Efficient Feature Extraction for Person Re-Identification via Distillation

Francisco Salgado
United Technologies Research Center
United Technologies Corporation
Cork, Ireland
salgadfr@utrc.utc.com

Rakesh Mehta
United Technologies Research Center
United Technologies Corporation
Cork, Ireland
rakeshere@gmail.com

Paulo Lobato Correia
Instituto de Telecomunicações
Instituto Superior Técnico
Lisbon, Portugal
plc@lx.it.pt

***Abstract*—Person re-identification has received increasing attention due to the high performance achieved by new methods based on deep learning. With larger networks of cameras being deployed, more surveillance videos need to be parsed, and extracting features for each frame remains a bottleneck. In addition, the feature extraction needs to be robust to images captured in a variety of scenarios. We propose using deep neural network distillation for training a feature extractor with a lower computational cost, while keeping track of its cross-domain ability. In the end, the proposed model is three times faster, without a decrease in accuracy. Results are validated on two popular person re-identification benchmark datasets and compared to a solution using ResNet.**

***Index Terms*—person re-identification, cross-domain, distillation, convolutional neural networks**

I. Introduction

Person re-identification (re-ID) refers to the task of finding a person of interest (query) across images or videos captured by different cameras (gallery). Cameras are usually placed in uncontrolled environments, where the quality of the videos is far from ideal and subjects might be partially occluded.

The increasing demand of public safety and the growing size of camera networks result in more surveillance videos needing to be parsed. Given this demand, current state-of-the-art re-ID architectures are computationally expensive to run, as reported e.g., in [1], [2], using ResNet and Inception as the backbone Convolutional Neural Network (CNN). This highlights the need for smaller and faster re-ID models that are still highly accurate.

In a scenario where it is not possible to further tune the system once deployed, it is also important that the system is able to generalize to unseen environments without having its performance considerably reduced. The current literature does not often consider the accuracy of the proposed algorithms when presented with images captured in other environments. For instance, while the model proposed in [1] achieves state-of-the-art results on the Market-1501 [3] dataset , with 83.40% rank-1 accuracy and 66.88% mAP, when tested on another popular dataset, DukeMTMC-reID [4], its performance drops significantly, achieving only 18.72% rank-1 and 9.20% mAP.

This paper addresses the above problems, and its key contributions are: (1) distillation is applied to develop a faster and more compact model, with a better trade-off between accuracy and speed; (2) distillation is used to improve the generalization power of person re-ID algorithms; (3) the proposed model is tested on a cross-domain scenario. To the best of our knowledge, this is the first work studying the performance of a person re-ID model when tested on a new domain.

The paper is organized as follows: Section II briefly reviews the relevant state-of-the-art publications. The proposed methodology is presented in Section III and Section IV discusses implementation details, provides information about the databases and metrics used for evaluation, and reports the results obtained for each proposed topology. Finally, conclusions are drawn in Section V.

II. Related Work

Person re-ID is a very active area of research [5]–[9]. For many years, systems relying on manual feature selection were the most popular approaches to find matching footage of a person across an array of cameras [5], [10]–[14]. However, in recent years and with the rise in popularity of deep learning, their performance has been surpassed. Convolutional neural networks are now the state-of-the-art approach for re-ID [15]. For a survey on recent approaches see [5], [15].

Two deep-learning approaches are commonly used to address the re-identification problem. In the first one, a CNN is fine-tuned as a classifier to optimize its performance on one of the existing re-ID datasets, typically by minimizing the cross-entropy loss between the predicted labels and the ground-truth labels in the dataset [2], [15], and the activations computed before the fully connected classification layer are the features to be used for matching. The second approach relies on using different cost functions, such as the contrastive or the triplet loss, aiming at directly learning a data representation that brings together feature points corresponding to images of the same identity, while separating those extracted from images of different people [1], [16].

