https://doi.org/10.2352/EI.2023.35.8.IQSP-308

Investigating pretrained self-supervised vision transformers for reference-based quality assessment

Kanjar De, Embedded Intelligent Systems Laboratory, Luleå University of Technology, 97187 Luleå, Sweden

Abstract

Reference-based image quality assessment techniques use information from an undistorted reference image of the same scene to estimate the quality of a distorted target image. The main challenge in designing algorithms for quality assessment is to incorporate the behavior of the human visual system into the algorithms. The advent of deep learning (DL) techniques has garnered sufficient interest among researchers in the field of image quality assessment. The common limitation of applying deep learning for image quality assessment is its dependence on a large amount of subjective training data. Recent advances in the field of patch-based self-supervised vision transformers have achieved remarkable results for tasks like object segmentation, copy detection, etc. and other downstream computer vision tasks. In this paper, we study how the distance between the pre-trained self-supervised vision transformer features applied on pristine and distorted images is related to the human visual system. Experiments carried out in two publicly available image quality databases (namely TID2013, and MDID2016) have yielded promising results that can be further exploited to design perceptual reference-based image quality assessment methods.

Introduction

Objective image quality assessment (IQA) is one of the most challenging problems for the research community due to the subjective nature of the problem. Human visual system (HVS) is trained to identify the difference between distorted and pristine images, but designing algorithms to mimic this task is extremely difficult. Subjective image quality assessment includes psychophysical experiments in controlled laboratory environments or crowd-sourced experiments. Subjective ratings can be available in the form of mean opinion scores (MOS), differential mean opinion scores (DMOS), etc. to name a few. Based on the need for reference images, IQA algorithms are classified into 3 categories. Full-reference [1, 2], reduced-reference [3, 4, 5] and no-reference IQA [6]. The peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) [7] are the most widely used measures for full reference image quality assessment. Reference-based IQA can be used in a variety of applications, such as the evaluation of enhancement and compression algorithms.

With the advancement in the field of deep learning (DL), there has been interest in the application of DL techniques in the area of image quality assessment [8, 9, 10, 11, 12, 13, 14, 15]. However, one of the biggest challenges of DL is the requirement of large datasets with images of varying image quality paired with subjective studies for validation. Dodge and Karam [16] conducted experiments involving humans and deep learning models for image classification and reported that humans are much better at classifying distorted compared to convolutional neural networks. Datasets like Imagenet-C [17] have been proposed to understand how DL algorithms work on poor quality images with a motivation to design self-correcting models with respect to image distortions [18].Transformers have become the most preferred model for most modern natural language processing tasks(NLP) [19, 20].Recently, Dosovitiskiy et al. have used the concept of transformer for NLP into vision transfomers (ViT) [21] by considering an image as a sequence of words of 16×16 patches. Vision transformers are becoming a powerful and popular tool for tasks like image classification, object detection, and semantic segmentation, and have recently emerged as a strong competitor to convolutional neural networks. Vision transformers have achieved one of the highest accuracies in the Imagenet data set [22]. Vision transformer models are designed in such a way that a 224×224 image is presented to the model (base) as a sequence of 16×16 words. The ViT architecture consists of three important components, namely patch embedding, feature extractor based on stacked transformer encoders, and the classification head. After the success of ViT, subsequent architectures such as SWIN Transformer [23], DEIT [24], XCIT [25] have been proposed, making it an active area among computer vision researchers. Transformers have shown promising results in the area of IQA [26, 27]. With more and more computer vision systems being deployed for critical applications in sectors like heavy industry, healthcare, defense, etc. to name a few, robustness of deep learning algorithms is an active area of research. Some work has been done in the area of how modern architectures, such as vision transformers, behave in distorted images [28, 29, 30]. Self-supervised learning is a sub-branch of machine learning algorithms where the technique learns from unlabeled sample data. Self-supervised learning has been very popular in the area of computer vision, and methods such as SimCLR [31] and momentum contrast [32] have achieved great success. Recently, Caron et al. [33] have shown that the features of the self-supervised vision transformer have achieved remarkable results in the field of downstream computer vision transformers, and excellent results have also been obtained from the k-nn classifiers. The method proposed by the authors was termed self-distillation without labels (DINO), and we used DINO models to investigate reference-based image quality assessment in this paper. Self-supervised learning using pre-trained transformer features [34] has shown promising results, and this has served as an important motivation to investigate self-supervised transformer features between pristine and distorted images. One of the main contributions of the paper is to investigate how the distance between features extracted from the self-supervised vision transformer model trained on Imagenet aligns with subjective image quality assessment. The rest of the paper is organized as follows. In Section we discuss the preliminaries followed by Section where we list the details of the experimental protocol for the study, and finally Section where we discuss the results.

