Adult Content Classification Through Deep Convolution Neural Network

Adi Nurhadiyatna*§, Septian Cahyadi‡, Febri Damatraseta† and Yan Rianto*
*Research Center for Informatics,Indonesian Institute of Sciences, Bandung, Indonesia
†Department of Information System Service, Swiss German University, Tangerang, Indonesia
‡Department of Computer Science, Pakuan University, Bogor, Indonesia
§Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

E-mail : adin005[at]lipi.go.id

***Abstract*—Adult content filtering is one of the main challenge in Indonesia to prevent children from accessing the adult content. Conventional web blocking and filtering through domain name server filtering is not enough to prevent the adult content distribution. Mobile phones, tablet, and personal computer can distribute the adult content through the offline way. In this case, a more sophisticated and autonomous system is needed that can detect the adult content automatically. To leverage this problem, a deep neural network is used to build a model that is able to detect adult content automatically. In this experiments, our model is able to detect adult content with an accuracy of $75,08\%$ and $69,02\%$ during the validation and testing process, respectively.**

***Keywords*—Adult Content filtering, deep neural network, convolutional neural network(CNN)**

I. Introduction

Adult content is prohibited in Indonesia as specified in The Law of Anti-Pornography (Undang-Undang anti pornografi). In several southeast Asia country, especially Malaysia, Brunei Darusalam, and Indonesia, adult content is strictly limited. Recently, in Indonesia the way to prevent the obtaining of such content on the Internet is by collecting all domain name server of website that contain the adult contents. This is effective, but it requires a lot of resources. Furthermore, domain name servers grows rapidly. This system is also supported by Ministry of Communication and Informatics (Kominfo), which provides "Internet sehat" to block adult content for family and children protection purpose. However, in the specific condition, a more robust and autonomous system is needed that can detect adult content, while we face file distribution from devices to devices.

Indonesia, with the largest number Muslim citizens has different levels of adult content threshold. For example, women with bikini will be a regular content in USA, Europe, Australia, and parts of Asia. In Indonesia, however, bikini will be categorized as an adult content for Indonesian citizen, especially children. This is a bias threshold and a controversial issue since the government introduced the pornographic and porn action law in 2008. Pornography in Indonesia is prohibited in terms of creating adult content, distribution of adult content, selling, and renting explicit adult material. This law means to prevent objectifying women, and will affect the traditional cultures such as how Balinese woman, Jaipong dancer, Guinea koteka and etc.

On the other hand, from 2011 to 2013 there were over 13.600 rape cases [1] in Indonesia and a few of the cases were committed by children that were affected by adult content [2]. This condition make us need more sophisticated approaches to prevent the children to access the adult content. Blocking the website will be good, and effective to prevent children to access adult content through Internet. However, the question is what if the content already in their gadgets (mobile phones, PC, tablet, etc)?

In 2016, one of the biggest adult content provider (PornHub) released a review about the traffic access to their website. 72% of the traffic was from both mobile phones and tablets, 50% of them are from Android, and 47% from Mac OS. Additionally, Indonesia raised its rank to top 45 of traffic in 2016 [3].

The conventional approaches still needed to prevent an online access. However, we also need to give an attention to offline distribution. In this research, we develop a model that can detect images and videos that contain adult content. Deep Neural networks are very popular in the field of image recognition, especially convolution neural networks (CNNs). In this paper, the neural network layer is replaced by bunch of image pixel parameters that are trainable through the neural network. We don't need predefined or handy-crafted features like in the past few years approach.

In this research, we develop a CNN model to prevent the adult image content to distribute through offline way. This model can be embed in the mobile, tablet, and also in the personal computer. We also create framework to create new model and also retrain from scratch, due to dataset update. In this research we only classify images and video is not considered in this experiment. We need a voting and a keyframe selection methods for video classification. We restrict our experiments to images classification and exclude video as our experiment object.

The paper is organized as follows. After this introduction, related work is briefly discussed in Section II. The training step and its data are presented in Section III, and Section IV briefly describes the evaluation method. The implementation and results are reported in Section V. The last section concludes the paper.

