IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 November 2023, accepted 23 November 2023, date of publication 27 November 2023,
date of current version 1 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3337108

RESEARCH ARTICLE

Designing Practical End-to-End System for Soft Biometric-Based Person Retrieval From Surveillance Videos

JAY N. CHAUDHARI[1], (Member, IEEE), HIREN GALIYAWALA[2], MINORU KURIBAYASHI[3], (Senior Member, IEEE), PAAWAN SHARMA[4], (Senior Member, IEEE), AND MEHUL S. RAVAL[1], (Senior Member, IEEE)

[1]School of Engineering and Applied Science, Ahmedabad University, Ahmedabad, Gujarat 380009, India
[2]Rydot Infotech Pvt., Ltd., Ahmedabad, Gujarat 380058, India
[3]Center for Data-Driven Science and Artificial Intelligence, Tohoku University, 41 Kawauchi, Aoba-ku, Sendai, Miyagi 980-8576, Japan
[4]ICT Department, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat 382426, India

Corresponding author: Mehul S. Raval (mehul.raval@ahduni.edu.in)

This work was supported in part by the Science, Technology, and Innovation (STI) Policy by the Gujarat Council of Science and Technology, Department of Science and Technology, Government of the Gujarat State, India, under Grant GUJCOST/STI/2021-22/3858; in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 22K19777; and in part by the Japan Science and Technology Agency (JST) Strategic International Collaborative Research Program (SICORP) under Grant JPMJSC20C3.

ABSTRACT Video surveillance improves public safety by preventing and sensing criminal activity, enhancing quick counteractions, and presenting evidence to investigators. This is effectively performed by firing a natural language query containing soft biometrics to retrieve a person from a video. State-of-the-art (SOTA) approaches focus on improving retrieval results; thus, the building blocks of any person retrieval system are not accorded due attention, putting novice researchers at a disadvantage. This study aims to provide a design methodology by showcasing the block-by-block construction of a person retrieval system using video and natural language. For each subsystem - natural language processing, person detection, attribute recognition, and ranking- we discuss the available design selections, provide empirical evidence, and discuss bottlenecks and solutions. We thereafter select and integrate the best choices to create an end-to-end system. We highlight the integration challenges and demonstrate that the proposed method achieves an average intersection over union and the true positive rate of $\geq$60%. This is the first study to provide practical guidance to researchers for fast prototyping of person retrieval with subsystem-level understanding and achieve SOTA performance.

INDEX TERMS Person attribute recognition, detection, retrieval, soft biometrics, visual-textual problem.

I. INTRODUCTION

Daily, unfortunate events such as riots, attacks, theft, and general strikes highlight the importance of public safety. Security agencies are responsible for nabbing culprits, and surveillance networks aid them during and after these events. The vital element is locating the responsible individual(s); therefore, person retrieval is crucial for accelerating investigations. Current manual video database searches are time consuming and ineffective, leading to the emergence of intelligent video surveillance techniques.

This study aims to develop an intelligent surveillance technique using computer vision (CV) and natural language to identify persons in surveillance videos. As illustrated in Figure 1, the proposed system finds the individual(s) in a video based on soft biometrics identified from the natural language description (NLD). Soft biometrics are non intrusive and easily observable physical traits, such as gender, height, clothing style, clothing colour, and gait analysis. A task similar to the proposed method is the problem of

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

person reidentification [1], [2]. The goal is to match the persons captured by different surveillance cameras. Given an image or video of a person from one camera, the same person is identified from other videos. However, in this study, we searched for individuals based on soft biometrics attributes. Given a set of soft biometric criteria, the task is to retrieve individuals from videos that match those criteria.

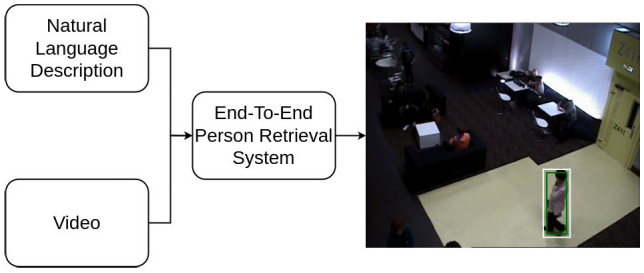


FIGURE 1. Soft biometrics-based end-to-end system for person retrieval. NLD: ***A woman in a pink shirt and black pants with a bag is walking toward the other end.*** The attributes or soft biometrics are in boldface. Image credit: AVSS Task - II frame [3].

However, there are several challenges encountered during person retrieval, which are described as follows:

1) **Integration challenges:** It is a complex system that addresses specific issues of natural language processing (NLP) and CV and establishes a correlation between their features.
2) **Semantic gap:** The method may fail to comprehensively represent a person's description and capture contextual information for effective retrieval in the video.
3) **Large volume of data:** Surveillance videos have large sizes and can contain several people, making searching for the person of interest difficult and computationally expensive.
4) **The dynamic nature of surveillance videos:** Surveillance videos are inherently dynamic, as people's movement and appearance change in the video footage. This introduces complexities during person tracking as individuals move across frames, change their poses, or alter their appearance through clothing or accessories. Such variations can hinder reliable identification and require robust algorithms that can handle the temporal aspects of the video data.

A. LITERATURE REVIEW

Deep learning approaches have found widespread application in person retrieval using soft biometric attributes. Typically, these models employ convolutional neural networks (CNNs) to extract essential features from the input. These extracted features are thereafter used to create embeddings or representations of individuals, facilitating efficient matching and retrieval. It is important to incorporate more attributes and account for their variations to enhance the retrieval accuracy [4]. For instance, by annotating 12 multiclass attributes of 110 individuals, the performance was improved by 21% over the baseline [5].

An important task is to detect a person efficiently using mask R-CNN, utilising a strong backbone network, for instance, Dense Net 161 for attribute classification, and using matching score (Hamming loss) to identify an individual [6]. Attributes can be detected sequentially through a linear filtering approach [7], however they can be inefficient because of noise propagation from one stage to another. This can be improved by adapting the detection and retrieval stages, for instance, adaptive torso patch extraction and bounding box regression [8].

Chen et al. [9] introduced the global and local image-language association to leverage semantic information in descriptions, achieving a top-1 accuracy of 43.58%. Zheng et al. [10] addressed ranking loss limitations by introducing instance level retrieval and instance loss, obtaining 44.40% top-1 rank accuracy using dual path convolutional image-text embedding architecture. Aggarwal et al. [11] employed attribute classification and deep coral loss to enhance representation learning, achieving a top-1 accuracy of 56.61% on CUHK PEDES dataset and setting a new SOTA.

Ding et al. [12] introduced a semantically self aligned network (SSAN), which extracts semantically aligned part level features, employs a multiview nonlocal network to capture body part relationships, and introduces a compound ranking (CR) loss for improved textual feature alignment. Jiang and Ye [13] introduced IRRA, a cross modal Implicit Relation Reasoning and Aligning framework. It learns relationship between visual, textual tokens, enhancing global image-text matching without extra supervision. Zhou et al. [14] proposed a text based person search model that extracts cross modal local relational global features in an end-to-end manner, enabling a fine-grained cross-modal alignment across these feature levels. It splits convolutional feature maps to capture local image features, extracts local textual features, employs a relation encoding module to learn implicit relational information, and uses a relation-aware graph attention network to fuse local and relational features for global representations of images and text queries.

The baseline method in unified person attribute recognition (UPAR) [15] allows the development of large-scale, generalisable attribute-based person retrieval. The approaches mentioned in [16] by Specker and Beyerer simultaneously improve single-attribute based recognition and retrieval using spatial projection and normalisation modules. A novel recurrent neural network with a gated neural attention mechanism established a baseline performance over CUHK PEDES, a large-scale person re-identification dataset containing NLDs with images [17].

We note that the existing SOTA method pushes boundaries for accurate person retrieval and discusses theoretical concepts appropriately. However, existing approaches do not discuss the design methodology, which is beneficial for novice researchers. They typically do not discuss the challenges faced by each building block in an end-to-end system. Although significant reproducible research is available, existing methods do not discuss the selection of

off-the-shelf tools that can quicken prototyping. Additionally, they do not test or suggest alternative tools for each design block.

Moreover, they fail to discuss integration challenges when different subsystems are interconnected owing to the current trend toward black-box designs. The lack of discussion on the methodology hinders reproducibility, benchmarking, and failure to provide practical guidance for constructing such systems. Most approaches employ discrete annotations (DAs) to provide limited information. Currently, there is a lack of a framework that enables seamless integration of existing models developed for DA datasets with NLD. This limits the use of existing models for datasets containing NLDs.

B. OUR CONTRIBUTION

This paper presents a systematic design methodology for constructing person retrieval system. It provides important insights and hints for designing a customised system while highlighting potential bottlenecks and difficulties. This paper presents practical guidance that allows practitioners to construct real world systems using quick prototyping.

C. PAPER ORGANISATION

The structure of the paper is outlined as follows: Section I, presents the literature review, which illustrates the evolution of the person retrieval problem. Section II outlines the proposed framework. Section III covers the datasets employed and the associated challenges. Experiments on NLP subsystem and vision subsystems are presented in Section IV. This section provides insights into the issues and challenges faced during subsystem development. Section V discusses the development of an end-to-end system for person retrieval, constructed by integrating the best subsystems. Section VI presents conclusions, limitations, and directions for future research.

II. PROPOSED FRAMEWORK

The proposed framework shown in Figure 2, comprises of two subsystems:- NLP and Vision. The NLP subsystem uses a methodology to extract DAs from the NLD of individuals. The vision subsystem comprises the detection and person attribute recognition (PAR) channels. The former extracts video frames and automatically locates all human figures. The latter uses identified humans and automatically labels specific attributes such as gender, age, clothing colour, and type. The labelled attributes and DAs were combined to generate a person ranking, with the individual with the highest score being the target.

