Exploring Augmentation Techniques to Classify Clothes in Low-Quality Images

Rebekah Jennifer
School of Computing
Dublin City University, Ireland
rebekah.manimaran2@mail.dcu.ie

Priya Dharshini Balaji
School of Computing
Dublin City University, Ireland
priya.balaji2@mail.dcu.ie

Graham Healy
Dublin City University
Insight Centre for Data Analytics, Ireland
lncs@springer.com

***Abstract*—Clothes are vital components in the field of e-commerce and forensic applications. Clothes are considered to be integral visual attributes because they can be used to describe people. Attribute-based identification systems are essential for forensic investigations because they help in identifying an individual. This requirement has created a need for an automation system that recognizes different apparel in images and videos. But most of the images/videos from CCTV or candid cameras lack resolution and quality. In this work, we propose an approach to classify and detect different clothing categories using data augmentation for real-world scenarios. Our method uses an annotated large scale dataset called DeepFashion to train and benchmark on clothing apparels of 10 types. This method helps to improve the accuracy of distinguishing garments that generally contains low resolution and unexpected posture of an individual. In this scenario, we evaluate the performance on a set of image collections extracted from the ILIDS Video – reIDentification dataset, YouTube pose dataset, Clothing Attribute dataset and creative common licensed google images and video frames.**

***Index Terms*—Clothing classification, Object detection, Data Augmentation, DeepFashion, Deep Learning, ResNet34, YOLO, EfficientNet**

I. INTRODUCTION

Clothing identification has always been a significant factor in identifying humans and recognizing different types of attire in the fashion industry. In recent times, apparels are created in a wide variety in fashion and e-commerce world. Therefore, clothing is a vital descriptor to describe a human being, e.g. ”the man with the red cap” or ”the woman in the striped top”. With the increase in online influencers and a massive amount of video content and movie collections, clothing images can be used for fashion similar image retrieval and outfit recommendations. At the same time, another crucial part when it comes to clothing recognition is analyzing real-time videos such as surveillance cameras, where clothing details can be used to pinpoint suspected criminals or missing individuals [10].

Previously, clothing identification and detection used a small volume of the dataset on their research studies. Due to the unrealistic situations with less amount of data and good quality of images, clothes were not classified based on real-world scenarios. The existing analysis of fashion images handled adopting gender, age, skin, colour and texture with region growing method of segmentation [10] and Voronoi image method [18].

Fig. 1 Data Augmentation Strategy 1 and 2

Based on these drawbacks, we propose a new approach that is effectively able to recognize the clothes on the low-quality images like those from a CCTV camera. Here, we introduce a method that adequately classifies and detects clothing apparel trained using DeepFashion in combination with data augmentation techniques.In Fig. 1, various data augmentation techniques used to train the clothing images are displayed.

In this, we perform extensive experiments on the clothing images that are under non-ideal conditions, explicitly when the image contains noise (e.g. when parts of the image are occluded or blurred). The main contributions involved in our work are: (1) We built a custom dataset from multiple public reusable sources to perform our evaluation and benchmark the state of art models. (2) We developed a classification approach using ResNet[19] and EfficientNet[20] for training on the DeepFashion[6] dataset. The predicted category type showed us the learned features of the clothes and the variation

of regularization between two CNN architecture (ResNet and EfficientNet). (3) We performed an object detection technique using YOLO to observe the localization of clothing items (4) Also, we show two different successful combinations of data augmentation strategies in the classification task that increased the accuracy by 5.3% in ResNet and 2.3% in EfficientNet. Likewise, the detection model also implements data augmentation techniques that have demonstrated to generalize well in real-world tasks. All these research methods applied with data augmentation strategies aim to overcome the deformation and quality of images. From the results, we show a comparative analysis and effectiveness between different deep learning architectures and data augmentation techniques.

II. RELATED WORK

In the earlier years, clothing identification was based on the features that were learned from the shapes of clothes. They included the use of handcrafted features combined with machine learning methods. Those tasks involved several sub-tasks such as localization of human figures, clothing segmentation and alignment and extraction of clothing representation. Wang et al. [1] use clothing shapes to build a Bayesian model to segment clothes. Di et al. [2] use several feature extractors types like SIFT, LBP, HOG to train linear SVMs for clothing style recognition and retrieval.

In a later period, many works were involved based on deep learning models to improve the accuracy of the attribute prediction. Lao et al. [3] use CNN models to outperform the SVM, random forests, and transfer models of Bossard et al [4]. Transfer learning (VGG, AlexNet, and ResNet) and data augmentation [5] have tackled the issues of clothing attribute prediction. The same work [5] has also tried clothing classification by hyper-parameter tuning and Bayesian Optimization Algorithm.

