FPEM: Face Prior Enhanced Facial Attractiveness Prediction for Live Videos with Face Retouching

Supplementary Material

1. Additional Implementation Details

In Sec. 5.2, we provided a brief overview of the implementation details, including model initialization and hyper parameters. In this section, we will elaborate further on the specifics of the experimental setup.

Data Preprocess. Each input image I is first resized to 224×224, 112×112, 160×160 and then fed into three image encoders in FPEM (E_I, E_{MS} and E_{FP}), respectively. It is important to note that, in order to maintain the original aspect ratio of each image, padding is applied to the shorter side before resizing. Meanwhile, data augmentation techniques are employed during the training stage. Specifically, the training images are horizontally or vertically flipped with a random probability of 0.5.

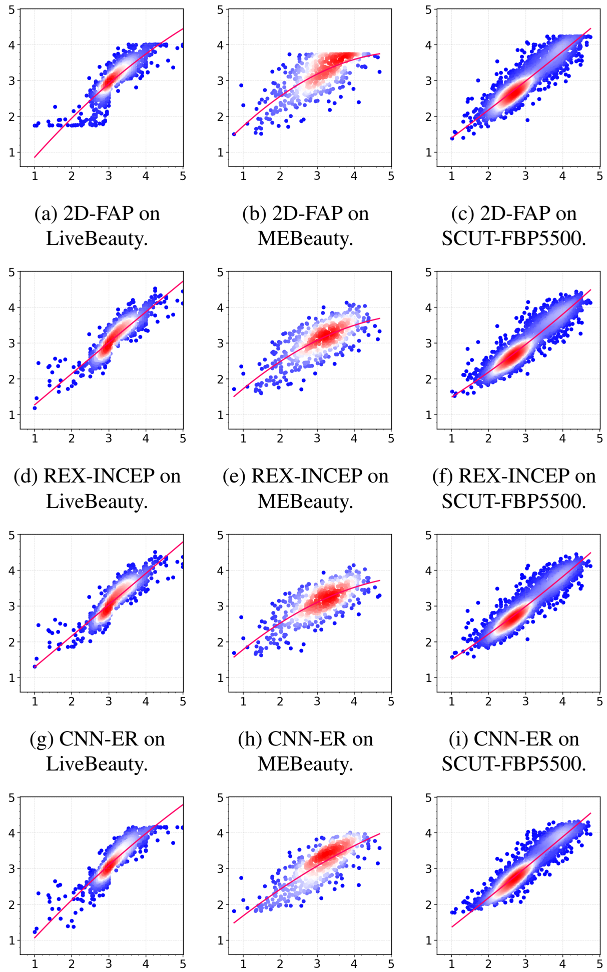
Learning rate. As mentioned in Sec. 5.2, the learning rate scheduler is set with linear warm-up and cosine annealing scheme, and the base learning rate varies with different training phase. Specifically, during the preliminary training phase, MAEM adopts 5e-6 as base learning rates. As for PAPM, since the initialization of the utilized Swin Transformer is inherited while the other layers are trained from scratch, different base learning rates are applied, with a base learning rate of 5e-4 for the Swin and 2e-5 for the other layers. During the selective fusion phase, a base learning rate of 5e-6 is applied.

Computation Complexity. The proposed model requires 140M parameters and 27.0G FLOPs (compared to 240M parameters and 35.7G FLOPs for CNN-ER). The two-stage training process takes less than 4 hours on a single V100 GPU, with peak memory usage of 5.3GB.

2. Additional Visualization Results

In Sec. 5.4, we provide the performance comparison of the proposed FPEM and other state-of-the-art FAP models on LiveBeauty, MEBeauty [24] and SCUT-FBP5500 [27]. In this section, we will illustrate additional visualization results.

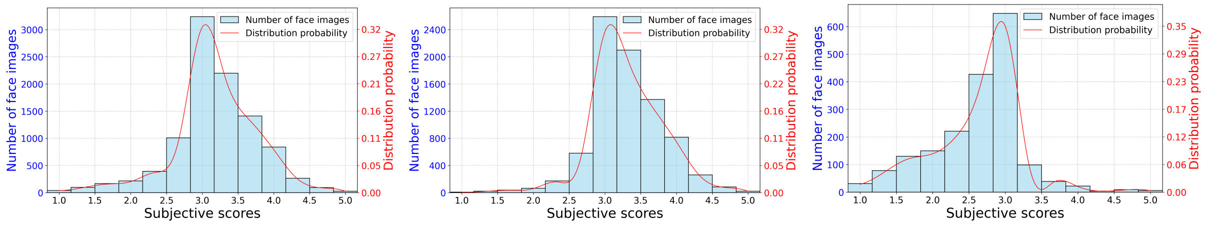
Fig. 9 presents scatter plots illustrating the relationship between the predicted attractiveness scores (vertical axis) obtained by several FAP methods (2D-FAP [32], REX-INCEP [2], CNN-ER [2] and FPEM) and the MOSs (horizontal axis) in the test sets from three benchmark datasets. Specifically, the MOSs of samples in MEBeauty are divided by two to align the MOS range with which of LiveBeauty and SCUT-FBP5500. The color of scatter points transitions from blue to red, indicating the point density from low to high. In each sub-figure, a second-order polynomial nonlinear fitting curve is superimposed. A superior model should exhibit a fitted curve closer to the diagonal line and less dispersion among the scatter points. Compared to other FAP methods, FPEM demonstrates better performance, particularly when evaluated on the MEBeauty.



