

SwinIQA: Learned Swin Distance for Compressed Image Quality Assessment

Jianzhao Liu, Xin Li, Yanding Peng, Tao Yu, Zhibo Chen

CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China

{jianzhao, lixin666, pyd, yutao666}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

Image compression has raised widespread interest recently due to its significant importance for multimedia storage and transmission. Meanwhile, a reliable image quality assessment (IQA) for compressed images can not only help to verify the performance of various compression algorithms but also help to guide the compression optimization in turn. In this paper, we design a full-reference image quality assessment metric SwinIQA to measure the perceptual quality of compressed images in a learned Swin distance space. It is known that the compression artifacts are usually non-uniformly distributed with diverse distortion types and degrees. To warp the compressed images into the shared representation space while maintaining the complex distortion information, we extract the hierarchical feature representations from each stage of the Swin Transformer. Besides, we utilize cross attention operation to map the extracted feature representations into a learned Swin distance space. Experimental results show that the proposed metric achieves higher consistency with human’s perceptual judgment compared with both traditional methods and learning-based methods on CLIC datasets.

1. Introduction

Image/Video compression plays a pivotal role in modern society. Currently, there are various compression methods including traditional codecs (e.g., HEVC/H.265 [12], VVC/H.266 [1]) and learning-based methods [15, 5], which aim to solve rate-distortion optimization (RDO) problem. In such a process, IQA of compressed images plays a vital role in guiding the optimization and verification of various compression algorithms.

Commonly used traditional IQA algorithms in image compression methods, such as PSNR (peak signal-to-noise ratio), are mainly utilized to measure the pixel-wise fidelity. Though they have low computational complexity, they are not well matched to perceived visual quality. Structure similarity (SSIM) index [13] measures the patch similarity between the reference and the distorted images, based



Figure 1: Illustration of the compression artifacts. (a) Reference image. (b) Distorted image generated by HEVC codec. (c) Distorted image generated by learning-based codec.

on the assumption that the human visual system (HVS) tends to perceive the local structures. It achieves more consistent results with human perceptual quality on popular datasets. Moreover, learning-based metrics also show impressive improvement[8, 10]. LPIPS [17] obtains the perceptual similarity judgment by calculating the l_2 distance between features extracted from deep convolutional neural networks (CNNs) pre-trained on ImageNet classification task. Similarly, DISTS [3] measures the texture and structure similarities between the VGG-based deep features to calculate the perceptual similarity of two images, which achieves the state-of-the-art (SOTA) performance on benchmark datasets.

Recently, Transformer has shown promising potential in computer vision area and outperforms CNN in various mainstream tasks such as image classification and object detection. Taking advantage of the self-attention layer, Transformer can capture long-range pixel interactions and aggregate the global information from the entire input sequence. Vision Transformer (ViT) [4] splits an image into patches and treats the image patches as tokens (words) to input to a Transformer following the same way in an NLP application. However, the complexity of ViT can increase quadratically with the number of image patches. To tackle this challenge, Swin Transformer [9] is designed by integrating the advantages of both CNNs and Transformers. By limiting self-attention computation to non-overlapping local windows, it has the advantages as CNN to process im-

ages with large size due to local attention mechanism. By allowing for cross-window connection it has the advantages as Transformer to model the long-range dependencies in the data.

Inspired by the success of Transformer, several researchers attempted to apply transformers in the IQA task. TRIQ[16] utilizes a shallow Transformer encoder on the top of a feature map extracted by CNN for blind image quality assessment. IQT [2] extracts the feature representations from a CNN backbone and then feeds the extracted feature maps into the transformer encoder and decoder in order to compare the reference and the distorted images. As shown in Fig. 1, the compression artifacts are usually non-uniformly distributed with diverse types and degrees, thus it is important to combine the local-global information to measure the perceptual quality of compressed images. In this paper, we propose a full-reference image quality assessment metric named SwinIQA, based on Swin Transformer. We demonstrate that the hierarchical features extracted from each stage of the Swin Transformer have strong representation ability towards the non-uniformly distributed compression artifacts. Besides, instead of calculating the l_2 distance or feature similarity between the reference and the distortion image features like LPIPS or DISTS, we utilize cross attention to map the extracted feature representations into a learned Swin distance space. Experiment results show that our SwinIQA achieves state-of-the-art performances on CLIC2022 validation set and CLIC2021test-subtest. Moreover, we also conduct experiments of different distance mapping strategies to verify the effectiveness of the cross attention operation when comparing the reference and the distorted features.

2. Approach

In this section, we will introduce the architecture of our SwinIQA first. Then we will introduce the training strategy of our method.

2.1. Network Architecture

The framework of SwinIQA is shown in Fig. 2. It consists of three parts: a Swin hierarchical feature extractor that extracts multi-scale local-global feature representations, a cross attention block that maps the pair of reference and distortion feature representations into a learned Swin distance space, and a MLP head which maps the learned Swin distance into a quality score.

