Headcount of the Crowd in a Congested Scene

Mayur Nair[1], Raghav Vasudeva[2], Aman Singh[3], Rana Gill[4]

[1,2,3,4] Chandigarh University, Gharuan, Mohali, Punjab 140413

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract— Crowd Counting and estimation of density is really challenging and an important problem if we visually analyze the crowd. Crowd Monitoring and Analyzing Crowd behavior has been an important aspect for every research field. A lot of already existing approaches use techniques based on regression on heat maps(density) to count people present in from a single frame. These techniques however cannot restrain an individual walking and further cannot approximate the original distribution of pedestrian in the locality. Whereas, detection-based techniques detect and restrain walking men's in the frame, but the efficiency of these techniques challenged when implemented in high-density crowd situations. To get the better of the limitations of above-mentioned problem, we have used the (Congested Scene Recognition) Neural Network. By using this type of Neural network, we are able to visualize the detection and form density map according to produce accurate outputs for the given scene. The experimental outcomes of the successfully showcases the effectiveness of the approach used.

Keywords— Deep Learning, Neural Network, CNN, Semantic Segmentation

1. Introduction

H OW much time would it take us to count down the number of people in the crowd? It happens to the best of us and remains a frustrating experience. But what if a computer visual model could keep a count of people in a matter of milliseconds?

This is what Semantic Segmentation [1] are capable of. While this is a simple example, there are multiple other applications of such algorithm.

So how Semantic Segmentation works? It uses image processing: it is any form of processing of information where the input is an image, such as frames from video-clips. In this process the output are features which could give information to depict an image [2], [3]. This feature extracted is particularly useful in this scenario where the density of the crowd may be high. The extracted feature is basically region or a structure. And using those boundaries / regions we split out image into labelled pieces. And these labelled pieces when combined forms an object. So, when we are talking about an object, we first look for the required parameters that appear in the image in some order [4].

An Input image when aggregated alongside its filter bank produce a texton map as a result After clustering. We shall be using Deep Learning in Semantic Segmentation here. So the model shall basically establish a relationship between the pixels.

1. Pixels having Same Color are more likely to have same label
2. Nearby pixels are more likely to have same label
3. The pixel amount pixels of "stem" are more likely to be leaves.

However, when analyzing a congested scene or a scene which has high density of object surrounding, using semantic segmentation isn't enough. This is where we make use of some preexisting models, such as:

i. Approach based on Detection
ii. Approach based on Regression
iii. Approach based on Density approximation
iv. CNN-based Approach
v. Dilated Residual Network
vi. Congested Scene Recognition Network(CSRNet)

A detailed explanation for each of these approaches is given below. Methods involving Neural Networks are often considered better because they have work upon training and testing [5]. As a result, these models tend to become better with time with more level of understanding for the same. Potential problems such as downscaling and decreasing resolution are also not an issue in such models [6], [7].

But they come with their own problems as well. Traditional approaches are more based on multi-scale architecture . There is no doubt in the fact that they provide high performance but the model they use also comes with two big disadvantages: Time for training is high and non-efficient branching based structure (e.g. Multi-column CNN).

Here, in this paper, we have used a deep network known as CSRNet for our application. It consists of convolutional layers for supporting element for incoming images with flexible images.

2. Problem Description

Big Events, Big Shows all attract Hugh Crowd. In such cases, counting the number of attendees can become overwhelmingly difficult. But keeping accurate records of the same is important as well. This helps the corporate to know which kind of events and products has the potential to attract more people and more potential customers[8].

There are other potential applications too. This model can be used to understand the traffic pattern in a locality and can be tinkered as needed. Therefore, this can be proven useful for Monitoring High-traffic area.

But the potential problem here becomes the high-density crowd image. Extracting all the parameter from potentially a very small part in the image can be a very difficult job[7]. We require a model that is able to Extract all minimum required parameters from the image and at the same keep the original resolution of the image intact. Our model should also be capable of doing all these tasks in as little time as possible to keep the whole experience smooth.