(a) 2D-FAP on LiveBeauty.
(b) 2D-FAP on MEBeauty.
(c) 2D-FAP on SCUT-FBP5500.
(d) REX-INCEP on LiveBeauty.
(e) REX-INCEP on MEBeauty.
(f) REX-INCEP on SCUT-FBP5500.
(g) CNN-ER on LiveBeauty.
(h) CNN-ER on MEBeauty.
(i) CNN-ER on SCUT-FBP5500.
(j) FPEM on LiveBeauty.
(k) FPEM on MEBeauty.
(l) FPEM on SCUT-FBP5500.

Figure 9. Scatter plots of the predicted scores vs. MOSs on three benchmark datasets.

Furthermore, we present several sample face images, their corresponding MOSs and predicted attractiveness scores in three benchmark datasets in Fig. 13, Fig. 14 and Fig. 15. Each face image caption is denoted as $\langle s1, s2, gt \rangle$, where $s1$ denotes the predicted score of the

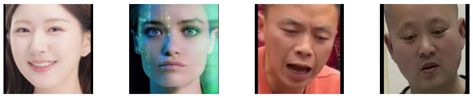


(a) MOS distribution of all samples.
Skewness=-0.470 Kurtosis=1.498

(b) MOS distribution of female samples.
Skewness=-0.083 Kurtosis=1.499

(c) MOS distribution of male samples.
Skewness=-0.281 Kurtosis=1.123

Figure 10. Illustration of the proposed LiveBeauty MOS distributions from different perspectives.



(a) <4.21,5.00> **(b) <3.09,3.74>** **(c) <2.00,2.60>** **(d) <1.37,2.20>**

Figure 11. Badcases, each image caption is denoted as $\langle pred, gt\rangle$.

model CNN-ER [2], $s2$ denotes the predicted score of the proposed FPEM, and gt denotes the ground-truth MOS. Since MEBeauty employs a face alignment method to preprocess the face images, some of the sample images displayed in Fig. 14 are slanted. It can be observed that the predicted results from FPEM closely align with the human-provided MOSs, which indicates that the proposed FPEM achieves superior human-comparable performance compared to the CNN-ER.

Failure cases are displayed in Fig. 11 while good cases are already shown in Fig. 13, Fig. 14, Fig. 15. Through analysis, we find that annotators tend to assign higher scores to celebrity faces (a). In addition, uncommon visual effects (b), exaggerated expressions (c), and extreme lighting conditions (d) also affect the model performance.

3. Additional Dataset Deatils

Rating Defination. The rating definition is as follows: 1 suggests "unappealing", describing messy makeup or multiple facial imperfections such as acne scars and wrinkles; 2 represents "less favorable", describing disharmonious makeup or some facial imperfections; 3 indicates "ordinary" with acceptable makeup or minimal facial imperfections; 4 denotes an "internet celebrity" appearance, featuring clean makeup or no visible facial imperfections; 5 describes a "celebrity" look with flawless makeup.

MOS Distribution. A more comprehensive MOS distribution of LiveBeauty is shown in Fig. 10. Skewness and kurtosis values of three MOS distributions are also presented. Fig. 10b indicates both higher skewness and higher kurtosis than Fig. 10c, demonstrating high attractiveness in females is more prevalent in the collected live videos.

Data Cleaning. During data cleaning, 3,436 annotations (1.72% of 200,000) were removed. Specifically, high-controversy images (exist annotations deviating >3 levels from the mode—the most common value among all annotations of the image—) were discarded directly while only those with moderate disagreement (exist annotations deviating =3 levels from the mode) were retained and then their outlier annotations will be removed by experts. Aggressively removing all controversial samples risks reduced diversity and overfitting. Our strategy retains challenging but realistic samples while eliminating clear outliers, aligning with best practices in subjective perception modeling.