As shown in Fig. 2, Swin Transformer builds hierarchical feature maps by merging multi-level deep features. Considering that the compression artifacts are usually non-uniformly distributed with diverse distortion types and degrees, we utilize Swin Transformer as the feature extractor to extract the multi-scale hierarchical representations. Given an image $I \in \mathbb{R}^{H \times W \times 3}$, we first extract interme-

diated features from each stage of the Swin Transformer and obtain a group of features $\{f_1 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 2C}, f_2 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 4C}, f_3 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 8C}, f_4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 8C}\}$. Then we upsample all the features to $\frac{H}{8} \times \frac{W}{8} \times 2C$ and concatenate the features along the channel dimension to get the final hierarchical feature representations $f \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 22C}$:

$$f = [f_1, Up(f_2), Up(f_3), Up(f_4)], \quad (1)$$

where $[]$ denotes the concatenation operation, Up means upsampling operation, *e.g.* bilinear upsampling.

For full-reference IQA task, given a reference image I_{Ref} and a distorted image I_{Dist} , their hierarchical feature representations are denoted as f_{Ref} and f_{Dist} , respectively. In order to better measure the perceptual distance of f_{Ref} and f_{Dist} , we adopt cross attention operation to map the feature representations of the reference image and the distortion image to a learned Swin distance space. The cross attention operation is defined by:

$$\begin{aligned} z' &= LN(MHSA(q, k, v)) \\ q &= (f_{Dis} - f_{Ref})^2 W_q, k = (f_{Dis} - f_{Ref})^2 W_k, \\ v &= (f_{Dis} - f_{Ref})^2 W_v, \\ z'' &= LN(MHSA(q', k', v') + z'), \\ q' &= f_{Ref} W'_q, k' = z' W'_k, v' = z' W'_v, \\ f_{mapped} &= LN(MLP(z'') + z''), \end{aligned} \quad (2)$$

where LN represents LayerNorm, $MHSA$ represents the standard multi-head self-attention module in a transformer. MLP consists of several Fully-connected layers. q, k and v denote the query, key and value respectively. It should be noted that in the second $MHSA$ module of the cross attention block, we use the reference feature f_{Ref} as the query. Finally, a MLP regression head is employed to regress the f_{mapped} in the learned Swin distance space to a perceptual quality score:

$$d = MLP(f_{mapped}) \quad (3)$$

2.2. Training Strategy

We first pretrain the SwinIQA on the KADID-10K [7] dataset, which contains MOS value for each of the distorted images. We adopt MSE loss for training:

$$L_{reg} = \|D(I_{Ref}, I_{Dist}) - s\|_2, \quad (4)$$

where D denotes the proposed SwinIQA which compares the perceptual distance of the image pair I_{Ref} and I_{Dist} . s denotes the ground-truth normalized MOS value ($s = 1.0 - MOS/5.0$ for KADID-10K). Higher s denotes larger perceptual distance and worse perceptual quality compared with the reference image.

Then we recruit datasets which employ two alternative forced choice (2AFC) test. It means that these

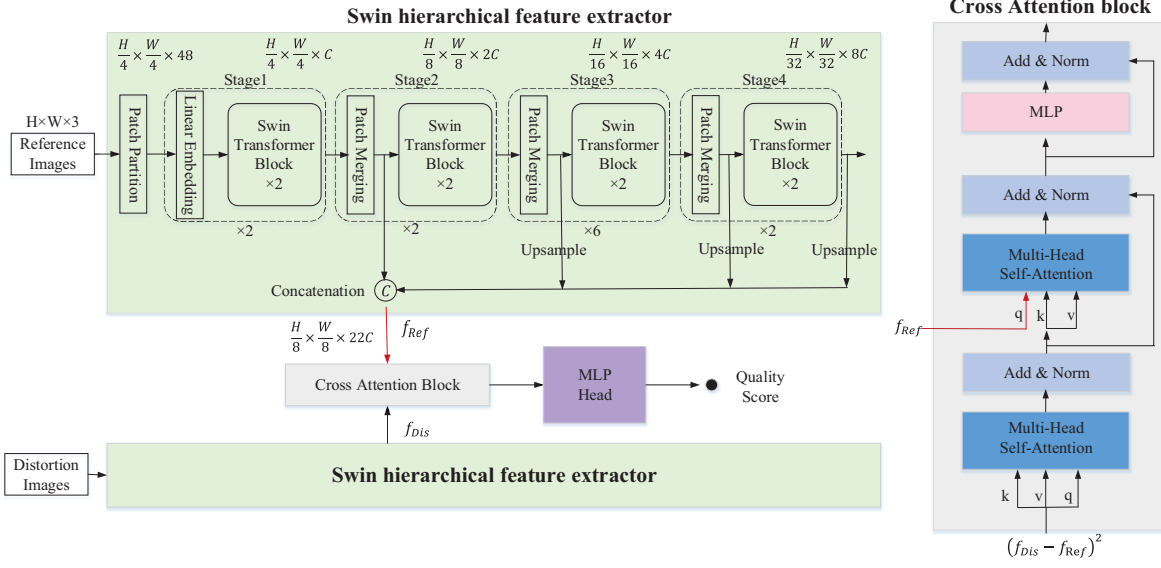


Figure 2: Framework of SwinIQA. Zooming in for better viewing.