III. DATASETS

Datasets form the foundation of deep learning-based systems, particularly in complex architectures with multiple independent subsystems. Ideally datasets should possess several key characteristics. First, the model must be accurate, consistent, and representative of real world scenarios. Balanced datasets are vital for reducing ambiguity and uncertainty during training. They must be well organised and properly annotated to accelerate the training process and optimise computational resources. Furthermore, a diverse dataset enables the model to learn from various contexts and variations. By extending these characteristics, an ideal person retrieval dataset may possess the following features:

1) **Image samples with NLDs**: The dataset should contain images of individuals and corresponding NLDs and DAs.
2) **Rich and varied NLDs**: The dataset should have diverse and detailed NLDs that cover various attributes of a person, such as clothing colour, patterns, accessories, and physical characteristics.
3) **Multiple sightings of the same person**: This allows the model to learn to match a person based on NLDs and DAs despite appearance variations.
4) **Natural language diversity**: The NLDs must represent various writing styles and language variations to help the model better generalise to real world text inputs.
5) **Ethical considerations**: The images and textual data are collected ethically, with proper consent and anonymity of individuals in the dataset.
6) **Balanced data**: The number of attributes with NLDs should be equal, avoiding biases in the training of the model.

Based on these criteria, we used following datasets.

A. DATASET FOR TRAINING DETECTION CHANNEL

CrowdHuman [18] is a benchmark dataset used for evaluating person detectors in crowds. It is large, richly annotated, and has a high diversity of subjects. The dataset contains 15,000, 4,370, and 5,000 images for training, validation, and testing, respectively. It has 4,70,000 human instances from the training and validation subsets, with 22.6 persons per image with various occlusions in the dataset. The dataset encompasses several scenes, including streets, parks, markets, and stadiums, along with occluded images, such as individuals partially obstructed by other people, objects, or the background. Figure 3 shows images from the CrowdHuman dataset [18], highlighting crowd densities and other complexities.

B. DATASET FOR ATTRIBUTE RECOGNITION CHANNEL

The dataset used for training the PAR model is derived by the combination of AVSS 2018 Challenge II [3] and Richly Annotated Pedestrian v1 (RAP v1) [19], henceforth known as AVSS+RAP dataset created by Galiyawala et al. [8]. The second dataset used is UPAR [15]. The AVSS and RAP datasets were merged and annotated for six attributes and corresponding forty six values [8]. This was performed to effectively train the PAR model, as the AVSS Challenge II dataset had extremly few images with poor resolution. The images from RAP v1 had a better resolution; merging them with the AVSS 2018 Challenge-II dataset increased the sample size and variability for generalisation. The datasets were as follows:

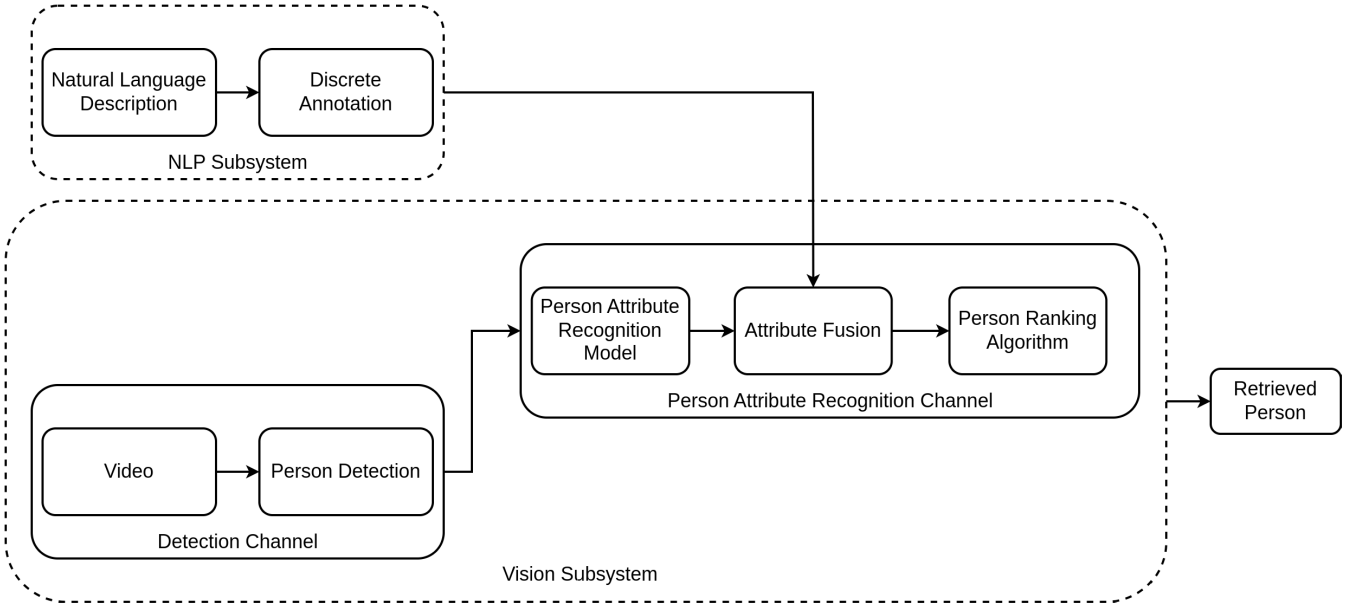


FIGURE 2. End-to-end system for person retrieval.



FIGURE 3. Samples from CrowdHuman dataset [18].

1) AVSS+RAP [8]: The AVSS 2018 Challenge II [3] dataset has 14,000 annotated images representing diverse scenes such as streets, parks, markets, and stadiums. Each image has 6 attributes, including gender, age, clothing colours and types, and accessories. The dataset also incorporates occlusions, such

as partial obstructions by people, objects, or the background. The RAPv1 [19] dataset is a widely used benchmark dataset for pedestrian attribute recognition. It comprises 41,585 images with rich annotations for various pedestrian attributes such as clothing colours, types, accessories, hairstyles, and gender.

2) UPAR: The UPAR [15] dataset is a large scale dataset for person attribute recognition and retrieval. It comprises 224,737 images with annotations for 40 binary values. The dataset is divided into a training set of 148,048 images, a validation set of 30,830 images, and a test set of 45,859 images. The UPAR dataset was created by combining four existing datasets: PA100K [19], PETA [20], RAPv2 [21], and Market1501 [22].

Figure 4 comprehensively depicts the challenges encountered with the UPAR dataset. Figure 4 (a) shows multiple individuals wearing similar clothing, and Figure 4 (b) illustrates the complexity of colour fidelity. Figure 4 (c) shows individuals blending into the background because of inadequate illumination, whereas Figure 4 (d) shows faulty viewpoints and excessive illumination. Figure 4 (e) shows an occluded scenario, where only a fraction of the person is visible in the image, and Figure 4 (f) shows a noisy image. UPAR adopts a binary gender representation with females as 1 and males as 0. Simultaneously, another dataset, the AVSS, has separate attributes for males and females. The UPAR annotation approach avoided unnecessary information, and produced efficient and concise attribute representations within the PAR framework.



FIGURE 4. Samples depicting the various challenges from the UPAR dataset [15].

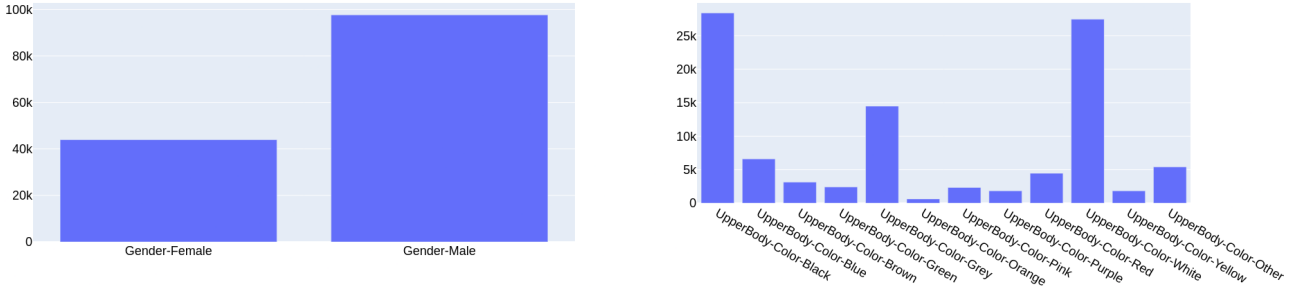
C. TEST DATASET FOR RETRIEVAL TASK

The proposed approach for PAR based person retrieval used the AVSS 2018 Challenge- II dataset introduced by [3]. This is the only publicly available video dataset comprising of video sequences captured by six stationary calibrated cameras, each with a resolution of 704 × 576 pixels. The primary purpose was to assess person retrieval algorithms in real world surveillance scenarios. The training set of the AVSS 2018 Challenge II dataset [3] contains unconstrained video sequences from 110 individuals, whereas the testing set comprises video sequences from 41 individuals. Each training sequence is annotated with 16 soft biometric attributes and 55 values, providing additional information regarding person's appearance, including clothing, accessories, and other characteristics. Each video sequence within the AVSS person [3] retrieval dataset captures individuals moving in and out of the camera view, and annotations are provided for a **single subject** in each video frame.

