https://doi.org/10.1007/s42979-024-02985-0

ORIGINAL RESEARCH

Visual Question Answering Models for Zero-Shot Pedestrian Attribute Recognition: A Comparative Study

Modesto Castrillón-Santana[1] · Elena Sánchez-Nielsen[2] · David Freire-Obregón[1] · Oliverio J. Santana[1] · Daniel Hernández-Sosa[1] · Javier Lorenzo-Navarro[1]

Received: 19 February 2024 / Accepted: 20 May 2024

Abstract

Pedestrian Attribute Recognition (PAR) poses a significant challenge in developing automatic systems that enhance visual surveillance and human interaction. In this study, we investigate using Visual Question Answering (VQA) models to address the zero-shot PAR problem. Inspired by the impressive results achieved by a zero-shot VQA strategy during the PAR Contest at the 20th International Conference on Computer Analysis of Images and Patterns in 2023, we conducted a comparative study across three state-of-the-art VQA models, two of them based on BLIP-2 and the third one based on the Plug-and-Play VQA framework. Our analysis focuses on performance, robustness, contextual question handling, processing time, and classification errors. Our findings demonstrate that both BLIP-2-based models are better suited for PAR, with nuances related to the adopted frozen Large Language Model. Specifically, the Open Pre-trained Transformers based model performs well in benchmark color estimation tasks, while FLANT5XL provides better results for the considered binary tasks. In summary, zero-shot PAR based on VQA models offers highly competitive results, with the advantage of avoiding training costs associated with multipurpose classifiers.

Keywords Pedestrian attribute recognition · Biometrics · Vision language models · Visual question answering

Introduction

Soft biometrics encompass various human attributes such as gender, age, clothing, accessories, or hairstyles. These characteristics alone do not definitively distinguish an individual. However, they can complement the information provided by biometric traits to enhance recognition performance [1], or their combined use may be robust enough for face verification and search in close-up facial images [2]. In surveillance scenarios, where the use of faces as unique biometric cue is challenging for state-of-the-art face recognizers, due to low resolution, unrestricted pose, and frequent occlusion [3, 4], leveraging such ancillary information would be highly beneficial.

Pedestrian Attribute Recognition (PAR) has often been adopted to search or retrieve a person in real-world surveillance images or video footage [5]. PAR leverages advanced computer vision, pattern recognition, and machine learning techniques to automate attribute recognition, enabling automatic systems to manage complex situations and enhance public safety, security, and human interaction. Additionally, PAR provides interpretable information for human observers,

Elena Sánchez-Nielsen, David Freire-Obregón, Oliverio J. Santana, Daniel Hernández-Sosa and Javier Lorenzo-Navarro contributed equally to this work.

✉ Modesto Castrillón-Santana
modesto.castrillon@ulpgc.es

Elena Sánchez-Nielsen
enielsen@ull.edu.es

David Freire-Obregón
david.freire@ulpgc.es

Oliverio J. Santana
oliverio.santana@ulpgc.es

Daniel Hernández-Sosa
daniel.hernandez@ulpgc.es

Javier Lorenzo-Navarro
javier.lorenzo@ulpgc.es

[1] SIANI, Universidad de Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain

[2] Universidad de La Laguna, 38200 San Cristóbal de La Laguna, Spain

Published online: 28 June 2024

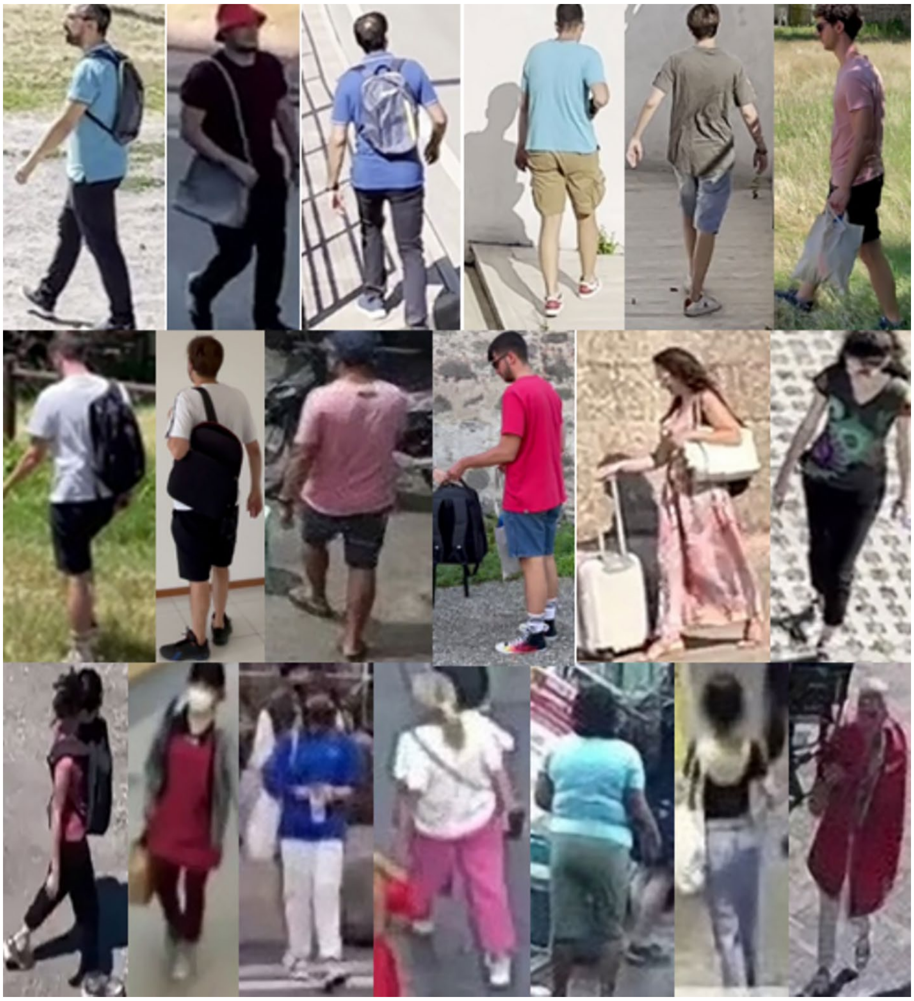


Fig. 1 A short excerpt of MIVIA dataset samples (image from [8])

as those attributes offer comprehensive semantic details that facilitate the creation of human categories.

