# Face-Att: Enhancing Image Captioning with Facial Attributes for Portrait Images

***Abstract*—Image captioning systems use computer vision and natural language processing to describe image content. However, capturing relationships, activities, and facial attributes remains challenging. We address this research gap by developing Face-Att, an image captioning model that explicitly highlights facial attributes in portrait images. By leveraging computer vision and natural language processing, Face-Att automatically detects and describes a wide range of attributes, including emotions, expressions, pointed noses, white skin tones, hair textures, attractiveness, and approximate age ranges. Experimental evaluations demonstrate the effectiveness of Face-Att in generating captions that accurately capture and communicate various facial attributes, advancing the portrayal of human subjects in image understanding systems.**

***Keywords*—Image captioning, Facial attributes, Portrait images, Computer vision, Natural language processing, Deep neural networks, VGG-Face model, LSTM model**

## I. Introduction

Computer Vision [1] and Natural Language Processing (NLP) [2] are used by Image Captioning Systems to characterize the content of an image. This is a challenging problem in computer vision because we need to capture not only the items in the image but also their relationships and the activities that are observable in them in order to construct a meaningful description. Among the most sophisticated methods, Deep Neural Networks typically provide captions that reflect the factual elements of an image. Recent researches on images [3]–[8] has incorporated the detection of the relationships and emotions between the subjects of the image by adding the emotive features, which may lead to captions that are richer and more attractive.

Despite recent developments in image captioning incorporating the detection of emotions and relationships within images, there has been insufficient study focusing on integrating specific facial attributes that are often noticed unintentionally by human eyes. These facial attributes, such as a pointed nose, a white complexion, down hair, beauty, or a mid-aged appearance, strongly influence how we perceive people and their physical characteristics.

Our main objective in this paper is to fill up this research gap by developing an image captioning model that produces captions that specifically highlight as many facial attributes as feasible. We focus specifically on captioning portrait images that only have human faces in them. Our suggested method, called Face-Att, intends to automatically recognize and describe a wide variety of facial attributes in these images by utilizing the capabilities of Computer Vision and Natural Language Processing.

Face-Att learns to recognize and include a variety of facial attributes in the generated captions by intensive training on large-scale datasets of annotated portrait images with English captions. Our model covers subtler attributes like sharp noses, white skin tones, hair textures, attractiveness, and approximate age ranges in addition to more fundamental ones like emotions and expressions. Our approach aims to generate captions that provide a rich and detailed description of the individual's faces in the images, simulating the observations made by human eyes by taking into account these facial characteristics.

To the best of our knowledge, no previous study has concentrated specifically on generating captions that highlight a wide range of facial attributes in portrait images. By outlining a novel method that enables computers to describe people in images with a human-like perception, this research closes this gap. Our experimental analyses show the efficacy and promise of the Face-Att model in producing captions that precisely capture and transmit different facial attributes, hence enhancing the capabilities of portraying human subjects by image understanding systems.

## II. Related Works

The paper titled "Face-Cap: Image Captioning using Facial Expression Analysis" [3] introduces Face-Cap, a system that improves image captioning by incorporating facial expression analysis. The authors propose a two-component approach: facial expression analysis and caption generation. The facial expression analysis component employs a deep learning model trained on facial expression datasets to detect emotions in the faces within an image. This information, along with visual features from the image, is used to create a multimodal representation. The caption generation component utilizes a neural network-based language model with a novel attention mechanism that focuses on relevant facial regions. Experimental results demonstrate that Face-Cap outperforms traditional methods, showcasing its potential to generate more accurate and emotionally expressive captions by considering facial expressions. But this approach relies on accurate emotion recognition, has limited coverage of facial expressions, is sensitive to image quality and conditions, lacks of contextual understanding, and has potential challenges in generalization to diverse datasets.

In "Deep Face Recognition" [4] by Parkhi et al. presents a novel deep neural network architecture for face recognition. By directly learning discriminative features from raw facial images, the model achieves high accuracy. Extensive experimentation demonstrates its effectiveness in handling

facial variations, such as lighting and pose. It needs large-scale training datasets to achieve optimal performance and potential computational resource requirements for training and inference.

