Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 1000–1005

Procedia Computer Science

www.elsevier.com/locate/procedia

Tenth International Conference on Information Technology and Quantitative Management (ITQM 2023)

Deep Learning Based Image Quality Assessment: A Survey

Jie Yang[a,c,d], Mengjin Lyu[c,d,f], Zhiquan Qi[b,c,d,1], Yong Shi[b,c,d,e,1]

[a] School of Mathematical Science, University of Chinese Academy of Sciences, Beijing 101408, China
[b] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 101408, China
[c] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
[d] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[e] College of Information Science and Technology, University of Nebraska at Omaha, NE 68182, USA
[f] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China

Abstract

Image quality assessment (IQA) is the problem of measuring the perceptual quality of images, which is crucial for many image-related applications. It is a difficult task due to the coupling of various degradation and the scarcity of annotations. To facilitate a better understanding of IQA, we survey the recent advances in deep learning based IQA methods, which have demonstrated remarkable performance and innovation in this field. We classify the IQA methods into two main groups: reference-based and reference-free methods. Reference-based methods compare query images with reference images, while reference-free methods do not. We further subdivide reference-based methods into full-reference and reduced-reference methods, depending on the amount of information they need from the reference images, and reference-free methods into single-input, pair-input, and multimodal-input methods, according to the form of input they use. The advantages and limitations of each category are analyzed and some representative examples of state-of-the-art methods are provided. We conclude our paper by highlighting some of the future directions and open challenges in deep learning based IQA.

Keywords: Image quality assessment; deep learning; review.

1. Introduction

Image quality assessment (IQA) is the problem of measuring the perceptual quality of images which can be affected by various factors such as distortion, noise, compression, blur, etc. It is an important task for many applications such as image processing, compression, enhancement, restoration, transmission, etc. IQA differs from aesthetic quality assessment in that it focuses on the distortion level of images, while the latter considers the artistic aspects of images.

IQA is a challenging problem due to the ambiguous evaluation criteria and perceptual features. Existing IQA methods aim to build a model that is consistent with the human visual system. This can be achieved by several

[1] Corresponding authors.
E-mail address: qizhiquan@foxmail.com and yshi@ucas.ac.cn.
*Jie Yang and Mengjin Lyu contributed equally to this work.

10.1016/j.procs.2023.08.080

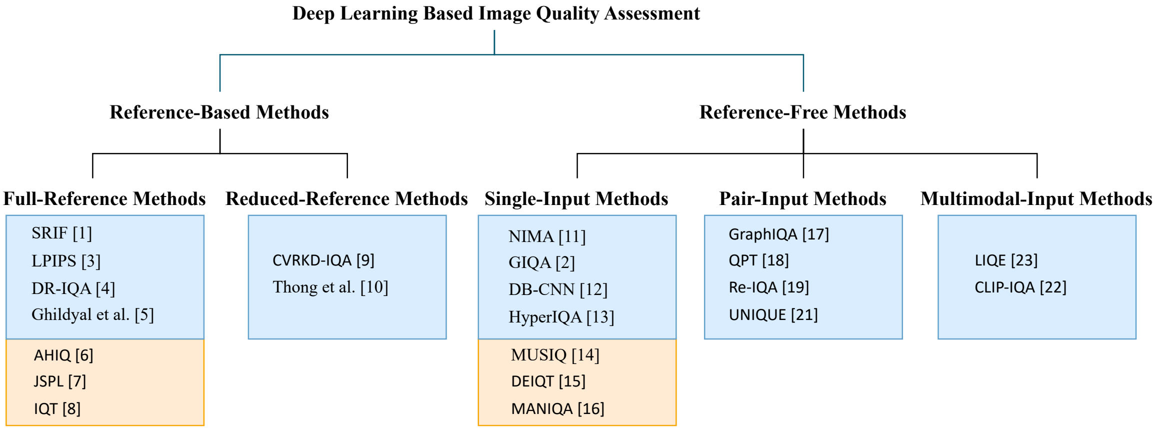

Fig. 1. The overview of the proposed taxonomy. Bluish and reddish blocks denote CNN-based and transformer-based methods, respectively. Best view in zoom.

