Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Supplementary materials for

Nandhini CHOCKALINGAM, Brindha MURUGAN, 2023. A multimodal dense convolution network for blind image quality assessment. *Front Inform Technol Electron Eng*, 24(11):1601-1615.
https://doi.org/10.1631/FITEE.2200534

1 Description of the dataset

The description of the three large-scale benchmark IQA datasets used in the proposed work is as follows:

LIVE: the LIVE dataset contains a total of 29 pristine images, which are distorted using five types of distortion, such as white noise (WN), Gaussian blur (GB), JPEG, and JP2K compression, and fast fading (FF) at different levels. Each distorted image is subjected to quality assessment, and its quality score, differential mean opinion score (DMOS), in the range [0, 100] is predicted in the lab environment by 161 observers, who are made available with a total of 779 distorted images in the dataset. The high DMOS indicates a poor quality image.

TID2013: TID2013 is an extended version of the TID2008 dataset with 3000 images constructed from pristine images with 24 types of distortions and 5 levels of distortion. This dataset contains more images compared to the LIVE, TID2008, CSIQ, and Wild-LIVE datasets. The MOSs are obtained in the range [0, 9] in the lab environment with the high MOS for good quality images.

KADID-10k: KADID-10k is a large-scale, artificially distorted IQA dataset obtained from 81 pristine images distorted using 25 distortions with 5 levels. It is three times larger than the TID2013 dataset. The MOSs are in the range [1, 5] and the quality labels are obtained using crowdsourcing. A high MOS corresponds to a good quality visual image.

2 Details of evaluation metrics

The proposed work uses the Spearman rank-order correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC) to evaluate the performance of the model.

SROCC: SROCC measures the degree of a monotonic relationship between the predicted quality score $\tilde{q}$ and the subjective quality score q_{sub}.

$$\text{SROCC}(\tilde{q}, q_{\text{sub}}) = 1 - \frac{6\sum(\text{rq}_i - \tilde{\text{rq}}_i)^2}{T(T^2-1)}, \tag{S1}$$

where T represents the total number of images in the evaluation dataset, and rq_i represents the rank of the ground truth quality score, and $\tilde{\text{rq}}_i$ is the rank of the predicted quality score for the i^{th} image in the dataset.

PLCC: PLCC measures the linear correlation between the predicted quality score $\tilde{q}$ and the subjective quality score q_{sub}.

$$\text{PLCC}(\tilde{q}, q_{\text{sub}}) = \frac{\text{cov}(q_{\text{sub}}, \tilde{q})}{\sigma(q_{\text{sub}})\sigma(\tilde{q})}, \tag{S2}$$

where $\sigma(q_{\text{sub}})$ and $\sigma(\tilde{q})$ represent the standard deviation between the predicted quality score $\tilde{q}$ and the subjective quality score q_{sub}, and $\text{cov}(q_{\text{sub}}, \tilde{q})$ is their covariance.

3 Local contrast normalization (LCN)

In several image processing applications, contrast normalization has been used as a preprocessing step to model nonlinear visual perception masking. The normalization not only addresses the saturation problem, but also strengthens the network's resistance to changes in illumination and contrast. Contrast is normalized over each small window rather than the entire image, and the pixel's local mean and variance can vary as a result of this local normalization. Because performance improves with a smaller normalization window, contrast normalization should be used locally for the NR-IQA tasks.

The preprocessing step used in this analysis assumes that the pixel's intensity value at location (m, n) is $I(m, n)$. Then, its normalized value $\tilde{I}(m, n)$ can be calculated as follows:

$$\tilde{I}\left(m,n\right)=\frac{I\left(m,n\right)-\mu(m,n)}{\sigma(m,n)+C}, \tag{S3}$$

$$\mu\left(m,n\right)=\sum_{k=-K}^{K}\sum_{l=-L}^{L}\omega_{k,l}I\left(m+k,n+l\right), \tag{S4}$$

$$\sigma\left(m,n\right)=\sqrt{\sum_{k=-K}^{K}\sum_{l=-L}^{L}\omega_{k,l}\left[I\left(m+k,n+l\right)-\mu\left(m,n\right)\right]^{2}}. \tag{S5}$$

The constant C is used to ensure numerical stability when the denominators are close to 0. $\mu(m, n)$ and $\sigma(m, n)$ are the adjacent local patch's local mean and standard deviation respectively. Theunit volume Gaussian window is given by $\omega = \omega_{k,l}$, in which k and l range from $k = -K, ..., K$ and $l = -L, ..., L$, respectively. The window sizes L and K are chosen as 3 for local normalization, which is smaller than the input patch size.

