Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19-21, 2024
Paper No. MVML 106
DOI: 10.11159/mvml24.106

[Case Study] Transfer Learning with Inflated 3D CNN for Word-Level Recognition for Azerbaijani Sign Language Dataset

Nigar Alishzade[1,2] and Gulchin Abdullayeva[2]
[1]French-Azerbaijani University under ASOIU
Baku, Azerbaijan
nigar.alishzada@ufaz.az
[2]MSERA Institute of Control Systems
Baku, Azerbaijan
gulchin.abdullayeva@isi.az

Abstract **-** Sign language is a non-verbal form of communication primarily used by the Deaf and Hard of Hearing (DHH). Sign language recognition (SLR) is the automatic recognition of sign language that treats each sign as a class. In related work, significant progress has been made using deep learning techniques. Large amounts of data are typically required to train deep learning models effectively. In the academic community working on SLR, different well-tuned models provide benchmarking on different sign language datasets, and many sign languages still lack corresponding datasets, making it challenging to train models for these languages. Transfer learning is a technique that utilizes a related task with abundant available data to solve a target task with insufficient data. Transfer learning has been successfully applied in computer vision and SLR. This study investigates how effectively transfer learning can be applied to isolated SLR using an inflated 3D convolutional neural network as a deep learning architecture. Transfer learning is implemented by pre-training a network and subsequently fine-tuning it on a small Azerbaijani Sign Language Dataset. This approach is promising because it leverages the knowledge gained from a related task with more abundant data to enhance the model's performance on the target SLR task.

Keywords**:** Sign Language Recognition, Transfer Learning, Inflated 3D CNN, Dynamic Sign Recognition

1. Introduction

SLR facilitates communication between sign language users and non-signers, promoting inclusivity and accessibility. The application of transfer learning in isolated SLR is a promising method, especially for languages with limited datasets. Transfer learning enables the model to utilize the knowledge obtained from a similar task with ample data to enhance performance on a specific task with scarce data [1]. In our case, we used a pre-trained deep learning model on a task with more available data and then fine-tuned it on the Azerbaijani Sign Language Dataset [3] which enhanced the model's ability to recognize signs effectively.

Our research focuses on the significance of pre-trained models in visual-based SLR, particularly highlighting the influence of 3D convolutional neural networks and LSTM layers in the chosen backbone model when dealing with dynamic signs. This approach allows us to leverage the knowledge learned from a large dataset, such as American Sign Language, and apply it to a target task with a smaller dataset, such as Azerbaijani Sign Language.

3D convolutional neural networks (CNNs) are well-suited for capturing spatiotemporal features, which is essential in recognizing sign language where both hand movements and gestures over time contribute to the meaning of signs. Initially successful in human action recognition tasks [3], [4], [5], 3D CNNs have since been adopted by the Sign Language Processing research community. Employing transfer learning and harnessing the capabilities of 3D CNNs offers a comprehensive approach to tackling the inherent challenges in isolated sign language recognition, especially for languages with limited datasets. This reflects a dedication to advancing the field through utilizing existing knowledge and customizing it to address the specific requirements of marginalized linguistic and cultural contexts.

By pre-training a deep learning model on a task with more available data and subsequently fine-tuning it on the Azerbaijani Sign Language Dataset, the model stands to gain a deeper understanding of the underlying spatiotemporal

features crucial for recognizing signs effectively within the specific linguistic and gestural context of the Azerbaijani Sign Language. The selection of 3D Convolutional Neural Networks for capturing spatiotemporal features aligns with the intricate nature of sign language recognition. In sign language, meaning is often conveyed not only through static hand gestures but also through the dynamic interplay of hand movements and gestures over time. The temporal aspect is a critical dimension for accurately interpreting the signs. Studies such as [6] have demonstrated the efficacy of 3D CNNs in extracting and understanding spatiotemporal patterns, making them well-suited for tasks that involve sequential data, such as sign language recognition.