978-1-5386-3978-8/17/$31.00

TABLE I: Dataset Classes

No	Group Classes	Classes	Subtotal Data
1	Animal	mouse,pig,rabbit, bee, bird, butterfly, camel, cat, chicken, crocodile, deer, dog, duck, elephant,fish,snake, turtle, cow, wolf, frog,goat,horse,kangaroo,lion,monkey,	20.804
2	Fruits	apples, banana, durian, grape, pinaple, rambutan	6.875
3	Transportation	airplane, car, motorcycle	3.271
4	Human	anime, baby, bikini, man, soldier, woman	11.243
5	Adult Contents	breast, penis, gay, hentai, interracial, jav, javcover, pornstar, vagina, western	19.365
	Total Data		**69.258**

TABLE II: Related Work Summary

No	Reference	Feature Detector	Feature Descriptor
1	Sandra et al. [4]	BossaNova	Binary Descriptor
2	Jonatas et al. [5]	ACORDE	Semantic Descriptor
3	Lopes et al. [6]	SIFT Blobs	Hue SIFT
4	Steel et al. [7]	Skin ROIs	Mask SIFT
5	Deselaers at al. [8]	SIFT based blobs Blobs	Difference of Gaussian
6	Ulges et al. [9]	Regular Grid	DCT
7	Zhang et al.[10]	Skin ROIs	Color, Texture, Intensity

II. Related Work

In recent years, adult content detection and filtering have become really important due to their presence on the Internet and gadgets. In several regions, this issue is not really important with regards to pornography laws that are applied in various country. There are countries Where citizens can access adult content freely because the law does not prohibit it. Pornography and adult content filtering become really interesting and indispensable in countries where the religion law is in force or where the law in a certain region is strongly influenced by the religion law. From the social and medical approaches, adult content and pornography may effect a bad condition and cause addiction. Even in the liberal countries, this content is often prohibited for children under 17.

Several approaches has been done [4]-[10] to solve this problem. However, in these approaches the main media to detect is a video that needs a keyframe and voting selection to assess whether it is adult content or not. In our case, there is no need for keyframe selection and voting since the main object is a single image. In [4], the Binary descriptor and mid level representation is used for the feature extraction. They combine both low level and mid-level representation to generate Bag of Word (BoW). Afterwards, an improvement of BossaNova [11] based method is used to generate BossaNova video descriptor which called BNVD. Nowadays deep neural network have evolved into a very power full approach in the pattern recognition field, especially in terms of classification since CNNs are the current actual state-of-the-art for the ImageNet challenge. In [5] a convolutional neural network is used to address this problem. In this work, features are extracted from the last convolutional layer from GoogleNet and from ResNets 52, 101, and 152. Basically, this work combined two sophisticated networks with adjusting the training strategy.

Table II shows us a brief summary of related work in term of adult content detection. In [12] they have a brief discuss about the wrong direction to decide a scene is belong to adult or not. The common way to decide is by assessing the skin exposure. However, there are a lot of adult content that shows really small skin surface, but contains perform adult activities. This is really relevant while the object is a video that has a lot of scenes to assess. In this research they introduce Bags of Visual Words (BoVW) to identify a video.

A. Model

In this experiment, ResNet18, ResNet32, ResNet50 [13], and VGG16 [14], were tested to compare the performance. We don't use pre-trained weight network and train all the network from scratch. However, it is still possible to perform transfer learning in the next training process. This model will be used for the mobile application and also in the personal computer, so we need a network as light as possible with competitive performance. VGG16 and ResNet are winners of ILSVRC14 [15] for ImageNet challenge with 1000 categories/classes. However, VGG16 has more than 249 million parameters and ResNet has 11 million parameters and 21 milion parameters for 18 layers and 34 layers to train. The number of parameters of VGG16 makes this model harder to implement in the mobile gadget for real-time application due to limited resources and higher computational time.