Preliminaries

One of the biggest challenges in designing image quality assessment techniques is that different types of distortion have different statistical properties. For example, the statistical properties of blurred images are different from those of images that are corrupted with additive noise, making it difficult to develop a single algorithm that can perform image quality assessment for all different types of distortion. With the advent of modern machine learning techniques, different approaches have been proposed to design IQA methods, but the biggest challenge of using fully supervised machine learning approaches is that these methods need a lot of good quality annotated training data to achieve human-level performance. Conducting subjective experiments in a laboratory-controlled environment is a slow and expensive process to generate sufficient data to train machine learning algorithms. Recent advances in self-supervised machine learning have enabled one to generate embeddings that are robust and have performed exceptionally well in downstream computer vision tasks. In this paper, we try to answer the following question: Given a pristine reference image and a distorted target image of the same scene, can self-supervised vision transformer features be able to predict the quality of the distorted image without being explicitly trained on any image quality assessment dataset? Caron et al. [33] have proposed DINO models that were based on self-supervised learning with knowledge distillation. Knowledge distillation was incorporated using a student-teacher network approach. The authors have followed the same image augmentation strategy as the boot strap your own latent (BYOL) [35] method plus added augmentations such as colour jittering, Gaussian blur and solarization and also multi-crop. These data augmentations play an important role as the model is exposed to images of the same class with different distortions, and the model is able to learn the same object with different image quality. We plot the different attention heads of a self-supervised DINO-ViTS/8 model in Fig. 1 with different image quality. We observe that for the image of the same scene, the attention weights are different as the quality of the image changes. The human visual system is able to identify the difference between two images on the basis of image quality. This observation motivates us to examine the self-supervised ViT models in detail, as during training images of the same class with different quality were introduced as part of the augmentation. Although the model was not explicitly trained to perform image quality assessment, it learned different embeddings for different qualities. In a subsequent section, we will discuss in detail how we use the features from the last hidden layers and conduct experiments to validate it against the human visual system to perform the task of reference-based image quality assessment.

Experiments

There are several public datasets available for research purposes, such as LIVE and its variants [37, 38], KADID-10K [39], CSIQ [40], VCLFER [41] to name a few. We have chosen two of the largest public data sets available for our investigation, which covers a wider group of distortions. Subjective ratings are available in the form of mean opinion scores (MOS) or differential mean opinion scores (DMOS). The details of the datasets used for this study are summarized in Table 1. As seen in Section , the attention maps of an image of the same scene with different image qualities are different. We tried to exploit this property to find the vector distance between the pristine and distorted images to check whether the results align with the human visual system. For this work, we used a simple pipeline to conduct experiments. We extracted features from a self-supervised vision transformer model for both the pristine reference and distorted target images, and then computed the distance between the two feature vectors. For this work, we used pre-trained self-supervised DINO models for feature extraction. We conducted separate experiments on the ViT-S and ViT-B architectures, where S and B represent small and base models, respectively. Furthermore, we examine the two separate configurations with patch size 8×8 and 16×16, respectively, for the ViT-S and ViT-B models. Due to the different patch sizes, the final vector dimensions are different for models with 8×8 patches and 16×16 patches. The vector obtained from the last hidden state of the DINO models serves as a feature vector for both pristine and distorted images. The resulting 2D feature vector is converted further to 1D for distance calculation. The overall pipeline of our analysis is shown in Figure 2. The dimensions of the feature vectors obtained by the different VIT architectures are reported in Table 2 .The distance measure between the vectors used for this study is listed in Section 3.

For all experiments, we have used the pretrained self-supervised vision transformer pretrained DINO models provided by the authors repository [1] (refer github for details of the hyperparameters) and the Hugging Face [42] interface for extracting the features using the feature extractor module. All experiments were carried out with Pytorch 1.10.2. For the distance calculation, we use scipy [43]. All experiments were performed with a NVIDIA RTX3070 GPU.