While complex CNN architectures have proven to be accurate for use in re-identification problems, they are cumbersome

This work was supported by Instituto de Telecomunicaes through FCT/MEC under the project UID/EEA/50008/2019 and by the SURVANT project which has received funding from the EU Horizon 2020 research and innovation programme under grant agreement No 720417.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

in practice due to high-end hardware requirements. Popular compression techniques for reducing the computational cost during inference range from pruning individual neuron weights [17], to pruning entire convolutional filters [18]–[20]. A more compact architecture can also be trained to replicate the output of a more accurate model. This is the process adopted in this paper, and it was originally proposed in [21].

III. Methodology

The architecture of the proposed system follows the standard pipeline of re-ID works [5]. The system takes as input images that have already been cropped to the bounding box containing only the person. These are then preprocessed and fed to the feature extractor to compute a feature vector (descriptors).

After the feature extractor computes descriptors for the query, the Euclidean distance is used as similarity metric to find the closest matches to the query image, by computing the distance to the descriptors previously stored in the gallery. To be successful, the model should keep the extracted feature points belonging to the same person close together, while pushing feature points from different people further apart.

Feature extraction is a key component to the success of person re-ID system, with most current approaches using CNN-based feature extractors. The backbone CNN is usually trained on the chosen re-ID dataset using the cross-entropy loss. During training, the softmax activation outputs the ID of the input image. At inference time, the fully-connected layer is often removed and the extracted features correspond to the output of the convolutional layers after the pooling layer.

A. Distillation

In order to develop a fast and computationally simple feature extractor the CNN backbone can be replaced by a lighter architecture with faster inference time. However, the resulting performance can be affected, as a simpler network might lack the ability to learn a powerful enough data representation.

Alternatively, distillation [21] allows to adopt a simpler architecture, denoted student network, and have its training guided by a more powerful pre-trained network, denoted teacher network. The student network can thus be trained using the available ground truth labels together with the teacher network predictions, which are used as soft labels. The adopted distillation-based architecture is represented in Figure 1.

The softmax activation function after the last layer of a CNN converts each logit z_i to a class probability q_i by taking into account the value of the other logits:

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \quad (1)$$

where T is denoted as temperature, taking the default value of 1. Increasing the value of the temperature parameter results in a probability distribution with higher entropy (the predictions are softened), as illustrated in Figure 2.

The relative probabilities of the classes, predicted by the teacher network, provide a considerable amount of information

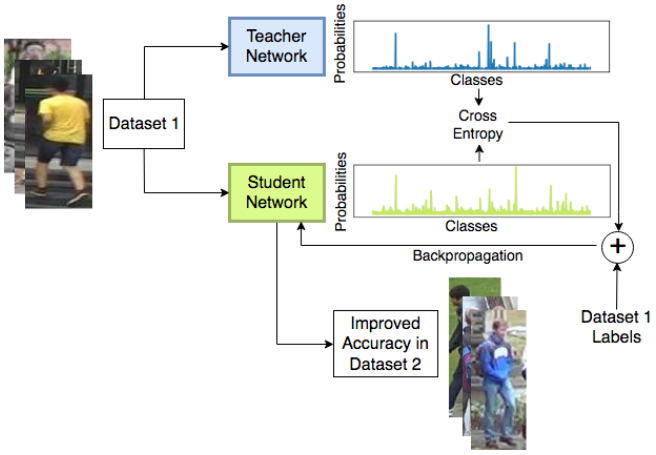
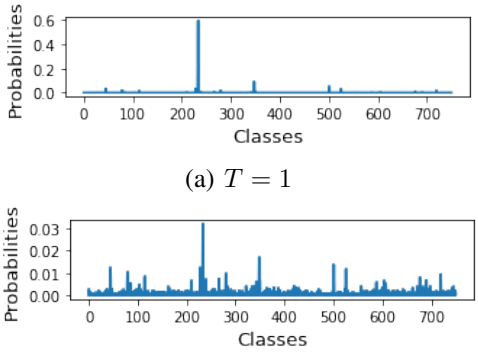


Fig. 1. Overview of proposed architecture.