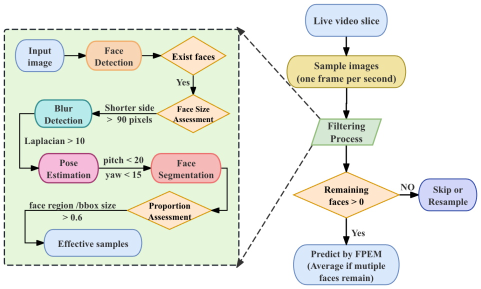
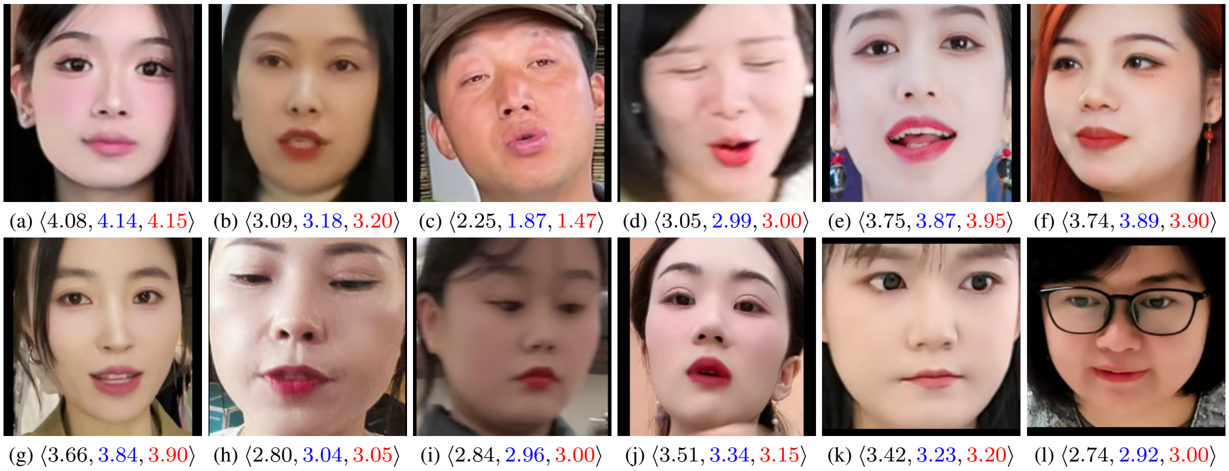


Figure 12. Application pipeline of our proposed FPEM in live streaming scenario.

4. Practical Application in Live Streaming

In this section, we further elaborate on the application of the proposed FPEM in a real-world live streaming platform. As illustrated in Fig. 12, input images are initially sampled from the live stream at a rate of one frame per second.

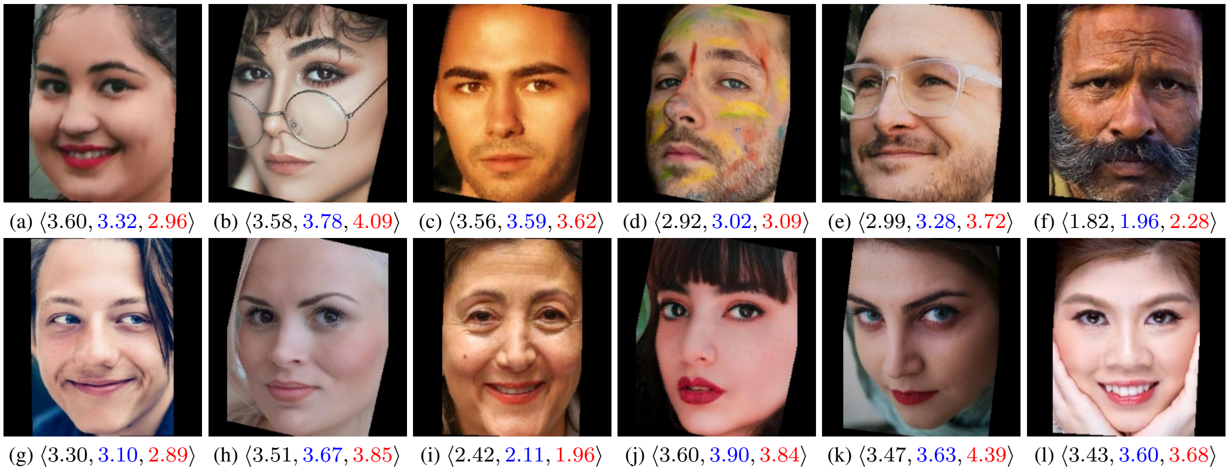
Each selected frame is first processed using a face detection method to obtain the bounding box of the detected face. If no face is detected or the detected face is too small, this frame will be skipped. In cases where multiple faces are detected within a single frame, each face is cropped based on its bounding box and subjected to the following filtering processes. Subsequently, the blur detection is applied to discard the extremely unclear faces. The pitch and yaw angles are predicted by a face pose estimation model, while faces with large angles of rotation will be excluded. With the help of a face segmentation method, faces that are significantly occluded will be filtered out. The remaining samples will be then fed into the FPEM, and the average score of these samples will serve as the predicted result for that slice of the live video.



