INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Phonetically-Aware Coupled Network For Short Duration Text-independent Speaker Verification

Siqi Zheng, Yun Lei, Hongbin Suo

Speech Lab, Alibaba DAMO Academy
{zsq174630,yun.lei, gaia.shb}@alibaba-inc.com

Abstract

In this paper we propose an end-to-end phonetically-aware coupled network for short duration speaker verification tasks. Phonetic information is shown to be beneficial for identifying short utterances. A coupled network structure is proposed to exploit phonetic information. The coupled convolutional layers allow the network to provide frame-level supervision based on phonetic representations of the corresponding frames. The end-to-end training scheme using triplet loss function provides direct comparison of speech contents between two utterances and hence enabling phonetic-based normalization. Our systems are compared against the current mainstream speaker verification systems on both NIST SRE and VoxCeleb evaluation datasets. Relative reductions of up to 34% in equal error rate are reported.
Index Terms: text-independent speaker verification, phonetic information, end-to-end framework

1. Introduction

Recent advancements in speaker verification technology resulted in a large number of successful applications. The majority of early applications have been long duration text-independent tasks, such as switchboard telephone calls[1]. As the popularity of products such as home-based virtual assistants and smart phones equipped with speech technologies increases, short duration text-dependent tasks have received growing attention [2][3]. However, short duration text-independent speaker verification still remains one of the most challenging tasks in the community and its large-scale application has yet to be seen.

One of the reasons that makes text-independent task intractable on short utterances is the significant amount of variations in terms of the content of the speech. When dealing with long utterances, we are able to collect enough number of frames pronouncing different phonemes. This allows the system to estimate a "good enough" speaker-related embedding even by simply taking the weighted average over the frames. In short utterances, however, there are not enough frames to cover a wide range of different phonemes. This introduces large variability to the problem. Ignoring the phonetic content can cause a large bias to the results. For example, different speakers saying similar contents often result in higher scores than utterances from the same speakers but with completely different phonetic contents. It is no longer appropriate to find a hyperplane in the speaker embedding space to discriminate between speakers without taking the phonetic information into account. In most applications to date, speakers are required to restrain the contents to certain phrases, hence reducing the problem to become text-dependent.

This paper aims to tackle this problem by extracting speaker-related embeddings that have been normalized by the similarities, or dissimilarities, between the frame-level phonetic contents of the two utterances. Before digging into the details, we first go over previous studies that lay the foundations and provide motivations for this work.

2. Related Works

The GMM-ivector framework with Probabilistic Linear Discriminant Analysis (PLDA) backends [4][5] have been standing as one of the mainstream approaches since proposed and is still demonstrating its effectiveness in long duration text-independent tasks. With enough input frames, the i-vector framework is able to provide good approximations of sufficient statistics, regardless of the speech content. However, for short utterances, the impact of phonetic content can no longer be ignored and it introduces large variability to the sufficient statistics estimated by the i-vector extractor.

The x-vector framework [6][7] has proven to be effective when compared to an i-vector system, especially on short utterances. Consisting of 5 layers of utterance-level time-delay structure, followed by a pooling layer and two layers of segment-level fully-connected layer, its simple architecture makes it efficient for large-scale deployment. The pooling layer makes the network flexible for random lengths of speech, which makes it scalable on different datasets. In addition, the system is well supplemented by the Linear Discriminant Analysis (LDA) and PLDA scoring back-ends.

In [8] Yun et al. proposed ASR-DNN, a successful implementation utilizing phonetic information to produce frame alignments for the i-vector extraction. The authors have demonstrated that it is beneficial to model Gaussian feature space using ASR senone posteriors, which are largely ignored by the unsupervised GMM-ivector paradigm. In this work the processes of aligning frames and extracting sufficient statistics are effectively decoupled, which allows the system to use different optimal features for each of the processes. Motivated by this work, we are interested in investigating whether phonetic information would also be favorable in the training of end-to-end speaker verification neural network.