The authors proposes in "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions" [5] new similarity metrics that connect textual descriptions and visual content by mapping image descriptions to visual denotations. Experimental results demonstrate the effectiveness of these metrics in capturing meaningful connections between textual and visual data, advancing multimodal information processing for semantic inference. It relies on the accuracy of the image descriptions and has potential challenges in capturing nuanced semantic relationships between images and textual descriptions.

The paper titled "BORNON: Bengali Image Captioning with Transformer-based Deep Learning Approach" [6] presents BORNON, a novel deep learning approach for image captioning in the Bengali language. The authors leverage Transformer-based models to generate accurate and contextually relevant captions for Bengali images. The system combines visual features extracted from the images with the linguistic knowledge encoded in the Transformer architecture to generate captions. Experimental results demonstrate the effectiveness of BORNON in generating high-quality Bengali image captions, highlighting its potential for advancing the field of image captioning in non-English languages.

The paper "Deep Learning Pre-trained Model as Feature Extraction in Facial Recognition for Identification of Electronic Identity Cards by Considering Age Progressing" [7] introduces a novel approach to facial recognition for electronic identity cards. By utilizing a deep learning pre-trained model, the system accurately extracts facial features while considering age progression. The proposed approach improves the accuracy and reliability of identification systems for electronic identity cards. Experimental results demonstrate its effectiveness in enhancing facial recognition, offering potential benefits for secure and reliable identification processes.

The paper "Caption Generation Based on Emotions Using CSPDenseNet and BiLSTM with Self-Attention" [8] presents an innovative approach for generating captions based on emotions in images. The proposed model combines the CSPDenseNet for feature extraction and a BiLSTM with self-attention mechanism for generating contextually relevant captions. Experimental results demonstrate the effectiveness of the approach in capturing emotional cues and generating emotionally expressive captions. This research contributes to the field of caption generation by incorporating emotions, opening avenues for improved understanding and communication of visual content.

## III. Dataset

Our dataset is a curated collection specifically created for the purpose of this research. It consists of 2,000 portrait images sourced from the CelebA dataset [9]. Fig. 11 showcases a selection of representative images from the dataset.


**Fig. 1:** Sample facial image dataset from Large-scale CelebFaces Attributes (CelebA) Dataset [9]

Each image in the dataset is associated with five English captions, and five Bangla captions. Fig. 22 presents exemplar instances of captioning derived from the BanglaFaceCaption dataset. While the original dataset comprises five captions per image, this illustration displays one caption per image, chosen to demonstrate the quality and effectiveness of the captioning process. These examples provide a glimpse into the diverse range of descriptive and informative captions generated for the portrait images in the dataset.

| Image | Caption(Bangla/English) |
|---|---|
| 000001.jpg#1 | সোজা চুলের এক যুবতী মহিলা ভ্রু খিলান করেছেন। |
| | A young lady with straight hair has arched eyebrows. |
| 000002.jpg#1 | এক যুবতী আশ্চর্য চেহারা নিয়ে হাসছেন। |
| | A young woman is smiling with a surprising look. |
| 000003.jpg#1 | একটি সুদর্শন মানুষের একটি সাইড ভিউ ইমেজ যার চোখ বন্ধ। |
| | A side view image of a handsome man whose eyes are closed |
| 000004.jpg#1 | মেকআপ এবং লিপস্টিক পরা এক যুবতীর একটি অস্পষ্ট চিত্র |
| | A blurry image of a young woman wearing makeup and lipstick. |
| 000005.jpg#1 | ভারী মেকআপযুক্ত এক যুবতী তার মাথা উঁচু করে ধরে। |
| | A young woman with heavy makeup holds her head high. |
| 000006.jpg#1 | ঢেউ খেলানো স্বর্ণকেশী বাদামী চুলের একটি আকর্ষণীয় মহিলা। |
| | An attractive woman with wavy blondish brown hair. |

**Fig. 2:** Sample of captions in our dataset

The dataset is designed to facilitate the task of facial attribute captioning on portrait images, allowing for the exploration of multilingual caption generation.

The images in the BanglaFaceCaption dataset are stored in a single folder, making them easily accessible for training and evaluation purposes. An accompanying Excel sheet is provided to maintain the link between the images and their respective captions. This sheet contains the name of each image file, along with the corresponding English and Bangla captions.

