Boosting Image Quality Assessment through Efficient Transformer Adaptation with Local Feature Enhancement

Supplementary Material

6. Introduction

This supplementary material presents: (1) details of experimental setting and implementation; (2) additional experimental analysis and quantitative results of the ablation study of LoDa; (3) more visualization of vision transformer (ViT) [8] and LoDa.

7. Details of Experimental Setting and Implementation

7.1. Evaluation Metrics

The detailed definitions of the two performance metrics (*i.e.*, SRCC, PLCC) we use in this paper are as follows:

$$SRCC = 1 - \frac{6\sum_{t=1}^{T} d_t}{T(T^2-1)} \tag{6}$$

where T is the number of distorted images, and d_t is the rank difference between the ground-truth quality score and the predicted score of image t.

$$PLCC = \frac{\sum_{t=1}^{T}\left(s_t - \bar{s}_t\right)\left(\hat{s}_t - \bar{\hat{s}}_t\right)}{\sqrt{\sum_{t=1}^{N}\left(s_t - \bar{s}_t\right)^2}\sqrt{\sum_{t=1}^{N}\left(\hat{s}_t - \bar{\hat{s}}_t\right)^2}} \tag{7}$$

where $\bar{s}_t$ and $\bar{\hat{s}}_t$ denote the means of the ground truth and predicted score, respectively.

The detailed definition of the 4-parameters logistic function for PLCC calculation is as follows:

$$\tilde{y}' = \frac{\beta_1 - \beta_2}{1 + exp(-(\tilde{y} - \beta_3)/|\beta_4|)} + \beta_2 \tag{8}$$

where $\tilde{y}$ is the predicted score, $\tilde{y}'$ is the predicted score after correction, and $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ are the 4-parameters.

7.2. Implementation Details

We implement the model by PyTorch and conduct training and testing on an NVIDIA RTX 4090 GPU. We first resize the smaller edge of images to 384, randomly crop an input image into multiple image patches with a resolution of 224×224, then horizontally and vertically augment them randomly to increase the number of data for training [53]. Particularly, the number of patches for training is determined depending on the size of the dataset, *i.e.*, 1 for FLIVE, 3 for KonIQ-10k, and 5 for LIVEC, the number of patches for testing is 15 for all datasets, and patches inherit quality scores from the source image. We create our model based on the ViT-B pretrained on ImageNet-21k with an image size of 224×224 and patch size of 16×16. We use ResNet50 [13] pretrained on ImageNet-1k for the CNN backbone and extract feature maps of the last four blocks as multi-scale features. we use average pooling to pool multi-scale features into a spatial size of 7×7. The dimension after the down projection is 64 and the number of heads used for cross-attention is 4. Moreover, we use AdamW optimizer with a weight decay of 0.01 and a mini-batch size of 128. The learning rate was initialized with 0.0003 and decayed by the cosine annealing strategy.

All experiments are trained for 10 epochs. By default, we select the evaluation of the last epoch. For each dataset, 80% images were used for training and the remaining 20% images were utilized for testing. We repeated this process 10 times for all experiments to mitigate the performance bias and the medians of SRCC and PLCC were reported.

	KonIQ-10k		KADID-10k	
$CNNs$	SRCC	PLCC	SRCC	PLCC
DenseNet121	0.931	**0.943**	0.913	0.917
EfficientNet	0.930	0.942	0.913	0.920
ResNet34	**0.933**	**0.943**	0.915	0.921
RepVGG	0.931	**0.943**	**0.928**	0.933
ResNet50 (**Ours**)	**0.932**	**0.944**	**0.931**	**0.936**
ResNet101	0.931	**0.943**	0.927	0.934
ConvNeXt-S	0.931	**0.944**	**0.931**	**0.937**

Table 8. Impact of different pretrained CNN architectures.

	KADID-10k		KonIQ-10k	
r	SRCC	PLCC	SRCC	PLCC
48	0.929	0.934	**0.934**	**0.945**
64	**0.931**	**0.936**	0.932	0.944
80	0.923	0.928	0.933	**0.945**

Table 9. Impact of the latent dimension r. The best performances are highlighted with **boldface**.