The main idea of this experiment has been evolved from Liu et al. [6], who constructed a unified dataset called DeepFashion. DeepFashion is a large-scale clothing dataset with comprehensive annotations that contains 800,000 images with a large number of attributes, landmarks and cross pose/cross-domain images. The authors used this dataset to support rapid development in clothing identification to develop a deep model called FashionNet, similar to VGG-16 [7]. This approach has been assessed based on top-k retrieval accuracy and loss functions. Models used in Where To Buy It (WTBI) [8] and Dual Attribute-aware Ranking Network (DARN) [9] have been compared with FashionNet which are trained on clothing images.

In recent times, due to the demand for forensic applications, clothes need to be classified in surveillance videos. Feris et al. [17] provided the underlying architecture, which involves background modelling for background subtraction to detect foreground objects while attribute detectors are added to improve the accuracy. Yang et al. [10] also provided solutions for addressing this problem with real-time surveillance footage. Bedeli et al. [11] use AlexNet to recognize clothing in surveillance videos.

Ruyue Liu [29] performed pedestrian clothing detection using YOLOV3. This model is a one-stage method with the extraordinary speed at the cost of accuracy to detect clothes on pedestrians, so the model worked in a way where it divided the categories into skirts and not-skirts to determine the position of skirts. Jan Cychnerski et al. [14] performed the classification and detection of clothing images to build an automatic e-commerce product tagging system. This model uses SSD, SqueezeNet, ResNet-50 along with background augmentation strategy.

Brian Lao and Karthik Jagadeesh [15] focused on four tasks in fashion classification. The first one is the multiclass classification of clothing type, the second one is clothing attribute classification, the third one is clothing retrieval of nearest neighbours and the last one is clothing object detection. In clothing type classification Apparel classification with style (ACS) dataset [26] has been used with an AlexNet CNN architecture, while in Clothing object detection Colourful Fashion (CF) dataset [27] has been used and trained based on a selective search on fine-tuned RCNN network. Besides, works like [16] adopted a modified version of Faster RCNN with ResNet 101 and a pruning mechanism using the DeepFashion dataset for fashion attribute detection. [13] Started initial steps towards clothing recognition in movies using AVA-fashion dataset [28]. They also experimented with consumer photos for clothing classification that are used to tag clothing items using the Amazon catalogue dataset. This method utilized ResNet 50 as a backbone architecture.

The main objective of this study is to use distinct and straightforward data augmentation methods to enhance clothing recognition in situations where images have low quality and resolution. Specifically, the evaluation of the study is divided into three main parts. The first part consists of clothing classification in good quality DeepFashion dataset and a custom testset. The second part is clothing detection in DeepFashion dataset and Custom Testset. All these parts are implemented by two appropriate data augmentation strategies, which involves a combination of augmentation methods performed in a balanced approach. The overall aim is to evaluate the performance of the clothing identification system in poor quality clothing images using ResNet-34 and EfficientNet-B3, which replicates real-world scenarios.

III. DATA

We use the DeepFashion dataset, a large-scale clothes dataset, that has over 800,000 diverse fashion images ranging from well-posed shop images to unconstrained consumer photos, and images from e-commerce sites. Each image in this dataset is labelled with 50 categories. This data and annotations belong to four benchmarks such as Category Classification, In-shop Clothes Retrieval, Consumer-to-shop, Clothes Retrieval and Fashion Landmark Detection.

A. Train dataset

Here, the Category and Attribute Prediction module is used, which is a large subset of DeepFashion dataset. It contains

289,222 clothes images under 50 clothing categories. Each image is annotated by bounding box and clothing type. On performing exploratory data analysis, an imbalance in the number of images per class was found. In order to produce a balanced dataset, the top 10 categories with more than 10k images are used. This structure helps to improve the balance of the model. Thus, 224k images spread across ten classes such as Dress, Sweater, Jacket, Tank, Top, Tee, Skirt, Shorts, Cardigan and Blouse are used for training.

B. Test dataset

The custom test dataset has 360 images on ten clothing categories. This is collected from various source combination of the ILIDS Video – re-Identification dataset, YouTube pose dataset, Clothing Attribute dataset, Creative Common Licensed google images and Creative Common Licensed video frames. These image collections have been chosen based on noise, low resolution, low lighting and low pixel count. These images were taken to reproduce real-life scenarios and have been annotated accurately using human annotators. There are 60 images from ILIDS dataset, 100 images each from the YouTube pose dataset, Clothing Attribute dataset, Creative Common Licensed google images and Creative Common Licensed video frames. The dataset is split in a way that there are at least 30 images per category.