4 Importance of GLCM features for influential performance

Fig. S1a represents the uniform patches (Patch-UNF1 and Patch-UNF2) and non-uniform patches (Patch-NUNF1 and Patch-NUNF2) extracted from the image on the left end. GLCM features are extracted from these patches along the positional direction ϕ. The detailed analysis of texture features used in the proposed study is given below to show that the features extracted using GLCM have an influence on performance.

Contrast/Inertia measures the sum of intensity variance squares between a pixel and its neighboring pixels over the entire picture. The weight $(i - j)^2$, increases exponentially, as the pixel distance increases from the diagonal. It helps in exposing the image structure by extracting the edge information. The uniform patches have a similar pattern and a significantly lower contrast value than the non-uniform patches, as illustrated in the contrast graph of Fig. S1b.

Correlation calculates the degree to which pixels are connected to their neighbors throughout the entire picture. Uniform patch 1 exhibits highest and consistent correlation values across various interpixel distances at different orientations. However, uniform patch 2 shows minor divergence for interpixel distances 2 and 3. The non-uniform patches, on the other hand, have greater deviations as shown in the correlation graph of Fig. S1b.

Dissimilarity measures the total intensity variance between pixels, with the weight $|i - j|$ increasing linearly as the pixel deviates from the diagonal, as opposed to an exponential increase observed in the contrast. The dissimilarity graph in Fig. S1b shows a pattern similar to that of the contrast, but there is a significant change in Patch-UNF2 that helps differentiate the texture changes effectively.

Energy measures the uniformity of the image. It gives the sum of squared values for the elements in the GLCM. The uniform image patch has few gray levels, which results in $M_(i, j)$ having a relatively higher value for energy, as illustrated in Fig. S1b.