approaches. The simplest one is comparison, which is also known as the reference-based method. This approach is based on the observation that humans can more easily rank the quality of image pairs than evaluate the quality of individual images. This kind of method [1, 2, 3, 4, 5, 6, 7, 8, 9] requires a well-aligned or partially aligned pristine-quality image with similar content to the query image as a reference. However, this approach has a major drawback, which is the difficulty of obtaining aligned image pairs with different quality levels in reality. Therefore, blind image quality assessment (BIQA), which quantifies the quality of images without referencing any pristine-quality image, has become the main research direction [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. NIMA [10] was the first method to apply deep convolutional neural networks to IQA and achieved satisfactory results. Since then, more advanced techniques, such as feature pyramid [25] and transformer [26], have been introduced to IQA. However, as the models become more complex, the trade-off between parameter size and data size becomes evident. Due to the difficulty of obtaining ground truth labels, researchers have tried to design more efficient data utilization methods. Some approaches [17, 18, 19] use existing data by cropping images into patches and forming positive and negative pairs to augment the datasets. Others [23, 22, 24] leverage auxiliary information to make better use of the data. LIQE [23] and CLIP-IQA[22] rely on a language and vision model to exploit the label semantics. KonIQ++ [24] incorporates the distortion recognition task to enhance learning efficiency. All these methods that do not require a pristine-quality image are called reference-free methods.

To gain a comprehensive understanding of IQA, we review recent deep learning-based methods and propose a taxonomy (see Figure 1). Our main contributions are as follows:

- We provide a comprehensive and systematic overview of deep learning based IQA methods, covering both reference-based and reference-free approaches.
- We introduce a novel taxonomy of deep learning based IQA, which can help to better understand the different techniques and challenges in this area.
- We identify some of the future directions and open problems in this field, which provides a reference value for future work.

The rest of our paper is organized as follows. Section 2 formally defines the problem of IQA. Section 3 categorizes different IQA methods according to our taxonomy. Finally, we conclude the paper and suggest some directions for future work in Section 4.

2. Preliminaries

The problem of IQA can be formally described by a set of parameters $\{I_q \in R^N,\ [I_r \in R^N],\ q \in R,\ q* \in R\}$ (see Figure2) where I_q denotes the query image, I_r represents the pristine-quality reference image that is only used

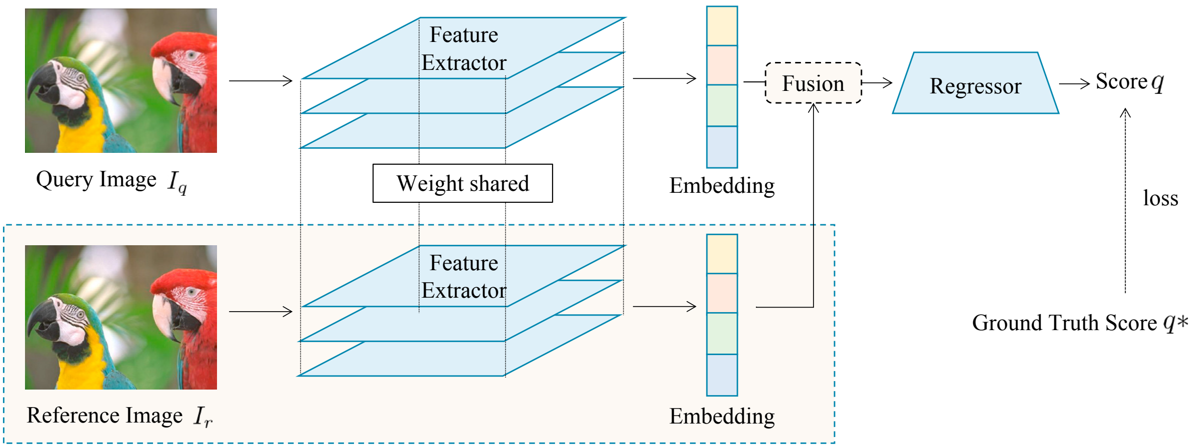

Fig. 2. The pipeline of deep learning based IQA methods. The modules in the dashed area are only used in reference-based methods.