Results and Discussion

We observe that the City block, Canberra and Euclidean distances between the reference and distorted images in the feature space have a strong correlation as visible from the scatter plots in Figures. 3, 4 for MDID2016 and TID2013 datasets respectively. As expected, the more distorted images have a higher distance from the reference image in the feature space. An important observation is that the distance measures included in this study have a strong correlation with the human mean opinion scores obtained from subjective experiments, as seen in all graphs for all DINO models for the three datasets used for our study. The self-supervised vision transformer-based features used for the experiments are not trained on any image quality dataset. The relationship is consistent across different ViT architectures which are based on 8×8 and 16×16 patches.

Evaluation Metrics

For the analysis in this paper, we use the following quantitative measures. To study monotonicity, we used Spearman's rank order correlation coefficient (SRCC) and Kendall's rank order correlation coefficient (KRCC) between the distance measures and the subjective ratings. We also calculate the Pearson's linear correlation coefficient (PLCC) between the distance measures and the subjective ratings. From Table 4, we observe that dis-

[1] `https://github.com/facebookresearch/dino`

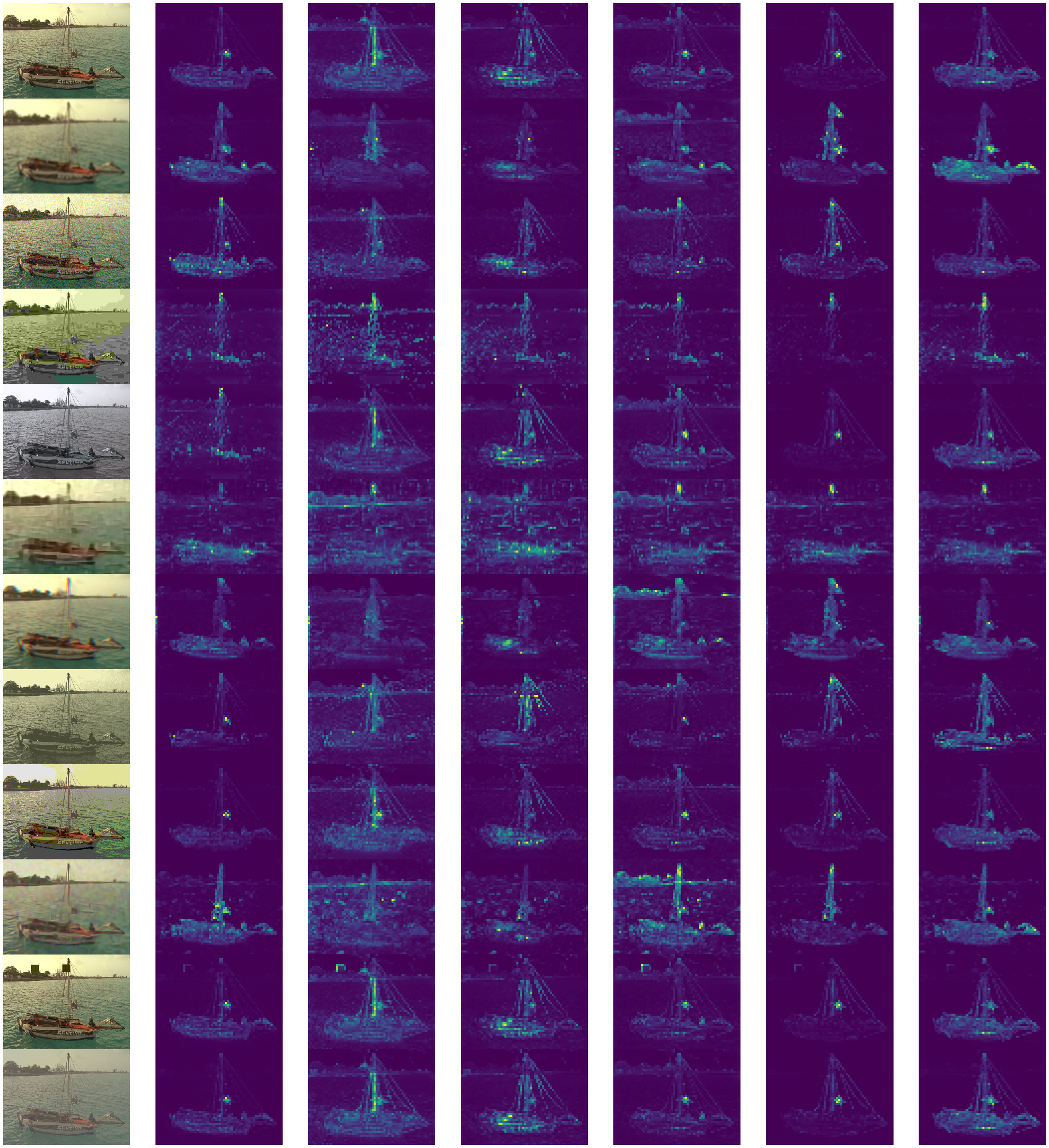


Figure 1: Attention heads of the DINO-ViTS/8 model from an image of the TID2013 dataset [36] with different qualities.

Name	# Reference Images	# Distorted Images	Subjective Study	# Distortion Types
TID2013	25	3000	Lab	24
MDID2016	20	1600	Lab	5

Table 1: Summary of the datasets used for the study

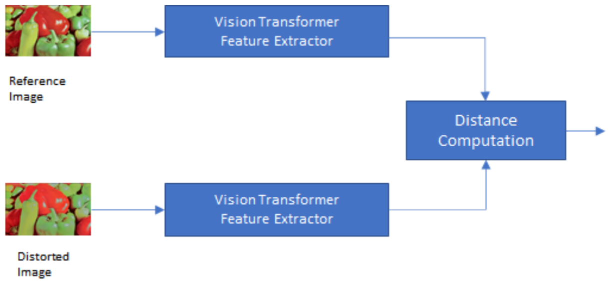

Vision Transformer Feature Extractor
Reference Image
Distance Computation
Vision Transformer Feature Extractor
Distorted Image

Figure 2: Pipeline for reference-based distortion analysis

Model	Architecture	Feature-vector size
dino-vits8	ViT-S/8	785 × 384
dino-vits16	ViT-S/16	197 × 384
dino-vitb8	ViT-B/8	785 × 768
dino-vitb16	ViT-B/16	197 × 768

Table 2: Dimensionality of the features obtained from the different models for this study

tance measures have a competitive correlation for three publicly available challenging IQA datasets with different subjective experiments and protocols. One of the key observations is that transformer features have decent correlation values for the MDID2016 and TID2013 databases. For the baseline, we compare with a perception-based image quality measure that combines contrast, luminance, structure, and the corresponding values reported by the authors of the respective datasets in Table 5. From Table 5, it is observed that the distance measures between pristine and distorted images extracted by the vision transformers give a better estimate of perceptual quality compared to a widely used image quality measure (SSIM). This motivates us to incorporate features based on self-supervised vision transformers to design IQA algorithms based on human perception.

