QPT V2: Masked Image Modeling Advances Visual Scoring

Anonymous Authors

ABSTRACT

Quality assessment and aesthetics assessment aim to evaluate the perceived quality and aesthetics of visual content. Current learning-based methods suffer greatly from the scarcity of labeled data and usually perform sub-optimally in terms of generalization. Although masked image modeling (MIM) has achieved noteworthy advancements across various high-level tasks (*e.g.* , classification, detection *etc.*). In this work, we take on a novel perspective to investigate its capabilities in terms of *quality- and aesthetics-awareness*. To this end, we propose **Q**uality- and aesthetics-aware **P**re**T**raining (QPT V2), the first pretraining framework based on MIM that offers a unified solution to quality and asthectics assessment. To perceive the high-level semantics and fine-grained details, pretraining **data** is curated. To comprehensively encompass quality- and aesthetics-related factors, **degradation** is introduced. To capture multi-scale quality and aesthetic information, **model** structure is modified. Extensive experimental results on 11 downstream benchmarks clearly show the superior performance of QPT V2 in comparison with current state-of-the-art approaches and other pretraining paradigms.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Computer vision tasks**; **Scene understanding**;

KEYWORDS

visual scoring, quality and aesthetics assessment, self-supervised learning, masked image modeling

1 INTRODUCTION

The aims of Image Quality Assessment (IQA), Visual Quality Assessment (VQA), and Image Aesthetics Assessment (IAA) are to appraise the quality and aesthetics of visual content, serving as critical components across a multitude of vision applications including video enhancement, transcoding, and transmission [34, 65, 105]. While being studied separately for a considerable period, these tasks present strong resemblance in various aspects. *All these tasks share the same core objective, that is, to mimic the Human Visual System (HVS), so as to generate accurate scores aligned with human perception* [27, 82, 83]. Moreover, the proliferation of User-Generated Content (UGC) [73, 77, 110] and AI-Generated Content (AIGC) [5, 93, 102, 103] has become a trend in recent years, which greatly contributed to the exponential growth of image and video data [3].

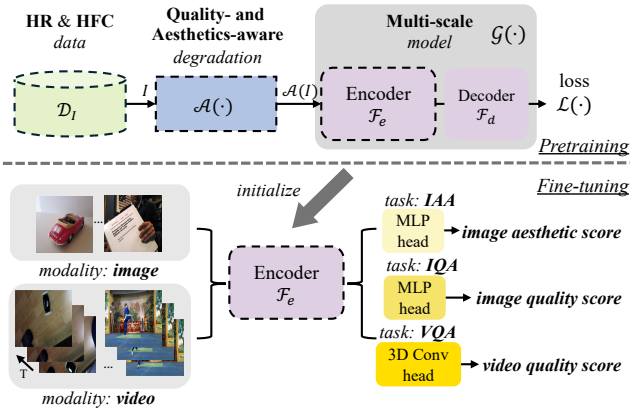


Figure 1: QPT V2: a new MIM-based pretraining paradigm for visual scoring. For pretraining, dataset $\mathcal{D}_I$ provides HR & HFC images, augmented by quality- and aesthetics-aware degradation $\mathcal{A}(\cdot)$. A multi-scale autoencoder $\mathcal{G}(\cdot)$ outputs the reconstructed images. Through finetuning of the encoder, it can solve visual scoring tasks like IQA, VQA, and IAA.

The complexity and interrelation of quality-related and aesthetics-related factors in emerging content are unprecedented, and analyzing single factors alone is insufficient to achieve a comprehensive perception of visual content aligned with human perception. In response to the aforementioned resemblance and trend, we refer to IQA, VQA, and IAA jointly as Visual Scoring (VS) for analysis.

