Categorization of Fashion Clothes from Wild Images using Object Detection and Segmentation based Models

MSc Research Project
MSc in Data Analytics

Rohan Indrajeet Jadhav
Student ID: x20169043

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes, Dr. Pramod Pathak

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rohan Indrajeet Jadhav
Student ID:	x20169043
Programme:	MSc in Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Paul Stynes, Dr. Pramod Pathak
Submission Due Date:	31/01/2022
Project Title:	Categorization of Fashion Clothes from Wild Images using Object Detection and Segmentation based Models
Word Count:	5190
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Categorization of Fashion Clothes from Wild Images using Object Detection and Segmentation based Models

Rohan Indrajeet Jadhav
x20169043

Abstract

Categorizing of clothes in the wild fashion images involves identifying the type of clothes a person wears from non-studio images such as a shirt, trousers, and so on. Identifying the fashion clothes from wild images that are grainy, unfocused, with people in different poses is a challenge. This research proposes the instance segmentation model outperforms compare to object detection model when it comes to clothes categorization from wild images. Faster RCNN is and object detection deep learning model and the segmentation model used is Mask RCNN which is based on RCNN (Region-Based Convolutional Neural Network). Both models are trained on the subset of DeepFashion2 dataset. The models are trained on 10k distinct fashion images using rich annotation files in file, 2k images are used for test purposes and 5k images for the validation. The results of two models presented in this paper based on the average precision, recall across the different IoU(Intersection over Union) metrics considering the clothes acquired in fashion image. Based on the results it is shown as the segmentation model Mask, the RCNN outperforms by 20% compared to object detection model Faster RCNN.

Keywords- Clothes Categorization, Segmentation, Object Detection, Faster-RCNN, MaskRCNN, Azure

1 Introduction

Fashion images are nothing but the images that intended to show clothing accessories. and such images that are taken in the wild holds different poses of people, grainy images, multiple persons in single image and different clothing items etc. Detecting contents of such image and categorizing the clothes such as trousers, shirts and skirts etc. is a challenge. However, the studio images are the fashion images which taken in studio having single person which has main intention to show clothes. Such clothes categorization has been developed using scratch and transfer learning by many researchers (Liu et al.; 2016), (Ren et al.; 2017a), (Feng et al.; 2018), (Kayed et al.; 2020).

On e-commerce-based platforms the demand for easy clothing retrieval based on search tags is growing on the other side the clothing categories are also increasing. In such a growing clothing fashion industry the faster clothes retrieval is the key element (Grana et al.; 2012). This can be achieved from only detecting the contents of fashion images irrespective of the fashion images that are studio or non-studio with multiple poses of person. Clothes detection from fashion images comes with three major challenges: first, clothes these days comes with variety of styles, cuttings, patterns, etc. Second, occlusion of images and deformations. Third, fashion images typically change based on different scenarios called non-studio images such as selfies, group photos, random people photos, and so on.

The aim of this research is to investigate to what extent the instance segmentation performs well in terms of average precision to categorize the clothes from input wild fashion image compared to object detection.The major contribution of this research is a novel that compares the segmentation model with object detection while categorizing the clothes from wild fashion images. The base model is referred RCNN which is a Region based Convolutional Neural Network model which primarily focuses on object detection of any images passes. The faster RCNN is and extended version of fast RCNN in which it used the Regional Proposed Network and the main use of RPN is train model which requires high quality regional proposal (Ren et al.; 2017b). Mask RCNN is the Convolutional Neural Network and state-of-the-art when it comes to image segmentation is required in images in this case detecting the shirts, trousers, and their different patterns and so on. Investigated the state-of-the-art performance by training the Mask RCNN an instance segmentation model (D and V; 2021) over Faster RCNN which is an object detection for improving the model accuracy in terms of average precision, recall considering the IoU and at what percent the object covered the area in fashion image(Zhou et al.; 2019).

This segmentation and detection model implementation is carried out on azure cloud platform by leveraging the GPU based compute node which are intended for deep learning and visualization tasks and other services such as azure Machine Learning, blob storage for the data storage purpose.[1]

This paper discusses further deep learning models for categorizing the clothes from fashion images in section 2 related work, the research methodology is explained in section 3, section 4 illustrates more on design components of both models. The implementation of this research is covered in section 5, section 6 is given as the evaluation and results of the model that are trained. Finally, section 7 shows the conclusion of this research.

2 Related Work

Categorization of fashion clothes from wild images is a challenge because while detecting the contents of fashion images need to deal with various patterns, different occlusions of images, number of availability of clothes and fashion. Extensive research has been devoted recently for detecting the clothes from any fashion images for categorizing the clothes.(Liu et al.; 2016),(Ge et al.; 2019), (Heilbron et al.; 2019), (Sharma et al.; 2018), (Grana et al.; 2012)

Researchers (Liu et al.; 2016) have proposed deep learning based novel model FashionModel which primarily focused on landmark detection and detecting the attributes of fashion images. The deep learning model's network used is same as the VGG-16 only the last convolutional layer has been designed specifically for clothes by introducing three different layers layer. First extracting the features from clothing images, second is for detecting local features of clothes based on clothing landmarks, lastly predicting the landmarks and clothes. FashionNet has been pre-trained on 300k images and have gained the best accuracy of 76.4% compared to DARN and WTBI. In combination with human

[1] https://studio.azureml.net/

joints the accuracy is dropped by 8 to 9% while FashionNet+100 and +500 have gained more 10 to 12%. There are other deep learning models are implemented named AlexNet, GoogleNet and ResNet. AlexNet is trained model with accuracy 44.10%, the ResNet was trained with 59.82% of accuracy and GoogleNet has the 64.40% this empirical analysis by researchers (Sharma et al.; 2018).