3 Literature survey

Semantic Segmentation remains a challenge because due to the pixel-level accuracy with multi-scale contextual reasoning it requires. However in recent times High accuracy gains in this type of segmentation have been obtained, thanks to Convolutional network that is trained by backpropagation. Specifically, shows us that the convolutional network architecture we used for traditional image processing and classification could be used for dense predictions as well [9]–[11]. Modern day classification models for images integrate multi-scale contextual information. Pooling and subsampling for prediction is what makes it possible. But all these things have to happen alongside full-resolution output. And this clearly is a conflicting demand. Recent works guide us through two approaches for the same. First being, Repeated Up-convolution to bring back the original resolution of the image while carrying the result extracted from down sampled layer. Second approach being, developing a dilated convolutional network that combines multi-scale contextual information of the input image without any loss in resolution or analyzing the rescaled image.

There are various approaches that can be taken for specifically out application, that is:

i. Detection-Based Approach

The solution to this problem was a detection-based approach using a retractable window, where it detected people and raised the count by 1. But this was not accurate as it seems because it required a well-trained network and feature extraction in a densely populated scene isn't as easy as it seems[12]. So, to overcome this, Researchers started detecting particular body part instead of the whole body.

ii. Regression-Based Approach

Since Detection-based Approach wasn't an ideal solution to a congested scene. This was when researchers tried to deploy approach based on regression to establish the relation among the extracted information in a cropped patch, and then calculate the count of the particulars. Other features are also used to generate low-level informatics [13]. Following the same approached models were proposed where Fourier Analysis and Scene Invariant Feature transform where made in use

iii. Density estimation-based Approach

A key feature in Image Extraction called Saliency was overlooked while using Regression-based solution. And this in turn can cause inaccurate results. Duan [14] proposed a solution to this problem by understanding the linear mapping in local as well as its subject density. Song [15] deployed Random Forest Regression technique to a non-linear map unlike Lempitsky.

iv. Convolution Neural Network Approach

This model is considered to be the best and most accurate model for congested regions. CNN refers to Convolutional Neural Network. In a CNN based model [16], [17]we built a regression model that is end-to-end. This is a hassle freeway. Rather than working on patches of cropped image, This uses the entire image as an input & generated a head count output directly. CSRNet, which we shall use, also deploys deep CNN for high-level of extraction(feature) and generating highly accurate density maps.

v. Dilated Residual Network

The main motto after considering Detection-Base Approach was to not change the spatial resolution in the convolutional network, for image classifier to work. Although down sampling the image isn't a bad option and it has been pretty successful as well, it can be proven harmful when detecting natural images. Natural images have high frequency count for objects and this if often important for understanding the scene [18]. It becomes difficult to understand the scene when the dominant object doesn't stand-out in the processed image. And making matters

even worse, if in the course of transmitting the image from one channel to the other, a loss in signal can affect the image. And as a result, recovering is difficult during training period. But, if somehow the spatial resolution is maintained throughout, backpropagation can be useful to analyze and learn to preserve important information for the less apparent subjects present in image. The Resnet and DRN are shown in figure 1.