Facilitated by the advancements of deep neural networks [13, 26, 41, 70, 81], learning-based methods [28, 43, 44, 72, 112] have surpassed traditional methods [10, 57, 78, 92] based on handcrafted features on multiple VS benchmarks [6, 9, 17, 20, 59, 86, 101]. They acquire features with strong expressiveness via regressing from the Mean Opinion Scores (MOS). However, one of the primary obstacles in solving VS lies in the limited size of labeled datasets [44, 46, 55, 74, 111]. Due to the high cost associated with collecting MOS through extensively annotated subjective studies, the scale of VS datasets is often only a fraction, ranging from one-tenth to even one-hundredth, of other high-level visual task datasets (*e.g.*, object recognition). At all events, the paucity of labeled data restricts the capabilities of data-driven deep learning methods.

To tackle this problem, some previous efforts increased data size by patch/frame-level augmentation [4, 35, 36, 46] or mixed-database training [39, 40, 94, 107]. However, the quality and aesthetics scores of local patches often differ from those of the entire content, and subjective differences are observed across datasets, thus hindering the achievement of promising results. On the other hand, a different research line [37, 76, 83, 89, 95] exploits knowledge valuable for VS from **datasets** and **model weights** of other domains, by tapping into the power of pretrained vision or vision-language (VL) models [63, 76, 89, 95, 98, 100, 106]. These works attempt to extract knowledge that is more quality- or aesthetic-aware from large-scale datasets by carefully designing **pretraining objectives**[44, 55], and

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

are then finetuned on downstream VS tasks. The pretraining objectives of existing works are mainly based on contrastive learning [67, 80], which can be viewed as a global self-supervised learning (SSL) approach, as it groups similar samples closer and diverse samples far from each other [25, 55, 111]. However, this "sample-level" discernment is insufficient for capturing local distortions and visual attributes [60]. Therefore, exploring more effective pixel-level discrimination may be beneficial for incorporating pretrained priors into downstream VS tasks.

Masked Image Modeling (MIM) [24], which learns representation by pixel-level reconstruction of the masked regions in the input, has demonstrated its impressive ability of semantic- and texture-aware perception in visual tasks [2, 61, 84, 109]. In this paper, we conduct a detailed exploration of MIM, in which we observe MIM can learn both sample-level and pixel-level information of the visual content, showing the potential to serve as a general pretraining recipe to VS tasks. As shown in Fig. 1, we propose QPT V2, the first pretraining framework based on MIM that offers a unified solution for VS tasks. To enhance the acquisition of prior knowledge by MIM for VS tasks, we propose further improvements and optimizations of the vanilla MIM from the perspectives of **data**, **degradation**, and **model**. *Regarding the realm of data*, we curate a dataset with high resolution (HR) and high foreground coverage (HFC), thereby aiding the pretext task of MIM. *Regarding the realm of degradation*, we propose an optimal strategy for applying degradations to the reconstruction target, exploring the type and composition of degradation to acquire prior knowledge of practical scenarios. *Regarding the realm of model*, we use a drop-in strategy to learn multi-scale representations by adaptively fusing features of different layers. Our main contributions can be summed as follows:

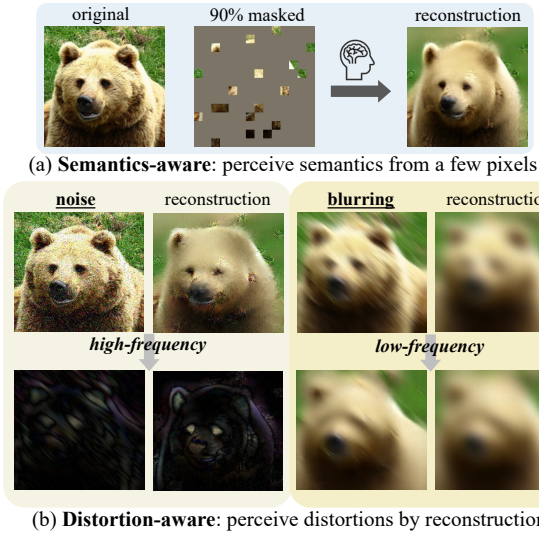
- To the best of our knowledge, we are the first to validate the capability of MIM in adeptly unifying downstream visual scoring tasks. We decompose MIM into three crucial components: data, degradation, and model, and individually investigate their respective influences.
- We propose QPT V2, which stands as the pioneering MIM-based pretraining framework, offering a unified solution for VS tasks. To enhance the acquisition of prior knowledge through MIM, we make targeted improvements in the aspects of data, degradation, and model.
- QPT V2 achieves state-of-the-art (SOTA) results on 11 benchmarks in IQA, VQA, and IAA, surpassing other pretraining paradigms as well. Extensive ablation studies prove the validity of each enhancement of MIM.

