Appendix A

De-duplication Experiments

A dataset with 300M images is almost guaranteed to contain images that overlap with the validation set of target tasks. In fact, we find that even for ImageNet, there are 890 out of 50K validation images have near-duplicate images in the training.

We use visual embeddings to measure similarities and identify duplicate or near-duplicate images. The embeddings are based on deep learning features. We find there are 5536 out of 50K images in ImageNet validation set, 1648 out of 8K images in COCO minival$*$, 201 out of 4952 images in Pascal VOC 2007 test set, and 84 out of 1449 images in Pascal VOC 2012 validation set that have near duplicates in JFT-300M. We rerun several experiments by removing near-duplicate images from validation sets and then comparing performance between baselines and learned models. We observe no significant differences in trends. Table 1, 2 and 3 show that the duplicate images have minimal impact on performance for all experiments.

	Original		De-duplication	
	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.
MSRA checkpoint	76.4	92.9	76.4	92.9
Random initialization	77.5	93.9	77.5	93.8
Fine-tune from JFT-300M	79.2	94.7	79.3	94.7

Table 1. Top-1 and top-5 classification accuracy on ImageNet validation set, before and after de-duplication. Single model and single crop are used.

	Original		De-duplication	
	mAP@0.5	mAP@[0.5,0.95]	mAP@0.5	mAP@[0.5,0.95]
ImageNet	54.0	34.5	54.0	34.6
300M	57.1	36.8	56.8	36.7
ImageNet+300M	58.2	37.8	58.2	37.7

Table 2. mAP@0.5 and mAP@[0.5,0.95] for object detection performance on COCO minival$*$, before and after de-duplication.

	VOC07 Detection		VOC12 Segmentation	
	Original	De-duplication	Original	De-duplication
ImageNet	76.3	76.5	73.6	73.3
300M	81.4	81.5	75.3	75.1
ImageNet+300M	81.3	81.2	76.5	76.5

Table 3. Object detection and semantic segmentation performance on Pascal VOC, before and after deduplication. (Left) Object detection mAP@0.5 on Pascal VOC 2007 test set. (Right) Semantic segmentation mIOU on Pascal VOC 2012 validation set.

We do not conduct de-duplication experiments of COCO testdev dataset for object detection and pose estimation as their groundtruth annotations are not publicly available.

Appendix B

Detailed and Per-category Results: Object Detection

In this section, we present detailed and per-category object detection results for Table 2 (Section 5.2) from the main submission, evaluated on the COCO test-dev split. In Table 4, we report detailed AP and AR results using different initializations. In Table 7, we provide per-category AP and AP@.5 results.

	AP	AP@.5	AP@.75	AP(S)	AP(M)	AP(L)	AR	AR@.5	AR@.75	AR(S)	AR(M)	AR(L)
ImageNet	34.3	53.6	36.9	15.1	37.4	48.5	30.2	47.3	49.7	26.0	54.6	68.6
300M	36.7	56.9	39.5	17.1	40.0	50.7	31.5	49.3	51.9	28.6	56.9	70.4
ImageNet+300M	37.4	58.0	40.1	17.5	41.1	51.2	31.8	49.8	52.4	29.0	57.7	70.5

Table 4. Object detection performance on COCO test-dev split using different model initializations.

Per-category Results: Semantic Segmentation

In Table 5, we report quantitative results on the VOC 2012 segmentation validation set for all classes (refer to Figure 5 (left), Section 5.3 in the main submission). Results are reported for different initializations. We observe more than 7 point improvement for categories like boat and horse.

Initialization	mIOU	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
ImageNet	73.6	93.2	88.9	40.1	87.3	65.0	78.8	89.9	84.3	88.8	37.2	81.6	49.3	84.1	78.9	79.3	83.3	57.7	82.0	41.7	80.3	73.1
300M	75.3	93.7	89.8	40.1	89.8	70.6	78.5	89.9	86.1	92.0	36.9	80.9	52.8	87.6	82.4	80.8	84.3	61.7	84.4	44.8	80.9	72.6
ImageNet+300M	76.5	94.8	90.4	41.6	89.1	73.1	80.4	92.3	86.7	92.0	39.6	82.7	52.7	86.2	86.1	83.6	85.7	61.5	83.9	45.3	84.6	73.6

Table 5. Per-class semantic segmentation performance on PASCAL VOC 2012 validation set.

Detailed Results: Human Pose Estimation

In Table 6, we present all AP and AR results for the performance reported in Table 7 (Section 5.4) in the main submission.

	AP	AP@.5	AP@.75	AP(M)	AP(L)	AR	AR@.5	AR@.75	AR(M)	AR(L)
CMU Pose [3]	61.8	84.9	67.5	57.1	68.2	66.5	87.2	71.8	60.6	74.6
ImageNet [26]	62.4	84.0	68.5	59.1	68.1	66.7	86.6	72.0	60.8	74.9
300M	64.8	85.8	71.5	62.2	70.3	69.4	88.4	75.2	63.9	77.0
ImageNet+300M	64.4	85.7	70.7	61.8	69.8	69.1	88.2	74.8	63.7	76.6

Table 6. Human pose estimation performance on the COCO test-dev split.

Table 7. Per-class object detection performance on COCO test-dev split using different model initializations.