Conclusion and Future Work

In this article, we observe that the features of the self-supervised vision transformer can demonstrate quality-related information. In this paper, we do not propose any explicit full reference image quality measure but instead we establish that the distance between feature vectors obtained from reference and distorted images of the same have shown promising correlation with the subjective human opinion scores, and this knowledge can further be leveraged in future to design reference-based image quality assessment techniques which mimic the human visual system more closely. Distances performed better than the structural similarity index measure, which is one of the widely used full reference image quality assessment measures. The images with poor perceptual quality had a greater distance from the pristine reference image in the feature space. One of the challenges in deep learning-based image quality assessment algorithms is the availability of data with quality variability and their corresponding human opinion scores to train them. The results presented in this study show that features extracted from pre-trained self-supervised vision transformer models for other tasks are in line with the human visual system and can be used to design perceptual metrics. The data sets used for this study are fundamentally different and use different types of distortion and even multiple distortions (MDID2016). One of the challenges in IQA is that different distortions have different statistical properties, and it is challenging to design a common IQA measure that works well on different types of distortion, but features based on vision transformers show positive results in that direction. The different augmentation strategies during the training of the DINO models expose these models to images of the same scene with different quality distortions, and this helps the models to distinguish a pristine image from a distorted one with a fair amount of accuracy. Training IQA specific datasets has a great limitation that the model trained on such datasets tends to work on the distortion specific to those datasets and fails on other unknown distortion types. One of the solutions could be quality-aware data augmentation and contrastive learning [34, 45]. The experimental results here motivate us to further explore transformer architectures trained in other different computer vision tasks and adapt to the field of image quality assessment, and more advanced transformer architectures [23, 24, 46, 47] have recently been proposed, making it a promising research area.