D. PRACTICAL CHALLENGES WITH DATASETS

In Section III, we enlist the characteristics of an ideal retrieval dataset; however several practical challenges are discussed as follows:

1) **Image samples:** Almost all datasets have an image gallery for person attribute extraction, and barring AVSS Challenge II, no annotated video dataset is available. This limits multiple sightings of the same person across the dataset.
2) **Lack of generalisation:** Most datasets have DAs; extremely few have sentences as annotations [23]. This limits model generalisation owing to the lack of varied textual inputs.
3) **Inaccurate annotations:** The ground truth bounding boxes in the CrowdHuman [18] dataset are not always annotated accurately, and sometimes they overlap. This makes it difficult for object detection algorithms to identify and track people in the crowd.
4) **Imbalance dataset:** The CrowdHuman dataset [18] is not balanced for the number of people in each image. Some images contain several people, while others contain only a few. This makes it difficult for object detection algorithms to generalise during different crowd densities. The AVSS+RAP [8] and UPAR dataset [15] are imbalanced in attributes. Figure 5 provides a clear visual representation of the significant attribute imbalance in the UPAR dataset. In particular, Figure 5a visually shows the imbalance between male and female attributes, with male attributes being more prevalent. The same applies to the distribution of upper body clothing colours and lower body colours.
5) **Limited scenarios:** The CrowdHuman dataset [18] contains images of people in outdoor scenes. There are few images of people in indoor scenes or crowded



(a) Gender attribute imbalance.

(b) Upper body clothing colour attribute imbalance.

FIGURE 5. **Imbalance of attributes in UPAR dataset [15].**

transportation settings. This can make it difficult for object detection algorithms to generalise to different types of scenes.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section discusses the experimentation with the NLP and vision subsystems. It provides practical guidance by suggesting design selections, corresponding results, practical difficulties, and making the best selection based on reasoning.

A. NLP SUBSYSTEM

Person retrieval is the task of identifying a person who matches the textual description. This task is challenging because NLD is often ambiguous and can refer to multiple people. For instance, the description "*a tall man with brown hair*" could refer to several people in a crowded scene. The use of DA reduces ambiguity and makes retrieval more accurate and efficient. Keywords can be extracted from NLD to obtain DA and use them as a query for the video. Therefore, in this study, NLD is converted to DA in several ways. However, the challenges associated with converting NLD to DA are as follows:

1) **Ambiguity and variability:** Textual descriptions can often be ambiguous or exhibit variations in language use. Existing methods may struggle to handle such cases accurately, leading to potential errors or inconsistencies in the generated annotations. For instance, skin colour can be perceived as a noun and verb as well as a noun and adjective when given as *skin coloured pants or shirt.*
2) **Dependency on training data:** Creating comprehensive and representative training datasets can be time consuming and resource intensive.
3) **Difficulty wi th unstructured text:** When handling unstructured text, such as social media posts, where the language can be informal, contain abbreviations, slang, or misspellings, extracting meaningful annotations from it can be a complex task.
4) **Limited generalisation:** Existing NLP methods struggle to generalise appropriately to new or unseen textual descriptions that differ significantly from the training data. They may exhibit limited adaptability to evolving language patterns, leading to reduced accuracy and effectiveness in annotation generation.

Considering these challenges, we used pre-trained large langauge models (LLMs) for convert NLD to DA, as discussed further.

1) DA USING SPACY

SpaCy [24] is an open source library for NLP tasks such as tokenisation, part-of-speech tagging, morphing, lemmatisation, named entity recognition, and sentence segmentation. The NLD was given as input and passed through two system functions, a SpaCy matcher and customised adjective noun extraction. The SpaCy matcher is a function that derives the adjective noun pair, and stores it in the matcher list, for instance, tall man, white shirt. The customised adjective-adjective noun extraction method extracts the adjective-adjective pair and appends to a custom list, for instance: black and blue jeans. The matcher and custom list are coupled into the final list. The list passes a function that maps the adjectives and nouns present into binary attributes according to the index. The following example explains the NLD to the DA extraction process: **NLD:** *A tall male with red and purple short-sleeved clothing wearing black jeans is sitting on the sofa.*

Matcher list: *tall male, black jeans*
Custom list: *red and purple*
Final List: *tall male, black jeans, red and purple.*

The limitations of the method described in Figure 6 are that it requires a preprocessing step, is unreliable for large datasets, and has high computational complexity. Bidirectional Encoder Representations from Transformers (BERT) models solvethese problems by establishing word correlations and enhancing comprehension.

2) BERT BASED MODEL

Bidirectional encoder representations distilled from transformers (DistilBERT) are SOTA LLMs used in NLP

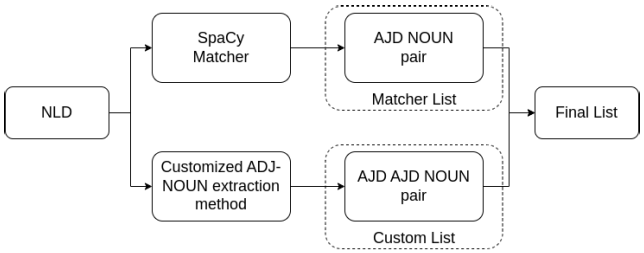


FIGURE 6. Algorithm of NLD to DA conversion.

tasks [25]. They use knowledge distillation to reduce the model size without compromising the language understanding capabilities. A robust optimised BERT pretraining approach (RoBERTa) [26] is an advanced NLP model based on the BERT architecture. RoBERTa refined the pretraining process of BERT by using larger batch sizes and training data, removing the next sentence prediction function, and using more training steps. These enhancements enable RoBERTa to perform better at converting NLD into DA.

The proposed method uses DistilBERT and RoBERTa models trained on a dataset created using the ChatGPT API [27]. DAs from AVSS 2018 Challenge II were used to generate sentences. DA was used as an input prompt to generate sentences using the ChatGPT API. This process generated 23,000 sentences. However, during sentence generation, 12% of the sentences had identical descriptions owing to the similarity of the annotations. Duplicate sentences were removed from the dataset, and training was performed using the remaining diverse set of sentences. The DistilBERT and RoBERTa model parameters for training are listed in Table 1.

TABLE 1. Parameters for training BERT based models.

Parameters	DistilBERT	RoBERTa
Number of labels	46	46
Batch size	8	8
Gradient accumulation steps	16	16
Learning rate	3e-04	1e-04
Epochs	25	10
Maximum sequence length	75	75

TABLE 2. Performance of BERT based models for NLD to DA conversion.

Model	LRAP
DistilBERT	**82.18**
RoBERTa	80.85

Table 2 summarizes a comprehensive evaluation of the performance of the DistilBERT and RoBERTa models using the label ranking average precision (LRAP) metric [28]. Here, each observation is associated with multiple labels, and their order is also important, therefore LRAP is used, which quantifies the ability of the classifier to assign higher scores to the correct labels than to the false ones. From Table 2, DistilBERT performs better than the RoBERTa model.

The formula for LRAP is given by Equation 1.

$$\text{LRAP} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i^{\text{true}}|}\sum_{j\in Y_i^{\text{true}}}\frac{|\text{pos}_i \cap Y_{i,j}^{\text{pred}}|}{|\text{pos}_i|} \tag{1}$$

where n denotes the number of instances in the dataset. Y_i^{true} denotes the set of true positive labels, for instance i. $Y_{i,j}^{\text{pred}}$ denotes the set of predicted labels for instance i at rank j. pos_i denotes the set of positive labels for instance i.

BERT-based models require extensive training and memory resources, including graphics processing unit (GPU) usage; therefore, their substantial size remains challenging. Utilising the ChatGPT API of open AI [27] provides a more feasible way to address these constraints.

3) CHATGPT API

The person description in the previous section was generated using the ChatGPT API from the DAs provided in the datasets. The sentences followed a similar pattern with limited context and lacked diversity, which BERT-based models could easily exploit. Therefore, in this section, we provide manually modified sentences to break the monotony of the model and apply them to ChatGPT API based DA extraction method. Figure 7 shows the algorithm used to extract the DA from NLD using the ChatGPT version gpt-3.5-turbo. For instance:

NLD: " *A tall male with red and purple short sleeved clothing wearing black jeans is sitting on the sofa*".

Gender: male
upper body clothing type: short-sleeved clothing
upper body clothing colour: red and purple
lower body clothing type: jeans
lower body clothing colour: black.

If the attributes are in the required format, they are directly passed to the attribute dictionary; if not, these attributes are passed into a synonym search list and converted into the required format. The sentence contains a visual description of a person. It is necessary to analyse the sentence and extract its attributes.

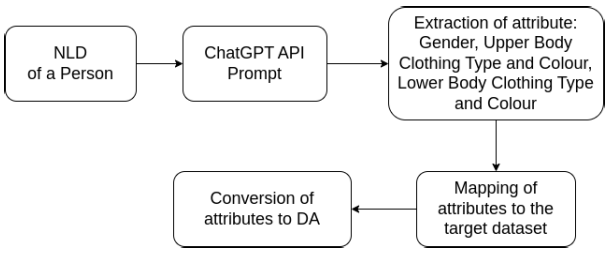


FIGURE 7. Algorithm for extraction of DAs using ChatGPT API.

The API prompt is as follows: *"The sentence contains a visual description of a person. It is necessary to analyse the sentence and extract its attributes. The extracted attributes should be mapped as follows: type of upper body clothing to upper_cloth_attributes; colour of upper body clothing to upper_cloth_color_attributes; type*

of lower body clothing to lower_cloth_attributes; colour of lower body clothing to lower_cloth_color_attributes; colour of shoes to shoes_color; gender to gender_attributes."

Table 3 lists the ground truth values for the five attributes in a sentence describing test subject ID 005. Overall, the AVSS Test-II dataset had 16 attributes with 55 values, but the test subjects listed in Table 3 had only five of them; so they are listed in the table. The performance of NLP models with variations in sentence style is discussed in Tables 4, 5, and 6. They provide insights into the effectiveness of the models in handling simple, indirect, and complex sentences.

