2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)

An Automatic System for Identifying and Categorizing Tribal Clothing Based on Convolutional Neural Networks

Ashraful Islam
Department of Computer Science and Engineering
Sylhet Engineering College
Sylhet,Bangladesh
ashraful.islam@sec.ac.bd

Tuhin Chowdhury
Department of Computer Science and Engineering
Sylhet Engineering College
Sylhet,Bangladesh
armoytuhin@gmail.com

Mehrab Hossain
Department of Computer Science and Engineering
Sylhet Engineering College
Sylhet,Bangladesh
hmehrab786@gmail.com

Nafiz Nahid
Department of Computer Science and Engineering
Sylhet Engineering College
Sylhet,Bangladesh
nafiz234.NN@gmail.com

Ariful Islam Rifat
Department of Computer Science and Engineering
American International University,
Dhaka Bangladesh
aarifulrifat548@gmail.com

Abstract*—The quantity of internet businesses providing tribal clothes is constantly increasing, and people tend to exaggerate how often they shop at such sites. However, we are concerned about the authenticity of the outfits. The study recommends using Convolutional Neural Networks (CNN) to automatically identify and categorize authentic images of particular tribal dresses used by some Bangladeshi tribes into predetermined categories. The study's impetus comes from the expansion of commerce and the desire to spread these traditional clothes over the globe. In order to categorize the clothing, we obtained images from actual tribal residences, shops, and a few online marketplaces. To that end, we made an effort to provide a dataset we've labeled "TribalBd," which has 680 samples, including six different classes. Then, use the YOLOv5, YOLOv6, and YOLOv7 models to put these datasets for detection and classification on our CNN. As a means of evaluating the efficacy of our model, we have experimented with a number of different CNN topologies and tweaks. We put the model through its tests with YOLOv6 and YOLOv7. YOLOv5 achieved the best results among these models. The final result shows that the YOLOv6 model gives 86.24%, the YOLOv7 model gives 71.28% accuracy whereas YOLOv5 gives 89.97% accuracy in classifying the images in the training and testing sets which are best compared to the other two models.

Keywords— Tribal dress, Monipuri, Chakma, Convolution Neural Network (CNN)

I. INTRODUCTION

Bangladesh has an extensive range of cultural variety. The Indian subcontinent has long been home to numerous tribal tribes. They gave rise to the roughly 50 tribal groups now visible in Bangladesh. Significant tribes include the "Manipuri," "Chakma," "Garo," "Sawtal," etc. The majority of these ethnicities reside in the country's hilly regions. This community participates in several unique cultural and religious events where members dress traditionally. Because communication and transportation have improved, tourists visit these places more often than they used to. Particularly in the area where they sell hand-made traditional clothing, they demonstrate an interest in learning more about their culture.

Recent years have seen an upsurge in interest in these traditional dresses, causing a ripple effect through the clothing industry. To take advantage of the growing popularity of the internet and the opportunities it presents for businesses to advertise their wares, many retailers now offer their consumers the option of shopping for apparel online. However, most purchasers have been cheated into purchasing unsatisfactory products. The traditional tribal clothes of Bangladesh cannot be categorized, sorted, or recognized by any such system at this time. If there were a technology that could automatically recognize traditional dress based on photos, it would be beneficial for both traditional dress enthusiasts and authentic producers. It is a remarkable way to share these genuine traditional garments with the globe. With just a picture, they could learn what each outfit was called, find the item that interests them, and then purchase it.

Among the many tribes that make up Bangladesh, the Manipuri and the Chakma are among the most well-known. They are also renowned for their clothing, as the women's clothing they sell is extremely popular in Bengali. The "Manipuri Sari" is famous among women because of its high-quality construction, spectacular design, and comforting feel. Bangladeshis love this tribal product so much that it has become the country's most outstanding seller. In addition, several "Chakma" dresses are also quite fashionable among Bangladeshi ladies. Within the Chakma culture, "Chakma Pinon" and " Chakma Hadi" shown as Fig: 1 are the most popular forms of clothing. As of late, it has also been widely accepted among other groups in our country. In this study, we utilized several dresses from these two groups. Recent studies have shown that deep learning outperforms traditional machine learning techniques, especially for picture recognition from massive datasets. Our study utilizes a Convolutional Neural Network (CNN) to categorize the various clothing worn by indigenous peoples in Bangladesh.