PAR is an active research field, evidenced by the growing number of scholarly papers on this topic and the active engagement in recent international competitions [6–8]. This surge of interest underscores the broad range of PAR applications, including human–machine interaction, retail analytics, smart cities, and surveillance scenarios. However, several challenges persist. Ethical and privacy considerations play a crucial role in dataset preparation, and a significant domain gap often exists between test and real-world deployment scenarios [9].

In recent years, zero-shot approaches have become increasingly prevalent in the literature [9, 10]. In this sense, successfully applying zero-shot strategies could shift the established paradigm. A zero-shot Visual Question Answering (VQA) approach has recently demonstrated remarkable competitiveness in the PAR problem [11]. That approach leveraged a pre-trained BLIP-2 for a VQA benchmark [12], outperforming all competitors in the PAR Contest-CAIP23 test set [8]. This achievement was attained without relying on the provided training data, as no training or fine-tuning was necessary. The results indicated that the VQA-based proposal consistently delivered robust performance across diverse datasets, spanning the five proposed benchmark tasks.

In our current study, we expand upon the preliminary evaluation covered in our earlier work [11], assessing a single VQA model. This time, we consider up to three distinct VQA models for the PAR Contest-CAIP23 benchmark. Our evaluation now encompasses results from the fully annotated validation and partially annotated training sets. The contributions of our research are threefold:

- Extensive Model Evaluation: we rigorously evaluate multiple VQA models on the PAR Contest-CAIP23 benchmark, comprehensively comparing their performance.
- Model Selection: we determine the most suitable VQA model for each specific task within the benchmark,

ensuring optimal performance across various attribute recognition scenarios.
- Visual Error Analysis: unlike our previous work [11], we delve into a detailed visual analysis of classification errors, shedding light on areas for improvement and potential refinements.

PAR Contest-CAIP23 Benchmark

As mentioned, this paper extends the proposal presented in [11], significantly influencing the final ranking at the PAR Contest-CAIP23 benchmark [8]. For this competition, the organizers curated the MIVIA dataset comprising 105,244 pedestrian images, divided into training and validation sets. Specifically, the training set contains 93,082 samples, while the validation set comprises 12,162 samples, see Fig. 1. Those samples were collected from existing datasets, including PETA [13] and RAP v2.0 [14]. However, most of the dataset consists of private samples meticulously collected and annotated by the organizers.

The validation set contains fully annotated samples, while the training set includes partially annotated samples. Summarizing, each sample may have up to five annotated features, with numeric labels denoting these attributes. A negative label indicates a non-annotated feature. The different features annotated are the following:

- Upper clothes: Each sample in the dataset is annotated with a single label corresponding to the dominant color present in the upper part of the body's clothing. The annotations include eleven possible colors, each associated with a numeric label (in brackets): black (1), blue (2), brown (3), gray (4), green (5), orange (6), pink (7), purple (8), red (9), white (10), and yellow (11). Colors beyond this set are not considered in the dataset annotations, nor are color combinations, e.g. *red and black*.
- Lower clothes: Following the same convention as for the upper clothes, a single label corresponds to the main color present in the clothes in the lower part of the body.
- Gender: Gender of the foreground person, considering two labels, i.e., male (0) and female (1).
- Bag: Indicates the presence or absence of a bag accessory considering two labels, i.e., absence (0) and presence (1).
- Hat: Indicates the presence or absence of a hat accessory considering two labels, i.e., absence (0) and presence (1).

See Fig. 2 for an illustrative example of the ground truth annotation provided in the MIVIA dataset. More details about the sample distribution of this dataset can be found in the paper describing the PAR Contest-CAIP23 benchmark [8].

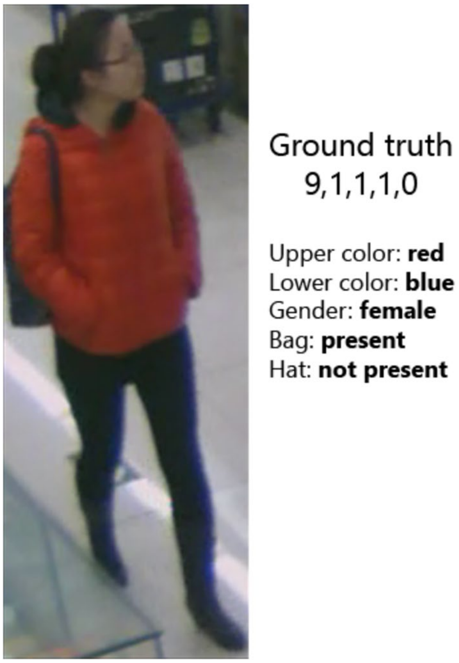


Fig. 2 Example of annotated image from the MIVIA validation set

Overall Description

Artificial Intelligence (AI) technologies and intensive learning applications have led to significant advancements in PAR in recent years. The availability of large annotated datasets has facilitated the training of deep neural networks for solving the PAR problem.

Interestingly, our approach diverges from the predominant trend in PAR literature [7], which focuses on designing architectures and training them using annotated data. Instead, we take a different path: we explore zero-shot performance by leveraging existing pre-trained VQA models for PAR. By doing so, we avoid the traditional training step with available PAR datasets, opening up new possibilities for efficient and effective attribute recognition.

VQA models are based on the rise of Large Language Models (LLMs). These LLMs are AI systems pre-trained on vast amounts of text data, showcasing remarkable language understanding, speech generation competence, and the ability to handle multi-domain tasks without fine-tuning. Traditionally, LLMs have been trained unimodally, focusing solely on language. However, VQA offers a meeting point for vision-language tasks [15, 16]. Within computer vision tasks, we highlight the work by Radford et al. [17], who fine-tuned a vision encoder with an LLM, aligning visual and linguistic representation spaces. As a result, this model accurately performs VQA tasks related to different object classification, action recognition, or OCR tasks, achieving competitive results with zero-shot transfer. Remarkably, it

