Supplementary Material of DSL-FIQA: Assessing Facial Image Quality via Dual-Set Degradation Learning and Landmark-Guided Transformer

Wei-Ting Chen[1,2†] Gurunandan Krishnan[2] Qiang Gao[2] Sy-Yen Kuo[1] Sizhou Ma[2*] Jian Wang[2*♦]

[1]National Taiwan University [2]Snap Inc.

1. Social Impact and Ethical Considerations

This paper contributes to the advancement of Face IQA technology, which has widespread applications in digital media and social networking platforms. By ensuring a balanced representation of gender and skin tones in the dataset, this paper addresses critical issues of fairness and bias in AI, promoting more equitable facial analysis technologies.

However, if the facial image quality assessment (IQA) method fails, it could lead to the selection of incorrect facial quality images for training, subsequently affecting the accuracy of facial-related algorithms trained on these images. This situation might result in biases or errors in facial recognition, emotion analysis, or other applications based on facial images.

To address this issue, an effective approach is to double-check the images filtered through the Face IQA method to ensure their quality meets the expected standards. This can be achieved through manual review or by employing additional verification mechanisms. Such a double-checking mechanism helps reduce the risk of erroneously selected images, ensuring the quality of training data, thereby enhancing the reliability and effectiveness of algorithms trained on these data.

2. Comprehensive Generic Face IQA Dataset

2.1. Data Collection

To create a diverse and comprehensive dataset, we initially collected face images from the CelebA dataset. We utilized skin tone [2] and gender [1] detectors to analyze these images, ensuring a balanced representation of both gender and skin tones. This careful sampling approach was complemented by the addition of 1,028 images from Flickr, specifically chosen to enhance the diversity in terms of skin tones and occlusion. The combined dataset consists of about 40,000 images. Each image was aligned using Dlib's face landmark detection [11, 12, 15] according to FFHQ dataset [10] and subsequently rescaled to a uniform resolution of 512×512 pixels, ensuring consistency across the dataset.

2.2. Annotation Procedure

To ensure accurate and consistent subjective quality assessment of facial images, we provided annotators with a user-friendly and intuitive interface. This interface was designed to display one facial image at a time, accompanied by an input field for annotators to enter their Mean Opinion Score (MOS) for the image. To assist annotators in making accurate judgments, we included example images for each quality level alongside the interface. These examples served as references, aiding annotators in better discerning and assessing the quality of each image. Additionally, our system supports arbitrary zoom-in and out features for each image, allowing annotators to better assess the details. An illustration of the user interface used in our study is shown in Figure 1.

For the subjective scoring process, we adopted the standard 5-interval Absolute Category Rating (ACR) scale, comprising levels: Bad, Poor, Fair, Good, and Excellent. This scale was linearly mapped to a range of [0, 1.0], corresponding to the ACR scale as follows: Bad at 0-20%, Poor at 20-40%, Fair at 40-60%, Good at 60-80%, and Excellent at 80-100%.

To elevate the precision and uniformity of the evaluations, we crafted detailed guidelines alongside a collection of definitive gold-standard principles. These encompassed several facets of image analysis, such as the visibility of eyelashes, articulated through specific classification tiers:

- Excellent: No visible artifacts, whether viewed as thumbnails or in original size.
- Good: Artifacts discernible solely at original size.
- Fair: Minor artifacts noticeable in thumbnail views.

[†] Part of the work done during internship at Snap Research.