2. Related Work

We primarily focus on CNN-oriented computer vision techniques because SLR involves human action recognition. Sign languages consist of complex sets of human gestures and body articulations, so the majority of the techniques were adapted from this superclass. The study by Khurana et al. [7] offers a thorough overview of the current literature on sign language recognition, employing deep learning methods and transfer learning techniques.

In [8], the authors proposed Spatial-Temporal Graph Convolutional Networks for skeleton-based action recognition, which move beyond the limitations of previous methods by automatically learning both the spatial and temporal patterns from data.

The authors of [9] proposed a two-stream CNN framework for action recognition, incorporating both spatial and temporal information from RGB and optical flow images.

In [10], the authors used 2D CNNs for training and bootstrapped the weights into 3D. They devised a 2D-Inflated operation and a parallel 3D ConvNet architecture for converting pre-trained 2D ConvNets into 3D ConvNets, which avoids video data pre-training.

In the field of sign language recognition, studies have used transfer learning and convolutional neural networks with pre-trained models like InceptionV3 and ResNet-50 to improve accuracy [11]. These studies have shown promising results in recognizing various sign languages, including American Sign Language, British Sign Language, and Indian Sign Language [12, 13]. Furthermore, the utilization of multi-modal data sources, such as infrared, contour, and skeleton information, has been proven to enhance the performance of sign language recognition systems [1, 3].

The results of the experiments [14] give clear empirical evidence that transfer learning can be effectively applied to isolated SLR. The accuracy performances of the networks applying transfer learning increased substantially by up to 21% as compared to the baseline models that were not pre-trained on the MS-ASL dataset.

In [1] the authors proposed a Skeleton Aware Multi-modal Sign Language Recognition framework, which takes advantage of multi-modal information to improve recognition rate. Specifically, the framework incorporates a Sign Language Graph Convolution Network to model the dynamics of sign language gestures, and a Separable Spatial-Temporal Convolution Network to exploit skeleton features.

In [15], the authors introduced multi-scale spatiotemporal graph convolutional networks for isolated sign language recognition. These networks leverage the spatiotemporal characteristics of sign language gestures by incorporating graph convolutional networks at multiple scales.

In [16], the authors proposed a (2+1)D-SLR network based on (2+1)D convolution that can achieve higher accuracy with a faster speed. Because (2+1)D-SLR can learn spatio-temporal features from the raw sign RGB frames.

The authors of [17] proposed a Temporal Interaction Module to capture both spatial and temporal information in sign language videos. They suggest not only can it obtain higher accuracy, but also inference speed and parameters of the network can meet practical application scenarios, because the TIM-SLR network is only composed of 2D convolution and temporal interaction module (TIM).

In [18], the authors provided comparative evaluations of different deep learning architectures for word-level sign language recognition, including 3D CNNs.

In [19] and [20] authors provide different approaches for Azerbaijani Sign Language fingerspelling alphabet and also employ transfer learning.

Addressing the dynamic nature of signing, including temporal dependencies as well as spatiotemporal characteristics, has been an area of focus in various works aiming to enhance the accuracy and efficiency of sign language recognition systems [21].

In conclusion, sign language recognition research has made significant progress by utilizing convolutional neural networks, transfer learning, and multi-modal information. In our study, we aim to build upon these advancements. By leveraging transfer learning on 3D convolutional neural network architecture, and multi-modal data streaming we aim to improve the accuracy of isolated sign language recognition for our specific dataset.

3. Methodology

Our research focuses on the significance of pre-trained models in visual-based SLR. As a backbone, we have selected a 3D CNN model with weights obtained from training with kinetics depth data. For fine-tuning, we used a custom dataset collected by frontal set RGB cameras for Azerbaijani Sign Language recognition. Our findings show that utilizing transfer learning through pre-training on a large-scale dataset like kinetics depth data greatly improves the accuracy of sign language recognition for Azerbaijani Sign Language compared to traditional methods.

