Perceptual Deep Portrait Image Selection: Subjective and Objective Approaches

Maryam Karimi[1,*], Parvin Razzaghi[2]

Abstract— The efficiency of portrait image selection systems depends on the quality of face images, influenced by various factors. Real-time manual selection of high-quality portrait photos from multiple frames is often impractical, making automatic methods beneficial, particularly for large collections. However, existing automatic methods may not match human performance in portrait classification, often focusing on specific factors like emotional state or gaze direction. This work aims to simulate human choices in intelligent systems for portrait images. A large collection of facial images was gathered, and under subjective quality assessment, 200 images were evaluated by over 80 people. The results provided binary ground truth labels for the portraits. Subsequently, a deep classifier network using transfer learning and fine-tuning was proposed to objectively select good portrait images. The model achieved an accuracy of 0.83, surpassing other methods by at least 0.08. Additionally, F1, precision, and recall values of 0.9, 0.81, and 1 were obtained, exceeding other approaches by at least 0.05. Qualitative evaluations demonstrated the model's ability to distinguish good portrait images like humans, making it suitable for mobile phones, digital cameras, and other imaging systems.

Keywords*—Portrait Image Selection, Face Image Quality Assessment, Face Recognition, Face Detection, Face Image Subjective Dataset, Biometric Quality.*

1. Introduction

A portrait is a type of photo that includes one or two people against a local scenery background. It often happens that these photos are shared on social networks. Today, photography with the help of smartphones or digital cameras has become easier; however, taking photos does not always guarantee the quality of the photos taken. Therefore, most people are usually faced with a full gallery of portrait photos, from which choosing images is time-consuming and confusing. So, if digital cameras could select people's best moments from a large sequence of photos like a professional photographer, image filtering would be much faster, easier, and less expensive than manual selection.

Automatic selection of photos is not an easy task because the definition of a high-quality face image is not so clear. This concept depends on a combination of the personal preferences of the viewer, aesthetic features and structural quality of the image, illumination, exposure, focus, pose, and facial expressions. Therefore, Face image quality assessment can have different definitions for different purposes and applications [1,2].

Automatic face image quality prediction can be useful in many practical applications. Face image quality can also be used for quality-based portrait fusion when multiple face images (e.g., sequences of video

[1] M. Karimi is with the Department of Computer Science, Faculty of Mathematical Sciences, Shahrekord University, Shahrekord 88186-34141, Iran (e-mail: ma.karimi@sku.ac.ir , Tel: +9838 32324401, Mobile: +989131821250).
[2] P. Razzaghi is with the Department of Computer Science and Information Technology, Institute for advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran (e-mail: p.razzaghi@iasbs.ac.ir).

frames) are available. Video-based face recognition for surveillance scenarios is another application of quality-oriented face selection to reduce computation time and storage [1,2].

Unlike other biometric samples such as iris or fingerprints, the human vision system (HVS) is extremely advanced in assessing the quality of face images in recognizing people's faces, a common daily task. However, it is impossible to use human subjects in automatic systems. As far as we know, very few studies have been done on face quality evaluation by humans. A very little correlation between human measurements and face image quality measures was found in [3], while Hsu *et al.* [4] found some consistency between human perception and face recognition-based criteria [5]. Early works in face image quality mostly focused on simulation and segmentation on face image databases collected in the laboratory (e.g., FRGC [6], GBU [7], Multi-PIE [8]) that included some facial variations such as illumination and pose. Our previous work in [9] labels a portrait image as Unacceptable Condition (UC), Acceptable Condition (AC), or Best Condition (BC) to ensure the readiness of people in capturing time, employing face detection, blink detection, and iris detection steps. But, since the quality of a portrait image includes a wide range of factors and only the human vision system can understand such a complex concept, extraction of hand-crafted features does not provide enough accuracy in this scenario. Thus, in this paper, using an end-to-end deep neural network, quality-aware features are automatically learned that can classify portrait images as "good" or "bad", similar to human subjects do. In summary, the contributions of this work are as follows:

a. Collecting a large set of portrait images, conducting a subjective study by crowdsourcing users' opinions about the quality of the images, and finally labeling the images by voting [10].
b. Using an end-to-end deep classifier neural network, automatic extraction of quality-aware features from face images.
c. Presenting a deep model for automatic face image quality labeling trained on facial features extracted from our subjectively labeled database.
d. Using VGG16 as an ideal choice for selecting portrait images due to its effective use of small filters and deep architecture, which captures complex facial features, pre-trained weights and robustness to optical changes.
e. In addition to the proposed model, various deep networks widely used in image classification were also set up, learned, and tested separately. The results of the proposed network are superior to other methods in terms of various evaluation criteria.

The rest of this paper is organized as follows. Section 2 describes previous works, including the contribution of the models and the reasons for proposing our model. The details of image collection and the subjective study are provided in section 3. Section 4 fully describes the deep network, including our technique and model. Section 5 shows and discusses the results of our model in comparison with state-of-the-art classification networks. Finally, after a discussion in section 6, section 7 summarizes and concludes this paper.

2. Related Work

Many studies have analyzed the performance of face recognition algorithms with respect to various parameters such as pose, brightness, expression, resolution, and others [11,12]. Earlier works on face quality assessment are based on recognition performance degrading when faces deviate from bounded conditions. For example, [13] measures luminance distortion having a reference image for an adaptive fusion of face representations, [14] investigates the Structural Similarity Index Measure (SSIM) for face media fusion, and [15] proposes a reference-based model for the selection of high-quality frames from a

video. While full-reference face quality assessment is easy to interpret, generalization of these methods to images for which references are unavailable is impossible.

Goswami *et al.* in [16] and [17] introduced the highest visual entropy and entropy-based feature richness in the wavelet domain to achieve high verification in video. It should also be mentioned that recent works [18], [19]. Although trained specifically for face verification or recognition, a byproduct of the training process is a measure of face quality since the weights or coefficients learned to combine multiple faces into a single representation reflect the quality of a face for recognition. In recent learning-based approaches, the quality of the face image is first defined in a way that can be a genuine score [4], [20], [21], or a binary value [22], [23], [24]. A neural network in [4] combines 27 quality measures, including exposure, focus, pose, illumination, etc. A mapping from illumination features to the score space is learned for match pairs using multi-dimensional scaling in [20]. In [21], a CNN predicts the matching score from the LBP/HOG features and mutual subspace method. A PCA-LDA classifier in [22] is learned on face regions after normalizing the size, orientation, and illumination to distinguish between low- and high-quality portraits. The contrast, brightness, sharpness, focus, and illumination features are the features used by a neural network to classify face images [23]. The feature vector utilized by an AdaBoost classifier in [24] includes pose, blurriness, brightness, and color mismatch. While these methods define target quality values for face images, the method in [25] learns a model to rank face images by giving five image descriptors. In [26], a retrained fine-tuned inception model is presented with fully connected and regression layers that provide the five main cores, including vivid color, color harmony, lighting, balance of elements, and depth of field. This method produces a separate score for each of these five attributes in portrait images. The work proposed in [1], defines face image quality as a measure of the use of a face image for automatic face recognition. This method uses an SVR model trained on face features extracted using a deep convolutional neural network (ConvNet) to predict face image quality. The results show that this measure has been able to improve the performance of face matchers to a great extent. In [27], 25 different quality measures are evaluated on three face image databases using three open-source face recognition solutions. The results show the artificial features lack general stability and are significantly worse than overall face-specific quality metrics. Recently, focusing on difficult examples near the classification boundaries, a lightweight quality classification network in [28] is trained by performing knowledge distillation in the quality evaluation branch of a face recognition system. The MagFace method in [29] introduces a class of losses whose value is measured by the probability of recognizing a given face. This method introduces an adaptive mechanism to learn the distribution of within-class features by centering the class around easy samples while pushing hard samples away. The quality of face samples is estimated by learning a CR_FIQA model on the allocations of training feature representations in angular space with respect to their class centers method [30]. An attention-based end-to-end network for facial expression recognition is proposed in [31] consists of a feature extraction module, an attention module, a reconstruction module and a classification module. Another work in [32] presents a novel deep learning model that leverages generative priors for accurately predicting face image quality, marking a first in the field. Experimental results highlight the model's superior performance and the significant value of the accompanying dataset for face IQA applications. The QMagFace [33] enhances face recognition by integrating quality-aware comparison scores with a magnitude-aware angular margin loss, improving performance in unconstrained conditions. This approach achieves state-of-the-art results across various benchmarks, effectively addressing challenges like cross-pose and cross-age recognition. The study in [34] introduces a method for generating pixel-level quality explanation maps in face recognition systems, identifying regions with varying utility for recognition. It offers interpretable feedback on image quality, enhancing system performance and clarifying acceptance or rejection decisions based on quality factors.