datasets only contain labels describing which of two distorted images is more similar to a reference. Given a triplet $(I_{Ref}, I_{Dist1}, I_{Dist2})$, we should compute $d_1 = D(I_{Ref}, I_{Dist1})$ and $d_2 = D(I_{Ref}, I_{Dist2})$ to decide which image is of higher fidelity compared with the reference image. Following the work of LPIPS [17], given two distances d_1 and d_2 , we utilize a small judgment network \mathcal{G} to map the distance feature $[d_1, d_2, d_1 - d_2, d_1/d_2, d_2/d_1]$ to a predicted judgment score $\hat{h} \in (0, 1)$. The architecture uses two 32-channel $FC - ReLU$ layers followed by a 1-channel FC layer and a sigmoid function. We adopt Binary Cross Entropy (BCE) loss for training:

$$L_{bce}(I_{Ref}, I_{Dist1}, I_{Dist2}, h) = -h \log \mathcal{G}(D(I_{Ref}, I_{Dist1}), D(I_{Ref}, I_{Dist2})) - (1 - h) \log(1 - \mathcal{G}(D(I_{Ref}, I_{Dist1}), D(I_{Ref}, I_{Dist2}))), \quad (5)$$

where $h \in (0, 1)$ is the ground-truth judgment label. The total training loss is composed of two parts:

$$L_{total} = L_{bce} + \lambda_{reg} L_{reg}, \quad (6)$$

where λ_{reg} is the hyper parameter that balances the weight of the two loss items.

The final predicted results can be given by:

$$h^* = \begin{cases} 0, & D(I_{Ref}, I_{Dist1}) \leq D(I_{Ref}, I_{Dist2}) \\ 1, & D(I_{Ref}, I_{Dist1}) > D(I_{Ref}, I_{Dist2}) \end{cases} \quad (7)$$

And the judgment accuracy can be calculated by:

$$Acc = \frac{\sum_{i=1}^N (h_i^* == h_i)}{N}, \quad (8)$$

where N is the total number of the triplet $(I_{Ref}, I_{Dist1}, I_{Dist2})$.

3. Experiments

3.1. Datasets

We summarize the datasets we use for pre-training, training and testing in Table 1. During the pre-training stage, we only use KADID-10K[7] dataset for training. Specially, CLIC datasets consist of images generated by various compression methods including traditional codecs (e.g., HEVC/H.265 [12], VVC/H.266 [1]) and learning-based methods [15, 5]. In order to cover the distortion types as comprehensively as possible, we select three another datasets: PIPAL[6], BAPPS[17] and PieAPP[11], which include both traditional distortions and algorithm outputs to join in the training process. We split 109,896 triplets out of the CLIC2021Test for training (i.e., CLIC2021Test-subtrain) and the remaining 12,211 triplets for testing (i.e., CLIC2021Test-subtest). We also use CLIC2022Val dataset for testing.

3.2. Implementation Details

To balance the performance and the computational complexity, we adopted Swin-T as the backbone which consists of 4 stages (layer numbers=2, 2, 6, 2). The linear embedding dimension C of stage one was set to 96. The patch size was set to 4 and window size was set to 7. SwinIQA was first pre-trained by optimizing the objective in Eq. 4. We trained the network on KADID-10K for 50 epochs, with a batch size of 48 and a learning rate of $1e^{-4}$. The training of the SwinIQA was carried out by optimizing the objective

Table 1: Summarization of datasets we use for pre-training, training and testing.

	Dataset	Num Distort.	Distort. Types	Distort. Images/Patches	Judgment Type
Pre-training	KADID-10K[7]	25	traditional	10.1k	MOS
Training	PIPAL[6]	40	trad.+alg.outputs	29k	MOS(Elo system)
	BAPPS(2AFC-Distort)[17]	425	trad.+CNN	321.6k	2AFC
	BAPPS(2AFC-Real alg)[17]	-	alg.outputs	53.8k	2AFC
	PieAPP[11]	75	trad.+alg.outputs	20.3k	2AFC
	CLIC2021Test-subtrain	-	codec outputs	109.9k	2AFC
Testing	CLIC2021Test-subtest	-	codec outputs	12.2k	2AFC
	CLIC2022Val	-	codec outputs	5.2k	2AFC

Table 2: PLCC and SROCC performance on KADID-10K of different distance mapping strategies.