[*] Co-corresponding authors

[♦] Project lead

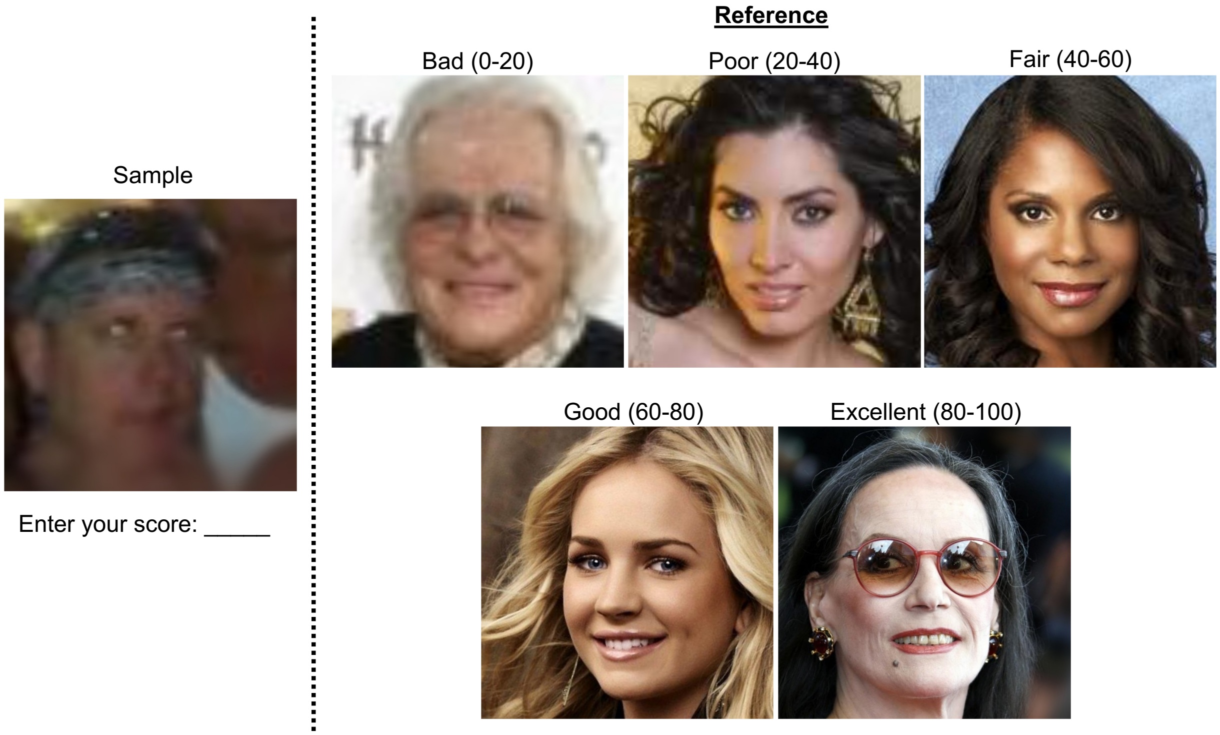

Figure 1. **User Interface of the Subjective Generic Face IQA Study.** Participants assess each image's visual quality by entering the scores in a toolbox.

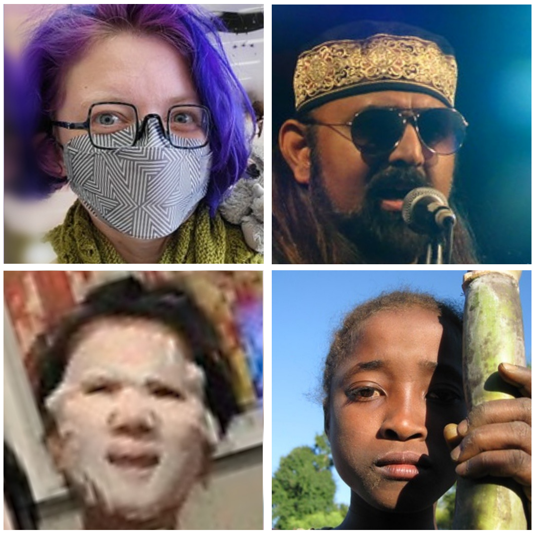

Figure 2. **Examples of occluded images in CGFIQA-40K dataset.**

Reference images from the GFIQA-20k dataset were instrumental in guiding the annotators.

Additionally, our guidance provides a structure for using midpoint scores when an image does not clearly fit into a single category. For instance, if an image falls between the "Poor" and "Fair" categories, a midpoint score of 40 is recommended.

We curated a collection of 35 images carefully selected by experts, where each of the five quality intervals is represented by seven images. Three images from each level were used as golden samples, which were provided to guide each annotator along with the rating guidelines. Additionally, we conducted a pre-annotation training using the remaining 20 images, with four images from each quality level (It is unknown to the annotators that they were evenly distributed). Annotators were required to achieve an accuracy of at least 80% in this test to complete their training. To clarify, an annotator's assessment was considered correct if their assigned Mean Opinion Score (MOS) was within a margin of ±15 points from the ground truth MOS score. If this criterion was not met, they were asked to revisit the guidelines and 15 example images and then retake the test until they reached the accuracy threshold. Importantly, annotators were not informed of the correct answers to the test questions throughout the process.

