IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 April 2024, accepted 24 April 2024, date of publication 29 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3394870

RESEARCH ARTICLE

Anchor-Net: Distance-Based Self-Supervised Learning Model for Facial Beauty Prediction

JIHO BAE, SEOK-JUN BUU, (Member, IEEE), AND SUWON LEE, (Member, IEEE)
Department of Computer Science, Gyeongsang National University, Jinju-si 52828, Republic of Korea

Corresponding author: Suwon Lee (leesuwon@gnu.ac.kr)

This work was supported in part by the Research Grant of Gyeongsang National University, in 2022; and in part by the Regional innovation strategy (RIS) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) under Grant 2021RIS-003.

ABSTRACT In today's society, beauty is more than just aesthetics, it has a profound impact on many aspects of life, including social interactions, self-confidence, and job opportunities. To quantify this beauty, the field of Facial Beauty Prediction (FBP) is gaining traction. In the field of FBP, traditional methods often fall short due to their reliance on absolute beauty scores, which do not fully capture the subjective nature of human aesthetic perception. This study presents a novel approach to address this gap through the development of Anchor-Net, a self-supervised learning model that predicts differences in relative beauty scores by comparing images. The objective of this research is to offer a more nuanced understanding of facial beauty by employing a reference image (anchor) alongside a prediction image, thereby aligning closer with how humans perceive aesthetic differences. To construct Anchor-Net, we first developed a Base model that predicts beauty scores using a model pre-trained with VGGFace2. This Base model was then adapted into Anchor-Net, which is designed to train on the difference in beauty scores between a reference image and a prediction image. Our methodology involved two transfer learning steps to leverage the strengths of pre existing models while tailoring them to our specific research problem. The experimental validation of Anchor-Net was conducted on the SCUT-FBP5500 benchmark dataset, utilizing a 6:4 training-testing split and 5-fold cross-validation to ensure robust testing of the model's predictive capabilities. The results demonstrate that Anchor-Net outperforms other state-of-the-art deep learning algorithms on all metrics: Pearson Correlation (PC), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Anchor-Net outperformed other models with a PC of 0.0021, MAE of 0.0055, and RMSE of 0.0065 on a 6:4 training-test split. On average, it achieved a PC of 0.0034, MAE of 0.0155, and RMSE of 0.0135 on 5-fold cross-validation. This research proposes a novel approach to FBP and suggests a broader application of relative comparison methodologies in fields where absolute measurements fall short.

INDEX TERMS Anchor, convolution neural network, deep learning, facial beauty prediction, SCUT-FBP, self-supervised learning.

I. INTRODUCTION

The quest to understand and quantify beauty has been a perennial human endeavor, transcending the bounds of mere aesthetics to significantly influence various facets of daily life, including social dynamics, self-esteem, and professional opportunities [1], [2], [3], [4], [5], [6]. In contemporary society, the interplay between appearance and these life aspects has intensified, catalyzing the development of facial beauty prediction (FBP) models at the confluence of computer vision and psychology [7], [8]. These models aim to encapsulate and predict perceptions of facial attractiveness, tapping into our inherent predisposition to evaluate physical appearance.

Historically, the evolution of FBP methodologies has mirrored broader technological progressions. Initial approaches predominantly harnessed feature-based analysis, focusing on quantifiable facial attributes such as symmetry, proportion, and other geometric considerations [9], [10], [11], [12], [13]. These metrics, though pioneering, offered a somewhat

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

TABLE 1. Comparative analysis between deep learning models using the proposed Anchor-Net.

Approach	Model	Method	Expediency	Impairments	Comparison with proposed Anchor-Net
Ensemble	CNN-ER [21]	6 CNN Model loss Ensemble	Leverages diverse models for robustness, Improves accuracy with multiple losses, Captures a wide range of feature	Increased complexity and potential overfitting, hard to interpret, lack of relativity feature in beauty perception	Ensemble with multiple anchors to capture relative features, interpretability through anchor image analysis
	FLAC-NET [22]	3 loss Ensemble			
	EN-CNN [23]	3 CNN model Ensemble			
Semi-supervised	Semi-supervised (NFME) [24]	Graph-based semi-supervised Learning	Reinforce the training of the model without additional labelled face images.	Limitations of computing similarity graphs before estimating beauty prediction models	Self-supervised learning method that learns by creating new labels, enabling supervised learning methods to capture additional features such as relative beauty features
	Semi-supervised (MSMFME) [25]	Multi-View Graph Fusion for Semi-Supervised Learning			
Ranking based	R2-ResNeXt [26]	Ranking network	Rank-based relative beauty feature extraction with pairwise data	Ranked features between images do not fully represent the relative features of two faces	Captures relative features of beauty by comparing the image directly to the anchor image
	R3CNN [27]	Siamese network			

constrained view of beauty, predicated on static and universal standards.

The advent of deep learning and neural networks heralded a new era in this domain, introducing models with the ability to assimilate and predict beauty based on comprehensive data sets [13], [14], [15], [16], [17], [18], [19], [20]. Trained on extensive facial imagery, these advanced algorithms have showcased an unparalleled aptitude for approximating human judgment of beauty, thus marking a significant milestone in the field's evolution. Despite their technological sophistication, a critical limitation persisted: the reliance on absolute beauty scores [15], [24], [28], [29]. Such scores, by their very nature, encapsulate a monolithic and somewhat reductive perspective on beauty, sidelining the nuanced and inherently subjective dimensions that characterize human aesthetic judgment.

Conventional models, anchored in absolute scoring mechanisms, inadequately reflect the dynamic, comparative, and contextual framework within which humans perceive and assess beauty [29], [30], [31], [32]. This limitation underscores the exigency for innovative FBP models that can navigate the complexities of relative beauty perception, thereby offering predictions that are not only refined but also resonate more authentically with diverse and evolving beauty standards.

In response to this imperative, the present study introduces Anchor-Net, a self-supervised learning model meticulously crafted for the realm of FBP, predicated on a nuanced understanding of relative human beauty perception. Distinctively, Anchor-Net eschews the traditional absolutes in favor of a comparison-based paradigm, wherein the beauty of a target facial image is evaluated relative to a pre-selected reference image, or 'anchor'. This methodological pivot not only aligns more closely with the inherent comparative nature of human aesthetic evaluation but also enriches the model's predictive fidelity by embracing the spectrum of beauty as a relational concept. The primary contributions and innovations of Anchor-Net are outlined below.

- Comparison-based Methodology: Anchor-Net introduces a novel comparison-based methodology for FBP. This approach contrasts a target image against a reference 'Anchor,' mirroring human comparative perception and offering a more nuanced understanding of facial beauty.
- Advanced Two-Step Transfer Learning: We employ a sophisticated two-step transfer learning process, first developing a Base model using VGGFace2 [33] and then adapting it to Anchor-Net. This enhanced the model's capability to assess relative beauty differences.
- Superior Performance on Benchmarks: This has been extensively tested on the SCUT-FBP5500 dataset [15] using a 6:4 training-testing split and 5-fold cross-validation. The model outperformed contemporary deep learning algorithms, excelling in key metrics such as the PC, MAE, RMSE.

This paper not only propounds a novel perspective on FBP but also beckons a broader reconsideration of how relative comparison methodologies might be leveraged across disciplines where absolute measurements fall short, heralding a paradigmatic shift in our approach to understanding beauty.