2 RELATED WORK

2.1 Visual Scoring

Visual scoring necessitates precise scoring of visual content in terms of quality (*e.g.* , IQA, VQA) and aesthetics (*e.g.* , IAA). In this work, we focus on Non-reference QA (*e.g.* , NR-IQA and NR-VQA), since the availability of pristine data is too hard in the real world. At the early stage, handcrafted features based on natural scene statistics (NSS) dominate the realm of VS [56, 57, 69]. Later, data-driven methods enhanced the performance significantly with the rise of deep learning [15, 35, 45, 83, 89]. Nonetheless, they rely heavily on label-intensive supervision. Previous works attempt to solve this problem by data augmentation [4, 35, 36, 99], mixed-database training [40, 46, 74, 94, 107], rank-based learning [47, 50] and general knowledge transfer [15, 83, 89, 108].

Several researches focus on extracting quality or aesthetics information by large-scale pretraining. Among them, CONTRIQUE [55] learns distortion-related information on images with synthetic and realistic distortions based on contrastive learning. Similarly, Re-IQA [66] re-engineers the MoCo-v2 [8] framework and applies intricate data augmentations to learn quality-aware features. Moreover, QPT [111] introduces a diverse array of degradations and composites to mimic real-world distortions, which greatly expands the pretraining data volume. Different from them, we devise a pretraining framework based on MIM to learn effective quality- and aesthetics-related representations.



(a) **Semantics-aware**: perceive semantics from a few pixels

(b) **Distortion-aware**: perceive distortions by reconstruction

Figure 2: Semantics- and distortion-awareness of the pixel-based MIM (a) MIM has the ability to understand the semantics; (b) Pixel-based MIM can reconstruct the distortions applied original images, the left column and the right column are high and low frequency intervals, respectively.

2.2 Masked Image Modeling

Masked modeling learns representation by reconstructing a masked portion of the input. Driven by the success of BERT [12] in NLP, MIM has become an representative SSL method in computer vision [2, 24, 61, 84, 109]. As a pioneer work, BEiT [2] proposes to reconstruct the features of DALL-E [64]. MAE [24] directly reconstructs raw pixels of the masked areas, which greatly simplifies the whole pretraining pipeline. Some studies prove that pixel-based MIM is biased towards reconstructing low-level details, thus hindering the performance on high-level tasks [2, 51, 60]. As a result, following works introduce more complicated reconstruction target rather than using raw pixels [16, 33, 87, 88]. While previous MIM studies mainly focus on high-level tasks, in this paper, we make the first attempt to adapt MIM to visual scoring.