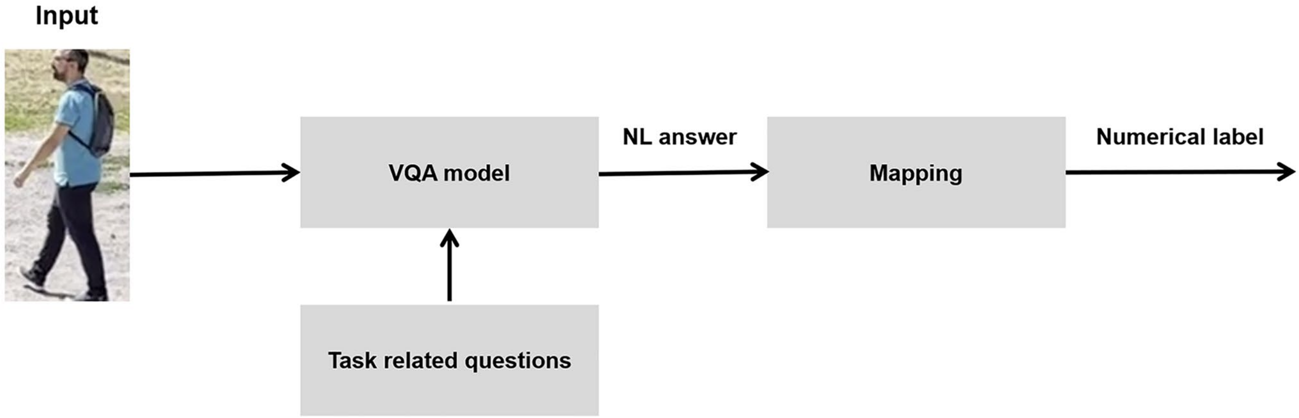


Fig. 3 The processing pipeline takes an input image that, according to the specific task in hands, poses one or more questions to the evaluated VQA model, which provides a short answer in natural language (NL) that is later mapped to a valid numerical label included in the annotation

outperforms a supervised ResNet-50 trained on millions of domain-specific samples, significantly reducing computational costs. Thanks to the generalization capabilities of LLMs, their model can answer questions related to visual information without requiring domain-specific training. The authors concluded that deep models trained on large datasets containing millions of image-text pairs can effectively associate visual concepts with their corresponding textual descriptions. More recently, a pre-trained Vision Language Model (VLM) played a pivotal role in developing the WISE Image Search Engine (WISE) [18]. Inspired by the work by Radford et al. [17], Sridhar et al. applied a pre-trained OpenCLIP VLM, followed by a nearest neighbor search in the resulting high-dimensional feature space. This innovative approach enables content-based image search, yielding relevant results.

In our proposed pipeline, see Fig. 3, we evaluate the efficacy of various pre-trained models without the need for additional training on the datasets provided by the contest organizers. Our primary goal is to assess the potential of using pre-trained VQA models for PAR. Initially, our plan for the PAR Contest-CAIP23 was to study a range of various VQA models. However, due to contest deadlines, we evaluated a single BLIP-2 model [12] for this purpose. Following the impressive results of the zero-shot transfer strategy in the PAR Contest-CAIP23, we have expanded our investigation. In the present work, we adopt and evaluate three different VQA models. Our assessment covers a broader spectrum, exploring additional VQA techniques and their adaptation to the specific challenges posed by PAR across various proposed tasks. In this sense, we delve into zero-shot VQA methods [19], eliminating the need for ground-truth question-answer annotations.

Recently, VQA has garnered significant attention from the research community after competitive results were reported for different computer vision tasks. VQA is a powerful intersection between computer vision and natural language processing [16]. Unlike image captioning, where image semantic information is extracted and expressed for human understanding, adopting a VQA strategy, the information is extracted according to the observer's needs, who may pose a question to obtain targeted information or even create an adaptive interaction tree based on previous queries. In a VQA system, the information within an image is compared with a set of questions expressed in natural language. Among the diverse applications identified by Barra et al. [16] for VQA, surveillance and biometrics are valid real-world scenarios [20]. In our approach, we meticulously design clear and direct questions tailored to the benchmark

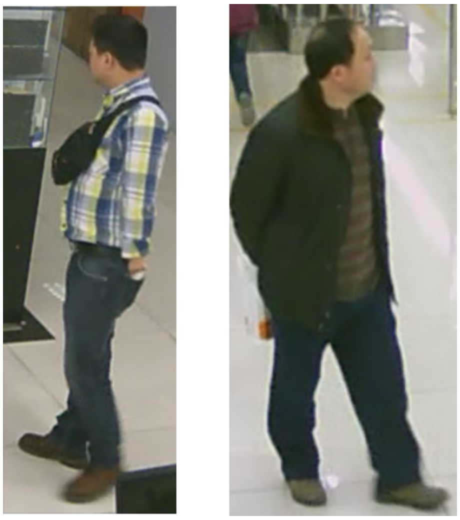


Fig. 4 MIVIA validation set samples with (left) upper clothes color annotated as a single color, triggering a combined color answer by the BLIP-2 OPT model, i.e., "blue and white", and (right) upper clothes color with different jacket and shirt color. The annotated label refers to the jacket color

tasks, ensuring their feasibility based on evaluation against the contest validation set.

As mentioned, we conducted our assessment and analysis with three different VQA models:

- OPT: a pre-trained BLIP-2 language model [12] that adopts a pre-training strategy based on off-the-shelf frozen pre-trained image encoders and a frozen Open Pre-trained Transformers (OPT) language model [21], who was trained on a large-scale corpus of text data and fine-tuned for VQA with the Visual Transformer (ViT) base backbone [22]. This was the model adopted in [11].
- FLANT5XL: a pre-trained BLIP-2 language model [12] that uses Flan-T5 model [23] as frozen language model. Flan-T5 has achieved strong few-shot performance even compared with a much larger model. Flan-T5 XXL leads the SOTA on the VQAv2 benchmark, followed by Flan-T5 XL. Given our hardware possibilities, the latter is the one evaluated in the paper.
- PNP: a pre-trained plug-and-play (PNP) model [24] that proposes a modular framework for zero-shot VQA that does not require training the pre-trained language models for the specific vision task. This provides context by previous image captioning, gluing pre-trained language and vision models together. PNP-VQA is currently the second LLM in the SOTA on the VQAv2 benchmark.

Contextual Queries

We have adopted the VQA strategy to contextualize the expected answers to solve the five contest tasks mentioned above: upper color, lower color, gender, bag, and hat. While image captioning outputs may be suitable for general descriptive tasks, they need to address the specific subtasks posed in the contest.

