Appendix

A. Additional Dataset Details

Hashtag Filtering and Canonicalization. We considered the set of all hashtags $\mathcal{H}$ posted by US users more than once in public posts as our candidate set. We design a many-to-one function to map a hashtag to WordNet synsets [22], $s : \mathcal{H} \rightarrow 2^{\mathcal{S}}$, where $\mathcal{S}$ is the set of WordNet synsets, and $2^{\mathcal{S}}$ is the power set of $\mathcal{S}$. s is defined as the `get_synsets()` Python function in Listing 1. We filter out hashtags which map to $\emptyset$, and consider all hashtags which map to the same set of synsets as the same label. For instance, `#eggplant` and `#aubergine` map to the same target, whereas `#newyork` is filtered out. We finally convert the output of $f(h)$, a set of synsets, to a text string, which we refer to as a *canonical hashtag*. We refer to the set of all canonical hashtags, which is our output vocabulary, by C.

```
from nltk.corpus import wordnet

MIN_LEN = 3
ALLOWED_SENSES = {
  "noun.animal",
  "noun.artifact",
  "noun.food",
  "noun.object",
  "noun.plant",
  "noun.event",
}


def get_synsets(hashtag):
  if len(hashtag) < MIN_LEN:
    return set()

  candidates = {wordnet.morphy(hashtag, wordnet.NOUN)}
  for i in range(MIN_LEN, len(hashtag) - MIN_LEN + 1):
    candidate = hashtag[:i] + "_" + hashtag[i:]
    candidates.add(wordnet.morphy(candidate))

  synsets = set()
  for word in candidates:
    if word is None:
      continue
    for synset in wordnet.synsets(word):
      if synset._lexname in ALLOWED_SENSES:
        synsets.add(synset)
  return synsets
```

Listing 1. Hashtag-to-synset mapping code in Python.

Image Sampling. We follow [49] and down-weight the relative weight of frequent hashtags. For deciding our sampling weights for images, we assign importance factors to each image based on the (canonical) hashtags associated with it. For a hashtag $h \in \mathcal{C}$, its importance factor, I_h, is defined as $f(h)^{-1/2}$, where $f(h)$ is the hashtag's frequency. For an image i, with associated hashtags $\{h_i^j\}$, we define the image's importance factor as $I_i = \max I_{h_i^j}$. Next, we partition the hashtags into two sets – the head, which contains hashtags which occur in at least 5000 images, and the tail which contains the remaining infrequent hashtags (see Figure 1). An image is considered a tail image iff it contains at least one tail hashtag.

We sample images from the set of all images available to us, $\mathcal{I}$, using the probability distribution $p_i = cI_i \ \forall i \in \mathcal{I}$, where c is a normalization constant. We continue sampling images independently until we reach our target dataset's total samples. For a target number of samples, M, we sample αM samples from the head and $(1-\alpha)M$ samples from the tail using this sampling procedure (we chose $\alpha = 0.7$). We note that because the tail is heavily upsampled, the number of unique images in a single epoch is smaller than the total samples M.

The deviations from [49] in dataset collection were ablated by pre-training on datasets of 100 million samples and evaluating linear classifier performance on ImageNet-1k, see Table 4 for details. Per the results in the table, our changes boost transfer performance on ImageNet-1K by 2.1% when pretraining a ResNeXt-101 32x8d on a 100M dataset. This number might change as we increase the size of the dataset or model.

Deduplication. [49] performed an extensive deduplication experiment, which suggests that the percentage of images in common evaluation datasets that appears on Instagram is very small ($< 0.5\%$) and, in fact, smaller than the overlap between those evaluation datasets and the ImageNet training set that is commonly used for model pre-training. While our sampling methodologies may differ, based on those observations, we chose not to repeat the deduplication experiments.

Approach	Hashtag vocabulary size	Head-tail sampling ratio (α)	IN-1k Accuracy
Baseline [49]	17K	-	74.9
+ Head-tail sampling	17K	0.7	76.6
	17K	0.5	76.3
	17K	0.3	75.5
+ Larger hashtag vocab.	27K	0.7	77.0

Table 4. **Dataset ablations**. Ablation study on training set collection using ResNeXt-101 32x8d models trained on IG datasets with 100M unique images; we report the linear classifier accuracy on ImageNet-1k. The baseline approach follows the dataset collection approaches in [49] and reproduces the results in that paper. Partitioning the hashtags and over-sampling the tail ($\alpha = 0.7$) improves transfer accuracy significantly, but excessively oversampling the tail ($\alpha = 0.3$) worsens it. Increasing the hashtag vocabulary size improves transfer accuracy.

B. Model Complexity and Speed

Table 5 presents the resolution, FLOPs, number of parameters, number of activations, and train and test throughputs of all models used in our study.

