

A Brief Review for Compression and Transfer Learning Techniques in DeepFake Detection

Andreas Karathanasis, John Violas, Ioannis Kompatsiaris, Symeon Papadopoulos

Information Technologies Institute, Centre for Research & Technology Hellas, Thessaloniki, 57001, Greece

Email: andrew.karathanasis@iti.gr, violas@iti.com, ikom@iti.gr, papadop@iti.gr

Abstract—Training and deploying deepfake detection models on edge devices offers the advantage of maintaining data privacy and confidentiality by processing it close to its source. However, this approach is constrained by the limited computational and memory resources available at the edge. To address this challenge, we explore compression techniques to reduce computational demands and inference time, alongside transfer learning methods to minimize training overhead. Using the Synthbuster, RAISE, and ForenSynths datasets, we evaluate the effectiveness of pruning, knowledge distillation (KD), quantization, fine-tuning, and adapter-based techniques. Our experimental results demonstrate that both compression and transfer learning can be effectively achieved, even with a high compression level of 90%, remaining at the same performance level when the training and validation data originate from the same DeepFake model. However, when the testing dataset is generated by DeepFake models not present in the training set, a domain generalization issue becomes evident.

I. INTRODUCTION

Deepfake detection often involves computationally intensive deep learning models that process high-dimensional images. Without compression, these models can be prohibitively large and slow, making real-time detection impractical, especially in edge computing environments [1]. By reducing model size while preserving accuracy, compression techniques allow for efficient inference, lower latency, and reduced energy consumption [2], making advanced deepfake detection accessible to a broader range of devices and applications.

Transfer learning can play a pivotal role in deepfake detection by leveraging pre-trained models developed for related tasks, such as image recognition or facial analysis, to save time and computational resources [3]. Deepfake datasets are often diverse and complex, requiring significant effort to train a model from scratch. Transfer learning allows models to reuse learned features from large, general purpose datasets and adapt them to the specific task of deepfake detection, even with limited labeled data [4]. This approach enhances detection accuracy, accelerates development, and helps tackle the challenge of detecting evolving deepfake techniques.

Although numerous studies explore deepfake detection, model compression techniques, and transfer learning independently, there is a lack of research on combining these approaches in deepfake detection models. To address this gap, we investigate the use of techniques such as pruning, KD, quantization, fine-tuning, and adapters in deepfake detection. Our work includes an analysis of their applications and an experimental evaluation across three datasets.

The structure of this paper is organized as follows: Section

II offers a concise overview of the techniques employed. Section III details the methodologies explored for achieving compression and transfer learning in DeepFake detection. Section IV presents the experimental results along with a discussion of their implications. Finally, Section V outlines our conclusions and directions for future research.

II. RELATED WORK & BACKGROUND

Deepfake detection refers to the identification and classification of manipulated digital media, such as images and videos, that are generated using deep learning techniques like Generative Adversarial Networks and Diffusion Models [5]. These manipulations often include face swaps, facial reenactments, or synthetic generation of human likeness, posing threats to privacy, authenticity, and public trust [6]. The survey [7] provides a comprehensive analysis of both the creation and detection of deepfakes, emphasizing the challenges posed by their increasing sophistication. Furthermore it explores potential legal implications that could facilitate the reduction of DeepFake material in public media platforms. While deep learning-based approaches can tackle the challenge of DeepFake detection, they typically demand substantial computational resources for both training and testing, making them unsuitable for edge computing applications. At the same time, as highlighted by [8], although the creation of deepfakes has become increasingly accessible due to open-source tools and pre-trained models, producing high-quality results still relies heavily on large datasets and, more importantly, significant computational resources. To prevent a potential “war” for resources, the development of resource-efficient detection techniques is imperative.