Table 4 lists the extraction of DA from a simple sentence. The sentence is straightforward; thus, by comparing the predictions with the ground truth, we observed that all models performed well. DistilBERT and RoBERTa models made errors on one attribute, whereas SpaCy and ChatGPT API extracted the DA correctly.

Table 5 summarises the performance for indirect sentences. Here, except for ChatGPT API, all other models made errors. The DA with SpaCy and RoBERTa could not find the true value of "Male"; when "He" is provided in the sentence.

Table 6 lists the cases with the most complex sentences. Here, except for ChatGPT API, all other models made errors. The DA with SpaCy and RoBERTa could not map a value to "Male" if the word "He" appears in the sentence. Moreover, except for ChatGPT, every other model skipped the second colour mentioned in the sentence for upper body clothing.

Tables 4, 5, 6 summarise the influence of imbalanced attributes. The prevalence of red and purple colours was notably lower than that of other colours such as black and white in the dataset. This imbalance was mainly observed in the BERT-based model. Given the uneven distribution of attributes in the training dataset, this discrepancy inevitably persists into sentence generation. The key challenges encountered during the customisation of the models are as follows:

1) **Compound adjectives:** Extracting adjective-noun pairs with compound adjectives (for instance, *A male wearing a **black and white** Tshirt.*) posed a problem. The matcher only captured the latter adjective, creating a completely altered pair. Users were instructed to use an underscore between the adjectives in their queries to address this issue.
2) **Presence of homonyms:** Within AVSS, the attribute key "skin" was used to refer to various colour related attributes, but it could also function as a noun. The key was modified to "*skin_coloured*" to avoid confusion and maintain consistency.
3) **Computational complexity:** Despite being more compact than the original BERT, DistilBERT still demands significant computational resources, particularly during training and inference. Implementing these models on resource constrained devices or in large scale applications is challenging.
4) **Prolonged training:** Optimisation of RoBERTa in pretraining requires prolonged training with exhaustive data compared to other models.
5) **Limited context window:** DistilBERT, similar to BERT, has a limited context window, restricting its ability to consider extremely long texts or documents in a single pass, potentially missing out on critical context.
6) **Cost:** Cost is incurred on the API calls while using the ChatGPT API. It can quickly accumulate depending on the usage.
7) **Transparency:** The ChatGPT API does not offer the same model access and customisation level compared to using BERT directly.
8) **Dependency on cloud resources:** ChatGPT API responses depend on the Internet connection and load on the server, which may introduce some latency.

Considering Tables 4, 5, and 6, ChatGPT is superior to other models owing to its higher word processing capability. Therefore, despite its dependency on the API cost, we shortlisted ChatGPT API-based DA extraction as an NLP subsystem while constructing the end-to-end system.

B. DETECTION CHANNEL

The detection channel is important for locating a person in a video frame and forwarding it to the PAR model. One widely used object detection technique for single-shot detection is the YOLO (You Only Look Once) model. Notably, YOLO V7 [29] demonstrated a promising performance on ImageNet [30], a benchmark dataset for object recognition tasks. Following this breakthrough, Ultralytics introduced YOLO V8 [31], which demonstrated impressive results on the ImageNet dataset.

The proposed method trained YOLO V7 and YOLO V8 on the CrowdHuman dataset [18]. The following section comprehensively describes the experimental procedures and findings of person detection using YOLO V7 and YOLO V8. The CV challenges associated with person detection are as follows:

1) **Occlusion:** In crowded environments, individuals are frequently occluded by other people, objects, or the surrounding environment. This occlusion hinders the visibility of individuals, making it arduous to identify and track people accurately. Figure 8 shows a person's occlusion behind the lamp.
2) **Variety:** The diversity among people in body shapes, sizes, clothing, and carried objects adds complexity to person detection (refer to Figure 8). Object detectors must be trained on a wide range of appearances to ensure reliable identification of individuals in crowded scenes. Handling such variability requires robust models capable of capturing details of human appearances.
3) **Background clutter:** Crowded scenes are typically characterised by high levels of background clutter, including structures such as trees, buildings, and various objects. The presence of such elements can create visual distractions and make it challenging to distinguish individuals from the background. Effective person detection algorithms must be able to

TABLE 3. AVSS Test - II dataset's [3] DA ground truth for the test subject 005.

Attributes	Gender	Upper Body Clothing	Upper Body Clothing Colour	Lower Body Clothing	Upper Body Clothing Colour
Value	Male	Short-sleeved	Red and Purple	Jeans	Black

TABLE 4. Performance of NLP models for the simple sentence describing test subject 005: *A tall male with red and purple short-sleeved clothing wearing black jeans is sitting on the sofa.*

Methods	Gender	Upper Body Clothing	Upper Body Clothing Colour	Lower Body Clothing	Lower Body Clothing Colour
DA with SpaCy	Male	Short-sleeved	Red and Purple	Jeans	Black
DistilBERT	Male	Others	Red and Purple	Jeans	Black
RoBERTa	Male	Short-sleeved	Purple	Long Trousers	Black
ChatGPT API	Male	Short-sleeved	Red and Purple	Jeans	Black

TABLE 5. Performance of NLP models for the indirect sentence describing test subject 005: *He is wearing a red short-sleeved cloth. The cloth is also having purple texture. He seems to wear a black jeans with a leather belt.*

Methods	Gender	Upper Body Clothing	Upper Body Clothing Colour	Lower Body Clothing	Lower Body Clothing Colour
DA with SpaCy	NA	Sleeve	Red	Jeans	Black
DistilBERT	Male	Others	Red and Purple	Long Trousers	Gray
RoBERTa	NA	Short-sleeved	Purple	Others	Black
ChatGPT API	Male	Short-sleeved	Red and Purple	Jeans	Black

TABLE 6. Performance of NLP models for a complex sentence describing test subject 005: *The person seems to be masculine. It may be a red or purple short-sleeved T-shirt, and may have both colours. He may be wearing a dress such as jeans of black colour. The person also had a backpack.*

Methods	Gender	Upper Body Clothing	Upper Body Clothing Colour	Lower Body Clothing	Lower Body Clothing Colour
DA with SpaCy	NA	Short-sleeved	Red	Jeans	Black
DistilBERT	Male	Others	Red	Others	Black
RoBERTa	NA	Other	Purple	Dress	Gray
ChatGPT API	Male	Short-sleeved	Red and Purple	Jeans	Black

differentiate people from the surrounding clutter to avoid false detection or missed individuals.

4) **Lighting conditions:** The lighting conditions in crowded scenes can vary significantly, spatially and temporally. Shifting shadows, uneven illumination, and varying light intensities can adversely affect the performance of object detectors. Adapting to different lighting conditions and handling variations in contrast and brightness is essential for robust person detection in crowded scenes. Figure 8 shows a sample from the AVSS Challenge II dataset under poor lighting conditions and a person merging with the background.
5) **Scale variation:** People in crowded scenes can appear at different distances from the camera, resulting in significant scale variations. This makes it challenging to detect people accurately across different scales and distances.
6) **Interactions and dynamics:** People in crowded scenes may be engaged in complex interactions, such as walking closely, hugging, or crossing paths leading to occlusions, multiple persons, and partial visibility of person in the single bounding box, making it noisy for the PAR model. Figure 8 shows crowded scenes with complex interactions and various poses.
7) **Real time processing:** In scenarios where real time processing is required, such as video surveillance, the person detection algorithm needs to operate at 25 frames per second. Achieving real time performance while maintaining high accuracy is a significant challenge.

1) YOLO V7 VS YOLO V8

In this section, we compare the two leading object detection models, YOLO V7 and YOLO V8, through a performance evaluation of the CrowdHuman dataset. The dataset was divided into training and test datasets. During the experimental phase, both the models were trained for 100 epochs. The model performance was evaluated using the mean average precision (mAP), which considers the precision and recall values across different levels of confidence thresholds and is defined as follows:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{2}$$

where n represents the number of object categories and AP denotes the average precision for category i.



FIGURE 8. Sample from the AVSS Challenge II dataset with a crowded scene, a person merging with the background is highlighted by a white ellipse, a person is obscured by a lamp, and a green ellipse emphasises the occlusion.

A comprehensive experiment was conducted by changing hyperparameters such as the learning rate, batch size, and number of epochs. The learning rate significantly improved the training process by decreasing it to 0.001 from 0.01. The validation accuracy increased linearly, and a similar pattern was observed for the training and validation loss. Table 7 lists the training parameters employed for the best performance of the YOLO V7 and YOLO V8 models.

TABLE 7. Training parameters of the YOLO models.

Parameters	YOLO V7	YOLO V8
Workers	8	8
Optimiser	SGD	Auto
Momentum	0.9	0.9
Epochs	100	100
Batch size	16	16
Learning rate	1.00E-03	1.00E-03
Weight Decay	0.0005	0.0005
Image size	640	640

Table 8 summarises a quantitative evaluation of training on the CrowdHuman dataset [18] based on two metrics: mAP50 and mAP50-95. The results indicate that the YOLO V7 models are better than the YOLO V8 models. However, it is important to test the detection performance and inference on the dataset. Therefore, we assessed the performance of all four models on the target AVSS Test-II dataset [3].

TABLE 8. Quantitative metrics for training on CrowdHuman dataset [18].

Model	mAP50	mAP 50-95	Parameters (M)
YOLO V7 tiny	0.81	0.48	06.2
YOLO V7 base	0.83	0.52	37.1
YOLO V8n	0.73	0.46	03.0
YOLO V8s	0.78	0.51	11.1

The sample results are summarised in Table 9 and Figures 9 and 10. All detectors had false negatives, as listed in Table 9 while detecting the number of persons and heads. It can also be observed that false negatives and inference times were greater in YOLO v7 models than in YOLO v8 variants. Based on the mAP, false negatives, inference time, and post processing time, we selected YOLO V8s as the model for use in an end-to-end system.

