ICIC Express Letters
Volume **18**, Number **11**, November **2024**
 ISSN 1881-803X
pp. **1145–1152**

IMBALANCE PROBLEM IN IMAGE QUALITY ASSESSMENT

Huu Thang Nguyen[1], Anh Chung Hoang[2], Manh Cuong Bui[2]
and Tuan Linh Dang[2,*]

[1]Research and Development Department
Pixelz Inc
No. 107, Nguyen Phong Sac, Cau Giay, Hanoi 100000, Vietnam
thangnh@pixelz.com

[2]School of Information and Communications Technology
Hanoi University of Science and Technology
No. 1, Dai Co Viet, Hanoi 100000, Vietnam
{ chung.ha204901; cuong.bm204871 }@sis.hust.edu.vn
*Corresponding author: linhdt@soict.hust.edu.vn

Received January 2024; accepted April 2024

Abstract. *Image Quality Assessment (IQA) is crucial in computer vision, yet label imbalance poses a persistent challenge, impacting model evaluation and generalization. This study proposes a novel approach to address IQA label imbalance by integrating Contrastive Language-Image Pre-training (CLIP) with Label Distribution Smoothing (LDS) and Feature Distribution Smoothing (FDS). Our method not only introduces an innovative pipeline for IQA but also effectively mitigates label imbalance. Experimental results on the KonIQ-10k dataset, yielding approximately 6.945 Mean Absolute Error (MAE), and on our internal dataset, achieving 0.345 MAE for target values with limited samples, demonstrate significant improvements, with a notable reduction in bias towards regions with abundant samples. This research contributes to advancing IQA methodologies and offers promising directions for future investigation.*
Keywords: IQA, Contrastive Language-Image Pre-training (CLIP), Deep regression, Imbalanced dataset, Label Distribution Smoothing (LDS), Feature Distribution Smoothing (FDS)

1. **Introduction.** In the dynamic field of computer vision, deep learning has been a driving force behind significant advancements, particularly in the area of Image Quality Assessment (IQA). This field is crucial for evaluating the perceptual quality of images and plays a vital role in domains such as surveillance, where accurate image analysis is paramount. It also has a significant impact on the user experience in advertising and social media platforms, where the quality of visual content can influence user engagement.

However, like many challenges in deep learning, addressing imbalanced datasets in the IQA presents a significant difficulty. The issue occurs when some classes in the dataset have fewer images compared to others, significantly impairing the model's performance to generalize effectively. Effectively addressing this imbalance is crucial in order to greatly enhance the accuracy of IQA models, particularly in applications where detailed evaluations of image quality are absolutely vital.

This paper introduces a new research trend by proposing a pipeline that utilizes a custom CLIP backbone and various methods to address the issue of imbalanced datasets. This innovative approach not only improves the accuracy of IQA models but also serves as a foundation for other research groups to further refine and enhance their work.

In practical terms, the advancements presented in this paper have the potential to revolutionize the way images are selected and recommended on e-commerce platforms.

DOI: 10.24507/icicel.18.11.1145

IQA can help in automatically selecting or generating images that are most appealing to users, based on learned human visual preferences. By accurately assessing the quality of images, our model can enhance image recommendation systems that could improve customer satisfaction and drive sales.

The manuscript is structured as follows. Section 2 presents an overview of relevant work. Section 3 details the proposed method. Section 4 outlines the experimental results. Finally, Section 5 presents the conclusion of the paper.

2. **Related Works.**

2.1. **IQA datasets.**

KonIQ-10k [1] is the dataset containing 10,173 in-the-wild images collected in crowdsourcing environments. Over 1,459 individuals meticulously employed 120 rigorous criteria to assess the exceptional quality of each and every image. We can create models with more accuracy and generalizability thanks to various ratings.

