IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 June 2024, accepted 21 July 2024, date of publication 1 August 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3436592

RESEARCH ARTICLE

FMSNet: A Multi-Stream CNN for Multi-Stereo Image Classification by Feature Map Sharing

FERIT CAN[1] **AND CAN EYUPOGLU**[2]
[1]Department of Computer Engineering, Atatürk Strategic Studies and Graduate Institute, National Defense University, 34334 Istanbul, Türkiye
[2]Department of Computer Engineering, Turkish Air Force Academy, National Defense University, 34149 Istanbul, Türkiye

Corresponding author: Ferit Can (feritcan@gmail.com)

ABSTRACT Convolutional Neural Networks (CNNs) have achieved significant success in image classification and object detection. CNN models generally consist of a single-stream and process single image data at once. In addition, multi-stream (or multi-modal) models have recently begun to be proposed that allow the processing of more than one input at the same time. The data can be an image, video, voice, or any other sensor data. Multi-modality may help us extract some hidden features of the same object. Furthermore, several new studies examine sharing feature maps between different streams of the same CNN. However, systematic studies that can adequately demonstrate the contribution of multi-modality and feature map sharing features to performance have not yet been conducted. Processing power and lack of available datasets are among the important factors that negatively affect progress. In this study, the contributions of multi-modality and feature map sharing (FMS) to increase the performance in object recognition are examined in detail. For this purpose, a new dataset and a new multi-modal multi-feature map sharing CNN model, which we call FMSNet, are developed. The proposed model achieved a 3.06% higher accuracy rate than its non-FMS counterpart, DenseNet-201, exceeding most of the state-of-the-art single-stream CNN models.

INDEX TERMS Artificial intelligence, convolutional neural networks, feature map sharing, four-stream, multi-modal, image classification.

I. INTRODUCTION

Computer vision deals with the problem of gaining meaningful information out of image data. Although this seems like a very easy problem to solve for humans, it is quite difficult for computers since all they can see and interpret are the numbers that represent the colors.

Image classification is considered as the main task of computer vision. To perform this task, image data are input, analyzed, and one of the predefined classes is decided by a classification algorithm. Other tasks are ‘‘semantic segmentation’’, ‘‘image classification and localization’’, ‘‘object detection’’ and ‘‘instance segmentation’’. Whereas ‘‘image classification and localization’’ deals with a single object; ‘‘object detection’’ and ‘‘image segmentation’’ tasks aim to find multiple objects in an image data. An improvement in the main task, which is ‘‘image classification’’, means an improvement in all other tasks.

Studies in the field of computer vision date back to the 1950s, when the first Convolutional Neural Network (CNN) algorithm was designed [1] and gained great momentum in 2012 with another CNN model with success in a well-known benchmark competition [2]. In the following years, many other CNN algorithms were designed in a similar fashion, trying to improve the performance in terms of speed and accuracy. These algorithms are typically designed by stacking convolutional layers that perform some specific tasks.

After reaching a certain limit of accuracy, the concept of ‘‘ensemble’’ emerged to further increase performance. With this method, different CNN algorithms are trained separately with the same or different datasets. To obtain results or inferences, the test data are input to each trained CNN separately. The results or inferences of these CNNs are compared either by averaging out or selecting the maximum value among all.

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

This method provides approximately +2/3% accuracy and is still widely used.

Our senses enable us to perceive the environment in which we live. It is well known that these senses are taste, sight, touch, smell, and hearing. We obtain information about the environment by evaluating the data received from these senses. All of these data influence our rational processes, whether we get them from one or more senses.

Even the absence of data provides information. By combining the data we perceive with our different senses, we obtain more information about the object. The approach of training CNN algorithms on only one source of data is prevalent, but, based on this idea, designing CNN architectures using the concept of multi-modality emerged.

In general, modality is every means that provides information about the environment. Therefore, a research problem or method is characterized as multimodal when it involves more than one such source of information or data.