by reference-based methods, q and $q*$ are the predicted and ground truth quality scores, respectively. In synthetic datasets, I_q is generated by applying one or more types of distortion, such as Gaussian noise, motion blur, and color jitter. In real scenes, the distortion can be more complex and intertwined. Device factors such as camera properties and lens deformation should also be taken into account. $q*$ is the mean opinion score (MOS) obtained from human ratings. IQA methods aim to design a model that can predict a score q for the query I_q that is close enough to the ground truth $q*$.

3. Methods

Image quality assessment (IQA) methods can be classified into reference-based and reference-free methods, depending on whether they use reference images or not. Reference-based methods can be further divided into full-reference and reduced-reference methods, depending on the amount of information they require from the reference images. In this section, we review some of the recent advances in reference-based IQA methods.

3.1. Reference-based methods

3.1.1. Full-reference methods

Full-reference methods employ a common feature extractor to obtain features from both the query image I_q and the reference image I_r. These features are then compared by a fusion block, which outputs a similarity map. Finally, a regression module is used to generate the final quality prediction. The general framework of full-reference methods is illustrated in Figure2.

Based on the pipeline, several extensions and improvements have been proposed. SRIF [1] uses a multi-level pyramid feature descriptor to capture information at different scales. DR-IQA [3] trains a degradation-tolerant embedding by aligning the features of pristine-quality images and their degraded counterparts. The degradation-tolerant feature is complementary to the quality-sensitive feature required by IQA and thus can facilitate the learning process. Considering that previous methods are sensitive to the alignment between query images and reference images, Ghildyal et al. [4] experiment and analyze the tolerance level to shift of different neural network components and develop a network with stronger shift robustness.

Taking advantage of the transformer [26], which can better model contextual information, IQT [7] adds it after feature fusion to further extract the distortion features. JSPL [6] uses an attention module to reweight the distance map between the query image and reference image and forces the network to focus more on informative regions. AHIQ [5] uses ViT [27] to capture spatial relationships and a shallow CNN to compensate for details.

The aforementioned methods assume that the input image pair is perfectly aligned, which is difficult to achieve in reality. They have limited applicability in practical scenarios. This limitation hinders the development of such methods.

3.1.2. Reduced-reference methods

To reduce the alignment dependency, some methods attempt to relax the constraint. CVRKD-IQA [8] adopts knowledge distillation to achieve this. It trains the teacher network as a full-reference method, but in the student network, the input reference images are allowed to have different content from the query images. With the same network structure, CVRKD-IQA aims to make the student network learn content-tolerant features. Thong et al. [9] also group image pairs with diverse content and expect the model to learn the impact of image content on quality scores.

Although these reduced-reference methods loosen the content requirements, they still need a high-quality image as the reference, which limits their applicability in real-world settings.

3.2. Reference-free methods

Reference-free methods completely eliminate the dependency of reference-based methods on high-quality reference images. They directly extract features from query images and regress them to obtain the final scores. Depending on the form of input, reference-free methods can be further classified into single-input methods, pair-input methods, and multimodal-input methods.

3.2.1. Single-input methods

Single-input methods use the original query image I_q or its patches as input. To obtain representative and discriminative descriptors of distortion without the guidance of reference, methods usually employ some techniques to do so.

NIMA [10] first explores transferring different CNN networks to IQA tasks, such as VGG [28], Inception v2 [29], and MobileNet [30], and verifies their effectiveness. GIQA [11] transforms the regression problem of quality scores into several binary classification problems under multiple thresholds to enhance robustness to label noise. In this case, each classifier only needs to answer whether the score of the image is greater than its threshold. DB-CNN [12] uses a tailored CNN designed for synthetic distortions and VGG-16 [28] to extract features, respectively. After that, it uses bilinear pooling to fuse and augment two features. HyperIQA [13] develops a hyper network that uses the semantic feature of input to generate weights for the quality prediction network adaptively.