II. RELATED WORKS

Early research on FBP focused primarily on hand-crafted features, such as geometric and textural features [19], [34], [35]. To describe geometric information, they empirically specified landmarks on the face and used distances or ratios between landmarks [9], [10], [11], [12], [13], or shallow

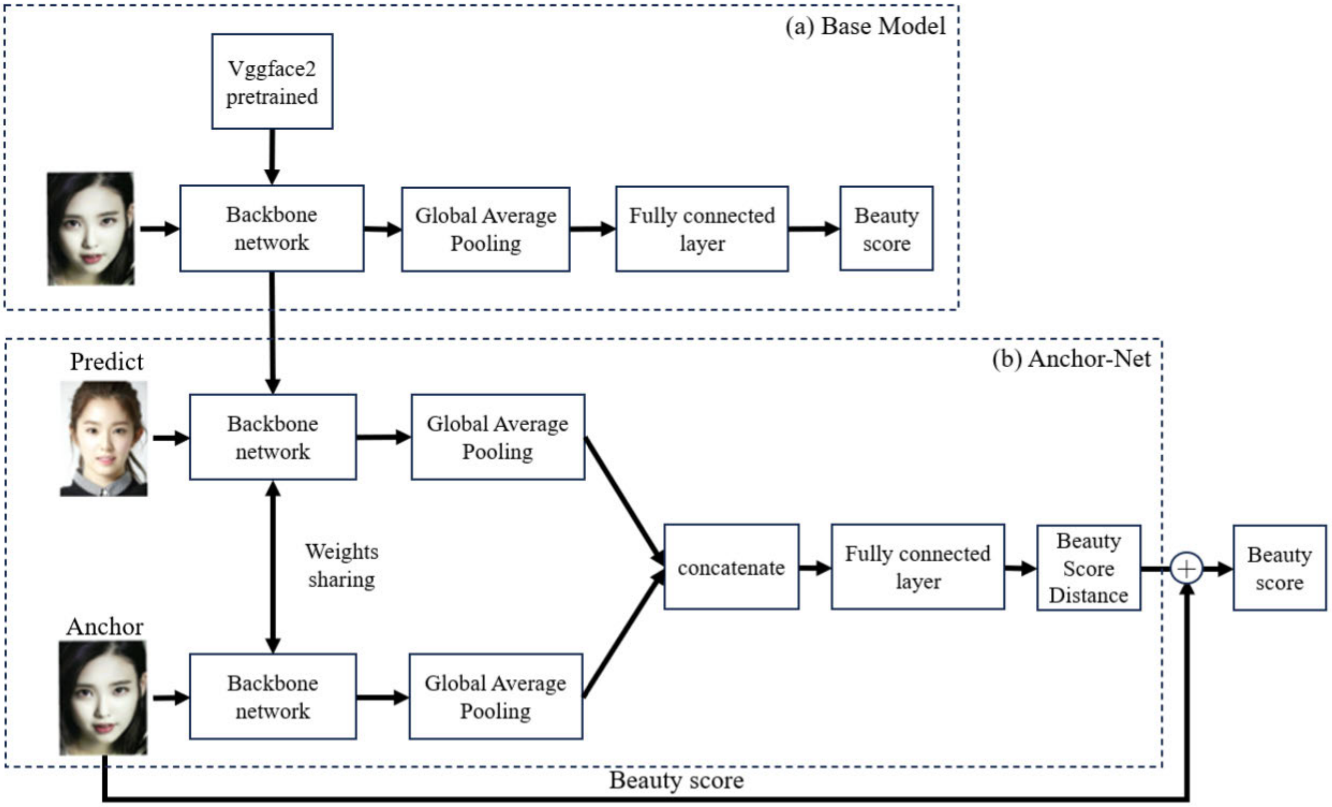

FIGURE 1. Overview of the proposed method.

predictors such as linear regression, Gaussian regression, and support vector regression [15], [35], [36]. To describe facial details, they used a variety of texture features such as Gabor-SIFT [6], [34], [37], [38]. However, their reliance on manually selected features limited their ability to capture the multifaceted beauty of the face.

The evolution of deep learning technologies marked a pivotal shift in FBP research. Deep learning's ability to autonomously extract rich facial features without the need for manual feature selection represented a significant advancement. Yet, while these models offered enhanced predictive power, they often required vast datasets for training and were not inherently designed to account for the relative and subjective aspects of beauty perception [30], [31], [32].

Table. 1 shows deep learning models for FBP prediction compared to Anchor-Net. The ensemble methodology emerged as a promising direction for overcoming some of the limitations of singular model approaches [21], [22], [23], [39], [40]. By leveraging a combination of multiple loss functions [21], diverse convolutional neural network (CNN) models [22], and an amalgamation of different models and loss functions [23], ensemble-based methods demonstrated superior performance on benchmark datasets like SCUT-FBP5500. Despite these advances, ensemble methods still suffer from issues of interpretability and computational efficiency, and are insufficient to capture relative aesthetic features. Anchor-Net, on the other hand, retains the robustness and improved performance of traditional ensemble methods, but is able to capture relative feature using multiple anchors and demonstrates interpretability through anchor images.

Semi-supervised learning models [24], [25], [28], [41], [42], including NFME [24] and MSMFME [25], underscore the value of graph-based and multi-view graph fusion techniques in reinforcing model training without additional labeled images. Despite their innovative approach, these models face limitations in computing similarity graphs before estimating beauty predictions, indicating a gap in capturing the relative aspects of beauty. Anchor-Net's self-supervised learning method bridges this gap by creating new labels and enabling the capture of additional features, such as relative beauty, thereby extending the capabilities of supervised learning methods in a more efficient and effective manner.

Ranking-based models like R2-ResNeXt [26] and R3CNN [27] introduce a novel perspective on FBP by extracting rank-based relative beauty features using pairwise data. While this approach offers insights into relative beauty, ranking features falls short in fully representing the comparative features of two faces. Anchor-Net transcends this limitation by directly comparing the target image to

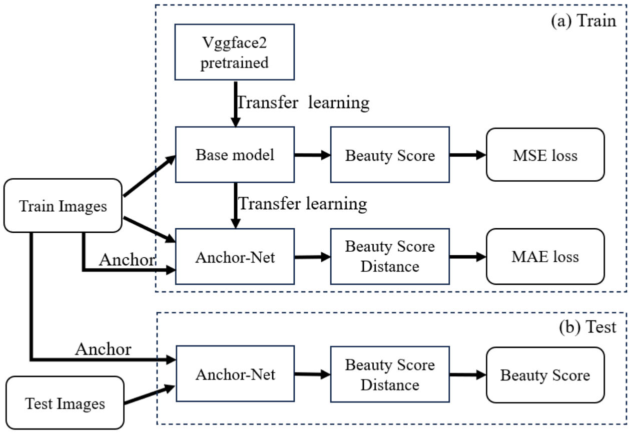

FIGURE 2. Flowcharts of a proposed method.

an anchor image, capturing the nuanced relative features of beauty more accurately and effectively.

In summary, Anchor-Net emerges as a pioneering solution in the FBP domain by innovatively combining the strengths of deep learning, ensemble, and relative-based methodologies. It not only overcomes the inherent limitations of these approaches but also sets a new benchmark for predictive accuracy, relativity in beauty perception.

III. PROPOSED METHOD

The traditional FBP problem, like a typical deep learning regression model [15], takes a single image as input and trains the model to predict the beauty score directly. On the other hand, a triplet network [43] does not directly learn the label of a deep learning model, but indirectly predicts the label of the model by embedding the input image. Inspired by this, Anchor-Net does not learn the beauty score directly, but predicts the beauty score by finding the difference between the beauty score of the Anchor (reference) image and the Prediction image, just as humans evaluate beauty by comparing faces with others. Fig. 1 presents an overview of the proposed method. Then, Fig. 2 shows the training and testing flow of the proposed method.

This Section is organized as follows: Section III-A introduces two-step transfer learning to effectively learn Anchor-Net. Section III-B introduces the structure and training method of Anchor-Net, and Section III-C introduces Anchor Sampling method for selecting Anchor images. Finally, in Section III-D, we introduce the Ensemble method to improve the performance and stability of the model.

A. TWO-STEP TRANSFER LEARNING

Transfer learning proceeds in two steps. Fig. 2 (a) shows the flow of two-step transfer learning. In the first step, the VGGFace2 dataset is used to train the backbone for the model [33], and the learned weights are used to construct the Base model. In the second step, the Base model is trained to predict Beauty Score using the SCUT-FBP5500 dataset and the trained backbone used in the Base model is used as the backbone for the Anchor-Net model.

TABLE 2. The structural details of the Base model.