IV. OUR APPROACH

Here, we evaluate multiple CNN pipelines trained on data augmentation strategies to predict the out of domain images (e.g. from CCTV). We use Transfer learning method where a model developed for a task is reused as the starting point for the second task of another model on clothes recognition. To initialize the weights of the neural network and to fine-tune them by updating weights on the DeepFashion dataset, we use ImageNet weights for classification models and COCO dataset weights for detection models. Using data augmentation on the images increases accuracy and reduces the error rate on the prediction. The data augmentation techniques used here explores various possibilities of a human pose, their angles, and the quality of the image in real-time. This style assists the features extracted from the convolutional neural network for efficient image classification and recognition. The output of the multiclass classification neural network and object detection framework is evaluated based on the ground truth labels and bounding box values for the ten clothing categories.

A. Data Augmentation

Data augmentation is a process of expanding the size, quality and diversity of data using various image manipulation techniques. When imagining real-life situations, we can understand how a human experience different kind of scenarios that are compared to CCTV footages and movie clips. With this augmented approach, more information can be extracted from the original image to create a better deep learning model architecture. Existing data augmentation methods are simple geometric transformations which are easier to implement and has shown proven results using various datasets in cropping, flipping and rotation [22]. In our method, we are exploring the effectiveness of distinct augmentation strategies, including the simple techniques for the clothes recognition task [12]. Here, we use multiple combinations of data augmentation techniques with different probabilities of implementation on the training images. Fastai library [25] is used for the classification task created by Jeremy Howard, and Rachel Thomas and Imgaug library [24] is used for the detection task created by Alexander Jung. Both these libraries are used to implement data augmentation for training the DeepFashion images. Augmentation method is shown to help generalization from computer models to the real-world task. The classification and detection performance are measured on testing DeepFashion dataset and a custom dataset using Top-K accuracy metric.

In our work, we first perform each data augmentation techniques individually to understand the nature and effectiveness of the method which can be studied from Table I. Later from the experiments conducted by Shijie et al. [23], it has been proved that an appropriate pairwise and triple combination method performs better than any individual ones. Also, the study shows that the overall performance of the triple combination is advanced than that of pair combinations. Therefore, we implemented a balanced approach to make this into a combination of two extensive augmentation strategy that has an effect on network learning patterns.

Data Augmentation	Top-1	Top-3	Top-5
Cutout	0.469	0.716	0.844
Symmetrc wrap	0.50	0.73	0.86
Contrast	0.486	0.744	0.86
Zoom	0.466	0.716	0.836
Perspective wrap	0.502	0.736	0.883
Brightness	0.425	0.679	0.848
Flip	0.439	0.667	0.823
Jitter	0.454	0.70	0.825

TABLE I Comparing accuracies of various data augmentation techniques in the custom testset

1) Data Augmentation Strategy I: In strategy I, we try to solve real-life low-quality image problems such as extremely low and high lightning, noise and inclined images. These are addressed using data augmentation techniques: (a) Brightness, (b) Jitter, (c) Perspective warp and (d) Flip. All these techniques are implemented with 0.50 probability so that there are multiple combinations of modified and unmodified versions of retaining the original image. (a) Brightness is an attribute of high and low colour intensities of visual perception which appears to have random darkening and lightening images. In our clothing images, we use brightness from 0.3 to 0.8 where 0.3 is the brightest to 0.8 being the darkest filter at a probability 0.50, i.e. random picture would be augmented with a random value between 0.3 to 0.8. (b) Jitter replaces each pixel with random factors from a neighbourhood of the specified radius between -0.03 to 0.01, thus producing a blurred effect in the image colour space. This technique relates with low clarity CCTV footage when

zoomed on a specific person. (c) Perspective Warp is a new image transformation used in circumstances where a camera is placed above or side on a CCTV footage or movie frames, and sometimes the clothing attire gets warped. It performs the slightest rotation with an angle between -0.5 to 0.5 and it preserves the labels of the clothing image. (d) Flipping is a common technique where we turn over the images to the horizontal axis to produce a diversity in clothes which is like a mirror. Here we create a flip using 0.5 probability.