While most of the state-of-the-art face image quality assessment methods are based on their ability to match, detect, or recognize faces, human perception-based quality evaluation is the most reliable method that has been less studied. Also, although most methods seek to manually extract effective features in face quality evaluation, it is necessary to provide methods that can automatically learn more effective features. In this article, it is tried to consider these two categories effectively.

3. Portrait Image Data Collection

As mentioned earlier, we want to classify portrait images that people usually take with their smartphones or digital cameras. Our goal is to create an image selection method for classifying these images as humans do. These images should be similar to pictures that people usually keep in the gallery of their smartphones, use in their profiles on social media, send to their family and friends, post, or share somewhere. To this end, we used the Helen dataset [35]. Helen's dataset includes portraits and selfies of people of different ages, genders, scenes, and gestures. We selected 200 images from the Helen database, including males and females in different age ranges. In addition, selected images differed in the background, lighting conditions, magnification, and shooting angle. Some samples of the image set can be observed in Fig.1.

3.1. Subjective Quality Labels

Labeling the entire dataset based on human decisions is the most reliable approach as we aim to classify images as humans do. Thus, we designed subjective tests to collect people's opinions as ground truth for our objective classification method. Because of the inherent ambiguity in defining facial image quality for viewers, framing an appropriate prompt to ask a human to assess a portrait image quality is challenging. For example, if you are asked to rate an image of a single face on a scale of 0 to 10, there is no concept of what the different levels of quality mean. Unless a specific definition of each grade is provided to the user. In addition, biased, inflated or conservative scoring also affects the final results in this method. To solve this problem, subjective methods of paired comparison are usually suggested, in which there is a need for a huge number of paired comparisons, which is not in the capacity and patience of the participants. But if the meaning of quality assessment for people simply falls into two categories, “good” and “bad”, it would be easier to choose for the viewers and would take less time than rating-based or pairwise comparison tasks. For these reasons, we choose the binary labeling method, that is, every time we show an image, and we ask the subject if, in your opinion, this face image is “good” or “bad” in terms of people's attention, readiness and gestures at the shooting time. The Results of votes in each test are displayed in Fig. 2. We collect data from our tests and label each image based on the largest number of "Good" or "Bad" labels. If the number of "Good" votes is more than "Bad" votes for an image, we label that image as a good image, and if the number of "Bad" votes is more, we label that image as a bad image [10]. Our dataset is publicly available for research [10].

3.2. Participants and Test Environment

Labeling all 200 images could be tedious and time-consuming for participants. In addition, participating in a long test can tire participants, and inaccuracies caused by their fatigue can affect the results. To put the viewers at ease and collect the most accurate data, we divided images into four groups of fifty. We designed an online test for each group of images and asked people to take our online test. Designing online tests helped us collect data from people of different ages and genders without special tools or environments. People could take the available test at any time of the day with their digital devices such as mobile phones, PCs, or laptops. These online tests include a complete description on their first page that guides human subjects on how to participate in the test and on what basis to label the images. After that, participants could view 50 different images one by one. On the first page of each test, we explained to participants that if they think the image is great and people in the image were ready at the time the photo was taken, select the

"good" label. On the other side, subjects were asked to select the "bad" label if they thought the person in the image was not ready to be photographed or if the image was taken in inappropriate conditions, such as a bad camera angle or inappropriate human pose. In each test, every image is located on a separate page, and two buttons, including "good" and "bad" are at the bottom of the page. A "next" button is also provided such that participants can choose it and view the next image when they choose either "good" or "bad" labels. In each test, selecting the appropriate label did not have any time limitations, and participants could choose one of the "good" or "bad" labels whenever they wanted. In addition, there was no time limitation for the entire test. We shared the links to our online tests in various groups on social networks approximately two days apart and asked people to attend our tests. Participation in the tests was unpaid and completely voluntary. More than 80 people participated in each subjective test, so the same labels were collected for each portrait image. The target groups in social media that we asked to take part in the tests were university students, instructors, and professors. Thus, almost all participants are probably between 18 and 50 years old with normal vision.

3.3. Outlier Detection

Outliers, as inappropriate data in the dataset, can affect the results and final accuracy. Thus, detecting and removing outliers is important in subjective tests [1], [36], [37]. After reviewing the results from the online tests, we observed some images with approximately equal numbers of "good" and "bad" votes. We figured out that people have different viewpoints in these cases, and humans cannot easily classify these images as good or bad. Since we attempt to use deep learning approaches to classify our dataset, having these images might affect the training process and confuse models. Thus, we detected them as outliers and removed them from the dataset.

To this end, we define a distance metric D_i for each image i in eq. (1) which is the absolute error between good and bad votes, gv_i, bv_i, normalized by their total number. If the distance value exceeded a defined threshold, we considered that image an outlier. Considering the obtained distance values for all images, we set the threshold as 0.1 to identify and discard the most complex images from the dataset. The complete information on our online tests, including the number of participants, viewers (People who have only viewed the tests or left them incomplete, we did not count their results in the tests), and outliers, can be found in Table. 1.

$$D_i = \frac{|gv_i - bv_i|}{gv_i + bv_i}$$
$$Status_i = \begin{cases} Acceptable, & D_i \geq 0.1 \\ Outlier, & D_i < 0.1 \end{cases} \quad (1)$$

4. Proposed Method

Convolutional Neural Network (CNN) is a deep neural network widely used for image domain [37]. It contains an input, a number of hidden, and output layers. The first layer of a CNN tries to extract the basic features of an input image, and the other layers focus on more complicated features. Based on the inherent structure of CNNs, they have shown remarkable performance in many computer vision research areas, such as image classification [38]. In this paper, we use CNN together with transfer learning and fine-tuning

techniques. Our proposed method is illustrated in detail in Fig. 3. In this method, after a preprocessing stage, a deep CNN for portrait image selection is trained to be used for test images.

4.1. Preprocessing and Augmentation

In this stage, preprocessing is done to standardize the image size and luminance range. We normalized each RGB image by dividing their pixel values to 255. Also, all the images are scaled to the size 224×224. Next, image augmentation is applied to generate some additional data. Data augmentation is a major step to obtaining a well-trained deep learning structure and avoiding over-fitting issues. Since the images in our dataset are complex due to differences in subject pose and emotional state, shooting angle, and background, their classification would be a sophisticated task. In addition, as presented in Table 1, not many images in our dataset can be used for both the training and test phases. Thus, the network requires more training data to have enough accuracy. However, due to the difficulty in finding proper images and enough participants for new subjective tests, we overcame this problem by using the image augmentation technique. Image augmentation provides artificial images by combining transformations such as rotation, shifting, zoom, cropping, etc., on the training image set. Using data augmentation technique can prevent overfitting and increase our model's performance [39]. To prevent extra space occupation, in our work, data augmentation is defined as an untrainable layer in our model, and four transformations, including random rotation, random flip, random contrast, and random zoom, were considered. The rotation, zoom and contrast range threshold is defined as 10%. It means our image augmentation layer can manipulate all input images with any of these transformations up to 10% compared to their original versions. One or more modifications at a time can be used to create each of the synthetic images in our method. When the training phase starts, in each epoch, the image augmentation layer applies these transformations to the input images to generate the artificial ones for training. Some samples of the augmented images from a training image can be observed in the fig.4.