Name	Output Shape	Parameters	Connected to
Input_0	350×350×3	-	-
Backbone (Resnet50)	2,048	23,561,152	Input_0
global_average_ pooling2d_0	2,048	-	Backbone (Resnet50)
Dense_0	1,024	2,098,176	global_average_ pooling2d_0
Dense_1	512	524,800	Dense_0
Dense_2	256	131,328	Dense_1
Dense_3	128	32,896	Dense_2
Dense_4	64	8,256	Dense_3
Dense_5	32	2,080	Dense_4
Dense_6	16	528	Dense_5
Dense_7	1	17	Dense_6
Total		26,359,233	

Fig. 1 (a) shows the overall structure of the Base model. The Base model is used as the backbone for Anchor-Net. The backbone network was constructed by a CNN such as vgg16 or resnet50 pre-trained with the VGGFace2 dataset [33]. The backbone network is then linked with global mean pooling and a fully connected layer to create the Base model. The Base model was trained with facial beauty scores as labels, which is similar to how a typical FBP model is trained [15]. Table. 2 presents the detailed structures of the Base model. Fine-tuning was performed during the Base model training process. Algorithm. 1 outlines the complete training process of the Base model using Resnet50 as the backbone network. The weights of the backbone network were frozen in training loop 1, and only fully connected layers were trained. The weights of the backbone network were unfrozen in training loop 2, and all weights were trained. We used Adam as the optimizer, a learning rate of 0.0001, and mean squared error (MSE) as the loss function.

B. ANCHOR-NET LEARNING

The backbone network trained by the Base model constructs an Anchor-Net with two images as inputs, as shown in Fig. 1 (b). backbone networks are set up to receive two inputs, and the outputs of each backbone network are connected with a global average pooling layer, concatenated, and linked with a fully connected layer to create a model. Table. 3 presents the detailed structures of the Anchor-Net model. Anchor-Net takes a prediction image and an anchor image as input and trains to output the Beauty Score Distance between the two images as a label. The prediction image is the image for which the model will predict a beauty score. The Anchor image is a reference image trained in the Base model that is used to make predictions by comparing the Prediction image to the beauty score. To get the beauty score of the Prediction image, we find the distance from the Anchor image that the model is already familiar with. In this process, The model can capture the features of the Prediction image that the model is not familiar with, as well as the comparison features of the Anchor image and the Prediction image. Beauty Score Distance, which is the model's label, was calculated by subtracting the facial beauty score of the Anchor image from the facial beauty score of the Prediction image.

Algorithm. 2 outlines the training and validation processes of the Anchor-Net. During training, the training image was used as the Prediction image, and the Anchor image was selected by Anchor Sampling from the training image at every epoch. The predicted Beauty Score Distance was obtained through Anchor-Net, and the beauty score of the Anchor image was subtracted from the beauty score of the Prediction image to obtain the true Beauty Score Distance. The loss function was then trained using the MAE of the predicted and true Beauty Score Distance.

```
Algorithm 1 Base Model Training
  Input:
    $X$: Training images
    $Y$: Corresponding beauty scores for $X$
    $e_1$: Num. of pretraining epochs
    $e_2$: Num. of fine-tuning epochs
    $L_{\text{MSE}}$: MSE loss for training
  Output:
    $M$: Trained Base model for beauty score prediction
  Initialization:
  Pretrain ResNet50 with VGG Face2 data
  Freeze ResNet50 weights, initialize FC layers
  Training Loop1:
  for $i = 1$ to $e_1$ do
    for $(x, y) \in (X, Y)$ do
      $\hat{y} = \text{FC}(\text{ResNet50}(x))$
      $L = L_{\text{MSE}}(\hat{y}, y)$
      Update FC weights to minimize $L$
    end for
  end for
  Training Loop2:
  Unfreeze ResNet50 weights
  for $i = 1$ to $e_2$ do
    for $(x, y) \in (X, Y)$ do
      $\hat{y} = \text{FC}(\text{ResNet50}(x))$
      $L = L_{\text{MSE}}(\hat{y}, y)$
      Update FC and ResNet50 weights to minimize $L$
    end for
  end for
  return $M$
```

TABLE 3. The structural details of Anchor-Net.

Name	Output Shape	Parameters	Connected to
Input_1	350×350×3	-	-
Input_2	350×350×3	-	-
Backbone (Resnet50)	2,048	23,561,152	[Input_1, Input_2]
global_average_ pooling2d_1	2,048	-	Backbone (Resnet50)(0)
global_average_ pooling2d_2	2,048	-	Backbone (Resnet50)(1)
Concatenate_0	4,096	-	[global_average_ pooling2d_1, global_average_ pooling2d_2]
Dense_8	1,024	4,195,328	Concatenate_0
Dense_9	512	524,800	Dense_8
Dense_10	256	131,328	Dense_9
Dense_11	128	32,896	Dense_10
Dense_12	64	8,256	Dense_11
Dense_13	32	2,080	Dense_12
Dense_14	16	528	Dense_13
Dense_15	1	17	Dense_14
Total		28,456,385	

During the Anchor-Net training, the weights of the backbone network were frozen. The backbone network extracts the features for facial beauty from images. The fully connected layer learns the differences in the features of the facial beauty scores extracted from the two images.

For validation, the test image is used as the prediction image, and the Anchor image is the data selected by Anchor Sampling from the training image, as shown in Fig. 2 (b). Anchor-Net's forward pass calculates the Beauty Score Distance. The beauty score of the Anchor image was added to the predicted Beauty Score Distance to predict the beauty score of the test image, as shown in Validation part of Algorithm. 2.

The training mechanism of the Anchor-Net is to learn the model based on the difference in facial beauty between various reference and predict images with which the model is familiar, enabling it to calculate the difference in beauty.

C. ANCHOR SAMPLING

Anchor-Net uses Anchor images to predict beauty scores, making the setting of the Anchor images crucial. This study introduces an Anchor sampling method for sampling Anchor images. Anchor sampling employs two methods: race/gender and error sampling.

1) RACE/GENDER SAMPLING

For Anchor-Net's high performance, it is essential to compute consistent Beauty Score Distance between the images. Our experiments and empirical evidence showed that the race/gender classification ensured consistent differences in beauty scores. The SCUT-FBP5500 dataset [15] was divided into four categories: Asian male (AM), Asian female (AF), Caucasian male (CM), and Caucasian female (CF). From these datasets, we sampled the Anchor images belonging to the same race/gender class as the predict images.

Algorithm 2 Anchor-Net Training and Validation

```
Input:
    $X_{\text{train}}$: Training images
    $Y_{\text{train}}$: Corresponding beauty scores for $X_{\text{train}}$
    $A_{\text{sampling}}$: Anchor sampling method
    $X_{\text{anchor}}$: Reference images sampled from $X_{\text{train}}$ at each epoch
    $Y_{\text{anchor}}$: Corresponding beauty scores for $X_{\text{anchor}}$
    $X_{\text{test}}$: Testing images
    $Y_{\text{test}}$: Corresponding beauty scores for $X_{\text{test}}$
    $e$: Number of training epochs
    $L_{\text{MAE}}$: MAE loss for Anchor-Net learning
Output:
    $\hat{Y}_{\text{test}}$: Predicted beauty scores for $X_{\text{test}}$ by Anchor-Net
Initialization:
Initialize ResNet50 with weights from trained Base model
Freeze ResNet50 weights, initialize FC layers
Main Training Loop:
for epoch $= 1$ to $e$ do
    $X_{\text{anchor}}, Y_{\text{anchor}} = A_{\text{sampling}}(X_{\text{train}}, Y_{\text{train}})$
    for $(x, a, y, a_y) \in (X_{\text{train}}, X_{\text{anchor}}, Y_{\text{train}}, Y_{\text{anchor}})$ do
        $o_{\text{train}} = \text{ResNet50}(x)$
        $o_{\text{anchor}} = \text{ResNet50}(a)$
        $o_{\text{concat}} = \text{Concatenate}(o_{\text{train}}, o_{\text{anchor}})$
        $\hat{d} = \text{FullyConnected}(o_{\text{concat}})$
        $d_{\text{true}} = y - a_y$
        $L = L_{\text{MAE}}(\hat{d}, d_{\text{true}})$
        Update FC weights to minimize $L$
    end for
end for
return Trained Anchor-Net model
Validation:
$\hat{d}_{\text{test}} = \text{Trained Anchor-Net model}(X_{\text{test}}, X_{\text{anchor}})$
$\hat{Y}_{\text{test}} = \hat{d}_{\text{test}} + Y_{\text{anchor}}$
return $\hat{Y}_{\text{test}}$
```

2) ERROR SAMPLING

For an image's FBP using Anchor-Net, the facial beauty score is calculated by adding the true facial beauty score of the Anchor image to the Beauty Score Distance, which is the output of the Anchor-Net. Therefore, the performance of FBP depends on the facial beauty score of the anchor image predicted by the Base model. Consequently, we use a training image with a small prediction error in the Base model as an Anchor image. N training images with low prediction errors were extracted for each race/sex (AM, AF, CM, and CF) from the Base model and used as anchor images.