Using fashion images from DeepFashion and AmazonCatalogue for clothes detection (Heilbron et al.; 2019) have gained the accuracy comparing CatalogueFashion and CatalogueFashion-10x. These two models are implemented with AC dataset having 71.0% of accuracy, 76.% respectively. Researchers have obtained the state-of-the-art performance by training the model by combing the fashion images which are catalogued and from DeepFashion helped in boosting the accuracy to 89.9%. Researchers (Ferreira et al.; 2019) also have proposed visual semantic attention model [VSAM] which is exploited by pose extraction and feature space. VSAM outperforms the state-of-the-art by giving the boost in accuracy of existing work done so far. VSAM has outperformed with f1 accuracy of 80.70

Researchers (Kayed et al.; 2020) have also proposed the clothes categorization CNN based model named LeNet-5 which has outperformed the accuracy and state-of-the-art by giving the f1 score of 98%. LeNet-5 is network that has been trained using MNIST images and CNN based model with seven layers including the output layer. Still there is gap expressed about using the real-world fashion images to categorise the clothes and the images that are taken in wild. Among all these researchers of clothes categorisation, the key element is attribute identification. By focusing on the attribute detection researchers (Jia et al.; 2018) have proposed a model which is the extension of Faster RCNN implemented using ResNet-101 and ROI by slight change DeepFashion images about bounding boxes removal. Evaluation is carried out through Average Precision and CorLoc value which is driven by ratio of detected groundtruth over total number of groundtruths.

There are two major categories are shown ahead for categorization purpose one is object detection and other is segmentation out of which for object detection much research has been devoted. Researchers (Feng et al.; 2018) have proved the approach object detection system using YoLo-opt which is originated from YoLo-V2. Performed the comparison bet wen the tradition CNN (Convolutional Neural Network) model with other machine learning algorithms alongside the semantic segmentation. The proposed system achieved the accuracy of 83% in which used the street beats dataset and the pre-trained darknet convolutional network-based model. (Xiang et al.; 2020) also have introduced the extended version of RCNN for the recognition task. Inception-ResNet V1 model is used with L-softmax layer is applied for categories prediction. They have achieved the precision of 73% and the recall at 83%.

There are other models which are based on object detection, the fast RCNN which is RPN (Regional Proposal Network) model used by (Ren et al.; 2017a). Baseline to this is Convolutional Neural Network which helps more precise object detection Authors have come with encouraging results using RPN and fast RCNN on PSACAL VOC dataset as well and MS COCO. The mAP for this model using the PASCAL VOC dataset achieved the average precision of 58% and while using MS COCO the precision is increased by 20% once replaced the network to VGG 100 layers to ResNet.

Humans are the one who is look forward to different fashion clothes, the other way human body segmentation can help in understanding the clothes that a person wearing. Similar way (Zhang et al.; 2020) have come up with approach that involved the human body segmentation. This research mainly carried in three steps first dataset preparation in which they have their own dataset prepared. Second is complete human body segmentation which is implemented in combination of deep learning approach. Third is the convolutional network that predicts the clothes category.

In all such researchers one of major key point comes about the video and its object tracking in it ultimately this also related to object detection from image. To do carry out the object tracking system (Noor et al.; 2021) came with approach using existing segmentation based SiamMask model usage. The main steps to do the changes are carried as run the detectron2 and YoLov4 to find the object and hold the scores along with video ID, number of objects in it. Secondly the tracking comes that the output of detectron2 is fed to SiamMask to pick an object with highest score and run the tracker until the last frame of video. Researchers (Chao et al.; 2018) also proposed and TAL-Net network for video localization which is implemented from Faster RCNN. They have exploited the architecture of Faster RCNN targeting the field alignment, feature extraction and feature fusion.

In fashion related tasks it has been proven that each task with higher complexity requires higher volume images that shows the variety of clothes wearing human beings. There are variety of dataset available, but the DeepFashion is the dataset which has the limitation such single clothing items, complex to retreive the landmarks of images, no per pixel masking due to which (Ge et al.; 2019) have come up with extended version of DeepFashion dataset named DeepFashion2 in which all the flaws are removed and supplied the rich annotated fahion dataset that can be used in various fashion tasks. Researchers have come up with two strong models that can shows the object detection and segmentation is carried out on DeepFashion2 dataset. They have implemented the Match RCNN and Mask RCNN to benchmark the use of this dataset. The Mask RCNN is achieved only 0.53 of average precision on 13 distinct categories of fashion clothes.

All the fashion related tasks may be object detection or segmentation and in videos the object tracking requires high processing tool. These days all of the process and flows are deployed on cloud environment. (Pliuhin et al.; 2019) have shown the use of azure cloud supplied service called azure machine learning studio. AML studio allows user to create all types of models and deploy those using pipelines provided on real-time input data. Authors also have concluded that the use of AML studio by developing their own model. Used can configure the cluster as per the requirement and deploy those cluster to build the own model. Provides high storage and optimized memory along with GPU cores and use can extend it whenever it is needed which is key requirement in all deep learning model implementation.

In conclusion, it is seen that much work has been done on categorizing the clothes from studio fashion images that are not grainy, unfocused as well as group photos with different poses. The current research shows that categorization has been deployed using many models such as CNN, ResNet, LetNet, and so on. Through this research, the old dataset such as MINST, MS COCO, deepfashion has been extensively used. Nowadays the fashion styles are increasing due to the variety of clothes availability the research is not carried to categorize the clothes from wild images, the deepfashion2 dataset has been published in 2019 challenge which will cover all the parts of the wild image such as occlusion, zoom-in, scale, and viewpoint. This research proposes the use of the deepfashion2 dataset and categorizes the clothes from wild fashion images using object detection and segmentation. As experimentation, Faster RCNN as object detection and Mask RCNN as segmentation will be trained on deepfashion2 dataset and accurate performance evaluated.