As illustrated in Fig. 1, multi-modal models incorporate more than one data source, and each source can make use of (same or) different data types, such as text, image, voice, or video at the same time. These data are semantically in correlation and can provide complementary information to each other [3]. The multi-modality enables CNN to extract hidden features, thus improving performance dramatically.

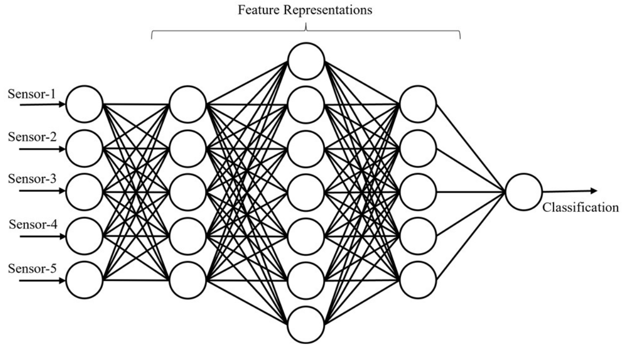

FIGURE 1. Multi-modal architecture.

We can perceive 3D depth by our brain processing two images received from our eyes. Each eye has a different angle of viewpoint, which makes 3D perception possible. There are studies focused on computing the depth information of a scene by using images from two different viewpoints. This is called the stereo vision problem, and can be achieved by having two cameras located at a pre-defined distance from each other. Disparity mapping calculations are performed, and effective results are obtained [4]. If more than two cameras can be utilized to compute depth information, then it is called multi-view stereo. But why do we not use this relation to separate one object from other objects?

In this context, our contributions are mainly three-fold.

- First, we created a multi-stereo dataset systematically by using five cameras, the images of which overlap to enable the CNN to extract hidden features. There are many multi-view stereo datasets in the literature, but these include images either taken randomly or not overlapping in a systematic manner. Our dataset enables users to evaluate the contribution of multi-modality and FMS to a CNN model.
- The second contribution is the multi-modal CNN architecture we developed, which can extract features (like depth, edges, corners, color intensity differences, etc.) by processing multi-stereo images and sharing the features inside the CNN structure itself. We assume that more hidden features can be extracted by using the correlations of all 4 images of the same object.
- The last contribution is putting forward how the feature map sharing itself contributes to the learning process. For this purpose, we utilized a well-known CNN model, which is DenseNet-201, as our backbone algorithm to develop FMSNET. We trained the DenseNet-201 and FMSNET separately, which could then enable us to compare and present how adding FMS feature actually affects the overall learning process and so the accuracy.

Experimental results show the proposed CNN achieved a 3.06% higher accuracy rate than its non-FMS counterpart, DenseNet-201, and exceeded most of the state-of-the-art single-stream CNN models.

The rest of the paper is organized as follows. In Section II, we review the related studies on multi-modality and multi-modal CNN models. In Section III, we introduce the multi-stereo dataset we created. In Section IV, the testing environment is presented. In Section V, the proposed CNN model is shared in detail. The performance is evaluated and compared to state-of-the-art performance in Section VI. Finally, conclusions and future works are summarized in Section VII.

II. RELATED WORK

Applying multi-modality to a CNN model is not a new concept. In [5], a two-stream CNN model is proposed that incorporates spatial and temporal networks for action recognition in video. Each stream is implemented using a CNN architecture. They fused these two separate streams after the last softmax layers by combining the scores using either averaging or a linear SVM. They decompose video data into spatial (single frame) and temporal components (multi-frame optical flow) to input into these separate CNN streams. They outperform all existing deep architectures at the time without implementing FMS.

Feichtenhofer et al. [6] proposed a similar two-stream CNN model for action recognition in video data. They investigated how and where to fuse two different CNN streams. For spatial fusion, sum, max, concatenation, conv, and bilinear fusion methods are implemented and tested. These methods can be applied at different points in the network, namely early-fusion, late-fusion, or multiple-layer fusion. This is one of the very early studies that discusses the FMS. Video data is used, as in previous study, to train both CNN streams.