Model	Resolution	Flops (B)	Params (M)	Acts (M)	Train (images/sec.)	Test (images/sec.)
EfficientNet L2	475	172.6	480.3	609.9	49	293
EfficientNet L2	800	479.9	480.3	1707.4	19	108
EfficientNet B8	672	63.7	87.4	442.9	103	480
EfficientNet B7	600	38.4	66.3	289.9	157	652
EfficientNet B6	528	19.5	43.0	167.4	246	849
ViT G/14	224	484.2	1844.4	275.4	- †	379
ViT G/14	518	2826.1	1846.3	2639.0	- †	56
ViT H/14	224	167.5	632.0	139.4	246	960
ViT H/14	392	545.9	632.7	638.0	56	242
ViT H/14	518	1018.8	633.5	1523.9	19	116
ViT L/16	224	61.7	304.3	63.5	701	2092
ViT L/16	384	191.5	304.7	270.2	177	567
ViT L/16	512	362.9	305.2	656.4	70	255
ViT B/16	224	17.6	86.6	23.9	2247	3861
ViT B/16	384	55.6	86.9	101.6	549	1161
ViT L/32	224	15.4	306.5	13.3	3176	4431
ViT L/32	384	54.4	306.6	43.9	921	1439
RegNetY 128GF	224	127.7	644.8	71.6	191	879
RegNetY 128GF	384	375.2	644.8	210.2	69	307
RegNetY 32GF	224	32.6	145.0	30.3	607	2824
RegNetY 32GF	384	95.1	145.0	88.9	248	976
RegNetY 16GF	224	16.0	83.6	23.0	989	4562
RegNetY 16GF	384	47.0	83.6	67.7	440	1401

Table 5. **Model complexity and speed.** Complexity and speed of models with an ImageNet-1k head at relevant resolutions. We measure train and train and test speed on a single node with 8 V100 32GB GPUs, maximizing the batch size for each model. Although EfficientNets have very few FLOPs, they produce a large amount of activations resulting in much slower train / test speeds. Training speeds measured for convolutional networks using SGD, and for ViTs using AdamW [47]. †We were unable to train a ViT G/14 using our setup, even with a batch size of 1.

Model	Res.	FLOPs (B)	Param. (M)	Act. (M)	Throughput Train	Throughput Test	Classification accuracy IN-1k	IN-5k	IG-0.7B → IN-1k
ResNeXt-101 32x4d	224	8.0	49.0	21.3	2,222	5,214	<u>79.1</u>	<u>50.9</u>	<u>80.0</u>
DenseNet-264	224	5.9	33.4	8.5	1,813	5,116	76.6	47.9	78.4
EfficientNet B3	300	1.9	12.2	23.8	1,802	2,979	78.5	49.3	77.9
RegNetY 8GF	224	8.0	39.2	18.0	1,770	4,562	**79.8**	**51.4**	**80.8**

Table 6. Overview of the convolutional models we evaluated for our experiments. ResNeXt and DenseNet models were augmented with squeeze-and-excitation (SE [34]) layers. We evaluate the classification accuracy of the models in three settings: (1) training on ImageNet-1k; (2) training on ImageNet-5k; and (3) pre-training 1 epoch on 1B examples of IG-0.7B followed by linear classifier evaluation on ImageNet-1k. We find that the RegNetY model performs best in all settings. The best result on each dataset is **boldfaced**; the second-best result is <u>underlined</u>. Higher is better.

C. Model and Hyperparameter Selection

C.1. Effect of Convolutional Model Family

We performed experiments investigating the performance of four different model families in weakly-supervised pre-training: ResNeXt [74], RegNetY [58], DenseNet [35] and EfficientNet [65]. As recent model families like EfficientNet and RegNetY use squeeze-and-excitation (SE) layers [34] for improved accuracies [34, 58], we also use these in our implementations of DenseNet and ResNeXt.

Since we trained our models at scale, our goal was to identify the most efficient model family in terms of the accuracy achieved with a fixed training budget. In line with this goal, instead of finding models with comparable numbers of FLOPs or parameters, which have been shown to correlate poorly with training speed [18, 58], we instead measured image throughput during training. We also include the test time throughput as well since it is a useful inference time constraint to consider.

To keep the experiments tractable, we used medium-sized models of each model family. Table 6 lists the candidate models for each of the families we used for our comparison; these models were selected to have similar training speeds (in terms of images processed per second). We note that the test throughputs were also similar except for the EfficientNet model which uses a higher resolution than the other models.

We tested the models in three settings: (1) training and testing on ImageNet-1k, (2) training and testing on ImageNet-5k, and (3) pre-training on IG-1B followed by a linear classifier trained and tested on ImageNet-1k. The results of our experiments are presented in Table 6. The results show that for a similar training budget, the RegNetY model family outperforms the other model families on all three datasets, while also having a competitive inference throughput. For that reason, we focused on RegNetY models in all subsequent experiments.

C.2. Effect of Dataset Size

During pre-training, usually the focus is on the the total number of unique images in the dataset, which we will refer to as the dataset *size* [20, 49, 73, 76]. In our setup, due to the upsampling of the least frequent hashtags, our final dataset is defined by an additional parameter – the dataset's *samples*, which we define as the total number of image-label pairs, counting duplicates. Table 7 shows the effect of a dataset's number of unique images vs the total samples seen during training. For the IG dataset in the smaller test regimes we explored, the total samples seen determined the model performance across a variety of dataset sizes for different model families (convolutional networks, transformers) and model capacities, rather than the number of unique images seen. We hypothesize that this is because in this regime the model has not yet saturated. It does suggest that the total number of samples seen during training is important to consider when comparing large datasets with a small number of epochs.

