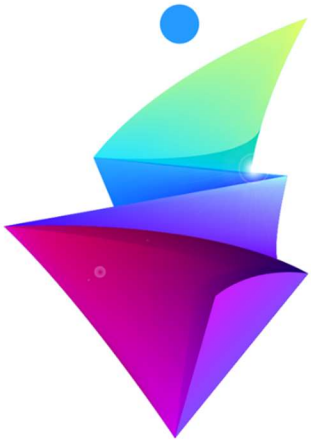
Department of Electronic and Information Engineering

Final Year Project (2022/23)



HUMANS
Holistic Uncover the Missing
by Artificial Neural System

[3]

Final Report

Student Name:	Pan Yuchen (Peter)
Student ID:	19079494D
Programme Code:	42477
Supervisors:	Dr. Bonnie Law Mr. Andy Yip Ms. Tracy Lee
Submission Date:	17th Apr 2023

Abstract

This paper presents a comprehensive study on utilizing cutting-edge machine learning models to address challenges in elderly garment recognition and synthesis. We focus on face detection, face recognition, garment recognition using CLIP, and full-body image synthesis using StyleGAN2, underpinned by rigorous literature review and methodology. We propose a novel dataset, the WISE dataset, featuring 12,030 full-body images predominantly of elderly individuals, which we use for training and evaluation. We explore data augmentation and fine-tuning techniques for the SCRFD model in face detection and the ArcFace model in face recognition, demonstrating improvements in mean average precision and accuracy. Our work extends the application of OpenAI's CLIP model for natural language-based garment retrieval, where we achieve superior performance through fine-tuned model on background-removed images. We employ the StyleGAN2 model to synthesize full-body images of elderly individuals with a Fréchet Inception Distance (FID) of 16.979, while assessing disentanglement using GanSpace and examining textual guidance via StyleCLIP. Through this research, we provide valuable insights and resources for the development of AI applications tailored to find missing elderly, facilitating advancements in the recognition and synthesis of garments for this demographic.

Contents

List of Figures

List of Tables

Acknowledgment

I wish to convey my profound gratitude to all those who have contributed to the triumphant completion of this project. Primarily, I am deeply beholden to Professor Bonnie, whose steadfast support, guidance, and commitment were pivotal in molding my research. Our weekly meetings, enriched by her invaluable insights, navigated me towards the right direction and enabled me to tackle the core issues of the project. Furthermore, Professor Bonnie's benevolence, empathy, and genuine enthusiasm for venturing into uncharted territories significantly enhanced the project and cultivated a thriving environment for growth and learning.

Additionally, I sincerely appreciate the distinguished officers of HKPF, Mr. Andy and Ms. Tracy. Their collaboration was indispensable in shaping the project's trajectory. Their expertise and guidance notably augmented the quality and relevance of the research, and their assistance merits the highest praise.

I am also tremendously grateful for the munificent support of Google TRC, which granted us the opportunity to perform our computationally demanding deep learning model training on GCP's TPU. This bountiful resource empowered us to transcend the limitations of our research and accomplish results that would have otherwise remained beyond our reach. Moreover, I wish to recognize Prof. Mak for imparting his wisdom and offering invaluable counsel for this project. His expertise and input were vital in guaranteeing the project's success.

In summary, I am truly blessed to have been encircled by such a devoted and supportive assembly of individuals. It is with immense gratitude that I recognize their invaluable contributions to this project.

1. Introduction

The rapid urbanization and increasing number of elderly citizens in Hong Kong have intensified the challenge of locating missing persons, particularly among the elderly population. In response to this critical issue, the Hong Kong Police Force has initiated this FYP, HUMANS (Holistic Uncover Missings by Artificial Neural System). This project, co-directed by Prof. Bonnie, Mr. Andy, and Ms. Tracy, aims to harness the power of deep learning models to assist in the swift and accurate identification of missing individuals, ultimately improving the efficiency and effectiveness of search efforts.

HUMANS is designed as a **two-pronged** approach to address the complexities of locating missing persons. The first component focuses on face recognition, leveraging cutting-edge face detection and face recognition models, which have been meticulously adapted and fine-tuned to suit this specific application scenario. The second component is centered on clothing recognition, aiming to match pedestrians' attire based on natural language input. To achieve this, we explore two distinct methods: (1) employing CLIP's [2] multi-modal capabilities to directly detect similarities between natural language descriptions and pedestrians' full-body clothing; and (2) using text input to guide image generation models, such as StyleGAN [4], in simulating corresponding full-body images of individuals and subsequently searching for similar pedestrians through image-to-image matching.

This paper is organized as follows: The Literature Review section provides a comprehensive overview of the face detection mode – SCRFD [5], face recognition model – ArcFace [6], CLIP model, StyleGAN model, and several multi-modal clothing databases, highlighting the novelty of our approach within the broader research landscape. In the Methodology section, we outline the data augmentation techniques for fine-tuning face detection and recognition models, the creation of a large-scale dataset using a deep learning filtering pipeline, and the implementation of the two natural language-based clothing matching methods.

In the Result Analysis section, we thoroughly examine the improvements in face detection and face recognition models after fine-tuning, and provide a comprehensive summary and analysis of our collected dataset, WISE, as well as the two natural language-based clothing matching methods. Through this analysis, we demonstrate the effectiveness of our proposed solutions and their potential for real-world applications in aiding the Hong Kong Police Force and other relevant stakeholders in their quest to quickly and accurately locate missing persons.

In conclusion, the HUMANS project offers an innovative and promising approach to tackling the pressing issue of locating missing individuals, particularly the elderly population. By harnessing the power of deep learning models and combining face and clothing recognition techniques, we provide a comprehensive and efficient solution that can be seamlessly integrated into existing police work systems. Our research contributes to the ongoing development of intelligent search and rescue systems, with the potential to greatly benefit the Hong Kong Police Force and other organizations involved in locating missing persons.

2. Literature Review

This section provides a concise overview of the critical methods and datasets used or relative to our project in face detection, face recognition, garment recognition using CLIP, and full-body image synthesis. The section is organized into four subsections, starting with face detection and the SCRFD model, followed by face recognition focusing on the ArcFace model. Next, we discuss the potential of the CLIP model in garment recognition, and finally, we delve into full-body image synthesis, emphasizing the StyleGAN model and its applications.

2.1. Face Detection

Face detection is a critical component of this project, serving as the foundation for face recognition. In the literature, various face detection models have been developed. The face detection model in this project is required to meet a set of criteria, including efficiency, accuracy, and adaptability to different circumstances. To address these requirements, the model will be fine-tuned with an augmented dataset containing numerous small faces.

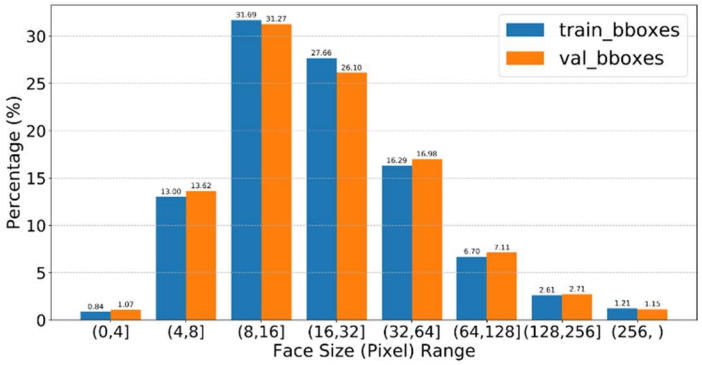


Figure 1. Face Size in WIDER FACE: Distribution of face hight from training (blue bar) and validation (orange bar) dataset.

2.1.1. WIDER FACE Dataset

The WIDER FACE dataset [7] is a challenging benchmark and a training dataset for face detection research due to its large number of small-scale faces and its wide

variety of face poses and partial occlusions. As shown in Figure 1, small faces make up over half of the dataset, while their recognition accuracy is exceptionally low (Figure 2). We will utilize these features in the dataset with data augmentation to fine-tune the face detection model.

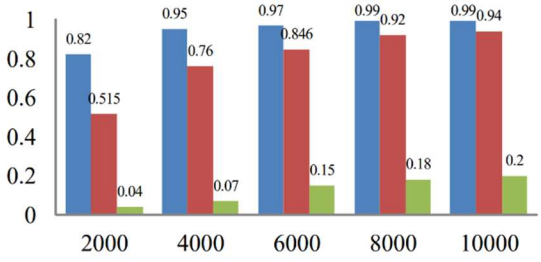


Figure 2. Face Detection Accurate Rate Baseline for Different Face Sizes (*X-axis denotes for average number of proposals per image.)*

2.1.2. SCRFD

Several face detection models have been proposed to handle the diverse and challenging WIDER FACE dataset. TinaFace [8] is one such model, which has demonstrated top-performing results with an AP score of 0.8143 in WiderFace's Hard face detection category. However, despite its high accuracy, the model's substantial computational requirements (Figure 3) make it unsuitable for certain applications, particularly those with limited computational resources. As a result, researchers have sought alternative models that maintain high accuracy while re

Method	Backbone	Easy	Medium	Hard	#Params(M)	#Flops(G)	Infer(ms)
TinaFace [41] (Arxiv20)	ResNet50	95.61	94.25	81.43	37.98	172.95	38.9
SCRFD-10GF	Basic Res	95.16	93.87	83.05	3.86	9.98	4.9

Figure 3. Comparison of accuracy and efficiency between TinaFace and SCRFD on WIDER FACE dataset

SCRFD [5] is one such model, which builds upon the success of TinaFace while significantly reducing computational requirements through various optimizations. This model represents a promising alternative, which significantly improved its efficiency without sacrificing accuracy.

2.2. Face Recognition

2.2.1. ArcFace

ArcFace is a deep-learning face recognition model demonstrating superior performance due to its uncomplicated architecture. Its architecture consists of two main components: the *backbone* and the *head*. The *backbone* is a variant of ResNet[9], while the *head* is a single fully connected layer used for classifying different individuals during the training phase and discarded (removed) in tests and actual interference scenarios. ArcFace employs the PartialFC [10] technique in the head to significantly reduce GPU memory consumption and enhance training efficiency for large datasets. The model's

exceptional performance is primarily attributed to its aggressive loss function [11]. This section will thoroughly analyze the architecture, including the structure and loss function, as well as the residual blocks and the ResNet-IR backbone.

A. Residual Block

The Residual Block serves as the foundational component of ResNet. Before examining the ResNet backbone, this section will carefully investigate the unique residual block configuration in ArcFace.

i.) Original Residual Block

The Residual Block is central to the success of ResNet. According to He et al. [9], the Residual Block superimposes its input onto its output via a residual connection (or skip/shortcut connection), enabling the block to learn only the residue. Equation 1 defines the formula for a residual block:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$$

Here, x and y represent the input and output of a Residual Block, respectively. For a Residual Block comprising two convolutional layers, $\mathcal{F}$ is defined as $\mathcal{F} = \mathrm{W}_2\sigma(\mathrm{W}_1\mathrm{x})$, where σ denotes the ReLU [12] activation function. Additionally, a second ReLU is applied to the output y.

ii.) Improved Residual Block

Using the original Residual Block, Deng et al. [11] proposed a modified Residual Block composed of two convolutional layers. The structure of the improved Residual Block (referred to as "IR") is BN-Conv-BN-PReLu-Conv-BN, as depicted in Figure 4. Improved Residual Block. Compared to the original Residual Block, IR exhibits the following characteristics:

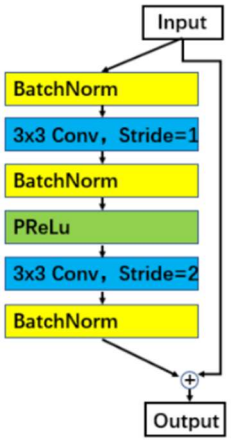


Figure 4. Improved Residual Block

- First, the ReLU initially applied to the merged output y is removed, preventing the blocking of negative propagation on the mainstream due to the ReLU placed on the mainstream [13].
- Second, IR incorporates batch normalization before the first convolutional layer, accelerating convergence and fully exploiting each Residual Block's potential [14].
- Third, the ReLU activation function is replaced with PReLU to achieve better performance with a negligible increase in computational cost [15].

B. ResNet-IR Backbone

ArcFace modifies not only the micro level (the Residual Block) but also the macro level, based on a detailed benchmark comparison [16]. This section will first discuss the original ResNet structure and its variants before examining ArcFace's macro-level modification, the modified backbone's performance, and the project's choice of backbones in terms of their varying depths.

i.) ResNet and Its Variants

The development of ResNet represents a groundbreaking achievement in deep learning, and its paper [9] received the CVPR Best Paper Award in 2016. By the end of 2022, the paper had accumulated 147k citations on Google Scholar. Inspired by ResNet, numerous pioneers in the fields of computer vision and pattern recognition have innovated several ResNet variants. These variants include Inception-ResNet [17], which stems from the addition of residual connections to the Inception network architecture [18]; DenseNet [19], which employs residual connections more radically than ResNet; and SE-ResNet [20], which incorporates global pooling into the Residual Block. Although these variants typically outperform ResNet, they do so at a significantly higher computational cost.

ii.) ResNet with IR

For the bottleneck residual block structure with three convolutional layers in the original ResNet whose depth is greater than 50, ArcFace replaces it with IR. Consequently, in ArcFace, each Residual Block contains only two convolutional layers. Thus, the backbone structure of the original ResNet has been adjusted accordingly (denoted as ResNet-IR). Table 3 and Table 4 display the backbone structure before and after modification.

iii.) Comparison of Different Backbones

Deng et al. [11] conducted a detailed comparison experiment of various backbones. The results show the outstanding efficiency of ResNet-IR while maintaining a very high accuracy rate.

Although ResNet represents a qualitative leap in addressing gradient vanishing and explosion, degradation still occurs in ultra-deep models. An experiment [13] shows the verification accuracy of ResNet-200 is lower than that of ResNet-152. Consequently, even within the range where performance degradation does not occur, increasing the number of layers gradually results in diminishing accuracy gains, or diminishing marginal benefits.

Considering the accuracy of the face recognition model, the computational speed, and the characteristics of ResNet itself, this project selects the ResNet-IR with a depth of 50 as the backbone.

2.2.2. Glint360k Dataset

The Glint360k dataset [10], later face-aligned by the Insightface Team [21], is utilized for fine-tuning the ArcFace model to achieve better results. This dataset is extensive, consisting of 17,091,657 photos corresponding to 360,232 individuals. On average, there are 44.45 images for one person. Due to its large size, data augmentation is expected to expand the dataset significantly, which will be discussed in the following section.

