Supplementary Material for Pretraining Helps When Capacity Allows: Evidence from Ultra-Small ConvNets

Srikanth Muralidharan Heitor R. Medeiros Masih Aminbeidokhti
Eric Granger Marco Pedersoli
LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
International Laboratory on Learning Systems (ILLS)

1. Small Model Generation

1.1. Smaller variants for EfficientNet

EfficientNet-B0 comprises a stem module, seven subsequent blocks, a convolutional head, and final classification components. In this section, we describe our process of scaling down EfficientNet-B0, which contains millions of parameters, to ultra-small networks with only tens of thousands of parameters.

We adopt two protocols for constructing these compact models. First, we downscale the width and depth of EfficientNet without modifying its block semantics. In the second protocol, we alter the semantics of the EfficientNet blocks by modifying the squeeze-and-excite modules.

1.1.1. Depth-Width Downscaling of EfficientNet

We downscale the width and depth of EfficientNet using its original scaling law conventions while maintaining constant input dimensions. Since our end task is object detection, we keep the input resolution unchanged to avoid degrading its performance. Specifically, EfficientNet introduces a compound scaling method that uniformly scales the network depth, width, and resolution using a set of predetermined constants [11]. We modify this compound scaling method to keep the resolution constant, while changing width and depth for our models. Our scaling operation is performed as follows:

$$\begin{aligned} \text{Depth}: \quad & d = \alpha^{\phi}, \\ \text{Width}: \quad & w = \beta^{\phi}, \end{aligned} \tag{1}$$

subject to the constraint:

$$\alpha \cdot \beta^2 \approx 2, \tag{2}$$

where ϕ is the compound scaling coefficient that controls overall model size, and α, β are constants determined via grid search to balance performance and efficiency. It is important to note here that the key distinction between compound scaling used in EfficientNet and ours is the absence of resolution scaling factor γ. The compound scaling done this way therefore enables us to stay as close to EfficientNet's original compound scaling as possible while retaining focus on our end task which is object detection in non-RGB visual domains. We start from EfficientNet-B0 and scale down until seven levels below it. Therefore the value ϕ in our case ranges from -7 to -1.

1.2. Smaller variants for MobileNet-V3

We start from Mobilenet-V3 small [1] and downscale upto six levels after the small model. With each downscaling, for each model, we roughly cut the number of parameters by roughly 50% compared to the model a level higher to it. We downsize both the width and the depth using the scaling framework from Mobilenet-v3. Further, for each downscaling we use same multiplier parameter for both width and depth multiplier.

2. Pre-training details

(b) ImageNet Pre-training Details: For the ImageNet pretraining, we pretrain our EfficientNet model families as well as MobileNetV3 model families using FFCV framework [3]. We train for 32 epochs with base learning rate of 0.05 and use input resolution of 384×384 and a total batch size of 392. We use SGD optimizer, with the weight decay of 0.0001 and momentum of 0.9. We use cyclic scheduling and keep resolution fixed throughout the training. We train our pre-training models at fixed resolutions throught training. The augmentation pipeline comprises a random resized crop, followed by a random horizontal flip with probability of 0.5 for data augmentation.

(b) COCO Pre-training Details: For COCO detection pretraining, we adopt the single-stage FCOS style object detector from Nanodet [9]. We train for 32 epochs with base learning rate of 0.005 and use input resolution of 480×384 and a total batch size of 150. We adopt a PANet [5] style feature pyramid network (FPN) as the