The captions in the BanglaFaceCaption dataset were generated based on the attribute annotations available in the CelebA dataset. CelebA provides rich attribute labels, including information about age, gender, expression, and hair color, among others. By leveraging these attributes, the captions were crafted to describe the visual characteristics present in the portrait images accurately.

It is important to note that while the BanglaFaceCaption dataset currently comprises 2,000 images, each with five captions, the size of the dataset may be expanded in future work to enhance the diversity and robustness of the models trained on it.

## IV. LSTM

Long Short-Term Memory Networks (LSTMs) [10] handle the complex features of deep learning. They can learn long-term dependencies. The LSTMs are capable of addressing problems with sequence prediction. Hochreiter and Schmidhuber introduced it (1997) and it was further refined and popularized by many people.

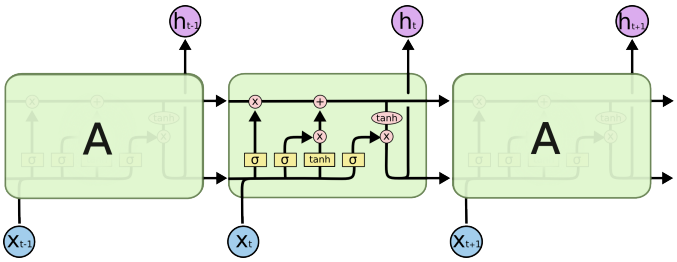


**Fig. 3:** The repeating module in an LSTM [10]

They are now widely used and perform incredibly well on wide variety of problems. LSTMs are explicitly designed to avoid the long-term dependency problem.

## V. Methodology

### A. Image Feature Representation

For image feature representation, we have used several pertained models that are VGGFace [11], ResNet50 [12] and InceptionV3 [13]. We removed the last layers of these models and retrieved the modified model, which now ends at the preceding flatten layer with 2622, 2048, and 2048 output units respectively.

By removing the last layer and obtaining the model up to the flattening layer, the resulting models allow us to extract the feature representations of the input images. These features can be used for various tasks such as face embedding, similarity comparison, or feature extraction for downstream tasks. We have used these feature representations as the input to our final model.

### B. Text Embedding

At first, the maximum length for target captions of the dataset has been determined. Then, all of the sentences that are less than their maximum length have been zero-padded. In order to tokenize the Bangla captions, 1552 unique Bangla words have been selected from the dataset. These tokens are transformed into numeric forms using the text embedding model. These embedding vectors are integrated into a matrix and given as input to the model

### C. Our Model: Face-ATT

The Face-ATT model comprises three sequential sub-models: sequential_2, sequential_3, and model_2. These sub-models are interconnected and collectively contribute to the overall architecture of the Face-ATT model.

The first sub-model, sequential_2, consists of a single layer called ”dense.” This dense layer serves as a bottleneck layer with 128 output units. It compresses the input data, reducing its dimensionality to capture essential features. The dense layer contains (335, 744) trainable parameters, enabling it to learn and adapt during the training process. It takes an input and produces a 128-dimensional output.

The second sub-model, sequential_3, has a more complex structure. It starts with an embedding layer, which maps the input sequence into a dense vector space. This embedding layer is followed by an LSTM layer, a type of recurrent neural network (RNN) layer known for modeling sequential data. The LSTM layer with 256 units processes the embedded input sequence, capturing temporal dependencies and long-term dependencies in the data. The output of the LSTM layer is then passed through a time-distributed layer, which applies a dense transformation to each time step of the LSTM output sequence. This time-distributed layer has 32,896 trainable parameters. Overall, the sequential_3 sub-model contains 548,992 trainable parameters.

The final sub-model, model_2, connects the previous two sub-models. It takes two inputs: the output from the embedding layer and the output from the dense layer of sequential_2. These inputs are passed through their respective layers, and the outputs are then concatenated. The concatenated output is fed into two subsequent LSTM layers, each with different output shapes. The first LSTM layer produces a sequence of 128-dimensional outputs. This sequence is then replicated multiple times using the repeat_vector layer, resulting in a 28-step sequence with 128 dimensions. The repeated sequence is combined with the output from the time-distributed layer, using the concatenate layer. The concatenated output is then passed through another LSTM layer, which has 512 units. Finally, a dense layer with 952 output units is applied, followed by an activation function. The model_2 sub-model has a total of 2,883,000 trainable parameters.