We have experimented with a wide range of possible optimizers to determine the most efficient CNN design regarding resource consumption. E-commerce sites will use our system.

2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) | 978-1-6654-5635-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICERECT56837.2022.10060409

Chakma Sari

Monipuri Maflar

Monipuri Fanek

Monipuri Sari

Fig. 1. Class demo

In Fig1: we can see some classes from our dataset TribalBD.

A. OBJECTIVE

- Detect some specific traditional dresses from images.
- Classify those dresses into several classes.
- Developing a unique dataset named "TribalBD" on the tribe's traditional clothing.
- Preventing from buying and selling fake traditional clothing.
- Finally, to introduce these handmade traditional tribal dresses internationally.

II. LITERATURE REVIEW

Fashion Classification refers to categorizing garments based on photographs alone. Internet retail, social media, and even the law find a new use for this classification technique. To better find fashion products that are similar to those in a query image, the authors of [1] focused on improving the fashion classifier. This article's authors break down the fashion classification issue into four distinct parts. We may divide the task of categorizing garments into four distinct categories like Dress categorization, dress feature extraction, dress information extraction, and dress image recognition.[2] Multiclass categorization, known as Clothing Type, labels the query image with a description of the clothing type (such as a T-shirt or a jacket) that can be seen inside the image. Multiple properties, including garment colour or pattern, were classified as part of the clothing attribute set for the query image. The Clothing Retrieval system found the clothing items most comparable to the one in the query image.

The portions of the test image that contained the clothing object were found after an analysis of the image. It generated a rectangle that perfectly matched the attire in the picture. The default settings of CaffeNet were used, and the authors made changes to boost performance. The test set's clothing type classification has a 46.0% accuracy rate for the stated result. The Caffenet model, which was instructed to employ frozen layers everywhere outside the test set, was the one to which this specifically applied. Nevertheless, performance on the test set increased to 50.2% after fine-tuning all layers. The average accuracy for retrieving garments was found to be 74.5 percent, and the K-nearest neighbor algorithm was used Instead of using the raw pixels from the image, they used the second entirely connected layer, which is a considerably more limited strategy. The test set's overall accuracy was 40.2%. The greatest validation accuracy for clothing object detection during phase one of training was 91.25 percent. Accuracy rose to 93.4% during the second training phase, and not as a separate article .

The ACS dataset was utilized by Brossard et al. (2012) to gather characteristics such as color information, the Histogram of Oriented Gradients (HOG), Speeded Up Robust Features (SURF), and Local Binary Patterns (LBP)[3]. In order to conduct multiclass classification, Brossard used these characteristics in conjunction with One vs. All SVM, random forests, and transfer forests, reaching average accuracy levels of 35.03%, 38.29%, and 41.36%, respectively. We outperform these accuracy levels on our own data set using CNN.

In the study,[4] concentrated on categorizing mobile phone-taken pictures of clothing. The authors suggested a framework that considered the different levels of fashion items. To accomplish this, they employed "layered deep convolutional networks." Hoodies are positioned lower on the hierarchy of "tops" following their design. One of the biggest e-commerce platforms in Japan, ZOZOTOWN, has gathered 811,226 clothing image sets, which include image data and categories. After processing the data, they had 129,749 image sets, which they put into three main categories and ten subcategories and labelled. They added another set of data on top of these to enable robust classification of random smartphone photos. In total, 160 images of clothing were provided by 20 volunteer participants, who used their smartphones to take the pictures. Two different CNN models were used to train the primary category and the subcategory. In terms of the outcome, they achieved 55.9% mean accuracy in the subcategory and 92.7% mean accuracy in the main category.

Wang, Yang, Mao, Huang, Huang and Xu [5] has suggested a procedure, that made the CNN-RNN framework by combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The CNN-RNN framework sorts multi-label images into groups using both CNN and RNN. A hybrid image and label embedding space were employed to simulate the co-occurrence dependency of labels in their solution, which relied on CNNs and RNNs

Parsing clothing in fashion photographs [6] suggested a technique for identifying apparel that uses two to three classifiers for each pixel. The ultimate prediction is then created by combining all the results. Only 685 of the 158,235 photos in the dataset, which the authors made, were used to validate the system. In paper[7] they focused on the over grown popularity of the dresses over internet. Our tribal dresses also gain some popularity through interenet.