For automated categorization of Age-Related Macular Degeneration (AMD), a two-stream CNN is proposed in [7].

TABLE 1. Number of images in each class per camera.

Class Dataset	Automobile	Cat	Dog	Bus	Tree	Bicycle	Building	Traffic Lights
Training (80%)	529	493	479	688	517	494	554	492
Validation (10%)	66	62	60	86	65	62	69	61
Test (10%)	66	62	60	86	65	62	69	61
TOTAL	661	617	599	860	647	618	692	614

An ophthalmologist uses color fundus photography (CFP) and Optical Coherence Tomography (OCT) for a diagnosis.

These two types of images are input to each stream of the proposed CNN model, which is adopting the ResNet-18 [8] architecture (pretrained on ImageNet [9]) as its backbone. The two streams are fused at the last fully connected layer, making a four-class prediction (probability of the eye being normal, dryAMD, Polypoidal Choroidal Vasculopathy (PCV) or wetAMD). They outperform the prior AMD categorization methods.

There is no doubt that the use of images taken in different modes from the same viewing angle in a multi-model CNN architecture helps to obtain better feature maps by providing more information.

Multi-modal CNN models have made progress in many other areas. For traffic speed prediction, Ke et al. [10] implemented a two-stream CNN architecture where a multi-channel speed matrix and multi-channel volume matrix are input. The streams are flattened and concatenated into one speed-volume vector and passed to fully connected layers.

Guo et al. [11] proposed a four-stream CNN model for improving glioma classification accuracy by using MRI images of four modalities. Streams are fused at the inferring stage by implementing element-wise addition of tensors. Each CNN stream is designed based on the DenseNet model proposed in [12]. The DenseNet model connects each layer to every other layer in a feed-forward fashion. In this model, for each layer, the feature maps of all preceding layers are used as inputs, and its own feature maps are input into all subsequent layers.

Jo and Kwak [13] proposed a novel four-stream model of Bidirectional Long Short-Term Memory (Bi-LSTM) and CNN for the diagnosis of depression from audio and text information. Late fusion is performed on the softmax scores of the four CNN streams to diagnose depression.

These previous studies benefit from multi-modality, which enables the CNN to extract more hidden features and finally increase the classification accuracy. However, they fuse the network at only the inferring stage, which is called late fusion, or only at the beginning, which is called early fusion, as discussed before. There was no research to find out the effects of multiple FMS and, therefore multiple fusions between the streams.

The most closely related study to ours is the two-stream CNN architecture proposed in [14], which is called HyperDense-Net. HyperDense-Net extends the DenseNet architecture one step forward by sharing the feature maps multiple times, not only in the same stream but also in the other stream. Therefore, the CNN can learn more complex correlations or features between the modalities at all levels of abstraction.

In terms of datasets, there are stereo datasets, which contain pairs of color images captured by the dual-lens system with two color cameras. These datasets can be used to achieve multi-modality; however, the contribution of multi-modality to the CNN may not be put forward since there is not any other image of the object to train a single-stream CNN model and compare [15], [16], [17], [18].

There are also multi-view datasets that contain more than two color images of the same object, which are captured by multiple color cameras [19], [20], [21], [22]. These datasets cannot be used for our study because they contain images either taken from random viewpoints or that do not have any overlapping areas (of the object) in order to let CNN extract hidden features by making use of the correlations.

III. DATASET

As discussed in the previous section, there is no available dataset in the literature to utilize in our study. For this reason, we created a dataset by building a rig, consisting of five cameras (Logitech C310 HD 720p) placed in parallel on the horizontal axis. Each camera can provide images in 19 different resolutions with two different codec options.

The cameras are labelled from one to five for ease of use. The distances between the lenses are shared in Fig. 2, so that the dataset can be used in different studies (e.g., disparity mapping, depth estimation).