C.3. Effect of Scaling Parameters

Due to the inherent noise in the learning signal, weakly supervised pre-training requires substantial scale to obtain optimal results. We performed experiments studying the effect on the transfer performance of two key scaling param-

Dataset Name (size)	Samples	Epochs	IN-1k transfer accuracy RegNetY 8GF	RegNetY 32GF	ViT B/16	ViT L/16
IG-0.2B	250M	8	81.5	83.7	80.5	83.2
IG-0.4B	500M	4	81.5	83.8	80.7	83.5
IG-0.7B	1B	2	81.4	83.8	80.3	83.5
IG-1.4B	2B	1	81.3	83.7	80.5	83.4

Table 7. **Effect of dataset size**. We compute the linear classifier accuracy of various models on ImageNet-1k to study the effect of unique images. Every data point corresponds to the same number of total samples trained (2 billion), but the dataset size varies.

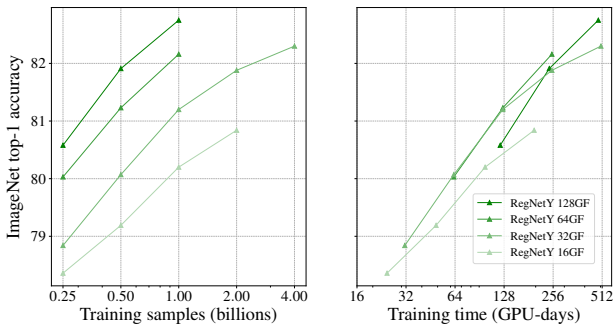

Figure 3. **Scaling model and dataset sizes.** ImageNet top-1 linear classifier accuracy for various model sizes as a function of the number of pre-training samples (**left**) and the training budget (**right**). As we go larger in model size, the models become more efficient in utilizing a given number of pre-training samples, and additional training samples improve performance. Training time calculated by dividing the total samples with the training speeds from Table 5.

eters: (1) model scale and (2) training set scale. To vary the model scale, we train RegNetY models that were independently optimized for use, starting from 16 GFLOPs, up to 128 GFLOPs. We followed [58] and searched for each of the models instances on ImageNet-1k. To vary the training set scale, we use IG datasets of varying sizes. We train all models for one full epoch, and measure linear classifier performance on ImageNet-1k.

The results of our experiments are presented in Figure 3. We present the transfer accuracy as a function of both the total samples seen and the total training time in GPU-days, for four different models. The results presented in Figure 3 are largely in line with those of [49, 76]. Specifically, transfer accuracy improves for both larger models and for longer training regimes. Akin to [49], we find that the larger models benefit from more training samples than their smaller counterparts: the slope of the accuracy curve of RegNetY 128GF is steeper than that of RegNetY 16GF. Thus, for a large enough training budget it makes sense to use a larger model rather than a smaller model trained on more samples.

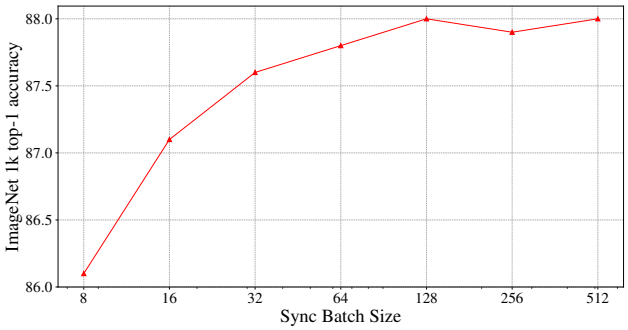

Figure 4. **Effect of sync batch norm while fine-tuning.** ImageNet top-1 accuracy while fine tuning a RegNetY 128GF continues to increase as we increase the sync batch size. The model was always trained with a mini batch size of 512, while varying the batch sizes for sync batch norm. Results reported without EMA.

D. Training Details

All our models were trained with *Classy Vision* [2]. For the transfer results for other works in Table 1, we used the *timm* library [71] to get pre-trained checkpoints. In this section we share details about our fine-tuning setup for Table 1, viz., the learning rate used (Table 8), and the utility of using synchronized batch normalization for convolutional networks (Figure 4).

Hashtag-to-class mapping in zero-shot experiments. Because both the ImageNet and IG-3.6B datasets have target sets drawn from English nouns, we can construct a many-to-many mapping from Instagram hashtags to ImageNet classes. We first map each hashtag to all WordNet synsets of the hashtag, and then map each ImageNet class to the most similar hashtag(s) based on the maximum path similarity score in WordNet [22] between any of the the hashtag synsets and the ImageNet class synset. As the hashtags are nouns or compound nouns, they can have multiple meanings: for example, `#crane` refers to both the bird and the building structure. However, the synset of `crane` referring to the bird and synset of `crane` referring to the structure are two distinct ImageNet classes. In this situation, we map both synsets to `#crane`. Likewise, a synset can represent a concept specified by multiple words and therefore by multiple hashtags, for example, the synset {`porcupine, hedgehog`} matches both `#porcupine` and `#hedgehog`. In this case, we map the synset to all corresponding hashtags.