4.2. Transfer learning with fine-tuning for portrait selection

Training a large CNN from scratch is a complex task because many parameters need to be tuned, and it requires a large dataset to train and avoid overfitting [40]. Transfer learning [41, 42] is a useful technique in machine learning that aims to improve the performance of a model, especially when the dataset is insufficient [43, 44]. Transfer learning is usually used to improve the performance of a machine learning model by using knowledge obtained from another model, which is trained for an almost similar task [40, 41, 42]. For instance, using transfer learning techniques can improve the accuracy of a cat detection model by using information learned from a dog detection model. In [40, 45, 46, 47] it is shown that using transfer learning for two unrelated tasks can also be effective and useful.

When a custom classifier with dense layer(s) is added to a pre-trained model, it should be fine-tuned because their primary weights are initialized randomly. In addition to the newly added classifier, we can unfreeze some or all layers from the pre-trained model and fine-tune all trainable layers end to end. Fine-tuning the entire network might be more effective when we use transfer learning for two irrelevant tasks [39]. The training process in the fine-tuning phase is also important because the weights of new added and unfrozen layers should be both optimized [48].

In this work, we use VGG16 [49] as a pre-trained model (trained with ImageNet, which is a large dataset) and add a custom classifier instead of its top layers. VGG16 is an effective choice for portrait image selection due to its architecture, utilizing small 3×3 filters for detailed feature extraction. Its 16-layer depth allows for learning complex features, and pre-training on large datasets enhances its performance. The model's simplicity means fewer hyperparameters, making it easier to optimize. Additionally, VGG16

is robust to lighting variations and can be fine-tuned for specific datasets, maximizing its effectiveness. VGG16 is a powerful CNN architecture for image classification, including six main blocks. The first five blocks consist of convolutional and pooling layers, and the last block contains flatten and dense layers. We discard the top layers of this architecture (the last block that contains Flatten and Dense layers) and add our custom classifier instead. More details of our custom-added layers are shown in Table 2. To design our custom classifier, we add a GlobalAveragePooling2D layer as the first layer to minimize the number of parameters. Then, a Dense layer is connected to it. After that, a Dropout layer with a rate of 0.5 is added to reduce overfitting. In the end, we add another dense layer as our output layer to classify images into one of two classes, including “Good Image” and “Bad Image”. According to Fig. 5, the main architecture and the workflow in this paper are as follows:

1- Taking convolutional and pooling layers from the VGG16 pre-trained model
2- Freezing all convolutional and pooling layers
3- Adding our proposed classifier on top of the VGG16 pre-trained model
4- Training our proposed classifier
5- Unfreezing all frozen convolutional and pooling layers in the VGG16 pre-trained model
6- Training the entire network end-to-end

5. Experimental Results and Analysis

The goal of this work is twofold: (1) determine the target, or "ground truth", quality values of a face image database, and (2) use this face image database with target quality labels to train a model to predict these labels using features automatically extracted from an unseen test face image. In this section, we are going to evaluate our model and report the experimental results. First, we need to define some setups that have been made for the dataset and the model.

5.1. Experimental Setup

After the outlier removal, we considered 188 images for the train and test to separate the train and test sets; we randomly selected 80% of images for the train and the remaining 20% for the test set. Since our dataset is completely unbalanced and less than 30% of images are labeled as "bad", the model may not be trained well or evaluated fairly if it is not fairly divided. To create appropriate train and test sets, we tried to have the same rate of good and bad images in both the train and test sets. To achieve this goal, we allocate about 30% of each of the two test and training sets to bad images and 70% of the rest to good images. A precise number of images existing in train and test sets can be observed in Table. 3. In addition, Fig. 6 depicts the status of our entire dataset, train set, and test set.

As it is explained in Section 4, our model has two training steps. At first, we consider 20 epochs with a learning rate of (1.00E-03), and the Adam optimization method [50] to train our custom classifier with transfer learning. We use the fine-tuning technique in the second training step to improve the model's performance. Adam is our optimizer with a very small learning rate (1.00E-05). In this phase, we considered the maximum number of epochs to be 100, but to control the training process and over-fitting avoidance, some callbacks are applied to get the best-trained version of the model:

a. Reduce learning rates in the case of failure over five consecutive epochs.
b. Stopping the learning process if there is no improvement in ten consecutive epochs.
c. Saving the best version of the trained model in each epoch and replacing it with the previous version.

5.2. Class weighting

When we plan to train a classification model in a supervised manner, the model should be trained from all classes. In supervised machine learning, when a model starts training on a dataset, it tries to reduce the error in each epoch. If we use an unbalanced dataset, the model might focus on the larger class because it has a greater effect on reducing the error. Therefore, the model may not be trained well, and it is more likely to assign the most repeated class label to a typical test image [51]. In other words, there is a bias towards the heaviest class compared to other classes. Since barely 30% of our entire dataset is made up of "bad" labeled images, our model might focus on "good" images. To overcome the imbalance caused by inappropriate image distribution in the dataset, we assign a higher weight to the minority class and a lower weight to the majority class. For this purpose, we determine the weight of each class based on eq. (2) where, W_i is the weight of class i. $N_{samples}$ is the number of all images in the dataset. $N_{classes}$ is the number of classes ($N_{classes} = 2$). $N_{samples_i}$ is the number of data samples in class i. According to this equation, the weights of classes are determined to be 1.7558 and 0.6991 for "bad" and "good" classes, respectively, to prevent any bias and train our model fairly.

$$W_i = \frac{N_{samples}}{N_{classes} \times N_{samples_i}} \tag{2}$$

5.3. Evaluation Metrics

Before explaining the evaluation metrics, it is important to define four main concepts that refer to how the model predicts the test samples and indicate whether the model correctly identifies their classes. Our work has two classes: "good" and "bad" images. Based on our classes, we can define them as follows:

TP (True Positive) and TN (True Negative) are, respectively, the number of good and bad images that the model predicted correctly. FP (False Positive) and FN (False Negative) are the numbers of images that are wrongly predicted as good and bad images, respectively.

To evaluate our proposed model, we use six commonly used metrics for classification methods: accuracy, precision, recall, F1 score, AUC-ROC, and AUC-PR.