Given the contest subtasks, and after manual observation of the answers obtained with the validation set for different possible questions for each subtask (see below), the chosen questions for each subtask are the following:

- **T1**. Upper color
  - What color is the person's shirt?
  - Does the person wear a jacket?
  - What color is the person's jacket?
- **T2**. Lower color
  - What color is the person's trousers?
- **T3**. Gender
  - Is the person male or female?
- **T4**. Bag presence
  - Does the person wear a bag?
- **T5**. Hat presence
  - Does the person wear a hat?
  - Does the person wear a cap?

As previously mentioned, those questions were carefully crafted after a prompt engineering process during VQA interactions with some validation set samples. VQA is well known for providing short answers in natural language. However, each short answer generated by a VQA model must later be *mapped* to the numerical labels expected by the contest organizers. The mapping is straightforward for the binary tasks (i.e., T3, T4, and T5) when considered in their context. For instance:

Table 1 PAR Contest-CAIP23 validation set accuracy/precision/recall/F1-score results

Task	OPT	PNP	FLANT5XL
T1	**0.804/0.804/0.804/0.801**	0.091/0.298/0.092/0.116	0.602/0.820/0.602/0.620
T2	**0.842/0.844/0.842/0.839**	0.093/0.468/0.093/0.145	0.493/0.587/0.493/0.522
T3	0.909/0.915/0.754/0.827	0.287/0.287/1.000/0.446	**0.958/0.959/0.891/0.924**
T4	0.494/0.269/**0.980**/0.422	0.349/0.218/0.946/0.354	**0.885/0.748**/0.595/**0.661**
T5	0.566/0.181/**0.987**/0.307	0.128/0.099/0.992/0.181	**0.923/0.688**/0.374/**0.484**

There are 12,162 annotated samples. In bold, the highest value obtained for each metric and task

Table 2 PAR Contest-CAIP23 training set accuracy/precision/recall/F1-score results

Task	#	OPT	PNP	FLANT5
T1	35,846	**0.815/0.813/0.815/0.811**	0.090/0.233/0.090/0.112	0.572/0.807/0.572/0.596
T2	60,758	**0.832/0.832/0.832/0.828**	0.092/0.440/0.092/0.141	0.431/0.569/0.431/0.451
T3	85,411	0.914/0.859/0.824/0.841	0.277/0.277/1.000/0.434	**0.966/0.944/0.933/0.938**
T4	65,683	0.517/0.242/**0.946**/0.385	0.366/0.197/0.958/0.326	**0.888/0.685**/0.565/**0.614**
T5	78,270	0.555/0.215/**0.984**/0.352	0.173/0.129/0.996/0.229	**0.954/0.859**/0.746/**0.799**

There are 93,081 partially annotated samples. In bold, the highest value obtained for each metric and task

- The answer to the question posed for T3 directly corresponds to the gender subtask (e.g., female or male).
- Similarly, a positive response (i.e., not containing “no”) to either of the questions for tasks T4 and T5 indicates the presence of the related accessory.

Tasks T1 and T2 exhibit slightly different behavior. While the annotation labels encompass eleven distinct color values, VQA models cannot provide answers solely from this predefined set. VQA models may offer other colors as answers and even multi-color responses are possible. We engaged in interactions with the validation set to handle colors not explicitly covered by the eleven annotation labels. During this process, we encountered a short collection of answers that fell outside the original eleven labels. These additional colors included grey, khaki, tan, and teal. We decided to map grey and tan to gray, khaki to yellow, and teal to black.

Of course, the VQA model may unexpectedly provide other colors. In cases where the model’s response is not among the considered situations, we assign a random color from the original set of eleven labels. Furthermore, when the VQA model’s answer comprises a combination of colors (e.g., “blue and white,” as shown in Fig. 4, left), we prioritize the first color from the list of eleven possible labels or their mapped equivalents.

After considering these aspects, T2 is resolved with a single question, while T1 requires up to three questions. This is because, when interacting with the validation set, we observed that the answer to the first question often did not match the annotated label when the individual wore a jacket or similar garment. This discrepancy arises because the VQA model may provide the color of the shirt. In contrast, the annotated color refers to the jacket or similar clothing