2.3. Garment Recognition By CLIP

In this section, we discuss the utilization of CLIP for garment recognition, which serves as the second component of our HUMANS project. CLIP's multi-modal capabilities enable it to match pedestrians' attire based on natural language input. We first review traditional image retrieval methods and their limitations, followed by an examination of deep learning image retrieval methods. We then provide an overview of the CLIP model, detailing its architecture, pre-training, and strengths in handling both images and natural language.

2.3.1. Traditional Image Retrieval Methods

Image content-based retrieval has a long history of development, and typical approaches in the traditional non-deep learning approach include Bag-of-Visual-Words (BoVW) [22], Scale-Invariant Feature Transform (SIFT) [23], and Global Features (e.g., GIST [24] and Color Histograms [25]). These techniques primarily rely on quantizing visual features into a fixed set of visual words or using handcrafted features that capture specific image properties, followed by a similarity measure such as Euclidean distance or cosine similarity. While these methods have demonstrated reasonable performance in the past, they struggle with the challenges of scale, viewpoint, and illumination variations in images. Additionally, these traditional methods lack the ability to understand human natural language description directly due to their reliance on handcrafted features and their lack of integration with natural language processing techniques, which are necessary for comprehending textual input.

2.3.2. Deep Learning Image Retrieval Methods

Deep learning methods have gained significant attention in image retrieval tasks because they can learn complex and high-level features from data. Some popular deep learning methods for image retrieval are based on Convolutional Neural Networks (CNNs) [26, 27], Recurrent Neural Networks (RNN) [27], and Triplet Networks [28].

A. CLIP

We chose to use CLIP for our garment recognition tasks due to its remarkable ability to understand and match both images and natural language. CLIP surpasses alternative deep learning models like fine-tuned ResNet-50 in many datasets for the task of zero-shot image classification, as shown in Figure 5. In zero-shot image retrieval, CLIP outperforms VSE++ [26] and SCAN [27] on the MS-COCO dataset [29]. Furthermore, CLIP has demonstrated superior performance to Deep Ranking [28] on fine-grained image retrieval tasks.

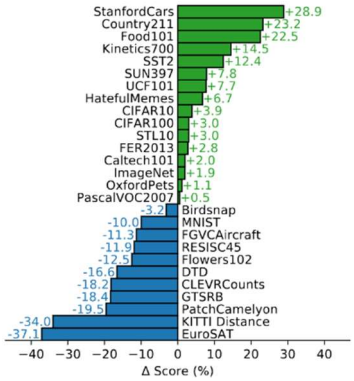


Figure 5: Zero-Shot CLIP vs. Linear Probe on ResNet50 [2].

The base line (Score 0%) is the classification accuracy of fine-tuned (on each dataset) ResNet-50 model, while the green and blur bar represent how much the one-shot classification of CLIP model surpass or below the base line.

In terms of applications, its capabilities in image retrieval based on image content have been demonstrated by a search engine built by AI enthusiasts using CLIP combined with the Laion5B database [30] (currently the most extensive public image dataset containing 5 billion images). In the following, we will give a short analysis of the structure of CLIP and how it is trained.

B. CLIP Model Architecture

This section briefly introduces the model architecture of CLIP, focusing on its vision and language models, which are essential for understanding the underlying mechanisms that enable the model's high performance in image retrieval tasks.

i.) Vision Model: Vision Transformer (ViT)

The vision model in CLIP is based on the Vision Transformer (ViT) architecture [31]. Unlike traditional convolutional neural networks (CNN), ViT treats an image as a sequence of fixed-size, non-overlapping patches, as shown in Figure 6. These patches are linearly embedded into flat vectors and combined with position embeddings to preserve spatial information. The resulting sequence of vectors is then fed into a standard Transformer architecture [32], which processes the information through self-attention mechanisms and learns to recognize patterns and objects within the image.

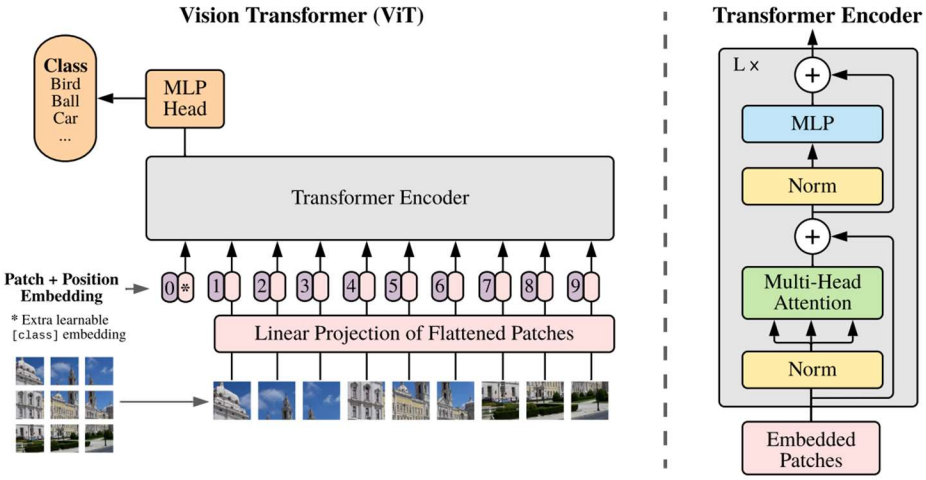


Figure 6: The main principle and structure of ViT [31]

ii.) Language Model: Transformer

CLIP's language model is based on the Transformer architecture [13], which has become the de facto standard for natural language processing tasks. The current most notable chatbot, ChatGPT and Bard, are all based on Transformer. The Transformer model leverages self-attention mechanisms to process input sequences in parallel rather than sequentially as in traditional recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs). The sequential process of RNNs and LSTMs leads to increased training times and difficulty in learning long-range dependencies due to vanishing gradients. In contrast, the Transformer's attention mechanism enables it to be computationally more efficient and better equipped to capture long-range dependencies in data.

C. Pre-Training of CLIP

The CLIP model is trained on a large-scale dataset containing image-text pairs from the web. The dataset includes 400 million image-caption pairs sourced from web pages. The images and captions are preprocessed and tokenized before being fed into the model.

Due to the strong understanding of both verbal and graphic content and its good matching ability between the two domains, we choose the pre-trained CLIP model (ViT-L/14@336px in particular, a relatively large model from CLIP) as the baseline for matching pedestrian clothing with natural language description (garment recognition).

2.4. Full-body Image Synthesis

In this section, we discuss various deep learning methods for generating full-body images, with a focus on StyleGAN and its applications. These methods enable the generation and manipulation of realistic images, which play a crucial role in our approach to matching missing individuals based on their clothing descriptions. We also examine the availability and characteristics of full-body image datasets, which are essential for training and fine-tuning the generative models for our specific application.

2.4.1. StyleGAN and Its Applications

A. StyleGAN2

StyleGAN2 [4] is a deep learning-based generative model for image synthesis introduced by Nvidia in 2019. It improves upon the original StyleGAN model by providing more stable training, better disentanglement of image features, and better image quality. The architecture of StyleGAN2 consists of a generator network and a discriminator network. The generator network generates new images from a random noise vector, while the discriminator network distinguishes between authentic and generated images. StyleGAN2 also uses adaptive instance normalization (AdaIN) [33] to control the style of the output images.

B. GanSpace: Interpretable Latent Space Analysis

GanSpace [34] is a technique for visualizing and manipulating the latent space of a StyleGAN2 model. It disentangles the various features of the latent space and provides users with intuitive controls for manipulating these features. GanSpace first analyzes the latent space (w^+ in our case) of a StyleGAN2 model and identifies the features that contribute to generated images. These features are visualized using principal component analysis (PCA), allowing users to see the dimensions of the latent space and their relationships to generated images

C. StyleCLIP: Text-guided Image Manipulation

StyleCLIP [35] combines the generative ability of StyleGAN and the multi-modal ability of CLIP, allowing users to control the style and content of generated images using natural language descriptions. StyleCLIP takes textual input from the user and encodes it into a latent vector using CLIP. This vector is fed into the StyleGAN2 generator network to produce an image matching the user's description.

D. Other DL Methods

i.) Variational Quantum Eigensolvers (VQE)

Variational Autoencoders (VAEs) [36] are generative models that probabilistically map latent spaces to data spaces. VAEs' encoder and decoder components learn probabilistic latent space distributions and data sample generation. Based on VQE, Jiang, et al. [37] successfully built a fashion full-body image generative network Test2Human. However, VAE-generated images are generally inferior to those produced by StyleGAN, exhibiting increased blurriness and fewer fine-grained details [38].

ii.) Diffusion Generative

Diffusion generative models are a class of deep learning-based generative models that learn to produce samples from complex data distributions by simulating a diffusion process. These models employ a denoising score-matching network, which predicts the score function of the data distribution given noisy observations, along with a forward diffusion process.

2.4.2. Full-body Images Dataset

To develop a StyleGAN capable of generating full-body images of elderly individuals based on textual input, a substantial collection of high-quality, full-body images of this demographic is required. However, no such datasets are currently publicly available within the academic community. By broadening the scope, datasets containing primarily younger individuals in full-body portraits can be considered, as detailed in the following sections.

A. SHHQ

The Stylish Humans HQ (SHHQ) [39] dataset encompasses 231,176 high-quality, full-body images, along with corresponding human-segmentation and clothing-body segmentation masks, as illustrated in Figure 7. This dataset also includes labeled clothing attributes, which prove beneficial for tasks related to human generation and editing. Regrettably, only a subset of this dataset, consisting of 40,000 images, is currently available to the public.

Figure 7. Image Samples of SHHQ [39]

B. FashionGen

FashionGen is a dataset comprising 293,008 high-definition (1360 x 1360 pixels) fashion images, accompanied by item descriptions crafted by professional stylists [3]. These images are divided into 260,480 for training and 32,528 for validation. Each item is photographed from multiple angles against a standardized background under consistent lighting conditions and is annotated with stylist-recommended matching items. Unfortunately, only a low-resolution (256 x 256 pixels) version of this dataset is presently accessible via a third-party shared file.

Fashion-Gen is a dataset of 293,008 (260,480 for training, 32,528 for validation) high definition (1360 x 1360 pixels) fashion images paired with item descriptions provided by professional stylists [40], as shown in . Each item is photographed from a variety of angles against a standardized

background under consistent lighting conditions and annotated with matching items recommended by a stylist. Unfortunately, only a low resolution (256 × 256 pixels) version of this dataset is currently publicly available through a third paity shared file.

Figure 8. Samples of Fashion-Gen [40]. Left Images are photographed from different angles and correspond to the text descriptions on the right side

C. DeepFashion-MultiModal

The DeepFashion-MultiModal dataset [37] is a large-scale, high-quality human image dataset. The dataset contains 11,484 high-quality images at 1024 x 512 resolution, collected from the high-resolution version of the DeepFashion dataset and further cleaned and selected to ensure they are full-body and of good quality. The dataset contains texture annotations, human parsing masks with diverse fashion attribute classes, and human poses. Both the textual attribute annotations and human parsing masks are manually labeled. The dataset is annotated with attributes for clothes shapes and textures, including the length of upper and lower clothes, the presence of fashion accessories, and the shapes of the upper clothes' necklines.

3. Methodology

In this section, we present the methodology employed in our study, which is organized into several subsections for clarity and ease of understanding. The methods, models, and datasets discussed here, such as Face Detection (SCRFD Model), Face Recognition (ArcFace Model), CLIP, and StyleGAN, are widely used and familiar to researchers in the field.

3.1. Face Detection focuses on detecting faces in images, with emphasis on data augmentation techniques (3.1.1.) and fine-tuning of the SCRFD model (3.1.2.).

3.2. Face Recognition delves into recognizing individuals based on their facial features. The section details data augmentation methods for face recognition (3.2.1.) and fine-tuning of the ArcFace model (3.2.2.).

In 3.3., we describe the web scraping process for building the WISE dataset. This includes the filtering pipeline (3.3.1.) and data collection from Google Image, Twitter API (3.3.2.), and YouTube (3.3.3.).

3.4. Constructing the Eloquent ElderWear (EEW) Benchmark section explains the creation of a benchmark for our study, covering background subtraction (3.4.1.), manual labeling (3.4.2.), and the evaluation of top-k accuracy (3.4.3.)

In 3.5., we discuss the use of the CLIP model for natural language garment retrieval. This involves the selection of a baseline model (3.5.1.), fine-tuning the CLIP model (3.5.2.), and evaluating its performance using various metrics (3.5.3.).

Finally, 3.6. presents the synthesis of full-body garment imagery using StyleGAN2 and fine-tuned CLIP models. The section covers pre-training StyleGAN2 on the SHHQ dataset (3.6.1.), integrating CLIP for textual guidance (3.6.2.), color extraction and modification using OpenAI's Chat API (3.6.3.), and fine-tuning StyleGAN2 on the WISE dataset (3.6.4.).

3.1. Face Detection

In this section, we present the methodology for enhancing the face detection model to address the specific requirements of the HUMANS project. We begin by discussing the data augmentation

techniques employed, followed by the fine-tuning of the SCRFD model, a state-of-the-art face detection algorithm.

3.1.1. Data Augmentation for Face Detection

There are three special needs in this project's actual application scenario, i.e., masked faces, blurred faces, and dim light environments. Considering the lower difficulty of face detection compared to face recognition, intuitively speaking, face detection often requires only the recognition of the general outline features of a face to determine. In contrast, face recognition requires very detailed features to extract accurate embedding features for face comparison. Therefore, face detection is more robust to blurred and unclear faces in darker environments than face recognition.

Thus, in the data augmentation of face detection, blurred faces and dim light environments will not be considered, and only the masked faces will be focused. This project uses the technique proposed by Anwar and Raychowdhury [41] and their open-source implementation called MaskTheFace, which is modified in this project. MaskTheFace is a maturing and efficient solution for adding different styles and angles of masks to faces.

MaskTheFace provides rich combinations of different mask types, colors, and textures. In order to reduce the dependence of the model on specific masks in fine-tuning the face detector model, this project makes full use of the different properties of masks provided in MaskTheFace. Figure 9 shows the different mask types corresponding to their probability. The gas mask is not used in this project because it is too rare. In addition, the color of the mask is entirely random and has a probability of 0.1 that a random texture or pattern will be added.