3 Methodology

The research methodology primarily consists of three stages, Data preparation, Train and Tune the model, and finally Evaluation of the model and its monitoring. Figure 1 shows the details of all these steps.

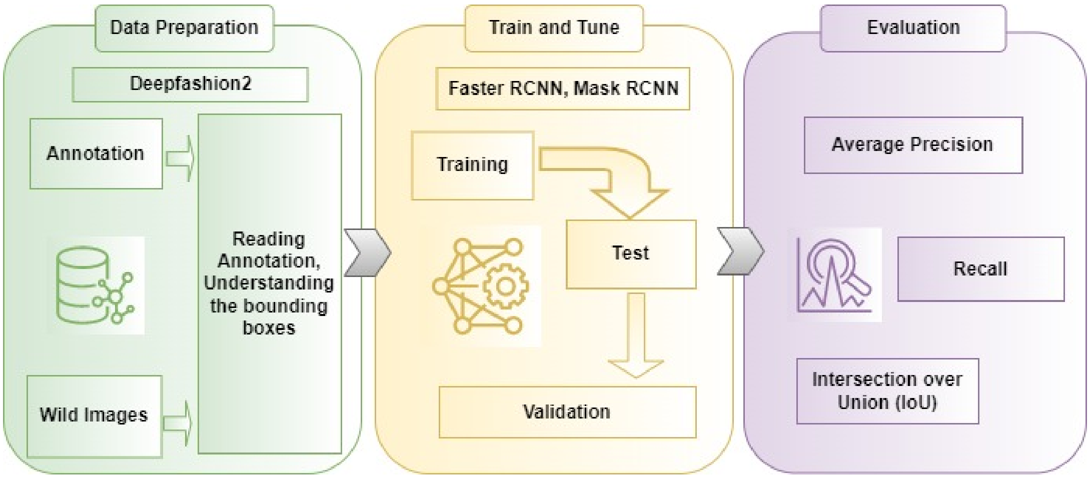

Figure 1: Research Methodology

The first step of research methodology is data preparation which involves reading the input dataset that Deepfashion2 which has rich annotated fashion images. This dataset is consisting of json files for every fashion image. JSON file has information such as pair id, image id, the scale of images, zoom-in, bounding boxes, occlusion of image. The fashion images are also included, and both sets of data is getting passed to the next step that is training and tuning the model. The details of dataset is exploited in section 3.1.

As part of the next step in which models are getting trained using the data fed in the forms of JSON for reading the annotation and actual image. JSON is the final file in which all annotations are consolidated into a single JSON file. This phase two main deep learning models Faster RCNN and the Mask RCNN. The Faster RCNN is a clothes detection methodology in which 10k train images are used to train, 5k images for validation purposes and 2k images are used for testing which is getting used in the next step. The Mask RCNN is a clothes segmentation methodology in which the same set of images and JSON files are used for training the model.

Finally, the test step involves testing the trained models by feeding certain wild images and seeing how the clothes detection and clothes segmentation works by using deep learning models The results and evaluation of models are defined using precision, recall, and IoU matrix which has been illustrated in section 6. These are the metrics that are used for evaluating the performance of both models.

3.1 Dataset Description

The dataset used for implementation of clothes categorization using object detection and segmentaion models is subset of deepfashion2 dataset which is an extended version of the deepfashion dataset. This dataset is intended to eliminate the flaws from deepfashion dataset [2]. The key characteristics of this dataset are:

[2] https://github.com/switchablenorms/DeepFashion2

1. The large sample set of different fashion clothes
2. This is a versatile dataset that has been built considering the different tasks in fashion images.
3. 13 distinct categories of clothes
4. Rich annotation JSON files for each of the fashion images.
5. Divided into train, test, and validation datasets[3].

Each fashion image is a unique six-digit number to find its annotations in JSON files. Each JSON file has below information that the cloth detection and segmentation is implemented accurately.