3) ANCHOR SAMPLING

Anchor sampling was performed using the race/gender and error sampling methods described in Section III-C1 and Section III-C2. The training images were divided into AM, AF, CM, and CF classes, and N images with low errors from the Base model were selected for each class. An anchor image is randomly selected from N images of the same class as the Prediction image. Anchor Sampling was performed on all predicted images at every epoch.

D. ENSEMBLE METHOD

Even with Anchor Sampling for Anchor image selection, the model's performance in test image predictions may be unstable because of the Anchor images' randomness. Ensuring a consistent beauty score distance is crucial even when repeatedly predicting the beauty score for one image. We incorporated an ensemble method to enhance the performance and address this issue.

Using multiple Anchor images for a single image can provide a more reliable and higher accuracy beauty score prediction.This also has the effect of data augmentation. The test image's beauty score was obtained by averaging the beauty score prediction results of the various anchor images selected through anchor sampling. FBP performance and reliability are improved using multiple Anchor images for each image, ensuring a consistent beauty score distance. In our experiments, we used about 20 Anchors for one image to improve performance.

IV. EXPERIMENTS AND ANALYSIS

In Section IV, we validate and discuss the performance of the Anchor-Net distance-based self-supervised learning model. Section IV-A introduces the SCUT-FBP5500 dataset and its benchmark methodology for FBP, which is used to validate the performance of Anchor-Net. Section IV-B presents the performance evaluation metrics of FBP. Section IV-C evaluates the performance of FBP with and without pretraining and by the backbone used in Anchor-Net. Section IV-D evaluates the performance of the Anchor Sampling and emsemble method. In Section IV-E, we compare our performance with state-of-the-art deep learning-based methodologies. In Section IV-F, we present a case analysis of the regression results of Base model and Anchor-Net. Finally, Section IV-G discusses the implications of Anchor-Net.

A. SCUT-FBP5500 DATASET BENCHMARK

The SCUT-FBP5500 dataset contains 5,500 frontal face images of various races, genders, and ages. Fig. 3 shows an example of facial beauty scores by race and gender in SCUT- FBP5500. For each gender/race, such as AM, AF, CM, and CF, a facial attractiveness score ranging from 1 to

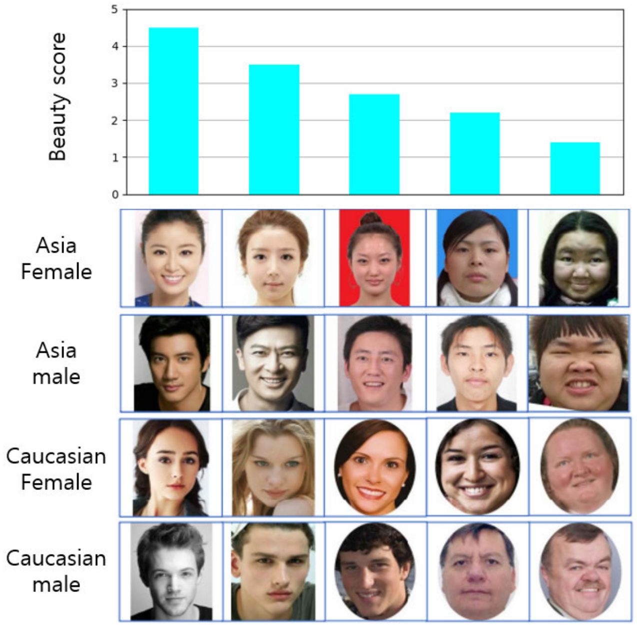

FIGURE 3. Example of facial beauty score by race and gender on SCUT-FBP5500.

5 is provided, with higher scores indicating more attractive images. Sixty workers rated the scores, and the average score of the 60 labels was set as the ground truth.This paper validates the performance of FBP using two methods proposed in [15]. The first method is 5-fold cross-validation, which divides the training and test data by 80:20 (4400 images for training, 1100 images for testing) for each fold. The second method is to divide the training and test data by 60:40 (3300 images for training, 2200 images for testing). This is called 6:4 training-test split.

B. EVALUATION METRICS

This study evaluated FBP using the PC, MAE, and RMSE. For data consisting of N images, these metrics are formulated as follows: eq. (1), eq. (2), eq. (3)

$$PC = \frac{\sum_{i=1}^{N}(y_i - \tilde{y})(p_i - \tilde{p})}{\sqrt{\sum_{i=1}^{N}(y_i - \tilde{y})^2}\sqrt{\sum_{i=1}^{N}(p_i - \tilde{p})^2}} \tag{1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - p_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2} \tag{3}$$

where, y_i and p_i represent the ground truth label and the predicted score of the i-th image, respectively. $\tilde{y}$ is the average score of all ground truth labels and $\tilde{p}$ is the average of the predicted scores. The higher the PC and the smaller the MAE and RMSE, the better the FBP performance.

TABLE 4. Comparison of the backbone model of the SCUT-FBP5500 with Anchor-Net in terms of PC, MAE and RMSE using a 6:4 training-testing split.

Model	Pretrained	PC	MAE	RMSE
AlexNet (Base model)	-	0.7240	0.3673	0.4771
Vgg16 (Base model)	Imagenet	0.8765	0.2550	0.3395
Vgg16 (Base model)	VGGFace2	0.9084	0.2190	0.2871
Resnet18 (Base model)	Imagenet	0.8770	0.2544	0.3305
Resnet50 (Base model)	Imagenet	0.8794	0.2467	0.3242
Resnet50 (Base model)	VGGFace2	0.9131	0.2136	0.2781
AlexNet+Anchor-Net	-	0.7284	0.3639	0.4753
Vgg16+Anchor-Net	Imagenet	0.8865	0.2411	0.3152
Vgg16+Anchor-Net	VGGFace2	0.9003	0.2257	0.2957
Resnet18+Anchor-Net	Imagenet	0.9012	0.2244	0.2943
Resnet50+Anchor-Net	Imagenet	0.9052	0.2201	0.2897
Resnet50+Anchor-Net	VGGFace2	0.9196	0.2034	0.2670

C. EVALUATION OF DIFFERENT BACKBONE ARCHITECTURES FOR ANCHOR-NET

We initially evaluated several CNN models (AlexNet, Vgg16, Resnet18, and Resnet50) as our Base model and the Anchor-Net model as a backbone using the SCUT-FBP5500 benchmark with a 6:4 training-testing split. In addition, some backbone networks utilize pre-trained weights from two datasets, ImageNet and VGGFace2. Anchor-Net does not perform anchor sampling but instead employs random sampling within the batch size at each epoch for training and prediction. Table. 4 shows the performances of the Base model and the Anchor-Net model using the y backbone. Among the Base models, AlexNet without pre-training exhibits the lowest PC, MAE, and RMSE performances. Amongst the models pre-trained with the ImageNet dataset, Vgg16 shows the lowest performance in PC, MAE, and RMSE, and Resnet50 demonstrates the best performance. The vgg16 model pre-trained with the VGGFace2 dataset shows approximately 0.03 PC, 0.04 MAE, and 0.05 RMSE better performance than the model trained with the ImageNet dataset. The Resnet50 model pre-trained with the VGGFace2 dataset exhibits approximately 0.03 PC, 0.03 MAE, and 0.05 RMSE better performance than the model trained with the ImageNet dataset.