Figure 9. Different mask types [1] and corresponding proportionality

Set	Total	Original (Unmasked)	Masked
Training	12,880	6440	6440
Validation	3,226	1,613	1,613
Test	16,097	8,049	8,048

Table 1. Statistics of data augmentation

Based on the above technique and the additional annotation of the five key points of the face of WIDER FACE provided by Deng, et al. [42]. Half of the original training set, which contains 12,880 images, is added with masks, i.e., 6440 images are the original images, and 6440 images are the images with simulated masks. The validation set and the test set are augmented using the same strategy. The

resulting distribution of the training set, validation set, and test set are shown in Table 1. The effect of the simulated masked images is shown in Figure 10.

After creating a suitable dataset, suitable models will be selected. Then transfer learning will be performed on the improved dataset so that the face detection model can be adapted to the actual application scenario.

Figure 10. Examples of masked images from WIDER FACE [7]

3.1.2. Fine-tuning of SCRFD

Based on the WIDER FACE dataset after data augmentation, SCRFD underwent 10,000 epochs of fine-tuning. The mean average precision (mAP) obtained for the validation process in

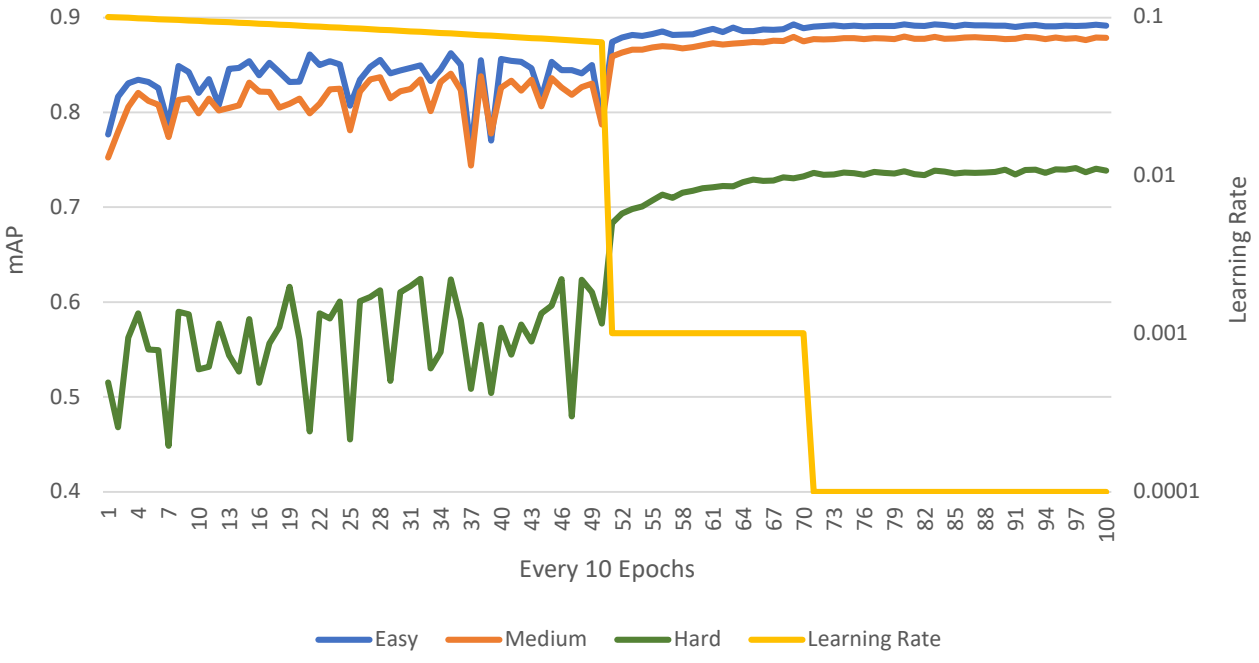


Figure 11. Val mAP and LR during the fine-tuning process of SCRFD

easy/medium/hard categories and the learning rate during the training process are illustrated in Figure 11. It can be observed from the graph that when the learning rate sharply decreased from approximately 0.06 to 0.001, all three groups of mAP achieved significant improvements. However, this phenomenon did not appear when the learning rate further dropped to 0.0001. This observation, along with the subsequent stabilization of mAP, corroborates that the training of this model has been essentially completed.

3.2. Face Recognition

This section delves into the adaptation of the face recognition model for the HUMANS project. We discuss the data augmentation techniques tailored for face recognition, followed by the fine-tuning of the ArcFace model, a cutting-edge face recognition algorithm.

3.2.1. Data Augmentation for Face Recognition

A. Data Augmentation Methods

The data augmentation techniques are tailored to the specific requirements of real-world scenarios, encompassing mask addition, motion blur, and brightness and contrast adjustments.

i.) Adding Masks

A facial masking method [41] incorporates masks into the original dataset. Randomly assigned mask styles, colors, and patterns diminish the face recognition model's reliance on masks, aligning with data augmentation strategies employed in face detection

ii.) Motion Blur

As face recognition heavily depends on precise facial feature extraction, it is intuitively more vulnerable to blurred faces or low-light conditions. Consequently, the data augmentation for face recognition should encompass blurred faces as supplementary cases.

A box blur kernel is employed for simplicity, with motion blur divided into vertical and horizontal components. In real-world scenarios, the relative motion of vehicles and pedestrians is predominantly horizontal, surpassing the vertical displacement caused by vehicle jitter. Accordingly, motion blur is primarily vertical. The kernel sizes for vertical and horizontal motion blur range from 1 to 3 and 3 to 15 integers, respectively, with the effects of different combinations illustrated in Figure 12. Blurred degree regarding different kernel sizesFigure 12.

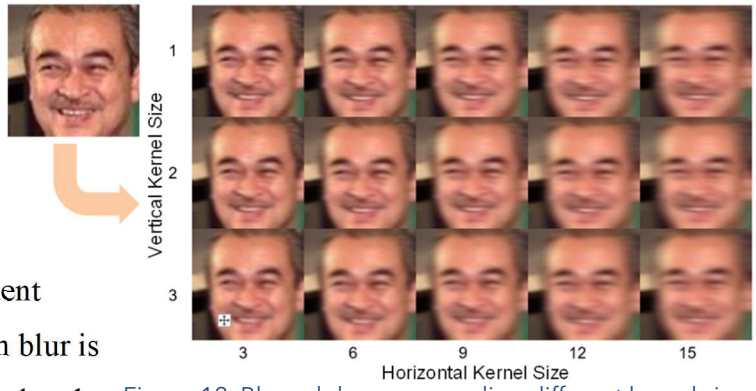


Figure 12. Blurred degree regarding different kernel sizes

iii.) Adjusting Brightness and Contrast

Both brightness and contrast alterations are random. Brightness varies between 0.2 to 0.6 of the original, while contrast ranges from 0.4 to 0.8 of the original images. The effects of different combinations are depicted in Figure 13.

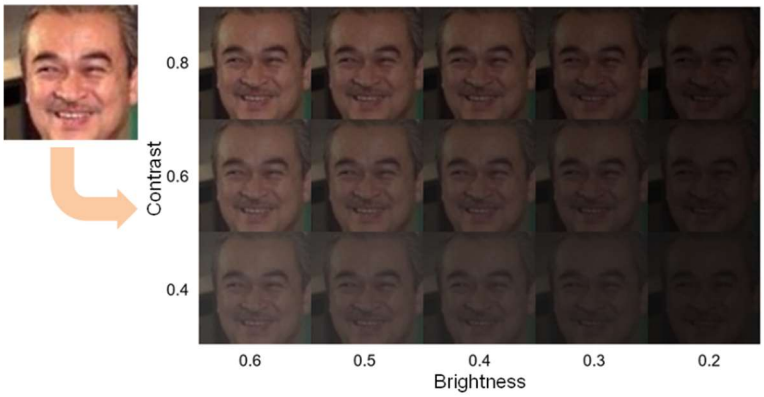


Figure 13. Effects of different brightness and contrast

B. Training Dataset

Training a face recognition deep learning network necessitates extensive data. The data augmentation methods detailed earlier are projected to expand the dataset sixfold. Given the augmented dataset's size, selecting a subset for training is essential.

Two strategies can be considered to address the dataset's substantial size: reducing the number of individuals or reducing the number of images per person. This project adopts the former approach, retaining all images after data augmentation for each person but constraining the number of individuals.

Sub-dataset	Image Amount
Original	5,000,000
Masked	3,552,894
Blurred	5,000,000
Blurred-Masked	3,552,894
Darkened	5,000,000
Darkened-Masked	3,552,894
Total	25,658,682

Table 2. Image amount of each data augmentation categories

The first 73,382 individuals from the Glint360k dataset are selected, encompassing 5M photos. Following data augmentation, the training set comprises 25,658,682 images (Table 2), surpassing the original Glint360k dataset in image quantity.

To further mitigate the inefficiencies resulting from the large dataset, it is divided evenly and randomly into ten sub-datasets. During training, only one sub-dataset is utilized per epoch. At the end of each epoch, the subsequent sub-dataset is loaded into the PyTorch dataloader, allowing periodic use of all sub-datasets. This method ensures that the face recognition model has ample training data and trains relatively quickly while benefiting from the augmented dataset to enhance performance.

C. Validation Dataset

The LFW dataset serves as the validation dataset, using the aligned version provided by the InsightFace Team [21] and following LFW's evaluation protocol [43]. The validation dataset comprises image pairs, with one original and one augmented photo (or predominantly augmented) per pair.

Four distinct validation datasets are created to evaluate model performance from various perspectives:

The first dataset simulates the actual application environment, with 80% masked augmented images, a 0.3 probability of blurred images, and a 0.2 probability of darkened images.

The second, third, and fourth datasets each address one of the three real-world characteristics (masked, blurred, and darkened). In these pairs, one image is the original, while the other is augmented using the corresponding method for that dataset.

Each validation set has 24,000 same-person and 24,000 different-person image pairs, totaling 96,000 images per dataset. This is eight times larger than the official LFW validation pairs [43], ensuring a thorough, fair comparison of model performance across scenarios.

3.2.2. Fine-tuning of ArcFace

The ArcFace model was trained using two RTX 3070 GPUs on a single host, undergoing a total of 60 sub-epochs. The first epoch focused on training the freshly initialized head, the final layer of the model, which is a fully connected layer linking the 512 features of the backbone to the 73,382 individuals in the output layer.

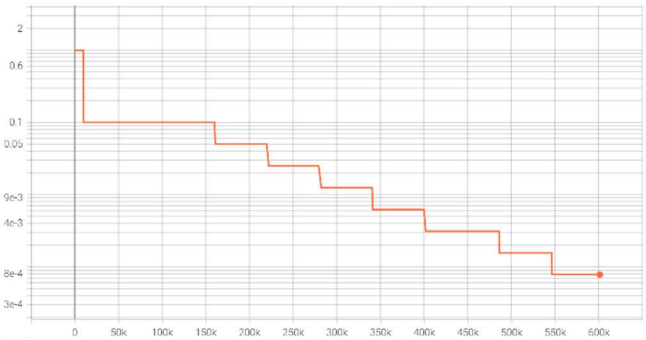


Figure 15. Learning Rate regarding training steps

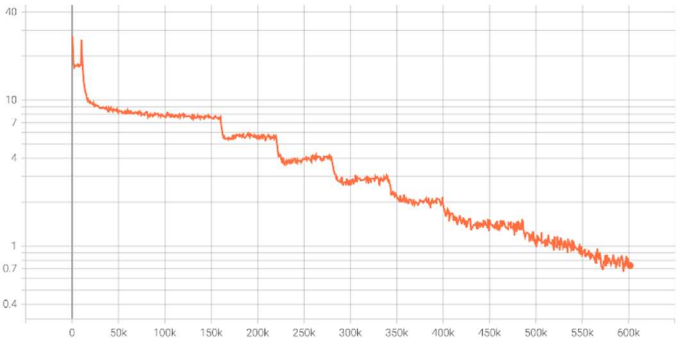


Figure 14. Training Loss corresponding to training steps

Our training approach handles approximately 2.56 million images in each epoch. Each GPU processes 128 images per step, resulting in a total of 256 images per step. Consequently, an epoch comprises approximately 10,000 steps. The 600,000 steps on the x-axis can be roughly equated to the 60th epoch.

Figure 15 presents the learning rate variations throughout the training process, with the x-axis representing the training steps. During the initial epoch, the learning rate was set to 1, enabling rapid head convergence. From the second epoch onwards, all training layers were unfrozen, and the learning rate was adjusted to 0.1. Subsequently, PyTorch's learning rate scheduler, employing a plateau strategy, assumed control. The learning rate was halved whenever a plateau was detected. Seven halving operations were executed throughout the training process, reducing the learning rate from 0.1 to 7.8125e-4.

Figure 14 illustrates the training loss variations, demonstrating significant loss reduction throughout the training process. Comparing this with Figure 15 reveals that learning rate adjustments considerably expedite the learning process.

It is important to note that a decrease in training loss does not guarantee an increase in test accuracy. Despite the substantial size of our dataset, overfitting remains a potential concern. To calculate test accuracy, we employed a mixed version of the validation sets as model input. As the probability of overfitting is minimal at the onset of training, test accuracy calculations begin from the 20th epoch. If overfitting occurs before the 20th epoch, this trend can be identified in the test accuracy from the 20th to the 60th epoch. Figure 16 depicts the test accuracy corresponding to epochs 20 through 60, indicating that while test accuracy appears to plateau towards the end, no signs of overfitting are evident.

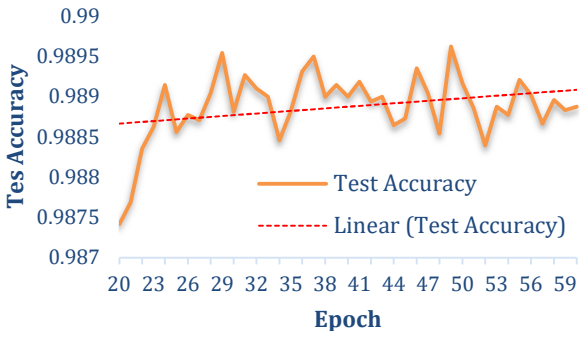


Figure 16. Test Accuracy during the training process corresponding to 20-60 epochs

3.3. Web Scraping for Building WISE Dataset

In the previous section, we discussed how StyleGAN-human is trained on SHHQ, a dataset primarily consisting of images from fashion promotions and sales. Consequently, the images generated by StyleGAN-human predominantly feature young female individuals dressed in fashionable clothing, which is inconsistent with the appearance of elderly people. However, there is no public dataset containing full-body images of elderly individuals as the main subjects. To address this issue, we need to generate a dataset featuring elderly individuals.