3.1. The Dataset and Data Preparation

We assembled a dataset tailored for the identification of Azerbaijani Sign Language, comprising RGB video recordings obtained from front-facing cameras. Initially recorded in sentence form, the videos were later trimmed into individual words. The selection process involved choosing only those videos with at least 20 samples for labeling and then focusing on the top 100 classes (words) for further analysis. Subsequently, these videos were divided into training and testing sets using a randomized split ratio of 9:1. Certain preprocessing techniques such as center-zoom and normalization techniques were applied to the RGB video data to enhance the quality and consistency of the input.

Addressing imbalanced datasets is a critical aspect of training robust machine learning models. In our case, where one class constituted more than 10% of the entire dataset, an imbalanced distribution could lead to biased model predictions. To counteract this imbalance, class weights are widely employed during the cost-sensitive training. By calculating and incorporating the class weights, the model was provided with a mechanism to appropriately weigh the contribution of each class during training. This approach ensures that the model is not disproportionately influenced by the larger class and helps it generalize more effectively across the entire dataset. We calculates class weight as

$$Class\_weight\_of\_the\_specific\_class = number\_of\_the\_samples\_in\ the\_largest\_class \,/\, number\_of\_the\_samples\_for\_this\_specific\_class$$

Multiplying the output of each class by its respective class weight serves as a guidance mechanism, signaling the relative importance of each class based on its representation in the data to the model. This nuanced training strategy is pivotal for achieving a balanced and unbiased model, enabling more accurate predictions across all classes, particularly in scenarios where class imbalances are prominent. We continue handling inbalanced data distribution problem by applying *cost-sensitive training.*

The data preprocessing steps involved saving label-to-one-hot encoding dictionaries, paths for training and testing, and class weights for streaming. To prevent memory issues, the video samples were not fully loaded into RAM. Instead, we utilized data streaming to read the samples.

3.2. Multimodal Data Streaming

Multimodal training entails incorporating multiple modalities to enhance the model's feature representation. By integrating various streams, the model can process diverse types of information concurrently. In our approach, we an RGB stream with TVL1 and Farneback as optical flow modalities to enable a more comprehensive feature for robust recognition. This is particularly important as RGB videos contain visual redundancy that may pose challenges single-stream models in extracting key information efficiently. Our multimodal training method, which integrates both and optical flow modalities, allows our model to leverage spatial information effectively at a higher level, leading to improved accuracy and robustness in recognizing Azerbaijani Sign Language signs. We also adjusted input layer parameters to accommodate differences in shape and batch size across dataset streams while leveraging complementary fusion of information from different streams for enhanced performance. Additionally, we utilized pre-trained 3D Convolutional Neural Networks trained on large-scale datasets like Kinect depth data to benefit from their learned representations and capture more complex features when fine-tuned for sign language data.

3.3. Model Architecture

The model architecture is based on a deep Convolutional Neural Network that utilizes 3D convolutions to capture spatiotemporal information across video frames. The model takes as input a sequence of video frames, either RGB or optical flow modalities, and applies 3D convolutions to extract spatial and temporal features simultaneously. The input layer is being modified to match the shape and batch size of the input data from different modalities *(((n_frames,),output_size (num_channels,)) which is (20,224,224,3) for RGB and Farneback an (20,224,224,2) for TVL1 Optical flow*. In the Fig. 1. the top layers of the model demonstrated:

Tables and figures should be placed close to their first citation in the text. All figures and tables should be numbered. Table headings should be centred above the tables. Figure captions should be centred below the figures. Refer to the figure below for a sample.

```
if include_top:
    x = self.Avg3DPool((2, 7, 7), strides=(1, 1, 1), padding='valid', name='top.global_avg_pool')(x)
    x = self.Dropout(dropout_prob,name = 'top.dropout')(x)
    x = self.conv3d_bn(x, classes, 1, 1, 1, padding='same',
            use_bias=True, use_activation_fn=False, use_bn=False, name='top')
    num_frames_remaining = int(x.shape[1])
    x = self.Reshape((num_frames_remaining, classes),name = "top.reshape")(x)

    # logits (raw scores for each class)
    x = self.Lambda(lambda x: K.mean(x, axis=1, keepdims=False),
            output_shape=lambda s: (s[0], s[2]))(x)
    x = self.activation('softmax', name='top.prediction')(x)
```