Prior to this work, several studies have focused on trying to integrate phonetic information into x-vector-based systems. Approaches such as performing multi-task training to suppress phonetic-related information, and training the x-vector system with concatenation of ASR bottleneck features, are investigated [9][10][11]. However, the relative gains of these approaches are marginal, for the additional cost and complexity of extracting phonetic information. Most of these studies have opted for softmax loss function, or variations of softmax. This paper tries to elaborate and reason why an end-to-end framework with triplet loss function would be a more suitable choice for normalizing phonetic information.

http://dx.doi.org/10.21437/Interspeech.2020-1306

3. Systems

3.1. Phonetically-Aware Coupled Network

This paper presents a simple coupled structure of time-delay neural network. It takes both filter bank and ASR bottleneck features as inputs. The network consists of three stems. The structure of each stem, when viewed alone, is similar to a TDNN network with several layers of 1D convolution on the frame level followed by the pooling layer and fully connected layers. We keep the structure of each stem simple to guarantee that the improvement is not masked by the overly sophisticated network structure and the gain solely comes from contributions of phonetic information and the coupling design.

As shown in Figure 1, the acoustic stem takes the 40-dimensional filter bank features as input. The phonetic stem takes the 100-dimensional ASR bottleneck features as input. The two stems are coupled together through a central coupled stem. On the first layer, the coupled stem takes the concatenation of filter bank and ASR bottleneck features as input. And on each of the subsequent convolution layers, it accepts the output from the previous corresponding layers of acoustic stem and phonetic stem, as well as outputs from its own previous layers.

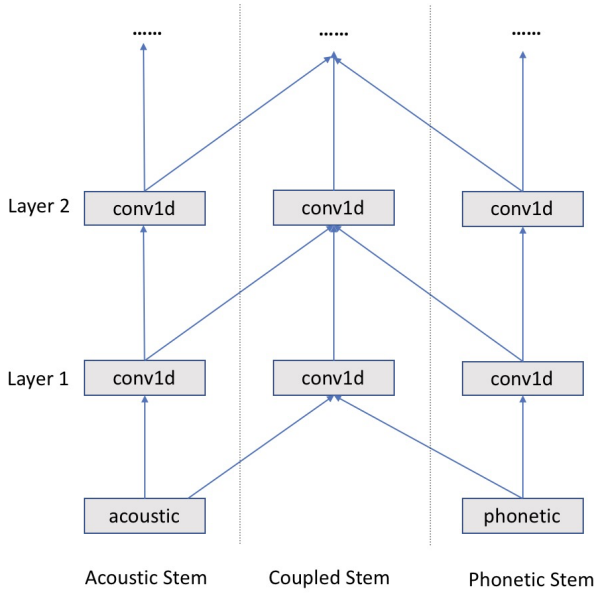

Figure 1: *An illustration of the first two layers of PacNet. The output from Layer 1 from each of the three stems are concatenated and fed to Layer 2 of Coupled Stem.*

The phonetically-aware coupled network (PacNet) introduces frame-level phonetic supervision on the filter bank features. The kernel size, stride and padding configurations on each of the convolution layers from the three stems are kept consistent to ensure accurate frame alignment between ASR bottleneck features and filter bank features.

The structure of the network is designed in a way such that each of the three stems focuses on learning different information from the inputs. The acoustic stem serves the purpose of extracting speaker embedding directly from acoustic features. The phonetic stem is trained to find the transformation of ASR bottleneck feature that provides optimal frame-level supervision. Coupling the two together, the central coupled stem learns to extract more speaker-related information from the acoustic feature of each frame, given the phonetic transformation of that frame.

Table 1: *Network configurations of PacNet*

Layer 7	Linear	In=1024 Out=1000		
Layer 6	Pooling	In=1024 Out=1024		
Layer 5	Conv1d	In=2048 Out=1024		
Layer 4	Conv1d kernel=5	Out=512 In=512	Out=1024 In=2048	Out=512 In=512
Layer 3	Conv1d kernel=5	Out=512 In=512	Out=1024 In=2048	Out=512 In=512
Layer 2	Conv1d kernel=5	Out=512 In=512	Out=1024 In=2048	Out=512 In=512
Layer 1	Conv1d kernel=5	Out=512 In=40	Out=1024 In=140	Out=512 In=100
Stem		Acoustic	Coupled	Phonetic

3.2. Why Triplet Loss

Due to the absence of direct comparison on phonetic contents between the utterance pairs, neural networks with softmax-based objective functions cannot maximize the additional benefits of phonetic supervision. This may be one of the reasons that previous efforts on integrating phonetic information have not shown remarkable gains. In this work triplet loss function is considered and being used to perform normalization based on phonetic information, in contrast to the softmax counterpart and its variations. As mentioned in previous section, the anchor-negative sample pairs with similar phonetic information tend to sit closer than they should be in speaker embedding space. PacNet is trained with the objective to recognize and decouple contributions from acoustic features and phonetic features. Triplet loss function can immediately punish such instances by suppressing contributions from phonetic information and pushing the negative samples further from the anchors. On the other hand, positive anchors with drastically different speech content are forced closer to each other by encouraging non-phonetic contributions.