Generating such a dataset is challenging, as the majority of full-body images in existing datasets focus on young people in the context of fashion. Thus, it is difficult to extract a sufficient number of full-body images of elderly people from existing public datasets for training StyleGAN. We decided to employ data crawling techniques to gather the necessary images.

Our raw image data comes from four sources grouped into three categories, they are 1. Google Image search and Twitter's API; 2. 2M YouTube videos (our main source); 3. a small subset of Laion5B dataset fetched by CLIP.

Among the millions of original image resources, we need to use a rigorous and meticulous screening mechanism based on multiple methods to select only a few clear full-body images of elderly people with almost no obstruction that can be used by us. To achieve this, we have developed an entire

set of filtering pipelines mainly based on Deep Learning models. The proposed pipeline includes the following modules: 1. Face recognition; 2. Three age estimations; 3. Cartoon detection; 4. Two Human Segmentation; 5. Pose Estimation; 6. Blurry detection (traditional method); 7.Image Quality Assessment; and 8. Super resolution. The following section will provide a detailed analysis and description of this pipeline as well as customized content for different data sources.

3.3.1. Filtering Pipeline

This pipeline has a high level of modularity, where each module operates independently and only transfers images that meet its own requirements to the next module in line.This pipeline contains ten DL models but runs efficiently.In our experiments, steps one through five can run simultaneously on an RTX3070 graphics card while maintaining a high-speed operation for long periods at sixty images per second across two RTX3070 cards.

A. Face Detection

Face detection's importance and relative efficiencies make it suitable as the first mode in out pipeline. It serves several purposes: 1. Detect the presence of a face in the picture (crucial for quickly filtering invalid video frames); 2. Provide face position information for age estimation; 3. Filter images that are unlikely to contain a full body image based on face position.

Figure 17. Illustraion of the accuracy of the face detection model

Here we utilized RetinaFace [42], a face detection model based on the Retina network [44] provided by the InsightFace Team, whose ability is demonstrated in Figure 17, to achieve the first two purposes directly. For the third purpose, we defined the height of the space below the face detection as h_b and the height of the face as h_f. Consider that the average ratio of head and body length for adults is roughly 1:7 [45], and based on the principle of perspective and the fact that the height of the lens is usually close to the head, we can conclude that the height of the body below the head of the person in the photo is more prone to shrink. Thus we define the threshold as 5, and only the images that meet the following condition can pass: $h_b / h_f \geq 5$ (2). As shown in Figure 18, The image is filtered out since, from the position of the face bounding box, it is highly unlikely that the space below the face will fit a human body.

B. Age Estimation

In order to make the dataset mainly composed of elderly people, age estimation methods must be applied. In practice, we have found that current age estimation methods generally perform poorly, such as in images 1 to 3 in Figure 19. which may be due to significant differences in facial clarity and angle among the images, as well as large distribution differences between the actual application scenarios and the datasets used to train these models. Assuming that each model's predicted age is a median probability distribution following a normal distribution, we can consider that adverse factors in practical applications have caused an increase in σ for age estimation probabilities. To ensure that we can obtain as many images with high credibility as possible, we need to use multiple estimation models and set lower ages than intended as common thresholds for multiple models. As long as one model's estimation falls below this threshold, it is considered that the age reflected by the face in the image does not meet our requirements.

Figure 18. Illustration of filtering out Hb/Hf <

If we donate all three networks' estimations results as E_1, E_2, and E_3, and the threshold as T, then the filter is the implementation of the following logic:

$$E_1, E_2, E_3 \in R, E_1 > T \wedge E_2 > T \wedge E_3 > T \quad (3)$$

Finally, we simultaneously used three age estimation models [6, 46, 47], and selected a lower age threshold (>40 years old). Only results like image 4 in Figure 19 could pass the detection. In addition, this facial-based age estimation method has its limitations because features outside of the face, such as hair and clothing, also contain valuable information about a person's age. These pieces of information become particularly important when they cannot be fully revealed or are completely obscured on the face. For example, in image 5 of Figure 7, the elderly woman in the nursing home had significant prediction errors due to her face being covered by mosaic but her gray hair could have been used as part of the evaluation criteria. At the same time, this method requires that most of a person's

Figure 19. Age Estimation

1: Inaccurate estimation of model 1.
2: Inaccurate estimation of model 1 and 2.
3: Inaccurate estimation of model 1 and 3.
4. Successfully passed all three models.
5. Due to mosaic, no valid estimation.

face must be visible; otherwise, it will not be able to predict accurately or with high precision. This leads to many images where faces are not facing forward being unable to pass through our pipeline and thus resulting in loss of some effective images that we were unable to explore.

C. Cartoon Character Detection

Cartoon character detection, also known as cartoon and real photograph classification, is a critical yet often neglected aspect of image analysis. A considerable number of cartoon characters can be found online, and if not properly filtered, these characters may be misidentified as real humans by neural networks. Although certain traditional methods based on color count and other features can distinguish between cartoons and real photographs, elementary neural networks can significantly outperform these methods in terms of accuracy [48].

For this study, we combined two datasets: the COCO 2017 dataset [49] and the Cartoon Classification dataset [50]. The distribution of the training and test sets for these datasets is presented in Table 3.

	Training	Test
Cartoon Classificatio	119,633	18,110
COCO 2017	107,735	14,689
Ratio (CC/COCO)	1.11	1.23
Total (CC+COCO	227,386	32,799

Table 3

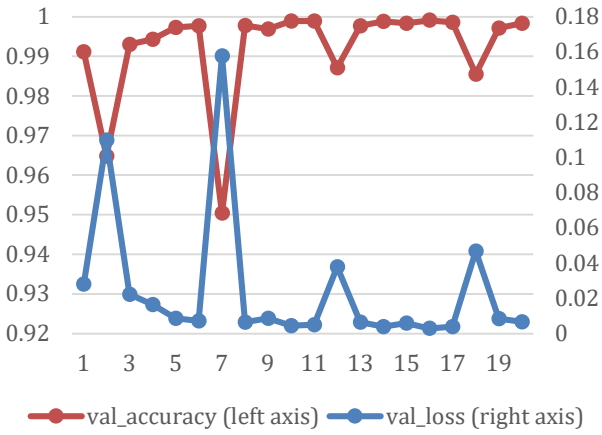


Figure 21. Training Progress of Cartoon Detector

We developed an Xception-based model [51] and trained it using the combined dataset. The loss function during the training process is depicted in Figure 21. The optimal model was achieved at epoch 16, with an accuracy of 0.9991, and thus, we utilized the model from epoch 16 for our pipeline.

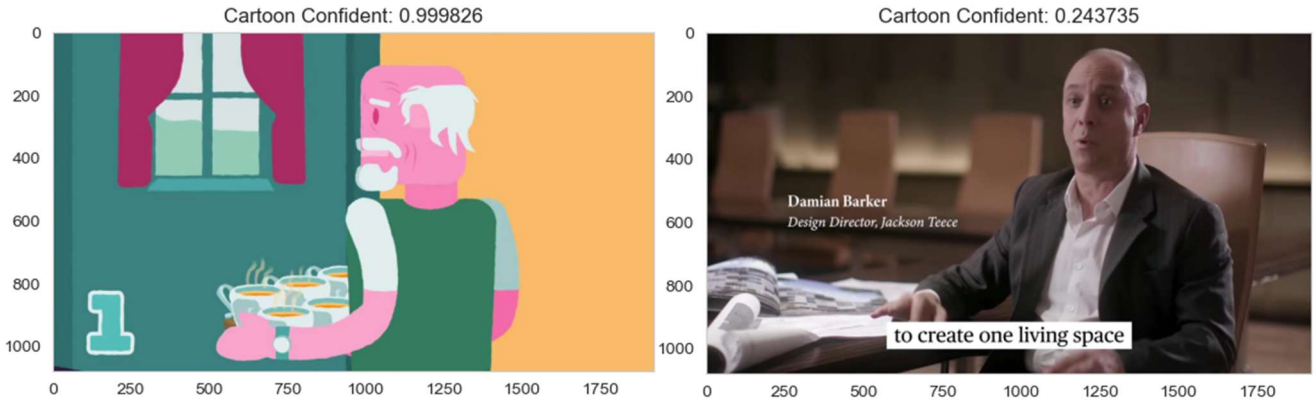


Figure 20

The trained model is able to recognize whether an image is a cartoon or not Figure 20, which helps us filter out a large number of invalid cartoon characters. However, the downside is that after testing, we found that this model does not have the ability to recognize PPTs, even though both PPTs and cartoons are non-realistic photos. As virtual meetings and online courses are prevalent now, many images consist of PPTs as the main subject and a small part of the speaker's upper body video. If we can add PPT recognition as part of our training for the model, we will be able to more efficiently filter out videos and images that cannot be utilized.

D. Human Segmentation & Pose Estimation

When this pipeline is able to determine the age of the character and exclude cartoon characters, it needs to be able to accurately judge: 1. Whether the character's body is fully presented in the image (this is crucial for training high-quality StyleGAN), and 2. Whether the character is standing upright (since pedestrians are mostly standing, our generated images should also be of people in an upright position). For the first point, we propose a solution that uses both human segmentation and pose estimation. This solution determines whether pose estimation results are completely contained within human segmentation Figure 22. For the second point, we use thigh-to-calf height ratio from pose estimation for judgment Figure 23.

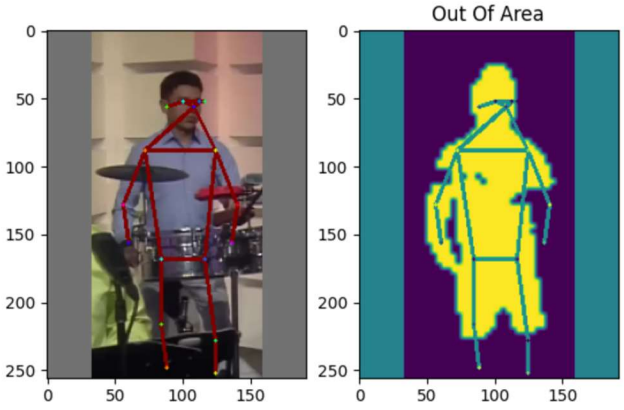


Figure 22. Determine if a person is obscured or out of frame by human pose and human

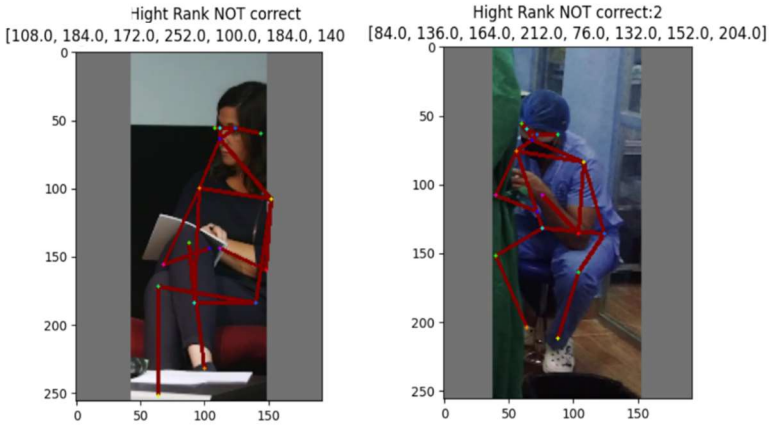


Figure 23. Determine sitting positions by human pose keypoints

For human segmentation, we employ YOLOv8 deep learning model developed by Ultralytics which builds upon previous YOLO's success of efficiencies (crucial for our heavy pipeline) with improved performance and flexibility [23]. As shown in Figure 24, this model can distinguish different individuals in complex environments despite its not-so-accurate boundary delineation during segmentation and suboptimal performance under overly complex conditions due to YOLO's non-maximum suppression mechanism, but it suffices for our requirements when used together with pose estimation.

In the field of pose estimation, the well-known open-source model, CMU's OpenPose [52], has gained significant attention. However, due to this experiment requires a precise ratio as a basis for judgment, we have opted for the more accurate HRNet model, which demonstrates superior performance on the COCO dataset [29] by achieving a mean Average Precision (mAP) of 78.9, as opposed to OpenPose's mAP of 61.8 [53].

In short, the method we propose does not require training new models for specific characters. Instead, it combines existing mature deep learning models from different fields to create new characters and achieves this through simple calculations based on the model's output. The effectiveness of this method that does not require training new models is reflected in the final dataset, which consists almost entirely of standing full-body visible characters.

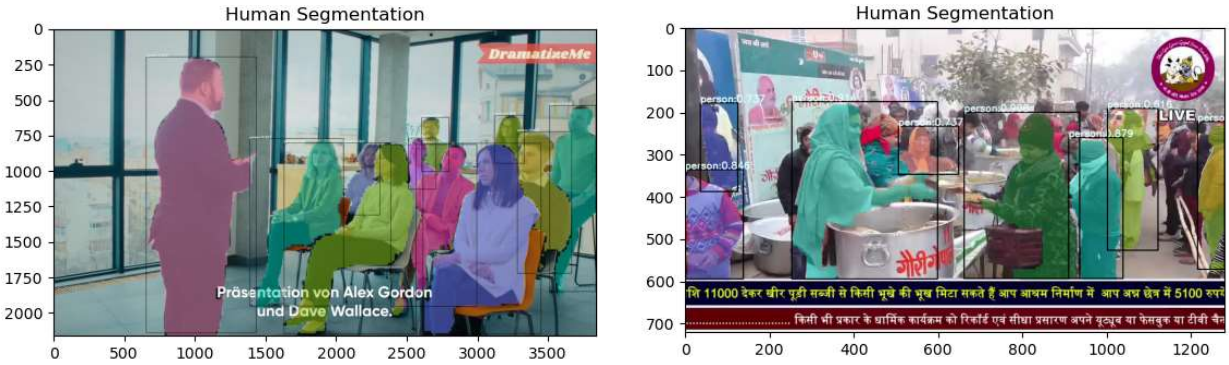


Figure 24

E. Blurry Detection & Image Quality Assessment

To train a high-quality image generation model, the quality of the images is crucial. As the quality of an image has regional properties, we need to perform person-based image quality detection based on cropped images obtained from the previous step. To ensure both efficiency and quality in detection, we use a combination of traditional and DL methods. The person's image will first undergo convolution with a Laplacian operator:

$$L = \frac{1}{6}\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

which can measure the rate of change of the image intensity. Its effect is shown in Figure 25, where whiter areas represent more drastic changes.