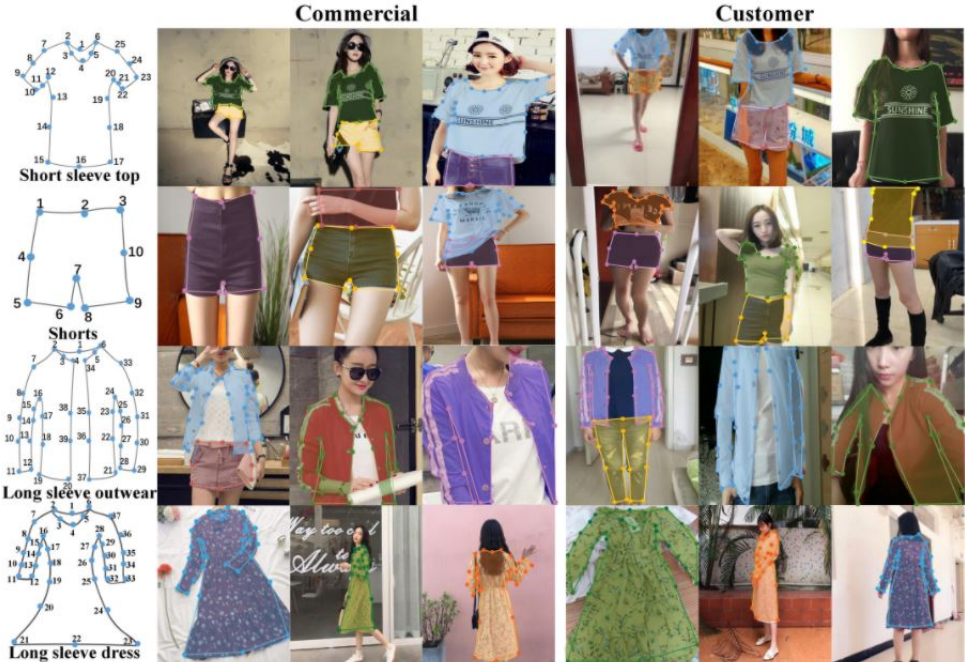

Figure 2: Dataset showing the landmarks

- *Source*: depicts the fashion is from store or user.
- *Category*: This info gives the category of cloth, here it has 13 distinct categories
- *Category ID*: This stands for the numerical value for the category.
- *Style*: the fashion images whose pair_id is the same that has the same style just bit of changes in clothes like color or shape.
- *Bounding box*: a bounding box that compels the object from fashion images. Shows in coordinates.
- *Landmarks*: Shown in figure 2. Each category has a different landmark.
- *Segmentation*: shows the polygon single clothing fashion compels 1 polygon
- *Scale*: Small, medium, or large-scaled image
- *Occlusion*: small occlusion has 1, the medium has 2 and large occlusion stands for 3.

[3] https://github.com/switchablenorms/DeepFashion2

- *Zoom-in*: stands for the at what scale the fashion is zoomed-in three values 1,2 and 3 shows small zoom, no-zoom, and large zoom

- *Viewpoint*: Depicts the front viewpoint, side, or back viewpoint of fashion clothes in the image.

4 Design Specification

The high-level architecture of clothes detection and clothes segmentation is shown in figure 3. The key components of both models are reading the fashion image, read the annotations for the same, apply the models and the finally it will show the clothes category from given fashion image in the output component.

Crucial element in the architecture is reading the annotations of each of the fashion image that is passed to train the model further. Annotations are nothing but landmarks for identifying the object from image (Kim et al.; 2021) which stored in the json format, so each image must have its own annotation files. This all data has been fed from subset of deepfashion2 dataset which is a rich annotated fashion dataset having more than 300k images to train and use for various fashion tasks. Key elements of both architecture is explained further with details of each model.

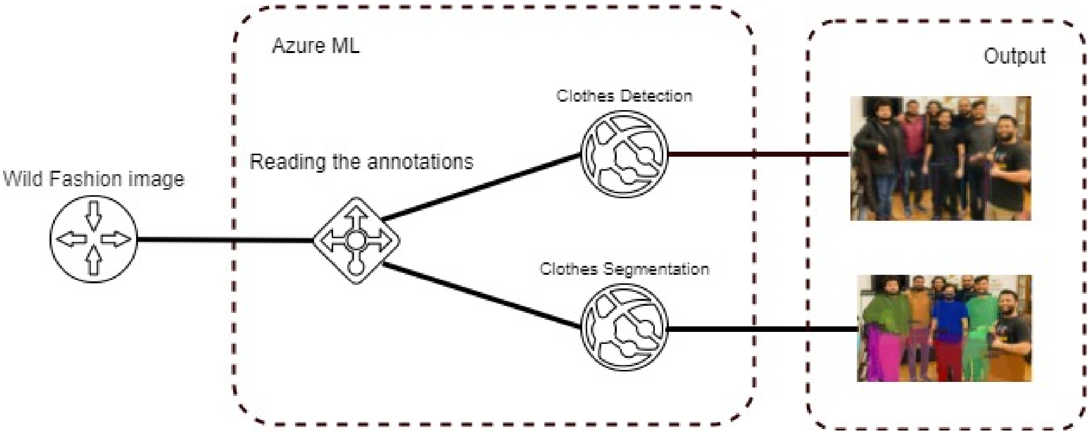

Figure 3: Architectural Flow

4.1 Clothes Detection

As part of clothes detection the deep learning model used is Faster RCNN which is and extended version of RCNN. All the basis of these models are CNN that is convolutional neural networks. As this is powerful API getting used all over when it comes to visualization tasks. All configuration details and implementation is explained in section 5.

Finally, the object detection is shows in output image, rectangular shapes are the bounding boxes under which the object is labelled based on the category. In the output component the boxes in first image shown in figure 3 are the clothes detected with labels.

4.2 Clothes Segmentation

Reading the annotation file is quite similar part as the dataset provided it has the landmarks and bounding boxes itself on each fashion images. Here in segmentation the Mask

RCNN model implemented, and it requires the landmarks rather than bounding boxes. Mask RCNN is implemented with 100 layered and used the weighs from the same API provided by Facebook.

Looking at the output image of segmentation it clearly shows that how the model has accurately looked at the surface of each object from image that is clothes from fashion images. Output component in figure 3 has two output images which clearly shows the how object detection and segmentation work.