C. PAR CHANNEL

The attribute channel comprises three essential components: the PAR model, attribute fusion (AF), and the person ranking algorithm (PRA).

1) **PAR Model:** It extracts various attributes from the given person's image and uses multitask deep learning techniques, typically based on a feature extractor, to automatically learn discriminative features from the images. These learned features are then used to predict different attributes associated with individuals, such as gender, age, clothing style, and other appearance characteristics. It is trained using backbone networks such as ConvNeXt [32], CoAtNet [33], and ResNet-50 [34] and uses 6 attributes for the AVSS+RAP dataset and 12 attributes for the UPAR dataset.
2) **Attribute Fusion:** It handles the element wise multiplication of the extracted DAs and the predicted attributes.
3) **Person Ranking Algorithm:** It evaluates and ranks individuals based on their attribute scores. Once the PAR model has predicted attributes for each person in the images, the PRA calculates a score based on the attributes from attribute correlation. This score can be used to rank individuals according to specific criteria.

1) PERSON ATTRIBUTE RECOGNITION MODEL

The multiattribute PAR model was designed for simultaneous learning of multiple attributes. The challenges associated with the PAR models are as follows:

1) **Large variability in pose and appearance:** In surveillance settings, people can appear in different poses, wear different clothing, and have various appearances owing to changes in lighting and camera angles. This variability makes it difficult for the model to generalise across different instances of the same attribute.
2) **Occlusions and low resolution:** Surveillance videos may suffer from occlusions. Moreover, the low resolution of the footage can make it hard to discern the finer details necessary for accurate attribute prediction.
3) **Imbalanced data:** In real world surveillance data, the distribution of different attributes may be imbalanced. As shown in Figure 5, colour attributes such as purple and pink might have rare occurrences, making it challenging for the model to learn from limited positive samples.
4) **Domain adaptation:** Models trained on one surveillance dataset may not generalise well to other datasets owing to domain shifts. This challenge requires

TABLE 9. Performance metrics for a video frame shown in Figures 9 and 10 from AVSS dataset [3]. The total number of people in the frame is 12.

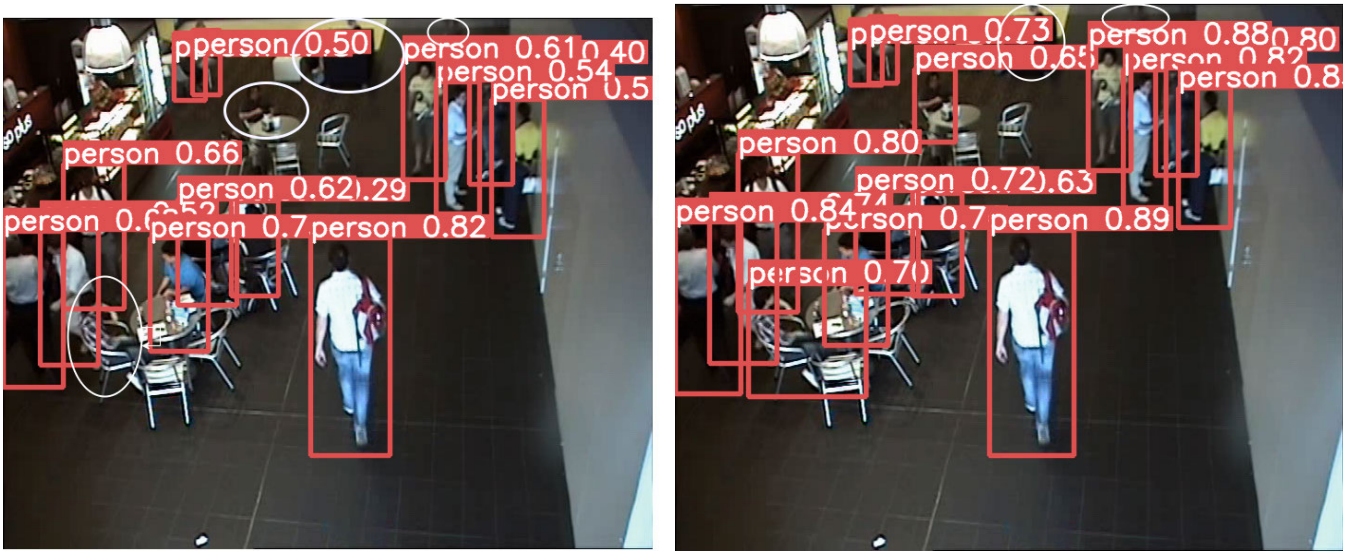
Model	Persons detected	Inference(ms)	Post Processing(ms)
YOLO V7 tiny	08	296.5	2.5
YOLO V7 base	11	296.8	2.5
YOLO V8n	09	040.8	5.5
YOLO V8s	11	099.1	7.1



(a) Performance of YOLO V7 tiny on a frame from dataset.

(b) Performance of YOLO V7 base on a frame from dataset.

FIGURE 9. YOLO V7 tiny and YOLO V7 base fail to detect five and two heads respectively, as listed in Table 9, and it also failed to detect the farthest person sitting on the sofa encircled by a white ellipse on a sample from AVSS dataset [3].



(a) Performance of YOLO V8n on a frame from dataset.

(b) Performance of YOLO V8s on a frame from the dataset.

FIGURE 10. YOLO V8n fails to detect two heads and two persons. The false negatives of the model are highlighted by an ellipse on a sample from the AVSS dataset [3].

strategies for domain adaptation to improve the model's performance on unseen data.

5) **Handling noisy data:** Surveillance videos can contain various noise sources affecting the model's performance, such as motion blur, camera jitters, or environmental factors.
6) **Attribute correlations:** Some attributes may be correlated, implying that the presence or absence of

one attribute might influence the likelihood of others. Capturing and managing these correlations effectively is vital for accurate predictions.

The architecture of the PAR model is shown in Figure 11. The PAR model is a multitasked deep learning model that employs a backbone for feature extraction. The model achieved attribute recognition by training fully connected layer networks for each attribute and generating a binary vector as the output. The imbalanced nature of the attributes in the dataset is handled by focal loss [35], and each fully connected layer predicts the probabilities of the individual attributes. These probabilities are passed through the sigmoid activation function, which converts the probabilities to zero if they are less than 50%; otherwise, they are 1.

We trained the PAR model with several backbones. The top three trained models based on the F1 score are listed in Table 10 and Table 11 for the AVSS+RAP and UPAR datasets, respectively. As summarised in Tables 10 and 11, ConvNeXt-Large [32] achieved the best F1 score for both datasets during training and testing. It leverages the power of spatial and depthwise convolutions, yielding a better performance. Therefore, we selected ConvNeXt-Large [32] as the model for the end-to-end system.

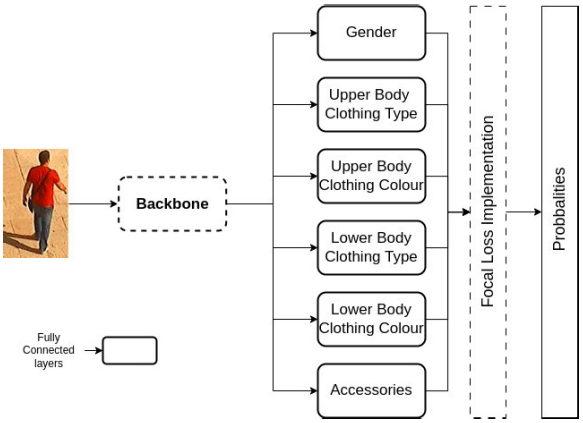


FIGURE 11. Architecture for PAR model. 40 probabilities are obtained for UPAR Dataset and 46 probabilities for AVSS+RAP dataset.

2) ATTRIBUTE FUSION

The attributes predicted by the PAR model were compared with the DA by performing element-wise multiplications. This filters incorrectly predicted attributes, eliminates false positives, and allows the model to focus on the target person, resulting in efficient memory management and improved performance.

3) PERSON RANKING ALGORITHM

Several distance metrics were used to rank the person in AVSS 2018 Challenge-II dataset. These include cosine similarity, Hamming distance, a hybrid of cosine similarity and Hamming distance, and shallow neural networks. Figure 12 shows the shallow neural network designed to predict the ranking score of each detected person. The network architecture comprises two hidden layers with 128 and 64 units, respectively, using the 'GELU' activation function. Additionally, there is an output layer with a single unit employing the 'sigmoid' activation function.

Overall, the attributes of the person detected in the frame were provided as inputs to the shallow network. The network thereafter calculated a score for each individual based on the predicted attributes. Once all the individuals in the frame are scored, the person with the highest score is identified as the target. It is important to note that the accuracy of person ranking using a shallow network relies heavily on the quality of attribute predictions by the PAR model. If attribute predictions are inaccurate or unreliable, the ranking results may not be accurate or consistent.

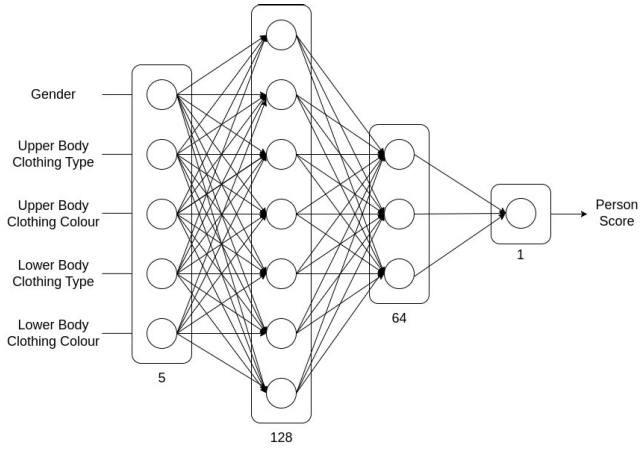


FIGURE 12. Shallow neural network for person ranking.