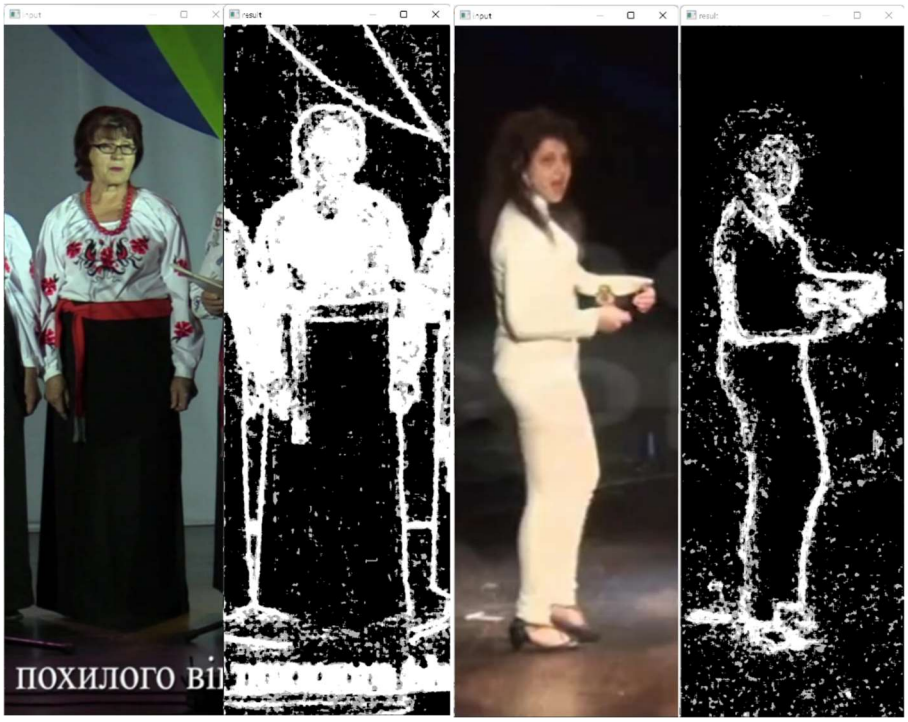


Figure 25

This method is fast but limited as it can only detect one aspect of image quality

evaluation: blurriness. Additionally, its ability to detect blur is restricted because it only focuses on changes in adjacent pixel values, ignoring low levels of adjacent pixel variation even in clear photos (such as black skirts of the upper case in Figure 25).

To overcome these limitations, we employ DL methods for more human-like assessment of image quality.which can measure the rate of change of the image intensity.

For this purpose, we employ a method known as the Multi-dimension Attention Network for No-Reference Image Quality Assessment (MANIQA) [54]. It is a no-reference image quality assessment (NR-IQA) algorithm leverages a synergistic combination of the Vision Transformer (ViT), Transposed Attention Block (TAB), and Scale Swin Transformer Block (SSTB) to extract features and enhance channel interaction among them. From Figure 27, we can see that the score given by MANIQA (higher scores indicating better image quality) is very close to human perception.

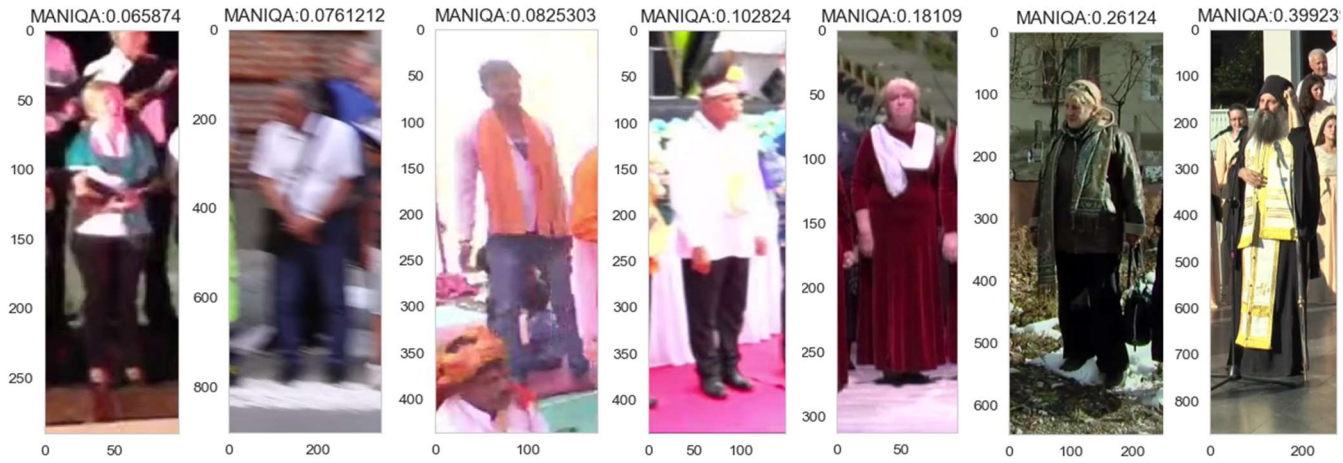


Figure 27. Illustration of MANIQA scores

By applying the MANIQA method to all images, we obtained the score density distribution depicted in Figure 26. This distribution aligns well with the natural distribution present in our dataset with. Consequently, we eliminated all images exhibiting a MANIQA index below then μ – σ (0.3).

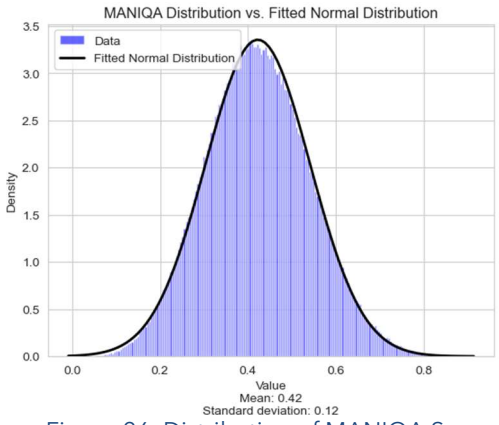


Figure 26. Distribution of MANIQA Score

F. Super Resolution

The quality of the image is crucial for training the image generation network. As most images in our dataset have a height under 1024, they do not meet our high-pixel training requirements. The traditional method of simply enlarging images by establishing a functional relationship between pixels to determine new pixel values will result in a serious decline in image quality [55].To address this issue, we employed the Super Resolution technique called Real-ESRGAN [56] to augment pixel density and resolution. Real-ESRGAN's training methodology incorporates a high-order degradation modeling process to simulate complex real-world degradations while utilizing sinc filters to counteract common ringing and overshoot artifacts.

This method has good restoration ability for images with low blur degree, not only limited to solid color clothes Figure 29, but also commendable in enhancing patterns. However, this method's restoration of extremely blurry images is very poor, such as Figure 28. This is understandable because super resolution is based on the information in the original image for restoration and not self-drawing. Thus, we posit that the Real-ESRGAN approach, grounded in the extensive learning of high-definition image networks, allows us to extrapolate potential high-resolution details from low-resolution images. This method elevates the effective learning content of our dataset compared to mere image scaling.

Figure 29

Figure 28

G. Model Optimization

We combine the above six machine learning models into a pipeline and apply it to tens of millions of web image crawling tasks. As you can imagine, our requirements for model speed are more demanding than real-time application scenarios. Some frameworks optimized for the interference phase, such as ONNX-runtime and TensorRT, are extremely capable and can often be several times faster than frameworks used in training models, such as PyTorch and TensorFlow [57, 58]. Therefore, we transformed all the models that were not originally ONNX and TensorRT.

3.3.2. Google-Image/Twitter-API

Google Image search and Twitter are both rich sources of photographs, making them ideal for data collection. In this study, we utilized web crawlers to gather images from Google Image search results using search terms related to seniors. Similarly, a substantial number of valid images were obtained from Twitter by employing Twitter's public and efficient search API with search terms associated with older individuals.

3.3.3. Youtube

YouTube is an extensive repository of image data, with the platform claimshttps://blog.youtube/press/ more than 500 hours of video uploads every minute. Given that the quantity of images acquired in the previous phase was insufficient for training a Generative Adversarial Network (GAN) that demands a vast amount of high-quality data, we shifted our focus to YouTube. However, this data collection process proved time-consuming due to the inferior quality of images in videos (e.g., motion blur, out-of-focus) compared to still photographs. Moreover, videos predominantly feature one or several individuals speaking, with a considerable portion of the frame occupied by busts, resulting in significantly lower information density than the image data acquired in the previous stage.

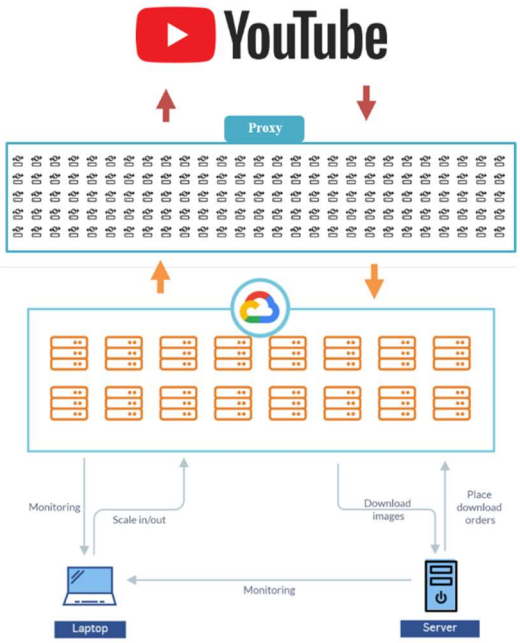
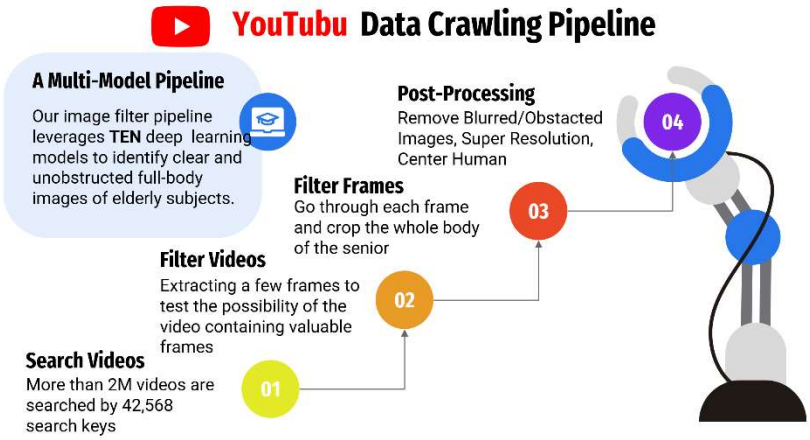


Figure 30. Youtube Data Crawling Pipeline

Gathering data from YouTube involved four steps: video search, video filtering, frame filtering, and post-processing (Figure 30). The following subsections provide a detailed description of each step.\



A. Video Searching

The most frequently appearing search terms in the word cloud Figure 31. I18n: WordCloud of our YouTube Search Keys include: 'asia/asian', 'senior', and 'geriatric/geriatrica/geriatrik/'. This indicates that our search keywords are primarily focused on elderly people from Asia.

To expand the scope of our search while maintaining its relevance to the elderly population, we employed the following methodology to generate over 42,000 search terms:

- Manually crafted a set of phrases pertaining to the elderly population in English (e.g., elderly population, nursing home, my grandparents, etc.)
- Leveraged ChatGPT to generate analogous terms in English
- Translated the aforementioned terms into 113 languages using Google Translate API.

Figure 31. I18n: WordCloud of our YouTube Search Keys

Considering YouTube delivers region-specific search results, we assumed that conducting searches in the most prevalent regions for each language would yield the richest outcomes. Consequently, we utilized the I18nRegions and I18nLanguages lists provided by the YouTube API, in conjunction with the ISO 639-1 [59], to execute a comprehensive global multilingual search. Ultimately, we obtained over 2 million YouTube video IDs.

B. Video Filtering

Due to the abundance of YouTube videos, such as talk shows, PowerPoint-based courses, animations, etc., that are usually totally useless for our purposes, we implemented a filtering process focusing on the video level. The procedure entailed the following steps:

- Extract keyframes from three 30-second clips at 20%, 50%, and 80% of the video length
- Verify the presence of a human face that satisfies formula 1.
- Determine if the content is animated
- Ascertain if any detected faces belong to individuals aged 40 or older

If at least one of the extracted frames passed the aforementioned filters, we deemed the video likely to contain frames meeting our criteria and thus approved it for further processing.

C. Frame Filtering

Upon a video's successful completion of the video filtering stage, we downloaded the entire content and extracted its keyframes. Recognizing the high temporal correlation among keyframes, we selected one every senven frames for input into the Basic Pipeline. If a frame contained a full-body image of an elderly individual, we then input the first and last three keyframes surrounding that frame into the pipeline for additional filtering.

Following this procedure, we obtained roughly 380,000 coarsely filtered full-body images of elderly individuals from YouTube, which were then combined with images extracted from Google Image/Twitter API and Laion5B for final post-processing.

3.4. Constructing the Eloquent ElderWear (EEW) Benchmark

To rigorously evaluate the proposed clothing recognition solutions, a test set consisting of full-body images of elderly individuals paired with their corresponding natural language descriptions is required. As no such dataset currently exists, we constructed a new dataset, called Eloquent ElderWear (EEW), tailored to this specific purpose.

In practical application scenarios, we believe that two car-mounted cameras in the bustling city of Hong Kong would capture a non-repeating number of pedestrians on the order of three per minute. To simulate this, we designed an experiment to perform image retrieval based on natural language descriptions of clothing within a small dataset comprising 1000 images.

Thus, we first randomly selected 1000 images from the WISE dataset. From these images, we removed their backgrounds and manually annotated 106 garment image captions.

3.4.1. Background Subtraction

Complex backgrounds, particularly when other individuals are present in the images, can distract the CLIP model's attention, potentially leading to lower accuracy. Our experimental results support this observation (Table 7). Consequently, we first performed background subtraction on the human images.

