Towards Light-weight Transformer-based Quality Assessment Metric for Augmented Reality

Aymen Sekhri[1,2] Seyed Ali Amirshahi[2] Mohamed-Chaker Larabi[1]

[1] CNRS, Université de Poitiers, XLIM, Poitiers, France

[2] Norwegian University of Science and Technology, Gjøvik, Norway

{aymen.sekhri, chaker.larabi}@univ-poitiers.fr, s.ali.amirshahi@ntnu.no

Abstract

This Paper introduces transformAR, a lightweight transformer-based model for objective quality assessment in AR applications. This approach utilizes pre-trained vision transformer-based encoders to capture image content information, computes distance vectors for quantifying distortions, and employs cross-attention-based decoders to model perceptual quality features. The model integrates adapted regularization techniques and label smoothing to mitigate overfitting. Experimental results demonstrate the effectiveness of transformAR, surpassing existing state-of-the-art methods.

Keywords

Augmented Reality, Vision Transformer, Image Processing, Image Quality Assessment.

1 Introduction

Augmented Reality (AR) overlays computer-generated information onto the real world via devices like smartphones and head-mounted displays, enhancing experiences in fields such as navigation, education, entertainment, and healthcare [1]. Ensuring high Quality of Experience (QoE) requires objective quality assessment methods that account for various factors impacting visual perception [2, 3]. Image quality is crucial for QoE [4], typically measured through complex and time-consuming psychophysical experiments, resulting in scarce subjective datasets for AR Image Quality Assessment (AR-IQA).

As most previous studies addressed geometric and textural degradation in 3D meshes and point clouds [5, 6]. However, Duan et al. [7] introduced CFIQA (Confusing Image Quality Assessment) and ARIQA datasets to simulate AR scenarios. They demonstrated that classical 2D metrics like PSNR, SSIM [8], and VIF [9] are ineffective for AR, necessitating advanced metrics. They explored LPIPS [10] using CNN-based feature extractors like SqueezeNet [11], AlexNet [12], and VGG [13]. Duan et al. also proposed CFIQA, a CNN-based model using VGG and ResNet [14]. This model processes features from reference and superimposed images at each convolution layer, generating distance feature maps that are refined by channel and spatial attention mechanisms to predict quality scores. The ARIQA model extends this approach with two superimposed images from the same reference but different quality scores. ARIQA+ incorporates edge detection features. However, CNN-based models face limitations due to local connectivity and translation invariance, restricting their ability to capture global patterns [15].

To address these limitations, we propose a transformer-based AR quality assessment metric, therefore, our contributions include :

- Adapting a lightweight encoder-decoder transformer framework to capture global quality features with minimal data.
- Introducing label smoothing for quality scores to reduce model overconfidence.
- We account for perceptual confusion by feeding the model with background and foreground images in addition to the fused content.

This approach mimics human observers by considering both global and localized regions for accurate quality perception [16].

2 Proposed Method

Our approach, transformAR, adapts the Vision Transformer (ViT) architecture [17, 18] for AR quality assessment. It comprises content-aware encoders, quality-aware decoders, and regressors. Below, we provide an overview of each component.

2.1 Content-aware Encoders :

We use three frozen ViT encoders with self-supervised pretrained weights via a method called, DINO [19]. These encoders capture global content information from both reference and distorted images. ViT divides an input image into non-overlapping patches, which are then processed into vectors. Using self-attention, ViT focuses on different parts of the image, enhancing feature extraction [18].

Due to data scarcity, using the full 12 transformer blocks in DINO led to overfitting. We found that using only the first two blocks was sufficient to map the input image into useful representations for quality assessment. The dataset