2) Data Augmentation Strategy II: In our second strategy of data augmentation, we used techniques: (e) Cutout, (f) Contrast, (g) Symmetric warp and (h) Zoom. These techniques are used similarly like strategy 1 for the replication of real-time scenarios in surveillance videos and movie clips. (e) The cutout is a new art for data augmentation that is inspired by regularization [21]. The cutout is like random erasing which are used for occlusion where some parts of the image are unclear. Through this technique, random squares or patch regions called as cut holes are formed by masking it with random values on the clothing images in the training phase. When performing cutout in clothing images, it enables the network to learn more descriptive characteristics of the entire image rather than focusing on small subsets of visual features which may not always be present in real-time. In order to proceed with modified and unmodified versions of images, the cutout is applied by a constraint of 0.50 probability. (f) Contrast applies scale values to the clothing image from 0.5 to 1.5, where the colour channels of the image are adjusted to transform the picture from less-contrast to super-contrast. (g) Symmetric warp transformation is excellent for simulating images in different angles of value from -0.3 to 0.3 with a tilt in forward and backward, left and right. This is allied to perspective warp, where it gives an illusion that the clothing image is viewed from different perspectives than initially seen. (h) Zoom changes the image smoothly from long shot to a close-up and vice-versa. Here we use a random zoom value from 1. to 1.4 with a probability of 0.5. Therefore, all these appropriate techniques serve a different purpose in the augmented model.

V. IMPLEMENTATION

The implementation consists of two methods. One is a classification task where the architecture used is standard ResNet-34 and EfficientNet-B3. The other is an object detection task where the YOLO V3 algorithm has been employed. All models were trained with 12 GB Nvidia Tesla T4 graphical processing unit (GPU) utilizing the Pytorch (https://pytorch.org) package in Google Cloud Platform.

A. Classification Models

1) ResNet34: One of the effective models used in ResNet architectures is ResNet34. ResNet34 is a 34-layer deep network consisting of multiple convolutions and pooling step which is then followed by the same pattern in four more layers for different dimensions. Here, gradients can directly flow from the initial layers to the other layer by skipping the layers in between. This enables the network to be robust from the vanishing gradient problem and deals with the degradation of accuracy appearing in conventional neural networks. In our study, we use ResNet from fastai library implemented by PyTorch torchvision module. Here ReLu activation function and Batch Normalization are used in conjunction with the convolution operations.

We initiated experiments to classify the ten categories on the 224k clothing images in the DeepFashion dataset. We use a two-stage process to train the clothing images in the model. Firstly, we train for five epochs on the last few layers of the model. Then, we unfreeze the model to train all the layers of the model by applying the discriminative learning rate to train at optimal capacity. In the first layer, the model tries to detect groups of pixel and simple gradient lines, by the time it reaches the 34^{th} layer of the network the model identifies specific shapes of the clothes of repeating patterns in a 2-dimensional clothing image. Later, the model is trained for another five epochs with learning rate 1e-06 for first few layer parameters and 1e-05 for the last layers of the network. The metrics obtained to estimate the effectiveness of the model are Top-1 and Top-N accuracy. Top-1 accuracy corresponds to the maximum confidence for the predicted class. Top-3 and Top-5 accuracy are the value of confidence of the predicted class if the target label falls in the top N values of your predictions. This approach is evaluated on DeepFashion testset of 45k images and Custom testset of 360 images trained on original DeepFashion images, Data augmentation strategy 1 and Data augmentation strategy 2. This change in approach with/without augmentation techniques shows the effectiveness of how a model can classify and generalize well. By using data augmentation strategy 1 and 2, some of the training images get altered by utilizing a combination of image manipulation methods in the clothing recognition task. This has resulted in an optimized and fine-tuned model with better and improved accuracy, where generalization performance is high. The difference between DeepFashion images and Custom images are DeepFashion images are of studio type under white background conditions and individuals with straight postures. However, our custom dataset shows the real-world CCTV images and video frames.In Fig. 3, predicted Top-1 accuracies of the ResNet-34, trained on the custom dataset is displayed. It is observed that DA-1 performs better over the others.

2) EfficientNet: EfficientNet is a CNN architecture developed by Google and trained on ImageNet. EfficientNet comes with a family of model architecture versioned from B0 to B7. It demonstrates the effectiveness of the architecture by uniformly scaling up the width, depth and resolution. Width is the channels of feature maps per layer, depth is the layers of the network, and resolution is the size of the input images. This neural network proposes a compound scaling method to scale up all three parameters with a fixed constant ratio. In our experiments, we are using EfficientNetB3, which scales the 25 layers and channels of the model based on