Model	Pretraining	mAP-ID	mAP50-ID	mAP-OOD	mAP50-OOD
B-0	Coco	0.457 (0.004)	0.854 (0.002)	0.328 (0.025)	0.619 (0.049)
B-0	Imagenet	0.439 (0.000)	0.848 (0.000)	0.326 (0.000)	0.666 (0.000)
B-0	None	0.410 (0.000)	0.799 (0.000)	0.259 (0.000)	0.547 (0.000)
B-1	Coco	0.443 (0.005)	0.831 (0.005)	0.330 (0.012)	0.642 (0.018)
B-1	Imagenet	0.408 (0.014)	0.795 (0.024)	0.304 (0.006)	0.634 (0.026)
B-1	None	0.415 (0.007)	0.801 (0.011)	0.300 (0.027)	0.607 (0.054)
B-2	Coco	0.410 (0.006)	0.792 (0.008)	0.310 (0.008)	0.655 (0.014)
B-2	Imagenet	0.388 (0.010)	0.777 (0.015)	0.299 (0.007)	0.654 (0.023)
B-2	None	0.378 (0.007)	0.752 (0.016)	0.256 (0.006)	0.548 (0.019)
B-3	Coco	0.367 (0.003)	0.761 (0.015)	0.275 (0.002)	0.617 (0.010)
B-3	Imagenet	0.360 (0.002)	0.742 (0.003)	0.245 (0.020)	0.547 (0.048)
B-3	None	0.358 (0.005)	0.737 (0.010)	0.238 (0.006)	0.531 (0.009)
B-4	Coco	0.346 (0.003)	0.748 (0.009)	0.189 (0.006)	0.469 (0.013)
B-4	Imagenet	0.342 (0.005)	0.751 (0.012)	0.146 (0.016)	0.406 (0.053)
B-4	None	0.349 (0.007)	0.738 (0.012)	0.187 (0.033)	0.442 (0.084)
B-5	Coco	0.357 (0.003)	0.760 (0.002)	0.190 (0.006)	0.457 (0.011)
B-5	Imagenet	0.340 (0.003)	0.750 (0.005)	0.152 (0.006)	0.391 (0.012)
B-5	None	0.346 (0.018)	0.730 (0.036)	0.194 (0.019)	0.448 (0.045)
B-6	Coco	0.328 (0.003)	0.716 (0.006)	0.126 (0.013)	0.322 (0.035)
B-6	Imagenet	0.326 (0.004)	0.726 (0.010)	0.162 (0.020)	0.403 (0.050)
B-6	None	0.339 (0.007)	0.731 (0.015)	0.199 (0.020)	0.490 (0.056)
B-7	Coco	0.322 (0.002)	0.709 (0.009)	0.204 (0.006)	0.519 (0.018)
B-7	Imagenet	0.313 (0.007)	0.710 (0.012)	0.152 (0.008)	0.400 (0.018)
B-7	None	0.331 (0.010)	0.735 (0.009)	0.151 (0.067)	0.365 (0.147)

Table 1. Performance comparison on LLVIP dataset [2] for EfficientNet family models.

Model	Pretraining	mAP-ID	mAP50-ID	mAP-OOD	mAP50-OOD
S-0	Coco	0.391 (0.011)	0.763 (0.019)	0.263 (0.009)	0.590 (0.011)
S-0	Imagenet	0.397 (0.000)	0.807 (0.000)	0.272 (0.000)	0.596 (0.000)
S-0	None	0.359 (0.000)	0.751 (0.000)	0.237 (0.000)	0.515 (0.000)
S-1	Coco	0.383 (0.001)	0.763 (0.002)	0.240 (0.024)	0.523 (0.054)
S-1	Imagenet	0.370 (0.005)	0.757 (0.008)	0.257 (0.017)	0.591 (0.019)
S-1	None	0.358 (0.011)	0.733 (0.013)	0.221 (0.016)	0.513 (0.027)
S-2	Coco	0.377 (0.002)	0.781 (0.006)	0.222 (0.026)	0.521 (0.050)
S-2	Imagenet	0.364 (0.004)	0.764 (0.002)	0.243 (0.005)	0.559 (0.020)
S-2	None	0.351 (0.020)	0.722 (0.040)	0.224 (0.031)	0.521 (0.065)
S-3	Coco	0.348 (0.003)	0.732 (0.013)	0.207 (0.020)	0.490 (0.047)
S-3	Imagenet	0.344 (0.001)	0.730 (0.004)	0.189 (0.024)	0.457 (0.049)
S-3	None	0.322 (0.005)	0.687 (0.011)	0.222 (0.009)	0.528 (0.003)
S-4	Coco	0.341 (0.003)	0.731 (0.008)	0.234 (0.007)	0.577 (0.009)
S-4	Imagenet	0.321 (0.005)	0.709 (0.015)	0.210 (0.006)	0.505 (0.011)
S-4	None	0.311 (0.009)	0.671 (0.015)	0.176 (0.015)	0.439 (0.026)
S-5	Coco	0.343 (0.005)	0.743 (0.007)	0.131 (0.008)	0.320 (0.022)
S-5	Imagenet	0.317 (0.004)	0.709 (0.013)	0.139 (0.027)	0.340 (0.064)
S-5	None	0.329 (0.015)	0.717 (0.025)	0.214 (0.006)	0.519 (0.012)
S-6	Coco	0.299 (0.001)	0.675 (0.005)	0.214 (0.008)	0.519 (0.012)
S-6	Imagenet	0.287 (0.010)	0.663 (0.026)	0.211 (0.012)	0.538 (0.021)
S-6	None	0.312 (0.007)	0.689 (0.014)	0.180 (0.009)	0.453 (0.029)