Distance	Type	Formula
Manhattan	L1	$\sum_i \lvert X_i - Y_i \rvert$
Canberra	L1	$\sum_i \frac{\lvert X_i - Y_i \rvert}{\lvert X_i \rvert + \lvert Y_i \rvert}$
Euclidean	L2	$\sum_i (\lvert X_i - Y_i \rvert^2)^{\frac{1}{2}}$

Table 3: Distance measures calculated between the reference and target images

References

[1] HR Sheikh and AC Bovik. A visual information fidelity approach to video quality assessment. In *1st Intl. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, volume 7, pages 2117–2128. sn, 2005.

[2] W Xue, L Zhang, X Mou, and AC Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE trans. image processing*, 23(2):684–695, 2013.

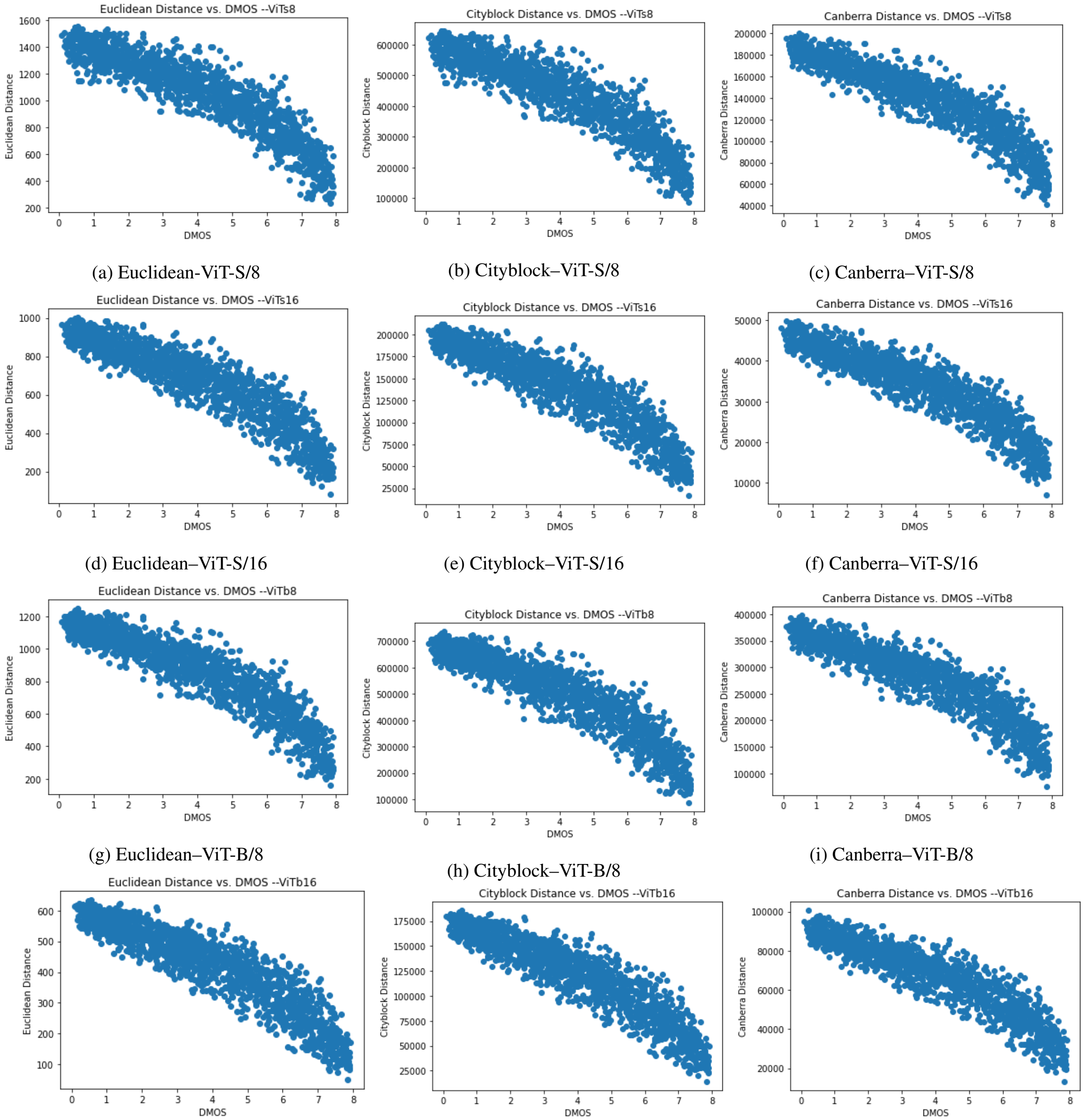
[3] A Maalouf, MC Larabi, and C Fernandez-Maloigne. A grouplet-based reduced reference image quality assessment. In *2009 Intl. Workshop on Quality of Multimedia Experience*, pages 59–63. IEEE, 2009.