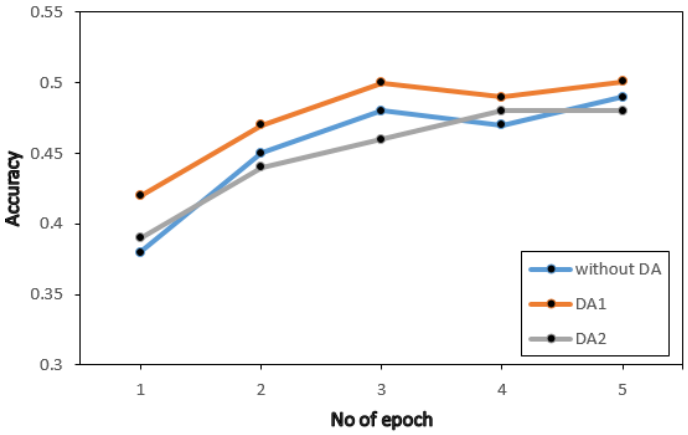

Fig. 2 Comparing Top-1 accuracy vs epoch of Resnet34 on custom dataset

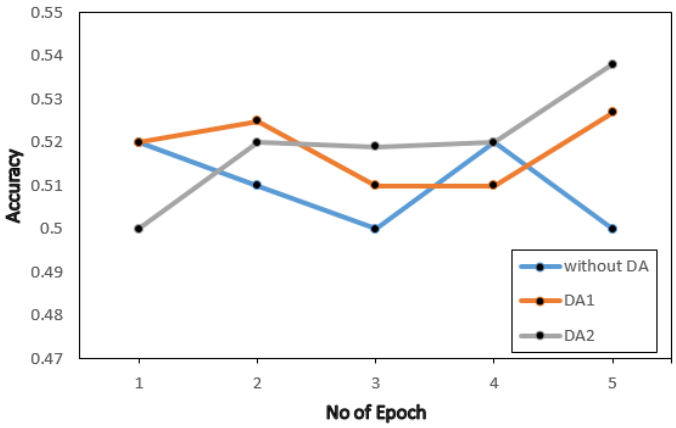

Fig. 3 Comparing Top-1 accuracy vs epoch of EfficientNet on custom dataset

input image size. This scaling leads to high performance with low computational cost and fewer parameters. It uses RMS optimizer with decay and momentum as 0.9 along with learning rate and swish activation function. Also, stochastic depth, weight decay, fixed auto augment policy and dropout are used in the multiple stages and layers of the convolutional neural network. Since DeepFashion contains high-resolution images, the network needs more layers and more channels to capture the fine-grained patterns of the clothing image. Hereabouts, depth captures more complex and more vibrant features along with generalization, width captures fine-grained features and becomes easy to train, and resolution captures detailed microscopic patterns.

Since EfficientNet scales upon three dimensions – depth, width and resolution, it outperforms the typical ResNet-34 architecture. This pre-trained neural network is trained in the same way as ResNet-34 network with ten epochs and a discriminative learning rate. Thus, it sets new records for both accuracy and computational efficiency. This pipeline is proceeded using original DeepFashion images, data augmentation strategy 1 and augmentation strategy 2. Therefore, It has been shown from the results that EfficientNet tends to capture more delicate details of the clothes and becomes easier to train.In Fig. 4, predicted Top-1 accuracies of the EfficientNet, trained on the custom dataset is displayed. It is observed that DA-2 performs better over the others.

B. Detection Model

1) YOLO: YOLO or "You Only Look Once" algorithm is one of the fastest CNN object detection algorithms. This algorithm functions as a single Neural network that divides the image into regions and predicts bounding boxes and probabilities for each region. Here, we use YOLO version 3, which uses Darknet-53 architecture. It contains 53 convolutional layers with an additional 53 layers added on top of one another, thus making it 106 fully connection convolutional layers, each followed by Batch Normalization and Leaky ReLU activation function. Detection is done at three different stages at 81st, 94th, and 106th layer. Thus, different layers of detections help to address the issue of detecting the diverse size of objects. The 94th layer is responsible for detecting large objects, whereas the 106th layer detects the smaller objects, with the 81st layer detecting medium objects.

To implement YOLO-V3, We adjusted the configuration values from the original paper[30] to work with the DeepFashion dataset and to improve the performance of the model. For training, pre-trained weights of COCO dataset is used. The batch size is 64, which is the number of images with labels in the forward pass to compute the gradient value and update the weights through backpropagation. We use multiple metrics to evaluate this model, Intersection over Union (IOU) metric evaluates the predicted bounding boxes. The IOU is calculated by dividing the area of intersection over the area of union between the actual (ground-truth) bounding box and the predicted bounding box. The IOU ranges from 0 to 1. The model is aimed to have a bounding box such that the IOU score is close to 1.

$$IOU = \frac{Area\,of\,Intersection}{Area\,of\,Union}$$

Next, we calculate the Average Precision (AP) and Mean Average Prediction (mAP) these metrics estimate the performance of object detection models. To calculate AP, we need Precision and Recall. The precision is calcutated by, the proportion of true positive (TP) and the total number of predicted positives (TP+FP). The recall is given by the ratio of true positive (TP) and total of actual positives(TP+FN).

