

An Analysis of Pre-Training on Object Detection

Hengduo Li Bharat Singh Mahyar Najibi Zuxuan Wu Larry S. Davis

University of Maryland, College Park

{hdli, bharat, najibi, zxwu, lsd}@cs.umd.edu

Abstract

We provide a detailed analysis of convolutional neural networks which are pre-trained on the task of object-detection. To this end, we train detectors on large datasets like OpenImagesV4, ImageNet Localization and COCO. We analyze how well their features generalize to tasks like image classification, semantic segmentation and object detection on small datasets like PASCAL-VOC, Caltech-256, SUN-397, Flowers-102 etc. Some important conclusions from our analysis are—1) Pre-training on large detection datasets is crucial for fine-tuning on small detection datasets, especially when precise localization is needed. For example, we obtain 81.1% mAP on the PASCAL-VOC dataset at 0.7 IoU after pre-training on OpenImagesV4, which is 7.6% better than the recently proposed DeformableConvNetsV2 which uses ImageNet pre-training. 2) Detection pre-training also benefits other localization tasks like semantic segmentation but adversely affects image classification. 3) Features for images (like avg. pooled Conv5) which are similar in the object detection feature space are likely to be similar in the image classification feature space but the converse is not true. 4) Visualization of features reveals that detection neurons have activations over an entire object, while activations for classification networks typically focus on parts. Therefore, detection networks are poor at classification when multiple instances are present in an image or when an instance only covers a small fraction of an image.

1. Introduction

For several computer vision problems like object detection, image segmentation and image classification, pre-training on large scale datasets is common [32, 14, 10]. This is because it leads to better results and faster convergence [62, 24, 10, 35, 18]. However, the effect of pre-training in computer vision is often evaluated by training networks for the task of *image classification*, on datasets like ImageNet [9], Places [62], JFT [49], Instagram [35]

etc., but rarely for object detection. It can be argued that the task of object detection subsumes image classification, so a network good at object detection should learn richer features than one trained for classification. After all, this network has access to an orthogonal semantic information, like the spatial extent of an object. However, it can also be argued that forcing a network to learn position sensitive information may affect its spatial invariance properties which help in recognition. To this end, we provide a comprehensive analysis which compares pre-training CNNs on object detection and image classification.

We pre-train a network on the OpenImagesV4 [27] (hereafter referred to as OPENIMAGES) dataset on the object detection task and fine-tune it on tasks like semantic segmentation, object detection and classification on datasets like PASCAL-VOC [12], COCO [30], CALTECH-256 [15], SUN-397 [53], and OXFORD-102 FLOWERS [38]. For a stronger evaluation, we also pre-train on the ImageNet classification dataset [9] with bounding-box annotations on 3,130 classes [46] (hereafter referred to as IMAGENET-LOC as opposed to IMAGENET-CLS for ImageNet Classification dataset without bounding boxes) and the COCO dataset [30] which helps us in evaluating the importance of the number of training samples. We then design careful experiments to understand the differences in properties of features which emerge by pre-training on detection *vs.* classification.

Our experimental analysis reveals that pre-training on object detection can improve performance by more than 5% on PASCAL-VOC for object detection (especially at high IoUs) and 3% for semantic segmentation. However, detection features are significantly worse at performing classification compared to features from IMAGENET-CLS pre-trained networks ($\sim 8\%$ on CALTECH-256). We also find that if features (like average pooled Conv5) are similar in the object detection feature space, they are likely to be similar in the image classification feature space, but the converse is not true. Visualization of activations for object detection shows that they often cover the entire extent of an object, so are poor at recognition when an object is present in a small part of an image or when multiple instances are present.

2. Related Work

Large Scale Pre-training The initial success of deep learning in computer vision can be largely attributed to transfer learning. ImageNet pre-training was crucial to obtain improvements over state-of-the-art results on a wide variety of recognition tasks such as object detection [29, 31, 39, 45, 7, 29], semantic segmentation [4, 19, 5, 32, 59], scene classification [62, 20], action/event recognition [51, 60, 54, 52] etc. Due to the importance of pre-training, the trend continued towards collecting progressively larger classification datasets such as JFT [49], Places [62] and InStagram [35] to obtain better performance. While the effect of large-scale classification is extensively studied [42, 10], there is little work on understanding the effect of pre-training on object detection.

Transfer Learning The transferability of pre-trained features has been well studied [1, 55, 22, 24, 6, 49]. For example, [1] measured the similarity between a collection of tasks with ImageNet classification; [6] studied how to transfer knowledge learned on large classification datasets to small fine-grained datasets; [24] addressed relationship among ImageNet pre-training accuracy, transfer accuracy and network architecture; [57] proposed a computational approach to model relationships among visual tasks of various abstract levels and produced a computational taxonomic map. However, the visual tasks in [57] did not involve object detection although object detection is one of the few tasks other than image-classification for which large-scale pre-training can be performed. We study the transferability, generalizability, and internal properties of networks pre-trained for object detection.