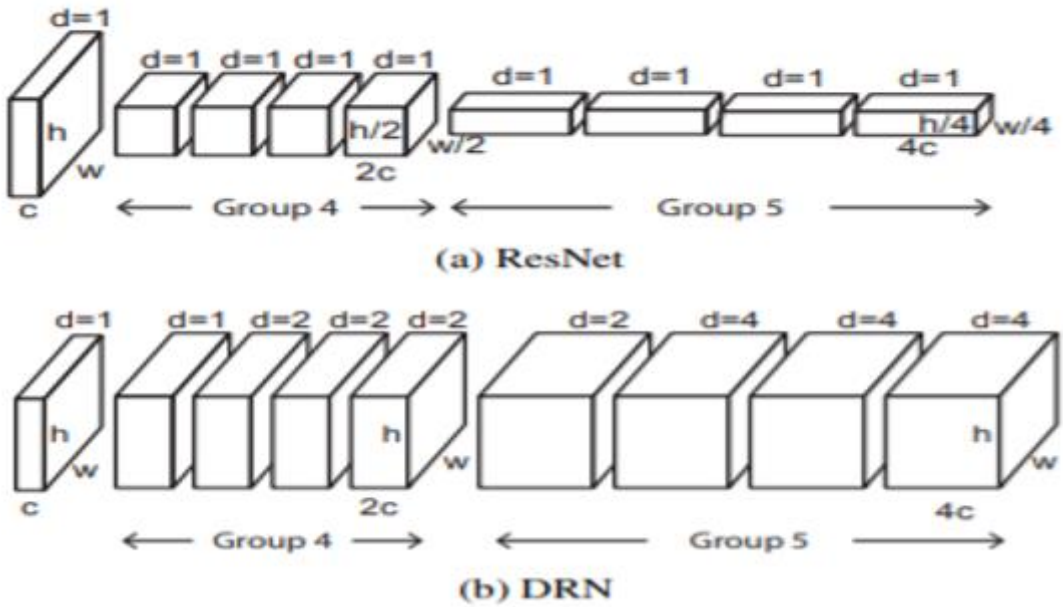


Figure 1. ResNet vs DRN

A Dilated Residual Network [19], [20] outperforms the non-DRN without generating extra complexity to the model. The DRN has also been successful in removing the artifacts that was introduced by the dilation. And as a result, the performance of this model significantly increased. There is further increase in accuracy as compared to downstream applications. The figure below shows a comparison between traditional method and DRN respectively.

vi. Congested Scene Recognition Network (CSRNet)

A Congested Scene Recognition Network [21], [22] also known as CSRNet is a data-driven DL method that is capable of understanding highly congested scenes and is capable of accurately count desired objects in the scene. This is mainly composed of two important and fundamental components: A Convolutional Neural Network for extraction(features) and for back-end, for backend operations. The MAE for CSRNet was observed to be 47% less than that of existing approaches.

Limitations of Existing approaches

Switch CNN [23] that was proposed by Deepak Babu Sam, uses a density level classifier to select different regressors for particular input patches. These both use Multi-column-based architecture. There is no doubt in the fact that they have developed a high performing design but then it also comes with two big disadvantages: Time for training is high and non-efficient branching based structure. Multi-column CNNs also come embedded with redundant structures. The next disadvantage for both the above-mentioned solution is the fact that it required density level classifier before sending pictures to MCNN[24]–[26].

4 Implementation

We shall be using CSRNet here. CSRNet stands for Congested Scene Recognition Neural Network. This helps in capturing the low as well as high-level feature. It captures the high-level semantics needed for crowd counting. The Model proposed is a plug and play model making it easier to employ into potential applications[27], [28].

Using the Ground Truth Value, we generate an original image and density map as shown in figure 2 and figure 3. Ground truth value is the theoretical value which is collected on site/actual location which is further used to check the machines implemented results for accuracy against the real-world data.

Figure 2. Original Image

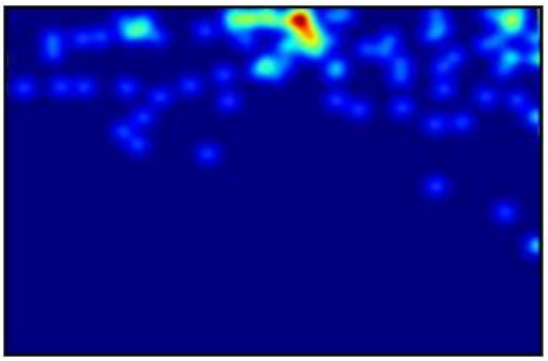


Figure 3. Density Map for the Original Image

Now, after getting the crowded images from various locations/sites as an input and observed ground truth value using visual analysis, we train our Data Model using MATLAB to label data, so that after machine implementation, the results which we get can be compared with the ground truth values to get more accurate results as an output[29].

So, after studying and analyzing the scene in the image the output for the same shall look something like the figure given below. The given figure shows the visualization of the detection result. The green squares in the scene depicts the required objects for count. It can be observed that even in places where the count of people is high in a particular part in an image, analyzation is still successfully done.

