# Dynamic Hand Gesture Recognition Using Temporal-Stream Convolutional Neural Networks

Fladio Armandika
*Department of Informatics*
*Universitas Jenderal Achmad Yani*
Cimahi, Indonesia
fladioarmandika@gmail.com

Esmeralda Contessa Djamal*
*Department of Informatics*
*Universitas Jenderal Achmad Yani*
Cimahi, Indonesia
corr author : esmeralda.contessa@lecture.unjani.ac.id

Fikri Nugraha
*Department of Informatics*
*Universitas Jenderal Achmad Yani*
Cimahi, Indonesia
fikri.nugraha@student.unjani.ac.id

Fatan Kasyidi
*Department of Informatics*
*Universitas Jenderal Achmad Yani*
Cimahi, Indonesia
fatan.kasyidi@lecture.unjani.ac.id

***A*bstract—Movement recognition is a hot issue in machine learning. The gesture recognition is related to video processing, which gives problems in various aspects. Some of them are separating the image against the background firmly. This problem has consequences when there are incredibly different settings from the training data. The next challenge is the number of images processed at a time that forms motion. Previous studies have conducted experiments on the Deep Convolutional Neural Network architecture to detect actions on sequential model balancing each other on frames and motion between frames. The challenge of identifying objects in a temporal video image is the number of parameters needed to do a simple video classification so that the estimated motion of the object in each picture frame is needed. This paper proposed the classification of hand movement patterns with the Single Stream Temporal Convolutional Neural Networks approach. This model was robust against extreme non-training data, giving an accuracy of up to 81,7%. The model used a 50 layers ResNet architecture with recorded video training.**

***Keywords—Hand-gesture recognition, Temporal Stream CNN, Convolutional Neural Networks***

## I. INTRODUCTION

Video is a collection of pictures in a sequence per unit of time. It has a configuration that is frames per second (fps), which represents the number of images in groups of the time. In the field of artificial intelligence, video can be used to recognize objects or movements. Human activity recognition has become a popular area in the field of artificial intelligence. Identifying human activity from a video such as hand gestures has been used in many fields, such as surveillance systems, remote control of robots, video games, and self-driving cars.

Previous studies have performed video classification in various ways, such as HOG3D [1], SIFT3D [2], HOF [3], and Deep Learning. The deep Learning method represents a Spatio-temporal feature that operates on individual frames or several frames using machine learning methods such as Convolutional Neural Network (CNN) by combining static and motion information in one stream on YouTube Dataset[4]. Other studies perform static and motion information in two streams and then merge (fusion) at the end of the network [5]. It resulted in an accuracy of Top 1 in training data of 99.9% and Top 1 in test data of 84.5% using the ExceptionNet architecture in the temporal stream, which is better than the spatial flow. The combination of the two results in better accuracy but produces more substantial computations with longer processes. Temporal Stream provided better accuracy when compared to Spatial Stream. CNN extracted spatial features in an unsupervised way from individual images and then used a Gaussian filter for early detection of each frame to avoid noise from a stable to quasi-periodic frame transition in unstable frames at the time of combustion in the detected fire image[6]. The results showed better with static and motion separations using validation by measuring Euclidean Distance on output data with target data.

Previous research on the hand gestures recognition from the temporal interframe pattern feature by combining two categories of hand gestures, which is static and dynamic data from hand gestures [7]. Labeling is done using Bayesian Network to reduce errors in labeling temporal datasets on each frame. Then the recognition used several classifiers such as SVM, Bayesian NW, Random Forest, and C4.5 Tree, which produced the highest accuracy of 98.1% with a recognition scenario using two and three classifiers at once. In the experimental process, the dataset used is only 16 static hand gesture data and seven dynamic hand gestures, so it is necessary to experiment with a dataset with a more significant number and use a better classifier in recognition of large amounts of data.