Among the Anchor-Net models, the Alexnet+Anchor-Net model exhibited the lowest performance, and the Resnet50+Anchor-Net model pre-trained with the VGGFace2 dataset exhibited the highest performance. Except for the Vgg16+Anchor-Net model, which was pre-trained with the VGGFace2 dataset, all models performed better

TABLE 5. Comparison of Anchor-Net performance average by sampling method and ensemble in terms of PC, MAE, RMSE using 5-fold cross-validation of SCUT-FBP5500.

Pearson Correlation (PC)	Pretrained	Sampling	1	2	3	4	5	Average
Resnet50 (Base model)	VGGFace2	-	0.9132	0.9225	0.9279	0.9279	0.9326	0.9248
Resnet50+Anchor-Net	VGGFace2	Random	0.9258	0.9233	0.9276	0.9299	0.9351	0.9283
Resnet50+Anchor-Net	VGGFace2	Race/Gender	0.9275	0.9258	0.9296	0.9311	0.9355	0.9299
Resnet50+Anchor-Net	VGGFace2	Error	0.9279	0.9277	0.9291	0.9317	0.9382	0.9309
Resnet50+Anchor-Net	VGGFace2	Race/Gender + Error	0.9298	0.9311	0.9323	0.9353	0.9382	0.9333
Resnet50+Anchor-Net +Ensemble	VGGFace2	Race/Gender + Error	0.9306	0.9318	0.9330	0.9357	0.9388	0.9339
Mean Absolute Error (MAE)	**Pretrained**	**Sampling**	**1**	**2**	**3**	**4**	**5**	**Average**
Resnet50 (Base model)	VGGFace2	-	0.2184	0.2143	0.1978	0.1978	0.1901	0.2036
Resnet50+Anchor-Net	VGGFace2	Random	0.1974	0.1997	0.1953	0.1968	0.1865	0.1951
Resnet50+Anchor-Net	VGGFace2	Race/Gender	0.1939	0.1940	0.1910	0.1942	0.1849	0.1916
Resnet50+Anchor-Net	VGGFace2	Error	0.1956	0.1928	0.1910	0.1927	0.1821	0.1908
Resnet50+Anchor-Net	VGGFace2	Race/Gender + Error	0.1924	0.1899	0.1900	0.1885	0.1805	0.1882
Resnet50+Anchor-Net +Ensemble	VGGFace2	Race/Gender + Error	0.1912	0.1891	0.1891	0.1877	0.1794	0.1873
Root Mean Squared Error (RMSE)	**Pretrained**	**Sampling**	**1**	**2**	**3**	**4**	**5**	**Average**
Resnet50 (Base model)	VGGFace2	-	0.2827	0.2764	0.2598	0.2598	0.2480	0.2653
Resnet50+Anchor-Net	VGGFace2	Random	0.2608	0.2679	0.2613	0.2559	0.2449	0.2581
Resnet50+Anchor-Net	VGGFace2	Race/Gender	0.2562	0.2595	0.2579	0.2539	0.2450	0.2545
Resnet50+Anchor-Net	VGGFace2	Error	0.2570	0.2591	0.2584	0.2510	0.2382	0.2527
Resnet50+Anchor-Net	VGGFace2	Race/Gender + Error	0.2540	0.2511	0.2565	0.2445	0.2389	0.2490
Resnet50+Anchor-Net +Ensemble	VGGFace2	Race/Gender + Error	0.2528	0.2500	0.2555	0.2438	0.2378	0.2479

than the existing Base model. The models pre-trained with the VGGFace2 dataset generally exhibited the best performance, and the Anchor-Net model demonstrated improved performance in most backbone networks compared to the Base model.

Resnet50+Anchor-Net model pre-trained with the VGGFace2 dataset exhibited the highest performance. This statement is in line with previous research that has demonstrated the potential of residual networks for the FBP problem [15], [44]. Therefore, we employ the Resnet50+Anchor-Net model pre-trained with the VGGFace2 dataset in the following experiments.

D. EVALUATION OF DIFFERENT SAMPLING AND ENSEMBLE FOR ANCHOR-NET

1) EVALUATION METHODS AND DETAILS

Anchor-Net uses anchor images for training and prediction, making the selection of anchor images vital. We evaluated the model's performance using Anchor-Net by sampling with the Base model Resnet50, pre-trained with VGGFace2, demonstrating the best performance in the previous experiment.

The same model was evaluated using four sampling methods: Random, Race/Gender, Error, and Race/Gender+Error. Additionally, we added a model with Race/Gender+Error and an ensemble to the experiment. The Random method uses a randomly selected image from the training data as the anchor image for each epoch. The Race/Gender method is the same as Race/Gender sampling described in Section III-C1. At each epoch, an anchor image of the same race/gender as the Prediction image was selected for training.

The Error method is the same as that described in Section III-C2. It selects N images with less error from the results predicted by the Base model from the training image. Then, it selects an anchor image corresponding to the predicted image at each epoch for training. In this experiment, we used 45 images with low error for each race/gender in the base model for sampling. The Race/Gender+Error method is the same as the Anchor sampling method in section III-C3. One Anchor image with the same race/sex as the Prediction image in each epoch was selected from the 180 low-error images.

Finally, the model was sampled with Race/Gender+Error and used the ensemble method. The beauty score predicted by the Anchor images sampled 20 times was calculated.

When predicting test data using Anchor-Net, the model's performance can vary significantly depending on the reference image, making it challenging to evaluate the sampling method's performance. Therefore, this experiment calculates the performance of each cross-validation pair as the average of 20 PC, MAE, and RMSE results using each sampling

TABLE 6. Compare Anchor-Net performance with stae-of-the-art methods using 5-fold cross-validation on SCUT-FBP5500.

Model		1	2	3	4	5	Average
Alexnet [15]	PC	0.8667	0.8645	0.8615	0.8678	0.8566	0.8634
	MAE	0.2633	0.2605	0.2681	0.2609	0.2728	0.2651
	RMSE	0.3408	0.3449	0.3538	0.3438	0.3576	0.3481
Resnet-18 [15]	PC	0.8847	0.8792	0.8929	0.8932	0.9004	0.8900
	MAE	0.2480	0.2459	0.2430	0.2383	0.2383	0.2419
	RMSE	0.3258	0.3286	0.3184	0.3107	0.2994	0.3166
ResneXt-50 [15]	PC	0.8985	0.8932	0.9016	0.899	0.9064	0.8997
	MAE	0.2306	0.2285	0.226	0.2349	0.2258	0.2291
	RMSE	0.3025	0.3084	0.3016	0.3044	0.2918	0.3017
PI-CNN [7]	PC	-	-	-	-	-	0.8978
	MAE	-	-	-	-	-	0.2267
	RMSE	-	-	-	-	-	0.3016
ResNet-18 based AaNet [45]	PC	-	-	-	-	-	0.9055
	MAE	-	-	-	-	-	0.2236
	RMSE	-	-	-	-	-	0.2954
Semi-supervised (MSMFME) [25]	PC	-	-	-	-	-	0.9113
	MAE	-	-	-	-	-	0.2210
	RMSE	-	-	-	-	-	0.2870
ResneXt-50-R3CNN [27]	PC	0.9143	0.9066	0.9136	0.9146	0.9217	0.9142
	MAE	0.2109	0.2152	0.2126	0.2130	0.2085	0.2120
	RMSE	0.2767	0.2895	0.2837	0.2804	0.2701	0.2800
CNN-ER [21]	PC	0.9232	0.9204	0.9264	0.9292	0.9257	0.9250
	MAE	0.2026	0.2016	0.2029	0.1990	0.1984	0.2009
	RMSE	0.2667	0.2710	0.2675	0.2583	0.2615	0.2650
FLAC-NET [22]	PC	**0.9325**	0.9220	**0.9357**	0.9259	0.9362	0.9305
	MAE	0.1953	0.2049	0.1990	0.2277	0.1869	0.2028
	RMSE	**0.2486**	0.2727	0.2558	0.2879	0.2419	0.2614
Resnet50+Anchor-Net (ours)	PC	0.9298	0.9311	0.9323	0.9353	0.9382	0.9333
	MAE	0.1924	0.1899	0.1900	0.1885	0.1805	0.1882
	RMSE	0.2540	0.2511	0.2565	0.2445	0.2389	0.2490
Resnet50+Anchor-Net+Ensemble (ours)	PC	0.9306	**0.9318**	0.9330	**0.9357**	**0.9388**	**0.9339**
	MAE	**0.1912**	**0.1891**	**0.1891**	**0.1877**	**0.1794**	**0.1873**
	RMSE	0.2528	**0.2500**	**0.2555**	**0.2438**	**0.2378**	**0.2479**

method for each cross-validation pair, except for the model using the ensemble method.

