International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Sign Language Interpreter AI using machine learning algorithm

Piyush Pardesi[1], Prince kumar[2], Dr.Madhumita.K[3]

Department of Computing Technologies,
SRM Insitute of Science and Technology,
Kattankulthur,Chennai, India.
[1]pp3055@srmist.edu.in, [2]pk0201@srmist.edu.in, [3]madhumik1@srmist.edu.in

Abstract

For people who are deaf or hard of hearing, sign language interpreters are essential in removing communication barriers. Their main duty is to mediate communication between persons who use sign language and others who rely on verbal or written language. This crucial function is present in a variety of contexts, such as meetings for business, court cases, medical appointments, educational settings, and daily encounters. Sign language interpreters have a thorough knowledge of both sign language and the intended spoken or written language. They are adept at communicating not only the words but also the subtleties, feelings, and cultural context of the conversation. Thorough training, certification, and ongoing professional development are required for this skill. The value of sign language interpreters goes beyond simple translation; They serve as interpreters and fill gaps in accessibility and inclusiveness. Interpreters give deaf and hard-of-hearing people the same access to social interactions, economic possibilities, healthcare, and other chances as hearing people by facilitating successful communication. Sign language interpreters act as catalysts for a more open and just society in a world that cherishes variety and inclusivity. Their dedication to facilitating communication increases respect for one another, lessens prejudice, and builds a feeling of community for the deaf and hard-of-hearing population.

Keywords: sign language interpreters, Hard of hearing, Verbal language

1. INTRODUCTION (HEADING 1)

A sign language interpreter is a highly skilled and essential specialist who acts as a crucial connection in facilitating successful communication between those who are deaf or hard of hearing and those who do not utilize sign language. These committed people are able to bridge the gap between two different linguistic and cultural worlds because they have a certain set of abilities and knowledge.

Sign language interpreters are crucial in guaranteeing that communication is a fundamental right for everyone in a world that increasingly values inclusivity and accessibility. Their skill extends beyond basic translation; they are skilled at communicating not only the words but also the feelings, nuances, and cultural context of the message This level of proficiency requires rigorous training, certification, and a deep commitment to continuous professional development. Sign language interpreters are found in a wide range of settings, from educational institutions and healthcare facilities to legal proceedings, business meetings, and community events. They are not just conveyors of language; they are advocates for equality, empowerment, and understanding for the deaf and hard-of-hearing community.

2. RELATED WORK

To address this issue, researchers are making every effort. Numerous studies have been conducted. However, they still need to be developed in order to be a decent solution to the issue with this topic. We're here to talk about other people's research. And for the primary research to be discussed, we choose [5], [6], and [8].

2014 saw Pigou do research on the recognition of sign language. He applied his own model. He began by using CLAP14 for his dataset [3]. The CLAP14 dataset contains 20 renditions of Italian sign language made by 27 individuals while they were in various environments and moving in various ways and densities of clothing. The Microsoft Kinect platform created the dataset. In the CLAP14 development set, he used 6600 videos, 4600 for the training set, and 2000 for the validation set. He also engaged in some preprocessing. He initially clipped the upper torso and the topmost hand using the given joint information. Four results will be produced by the preprocessing, including hands, upper bodies, and both grayscale and 64x64x32 (HxWxFrame) depths. Additionally, he applied median filtering, user index background removal, and thresholding to lower the noise in depth maps.

His suggested model for a convolutional neural network has six layers, including input and output. He needs two inputs for his input layer. The shape would therefore be 2x2x64x64x32 (Grayscale, Depth, Hand, Upper Body, Height, Width, Frames). All of the kernels in his suggested model are 2D because it is not a 3D convolutional neural network. Retrodicted Linea Units (ReLU) are used by all neurons. Convolutional neural networks are divided into three layers in the suggested model, which is followed by fully linked or classical artificial neural networks. He achieved 91.70% accuracy and 8.30% error rate using his suggested methodology.

J. Huang conducted research on the same subject in 2015. He applied his suggested model, a 3D Convolutional Neural Network, nevertheless. Using Microsoft Kinect, he created his own dataset.