(a) $T = 1$

(b) $T = 3$

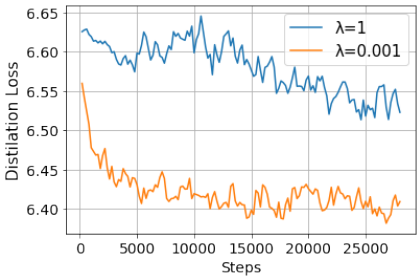
Fig. 2. Influence of the temperature parameter T on the output of a softmax activation. In this example, each class corresponds to one of the 751 IDs present on the Market-1501 dataset.

about what classes are seen as similar, and how the complex model tends to generalize, allowing the simpler student network to learn a model that approximates reasonably well the original one, learnt using a deeper network. Moreover, increasing the entropy of the soft-labels can be seen as as a regularizer, with the additional noise helping to prevent model overfitting and contributing to improve the model accuracy.

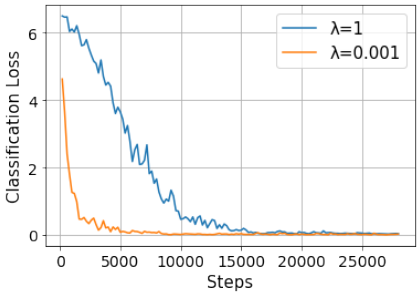
The cost function for training the student network becomes:

$$L_{student} = H(p_{teacher}(T = T_0), p_{student}(T = T_0)) + \lambda H(ground\ truth, p_{student}(T = 1)) \quad (2)$$

The first term, the distillation loss, is the cross-entropy with the soft targets, being computed using the same (high) temperature T_0 in the softmax of the student model that was used by the teacher model to generate the soft targets. The second term is the cross entropy with the ground truth labels, being computed with temperature of 1. λ defines the contribution of the cross-entropy loss using the ground truth labels to the total loss function and should be selected so that both terms converge together, as illustrated in Figure 3.



(a) Distillation loss



(b) Classification loss

Fig. 3. Selection of the λ parameter in the student network loss function. When λ is too high ($\lambda = 1$), the classification loss leads the training and prevents the sum of both losses to converge to a lower value.

During training, only the student network weights are updated to minimize the loss function, as the goal is to have the student network mimic the teacher network.

IV. Experiments

This section presents the results obtained using the proposed methodology.

A. Datasets

Two datasets are selected for training and evaluation. Person re-identification datasets come with a training split, and query and gallery splits used for validation, the later containing images of people not present in the training split. For each query image, the goal is to find all the images in the gallery set corresponding to the same person.

The Market-1501 (Market) [3] and DukeMTMC-reID (Duke) [4] datasets were chosen due to their popularity. Market contains 32668 images of 1501 different people, while Duke has 36441 images of 1812 different people, corresponding to an average of 17 and 24 training images of each person, respectively. The first dataset was captured across 6 cameras and has 2798 distractor images in the gallery set with no corresponding images in the query set, while the second was captured by 8 cameras and has 3463 distractors, providing a more challenging setup.

Only one dataset is used for training in each experiment. The model performance is evaluated in the dataset where it was trained, but also a generalization, or cross-domain, evaluation is done, by assessing the model's performance on the other dataset. This simulates the condition where the model is trained with a given set of data and, when deployed, it will be used with a different setup and population of users.

B. Evaluation Metrics

Two metrics commonly used to evaluate the performance of re-ID algorithms are adopted: the rank-1 accuracy and the Mean Average Precision (mAP), i.e., the average of the maximum precision at different recall values [5]. A query is given rank-1 when the first subject returned by the system is the correct one. If multiple ground truth images have the same label in the gallery, rank-1 may not be discriminative enough to distinguish two systems and the mAP should also be used.