Understanding CNNs Towards understanding the superior performance of CNNs on complex perceptual tasks, various qualitative [44, 61, 41, 56, 48, 37, 58, 11, 36] and quantitative [40, 26, 13, 2] approaches have been proposed. A number of previous works explain the internal structure of CNNs by highlighting pixels which contribute more to the prediction using gradients [44], Guided BackPropagation [58, 41], deconvolution [48], etc. Other methods adopt an activation maximization based approach and synthesize the preferred input for a network neuron [37, 36, 56, 11]. Attempts have also been made to interpret the properties of CNNs empirically by investigating what it learns and is biased towards. While [40, 26] suggest that deep neural networks implicitly learn representations of shape, recent work [13, 2] indicates that CNNs trained for image classification task are biased towards texture. [13] further indicates the advantage of a shape-based representation by training CNNs on a stylized version of ImageNet.

Training From Scratch While most modern detectors are pre-trained on the ImageNet classification dataset [29, 45, 7, 29, 31, 39], effort has also been made to deviate from the conventional paradigm and train detectors from

scratch [43, 28]. [43] proposed a set of design principles to train detector from scratch. [18] demonstrated that with a longer training schedule, detectors trained from scratch can be as good as ImageNet pre-trained models on large datasets (like COCO). However, pre-training is still crucial when the training dataset is small (like PASCAL-VOC).

3. Discussion

A detailed analysis of detection pre-training is lacking in the existing literature. This is primarily because COCO [30] is still small compared to IMAGENET-CLS (by 10× images, 10× categories), so there is an unknown variable about the scale of the dataset. While IMAGENET-LOC also contains bounding-boxes for objects, detection in it is not challenging as images typically only contain a single object (which are often large, making localization trivial in many cases), so in this case, it is unclear if the network is learning instance level features. Recently, due to a massive data collection effort, a new dataset called OPENIMAGES was released which contains bounding-box annotations for 15 million instances and close to 2 million images. This is the first dataset which provides an orthogonal semantic information at the scale of ImageNet. Therefore, it allows us to fairly compare networks pre-trained on large scale object detection with large scale image classification when fine-tuning on standard computer vision datasets in which the number of annotations is lower by one or two orders of magnitude.

4. Analysis

We perform pre-training on multiple detection datasets and compare it with IMAGENET-CLS pre-training for different computer vision tasks like object detection, image classification and semantic segmentation. For detection pre-training, our experimental setup is as follows. All our detection networks are pre-trained first on IMAGENET-CLS. They are then trained on detection datasets like OPENIMAGES [27], IMAGENET-LOC [9, 46] and COCO [30]. The SNIPER [47] detector is trained on all the datasets. We use multiple pre-training datasets for two reasons - 1) To thoroughly evaluate our claims about pre-training for the detection task 2) Since the datasets contain a different number of classes and training examples, it also provides an indication of the magnitude of improvement one can expect by pre-training on detection datasets of different sizes.

Datasets Here we briefly introduce the target datasets used in our fine-tuning experiments. For the *object detection* task, we fine-tune on the PASCAL-VOC dataset [12]. We use the VOC 07+12 trainval set for training and the VOC 07 test set for evaluation. For the *semantic segmentation* task, we follow [8, 17, 33, 3] and use VOC 2012 plus additional annotations provided in [16]. For *image classification*, we fine-tune on CALTECH-256 [15], SUN-397 [53]

Dataset	#Class	#Images	#Objects
IMAGENET-CLS [9]	1000	1.28M	-
OPENIMAGES [27]	500	1.74M	14.6M
IMAGENET-LOC [9, 46]	3,130	0.92M	1.06M
COCO [30]	80	0.14M	0.89M
CALTECH-256 [15]	257	15.4K/15.2K	-
SUN-397 [53]	397	19.9K/19.9K	-
OXFORD-102 FLOWERS [38]	102	2.0K/6.1K	-
PASCAL-VOC Det [12]	20	16.6K/5.0K	40.1K/12.0K
PASCAL-VOC Seg [12]	21	10.6K/1.4K	-/-

Table 1: Source and target datasets examined. x/y denotes x for training set and y for evaluation set.

and OXFORD-102 FLOWERS [38]. We use the trainval and test sets directly in CALTECH-256 and OXFORD-102 FLOWERS; for SUN-397 we follow [24] and use the first split for training and evaluation. The number of classes, images and instances in each of these datasets are mentioned in Table 1.

Architecture We briefly describe the architecture of the two detection heads (Faster-RCNN and R-FCN) which are used for training. On OPENIMAGES detector after Conv5 (2048,14,14) we have the following layers: ConvProj (256,14,14), FC1 (1024), FC2 (1024), Output (501), Regression (4). A fully connected layer projects the (256,14,14) blob to a 1024 dimensional vector, thus spatial information is preserved for the blob. The Output (501) and Regression (4) layers are connected to FC2. The same architecture is used for the COCO detector, except that the Output layer has 81 dimensions. For the IMAGENET-LOC detector, the architecture is as described in [46]. In this architecture, classification and detection are decoupled and performed independently. For classification, Conv5 features are average pooled and a fully connected layer projects these 2048 dimensional features to a 1024 vector, on which a 3130 dimensional classifier is applied. Detection is performed using a R-FCN head on the Conv5 features which are first projected to 1024 dimensional features.