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The final step to calculating the AP score is to take the average value of the precision across all recall values. Since precision

and recall are between 0 and 1, AP values also fall under the same range. Additionally, we calculate the mean of all the AP's, resulting in the mAP value. During the testing phase, we changed the batch size and subdivisions to 1, and the rest of the configuration is the same as the training phase.

Jacket - 0.85 Skirt - 0.65 Jacket - 0.79 Dress - 0.95

Fig. 4 Accurate Predictions from the ILIDS, Youtube pose, and create commons dataset after performing data-augmentation

Dress - 0.85 Blouse - 0.52 Jacket - 0.65 Tee - 0.17

Fig. 5 Misclassified Predictions from the ILIDS, Youtube pose, and create commons dataset after performing data-augmentation

As illustrated in Fig.4, we show the correct classification with their bounding boxes and predicted accuracies. This demonstrates how our object detector model performs within certain classes. Similarly, Fig.5 displays misclassified clothing images. As we can interpret that Jacket is misclassified with Dress, Tee is misclassified with Blouse and so on.

VI. EXPERIMENTS AND RESULTS

This section provides quantitative evaluations of different methods on the Clothing Classification and detection models. For an initial value, we perform a prediction with the sklearn dummy classifier on the balanced dataset, and this model produced an accuracy of about 0.167 This is a baseline result to compare against the other real classifier models. We conducted multiple experiments to check if the data augmentation influenced the detection of low-quality clothing images. There are two parts to these experiments. First, the models are tested on the DeepFashion images and then on the custom images. With this, we compare the effect of data augmentation on the low-quality images and high-quality images.

We implemented the clothing classification on ResNet34 Architecture. While testing on the DeepFashion dataset, Top-1 accuracy decreased when data augmentation strategies are applied while compared to the original DeepFashion testset. It happened because the images manipulated on the training stage does not affect the studio-type white background DeepFashion testset. At the same time, the Top-3 and Top-5 accuracies stayed pretty much the same. We can infer that the data augmentation techniques have minimal effect on the original DeepFashion testset. On examining more on these predictions, specific categories like Blouse gets misclassified with Tee, and Top gets misclassified with Blouse. This misclassification is because there are some images whose shapes, size and patterns look similar to one or more classes. Also, some misclassifications happened because of the multiple attributes in a single image, for an image can have a top, and a skirt, the model was able to predict the top or the skirt alone. To resolve this issue, we also introduce the Top-3 and Top-5 accuracy to improve the prediction rate. On the contrary, the Top-1 accuracy of the custom dataset increased by 7.5% when compared with the custom data without augmentation and with data augmentation strategy 1.

When the effectiveness of each clothing category is analyzed in data augmentation strategy 1, it's proven to show better accuracy. Misclassification of Blouse, dress, jacket, sweater and tank is significantly reduced on custom dataset after training on augmentation techniques. The identified problem with Resnet-34 is that Top clothing category did not correctly classify in the custom dataset as the patterns and depth of the network is getting confused with Blouse and Tee. Overall, our tests showed a significant increase when data augmentation techniques are applied to ResNet-34. Table II summarizes the performance of different pipelines on the ResNet-34 model of Clothing Classification. Here DA specifies Data Augmentation and Top-1, Top-2, Top-3 specifies accuracies.

Dataset	Description	Top-1	Top-3	Top-5
DeepFashion	Without DA	0.743	0.908	0.964
DeepFashion	With DA Strategy 1	0.727	0.910	0.968
DeepFashion	With DA Strategy 2	0.715	0.902	0.964
Custom	Without DA	0.491	0.747	0.841
Custom	With DA Strategy 1	0.544	0.770	0.863
Custom	With DA Strategy 2	0.566	0.783	0.883

TABLE II ResNet34: Comparing clothing classification pipelines and data augmentation techniques

Next, we executed the same experiment on the EfficientNet model. The classification accuracy is 0.755 for DeepFashion testset, which is the highest prediction accuracy on the original dataset. Furthermore, on testing with the data augmentation techniques, we were able to produce notable results. The acquired results show that the category labels significantly predict correctly than in ResNet architecture. After applying data augmentation strategy 1, Blouse and

tank demonstrated significant improvement in classification accuracy. Later by using data augmentation strategy 2, we can see only a few improvements in prediction accuracy compared to strategy 1. The problem faced in ResNet has been solved in EfficientNet by predicting the correct label for the Top clothing category. This shows that the compound scaling technique of EfficientNet is useful in recognizing intricate patterns and similar shapes of the clothing.