5 Implementation

Clothes categorization from wild fashion images is implemented using two major deep learning-based models first is clothes detection in which Faster RCNN is implemented and second is clothes segmentation in which Mask RCNN is implemented. Underlying model building implementation for both of the models is shown in figure 4

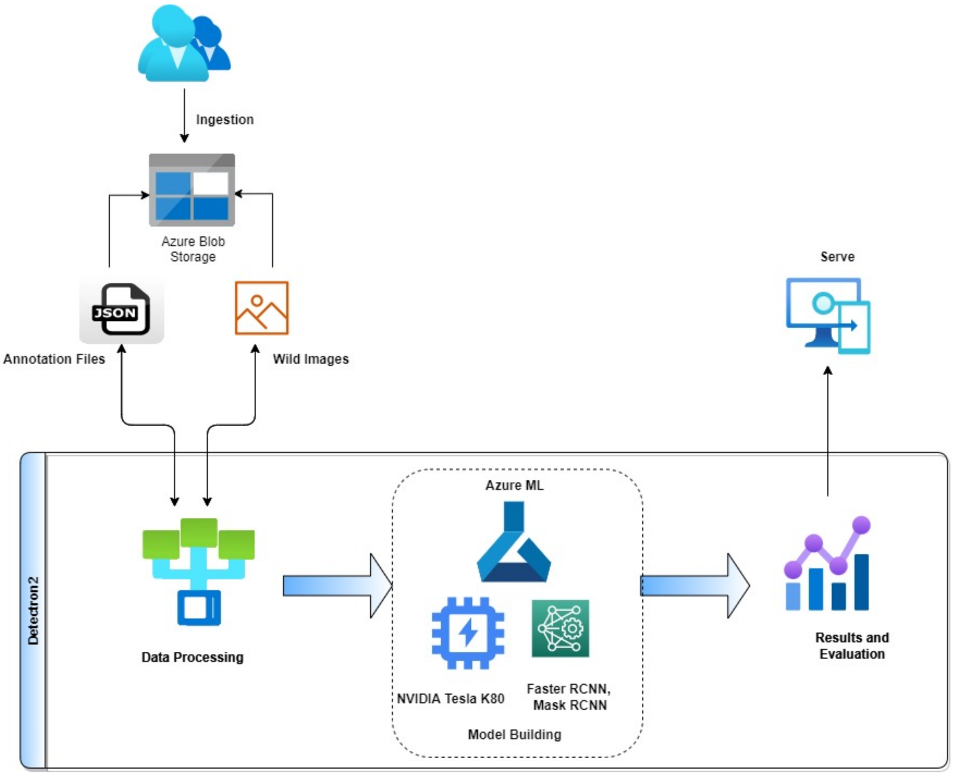

Figure 4: Architecture of Categorization of Clothes from Wild Images.

The first phase that comes in the picture is data storage, all the datasets including JSON files and fashion images are stored on blobs the service provide by azure [4]. One container is created and json files for each of the fashion images are located with naming convection provided in the deepfashion2 dataset. For example, in the JSON file of name 000001.json the respective image for this json is 000001.jpg. This is needed because the dataset registration happens in further process before model building the consolidated JSON must be passed to the model. The more details are provided about dataset in section 5.1

[4] https://azure.microsoft.com/en-us/services/storage/blobs/

The next step that comes into the picture is data pre-processing, as mentioned earlier this phase is focusing on reading the image and JSON file for each fashion and creating the consolidated json file. This json will be having a detailed annotation of each image with its unique ID and it has other crucial information such as height, the width of images, bounding boxes, segmentations, landmarks, pair_id, number of key points. This all information is required while training the deep learning model using detectron2 (Ding et al.; 2021). Here it will create the two consolidated JSON files one for train dataset of 10k images and one for validation dataset of 5k fashion images.

The next component after data processing is to build the model is an actual train the model on the train dataset which is of 10k fashion images and its one consolidated JSON file. Both clothes detection and clothes segmentation are carried out using the detectron2 API which is developed by Facebook for visual recognition tasks. As this requires a CUDA environment to train the models that leveraged the azures GPU-based compute node NVIDIA Tesla K80[5]. The whole implementation is carried out using service provided by azure is Azure Machine Learning using pyhon3 kernel. Both model details are explained in section 5.2 and 5.3. Along with the configuration used is azure ML studio an azure cloud provided service. Python3 kernel and langauge is used to implement both model. The compute node which is spinned to run these models is of 20 RAM, 256 GB of disk space located on blob storage and 2 core GPU.

Once the model is trained the next phase comes in picture to test the trained model on random test images from test dataset of 2k images provided in deepfashion2 dataset. This has been tested on random other wild images also which has been discussed more in section 6. Along with this model, evaluation is carried using Average Precision metrics, IoU[Intersection over Union], and the recall. These are the metrics that are majorly considered into account when deep learning deals with object detection or object segmentation. IoU is nothing but the actual vs predicted bounding box of a fashion image.

5.1 Faster RCNN

The faster RCNN is an object detection model in which its key networks are backed first RPN which is a Region Proposed Network used to generate the regional proposal based on the bounding boxes. Second is a network of fast RCNN which helps to detect the objects from the region passed by RPN. The third is the classification layer that is the output layer to classify the fashion clothes. Finally, the regressor layer which gives a more precise bounding box of an object here the clothes.

Detectron2 supplies the model zoo library from which the pre-trained models are used using the yaml file provided in the GitHub repository [6], faster_rcnn_R_50_FPN_1x. yaml is the model is used in this implementation. R50 shows the ResNet-50 model is used in this faster RCNN based on detectron2.

5.2 Mask RCNN

Originally the Mask RCNN is built on top of Faster RCNN, but the major difference is mask RCNN is works based on the object segmentation which gives more precise detection. In this implementation as well Mask RCNN has trained the same dataset,

[5] https://www.nvidia.com/en-gb/data-center/tesla-k80/

[6] https://github.com/facebookresearch/detectron2/blob/main/MODEL$_{Z}OO.md$

which is used for faster RCNN, but this model outperformed in terms of accuracy and the result of the fashion images which are in the wild. The mask RCNN works in two major steps first, produces the proposals of the region where the object might have been present, and second is an actual prediction of an instance that has been detected. The backbone of the mask RCNN is FPN based deep neural network that connection from ConvNet, ResNet, or VGG which are strong CNN-based models (Anantharaman et al.; 2018).