CNN-based models have input size limitations, which require cropping or resizing the input image, which may affect the quality of the input itself. Based on this observation, MUSIQ [14] constructs a network based on ViT [27]. It designs a 2D hash position encoding and learnable scale encoding, so that the input size and receptive field of the model are not constrained, and the information between space and scale can be better captured. DEIQT [15] treats the refinement process from classification to IQA as an interpretation. It applies the transformer decoder to the classification token to efficiently acquire quality-aware features. MANIQA [16] first inputs the image patches into ViT [27] to extract features. Then it applies self-attention across channel dimensions to encode the global context and adds a scale factor after the Swin transformer [31] layer to enhance the local interaction. The final score of the query image is the weighted score sum of each patch.

3.2.2. Pair-input methods

Besides reference images, manual annotations $p*$ are also difficult to obtain. How to make better use of data has become an important factor affecting model performance. Some methods [17, 18, 19] address this issue by grouping image pairs and applying metric learning. GraphIQA [17] treats images I_q with the same distortion type as positive pairs and images with different distortion types as negative pairs. Each anchor image and its positive and negative pairs are sent to the shared-weights parallel node builder and edge builder to generate graph features and then decoded by CNN networks to obtain distortion type and level predictions. Triplet loss is used to make the distortion type of positive pairs as close as possible and negative pairs as far away as possible. QPT [18] constructs image pairs based on the fact that patches from the same image should have similar quality scores. The pair groups depend on the distortion type and content of the patches. Patches from the same image are positive pairs. Patches from different images with the same content but different degradation are negative pairs in terms of degradation. Patches from different images with different content are negative pairs in terms of content. Re-IQA [19] uses a similar method to create image pairs. It treats patches from the same images as positive pairs and patches from the query images and its augmented images as negative pairs. In addition to metric learning methods, image pairs can also be constructed for rank learning. UNIQUE [21] randomly samples pairs of images from the dataset to use their relative ranking information of MOSs $q*$. It uses fidelity and hinge losses to optimize the whole model.

3.2.3. Multimodal-input methods

Besides visual features, text features are also a rich source of information. LIQE [23] formats ground truth labels as textual templates with the form 'a photo of $a(n)$ $\{s\}$ with $\{d\}$ artifacts, which is of $\{c\}$ quality' where $a(n)$ is the object number, s is the scene category, d is the distortion type, and c is the image quality. $c \in C = \{1, 2, 3, 4, 5\} = \{\text{``}bad\text{''}, \text{``}poor\text{''}, \text{``}fair\text{''}, \text{``}good\text{''}, \text{``}perfect\text{''}\}$. It uses the CLIP [32] model to extract the text and visual features. The cosine similarity of these features is used to predict the quality score. CLIP-IQA [22] uses $c_1$ ($c_1 = \{\text{``}badphoto\text{''}\}$) and c_2 ($c_2 = \{\text{``}goodphoto\text{''}\}$) as the opposite text inputs and image I_q as visual input of CLIP [32]. The final score is calculated as the softmax of cosine similarity of visual feature and prompt feature.

4. Conclusion

With the development of deep learning and the introduction of various techniques, IQA has made significant progress. However, the complexity of real-world conditions and the scarcity of annotation data are still challenges that IQA needs to face. We can glimpse some of the future directions of IQA:

- Continuously searching for alternative ways to manual annotation. Only by breaking free from the limitations of data volume on model performance can the model have a better predictive ability and generalization ability.
- Introducing multimodal information. Labels contain rich semantic information. Making good use of this semantic information can greatly improve the model with limited data.
- Multi-task assisted learning. IQA and many other tasks are complementary. Learning them simultaneously can play a mutually reinforcing role, such as KonIQ++ [24] predicts image quality by jointly recognizing distortion type, and LIQE [23] learns scene category, distortion type, and image quality at the same time.