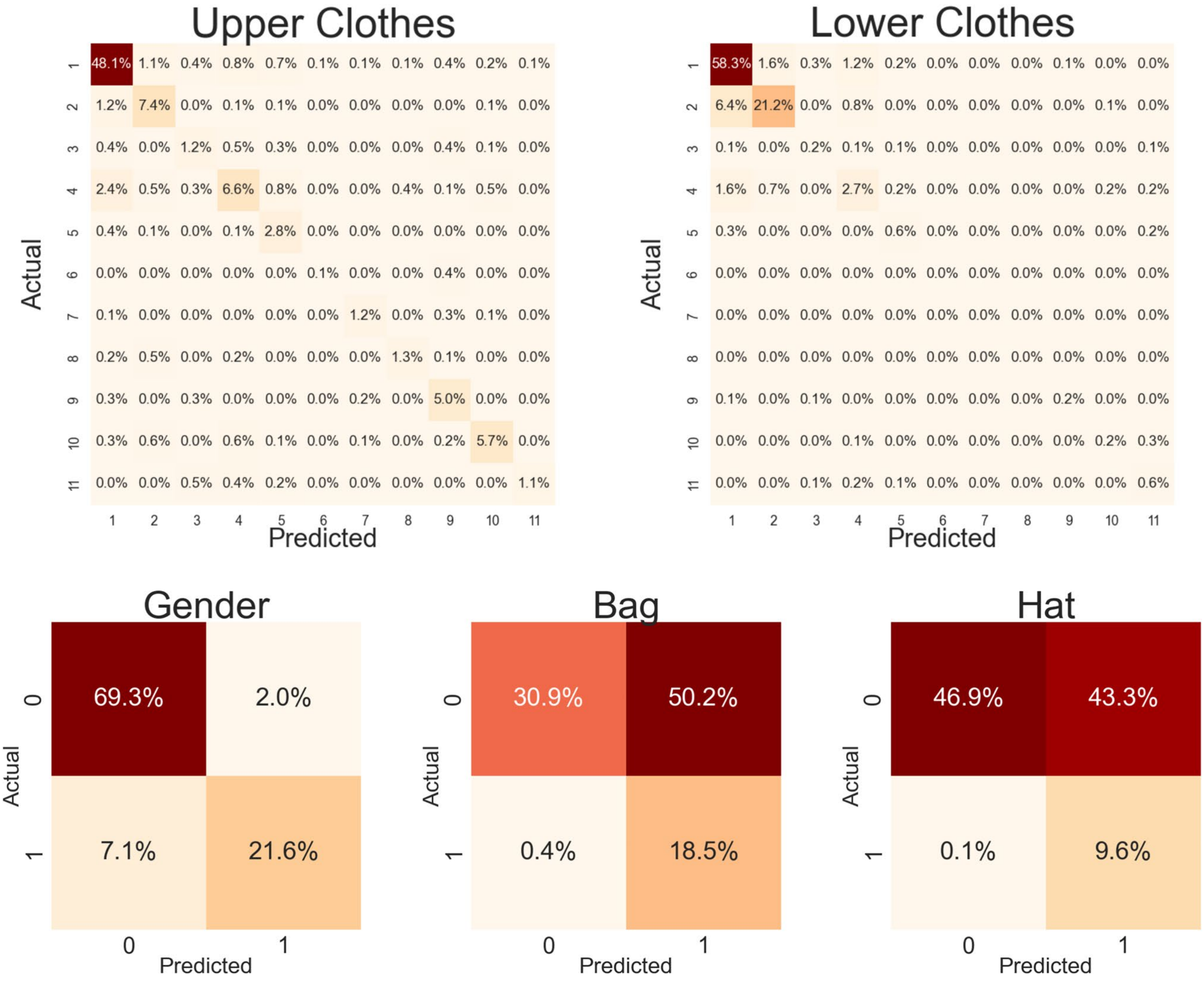


Fig. 5 MIVIA validation set confusion matrices for OPT. For T4 and T5, there is an evident large number of FPs. For T3, it is observed a bias for the female class. The meaning of the numerical labels associated for each task are defined in Sect. “PAR Contest-CAIP23 Benchmark”

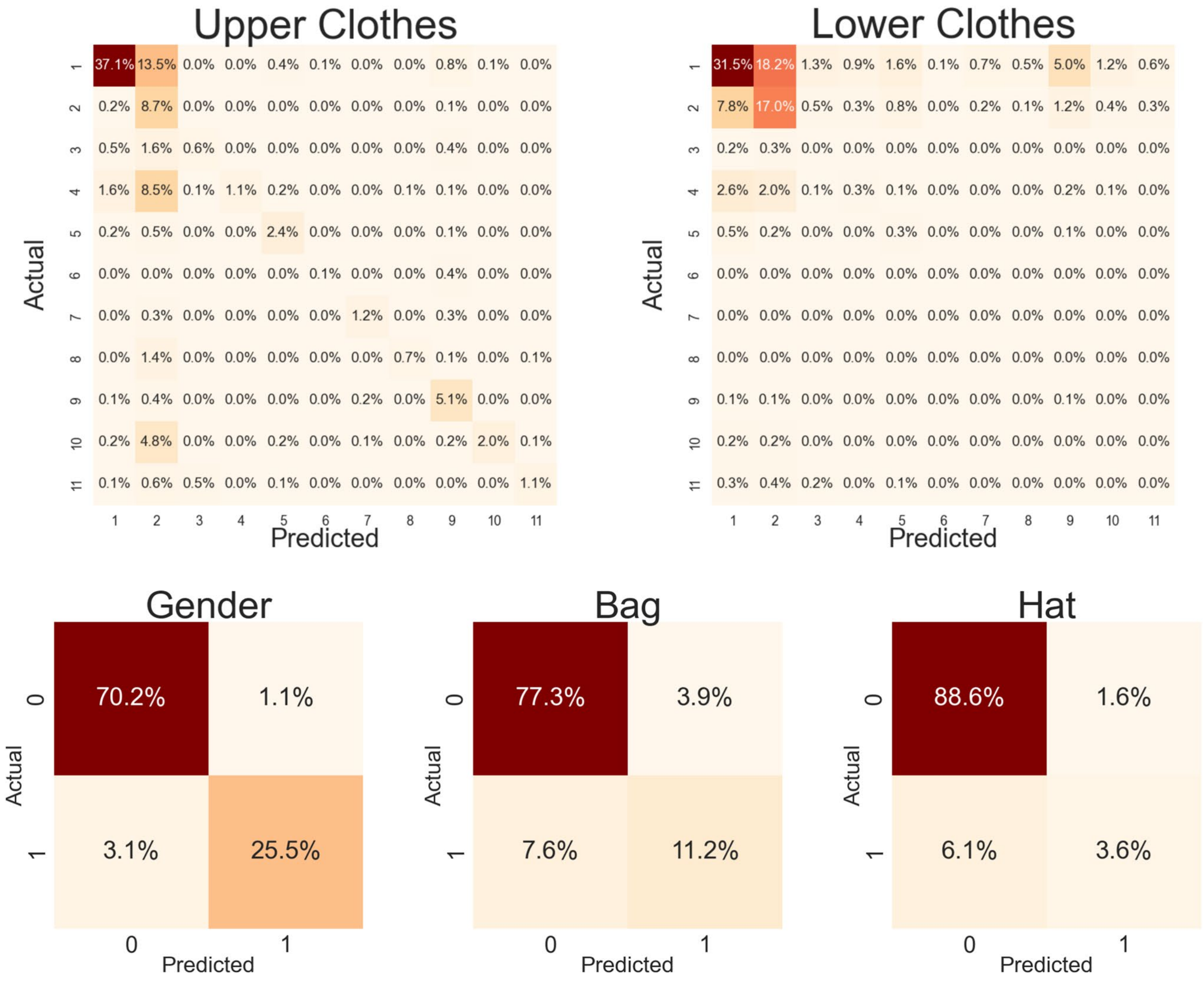


Fig. 6 MIVIA validation set confusion matrices for FLANT5XL. For T1 and T5, it is observed a larger sparse distribution of the predicted colors. The meaning of the numerical labels associated for each task are defined in Sect. "PAR Contest-CAIP23 Benchmark"

item when it is present, as illustrated in the relevant sample depicted in Fig. 4 right:

```
if person wears a jacket then
   color of upper body clothes = jacket color
else
   color of upper body clothes = shirt color
endif
```

Results

This section summarizes the results achieved in both the public datasets made available to the PAR Contest-CAIP23 participants, i.e., training and validation datasets, discussing the performance of the evaluated models. In addition, the last subsection presents some examples of classification errors obtained for each task leader model.

Training and Validation Sets

As mentioned earlier, the combined training and validation sets comprise more than 100,000 samples. However, only the validation samples are fully annotated, while the training images may lack none or any annotated labels.