[4] W Xue and X Mou. Reduced reference image quality assessment based on weibull statistics. In *2nd Intl. Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2010.

[5] R Soundararajan and AC Bovik. Rred indices: Reduced reference entropic differencing for image quality assessment. *IEEE Trans. on Image Processing*, 21(2):517–526, 2011.

[6] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. on image processing*, 21(12):4695–4708, 2012.

[7] Z Wang, AC Bovik, HR Sheikh, and EP Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE trans. on image processing*, 13(4):600–612, 2004.



(a) Euclidean-ViT-S/8

(b) Cityblock–ViT-S/8

(c) Canberra–ViT-S/8

(d) Euclidean–ViT-S/16

(e) Cityblock–ViT-S/16

(f) Canberra–ViT-S/16

(g) Euclidean–ViT-B/8

(h) Cityblock–ViT-B/8

(i) Canberra–ViT-B/8

(j) Euclidean–ViT-B/16

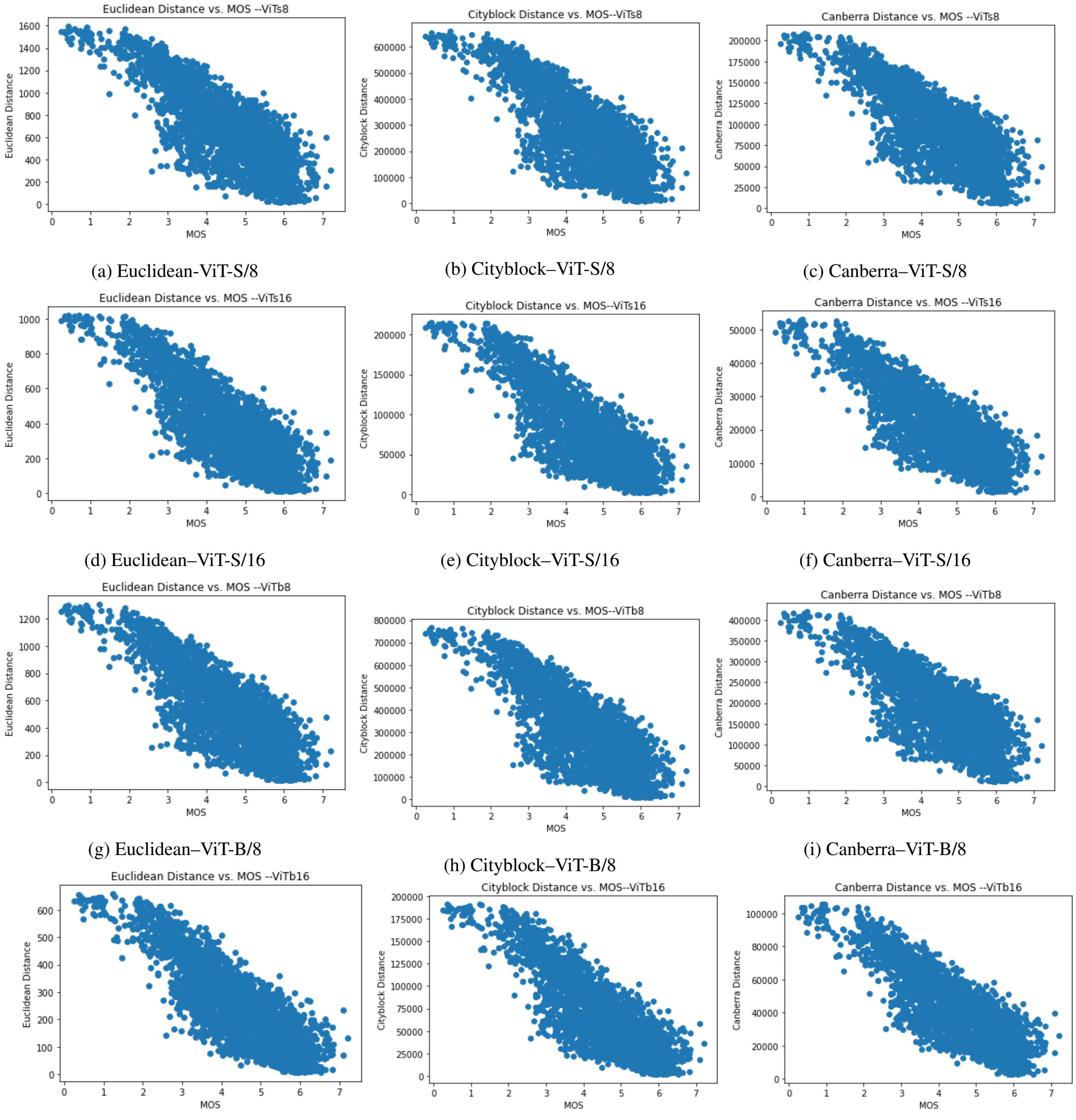
(k) Cityblock–ViT-B/16

(l) Canberra–ViT-B/16

Figure 3: Scatter plots to demonstrate the relationship between distance measures and subjective human opinion scores for the MDID2016 [44] dataset

Dataset	Model	Euclidean			Cityblock			Canberra		
		SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC
MDID-2016	dino-vits8	0.92	0.75	0.91	0.92	0.75	0.91	0.94	0.78	0.92
	dino-vits16	0.92	0.75	0.91	0.92	0.76	0.92	0.93	0.76	0.92
	dino-vitb8	0.93	0.77	0.91	0.93	0.76	0.91	0.93	0.77	0.91
	dino-vitb16	0.93	0.76	0.91	0.93	0.76	0.92	0.93	0.77	0.92
TID-2013	dino-vits8	0.78	0.58	0.81	0.76	0.57	0.80	0.75	0.56	0.79
	dino-vits16	0.80	0.6	0.83	0.79	0.59	0.83	0.78	0.59	0.82
	dino-vitb8	0.78	0.58	0.81	0.76	0.57	0.8	0.75	0.56	0.79
	dino-vitb16	0.82	0.62	0.85	0.81	0.61	0.84	0.8	0.61	0.83

Table 4: Quantitative measures to establish the relationship between distance measures and human opinion scores