- **Accuracy:** Accuracy is an important metric in machine learning that shows the general performance of a trained model. You can find it in eq. (3). Accuracy would be a good metric when we trained a model using a balanced dataset. Since this metric only compares true predictions to all predictions, a model trained on an unbalanced dataset may show good accuracy but is not well trained. In this case, the model may focus on the larger class and ignore the smaller one. Thus, accuracy shows a good number but does not mean good performance. Since our dataset is quite unbalanced, using accuracy cannot be sufficient sole.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

- **Precision:** As shown in eq. (4) precision is calculated by dividing the number of positive samples that are predicted correctly by all items predicted to be positive. Precision indicates how well our model can predict positive samples.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

- **Recall:** This metric shows the ratio of positive samples that were predicted correctly by the model to the summation of positive samples that were predicted correctly and negative samples that were predicted incorrectly. Compared to Precision, Recall is used when we want to consider FN instead of FP. It demonstrates how many positive samples are detected correctly.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

- **F1:** Since Precision and Recall are both important, we use the F1 metric (F measure) to have them together. As it is shown in eq. (6), F1 is calculated according to both precision and recall. Thus, a high F1 means that both precision and recall were approximately high.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

- **AUC-ROC:** The "Area Under the Curve" (AUC) of "Receiver Operating Characteristic" (ROC) is another evaluation metric that can be used to show the performance of a model. This score will be obtained by calculating the area under the ROC, a curve that visualizes true positive rate (TPR) and false positive rate (FPR). AUC-ROC score is widely used in binary classification to show how well a model can separate each class. This score will be between 0 and 1; the higher score means that the model can distinguish each class sample more accurately. A score of 0.5 means that the model cannot separate the classes and randomly classify each sample or choose one class for all samples. Therefore, a good score is something between 0.5 and 1. As explained in [52], AUC-ROC is not the best metric for evaluating unbalanced data. Since our data set is highly unbalanced, using AUC-ROC solely is insufficient.

- **AUC-PR:** The "Area Under the Curve" (AUC) of "Precision Recall" (PR), which is also called average precision, is also used to evaluate the performance of a trained classification model. Similar to AUC-ROC, this score would be between 0 and 1. This metric is more proper to evaluate a model with unbalanced data.

5.4. Quantitative Evaluation

We examined the trained model on the test image set when the training process was finished. The obtained test results are mentioned in Table. 4, in which some well-known image classifiers such as GoogleNet [53], AlexNet [54], and Vision Transformer [55] are compared with the proposed approach in terms of accuracy, precision, recall, F1 score, AUC-ROC, and AUC-PR. Using the transfer learning technique, we also test our custom classifier with some changes on VGG16 [49] and ResNet50 [56]. We slightly changed the custom classifier on these networks to achieve better results. To this end, a flattened layer was added to the middle dense layer instead of GlobalAveragePooling, in addition to three regularizers, including kernel, bias, and activity regularizers. As it can be observed in Table. 4, the performance results obtained by the proposed method are superior to others in terms of all the evaluation criteria.

VGG16 tends to excel over other models in selecting portrait images due to its simple and uniform architecture. It employs small 3×3 convolutional filters arranged in a deep structure, which effectively captures complex features. The depth of VGG16 enables it to learn hierarchical representations, crucial for recognizing subtle details in portraits. Additionally, the model benefits from pre-trained weights on large datasets, enhancing its performance on smaller, specific datasets. With fewer hyperparameters, VGG16 is

easier to fine-tune and train. Its design allows for robust handling of variations in lighting and angles, making it suitable for portrait recognition. Overall, the combination of effective feature extraction and task-specific optimizations contributes to VGG16's superior results in this domain.

The Precision-Recall (PR), and Receiver Operating Characteristic (ROC) curves are depicted in Fig. 7.a and 7.b respectively to illustrate the performance of the proposed portrait selector in terms of precision and recall, as well as its ability to distinguish between positive and negative classes.

5.5. Qualitative Evaluation

Although we presented the numerical results of the method on the unseen test set in subsection 5.4, in this section, we test this model on some different challenging images. We collect 35 non-copywrite portrait images from [57, 58] and label them using the proposed model. These images and their output labels are depicted in Fig. 8. Considering that the human psycho-visual system provides the ground truth for each portrait image, the judgment of the ability of this method is left to the viewers. It is emphasized that the efficiency of this method is high; however, the labeling of very few images is not done correctly, which are marked with red frames. Our model is currently more sensitive to detecting people's readiness and poses at the time of imaging and is not biased towards people's gaze direction and emotional state, which is the result we tried to achieve. However, as mentioned previously, the definition of a good portrait image is unclear, and people have different viewpoints about it.

6. Discussion and Future Work

In this work, we tried to train a classification model to separate good and bad portrait images from the way humans do. Although the results obtained from our proposed model are promising, some limitations and inadequacies must be addressed in future research. We tried to train our model with the average votes collected from human subjects to be sure that our model would have a human-like performance. Since people have different perspectives and attitudes may have different definitions of a good portrait image. Therefore, finding a way to customize a model to select good images based on each user preference will be challenging for future work.

Defining a good or bad portrait image is very controversial and requires deep psychological research beyond this work's scope. As an open problem for future work, psychological research could be conducted to investigate what criteria people consider to choose a portrait photo as a good photo or not.

Since training a model for our purpose is a heavy task and our dataset contains complex images, more images are needed for a better training step. Also, based on the comparison results, some more modifications on our network or the comparing ones may result in better classification performances.

6. Conclusion

This article investigated whether a classification model could separate good and bad portrait images as humans do. To this end, diverse portrait images were collected and labeled under a subjective test. After preprocessing and augmentation on the image set, a CNN architecture using transfer-learning with fine-tuning on a pre-trained VGG-16 is trained end to end for binary classification of face images. Numerical and visual results showed the efficiency of this model for selecting face images is very high, and it can be used in different devices such as smartphones or digital cameras. Our results showed that intelligent systems can perform almost as well as humans in selecting portrait images, indicating potential future research opportunities in this field.

Conflict of Interest:

The authors declare no conflict of interest.

Data Availability Statement:

The datasets generated during and analyzed during the current study are available in https://github.com/mkarimid/Portrait-Image-Selection-Dataset.

References

[1] Best-Rowden L., and Jain A.K., "Learning face image quality from human assessments," *IEEE Trans. Information Forensics and Security*, 13 (12), pp. 3064-3077, (2018), DOI: https://doi.org/10.1109/TIFS.2018.2799585 .

[2] Schlett T., Rathgeb C., Henniger O., et al., "Face image quality assessment: A literature survey," *ACM Computing Surveys (CSUR)*, 54(10), pp.1-49, (2022), DOI: https://doi.org/10.1145/3507901 .

[3] Adler A. and Dembinsky T., "Human vs. automatic measurement of biometric sample quality," In Canadian Conf. on Electrical and Computer Engineering, pp. 2090-2093, (2006), DOI: https://doi.org/10.1109/CCECE.2006.277715 .

[4] Hsu R.-L., Shah J., and Martin B., "Quality assessment of facial images," In *Biometrics Symposium: Special Issue on Research at the Biometric Consortium Conf.*,pp. 1-6, (2006), DOI: https://doi.org/10.1109/BCC.2006.4341617 .

[5] Scheirer W. J., Flynn P. J., Ding C., et al., "Report on the BTAS 2016 video person recognition evaluation," *in IEEE Conf. on Biometrics Theory, Applications and Systems*, (2016), DOI: https://doi.org/10.1109/BTAS.2016.7791198 .

[6] Phillips P. J., Flynn P. J., Scruggs T., et al., "Overview of the face recognition grand challenge," in *IEEE Conf. Computer Vision and Pattern Recognition*, (2005), DOI: https://doi.org/10.1109/CVPR.2005.268 .

[7] Phillips P. J., Beveridge J. R., Draper B. A., et al., "The good, the bad, and the ugly face challenge problem," Image and Vision Computing, 30(3), pp. 177–185, (2012), DOI: https://doi.org/10.1016/j.imavis.2012.01.004 .

[8] Gross R., Matthews I., Cohn J., et al., "Multi-PIE," Image and Vision Computing, 28(5), pp. 807–813, (2010), DOI: https://doi.org/10.1016/j.imavis.2009.08.002 .

[9] Entezami E., and Karimi M., "Automatic Portrait Image Selection for Smart Phones," In *2020 6th Iranian IEEE Conf. Signal Process. Intelligent Systems (ICSPIS)*, pp. 1-5, (2020), DOI: https://doi.org/10.1109/ICSPIS51611.2020.9349559 .