Techniques such as pruning, KD, and quantization have been proposed for model compression across various tasks and modalities. Pruning involves reducing the complexity of a neural network by removing less significant parameters, such as weights or neurons, while retaining its predictive power [9]. KD transfers knowledge from a larger, more complex model (the teacher) to a smaller model (the student) [1]. The student model learns to mimic the teacher by approximating its predictions or output probabilities [10]. Quantization simplifies models by reducing the precision of their numerical representations, such as converting 32-bit floating-point values to 8-bit integers [9]. This significantly reduces memory and computational requirements with minimal accuracy loss.

Fine-tuning is a key technique that can be applied after compressing deep learning models to restore any lost performance and optimize their accuracy [11]. Additionally, adapters,

which are lightweight, trainable modules inserted into pre-trained models, offer an efficient method for transfer learning by enabling the model to adapt to new tasks or datasets without extensive re-training [12]. However, determining the most effective approach between these techniques for deepfake detection remains an open question. This research addresses the gap both theoretically in Section 3 and experimentally in Section 4, providing an experimental comparison.

III. APPROACHES FOR COMPRESSION & TRANSFER LEARNING

We explore a baseline method, three compression and six transfer learning approaches, which are briefly presented below. For a comprehensive overview of these methods, we direct readers to [13]. The baseline and the compression are applied in one dataset. Transfer learning utilizes two datasets: a) the dataset used to train a model on a general or related task, which is called the *source* dataset, and b) the one used to fine-tune the model for a specific task, i.e. the *target* dataset.

- (1) **Baseline (BL):** The baseline approach involves straightforward training of the original, large deepfake detection model.
- (2) **Compression with Pruning and Fine-Tuning (CPF):** The first compression approach involves pruning the originally trained BL model, followed by a fine-tuning process. In this approach, the pruned model undergoes training across all layers for a few epochs to reduce the performance degradation caused by pruning.
- (3) **Compression with Knowledge Distillation without Fine-Tuning (CKD):** In this approach, the BL model serves as the teacher model, while a smaller model is used as the student model. Notably, no fine-tuning is performed in this process.
- (4) **Compression with Quantization (CQ):** In this approach, we quantized the weights of the linear layers of the BL model, reducing their precision from float32 (the PyTorch default) to int8, the smallest supported precision.
- (5) **Transfer Learning with Fine-Tuning (TL+FT):** In this approach, the BL model which is trained on the source dataset is fine-tuned on selected layers using the target dataset
- (6) **Transfer Learning with Fine Tuning and Quantization (TL+FT+Q):** The transferred large model, undergoes the CQ method and it's subsequently quantized in its linear layers to achieve compression. Fine-tuning facilitates adaptation to the target data, while quantization addresses model compression.
- (7) **Transfer Learning with Pruning and Fine-Tuning (TL+P+FT):** The pruned models, initially trained and fine-tuned on the source data as outlined in CPF method, are afterwards, further fine-tuned on the target data, in some selected convolutional layers. The pruning process addresses model compression, while the additional fine-tuning on the target data enables transfer learning.
- (8) **Transfer Learning with Knowledge Distillation (TL+KD):** In this approach, the BL model serves as the teacher model, aiming to transfer general, low-complexity knowledge learned from the source data

that is also relevant to the target data. The student model is a smaller, completely untrained model, and the target data is used during the distillation process.

Transfer Learning with Pruning and Adapter (TL+P+A): In this approach, the pruned and transferred, TL+FT+Q, models, undergo architectural modifications with the addition of an adapter layer. This leverages the concept of depthwise separable convolution to minimize its size and is exclusively trained on the target data, while all other layers remain frozen.

Transfer Learning with Knowledge Distillation and Adapter (TL+KD+A): This approach builds on the distilled models used for transfer learning, as outlined in TL+P+FT, by incorporating adapter layers into their architecture. While these adapters are identical to those described in method (8), their placement differs: instead of being positioned in the middle of the layers, as in the pruned approach, they are added after the last convolutional layer (specific to each architecture). The adapter is then trained independently.