Initialization →	ImageNet		300M		ImageNet+300M	
	AP@.5	AP	AP@.5	AP	AP@.5	AP
person	71.5	47.7	73.1	49.8	72.7	49.9
bicycle	48.9	26.4	54.9	30.0	52.7	29.9
car	55.7	34.7	58.3	36.9	59.3	37.1
motorcycle	56.5	36.7	61.6	40.5	59.9	39.6
airplane	67.9	52.0	70.1	55.0	70.4	54.7
bus	77.7	62.5	79.5	64.6	79.0	64.2
train	66.8	59.2	69.7	62.8	69.7	62.1
truck	46.3	29.9	49.7	33.0	52.2	34.5
boat	30.6	19.4	32.5	22.1	32.1	22.3
traffic light	48.9	22.7	49.8	24.3	49.1	24.6
fire hydrant	75.3	59.1	74.4	59.3	74.9	59.5
stop sign	83.2	63.6	84.4	63.8	85.6	66.4
parking meter	62.2	37.5	64.9	38.5	64.5	37.6
bench	38.1	19.6	39.3	20.1	40.6	21.4
bird	60.2	29.4	61.9	33.0	63.3	34.2
cat	64.2	58.1	68.0	61.9	67.9	62.4
dog	62.6	52.9	66.1	56.2	66.9	57.3
horse	67.2	53.5	70.8	57.0	71.3	57.0
sheep	64.4	43.6	64.8	45.4	66.7	46.3
cow	70.7	45.4	71.9	47.4	73.3	48.9
elephant	75.1	64.1	77.3	66.4	76.1	65.5
bear	70.5	66.9	74.5	69.8	72.7	70.0
zebra	71.0	59.3	71.5	60.4	71.3	61.0
giraffe	75.3	67.4	75.9	69.0	75.9	69.3
backpack	19.6	12.8	19.5	14.7	18.5	15.1
umbrella	46.2	28.9	50.7	32.3	50.4	32.8
handbag	14.7	9.7	13.7	10.9	16.1	12.0
tie	50.8	26.3	53.2	27.9	51.5	28.4
suitcase	40.4	26.7	44.4	30.3	46.9	32.5
frisbee	53.4	43.8	55.3	48.6	58.6	48.3
skis	1.5	18.1	3.0	20.0	2.3	20.7
snowboard	45.7	29.3	47.0	33.3	43.9	32.1
sports ball	41.8	35.6	48.7	37.6	42.3	38.6
kite	39.4	37.5	33.9	38.9	35.9	40.0
baseball bat	8.3	23.4	6.7	25.1	9.9	27.5
baseball glove	35.6	27.4	33.7	31.2	41.9	31.8
skateboard	42.2	40.0	48.6	44.7	49.2	44.4
surfboard	48.5	31.1	51.7	32.8	52.4	33.9
tennis racket	53.1	42.6	55.1	44.1	55.4	45.1
bottle	61.2	28.6	61.8	30.5	61.6	30.8

Initialization →	ImageNet		300M		ImageNet+300M	
	AP@.5	AP	AP@.5	AP	AP@.5	AP
wine glass	53.8	30.2	56.3	33.3	58.7	34.7
cup	64.7	32.5	67.5	35.6	68.4	35.9
fork	45.7	23.2	45.1	26.5	50.1	27.8
knife	29.9	12.8	37.1	15.7	37.2	16.4
spoon	13.0	10.0	11.4	11.7	11.6	13.3
bowl	49.4	32.1	53.6	35.4	52.2	35.4
banana	38.1	18.7	39.8	20.4	40.0	21.1
apple	49.4	19.1	50.1	20.8	51.5	21.7
sandwich	44.0	29.6	45.2	31.3	47.8	34.1
orange	48.7	25.0	50.7	26.2	49.0	26.1
broccoli	30.6	22.9	32.5	24.8	31.9	24.6
carrot	25.9	14.0	28.6	16.1	21.5	16.4
hot dog	43.7	21.8	46.5	24.8	48.2	25.8
pizza	67.9	51.1	69.0	52.3	68.7	52.8
donut	60.2	40.1	64.8	43.9	66.8	46.4
cake	42.7	25.5	46.4	28.1	46.5	29.1
chair	33.0	21.1	36.7	24.0	35.9	24.4
couch	41.3	36.2	44.5	38.9	44.9	39.4
potted plant	25.6	20.1	27.3	21.9	30.0	23.4
bed	44.5	40.6	45.6	41.7	47.2	43.4
dining table	33.9	25.3	36.3	27.5	36.8	27.6
toilet	61.1	54.8	61.8	56.1	63.3	57.4
tv	61.8	50.0	63.0	51.9	63.7	52.7
laptop	65.8	54.5	68.3	56.6	68.9	57.5
mouse	72.1	44.4	72.0	47.6	75.6	47.3
remote	56.4	22.1	55.8	24.4	59.1	26.0
keyboard	57.1	45.4	57.5	45.9	61.4	48.3
cell phone	54.0	23.4	58.5	26.1	57.5	26.7
microwave	53.9	50.3	53.7	50.5	58.7	53.1
oven	40.9	31.7	41.9	33.5	43.2	34.6
toaster	32.6	14.7	39.9	20.5	32.9	20.1
sink	43.2	31.0	44.8	34.4	44.0	33.9
refrigerator	48.6	42.3	51.7	44.6	52.4	46.1
book	15.2	7.4	18.7	8.8	21.3	9.8
clock	56.7	43.7	56.7	45.3	55.8	45.1
vase	57.1	32.3	61.5	35.9	61.4	36.5
scissors	31.1	20.8	38.9	25.2	34.8	25.9
teddy bear	50.4	35.4	54.7	40.2	54.7	40.4
hair drier	2.3	1.0	4.8	1.8	4.0	1.9
toothbrush	48.5	34.3	50.7	36.7	51.2	37.4