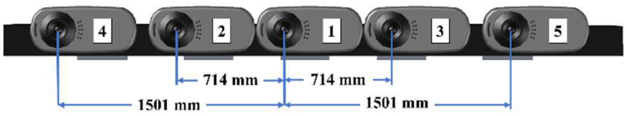

FIGURE 2. Rig setup consisting of five cameras.

A simple interface is designed to facilitate the process, as shown in Fig. 3. When the "snapshot" button is pressed, snapshots obtained from all five cameras are saved in five separate folders with the same file name. The naming convention for each snapshot is "snapshotYYYYMMDDHHMMSS.png".

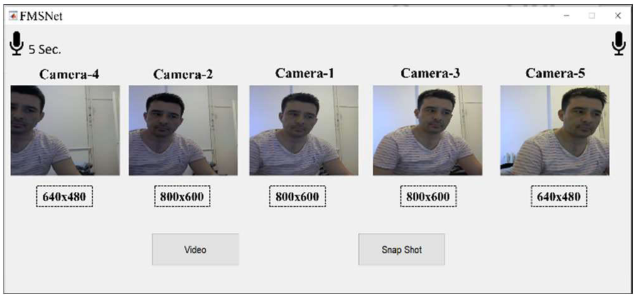

FIGURE 3. Interface.

When the "video" button is pressed, video data obtained from all five cameras and audio data obtained from the microphones of the cameras on the far right and left are saved as a file in separate folders. Both video and audio are recorded for the duration of five seconds. The naming convention for each video data is "videoYYYYMMDDHHMMSS.avi" and for each audio data, it is "audioYYYYMMDDHHMMSS.wav".

An important point here is to be able to capture image and video data simultaneously. As a result of our examination, we determined that there is a maximum delay of approximately half a second between the first and fifth cameras.

Mobility is important to be able to obtain outdoor images and videos. For this reason, the cameras are connected to a laptop. The final hardware set-up is shown in Fig. 4.

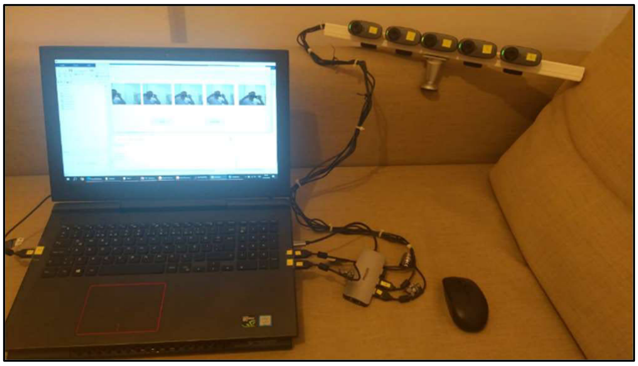

FIGURE 4. Hardware set-up to create the dataset.

Cameras labelled; 1, 2, and 3 have the resolution of 800×600 pixels, whereas 4 and 5 have the resolution of 640×480 pixels. The reason for this difference in resolutions is the bottleneck of the available "USB Controller Bandwidth" provided by the hardware configuration (the laptop). Higher resolutions and simultaneity can be achieved with different hardware configurations. However, the horizontal field of view of each camera, which is 60^0, is more important and considered adequate for this study.

The number of classes in the dataset consisting of "Automobile", "Cat", "Dog", "Bus", "Tree", "Bicycle", "Building", and "Traffic Lights" is eight.

TABLE 2. Examples from the dataset.

Camera-5	Camera-3	Camera-1	Camera-2	Camera-4
Traffic Lights				
Cat				
Bicycle				

Table 2 shows some examples of data from the "Traffic Lights", "Cat", and "Bicycle" classes in the dataset.

The dataset is divided into three groups: "training", "validation", and "test". Table 1 shows the number of images per camera in each class utilized in this study.

Images obtained from cameras 2, 3, 4, and 5 are used for training the proposed CNN, and images obtained from Camera-1 are used for the evaluation purposes.