These approaches are different in terms of performance, training time, and inference time. Taking into consideration that synthetic images in a dataset can be generated from multiple types of generative models (i.e., GANs, VAEs, diffusion models), serve different purposes (i.e., entertainment, realism enhancement, data augmentation), and have different levels of manipulations (i.e., splicing, object removal, face swapping), we experimentally compared them with three datasets as it will be described in Section IV.

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup

We used three datasets for our experiments. The first comprises DeepFake images from the SynthBuster test set [14]¹ and authentic images from the RAISE dataset². The SynthBuster contains 9,000 AI-generated images (1,000 each) from nine DeepFake models: DALL-E 2, DALL-E 3, Adobe Firefly, Midjourney v5, and five versions of Stable Diffusion. RAISE contains 8,153 authentic images. Images are split into 60% for training and 40% for testing, with both sets containing a mix of authentic images and those generated by all nine DeepFake models.

The second dataset is ForenSynths [6]³, which includes authentic images and DeepFake images generated from 13 different CNN-based GAN models. The training set contains 720,119 images, the validation set 8,000, and the test set 90,310. For our experiments, we focused exclusively on human face images, as deepfakes are often used to impersonate prominent individuals, leading to potential harm, confusion, and security threats [15], [16], [17]. It is important to emphasize that the DeepFake images in the training set are exclusively generated by the ProGAN model, while the test set includes images not only from ProGAN but also from other models

¹<https://zenodo.org/records/10066460>

²<http://loki.disi.unitn.it/RAISE/download.html>

³<https://github.com/PeterWang512/CNNDetection>

like StarGAN and WhichFaceIsReal, which are absent from the training set.

To evaluate the transfer learning approach, we utilized a simple "dogs vs. cats" dataset⁴ consisting of 25,000 images to train the original models. These were later employed to accelerate the training process using techniques such as Fine-Tuning, KD, and adapter layers. All images from the datasets were resized to 224×224 pixels. Experiments were conducted in Python 3, utilizing libraries such as PyTorch, PIL, and scikit-learn. They were executed on Kaggle notebooks, using two NVIDIA T4 GPUs, each with 16 GB of memory.

The experiments involving the aforementioned datasets utilized a primary VGG-based model consisting of approximately 4.5 million parameters. The code for the primary BL model, the compressed and transferred models are available in [18].

B. Outcomes

Figures 1 and 2 present the results of the experiments using compression approaches, while Figure 3 illustrates the experiments with transfer learning approaches. The acronyms in the legend above the box plots correspond to the methods outlined in Section III. Evaluations are presented across compression rates ranging from 60% to 90% relative to the original model size in terms of parameters. We present the Accuracy, F1-Score, compression time and transfer learning time. We do not provide compression time for the quantization, as converting from float to int is almost negligible. The red lines in the Baseline methods indicate the training time of the original models. Figure 1 presents the outcomes on the Synthbuster dataset. The findings reveal that the pruned models achieved performance comparable to the original larger model. However, KD demonstrated slightly superior results, with some student models even outperforming the teacher model (baseline approach).

Figures 2 shows the results obtained on the ForenSynths dataset. Since the test images were generated by different synthetic image generators, we evaluated the images from each generator separately. When testing with images generated by ProGAN, which also produced the training dataset, all methods achieved very high accuracy, even at high compression levels. However, performance declines with images from StarGAN, and the degradation becomes more pronounced with WhichFaceIsReal and DeepFake generators. This occurs because StarGAN, like ProGAN, is also a GAN-based model with several similarities; however, the synthetic generation processes used in WhichFaceIsReal and DeepFake differ significantly. Nonetheless, it is worth noting that the compression methods, including pruning, KD, and quantization, deliver accuracy comparable to the uncompressed baseline models across all cases.