4.1. Object Detection

Baseline Configuration and Results For our object detection experiments, we train our detectors (SNIPER with ResNet-101) on 3 datasets: OPENIMAGES, COCO and IMAGENET-LOC. Our OPENIMAGES model obtains 45% mAP (at 0.5 overlap) on the validation set. It is trained at 2 scales, (480,512) and (768, 1024) without negative chip mining. Inference is also performed at these two scales only. For the COCO model, training and inference is performed at 3 scales (480,512), (800, 1280) and (1400,2000) and the detector obtains an mAP of 46.1% (COCO metric) on the test-dev set. The IMAGENET-LOC model ob-

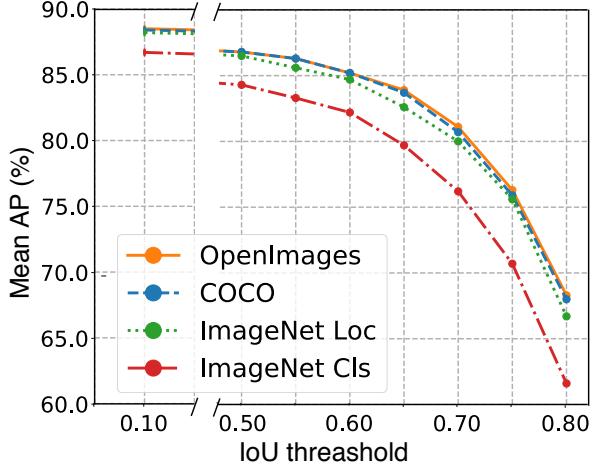


Figure 1: Detection performance (mAP %) at different IoUs on PASCAL-VOC 2007 [12] test set of detectors pre-trained on different datasets. Typically, localization errors are ignored at 0.1 IoU threshold.

tains 37.4% mAP (at 0.5 overlap) on the ImageNet *Detection* dataset (not IMAGENET-LOC). This detector was only trained at a single scale of (512,512) on IMAGENET-LOC without any negative chip mining. Inference is also performed only at a scale of (512,512) as compared to others, this dataset contains relatively bigger objects.

Fine-tuning on PASCAL-VOC We fine-tune these pre-trained models on PASCAL-VOC [12] using the same set of scales as COCO for both training and inference. Detection heads of the models pre-trained on detection datasets are re-initialized before fine-tuning. Following [47], we train the RPN for 2 epochs first for negative chip mining. Then training is performed for 7 epochs with learning rate step-down at the end of epoch 5. Horizontal flipping is used as data augmentation. The results are shown in Table 2. As a reference, the recently proposed Deformable ConvNet-V2 [63] obtains 73.5% at an overlap of 0.7 while the OPENIMAGES/COCO/ImageNet-3k models obtain 81.1%, 80.7% and 80% mAP at 0.7 overlap. Our

Method / Pre-trained Dataset	mAP@0.5	mAP@0.7
DCNv1 [8]	81.9	68.2
DCNv2 [63]	84.9	73.5
IMAGENET-CLS [9]	84.6	76.3
IMAGENET-LOC [9, 46]	86.5	80.0
COCO [30]	86.8	80.7
OPENIMAGES [27]	86.8	81.1

Table 2: Baseline and our results on PASCAL-VOC 2007 [12] object detection dataset.

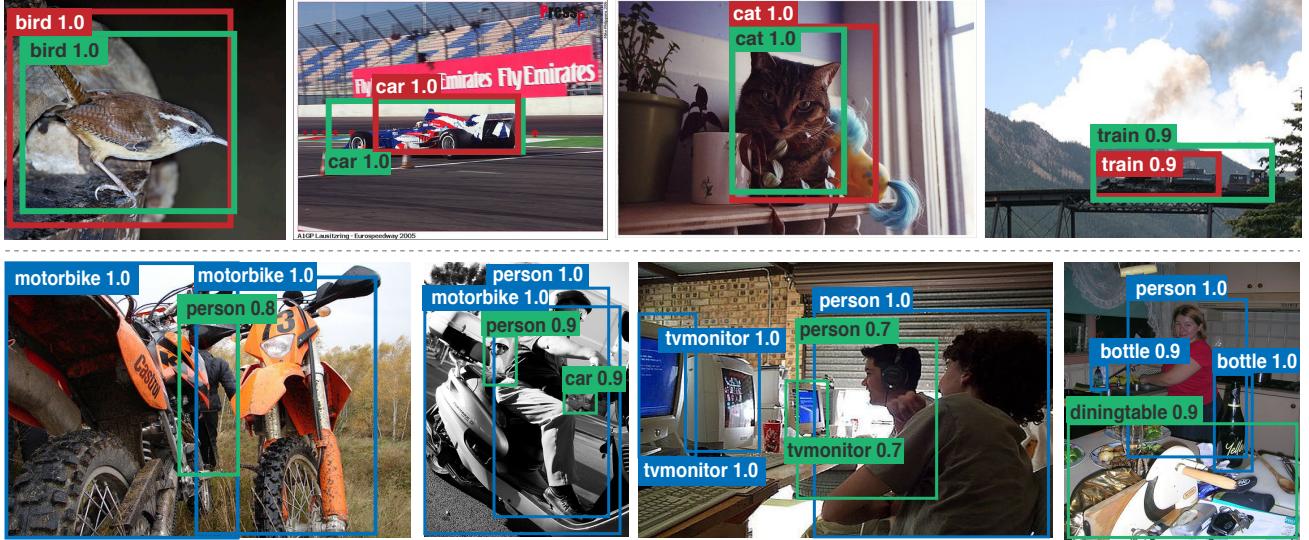


Figure 2: Qualitative results on PASCAL-VOC 2007 from detectors pre-trained on IMAGENET-CLS [9] and OPENIMAGES [27]. **Above:** OPENIMAGES pre-trained detector shows better localization ability. Green and red boxes are from OPENIMAGES pre-trained and IMAGENET-CLS pre-trained detectors respectively. **Below:** OPENIMAGES pre-trained detector handles occlusion cases better. Blue boxes are correct predictions from both detectors while green boxes are occluded objects successfully detected only by the OPENIMAGES pre-trained detector.