IV. TESTING ENVIRONMENT

The hardware used in this study to design, train, test and evaluate the CNNs is shown in Table 3. GPU, HDD/SSD, and RAM are the main components that affect training time and ability to design and train more complex CNNs.

TABLE 3. Hardware.

Processor	Intel(R) Core (TM) i9-13900K 3.00 GHz
GPU	NVIDIA GeForce RTX 4090
RAM	32 GB
Operating System	x64, Windows 11 Pro
SSD	Samsung SSD 990 Pro 1TB

Various popular software such as Caffe, Keras, PyTorch, and TensorFlow are available for building CNN architectures. In this study, we used Deep Network Designer tool of MATLAB R2023b software.

V. PROPOSED METHOD

This section provides detailed information about the proposed model.

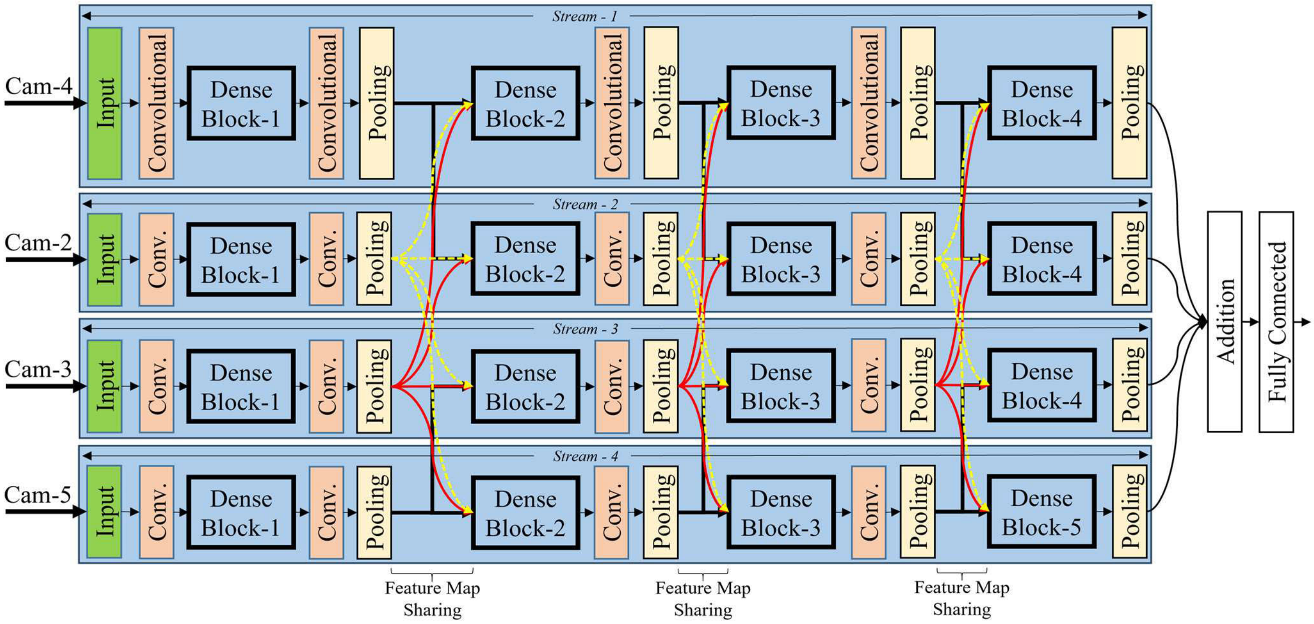

FIGURE 5. FMSNet: Four-stream CNN architecture for multi-stereo image classification by feature map sharing.

A. FMSNET CONFIGURATION

The configuration of our multi-stream CNN is presented graphically in Fig. 5. Each of the four streams is essentially an untrained DenseNet-201 architecture. We applied;

• "Feature Map Sharing" after Dense Blocks - 1, 2, and 3. The sharing of the features is depicted by arrows with different colors.

There are a number of methods to fuse layers between different networks/streams like sum, max, concatenation, conv and bilinear fusions [6].