He made no mention of the nation or standards of any sign language. He received 25 classes of vocabulary that are frequently used in daily life. Nine signers, each signing three times, perform each word. There would be 27 videos each word, or 25x9x3 = 675 in total. For his training set, he randomly chose 18 films for each word.

Kinect records the information by concurrently capturing a colour image, He made no preparations beforehand. Let's go on to his suggested model now. His suggested model has eight layers, including an input layer and an output layer. His model's input shape is 5 x 64 x 48 x 9 (channels x height x width x frames). According to the films, he chose 9 frames that were in the middle. The 5 channels are depth, body skeleton, colour-R, colour-B, and colour-G. The model comprises four layers of 3D Convolutional Networks, all of which have 3D kernels. His average accuracy with this technique was 94.2%.

J. Carriera brought his idea to the area of action recognition from 2017 to 2018. Which is beneficial for the subject. Inception V1 [11] is now a 3D Convolutional Neural Network model, which is his innovation. Additionally, he applied pretraining, commonly referred to as transfer learning, to his studies.

His pretrained dataset included the Kinetic Dataset [7] and ImageNet [2]. Additionally, they used the two datasets UCF- 101 [12] and HMDB-51 [13] to train the pre-trained model. Numerous model combinations from the RGB model, flow model, pre-trained Kinetic, and pre-trained ImageNet are used in his research. The best accuracy is 98.0% on the UCF- 101 dataset and is 80.9% on the HMDB-51 dataset.

According to our assessment, [6] was decently incorporated into sign language and performed a great job.

As we've already discussed, Action Recognition and Sign Language have a similar trait. So, if that is appropriate for Sign Language, why is that appropriate? Then why not? After this investigation, we are hopeful that we can answer that kind of query.

3. ARCHITECTURE DIAGRAM

1. Building a database

Create an extensive database with a wide range of sign language movements to start. This database provides the artificial neural network with a vast store of visual data for learning and recognition, acting as its training environment.

2. Processing of images:

take pictures with sensors—such as cameras or depth sensors—and process the images that are captured. The raw data is prepared for further analysis using image processing techniques such as scaling, normalization, and noise removal.

3. Extraction by Hand:

Identify and extract the region of interest, which are the hands making sign gestures, by using algorithms. The efficiency of gesture recognition is increased by this stage,which isolates the pertinent components for further analysis.

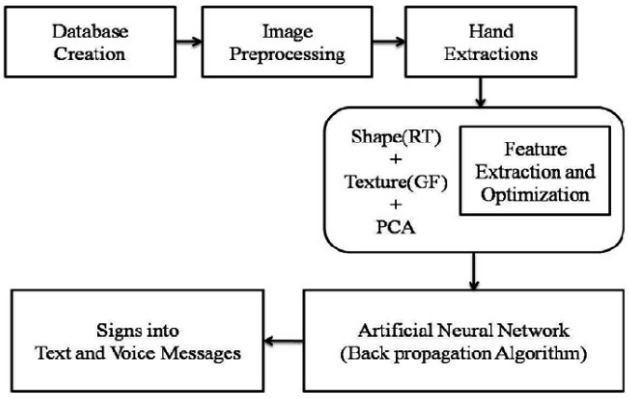


Figure III: Architecture Diagram of Autonomous Vehicles

Optimization and Feature Extraction:

Highlight important elements of the sign language gestures by extracting significant features from the previously processed photos. The efficiency and effectiveness of feature representation are improved by optimization approaches like dimensionality reduction or normalization.

Sign up for text and voice messages:

Use artificial intelligence methods to understand the retrieved features, perhaps utilizing deep learning models. Recognized sign language motions can be translated into text messages or artificial voice output

for conversation.
a synthetic neural network:

Create and train a neural network using artificial intelligence using the pre-processed data. This network—typically a convolutional or recurrent neural network—learns the correlations and patterns found in sign language gestures to enable precise translation and identification.