Algorithm 1 Residual Unit Function

```
1:  ResnetUnit(input_layer):
2:      p1 = BatchNormalization(input_layer)
3:      p2 = relu(p1)
4:      p3 = conv2d(p2)
5:      p4 = batch_norm(p3)
6:      p5 = relu(p4)
7:      p6 = conv2d(part5)
8:      #Addition between input and p6
9:      output = input_layer + p6
10:     #return function from addition operation
11:     return output
```

In this experiment, we compare two popular network as basic model for the real time implementation. VGG16, and Resnet will be compared to know which one more appropriate for this case. However, in this experiment we will examine Resnet deeper than VGG16 due to parameter number. In this network, there are few main operation such as: Batch Normalization (BN), Rectified Linear Unit (Relu), and addition from input layer. This network proposed to solve the vanishing gradient which usually got vanished after backward operation. In this network the gradient sent by additional operation from input layer with side network called Residual Network. As we can see in algorithm 1, there are three main operation in

Algorithm 2 Batch+Normalization function

Input : x over mini-batch: $B = \{x_{1..m}\}$
Output : $\{y_i = BN_{\gamma,\beta}(x_i)\}$
learning parameters : γ, β
mean of mini-batch
$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$
variance of mini-batch
$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$
normalization
$x_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$
scale and shift
$y_i \leftarrow \gamma x_i + \beta \equiv \hat{B}N_{\gamma,\beta}(x_i)$

residual unit. The first operation convolution, where the image will convoluted with the main convolution kernel. The next operation is batch normalization which explained in algorithm 2, and as an activation function Relu is used. Relu is a threshold operation which set input as zero if the input less than zero. The equation for this operation is as follow :

$$f(x) : \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} \quad (1)$$

Let $f(x)$ as Relu function and x as input value, Relu function will be $fx = max(0, x)$.

III. Training

In this section, training process will briefly described. In this research we have collected all data manually for training and validation process. In the testing process, we used a part of the NPDI database [4]. This database contains non-porn easy, non-porn hard, and porn labeled data. However, in our case we only use porn data to test our model. The reason behind this decision is that in the NPDI dataset, breastfeeding is categorized as non-porn hard. In this experiment we use certain hyper-parameter setup. Table III shows us the setup description for every network. We use a method for stochastic optimization (ADAM) optimizer based on [16].

A. Dataset

1) Manual Collection: Google image search was used to download images for a specific query. To collect all these data, we use 50 specific query and separate into different folder that indicate its classes. In this data, there are 5 (five) main classes such as: animals, fruits, transportation, human and adult content. From five main classes, we divide into 50 secondary classes. 50 classes built including 5 specific adult content classes. To verify our dataset, humans performed manual annotation. If there is an image that was missplaced into the wrong folder class, then the investigator deleted it or replace it. This dataset contain at least 1000 images for each class for training and validation except for adult content classes. We provide more than 2000 images per adult content classes. In table I we can see 5 group classes and 50 specific classes. In the next research we intend to enrich the dataset to cover all possibility and gain the network capability in term of adult content detection.

In this experiments, more than 60.000 images with size 224×224 were used to train the model. We use the same image dimensions as VGG16 style to fit the general experiment that was done by the other image recognition research. We divide these dataset into training and validation with 75% and 25% from total data respectively. To test our model, the nudity Database was used to verify our model performance.

2) NPDI Dataset: The NPDI database is a limited access database for research purpose only. This database contains 80 hours of 400 pornographic and 400 non-pornographic videos. This database was collected and annotated by researchers from Federal University of Minas Gerais (UFMG) in collaboration with University of Campinas, Brazil. In this database, three classes are provided: porn labeled, non-porn easy, and non-porn hard. The porn class consists of 400 videos and 200 videos for each of non-porn easy and non-porn hard classes. Non-porn easy means this video/frame consist a hard scene to categorized, e.g. wrestling scene where all wrestler wear no shirt. This scene alike with porn scene. This database will be used as testing data in this research.