To utilize the resulting many-to-many mapping between hashtags and ImageNet classes, we need to aggregate the model (hashtag) predictions into predictions over the ImageNet classes. For the RegNetY models, we first map the prediction value for a hashtag to all ImageNet classes that the hashtag maps to. When Platt scaling is used, we *sum* all the resulting values for an ImageNet class to aggregate

Model	Pre-training	Learning rate				
		IN-1k	IN-5k	iNat	Places	CUB
Supervised pre-training						
EfficientNet L2 [73]	JFT 300M	–	2.0E−1	2.0E−1	9.6E−2	2.0E−1
EfficientNet B7 [73]	JFT 300M	–	2.0E−1	2.0E−1	2.0E−1	2.0E−1
EfficientNet B6 [73]	JFT 300M	–	2.0E−1	2.0E−1	2.0E−1	2.0E−1
EfficientNet B8 [72]	IN-1k	–	9.6E−2	4.0E−1	9.6E−2	2.0E−1
EfficientNet B7 [72]	IN-1k	–	9.6E−2	4.0E−1	9.6E−2	2.0E−1
EfficientNet B6 [72]	IN-1k	–	9.6E−2	4.0E−1	9.6E−2	2.0E−1
ViT L/16 [20]	IN-21k	–	–	4.8E−2	2.4E−2	4.8E−2
ViT B/16 [20]	IN-21k	–	–	4.8E−2	2.4E−2	4.8E−2
ViT L/32 [20]	IN-21k	–	–	4.8E−2	2.4E−2	4.8E−2
Weakly supervised pre-training						
ViT H/14	IG 3.6B	6.0E−3	2.4E−2	1.2E−2	3.0E−3	6.0E−3
ViT L/16	IG 3.6B	6.0E−3	2.4E−2	1.2E−2	3.0E−3	3.0E−3
ViT B/16	IG 3.6B	6.0E−3	2.4E−2	1.2E−2	3.0E−3	3.0E−3
RegNetY 128GF	IG 3.6B	6.0E−3	2.4E−2	1.2E−2	6.0E−3	3.0E−3
RegNetY 32GF	IG 3.6B	6.0E−3	1.2E−2	1.2E−2	1.2E−2	6.0E−3
RegNetY 16GF	IG 3.6B	6.0E−3	1.2E−2	1.2E−2	1.2E−2	6.0E−3

Table 8. Base learning rate used for the transfer results in Table 1.

them. When Platt scaling is not used, we instead *average* the predicted values for a class. For the ViT models, we achieved better results with a different aggregation method: we map $1/N$ of the prediction value for a hashtag to all N ImageNet classes that the hashtag maps to, and take the maximum over all the resulting values for each class.

E. ImageNet Robustness Experiments

A potential advantage of weakly supervised pre-training is that the resulting models have observed more training images. This may lead the model to be more robust to variations in the image content. To evaluate the robustness of our models under small variations in visual content, image distribution, or labeling, we performed additional transfer-learning experiments using three ImageNet-like datasets: (1) ReaL ImageNet [7], (2) ImageNet v2 [60], and (3) ObjectNet [4]. We fine-tune pre-trained models on the ImageNet-1k dataset and test them directly on the three evaluation datasets.

Table 9 presents the results of this experiment. While the highest accuracies are obtained by large vision transformers (ViT) trained on 3 billion labeled images (JFT 3B), our weakly pre-trained RegNetY and ViT models are very competitive: our largest models are the runner-up on each of the datasets. In terms of differences in robustness, however, the results are inconclusive: validation accuracy on ImageNet-1k appears to be a good predictor for accuracy on the other tests sets across models and training regimes.

F. Broader Impact

This section presents a more detailed account of the experiments presented in the main paper, which aim to understand: (1) how well our models perform on photos taken in non-English speaking countries, and (2) the associations our hashtag-prediction models learn with photos of people with