The process of creating a comprehensive database of sign language gestures, which serves as a foundation for training, is the first step in the architecture diagram for sign-to-text translation. Visual inputs that are captured are carefully processed, including noise reduction and normalization. Algorithms for hand extraction separate relevant components, optimizing further analysis. By distilling the important parts of gestures, feature extraction improves the effectiveness of representation. The interpretation of sign language by an artificial neural network trained on the analysed data enables the translation of movements into text or voice generated by computers. This methodical technique guarantees a smooth transition from data collection to language output. A robust system for overcoming communication barriers between sign language users and others, the architecture is a sophisticated blend of image processing, machine learning, and linguistic translation.

4. METHODS

A.
ata Collection

The gathering of data is a crucial component of this study.

Because the dataset has a big impact on the outcome. Due to this, we made use of the open-source dataset LSA64 [9], which is what you can see in Figure 1. Each word was signed five times by each signer using ten different languages. Therefore, there would be 500 videos overall..

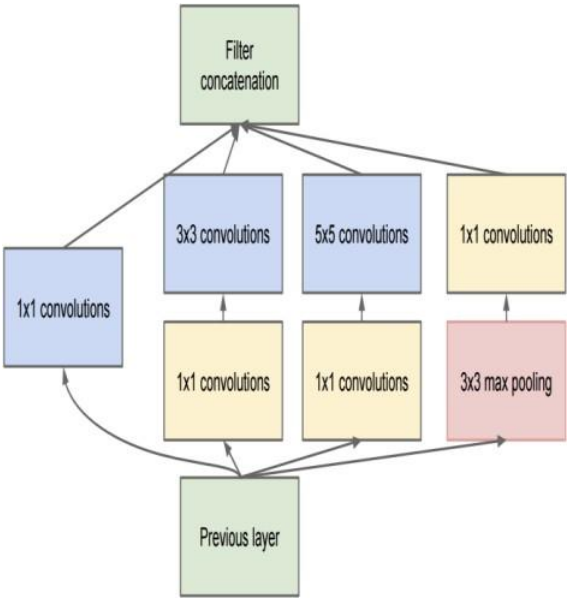
Fig. 1. LSA64 Dataset example

B. Data Processing

In the beginning, we convert the movie into a series of photos and create various numbers of frames. Following that, we added extra frames to the average frame for all videos. There are 126 frames on average. For every frame that has fewer than 126 frames, we add a blank screen that is RGB (0,0,0), and we remove any video that has more than 126 frames After converting, we used bilinear interpolation to resize the frames into 224x224 height and width. Then, pixel values were scaled from -1 to 1, normalizing them.
So, (1, 126, 224, 224, 3) would be the numpy form. Using the h5py library, the numpy were being stored in an HDF5 file.

C. 3D Convolutional Neural Network Architecture

This process is referred to as i3d inception or inflated inception for the 3D CNN architecture we utilized [6]. because the inception v1 model [11] is the foundation of this concept. A 3D CNN version of Inception v1 was being developed. This architecture was chosen because it can enhance the findings of earlier studies that used ResNet-50 models [4], C3D Ensamble [14], and Two-Stream Fusion + IDT [15].
67 convolutional layers, including input and output, make up I3D Inception. There are 9 modules for



conception. The specifics of Fig. 2 displays the inception module. We divide the dataset into a 6:2:2 ratio for training. The training set would consist of 300 videos, the validation set of 100 videos, and the testing set of 100 videos. We give each signer one of three videos for the training set. Then there will be 300 videos, or 3 x 10 x 10 (words). One word multiplied by ten words gives 100 videos for validation. The testing set isidentical to the validation set, but it contains unique videos rather than duplicates.