	KADID-10k		KonIQ-10k	
h	SRCC	PLCC	SRCC	PLCC
2	0.929	0.934	0.932	0.944
4	**0.931**	**0.936**	**0.932**	**0.944**
8	0.929	0.935	0.933	0.944

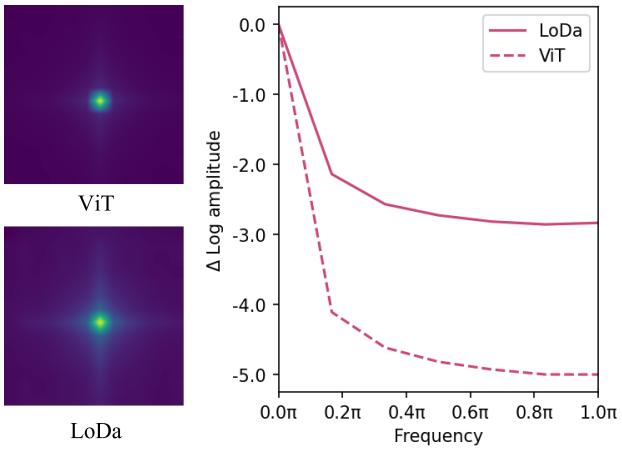
Table 10. Impact of the number of heads h in cross-attention.

Method	LIVE		TID2013		KADID-10k		LIVEC		KonIQ-10k		SPAQ		FLIVE	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
ResNet50	0.933	0.944	0.785	0.829	0.887	0.896	0.822	0.848	0.910	0.922	0.917	0.921	0.556	0.650
ViT-B	0.967	0.973	0.744	0.808	0.889	0.899	0.774	0.800	0.874	0.891	0.918	0.922	0.494	0.538
LoDa	**0.975**	**0.979**	**0.869**	**0.901**	**0.931**	**0.936**	**0.876**	**0.899**	**0.932**	**0.944**	**0.925**	**0.928**	**0.578**	**0.679**

Table 11. Performance comparison of LoDa and its CNN, Transformer backbone.

N	KADID-10k		KonIQ-10k	
	SRCC	PLCC	SRCC	PLCC
3	0.923	0.929	0.929	0.941
6	0.927	0.933	**0.932**	0.943
12	**0.931**	**0.936**	**0.932**	**0.944**

Table 12. Impact of the number of heads h in cross-attention.



(a) (b)

Figure 8. Fourier analysis of features of ViT and LoDa on KonIQ-10k. (a) Fourier spectrum of ViT and LoDa. (b) Relative log amplitudes of Fourier Transformed feature maps. (a) and (b) show that LoDa captures more high-frequency signals.

8. More Ablation Study and Discussion

8.1. Comparison of LoDa and backbones

The base results of using ResNet50 and ViT-B alone on all datasets are shown in Table 11. It can be observed that compared to the backbones, LoDa significantly enhances its performance across all these datasets.

8.2. Studies on Diverse CNNs

We conduct ablation studies on several CNNs as shown in Table 8. Most models are trained on ImageNet-1k, except for ConvNeXt-S pretrained on ImageNet-21k and finetuned on ImageNet-1k. The transformer model is fixed to ViT-B pretrained on ImageNet-21k. We present the CNNs from top to bottom in ascending order of their performance on ImageNet Validation. For *KonIQ-10k*, our method yields consistent results over various CNNs. For *KADID-10k* with more diverse local distortions, superior performance would require CNN of higher proficiency to extract more local distortion features. This aligns with the discussion in Section 4.5, indicating that the multi-scale distortion features enhance LoDa's ability to address local distortions.

8.3. Latent Dimension

Due to the potentially overwhelming number of parameters and computational overhead caused by the large dimension of ViT [8] (768 for ViT-B), inspired by the concept of adapters in the field of NLP [17], we propose to down project the ViT tokens and multi-scale distortion tokens to a smaller latent dimension r. We study the effect of the latent dimension r on KonIQ-10k [15] and KADID-10k [26] datasets. Results are shown in Table 9. From the table, we can observe that on the KonIQ-10k dataset, our model is slightly affected by the effect of latent dimension r, and on the KADID-10k dataset, our model performs the best when r is 64. Therefore, we empirically set r to 64 by default.

8.4. Number of Heads in Cross-attention

We run an ablation study on different numbers of heads in cross-attention when the latent dimension is set to 64. As shown in Table 10, when the latent dimension is fixed, the number of heads in cross-attention has little effect on our model. Thus, we set the number of heads to four so that our model performs slightly better on the KADID-10k dataset.

8.5. Number of Interactions

In the paper, we empirically fuse the ViT tokens with multi-scale features in all of the ViT encoder layers. However, it's essential to acknowledge that this choice is a result of our empirical decision-making process, in fact, we can fuse them only in part of ViT encoder layers. Thus, we run an ablation study with different numbers of interactions Non KonIQ-10k and KADID-10k datasets. In this ablation study, we distribute the ViT encoder layers into N blocks, with each block containing L/N encoder layers, where L denotes the total number of encoder layers. Then, we only fuse the ViT tokens with multi-scale distortion features in each block instead of each layer. Results are shown in 12. It can be observed that our model's performance improves

with an increased number of interactions. Notably, it is worth observing that even with just half of the interactions, our model yields excellent performance outcomes.