baseline network which is only trained on IMAGENET-CLS obtains an mAP of 76.3%. Thus, pre-training on larger detection datasets can improve performance on PASCAL by as much as 4.8% at an overlap of 0.7. However, such large improvements do not translate to lower overlap thresholds. For example, the difference in mAP between IMAGENET-CLS and the OPENIMAGES model at an overlap of 0.5 is only 2.2%. We plot the mAP for all the detection models at different overlap thresholds in Fig 1. This clearly shows that pre-training for detection helps to a large extent in improving localization performance on PASCAL. We also observe this phenomenon when we use the OPENIMAGES model to fine-tune on the COCO dataset. For example, the performance at an overlap of 0.5 on COCO with IMAGENET-CLS pre-training is 67.5 and at 0.75 it is 52.2. When OPENIMAGES pre-training is used, the performance at 0.5 improves by 0.7%, but results at 0.75 improve by 1.4%.

Pre-training helps at Higher IoU While the COCO result that mAP at 0.75 improves more than mAP at 0.5 after fine-tuning from an OPENIMAGES model was presented in SNIPER [47], here we show that this is indeed a systematic pattern which is observed when pre-training is performed on large scale detection datasets. When the size of the detection dataset is small (like PASCAL-VOC), localization at higher overlap thresholds can significantly benefit from pre-training on large detection datasets. Another pattern we observe here is that the number of samples in the pre-trained dataset did not affect the fine-tuning performance to a large extent (differences are within 1%). The important factor

was whether the network was pre-trained on a reasonably large detection dataset ($> 1\text{M}$ training instances) or not.

Qualitative Results and Error Analysis We show qualitative results on the PASCAL-VOC dataset for OPENIMAGES and IMAGENET-CLS pre-training. Fig 2 shows that localization for the OPENIMAGES model is better. Following [30], we evaluate detectors pre-trained on different aforementioned datasets at different IoU thresholds including 0.1, where localization errors are typically ignored. The small gap between mAP@0.1 and higher IoUs like 0.5 indicates that large localization errors are rare. We also observe in Fig 2 that the OPENIMAGES model handles occlusion cases better. To further verify the observation on performance improvement under occlusion, we also analyze the errors using the object detection analysis tools in [21, 30]. Quantitative results are mentioned in Table 3 which demonstrate that the OPENIMAGES network is indeed better under occlusion.

% missed object	occluded Low	occluded Medium
IMAGENET-CLS [9]	14.7%	15.7%
OPENIMAGES [27]	10.1%	10.8%

Table 3: Percentage of missed objects under low and medium occlusion levels in the PASCAL-VOC 2007 [12] test set. Results are obtained using the object detection analysis tool in [21]

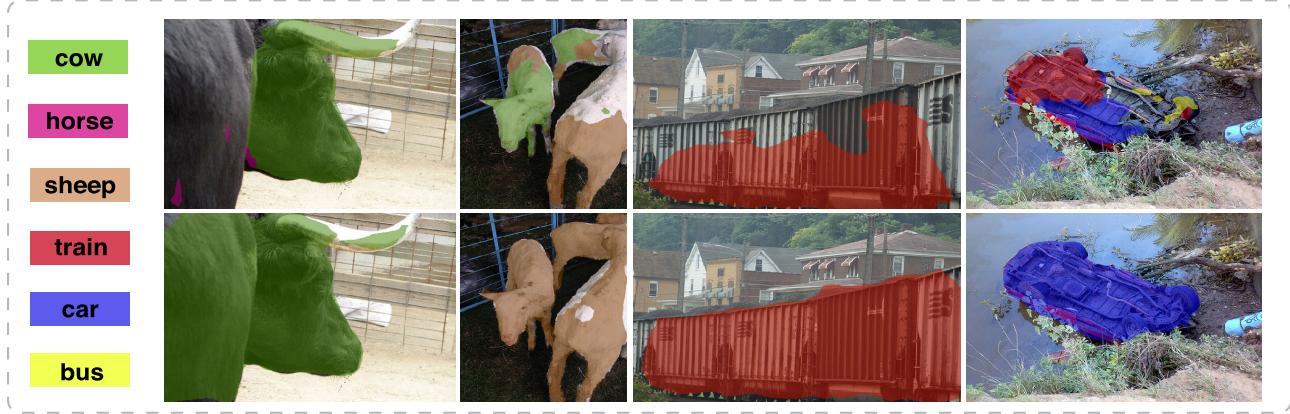


Figure 3: Qualitative results of semantic segmentation from networks pre-trained on IMAGENET-CLS [9] (**Above**) and IMAGENET-LOC [9, 46] (**Below**). The IMAGENET-LOC model is better at covering entire objects while the classification pre-trained model is more likely to mis-classify pixels on some parts of an object.