In total, we engaged 20 annotators for this study. On average, each annotator spent approximately 30 seconds assessing the quality of each image. This arrangement ensured both the ratings' efficiency, quality, and consistency. These detailed guidelines and scoring mechanisms ensured that participants could accurately and consistently assess image quality, thereby enhancing our dataset's overall quality and reliability.

2.3. Dataset Overview

In this section, we delve into the CGFIQA-40k dataset, which is comprised of 40,000 face images, each meticulously annotated with a Mean Opinion Score (MOS). This dataset represents a comprehensive collection, covering a broad spectrum of image quality with MOS values ranging from 0 to 1.

The CGFIQA-40k dataset is specifically curated to focus on facial images, showcasing various visual qualities, including several images with occlusions. As illustrated in Figure 2, these occluded images are integral to the dataset, contributing to its diversity and providing edge cases for robust model training. We have included image samples from different categories - Excellent, Good, Fair, Poor, and Bad to demonstrate the overall diversity. From each category, as shown in Figure 3, six images have been carefully selected to represent the range of qualities within that category. These images and their respective MOS values are displayed in the accompanying figures, illustrating the perceptual quality differences across categories.

Furthermore, we present a histogram of the MOS distribution in Figure 4 for the entire dataset. This histogram provides a clear overview of the quality distribution of the images, highlighting the frequency and range of different quality levels within the dataset.

3. Implementation Details

3.1. Evaluation Criteria

In our evaluation, we use two well-established metrics to assess the performance of our model: Spearman's Rank-Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC).

PLCC measures the linear correlation between actual and predicted quality scores, indicating how closely the predictions align with real values on a linear scale. It is sensitive to numerical differences between scores.

SRCC, in contrast, evaluates the monotonic relationship between two datasets. It focuses on rank order rather than numerical values, offering robustness against outliers and skewed distributions. Both metrics range from -1 to 1, where 1 signifies perfect correlation, -1 indicates perfect inverse correlation, and 0 means no linear correlation. Higher absolute values indicate better performance, with positive values showing consistency with the ground truth.

For the PLCC, given s_i and $\hat{s}_i$ as the actual and predicted quality scores for the i-th image, and μ_{s_i} and $\mu_{\hat{s}_i}$ as their means, with N as the number of test images, it is defined as:

$$\text{PLCC} = \frac{\sum_{i=1}^{N}(s_i - \mu_{s_i})(\hat{s}_i - \mu_{\hat{s}_i})}{\sqrt{\sum_{i=1}^{N}(s_i - \mu_{s_i})^2}\sqrt{\sum_{i=1}^{N}(\hat{s}_i - \mu_{\hat{s}_i})^2}}, \quad (1)$$

For SRCC, where d_i is the rank difference of the i-th test image in the ground truth and predicted scores, it is given by:

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \quad (2)$$

Both PLCC and SRCC provide insights into the model's performance, with higher values indicating better accuracy and consistency with the ground truth.

3.2. Training Details

Degradation Encoder. The Degradation Encoder is tailored to extract and encode degradation features inherent in the input face images. Our architecture employs a CNN comprising six 3×3 convolution blocks. Each block incorporates batch normalization and is succeeded by a leaky ReLU activation. After feature extraction, these features are processed through a two-layer MLP to produce the final degradation representation vector. We use the Adam optimizer with a learning rate of 3×10^{-5} across 300 epochs for training. Our training data is divided into two distinct sets. The first set, labeled as *Set* $\mathcal{S}$, consists of m images, as mentioned in Section 3.2 of the main paper. These are derived from 5000 high-quality images from the FFHQ dataset [10], resized to 512×512. The images in this set are subjected to 15 different synthesized degradations, while one image remains undegraded, resulting in a total of 16 images (*i.e.,* $m = 16$). The synthesized degradations encompass a variety of conditions such as low-light, high-light, blur, defocus, 2x downsample, Gaussian noise, Gaussian blur with kernel sizes from 3 to 31, JPEG compression quality ranging from 1 to 30, motion blur, sun flare, ISO noise, shadow, and zoom blur. The low-light and high-light degradations are implemented using the torchvision library, whereas the other degradations are applied using albumentations [3]. The second set, designated as *Set* $\mathcal{R}$, includes n images, amounting to 256 as specified in Section 3.2.2 of the main paper. This set is dynamically curated by selecting from the GFIQA-20k dataset, ensuring that each subset of 256 images contains at least one high-quality face image with a Mean Opinion Score (MOS) greater than 0.9. The temperature parameter θ is 1.0. Notably, both sets undergo resampling in each iteration to ensure a diverse training experience. This module comprises a total of 1.27×10^6 parameters. The training process was conducted on a single NVIDIA A100 GPU, equipped with 80GB of memory, using the PyTorch framework. The entire training was completed in roughly 12 hours.