M. M. Tanzim Nawaz, Rasik Hasan, Md. Abid Hasan, Mahadi Hassan, and Rashedur M. Rahman[8] work on some traditional Bangladeshi dresses. Their dataset contains around 1500 images collected from several online shops and local stores. They gain a maximum of 92.05% accuracy using the CNN-based Inception V3 architecture.

III. Methodology

Deep Learning uses artificial neural networks (ANNs) called CNNs or ConvNets analyze images[9].In addition to being known as "shift invariant" or "space invariant" artificial neural networks (SIANN), CNNs are also known as "shared-weight" Convolutional Neural Networks (convolution kernels or filters). Researchers test CNNs on photos that haven't been touched up, so they don't have to decide which features should be prioritized. Because CNN's only perform convolution on tiny regions of the input space and communicate parameters among themselves, they require fewer parameters than fully linked networks. It also takes a different approach to regularization by using the hierarchical patterns in the data and putting together increasingly complex patterns by employing smaller and easier patterns that are imprinted in their filters. Therefore, CNNs are near the bottom of a scale measuring connectedness and complexity. For classification, different CNN-based architectures have been made and put into two groups: Multi-Stage (RCNN, Fast RCNN, Faster RCNN) and Single-Stage (SSD, YOLO). Indeed, in this work, we used three CNN models from the YOLO architecture family, such as YOLOv5, YOLOv6, and YOLOv7, whereas from these three, YOLOv5 performs relatively better than the other two.

A. *Our Used Model*

Because it is a single-stage object detector with the standard three-part architecture, we based our work on the YOLO v5 model which has three parts [10]

1. *Model Backbone:* Using the provided input image, the BackBone model primarily pulls out the most crucial characteristics. The CSP-Cross Stage Partial Networks serve as the framework for creating image features in YOLOv5 [11]. The CSP solves problems with the same gradients in other, more extensive ConvNet backbones. This means that there are fewer parameters and fewer FLOPS. DenseNet is the foundation for the CSP models. By connecting layers in a CNN with DenseNet, we may lessen the impact of the vanishing gradient problem, boost feature reuse, and cut down on the number of network parameters.Here, Fig 2: Examples of the Cross Stage Partial DenseNet (b) (CSPDenseNet). The base layer's feature map is divided into two pieces by CSPNet, one of which will pass through a heavy block and a transition layer before being mixed with the transmitted feature map at the following stage [11].

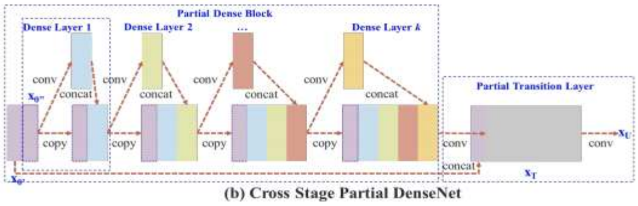

Fig. 2. Cross-Stage Partial DenseNet

2. *Model Neck:* Its aim is to produce feature pyramids. These features enable models to scale objects successfully in general. The ability to recognize the same thing in various sizes and scales is helpful. Fig3: [12]: Feature network design[13] – (a) In FPN [14], a top-down pathway combines multi-scale characteristics from levels 3 to 7 (P3-P7). In PANet [15], a bottom-up route is added on top of FPN. In NAS-FPN [16], a neural network model search is used to find an unusual functionality topology of the network, and then the same block is used repeatedly. In our BiFPN, there is improved accuracy and performance exchange.

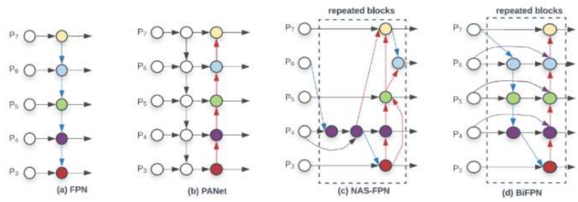

Fig. 3. Feature network design

3. *Model Head:* Three alternative scales of feature maps are generated to achieve multiscale prediction, improving the model's prediction of small, medium, and oversized items.For context, the model has 191 layers and 7.46816e+06 parameters and gradients.

B. *Workflow Diagram*

We split the data into three parts a test, train, and validation.Fig 4: shows the steps we followed in this work. Following that, we perform some processing on the raw images, such as resizing, rotating, rescaling, and cropping, and then we send those images to the model so that it can train in our system using those photos. We used three different versions of the YOLO architecture for this classification: YOLO v5, YOLO v6, and YOLO v7.