Intuitively, it shows that from the top-10 balanced categories of DeepFashion dataset blouse and top are complicated clothing attributes to classify. However, dress, shorts, sweater, tees, in turn, were simpler to classify. Also, we observe several clothing categories with noisy data has accuracy gain after data augmentation methods. The prediction accuracy of EfficientNet is tabulated in Table III.

Dataset	Description	Accuracy	Top-3	Top-5
DeepFashion	Without DA	0.755	0.919	0.968
DeepFashion	With DA Strategy 1	0.746	0.916	0.968
DeepFashion	With DA Strategy 2	0.744	0.914	0.968
Custom	Without DA	0.500	0.775	0.869
Custom	With DA Strategy I	0.527	0.794	0.900
Custom	With DA Strategy II	0.538	0.802	0.894

TABLE III EfficientNet: Comparing clothing classification pipelines and data-augmentation techniques

The best-attained accuracy on DeepFashion testset is 75.5% of EfficientNet without any data augmentation techniques. Likewise, the best-published accuracy on Custom testset is 56.6% of Resnet with data augmentation strategy 2 (contrast, cutout, symmetric warp, zoom).

Subsequently, when we had the classification models, we wanted to see how the dataset performed with the object detection task. We implemented object detection utilizing the bounding box approach using YOLO-v3, where the average precision values for each category is tabulated in Table IV.

Clothing Category	Without DA	With DA-1
Jacket	18.59	23.94
Dress	58.91	60.12
Skirt	19.97	25.85
Top	15.92	19.34
Sweater	48.48	50.58
Shorts	30.51	43.52
Cardigan	48.44	41.52
Blouse	32.91	35.47
Tank	49.98	28.66
Tee	47.09	55.91

TABLE IV YOLO-V3 : Comparing Average Precision (AP) values of DA-1 and Without DA on custom dataset

In Table IV, we can observe the comparison of AP values between weights trained on the DeepFashion dataset and Data Augmentation strategy 1. The minimum threshold for prediction is set to be 0.50. On illustration, we can notice that the AP value has predominantly increased on almost all the categories after testing the custom dataset with the data-augmented weights. There is a significant increase in the prediction of Shorts and Tee, while types like Tank and Cardigan has the slightest change. We get the mAP estimation of 0.35 without data augmentation and 0.37 with data augmentation strategy 1 on Custom testset.

Dataset	Description	F1-score	Avg IOU	mAP
DeepFashion	Without DA	0.61	53.57%	0.61
Custom	Without DA	0.38	40.61%	0.35
DeepFashion	With DA Strategy I	0.6	53.09%	0.60
Custom	With DA Strategy I	0.40	43.95%	0.37

TABLE V Comparing object detection pipelines and data-augmentation technique using YOLO-V3.

For further examinations, F1 score, Avg IOU and mAP are tabulated in Table V.

VII. DISCUSSION

As the image processing and deep learning evolve, the need for clothes classification in commercial use will increase. These techniques, with the help of smart sensors and internet of things, will help to detect real-time clothing in CCTV cameras, movies and Creative commons videos. Our objective is to compare the prediction accuracy after training the DeepFashion images on various data augmentation techniques and test it on independent custom data.

Data Preprocessing: There is a massive imbalance in the Deepfashion dataset. This dataset consist of 50 categories and most of the categories contains hundreds of images, while some of them have more than 40,000 images. The imbalanced nature makes it hard to train the insignificant categories. So, we handled the imbalance by including only the top 10 category types of clothing with more than 10,000 images.

Object Detection: The DeepFashion images were converted into YOLO format with specific python scripts. Here, we combined image location, Category and bounding box values and converted the original [x1,y1,x2,y2] format of bounding box values into the desired YOLO format. Due to the vast volume of data, training consumed a considerable amount of time, to reach accuracy with satisfactory values.

VIII. CONCLUSION

This work presents the effectiveness of data augmentation strategies to classify and detect images in real-life scenarios. It surpasses the existing simple transformation in terms of rich image manipulation methods. To demonstrate the advantages of augmentation, we implemented them with ResNet, EfficientNet and YOLO architectures. Our experiments were conducted to provide a solution to real-time clothing classification and detection that can be used in forensic investigations. Our findings reveal that data augmentation techniques can be effectively used on videos and images with low quality to improve the prediction and accuracy.