Landmark Detection Network. We used a commercial

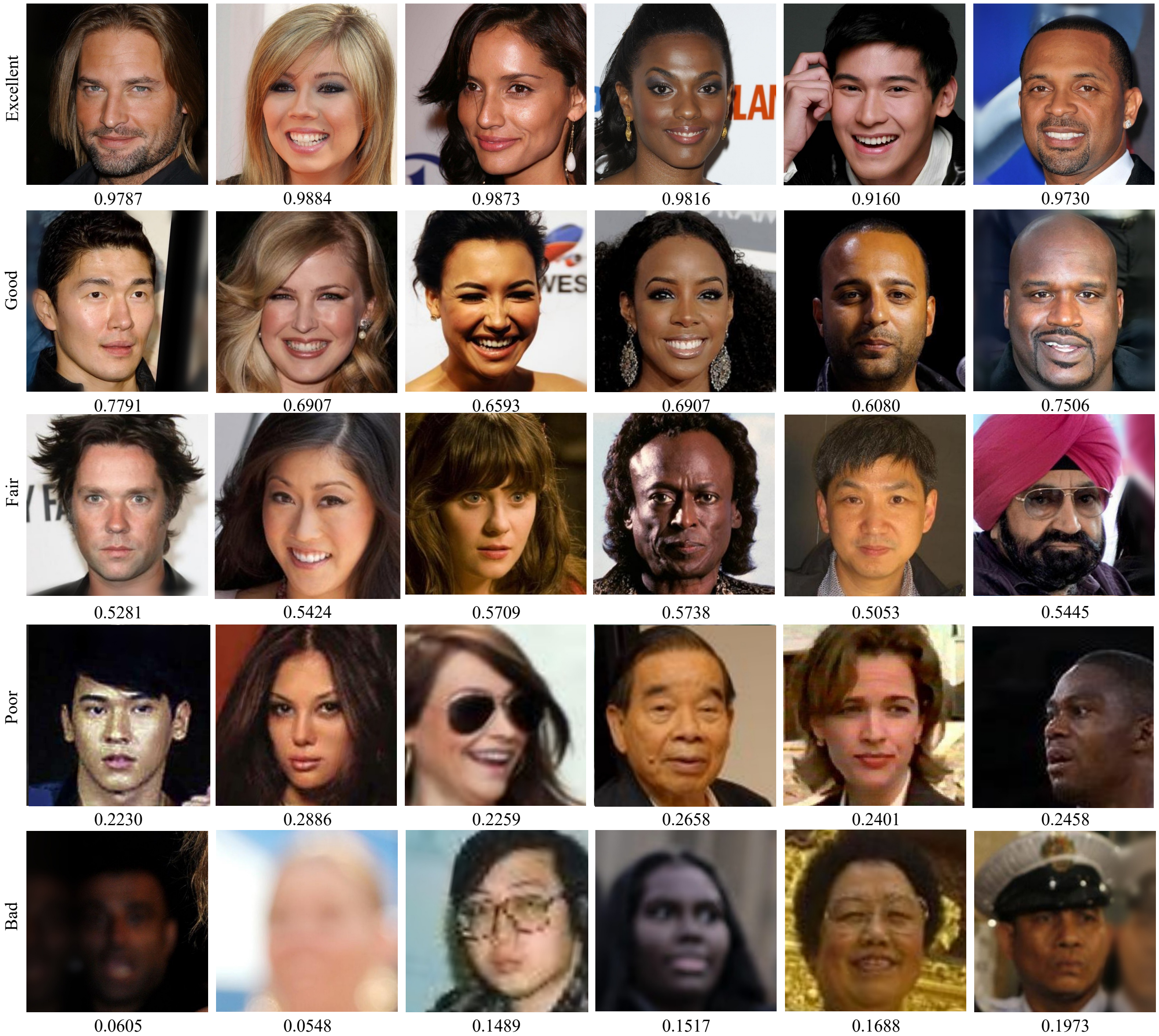

Figure 3. **Sampled face images from the CGFIQA-40k dataset.** These images showcase the diversity of visual quality across five categories: Excellent ((0.8,1]), Good ((0.6,0.8]), Fair ((0.4,0.6]), Poor ((0.2,0.4]), and Bad ([0,0.2]). Each category is represented by six randomly selected images, annotated with their corresponding Mean Opinion Scores (MOS).