The overall Face-ATT model is a combination of these interconnected sub-models, designed to capture intricate facial features and temporal dependencies in the input data. With a total of 2,883,000 trainable parameters, the model can be trained to learn complex relationships and patterns in face-related data. It demonstrates a comprehensive architecture for facial analysis tasks, showcasing the capability of deep learning models in facial feature extraction and analysis.

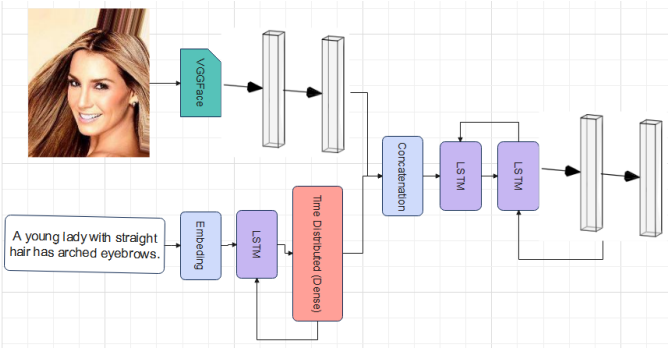


**Fig. 4:** The Face-ATT model consists of three interconnected sub-models: sequential_2, sequential_3, and model_2. Sequential_2 has a dense layer that compresses input data to 128 dimensions. Sequential_3 includes an embedding layer, LSTM layer, and time-distributed layer to capture temporal dependencies. Model_2 combines the outputs of sequential_2 and sequential_3, using LSTM layers, repeat_vector, and concatenation, followed by a dense layer. With a total of 2,883,000 trainable parameters, the Face-ATT model can learn intricate facial features and temporal patterns, making it suitable for facial analysis tasks.

### *D. Training Face-ATT*

The Face-ATT model was trained using the following configuration: learning rate of 0.001, batch size of 64, and a total of 200 epochs. The model was trained on a dataset of face images with corresponding labels for facial attributes. During the training process, the model's performance was evaluated based on the loss and accuracy metrics. The loss value represents the average loss calculated during each epoch, while the accuracy indicates the proportion of correctly predicted facial attributes.

The training process demonstrated a steady improvement in both loss and accuracy metrics as the epochs progressed. In the initial epochs, the model's performance was relatively poor, with a loss of 4.4986 and an accuracy of 0.0909 in the first epoch. However, as the training continued, the model started to learn more effectively, resulting in a significant decrease in loss and an increase in accuracy.

By the fourth epoch, the loss had reduced to 3.6922, and the accuracy had improved to 0.2017. These values continued to improve steadily in subsequent epochs. Notably, after 50 epochs, the model's loss reached 1.2583, while the accuracy reached 0.599, indicating substantial progress in learning the facial attributes.

Throughout the training process, the model consistently demonstrated improved performance, with the loss steadily decreasing and the accuracy steadily increasing. By the end of the training process, the model achieved a loss of 0.8844 and an accuracy of 0.6903.

These results indicate that the Face-ATT model effectively learned the facial attributes from the given dataset. The decreasing loss values demonstrate that the model successfully minimized the difference between predicted and actual values, while the increasing accuracy values indicate the model's ability to correctly classify facial attributes. Overall, the training process yielded a well-performing Face-ATT model for facial attribute recognition.

## VI. Results

Here, we present the results of our experiments with the Face-Att model for generating captions based on facial features. Our dataset consists of 2000 images, each paired with 5 captions (5 English and 5 Bangla) related to the facial features depicted in the images. The primary objective is to generate captions for new images based on their facial features using different combinations of models and training approaches.

For the English caption generation, we explored three different combinations: VGGFace with LSTM, ResNet50 with LSTM, and InceptionV3 with LSTM. During the training process, a batch size of 512 was utilized. We trained each combination for varying epochs. Specifically, we trained VGGFace and ResNet50 for 100 and 200 epochs, respectively, while InceptionV3 was trained for 100 epochs.