Softmax loss, on the other hand, does not allow for such direct comparisons on phonetic contents of the two utterances. For each learning sample, neural network performs forward propagation and computes its cross entropy without having to know the acoustic and phonetic information from any other samples. Hence it is hard for the network to directly discern whether the phonetic contents have caused biased scores between any two utterances. This provides a plausible explanation why previous trials to incorporate phonetic information into softmax-based x-vector systems have not been as successful as expected.

Both phonetic and channel information are encoded in x-vectors. Dimensionality reduction techniques such as LDA and PLDA take the major responsibility for compensating channel and content mismatch between utterances, and they have been proven to be very effective on x-vector systems. However, neither LDA nor PLDA makes any nonlinear transformations on the speaker embedding space. They accept the speaker embedding as is and compensate for what the embedding could not account for without trying to perform any actions of nonlinear projections or transformations. Since the x-vector embedding is a speaker representation of the entire utterance, most of the detailed frame-level information has been lost before go-

ing through LDA/PLDA. As a result, no frame level phonetic normalization is carried out. Therefore, it is of our interest to investigate the effects of letting the neural network to learn to handle all the work of compensation for content and channel mismatch.

4. Experiments

4.1. Corpus

The NIST SRE training corpus consists of 57,517 utterances from 5,767 speakers in NIST SRE 04-10 corpus. The performance is evaluated on SRE10 Evaluation set. Since we are interested in investigating the effects of our system on short utterances, only the first 10 seconds from both enrollment and test utterances are used for evaluation. In addition, the Fisher corpus [12] of conversational telephone speech is used to train an ASR network for bottleneck feature extraction.

The systems are also evaluated on the VoxCeleb dataset. The training set contains 1,251 speakers and the test set contains 37 thousand trials from 40 different speakers [13]. Since some of the utterances are over 30 seconds, we truncate up to 4 seconds of non-silence frames for all utterances in evaluation set.

4.2. Baseline

Our systems are compared against both GMM-ivector and x-vector baselines. For both baseline systems, we followed the network configurations and hyper-parameters tuned in Kaldi SRE recipe [7]. A 2048-dimensional GMM is used and 400-dimensional i-vectors are extracted. For x-vector system, 512-dimensional embedding are extracted from TDNN structure. An LDA is applied to reduce the dimension to 150. A PLDA scoring back-end is used for both frameworks.

4.3. Triplet Loss Framework

A TDNN-based triplet loss network is trained in comparison with PacNet. The baseline triplet loss network takes 40-dimensional filter bank features as input. Speaker embeddings of dimension 1000 are extracted. New sets of training triplets are selected at every epoch. For the selection of triplets, we adopt the semi-hard random negative selection strategy with a margin of 0.2 [14]. Euclidean distances are used as a the score measure. The triplet training samples are all truncated to be of length 10 seconds. At every iteration, a random segment of 10 seconds are trimmed from the original audio. In addition, to improve training efficiency, triplets are selected only from the pool of same gender. In all our experiments, no pre-training on softmax have been used.

4.4. PacNet

PacNet adopts the same triplet loss training scheme as described in the previous section. It takes a 40-dimensional filter bank features and 100-dimensional ASR bottleneck features as input. The bottleneck features are extracted from the ASR network trained on 40-dimensional high resolution MFCC features. The coupled network configurations from Table 1 are used. Euclidean distances are computed from the 1000-dimensional embeddings.

5. Results and Discussions

As displayed in Table 2, PacNet achieves an equal error rates of 9.8% on NIST SRE10 evaluation corpus, while the i-vector and x-vector systems with LDA/PLDA scoring backends resulted in 14.66% and 14.26% equal error rates, respectively. The results also show that the major improvements come from the integration of phonetic information and the coupled design, as the ordinary triplet loss network with only acoustic features reports an EER of 13.7%.