Other research on hand gesture recognition performs 2D and 3D Convolutional Neural Network through video. Both are used to represent the temporal features of video-shaped datasets from VIVA HGD [8]. Three dimensions CNN can do the mapping from hand gesture videos to temporal patterns that represent dynamic hand gestures, whereas 2D CNN is used to get vector features as the representation of static hand gestures. From the experimental results, 2D CNN has better recognition capabilities compared to 3D CNN. 2D CNN accuracy with GoogleNet reached 72.5%, while 3D CNN with C3D models reached 68%. However, in this research, the recognition accuracy still needs to be improved by implementing DenseNet, which can be performed on 2D CNN.

Other studies of hand gesture recognition had a problem with a complex background. This complex background dramatically influences the performance to recognize hand gestures from the video [9]. This problem can be dealt with by analyzing the pattern of the Spatio-temporal domain feature extraction. Besides, coupled with color HSV analysis to improve hand recognition capabilities. HSV color space can segment to separate noisy background with the visual perception that represents individual objects. The recognition accuracy reaches 98.33% using Manhattan distance to measure the feature vector between the training data and the test data. However, there are still failures in recognition caused by ambiguity, so a better method is needed in the process of hand recognition.

Other research on Sign Language Recognition (SLR) used 3D Convolutional Neural Network. The feature used for SLRs is the Spatio-temporal feature extracted from the CNN feature layer without prior knowledge to find out the type of functions [10]. In addition to the Spatio-temporal features, features such as color information, depth clue, body joint, and trajectory information are used to improve the recognition performance. Microsoft Kinect was used to obtain hand gesture patterns that represent Sign Language. The accuracy obtained in the experimental results is 94.2% with a multi-channel configuration. However, in the classification process, there are still recognition errors caused by noise arising from the extraction process, so it needs a better analysis of feature selection.

Other studies for hand gesture recognition used Dynamic Time Warping (DTW), which is often used for speech recognition [11]. The difference is the sampling rate used for speech is lowered to 10 Hz in hand gesture recognition. Before the introduction of video data, the Spatio-temporal feature extracted has a higher fluctuation in time of axis compared to the swing of the sound pattern, from the experimental results obtained an accuracy of 89.6% from 12 classes of hand gestures. However, fluctuations in patterns occur when the Spatio-temporal feature is extracted, so that it would be better if the dataset being built focuses on static hand gestures.

Other studies focused on using temporal features to recognize gestures fully. This temporal feature is extracted by segmenting the total gesture activity that occurs in a video sequence [12]. After being added, it will enter the decision-making stage, of which the temporal segment meets true local minima. The accuracy obtained from the experiment was 93% by dividing the number of filtered temporal transitions by the number of developments that exist in the database of body gestures taken through Microsoft Kinect. This method can be used in the future for preprocessing in hand gesture recognition, specifically for Sign Language Recognition (SLR). However, the level of noise reduction of features can still be increased by using other extraction methods such as Histogram of Oriented Gradient (HOG).

Other studies regarded the introduction of dynamic hand gestures based on depth information[13]. Depth information itself was obtained from several components such as depth image, trajectory, hand shape changes. Besides, the temporal flow feature is extracted using a Temporal Pyramid Algorithm that can accurately represent and combine parts of elements in different dimensions. Support Vector Machine is used for the classification process, with a recognition rate of 95%. Furthermore, to simplify the representation of hand gestures, that needs a way to extract temporal features on the fingers to get better hand gestures.

Other related research to recognize hand gestures was to utilize the ability of the Wrist-worn inertial sensor [14]. The sensor is used to extract the hand gesture features, both spatial and temporal. The feature extraction used Dynamic Time Warping combined with Riemannian distance to measure the similarity between hand gestures in each trajectory read by the sensor. Based on that, the accuracy of its introduction reaches 99.2% using the Adaptive DBA Algorithm. Besides excellent recognition skills, there are significant obstacles because specific sensors were needed so that they are expected to be applied to sensors that have more general specifications and are easily accessible.