Model	Pre-training	Classification accuracy			
		IN-1k	ReaL-IN	IN-v2	Obj. Net
Supervised pre-training†					
EfficientNet L2 [73]	JFT 300M‡	*88.4*	90.6	80.2	68.3
EfficientNet B7 [73]	JFT 300M‡	*86.9*	90.1	78.1	61.2
EfficientNet B6 [73]	JFT 300M‡	*86.4*	89.8	76.7	60.0
EfficientNet B8 [72]	IN-1k	*85.5*	89.6	76.1	54.5
EfficientNet B7 [72]	IN-1k	*85.2*	89.5	75.7	53.3
EfficientNet B6 [72]	IN-1k	*84.8*	89.4	75.5	51.9
ViT G/14 [76]	JFT 3B	***90.5***	***90.8***	***83.3***	***70.5***
ViT L/16 [76]	JFT 3B	*88.5*	*90.4*	*80.4*	–
ViT H/14 [20]	JFT 300M	*88.6*	*90.7*	–	–
ViT L/16 [20]	JFT 300M	*87.8*	*90.5*	–	–
ViT H/14 [20]	IN-21k	*85.1*	*88.7*	–	–
ViT L/16 [20]	IN-21k	85.2	*88.4*	74.8	56.6
ViT B/16 [20]	IN-21k	84.2	*88.4*	73.5	52.6
ViT L/32 [20]	IN-21k	81.5	*86.6*	71.2	47.2
Weakly supervised pre-training					
ViT H/14	IG 3.6B	88.6	90.5	81.1	69.5
ViT L/16	IG 3.6B	88.1	90.6	80.3	66.2
ViT B/16	IG 3.6B	85.3	89.1	75.6	55.2
RegNetY 128GF	IG 3.6B	88.2	90.7	80.4	68.5
RegNetY 32GF	IG 3.6B	86.8	90.2	78.2	62.5
RegNetY 16GF	IG 3.6B	86.0	89.9	76.9	59.0

Table 9. Classification accuracy of models pre-trained on the specified pre-training dataset followed by finetuning on ImageNet-1k. Accuracy is measured on four ImageNet-like datasets: (1) ImageNet-1k itself, (2) ReaL ImageNet [7], (3) ImageNet v2 [60], and (4) ObjectNet [4]. The best result on each dataset is **bold-faced**; the second-best result is underlined. Numbers that are adopted from the original paper are *italicized*. Higher is better. †It is unknown how much manual curation was performed in the annotation of JFT datasets. ‡Pre-training data also includes IN-1k.

varying characteristics. In this section we share and discuss all the experimental results in more detail. As a reminder, all results presented below are for the hashtag-prediction models; no fine-tuning is employed.

F.1. Analyzing Hashtag Prediction Fairness

Following prior work [17], we analyzed how well the RegNetY 128GF model works on photos taken across the world. We first repeated the analysis of [17] on the Dollar Street dataset. To this end, we use the hashtag-prediction model in a zero-shot fashion: we manually define a mapping from hashtags to the 112 classes in the Dollar Street, and task the model with predicting only hashtags that are mapped to a class. We measure the accuracy of the model's predictions per country, and display the results on a world map in which colors correspond to accuracies in the left plot in Figure 5 (red is 40% correct; green is 70%).

Although the absolute numbers are lower because the image-recognition model operates in zero-shot mode (the average accuracy over all countries is 48.0%), qualitatively, the observations we obtain are in line with prior work [17]: observed recognition accuracies are higher in the US and Europe than in most other countries.

Because the Dollar Street dataset may itself have issues,

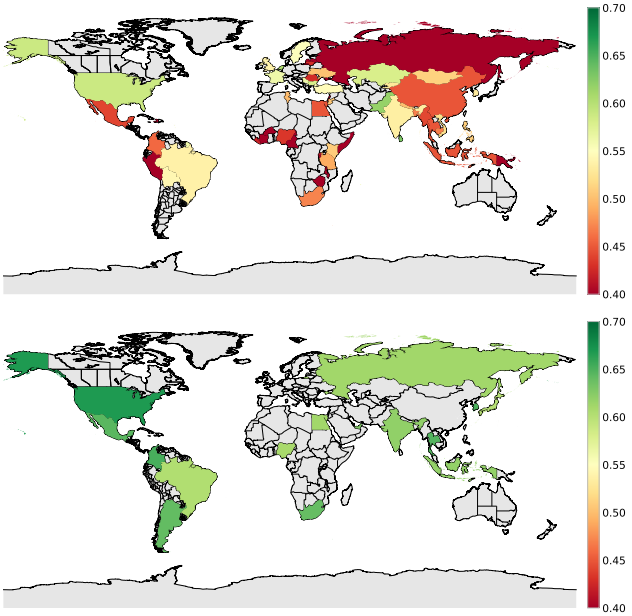

Figure 5. Recognition accuracy per country of our zero-shot classifier on the Dollar Street dataset (**top**) and a proprietary dataset (**bottom**). The accuracy on all images is 48.0% on the Dollar Street dataset and 63.3% on the proprietary dataset.

we repeated the analysis on a proprietary dataset that contains millions of images labeled for visual concepts and their country of origin. The resulting world map is shown in the right plot in Figure 5. The results suggest that the range of accuracy values is relatively tight (approximately 5%) on this large proprietary dataset.

Following common practice [75], we also measure the percentage of classes for which the ratio between the class-recognition accuracy in country 1 and country 2 is smaller than 0.8. The results of this analysis are shown in the heat map in Figure 6. If an entry in the heat map is yellow, then the model recognizes a substantial percentage (up to 35%) of classes substantially worse in the "row country" than in the "column country".

The results in the figure suggest that the hashtag-prediction model performs better in the US and worse in Egypt and Nigeria. There is also a notable difference between the accuracy map in Figure 5 and the heat map in Figure 6. The accuracy map suggests that the hashtag-prediction model performs worst in Brazil and Japan, whereas the heat map suggests the lowest accuracy is obtained in Egypt and Nigeria. This result may be due to variations between countries in the distribution of per-class accuracy discrepancies and/or due to variations in the concept distribution per class.