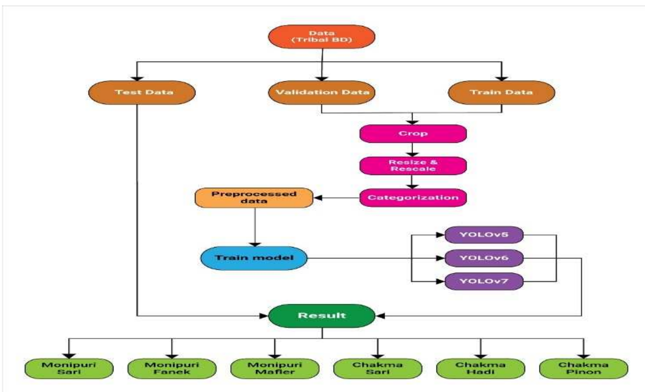

Fig. 4. Feature network design

C. Dataset

In our "TribalBD" dataset we divided the six items of clothing into two groups that best represent Bangladeshi traditional tribal dress (Monipuri, Chakma). The following are listed in chronological order as we can see in Figure 5: 1) Monipuri Sari, 2) Monipuri Fanek, 3) Monipuri Maflar, 4) Chakma Sari, 5) Chakma Hadi, and 6) Chakma Pinon. No dataset with photos of these things in advance was available for our work. So, it is necessary for us to collect pictures from numerous places, stores, and internet retailers. We managed to gather a total of 608 images. 75% of the data is used for training, 15% for validation, and 10% for testing. As a result, we used 68 images for the test set, 110 for the validation set, and 508 for the training set .Details distribution shown in table 1.

Monipuri Maflar

Monipuri Sari

Monipuri Fanek

Chakma Pinon

Chakma Sari

Chakma Hadi

Fig. 5. Image of all Six Classes

TABLE I. CATEGORY WISE THE IMAGES ON DATASET

Category	***Training Set***	***Validation Set***	***Test***	***Target label***
Monipuri Sari	169	35	17	Monipuri_Sari
Monipuri Fanek	163	12	8	Monipuri_Fanek
Monipuri Maflar	55	12	7	Monipuri_Maflar
Chakma Sari	108	33	18	Chakma_Sari
Chakma Hadi	68	11	10	Chakma_Hadi
Chakma Pinon	91	14	12	Chakma_Pinon

IV. IMPLEMENTATION & RESULT

Python deep learning is used to realize the design. For example, Facebook developed and made available the open-source machine learning algorithm known as Pytorch. Our system was tested on our "TribalBd" dataset, which consists of a training set, test set, and validation set.

To conduct the tests, we utilized photos that were 416 x 416 x 3. The YOLOv5 architecture is defined by its specific configuration, which includes

```
nc: 80  # number of classes
depth_multiple: 0.33  # model depth multiple
width_multiple: 0.50  # layer channel multiple
anchors:
  - [10,13, 16,30, 33,23]  # P3/8
  - [30,61, 62,45, 59,119]  # P4/16
  - [116,90, 156,198, 373,326]  # P5/32
```

Anchor

```
# YOLOv5 v6.0 backbone
backbone:
  # [from, number, module, args]
  [[-1, 1, Conv, [64, 6, 2, 2]],  # 0-P1/2
   [-1, 1, Conv, [128, 3, 2]],  # 1-P2/4
   [-1, 3, C3, [128]],
   [-1, 1, Conv, [256, 3, 2]],  # 3-P3/8
   [-1, 6, C3, [256]],
   [-1, 1, Conv, [512, 3, 2]],  # 5-P4/16
   [-1, 9, C3, [512]],
   [-1, 1, Conv, [1024, 3, 2]],  # 7-P5/32
   [-1, 3, C3, [1024]],
   [-1, 1, SPPF, [1024, 5]],  # 9
  ]
```

Backbone

```
# YOLOv5 v6.0 head
head:
  [[-1, 1, Conv, [512, 1, 1]],
   [-1, 1, nn.Upsample, [None, 2, 'nearest']],
   [[-1, 6], 1, Concat, [1]],  # cat backbone P4
   [-1, 3, C3, [512, False]],  # 13

   [-1, 1, Conv, [256, 1, 1]],
   [-1, 1, nn.Upsample, [None, 2, 'nearest']],
   [[-1, 4], 1, Concat, [1]],  # cat backbone P3
   [-1, 3, C3, [256, False]],  # 17 (P3/8-small)

   [-1, 1, Conv, [256, 3, 2]],
   [[-1, 14], 1, Concat, [1]],  # cat head P4
   [-1, 3, C3, [512, False]],  # 20 (P4/16-medium)

   [-1, 1, Conv, [512, 3, 2]],
   [[-1, 10], 1, Concat, [1]],  # cat head P5
   [-1, 3, C3, [1024, False]],  # 23 (P5/32-large)

   [[17, 20, 23], 1, Detect, [nc, anchors]],  # Detect(P3, P4, P5)
  ]
```