The Smartphone Photography Attribute and Quality (SPAQ) database, as presented in [2], consists of 11,125 images taken by 66 different smartphones. The vast database encompasses a diverse array of human evaluations for each photograph, providing comprehensive assessments of image quality and various attributes including brightness, colorfulness, contrast, noise, and sharpness. Moreover, the dataset includes labels that classify images into different scene categories. These categories comprise animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and more.

Nevertheless, these datasets exhibit an imbalance, with a notably smaller number of high-quality and low-quality images. This inherent imbalance in their nature reflects the likelihood that normal-quality images are more commonly captured. As part of our research, we will examine the distribution of the dataset and work on implementing robust methods to tackle the imbalance issues within the IQA dataset.

2.2. **Imbalanced regression.**

2.2.1. *Label distribution smoothing.* Label Distribution Smoothing (LDS) [3] is designed to handle imbalanced data by estimating the dataset's effective label density distribution. The development of LDS was prompted by experiments that showed the power of using the original label density distribution of a dataset as a weighting loss function in imbalanced classification tasks. However, it may not be practical in imbalanced regression tasks due to the hidden information between nearby target values, such as images with a similar Mean Opinion Score (MOS).

Label distribution smoothing addresses this problem by using kernel density estimation to effectively estimate the label density distribution of the dataset. A symmetry kernel, such as the Gaussian kernel, is convoluted with the original density distribution, resulting in a smooth distribution that allows nearby labels to share information in data samples. Various strategies are available to utilize the weighted loss function effectively. To enhance the accuracy of the results, techniques such as inverting or calculating the square root of the estimated density obtained from LDS for every target value could be used.

2.2.2. *Feature distribution smoothing.* Feature Distribution Smoothing (FDS), introduced in [3], along with LDS, is a technique to handle imbalanced deep regression. FDS operates in feature space, unlike LDS, which operates in label space. The idea behind this approach is that data points near the target value should show similarities in their feature embeddings. For example, in IQA, images in nearby MOS should have a close distance in their feature embedding.

The technique modifies the feature mean and covariance distribution using kernel smoothing to smoothen the feature statistics distribution. Then, the whitening and re-coloring procedure introduced in [4] is applied to the feature embedding of each data point to calibrate the feature. Feature distribution smoothing can be integrated seamlessly into

network architectures by adding the FDS calibration layer after the feature extractor. The combination of LDS and FDS has demonstrated its effectiveness in various imbalanced regression tasks [3], such as age prediction and depth estimation.

2.3. **IQA method.**

2.3.1. *Traditional methods.* Traditional IQA models are usually divided into two stages: identifying distortion types and evaluating quality according to distortion types. These models often use handcrafted features from the frequency domain or the spatial domain of images, and use machine learning methods such as support vector machine to classify distortion types and regress quality scores. Examples of these models are BIQI [5], DIIVINE [6], BLIINDS [7], BLIINDS-II [8], BRISQUE [9], and NIQE [10].

However, these models have some limitations that they cannot have the deep understanding of the nature of the image and the distortion types, i.e., these features only reflect the objective quality of the image and cannot capture the human preference or priority when viewing the image.

2.3.2. *CLIP for IQA.* Traditional deep learning methods for NR-IQA usually require designing network architectures that are suitable for the type of noise and impact on the image quality, as well as pretraining the network with labeled IQA dataset. However, these cannot exploit the ability of a deep learning model that is pre-trained on a large and diverse dataset, enhancing the generalization ability for real-world situations.

Some studies have exploited the potential of CLIP [11] for the IQA problem. For example, [12] proposed a new IQA method, which can be explained, using multiple pairs of contrasting words corresponding to the features describing the image quality. This method not only can estimate the quality score of the image, but also identify the causes leading to that score. Experimental results show that this method outperforms existing zero-shot IQA methods in terms of accuracy and can evaluate the causes of image quality degradation. [12] explored CLIP for assessing the appearance and feeling of images. They proposed a new IQA method, using some suggestive words to describe the desired appearance and feeling of the image, and using the contrast between the image and the suggestive words to estimate the image quality. They also provided many experiments on controlled datasets and standard IQA datasets.