The results for the different VQA models evaluated on the training and validation sets are summarized in Tables 1 and 2. We utilized scikit-learn (sklearn) to compute accuracy, precision, recall, and F1-score metrics. Given the class imbalance in multiple classification problems, we used the weighted average. It is important to note that the number of columns differs between the two tables because training samples may not be completely annotated. Therefore, we have also included the total number of samples for each task to compute the corresponding metrics.

Moreover, we decided not to combine the results in a single table. Additionally, as the validation set guided the question selection process, there might be a bias in the combined results. Considering that the training set contains significantly more samples, we have included them to better illustrate the generalization observed in the evaluated VQA models.

A first observation of the best-performing values achieved for each task and set suggests similar behavior, with slightly better results observed for the training set in three tasks: T1, T3, and T5. However, the PnP-VQA model is the least suitable for the PAR tasks, as it significantly underperformed against the chosen benchmark. A closer examination reveals distinct behavior across the five tasks of both BLIP-2 models. Interestingly, there is no winner, as the OPT model excels in the color estimation tasks (T1 and T2), while the FLANT5XL model performs better for the binary tasks (T3–T5). An exception to this trend is the recall metric, which favors OPT for tasks T4 and T5 when used as LLM. As we will show later in the confusion matrices, OPT reports a more significant number of positive detections of bags and hats but also increases the number of false positives, i.e., decreasing precision. Subsequently, let us summarize the findings:

- Color estimation tasks, which involve considering eleven possible annotated colors, yield promising results using the OPT model, with an F1-score exceeding 0.8 for both sets and tasks.
- T3, related to gender classification (treated as a binary problem), achieves the highest F1-score, surpassing 0.92 for both datasets using the FLANT5XL model.
- The last two tasks, T4 and T5, exhibit the lowest F1-score among the five tasks. However, FLANT5XL achieves the highest F1-score, with T4 hovering around 0.6 for both datasets. Notably, T5 shows a more pronounced difference: the F1-score for the training dataset (which has a larger number of samples) is significantly higher.

Table 3 Average processing time in seconds after analyzing 1000 samples in a NVIDIA A40 GPU

Question	OPT	PNP	FLANT5XL
Does the person wear a jacket?	0.0149	0.0993	0.0518
What color is the person jacket?	0.0131	0.0984	0.0265
What color is the person shirt?	0.0130	0.0982	0.0230
What color is the person trousers?	0.0130	0.0978	0.0270
Is the person male or female?	0.0156	0.0989	0.0317
Does the person wear a bag?	0.0127	0.0978	0.0539
Does the person wear a hat?	0.0150	0.0982	0.0543
Does the person wear a cap?	0.0148	0.0986	0.0545

After analyzing the reported metrics, the reader is invited to explore the corresponding confusion matrices for the models that achieved the best performance in at least one subtask: OPT and FLANT5XL. Figures 5 and 6 present the matrices for all five subtasks for both models, considering the validation set. We have specifically chosen the validation set because the number of annotated samples per task is identical.

The first two matrices in both figures correspond to the color-related tasks (T1 and T2). To facilitate their understanding, the reader can refer to "PAR Contest-CAIP23 Benchmark" for the meaning of the eleven labels, which represent a single color alphabetically sorted (i.e., black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow). Notably, there is a higher concentration of annotated colors for the lower body; black and blue dominate, followed at a distance by gray. The upper body has a more significant presence of those three colors, with black being the most prevalent.

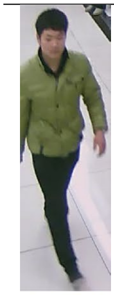
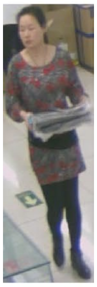
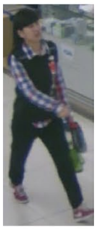
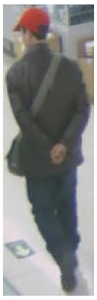
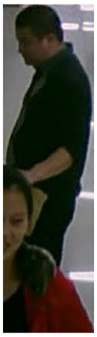
The OPT model's diagonal corresponding to upper body colors is relatively clean, except for label 6 (orange), which is mainly classified as label 9 (red). Additionally, labels 3 and 7 (respectively brown and pink) are occasionally classified as red, which may be due to the acquisition conditions, creating an illusion even for humans [25]. The most common label for upper body clothes, black, is primarily confused by blue, gray, and green. Similarly, blue is mistaken for black, gray is confused with black, and pink is occasionally classified as red. These confusions appear reasonable, given the limited number of annotated labels.

For the lower clothes color estimation, the confusion matrix focuses more on the first four colors: black, blue, brown, and gray. Only green and white appear infrequently. The reduced number of possible colors limits the range of classification errors. This concentration of colors likely contributes to the higher metrics obtained for T2 using the OPT model in both datasets.

The observation of both color estimation matrices for FLANT5XL, see Fig. 6, exhibits a worse behavior. For upper colors, FLANT5XL in T1, black, blue, and red attract most prediction errors. For lower colors, the most prevalent labels (black and blue) are frequently misinterpreted across the entire range of possible labels. To summarize, for T1 and T2, the OPT models provide very competitive results. In the following subsection, after discussing the results for other tasks, we will discuss some classification errors.

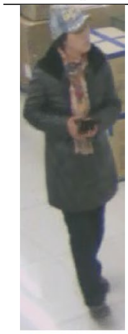
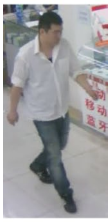
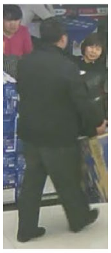
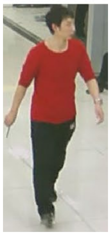
Focusing on the three remaining subtasks, observing the matrices in the lower part of Figs. 5 and 6 reveals differences for each task and model. For the gender classification

Table 4 Excerpt of OPT classification error examples for T1 and T2 subtasks

Sample	Task	Annotation	VQA Answer	Observations
	T1	black	green	Annotation error
	T1	black	red and gray	Annotation error
	T1	black	blue and white	What color is the person shirt? blue and white Does the person wear a jacket? no. What color is the person jacket? black
	T2	black	blue	Annotation error?
	T2	black	red	Color taken from occluding person

task, where most samples are annotated as males, the OPT model exhibits a larger error in classifying females, suggesting a bias in the model. This behavior was already discussed in [11] for the validation set. While there is an inherent imbalance between the number of males and females in both training and validation sets, a closer observation suggests that the model exhibits a bias, misclassifying female samples more frequently. However, this behavior is significantly reduced in the FLANT5XL model, even though there

