Supplementary Material for Compatibility-aware Heterogeneous Visual Search

Rahul Duggal* Hao Zhou Shuo Yang Yuanjun Xiong
Wei Xia† Zhuowen Tu Stefano Soatto

AWS/Amazon AI

rduggal7@gatech.edu {zhouho, shuoy, yuanjx, wxia, ztu, soattos}@amazon.com

A. Implementation Details

A.1. Training, Validation and Testing Dataset

For searching the best query architecture, we carve out a small validation split from the original training set. For the face tasks, we set aside 5% from the training set of IMDB [5] while for the fashion tasks, we set aside 10% from the training set of DeepFashion2 [1]. The remaining portions of the training sets are actually used to train all our embedding models (query supernet, gallery model, final query models). After a super-network is trained, we evaluate the performance of each candidate architecture (we refer to it as a sub-network) on the held out validation split. The final results presented in this paper are reported on the original validation portions of IMDB and DeepFashion2.

A.2. Designing and Training the Super-network

For each computational tier (330, 230, 100 Mflops), we train a different super-network. For the 300 Mflops tier, our super-network is the same as that in [2]. For the 230 and 100 Mflops tiers, we reduce the channel widths by $0.75\times$ and $0.5\times$ in each layer. The super-network is trained through a sampling process: In each batch, a new architecture (we call this a sub-network) is sampled and only the weights corresponding to it are updated. For sampling a sub-network, we use the parameter free *uniform sampling* method. This means that, for each layer, the chosen block (includes four choices from 0-3) and channels width (includes ten choices from 0-9) are sampled uniformly. We notice that the super-network fails to converge if the sampling process is started from the first epoch. To solve this, we use a warm-up phase of 10 epochs wherein the the super-network is trained without sampling. During the warm-up phase, the output of all four blocks in each layer are combined through averaging and the largest channel width is used.

A.3. Details of the evolutionary search

We reuse the same hyper-parameters from [2] for the evolutionary search step. Specifically, we search for 20 generations, each with a population size of 50, crossover size of 40, mutate chance of 0.1 and random select chance of 0.1. To guide the evolutionary search for finding the most compatible architectures, we use reward $\mathcal{R}_3$ from Tab.1 in the main paper. For the face tasks, we compute this reward on the IMDB "validation" split using the 1:1 verification metric of TAR@FAR=10^{-3}. For the fashion tasks, we compute this reward on the DeepFashion2 validation split using the top-50 metric. Note that our rewards metrics (TAR@FAR=10^{-3} for face, top-50 for fashion) are different from the target metrics (TAR@FAR=10^{-4}/TNIR@FPIR=10^{-1} for face, top-10 for fashion). This is mainly because the validation split is smaller (than the test split), and thus target metric (*e.g.* top-10 accuracy) is noisy compared to the validation metric (*e.g.* top-50 accuracy).

B. Additional Results under Different Evaluation Metrics

Due to space limits, in the main paper, we present one evaluation metric per task. In this section, we present the full metric results according to IJB-series and DeepFashion2 benchmark standard for reference. More specifically, in Sec. B.1 we show top-k search accuracy on face retrieval task. In Sec B.2, we evaluate our CMP-NAS on face verification task at additional operating points; In Sec B.3, we show the results of the proposed method using top-1, top-10 and top-20 retrieval accuracy on fashion retrieval task. All these additional results further demonstrate that (1) With CMP-NAS, the compatibility rule holds; (2) The architectures searched with CMP-NAS outperform other baselines for both homogeneous and heterogeneous search accuracy.

*Currently at the Georgia Institute of Technology. Work conducted during an internship with Amazon AI.

†Corresponding author

Query Model	MFlops	Homogeneous Acc. Top-k with k=			Heterogeneous Acc Top-k with k=		
		1	5	10	1	5	10
ResNet-101	7597	91.1	95.0	96.1	-	-	-
MobileNetV1	579	80.0	88.9	91.5	83.5	91.4	93.7
MobileNetV2	329	85.8	92.2	94.2	88.1	93.8	95.2
ProxylessNAS	332	86.3	92.5	94.4	88.5	93.9	95.4
CMP-NAS-a(Face)	**327**	**89.7**	**94.2**	**95.5**	**90.7**	**94.7**	**96.1**
MobileNetV3	226	85.6	92.1	94.0	88.0	93.5	95.2
CMP-NAS-b(Face)	**216**	**88.2**	**93.5**	**95.2**	**89.8**	**94.5**	**95.9**
MobileNetV1(0.5x)	155	74.1	77.5	85.3	77.5	88.3	91.3
ShuffleNetV2	149	81.6	89.8	92.2	85.0	92.0	94.1
ShuffleNetV1(g=1)	148	81.3	89.7	92.1	85.1	92.1	94.0
MobileNetV2(0.5x)	100	80.0	88.5	91.3	83.6	90.9	93.3
CMP-NAS-c(Face)	**94**	**84.3**	**91.4**	**93.4**	**86.9**	**93.1**	**94.9**