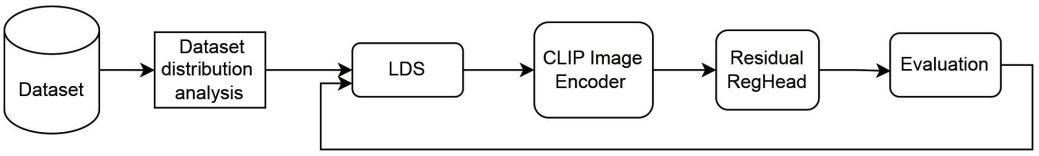
Therefore, acknowledging the limitations of traditional IQA methods in grasping the intrinsic nature of images and types of distortion, which merely reflect objective image quality without capturing human preferences, our work adopts the CLIP approach to better reflect human perception of image quality and gain more generalization on inferencing.

3. **Proposed Method.**

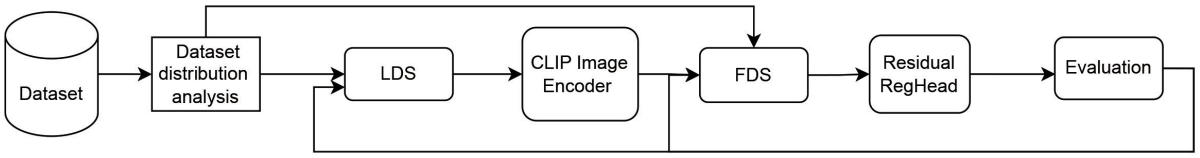
Our study unfolds in two primary stages for each dataset. In the first stage, we delve into the LDS method using the CLIP image encoder as a backbone, analyzing data distribution and tackling imbalances. Following successful initial experiments, we refine the LDS application by tuning bin size and Gaussian kernel parameters of it to create a smoothed label density distribution. We then apply a weighted loss function based on this distribution, focusing on training the regression head of the model CLIP. In the second stage, we proceed to reuse the LDS label smoothed distribution that were set in the previous step, and proceed to train further with FDS.

This study thoroughly analyzes the KonIQ-10k dataset and an internal dataset, both of which are image quality assessment datasets that depict real-world scenarios. We assess the extent of the data imbalance and its influence on the performance of image quality assessment models by examining the data distribution. By doing that, we implement potent methods to address the data imbalance problem in the dataset, including LDS and FDS, which are expected techniques for tackling label and feature imbalances. Simultaneously, in testing the methods to deal with the data imbalance mentioned above, we also

experimented with the model CLIP. This powerful image encoder can analyze vast sets of image-text pairs from the current dataset to assess its versatility in evaluating image quality. The details of the model CLIP settings on the IQA dataset are presented in Section 3.1, and the data analysis and evaluation are presented in Section 3.2. Overall, the process of conducting data research and applying methods is illustrated in Figure 1.



(a) The overall pipeline of LDS method process



(b) The overall pipeline of combining LDS and FDS method process

FIGURE 1. The total stages of imbalance handling methods used

3.1. **Vanilla CLIP model for IQA.** Following [3], we use the term *Vanilla Model* as models that do not adopt any imbalance-handling method. We train a vanilla CLIP model with transfer learning from the pretrained image encoder `vit_base_patch32_224_clip_laion2b` from Laion [13] because CLIP is a foundation model of computer vision tasks trained with a very large dataset of up to hundreds of millions of image-text pairs and it can store the semantics of text.

We also apply additional residual connections to resolving the gradient vanishing problem when stacking multiple layers on top of each other. Thus, the embedding feature vectors of images can be learned more efficiently. Specifically, we remove the text encoder layer of CLIP, and only use the image encoder of CLIP. That is, we feed the image encoding vector of CLIP, which usually has 768 dimensions, as the input. Then, we freeze the image encoder layer, to use all the knowledge that has been learned before, to extract the semantic features from the image quality assessment dataset, and use a new projection layer to map the semantic features to the output for the fully-connected layer. The detailed architecture is described through the figure (Figure 2).