While implementing the Mask RCNN on deepfashion2 used the model zoo library and the pre-trained model mask_rcnn_R_101_FPN_3x. yaml which uses the ResNet101 internally. The FPN here runs on RoI that Region of Interest which works on pixel and finds out the surface of each object rather than the boxes of an object. In fashion clothes segmentation it is very necessary to find out the RoI based on pixel which helps the segment the instance of cloth from whole fashion images. In section 4 explained the key difference of Mask RCNN and Faster RCNN with its bounding boxes and landmark estimation.

6 Results and Discussion

This research aims to categorize fashionable clothes in the wild using object detection and instances segmentation. Using a subset of deepfashion2 dataset trained the Faster RCNN and Mask RCNN able to categorize the clothes in different 13 categories. Both models are trained on 10k images, test data used of 2k images, and for validation set used 5k images out of the whole dataset of deepfashion2. Here to evaluate both models used the AP [Average Precision metrics]. [7]

AP is nothing but and value between 0 to 1 which gives an average precision value for respective recall. When it comes to object detection or instance segmentation the IoU [Intersection over Union][8]. IoU is calculated between the two borders that are nothing but ground truth and Prediction. Ground truth is the actual object/segment Vs the predicted object/segment by model. Details of each model's evaluation are explained further. In clothes detection flow RoI is considered as positive when the IoU with ground truth is larger than 0.5 further results show details of different IoU and different area coverage-based results.

6.1 Faster RCNN

The aim of this experiment is to train the Faster RCNN model on deepfashion2 dataset and categorize the clothes. This model is trained using 50 layers as R50 FPN used but looking at average precision it did not gain much accuracy in terms of object detection. Table 1 shows the AP metrics for the bounding boxes.

Table 1: Evaluation results for Bounding boxes

AP	**AP50**	**Ap75**	**APs**	**Apm**	**APl**
15.118	24.174	16.610	40.00	15.070	15.217

[7] https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

[8] https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

Table 1 contains AP 15.118 is averaged over all the classes of 13 which is a major AP, this is much likely less and that is expected as we used a subset of deepfashion2 due to which classes are imbalanced. This AP is majorly considered as this shows the AP of images whose IoU is 10 likewise whose IoU is 50 shows the AP50 and seems like this good IoU and its average precision too. AP75 is an excellent IoU but the average precision is less for the same. APs show the AP for the images in which objects acquired the small area, same way APm is average precision for the fashion images that are acquired the object in image medium area and finally the APl for those who acquired the larger area. The average precision for small area images is quite good compared to APm, APl.

Table 2 contains the AP's based on each category of clothes. It clearly shows that the dataset did not have the fashion images of category short_sleeved_outerwear and sling. This is why the AP is showing 0 for these two categories.

Table 2: Average Precision per class

category	**AP**	**category**	**AP**	**category**	**AP**
short_sleeved_shirt	31.492	long_sleeved_shirt	20.139	short_sleeved_outwear	0.000
long_sleeved_outwear	9.827	vest	8.321	sling	0.000
shorts	15.855	trousers	39.647	skirt	30.478
short_sleeved_dress	14.425	long_sleeved_dress	4.593	vest_dress	18.296
sling_dress	3.465				

Tested the model by passing the wild images in which group of people are standing with different fashion clothes figure 5 shows the result for the same.

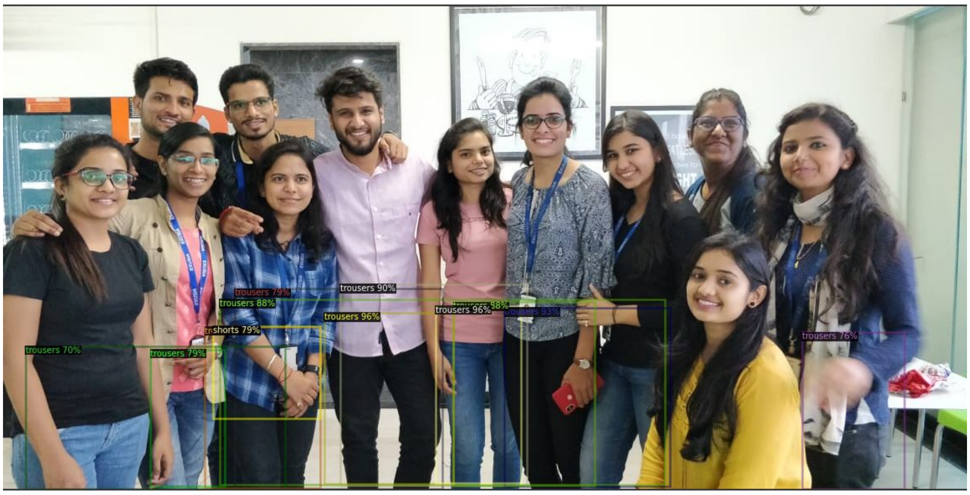

Figure 5: Wild image tested for clothes detection

Table 3: Average Recall at IoU=0.50

Area	all	small	medium	Large
Average Recall	0.669	-1	0.453	0.676

The table 3 contains the recall result in which it has covered with IoU at 50 and the different area wise that small, medium, and large. Average Recall for all areas at IoU 50 is 60% for the small is negative due to imbalance of data and the medium as well large gives significant Average Recall of 45% and 67% respectively.

6.2 Mask RCNN

The aim of this experiment is to train the Mask RCNN model on deepfashion2 dataset and categorize the clothes. This model is based on instance segmentation in which the landmarks are a key part of training the model. It is an extended version of Faster RCNN, so it has outperformed in terms the accuracy [AP], loss even in random image tests. Used the 100 layered FPN based mask RCNN which is provided under model zoo library of detectron2. As this model deals with the bounding box as well as the segmentation here are the AP metrics for both. Used the same dataset as used in Faster RCNN for setting up the benchmark between both models that is 10k images for train, 2k for the test, and 5k for validation.