Performance of different dimensions of embeddings extracted from PacNet are compared. The performance is almost identical as the embedding dimension increases from 200 to 1000.

Table 2: *Performance comparison on NIST SRE10 corpus, trained on 5,767 speakers, under 10s-10s condition.*

System	Embedding Dimension	EER(%)
i-vector	400	14.66
x-vector	512	14.26
End-to-end triplet	200	13.7
PacNet	200	9.9
PacNet	500	9.9
PacNet	1000	**9.8**

The effects of the number of layers being coupled in PacNet are examined. The results are shown in Table 3. When $L = 1$, only the first layer is coupled. When $L = 2$, Layer 1 and Layer 4 in Table 1 are coupled. When $L = 4$, all convolution layers 1-4 are coupled together. A slight improvement is observed as the number of coupled layers in PacNet increases. The results suggest that the coupled structure has a positive effect on leveraging the additional power of phonetic information.

Table 3: *Performance comparison on different number of coupled layers in PacNet.*

Number of coupled layers	EER(%)
L=1	10.5
L=2	10.2
L=4	**9.8**

Despite the significant gain observed on NIST SRE corpus, the results on VoxCeleb are less encouraging. The baseline end-to-end triplet system reported a reduction in performance, comparing to the x-vector baseline. This is not surprising, considering the complex nature of VoxCeleb corpus, for the triplet framework has been known for its difficulties to train, as well as its deficit of capability to generalize to complicated situations. Therefore, even though the additional phonetic information and PacNet have introduced a decent amount of improvement against the vanilla triplet framework, the overall gains have been refrained by the performance of triplet loss objective function.

Table 4 shows performance of various systems on VoxCeleb1 Evaluation set. For fairness, the Voxceleb 1 and 2 training datasets are used as provided and no data augmentation from other corpus has been applied. As we can see, the i-vector system has a significantly higher EER than the other three. This again suggests the deficiency of i-vector framework on short utterances. However, we noticed that the performance of end-to-end triplet loss framework does not match that of an x-vector

framework. This may be caused by the complicated scenarios of Voxceleb dataset. Our results indicate that a softmax-based system with PLDA back-end is more robust on dataset with diverse scenarios mixed with noises, different recording devices, and different languages. Since triplet-based training places more focus on the detailed comparison between the selected samples, it is more sensitive to the channel and language mismatch in training and evaluation corpus.

PacNet outperforms the vanilla triplet loss system(4.95% vs. 5.85%). However, it only beats the x-vector system by an insubstantial margin, unlike it did in the NIST SRE dataset. We believe that this is related to the nature of VoxCeleb dataset. Collected from open source media under unconstrained conditions, the VoxCeleb dataset provides a closer approximation to the real-world problems. However, since the collection process is not controlled, the speaker-related information are deeply intertwined with channel-related information. For example, the dataset covers a wide range of different languages without having any language labels. In evaluation stage, the language factor helps to differentiate against other speakers who speak different languages. It is difficult to prove whether the trained model is actually trying to recognize speakers' voices, or partially trying to identify other channel-related information such as recording devices, languages, background noises, etc.

The NIST SRE dataset, on the other hand, minimizes the non-speaker-related effects. Following a controlled data collection procedure, the train set and evaluation set consist of only telephone data. Therefore, we are more confident to conclude that the trained model places focus on identifying the speakers' voices, rather than other factors. Therefore, the triplet-based frameworks are more likely to learn useful information from the speakers. The results observed from NIST SRE are more effective indicators for the comparison of different systems in our settings.

Removing the domain mismatch between ASR bottleneck feature extractor and VoxCeleb evaluation set could also lead to significant enhancement in performance. Currently, the ASR bottleneck features used in the experiment are extracted from a network trained on 8,000 Hz Fisher's conversational telephone corpus. There is an inevitable channel mismatch with the Voxceleb corpus, which consists of 16,000 Hz web audio data. A better performance from PacNet should be expected if the ASR features are extracted from a more related channel domain.