Table 2. Performance comparison on LLVIP dataset [2] for MobilenetV3 family models.

Model	Pretraining	mAP-ID	mAP50-ID	mAP-OOD	mAP50-OOD
B-0	Coco	0.269 (0.003)	0.593 (0.005)	0.147 (0.008)	0.373 (0.022)
B-0	Imagenet	0.249 (0.000)	0.560 (0.000)	0.122 (0.000)	0.303 (0.000)
B-0	None	0.256 (0.000)	0.560 (0.000)	0.113 (0.000)	0.277 (0.000)
B-1	Coco	0.259 (0.002)	0.568 (0.006)	0.108 (0.001)	0.278 (0.004)
B-1	Imagenet	0.245 (0.001)	0.545 (0.003)	0.096 (0.007)	0.252 (0.023)
B-1	None	0.239 (0.000)	0.521 (0.002)	0.070 (0.027)	0.169 (0.072)
B-2	Coco	0.226 (0.002)	0.508 (0.004)	0.086 (0.004)	0.223 (0.011)
B-2	Imagenet	0.214 (0.002)	0.489 (0.005)	0.076 (0.007)	0.190 (0.013)
B-2	None	0.203 (0.002)	0.458 (0.006)	0.054 (0.027)	0.137 (0.068)
B-3	Coco	0.196 (0.002)	0.443 (0.003)	0.052 (0.002)	0.129 (0.003)
B-3	Imagenet	0.182 (0.002)	0.424 (0.006)	0.044 (0.002)	0.120 (0.006)
B-3	None	0.181 (0.004)	0.413 (0.012)	0.043 (0.012)	0.113 (0.036)
B-4	Coco	0.175 (0.001)	0.402 (0.003)	0.049 (0.003)	0.142 (0.006)
B-4	Imagenet	0.163 (0.000)	0.381 (0.002)	0.032 (0.002)	0.090 (0.004)
B-4	None	0.161 (0.003)	0.371 (0.006)	0.036 (0.013)	0.098 (0.034)
B-5	Coco	0.184 (0.002)	0.423 (0.005)	0.045 (0.002)	0.117 (0.008)
B-5	Imagenet	0.168 (0.002)	0.393 (0.005)	0.029 (0.005)	0.075 (0.012)
B-5	None	0.162 (0.005)	0.376 (0.010)	0.037 (0.010)	0.103 (0.027)
B-6	Coco	0.171 (0.003)	0.399 (0.006)	0.033 (0.002)	0.091 (0.006)
B-6	Imagenet	0.160 (0.002)	0.380 (0.006)	0.030 (0.007)	0.089 (0.017)
B-6	None	0.160 (0.004)	0.371 (0.006)	0.028 (0.005)	0.078 (0.015)
B-7	Coco	0.169 (0.001)	0.394 (0.004)	0.036 (0.003)	0.104 (0.010)
B-7	Imagenet	0.157 (0.001)	0.371 (0.005)	0.026 (0.003)	0.069 (0.007)
B-7	None	0.159 (0.005)	0.369 (0.010)	0.036 (0.010)	0.099 (0.027)

Table 3. Performance comparison on FLIR dataset[12] for EfficientNet family models.