Fig. 1: Top layers of I3D CNN

Before I3D, we utilized a model with late diffusion that involved generating feature maps for each frame using the pretrained MobileNet model. These sequential feature maps were then inputted into the LSTM model to address the final prediction problem. However, this approach was limited in capturing frame interactions as it only relied on late feature maps from the model, resulting in some loss of information about connections and interactions between frames. The core issue lay in capturing temporal features essential for identifying dynamic patterns. The I3D model tackled this limitation by incorporating continuous fusion techniques.This allowed for the extraction of both spatial and temporal features, capturing the dynamic patterns present in the target dataset more effectively.

3.4. Training

During the training phase, we followed a two-step process. Initially, we trained the top layer of the model (Fig. 1.) to fit the new initialized parameters to the previously trained parameters. This process helps prevent the top untrained parameters from interfering with the previously trained parameters. Then, we unfreeze the rest of the backbone layers and fine-tune the entire model using the target dataset.

The number of epochs during experiments typically varied between 50-100 depending on the convergence of the training loss and validation accuracy. During the first step of the training phase, we froze the backbone layers and only

trained the top layer with a lower learning rate to focus on fine-tuning the new parameters. In the second step, we unfreezed the backbone layers and used a higher learning rate to fine-tune the entire model. After training, we evaluated the performance of our model using the F1 score. Fig. 2. and Fig. 3. show trainig and testing evaluation, accordingly:

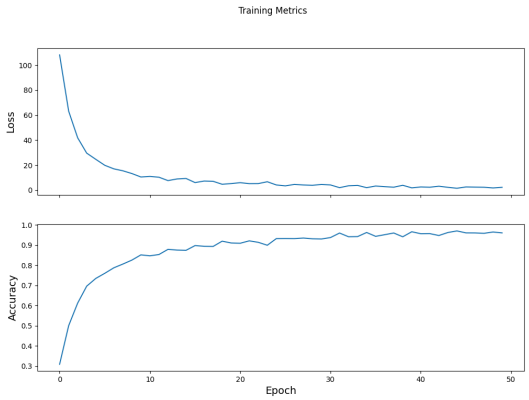

Fig. 2. Training progress evaluations

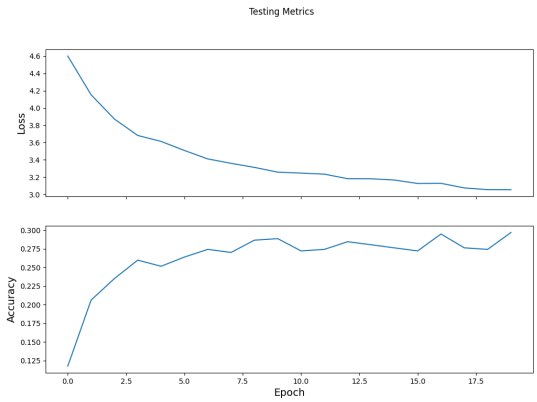

Fig. 3. Testing progress evaluations

3.5. Testing and evaluation metrics

To evaluate the performance of our model in recognizing Azerbaijani Sign Language gestures, we conducted extensive experiments and measured the F1 score as the evaluation metric. The experiments involved testing our model on a dataset of Azerbaijani Sign Language gestures and comparing the results with the ground truth labels. Our model achieved an F1 score of 84.85% on the test dataset, demonstrating its effectiveness in accurately recognizing Azerbaijani Sign Language gestures. The Fig. 4. shows the F1 scores for the top 10 frequent words (classes). Fig. 5. shows the same results for the top 20 classes:

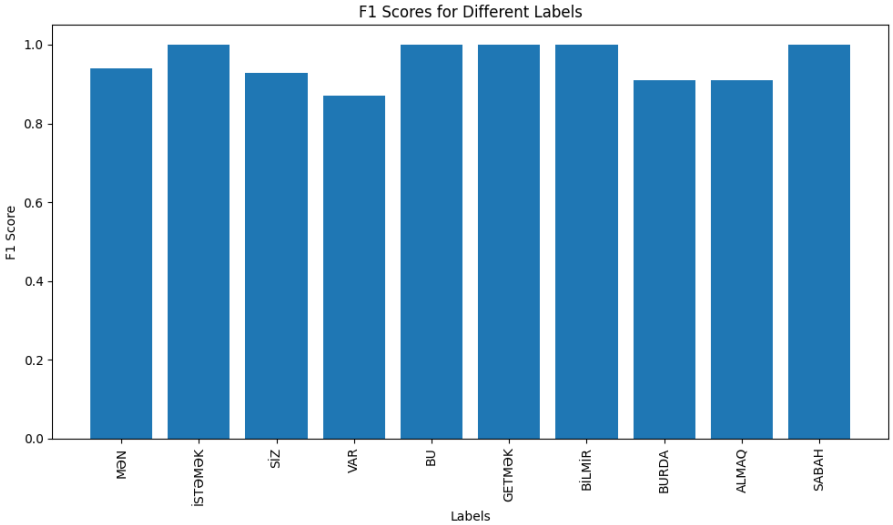

Fig. 4. F1-scores for top 10 classes

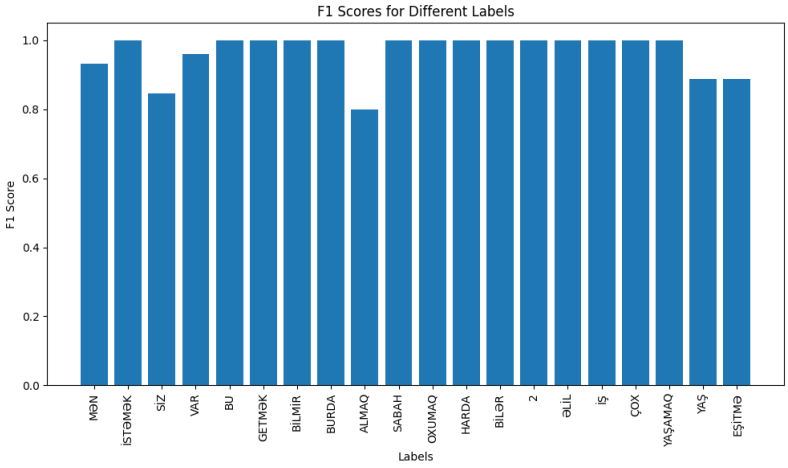

Fig. 5. F1-scores for top 20 classes

4. Conclusion

Our research highlights the potential of using pre-trained 3D convolutional neural networks to recognize sign language, specifically focusing on Azerbaijani Sign Language. By implementing transfer learning and fine-tuning methods, we achieved high accuracy in identifying gestures. To further improve our findings, we plan to expand the dataset by including a wider range of sign language gestures.

Transfer learning significantly improved final accuracy and training behavior across all experiments. It led to faster convergence, requiring fewer training epochs while demonstrating better initial accuracy on the validation dataset – indicating enhanced generalization capability from the outset.

Furthermore, our plans involve evaluating attention-based convolutional neural network architectures and comparing their performance with the pre-trained 3D CNN used in this study. Ultimately, our paper proposes an approach for sign language recognition employing pre-trained 3D convolutional neural networks. Our methodology delivered promising results in accurately recognizing Azerbaijani Sign Language gestures with an overall F1 score of 84.85%.

References

[1] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021, March 15). Skeleton Aware Multi-modal Sign Language Recognition. Cornell University. https://doi.org/https://doi.org/10.48550/arxiv.2103.08833

[2] Jamaladdin Hasanov, 2023. Data loader code for AzSL. https://github.com/ADA-SITE-JML/azsl_dataloader

[3] Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., & Mak, B. (2022, November 2). Two-Stream Network for Sign Language Recognition and Translation. Cornell University. https://doi.org/https://doi.org/10.48550/arxiv.2211.01367

[4] Ji, Shuiwang & Xu, Wei & Yang, Ming & Yu, Kai. (2010). 3D Convolutional Neural Networks for Human Action Recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 35. 495-502. 10.1109/TPAMI.2012.59.