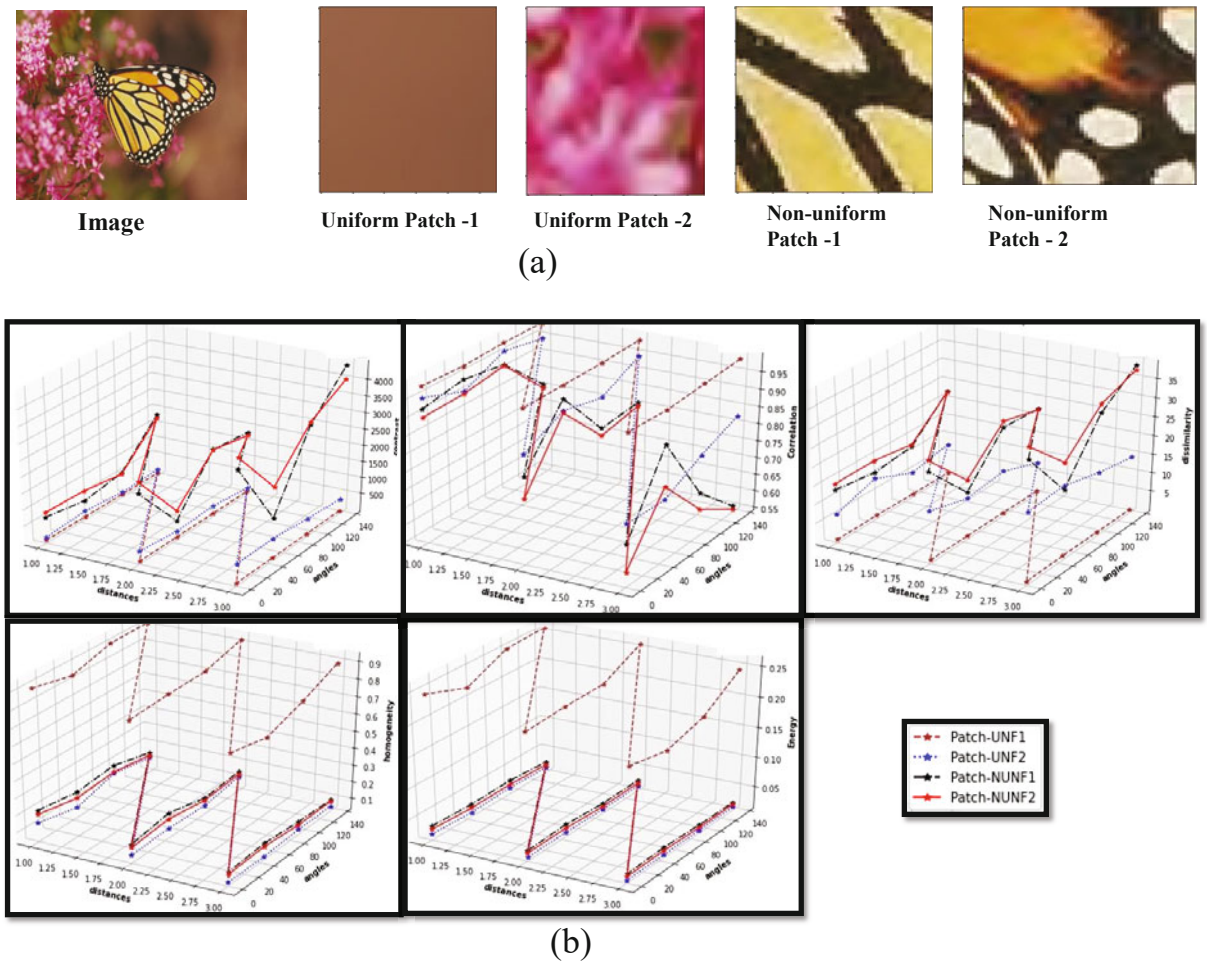

Fig. S1 Uniform and non-uniform patches with their GLCM property: (a) image with patches; (b) GLCM property

Homogeneity, due to its inverse difference moment $\frac{1}{1+|i-j|}$, shows a significant reduction in the value for the non-uniform region. Energy and homogeneity have a similar pattern in the graph, but they have different values.

The graphs in Fig. S1b show that the GLCM feature block employed in the proposed work is capable of extracting significant texture information and is best suited to describe structural information.

5 Visualization of distortion specific prediction

The predicted quality scores for DSC-Net with the LIVE dataset for different distortion types are shown in Fig. S2. The predicted DMOS value positively correlates with the subjective score for all distortion types. The scatter plot for JP2K and GB distortion contains some points that are scattered away from the line of best fit, implying weak co-relation for these specific distortions. The DSC-Net has the ability to learn Gaussian Noise and JPEG features in an effective way.

6 Effect of patch size

The input image size of the DenseNet should be 224×224 pixels. The IQA dataset used in the proposed implementation contains images of varying resolution, and the number of images in the dataset is limited. Because rescaling or resizing the image will affect the image quality, the proposed implementation augments