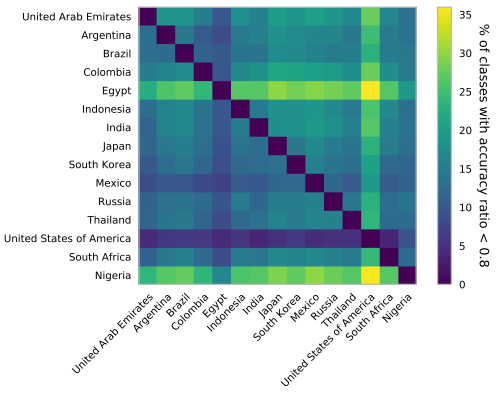

Figure 6. Percentage of classes for which recognition accuracy is substantially higher in one country (**rows**) than in another country (**columns**). We use the 80% rule to assess whether one accuracy is "substantially higher" than the other.

F.2. Analyzing Associations in Hashtag Predictions

We performed experiments in which we analyze the associations our hashtag-prediction models make for photos of people with different apparent skin tone, different apparent age, different apparent gender, and different apparent race. We present the results of each experiment separately below.

Apparent Skin Tone We first evaluated potentially troubling associations in hashtag predictions by apparent skin tone. To this end, we used a proprietary dataset that contains 178,448 Instagram photos that were annotated using the Fitzpatrick skin tone scale [23]. We ran all these photos through our RegNetY 128GF hashtag-prediction model, asking it to predict the five highest-scoring hashtags for each photo. We maintain per-skin tone statistics on how often a hashtag in the vocabulary is predicted for a photo of an individual with that skin tone. Next, we inspect differences in the hashtag prediction rate between different skin tones. For each skin tone, we identify the hashtags with the largest absolute difference in hashtag prediction rate compared to the average prediction rate for the other five skin tones. We also compute the associated relative difference in hashtag prediction rate. We show the resulting hashtags for skin tone 1 (lightest skin tone) and skin tone 6 (darkest skin tone) in the top row of Figure 7.

The results in the figure reveal several associations that may not be unexpected: for example, `#redhead` is more commonly predicted by the model for photos of people with a light skin tone, whereas `#black` is more often predicted for people with a dark skin tone. The analysis also reveals associations that are more difficult to explain: do people with lighter skin tones wear more `#headbands` or `#bandanas`? It is also unclear to what extent the associations we find are learned by the model and to what extent

they reflect characteristics of the evaluation data.

Apparent Age We performed a similar analysis of associations between predicted hashtags and apparent age groups. For this evaluation, we used the UTK Faces dataset [78], which provides apparent age labels. People were grouped into age buckets with a range of 10 years ($0-10$, $10-20$, $20-30$ years, *etc.*). We performed the same analysis as before. The second row of Figure 7 shows the most common hashtag predictions for two different (apparent) age groups.

Some associations that the analysis reveals are not unexpected: for example, predicting `#baby` or `#kid` for age group $1-10$ years or predicting `#elder` for age group $80-90$ years. The results also show that there may be discrepancies in the meaning of words and hashtags: `#rip` is in the hashtag dictionary because one may have a rip in their shirt but it is commonly used on Instagram as abbreviation for "rest in peace", which is more likely to apply to people of age. Other disparate associations appear unfortunate, such as the association of `#spermbank` with photos of people aged $0-10$ years.

Apparent Gender We performed the same analysis on the UTK Faces dataset [78] by apparent gender. Due to limitations of the evaluation dataset, we restricted our analyses to males and females but did not consider non-binary genders. The results are presented in the third row of Figure 7.

The results suggest that the model has learned certain gender-specific stereotypes, for example, associating men with `#football` and `#basketball` more frequently or associating women more frequently with `#makeup` and `#bikini`. The associations revealed by the analysis vary in how problematic they are: for example, men may not be excited that they are more frequently associated with `#mugshot` – and in some cases, such an association could be harmful. We will return to this example below.

Apparent Race For better or worse (see below), the UTK Faces dataset [78] also contains annotations of apparent race. We repeated the same hashtag prediction analysis for the groups defined in UTK Faces (Indian, Asian, Black, White, Other) as well. We present the results of this analysis in the fourth row of Figure 7.

The results analysis suggest a variety of disparate associations, some of which are more problematic than others. Likely the most troubling association suggested by the analysis is the association of photos of Black people with `#mugshot` and `#prison`. Because of the sensitivity of this type of association, we investigated it more in-depth. First, we performed a visual analysis of the photos for which the hashtag-prediction model predicted `#mugshot` or `#prison` among its top-5 predictions. This inspection revealed that a small percentage of the photos in the UTK Faces dataset are, indeed, mug shots. Specifically, some of the images in the dataset appear to have been sourced from `http://mugshots.com/`. This observation raises an important question: *Are the associations our analyses identify due to associations that the model has learned, due to biases in the evaluation data, or both?* This question is difficult to answer without collecting additional annotations.