Table 12 lists the hyperparameter used to train the shallow neural network. The data for training the shallow network were prepared by assigning a ground truth value of 1 to the attributes correctly predicted by the PAR model. By contrast, attributes with at least three incorrect predictions were assigned a false value of 0. This step ensures the creation of a dataset that captures the accuracy of attribute predictions.

The performance of the PAR channel was evaluated using the intersection over union (IOU) and true positive rate (TPR). IOU describes the overlap between the bounding boxes of the person retrieved by the PAR model and the ground truth. Here, we computed the average IOU, that is, the average of the IOUs for all persons detected in a frame. Similarly, the TPR was calculated as follows:

$$TPR = \left(\frac{\text{No. of frames with correct retrieval}}{\text{Total frames}}\right) \times 100 \quad (3)$$

The performance of the PAR model is summarised in Table 13. We can observe that allowing the shallow neural network to learn the weights of different attributes significantly improves the retrieval. The other metrics merely compute the distances between the ground truth and the predicted vectors and are not flexible in person ranking.

TABLE 10. Results for attribute recognition on AVSS+RAP dataset.

Backbone	Training accuracy	Testing accuracy	Precision	Recall	F1-Score
CoAtNet	**99.44**	**93.29**	**77.19**	**75.35**	**76.26**
ConvNeXt Large	97.25	93.47	79.69	72.98	76.19
ResNet-50	87.48	87.17	87.31	73.72	75.93

TABLE 11. Results for attribute recognition on UPAR dataset.

Backbone	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
ConvNeXt Large	**99.87**	**97.25**	**92.28**	**91.56**	**91.92**
CoAtNet	98.18	96.23	91.37	86.34	88.78
ResNet-50	96.9	95.19	89.42	81.95	85.52

TABLE 12. Parameters for training the shallow neural network.

Parameter	Value
Learning Rate	1.00E-02
Epsilon	1.00E-07
beta_1	0.9
beta_2	0.999
Epochs	10
Optimiser	Adam

TABLE 13. Results of person ranking strategies on AVSS Challenge II dataset.

Person ranking Algorithm	Avg. IOU	TPR
Shallow network	**0.65**	**68%**
Cosine similarity	0.40	48%
Hamming distance	0.45	41%
Hybrid method	0.35	23%

The advantages of using a shallow network for person ranking are its simplicity and efficiency. Therefore, considering performance, we a used shallow network for ranking in the end-to-end system.

V. END-TO-END FRAMEWORK

Based on the selections in Section IV, we constructed an end-to-end framework using the following components.

- NLP: ChatGPT API for NLD to DA conversion.
- Person detection using YOLO V8s trained on CrowdHuman dataset.
- PAR uses ConvNeXt-Large as the backbone trained on AVSS+RAP, UPAR.
- Ranking using the shallow neural network.
- End-to-end framework is tested on the AVSS Challenge-II test dataset.

A. NUMBER OF ATTRIBUTES IN RETRIEVAL

Figure 13 illustrates the end-to-end framework used for person retrieval.

AVSS 2018 Challenge-II [3] has 55 values, and UPAR [15] has 40 values; however, only some of them play a dominant role in the retrieval process [36]. For instance, we observe that an attribute such as *age* is difficult to predict when there is a back view or a partial view of the person. Similarly, accessories such as purses and watches were difficult to recognise. In contrast, acting as noise may contribute to misinterpretations. Therefore, we employ a retrieval process using five attributes: *gender, upper body clothing type and colour, lower body clothing type, and colour.* The AVSS 2018 Challenge-II test dataset provides two clothing colour annotations for the upper body: torso colour-1, torso colour-2, and similarly for the lower body: leg colour-1, leg colour-2. In the cases with two-colour attributes, we considered the colour with the highest score in the network.

B. SUCCESSFUL RETRIEVAL

Figure 14 illustrates the frames in which the person retrieval process was successful across frames of various difficulty levels. The notation "TS.10, F.44 (very easy)" refers to Test Sequence 10 with frame number 44 and a very easy difficulty level frame. Individuals with a detection score of less than 0.35 are excluded during retrieval. The green bounding box in the frame shows the ground truth, and the white bounding box shows the rank-1 prediction.

Figure 14a depicts a scenario from TS:003, F44, characterised by a **very easy difficulty level**, where the target person is visible, the illumination is adequate, and there are few other individuals in the frame. The camera angle was a corner top view, allowing for a comprehensive view of a person's attributes and correct retrieval. Figure 14b shows a person from TS:008, F71, presenting a **medium difficulty** situation. The target person was visible, the illumination remained fair, and only a few other people were in the frame. The front top camera angle offers a clear view of a person's attributes.

In Figure 14c, we encounter a person from TS:023, F72, at the **hard difficulty level**. Despite the target person being visible, the illumination was poor, and the frame was crowded with numerous individuals. Moreover, inadequate lighting and camera angles can affect the person's appearance. Interestingly, the attributes annotated in the dataset indicate the upper body clothing colour as brown, but appears yellow when viewed in the frame. Finally, Figure 14d portrays a person from TS:023, F72, under **very hard difficulty** conditions. Here, the target person almost merges with the background because of the challenging illumination conditions and a crowded frame with multiple people.

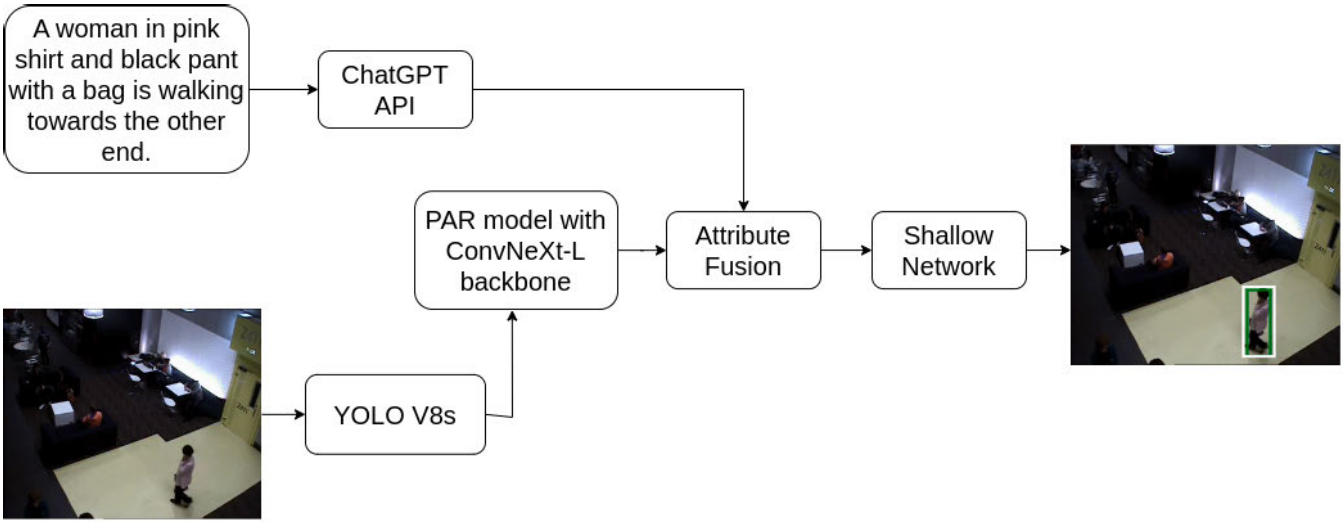


FIGURE 13. End-to-end system with various subsystems for person retrieval. The white bounding box in the frame shows the retrieved bounding box, and the green bounding box shows the ground truth bounding box.

Despite these challenges, the system achieved a TPR of more than 80 % on these frames.

C. FAILURES IN RETRIEVAL

The end-to-end system failed in certain scenarios, as shown in Figure 15. Figure 15a shows that the system fails to detect the target person because the video frame contains multiple persons with similar clothing styles and colours. The target person ranked third position for 95 % of frames. Figure 15b shows the false positive when another person was detected as the target person. One reason for this retrieval is also the noisy images in the training dataset. The TPR for the test participant was 8 %. As shown in Figure 15c, the test subject (in the green bounding box) merged with the background, making it challenging for the PAR model to predict the correct attributes. The TPR in this subject was only 10 %. Figure 15d shows crowded scene in which the target person is surrounded by several people, resulting in an occluded view. The person was correctly predicted in frames without occlusions. However, the overall TPR was only 12 %.

D. END-TO-END SYSTEM RETRIEVAL

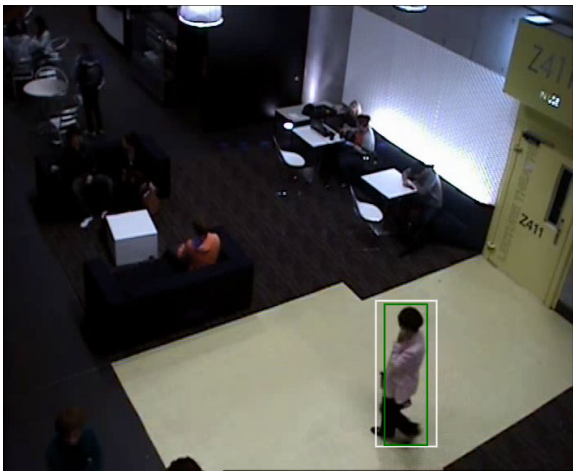
We experimented with several combinations of detectors and PAR models and selected the combination that resulted in TPR $\geq$ 50%. We experimented with two datasets - AVSS+RAP and UPAR. The metrics used for testing the retrieval efficacy were - IOU, IOU $\geq$ 0.4, and TPR. Table 14 summarises a concise overview of the retrieval of the AVSS+RAP dataset using the top two approaches. Notably, the YOLO V8s + ConvNeXt-Large system exhibited superior retrieval performance.