C. Implementation details

ResNet-50 [22] and MobileNet [23] have been selected for teacher and student networks, respectively. ResNet-50 has been proven to learn suitable representations for re-ID [15], providing a good benchmark for comparison of results. MobileNet lacks skip connections and replaces standard convolutions with depth-wise separable convolutions, resulting in faster inference times, having shown good results for distillation. Additionally, MobileNet's depth multiplier, α, allows to further reduce complexity, multiplying the number of convolutional filters on each layer by α. For this study, $\alpha = 1.0$ and $\alpha = 0.25$ are considered, corresponding to the default and the smallest MobileNet models available pre-trained.

For baseline results, networks were trained for 30 epochs on both datasets. The optimizer used was the Stochastic Gradient Descent, with a batch size of 16, learning rate of 0.001, momentum of 0.9 and learning rate decay of 0.1 every 20.000 batches. For the distillation experiments, the learning rate was set to 0.02 and models were trained for 25 epochs.

D. Data Preprocessing

Before feeding the training images to the feature extractor, a preprocessing step normalizes the available images according to the CNN architecture [22], [23].

Images are resized so that their largest dimension is 256 pixels and the original aspect ratio is kept.

Since the number of images of each person available in the re-ID datasets is limited, data augmentation is used to provide more data for training the network, considering:

- Random horizontal flips (baseline);
- Random desaturation of images, up to 10%;
- Random rotations, up to 5 degrees;
- Random cropping.

Table I studies the impact of each transformation. Only Market-1501 was chosen as the training dataset, being representative enough since both datasets having similar dimensions and number of images per class, which should correspond to similar intra-class variability.

We can observe that random rotations alone lead to the highest rank-1 and mAP on the second dataset. Randomly desaturating the input image results in a decrease of accuracy on both datasets.

In light of these results, we apply random horizontal flips and rotations for data augmentation for the rest of this paper.

TABLE I
DATA AUGMENTATION: RESNET TRAINED ON MARKET-1501.

	Market		Duke	
	Rank-1	**mAP**	**Rank-1**	**mAP**
H. Flip (Baseline)	67.93	41.61	26.80	12.70
H. Flip + 0.9 Sat	56.18	27.94	21.54	8.81
H. Flip + Rot.	**69.51**	43.20	**27.15**	**13.18**
H. Flip + Crop	68.71	**43.97**	25.31	12.38
H. Flip + Crop + Rot.	68.68	43.30	27.10	12.97

E. Baseline Results

Both networks are first initialized with their ImageNet pretrained weights [24] and fine-tuned on both datasets without distillation to set a benchmark. The obtained performance results are presented in Table II.

TABLE II
FINETUNING ON MARKET-1501 AND DUKEMTMC-REID

		Market		Duke	
Trained on		**Rank-1**	**mAP**	**Rank-1**	**mAP**
Market	ResNet	69.51	43.20	**27.15**	**13.18**
	MobileNet 1.0	**75.80**	**51.60**	15.17	6.65
	MobileNet 0.25	74.29	48.80	17.68	7.74
Duke	ResNet	37.0	13.99	**64.45**	**44.12**
	MobileNet 1.0	**37.65**	**14.0**	62.66	40.77
	MobileNet 0.25	35.48	12.85	62.16	39.05

As expected, both models achieve higher rank-1 accuracy and mAP on the training dataset than on the second dataset. When trained on Market-1501, MobileNet achieves slightly higher performance when compared to the ResNet model, despite its simpler architecture. This is due to models being evaluated in a retrieval scenario, which is not being directly optimized when training as a classifier. Nonetheless, its performance is significantly lower on the DukeMTMC-reID dataset. On the other hand, both models achieve very similar performance on both datasets when trained on the DukeMTMC-reID dataset, as this is a more challenging dataset.

While the performance drop is expected when testing a system on a different dataset than the one it was trained on, a good feature extractor should not see its performance dropping drastically. In other words, it should generalize well to unseen environments.

Due to the skip connections and higher number of layers, ResNet is more robust to changes in the data. In contrast, MobileNet is learning a simpler representation of the training data, resulting in worse cross-domain performance.

