VNU Journal of Science: Comp. Science & Com. Eng, Vol. 38, No. 1 (2022) 68-75

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

VLSP 2021 - SV challenge: Vietnamese Speaker Verification in Noisy Environments

Vi Thanh Dat[1], Pham Viet Thanh[2], Nguyen Thi Thu Trang[2,*]

[1]*Sea Company Limited, Hanoi, Vietnam*
[2]*Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam*

Received 27 December 2021
Revised 5 April 2022; Accepted 5 May 2022

Abstract: The VLSP 2021 is the eighth annual international workshop whose campaign was organized at the University of Information Technology, Vietnam National University, Ho Chi Minh City (UIT-VNU-HCM). This was the first time we organized the Speaker Verification shared task with two subtasks SV-T1 and SV-T2. SV-T1 focuses on the development of SV models with limited data, and SV-T2 focuses on testing the capability and the robustness of SV systems. With the aim to boost the development of robust models, we collected, processed, and published a speaker dataset in noisy environments containing 50 hours of speech and more than 1,300 speaker identities. A total of 39 teams registered to participate in this shared task, 15 teams received the dataset, and finally, 7 teams submitted final solutions. The best solution leveraged English pre-trained models and achieved 1.755% and 1.950% Equal Error Rate for SV-T1 and SV-T2 respectively.

Keywords: Speaker Verification, Vietnamese, Data Processing, Speech Processing.

1. Introduction

Speaker verification (SV) is a task of verifying whether an input utterance matches the claimed identity. Over the years, speaker verification has become a much more active field and gained huge advances, attributed to the development of neural networks [1, 2].

In order for researchers to exchange ideas to improve the current state of speaker verification systems and solve the remaining problems, several speaker verification shared tasks have been organized globally. One of the largest and most popular shared tasks is VoxSRC [3]. The challenge has been held for 3 years since 2019. Within the shared task, many interesting methods have been proposed, which contributes to the development of the current speaker verification systems. However, since the dataset consists mostly of utterances from the USA and UK, it is difficult to tell whether these methods will perform well on country-specific data, especially in low-resource settings.

* Corresponding author.
E-mail address: trangntt@soict.hust.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.333

With the aim to leverage the development of speaker verification in Vietnamese, we have organized the Vietnamese speaker verification shared task in the framework of the eighth workshop of Vietnamese Language and Speech Processing - VLSP 2021. This is the first time we organized the shared task at VLSP. By organizing the challenge, we hope to provide the community with a Vietnamese speaker verification dataset and to set up a benchmark for Vietnamese speaker verification. This year's challenge includes two evaluation tasks, in which there is a private test set dedicated to each task. Participants can join in one of the tasks, or both of them:

- Task-01 (SV-T1): Focusing on the development of SV models with limited data. For this task, participants can only use the provided training dataset for model development. Any use of additional data for model training is prohibited.
- Task-02 (SV-T2): Focusing on testing the robustness of SV systems. In this task, the majority of evaluation data consists of hard-sampled pairs. Participants can use the provided training set and any additional data.

Figure 1. Participants in Speaker Verification task - VLSP 2021.

For speaker verification systems to be applicable in the real world, they have to work well in various conditions, especially in noisy environments [4, 5]. With that thought in mind, we have built the shared task dataset with speech collected from different environments and different region-specific accents of Vietnam. Throughout the challenge, there have been several interesting methods proposed by the participants which achieved remarkable results.

The rest of this paper is organized as follows. Section 2 provides information about participants and the phases of the challenge. In Section 3, we discuss the process of data preparation for the shared task. Evaluation results are described in Section 4. Lastly, we draw conclusions and discuss future works in Section 5.

2. Participants

For the SV shared task this year, each team has to go through two main phases: the data cleaning phase, which will be described in Section 3.2, and the competition phase. In order to receive the dataset and participate in the competition, registered teams had to go through the data cleaning phase.