2) ANALYZE EVALUATION RESULT

Table. 5 presents the results of the 5-fold cross-validation regarding PC, MAE, and RMSE for the sampling and ensemble methods. The Anchor-Net model outperformed the Base model regarding the PC, MAE, and RMSE. The Random method demonstrated the worst performance compared to the other sampling methods regarding the PC, MAE, and RMSE metrics for most cross-validation pairs. The Race/Gender model performs better than the Random model. It performs better than the Error model for some cross-validation pairs, but on average, it performs worse than the Error method. The beauty score differences for race and gender separately ensure more consistent beauty score differences than a randomized approach. The Error method also performed better than the Random method for all the cross-validation pairs. Using an image with fewer errors in the Base model as an anchor image compensates for the error caused by adding the beauty score of the anchor image during the model prediction process. Therefore, the Race/Gender+Error method combines the two methods and shows the highest average performance among all sampling methods except the ensemble method. Sampling with Race/Gender+Error followed by ensemble demonstrates the best performance for all cross-validation pairs and metrics.

E. COMPARISON ANCHOR-NET PERFORMANCE WITH STATE-OF-THE-ART METHODS

We compared the proposed methods with state-of-the-art methods using a 6:4 training-testing split and 5-fold cross-validation on the SCUT-FBP5500 Benchmark. We tested the Resnet50+Anchor-Net model and its ensemble, the Resnet50+Anchor-Net+Ensemble model, which performed the best in the previous experiment.

Table. 6 compares our method with state-of-the-art methods using 5-fold cross-validation and indicates its superior performance on average for PC, MAE, and RMSE. Although some metrics show that our method performs worse than the

TABLE 7. Compare Anchor-Net performance with stae-of-the-art methods using a 6:4 training-testing split on SCUT-FBP5500.

Model	PC	MAE	MAE
LR [15]	0.5948	0.4289	0.5531
GR [15]	0.6738	0.3914	0.5085
SVR [15]	0.6668	0.3898	0.5132
AlexNet [15]	0.8298	0.2938	0.3819
Resnet-18 [15]	0.8513	0.2818	0.3703
ResneXt-50 [15]	0.8777	0.2518	0.3325
CNN-ER [21]	0.9207	0.2032	0.2683
Resnet50 (Base model)	0.9131	0.2136	0.2781
Resnet50+Anchor-Net (ours)	0.9224	0.1984	0.2625
Resnet50+Anchor-Net+Ensemble (ours)	0.9228	0.1977	0.2618

state-of-the-art methods for cross-validation pairs, the overall performance across different training/test data demonstrates that our method outperforms the other methods. The model without the ensemble method also outperformed the state-of-the-art methods, validating our method for the Anchor-Net model.

Table. 7 compares our method with state-of-the-art methods using a 6:4 training-testing split, showing superior performance for all metrics. This shows the effectiveness of the Anchor-Net distance-based self-supervised learning model.

F. REGRESSION ERROR CASE ANALYSIS

In this section, we delve into a comprehensive examination of the predictive performance discrepancies between Anchor-net and Base model. The Base model has the same structure as a general deep learning model and is suitable for case comparison analysis with Anchor-Net. This analysis is predicated on observations from eight subfigures, labeled (a) through (h) in Fig. 4, alongside a detailed exploration presented in Table. 8.

Our first observation pertains to instances where both models in their predictive accuracy. Specifically, Fig. 4a and Fig. 4b illuminate scenarios where each model demonstrated significant errors. A notable trend in these cases is that the inaccuracy of the model was higher in some cases, even though there was more data for East Asian images. This phenomenon suggests a potential area of bias or underrepresentation within the training datasets, underscoring the imperative for more inclusive data sampling in future.

Conversely, Fig. 4c and Fig. 4d showcase instances where Anchor-Net significantly outperformed the Base model. This improvement in performance is attributed to the presence of similar images within the Anchor-Net's Anchor pool, as illustrated by Fig. 4g and Fig. 4h, which served as anchor images for Fig. 4c and Fig. 4d, respectively. The parallels between Fig. 4c and Fig. 4g are primarily found in their distinctive makeup styles and pronounced eyeliner, contributing to their memorable visual impact. Similarly, Fig. 4d and Fig. 4h share notable similarities in nasal structure and overall facial impression, which facilitated the enhanced performance of Anchor-Net over the base model in these cases. This improvement underscores the importance of a diverse and representative anchor image Sampling for refining predictive accuracy.

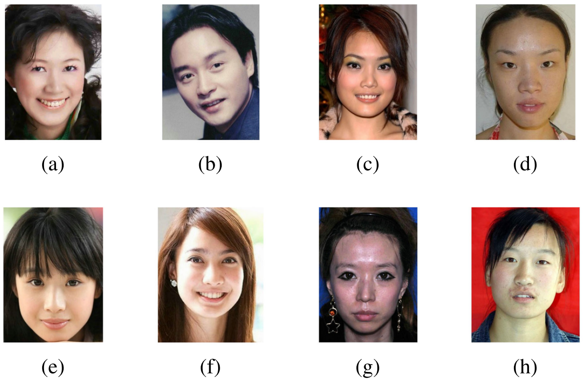

FIGURE 4. Case image Anchor-Net and base model.

The cases represented by Fig. 4e and Fig. 4f stand in stark contrast to the previous examples. In these instances, the absence of visually or structurally similar images in the Anchor-Net's database resulted in a deterioration of performance compared to the base model. This observation highlights a critical limitation of the Anchor-Net approach: the reliance on a sufficiently varied and comprehensive anchor images to support accurate prediction. The lack of analogous facial impressions or features within the anchor image pool for these cases illuminates a gap in the model's ability to generalize from its available references.

This analysis unequivocally emphasizes the pivotal role of facial similarity in the context of FBP problems. The discernible impact of anchor image diversity on predictive performance not only advocates for the expansion of anchor image databases but also calls for a more nuanced understanding of how facial features influence prediction models.

G. DISCUSSION

1) MODEL PERFORMANCE INSTABILITY CAUSED BY ANCHOR IMAGES

Fig. 5 shows the performance distribution of the three metrics (PC, MAE, and RMSE) when predicting the test image using Anchor sampling methods. This illustrates the performance distribution of the facial beauty score predicted using the anchor images selected through Anchor sampling for 100 iterations.

Typically, deep learning models use fixed weights for prediction. Therefore, the model's performance does

TABLE 8. Case Analysis of Anchor-Net and Base model.

Case	Image	Similar Image	True Beauty Score	Anchor Pred Beauty Score	Base Pred Beauty Score
Both models predicted incorrectly	(a)	-	2.23	3.25	3.34
	(b)	-	4.48	3.40	3.42
Anchor-Net predicted better than the Base model	(c)	(g)	3.38	3.45	3.91
	(d)	(h)	1.78	1.80	1.29
Base model predicted better than the Anchor-Net	(e)	-	3.75	3.18	3.55
	(f)	-	3.51	3.11	3.48

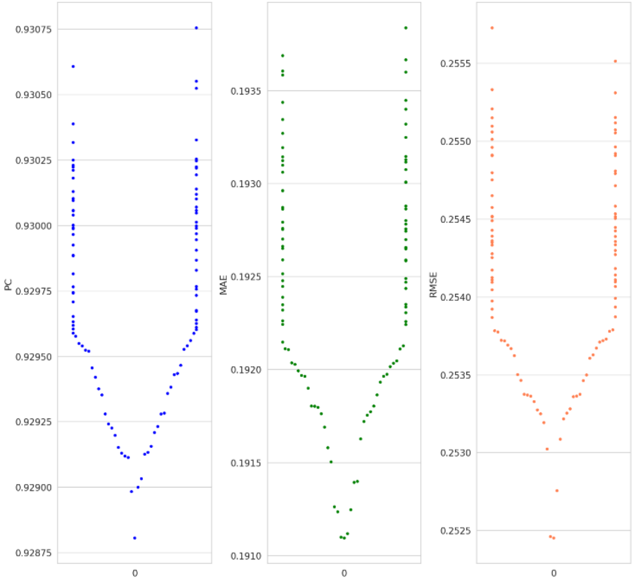

FIGURE 5. Distribution of all evaluation metrics (PC, MAE, RMSE) for 100 iterations of anchor sampling.