Acknowledgments

This work is supported by grants from Key Projects of National Natural Science Foundation of China (No.71932008, 72231010).

References

[1] W. Zhou, Z. Wang, Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 934–942.
[2] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
[3] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, J. Luo, Learning conditional knowledge distillation for degraded-reference image quality assessment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10242–10251.
[4] A. Ghildyal, F. Liu, Shift-tolerant perceptual similarity metric, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII, Springer, 2022, pp. 91–107.
[5] S. Lao, Y. Gong, S. Shi, S. Yang, T. Wu, J. Wang, W. Xia, Y. Yang, Attentions help cnns see better: Attention-based hybrid image quality assessment network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1140–1149.
[6] Y. Cao, Z. Wan, D. Ren, Z. Yan, W. Zuo, Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5851–5861.
[7] M. Cheon, S.-J. Yoon, B. Kang, J. Lee, Perceptual image quality assessment with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 433–442.
[8] G. Yin, W. Wang, Z. Yuan, C. Han, W. Ji, S. Sun, C. Wang, Content-variant reference image quality assessment via knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 3134–3142.
[9] W. Thong, J. C. Pereira, S. Parisot, A. Leonardis, S. McDonagh, Content-diverse comparisons improve iqa, arXiv preprint arXiv:2211.05215.
[10] H. Talebi, P. Milanfar, Nima: Neural image assessment, IEEE transactions on image processing 27 (8) (2018) 3998–4011.
[11] S. Gu, J. Bao, D. Chen, F. Wen, Giqa: Generated image quality assessment, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, 2020, pp. 369–385.

[12] W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind image quality assessment using a deep bilinear convolutional neural network, IEEE Transactions on Circuits and Systems for Video Technology 30 (1) (2018) 36–47.
[13] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, Y. Zhang, Blindly assess image quality in the wild guided by a self-adaptive hyper network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3667–3676.
[14] J. Ke, Q. Wang, Y. Wang, P. Milanfar, F. Yang, Musiq: Multi-scale image quality transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5148–5157.
[15] G. Qin, R. Hu, Y. Liu, X. Zheng, H. Liu, X. Li, Y. Zhang, Data-efficient image quality assessment with attention-panel decoder, arXiv preprint arXiv:2304.04952.
[16] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, Y. Yang, Maniqa: Multi-dimension attention network for no-reference image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1191–1200.
[17] S. Sun, T. Yu, J. Xu, W. Zhou, Z. Chen, Graphiqa: Learning distortion graph representations for blind image quality assessment, IEEE Transactions on Multimedia.
[18] K. Zhao, K. Yuan, M. Sun, M. Li, X. Wen, Quality-aware pre-trained models for blind image quality assessment, arXiv preprint arXiv:2303.00521.
[19] A. Saha, S. Mishra, A. C. Bovik, Re-iqa: Unsupervised learning for image quality assessment in the wild, arXiv preprint arXiv:2304.00451.
[20] H. Zhu, L. Li, J. Wu, W. Dong, G. Shi, Metaiqa: Deep meta-learning for no-reference image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14143–14152.
[21] W. Zhang, K. Ma, G. Zhai, X. Yang, Uncertainty-aware blind image quality assessment in the laboratory and wild, IEEE Transactions on Image Processing 30 (2021) 3474–3486.
[22] J. Wang, K. C. Chan, C. C. Loy, Exploring clip for assessing the look and feel of images, arXiv preprint arXiv:2207.12396.
[23] W. Zhang, G. Zhai, Y. Wei, X. Yang, K. Ma, Blind image quality assessment via vision-language correspondence: A multitask learning perspective, arXiv preprint arXiv:2303.14968.
[24] S. Su, V. Hosu, H. Lin, Y. Zhang, D. Saupe, Koniq++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects, in: The 32nd British Machine Vision Conference, 2021.
[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308.
[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, CoRR abs/1704.04861. arXiv:1704.04861.
URL `http://arxiv.org/abs/1704.04861`
[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.