Other studies have performed experiments on the Deep CNN architecture to detect action on sequential images. The challenge is to get information that balances each other on static and motion between frames. The study used Two-Stream CNN architecture, which combines spatial and temporal networks. In the CNN method, several frames of learning are processed with Dense Optical Flow, which can help produce excellent performance even with limited training data. Besides being used to detect moving objects, single-stream temporal can be used to cut videos based on moving objects. Besides CNN, Skeleton-based Recurrent Neural Networks can also be used to recognize gestures [15] or by using Kinect RGB to capture skeletal joints[16]. In Temporal Learning, frames are stacked by adding dimensions to the input data. Nevertheless, the disadvantage is not processing movements between frames, so accuracy can still be improved [17].

While other studies used Attention-based Temporal Weighted CNN, a model with a sequence to get dominant features on a set of frames is then filtered with CNN to get more targeted results and focus on filtering on more dominant features [18].

This research proposed a hand gesture recognition using single-stream temporal CNN. We used training 160 data recorded following recording scenarios. Video data trained and tested using the Convolutional Neural Network with the ResNet-50 architecture. The identification is divided into four classes, mainly "Left", "Right", "Phone" and "Grab".

## II. METHODS

### A. Data Acquisition

Training data is recorded using a cellphone camera that has a size of 320 x 240 with 20 frames per second (fps). The video is recorded with normal lighting conditions, and the hand is not using a glove and is not blocked by other objects and background colors that are not more dominant. After that, the video recording will be done trimming or cutting the duration to one second for each movement and saving it into one video.

Fig. 1. Hand movement of four classes

The trimming results were 160 videos and are grouped based on the same movement in one folder, as in Fig. 1.

### B. Convolutional Neural Network

Convolutional Neural Network (CNN) is one of the Deep Learning and neural networks method that are most commonly used to analyze visual image data. Compared to other image classification algorithms, Convolutional Neural Network (CNN) uses very little preprocessing compared to others, connecting filter networks that are usually manually engineered in other systems [19].

In the previous literature, CNN was used to analyze sequential image data by dividing two streams, which are spatial and temporal stream [3]. Other studies using CNN with Resnet-34 architecture[16] and compare it with different architectures such as VGG[20]. In addition to using temporal

segmentation, Two-Stream CNN can be used by dividing global and local features [21].

The initial stage of CNN is convolute the input frame that has been processed. The convolution process utilizes what is called a filter. Like images, filters have a height, width, and thickness (h x w x d). This filter is initiated with a specific value (randomly or using a precise technique), and the importance of this filter is a parameter that will be updated in the learning process. The Max Pooling Layer used the maximum of each neuron in the layer. The approach multiplies by dot product each feature with a filter [22]. The Convolution Layer uses the ReLU activation function. Rectified Linear Unit (ReLU) is one of the activation functions used in other studies in Nonparametric Regression[23]. At the end of the convolution, the output will be carried out on Average Pooling on each feature map of the final result of the convolution so that the output dimension can look like Fig 2.


Fig. 2. Average pooling at the end of convolution

Then Flatten is carried out at the convolution output. Flatten used the process of converting a feature map matrix into an input vector. Making vectors as dense input layer or fully connected network for the classification or identification process[24].

The network output results will be converted into probabilities by taking from the exponents of each output using (1). The highest chance determines which class the video is categorized.

$$Softmax(x_i) = \frac{\exp(x_i)}{\sum_i e^{x_i}} \quad (1)$$

Using Backpropagation is each input neuron is multiplied by the weight and summed with a bias to be used as input to the hidden layer value.

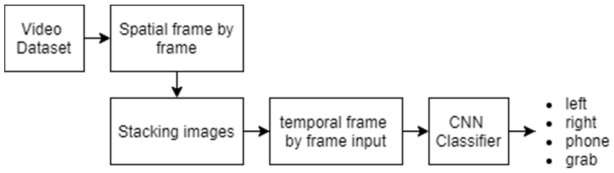


Fig. 3. Single-Stream CNN of hand gesture classification

The video will be converted into images per frame, then temporal feature extraction will be performed on several frames, and the results will be an input of Temporal Stream CNN. The converter shows in Fig. 4. Temporal Stream is a stream that processes multiple video frames simultaneously. One way to process temporal input is by stacking each network output each time [25]. Every two frames are stacked into one frame. So the number of inputs would be n-1. Previous research using Optical Flow calculations on several structures at once. Therefore, this study performs an average of two frames as input to artificial neural networks.