A fusion function $f : x_t^a, x_t^b, x_t^c \rightarrow y_t$ fuses three feature maps $x_t^a \in \mathbb{R}^{H \times W \times D}$, $x_t^b \in \mathbb{R}^{H' \times W' \times D'}$ and $x_t^c \in \mathbb{R}^{H'' \times W'' \times D''}$, to produce an output feature map $y_t \in \mathbb{R}^{H''' \times W''' \times D'''}$, at time t, where H is height, W is width, D is number of channels, and superscript a, b, c and d are networks/streams to be fused as depicted in Fig. 5.

For the purpose of "Feature Map Sharing", we applied depth concatenation by taking the outputs from own and the neighboring two pooling layers and stacking them along the channel dimension. Fusion function:

$$y^{cat} = f^{cat}\left(x^a, x^b, x^c\right) \tag{1}$$

concatenates the three feature maps at the same spatial locations i, j across the future channels d, where $y \in \mathbb{R}^{H \times W \times 3D}$:

$$y_{i,j,2d}^{cat} = x_{i,j,d}^a \quad y_{i,j,2d-1}^{cat} = x_{i,j,d}^b \tag{2}$$

So, for stream-1; a is 1 (own stream), b is 2 (neighboring stream) and c is 3 (other neighboring stream). "Future Map Sharing" is employed in the same manner to the other layers and streams.

• Addition layer, which performs element-wise addition to fuse the four streams.

$$y^{sum} = f^{sum}\left(x^1, x^2, x^3, x^4\right) \tag{3}$$

computes the sum of the four-feature maps at the same spatial locations i, j and dimension d :

$$y_{i,j,d}^{sum} = x_{i,j,d}^a + x_{i,j,d}^b + x_{i,j,d}^c + x_{i,j,d}^d \tag{4}$$

where $1 \leq i \leq H$, $1 \leq j \leq W$, $1 \leq d \leq D$ and x^a, x^b, x^c, x^d, $y \in \mathbb{R}^{H \times W \times D}$.

Image input size is set to 227,227,3 and classification layer output is set to eight. All other layers are used as they are in the DenseNet-201 architecture.

B. TRAINING

The hyper-parameters play an important role in the success of CNN training. Appropriate hyper-parameter selection may vary depending on the architecture of the CNN being trained and the type of data to be trained. The hyper-parameter values used in our study are presented below.

1) LEARNING RATE

Whereas a constant learning rate can be used throughout the entire training, a learning rate that is high at the beginning of the training and decreases steadily or exponentially as the training progresses (decaying learning rate) can also be preferred. The value of 0.001 is used as the learning rate in our CNN and kept constant throughout the training.

2) OPTIMIZATION ALGORITHM

Stochastic Gradient Descent (SGD) algorithm is used for optimization purposes.

3) MOMENT VALUE

For our CNN not to be negatively affected by local minima during backpropagation, the moment value is set to 0.9.

The dataset is resized to 227,227,3 and normalized before training. Images obtained from the four adjoining cameras are inputted into the four separate streams all at once for each of the forward passes. At each iteration, a mini batch of 16 images is used. Training is performed from scratch for the duration of five epochs.

VI. EXPERIMENTAL RESULTS, EVALUATION AND DISCUSSIONS

There is unfortunately no benchmark competition or similar study in literature to compare and evaluate the results of the proposed CNN. To validate the effectiveness of the proposed model, we compare it with four state-of-the-art CNNs that have proven themselves in various benchmark competitions by providing high accuracy rates. These CNNs are DenseNet-201, GoogLeNet [23], InceptionResNet(v2) [24], and Inception(v3) [25]. For training, the images obtained only from Camera-1 are used, which are resized according to the requirements of the related CNN and normalized before the training.

After the pre-processing of the dataset, these CNNs are trained from scratch by using the same hyper-parameters, training images and training options (learning rate, optimization algorithm, minibatch size, epoch, and moment values) as in our proposed CNN. After training, CNNs are tested with the test data of Camera-1. Experimental results are presented in Table 4, where we reported the number of parameters, training time, query response time and accuracy rate of each CNN.