Model	Pretraining	mAP-ID	mAP50-ID	mAP-OOD	mAP50-OOD
S-0	Coco	0.216 (0.000)	0.485 (0.000)	0.088 (0.000)	0.225 (0.000)
S-0	Imagenet	0.202 (0.001)	0.463 (0.002)	0.059 (0.003)	0.157 (0.006)
S-0	None	0.184 (0.003)	0.416 (0.002)	0.043 (0.006)	0.118 (0.014)
S-1	Coco	0.195 (0.001)	0.447 (0.004)	0.071 (0.006)	0.186 (0.018)
S-1	Imagenet	0.181 (0.003)	0.419 (0.004)	0.055 (0.003)	0.155 (0.014)
S-1	None	0.177 (0.002)	0.404 (0.007)	0.039 (0.005)	0.107 (0.015)
S-2	Coco	0.182 (0.002)	0.416 (0.000)	0.052 (0.002)	0.143 (0.003)
S-2	Imagenet	0.172 (0.003)	0.395 (0.005)	0.047 (0.001)	0.127 (0.002)
S-2	None	0.174 (0.003)	0.393 (0.006)	0.036 (0.007)	0.101 (0.024)
S-3	Coco	0.164 (0.001)	0.378 (0.003)	0.050 (0.003)	0.141 (0.010)
S-3	Imagenet	0.156 (0.003)	0.358 (0.004)	0.023 (0.004)	0.067 (0.009)
S-3	None	0.163 (0.003)	0.373 (0.008)	0.036 (0.005)	0.102 (0.013)
S-4	Coco	0.164 (0.001)	0.381 (0.003)	0.037 (0.001)	0.121 (0.003)
S-4	Imagenet	0.150 (0.001)	0.354 (0.004)	0.031 (0.004)	0.090 (0.012)
S-4	None	0.150 (0.007)	0.348 (0.014)	0.040 (0.003)	0.115 (0.007)
S-5	Coco	0.159 (0.001)	0.373 (0.002)	0.030 (0.001)	0.085 (0.004)
S-5	Imagenet	0.141 (0.001)	0.338 (0.004)	0.022 (0.002)	0.067 (0.007)
S-5	None	0.141 (0.001)	0.322 (0.004)	0.035 (0.005)	0.101 (0.014)
S-6	Coco	0.143 (0.001)	0.339 (0.004)	0.040 (0.002)	0.114 (0.004)
S-6	Imagenet	0.136 (0.002)	0.323 (0.004)	0.023 (0.003)	0.069 (0.005)
S-6	None	0.140 (0.002)	0.330 (0.005)	0.023 (0.003)	0.066 (0.007)

Table 4. Performance comparison on FLIR dataset[12] for MobileNetV3 family models.

Model	Pretraining	mAP-ID	mAP50-ID	mAP-OOD	mAP50-OOD
B-0	Coco	0.705 (0.002)	0.948 (0.001)	0.663 (0.004)	0.893 (0.007)
B-0	Imagenet	0.709 (0.002)	0.950 (0.001)	0.655 (0.011)	0.880 (0.013)
B-0	None	0.698 (0.003)	0.943 (0.003)	0.665 (0.010)	0.909 (0.017)
B-1	Coco	0.711 (0.003)	0.947 (0.001)	0.662 (0.004)	0.894 (0.007)
B-1	Imagenet	0.711 (0.001)	0.947 (0.000)	0.657 (0.008)	0.891 (0.010)
B-1	None	0.691 (0.003)	0.941 (0.001)	0.645 (0.015)	0.886 (0.020)
B-2	Coco	0.629 (0.003)	0.923 (0.000)	0.607 (0.002)	0.877 (0.002)
B-2	Imagenet	0.625 (0.001)	0.922 (0.002)	0.600 (0.010)	0.863 (0.012)
B-2	None	0.610 (0.001)	0.906 (0.005)	0.596 (0.006)	0.860 (0.007)
B-3	Coco	0.584 (0.011)	0.891 (0.008)	0.581 (0.004)	0.868 (0.003)
B-3	Imagenet	0.582 (0.003)	0.894 (0.004)	0.555 (0.008)	0.825 (0.008)
B-3	None	0.524 (0.089)	0.839 (0.072)	0.525 (0.077)	0.801 (0.080)
B-4	Coco	0.554 (0.003)	0.873 (0.003)	0.563 (0.013)	0.855 (0.014)
B-4	Imagenet	0.546 (0.003)	0.871 (0.000)	0.545 (0.006)	0.832 (0.015)
B-4	None	0.558 (0.002)	0.874 (0.000)	0.556 (0.008)	0.831 (0.014)
B-5	Coco	0.557 (0.002)	0.879 (0.000)	0.544 (0.012)	0.827 (0.021)
B-5	Imagenet	0.550 (0.006)	0.869 (0.007)	0.541 (0.002)	0.820 (0.006)
B-5	None	0.563 (0.006)	0.868 (0.005)	0.554 (0.025)	0.825 (0.034)
B-6	Coco	0.559 (0.004)	0.872 (0.004)	0.540 (0.015)	0.810 (0.025)
B-6	Imagenet	0.549 (0.001)	0.865 (0.003)	0.560 (0.003)	0.845 (0.002)
B-6	None	0.562 (0.002)	0.876 (0.000)	0.553 (0.011)	0.833 (0.018)
B-7	Coco	0.561 (0.002)	0.870 (0.001)	0.539 (0.016)	0.804 (0.022)
B-7	Imagenet	0.562 (0.000)	0.875 (0.003)	0.548 (0.004)	0.830 (0.005)
B-7	None	0.536 (0.039)	0.856 (0.029)	0.537 (0.022)	0.821 (0.013)