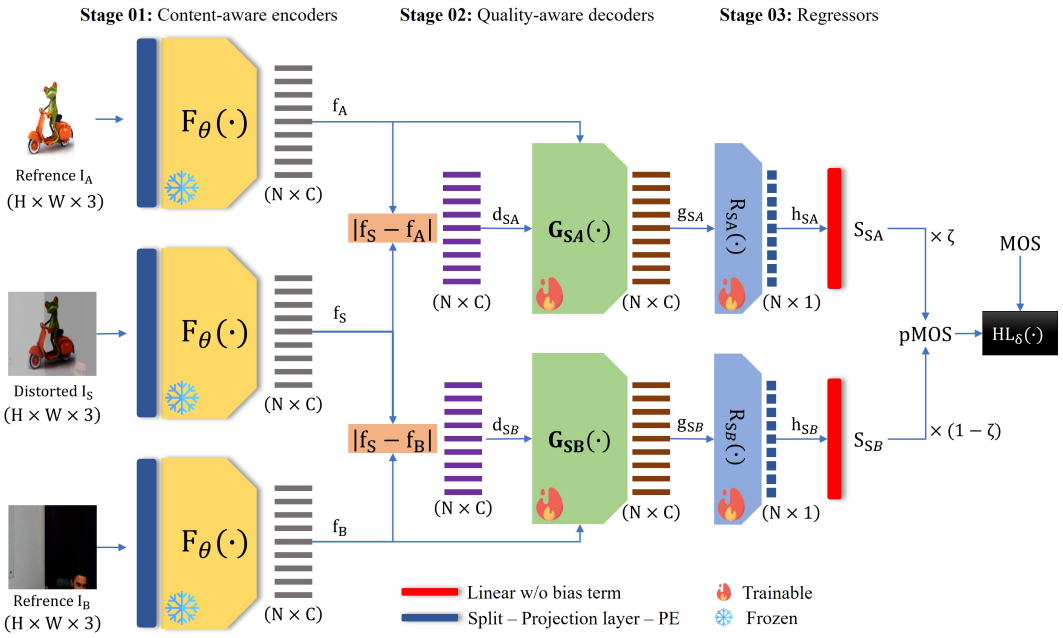

FIGURE 1 – *Illustrates our proposed architecture, showing all components.*

includes three input images : the superimposed image I_S or the distorted image, the background image I_B, and the AR image I_A or the foreground. I_S is calculated by :

$$I_S = \lambda \circ D(I_A) + (1 - \lambda) \circ I_B, \tag{1}$$

where $D(\cdot)$ denotes distortions and λ is the mixing value. Each encoder $F(\cdot)$ generates vectors $f_* = \{f_{*_i}\}_{i=0}^{N}$ for each image, with $* = \{S, A, B\}$. We use the l_1 distance to compute the sequence of distance vectors between superimposed and reference vectors :

$$\begin{cases} d_{SA_i} = |f_{S_i} - f_{A_i}| \\ d_{SB_i} = |f_{S_i} - f_{B_i}| \end{cases} \tag{2}$$

Here, d_{SA} and d_{SB} denote the distances between the representations of I_S and I_A, and I_S and I_B, respectively, which serve as inputs to the decoders described in the next section.

2.2 Quality-aware Decoders

We adapt a transformer decoder [17] without the masked self-attention mechanism. Instead, cross-attention (CA) is used, where queries come from the reference vectors and keys/values from the distance vectors. Decoders $G_{SA}(\cdot)$ and $G_{SB}(\cdot)$ use CA :

$$\begin{cases} \text{CA}(Q_{f_A}, K_{d_{SA}}, V_{d_{SA}}) = \text{softmax}\left(\frac{Q_{f_A} K_{d_{SA}}^T}{\sqrt{d_k}}\right) V_{d_{SA}} \\ \text{CA}(Q_{f_B}, K_{d_{SB}}, V_{d_{SB}}) = \text{softmax}\left(\frac{Q_{f_B} K_{d_{SB}}^T}{\sqrt{d_k}}\right) V_{d_{SB}} \end{cases} \tag{3}$$

where Q_{f_A} and Q_{f_B} are queries from f_A and f_B. Keys/values are from d_{SA} and d_{SB}, and d_k is the key vector dimension. The output embeddings are normalized, followed by a skip connection and projected via a multi-layer perceptron, producing vectors $g_{SA} = G_{SA}(f_A, d_{SA})$ and $g_{SB} = G_{SB}(f_B, d_{SB})$. This captures long-range quality information based on the distortions and visual confusion information that come from the distance vectors.