Fig. 2. Details of inception module

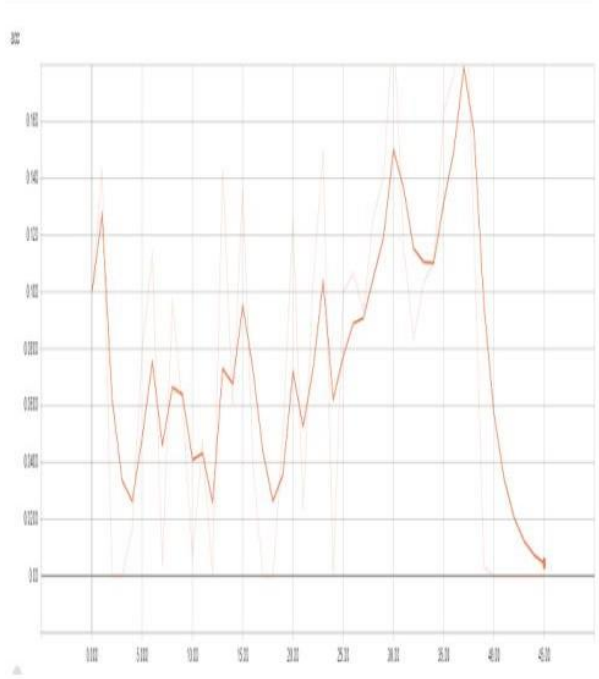


Fig. 3. 10 Signer with 10 classes (500 videos) result.

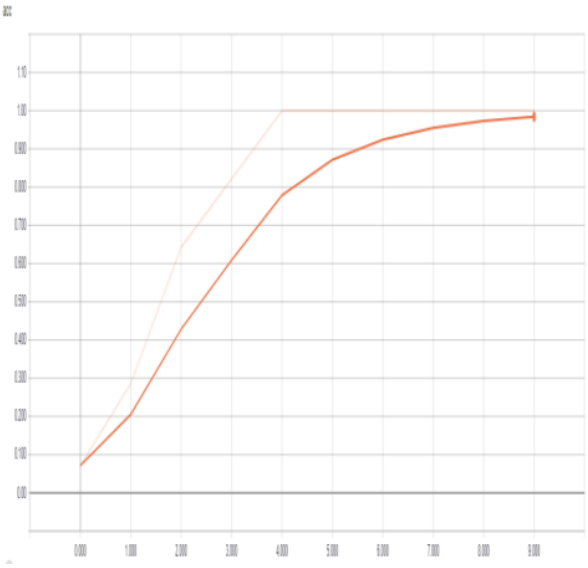


Fig. 4. 1 Signer with 10 classes (50 videos) result.

TABLE I. THE COMPARISON OF TESTING RESULT FOR EACH TRAINING

Signer	Classes	Testing Accuracy
1 Signer	*10 Classes*	100%
2 Signer	*20 classes*	75%
2 Signers	*10 classes*	25%
4 Signers	*40 classes*	0%

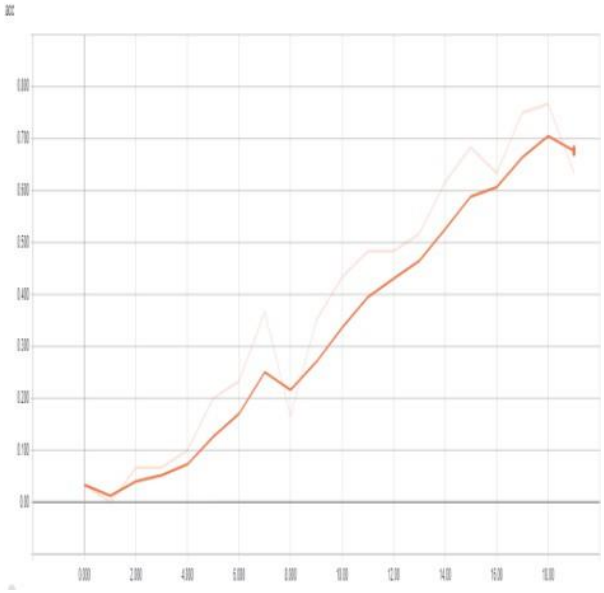
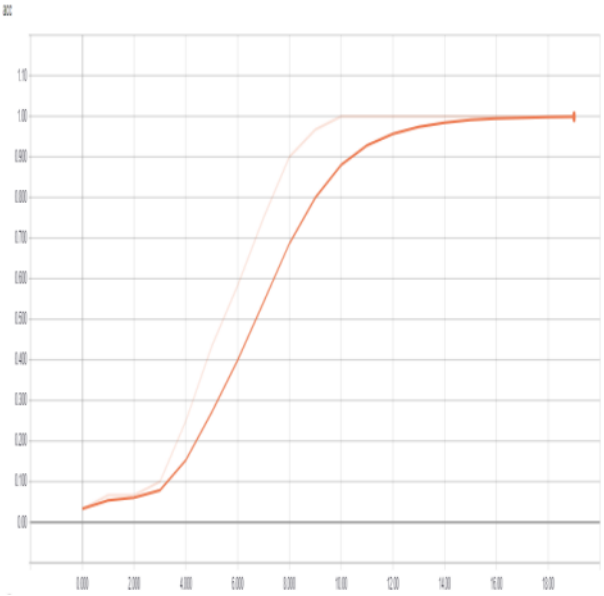