implementation of [5] which outputs 1313 landmarks by fitting the 3DMM model [7] on the initially detected 68 landmarks. We have observed that the original face landmark detection algorithm does not perform well on low-quality images. However, when fine-tuned specifically for low-quality images, it significantly improves performance, as shown in Figure 5. These low-quality images are synthesized based on the image degradation model [9] on the current landmark detection dataset.

GFIQA Network. The GFIQA Network, informed by the features extracted by the Degradation Encoder, endeavors to predict the Mean Opinion Score (MOS) for input face images. Our network architecture combines a hybrid CNN-Transformer backbone, comprising a VGG-19 model pre-trained on ImageNet [13], and a Vision Transformer (ViT) backbone [6], also pre-trained on ImageNet. This setup is further enhanced with two Swin Transformer blocks [14], a channel attention layer [8], a transformer decoder, and two MLP regression layers. The ViT backbone, tailored for an input size of 384 × 384, processes the image by dividing it into multiple 16 × 16 pixel patches, ensuring detailed and

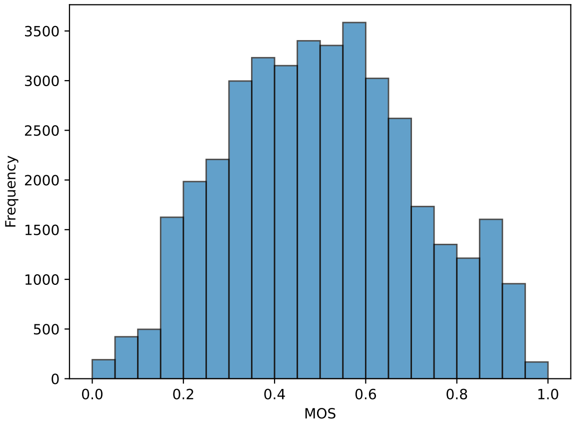

Figure 4. **Distribution of the CGFIQA-40K dataset in terms of MOS scores.**

comprehensive image analysis. During training, we employ a batch size of 16, and all input images undergo random cropping from 512 × 512 to 384 × 384. Additionally, data augmentation in the form of random horizontal flipping is applied to enhance the model's generalization capability. The learning rate is set at 10^{-5} across 100 epochs, and we use the Adam optimizer. The ϵ in $\mathcal{L}_{char}$ is 10^{-3}. The module consists of 2.51×10^{8} parameters in total. The network was trained on an Nvidia A100 GPU, which has 80GB of memory, using the PyTorch framework. The entire training process was completed within 20 hours.

Clarification. To clarify, in our system, both the degradation extraction network and the landmark detection network process the entire image (512 × 512 pixels) to predict landmarks and extract degradation representations. However, for the GFIQA network, we adapt to the input size requirements of the pre-trained Vision Transformer (ViT), which is 384 × 384 pixels in our implementation. To accommodate this, we crop the facial image into several overlapping 384 × 384 patches, each serving as an individual input for the ViT. This ensures that the total coverage area of all patches encompasses the original input image.

In the main paper, particularly in Fig. 2, we simplified the explanation by omitting the step of cropping the facial image into multiple patches. Moreover, the images outlined in red in the GFIQA Network section are intended to illustrate how the ViT divides the input image (384 × 384) into several patches for feature extraction between patches.