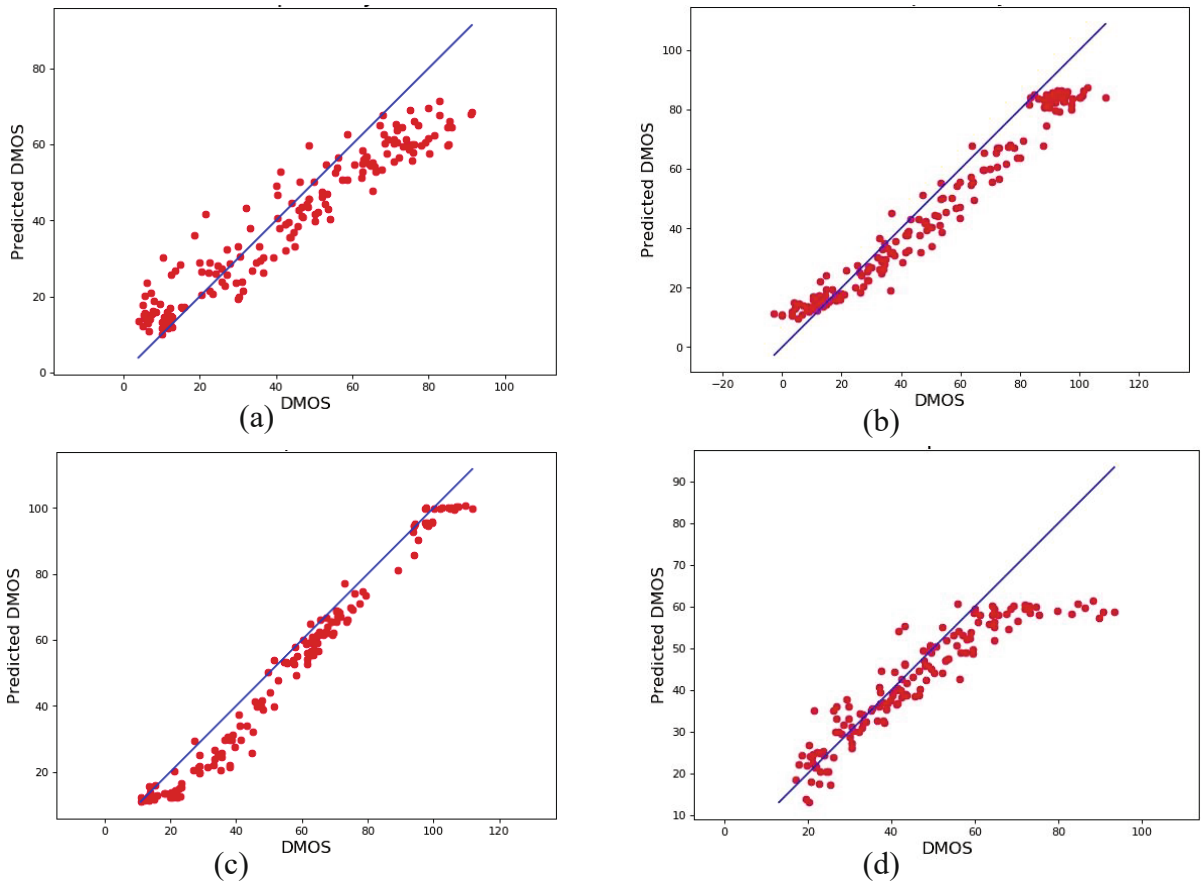

Fig. S2 Scatter plot predicted DMOS vs. DMOS using DSC-Net on the LIVE dataset: (a) JP2K; (b) JPEG; (c) white Gaussian noise; (d) Gaussian blur

the dataset by cropping the images with small patches. The patch size of 224 will reduce the augmented dataset size, and the trainable parameters of the model will increase significantly. Moreover, most of the CNN-based NR-IQA algorithms use the patch size 32×32, due to a shallow network architecture. Therefore, the proposed DSC-Net network is trained with images cropped into patches of various sizes such as 32, 56, and 64 on the LIVE dataset. However, patch size 56 provides the best result in terms of SROCC as shown in Fig. S3. So patch size 56 is used for the rest implementation.

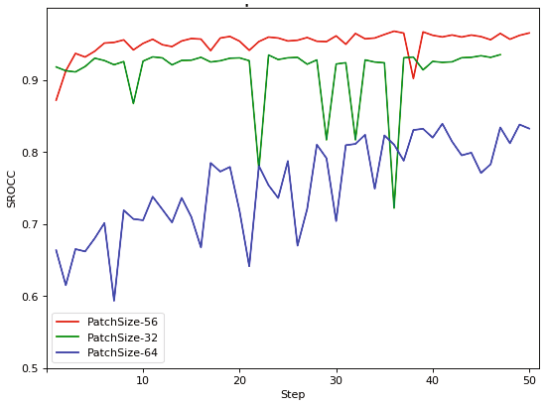

Fig. S3 SROCC on the LIVE dataset with different patch sizes

7 Performance of pre-trained DenseNet

The experiment is conducted by extracting features using the pre-trained DenseNet with 121 layers, which learns the weights from ImageNet datasets. The feature vectors from DenseNet are connected using two fully connected layers and is used to learn the quality score. The 10 random crops of size 224×224 from each image are used to train the DenseNet, and the performance on LIVE, TID2013, and KADID-10k datasets are given in Table S1. DSC-Net has a PLCC improvement of 15% over the pre-trained features on the LIVE dataset. The proposed MDSC-Net achieves about 26% and 5% SROCC improvement over the pre-trained DenseNet on the TID2013 and KADID-10k databases respectively. This proves the effectiveness of the proposed MDSC-Net architecture over the pre-trained DenseNet.