Fig. 5. 2 Signer with 10 classes (100 videos) result.



5. RESULT AND DISCUSSION

The first training had poor results; with 45 epochs, the training accuracy ranged from 0.00% to 20.00%. Because we believed something went wrong, we stopped that.

Then, with 10 students, we tried using a single signer. 500 videos were used because of 10 (words) multiplied by 5 (executed) and 10 (signers). and achieved accuracy of 100.00%, as illustrated in Figs. 3, 4.

These results led us to believe that our model was overfitting. It acquired too much knowledge before it could even classify the signer. Then we want to confirm our hypothesis. So, With a modified dataset structure, we retrained. Currently, we used 2 signers for 10 classes with 100 videos overall, 2 signers for 20 classes with 100 videos total, and 4 signers for 40 classes with 200 films total. The training accuracy

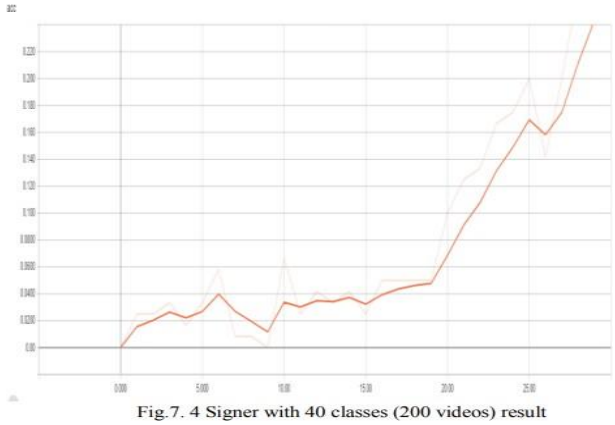


Fig.7. 4 Signer with 40 classes (200 videos) result

for the two signers was between 50.00% and 80.00% for the first ten classes, 100.00% for the following ten courses, and 20.00% for the following twenty classes (see Figs. 5, 6, 7).Fig. 6. 2 Signer with 20 classes (100 videos) result.

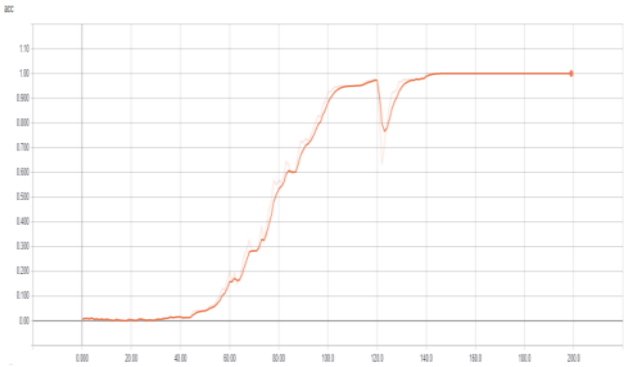


Fig. 8. 10 Signer with 100 classes (500 videos) training accuracy.