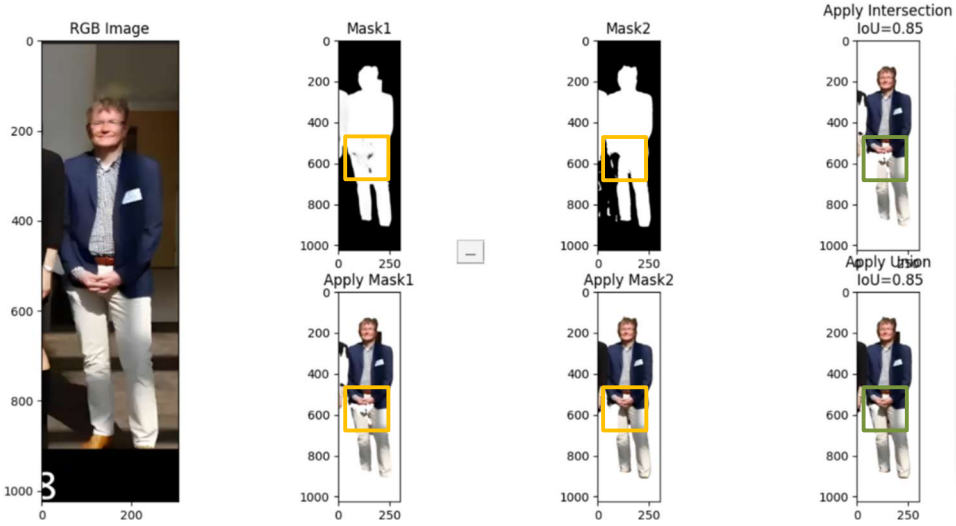


Figure 33

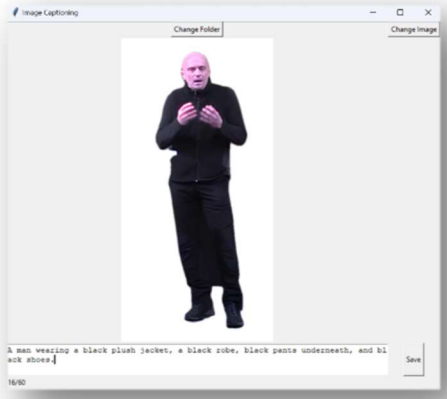


Figure 32. Image Caption Interface

While testing human segmentation models, we found that they often produce either too many or too few segmentations for unclear or structurally complex images. Excessive segmentations can introduce interference, whereas insufficient segmentations may exclude valuable information. To mitigate these issues, we employed two distinct human segmentation algorithms [60, 61]. Algorithm [61] was previously used to construct the SHHQ dataset and demonstrated satisfactory results. Algorithm [60] is a popular model widely adopted by the GitHub community. We combined the results from both algorithms by taking their union, as depicted in Figure 33. This approach effectively avoids the exclusion of essential body parts while reducing interference.

3.4.2. Manual Labeling

We first developed a simple image captioning GUI, as shown in Figure 32, which consists of an image box and a text box to input the corresponding garment description.

Considering the human bias when describing garments, we try to gather descriptions from different people. Finally, we gathered 106 natural language descriptions corresponding to 101 images (with 5 images having two corresponding image captions) from 10 students. The contribution of each student is listed in Table 4.

Student ID	Number of Descriptions	Number of Described Images
1	22	22
2	15	15
3	12	12
4	13	13
5	10	10
6	8	8
7	7	7
8	6	6
9	6	6
10	7	7
Total	106	101 (unique)

Table 4

3.4.3. Top-k Accuracy

As clothing identification can be relatively ambiguous, we employed the Top-k approach, returning the k most similar results for professional reference. Small k values may significantly decrease matching accuracy, whereas large k values can overwhelm professionals with numerous potential matches, reducing the system's overall effectiveness. Therefore, we selected k values of 3, 10, and 25 for detection.

3.5. Natural Language Garment Retrieval by CLIP

In order to harness the full potential of the CLIP model for the natural language garment retrieval task, we propose a fine-tuning process that optimizes the multi-modal capabilities of CLIP. This approach facilitates more accurate and efficient retrieval of relevant images based on the similarity between candidate images and the search language description.

3.5.1. Baseline Model Selection

To ensure efficiency and optimal performance, we selected OpenAI's smallest and fastest model, ViT-B/32, as our baseline model. This model is well-suited for this task, as it is computationally efficient, enabling rapid processing and retrieval of relevant images among many alternatives.

3.5.2. Fine-Tuning Process

We fine-tuned the ViT-B/32 model using the FashionGen dataset [40], which consists of 260,480 garment-description pairs in the training set and 32,528 garment-description pairs in the test set. The dataset provides a rich and diverse collection of images and their corresponding textual descriptions, making it ideal for training the model to recognize and retrieve garments based on natural language queries.

We employed a distributed training strategy using eight TPUs. Each mini-batch consisted of 2048 photo-text pairs and the fine-tuning process was conducted over ten epochs (160 steps). Random horizontal flipping is employed to prevent overfitting and improve generalization. This relatively small number of training epochs was designed to maximize the model's ability to learn from the dataset while maintaining efficiency and minimizing training time.

3.5.3 Evaluation and Performance Metrics

Throughout the fine-tuning process, we monitored the evaluation loss to gauge the model's performance. As shown in Figure 34, the evaluation loss decreased consistently over the training period, reaching a final value of 0.3187. This indicates that the model has successfully learned to identify and retrieve relevant images based on the textual descriptions provided.

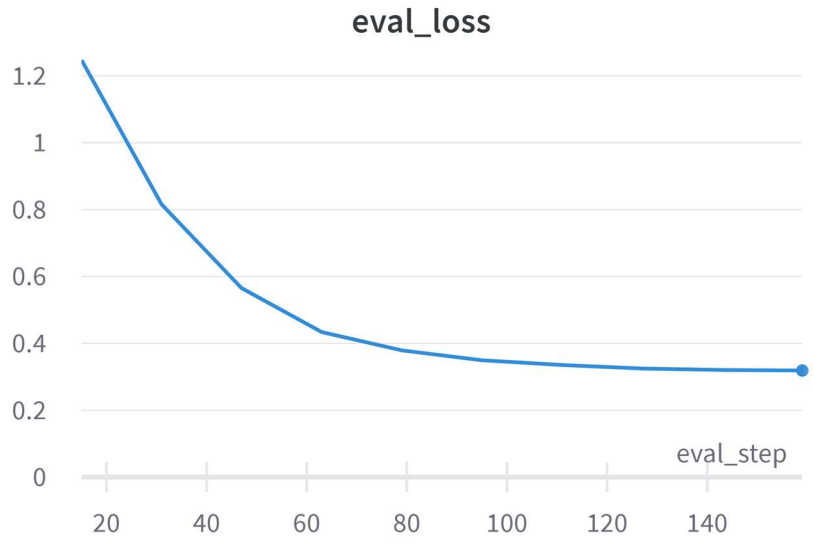


Figure 34

3.6. Textual-Driven Synthesis of Full-Body Garment Imagery: Leveraging StyleGAN2 and Fine-tuned CLIP ModelsStyleGAN2

This section explores the synthesis of full-body garment imagery using a combination of StyleGAN2 and fine-tuned CLIP models. We detail our pre-training and fine-tuning processes and our attempts to address color-matching issues and integrate textual guidance into the generation of accurate full-body garment images.

3.6.1. Pre-training StyleGAN2 on the SHHQ Dataset

We pre-trained a StyleGAN2 model using the JAX/FLAX framework on the SHHQ dataset [39]. The model was trained on 8 TPUs with a mini-batch configuration set to 16. The pre-training was conducted for 230K steps, equivalent to 92 epochs. The Fréchet Inception Distance (FID) scores during the training process are illustrated in Figure 35. The lowest FID score achieved by the model was 14.515, which occurred at step 230K.

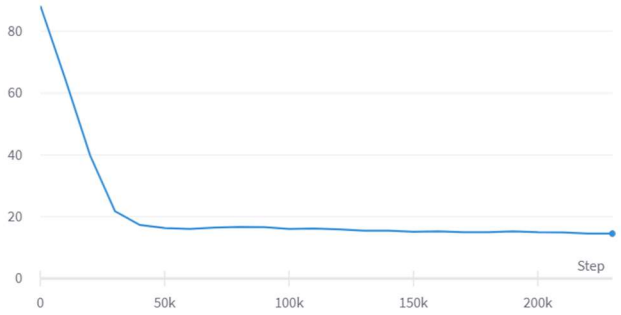


Figure 35

3.6.2. Integrating CLIP for Textual Guidance

With the pre-trained StyleGAN2 models available, we integrated fine-tuned CLIP models to apply textual guided image generation techniques for preliminary performance testing. During the testing phase, we discovered that the CLIP model (even after fine-tuning) encountered significant issues with incorrect color and position matching. For instance, when prompted with "A woman wears a short blue T-shirt with short red pants," the generated image displayed a red top and blue-black pants instead, as shown in Figure 36, highlighting the model's inability to accurately localize the colors.

Figure 36

We hypothesize that this problem may originate from the structure of the CLIP model itself or, more likely, from the training dataset, as the datasets typically used for training image captions focus on object and relationship

descriptions but may neglect color descriptions. The FashionGen dataset [40], which we used for fine-tuning, also lacks sufficient color descriptions.

3.6.3. Color Extraction and Modification using OpenAI's Chat API

To address the aforementioned color matching issue, we propose a feasible method that involves extracting clothing colors from natural language descriptions using OpenAI's chat API. We then modify the upper and lower garment colors using basic color modification techniques. Although this method exhibits significant limitations, such as the generation of unrealistic colors (as shown in Figure 37), it can distinguish between upper and lower garments. However, it struggles to manage complex clothing designs.

Figure 37

3.6.4. Fine-tuning StyleGAN2 on the WISE Dataset

Subsequently, we fine-tuned the pre-trained StyleGAN2 model on an elderly full-body image dataset called WISE. The fine-tuning process involved 370K steps, equivalent to 49 epochs, with a mini-batch size of 16. The FID scores during the fine-tuning process are depicted in Figure 38. Although the final FID score of 16.979 was not the lowest score throughout the process, a visual inspection revealed that the images generated by this model were optimal compared to others. Therefore, we selected the model obtained at 370K steps for further experimentation.

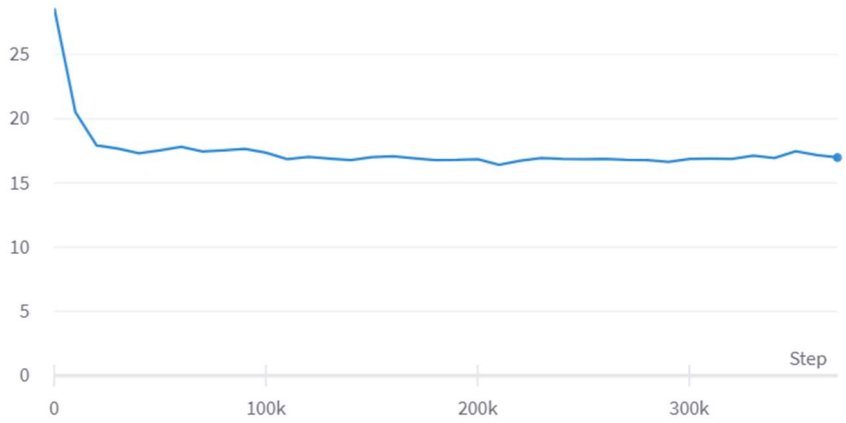


Figure 38

4. Result Analysis

In this section, we present a comprehensive analysis of the results obtained from our experiments. The structure of this section is organized as follows: Section 4.1 discusses the fine-tuning results of face detection, while Section 4.2 presents the fine-tuning results of face recognition. In Section 4.3, we introduce the WISE dataset and provide a detailed quantitative analysis of its characteristics. Section 4.4 describes the Eloquent ElderWear (EEW) benchmark, which offers insights into the complexity and richness of information in image captions. Section 4.5 focuses on the enhanced garment retrieval by fine-tuned CLIP, comparing the performance of the original CLIP model with the fine-tuned model. Lastly, Section 4.6 delves into the application of StyleGAN2 for elderly garment generation, assessing full-body image synthesis performance, disentanglement, and textual guided garment generation with StyleCLIP.

Our analysis covers a wide range of aspects, from model performance improvements through fine-tuning to the construction and evaluation of novel datasets tailored for the elderly population. We also investigate the potential of state-of-the-art models, such as CLIP and StyleGAN2, to address specific tasks within the context of locating missing persons and generating realistic garment images. By providing a detailed and structured account of our findings, we aim to offer readers a clear understanding of our work and its implications for future research and development in AI applications for the elderly population.

4.1. Fine-tuning Result of Face Detection

After 10,000 epochs of fine-tuning, the model's mean average precision (mAP) improved for all three splits compared to the pre-trained model. The results are listed in Table 5.

Split	Pre-trained	Fine-tuned
Easy	0.869026079	0.891406863
Medium	0.849931985	0.878591500
Hard	0.737039513	0.738500043

Table 5. mAP Comparision of Pre-trained and Fine-tuned SCRFD on WIDER FACE

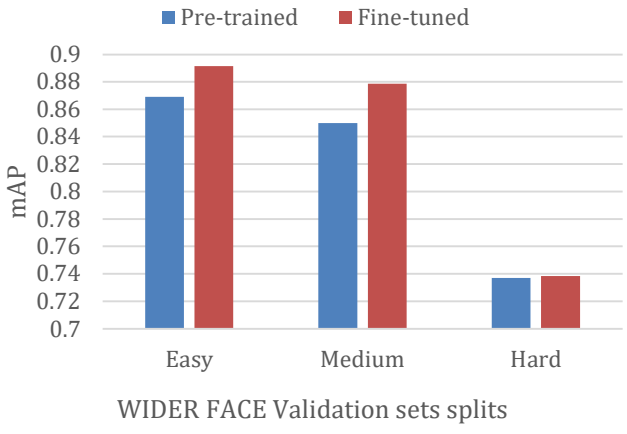


Figure 39. Transfer Learning Comparison of SCRFD on WIDER FACE

Split	Pre-trained	Fine-tuned
Easy	0.869026079	0.891406863
Medium	0.849931985	0.878591500
Hard	0.737039513	0.738500043

Table 5

The fine-tuning process of the SCRFD model on the augmented WIDER FACE dataset yielded improvements in the model's performance, particularly for the easy and medium splits. Although the improvement in the hard split was not as pronounced, it still indicates the model's adaptability to challenging scenarios involving small faces.