4.2. Semantic Segmentation

Baseline Configuration and Results We fine-tune detection networks for the semantic-segmentation task on PASCAL-VOC 2012. Following [8], we use Deformable ConvNets [8] as our backbone in DeepLab [4] throughout our experiments. In training and inference, images are resized to have a shorter side of 360 pixels while keeping the larger side less than 600 pixels. Baseline results are shown in Table 4.

Detection Pre-Training Helps Segmentation The results after fine-tuning are shown in Table 4. These results show that networks which are trained for object detection obtain a 3% improvement in performance compared to image classification. We evaluate this for the OpenImages dataset and also for IMAGENET-LOC dataset.

Error Analysis We also perform experiments to understand where these improvements occur. Specifically, we study if the improvements from detection pre-trained net-

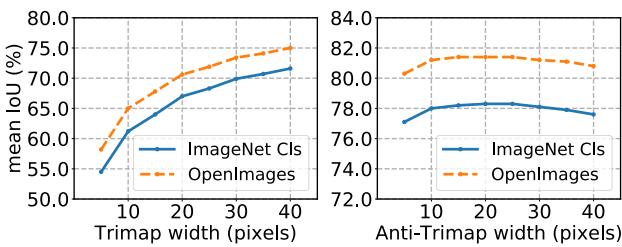


Figure 4: Results of Trimap (**left**) and Anti-Trimap(**right**) experiments. Segmentation performance on pixels inside an x -pixel-wide trimap band (near object boundary) and outside the trimap band (away from object boundary) are evaluated respectively.

Method / Pre-trained Dataset	mIoU
DCNv1 [8]	75.2
IMAGENET-CLS [9]	75.7
IMAGENET-LOC [9, 46]	78.3
OPENIMAGES [27]	78.6

Table 4: Baseline and our fine-tuning results on PASCAL-VOC 2012 [12] semantic segmentation dataset.

works are due to better segmentation at boundary pixels or not. For this we evaluate the accuracy at boundary pixels with the “trimap experiment” [23, 25, 4, 5] and non-boundary pixels called the “anti-trimap experiment”. The boundary pixels are obtained by applying morphological dilation on the “void” labeled pixels which often occurs at object boundaries.

We perform two types of evaluations. 1) Accuracy at pixels which are within a distance x from an object boundary 2) Accuracy at pixels of an object or background, not in (1). The first evaluation compares the accuracy at boundary pixels while the second one compares the accuracy for pixels which are not at the boundary. The results for these experiments are shown in Fig. 4 (using OPENIMAGES). These results show that the gap in performance remains the same as the size of boundary pixels is increased (instead of reducing, if one model was more accurate at boundaries). The same is true for pixels which are far away from the boundary. There is still a significant gap in performance for pixels which are not at the boundary. Thus, from these experiments, it is clear that improvement in performance is not due to better classification at boundary pixels.

Qualitative and Semantic Analysis We provide some

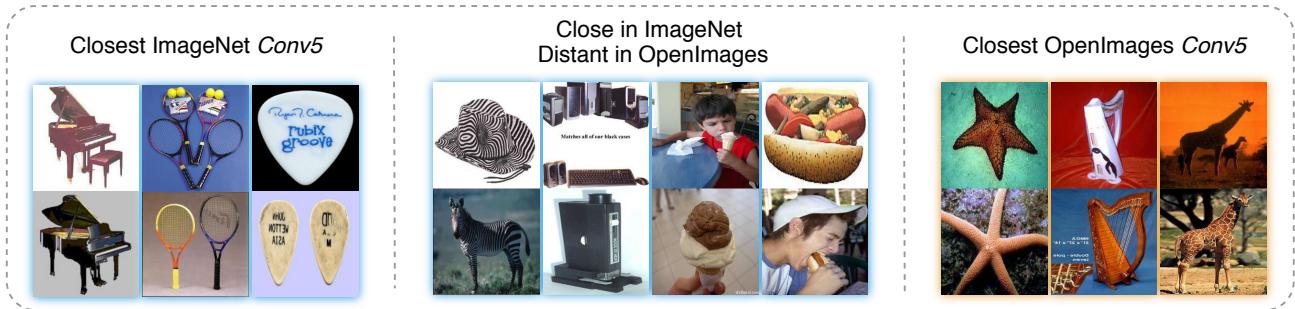


Figure 5: Qualitative results of our feature space analysis. We extract Conv5 features of CALTECH-256 [15] dataset with networks pre-trained on IMAGENET-CLS [9] and OPENIMAGES [27] without fine-tuning, then qualitatively analyze the two networks in feature space based on ℓ_2 distance. **Left/Right:** Image pairs that are closest in feature space of IMAGENET-CLS/OPENIMAGES pre-trained network (using inner-product). **Middle:** Image pairs that are close in IMAGENET-CLS pre-trained network but distant in OPENIMAGES pre-trained network.

qualitative examples for segmentation predictions in Fig 3 (using IMAGENET-LOC and IMAGENET-CLS). From these examples, we find that the network pre-trained on classification is unable to cover entire objects as it is weak at understanding instance boundaries - like in the case of the cow in Fig 3. Detection pre-training provides a better prior about the spatial extent of an instance which helps in recognizing parts of an object. It also helps more for object classes like sheep (+7.5%), cow (+6.5%), dining-table (+5.6%). These classes typically have a multi-modal distribution in appearance (like color and shape distribution). On the other hand, classes like Potted Plant which have a consistent shape and appearance, obtain no improvement in performance when detection pre-training is used.