Table 4: Evaluation results for Bounding boxes in Mask RCNN

AP	**AP50**	**AP75**	**APs**	**APm**	**APl**
21.050	29.749	23.881	nan	23.649	21.386

The table 4 contains Average precision for bounding boxes in mask RCNN . APs showing NaN due to there were no fashion images in which clothes covering small areas. APm for the images that acquire the medium area is 23.6 wherein for a large area is 21.3. The AP50 is larger than all of the other AP's that is 29 and AP75 is 23 this all belongs to IoU metrics. Looking at the results can say that Mask RCNN's AP for bounding boxes is much good but still, there is a gap due to imbalanced classes

Table 5: Average Precision per class in Mask RCNN bounding boxes

category	**AP**	**category**	**AP**	**category**	**AP**
short_sleeved_shirt	43.258	long_sleeved_shirt	27.411	short_sleeved_outwear	0.000
long_sleeved_outwear	11.193	vest	16.855	sling	0.000
shorts	24.501	trousers	49.330	skirt	37.733
short_sleeved_dress	20.556	long_sleeved_dress	7.689	vest_dress	27.779
sling_dress	7.350				

Table 5 contains the per-category AP's for the bounding boxes in Mask RCNN.Table 6 contains the average precision metrics for the segmentation in Mask RCNN. All the AP's regarding area coverage and IoU here in the segmentation has been decreased, the major fall in APm, this deals with segmentation in which all the landmarks must be trained in detail.

Table 6: Evaluation results for segmentation in MaskRCNN

AP	**AP50**	**AP75**	**APs**	**APm**	**APl**
19.903	29.159	22.219	nan	14.808	20.812

Table 7: Average Precision per class in Mask RCNN Segmentation

category	**AP**	**category**	**AP**	**category**	**AP**
short_sleeved_shirt	42.332	long_sleeved_shirt	25.482	short_sleeved_outwear	0.000
long_sleeved_outwear	6.656	vest	16.194	sling	0.000
shorts	22.566	trousers	45.276	skirt	38.499
short_sleeved_dress	20.520	long_sleeved_dress	7.826	vest_dress	26.348
sling_dress	7.044				

Table 7 contains the average precisions for all categories which is evaluated based on the segmentation from Mask RCNN.

Figure 6 shows the wild fashion images test for segmentation model that is Mask RCNN. This is the same image passed to FasterRCNN model and it categorizes more clothes compare to FasterRCNN.

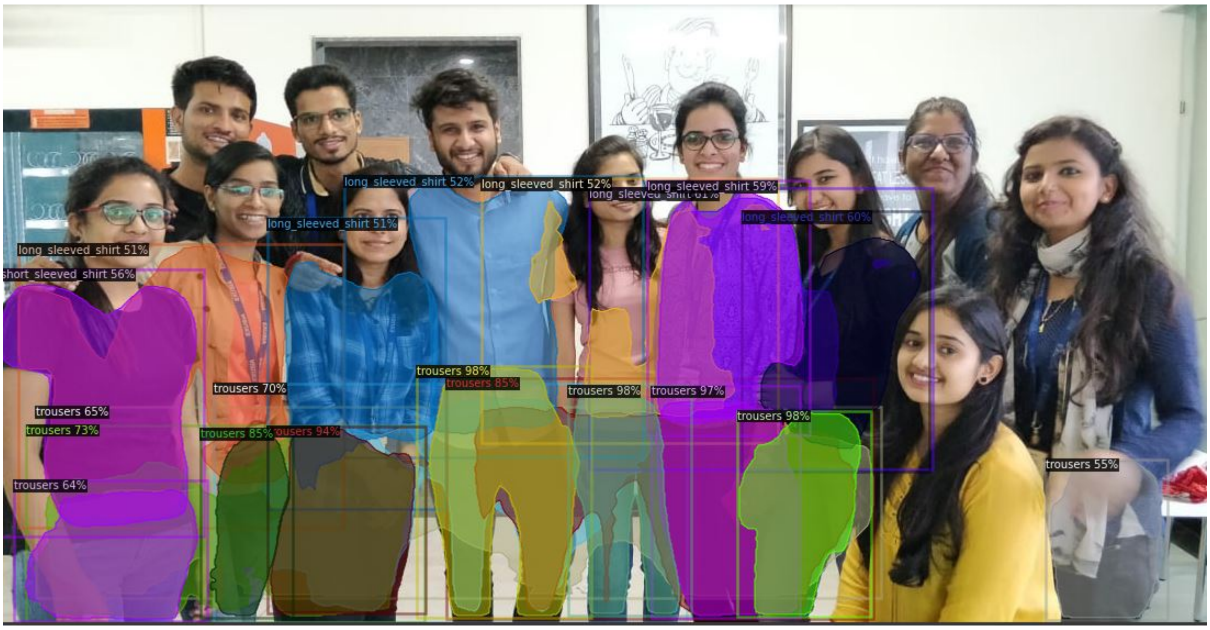

Figure 6: Wild image tested for clothes segmentaion

Most of the key clothes this mask RCNN has been detected and even the person who is acquiring small area in the whole wild image. The key part is that in other test images as well this model did not misclassify the clothes which can be seen in Faster RCNN that is object detection models.Table 8 contains the average recall values of results for the fashion images acquiring the area as all, small and large from fahion image.