Table 5 Excerpt of FLANT5XL classification error examples for T3 subtask

Sample	Task	Annotation	VQA Answer	Observations
	T3	female	male	Misclassified
	T3	female	male	Annotation error?
	T3	male	female	A female face is visible in the background
	T3	male	female	Misclassified

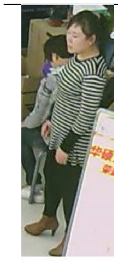
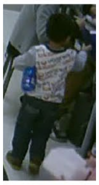
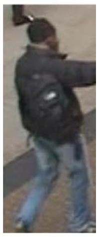
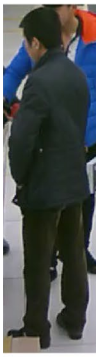
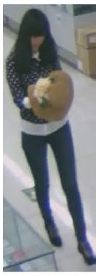
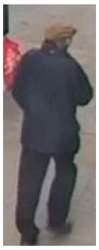
is still an imbalance in the number of classification errors for females, despite their lower presence in both sets, as mentioned above.

For the other two tasks, the presence of the accessories needs to be balanced in the validation set. Therefore, observing how classification errors are distributed is crucial to assess the model's robustness. As described in the same study [11], the OPT model can detect the presence of hats and bags precisely, but at the cost of producing many false positives. In contrast, the FLANT5XL model reduces the number of false positives and decreases the number of true positives, particularly when detecting the presence of hats. A final observation regarding both accessory tasks is that those items are often located near the image boundaries. Consequently, any of the evaluated VQA models may lack sufficient contextual information to determine the presence of those elements accurately. Unfortunately, we do not have the original images to explore this possibility further. Additionally, adding image context could introduce distractors, such as other individuals. In the next section, we present some examples of classification errors, evidencing that circumstance.

Before concluding this section, we will first delve into the processing costs of the evaluated VQA models. Table 3 presents the average processing time for each question using an NVIDIA A40 GPU. OPT demonstrates the highest speed among the models, while PNP exhibits the slowest performance. FLANT5XL falls in between these extremes. Notably, the processing time varies based on the type of question. Specifically, color-related questions are handled most swiftly by the OPT model (except queries related to wearing a bag) and FLANT5XL.

In summary, OPT is better suited for solving color-related questions and significantly faster than any other evaluated model. On the other hand, FLANT5XL is up to four times slower but provides better results for binary tasks. The achievements in color estimation are particularly impressive, considering that this is not a binary classification problem;

Table 6 Excerpt of FLANT5XL classification error examples for T4 and T5 subtasks

Sample	Task	Annotation	VQA Answer	Observations
	T4	no	yes, a bag	Confused by the background?
	T4	no	yes, he does	Carrying a bottle?
	T4	yes	no, he does not	Misclassified, should a question related to backpacks be included?
	T5	no	yes, he does	Confused by the background?
	T5	no	yes	She has a hat in ther hands
	T5	yes	no	Misclassified

it involves eleven possible classes. As demonstrated below in the following subsection, some misclassification results reveal that human observers' perception of color in images does not always align perfectly with the provided annotations, especially in scenarios involving multiple individuals within the same image.

Classification Errors

In this subsection, we aim to illustrate the classification errors achieved by the best VQA model for each subtask. Consequently, we direct our attention to errors reported by the OPT model for T1 and T2, as well as errors reported by the FLANT5XL model for T3, T4, and T5.

Table 4 presents a concise collection of samples incorrectly classified according to the annotated labels for T1 or T2. Among the T1 errors, we deliberately chose some perceptually strong misclassifications, as confusing black with (dark) blue may provide less informative insights. The first two reported classification errors are likely due to annotation errors, as the VQA model provides more accurate labeling. However, the reason for the third error example may lie in the posed questions. Precisely, the annotated color corresponds to what appears to be the woman's vest. Interestingly, even when no jacket presence is reported by the VQA model when asked about the color of the jacket, the answer matches the vest color.

For T2, most classification errors are related to colors that humans can easily confuse. A noteworthy comment pertains to the last example included in the table. The cropped image contains two individuals, one partially visible that occludes the lower part of the person of interest. The VQA model's answer takes the color from the occluding person, resulting in a misclassification.

As previously mentioned, for T3, we exclusively present examples of misclassifications using the FLANT5XL model. A short collection of instances are displayed in Table 5. The first and last examples appear as misclassifications, whereas the second example seems to be a clear annotation error. In the third case, a woman looking directly at the camera in the background could influence the VQA model's answer.

Finally, let us investigate some classification errors for tasks T4 and T5, see Table 6. For T4, the first example features a cluttered background that could confuse the VQA model. In the second example, the person carries something (possibly a bottle), likely triggering a positive model answer. In the last T4 example, the adopted question was about the presence of a bag, not specifically a backpack, which is indeed present in the sample.

Regarding T5, the first example again contains a cluttered background, cropped in the image, which might impact the VQA model's answer. The second example is intriguing: the person wears a hat on their hand but not on their head. Consequently, the annotation is erroneous, yet the model answered positively. Lastly, the final example depicts a back view, which appears to be a clear misclassification (Table 6).

Conclusions

In recent years, deep learning has revolutionized computer vision models. However, these models often require a large number of annotated samples and complex architecture layers. While this has improved performance, it has also sacrificed some level of explainability for human understanding.

Integrating other modalities, such as LLMs, offers a fresh perspective in this field. By simplifying the need for extensive training with annotated datasets, we shift our focus away from architectural intricacies toward strategic approaches.

In our study, we adopt a zero-shot strategy to address the PAR problem. Specifically, we evaluate various VQA models across the five tasks included in the PAR Contest-CAIP23. We assess two state-of-the-art VQA models for these tasks alongside the contest-winning model. The consistent outstanding performance across different models underscores their suitability for specific tasks. Notably, the OPT model excels in color estimation, while FLANT5XL outperforms others in binary classification tasks.

Remarkably, this impressive performance is achieved through a zero-shot strategy, highlighting the potential impact of VLMs in computer vision. Their adaptability and flexibility hold great promise for addressing complex real-world vision challenges. As we look ahead, we anticipate significant shifts in how computer vision problems will be approached in the near future.