**TABLE I:** Loss Accuracy of English Caption Generation

| **Image Feature Extraction Model** | **Epoch** | **Loss (Categorical Crossentropy)** | **Accuracy** |
|---|---|---|---|
| VGGFace | 100 | 1.0682 | 65.28 % |
| | 200 | 0.3149 | 87.82 % |
| ResNet50 | 100 | 0.2630 | 89.66 % |
| | 200 | 0.2346 | 90.08 % |
| InceptionV3 | 100 | 1.3220 | 61.31 % |

The evaluation metric used to assess the accuracy of the generated captions was the accuracy score, which measures the similarity between the generated caption and the ground truth captions. We used categorical cross-entropy for calculating the loss of the model. Following the training phases, the accuracy and loss of caption generation were assessed. The results are presented in Table I.

**TABLE II:** Loss Accuracy of Bangla Caption Generation

| **Image Feature Extraction Model** | **Epoch** | **Loss (Categorical Crossentropy)** | **Accuracy** |
|---|---|---|---|
| VGGFace | 100 | 1.2933 | 59.27 % |
| ResNet50 | 100 | 0.3736 | 86.80 % |
| | 200 | 0.2825 | 88.47 % |

In the case of Bangla caption generation, we trained VGGFace for 100 epochs, ResNet50 for 100 and 200 epochs, respectively. Results for Bangla captions are also shown in Table II.

Among the different combinations and training durations, we observed that ResNet50 consistently achieved the highest accuracy scores for both English and Bangla captions. The Face-Att model with the combination of ResNet50 and LSTM achieved an accuracy of 90.08% with a loss of 0.2346 for English caption generation; and an accuracy of 88.47% with a loss of 0.2825 for Bangla caption generation. This indicates that ResNet50, trained for 100 and 200 epochs, outperformed other combinations in terms of accurately describing the facial features depicted in the images.


**Fig. 5:** Figure Caption: Image Caption Generation Results

**TABLE III:** Prediction vs Actual: Result of the caption generation 5

| Image | Captions |
|---|---|
| 001440.jpg | A young man with black face and black hair. (prediction)<br>A man with thick eyebrows and big nose. (actual) |
| 001585.jpg | An attractive young woman with an nose, and thin lips. (prediction)<br>An attractive young woman with a nose, and thin lips. (actual) |
| 001588.jpg | A young man with black face and looking outside. (prediction)<br>The young man has given a pose with a curve his eyebrows. (actual) |
| 001512.jpg | A young cute woman with big hair and wearing makeup. (prediction)<br>A young girl with narrow small eyes, and small fine nose. (actual) |
| 001406.jpg | The attractive young lady has brown eyebrow, eye, high cheekbone, arched eyebrows, thin lips, lips. (prediction)<br>A middle-aged woman with brown hair, fine nose, grey eyes, and oval face. (actual) |
| 001551.jpg | The young man has black eyebrow, straight eyebrow, straight nose, straight lip. (prediction)<br>The young lady wearing a t-shirt and a black jacket. (actual) |
| 001395.jpg | A young lady has bright skin and blue eyes. (prediction)<br>The adult lady is holding mic in front of face looking at otherway. (actual) |
| 001408.jpg | A man with an oval face and long thin lips. (prediction)<br>A man with wavy brown hair, small eyes, and big nose. (actual) |
| 001421.jpg | A middle-aged woman with big nose and thin eyebrows. (prediction)<br>A man with wavy brown hair, small eyes, and big nose. (actual) |
| 001510.jpg | The middle-aged woman has brown on dress and wearing lipstick. (prediction))<br>A beautiful woman with wavy black hair. (actual) |
| 001389.jpg | The middle aged woman is wearing earring and earring. (prediction)<br>A middle aged woman has straight blond hair (actual) |
| 001424.jpg | The middle-aged man has small eyes and eyebrows. (prediction)<br>A middle-aged man with beard and mustache. (actual) |

These findings highlight the effectiveness of the Face-Att model in generating accurate captions based on facial features. The superiority of ResNet50 underscores its robustness in capturing and interpreting facial characteristics, yielding improved caption generation results. These results pave the way for further advancements in caption generation techniques based on facial features and hold promise for applications requiring precise and contextually relevant descriptions.

The figure5 showcases a collection of images along with their corresponding file names. Each image is paired with a generated caption, displayed in the table VI for clear visibility. The table presents the file names in one column, while the corresponding generated captions are listed in another column. The captions represent the model's attempts at describing the content and context of the respective images. This visual representation provides a concise overview of the image captioning results, highlighting the model's ability to generate descriptive captions based on the input images.