4.3. Image Classification

We also compare the effect of pre-training for image classification by evaluating multiple pre-trained detection backbones like IMAGENET-LOC and COCO apart from OPENIMAGES. Diverse classification datasets like CALTECH-256, SUN-397 and OXFORD-102 FLOWERS are considered. Apart from fine-tuning for image classification, we also evaluate off-the-shelf features from detection and classification backbones.

Fine-Tuning on Classification Results for fine-tuning different pre-trained networks on classification datasets are shown in Table 5. These results show that pre-training on IMAGENET-CLS outperforms IMAGENET-LOC, OPENIMAGES, and COCO by a significant margin on all three classification datasets. Therefore, pre-training for object detection hurts performance for image classification. It is a bit counter-intuitive that a network which also learns about the spatial extent of an object is worse at performing image classification. To get a better understanding of the possible reasons, we evaluate features which are extracted from the

pre-trained image classification networks without any fine-tuning.

Conv5 features We average pool the Conv5 features extracted from networks pre-trained on OPENIMAGES and IMAGENET-CLS. Then we add a linear classifier followed by a softmax function to perform image classification. The results for different datasets are presented in Table 6. This shows that without fine-tuning, there exists a large performance gap between the features which are good for object detection *vs.* those which are trained for the task of image classification. The performance of the average pooled Conv5 features of IMAGENET-LOC and OPENIMAGES pre-trained networks is the same for classification on CALTECH-256. For COCO, the performance drops further by 2%, possibly because of the smaller number of classes in object detection.

Intermediate Detection Features Table 7 compares features extracted from different layers in the detection head of the OPENIMAGES pre-trained object detection network. We present results for classification on the CALTECH-256 [15] dataset when a linear classifier is applied to different features, including avg pooled Conv5 (2048), ConvProj blob (256,14,14), avg pooled ConvProj blob (256,4,4), avg pooled ConvProj blob (256,2,2), avg pooled (ConvProj) (256), FC1 (1024) and FC2 (1024) features. We find that FC1 is better than FC2. The avg pooled (ConvProj) (256) is better than avg

Pre-trained Dataset	CALTECH-256 [15]	SUN-397 [53]	OXFORD-102 FLOWERS [38]
IMAGENET-LOC [9, 46]	82.3	58.3	90.9
COCO [30]	79.8	57.8	91.4
OPENIMAGES [27]	82.2	59.5	92.6
IMAGENET-CLS [9]	86.3	61.5	95.0

Table 5: Results (Top-1 accuracy) for fine-tuning different pre-trained networks on classification datasets..

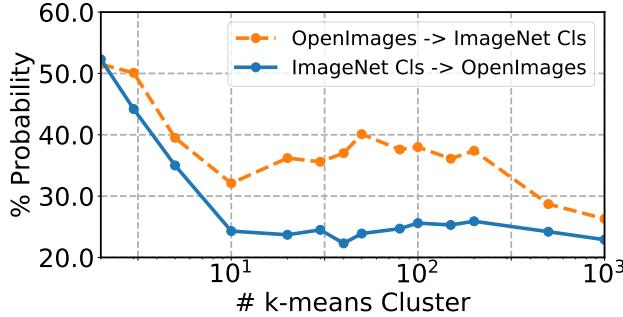


Figure 6: A pair of features within a cluster in one feature space is randomly selected and checked whether it is also assigned to the same cluster in the other feature space. We sample 100,000 pairs and measure the probability of success. The statistics indicate that similar features in OPENIMAGES space are more likely to be similar in IMAGENET-CLS space but if features are similar in IMAGENET-CLS space, it is less likely for them to be close in the OPENIMAGES space.

Pre-trained Dataset	IMAGENET-CLS [9, 46]	OPENIMAGES [27]
CALTECH-256 [15]	84.7	76.7
SUN-397 [53]	57.3	51.1
OXFORD-102 FLOWERS [38]	87.4	83.1

Table 6: Linear classification results (Top-1 Accuracy) using Conv5 features from IMAGENET-CLS and OPENIMAGES pre-trained networks.

pooled ConvProj blob (256,2,2), which is better than ConvProj blob (256,4,4). Therefore, it is evident that preserving spatial information hurts image classification. Although averaging is an operation which can be learned from a higher dimensional representation (like ConvProj blob (256,14,14)), it is also easily possible to overfit to the training set in a higher-dimensional feature space. We also find that as we approach the Output layer of detection, the per-

Feature	Top-1 Acc
Conv5	76.7
ConvProj blob (256,14,14)	69.7
ConvProj blob (256,4,4)	72.4
ConvProj blob (256,2,2)	73.3
ConvProj blob (256)	74.1
FC1 (1024)	71.6
FC2 (1024)	70.0

Table 7: Linear classification results on CALTECH-256 [15] using different features from the detection head of OPENIMAGES [27] pre-trained object detection network.