Table 8: Average Recall at IoU=0.50

Area	all	small	medium	Large
Average Recall	0.73	-1	0.64	0.73

The Average recall for the fashion images covering all aread has 73%, small is negative due to there were no images captured from given dataset which has small area covered.

Medium and large area covered recall is 64% and 73% which is much good compare to Faster RCNN.

All the results for Mask RCNN and Faster RCNN considering the different metrics such as Average Precision, IoU and the recall it shows that segmentation performs well while categorizing the clothes from wild fashion images. This is also exploited in the test images passed to trained model. Mask RCNN result can see that more clothes are categorized correctly compare to Faster RCNN. Noteworthy point is how the bounding boxes are shown in object detection that is faster RCNN and the landmarks in segmentation which is Mask RCNN.

7 Conclusion and Future Work

The aim of this research is to show that segmentation performs good compare to object detection model when it deals with categorizing the clothes from wild fashion images. This research proposes model implementation based on segmentation and object detection using pre-trained models. The dataset used to accomplish is deepfashion2. Results of both models demonstrated that Mask RCNN the instance segmentation outperforms in terms of recall, average precision. The limitation of this research is the dataset used a subset of deepfashion2 which is a huge dataset of train images around 300k fashion images with rich annotation to depict and implement the fashion-related tasks.

For the future work it can be improved by using all the datasets of deepfashion2 images wherein the object detection models can also give better clothes categorization. For object detection it can be improve by using other 101 layered FPN based faster RCNN which is provided by Facebook API detectron2 under the model zoo library. As the dataset is huge it can be run on a higher configured compute node having strong GPU capabilities. Furthermore the intelligent system on top of this research can be benefits in virtual try-on chaos which is currently a need on e-commerce providers. Also this can help in criminal detection based on the apparels, can be beneficial to understand the fashion trends.

Acknowledgement

I would like to thank you professor Dr. Paul Stynes and Dr. Pramod Pathak for their valuable guidance and encouragement to accomplish this research project. I continue my gratefulness towards the National College of Ireland for giving me this great opportunity.

References

Anantharaman, R., Velazquez, M. and Lee, Y. (2018). Utilizing mask r-cnn for detection and segmentation of oral diseases, *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2197–2204.

Chao, Y.-W., Vijayanarasimhan, S., Seybold, B., Ross, D. A., Deng, J. and Sukthankar, R. (2018). Rethinking the faster r-cnn architecture for temporal action localization, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1130–1139.

D, T. D. and V, K. (2021). Deep learning based object detection using mask rcnn, *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1684–1690.

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M., Belongie, S., Luo, J., Datcu, M., Pelillo, M. and et al. (2021). Object detection in aerial images: A large-scale benchmark and challenges, *IEEE Transactions on Pattern Analysis and Machine Intelligence* p. 1–1.
URL: *http://dx.doi.org/10.1109/TPAMI.2021.3117983*

Feng, Z., Luo, X., Yang, T. and Kita, K. (2018). An object detection system based on yolov2 in fashion apparel, *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pp. 1532–1536.

Ferreira, B. Q., Costeira, J. P., Sousa, R. G., Gui, L.-Y. and Gomes, J. P. (2019). Pose guided attention for multi-label fashion image classification.

Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X. and Luo, P. (2019). Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images.

Grana, C., Borghesani, D. and Cucchiara, R. (2012). Class-based color bag of words for fashion retrieval, *2012 IEEE International Conference on Multimedia and Expo*, pp. 444–449.

Heilbron, F. C., Pepik, B., Barzelay, Z. and Donoser, M. (2019). Clothing recognition in the wild using the amazon catalog, *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3145–3148.

Jia, M., Zhou, Y., Shi, M. and Hariharan, B. (2018). A deep-learning-based fashion attributes detection model.

Kayed, M., Anter, A. and Mohamed, H. (2020). Classification of garments from fashion mnist dataset using cnn lenet-5 architecture, *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*, pp. 238–243.

Kim, H. J., Lee, D. H., Niaz, A., Kim, C. Y., Memon, A. A. and Choi, K. N. (2021). Multiple-clothing detection and fashion landmark estimation using a single-stage detector, *IEEE Access* **9**: 11694–11704.

Liu, Z., Luo, P., Qiu, S., Wang, X. and Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1096–1104.

Noor, S., Waqas, M., Saleem, M. I. and Minhas, H. N. (2021). Automatic object tracking and segmentation using unsupervised siammask, *IEEE Access* **9**: 106550–106559.

Pliuhin, V., Korobka, V., Karyuk, A., Pan, M. and Sukhonos, M. (2019). Using azure machine learning studio with python scripts for induction motors optimization web-deploy project, *2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S T)*, pp. 631–634.

Ren, S., He, K., Girshick, R. and Sun, J. (2017a). Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6): 1137–1149.

Ren, S., He, K., Girshick, R. and Sun, J. (2017b). Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6): 1137–1149.

Sharma, N., Jain, V. and Mishra, A. (2018). An analysis of convolutional neural networks for image classification, *Procedia Computer Science* **132**: 377–384. International Conference on Computational Intelligence and Data Science.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050918309335*

Xiang, J., Dong, T., Pan, R. and Gao, W. (2020). Clothing attribute recognition based on rcnn framework using l-softmax loss, *IEEE Access* **8**: 48299–48313.

Zhang, X., Song, C., Yang, Y., Zhang, Z., Zhang, X., Wang, P. and Zou, Q. (2020). Deep learning based human body segmentation for clothing fashion classification, *2020 Chinese Automation Congress (CAC)*, pp. 7544–7549.

Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y. and Yang, R. (2019). Iou loss for 2d/3d object detection, *2019 International Conference on 3D Vision (3DV)*, pp. 85–94.