The competition phase of the competition is further divided into 2 stages: public test, private test. Public test stage is hosted on the AiHub platform for 33 days at the link https://aihub.vn/competitions/62. Each team can submit 10 submissions per day and immediately get their results to benchmark, finetune, and improve their solutions. After the Public test phase ends, participating teams have two days to submit their final submissions for the two private tests SV-T1 and SV-T2. Each team only gets 5 submissions per test in this stage. After that, final scoreboards and top teams for two tasks are announced; these teams proceed to write and submit technical reports. The scoring metric, final scoreboards, and solutions are further discussed in Section 4.

In recap, thirty nine teams registered for the Speaker Verification challenge. All of them were invited to participate in the data cleaning phase. After that, the total number of teams eligible for receiving the dataset and entering the competition phase is narrowed down to 15 teams. Out of which, 8 teams participated in the public test and 7 teams submitted their final results on private tests. Figure 1 illustrates the number of teams throughout different stages.

3. Data Building

As stated above, the purpose of the challenge is to promote the development of Vietnamese speaker verification systems in noisy environments. In this section, we discuss the process of building the VLSP 2021 speaker verification dataset from highly diverse existing datasets as well as talk shows and TV reportage channels on Youtube.

3.1. Data Building Pipeline

The overall data building pipeline is illustrated in Figure 2. To prepare for the final dataset, we collect data from different sources, including public datasets and Youtube data.

The chosen public datasets are ZaloAI , VLSP 2020 and VIVOS. ZaloAI is a dataset designed for experimenting with speaker verification, while VLSP 2020 and VIVOS are speech recognition datasets which contain speaker information. Additionally, we crawled external data from TV programs from various Youtube channels. These programs are from talk shows, game shows or TV reportage with varied background environments including inaudible chatter, laughs, street noise, school, music, ... Which makes identifying a person with his or her voice much more challenging compared to studio-recorded speech. The external data includes audio samples of different environments and different Vietnamese accents. Table 1 shows the information of each TV program.

Table 1. TV programs information:

TV program	Environments	Main Accent
Vietcetera	Talk show	Mixed
Bua Trua Vui Ve	Game show	Northern
Goc Hue Trong Toi	TV Reportage	Central
Khong Cay Khong Ve	Talk show	Mixed
Tu Tinh Luc 0h	Talk show	Mixed
Vuot Doc	TV Reportage	Southern

One problem of the external data is that there are no speaker identities. We address this by grouping the utterances into clusters, in which each cluster represents an identity. Firstly, we generate speaker embeddings of utterances using a pre-trained speaker verification model. Then the samples with cosine similarity scores higher than a pre-defined threshold are grouped into clusters.

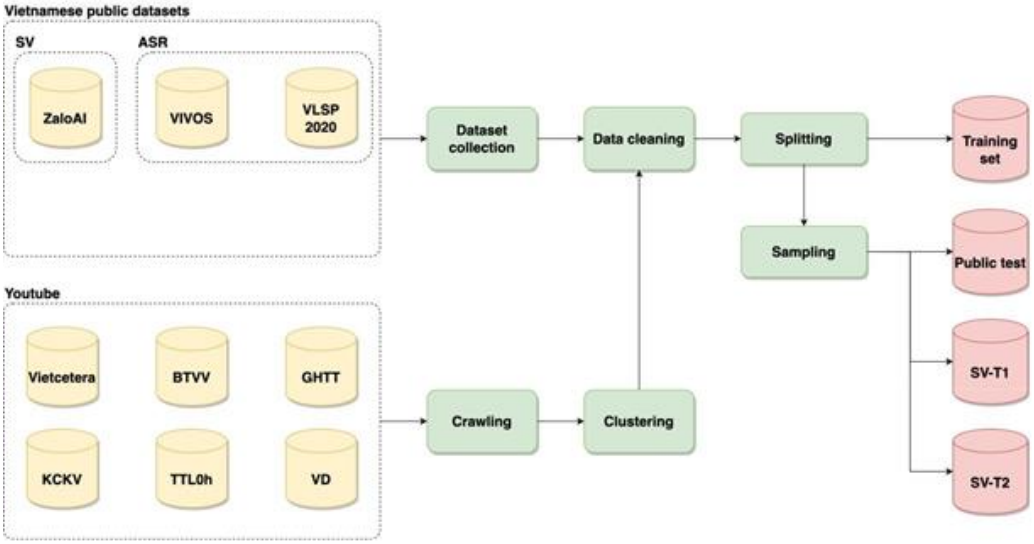

Figure 2. Overall processes of building the VLSP 2021 speaker verification dataset.