After the detection process is complete, a density map is designed for the some using the CNN regression network. The result for the same shall look something like in figure (v). The density map can be read in the same way you read a heatmap. The map turns red (hotter) where the density increases and turns blue (colder) where the density is less as shown in figure 4 and figure 5.

Figure 4. Density Map for the Original Image

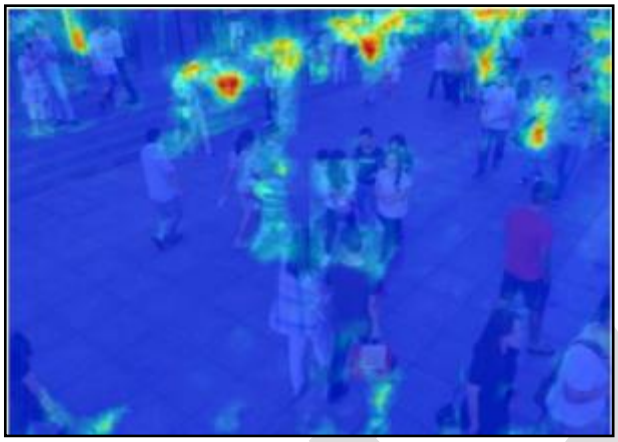
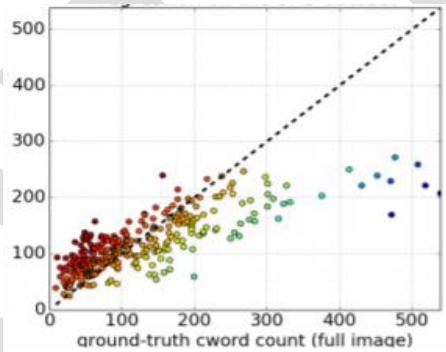


Figure 5. Density Map for the detected objects

The Ground Truth Count for the scene shall look something like the figure 6 mentioned below.

Figure 6. Density Map for the detected objects



.

2. Results

Original Count: 457

To find the accuracy of the system we shall calculate the Mean Absolute Error [MEA]. Mean Absolute error basically tells us the variation in difference between two variables. The MEA for the proposed system is derived as shown in figure 7 and figure 8.

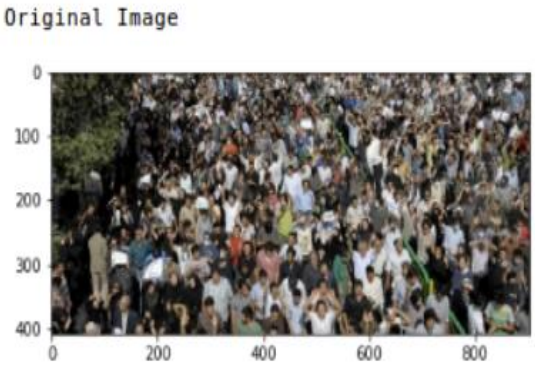


Figure 7. Original Input Image

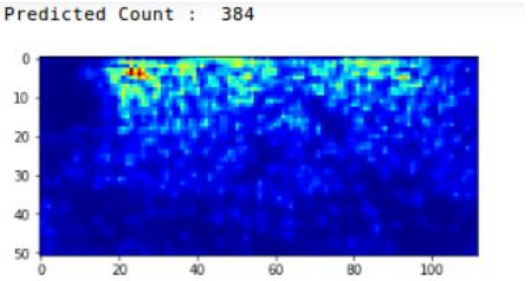


Figure 8. Density Map to predict Count

MEAN ABSOLUTE ERROR:

$$\frac{\sum_{i=0}^{n} \left| AP - PP \right|}{\sum_{i=0}^{n} |AP|} = 15.9\%$$

Where,
AP = Actual Population
PP=Predicted Population
n= number of training sets

The Mean Absolute Error comes out to be 15.9% error or 84.1% accuracy.