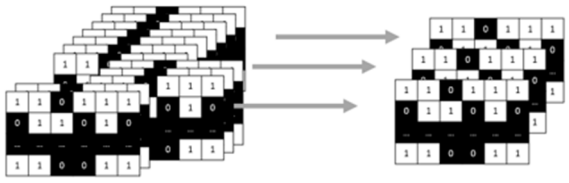

Fig. 4. Average Frames

They are stacking each frame. In other research, optical flow can combine several frames and represent movement from one frame to another [26]. The stream that is used to optimize motion recognition sometimes is different from Optical Flow. Other studies use Gunnar Farneback to convert spatial frames into segmented images based on the motion [27]. Pre-Trained Lite FlowNet models can also be used to estimate movement [28] or by stacking several flow layers [29].

### *C. ResNet*

This research used the ResNet architecture. Residual Neural Network Architecture consists of remaining blocks containing parallel states for residual streams. The residual block unit includes a connection similar to the residual block structure in the original ResNet with a Single Convolutional Layer and transient flow, which is the usual convolution layer.

$$r_{l+1} = \sigma(conv(r_l, W_{l,r\to r}) + conv(t_l, W_{l,t\to r}) + shortcut(r_l)) \quad (2)$$

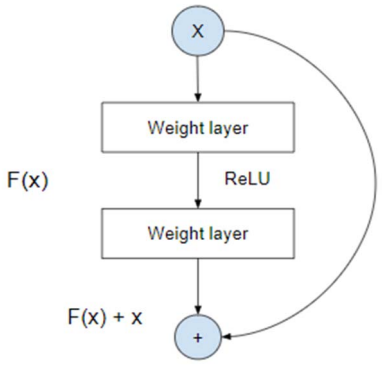


Fig. 5. ResNet

$$t_{l+1} = \sigma(conv(r_l, W_{l,r\to t}) + conv(t_l, W_{l,t\to r})) \quad (3)$$

Where,

$r_l$ = residual stream
$t$ = transient stream
$W$ = *block*

In ResNet architecture, the sum performs each two convolution F(x) result with the previous convolution (x), as shown in Fig. 5. In the mapping process has no parameters, and it is used to add the output from the previous layer to the next layer. In some cases, x and F (x) do not have the same dimensions. In other literature, the use of Resnet in Resnet can improve performance on the CIFAR-100 dataset [30]. In the mapping process, it has no parameters, and it is used to add the output from the previous layer to the next layer, as in Fig. 5. In some cases, x and the function do not have the same

dimensions [30]. Other studies using ResNet 50 Layer perform an accuracy of 96% [31].

## III. Result and Discussion

Hand movements video used 160 videos that consist of four classes with a one-second duration each video. Data were grouped according to data recording scenarios. The data was divided into two parts, 80% for training and 20% for testing. Training and testing used the Keras library. This study used the GPU Tesla K80 to accelerate training and testing performance.

The experiment used a 0.01 learning rate and 100 epochs to optimize updated models such as SGD and Adam. The results of the SGD and Adam optimization tests can be seen in Table I that used ResNet.

TABLE I. Accuracy Each Model

| Feature | Train | | Test | |
|---|---|---|---|---|
| | *Accuracy* | *Loss* | *Accuracy* | *Loss* |
| ResNet with SGD | 94,26 | 0,67 | 79,46 | 0,49 |
| ResNet with Adam | 96,00 | 0,59 | 81,72 | 0,41 |

Table I shows that the accuracy of Adam as optimization higher compared to the SGD as an optimizer. All models used ResNet.