Table 1: Extending Tab. 5 of the main paper. Evaluating CMP-NAS on the IJB-C 1:N face retrieval benchmark using two additional metrics: top-1, top-5 top-10 accuracy. Observe that the models discovered with CMP-NAS comprehensively outperform the baselines on both, homogeneous and heterogeneous accuracy.

Query Model	MFlops	Homogeneous Acc. TAR@FAR=			Heterogeneous Acc. TAR@FAR=		
		10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
Resnet-101(gallery)	7597	96.9	92.8	85.4	-	-	-
MobileNetV1(1x)	579	93.2	82.6	66.7	95.0	86.6	73.0
MobileNetV2(1x)	329	95.6	88.1	75.4	96.5	91.0	80.8
ProxyLess(mobile)	332	95.7	88.2	75.5	96.5	90.7	80.3
CMP-NAS-a(Face)	**327**	**96.7**	**91.5**	**81.6**	**97.1**	**92.7**	**84.5**
MobileNetV3	226	95.5	88.0	74.3	96.5	90.9	79.9
CMP-NAS-b(Face)	**216**	**96.3**	**90.2**	**79.0**	**96.9**	**92.2**	**82.8**
MobileNetV1(0.5x)	155	90.8	76.9	58.0	93.4	82.1	64.3
ShuffleNetV2(1x)	149	93.7	83.8	66.8	95.4	88.7	74.8
MobileNetV2(0.5x)	100	93.3	82.0	64.8	94.9	86.8	72.8
CMP-NAS-c(Face)	**94**	**95.1**	**86.6**	**71.5**	**96.1**	**90.2**	**78.3**

Table 2: Extending Tab. 6 of the main paper. Evaluating the models CMP-NAS-a,b,c(Face) on the 1:1 face verification task using IJB-C using additional operating points. The searched models outperform the baselines indicating they can generalize across tasks.

B.1. Additional Results on Face Retrieval

Tab. 1 extends Tab.5 in the main paper by including other popular metrics (top-1, top-5 and top-10) for the face retrieval task. Additionally, we include the homogeneous accuracy achieved by the models.

B.2. Additional Results on Face Verification

Besides face retrieval, face verification is another popular task in the "open-universal" problem of face recognition. in Tab. 2, we extend Tab. 6 of the main paper by showing the results on additional operating points (FAR=$10^{-2}, 10^{-3}, 10^{-4}$).

B.3. Additional Results on Fashion Retrieval

Tab. 3 extends Tab. 5 in our paper by showing the homogeneous and heterogeneous accuracy through the top-1, top-10 and top-20 metrics.

Query Model	MFlops	Homogeneous Acc. Top-k with k as			Heterogeneous Acc Top-k with k as		
		1	10	20	1	10	20
ResNet-101		39.4	65.1	72.0	-	-	-
MobileNetV1	579	34.7	60.5	67.8	36.3	62.3	69.1
MobileNetV2	329	32.4	58.0	65.9	33.9	60.4	67.9
ProxylessNAS	332	35.1	60.8	68.5	36.6	62.1	69.4
CMP-NAS-a(Fashion)	**314**	**39.0**	**65.4**	**72.4**	**39.3**	**65.6**	**72.5**
MobileNetV3	226	37.1	62.7	69.9	37.5	63.0	70.2
CMP-NAS-b(Fashion)	**211**	**38.2**	**64.0**	**71.2**	**38.4**	**64.9**	**72.2**
MobileNetV1(0.5x)	155	32.8	57.7	65.6	34.0	60.2	67.5
ShuffleNetV2	149	35.4	60.7	68.1	35.7	62.1	69.7
ShuffleNetV1(g=1)	148	34.4	60.5	68.1	35.3	62.6	69.8
CMP-NAS-c(Fashion)	**93**	**37.6**	**63.5**	**71.0**	**38.4**	**64.8**	**72.1**