TABLE 9. Compare Anchor-Net performance with best/worst result.

Model	PC	MAE	MAE
Resnet50 (Base model)	0.9132	0.2184	0.2827
Resnet50+Anchor-Net	0.9298	0.1924	0.2540
Resnet50+Anchor-Net +Ensemble	0.9306	0.1912	0.2528
Resnet50+Anchor-Net best result	0.9307	0.1910	0.2524
Resnet50+Anchor-Net worst result	0.9288	0.1938	0.2557

not change even if the prediction is repeated. However, for Anchor-Net, there was a significant difference between the maximum and minimum values of the anchor image, with PC = 0.0019, MAE = 0.0028, and RMSE = 0.33. This indicates that Anchor-Net needs improvement.

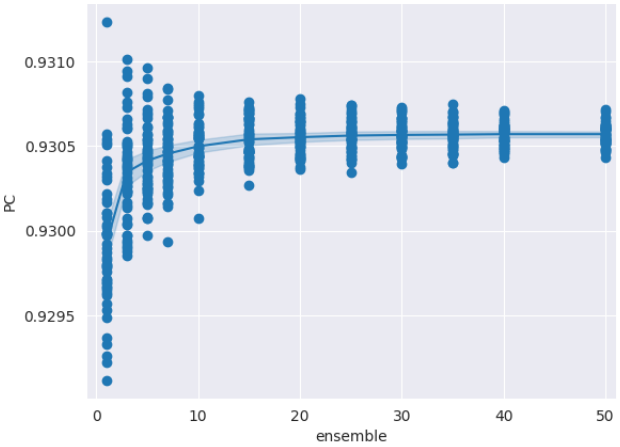

FIGURE 6. Distribution of model performance (PC) over the number of ensembles (Anchors).

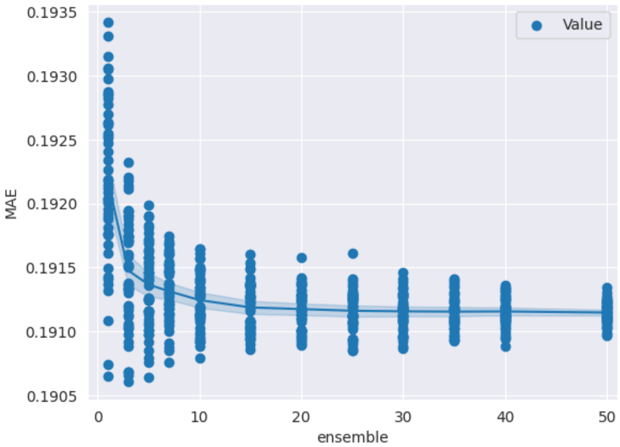

FIGURE 7. Distribution of model performance (MAE) over the number of ensembles (Anchors).

2) DISTRIBUTION OF MODEL PERFORMANCE OVER NUMBER OF ANCHOR

You can use the ensemble method to increase the number of Anchors used in your model. Also, The ensemble method can compensate for the variability in the model performance discussed in the previous section. Fig. 6, Fig. 7 and Fig. 8 show the model performance distribution as a function of the number of ensembles. For example, an Anchor(Ensemble) count of one is equivalent to not running an ensemble, whereas 10 is the average(Ensemble) of the model predictions from the 10 Anchor images.

When the number of ensembles was one, the variance of all the metrics (PC, MAE, and RMSE) was high, and the average performance was poor. The average performance improved as the number of ensembles increased, whereas the variance decreased. This demonstrates the power of data augmentation through ensemble methods. Once 20 ensembles were reached, the variance and average performance were almost

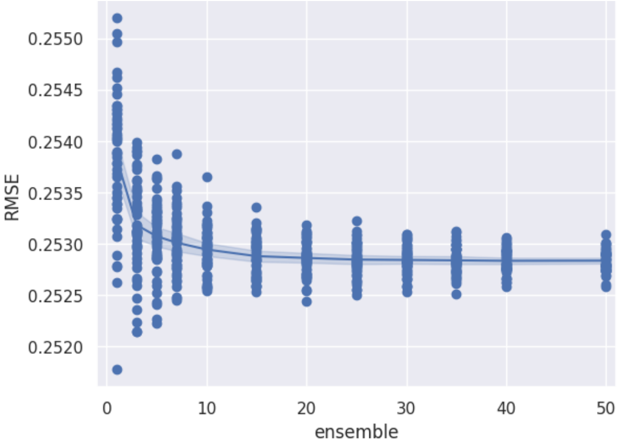

FIGURE 8. Distribution of model performance (RMSE) over the number of ensembles (Anchors).

consistent.This shows that 20 anchors is roughly the optimal number of anchors for computational efficiency.

3) ANCHOR IMAGE SELECTION

Table. 9 shows the best and worst performances of the existing Anchor-Net method in 100 predictions when predicting facial beauty using images selected by anchor sampling. The best performance result exhibited higher performance in PC, MAE, and RMSE than the ensemble method. In contrast, the worst performance result showed higher performance than the Base model but worse than other Anchor-Net methods. Improvement in the anchor-sampling method to select the appropriate anchor image for the prediction image is expected to compensate for the shortcomings of the existing Anchor-Net. This would be an interesting direction for future studies.

V. CONCLUSION

This paper introduces Anchor-Net, a self-supervised learning model, as a novel approach to quantifying aesthetics. The model can distinguish subtle differences in facial beauty perception by comparing the beauty scores of prediction and anchor images. By implementing a distance-based method with two-step transfer learning, it is possible to learn about the relative nature of beauty perception through comparisons rather than absolute metrics. Furthermore, an anchor sampling method was used to select suitable anchor images. Finally, the ensemble method was applied to enhance the model performance and ensure consistency.

Through extensive experiments, including 5-fold cross-validation and a 6:4 training-testing split on the SCUT-FBP5500 dataset, we demonstrated that Anchor-Net could be integrated into various backbone networks and validated the anchor sampling method. Our results also reveal that Anchor-Net outperforms state-of-the-art deep-learning-based methods regarding key performance metrics such as PC, MAE, and RMSE. In our discussion, we discovered that selecting suitable anchor images can further enhance the model's performance and stability, which is a worthwhile consideration for future research.

REFERENCES

[1] K. Dion, E. Berscheid, and E. Walster, "What is beautiful is good," *J. Personality Social Psychol.*, vol. 24, no. 3, p. 285, 1972.

[2] A. C. Little, B. C. Jones, and L. M. DeBruine, "Facial attractiveness: Evolutionary based research," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 366, no. 1571, pp. 1638–1659, Jun. 2011.

[3] L. Liang, L. Jin, and X. Li, "Facial skin beautification using adaptive region-aware masks," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2600–2612, Dec. 2014.

[4] J. Li, C. Xiong, L. Liu, X. Shu, and S. Yan, "Deep face beautification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 793–794.

[5] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Data-driven enhancement of facial attractiveness," in *Proc. ACM SIGGRAPH Papers*, 2008, pp. 1–9.

[6] D. Zhang, F. Chen, and Y. Xu, *Computer Models for Facial Beauty Analysis*. Switzerland: Springer, 2016.

[7] J. Xu, L. Jin, L. Liang, Z. Feng, D. Xie, and H. Mao, "Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN)," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1657–1661.

[8] L. Liang, L. Jin, and D. Liu, "Edge-aware label propagation for mobile facial enhancement on the cloud," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 1, pp. 125–138, Jan. 2017.