Fig. 6. Accuracy of test data

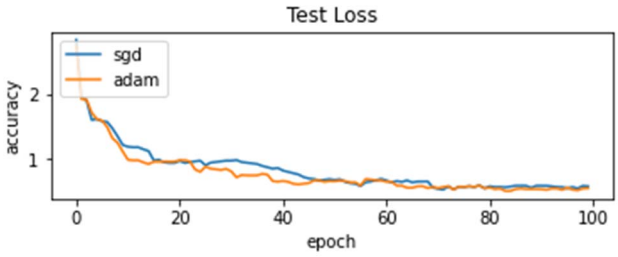


Fig. 7. Loss of test data

Fig. 6 showed the accuracy of the model with an optimizer using SGD. By using a 0.01 Learning Rate, we can see that significant fluctuation occurs at the beginning of learning. Loss values are volatile, with a tendency to increase its value. It indicates the difficulty of learning models to do learning correctly. Fig. 7 showed loss values increased slightly at epoch 30th. In models with the optimizer, Adam has an accuracy of 96% in training data and 81.7% in test data. The loss value continues to increase in the test data, while the accuracy of the test data is not too large but stable.

In testing, which is 20% of the total dataset, five results do not match up to correct labels on the Phone class. While the Right type has one incorrect conclusion, and the Phone class has the most incorrect results with the results found in the other classes, as shown in Table II.

TABLE II. Confusion Matrix

| True label | Predicted | | | |
|---|---|---|---|---|
| | *Grab* | *Left* | *Phone* | *Right* |
| Grab | 132 | 5 | 15 | 3 |
| Left | 10 | 142 | 5 | 1 |
| Phone | 5 | 2 | 138 | 5 |
| Right | 12 | 6 | 11 | 117 |

## IV. Conclusion

Convolutional Neural Network can be used to classify sequential images. Convolutional Neural Network by processing several frames at once, can produce good accuracy. By using the SGD optimizer, which has the characteristics of using random values in data sampling that will be studied by the learning model. So that it is a more dynamic learning model marked by fluctuations in the amount of Loss, which is more significant compared to Adam, we can analogize SGD like a machine that can learn by reading books. If the learning rate is the level of the page being studied, SGD will consider pages randomly jumping up and down and cause a tendency to change the value of Loss, which is more volatile. This result can be interpreted both in the number of data sets as much as in video data.

However, on the other hand, Adam works by regularly updating weights with random and threshold values for Loss. If we analogize machine learning by reading books, Adam has characteristics that are more adjusting to the situation when each page is learned by the model that produces a higher loss value than before. Then he will make adjustments by randomizing the page or sample data being studied. It is useful in this study by producing significant accuracy and loss values when compared to SGD. However, Adam updated that can converge to the configured Loss limit will be longer and require a longer epoch. In this study, we also found temporal features using CNN and stacking images in a two-frame configuration producing good accuracy. Further development in this research can be done by using stacked images as features to get better results in terms of accuracy and complexity of the motion that must be detected.

## Acknowledgment

Thanks to the Ministry of Technology Research for the financial support provided for this research in the “Penelitian Terapan Unggulan Perguruan Tinggi” 2020 grant with number B/87/E3/RA.00/2020.

## References

[1] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," *BMVC 2008 - Proceedings of the British Machine Vision Conference 2008*, 2008.

[2] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *Proceedings of the ACM International Multimedia Conference and Exhibition*, no. c, pp. 357–360, 2007.

[3] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[4] S. Ramasinghe and R. Rodrigo, "Action recognition by single stream convolutional neural networks: An approach using combined motion and static information," *Proceedings - 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015*, pp. 101–105, 2016.

[5] F. Nugraha and E. C. Djamal, "Video Recognition of American Sign Language Using Two-Stream Convolution Neural Networks," *The International Conference on Electrical Engineering and Informatics*, 2019.

[6] D. K. Jha, A. Srivastav, and A. Ray, "Temporal Learning in Video Data Using Deep Learning and Gaussian Processes," *International Journal of Prognostics and Health Management*, vol. 7, no. 022, p. 11, 2016.

[7] K. Hu, L. Yin, and T. Wang, "Temporal Interframe Pattern Analysis for Static and Dynamic Hand Gesture Recognition," *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3422–3426, 2019.

[8] M. Kurmanji and F. Ghaderi, "A Comparison of 2D and 3D Convolutional Neural Networks for Hand Gesture Recognition from RGB-D Data," *ICEE 2019 - 27th Iranian Conference on Electrical Engineering*, pp. 2022–2027, 2019.