3.2. **Dataset analysis.**

3.2.1. *The KonIQ-10k dataset.* After doing an investigation about the imbalance of the dataset, we find that the dataset was extremely on the state of imbalance. Beforehand, for convenience, since the domain of label is on $\mathbb{R}$, we set the bin size as 0.1. Figure 3 shows clearly that the labels of the dataset are skewed to one side in the distribution. Specifically, the labels are concentrated on the range 2.2-3.9, while the labels from 0-1 and from about 4.2 to 5 are zero-shot samples. The few-shot samples are in range 1-2 and 3.5-4.2.

As the vanilla model in Section 3.1 undergoes the early stages of training, it becomes clear that the loss in the labeled area of the rich label significantly decreases when the number of labels increases. In this area, the loss value is remarkably low. On the other hand, the two sides with fewer labels show a significantly high loss. What truly sets it apart is its extraordinary shape – a bell curve that slightly tilts to the right. On the right side, the label area may appear smaller, but it is surrounded by a cluster of labels. On the left side, there is a neighboring area with a label density that gradually increases, forming

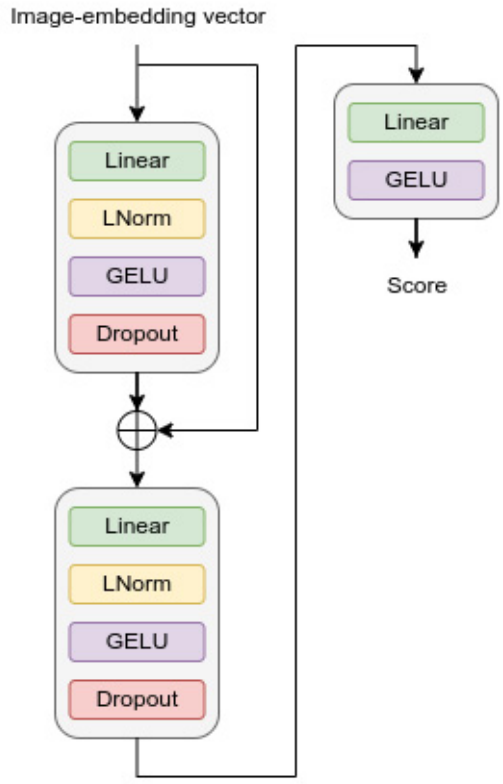


FIGURE 2. The residual MLP of regression head for CLIP

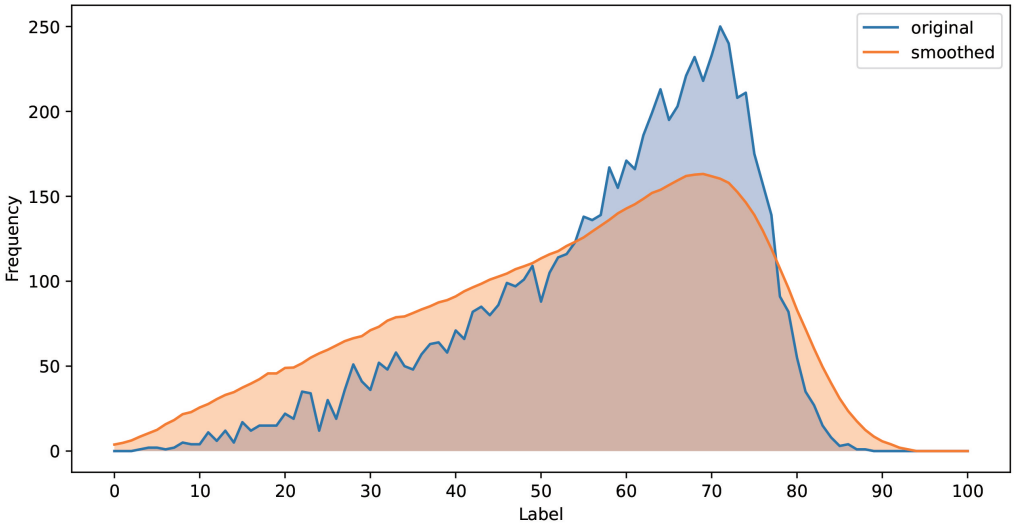


FIGURE 3. The KonIQ-10k dataset exhibits a right-skewed distribution. The original scale ranges from 0 to 100, but we have rescaled it to 0 to 5 for analysis, with a mode around 3.5 and most values falling between 2.2 and 3.9.

a steep slope. However, the label area in this neighbor is significantly smaller compared to other parts. It can be inferred from this that the neighboring labels have a significant impact on the nearby few-shot label areas. Using the same weight for these two regions is not a fair approach. In this case, we ultilize LDS to accurately determine the weight for each region.