Table 5. Performance comparison on LLVIP dataset [2] for EfficientNet family models.

Model	Pretraining	mAP-ID	mAP50-ID	mAP-OOD	mAP50-OOD
S-0	Coco	0.620 (0.007)	0.911 (0.002)	0.584 (0.009)	0.843 (0.015)
S-0	Imagenet	0.631 (0.004)	0.923 (0.004)	0.580 (0.007)	0.841 (0.012)
S-0	None	0.615 (0.002)	0.910 (0.005)	0.575 (0.012)	0.843 (0.018)
S-1	Coco	0.611 (0.003)	0.910 (0.002)	0.571 (0.024)	0.834 (0.030)
S-1	Imagenet	0.609 (0.003)	0.907 (0.004)	0.578 (0.016)	0.841 (0.021)
S-1	None	0.594 (0.001)	0.894 (0.003)	0.579 (0.002)	0.851 (0.003)
S-2	Coco	0.607 (0.002)	0.905 (0.001)	0.578 (0.004)	0.848 (0.005)
S-2	Imagenet	0.607 (0.003)	0.906 (0.002)	0.556 (0.017)	0.812 (0.022)
S-2	None	0.603 (0.008)	0.900 (0.002)	0.589 (0.009)	0.862 (0.015)
S-3	Coco	0.587 (0.001)	0.896 (0.002)	0.566 (0.018)	0.844 (0.030)
S-3	Imagenet	0.586 (0.003)	0.896 (0.001)	0.550 (0.011)	0.814 (0.018)
S-3	None	0.582 (0.003)	0.893 (0.001)	0.575 (0.002)	0.850 (0.008)
S-4	Coco	0.539 (0.000)	0.863 (0.006)	0.540 (0.013)	0.826 (0.017)
S-4	Imagenet	0.546 (0.004)	0.865 (0.005)	0.522 (0.004)	0.803 (0.001)
S-4	None	0.540 (0.006)	0.856 (0.005)	0.519 (0.010)	0.798 (0.012)
S-5	Coco	0.535 (0.002)	0.861 (0.004)	0.540 (0.015)	0.830 (0.019)
S-5	Imagenet	0.543 (0.008)	0.867 (0.007)	0.537 (0.004)	0.818 (0.008)
S-5	None	0.542 (0.004)	0.858 (0.002)	0.545 (0.003)	0.827 (0.004)
S-6	Coco	0.541 (0.005)	0.865 (0.008)	0.542 (0.012)	0.831 (0.019)
S-6	Imagenet	0.542 (0.002)	0.860 (0.001)	0.535 (0.007)	0.818 (0.019)
S-6	None	0.545 (0.008)	0.860 (0.006)	0.544 (0.011)	0.820 (0.014)

Table 6. Performance comparison on Distech dataset for MobileNetV3 family models.