[10] https://github.com/mkarimid/Portrait-Image-Selection-Dataset.

[11] Beveridge J. R., Givens G. H., Phillips P. J., et al., "FRVT 2006: Quo vadis face quality," Image and Vision Computing, 28(5), pp. 732–743, (2010), DOI: https://doi.org/10.1016/j.imavis.2009.09.005 .

[12] Beveridge J., Givens G., Phillips P. J., et al., "Factors that influence algorithm performance in the face recognition grand challenge," *Comput. Vis. Image Underst.*, 113(6), pp. 750–762, (2009), DOI: https://doi.org/10.1016/j.cviu.2008.12.007.

[13] Sellahewa H., and Jassim S. A., "Image-quality-based adaptive face recognition," *IEEE Trans. Instrumentation and Measurement*, 59(4), pp. 805–813, (2010), DOI: https://doi.org/10.1109/TIM.2009.2037989 .

[14] Best-Rowden L., Han H., Otto C., et al., "Unconstrained face recognition: Identifying a person of interest from a media collection," IEEE Trans. on Information Forensics and Security, 9(12), pp. 2144–2157, (2014), DOI: https://doi.org/10.1109/TIFS.2014.2359577 .

[15] Wong Y., Chen S., Mau S., et al., "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," *in IEEE CVPRW*, pp. 74–81, (2011), DOI: https://doi.org/10.1109/CVPRW.2011.5981881 .

[16] Goswami G., Bhardwaj R., Singh R., et al., "Mdlface: Memorability augmented deep learning for video face recognition," *In IEEE Int'l Joint Conf. on Biometrics*, pp. 1–7, (2014), DOI: https://doi.org/10.1109/BTAS.2014.6996299 .

[17] Goswami G., Vatsa M., and Singh R., "Face verification via learned representation on feature-rich video frames," *IEEE Trans. on Information Forensics and Security*, 12(7), pp. 1686–1698, (2017), DOI: https://doi.org/10.1109/TIFS.2017.2668221.

[18] Tran L., Yin X., and Liu X., "Representation learning by rotating your faces," *IEEE Trans. pattern analysis and machine intelligence*, 41(12), pp. 3007-3021, (2018) , DOI: https://doi.org/10.1109/TPAMI.2018.2868350 .

[19] Yang J., Ren P., Zhang D., et al., "Neural aggregation network for video face recognition," *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4362-4371, (2017), DOI: https://doi.org/10.1109/CVPR.2017.554.

[20] Aggarwal G., Biswas S., Flynn P. J., et al., "Predicting performance of face recognition systems: An image characterization approach," *In IEEE CVPRW*, pp. 52-59, (2011), DOI: https://doi.org/10.1109/CVPRW.2011.5981784 .

[21] Vignesh S., Priya K. M., and Channappayya S. S., "Face image quality assessment for face selection in surveillance video using convolutional neural networks," *In 2015 IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, pp. 577-581, (2015), DOI: https://doi.org/10.1109/GlobalSIP.2015.7418261 .

[22] Phillips P. J., Beveridge J. R., Bolme D. S., et al., "On the existence of face quality measures," *In IEEE Conf. on Biometrics: Theory, Applications and Systems*, pp. 1–8, (2013), DOI: https://doi.org/10.1109/BTAS.2013.6712715 .

[23] Abaza A., Harrison M. A., Bourlai T., et al., "Design and evaluation of photometric image quality measures for effective face recognition," IET Biometrics, 3(4), pp. 314–324, (2014), DOI: https://doi.org/10.1049/iet-bmt.2014.0022 .

[24] Kim H. I., Lee S. H., and Ro Y. M., "Face image assessment learned with objective and relative face image qualities for improved face recognition," *In IEEE Int'l Conf. on Image Processing*, pp. 4027-4031, (2015), DOI: https://doi.org/10.1109/ICIP.2015.7351562 .

[25] Chen J., Deng Y., Bai G., et al., "Face image quality assessment based on learning to rank," *IEEE Signal Processing Letters*, 22(1), pp. 90–94, (2015), DOI: https://doi.org/10.1109/LSP.2014.2347419 .

[26] Wettayakorn P., Traivijitkhun S., Phetchai P., et al., "A deep learning methodology for automatic assessment of portrait image aesthetic quality," In *2018 15th IEEE Int'l. Joint Conf. on Computer Science and Software Engineering (JCSSE)*, pp. 1-6, (2018), DOI: https://doi.org/10.1109/JCSSE.2018.8457381 .

[27] Fu B., Chen C., Henniger O., et al., "A deep insight into measuring face image utility with general and face-specific image quality metrics," In *the IEEE/CVF Winter Conference on Applications of Computer Vision,* pp. 905-914, (2022), DOI: https://doi.org/10.1109/WACV51458.2022.00119 .

[28] Chen K., Yi T., and Lv Q., "LightQNet: Lightweight Deep Face Quality Assessment for Risk-Controlled Face Recognition," *IEEE Signal Processing Letters*, 28, pp.1878-1882, (2021), DOI: https://doi.org/10.1109/LSP.2021.3109781.

[29] Meng Q., Zhao S., Huang Z., et al., "Magface: A universal representation for face recognition and quality assessment," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14225-14234, (2021), DOI: https://doi.org/10.1109/CVPR46437.2021.01400 .

[30] Boutros F., Fang M., Klemt M., et al., "CR-FIQA: face image quality assessment by learning sample relative classifiability," In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5836-5845, (2023), DOI: https://doi.org/10.1109/CVPR52729.2023.00565 .

[31] Li J., Jin K., Zhou D., et al., "Attention mechanism-based CNN for facial expression recognition. *Neurocomputing*," 411, pp.340-350, (2020), DOI: https://doi.org/10.1016/j.neucom.2020.06.014 .

[32] Su S., Lin H., Hosu V., et al., "Going the extra mile in face image quality assessment: A novel database and model," *IEEE Trans. on Multimedia*, 26, pp. 2671-2685, (2023), DOI: https://doi.org/10.1109/TMM.2023.3301276 .

[33] Terhörst P., Ihlefeld M., Huber M., et al., "Qmagface: Simple and accurate quality-aware face recognition," In *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (CVPR)*, pp. 3484-3494, (2023), DOI: https://doi.org/10.1109/WACV56688.2023.00348 .

[34] Terhörst P., Huber M., Damer N., et al., "Pixel-level face image quality assessment for explainable face recognition," *IEEE Trans. on Biometrics, Behavior, and Identity Science*, 5(2), pp.288-297, (2023), DOI: https://doi.org/10.1109/TBIOM.2023.3263186 .

[35] Le V., Brandt J., Lin Z., et al., "Interactive facial feature localization," In *European conf. on computer vision*, pp. 679-692, (2012), DOI: https://doi.org/10.1007/978-3-642-33712-3_49 .

[36] Sheikh H.R., Sabir M.F., and Bovik A.C., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. on image process.*, 15(11), pp.3440-3451, (2006), DOI: https://doi.org/10.1109/TIP.2006.881959 .

[37] Jiang Q., Shao F., Jiang G., et al., "Three-dimensional visual comfort assessment via preference learning," *Journal of Electronic Imaging*, 24(4), p.043002, (2015), DOI: https://doi.org/10.1117/1.JEI.24.4.043002 .

[38] Sahu M., Dash R., "A survey on deep learning: convolution neural network (CNN)," *In Intelligent and Cloud Computing*, pp. 317-325, (2021), DOI: https://doi.org/10.1007/978-981-15-6202-0_32 .

[39] Masi I., Trần A.T., Hassner T., et al., "Face-specific data augmentation for unconstrained face recognition," *Int'l Journal of Computer Vision*, 127(6), pp. 642-667, (2019), DOI: https://doi.org/10.1007/s11263-019-01178-0 .