3. Conclusion

In this paper, we have successfully implemented a head counting model using CSRNet for a congested cr owded scene. We were able to generate highly accurate density map and connect them to the original image to estimate the crowd count[30]–[34]. Dilated Convolution Model was used as well to generate better results without losing information. This model can be extended to be used with trees and vehicles based on the training. This can be highly useful for detecting construction sites in a highly congested satellite map as well.

References

1. L. Deng, M. Yang, Y. Qian, C. Wang, and B. Wang, "CNN based semantic segmentation for urban traffic scenes using fisheye camera," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2017, pp. 231–236.
2. O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother, "Analyzing modular CNN architectures for joint depth prediction and semantic segmentation," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2017, pp. 4620–4627.
3. C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10111 LNCS, 2017.
4. A. Sáez, L. M. Bergasa, E. Romeral, E. López, R. Barea, and R. Sanz, "CNN-based Fisheye Image Real-Time Semantic Segmentation," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2018, vol. 2018–June, pp. 1039–1044.
5. G. Sharma, S. Sharma, and S. Gujral, "A Novel Way of Assessing Software Bug Severity Using Dictionary of Critical Terms," in *Procedia Computer Science*, 2015, vol. 70, pp. 632–639.
6. M. Kaur and V. Wasson, "ROI Based Medical Image Compression for Telemedicine Application," in *Procedia Computer Science*, 2015, vol. 70, pp. 579–585.
7. A. Gupta, D. Singh, and M. Kaur, "An efficient image encryption using non-dominated sorting genetic algorithm-III based 4-D chaotic maps Image encryption," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 3, SI, pp. 1309–1324, Mar. 2020.
8. M. K. Gupta *et al.*, "Parametric optimization and process capability analysis for machining of nickel-based superalloy," *Int. J. Adv. Manuf. Technol.*, vol. 102, no. 9–12, pp. 3995–4009, Jun. 2019.
9. M. Ravanbakhsh, H. Mousavi, M. Nabi, M. Rastegari, and C. Regazzoni, "CNN-aware binary MAP for general semantic segmentation," in *Proceedings - International Conference on Image Processing, ICIP*, 2016, vol. 2016–August, pp. 1923–1927.