In addition, to enhance the balance between the labels, we also use FDS, to help the CLIP feature vector embedding become more balanced, because it has to learn on a

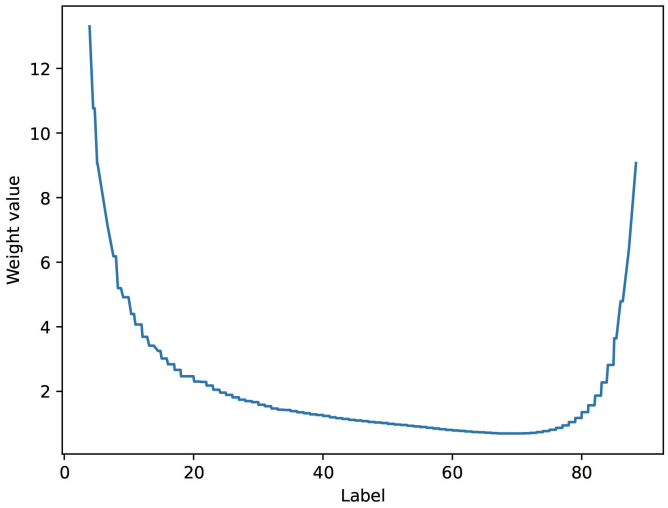


FIGURE 4. The weight value of labels after using LDS

dataset that is quite biased towards the mode of distribution, so calibrating the model's feature is necessary to help the model stabilize the features before passing them through the fully-connected layer to learn the exact score characteristics according to each region.

3.2.2. *The internal dataset.* Our internal dataset consists of 70,000 images, including food, locations, scenery, and people. These images have a standard size of 224×224 pixels. We assess the excellence of the images by relying on MOS ratings provided by a panel of annotators utilizing a scale ranging from 0 to 4. Higher scores ultimately signify superior image quality.

After labeling the data, our next step is to carry out an extensive analysis to evaluate the distribution and identify any imbalances in the dataset. In Figure 5, we can see that the dataset also faces a very serious problem of imbalance, with most of the sample points belonging to interval $[0, 2]$. Labels 3 and 4 are actually rare occurrences, and training the model without addressing the imbalance in the dataset will inevitably introduce bias towards these specific areas. This will concentrate the test loss in the $[3, 4]$ region, particularly affecting sample points with a score of 4. Training the model with such a dataset will undoubtedly lead to outstanding performance, as it will be capable of predicting labels consisting solely of numbers 1 or 2, resulting in a remarkable overall Mean Absolute Error (MAE) score. However, this expectation is nothing but a mere