4.2. Fine-tuning Result of Face Recognition

Split	Pre-trained	Fine-tuned
Mixed-Original Pairs	0.98087	0.98873
Masked-Original Pairs	0.98362	0.9911
Blurred-Original Pairs	0.97402	0.9901
Darkened-Original Pairs	0.99277	0.99023

Table 6. mAP Comparision of Pre-trained and Fine-tuned ArcFace on Augmentated LFW

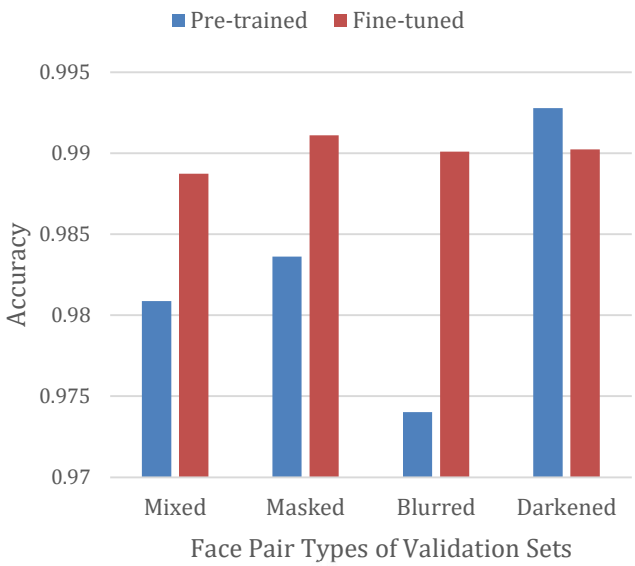


Figure 40. Transfer learning result comparisons upon different validation sets

The comparison of the fine-tuned face recognition model's performance against the pre-trained model is list in Table 4 and illustrated in Figure 40. The results demonstrate that transfer learning yields substantial improvements overall. In the evaluation of mixed-original pairs, which maintains the same proportion as the training set, the fine-tuned model exhibits a 0.8% absolute enhancement in accuracy relative to the pre-trained model. Notably, the fine-tuned model delivers considerable advancements in the recognition of both masked and blurred faces.

Conversely, the fine-tuned model experiences a minor decline in performance concerning darkened faces as compared to the pre-trained model. This observation may be attributed to the fact that the other two cases—masked and blurred faces—present greater opportunities for optimization. Consequently, the model might slightly compromise the accuracy of recognizing darkened faces to achieve marked improvements in the other two scenarios during the training process.

4.3. WISE Dataset

After two months of rigorous data collection and meticulous curation, we have successfully constructed the WISE dataset, which comprises 12,030 full-body images predominantly featuring elderly individuals. In this section, we present a comprehensive quantitative analysis of the WISE dataset, with a focus on image dimensions, age distribution, and comparison with the SHHQ dataset.

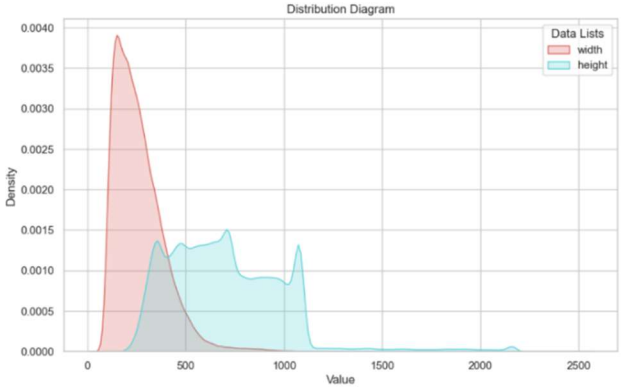


Figure 41. Roughly Filtered Images' Size Distribution

Initially, we examined the dimensions of the raw, unprocessed images in the dataset prior to applying super-resolution techniques. The average width of the images was found to be 262.6 pixels, while the average height measured 701.7 pixels. The distribution of both width and height dimensions is illustrated in Figure 41, which offers a detailed visual representation of the dataset's dimensional characteristics. Following the application of super-resolution algorithms, all images were standardized to a height of 1024 pixels, ensuring uniformity and improved clarity across the dataset.

Considering that the WISE dataset specifically targets the elderly population, evaluating the age distribution is a crucial aspect of our analysis. To construct the dataset, we employed three age estimation deep learning models [6, 46, 47] during the filtering process. However, in this result analysis

phase, we opted for a third-party age estimation API [62] to ensure fairness and accuracy when assessing the age distribution. It is important to note that age estimation models are subject to limitations and biases inherent in their training sets, which may impact the accuracy of their predictions. Consequently, we conducted a relative comparison between the age distributions of our WISE dataset and the SHHQ dataset to gauge the effectiveness of our dataset in capturing the elderly

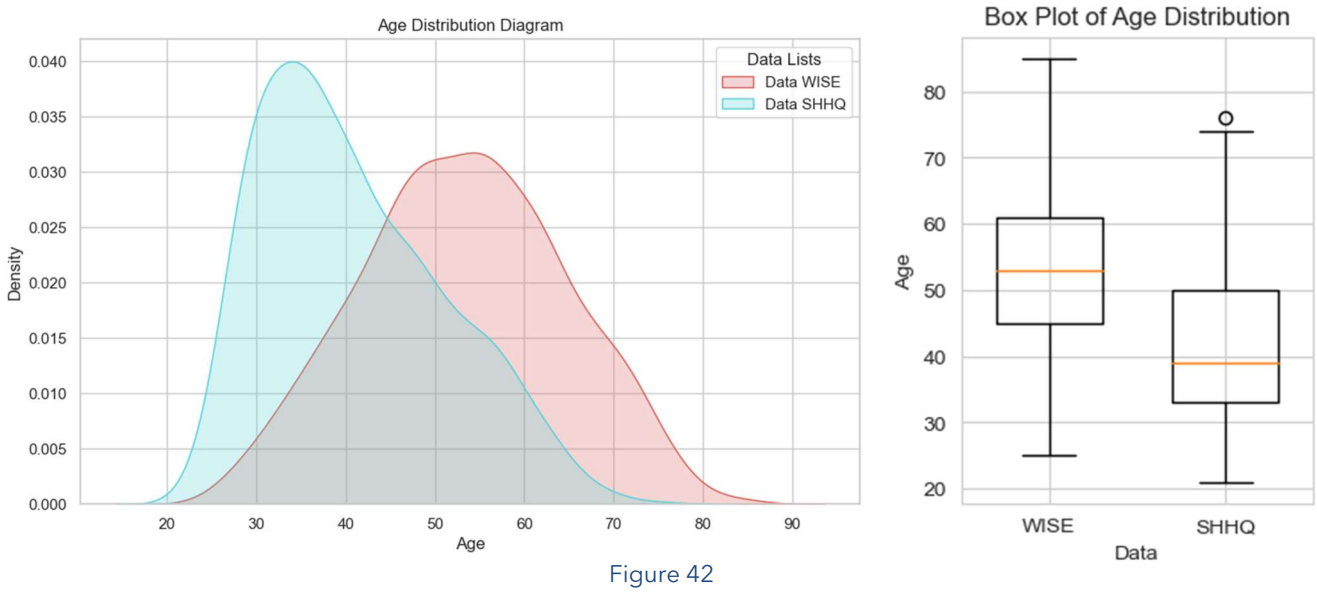


Figure 42

population.

Figure 42 depicts the age distributions of both datasets, demonstrating that the WISE dataset exhibits a significantly higher average age (51.2 years) compared to the SHHQ dataset [39]. This finding reaffirms the success of our data collection and filtering methodologies in effectively targeting the elderly demographic.

A random selection of representative samples from the WISE dataset is presented in Figure 43, with additional samples provided in *Appendix A: WISE Datset Samples*. In conclusion, the WISE dataset represents a highly reliable and effective collection of full-body images primarily featuring

Figure 43. Random Selection of Representative Samples From the WISE dataset

elderly individuals, making it a valuable resource for research and development in AI applications tailored towards the elderly population.

4.4. Eloquent ElderWear (EEW) Benchmark

The EEW dataset consists of 101 clothing images, each accompanied by at least one image caption. In total, the annotators provided 106 captions, with some images receiving multiple annotations to capture the diverse perspectives on the garments. The captions were analyzed to determine their complexity and richness of information.

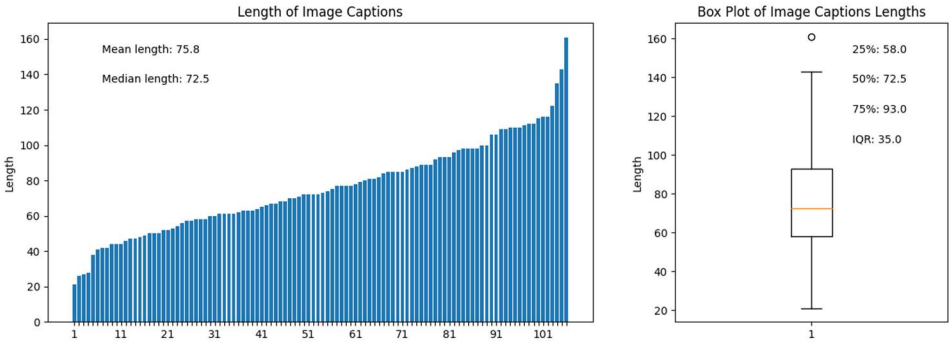


Figure 44

The word lengths of the 106 image captions are illustrated in Figure 44, highlighting the distribution of caption lengths. The average length of a caption is 75.8 words, demonstrating the level of detail provided by the annotators. The interquartile range (IQR) of caption lengths is 35, indicating a concentrated di stribution and relatively convs istent level of descrip tion across the dataset.

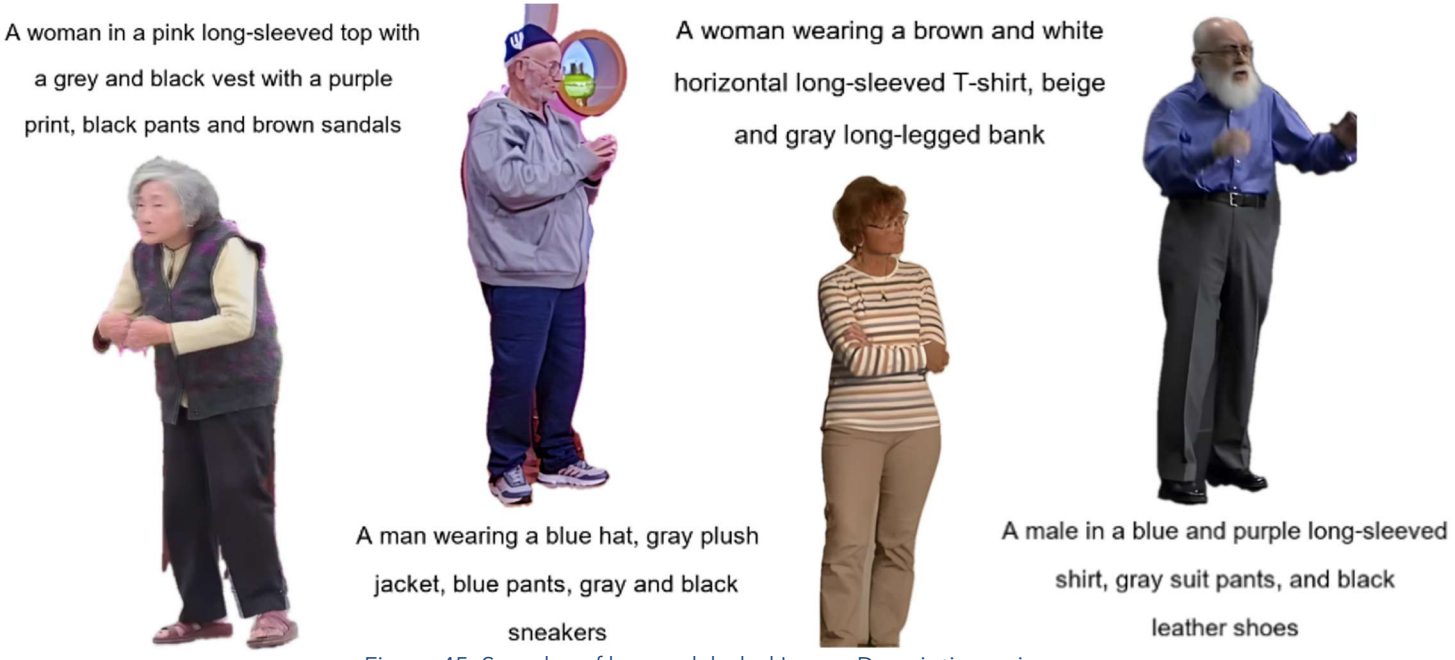


Figure 45. Samples of human-labeled Image-Description pairs

Figure 45 illustrates some samples from EEW. We plan to expand the EEW dataset in future work by incorporating additional images and annotations. This would further enrich the dataset and provide an even more comprehensive resource for research and development in AI applications tailored towards the elderly population.

4.5. Enhanced Garment Retrieval by Fine-tuned CLIP

In this section, we present a comprehensive analysis of the results obtained by comparing the performance of the original CLIP model to that of a fine-tuned CLIP model on the garment retrieval task. Our goal is to highlight the potential improvements in retrieval accuracy and robustness that can be achieved through fine-tuning.

4.5.1. Experimental Setting and Results

Our study sought to examine the performance of the pre-trained CLIP model on original images, pre-trained CLIP model with background-removed images (Baseline), and a fine-tuned CLIP model with background-removed images. The top-k precision for k=3, k=10, and k=25 is shown in Table 7.

T-ranked precision (among 1K images)	Pre-trained CLIP with original images	Pre-trained CLIP with background-removed images (Baseline)	Fine-tuned CLIP with background-removed images
T=3	0.1509	0.1792	**0.2358**
T=10	0.2642	0.2830	**0.3491**
T=25	0.4057	0.4811	**0.5094**

Table 7

4.5.2. Result Comparison

Upon comparing the top-k precision between the original CLIP model with original images, the baseline, and the fine-tuned CLIP model with background-removed image, it is evident that the fine-tuned CLIP model with background-removed image outperforms the other two across all k values. This demonstrates the effectiveness of fine-tuning the model for the specific task of garment retrieval in the context of locating missing persons.

We showcase a selection of exemplary results from our meticulously fine-tuned experiments compared to the baseline, where k equals 3, from Figure 46 to Figure 47. The uppermost text in each figure corresponds to the textual description employed for image matching. In the lower portion of each figure, the left and right parts are from the baseline and fine-runed model, respectively. The green and

red boxes signify accurate and inaccurate matches, respectively. Accompanying each box is the corresponding image, annotated with the CLIP model Score (expressed as a percentage), quantifying the degree of similarity between the image and the provided textual description. Despite the improved performance of the fine-tuned CLIP model, there are still several challenges that the model encounters, as delineated below:

• The model exhibits a limited capacity to accurately associate colors with distinct clothing items, often leading to confusion between colors and garments.