In this particular case, we decided to re-use the skin tone dataset we used earlier and measure how often `#mugshot` is predicted for the images in that dataset. While skin tone does not map to race very well, we would expect to observe at least some correlation between `#mugshot` prediction and skin tone if the model had learned this association. The results were quite the opposite: `#mugshot` was predicted 7 times (0.0078%) for images with Fitzpatrick skin tone 1 (lightest skin tone) but only once for skin tone 6 (darkest skin tone; 0.0023%). Combined with our visual inspection, this suggests that the problematic association we observed in the analysis on UTK Faces is most likely to be due to problems in the UTK Faces dataset itself than due to problems in the hashtag-prediction model. Having said that, we acknowledge that there are many caveats here, and that our experiments are not fully conclusive.

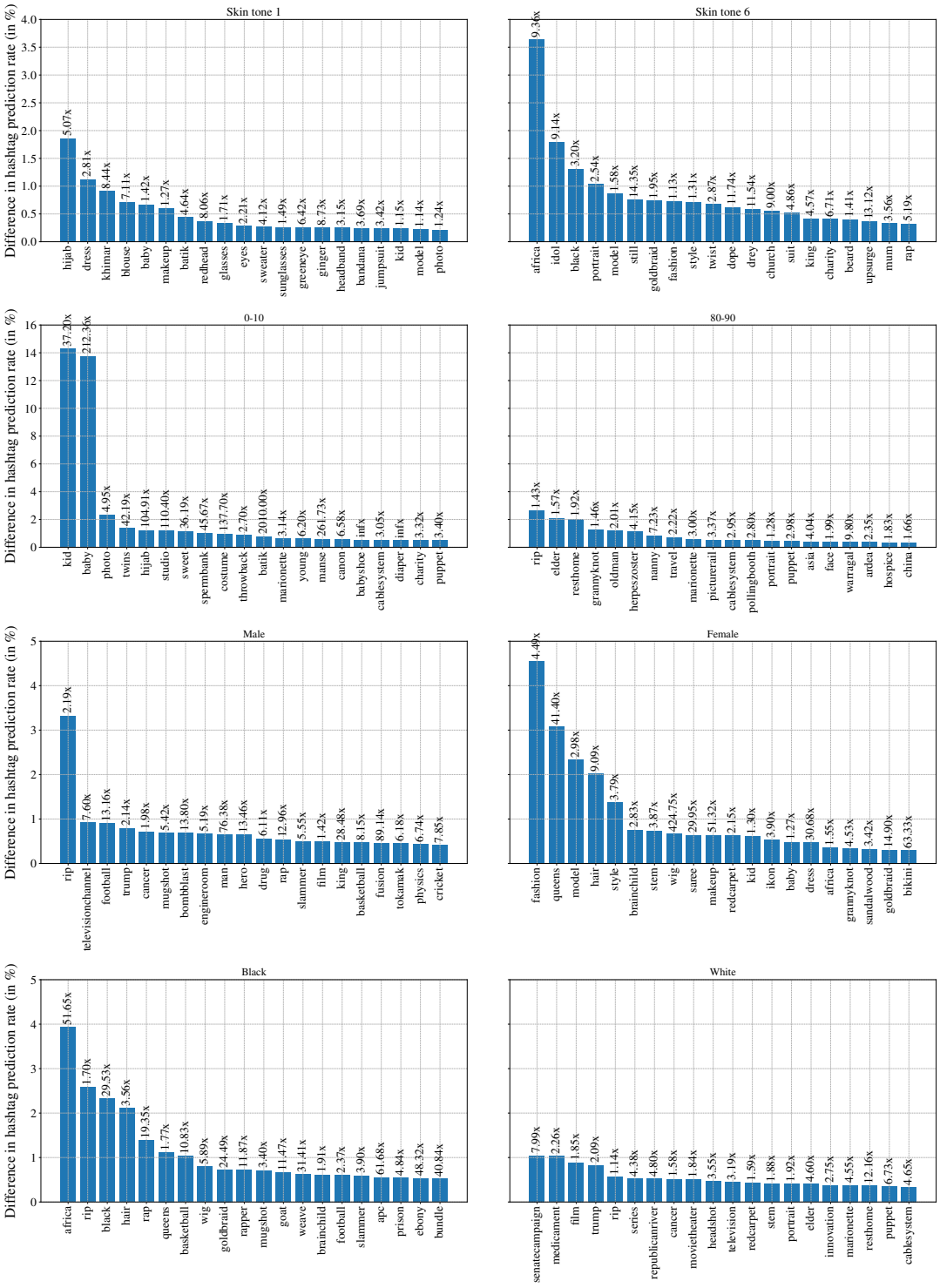

Figure 7. Differences in hashtag prediction rate for photos from various apparent subgroups. Absolute differences are sorted, and results for 20 hashtags with the largest difference are shown. Relative hashtag prediction differences are shown on top of the bars. From **top** to **bottom**: Differences for photos of people with (apparent) Fitzpatrick skin tone 1 and photos of people with other apparent skin tones (**left**); and between photos with skin tone 6 and other skin tones (**right**). Differences between photos with (apparent) age group 0-10 and other age groups; and between age group 80-90 and other age groups. Differences between photos of (apparent) women and photos of men; and between photos of men and women. Differences between photos of (apparent) Black people and people of other races; and between photos of White people and other races.