Table 4: *Performance on VoxCeleb1 evaluation corpus, under 4s-4s condition.*

System	EER(%)
i-vector	7.56
x-vector	5.11
End-to-end triplet	5.85
PacNet	**4.95**

Another interesting observation shows that the end-to-end triplet framework has demonstrated relatively strong robustness on very small training dataset. When the number of training data reduced to only $2,016$ male speakers on NIST SRE10 and evaluated on male-only trials, the performance of x-vector system dropped significantly, while PacNet and End-to-end triplet network are still able to achieve a reasonable level of accuracy, as shown in Table 5. When the number of training speakers reduces from $5,767$ to $2,016$, the EER for x-vector system increases from 14.26% to 19.2%, while for PacNet increases only from 9.8% to 10.7%. Since triplet loss function directly computes distances between individual samples, it is less sensitive to the size of training dataset. Softmax function and its variations, on the other hand, require a large enough training dataset in order to measure a valid probability distribution over the entire set of training data. Furthermore, the training of LDA and PLDA backends are also affected by the reduction of training data.

Table 5: *Performance comparison on NIST SRE10 corpus, trained on 2,016 speakers, under 10s-10s condition.*

System	EER(%)
x-vector	19.2
End-to-end triplet	14.2
PacNet	**10.7**

6. Conclusions

This paper demonstrates that integrating phonetic information into end-to-end neural networks can significantly increase performance. It is illustrated that a triplet loss training scheme is more fitting than softmax loss system for normalizing phonetic contents. The coupled network structure provides frame-level supervision, which has been proven to be effective on short duration text independent tasks.

More works are needed for the future. First, the system needs to be more flexible for the varied segment lengths in both training and testing sets. The second aspect worthy of investigation is the back-end scoring methods and score normalization techniques related to PacNet embeddings. In our experiments, neither LDA nor PLDA has proven to be superior over Euclidean distance on PacNet embeddings. It is uncertain whether the inclusion of phonetic information and triplet training scheme have accounted for all within-speaker variability that there is no additional information LDA and PLDA can mine.

In addition, we are not arguing that the network structure presented in this work is the optimal choice for learning unbiased speaker embedding under the supervision of phonetic information. Our goal is to present a simple structure that has been proven to be effective and open the gate for further discussions. Further improvements can be expected from more sophisticated designs of network structures.

The strengths and weaknesses of triplet loss function have been thoroughly discussed[15][16]. Its relatively weak capacity to generalize has led to inferior performance in some scenarios. Hence some researchers turned away from it and favored softmax-based objective functions instead. Even though it is shown that phonetic supervision can introduce significant improvement using triplet loss objective, we do not deny that the overall performance may be restrained by the drawbacks of triplet loss under certain situations. Therefore, other contrastive metric learning methods can be examined in order to fully utilize phonetic information.

Despite the related issues that still need to be addressed, we have become increasingly optimistic about the large-scale deployment of speaker verification on short duration text-independent tasks, for the substantial improvement from our experiments.

7. References

[1] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Adv. Sig. Proc.*, vol. 2004, no. 4, pp. 430–451, 2004. [Online]. Available: https://doi.org/10.1155/S1110865704310024

[2] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20*, 2018, pp. 4879–4883. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462665

[3] S. Zheng, G. Liu, H. Suo, and Y. Lei, "Autoencoder-based semi-supervised curriculum learning for out-of-domain speaker verification," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September*, 2019.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011. [Online]. Available: https://doi.org/10.1109/TASL.2010.2064307

[5] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV*, 2006, pp. 531–542. [Online]. Available: https://doi.org/10.1007/11744085_41

[6] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24*, 2017, pp. 999–1003. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0620.html

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20*, 2018, pp. 5329–5333. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461375

[8] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9*, 2014, pp. 1695–1699. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6853887

[9] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September*, 2018, pp. 2247–2251. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1226

[10] M. H. Rahman, I. Himawan, M. McLaren, C. Fookes, and S. Sridharan, "Employing phonetic information in DNN speaker embeddings to improve speaker recognition performance," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September*, 2018, pp. 3593–3597. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1804

[11] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. ernocky, "On the usage of phonetic information for text-independent speaker embedding extraction," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September*, 2019.

[12] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal, May 26-28*, 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/summaries/767.htm

[13] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24*, 2017, pp. 2616–2620. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0950.html

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12*, 2015, pp. 815–823. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298682

[15] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017. [Online]. Available: http://arxiv.org/abs/1703.07737

[16] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1320–1329. [Online]. Available: https://doi.org/10.1109/CVPR.2017.145