Table S1 Performance of pre-trained DenseNet on the LIVE, TID2013, and KADID-10k datasets

Dataset	PLCC	SROCC
LIVE	0.820	0.804
TID2013	0.723	0.639
KADID-10k	0.832	0.829

PLCC: Pearson linear correlation coefficient; SROCC: Spearman rankorder correlation coefficient

8 Visualization of patchwise training strategy

To prove the performance of the proposed MDSC-Net, experiments were conducted to visualize the patchwise training. This patchwise training can provide the best representation of the image as shown in Fig. S4, because all patches are included in the training and it is evident that MDSC-Net can boost the BIQA performance. During training, non-overlapping cropping of image patches is done, and each patch is assigned the quality score of the image. These patches from different images are trained in a single batch. During inference, given the image, all its patches are extracted and the prediction is done based on every single patch, and the predicted quality score of the image is the mean value of the predicted scores of those patches. To visualize the efficiency, the image from the KADID-10k database is shown in Fig. S4a. All the patches obtained from the image are shown in Fig. S4b, and because all patches are included, it represents the entire image, which boosts the network learning capability. The samples from intermediate features from DenseNet are shown in Figs. S4c and S4d, which show the feature extraction capability of the DenseNet. It is able to extract the more complex features efficiently from the entire image. Each patch detail is learned separately, which helps retrieve the local spatial details in a fine-grained manner. In addition to that, GLCM features are used to quantify the patterns and variations in intensity or color within an image. This helps identify different textures, such as a smooth or hard surface across the image patches. Hence it provides a complete representation of the image feature.

9 Convergence analysis

The loss convergence graph during training the DSC-Net model is as shown in Fig. S5. Fig. S5 shows that both training loss and validation loss have a similar plot. From the graph, a parallel plot is observable, which shows that the model has comparable performance on both training and validation datasets and that the loss is converging near 120 epochs. So, the training is stopped at 120 epochs.

To visualize the performance of the proposed approach during training, the PLCC and SROCC are recorded at the end of every epoch on the test set of the images. Fig. S6 shows that the performance of MDSC-Net is better than the performance of DSC-Net on the LIVE dataset. However, in the case of the TID2013 dataset, MDSC-Net has a higher performance overall.

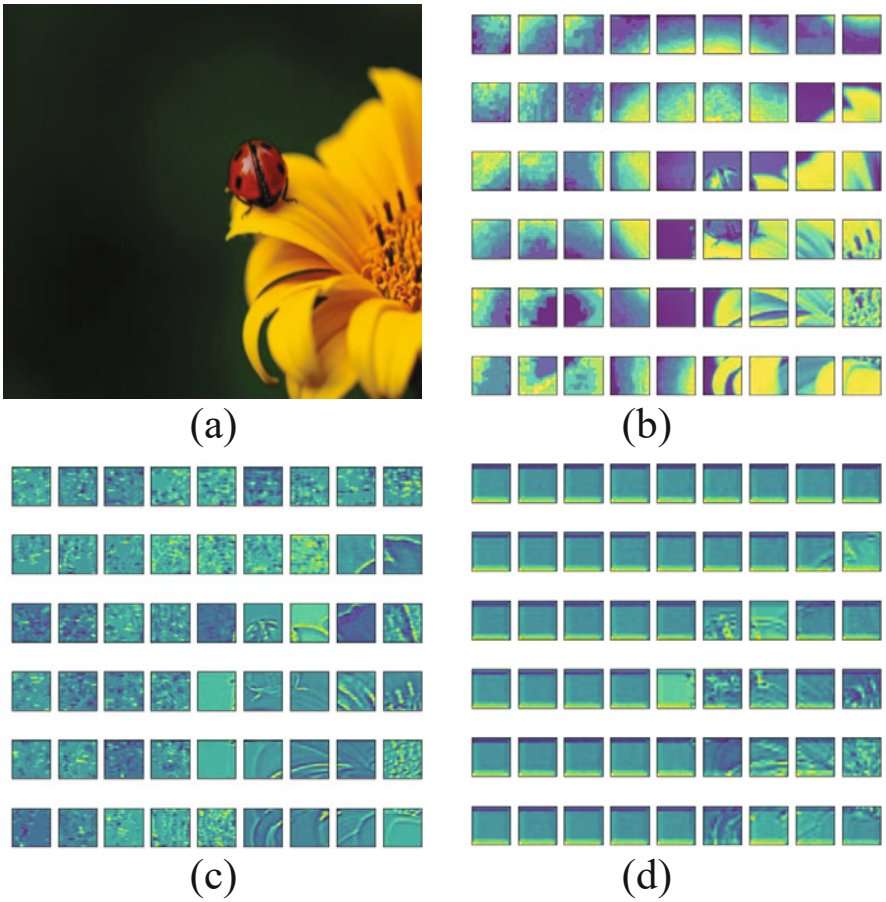

Fig. S4 Visualization of patchwise training that provides the best representation of the image: (a) image from the KADID-10k dataset; (b) patches obtained from the image; (c) intermediate features from the MDSC-Net