• The proficiency in discerning sleeve lengths is suboptimal, resulting in the underutilization of valuable information pertaining to length.

• The intrinsic limitations of linguistic descriptions render the differentiation of similar clothing items particularly arduous, especially in the case of black and white garments.

Figure 46

Figure 47

4.6. StyleGAN2 for Elderly Garment Generation

In this section, we present the results of our StyleGAN2 model for generating elderly garment imagery. We discuss the performance of our model in terms of image synthesis, disentanglement assessment, and the challenges faced in achieving accurate textual guided garment generation.

4.6.1. Full-body Image Synthesis Performance

After an intensive week-long training process, our StyleGAN2 model has displayed considerable proficiency in generating full-body images of elderly individuals, albeit with some imperfections. The final Fréchet Inception Distance (FID) achieved is 16.979. This FID value is relatively impressive given the intricate nature of full-body images, the varying backgrounds, diverse subject orientations and poses, and the substantial differences among them. Additionally, the training was conducted at a high resolution of 512×1024 pixels.

Compared to mainstream experiments employing 8×A100/V100 GPUs for over a week [4, 39, 63], our results utilizing Google Cloud Platform's (GCP) v3-8 TPU for a week of training are noteworthy. A selection of sample images is provided in Figure 48, with supplementary results in *Appendix D: Synthesized Samples of StyleGAN2 Fine-tuned on WISE*.

Figure 48

4.6.2. Disentanglement Assessment with GanSpace

Disentanglement is a crucial evaluation criterion for StyleGAN [34], as low entanglement levels facilitate model adjustment and expected image generation. GanSpace employs Principal Component Analysis (PCA) decomposition of the w+ space to provide a more intuitive understanding of disentanglement while directly applying PCA vector outcomes to modify images. Conducting GanSpace's PCA analysis on our final StyleGAN2 model, we acquired the results depicted in Figure 49, where both image sets are transitioning in the same PCA direction.

Figure 49

Our StyleGAN2 model exhibits a high degree of entanglement, as demonstrated in Figure 49, which displays multiple transformations along one direction, such as sitting to standing posture, body

rotation, shorts-to-pants conversion, and alterations in clothing color and style. The high entanglement in our model can likely be attributed to the absence of background subtraction in our training set. Contrary to typical StyleGAN training sets, a single image in our dataset may encompass numerous target objects (in our case, humans) within image backgrounds and other intricate elements. Consequently, the generator learns more elements, resulting in elevated entanglement.

4.6.3. Textual Guided Germent Generation with StyleCLIP

StyleCLIP's results are not entirely satisfactory, as the back-propagation guided by CLIP tends to render the character images peculiar before aligning with the desired outcomes. As shown in Figure 50 and Figure 51, we input clothing descriptions adjacent to the images, and after 100 iterations with a learning rate of 8e-3, we achieved the following results. While the model appears to be moving towards the language description, it remains insufficient. For instance, in Figure 50, the floral-printed dress exhibits a pattern resembling floral print but lacks an actual dress.

The elderly lady donned a flowy, floral-printed dress with a delicate lace collar and cuffs.

Figure 50

The elderly woman was dressed in a cozy, cable-knit sweater and a pair of soft, comfortable slcks.

Figure 51

We posit that this outcome may stem from a combination of the following factors:

- Insufficient training: Given that numerous large-scale StyleGAN models undergo approximately a week of training with 8×A100 GPUs, our training volume is inadequate. To enhance the results, we should continue training the model.

- Lack of background removal: The StyleGAN-Human [39] model uses the background-removed SHHQ dataset and attains satisfactory results. Background removal can diminish information interference, latent space complexity, and entanglement. In our experiments, the colors of background and upper outer garments in the two results display a degree of correlation. Non-white backgrounds also introduce interference for the CLIP model during inference.
- Dataset distribution issues: Our dataset, compiled from a diverse array of online resources, contains images with a broad range of poses, orientations, and occlusions. This variability prohibits StyleGAN from learning the distribution efficiently.

5. Limitations and Future Improvements

In this section, we discuss the limitations of our current study and propose potential future improvements to enhance the effectiveness of our research. By addressing these limitations and incorporating the suggested improvements, we aim to refine our methodology and provide an even more robust tool for locating missing persons.

5.1 Limitations

Despite the promising results and significant advancements demonstrated in our study, there are a few limitations that warrant further consideration:

- Lack of practical application scenario testing: While our research focuses on assisting the Hong Kong Police in finding missing persons, particularly elderly individuals, our study has not been extensively tested in real-world scenarios. Thus, the effectiveness of our models and methodologies in practice remains unclear.
- Dataset demographic bias: The WISE dataset is gathered internationally, which may limit the applicability of the models to deployed locally. This issue may impact the accuracy and robustness of our models when applied to Hong Kong.

5.2 Future Improvements

To further enhance the utility and effectiveness of our research, we propose several future improvements:

- Conditional Generation: By incorporating conditional generation techniques, we can enable more precise control over the synthesis of full-body images, thereby generating images that better match the desired clothing and personal characteristics of missing persons.\
- Transfer from 2D Images to 3D Humans [64]: By adopting the method proposed in [64], we can create more realistic and accurate 3D models of individuals and then generate the image characters of the missing persons from a variety of prospects, which can improve the performance of our system in real-world applications.
- Use Mean Average Precision (MAP) to reflect rank: In our current evaluation, we use top-k accuracy to assess our models' performance. However, adopting MAP as an evaluation metric would allow us to better understand the ranking of results, which is crucial when searching for missing persons.
- Practical application testing: We recommend conducting extensive testing in real-world scenarios to evaluate the effectiveness of our methodologies and models in practice. This will enable us to identify potential issues and further refine our models to better suit the needs of the Hong Kong Police in finding missing persons.

By addressing these limitations and incorporating the proposed future improvements, we believe that our research can be further refined to provide an even more effective and robust tool for the Hong Kong Police and other organizations working to locate missing persons, particularly elderly individuals.

References

[1] T. Mare *et al.*, "A realistic approach to generate masked faces applied on two novel masked face recognition data sets," *arXiv preprint arXiv:2109.01745,* 2021.

[2] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021: PMLR, pp. 8748-8763.

[3] M. Kryvaltsevich. "Vector Set Of Human, People Group Logos. Family, Business Teamwork, Friendship Concept Stock Vector - Illustration of body, modern: 138818924." @Dreamstime. https://www.dreamstime.com/vector-set-human-people-group-logos-family-business-teamwork-friendship-concept-d-origami-multicolor-men-logo-collection-image138818924 (accessed 18, Dec, 2022).

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," 2019, doi: 10.48550/arxiv.1912.04958.

[5] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," *arXiv preprint arXiv:2105.04714,* 2021.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition (v4)," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690-4699.

[7] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525-5533.

[8] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "Tinaface: Strong but simple baseline for face detection," *arXiv preprint arXiv:2011.13183,* 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[10] X. An *et al.*, "Partial fc: Training 10 million identities on a single machine," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1445-1449.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition (v1)," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690-4699.

[12] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[13] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021: IEEE, pp. 9415-9422.

[14] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5927-5935.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026-1034.

[16] Y. Feng, S. Yu, H. Peng, Y.-R. Li, and J. Zhang, "Detect Faces Efficiently: A Survey and Evaluations," *IEEE Transactions on Biometrics, Behavior, and Identity Science,* vol. 4, no. 1, pp. 1-18, 2021.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.

[21] "InsightFace," 2023. [Online]. Available: https://insightface.ai/.

[22] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Computer Vision, IEEE International Conference on*, 2003, vol. 3: IEEE Computer Society, pp. 1470-1470.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, pp. 91-110, 2004.

[24] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision,* vol. 42, pp. 145-175, 2001.

[25] M. J. Swain and D. H. Ballard, "Color indexing," *International journal of computer vision,* vol. 7, no. 1, pp. 11-32, 1991.

[26] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612,* 2017.

[27] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201-216.

[28] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1386-1393.

[29] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014: Springer, pp. 740-755.

[30] C. Schuhmann *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402,* 2022.

[31] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[32] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[33] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," 2017, doi: 10.48550/arxiv.1703.06868.

[34] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANSpace: Discovering Interpretable GAN Controls," 2020, doi: 10.48550/arxiv.2004.02546.

[35] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery," 2021, doi: 10.48550/arxiv.2103.17249.

[36] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114,* 2013.

[37] Y. Jiang, S. Yang, H. Qju, W. Wu, C. C. Loy, and Z. Liu, "Text2Human: Text-Driven Controllable Human Image Generation," *ACM transactions on graphics,* vol. 41, no. 4, pp. 1-11, 2022, doi: 10.1145/3528223.3530104.

[38] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," 2017, doi: 10.48550/arxiv.1710.10196.

[39] S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, "StyleGAN-Human: A Data-Centric Odyssey of Human Generation," vol. 13676, (Lecture Notes in Computer Science. Switzerland: Springer, 2022, pp. 1-19.

[40] N. Rostamzadeh *et al.*, "Fashion-Gen: The Generative Fashion Dataset and Challenge," 2018, doi: 10.48550/arxiv.1806.08317.

[41] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," *arXiv preprint arXiv:2008.11104,* 2020.

[42] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203-5212.

[43] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep,* vol. 14, no. 003, 2014.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.

[45] G. Mather, "Head–body ratio as a visual cue for stature in people and sculptural art," *Perception,* vol. 39, no. 10, pp. 1390-1395, 2010.

[46] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognition Letters,* vol. 140, pp. 325-331, 2020.

[47] nateraw. "nateraw/vit-age-classifier · Hugging Face." @huggingface. https://huggingface.co/nateraw/vit-age-classifier (accessed.

[48] C. Griffith. "Cartoon or Photo? – Image detection with Python." https://codecalamity.com/cartoon-or-photo-image-detection-with-python/ (accessed 2023).

[49] "T. Lin et al., “Microsoft COCO: Common Objects in Context,” ArXiv. /abs/1405.0312.."

[50] Volkandl. "Cartoon Classification." https://www.kaggle.com/datasets/volkandl/cartoon-classification (accessed.

[51] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2016, doi: 10.48550/arxiv.1610.02357.

[52] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE transactions on pattern analysis and machine intelligence,* vol. 43, no. 1, pp. 172-186, 2021, doi: 10.1109/TPAMI.2019.2929257.

[53] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693-5703.

[54] S. Yang *et al.*, "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment," 2022, doi: 10.48550/arxiv.2204.08958.

[55] A. C. Bovik, *Handbook of image and video processing*. Academic press, 2010.

[56] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data," 2021, doi: 10.48550/arxiv.2107.10833.

[57] "S. K. (2021, August 16). An empirical approach to speedup your BERT inference with ONNX TorchScript [Online]. Available: https://towardsdatascience.com/an-empirical-

approach-to-speedup-your-bert-inference-with-onnx-torchscript-91da336b3a41." (accessed.

[58] "M. Kostylev. (2022, May 2). Optimizing and deploying transformer INT8 inference with ONNX Runtime TensorRT on NVIDIA GPUs [Online]. Available: https://cloudblogs.microsoft.com/opensource/2022/05/02/optimizing-and-deploying-transformer-int8-inference-with-onnx-runtime-tensorrt-on-nvidia-gpus/." (accessed.

[59] ""ISO 639-1 - Wikipedia," Wikipedia, 2023. [Online]. Available: https://en.wikipedia.org/wiki/ISO_639-1. [Accessed: 03-Apr-2023].".

[60] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern recognition,* vol. 106, p. 107404, 2020.

[61] L. Chu *et al.*, "Pp-humanseg: Connectivity-aware portrait segmentation with a large-scale teleconferencing video dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 202-209.

[62] "Age Detector," 2022-04-29 2022. [Online]. Available: https://smartclick.ai/api/age-detector/.

[63] T. Karras *et al.*, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems,* vol. 34, pp. 852-863, 2021.

[64] A. Mir, T. Alldieck, and G. Pons-Moll, "Learning to Transfer Texture from Clothing Images to 3D Humans," 2020, doi: 10.48550/arxiv.2003.02050.

Appendix A: WISE Datset Samples

Figure 52. Image Samples of our WISE dataset

Appendix B:

A woman wearing a purple half-sleeve blouse and brown-grey knee-length shorts

Similarity Score: 26.25 Similarity Score: 24.09 Similarity Score: 24.06

Figure 55

A man in a white half-sleeve uniform (with blue and purple shoulders and collar), gray pants, and black shoes

Similarity Score: 24.95 Similarity Score: 24.62 Similarity Score: 24.61

Figure 53

A man wearing blue polo shirt with long sleeves, and black pants

Similarity Score: 23.91 Similarity Score: 23.45 Similarity Score: 23.36

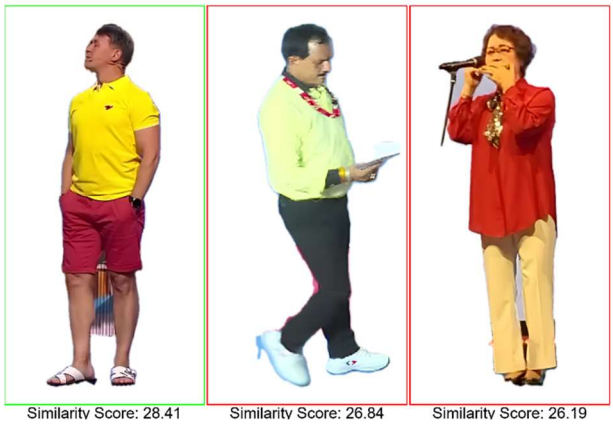
Figure 57

A man wearing a watch, white long sleeve, blue jacket, blue jeans, and Nike shoes

Similarity Score: 26.64 Similarity Score: 26.30 Similarity Score: 25.84

Figure 56

A man wearing a bright yellow half-sleeve polo shirt, red leggings, and white sandals



Similarity Score: 28.41 Similarity Score: 26.84 Similarity Score: 26.19

Figure 54

A man in a brown and gray open-collar half-sleeve shirt, long blue bleached jeans, and a pair of brown sneakers, wearing a silver watch

Similarity Score: 23.44 Similarity Score: 23.16 Similarity Score: 22.94

Figure 58

Appendix C: Synthesized Samples of StyleGAN2 Pre-trained on SHHQ

Figure 59. Synthesized Samples of StyleGAN2 Pre-trained on SHHQ

Appendix D: Synthesized Samples of StyleGAN2 Fine-tuned on WISE

Figure 60. Synthesized Samples of StyleGAN2 Fine-tuned on WISE