[5] J. Arunnehru, G. Chamundeeswari, S. Prasanna Bharathi, Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos Procedia Computer Science, Volume 133, 2018, Pages 471-477, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.07.059

[6] Y. Xu, Y. Feng, Z. Xie, M. Xie and W. Luo, "Action Recognition Using High Temporal Resolution 3D Neural Network Based on Dilated Convolution," in **IEEE Access**, vol. 8, pp. 165365-165372, 2020, doi: 10.1109/ACCESS.2020.3022407.

[7] S. Khurana, R. Sreemathy, M. Turuk and J. Jagdale, "Comparative Study and Performance Analysis of Deep Neural Networks for Sign Language Recognition using Transfer Learning," **2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)**, Bhilai, India, 2023, pp. 1-8, doi: 10.1109/ICAECT57570.2023.10118114.

[8] Yan, S., Xiong, Y., & Lin, D. (2018, April 27). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v32i1.12328

[9] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27.

[10] Y. Huang, Y. Guo and C. Gao, "Efficient Parallel Inflated 3D Convolution Architecture for Action Recognition," in **IEEE Access**, vol. 8, pp. 45753-45765, 2020, doi: 10.1109/ACCESS.2020.2978223.

[11] Novopoltsev, M., Verkhovtsev, L., Murtazin, R., Milevich, D. \& Zemtsova, I. Fine-tuning of sign language recognition models: a technical report. (2023)

[12] Sakshi Sharma, Sukhwinder Singh, ISL recognition system using integrated mobile-net and transfer learning method, Expert Systems with Applications, Volume 221, 2023, 119772, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.119772

[13] Hira Hameed, Muhammad Usman, Muhammad Zakir Khan, "Privacy-Preserving British Sign Language Recognition Using Deep Learning," **2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)**, Glasgow, Scotland, United Kingdom, 2022, pp. 4316-4319, doi: 10.1109/EMBC48229.2022.9871491.

[14] Töngi, R. (2021, February 25). Application of Transfer Learning to Sign Language Recognition using an Inflated 3D Deep Convolutional Neural Network. arXiv.org. https://arxiv.org/abs/2103.05111

[15] M. Vázquez-Enríquez, J. L. Alba-Castro, L. Docío-Fernández and E. Rodríguez-Banga, "Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 3457-3466, doi: 10.1109/CVPRW53098.2021.00385.

[16] Wang, Fei & Du, Yuxuan & Wang, Guorui & Zeng, Zhen & Zhao, Lihong. (2022). (2+1)D-SLR: an efficient network for video sign language recognition. Neural Computing and Applications. 34. 10.1007/s00521-021-06467-9.

[17] Fei Wang, Libo Zhang, Hao Yan, and Shuai Han. 2023. TIM-SLR: a lightweight network for video isolated sign language recognition. Neural Comput. Appl. 35, 30 (Oct 2023), 22265–22280. https://doi.org/10.1007/s00521-023-08873-7.

[18] D. Li, C. R. Opazo, X. Yu and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 2020, pp. 1448-1458, doi: 10.1109/WACV45572.2020.9093512.

[19] G. Abdullayeva, N. Alishzade, Transfer learning for Azerbaijani Sign Language Recognition. Journal of Informatics and Control Problems. https://doi.org/10.54381/icp.2022.2.08.

[20] J. Hasanov, Nigar Alishzade, Aykhan Nazimzade, Samir Dadashzade, Toghrul Tahirov, Development of a hybrid word recognition system and dataset for the Azerbaijani Sign Language dactyl alphabet, Speech Communication, Volume 153, 2023, 102960, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2023.102960.

[21] A. Mino, M. Popa and A. Briassouli, "The Effect of Spatial and Temporal Occlusion on Word Level Sign Language Recognition," **2022 IEEE International Conference on Image Processing (ICIP)**, Bordeaux, France, 2022, pp. 2686-2690, doi: 10.1109/ICIP46576.2022.9897770