3.2. Data Cleaning

After clustering the Youtube data, all datasets are combined and go through the data cleaning step. The process includes removing invalid utterances and validating the speaker identities, and is done via a web application. Each team is required to validate 500 pairs of utterances out of 6,000 pairs in total. Figure 3 shows the interface of the application. Utterances and speakers are removed or merged based on major votes on utterance pairs. For each pair, the team could choose one of the following five options: "Utterances come from the same identity", "Utterances come from different identities", "There are multiple identities in utterance 1", "There are multiple identities in utterance 2", and "There are multiple identities in both utterances"

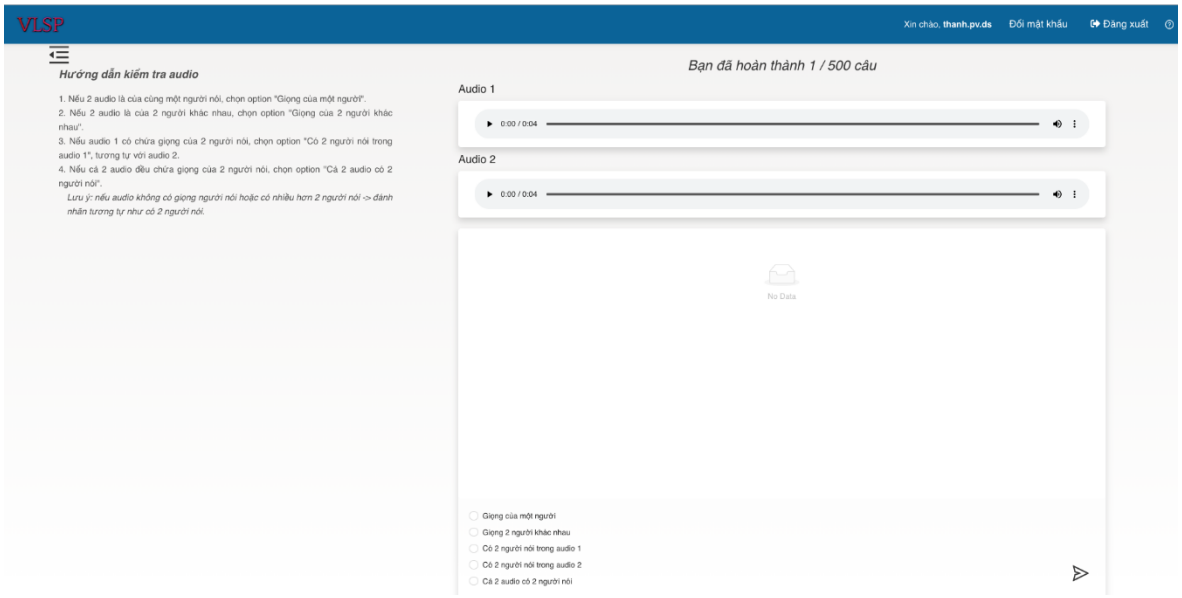

Figure 3. Online Tool for data cleaning.

3.3. Final Datasets

The final step of the data building pipeline is splitting the data into a training set and different test sets. In the case of the training set, we use audio samples from the public datasets and from the following Youtube audio sets: Bua Trua Vui Ve, Vietcetera and Goc Hue trong toi. Therefore, the test sets Bua Trua Vui Ve and Goc Hue Trong Toi are considered in-domain test sets with similar characteristics. Whereas Khong Cay Khong Ve, Tu Tinh Luc 0h, Vuot Doc are considered out-domain. To build the private test sets, as shown in Table 2, we only use newly collected Youtube datasets.