Mode	PLCC	SROCC
1	0.9521	0.9553
2	0.9213	0.9270
3	0.9451	0.9482
4	0.8698	0.8718
5	0.7713	0.7311

in Eq. 6 with the learning rate of $1e^{-4}$. The learning rate of the judgment network \mathcal{G} in Eq. 5 was also set to $1e^{-4}$. The value of λ_{reg} was set to 5.0. We randomly cropped the images to $224 \times 224 \times 3$ while training. During testing, we cropped the images into various patches and averaged the predicted distances of all patches to get more accurate results.

3.3. Discussion of different distance mapping strategies

In this section, we discuss the performance of 5 different distance mapping strategies.

- Mode 1: $f_{mapped} = (f_{Dis} - f_{Ref})^2 \otimes f_{Ref}$, where \otimes denotes the cross attention operation.
- Mode 2: $f_{mapped} = f_{Dist} \otimes f_{Ref}$.
- Mode 3: $f_{mapped} = (f_{Dist} - f_{Ref}) \otimes f_{Ref}$.
- Mode 4: $f_{mapped} = (f_{Dist} - f_{Ref})$.
- Mode 5: $f_{mapped} = f_{Dist} \odot f_{Ref}$, where \odot means the similarity distance used in DISTs[3].

The results on the KADID-10K testing set is shown in Table 2. From the table, we can see that the cross attention between $(f_{Dis} - f_{Ref})^2$ and f_{Ref} is the most effective mapping strategy when comparing the perceptual similarity of two image representations.

3.4. Comparisons with state-of-the-arts

We compare our method with two traditional methods (PSNR and MS-SSIM), two CNN-based methods (LPIPS

and DISTs), one Transformer-based method IQT[2] and last year’s champion method MMFN[10]. All the compared learning-based methods are retrained using the same datasets as SwinIQA. Given triplets $(I_{Ref}, I_{Dist1}, I_{Dist2})$, we record the predicted judgment (which distorted image is closer to the reference image I_{Ref}) given by each metric and compute the accuracy. We evaluate our performance on CLIC2022 validation set (CLIC2022Val) and the subset of CLIC2021 testing set (CLIC2021Test-subtest). The comparison results can be found in Table 3. Our method steadily outperforms other methods regarding the compressed images.

Table 3: Accuracy evaluation on CLIC2022Val and CLIC2021Test-subtest.

Methods	CLIC2022Val	CLIC2021Test-subtest
PSNR	0.572	0.510
MS-SSIM[14]	0.612	0.525
LPIPS[17]	0.761	0.749
DISTs[3]	0.762	0.752
IQT[2]	0.766	0.767
MMFN [10]	0.764	0.753
Ours	0.780	0.773

4. Conclusion

In this paper, we propose a full-reference image quality metric SwinIQA for compressed images. We employ Swin Transformer to extract the hierarchical feature representations. Then we utilize the cross attention operation to map the pair of reference and distorted image representations to the learned Swin distance space. Extensive experiments have demonstrated the effectiveness of the proposed SwinIQA for the perceptual quality assessment of compressed images.

Acknowledgement

This work was supported in part by NSFC under Grant U1908209, 62021001 and the National Key Research and Development Program of China 2018AAA0101400.

References

- [1] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 1, 3
- [2] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2021. 2, 4
- [3] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 1, 4
- [4] 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. 1
- [5] Yixin Gao, Yaojun Wu, Zongyu Guo, Zhizheng Zhang, and Zhibo Chen. Perceptual friendly variable rate image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1916–1920, 2021. 1, 3
- [6] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 3, 4
- [7] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *International Conference on Quality of Multimedia Experience*, pages 1–3. IEEE, 2019. 2, 3, 4
- [8] Jianzhao Liu, Wei Zhou, Jiahua Xu, Xin Li, Shukun An, and Zhibo Chen. Liqa: Lifelong blind image quality assessment. *arXiv preprint arXiv:2104.14115*, 2021. 1
- [9] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1
- [10] Yanding Peng, Jiahua Xu, Ziyuan Luo, Wei Zhou, and Zhibo Chen. Multi-metric fusion network for image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1857–1860, 2021. 1, 4
- [11] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 3, 4
- [12] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 1, 3
- [13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1
- [14] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 4
- [15] Yaojun Wu, Xin Li, Zhizheng Zhang, Xin Jin, and Zhibo Chen. Learned block-based hybrid image compression. *arXiv preprint arXiv:2012.09550*, 2020. 1, 3
- [16] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1389–1393. IEEE, 2021. 2
- [17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 3, 4