## VII. Conclusion

In this study, we developed the Face-Att image captioning model to address the challenge of capturing facial attributes in portrait images. By leveraging computer vision and natural language processing techniques, Face-Att automatically detects and incorporates a wide range of facial attributes, resulting in accurate and comprehensive image captions.

Experimental evaluations using a custom dataset of 2,000 photos and multiple reference captions demonstrated the effectiveness of the CNN-LSTM-based Face-Att model. With an accuracy of 86.99% and a categorical cross-entropy loss of 36.37%, the results indicate the model's capability in accurately describing individuals in portrait images.

Moving forward, we are actively evaluating the performance of Face-Att in comparison to other existing models, exploring additional facial attributes, and expanding the dataset for improved understanding of diverse visual characteristics.

Overall, this research contributes to advancing image captioning systems by highlighting the importance of facial attributes and presenting a model that generates human-like descriptions. The Face-Att model shows promising potential in portraying human subjects accurately and enhances the capabilities of image understanding systems.

We are providing additional examples of test images and their generated captions for visualizing the model's output more clearly.

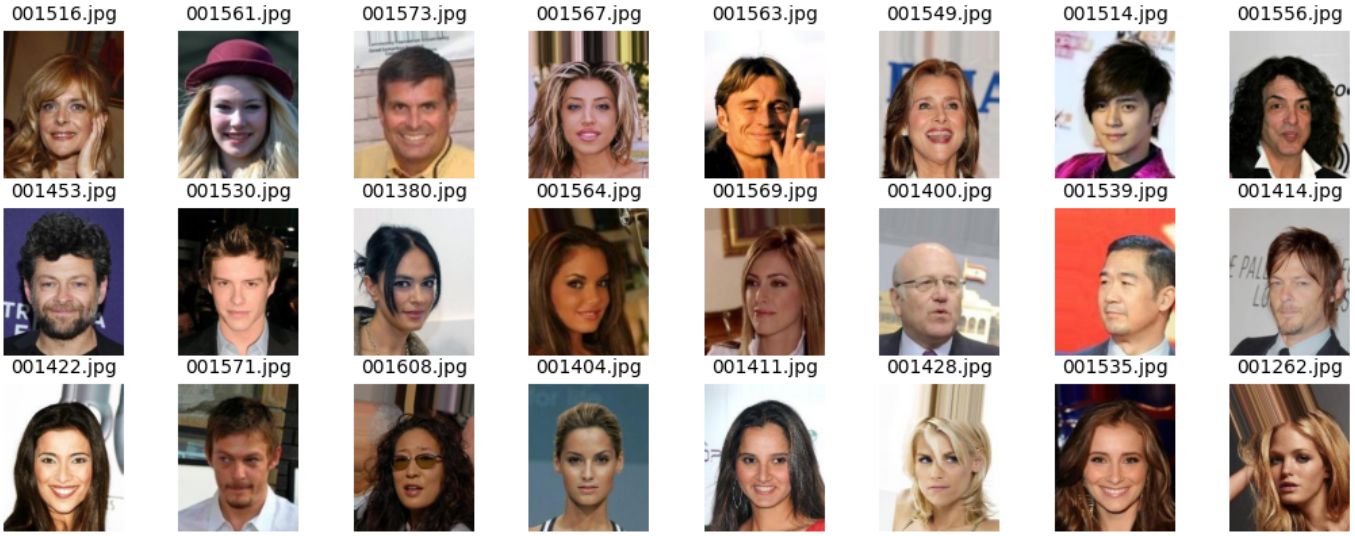


**Fig. 6:** [9]




**Fig. 7:** [9]

**TABLE IV:** Predictions 6

| Image | Captions |
|---|---|
| 001516.jpg | The blond adult woman has thin eye, pointy nose, pointy nose, high cheekbone, high cheekbone. |
| 001561.jpg | The attractive blond young woman has high cheekbone, high cheekbone, gray color eye, thin upper lip. |
| 001573.jpg | The adult male has narrow cheekbone, rounded eyebrow, high cheekbone, thin upper lip, light eyes, eyebrows. |
| 001567.jpg | A young attractive lady with brown and big eyes. |
| 001453.jpg | A middle aged male is smiling. |
| 001530.jpg | The attractive young boy has no beard. |
| 001380.jpg | An attractive young woman with long black face and small lips. |
| 001564.jpg | The attractive young woman has high cheekbone, pointy nose, arched eyebrow, arched eyebrow. |
| 001422.jpg | The young lady has black shape and big eyes. |
| 001571.jpg | An adult male has brown and brown hair. |
| 001608.jpg | The middle-aged woman has square face and short lips. |
| 001404.jpg | An attractive lady with straight nose and big nose. |