[40] Akhand MA., Roy S., Siddique N., et al., "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, 10(9), p:1036, (2021), DOI: https://doi.org/10.3390/electronics10091036 .

[41] Oquab M., Bottou L., Laptev I., et al., "Learning and transferring mid-level image representations using convolutional neural networks," *In IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1717-1724, (2014), DOI: https://doi.org/10.1109/CVPR.2014.222 .

[42] Yosinski J., Clune J., Bengio Y., et al., "How transferable are features in deep neural networks?," *Advances in neural information processing systems 27* (NIPS), (2014).

[43] Pan SJ., Yang Q., "A survey on transfer learning," *IEEE Trans. on knowledge and data engineering*, 22(10), pp.1345-59, (2009), DOI: https://doi.org/10.1109/TKDE.2009.191 .

[44] Torrey L., Shavlik J., "Transfer learning," *In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242-264, IGI global, (2010), DOI: https://doi.org/10.4018/978-1-60566-766-9.ch011 .

[45] Sharif Razavian A., Azizpour H., Sullivan J., et al., "CNN features off-the-shelf: an astounding baseline for recognition," *In IEEE conference on computer vision and pattern recognition workshops*, pp. 806-813, (2014), DOI: https://doi.org/10.1109/CVPRW.2014.131 .

[46] Bukar AM., Ugail H., "Automatic age estimation from facial profile view," *IET Computer Vision*, 11(8), pp.650-5, (2017), DOI: https://doi.org/10.1049/iet-cvi.2016.0486 .

[47] Chhikara P., Singh P., Gupta P., et al., "Deep convolutional neural network with transfer learning for detecting pneumonia on chest X-rays," *In Advances in bioinformatics, multimedia, and electronics circuits and signals*, pp. 155-168, (2020), DOI: https://doi.org/10.1007/978-981-15-0339-9_13 .

[48] Xiang Q., Wang X., Li R., et al., "Fruit image classification based on Mobilenetv2 with transfer learning technique," *In 3rd Int. Conf. on Computer Science and Application Engineering*, pp. 1-7, (2019), DOI: https://doi.org/10.1145/3331453.3361658.

[49] Simonyan K., and Zisserman A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, (2014), DOI:: https://doi.org/10.48550/arXiv.1409.1556 .

[50] Jais I.K.M., Ismail A.R., and Nisa S.Q., "Adam optimization algorithm for wide and deep neural network," *Knowledge Engineering and Data Science*, 2(1), pp.41-46, (2019), DOI: http://dx.doi.org/10.17977/um018v2i12019p41-46 .

[51] Byrd J., and Lipton Z., "What is the effect of importance weighting in deep learning?", In *Int'l Conf. on Machine Learning*, pp. 872-881, (2019), DOI: https://doi.org/10.48550/arXiv.1812.03372 .

[52] Saito T., Rehmsmeier M., "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, 10(3), p. e0118432, (2015), DOI: https://doi.org/10.1371/journal.pone.0118432 .

[53] Szegedy C., Zaremba W., Sutskever I., et al., "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, (2013), DOI: https://doi.org/10.48550/arXiv.1711.02846 .

[54] Krizhevsky A., Sutskever I., and Hinton G.E., "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, 60(6), pp.84-90, (2017), DOI: https://doi.org/10.1145/3065386 .

[55] Vaswani A., Shazeer N., Parmar N., et al., "Attention is all you need," *Advances in neural information processing systems*, 30, (2017), https://doi.org/10.48550/arXiv.1706.03762 .

[56] He K., Zhang X., Ren S., et al., "Deep residual learning for image recognition," In *IEEE Conf. on computer vision and pattern recognition (CVPR)*, pp. 770-778, (2016), DOI: https://doi.org/10.1109/CVPR.2016.90 .

[57] https://www.pexels.com/ , available 2024.

[58] https://unsplash.com/ , available 2024.

Biographies

Maryam Karimi received the B.S. degree in computer engineering from Amirkabir University of Technology, Tehran, Iran, in 2006, the M.Sc. degree in computer engineering from the Sharif University of Technology, Tehran, Iran, in 2009 and Ph.D. degree in computer engineering from the Isfahan University of Technology, Isfahan, Iran, in 2018. She is currently an Assistant Professor with the Department of Computer Science, Faculty of Mathematical Sciences, Shahrekord University, Shahrekord, Iran. Her teaching experience includes image processing, machine learning, deep learning, and artificial intelligence. Her research interests include subjective and objective image/video quality assessment and perceptual modeling.

Parvin Razzaghi is an assistant professor at the Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran. She completed a B.S. in computer science at Tabriz University, an M.S. degree, and a Ph.D. degree in computer engineering at the University of Isfahan, Iran. Her teaching experience includes graduate courses in computer vision, image processing, machine learning, advanced machine learning, and graphical models, as well as undergraduate courses in programming, artificial intelligence, and introduction to machine learning. She is interested in machine learning, deep learning, transfer learning, pattern recognition, and its application to computer vision and bioinformatics.

List of captions

Fig.1: Some samples of our image set [10] selected from the Helen dataset [31].

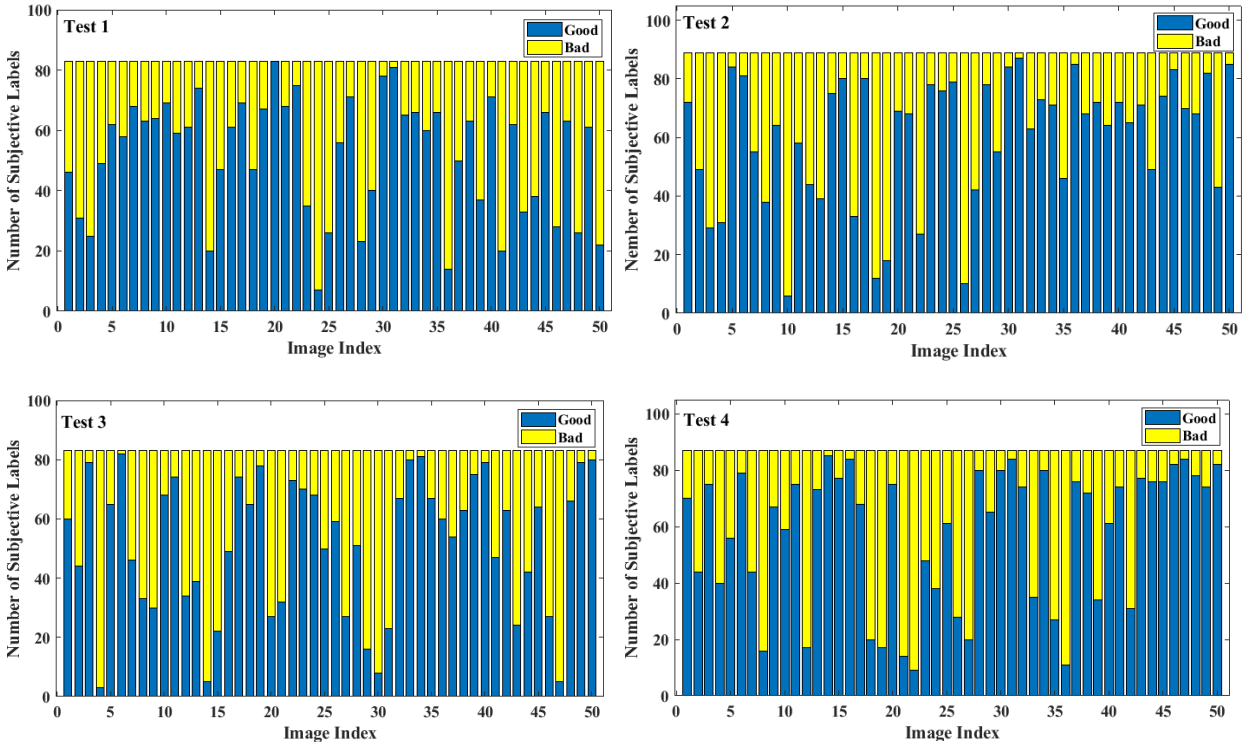


Fig 2: The number of “Good” and “Bad” labels for each image in the subjective tests