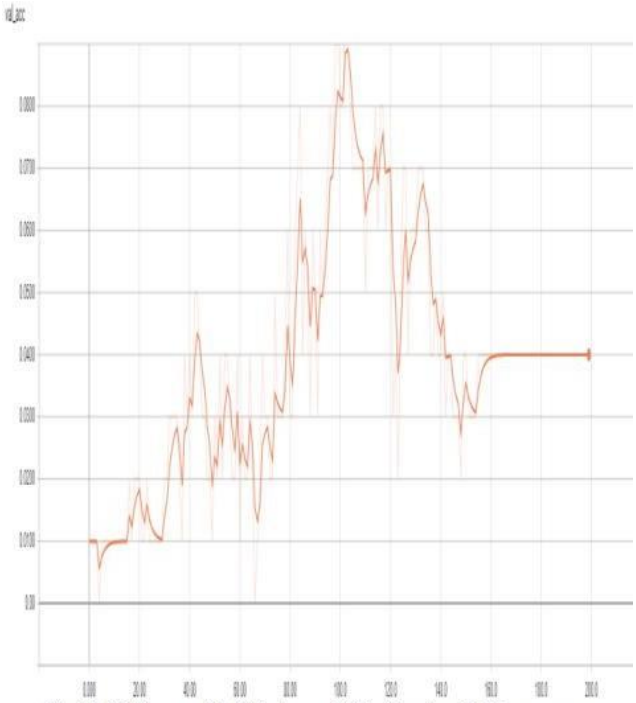


Fig. 9. 10 Signer with 100 classes (500 videos) validation accuracy.

CONCLUSION AND FUTURE WORK

We trained the model using 10 signers and 100 classes (500 films) with 200 epochs after numerous trainings because, in our opinion, i3d inception without modification is too overfit due to the results. The model has a good training accuracy but a very low validation accuracy. With this model, we are capable of a great deal more, including layer freezing, the removal of some inception modules, the removal of transfer learning, and the transformation of the completely linked layer into a different deep learning model. In our According to some, the completely connected layer is not really the issue. Since the convolutional neural network layer is the detector, we believe that we should pay greater attention to it.
The dataset itself may have contributed to these outcomes in a second way. The overfitting may be due to LSA64's different background lightning. Considering that the variations in background lighting might be a feature for machine learning. As a result, machine learning would be trained on the incorrect feature.

ACKNOWLEDGMENT

his research was supported by Research Grant No. 036/CR.RTT/IV/2018from Ministry of Research Technology and Higher Education of Republic of Indonesia.

Fig VI.2

A crucial part of creating and testing autonomous vehicles is traffic simulation. It entails building virtual worlds that replicate traffic patterns and human interactions in the real world. These simulations are crucial for the advancement and verification of autonomous car technology in a number of ways.

1) Visualizing Neural Network

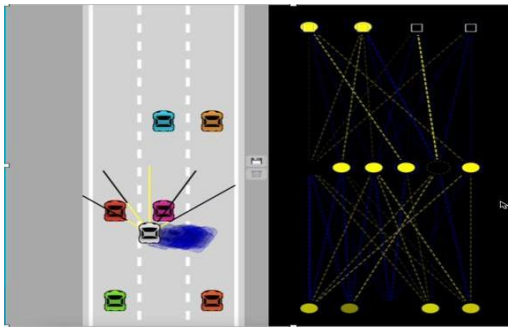


Fig VI.7

The car successfully overcomes all the hindrances in its way.

REFERENCES

1. L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Sign language recognition using convolutional neural networks," in Proc. Europea Conference on Computer Vision (ECCV) Workshops, 2014, pp. 572–578.
2. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
3. J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks," in Proc. IEEE International Conference on Multimedia and Expo (ICME), 2015, pp. 1–6.
4. J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4724–4733.
5. W. Kay et al., "The Kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.
6. M. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Neural sign language translation," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018,
7. pp. 7784–7793.
8. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
9. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei- Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
10. L. von Agris, D. Schneider, and K. Kraiss, "The significance of facial features for automatic sign language recognition," in Proc. IEEE 8th International Conference on Automatic Face & Gesture Recognition, 2008, pp. 1–6.
11. C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," Computer Vision and Image Understanding, vol. 114, no. 8, pp. 943– 955, 2014.
12. C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.