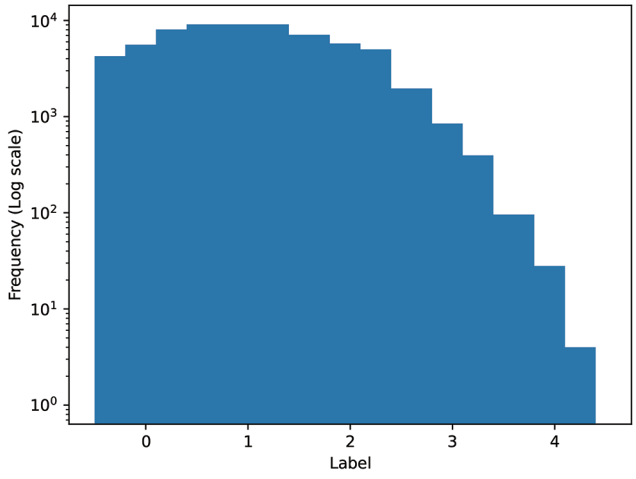


FIGURE 5. Distribution about internal dataset

hallucination. Effective techniques can be applied to address data imbalance and ensure unbiased results, as demonstrated with the KonIQ-10k dataset.

After analysis, we gain the combination of parameters for LDS and FDS as LDS: $r = 1$, $m = 100$, $w(x) = \sqrt{\frac{1}{x}}$, $k_{LDS} = \mathcal{N}(12, 10)$ and FDS: $u = 0$, $s = 1$, $k_{FDS} = \mathcal{N}(5, 2)$ where: $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ.

- For LDS, r, m, $w(x)$, k_{LDS} are symbols for the parameters resolution, max target, reweighting function, and kernel of LDS.
- For FDS, u, s, k_{FDS} are symbols for the parameters of the start update epoch, start smooth epoch and kernel of FDS.

4. Experiments.

4.1. Environment and hyperparamters settings.

In this study, we performed experiments on GPU NVIDIA P100, running each experiment for 200 epochs continuously. The images were resized to 224×224 to be able to use the backbone image encoder from CLIP directly. The experiment involved the Vanilla CLIP model, a fusion of the CLIP model with LDS, as well as a merger of the CLIP image encoder with both LDS and FDS.

We used Adam [14] as the optimizer. For the KonIQ-10k dataset, we used a learning rate of 0.1, while for the internal dataset, we used a learning rate of 0.01. This was because the label range of KonIQ-10k was from $[0, 100]$, which was wider and larger than the internal dataset, which was from $[0, 4]$. Other hyperparameters of Adam were the same in both datasets which were set as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, $\lambda = 0$, where β_1, β_2 are coefficients of computing averages of gradient in exponential moving average strategy.

4.2. KonIQ-10k.

Having analyzed the imbalance and the distribution of the dataset, we proceed to train experiments with sequentially adding components. To evaluate, we use the MAE (Mean Absolute Error) metric for each region. Overall, the methods for addressing the imbalance issue have been consistently effective across various datasets, notably the KonIQ-10k dataset.

TABLE 1. Ablation study on KonIQ-10k dataset. The highlight values show the best result. CLIP + LDS + FDS shows up to be the best in "All" and the most sample region $[60, 80]$ case. The optimal MAE for the whole dataset is with CLIP + LDS.

Model	$[0, 20]$	$[20, 40]$	$[40, 60]$	$[60, 80]$	$[80, 100]$	All
CLIP	8.2582	10.4142	9.3396	5.8455	2.9101	7.4405
CLIP + LDS	**6.9505**	**8.3953**	**8.5447**	5.8799	**2.703**	6.9477
CLIP + LDS + FDS	8.0852	9.6857	8.6172	**5.4826**	3.3175	**6.9449**

4.3. Internal dataset.

Table 2 shows results in our internal dataset. In our internal dataset, we have found an acceptable trade-off between the MAE of all labels and the MAE of specifically labels 3 and 4. This balance is achieved while also significantly improving the overall balance between labels 3, 4, and the dataset as a whole. The difference noticeably decreases when using the methods LDS and then combining both LDS and FDS.