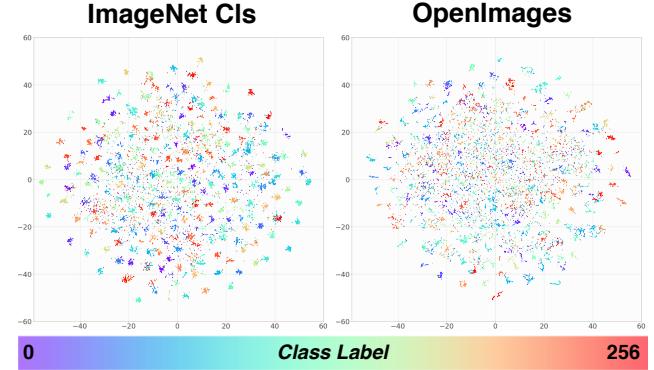


Figure 7: t-SNE [34] visualization of avg pooled Conv5 features from IMAGENET-CLS [9] and OPENIMAGES [27] pre-trained networks. The IMAGENET-CLS features are clustered while OPENIMAGES features are fragmented.

formance for image classification deteriorates.

Semantic and Feature Analysis In Fig 5 we show the most similar images (in the avg. pooled Conv5 feature space, using correlation) for IMAGENET-CLS and OPENIMAGES pre-trained networks on CALTECH-256. As can be seen, similar images from IMAGENET-CLS features can have multiple objects; however for OPENIMAGES, the most similar image pairs typically match in shape and size. To understand the relationship between OPENIMAGES and IMAGENET-CLS features, we perform K-means clustering with different numbers of clusters (from 2 to 1000). Then, given an image pair in the same cluster in an embedding (like OPENIMAGES), we check if the same image pair belongs to the same cluster in another embedding (like ImageNet) or not. We plot this probability in Fig 6. This plot shows that if features are similar in the OPENIMAGES space, they are likely to be similar in the IMAGENET-CLS space; however the converse is not true. Some example images which are close in the IMAGENET-CLS space but distant in the OPENIMAGES space are shown in the middle of Fig 5. This shows that objects of different scale and similar texture can be close in the IMAGENET-CLS space but far away in the OPENIMAGES space. We briefly describe how we define close and distant. An image pair is considered to be close if it is part of the same cluster when the number of clusters is large (> 1000). An image pair is considered to be distant if it is not part of the same cluster when the number of clusters is small (< 5).

We also show the t-SNE [34] visualization of avg pooled Conv5 features from IMAGENET-CLS and OPENIMAGES pre-trained networks before fine-tuning. We use Barnes-Hut-SNE [50] and set *perplexity* and *theta* to 10.0 and 0.5 respectively. Results in Fig 7 show that features from the same class are clustered and close to each other in the IMAGENET-CLS space; however, OPENIMAGES features

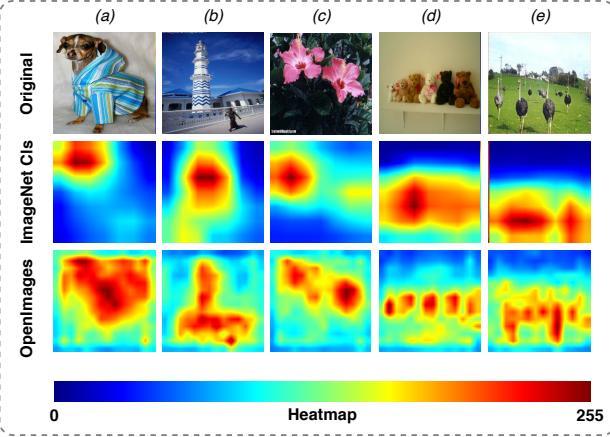


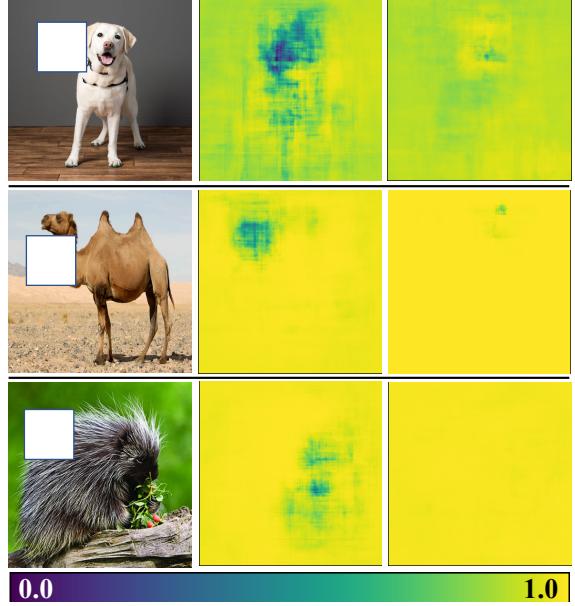
Figure 8: Activation visualization of networks pre-trained on IMAGENET-CLS [9] and OPENIMAGES [27]. Activation maps are averaged across channels, normalized and scaled to range $[0, 255]$, resized to input image size and then colorized.

are fragmented.

4.4. Visualization

Activation Visualization The previous quantitative and qualitative analysis suggests considerable differences between networks pre-trained on detection and classification datasets. To illustrate the differences in their internal representations, we first visualize the CNN activations (Conv5) and investigate which part of input images contribute more. As shown in Fig 8(a-b), the IMAGENET-CLS pre-trained activations tend to focus on discriminative parts. On the other hand, OPENIMAGES pre-trained models emphasize the entire spatial extent of the objects. Moreover, the latter exhibits an instance-level representation, especially when multiple objects are present such as Fig 8(c-e).