Head

Fig. 6. Model Configuration

YOLOv5 uses 3 different size anchor boxes. Anchor box size shows in Fig 6.1. It is scale and aspect ratio of specific classes object detection. FPN has three output scale in example: P3/8 for small, P4/8 for medium, P5/32 for large object detection. Backbone consists four different convolutional layers Fig 6.2 with four different arguments.

Adam optimizer was selected automatically, but we tried it with the default optimizer, a batch size of 16, a learning rate of 0.001, and 800 iterations of training to guarantee convergence.

A. Result

In Table 2: we can see YOLOv5 and YOLOv6 perform relatively close where the F1 difference is 0.016 and the mAP difference is 0.011.

TABLE II. MODEL PERFORMANCE

Model	***Precision***	***Recall***	***F1***	***mAP***
YOLOv5	0.754	0.886	0.814	0.873
YOLOv6	0.751	0.852	0.798	0.862
YOLOv7	0.729	0.657	0.691	0.712

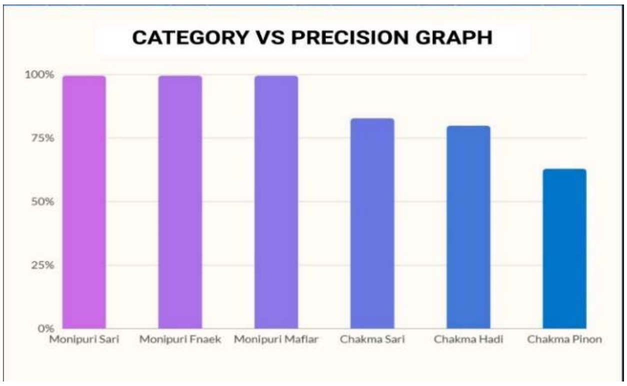

Fig. 7. Precision Graph for Category based on YOLOv5

We have 6 categories Fig 7: Monipuri Sari(99.5%), Monipuri Fanek(99.5%), Monipuri Maflar(99.5%), Chakma Sari(82.7%), Chakma Hadi(79.8%), Chakma Pinon(62.8%). The precision graph is given:

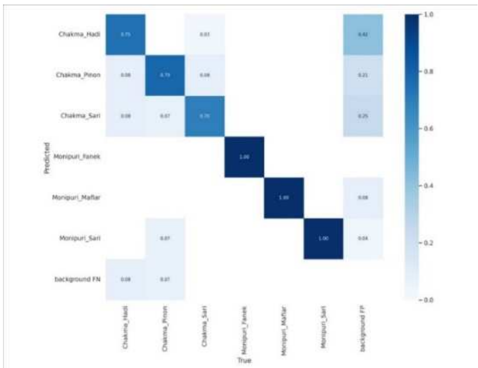

Fig. 8. Confusion Matrix for YOLOv5

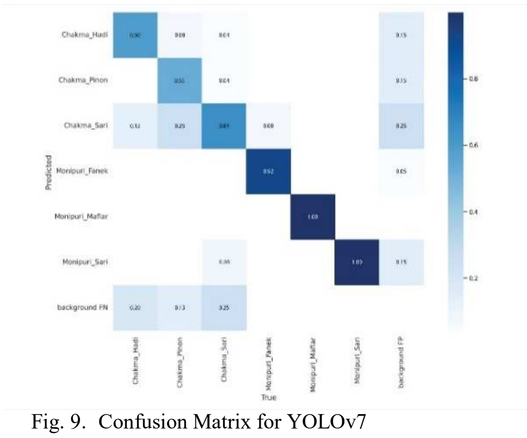

Fig. 9. Confusion Matrix for YOLOv7

From this confusion matrix Fig 8:, we get accuracy for each class. Accuracy is calculated as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \text{ [17]}$$

The accuracy for the six classes are Monipuri Sari (99.99%), Monipuri Fanek(99.99%), Monipuri Maflar(99.99%), Chakma Sari(74.2%), Chakma Hadi(81.2%), Chakma Pinon(83.04%). With this model, we see an overall accuracy of 89.97%.