5. Conclusion.

In this publication, we have explored the data imbalance issue in the most famous and standard IQA dataset, KonIQ-10k, in Section 3.2. We have applied various methods to mitigating the effect of data imbalance on the test loss, and also proposed a simple regression head to help fit the CLIP embedding vectors in IQA more accurately at Section 3.1.

TABLE 2. Ablation study on internal dataset. Highlighted values represent the best result. Our experiments demonstrate that the combination of LDS and FDS significantly reduce the MAE on images with groundtruth values in the range $(3, 4)$. While the overall MAE is increased slightly, it is due to the rebalancing effect towards low-sample data.

Model	3	4	3&4	All
CLIP	0.9282	1.8276	0.9548	**0.3297**
CLIP + LDS	0.4270	0.9321	0.4420	0.3844
CLIP + LDS + FDS	**0.3421**	**0.5934**	**0.3458**	0.3766

In addition, we have introduced a new IQA dataset to evaluate the effectiveness of these methods in an objective way in Section 4.3. Ensuring objectivity in evaluation is essential when dealing with imbalanced datasets, making it crucial to apply dataset imbalance-handling methods to addressing this issue of authenticity imbalance. In practice, we have verified the effectiveness of these methods, and obtained improvement on both datasets we have mentioned. The combination of CLIP, LDS, and FDS in the internal dataset consistently outperforms all other approaches, producing a remarkable MAE of 0.3766 for the whole dataset. The result of CLIP + LDS + FDS shows up to be balancing in KonIQ-10K. approximately 6.9449 of all samples.

Acknowledgment. This research is funded by the Hanoi University of Science and Technology (HUST) under project number T2022-PC-052.

REFERENCES

[1] V. Hosu, H. Lin, T. Sziranyi and D. Saupe, KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment, *IEEE Transactions on Image Processing*, vol.29, pp.4041-4056, 2020.

[2] Y. Fang, H. Zhu, Y. Zeng, K. Ma and Z. Wang, Perceptual quality assessment of smartphone photography, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] Y. Yang, K. Zha, Y. Chen, H. Wang and D. Katabi, Delving into deep imbalanced regression, *International Conference on Machine Learning*, pp.11842-11851, 2021.

[4] B. Sun, J. Feng and K. Saenko, Return of frustratingly easy domain adaptation, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.30, 2016.

[5] A. K. Moorthy and A. C. Bovik, A two-step framework for constructing blind image quality indices, *IEEE Signal Processing Letters*, vol.17, no.5, pp.513-516, 2010.

[6] A. K. Moorthy and A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, *IEEE Transactions on Image Processing*, vol.20, no.12, pp.3350-3364, 2011.

[7] M. A. Saad, A. C. Bovik and C. Charrier, A DCT statistics-based blind image quality index, *IEEE Signal Processing Letters*, vol.17, no.6, pp.583-586, 2010.

[8] M. A. Saad, A. C. Bovik and C. Charrier, DCT statistics model-based blind image quality assessment, *2011 18th IEEE International Conference on Image Processing*, pp.3093-3096, 2011.

[9] A. Mittal, A. K. Moorthy and A. C. Bovik, No-reference image quality assessment in the spatial domain, *IEEE Transactions on Image Processing*, vol.21, no.12, pp.4695-4708, 2012.

[10] A. Mittal, R. Soundararajan and A. C. Bovik, Making a "completely blind" image quality analyzer, *IEEE Signal Processing Letters*, vol.20, no.3, pp.209-212, 2013.

[11] 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, *International Conference on Machine Learning*, pp.8748-8763, 2021.

[12] J. Wang, K. C. K. Chan and C. C. Loy, Exploring clip for assessing the look and feel of images, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.37, pp.2555-2563, 2023.

[13] G. Ilharco, M. Wortsman, R. Wightman et al., *OpenCLIP*, Zenodo, DOI: 10.5281/zenodo.5143773, 2021.

[14] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *Proc. of the 3rd International Conference for Learning Representations*, 2015.