Model	Pretraining	mAP-OOD	mAP50-OOD
B-0	Coco	32.930 (4.330)	12.230 (1.580)
B-0	Imagenet	21.050 (0.550)	7.800 (0.100)
B-0	Coco	32.370 (5.340)	12.600 (1.850)
B-1	Imagenet	23.870 (3.790)	8.230 (1.430)
B-1	None	4.900 (2.660)	1.870 (1.020)
B-1	Coco	21.630 (3.270)	8.370 (0.980)
B-2	Imagenet	19.330 (2.160)	6.800 (0.940)
B-2	None	2.130 (2.050)	0.900 (0.700)
B-2	Coco	26.930 (1.730)	9.530 (0.590)
B-3	Imagenet	10.170 (1.320)	3.130 (0.210)
B-3	None	1.500 (1.770)	0.530 (0.610)
B-3	Coco	8.400 (6.520)	2.830 (2.110)
B-4	Imagenet	3.630 (1.440)	1.030 (0.450)
B-4	None	0.600 (0.780)	0.170 (0.240)
B-4	Coco	18.600 (2.870)	6.200 (1.350)
B-5	Imagenet	9.000 (0.830)	2.900 (0.220)
B-5	None	2.100 (1.000)	0.830 (0.120)
B-5	Coco	7.630 (1.130)	2.500 (0.360)
B-6	Imagenet	7.270 (1.960)	2.230 (0.760)
B-6	None	0.470 (0.330)	0.130 (0.120)
B-6	Coco	13.430 (2.490)	4.770 (1.010)
B-7	Imagenet	7.130 (2.380)	2.400 (0.820)
B-7	None	0.330 (0.340)	0.070 (0.090)

Table 7. Performance comparison on cross-dataset FLIR→LLVIP RGB on person class for EfficientNet family models.

neck with an output channel dimension of 96 operating on three output scales. For the detection head, we use NanoDetHead [9], consisting of two stacked convolutional layers with ReLU6 activation and batch normalization. The head uses an input and feature channel dimension of 96, operates with strides (s) of s=[8, 16, 32], and shares classification and regression features. It predicts bounding boxes with a maximum regression bin index (reg_max) of 7, using an octave base scale of 5 with 1 scale per octave. The loss function is composed of a Quality Focal Loss [4] for classification ($\beta = 2.0$, weight=1.0), a Distribution Focal Loss [4] for bounding box distribution regression (weight=0.25), and a GIoU Loss [10] for bounding box localization (weight=2.0). We use SGD optimizer, with the weight decay of 0.0001 and momentum of 0.9. We use cyclic scheduling and keep resolution fixed throughout the training. We train our pre-training models at fixed resolutions throught training. The augmentation pipeline comprises a random resized crop, followed by a random horizontal flip with probability of 0.5 for data augmentation. For data augmentation, we apply horizontal flip with a probability of 0.5, translation with a probability of 0.2, random scaling between 0.5 and 1.5, apply color augmentations. We train our models with batch size of 150, use AdamW [7] optimizer with linear warmup for 500 steps and CosineAnnealing [6] learning rate scheduler. We train Object detector in LLVIP for 32 epochs and in FLIR for 80 epochs. We use base learning rate of 0.0005.

3. Pre-training performance of Models

3.1. Imagenet classification

The top-1 and top-5 accuracy of all our models in our two model families at the end of this training is shown in Tab. 13.

3.2. COCO Object Detection

The mAP and mAP50 of all our models in our two model families at the end of COCO detection pre-training is shown in Tab. 14.

4. Detailed results

4.1. In-domain and out-domain results

In modality and RGB to IR cross-modality results: Here we report tabular results for LLVIP, FLIR datasets. LLVIP results are presented in tables 1 and 2 for EfficientNet and MobileNetV3 model families respectively. FLIR dataset's numbers are presented in tables 3 and 4. **Cross-dataset robustness results:** Here, we provide details of results on FLIR→LLVIP and LLVIP→FLIR in RGB domain in 7,8,9, and 10 respectively. **DomainNet benchmark results:** Here, we provide details of results obtained using EfficientNet and MobileNetV3 models in tables 11 and 12 respectively. **ID-viewpoint, OOD-viewpoint In-domain and out-domain results**: Here, we provide details of performance of the models on Distech dataset are reported in tables 5 and 6 for EfficientNet and MobileNetV3 models respectively.