10. X. Xu, G. Li, G. Xie, J. Ren, and X. Xie, "Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions," *Complexity*, vol. 2019, 2019.
11. R. Yasrab, N. Gu, and X. Zhang, "SCNet: A simplified encoder-decoder CNN for semantic segmentation," in *Proceedings of 2016 5th International Conference on Computer Science and Network Technology, ICCSNT 2016*, 2017, pp. 785–789.
12. A. Pemasiri, D. Ahmedt-Aristizabal, K. Nguyen, S. Sridharan, S. Dionisio, and C. Fookes, "Semantic segmentation of hands in multimodal images: A region new-based CNN approach," in *Proceedings - International Symposium on Biomedical Imaging*, 2019, vol. 2019–April, pp. 819–823.
13. K. Nguyen, C. Fookes, A. Ross, and S. Sridharan, "Iris Recognition with Off-the-Shelf CNN Features: A Deep Learning Perspective," *IEEE Access*, vol. 6, pp. 18848–18855, 2017.
14. M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning CNN–ELM for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, 2018.
15. J. Xie, C. Liu, Y.-C. Liang, and J. Fang, "Activity Pattern Aware Spectrum Sensing: A CNN-Based Deep Learning Approach," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 1025–1028, 2019.
16. M. Abdullah, M. Hadzikadicy, and S. Shaikhz, "SEDAT: Sentiment and Emotion Detection in Arabic Text Using CNN-LSTM Deep Learning," in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 2019, pp. 835–840.
17. J. Wang *et al.*, "CD-CNN: A partially supervised cross-domain deep learning model for urban resident recognition," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018, pp. 192–199.
18. B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, and A. Sharma, "Image denoising review: From classical to state-of-the-art approaches," *Inf. FUSION*, vol. 55, pp. 220–244, Mar. 2020.
19. H. Fu, H. Ma, G. Wang, X. Zhang, and Y. Zhang, "MCFF-CNN: Multiscale comprehensive feature fusion convolutional neural network for vehicle color recognition based on residual learning," *Neurocomputing*, vol. 395, pp. 178–187, 2020.
20. M. Aqqa and S. K. Shah, "CAR-CNN: A deep residual convolutional neural network for compression artifact removal in video surveillance systems," in *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2020, vol. 4, pp. 569–575.
21. H. Phan, P. Koch, L. Hertel, M. Maass, R. Mazur, and A. Mertins, "CNN-LTE: A class of 1-X pooling convolutional neural networks on label tree embeddings for audio scene classification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 136–140.
22. R. Yasrab, "ECRU: An encoder-decoder based convolution neural network (CNN) for road-scene understanding," *J. Imaging*, vol. 4, no. 10, 2018.
23. H. Zhenlong, Z. Qiang, and W. Jun, "The prediction model of air-jet texturing Yarn intensity based on the CNN-BP neural network," in *2018 3rd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2018*, 2018, pp. 116–119.
24. Y. Zhang, X. Zhang, J. Song, Y. Wang, R. Huang, and R. Wang, "Parallel Convolutional Neural Network (CNN) Accelerators Based on Stochastic Computing," in *IEEE Workshop on Signal Processing Systems, SiPS: Design and Implementation*, 2019, vol. 2019–October, pp. 19–24.
25. J. Gu, A. He, and X. Tian, "RC-CNN: Representation-consistent convolutional neural networks for achieving transformation invariance," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2019, vol. 2019–October, pp. 1588–1595.
26. .M. Y. Fikri *et al.*, "Clustering green openspace using UAV (Unmanned Aerial Vehicle) with CNN (Convolutional Neural Network)," in *Proceeding - 2019 International Symposium on Electronics and Smart Devices, ISESD 2019*, 2019.
27. M. Geese, R. Tetzlaff, D. Carl, A. Blug, H. Höfler, and F. Abt, "High-speed visual control of laser welding processes by Cellular Neural Networks (CNN)," in *Proceedings of the IEEE International Workshop on Cellular Neural Networks and their Applications*, 2008, p. 9.
28. M. Anguita, F. J. Fernández, A. F. Díaz, A. Cañas, and F. J. Pelayo, "Parameter configurations for hole extraction in cellular neural networks (CNN)," *Analog Integr. Circuits Signal Process.*, vol. 32, no. 2, pp. 149–155, 2002.
29. F. Hardalac, H. Yaşar, A. Akyel, and U. Kutbay, "A novel comparative study using multi-resolution transforms and convolutional neural network (CNN) for contactless palm print verification and identification," *Multimed. Tools Appl.*, vol. 79, no. 31–32, pp. 22929–22963, 2020.
30. W. Li, K. Liu, L. Yan, F. Cheng, Y. Q. Lv, and L. Z. Zhang, "FRD-CNN: Object detection based on small-scale convolutional neural networks and feature reuse," *Sci. Rep.*, vol. 9, no. 1, 2019.
31. H. Hasan, H. Z. M. Shafri, and M. Habshi, "A Comparison between Support Vector Machine (SVM) and Convolutional Neural Network (CNN) Models for Hyperspectral Image Classification," in *IOP Conference Series: Earth and Environmental Science*, 2019, vol. 357, no. 1.

32. P. Bagave, J. Linssen, W. Teeuw, J. K. Brinke, and N. Meratnia, “Channel state information (CSI) analysis for predictive maintenance using convolutional neural network (CNN),” in *DATA 2019 - Proceedings of the 2nd ACM Workshop on Data Acquisition To Analysis, Part of SenSys 2019*, 2019, pp. 51–56.
33. R. Assaf, I. Giurgiu, F. Bagehorn, and A. Schumann, “MTEX-CNN: Multivariate time series explanations for predictions with convolutional neural networks,” in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2019, vol. 2019–November, pp. 952–957.
34. K.-J. Wang and C. Y. Zheng, “Toward a Wearable Affective Robot That Detects Human Emotions from Brain Signals by Using Deep Multi-Spectrogram Convolutional Neural Networks (Deep MS-CNN),” in *2019 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019*, 2019.