Each test set has a different number of hard negative pairs. A hard negative pair is a pair of utterances from 2 different speakers having the same hand-labeled gender and regional accent.

Table 2. Number of pairs from each audio set in the test sets:

Audio Set	Public test	SV-T1	SV-T2
VLSP 2020	7,135	0	0
Bua Trua Vui Ve	6,286	11,820	6,217
Goc Hue trong toi	6,579	8,998	4,639
Khong Cay Khong Ve	0	7,112	11,532
Tu tinh Luc 0h	0	6,112	8,791
Vuot Doc	0	5,958	8,821

50% of the pairs of the public test set and SV-T1 private test set are hard negative pairs. For the private test set of SV-T2, the number is 80%, since the task focuses more on assessing the robustness of speaker verification systems. Embedding models generate high similarity scores for people with similar voices, which skews the score distribution and requires more robust models to achieve low Equal Error Rate.

Table 3 shows the statistics of the dataset after splitting.

Table 3. Dataset statistics:

Set	Hours	Speakers	Utterances	Pairs
Training	41.43	1,305	31,600	-
Public test	4.35	114	2,941	20,000
SV-T1	4.91	125	3,983	40,000
SV-T2	4.91	125	3,983	40,000

4. Evaluation

4.1. Evaluation Metrics

The widely-adopted metric Equal Error Rate (EER) (Figure 4) is chosen to measure performance in the competition. EER is the location in Receive Operating Characteristic (ROC) curve or Detection Error Tradeoff (DET) curve where the false acceptance rate and false rejection rate are equal. EER can be found by shifting the decision threshold. In general, the lower EER value, the higher accuracy of the system.

To conclude the final scoreboard for each task, a team's submission with lowest EER is chosen as its final solution.

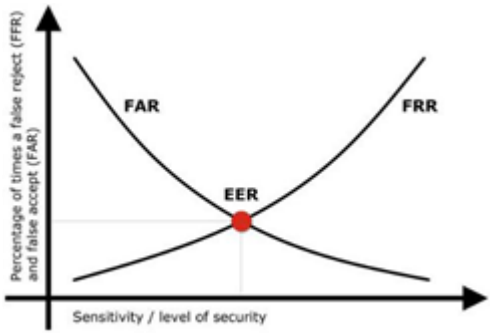

Figure 4. Equal Error Rate.

4.2. Evaluation Results

Final scoreboards

Table 4 and Table 5 show the best EER of teams for SV-T1 and SV-T2. We can see that Smartcall-ITS system has the best performance in both tasks with EERs of 1.755% and 1.950% respectively. Remaining teams scored EERs ranging from 5.305% to 11.605% which are substantially higher than the first-ranking solution. We will try to explain this performance gap by looking at teams' methods in Subsection 4.3.

Table 4. Final scoreboard for SV-T1:

Rank	Team	EER(%)
1	Smartcall-ITS	1.755
2	hynguyenthien	5.305
3	AssistantReg	5.800
4	EASV	5.955
5	anbn14	6.830
6	proptit	6.845
7	ffyytt	8.805

According to technical reports, top teams do not use separate systems or additional data for SV-T2 compared to SV-T1 due to time and data constraints. We can observe that systems exhibit a significant performance degradation between SV-T1 and SV-T2 (for example 8.55% degradation for Smartcall-ITS, 24.60% degradation for hynguyenthien, 23.88% for AssistantReg, and 8.56% for EASV). This proves that our hard negative sampling strategy works and that SV systems should be carefully calibrated to voices from the same region and gender if applied in real-life scenarios.