[9] H. Gunes, "A survey of perception and computation of human beauty," in *Proc. Joint ACM Workshop Hum. Gesture Behav. Understand.*, 2011, pp. 19–24.

[10] Y. Eisenthal, G. Dror, and E. Ruppin, "Facial attractiveness: Beauty and the machine," *Neural Comput.*, vol. 18, no. 1, pp. 119–142, Jan. 2006.

[11] A. Kagian, G. Dror, T. Leyvand, D. Cohen-Or, and E. Ruppin, "A humanlike predictor of facial attractiveness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2006, pp. 1–8.

[12] H. Mao, L. Jin, and M. Du, "Automatic classification of Chinese female facial beauty using support vector machine," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2009, pp. 4842–4846.

[13] D. Zhang, Q. Zhao, and F. Chen, "Quantitative analysis of human facial beauty using geometric features," *Pattern Recognit.*, vol. 44, no. 4, pp. 940–950, Apr. 2011.

[14] D. Xie, L. Liang, L. Jin, J. Xu, and M. Li, "SCUT-FBP: A benchmark dataset for facial beauty perception," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2015, pp. 1821–1826.

[15] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li, "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1598–1603.

[16] L. Liang, D. Xie, L. Jin, J. Xu, M. Li, and L. Lin, "Region-aware scattering convolution networks for facial beauty prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2861–2865.

[17] J. Gan, L. Li, Y. Zhai, and Y. Liu, "Deep self-taught learning for facial beauty prediction," *Neurocomputing*, vol. 144, pp. 295–303, Nov. 2014.

[18] S. Wang, M. Shao, and Y. Fu, "Attractive or not?: Beauty prediction with attractiveness-aware encoders and robust late fusion," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 805–808.

[19] Y. Ren and X. Geng, "Sense beauty by label distribution learning," in *Proc. IJCAI*, 2017, pp. 2648–2654.

[20] R. White, A. Eden, and M. Maire, "Automatic prediction of human attractiveness," *UC Berkeley CS280A Project*, vol. 1, p. 2, 2004.

[21] J. N. Saeed, A. M. Abdulazeez, and D. A. Ibrahim, "Automatic facial aesthetic prediction based on deep learning with loss ensembles," *Appl. Sci.*, vol. 13, no. 17, p. 9728, Aug. 2023.

[22] D. E. Boukhari, A. Chemsa, A. Taleb-Ahmed, R. Ajgou, and M. Bouzaher, "Facial beauty prediction using an ensemble of deep convolutional neural networks," *Eng. Proc.*, vol. 56, no. 1, p. 125, 2023.

[23] F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed, "Deep learning based face beauty prediction via dynamic robust losses and ensemble regression," *Knowl.-Based Syst.*, vol. 242, Apr. 2022, Art. no. 108246.

[24] F. Dornaika, K. Wang, I. Arganda-Carreras, A. Elorza, and A. Moujahid, "Toward graph-based semi-supervised face beauty prediction," *Expert Syst. Appl.*, vol. 142, Mar. 2020, Art. no. 112990.

[25] F. Dornaika and A. Moujahid, "Multi-view graph fusion for semi-supervised learning: Application to image-based face beauty prediction," *Algorithms*, vol. 15, no. 6, p. 207, Jun. 2022.

[26] L. Lin, L. Liang, and L. Jin, "R2-ResNeXt: A ResNeXt-based regression model with relative ranking for facial beauty prediction," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 85–90.

[27] L. Lin, L. Liang, and L. Jin, "Regression guided by relative ranking using convolutional neural network (R^3CNN) for facial beauty prediction," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 122–134, Jan./Mar. 2020.

[28] F. Dornaika, A. Elorza, K. Wang, and I. Arganda-Carreras, "Nonlinear, flexible, semisupervised learning scheme for face beauty scoring," *J. Electron. Imag.*, vol. 28, no. 4, 2019, Art. no. 043013.

[29] E. Siahaan, J. A. Redi, and A. Hanjalic, "Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal," in *Proc. 6th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 245–250.

[30] G. U. Hayn-Leichsenring, T. Lehmann, and C. Redies, "Subjective ratings of beauty and aesthetics: Correlations with statistical image properties in western oil paintings," *i-Perception*, vol. 8, no. 3, Jun. 2017, Art. no. 204166951771547.

[31] D. M. Sidhu, K. H. McDougall, S. T. Jalava, and G. E. Bodner, "Prediction of beauty and liking ratings for abstract and representational paintings using subjective and objective measures," *PLoS ONE*, vol. 13, no. 7, Jul. 2018, Art. no. e0200431.

[32] B. J. Schabel, L. Franchi, T. Baccetti, and J. A. McNamara, "Subjective vs objective evaluations of smile esthetics," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 135, no. 4, pp. S72–S79, Apr. 2009.

[33] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[34] H. Yan, "Cost-sensitive ordinal regression for fully automatic facial beauty assessment," *Neurocomputing*, vol. 129, pp. 334–342, Apr. 2014.

[35] K. Schmid, D. Marx, and A. Samal, "Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios," *Pattern Recognit.*, vol. 41, no. 8, pp. 2710–2717, Aug. 2008.

[36] F. Chen and D. Zhang, "A benchmark for geometric facial beauty study," in *Proc. Int. Conf. Med. Biometrics*. Berlin, Germany: Springer, 2010, pp. 21–32.

[37] J. Whitehill and J. R. Movellan, "Personalized facial attractiveness prediction," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–7.

[38] H. Altwaijry and S. Belongie, "Relative ranking of facial attractiveness," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 117–124.

[39] J. N. Saeed, A. M. Abdulazeez, and D. A. Ibrahim, "An ensemble DCNNs-based regression model for automatic facial beauty prediction and analyzation," *Traitement du Signal*, vol. 40, no. 1, p. 55, 2023.

[40] E. Vahdati and C. Y. Suen, "Female facial beauty analysis using transfer learning and stacking ensemble model," in *Proc. 16th Int. Conf. Image Anal. Recognit. (ICIAR)*. Waterloo, ON, Canada: Springer, 2019, pp. 255–268.

[41] F. Dornaika, A. Elorza, K. Wang, and I. Arganda-Carreras, "Image-based face beauty analysis via graph-based semi-supervised learning," *Multimedia Tools Appl.*, vol. 79, nos. 3–4, pp. 3005–3030, Jan. 2020.

[42] Y. Xiao, L. Zhang, B. Liu, R. Cai, and Z. Hao, "Multi-task ordinal regression with labeled and unlabeled data," *Inf. Sci.*, vol. 649, Nov. 2023, Art. no. 119669.

[43] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. 3rd Int. Workshop Similarity-Based Pattern Recognit. (SIMBAD)*. Copenhagen, Denmark: Springer, 2015, pp. 84–92.

[44] Y.-Y. Fan, S. Liu, B. Li, Z. Guo, A. Samal, J. Wan, and S. Z. Li, "Label distribution-based facial attractiveness computation by deep residual learning," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2196–2208, Aug. 2018.

[45] L. Lin, L. Liang, L. Jin, and W. Chen, "Attribute-aware convolutional neural networks for facial beauty prediction," in *Proc. IJCAI*, 2019, pp. 847–853.

JIHO BAE is currently pursuing the bachelor's degree with the Department of Computer Science, Gyeongsang National University, Jinju-si, South Korea. His research interests include artificial intelligence, computer vision, human–computer interaction, and augmented reality.

SEOK-JUN BUU (Member, IEEE) received the Ph.D. degree in computer science from Yonsei University, South Korea. Since 2023, he has been an Assistant Professor with the Department of Computer Science, Gyeongsang National University. He works primarily in the field of deep learning as a practical approach to solve a broad range of industrial problems in an empirical form. He is exploring how to leverage the domain knowledge and inject the real-world constraints into deep learning applications.

SUWON LEE (Member, IEEE) received the Ph.D. degree in computer science from KAIST, Daejeon, South Korea, in 2017. Since 2018, he has been a Professor with the Department of Computer Science, Gyeongsang National University, Jinju-si, South Korea. His research interests include artificial intelligence, computer vision, human–computer interaction, and augmented reality.