Mask-out visualization Besides visualizing activation maps, we further conduct “Mask-out” visualization to reveal the relationship between image parts and the final class prediction. Specifically, we shift a blank mask over the input image and measure the output confidence of the correct class. We conduct this experiment on the CALTECH-256 dataset. The classification layer for IMAGENET-CLS and the detection head for OPENIMAGES is replaced with a linear classification layer. In Fig 9, we show the classification probability at each pixel assuming that the center of the mask is placed at that location. We can see that for many locations (like the head of the dog or the camel), the classification score of the IMAGENET-CLS classifier drops to zero, which is not the case for OPENIMAGES. This is because it relies on the entire spatial extent of an object to make a prediction so the classification score is not sensitive to minor structural changes in the image. However,



(a) Image + Mask (b) ImageNet Model (c) OpenImages Model

Figure 9: Mask-out visualization. (a) A 60x60 blank mask is shifted over the image. The probability of the correct class at each mask position is shown as a probability map for the IMAGENET-CLS pre-trained and OPENIMAGES pre-trained models in (b) and (c) respectively. Unlike the OPENIMAGES model, the ImageNet-based classifier decides based on specific regions inside the image.

the IMAGENET-CLS pre-trained network classifies images based on discriminative parts and when a critical part is masked out, the classification score drops significantly.

5. Conclusion

We presented an extensive study on object detection pre-training. When fine-tuning on small detection datasets, we showed that pre-training on large detection datasets is very beneficial when a higher degree of localization is desired. Typically, detection pre-training is beneficial for tasks where spatial information is important such as detection and segmentation, but when spatial invariance is needed, like classification, it can hurt performance. Our feature-level analysis suggests that if detection features of an image are similar, it is likely that their classification features would also be similar while the converse may not hold. Visualization of activations indicates that detection networks focus more on the entire extent of an object while classification networks typically focus on parts. Thus, when minor structural changes are made to an image, detection networks would be robust compared to those trained for classification.

Acknowledgement This research was supported in part by the Office of Naval Research (ONR) under Grant N000141612713: Visual Common Sense Reasoning for Multi-agent Activity Prediction and Recognition and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of ONR, IARPA, DOI/IBC or the U.S. Government.

References

- [1] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Factors of transferability for a generic convnet representation. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1790–1802, 2016. [2](#)
- [2] W. Brendel and M. Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019. [2](#)
- [3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. [2](#)
- [4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. [2, 5](#)
- [5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. [2, 5](#)
- [6] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018. [2](#)
- [7] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. [2](#)
- [8] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. [2, 3, 5](#)
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [1, 2, 3, 4, 5, 6, 7, 8](#)
- [10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional acti-
- vation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014. [1, 2](#)
- [11] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. [2](#)
- [12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. [1, 2, 3, 4, 5](#)
- [13] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, 2019. [2](#)
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. [1](#)
- [15] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. [1, 2, 3, 6, 7](#)
- [16] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. [2](#)
- [17] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. [2](#)
- [18] K. He, R. Girshick, and P. Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018. [1, 2](#)
- [19] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. [2](#)
- [20] L. Herranz, S. Jiang, and X. Li. Scene recognition with cnns: objects, scales and dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–579, 2016. [2](#)
- [21] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. [4](#)
- [22] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. [2](#)
- [23] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. [5](#)
- [24] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*, 2018. [1, 2, 3](#)
- [25] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. [5](#)
- [26] J. Kubilius, S. Bracci, and H. P. O. de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016. [2](#)

- [27] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 1, 2, 3, 4, 5, 6, 7, 8
- [28] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 2
- [29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2
- [30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 3, 4, 6
- [31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2
- [32] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2
- [33] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
- [34] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7
- [35] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. 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
- [36] A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016. 2
- [37] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, 2016. 2
- [38] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 1, 2, 3, 6, 7
- [39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2
- [40] S. Ritter, D. G. Barrett, A. Santoro, and M. M. Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2940–2949. JMLR.org, 2017. 2
- [41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 2
- [42] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 2
- [43] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue. Dsod: Learning deeply supervised object detectors from scratch. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1919–1927, 2017. 2
- [44] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2
- [45] B. Singh and L. S. Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3587, 2018. 2
- [46] B. Singh, H. Li, A. Sharma, and L. S. Davis. R-fcn-3000 at 30fps: Decoupling detection and classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1081–1090, 2018. 1, 2, 3, 5, 6, 7
- [47] B. Singh, M. Najibi, and L. S. Davis. Sniper: Efficient multi-scale training. In *Advances in Neural Information Processing Systems*, pages 9333–9343, 2018. 2, 3, 4
- [48] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2
- [49] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1, 2
- [50] L. Van Der Maaten. Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*, 2013. 7
- [51] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2
- [52] X. Wang and A. Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018. 2
- [53] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 1, 2, 3, 6, 7
- [54] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 2
- [55] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances*

- in neural information processing systems*, pages 3320–3328, 2014. 2
- [56] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 2
- [57] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018. 2
- [58] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2
- [59] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2
- [60] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 2
- [61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2
- [62] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2018. 1, 2
- [63] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. *arXiv preprint arXiv:1811.11168*, 2018. 3