Table 5. Final scoreboard for SV-T2:

Rank	Team	EER(%)
1	Smartcall-ITS	1.950
2	EASV	6.465
3	hynguyenthien	6.610
4	AssistantReg	7.185
5	ffyytt	11.605

Performance breakdowns on different audio sets. Figure 5 and Figure 6 demonstrate the performance breakdowns on different audio sets for SV-T1 and SV-T2. Overall, we can see that all teams perform worse on out-domain audio sets (Khong Cay Khong Ve, Tu Tinh Luc 0h, and Vuot Doc) compared to in-domain sets (Bua Trua Vui Ve and Goc Hue Trong Toi). Out-domain sets can have different recording devices, background environments, everberation settings, etc. compared to the training set.

Contrastly, in-domain sets can have recognizable attributes which models potentially learned from the training set.

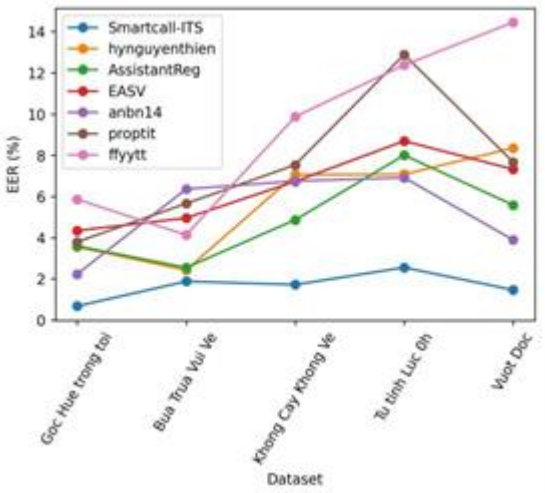

Figure 5. Performance breakdown by audio sets for submitting teams in T1.

We can see that the 1st-ranking Smartcall-ITS system has the best performance unanimously among the audio sets and low-performance degradation between in-domain and out-domain sets in both SV-T1 and SV-T2. On the contrary, other teams seem to have higher performance degradation and less consistent results between audio sets especially out-domain audio sets. For example, in SV-T1, hynguyenthien outperforms AssistantReg on Vuot Doc set but not Khong Cay Khong Ve set.

We did observe the false-negative pairs and found that various microphone distances can greatly distort one's voice thus making the systems faulty. Additionally, recording voices from playbacks and phone calls (in some shows, guests can participate via phone or video call) could also negatively affect the quality of voices. Finally, noisy background causes the most errors in the test sets. This puts an emphasis on developing good quality check services before SV systems.

Table 6. Method summary:

Team	Pre-train	Backbone	Pooling layer	Loss Function	Backend	Others
SmartCall-ITS	Yes	TDNN, ResNet34	ASP	AAM-softmax, AP loss	PLDA, Cosine	Fusion, AS-norm
EASV	No	X-vector Backbone, ECAPA -TDNN	ASP	Softmax, GE2E	Cosine	Fusion
hynguyenthien	Yes	Thin-ResNet34	GhostVLAD	Softmax, AM-Softmax	Cosine	-
AssistantReg	No	ResNet34S	NetVLAD	Softmax	Cosine	-

4.3. Method Summary

Table 6 summarizes the approach of top teams at SV task this year. All teams subscribe to X-Vector [6] architecture with various component choices. The chosen backbones include ECAPA-TDNN, ResNet 34 variations, and TDNN [7, 8, 9]. The chosen pooling layers are ASP, dictionary-based GhostVLAD, and NetVLAD [10, 11, 12]. Loss functions are of both types objective losses such as Softmax, AM-Softmax, AAM-Softmax [13, 14] and metric losses such as GE2E and AP [15, 16].

Compared to other teams, Smartcall-ITS and EASV utilize the recent ECAPA-TDNN and ResNet34 pre-trained models which were trained on the enormous English dataset VoxCeleb [3]. These models achieve sub 1% EER on English test set and transfer well to other languages. However, in the EASV solution, they only use ECAPA-TDNN as is to fuse with the results from X-vector without finetuning due to time constraints. Additionally, Smartcall-ITS also experimented with other techniques including PLDA and AS-norm to boost verification performance. This

explained the large performance gap between Smartcall-ITS and others.