After calculating the accuracy from YOLOv7 confusion matrix shown in Fig 9: we get 71.28% accuracy. We see that the True Positive value of every classes in YOLOv7 is lower than the YOLOv5. So, our obvious choice is YOLOv5.

Fig. 10. Test images with bounding box

In Fig 10: we can see Chakma_Sari and Chakma_Hadi has some similar patterns and due to some low-quality image, our model fails to differentiate those classes. This the reason for arising typo errors.

V. Conclusion & Future Work

In this study, we proposed a method to automatically identify and classify traditional dresses using three CNN models that can detect the tribal dress precisely. We used our raw data to train a deep learning model, and its accuracy was compared to that of other models. With the YOLOv5 model and a raw dataset (TribalBD), we get the best results of the three models we looked at here. In the future, we want to spend more time and effort on this research. The data set can be expanded by adding more tribal groups from Bangladesh, like the Marma, Garo, Shawtal, etc., and by adding more types of clothing that can help classify edible clothing.

References

[1] B. Lao and K. Jagadeesh, "Convolutional Neural Networks for Fashion Classification and Object Detection".

[2] S. Hiriyannaiah, G. M. Siddesh, and K. G. Srinivasa, "Deep visual ensemble similarity (DVESM) approach for visually aware recommendation and search in smart community," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2562–2573, Jun. 2022, doi: 10.1016/J.JKSUCI.2020.03.009.

[3] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. van Gool, "Apparel classification with style," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7727 LNCS, no. PART 4, pp. 321–335, 2013, doi: 10.1007/978-3-642-37447-0_25/COVER.

[4] K. Hori, S. Okada, and K. Nitta, "Fashion image classification on mobile phones using layered deep convolutional neural networks," *ACM International Conference Proceeding Series*, pp. 359–361, Dec. 2016, doi: 10.1145/3012709.3016075.

[5] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A Unified Framework for Multi-Label Image Classification." pp. 2285–2294, 2016.

[6] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3570–3577, 2012, doi: 10.1109/CVPR.2012.6248101.

[7] K. Yamaguchi, T. L. Berg, and L. E. Ortiz, "Chic or social: Visual popularity analysis in online fashion networks," *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pp. 773–776, Nov. 2014, doi: 10.1145/2647868.2654958.

[8] M. M. Tanzim Nawaz, R. Hasan, M. A. Hasan, M. Hassan, and R. M. Rahman, "Automatic Categorization of Traditional

Clothing Using Convolutional Neural Network," *Proceedings - 17th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2018*, pp. 98–103, Sep. 2018, doi: 10.1109/ICIS.2018.8466523.

[9] M. v. Valueva, N. N. Nagornov, P. A. Lyakhov, G. v. Valuev, and N. I. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Math Comput Simul*, vol. 177, pp. 232–243, Nov. 2020, doi: 10.1016/J.MATCOM.2020.04.031.

[10] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNET: A NEW BACKBONE THAT CAN ENHANCE LEARNING CAPABILITY OF CNN A PREPRINT," 2019, Accessed: Dec. 05, 2022. [Online]. Available: https://github.com/WongKinYiu/CrossStagePartialNetworks.

[11] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A New Backbone That Can Enhance Learning Capability of CNN." pp. 390–391, 2020.

[12] M. Tan, R. Pang, and Q. v. Le, "EfficientDet: Scalable and Efficient Object Detection," pp. 10781–10790, 2020, Accessed: Nov. 05, 2022. [Online]. Available: https://github.com/google/

[13] "11.04 PANet (Path Aggregation Network) - BoS - Deep Learning Bible - 3. Object Detection - 한글", Accessed: Nov. 05, 2022. [Online]. Available: https://wikidocs.net/163813

[14] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," pp. 2117–2125, 2017.

[15] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," pp. 8759–8768, 2018.

[16] G. Ghiasi, T.-Y. Lin, and Q. v. Le, "NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection," pp. 7036–7045, 2019.

[17] A. E. Maxwell, T. A. Warner, and L. A. Guillén, "Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 1: Literature review," *Remote Sens (Basel)*, vol. 13, no. 13, Jul. 2021, doi: 10.3390/RS13132450.