Similarly, Figure 3 present the evaluation results of transfer learning using the ForenSynths dataset. Testing on the set generated by ProGAN shows strong performance, whereas the other test sets exhibit significantly lower performance, highlighting a domain generalization challenge. When comparing methods, KD generally outperforms pruning and incorporating an adapter typically enhances performance in most cases.

C. Discussion

In Compression approaches we observed that even if the compression level is very high, reaching up to 90% compared to the initial size, in all cases, the accuracy of compressed models remain at the same level with the baselines. Furthermore, we observe that when the testing and training datasets come from the same deepfake generator, in almost all the cases, KD surpasses the pruning approach, while when we use a different deepfake generator the pruning approach leads to better results in all cases except from the very high compression level of 90%.

Quantization effectively maintained performance but did not achieve the same level of compression as KD and pruning. Additionally, quantized DeepFake models cannot benefit from accelerators like GPUs, limiting their practical advantage. In some cases, depending on the hardware, quantization might not have any benefit at all. However, for DeepFake detection on Android devices for example, where inference is typically done on optimized mobile CPUs, quantization remains valuable.

It is also worth noting that we conducted experiments with low-rank factorization, supported as a method by [19], which reduced model size to some extent but often caused significant performance degradation, limiting its utility. That is also the reason that the method's results are not included in this paper. Its effectiveness is context-dependent, and the compression achieved usually requires combining it with other methods to meet the needs of DeepFake edge computing.

Regarding transfer learning, we found that the adapter's position significantly affects performance. In deeper models, placing the adapter near the first or middle layers maximized effectiveness. In smaller models, such as student models in KD, placing the adapter at the end of the convolutional layers yielded the best results. Notably, adapters only improved feature extraction; those using linear layers showed no performance gains in some preliminary experiments we conducted.

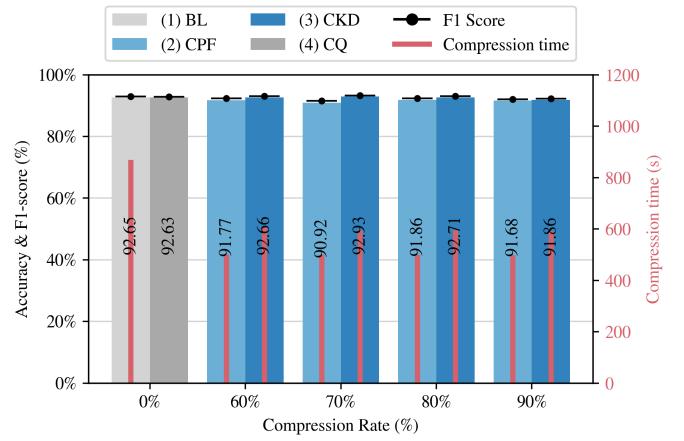


Fig. 1: Evaluation of Compression Techniques Using a Dataset Combining Deepfake Images from Synthbuster and Authentic Images from RAISE