References

[1] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[2] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021. 2, 4

[3] Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, and Aleksander Madry. ffcv. https://github.com/libffcv/ffcv/, 2022. commit xxxxxxx. 1

[4] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in neural information processing systems*, 33:21002–21012, 2020. 5

Model	Pretraining	mAP-OOD	mAP50-OOD
S-0	Imagenet	12.300 (4.580)	4.250 (1.710)
S-0	None	3.500 (3.350)	1.170 (1.020)
S-1	Coco	24.930 (3.970)	9.500 (1.680)
S-1	Imagenet	17.970 (3.450)	6.170 (1.390)
S-1	None	0.470 (0.330)	0.100 (0.080)
S-2	Coco	19.170 (1.350)	6.970 (0.410)
S-2	Imagenet	13.700 (2.550)	4.670 (0.970)
S-2	None	0.600 (0.450)	0.270 (0.190)
S-3	Coco	13.500 (2.890)	4.670 (1.110)
S-3	Imagenet	8.470 (3.920)	2.700 (1.180)
S-3	None	0.900 (0.370)	0.370 (0.190)
S-4	Coco	14.470 (2.400)	5.170 (0.760)
S-4	Imagenet	12.800 (3.680)	4.000 (0.860)
S-4	None	1.070 (0.450)	0.300 (0.080)
S-5	Coco	11.300 (0.880)	3.830 (0.310)
S-5	Imagenet	2.830 (3.020)	0.930 (1.040)
S-5	None	1.030 (0.540)	0.300 (0.280)
S-6	Coco	13.270 (2.520)	4.530 (0.920)
S-6	Imagenet	2.230 (0.900)	0.670 (0.260)
S-6	None	1.400 (0.290)	0.370 (0.170)

Table 8. Cross-domain performance for LLVIP→FLIR RGB on person class for MobileNetV3 family models.

Model	Pretraining	mAP-OOD	mAP50-OOD
B-0	Coco	9.730 (1.170)	3.430 (0.380)
B-0	Imagenet	7.130 (0.710)	2.500 (0.280)
B-0	None	4.800 (1.420)	1.830 (0.390)
B-1	Coco	8.030 (0.050)	2.770 (0.050)
B-1	Imagenet	6.830 (1.470)	2.600 (0.510)
B-1	None	5.170 (0.260)	1.970 (0.120)
B-2	Coco	6.870 (1.170)	2.570 (0.450)
B-2	Imagenet	4.530 (0.450)	1.670 (0.210)
B-2	None	4.130 (0.840)	1.530 (0.450)
B-3	Coco	3.700 (0.330)	1.400 (0.220)
B-3	Imagenet	3.600 (0.850)	1.270 (0.250)
B-3	None	2.870 (0.340)	1.100 (0.140)
B-4	Coco	3.700 (0.490)	1.400 (0.080)
B-4	Imagenet	3.330 (0.630)	1.200 (0.290)
B-4	None	2.170 (0.340)	0.970 (0.190)
B-5	Coco	2.630 (0.120)	1.070 (0.050)
B-5	Imagenet	3.130 (0.330)	1.170 (0.050)
B-5	None	1.830 (1.300)	0.700 (0.510)
B-6	Coco	2.770 (0.120)	1.000 (0.000)
B-6	Imagenet	2.300 (0.370)	0.930 (0.050)
B-6	None	2.900 (0.410)	1.170 (0.120)
B-7	Coco	2.130 (0.520)	0.730 (0.250)
B-7	Imagenet	2.230 (0.340)	1.030 (0.050)
B-7	None	1.970 (0.540)	0.870 (0.260)

Table 9. Performance comparison on cross-dataset LLVIP→FLIR RGB on person class for EfficientNet family models.