2.3 Regressors

Two regression modules $R_{SA}(\cdot)$ and $R_{SB}(\cdot)$ produce quality scores for each patch x_i from I_S : $h_{SA} = R_{SA}(g_{SA})$ and $h_{SB} = R_{SB}(g_{SB})$. Scores are aggregated using a linear layer with parameters W_{SA} and W_{SB}, resulting in final scores S_{SA} and S_{SB}. The predicted MOS (*pMOS*) is then computed as :

$$pMOS = \zeta S_{SA} + (1 - \zeta) S_{SB}, \tag{4}$$

where ζ is set to 0.51 based on iterative experimentation.

2.4 Training Procedure

Our training procedure includes key aspects such as loss function choice and addressing overfitting with regularization techniques. We selected the Huber loss [20] for its balance between robustness to outliers and sensitivity to small errors :

$$HL_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta, \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \tag{5}$$

To mitigate overfitting, we use elastic net regularization [21], which combines Lasso and Ridge methods. The overall loss function is :

$$L(y, \hat{y}, \beta) = HL_\delta(y, \hat{y}) + \lambda(\alpha \|\beta\|_1 + (1 - \alpha)\|\beta\|_2^2), \tag{6}$$

where β represents the learnable weights, λ controls the penalty terms, and α balances l_1 and l_2 penalties.

We also apply label smoothing, a technique traditionally used in classification to reduce model overconfidence,

(a) *AR image* I_A (b) *Distorted image* I_S (c) *Background image* I_B

FIGURE 2 – *Illustration of reference and superimposed (distorted) images from the used dataset [7].*

adapted here to regression. Given data scarcity and the overconfidence in predicting MOS, we introduce small random noises to the MOS :

$$y_\epsilon = y + \lambda_n \epsilon, \tag{7}$$

where λ_n is a uniform random value between [-1, 1], and ϵ is a normal distribution random number. This approach reduces overfitting and slightly improves performance.
For the implementation, we trained the model using the AdamW optimizer [22] with a learning rate of $1e^{-4}$, a batch size of 32 images, and for 150 epochs. A learning rate scheduler reduced the rate if no improvement was seen in 10 epochs. The implementation in PyTorch was run on an NVIDIA Tesla V100S-PCIE-32GB GPU.

3 EXPERIMENTAL RESULTS

We evaluated our approach on the ARIQA dataset (Figure 2), containing 560 superimposed images with associated MOS. Following [7], we divided the dataset into 50 folds, splitting each fold into 280 training and 280 testing samples without scene repetition, as detailed in Equations 8 and 9.

$$\mathcal{D} = \{[I_{A_i}, I_{B_i}, I_{S_i}, MOS_i]\}_{i=1}^{560} \tag{8}$$

$$\mathcal{X} = \{(\mathcal{T}_i, \mathcal{S}_i) \mid \mathcal{T}_i \cap \mathcal{S}_i = \emptyset\}_{j=1}^{50} \text{where} \bigcup_{j=1}^{50} \mathcal{S}_j = \mathcal{D} \tag{9}$$

3.1 Dataset

The ARIQA dataset includes 20 background images (10 indoor, 10 outdoor) and 20 AR images (web pages, natural images, and graphical representations). Each AR image has six degraded levels using JPEG compression, scaling, and contrast adjustment. Visual confusion is considered a distortion [7] with mixing thresholds $\lambda \in [0.26, 0.42, 0.58, 0.74]$, resulting in 560 stimuli. 23 participants evaluated the images using HTC VIVE Pro Eye VR headsets.

3.2 Comparison to State-of-the-Art

Table 1 compares our method to state-of-the-art approaches. Averaging metrics across 50 folds, our method outperforms others, including ARIQA+, with fewer parameters (15.32M).

TABLEAU 1 – *Comparison with state-of-the-art performance.*

Model	SRCC↑	KRCC↑	PLCC↑	# params
LPIPS [10]	0.7624	0.5756	0.7591	20.02 M
CFIQA [7]	0.7787	0.5863	0.7695	20.12 M
ARIQA [7]	0.7902	0.5967	0.7824	20.12 M
ARIQA+ [7]	0.8124	0.6184	0.8136	35.02 M
TransformAR (ours)	**0.8267**	**0.6359**	**0.8251**	**15.32 M**

3.3 Ablation Study

An ablation study using five folds from $\mathcal{X}$ evaluated the impact of removing components like the decoder, l_1-distance, label smoothing, elastic net, and Huber loss. Each component significantly impacted the model's performance (Table 2).