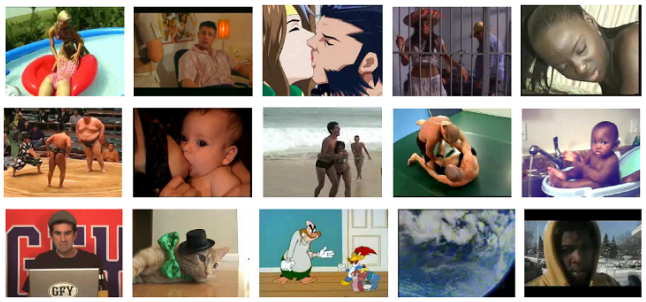

Fig. 1: *NPDI dataset samples [4]*

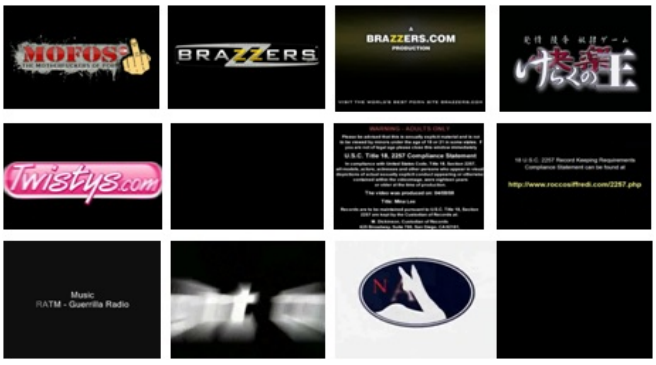

Fig. 2: *NPDI dataset samples [4]*

Figure 1 shows a few samples from the NPDI dataset that were extracted from video scenes. The first row is for porn class, the second row is for non-porn hard, and the last row is for non-porn easy. There are images limitation which can be shown in every publication and also its distribution. They provide both videos and extracted images with its own classes. In this dataset, there are three classes as mentioned above. In this paper we use only porn class for the testing process. However, main purpose of this database is to tackle adult video content, which means that the data was extracted from video

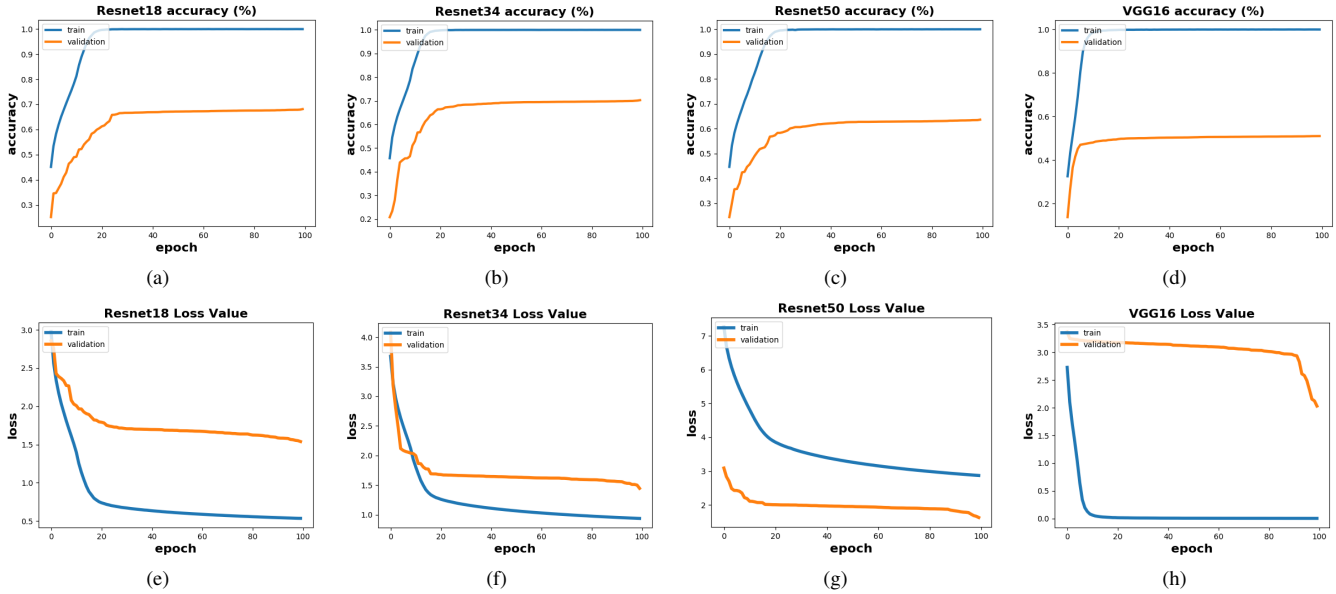

Fig. 3: *Training and Validation performance comparison, a-d) Accuracy for Resnet18, Resnet34, Resnet50 and VGG16, e-h) loss value for Resnet18, Resnet34, Resnet50 and VGG16*