(a) Euclidean-ViT-S/8 (b) Cityblock–ViT-S/8 (c) Canberra–ViT-S/8

(d) Euclidean–ViT-S/16 (e) Cityblock–ViT-S/16 (f) Canberra–ViT-S/16

(g) Euclidean–ViT-B/8 (h) Cityblock–ViT-B/8 (i) Canberra–ViT-B/8

(j) Euclidean–ViT-B/16 (k) Cityblock–ViT-B/16 (l) Canberra–ViT-B/16

Figure 4: Scatter plots to demonstrate the relationship between distance measures and subjective human opinion scores for the TID2013 dataset

Dataset	SRCC	KRCC	PLCC
TID2013	0.71	0.46	0.75
MDID2016	0.71	0.53	0.74

Table 5: Baseline- Structural Similarity Index Measure

[8] W Hou, X Gao, D Tao, and X Li. Blind image quality assessment via deep learning. *IEEE trans. on neural networks and learning systems*, 26(6):1275–1286, 2014.

[9] S Bosse, D Maniry, KR Müller, T Wiegand, and W Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. on image processing*, 27(1):206–219, 2017.

[10] J Kim and S Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *Proc. IEEE conf. on computer vision and pattern recognition*, pages 1676–1684, 2017.

[11] S Bianco, L Celona, P Napoletano, and R Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2018.

[12] M Zhang, L Zhang, W Hou, and J Feng. Blind image quality assessment with visual sensitivity enhanced dual-channel deep convolutional neural network. In *2020 12th Intl. Conf. on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2020.

[13] S Ahn, Y Choi, and K Yoon. Deep learning-based distortion sensitivity prediction for full-reference image quality assessment. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 344–353, 2021.

[14] Y Liu, S Li, and Q Chen. Progress of no-reference image quality assessment based on deep learning. In *14th Intl. Conf. on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 397–402. IEEE, 2022.