TABLE 4. Experimental results of the CNN models.

CNN	Parameters	Training Time	Query Response Time	Accuracy
FMSNet	118.7 M	122 min., 42 s.	0.14 s.	**86.44%**
DenseNet-201	18.1 M	12 min., 29 s.	0.03 s.	75.14%
GoogleNet	5.9 M	1 min 35 s.	0.01 s.	69.96%
InceptionResNet(v2)	54.3 M	16 min 16 s.	0.03 s.	82.49%
Inception(v3)	21.8 M	5 min 33 s.	0.01 s.	76.65%

Our proposed CNN outperformed all other CNN models owing to utilizing feature map sharing; however, the number of parameters, training, and query response time are increased significantly.

Although FMSNet is utilizing DenseNet-201 as a backbone model, it achieved 11.3% higher accuracy rate. The closest result to FMSNet is InceptionResNet(v2) with the accuracy rate of 82.49%, which still means 3.95% lower accuracy rate than FMSNet.

One may think that FMSNet is trained on images obtained from Cam2, Cam3, Cam4, and Cam5, whereas other CNNs are trained on images obtained only from Cam1. This means a considerable number of less training data and so inequality. In order to resolve this mismatch, we trained all CNNs from scratch with images obtained from Cam2, Cam3, Cam4, and Cam5 by using the same hyperparameters mentioned in previous section. The results are presented in Table 5.

TABLE 5. Experimental results of the CNN models (Trained with equal amount of training data).

CNN	Parameters	Training Time	Query Response Time	Accuracy
FMSNet	118.7 M	122 min., 42 s.	0.14 s.	86.44%
DenseNet-201	18.1 M	70 min., 52 s.	0.03 s.	83.38%
GoogleNet	5.9 M	11 min., 58 s.	0.01 s.	84.98%
InceptionResNet(v2)	54.3 M	111 min., 07 s.	0.03 s.	**86.96%**
Inception(v3)	21.8 M	32 min., 53 s.	0.01 s.	82.49%

The CNN model to achieve the highest accuracy rate is InceptionResNet(v2) with only 0.52% difference. FMSNet achieved once more a 3.06% higher accuracy rate than its non-FMS counterpart, DenseNet-201, in second in place. This clearly shows the contribution of the FMS to the learning process. Training times are increased due to the increased training data, whereas the query response times stay the same.

VII. CONCLUSION AND FUTURE STUDIES

In this study, we propose a novel multi-modal multi-feature map sharing CNN model (FMSNet) and a new dataset consisting of multi-stereo images that are overlapping in a systematic manner. The proposed model achieved a 3.06% higher accuracy rate than its non-FMS backbone CNN model, which is a single-stream DenseNet-201. Our model also gained better results than most of the state-of-the-art single-stream CNN models.

It is of no question that our brain receives multi-modal data, and processes these with neurons that have intricate connections. We assess that studies in the area will shift towards researching how feature map sharing (FMS), which brings about intricate connections between different CNN streams, should be configured.

FMS brings about a high number of parameters. This feature is highly dependent on computer resources like processing power (CPU, GPU/parallel programming) and RAM. Progress in hardware will enable further studies.

The lack of dataset is another challenge that hinders further studies. Existing multi-stereo datasets contain images that are randomly captured or do not systematically overlap. The dataset created in this study will be improved by adding voice and video data. In further studies, these data can be input in a CNN as another stream, which can significantly improve the feature extraction process. This dataset can also be used for depth estimation studies.

In this study, one of our main objectives is to put forward how the feature map sharing itself contribute to the learning process. In this context, utilizing the well-known CNN model, which is DenseNet-201, as a backbone helped to make comparison with FMSNET.