One of the possible next steps would be to use YOLO-V5 for object detection and explore with DeepFashion2 dataset to add multiple bounding boxes and details to the images. Also, through extensive experiments, we demonstrate the effectiveness of data augmentation techniques which may significantly facilitate future researches.

IX. REFERENCES

[1] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 1535–1542 IEEE, 2011.

[2] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing, style detection and retrieval. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, pages 8–13. IEEE, 2013.

[3] Lao B, Jagadeesh K. Convolutional neural networks for fashion classification and object detection [cited 2016 June 26].

[4] Bossard, Lukas, et al." Apparel classification with style." Computer VisionACCV 2012. Springer Berlin Heidelberg, 2013. 321-335.

[5] Han, Dongmei, Qigang Liu, and Weiguo Fan. "A new image classification method using CNN transfer learning and web data augmentation." Expert Systems with Applications 95 (2018): 43-56.

[6] Liu, Ziwei, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1096-1104. 2016.

[7] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv: 1409.1556 (2014).

[8] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In ICCV, 2015

[9] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In ICCV, 2015.

[10] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In Image Processing (ICIP), 18th IEEE International Conference on pages 2937–2940. IEEE, 2011.

[11] Bedeli, Marianna, Zeno Geradts, and Erwin van Eijk. "Clothing identification via deep learning: forensic applications." Forensic sciences research 3, no. 3 (2018): 219-229.

[12] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujście, 2018, pp. 117-122.

[13] Heilbron, Fabian Caba, Bojan Pepik, Zohar Barzelay, and Michael Donoser. "Clothing Recognition in the Wild using the Amazon Catalog." In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3145-3148. IEEE, 2019.

[14] Cychnerski, Jan, Adam Brzeski, Adrian Boguszewski, Mateusz Marmolowski, and Marek Trojanowicz. "Clothes detection and classification using convolutional neural networks." In 2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1-8. IEEE, 2017.

[15] Lao, Brian, and Karthik Jagadeesh. "Convolutional neural networks for fashion classification and object detection." CCCV 2015: Computer Vision (2015): 120-129.

[16] Jia, Menglin, Yichen Zhou, Mengyun Shi, and Bharath Hariharan. "A deep-learning-based fashion attributes detection model." arXiv preprint arXiv: 1810.10148 (2018).

[17] Chen Q, Huang J, Feris R, et al. Deep domain adaptation for describing people based on fine-grained clothing attributes. IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA: IEEE Press; 2015; p. 5315–5324.

[18] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi, "Yet another survey on image segmentation: Region and boundary information integration," in Proc. ECCV, May 2002, pp. 408–422.

[19] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[20] Tan, Mingxing, and Quoc V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." arXiv preprint arXiv:1905.11946 (2019).

[21] DeVries, Terrance, and Graham W. Taylor. "Improved regularization of convolutional neural networks with cutout." arXiv preprint arXiv:1708.04552 (2017).

[22] Perez, Luis, and Jason Wang. "The effectiveness of data augmentation in image classification using deep learning." arXiv preprint arXiv:1712.04621 (2017).

[23] Shijie, Jia, Wang Ping, Jia Peiyi, and Hu Siping. "Research on data augmentation for image classification based on convolution neural networks." In 2017 Chinese automation congress (CAC), pp. 4165-4170. IEEE, 2017.

[24] A. Jung. Imgaug. [online] available at: https://github.com/aleju/imgaug, 2015 [Accessed 16 August 2020].

[25] J. Howard, R.Thomas. Fastai. [online] available at: https://www.fast.ai/, 2016 [Accessed 16 August 2020].

[26] Bossard, Lukas, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. "Apparel classification with style." In Asian conference on computer vision, pp. 321-335. Springer, Berlin, Heidelberg, 2012.

[27] Liu, Si, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. "Fashion parsing with weak color-category labels." IEEE Transactions on Multimedia 16, no. 1 (2013): 253-265.

[28] Gu, Chunhui, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan et al. "Ava: A video dataset of spatio-temporally localized atomic visual actions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6047-6056. 2018.

[29] Liu, Ruyue, Zeyu Yan, Zhijie Wang, and Shenyi Ding. "An Improved YOLOV3 for Pedestrian Clothing Detection." In 2019 6th International Conference on Systems and Informatics (ICSAI), pp. 139-143. IEEE, 2019.

[30] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).