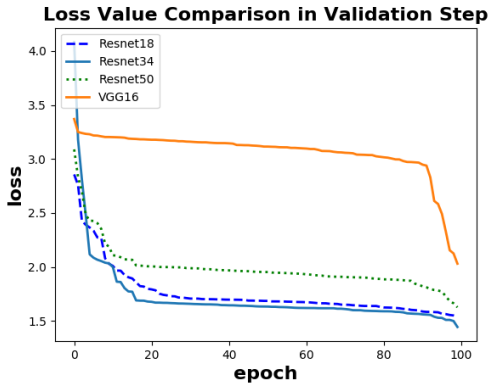

Fig. 4: *Loss value comparison for adult content detection*

TABLE III: Hyper-parameter setup

No	Hyper-parameter	Resnet	VGG16
1	Optimizer	ADAM	ADAM
2	Learning Rate (α)	1×10^{-2}	$\times 10^{-2}$
3	Decay of α (γ)	0.8	0.9
4	Batch Size	32	10
5	Loss parameter	Categorical Cross entropy	Categorical Cross entropy

TABLE IV: Performance Comparison

No	Network	Computational Time (second)	Description (hour)
1	Resnet_18	262.721	72,98
2	Resnet_34	345.851	96,07
3	Resnet_50	501.302	139,25
4	VGG16	518.654	144,07

keyframes. So, this data possesses a non-porn image in a porn class. For example, in the porn class they posses a credit scene which show the black screen with credits or any other text as shown in Figure 2. This condition will make our network confused while we only try to classify a single image. In this condition, we will separate this type of images to avoid image miss-classification in the testing process. The original frames amount of this class is 6387 images, including the credits scene. We found at least 250 images in the porn class that contain the same characteristic images like in Figure 2.

B. Computation Environment

Our experiment used Nvidia Titan X with 12 GB of RAM and Nvidia Tesla-K20 with 6 GB of RAM. These devices were installed under Linux Ubuntu 14.04 operating system. In training, validation, and testing processes, we used Keras [17] as the main library and TensorFlow [18] as computational backend. During the training process, it took over 36 GB of RAM with regards to the image dimensions.

IV. Experimental Result

The main challenge for nowadays conditions in term of adult contents filtering is data growth and also data diversity. In this experiments we used a dataset that is commonly accessed. This modular system can be retrained and we can add more data to train. In this case, we try to trade some accuracy for speed to get faster performance with limited computation resource. With regards to its training process, Table IV shows us the computational times for training and validation process. ResNet50 and VGG16 took over 6 days to reach 100 epoch, and 3 days and 4 days for ResNt18 and ResNet34, respectively.

Figure 3 shows all accuracy and loss values from ResNet

TABLE V: Performance Comparison

No	Network	Validation Loss	Accuracy (%)
1	Resnet_18	1,31	73,31
2	Resnet_34	1,23	**75,08**
3	Resnet_50	1,52	62,90
4	VGG16	2,45	59,60

and VGG16. ResNet 34 achieved the highest accuracy in validation process with 75% and loss value 1.23. We have a really big gap between training and validation process. This makes sense, since we have really diverse data. The main goal of our experiment is to get a good network for multy-platform implementation. Since mobile phone is one of our targets, we avoid a complex network to implement. For example, we limit ResNet till 50 layers and never use ResNet with 101 layers due to its complexity. In the training process, more complex networks will impact to training time as well as model size and resource availability. VGG16 has more than 1 billion parameters to train. To train the VGG16 network, we use 10 of batch size due to memory limitation in our GPU. Different with ResNet, we use 32 of batch size, with less than 20 million parameters.

Figure 4 shows us the loss value comparison between ResNet 18, ResNet34, ResNet50, and VGG16. This loss value taken during the validation process. As we can see, ResNet34 reached the lowest loss value and it outperformed the other networks. In our experiment, ResNet50 got lower accuracy and higher loss value compared to ResNet 34. This means that the network's depth does not assure the performance quality. Figure 3.g shows an anomaly condition where the training loss value is higher than the loss validation value. The other reason is epoch number, where we use only 100 epochs due to computational power limitation. Since we train our network with no pre-trained model and train from scratch, weight initialization is an important part that we miss. In the next work, pre-trained model will possibly be used.