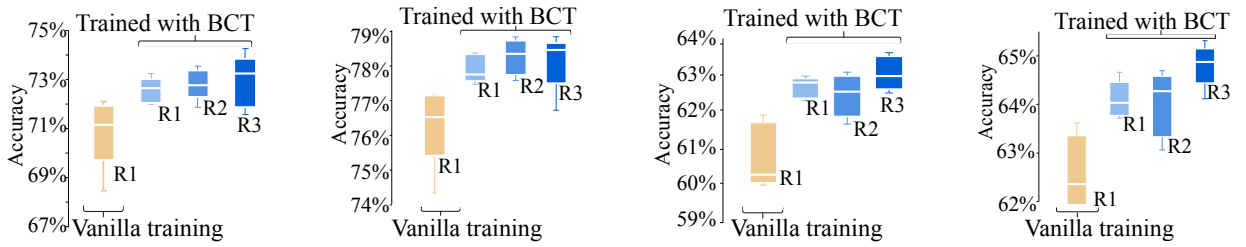
Table 3: Extending Tab. 5 of the main paper. Evaluating CMP-NAS on the Deepfashion2 fashion retrieval benchmark using additional metrics: top-1 and top-20 accuracy. We observe that the models discovered with CMP-NAS comprehensively outperform the baselines on both, homogeneous and heterogeneous accuracies.

Gallery model	Query Prune method	Prune Amt.	Train Scratch	Fine-tune	BCT	KD
ResNet-101	-	0%	87.9	-	-	-
ResNet-101	Magnitude [4]	30%	0.0	87.9	**88.5**	0.0
ResNet-101	Magnitude [4]	50%	0.0	87.3	**88.2**	0.0
ResNet-101	Magnitude [4]	70%	0.0	87.2	**87.9**	0.0
ResNet-101	Magnitude [4]	90%	0.0	86.5	**87.2**	0.0
ResNet-101	Channel [3]	30%	0.0	87.6	**88.4**	0.0
ResNet-101	Channel [3]	50%	0.0	87.5	**87.8**	0.0
ResNet-101	Channel [3]	70%	0.0	87.3	**87.9**	0.0
ResNet-101	Channel [3]	90%	0.0	86.3	**87.4**	0.0

Table 4: Extending Tab. 4 of the main paper. Comparing training methods for heterogeneous accuracy achieved on the 1:N face retrieval task. The query model ϕ_q is obtained via pruning filters from the first two layers of each residual block of the gallery model. We compare two different pruning methods [3, 4] at several pruning amounts. Observe that for all pruning methods and amounts, training the query model with BCT loss leads to (1) non-zero heterogeneous accuracy and (2) the highest heterogeneous accuracy.

C. Additional results for weight-level compatibility

Due to space limit, Tab 4 in the main paper compares different training methods for weight-level compatibility using query model achieved by pruning 90% of filters. Tab. 4 extends Tab 4 in the main paper by showing heterogeneous accuracy of query models achieved by pruning the gallery model to different levels.



(a) Homogeneous (Face) (b) Heterogeneous (Face) (c) Homogeneous (Fashion) (d) Heterogeneous (Fashion)

Figure 1: Extending Fig. 6 of the main paper. The figures are generated by averaging the best five architectures discovered by CMP-NAS (under 100 Mflops) when using different training strategies (Vanilla, BCT) and rewards ($\mathcal{R}_1 - \mathcal{R}_3$). In (a),(b) we plot the homogeneous and heterogeneous accuracy for the 1:N face retrieval task using the metric TNIR@FPIR=10^{-1}. In (c),(d) we plot the homogeneous and heterogeneous accuracy for the fashion retrieval task using the metric top-10. Observe that in all cases, BCT training works best among the training strategies while $\mathcal{R}_3$ outperforms all other rewards.

D. Comparing different rewards

Fig. 1 is an extension of Fig. 6 in the paper. We present the homogeneous and heterogeneous accuracy achieved by the best five query models searched using different rewards and training schemes on the face and fashion retrieval tasks. These complementary results further reinforce our conclusions: $\mathcal{R}_3$ generally works better than $\mathcal{R}_1$ and $\mathcal{R}_2$; Training the super-network with BCT outperforms vanilla training by a large margin.

References

[1] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, 2019.

[2] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *ECCV*, 2020.

[3] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *ICCV*, 2017.

[4] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *ICLR*, 2017.

[5] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *ECCV*, 2018.