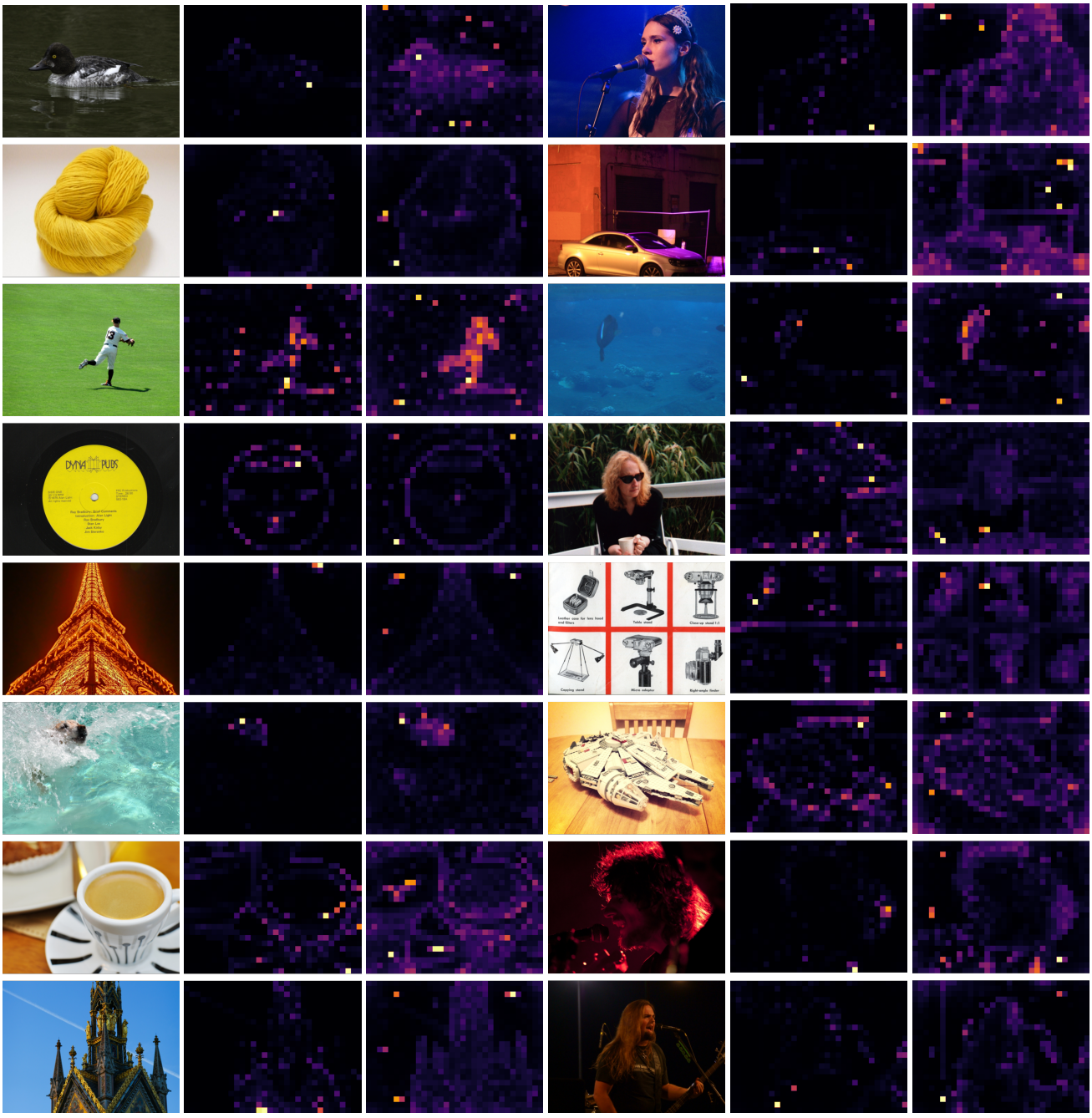
9. Visualization of Vision Transformer (ViT) and LoDa

9.1. Visualization of Fourier Analysis

In the paper, we show the Fourier analysis of features of ViT and LoDa on the KADID-10k dataset, here we additionally show the Fourier analysis of full-fintuned ViT and LoDa on the KonIQ-10k dataset (average over 128 images) in Figure 8. We can observe the same results on the KonIQ-10k dataset, it further demonstrates that LoDa captures more high-frequency signals and show the effect of our proposed method.

9.2. Visualization of attention maps

In the paper, we show the attention maps of ViT and LoDa, here we additionally show more attention maps of full-fintuned ViT and LoDa in Figure 9.



(a) images (b) ViT (c) LoDa (d) images (e) ViT (f) LoDa

Figure 9. Visualization of attention maps of features of ViT and LoDa.