As we mentioned in the dataset section, the testing process used the NPDI dataset. The certain class was chosen as our testing data. We only use porn class with total 6137 images that extracted and provided by this database. The main reason to choose only porn data is to ensure that our network can detect adult content on which it never learned before. Detecting adult contents means detecting image contain adult part and ignore other objects (whole object in the world). The testing performance shows that ResNet 34 can detect better than other network with 69% of accuracy. This is reasonable since in the training process ResNet34 outperformed other network.

V. Conclusion

From the performance comparison we can see that ResNet with 34 layers reaches the greatest accuracy and the lowest error rate. This network outperformed other networks although it has a lower depth than ResNet50 and VGG16. In the validation process ResNet34 reaches 75.08% of accuracy and in the testing process on the NPDI dataset Resnet34 reaches $69,02\%$ of accuracy for nude class only.

Acknowledgment

The authors would like to thank to Nikola Banić for insightful discussion and proofreading process to improve the paper quality, We also would like to thank to all reviewers.

References

[1] Badan Pusat Statistik, "Statistik Kriminal 2014," Katalog BPS:4401002, 2014.

[2] Metrotvnews, "Pelaku di Bawah Umur Kasus Pemerkosaan di Bengkulu Wajib Direhabilitasi,"http://news.metrotvnews.com/read/2016/05/04/523322/pelaku-di-bawah-umur-kasus-pemerkosaan-di-bengkulu-waji

[3] Pornhub Insight Review 2016, https:www.pornhub.cominsights2016year-inreview

[4] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, Arnaldo de A. Arajo. Pooling in Image Representation: the Visual Codeword Point of View. Computer Vision and Image Understanding (CVIU), volume 117, issue 5, p. 453-465, 2013.

[5] Jonatas Wehrmann, Gabriel S. Simes, Rodrigo C. Barros, Victor F. Cavalcante. Adult Content Detection in Videos with Convolutional and Recurrent Neural Networks. Neurocomputing, in press, 2017

[6] A. Lopes, S. Avila, A. Peixoto, R. Oliveira, A. Arau jo, A bag-of-features approach based on hue-SIFT descriptor for nude detection, in: European Signal Processing Conference (EUSIPCO), 2009, 11521156.

[7] C. Steel, The mask-SIFT cascading classifier for pornography detection, in: IEEE World Congress on Internet Security (WorldCIS), 2012, 139142.

[8] T. Deselaers, L. Pimenidis, H. Ney, Bag-of-visual-words models for adult image classification and filtering, in: IEEE International Conference on Pattern Recognition (ICPR), 2008, 14

[9] A. Ulges, A. Stahl, Automatic detection of child pornography using color visual words, in: IEEE International Conference on Multimedia and Expo (ICME), 2011, 16.

[10] J. Zhang, L. Sui, L. Zhuo, Z. Li, Y. Yang, An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain, Neurocomputing 110 (2013) 145152.

[11] S. Avila, N. Thome, M. Cord, E. Valle and A. de A. Arajo, "BOSSA: Extended bow formalism for image classification," 2011 18th IEEE International Conference on Image Processing, Brussels, 2011, pp. 2909-2912.

[12] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha. Pornography Classification: The Hidden Clues in Video Space-Time. Forensic Science International, volume 268, p. 46-61, 2016.

[13] He, K., Zhang, X., Ren, S., and Sun, J.,"Deep Residual Learning for Image Recognition." https://arxiv.org/abs/1512.03385, 2015.

[14] Simonyan, K., and Zisserman, A., "Very deep convolutional networks for large-scale image recognition." http:arxiv.orgabs pp.1409-1556, 2014.

[15] Russakovsky, O., Deng, J., Hao, S., Krause, J., Satheesh, S., Ma, S., Huang, Z., and Karpathy, A., Khosla, A., Bernstein, M., Berg A.C., and Fei-Fei, L.:"ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision (IJCV), vol.115, no. 3, pp. 211-252, 2015.

[16] Diederik Kingma, and Jimmy Ba: Adam: A Method for Stochastic Optimization, https://arxiv.org/abs/1412.6980v8, 2014.

[17] Chollet, François and others:"Keras," https://github.com/fchollet/keras, 2015.

[18] A. Martn , and others:"TensorFlow: Large-scale machine learning on heterogeneous systems," Software available from tensorflow.org., 2015.