Table. 1: Complete information on all subjective tests

Subjective Test	Number of Images	Number of Viewers	Number of participants	Number of Outliers
1	50	142	83	2
2	50	139	89	4
3	50	124	83	3
4	50	128	87	3

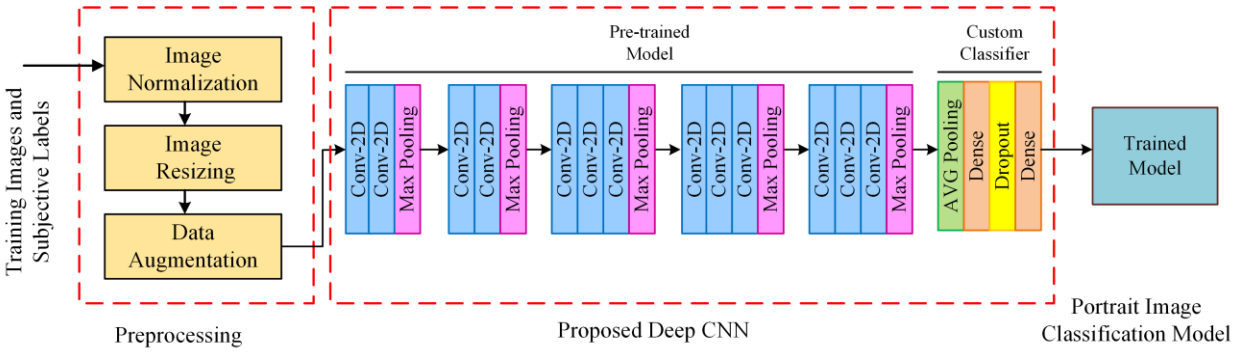


Fig. 3: Overview of the proposed image classification model for automatic portrait image selection.

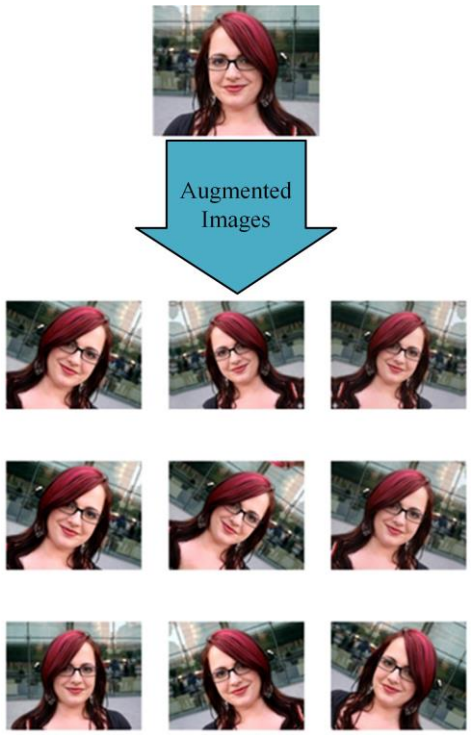
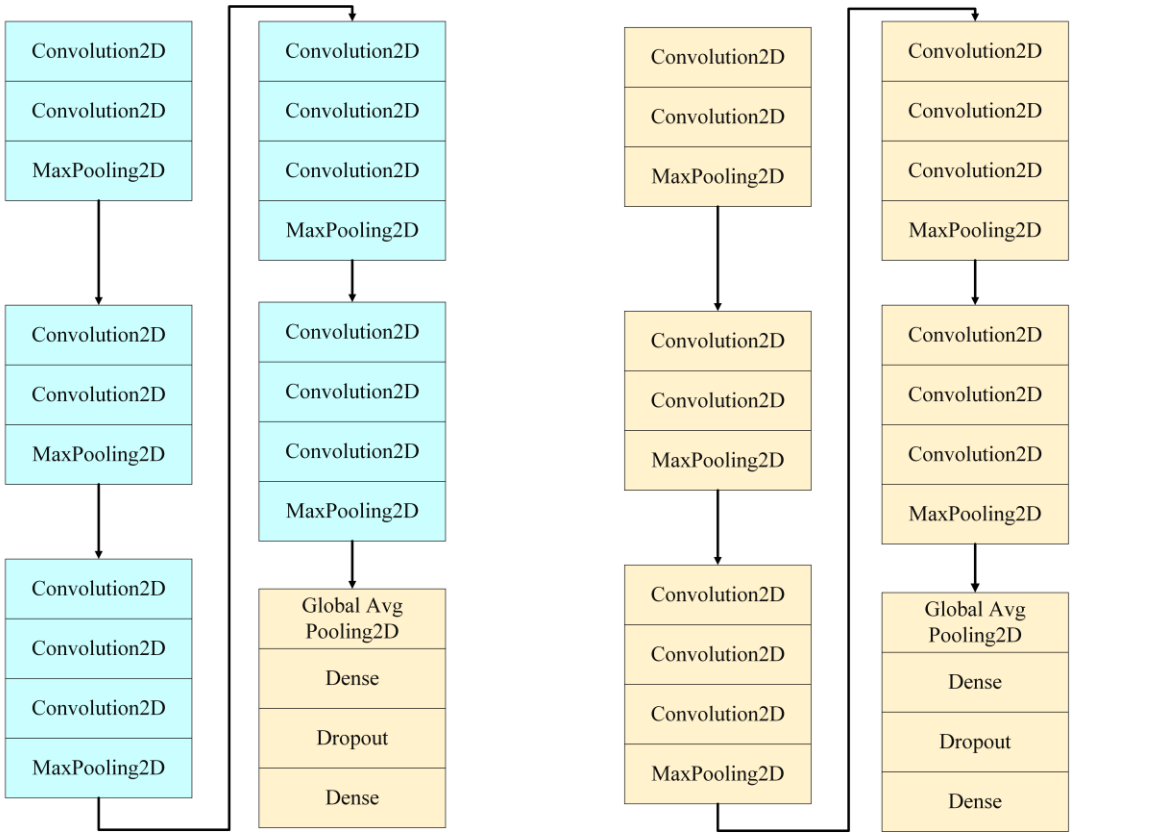


Fig 4: Some outputs of the augmentation process on a sample portrait image

Table 2: Detail of our custom classifier

Layers	Details
GlobalAveragePooling2D	
Dense	Number of Filters=256 Activation= ReLU
Dropout	Dropout Rate=0.5
Dense	Number of Filters=2 Activation=Sigmoid



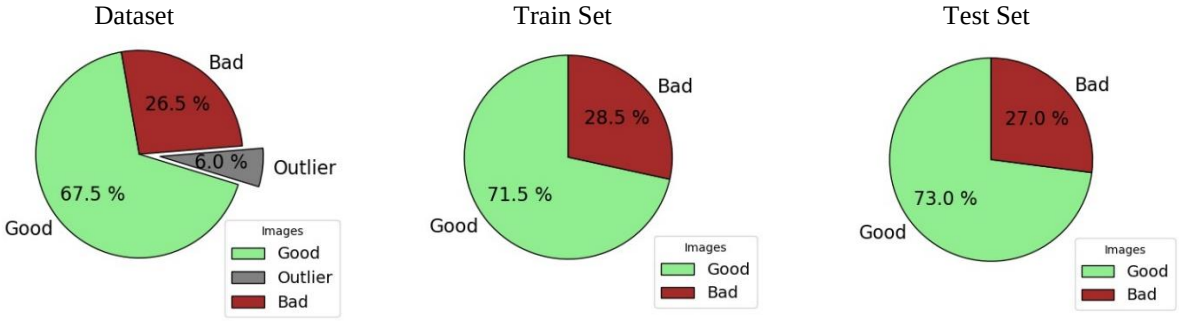
Phase1: Freezing all Convolutional and Pooling layers in VGG16 pre-trained model and training custom classifier

Phase2: Unfreezing all frozen layers and fine-tuning entire network end to end

Fig. 5: the training process of our portrait image classifier. The blue layers are frozen, and the yellow ones are trainable.