Acknowledgements This work is partially funded by the Spanish Ministry of Science and Innovation under projects PID2021-122402OB-C22, TED2021-131019B-10, as well as the ACI-ISI-Gobierno de Canarias and European FEDER funds under projects ProID2021010012, ULPGC Facilities Net, and Grant EIS 2021 04.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Jain AK, Dass SC, Nandakumar K. Soft biometric traits for personal recognition systems. In: International conference on biometric authentication. Berlin, Heidelberg: Springer; 2004. p. 731–8.
2. Kumar N, Berg AC, Belhumeur PN, Nayar SK. Describable visual attributes for face verification and image search. IEEE Trans Pattern Anal Mach Intell. 2011;33(10):1962–77.
3. Dietlmeier J, Antony J, Mcguinness K, O'Connor NE. How important are faces for person re identification? In: Proceedings international conference on pattern recognition. Milan: IEEE Computer Society; 2020.
4. Cheng Z, Zhu X, Gong S. Face re-identification challenge: are face recognition models good enough? Pattern Recognit. 2020;107:107422.
5. Li S, Xiao T, Li H, Zhou B, Yue D, Wang X. Person search with natural language description. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 5187–96. https://doi.org/10.1109/CVPR.2017.551.
6. Cormier M, Specker A, Junior J, Jacques C, Florin L, Metzler J, Moeslund TB, Nasrollahi K, Escalera S, Beyerer J. UPAR Challenge 2024: pedestrian attribute recognition and attribute-based person retrieval - dataset, design, and results. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. Waikoloa: IEEE Computer Society; 2023. p. 166–75.
7. Cormier M, Specker A, Junior J, Jacques C, Moritz L, Metzler J, Moeslund TB, Nasrollahi K, Escalera S, Beyerer J. UPAR challenge: pedestrian attribute recognition and attribute-based person retrieval - dataset, design, and results. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. Waikoloa: IEEE Computer Society; 2024.
8. Greco A, Vento B. Par contest 2023: pedestrian attributes recognition with multi-task learning. In: Tsapatsoulis N, Lanitis A, Pattichis M, Pattichis C, Kyrkou C, Kyriacou E, Theodosiou Z, Panayides A, editors. Computer analysis of images and patterns. Cham: Springer; 2023. p. 3–12.
9. Jia J, Huang H, Chen X, Huang K. Rethinking of pedestrian attribute recognition: a reliable evaluation under zero-shot pedestrian identity setting. 2021. arXiv preprint. arXiv:2107.03576.
10. Freire-Obregón D, De Marsico M, Barra P, Lorenzo-Navarro J, Castrillón-Santana M. Zero-shot ear cross-dataset transfer for person recognition on mobile devices. Pattern Recogn Lett. 2023;166:143–50.
11. Castrillón-Santana M, Sánchez-Nielsen E, Freire-Obregón D, Santana OJ, Hernández-Sosa D, Lorenzo-Navarro J. Evaluation of a visual question answering architecture for pedestrian attribute recognition. In: Tsapatsoulis N, Lanitis A, Pattichis M, Pattichis C, Kyrkou C, Kyriacou E, Theodosiou Z, Panayides A, editors. Computer analysis of images and patterns. Cham: Springer; 2023. p. 13–22.
12. Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. Preprint arXiv; 2023. https://doi.org/10.48550/arXiv.2301.12597.
13. DENG Y, Luo P, Loy CC, Tang X. Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM international conference on multimedia. MM '14. New York: Association for Computing Machinery; 2014. p. 789–92. https://doi.org/10.1145/2647868.2654966.
14. Li D, Zhang Z, Chen X, Huang K. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE Trans Image Process. 2019;28(4):1575–90.
15. Agrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, Parikh D, Batra D. VQA: visual question answering. Int J Comput Vis. 2015;123:4–31.
16. Barra S, Bisogni C, De Marsico M, Ricciardi S. Visual question answering: which investigated applications? Pattern Recognit Lett. 2021;151:325–31. https://doi.org/10.1016/j.patrec.2021.09.008.
17. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning (ICML), vol. 139; 2021. p. 8748–8763. http://proceedings.mlr.press/v139/radford21a.html.
18. Sridhar P, Lee H, Dutta A, Zisserman A. WISE image search engine (WISE). In: Wiki workshop, virtual event, May 11, 2023.
19. Kafle K, Kanan C. An analysis of visual question answering algorithms. In: IEEE international conference on computer vision (ICCV). Venice: IEEE Computer Society; 2017. p. 1983–91.
20. Toor AS, Wechsler H, Nappi M. Biometric surveillance using visual question answering. Pattern Recognit Lett. 2019;126:111–118. https://doi.org/10.1016/j.patrec.2018.02.013. Robustness, security and regulation aspects in current biometric systems.
21. Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab M, Li X, Lin XV, Mihaylov T, Ott M, Shleifer S, Shuster K, Simig D, Koura PS, Sridhar A, Wang T, Zettlemoyer L. OPT: open pre-trained transformer language models. arXiv preprint arXiv:2205.01068v4. 2022.
22. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. In: 9th international conference on learning representations, ICLR, Austria. 2021. https://openreview.net/forum?id=YicbFdNTTy.
23. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li E, Wang X, Dehghani M, Brahma S, Webson A, Gu SS, Dai Z, Suzgun M, Chen X, Chowdhery A, Narang S, Mishra G, Yu A, Zhao V, Huang Y, Dai A, Yu H, Petrov S, Chi EH, Dean J, Devlin J, Roberts A, Zhou D, Le QV, Wei J. Scaling instruction-finetuned language models. 2022. Preprint arXiv:2210.11416.
24. Tiong AMH, Li J, Li B, Savarese S, Hoi SCH. Plug-and-play VQA: zero-shot VQA by conjoining large pretrained models with zero training. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Findings of the association for computational linguistics: EMNLP 2022. Abu Dhabi: Association for Computational Linguistics; 2022. p. 951–967. https://doi.org/10.18653/v1/2022.findings-emnlp.67. https://aclanthology.org/2022.findings-emnlp.67.
25. Schlaffke L, Golisch A, Haag LM, Lenz M, Heba S, Lissek S, Schmidt-Wilcke T, Eysel UT, Tegenthoff M. The brain's dress code: how the dress allows to decode the neuronal pathway of an optical illusion. Cortex. 2015;73:271–5. https://doi.org/10.1016/j.cortex.2015.08.017.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.