								
WSP	ViT-L	45.0	8.24	17.96	2.02	0.0	10.76	16.04
MAE	ViT-L	47.5	10.98	15.57	1.49	0.0	10.92	13.52
MAE→WSP	ViT-L	47.1	9.15	17.00	2.12	0.0	10.47	13.85

Table 23: TIDE [8] analysis on LVIS for our MAE, WSP and MAE→WSP pretrained ViT-L models, evaluated with two different IoU thresholds. The error contributions are computed with the *progressive* mode of TIDE. Overall, MAE is stronger at finding and localisation of objects while WSP is stronger at classifying boxes. MAE→WSP strikes a balanced between both.

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. *arXiv preprint arXiv:2110.02095*, 2021.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *CVPR*, 2021.

[3] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022.

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023.

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[6] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019.

[7] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[8] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *ECCV*, 2020.

[9] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*. Springer, 2014.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[13] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[14] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[16] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *ICLR*, 2020.

[17] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020.

[18] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2020.

[19] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, 2021.

[20] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *CVPR Workshop*, 2019.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[22] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[24] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021.

[25] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

[26] WeiFu Fu, CongChong Nie, Ting Sun, Jun Liu, TianLiang Zhang, and Yong Liu. Lvis challenge track technical report 1st place solution: distribution balanced and boundary refinement for large vocabulary instance segmentation. *arXiv preprint arXiv:2111.02668*, 2021.

[27] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019.

[28] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *CVPR*, 2023.

[29] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2013.

[30] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[31] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020.

[32] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[33] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[35] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[36] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

[37] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

[38] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[39] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016.

[40] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022.

[41] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.

[42] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.

[43] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, AMustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[44] S Kornblith, J Shlens, and QV Le. Do better imagenet models transfer better? arxiv 2018. In *CVPR*, 2019.

[45] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[46] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022.

[47] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022.

[48] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnet: A composite backbone network architecture for object detection. *IEEE Transactions on Image Processing*, 2022.

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[50] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022.

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[52] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

[53] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[54] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, 2022.

[55] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[57] Karol J Piczak. Esc: Dataset for environmental sound classification. In *ACM MM*, 2015.

[58] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 1992.

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[60] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshops*, 2014.

[61] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[62] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[63] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS*, 2021.

[64] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[65] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.

[66] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *ICCV*, 2021.

[67] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019.

[68] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[69] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022.

[70] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*, 2022.

[71] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[72] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[73] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[74] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

[75] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[77] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010.

[78] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.

[79] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

[80] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. *arXiv preprint arXiv:2203.11926*, 2022.

[81] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022.

[82] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[83] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[84] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022.

[85] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022.

[86] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[87] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2022.

[88] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020.

[89] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022.

[90] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. *arXiv preprint arXiv:2211.12860*, 2022.