Table 15 lists the retrieval outcomes for the UPAR dataset, employing the top two approaches: YOLO V7 base with ResNet-50 and YOLO V8s with ConvNeXt-Large. Again, YOLO V8s + ConvNeXt-Large demonstrated a superior performance. Considering that the same backbone network was used in both approaches, the difference in results can be attributed to the detector performance. The YOLO v7 generated more false negatives than YOLO v8 models, resulting in a performance close to a random estimate of 50%. The YOLO V8s + ConvNeXt-Large performance for the two datasets was consistent for all metrics, and minor variations were due to the dataset's characteristics. In this case, the retrieval results for AVSS+RAP were better than UPAR; however, it is difficult to conclude and requires more evidence.

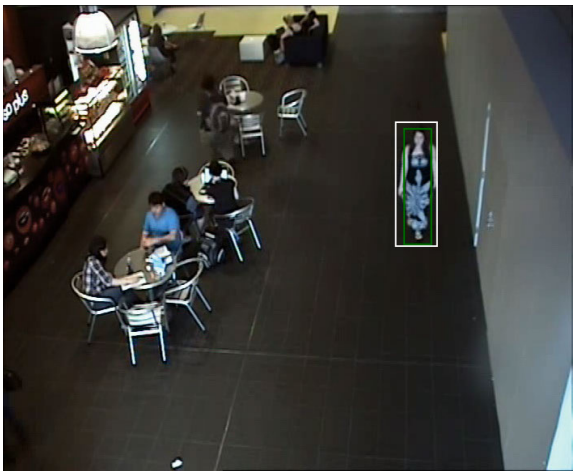
E. INTEGRATION CHALLENGES

The integration challenges encountered during the development of this end-to-end system are as follows:

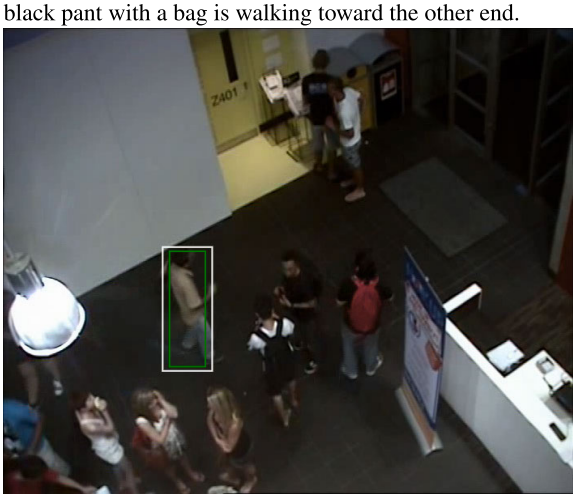
1) **Lack of annotated video datasets:** To the best of our knowledge, there is only one video dataset with DA - AVSS 2018 Challenge-II and no video dataset with both NLD and DA.
2) **Noise propagation across subsystems:** Every subsystem introduces noise at its respective level, potentially contributing to system failure. For instance, as summarised in Table 15, an inferior detector (YOLO V7 Base) will downgrade the performance.
3) **Number of attributes:** The addition of attributes such as age and accessories, which are extremely small spatially in the image, contributes to noise and leads to wrong prediction of attributes. This significantly impacts the model size, its parameters, and performance.
4) **Variations in the image scale:** Individual subsystems operate on specific resolution, but when integrated, they introduce varying image resolutions that can disrupt the end-to-end system's performance. For



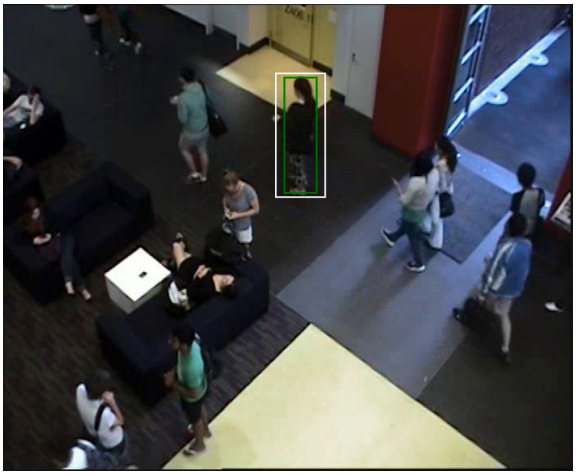
(a) TS:003, F44 (Easy). NLD: A woman in a pink shirt and black pant with a bag is walking toward the other end.



(b) TS:008, F71 (Medium). NLD: She is wearing a black coloured dress. The dress also has white coloured design.



(c) TS:23, F72 (Hard). NLD: A male seeming to be in late 20s, is walking toward his friend. He is wearing a brown short-sleeved shirt and grey jeans. He also has a grey pair of shoes.



(d) TS:31, F80 (Very Hard). NLD: She is wearing a black and white coloured sweater, also she has a black and white patched dress.

FIGURE 14. Successful retrieval of the person with only five attributes across various difficulty levels. The green bounding box in the frame shows the ground truth, and the white bounding box shows the rank-1 prediction.

TABLE 14. Results for person retrieval with PAR model trained on AVSS+RAP dataset.

Approach	Average IOU	IOU ≥ 0.4	True Positive Rate
YOLO V8s + ConvNeXt-L	**0.65**	**0.70**	**68** %
YOLO V7 base + ConvNeXt-L	0.48	0.56	57 %

TABLE 15. Results for person retrieval with PAR model trained on UPAR dataset.

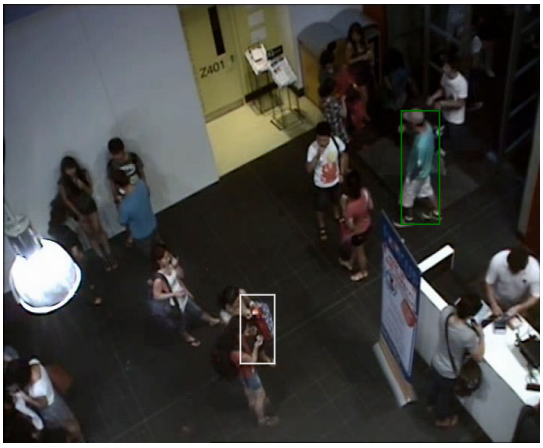
Approach	Average IOU	IOU ≥ 0.4	True Positive Rate
YOLO V8s + ConvNeXt-L	**0.62**	**0.73**	**64** %
YOLO V7 base + ResNet-50	0.45	0.51	55 %

instance, the resolution of AVSS Challenge II video frames is 574 $\times$ 706 pixels. The person detection model works at 640 $\times$ 640, yielding a person frame size of 5 $\times$ 5. The PAR model is trained to work on 224 $\times$ 224 pixels. Thus, the person frame is resized to the aforementioned resolution. These changes in resolution from each subsystem impact the performance.

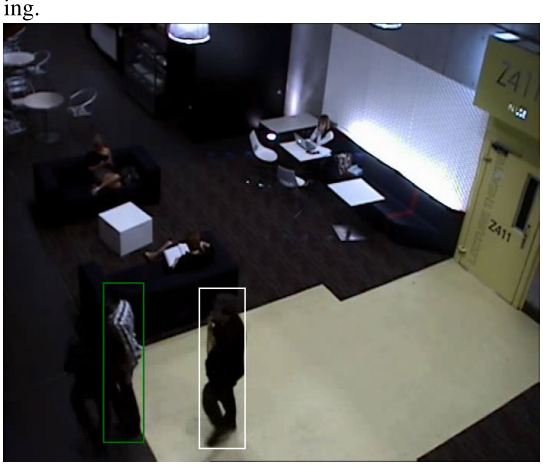
5) **Computational resources:** Detection models YOLO V7 and YOLO V8 integrated with PAR backbone ConvNeXt-Large or ResNet-50 are computationally expensive.



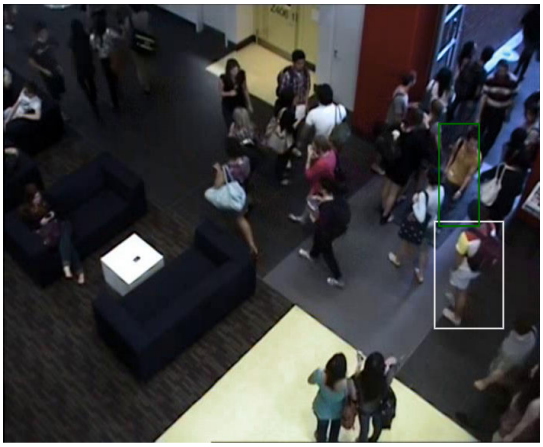
(a) TS:014, F0070. Reason: A person with similar clothing.



(b) TS:026, F71. Reason: Detector problem.



(c) TS:001, F38 Reason: Subject merging with the background.



(d) TS:40, F86. Reason: Occlusion because of crowd.

FIGURE 15. Failure cases of person retrieval. The green bounding box in the frame shows the ground truth, and the white bounding box shows the rank-1 prediction.