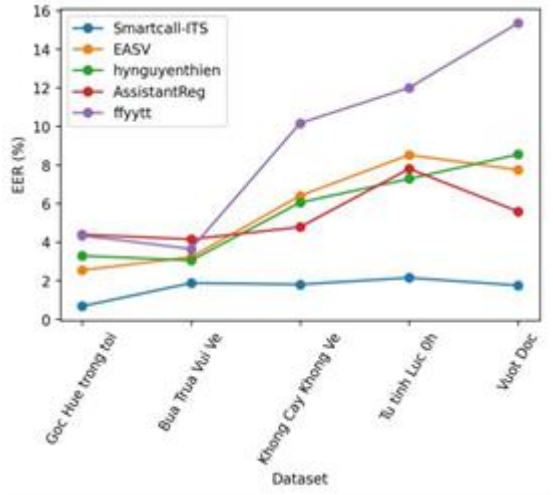

Figure 6. Performance breakdown by audio sets for submitting teams in T2.

5. Conclusion

In this paper, we have summarized the organization of the first speaker verification shared task at VLSP 2021. The shared task includes two subtasks SV-T1 and SV-T2. SV-T1 focuses on the development of SV models with limited data, and SV-T2 focuses on testing the capability and the robustness of SV systems. We have collected, processed, and published a speaker dataset in noisy environments containing 50 hours of speech and more than 1,300 speaker identities. A total of 39 teams registered to participate in this shared task and each team had to make contributions by cleaning the dataset. A total of 15 teams were eligible to receive the dataset and participate in the competition. Finally, 7 participants submitted final solutions. The best solution leveraged English pre-trained models and achieved 1.755% and 1.950% EER for SV-T1 and SV-T2 respectively. Errors including voice distortion and noisy environment put an emphasis on building a good quality check service for SV systems. For VLSP Campaign in 2022, we wish to organize one of the adjacent tasks of SV such as speaker identification or speaker diarization.

References

[1] M. McLaren, Y. Lei, L. Ferrer, Advances in deep neural network approaches to speaker recognition, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 4814–4818.

[2] Z. Bai, X.-L. Zhang, Speaker recognition based on deep learning: An overview, Neural Networks.

[3] J. S. Chung, A. Nagrani, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, A. Zisserman, Voxsrc 2019: The first voxceleb speaker recognition challenge, ISCA Challenges.

[4] J. Ming, T. J. Hazen, J. R. Glass, D. A. Reynolds, Robust speaker recognition in noisy conditions, IEEE Transactions on Audio, Speech, and Language Processing 15 (5) (2007) 1711–1723.

[5] M. I. Mandasari, M. McLaren, D. A. van Leeuwen, The effect of noise on modern automatic speaker recognition systems, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 4249–4252.

[6] D. Snyder, D. Garcia-Romero, G. Sell, Povey, S. Khudanpur, X-vectors: Robust dnn embeddings for speaker recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5329–5333.

[7] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: Proc. Interspeech 2020, 2020, pp. 3830–3834. doi:10.21437/Interspeech.2020-2650.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[9] V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts, in: Sixteenth annual conference of the international speech communication association, 2015.

[10] K. Okabe, T. Koshinaka, K. Shinoda, Attentive statistics pooling for deep speaker embedding, CoRR abs/1803.10963.

[11] Y. Zhong, R. Arandjelovic,´ A. Zisserman, Ghostvlad for set-based face recognition, in:

Asian conference on computer vision, Springer, 2018, pp. 35–50.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, Sivic, Netvlad: Cnn architecture for weakly supervised place recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.

[13] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5265–5274.

[14] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Signal Processing (ICASSP), IEEE, 2018, Additive angular margin loss for deep face pp. 4879–4883.

[15] L. Wan, Q. Wang, A. Papir, I. L. Moreno, in: Proc. Interspeech 2020, 2020, pp. 2977–2981.

[16] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, Conference on Computer Vision and Pattern S. Choe, C. Ham, S. Jung, B.-J. Lee, I. Han, In Recognition, 2019, pp. 4690–4699.