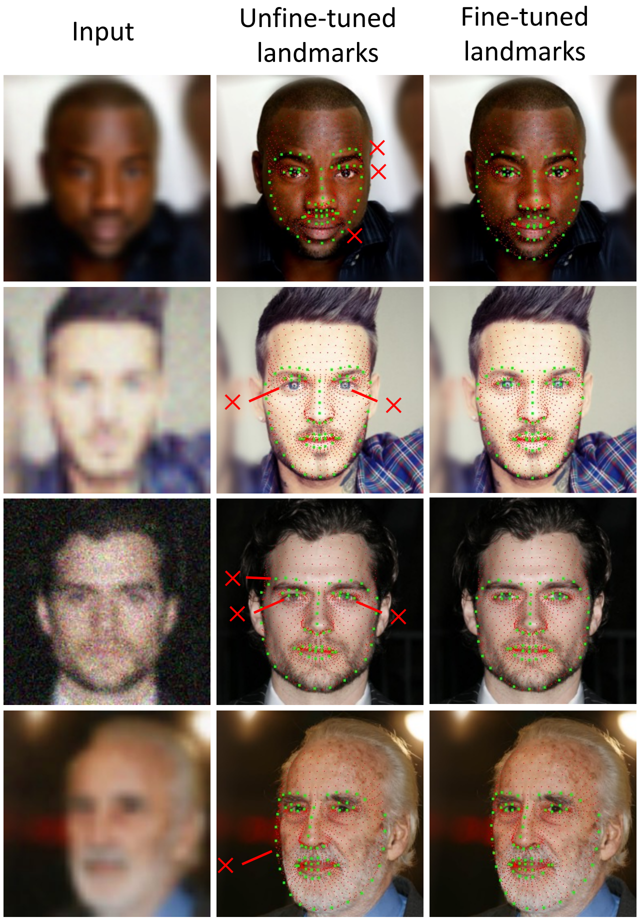

Figure 5. **Evaluating the Impact of Fine-Tuning on Landmark Detection in Poor-Quality Images.** The fine-tuned landmark detection algorithm can handle low-quality inputs (first column), as demonstrated in the third column of results. In contrast, the unfine-tuned algorithm has large errors, as evidenced in the second column (highlighted by the red crosses). The detected landmarks have been overlaid on the high-quality version of the input for better visualization. The basic 68 landmarks are represented by green dots, while the expanded set of 1313 landmarks is denoted by small red dots.

4. More Experimental Results

4.1. Cross-Dataset Validation

To explore the quality attributes of facial data, we conducted an experiment using our newly proposed CGFIQA-40k dataset and the existing GFIQA-20k [16] dataset to train models. In this experiment, we employed the StyleGAN-IQA model [16] and our method for training. The effectiveness of these models was then verified on the PIQ23 dataset [4], a benchmark for unseen face image quality assessment.

As shown in Table 1, we observed that models trained on

Dataset/Model	PLCC	SRCC
GFIQA-20k/StyleGAN-IQA	0.3323	0.3421
CGFIQA-40k/StyleGAN-IQA	0.3541	0.3643
GFIQA-20k/Ours	0.3947	0.4165
CGFIQA-40k/Ours	**0.4229**	**0.4653**

Table 1. **Performance Comparison of Zero-shot GFIQA on PIQ23 [4] Dataset.** This table compares the effectiveness of models trained on CGFIQA-40k and GFIQA-20k datasets. The results highlight the superior performance of models using CGFIQA-40k, underscoring its larger scale and balanced diversity in gender and skin tones.

Strategy	Naive	Patch-based	DSL
mAP	39.21	52.30	**72.1**

Table 2. **Comparative degradation retrieval accuracy using DSL, patch-based, naive methods under real-world degradations, quantified by mAP scores.**

our datasets, particularly the CGFIQA-40k, demonstrated superior performance on the PIQ23 dataset[1], an unseen face image quality dataset. This enhanced performance can be attributed to several key factors. Firstly, the CGFIQA-40k dataset is extensive in scale, encompassing a wide range of image qualities and scenarios. Secondly, and crucially, it offers a more balanced representation in terms of gender and skin tone compared to the GFIQA-20k dataset. This balanced representation ensures a more comprehensive and unbiased training process, leading to models that are better equipped to handle a diverse array of facial images in real-world applications. The results clearly highlight the advantages of our dataset, underscoring its potential in advancing the field of facial image quality assessment.

4.2. More Ablation Studies

Effectiveness of DSL. In our experiments, we compared two approaches to validate the effectiveness of our dual-set design in contrastive learning. The first approach, which we refer to as the “Naive method”, involves training a model exclusively on the synthetic set (*Set* $\mathcal{S}$). In this method, positive pairs are formed from images with identical synthetic degradations, while negative pairs are composed of images with different degradations. This approach, however, showed limitations in generalizing to real-world images due to its sole reliance on synthetic degradations.