[15] Z Pan, F Yuan, X Wang, L Xu, S Xiao, and S Kwong. No-reference image quality assessment via multi-branch convolutional neural networks. *IEEE Trans. on Artificial Intelligence*, 2022.

[16] S Dodge and L Karam. Human and dnn classification performance on images with quality distortions: A comparative study. *ACM Trans. on Applied Perception (TAP)*, 16(2):1–17, 2019.

[17] D Hendrycks and T Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proc. of the Intl. Conf. on Learning Representations*, 2019.

[18] T Borkar and LJ Karam. Deepcorrect: Correcting dnn models against image distortions. *IEEE Trans. on Image Processing*, 28(12):6022–6034, 2019.

[19] J Devlin, M Chang, K Lee, and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[21] 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*, 2020.

[22] J Deng, W Dong, R Socher, LJ Li, K Li, and F Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conf. on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[23] Z Liu, Y Lin, Y Cao, H Hu, Y Wei, Z Zhang, S Lin, and B Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, pages 10012–10022, 2021.

[24] H Touvron, M Cord, M Douze, F Massa, A Sablayrolles, et al. Training data-efficient image transformers and distillation through attention. In *Intl. Conf. on Machine Learning*, pages 10347–10357. PMLR, 2021.

[25] A Ali, H Touvron, M Caron, P Bojanowski, M Douze, A Joulin, I Laptev, N Neverova, G Synnaeve, J Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021.

[26] J You and J Korhonen. Transformer for image quality assessment. In *2021 IEEE Intl. Conf. on Image Processing (ICIP)*, pages 1389–1393. IEEE, 2021.

[27] M Cheon, S Yoon, B Kang, and J Lee. Perceptual image quality assessment with transformers. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 433–442, 2021.

[28] S Bhojanapalli, A Chakrabarti, D Glasner, D Li, T Unterthiner, and A Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF Intl. Conf. on Computer Vision*, pages 10231–10241, 2021.

[29] S Paul and PY Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2(3), 2021.

[30] Y Bai, J Mei, AL Yuille, and C Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34, 2021.

[31] T Chen, S Kornblith, M Norouzi, and G Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[32] K He, H Fan, Y Wu, S Xie, and R Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[33] M Caron, H Touvron, I Misra, H Jégou, J Mairal, P Bojanowski, and A Joulin. Emerging properties in self-supervised vision transformers. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, pages 9650–9660, 2021.

[34] P Chen, L Li, Q Wu, and J Wu. Spiq: A self-supervised pre-trained model for image quality assessment. *IEEE Signal Processing Letters*, 2022.

[35] JB Grill, F Strub, F Altché, C Tallec, P Richemond, E Buchatskaya, C Doersch, B Avila Pires, Z Guo, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[36] N Ponomarenko, L Jin, O Ieremeiev, V Lukin, K Egiazarian, J Astola, B Vozel, K Chehdi, M Carli, F Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

[37] HR Sheikh, MF Sabir, and AC Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. on image processing*, 15(11):3440–3451, 2006.

[38] D Jayaraman, A Mittal, AK Moorthy, and AC Bovik. Objective quality assessment of multiply distorted images. In *Proc. 46th asilomar conf. on signals, systems and computers (ASILOMAR)*, pages 1693–1697. IEEE, 2012.

[39] H Lin, V Hosu, and D Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh Intl. Conf. on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[40] EC Larson and DM Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010.

[41] A Zaric, N Tatalovic, N Brajkovic, H Hlevnjak, M Loncaric, E Dumic, and S Grgic. Vcl@ fer image quality assessment database. In *Proceedings ELMAR-2011*, pages 105–110. IEEE, 2011.

[42] T Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[43] P Virtanen, R Gommers, TE Oliphant, M Haberland, T Reddy, D Cournapeau, E Burovski, P Peterson, W Weckesser, J Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[44] W Sun, F Zhou, and Q Liao. Mdid: A multiply distorted image database for image quality assessment. *Pattern Recognition*, 61:153–168, 2017.

[45] PC Madhusudana, N Birkbeck, Y Wang, B Adsumilli, and AC Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 2022.

[46] H Touvron, M Cord, A Sablayrolles, G Synnaeve, et al. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, October 2021.

[47] S Atito, M Awais, and J Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021.

Author Biography

Kanjar De completed his Ph.D. from Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram in 2017 and a Master in Information Technology from International Institute of Information Technology, Bangalore, His research interests include Image Quality Assessment, applied machine learning, and colour imaging. and SPIE.