⁴<https://www.kaggle.com/datasets/salader/dogs-vs-cats>

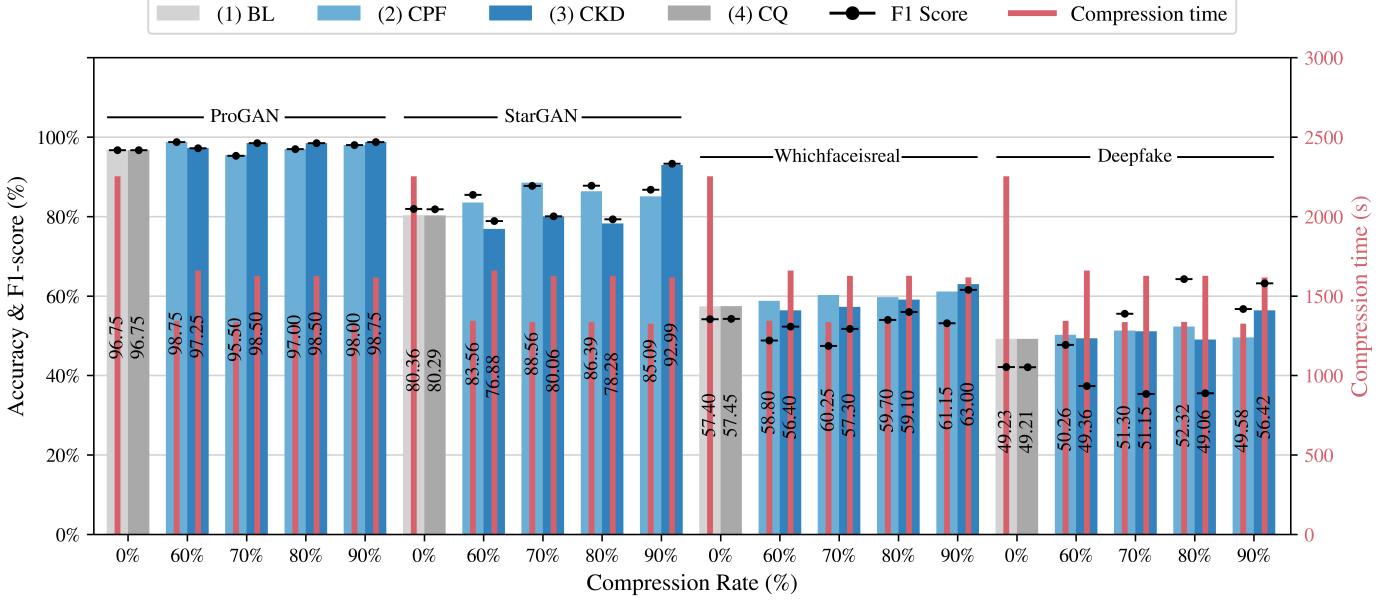


Fig. 2: Evaluation of Compression Techniques Using ForenSynths dataset

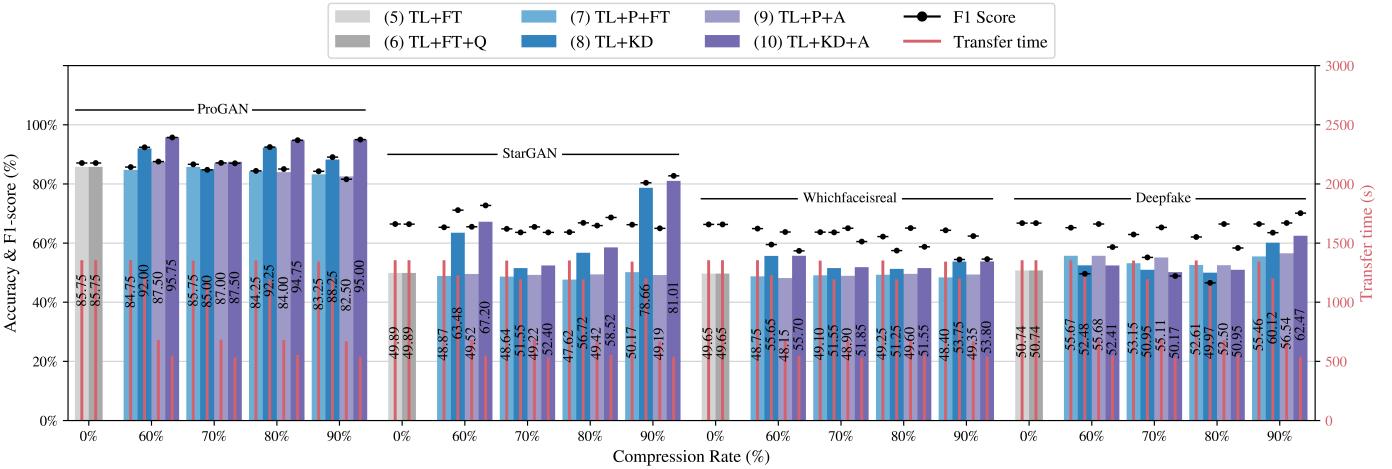


Fig. 3: Evaluation of Transfer Learning Techniques Using ForenSynths dataset

V. CONCLUSION & FUTURE WORK

In this paper we observe that even with a high-level compression of nearly 90% of the original model, performance remains consistent when the testing dataset originates from the same DeepFake generator as the training/compression dataset. However, there is a significant drop in performance when the evaluation dataset is generated by a different DeepFake generator. This highlights a domain generalization challenge combined with compression in DeepFake detection, which we aim to address in our future research.

ACKNOWLEDGMENT

This work was funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101120237 (ELIAS).

REFERENCES

- [1] J. Violos, S. Papadopoulos, and I. Kompatsiaris, "Towards Optimal Trade-Offs in Knowledge Distillation for CNNs and Vision Transformers at the Edge," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, Aug. 2024, pp. 1896–1900, iSSN: 2076-1465. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10715301>
- [2] M. Kinnas, J. Violos, I. Kompatsiaris, and S. Papadopoulos, "Reducing inference energy consumption using dual complementary CNNs," *Future Generation Computer Systems*, vol. 165, p. 107606, Apr. 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167739X24005703>
- [3] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, conference Name: IEEE Transactions on Knowledge and Data Engineering. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/5288526>

- [4] G. Ayana, K. Dese, A. M. Abagaro, K. C. Jeong, S.-D. Yoon, and S.-w. Choe, "Multistage transfer learning for medical images," *Artificial Intelligence Review*, vol. 57, no. 9, p. 232, Aug. 2024. [Online]. Available: <https://doi.org/10.1007/s10462-024-10855-7>
- [5] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25 494–25 513, 2022, conference Name: IEEE Access. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9721302>