(a) $\langle 4.08, 4.14, 4.15\rangle$ (b) $\langle 3.09, 3.18, 3.20\rangle$ (c) $\langle 2.25, 1.87, 1.47\rangle$ (d) $\langle 3.05, 2.99, 3.00\rangle$ (e) $\langle 3.75, 3.87, 3.95\rangle$ (f) $\langle 3.74, 3.89, 3.90\rangle$

(g) $\langle 3.66, 3.84, 3.90\rangle$ (h) $\langle 2.80, 3.04, 3.05\rangle$ (i) $\langle 2.84, 2.96, 3.00\rangle$ (j) $\langle 3.51, 3.34, 3.15\rangle$ (k) $\langle 3.42, 3.23, 3.20\rangle$ (l) $\langle 2.74, 2.92, 3.00\rangle$

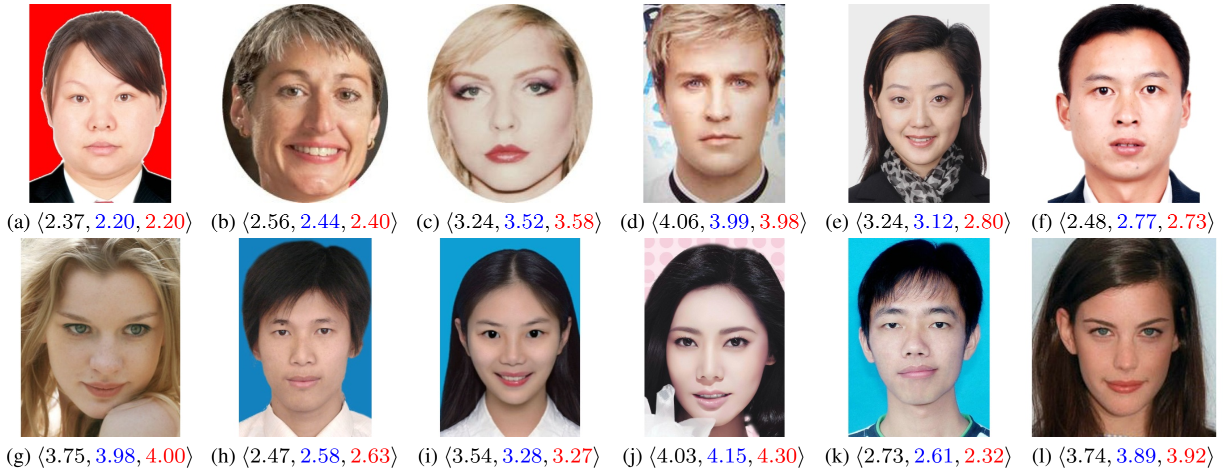
Figure 13. Prediction Results on LiveBeauty.



(a) $\langle 3.60, 3.32, 2.96\rangle$ (b) $\langle 3.58, 3.78, 4.09\rangle$ (c) $\langle 3.56, 3.59, 3.62\rangle$ (d) $\langle 2.92, 3.02, 3.09\rangle$ (e) $\langle 2.99, 3.28, 3.72\rangle$ (f) $\langle 1.82, 1.96, 2.28\rangle$

(g) $\langle 3.30, 3.10, 2.89\rangle$ (h) $\langle 3.51, 3.67, 3.85\rangle$ (i) $\langle 2.42, 2.11, 1.96\rangle$ (j) $\langle 3.60, 3.90, 3.84\rangle$ (k) $\langle 3.47, 3.63, 4.39\rangle$ (l) $\langle 3.43, 3.60, 3.68\rangle$

Figure 14. Prediction Results on MEBeauty.



(a) $\langle 2.37, 2.20, 2.20\rangle$ (b) $\langle 2.56, 2.44, 2.40\rangle$ (c) $\langle 3.24, 3.52, 3.58\rangle$ (d) $\langle 4.06, 3.99, 3.98\rangle$ (e) $\langle 3.24, 3.12, 2.80\rangle$ (f) $\langle 2.48, 2.77, 2.73\rangle$

(g) $\langle 3.75, 3.98, 4.00\rangle$ (h) $\langle 2.47, 2.58, 2.63\rangle$ (i) $\langle 3.54, 3.28, 3.27\rangle$ (j) $\langle 4.03, 4.15, 4.30\rangle$ (k) $\langle 2.73, 2.61, 2.32\rangle$ (l) $\langle 3.74, 3.89, 3.92\rangle$

Figure 15. Prediction Results on SCUT-FBP5500.