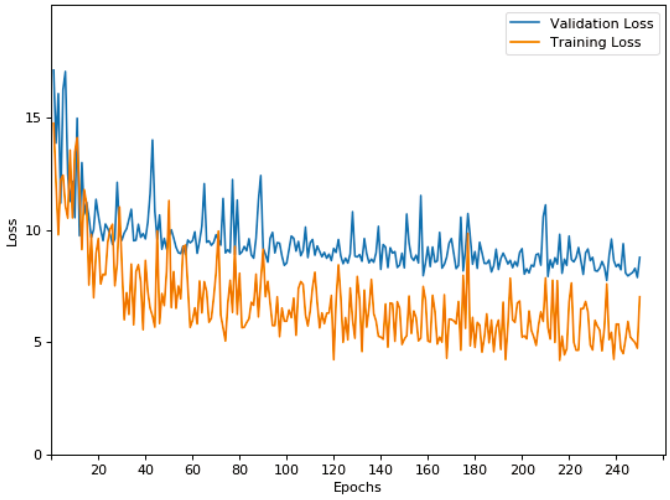

Fig. S5 Epoch vs. loss using DSC-Net

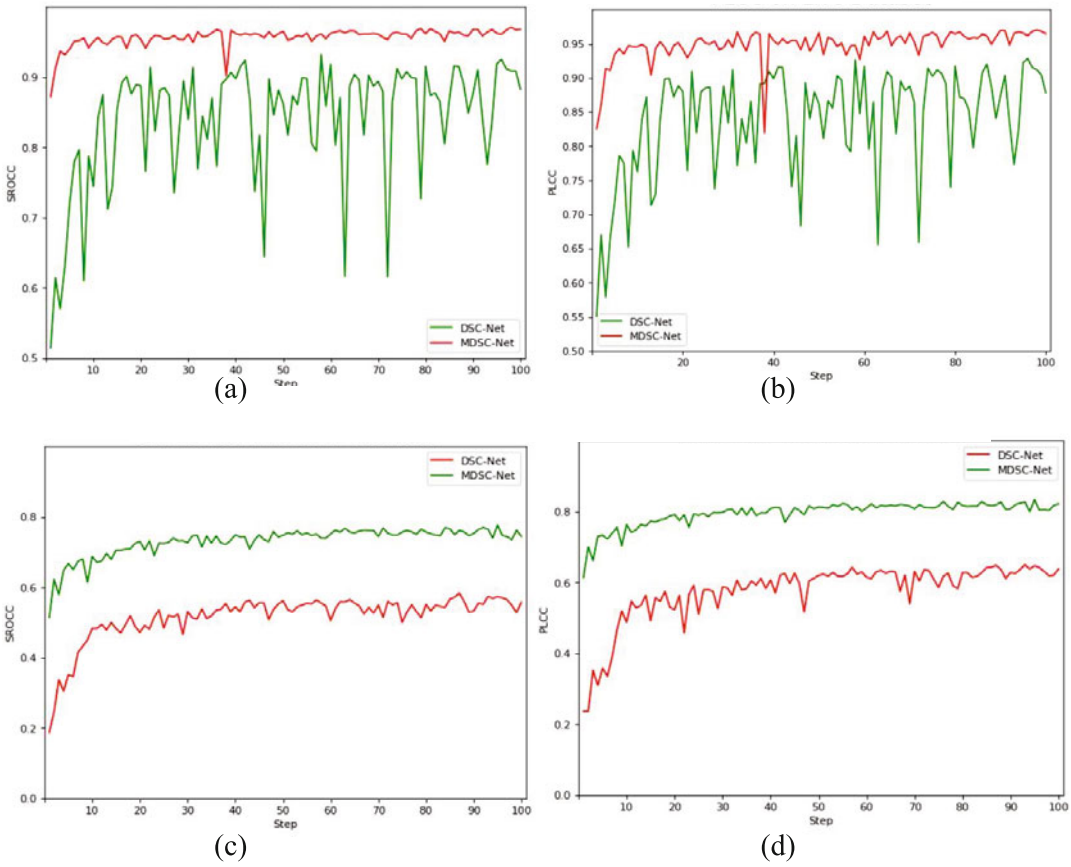

Fig. S6 Visualization of PLCC and SROCC during training: (a) SROCC on the LIVE dataset; (b) PLCC on the LIVE dataset; (c) SROCC on the TID2013 dataset; (d) PLCC on the TID2013 dataset