Model	Pretraining	mAP-OOD	mAP50-OOD
S-0	Coco	4.530 (0.250)	2.000 (0.140)
S-0	Imagenet	3.950 (0.650)	1.450 (0.250)
S-0	None	2.970 (1.190)	1.030 (0.460)
S-1	Coco	4.230 (0.540)	1.770 (0.190)
S-1	Imagenet	3.200 (0.400)	1.400 (0.200)
S-1	None	3.270 (0.560)	1.200 (0.140)
S-2	Coco	4.300 (0.080)	1.570 (0.050)
S-2	Imagenet	3.770 (0.480)	1.470 (0.260)
S-2	None	3.030 (0.120)	1.130 (0.050)
S-3	Coco	2.700 (0.160)	1.130 (0.050)
S-3	Imagenet	2.630 (0.170)	1.070 (0.050)
S-3	None	2.430 (0.400)	1.000 (0.160)
S-4	Coco	3.500 (0.570)	1.300 (0.360)
S-4	Imagenet	3.170 (0.210)	1.230 (0.120)
S-4	None	1.770 (0.980)	0.670 (0.400)
S-5	Coco	3.070 (0.390)	1.230 (0.090)
S-5	Imagenet	2.750 (0.150)	1.100 (0.100)
S-5	None	2.150 (0.050)	0.800 (0.100)
S-6	Coco	2.070 (0.450)	0.800 (0.140)
S-6	Imagenet	2.700 (0.240)	1.100 (0.080)
S-6	None	2.170 (0.170)	0.830 (0.210)

Table 10. Performance comparison on cross-dataset LLVIP→FLIR RGB on person class for MobileNetV3 family models.

Model	Pretraining	Accuracy
B-0	Imagenet	39.439
B-1	Imagenet	34.993
B-2	Imagenet	25.999
B-3	Imagenet	17.002
B-4	Imagenet	9.303
B-5	Imagenet	6.772
B-6	Imagenet	4.864
B-7	Imagenet	3.381
B-0	None	29.474
B-1	None	26.568
B-2	None	21.553
B-3	None	16.995
B-4	None	10.968
B-5	None	9.289
B-6	None	7.114
B-7	None	5.157

Table 11. Performance comparison on DomainNet [8] benchmark for EfficientNet family models.

[5] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[6] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[8] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 7, 8

[9] RangiLyu. Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model. https:

Model	Pretraining	Accuracy
S-0	Imagenet	28.181
S-1	Imagenet	24.059
S-2	Imagenet	19.437
S-3	Imagenet	14.363
S-4	Imagenet	11.427
S-5	Imagenet	8.104
S-6	Imagenet	4.586
S-0	None	22.025
S-1	None	18.735
S-2	None	15.613
S-3	None	11.349
S-4	None	9.588
S-5	None	8.247
S-6	None	4.773

Table 12. Performance comparison on DomainNet [8] benchmark for EfficientNet family models.

Model	**Top-1 Acc.**	**Top-5 Acc.**
EfficientNet		
B-0	0.673	0.882
B-1	0.635	0.857
B-2	0.554	0.798
B-3	0.461	0.720
B-4	0.354	0.605
B-5	0.316	0.563
B-6	0.259	0.497
B-7	0.210	0.426
MobileNetV3		
S	0.55	0.795
S-1	0.507	0.757
S-2	0.450	0.711
S-3	0.391	0.651
S-4	0.360	0.613
S-5	0.323	0.572
S-6	0.249	0.476

Table 13. **ImageNet top-1 and top-5 accuracy for EfficientNet and MobileNetV3 model families at input resolution 384×384 except for EfficientNet-B0 and MobileNet-V3 S**.

//github.com/RangiLyu/nanodet, 2021. 1, 5

[10] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[12] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pages 276–280. IEEE, 2020. 3

Model	**mAP**	**mAP50**
EfficientNet		
B-0	0.281	0.471
B-1	0.242	0.403
B-2	0.178	0.317
B-3	0.124	0.237
B-4	0.082	0.169
B-5	0.077	0.161
B-6	0.068	0.144
B-7	0.059	0.127
MobileNetV3		
S	0.209	0.365
S-1	0.16	0.29
S-2	0.132	0.246
S-3	0.105	0.203
S-4	0.088	0.177
S-5	0.073	0.15
S-6	0.055	0.119

Table 14. **COCO Detection mAP and mAP50 for EfficientNet and MobileNetV3 model families at input resolution 480×384 except for EfficientNet and MobileNetV3**.