In contrast, our dual-set model integrates both synthetic (*Set* $\mathcal{S}$) and real-world (*Set* $\mathcal{R}$) degradations. This model is trained to recognize and adapt to a broader range of degradation patterns, encompassing both controlled synthetic and naturally occurring real-world degradations. As a result, it demonstrated superior generalization capabilities, particularly in diverse real-world scenarios. The comparative performance of these two approaches is detailed in Table 2, highlighting the significant advantage of our dual-set approach in achieving more effective generalization in extracting degradation representation.

GFIQA-20k	w/o CA	w CA
PLCC/SRCC	0.9738/0.9733	**0.9745/0.9740**

Table 3. **Impact of Channel Attention on Model Performance.**

PLCC/SRCC	StyleGAN-IQA	MANIQA	DSL-FIQA w/o landmark	DSL-FIQA
GFIQA-20k	0.9673/0.9684	0.9614/0.9604	0.9725/0.9720	**0.9745/0.9740**
CGFIQA-40k	0.9822/0.9821	0.9805/0.9809	0.9855/0.9852	**0.9873/0.9880**

Table 4. **Impact of Landmark Guidance on Model Performance.**

Effectiveness of Channel Attention. By integrating a channel attention block, our method achieves a more precise feature focus, enhancing face quality assessment. This improvement leverages the well-documented advantages of attention mechanisms within the domain of image analysis, effectively emphasizing crucial channels. The comparative results, demonstrating the impact of incorporating channel attention into our approach, are detailed in Table 3.

Effectiveness of Landmark Guidance. We examine the impact of landmark guidance by conducting an experiment in which we omit the landmark detection component from DSL-FIQA. We then assess the performance on the GFIQA-20k [16] and CGFIQA-40k datasets, with the results detailed in Table 4. This evaluation demonstrates that incorporating landmark guidance improves the effectiveness of our method.

5. Discussion

In our Dual-Set Contrastive Learning (DSL) framework, we utilize the real-world image set ($\mathcal{R}$) to establish soft proximity mapping through the synthetic image set ($\mathcal{S}$). Theoretically, it is possible for two or more images in set $\mathcal{R}$ to have identical degradation representations.

However, it is important to note that the likelihood of this occurrence is extremely low due to the complex and variable nature of image degradation in real-world scenarios. In practice, even degradations that appear visually similar can have distinct characteristics influenced by various factors such as environmental conditions, lighting, and camera settings. Therefore, while the theoretical possibility of identical degradation representations in two images exists, it is practically negligible.

Additionally, we examine the t-SNE results presented in Figure 4 of the main paper. Initially, we observe that Gaussian noise, which is random and impacts the entire image, fundamentally contrasts with blurs and compressions that

[1] We test on device-exposure subset in PIQ23 dataset.

specifically affect image structure. This distinction likely causes Gaussian noise to appear separate from other degradations in t-SNE visualizations. Furthermore, JPEG compression and low resolution both lead to a loss of image detail, with the former eliminating high-frequency information and the latter decreasing the pixel count. This commonality in their impact on image clarity might result in similar patterns within the t-SNE visualizations.

References

[1] Gender-and-age-detection. https://github.com/smahesh29/Gender-and-Age-Detection, 2023. 1

[2] Peter J. Bevan and Amir Atapour-Abarghouei. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification, 2022. 1

[3] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 2020. 3

[4] Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Théo Cayla, Sira Ferradans, and Jean Ponce. An image quality assessment dataset for portraits. In *CVPR*, 2023. 5, 6

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[7] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM ToG*, 2020. 4

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 4

[9] Byungho Jo, Donghyeon Cho, In Kyu Park, and Sungeun Hong. Ifqa: Interpretable face quality assessment. In *WACV*, 2023. 4

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 3

[11] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 1

[12] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 1

[13] F-F ImageNet Li. Crowdsourcing, benchmarking & other cool things. *CMU VASC Semin*, 2010. 4

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[15] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016. 1

[16] Shaolin Su, Hanhe Lin, Vlad Hosu, Oliver Wiedemann, Jinqiu Sun, Yu Zhu, Hantao Liu, Yanning Zhang, and Dietmar Saupe. Going the extra mile in face image quality assessment: A novel database and model. *TMM*, 2023. 5, 6