F. Distillation

For this experiment, the MobileNet models are guided during training by ResNet, according to section III. Firstly, the two parameters T and λ used in the distillation loss function need to be determined.

MobileNet with a $\alpha = 1.0$ is trained as student network and it is initialized with its ImageNet pretrained weights. The trained ResNet model from the previous experiment is used as the teacher network. Data augmentation is equally applied to the input images of both networks. λ is first fixed to 0.001 and distillation is run with values of T set to 1, 3, 5, 10 and 15. The obtained results are presented in Table III.

TABLE III
INFLUENCE OF DIFFERENT TEMPERATURE, T, VALUES.

	Market		Duke	
	Rank-1	**mAP**	**Rank-1**	**mAP**
$T = 1$	75.06	49.39	15.80	6.60
$T = 3$	**80.29**	**56.67**	27.47	14.60
$T = 5$	77.55	54.16	**29.76**	**15.56**
$T = 10$	76.69	51.76	28.41	14.96
$T = 15$	75.95	51.78	28.71	15.20

The performance peaks at lower temperatures. As the goal is to produce a student network with the same or higher generalization ability as the teacher network, $T = 5$ is picked, due to its higher cross-domain performance.

With T fixed to 5, the λ values of 1, 0.1, 0.01, 0.001 and 0.0001 are tested. Results are presented in Table IV.

TABLE IV
INFLUENCE OF DIFFERENT λ VALUES.

	Market		Duke	
	Rank-1	**mAP**	**Rank-1**	**mAP**
$\lambda = 1$	61.73	33.22	15.66	6.40
$\lambda = 0.1$	77.05	52.28	22.58	10.46
$\lambda = 0.01$	77.29	53.79	28.14	15.41
$\lambda = 0.001$	**77.55**	**54.16**	**29.76**	**15.56**
$\lambda = 0.0001$	77.29	53.37	28.05	14.80

With these two experiments, the combination of the distillation parameters that ensures the highest cross-domain accuracy is determined to be $T = 5$ and $\lambda = 0.001$.

MobileNet is now trained on the two datasets through distillation. The obtained results are reported in Table V.

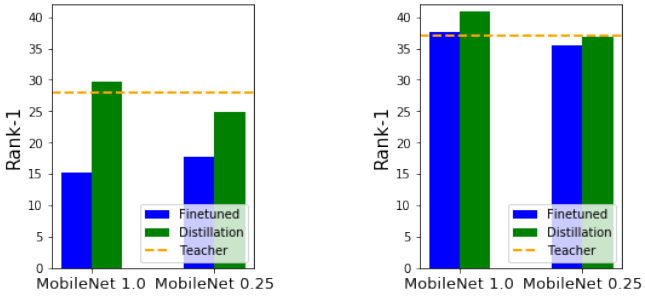
TABLE V
DISTILLATION ON MARKET-1501 AND DUKEMTMC-REID.

		Market		Duke	
Trained on		**Rank-1**	**mAP**	**Rank-1**	**mAP**
Market	Teacher (ResNet)	69.51	43.20	27.15	13.18
	MobileNet 1.0 distilled	**77.55**	**54.16**	**29.76**	**15.56**
	MobileNet 0.25 distilled	75.12	50.86	24.92	12.51
Duke	Teacher (ResNet)	37.0	13.99	64.45	44.12
	MobileNet 1.0 distilled	**40.97**	**16.86**	**70.51**	**50.52**
	MobileNet 0.25 distilled	36.88	14.08	65.66	45.40

It can be observed that the student network with $\alpha = 1.0$ outperforms the teacher network on both datasets. Despite having similar performance to ResNet when fine-tuned on the DukeMTMC-reID, MobileNet achieves an improvement in the two metrics on both datasets when trained with distillation.

The improvement in cross-domain accuracy resulting from distillation is plotted in Figure 4.