- [6] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.
- [7] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.
- [8] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.
- [9] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, Oct. 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231221010894>
- [10] S. Tsanakas, A. Hameed, J. Violas, and A. Leivadeas, "A light-weight edge-enabled knowledge distillation technique for next location prediction of multitude transportation means," *Future Generation Computer Systems*, vol. 154, pp. 45–58, May 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167739X23004867>
- [11] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "SpotTune: Transfer Learning Through Adaptive Fine-Tuning," 2019, pp. 4805–4814. [Online]. Available: <https://arxiv.org/abs/1811.08737>
- [12] Y.-L. Sung, J. Cho, and M. Bansal, "VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks," 2022, pp. 5227–5237. [Online]. Available: <https://arxiv.org/abs/2112.06825>
- [13] A. Karathanasis, J. Violas, and I. Kompatiari, "A Comparative Analysis of Compression and Transfer Learning Techniques in DeepFake Detection Models," *Mathematics*, vol. 13, no. 5, p. 887, Jan. 2025, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/2227-7390/13/5/887>
- [14] Q. Bammey, "Synthbuster: Towards detection of diffusion model generated images," *IEEE Open Journal of Signal Processing*, 2023.
- [15] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.
- [16] M. Westerlund, "The emergence of deepfake technology: A review. technology innovation management review, 9 (11), 39–52," 2019.
- [17] N. Misirlis and H. B. Munawar, "From deepfake to deep useful: risks and opportunities through a systematic literature review," *arXiv preprint arXiv:2311.15809*, 2023.
- [18] A. Karathanasis, "andreaskarathanasis/Compression-Transfer-of-DeepFake-Models," Jun. 2024, original-date: 2024-06-29T05:44:03Z. [Online]. Available: <https://github.com/andreaskarathanasis/Compression-Transfer-of-DeepFake-Models>
- [19] M. Masana, J. Van De Weijer, L. Herranz, A. D. Bagdanov, and J. M. Alvarez, "Domain-adaptive deep network compression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4289–4297.