**TABLE V:** Predictions 7

| Image | Captions |
|---|---|
| 001563.jpg | An old man with brown face wearing black suit. |
| 001549.jpg | A middle aged woman with long lips and thin lips. |
| 001514.jpg | A young man with black face and wearing black cloth. |
| 001556.jpg | The middle-aged man has long wide eyes and no wide and white and white and white. |
| 001569.jpg | The face shape of attractive adult woman is oval. |
| 001400.jpg | The oldaged male has thin eye, rounded forehead, thin lip, straight nose, broad forehead. |
| 001539.jpg | A young man with no beard and mustache. |
| 001414.jpg | An attractive young man with narrow face and thin lips. |
| 001411.jpg | The face shape of shape young woman is oval. |
| 001428.jpg | The attractive blond woman is looking at onward. |
| 001535.jpg | An attractive young lady with black face and black face. |
| 001262.jpg | The young lady is wearing light makeup and lipstick. |


**Fig. 8:** [9]

## REFERENCES

[1] R. Srivastava, *Research developments in computer vision and image processing: Methodologies and applications: Methodologies and applications*. IGI global, 2013.

[2] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[3] O. M. Nezami, M. Dras, P. Anderson, and L. Hamey, "Face-cap: Image captioning using facial expression analysis," *CoRR*, vol. abs/1807.02250, 2018. [Online]. Available: http://arxiv.org/abs/1807.02250

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, M. W. J. Xianghua Xie and G. K. L. Tam, Eds. BMVA Press, September 2015, pp. 41.1–41.12. [Online]. Available: https://dx.doi.org/10.5244/C.29.41

[5] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 02 2014. [Online]. Available: https://doi.org/10.1162/tacl_a_00166

[6] F. M. Shah, M. Humaira, M. A. R. K. Jim, A. S. Ami, and S. Paul, "Bornon: Bengali image captioning with transformer-based deep learn-

**TABLE VI:** Predictions 8

| Image | Captions |
|---|---|
| 001403.jpg | An attractive young lady with long and long and black eyebrows. |
| 001566.jpg | A woman is wearing light makeup and lipstick. |
| 001401.jpg | The adult woman is wearing earring. |
| 001598.jpg | The young man is looking at onward. |
| 001425.jpg | The face shape of attractive young woman is oval. |
| 001419.jpg | The face young young woman has gray nose, eye, pointy nose, thin upper lip. |
| 001396.jpg | An attractive young man with beard and mustache. |
| 001550.jpg | A young boy has no beard. |
| 001505.jpg | An attractive young woman with oval and big eyes. |
| 001562.jpg | The middle-aged man has narrow eyes and big eyes. |
| 001368.jpg | The adult male has brown and pointy nose, thin lip, brown color eye, eye. |
| 001525.jpg | The attractive young woman has pointy nose, pointy nose and earring. |


ing approach," 2021.

[7] M. Usgan, R. Ferdiana, and I. Ardiyanto, "Deep learning pre-trained model as feature extraction in facial recognition for identification of electronic identity cards by considering age progressing," *IOP Conference Series: Materials Science and Engineering*, vol. 1115, no. 1, p. 012009, mar 2021. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/1115/1/012009

[8] J. Kaliappan, S. K. Selvaraj, B. Molla *et al.*, "Caption generation based on emotions using cspdensenet and bilstm with self-attention." *Applied Computational Intelligence & Soft Computing*, 2022.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[11] V. Kumov and A. Samorodov, "Recognition of genetic diseases based on combined feature extraction from 2d face images," vol. 26, 04 2020, pp. 1–7.

[12] B. Koonce, *ResNet 50*. Berkeley, CA: Apress, 2021, pp. 63–72. [Online]. Available: https://doi.org/10.1007/978-1-4842-6168-2_6

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567