VI. CONCLUSION

This paper presented a comprehensive end-to-end algorithm tailored for person retrieval in surveillance scenarios. By harnessing ChatGPT API's capabilities, the algorithm addresses ambiguity and vagueness in sentences, thereby enabling the extraction of attributes from textual descriptions. YOLOv8 detection algorithm was used for person detection. The ConvNeXt backbone was utilised, and a shallow neural network provided effective person ranking, optimising the retrieval process. The retrieval process was robust, and with five attributes, it achieved an average IOU and TPR of $\geq$60%.

This paper provides a detailed methodology for designing a complex person retrieval system. Rather than following a black box approach, it discusses each building block and highlights the bottlenecks and their solutions. Practical guidance will lead to rapid and efficient prototyping using SOTA off-the-shelf tools.

We observed certain limitations in the developed system: it fails to exploit the cross-correlation between NLP and the vision subsystems fully; it is dependent on ChatGPT API for extracting DA; its large memory footprint prevents deployment on the end devices; and retrieval fails if more than three attributes are incorrectly retrieved. In the future, we will focus on adapting the developed framework to function efficiently in edge devices. Additionally, efforts will be made to explore and create a lightweight transformer-based framework as an alternative.

ACKNOWLEDGMENT

The authors would like to thank Harsh Tripathi and Dinesh Nariani for their support in the development of this study.

REFERENCES

[1] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 17–35.

[2] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.

[3] M. Halstead, S. Denman, C. Fookes, Y. Tian, and M. S. Nixon, "Semantic person retrieval in surveillance using soft biometrics: AVSS 2018 challenge II," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[4] M. Halstead, S. Denman, S. Sridharan, and C. Fookes, "Locating people in video from semantic descriptions: A new database and approach," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4501–4506.
[5] S. Denman, M. Halstead, A. Bialkowski, C. Fookes, and S. Sridharan, "Can you describe him for me? A technique for semantic person search in video," in *Proc. Int. Conf. Digit. Image Comput. Techn. Appl. (DICTA)*, Dec. 2012, pp. 1–8.
[6] T. Yaguchi and M. S. Nixon, "Transfer learning based approach for semantic person retrieval," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.
[7] H. Galiyawala, K. Shah, V. Gajjar, and M. S. Raval, "Person retrieval in surveillance video using height, color and gender," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.
[8] H. Galiyawala, M. S. Raval, and S. Dave, "Visual appearance based person retrieval in unconstrained environment videos," *Image Vis. Comput.*, vol. 92, Dec. 2019, Art. no. 103816.
[9] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 54–70.
[10] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, 2020.
[11] S. Aggarwal, R. V. Babu, and A. Chakraborty, "Text-based person search via attribute-aided matching," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2606–2614.
[12] Z. Ding, C. Ding, Z. Shao, and D. Tao, "Semantically self-aligned network for text-to-image part-aware person re-identification," 2021, *arXiv:2107.12666*.
[13] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for Text-to-Image person retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2787–2797.
[14] J. Zhou, B. Huang, W. Fan, Z. Cheng, Z. Zhao, and W. Zhang, "Text-based person search via local-relational-global fine grained alignment," *Knowl.-Based Syst.*, vol. 262, Feb. 2023, Art. no. 110253.
[15] A. Specker, M. Cormier, and J. Beyerer, "UPAR: Unified pedestrian attribute recognition and person retrieval," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 981–990.
[16] A. Specker and J. Beyerer, "Balanced pedestrian attribute recognition for improved attribute-based person retrieval," in *Proc. IEEE 13th Int. Conf. Pattern Recognit. Syst. (ICPRS)*, Jul. 2023, pp. 1–7.
[17] X. Lin, P. Ren, Y. Xiao, X. Chang, and A. Hauptmann, "Person search challenges and solutions: A survey," 2021, *arXiv:2105.01605*.
[18] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*.
[19] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.
[20] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 789–792.
[21] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019.
[22] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[23] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5187–5196.
[24] I. Montani et al., "Explosion/spaCy: v3.6.0: New span finder component and pipelines for Slovenian," Explosion.ai, Berlin, Germany, Tech. Rep. v3.7.2, Jul. 2023, doi: 10.5281/zenodo.8123552.
[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[27] OpenAI. (2022). *ChatGPT—Openai.com*. Accessed: Aug. 4, 2023. [Online]. Available: https://openai.com/chatgpt
[28] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, nos. 2–3, pp. 135–168, May 2000.
[29] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[31] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *Ultralytics YOLOv8*. [Online]. Available: https://github.com/ultralytics/ultralytics
[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.
[33] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3965–3977.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[36] H. J. Galiyawala, M. S. Raval, and A. Laddha, "Person retrieval in surveillance videos using deep soft biometrics," in *Deep Biometrics*. Cham, Switzerland: Springer, 2020, pp. 191–214.



JAY N. CHAUDHARI (Member, IEEE) received the Bachelor of Technology degree in electrical engineering from the Institute of Infrastructure, Technology, Research and Management, Gujarat, India, in 2019, and the Master of Engineering degree in automatic control and robotics from The Maharaja Sayajirao University of Baroda, Gujarat, in 2022. He is currently a Junior Research Fellow of the GUJCOST Project with Ahmedabad University. He holds one Indian patent (published) and has one publication in time series forecasting. His research interests include computer vision and the deployment of deep learning models on edge devices. He is a Young Professional Member of IEEE.



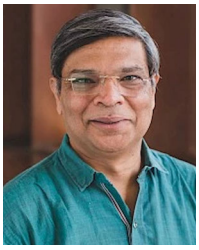
HIREN GALIYAWALA received the bachelor's degree in electronics and communication engineering from the Sarvajanik College of Engineering, Surat, in 2007, the master's degree in technology from the College of Engineering, Pune, in 2010, and the Ph.D. degree in engineering from Ahmedabad University, in 2022. With over 13 years of experience, he has a strong background in computer vision, application development, testing, and engineering education. He has a diverse professional history, having served as a System Engineer with IBM, an Assistant Professor with UTU, and a Senior Research Fellow of the BRNS Project with Ahmedabad University. He is currently a Senior Data Scientist with Rydot Infotech Pvt., Ltd., Ahmedabad. His research published in IEEE, Springer, and Elsevier, is widely cited. He has mentored engineers and students in AI solutions, led teams in developing solutions using deep learning, machine learning, computer vision, and natural language processing, and holds certifications, including ISTQB and IBM Certified Solution Designer.



MINORU KURIBAYASHI (Senior Member, IEEE) received the B.E., M.E., and D.E. degrees from Kobe University, Japan, in 1999, 2001, and 2004, respectively. He was a Research Associate and an Assistant Professor with Kobe University, from 2002 to 2007 and from 2007 to 2015, respectively. He was an Associate Professor with the Graduate School of Natural Science and Technology, Okayama University, from 2015 to 2023. Since 2023, he has been a Professor with the Center for Data-driven Science and Artificial Intelligence, Tohoku University. He has published more than 150 research papers in numerous international journals and conferences. His research interests include multimedia security, digital watermarking, cryptography, and coding theories. He is a TC Member of the IEEE SPS Information Forensics and Security. He received the Young Professionals Award from the IEEE Kansai Section in 2014 and the Best Paper Award from IWDW in 2015 and 2019. He is the Chair of the APSIPA TC of Multimedia Security and Forensics. He has been an Associate Editor of IEEE SIGNAL PROCESSING LETTERS, since 2022. He was an Associate Editor of *Journal of Information Security and Applications* (JISA), from 2014 to 2021.



PAAWAN SHARMA (Senior Member, IEEE) was born in Bikaner, Rajasthan, in January 1983. He completed his formal education in Ajmer, Rajasthan. He received the B.E. degree in ECE from the University of Rajasthan, Jaipur, in 2005, the M.Tech. degree in communication systems from SVNIT, Surat, in 2008, and the Ph.D. degree in engineering from the Homi Bhabha National Institute, Mumbai, in 2014. He was an Associate Professor with the Department of ICT, PDEU, and acts as the Dean of the IT Infrastructure, PDEU, Gandhinagar. He was with Wipro Technologies in the VLSI domain. He has guided/co-guided four Ph.D. scholars who have focused on computer vision, disaster management technology, and smart grid analysis. He has four Indian patents (published) and over 40 publications. His research interests include multidisciplinary, spanning applications/solution development in signal processing, embedded systems, pattern recognition, machine vision, and artificial intelligence. He is a Senior Member of ACM, ISSIA, and IAPR.



MEHUL S. RAVAL (Senior Member, IEEE) received the bachelor's degree in electronics and telecommunication engineering (ECE) from the College of Engineering, Pune, India, in 1996, the master's degree in electronics-digital systems, in 2002, and the Ph.D. degree in electronics and telecommunication engineering (ECE), in 2008. His commitment to academic excellence transcends borders, with notable research stints with Okayama University, Japan, and as an Argosy Visiting Associate Professor with the Olin College of Engineering, USA, in 2016. He further enriched his global academic footprint with a visiting professorship with Sacred Heart University, Connecticut, in 2019. He is currently a Distinguished Faculty Member with Ahmedabad University. He is also the Associate Dean of the Experiential Learning and a Professor, with a career spanning over 25 years, marked by expertise in computer vision and engineering education. His scholarly contributions extend to esteemed publications and respected reviewers for publishers, such as IEEE, ACM, Springer, Elsevier, IET, and SPIE. His research has garnered support from institutions, such as the Board of Research in Nuclear Science (BRNS) and the Department of Science and Technology of the Government of India. He actively mentors students and contributes to curriculum development in leading Indian universities. He holds prestigious titles as a fellow of the Institution of Electronics and Telecommunication Engineers (IETE) and the Institution of Engineers, India. His substantial contributions extend to leadership roles within the IEEE Gujarat section, including the IEEE Signal Processing Society (SPS) and the IEEE Computational Intelligence Society chapters.