More parameters mean long query response times. This will hamper the utilization of our proposed CNN especially in online platforms which requires instant query responses. Having known the importance and positive contribution of utilizing the FMS feature, other single-stream CNN models can be re-modelled or new CNN models can be created in future studies to increase accuracy and reduce the number of parameters, training, and query response time.

REFERENCES

[1] Y LeCun, E. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9. [Online]. Available: http://code.google.com/p/cuda-convnet/

[3] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[4] Y.-C. Du, M. Muslikhin, T.-H. Hsieh, and M.-S. Wang, "Stereo vision-based object recognition and manipulation by regions with convolutional neural network," *Electronics*, vol. 9, no. 2, p. 210, Jan. 2020, doi: 10.3390/electronics9020210.

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[6] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941, doi: 10.1109/CVPR.2016.213.

[7] W. Wang, X. Li, Z. Xu, W. Yu, J. Zhao, D. Ding, and Y. Chen, "Learning two-stream CNN for multi-modal age-related macular degeneration categorization," 2020, *arXiv:2012.01879*.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[10] R. Ke, W. Li, Z. Cui, and Y. Wang, "Two-stream multi-channel convolutional neural network (TM-CNN) for multi-lane traffic speed prediction considering traffic volume impact," 2019, *arXiv:1903.01678*.

[11] S. Guo, L. Wang, Q. Chen, L. Wang, J. Zhang, and Y. Zhu, "Multimodal MRI image decision fusion-based network for glioma classification," *Frontiers Oncol.*, vol. 12, Feb. 2022, Art. no. 819673, doi: 10.3389/fonc.2022.819673.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[13] A.-H. Jo and K.-C. Kwak, "Diagnosis of depression based on four-stream model of bi-LSTM and CNN from audio and text information," *IEEE Access*, vol. 10, pp. 134113–134135, 2022, doi: 10.1109/ACCESS.2022.3231884.

[14] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. Ben Ayed, "HyperDense-net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019, doi: 10.1109/TMI.2018.2878669.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3213–3223. [Online]. Available: www.cityscapes-dataset.net

[16] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8. [Online]. Available: http://vision.middlebury.edu/stereo/data/

[17] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.

[18] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4040–4048. [Online]. Available: https://www.blender.org/

[19] Z. Gao, H. Zhang, G. P. Xu, Y. B. Xue, and A. G. Hauptmann, "Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition," *Signal Process.*, vol. 112, pp. 83–97, Jul. 2015, doi: 10.1016/j.sigpro.2014.08.034.

[20] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," 2016, *arXiv:1604.02808*.

[21] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2649–2656.

[22] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A deep structured model with radius–margin bound for 3D human activity recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 256–273, Dec. 2015, doi: 10.1007/s11263-015-0876-z.

[23] in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 7–12.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–7. [Online]. Available: www.aaai.org

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

FERIT CAN was born in Buca, İzmir, Türkiye, in 1981. He received the B.S. degree in systems engineering from Turkish Military Academy, Ankara, Türkiye, in 2003, and the M.S. degree in computer engineering from National Defense University, Istanbul, Türkiye, in 2019, where he is currently pursuing the Ph.D. degree in computer engineering. His research interests include image classification, object detection, deep learning, and convolutional neural networks.

CAN EYUPOGLU received the B.Sc. degree (Hons.) in computer engineering and the Minor degree in electronics engineering from Istanbul Kültür University, Türkiye, in 2012, and the M.Sc. and Ph.D. degrees (Hons.) in computer engineering from Istanbul University, in 2014 and 2018, respectively.

From 2019 to 2021, he was an Assistant Professor with the Computer Engineering Department, Turkish Air Force Academy, National Defense University, Istanbul, Türkiye. He is currently an Associate Professor and the Head of the Department of Computer Engineering, Turkish Air Force Academy, National Defense University. He has published about 70 papers in various esteemed journals and conferences. His current research interests include image processing, artificial neural networks, machine learning, bioinformatics, and data privacy. He has been serving as a member of the reviewer board in nearly 40 prestigious academic journals. He is also on the editorial board of some reputable journals.