Distillation resulted in MobileNet models that match or even outperform the teacher network, at a much lower computational cost: while ResNet needs to perform 4.67×10^3 GFLOPs to compute the descriptors for a single image, MobileNet with



a) Trained on Market b) Trained on Duke

Fig. 4. Cross-domain performance of MobileNet when trained without (blue) and with distillation (green) from a more powerful model. Distillation resulted in a higher improvement when trained on the smaller dataset Market-1501.

$\alpha = 1.0$ and $\alpha = 0.25$ only perform 762.9 GLOPs and 55.83 GFLOPs, respectively.

Additionally, feature extraction with MobileNet models ($\sim$ 0.00263s) is 3 times faster than with ResNet ($\sim$ 0.00658s), when computed with batches containing a single image on a Nvidia GTX1070 card.

The same trend was observed with smaller datasets but results were not as relevant and due to space limitations were not included.

V. Conclusions

This paper proposes a method to train a faster person re-identification model via distillation. Different data augmentation techniques were studied to counter the small dimensions of the datasets. In the end, the model achieves the same performance as the more complex teacher model, being 3 times faster during inference.

The key outcomes are: 1) distillation allows a simpler architecture to learn a representation on re-identification datasets as powerful as the one learnt by a more complex architecture, even outperforming; 2) when presented with data captured in unseen operation scenarios, its accuracy does not drop as drastically, improving rank-1 up to 15%.

While this study was focused on the image-based re-ID problem, the resulting model can still be applied to video-based re-ID, e.g. using a recurrent network [7].

For future work, semi-supervised learning should be studied to make use of unlabelled surveillance footage. Current approaches rely on GANs, however improvements are still marginal [25].

References

[1] Jon Almazán, Bojana Gajic, Naila Murray, and Diane Larlus, "Re-id done right: towards good practices for person re-identification," *CoRR*, vol. abs/1801.05339, 2018.

[2] Yanbei Chen, Xiatian Zhu, Shaogang Gong, et al., "Person re-identification by deep learning multi-scale representations," 2018.

[3] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference on*, 2015.

[4] Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J. Radke, "Dukemtmc4reid: A large-scale multi-camera person re-identification dataset," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[5] Liang Zheng, Yi Yang, and Alexander G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[6] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.

[7] Jean-Baptiste Boin, Andre F. de Araújo, and Bernd Girod, "Recurrent neural networks for person re-identification revisited," *CoRR*, vol. abs/1804.03281, 2018.

[8] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue, "Multi-scale deep learning architectures for person re-identification," *CoRR*, vol. abs/1709.05165, 2017.

[9] Rui Zhao, Wanli Oyang, and Xiaogang Wang, "Person re-identification by saliency learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 356–370, Feb. 2017.

[10] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision – ECCV 2008*, David Forsyth, Philip Torr, and Andrew Zisserman, Eds., Berlin, Heidelberg, 2008, pp. 262–275, Springer Berlin Heidelberg.

[11] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier, "Person re-identification using kernel-based metric learning methods," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 1–16, Springer International Publishing.

[12] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3586–3593.

[13] F. M. Khan and F. Brmond, "Multi-shot person re-identification using part appearance mixture," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp. 605–614.

[14] S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2197–2206.

[15] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia I. Camps, and Richard J. Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," *CoRR*, vol. abs/1605.09653, 2016.

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017.

[17] Yann Le Cun, John S. Denker, and Sara A. Solla, "Advances in neural information processing systems 2," chapter Optimal Brain Damage, pp. 598–605. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[18] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz, "Pruning convolutional neural networks for resource efficient transfer learning," *CoRR*, vol. abs/1611.06440, 2016.

[19] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf, "Pruning filters for efficient convnets," *CoRR*, vol. abs/1608.08710, 2016.

[20] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung, "Structured pruning of deep convolutional neural networks," *CoRR*, vol. abs/1512.08571, 2015.

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[23] 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," *CoRR*, vol. abs/1704.04861, 2017.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[25] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," *CoRR*, vol. abs/1701.07717, 2017.