TABLEAU 2 – *Ablation study results on 5 folds.*

Model \ Criteria	SRCC↑	KRCC↑	PLCC↑
w/o decoder	0.6161	0.4408	0.6242
w/o l_1-distance	0.4365	0.4443	0.2960
w/o label smoothing	0.8269	0.6361	0.8221
w/o elastic net	0.8427	0.6567	0.8471
w/o Huber loss	0.8374	0.6525	0.8443
all combined	**0.8461**	**0.6582**	**0.8481**

4 CONCLUSION

This paper introduces an efficient and lightweight objective quality assessment metric for AR scenarios based on the transformer architecture. To address data scarcity, we simplified the model using two encoder blocks and one decoder block. Elastic net regularization and label smoothing were employed to enhance model robustness. Our proposed method surpassed widely used metrics like LPIPS and existing ARIQA metrics, achieving superior performance with significantly fewer parameters. The emerging field of AR quality assessment presents opportunities for advancement. Future research will focus on transformer-based approaches in realistic AR scenarios and developing specialized AR-IQA datasets to enhance objective quality metrics.

Références

[1] R. Vertucci, S. D'Onofrio, S. Ricciardi, et M. De Nino. History of augmented reality. Dans *Springer Handbook of Augmented Reality*, pages 35–50. Springer, 2023.

[2] International Telecommunication Union. Itu-t recommendation g.1036. `https://www.itu.int/rec/T-REC-G.1036-202207-I`, 2022. Accessed on October 27, 2023.

[3] A. Perkis, C. Timmerer, S. Baraković, et al. Qualinet white paper on definitions of immersive media experience (imex). *arXiv preprint arXiv :2007.07032*, 2020.

[4] J. Xu, C. Lin, W. Zhou, et Z. Chen. Subjective quality assessment of stereoscopic omnidirectional image. Dans *Advances in Multimedia Information Processing–PCM 2018 : 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*, pages 589–599. Springer, 2018.

[5] E. Alexiou, E. Upenik, et T. Ebrahimi. Towards subjective quality assessment of point cloud imaging in augmented reality. Dans *2017 IEEE 19th Int. Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017.

[6] J. Gutiérrez, T. Vigier, et P. Le Callet. Quality evaluation of 3d objects in mixed reality for different lighting conditions. *Electronic Imaging*, 32 :1–7, 2020.

[7] H. Duan, X. Min, Y. Zhu, et al. Confusing image quality assessment : Toward better augmented reality experience. *IEEE Transactions on Image Processing*, 31 :7206–7221, 2022.

[8] Z. Wang, A. Bovik, H. Sheikh, et E. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4) :600–612, 2004.

[9] H. Sheikh et A. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing (TIP)*, 15(2) :430–444, 2006.

[10] R. Zhang, P. Isola, A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[11] F. N Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, et K. Keutzer. Squeezenet : Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] A. Krizhevsky, I. Sutskever, et G. Hinton. Imagenet classification with deep convolutional neural networks. Dans *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[13] K. Simonyan et A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[14] K. He, X. Zhang, S. Ren, et J. Sun. Deep residual learning for image recognition. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, et A. Dosovitskiy. Do vision transformers see like convolutional neural networks ? *Advances in Neural Information Processing Systems*, 34 :12116–12128, 2021.

[16] M. Cheon, S. Yoon, B. Kang, et J. Lee. Perceptual image quality assessment with transformers. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, et I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] A. Dosovitskiy, L. Beyer, A.r Kolesnikov, D. Weissenborn, et al. An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, 2020.

[19] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, et A. Joulin. Emerging properties in self-supervised vision transformers. Dans *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[20] P. Huber. Robust estimation of a location parameter. Dans *Breakthroughs in statistics : Methodology and distribution*, pages 492–518. Springer, 1992.

[21] H. Zou et T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 67(2) :301–320, 2005.

[22] I. Loshchilov et F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv :1711.05101*, 2017.