Table 3: Number of good and bad images in train and test sets

Labels	Train set	Test set
"Good"	108	27
"Bad"	43	10
Total	151	37

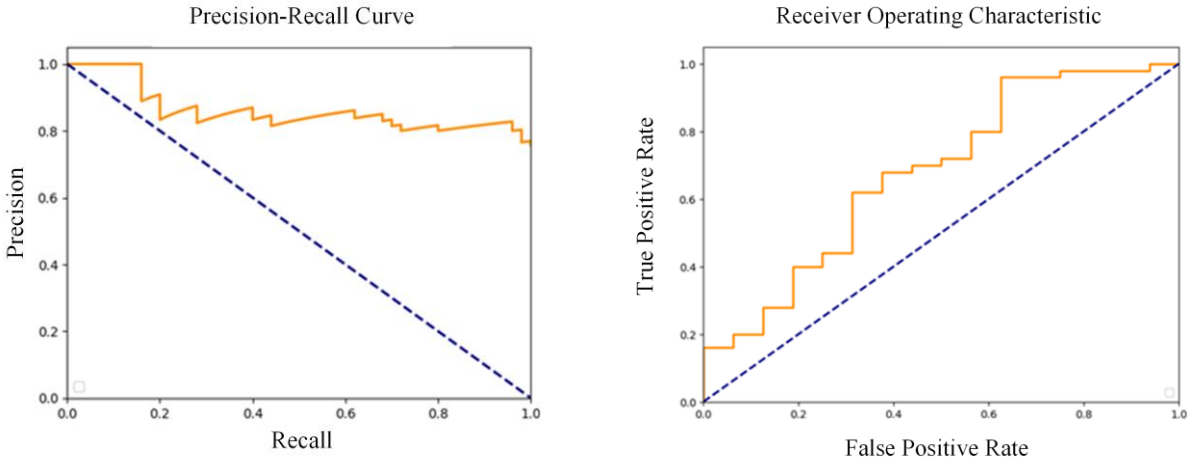


(a) (b) (c)

Fig. 6: Percentage of "Good" and "Bad" images in the (a) entire dataset, (b) train, and (c) test sets

Table 4: Performance comparison of different classification models in terms of accuracy, precision, recall, F1, AUC-ROC, and AUC-PR on the subjective portrait image dataset.

Classification Model	Accuracy	Precision	Recall	F1	AUC-ROC	AUC-PR
GoogleNet	0.75	0.75	1.00	0.85	0.55	0.75
AlexNet	0.72	0.72	1.00	0.84	0.50	0.72
Vision Transformer	0.64	0.75	0.77	0.76	0.53	0.74
VGG16 + Transfer Learning	0.64	0.76	0.74	0.75	0.57	0.75
ResNet50 + Transfer Learning	0.67	0.75	0.81	0.78	0.55	0.75
Proposed model	**0.83**	**0.81**	**1.00**	**0.90**	**0.70**	**0.81**



(a) **(b)**

Fig. 7: (a) Precision-Recall (PR) and (b) Receiver Operating Characteristic (ROC) curves for the proposed model

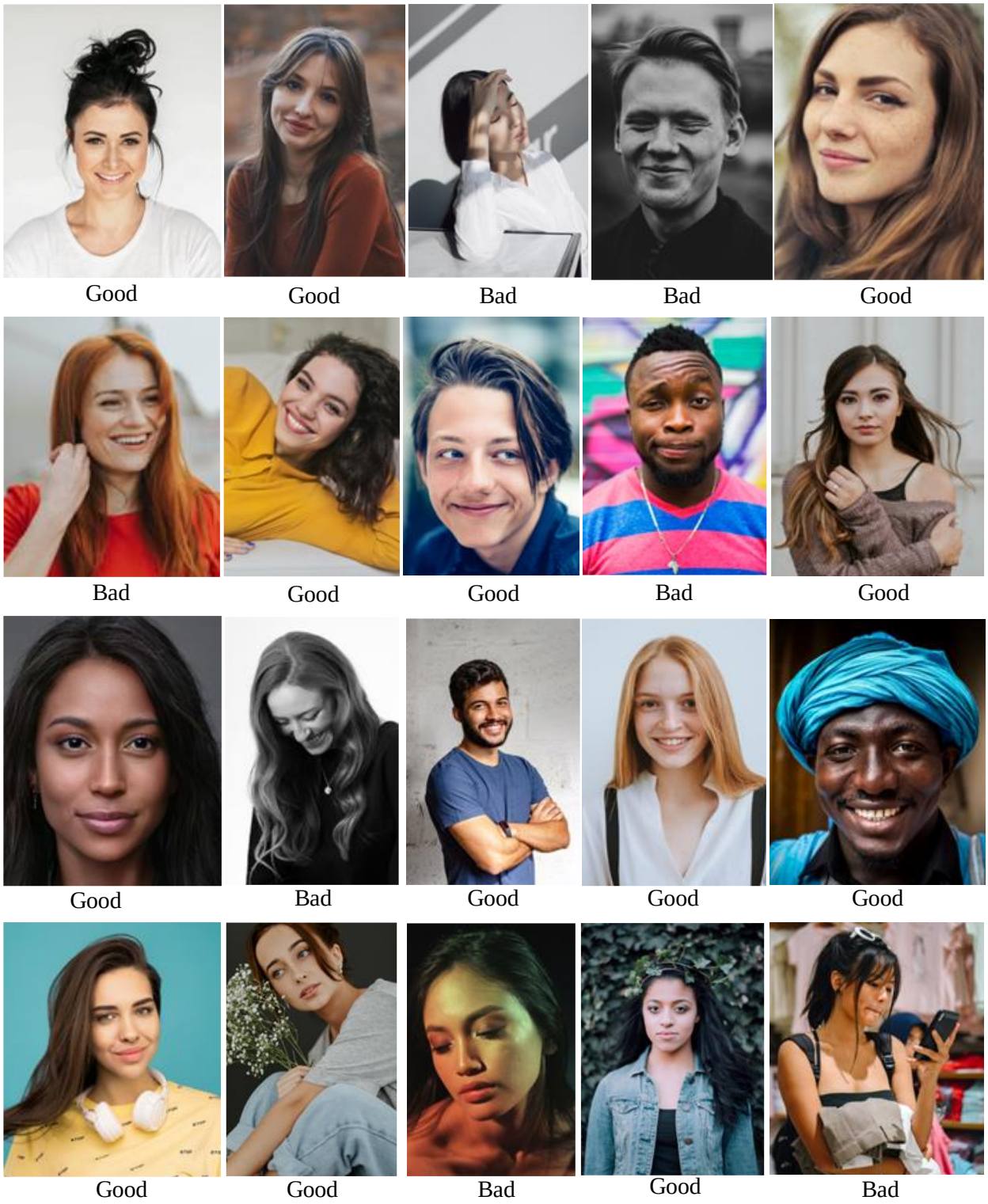


Fig. 8: Predicted labels for some non-copywrite portrait images out of the collected database