[9] E. Pabendon, H. Nugroho, A. Suheryadi, and P. E. Yunanto, "Hand Gesture Recognition System Under Complex Background Using Spatio Temporal Analysis," *Proceedings of 2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering, ICICI-BME 2017*, no. November, pp. 261–265, 2018.

[10] H. Jie, Z. Wengang, L. Houqiang, and L. Weiping, "Sign Language Recognition using 3D Convolutional Neural Networks," *2015 IEEE International Conference on Multimedia and Expo (ICME)*, 2015.

[11] Y. Zhu, H. Ren, G. Xu, and X. Lin, "Toward real-time human-computer interaction with continuous dynamic hand gestures," in *Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000*, 2000.

[12] T. L. Le, V. N. Nguyen, T. T. H. Tran, V. T. Nguyen, and T. T. Nguyen, "Temporal gesture segmentation for recognition," *2013 International Conference on Computing, Management and Telecommunications, ComManTel 2013*, no. 1, pp. 369–373, 2013.

[13] X. Bai, C. Li, L. Tian, and H. Song, "Dynamic Hand Gesture Recognition Based on Depth Information," *ICCAIS 2018 - 7th International Conference on Control, Automation and Information Sciences*, no. Iccais, pp. 216–221, 2018.

[14] Y. T. Liu, Y. A. Zhang, and M. Zeng, "Novel algorithm for hand gesture recognition utilizing a wrist-worn inertial sensor," *IEEE Sensors Journal*, 2018.

[15] C. Li, Y. Hou, P. Wang, and W. Li, "Skeleton-Based Action Recognition with Convolutional Neural Networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.

[16] S. Ahlawat, V. Batra, S. Banerjee, J. Saha, and A. K. Garg, "Hand gesture recognition using convolutional neural network," in *Lecture Notes in Networks and Systems*, vol. 56, 2019, pp. 179–186.

[17] I. Phueaksri and S. Sinthupinyp, "Convolutional Neural Network using Stacked Frames for Video Classification," *ACM International Conference Proceeding Series*, pp. 85–89, 2019.

[18] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, and N. Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," *14th International Conference on Artificial Intelligence Applications and Innovations (AIA I'2018)*, vol. 519, pp. 97–108, 2018.

[19] L. Xu, J. S. J. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, vol. 2, no. January, pp. 1790–1798.

[20] G. Strezoski, D. Stojanovski, I. Dimitrovski, and G. Madjarov, "Hand Gesture Recognition using Deep Convolutional Neural Networks," *ICT Innovations 2016*, 2016.

[21] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 2257–2264, 2018.

[22] D. C. Cires, U. Meier, J. Masci, and L. M. Gambardella, "Flexible, High Performance Convolutional Neural Networks for Image Classification," *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence Flexible*, pp. 1237–1242, 2003.

[23] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *IEEE conference on computer vision and pattern recognition*, 2017.

[24] J. Jin, A. Dundar, and E. Culurciello, "Flattened Convolutional Neural Networks for Feedforward Acceleration," *The International Conference on Learning Representations 2015*, no. 2014, pp. 1–11, 2014.

[25] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, no. i, pp. 1933–1941.

[26] A. Ali and G. W. Taylor, "Real-time end-to-end action detection with two-stream networks," *Proceedings - 2018 15th Conference on Computer and Robot Vision, CRV 2018*, pp. 31–38, 2018.

[27] G. Farneb, "Two-Frame Motion Estimation Based on," *Lecture Notes in Computer Science*, vol. 2749, no. 1, pp. 363–370, 2003.

[28] T. W. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8981–8989, 2018.

[29] A. J. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 9937–9945, 2019.

[30] S. Targ, D. Almeida, and K. Lyman, "Resnet in Resnet: Generalizing Residual Architectures," *Intelligent Computing Methodologies: 14th International Conference*, pp. 1–7, 2016.

[31] M.-K. C. Jung, Heechul Jung, "ResNet-based Vehicle Classification and Localization in Traffic Surveillance Systems," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 61–67, 2017.