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. In *arXiv:2110.02095*, 2021. 2

[2] A. Adcock, V. Reis, M. Singh, Z. Yan, L. van der Maaten, K. Zhang, S. Motwani, J. Guerin, N. Goyal, I. Misra, L. Gustafson, C. Changhan, and P. Goyal. Classy vision. https://github.com/facebookresearch/ClassyVision, 2019. 14

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 5, 6

[4] 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 7, 15

[5] Josh Beal, Hao-Yu Wu, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Billion-scale pretraining with vision transformers for multi-task visual representations. In *arXiv:2108.05887*, 2021. 2

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. 1, 7, 8

[7] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aaron van den Oord. Are we done with ImageNet?, 2020. 7, 15

[8] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. In *arXiv:2110.01963*, 2021. 3

[9] Tom B. Brown, Ben Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen M. Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing*, 2020. 1, 7, 8

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *arXiv:1807.05520*, 2018. 2, 5

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 2, 5

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *arXiv:2104.14294*, 2021. 1, 2, 5

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 1, 2, 5, 6

[14] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 5

[15] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *arXiv:1805.09501*, 2018. 4

[16] E. Denton, J. Weston, M. Paluri, L. Bourdev, and R. Fergus. User conditional hashtag prediction for images. In *Proc. KDD*, pages 1731–1740, 2015. 2

[17] T. DeVries, I. Misra, C. Wang, and L.J.P. van der Maaten. Does object recognition work for everyone? In *CVPR Workshop on Computer Vision for Global Challenges*, 2019. 7, 8, 15

[18] Piotr Dollár, Mannat Singh, and Ross Girshick. Fast and accurate model scaling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 924–932, 2021. 3, 13

[19] 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 *International Conference on Machine Learning*, pages 647–655, 2014. 2

[20] 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. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 4, 5, 13, 15

[21] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *arXiv:2104.11227*, 2021. 2

[22] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 3, 6, 12, 14

[23] T. B. Fitzpatrick. Soleil et peau. *Journal de Médecine Esthétique*, 2:33–34, 1975. 7, 16

[24] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *UAI*, pages 148–155, 1998. 6

[25] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, pages 12046–12055, 2019. 2, 8

[26] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *arXiv:1803.07728*, 2018. 2, 5

[27] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. 1, 2, 5, 6

[28] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 3

[29] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *arXiv:2006.07733*, 2020. 1, 2, 5

[30] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. In *arXiv:2103.07976*, 2021. 2

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2019. 1, 2, 5

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[33] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. In *arXiv:1709.01450*, 2017. 7

[34] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 13

[35] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3, 13

[36] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth. In *arXiv:1603.09382*, 2016. 4

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *arXiv:2102.05918*, 2021. 1, 2, 6, 7, 8

[38] A. Joulin, L.J.P. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–84, 2016. 1, 2

[39] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. In *arXiv:1705.06950*, 2017. 2

[40] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. In *arXiv:1704.03162*, 2017. 2

[41] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019. 4

[42] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *arXiv:1805.08974*, 2018. 1, 2

[43] C.H. Lampert. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009. 6

[44] A. Li, A. Jabri, A. Joulin, and L.J.P. van der Maaten. Learning visual n-grams from web data. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[45] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017. 7

[46] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3, 13

[48] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 3, 4

[49] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 1, 2, 3, 4, 8, 12, 13, 14

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2

[51] I. Misra and L.J.P. van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5

[52] R. Müller, S. Kornblith, and G.E. Hinton. When does label smoothing help? In *NeurIPS*, 2019. 3

[53] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999. 6

[54] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. 3, 4

[55] V.U. Prabhu and A. Birhane. Large datasets: A Pyrrhic win for computer vision? In *arXiv:2006.16923*, 2020. 3

[56] Filip Radenovic, Animesh Sinha, Albert Gordo, Tamara Berg, and Dhruv Mahajan. Large-scale attribute-object compositions. In *arXiv:2105.11373*, 2021. 2, 8

[57] 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 *arXiv:2103.00020*, 2021. 1, 2, 6, 7, 8

[58] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 3, 13, 14

[59] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 512–519, 2014. 2

[60] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 2, 7, 15

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 2

[62] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *arXiv:2104.10972*, 2021. 2

[63] 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. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 4

[64] Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[65] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3, 13

[66] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*, 2019. 4

[67] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[68] A. Veit, M. Nickel, S. Belongie, and L.J.P. van der Maaten. Separating self-expression and visual content in hashtag supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5919–5927, 2018. 2

[69] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 4

[70] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *arXiv 2012.00759*, 2020. 2

[71] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 14

[72] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020. 4, 5, 15

[73] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 4, 5, 13, 15

[74] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3, 4, 13

[75] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of AISTATS*, 2017. 16

[76] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021. 1, 2, 4, 5, 13, 14, 15

[77] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3, 4

[78] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 7, 8, 17

[79] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 4