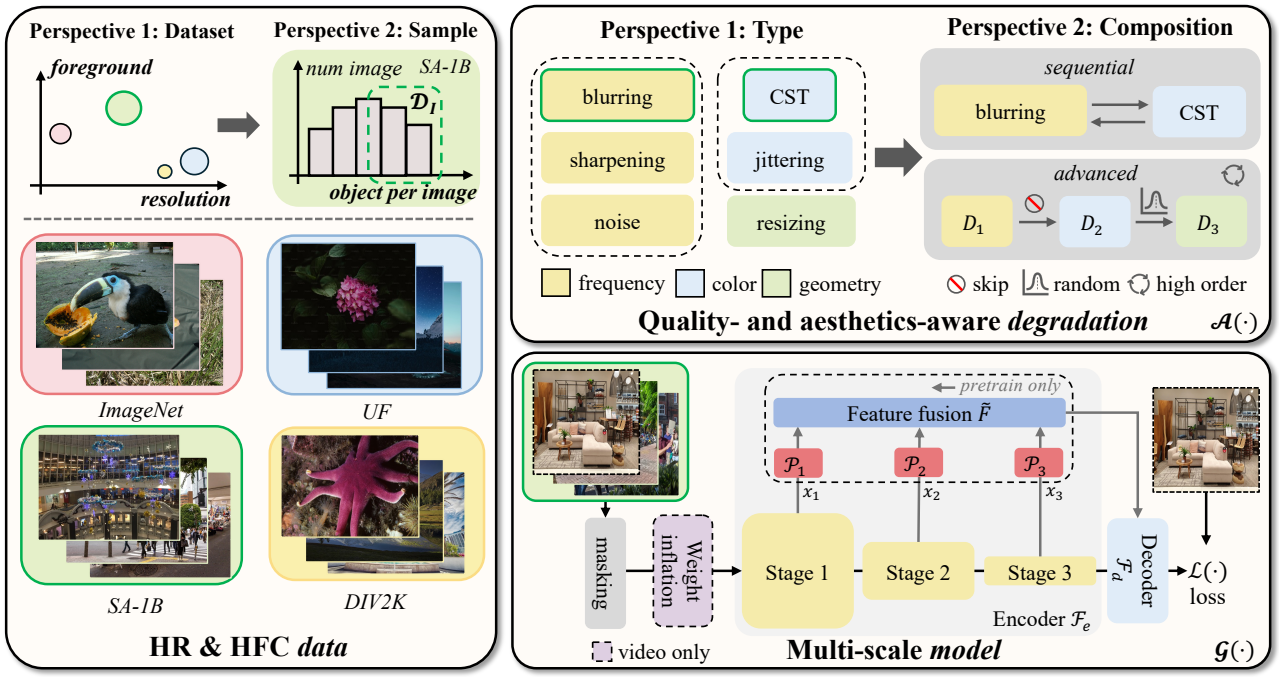


Figure 3: Overview of our proposed QPT V2. QPT V2 incorporates three improvements based on pixel-based MIM tailored for VS. To curate HR & HFC training data, we examine the resolution and foreground coverage of various datasets and samples. To determine quality- and aesthetics-aware degradation, we explore the degradation type and composition. To perceive distortion and aesthetics information in multi-scale fashion, we design a pretrain only feature fusion module based on hierarchical encoder.

3 METHODOLOGY

We first revisit MIM concisely in Sec.3.1, and then describe the motivation of QPT V2 in Sec.3.2. Last, the key designs incorporated in QPT V2 are elucidated.

3.1 A Revisit of Masked Image Modeling

There are three major steps in MIM: (1) split the image into visible and masked patches, (2) reconstruct the masked patches and (3) calculate the reconstruction loss.

Given the original image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, where H, W are the height and width of the image. **First**, specific degradations $\mathcal{A}(\cdot)$ (*e.g.* , resizing) are applied to the image, generating non-overlapping visible patches $\mathbf{I}_v$ and masked patches $\mathbf{I}_m$ with masking $\mathcal{M}$:

$$\begin{aligned} \mathbf{I}_v &= (1-\mathcal{M}) \odot \mathcal{A}(\mathbf{I}) \\ \mathbf{I}_m &= \mathcal{M} \odot \mathcal{A}(\mathbf{I}) \end{aligned} \tag{1}$$

Second, only the visible patches $\mathbf{I}_v$ are fed into the autoencoder $\mathcal{G}(\cdot)$ to reconstruct the masked patches $\hat{\mathbf{I}}_m$ as:

$$\hat{\mathbf{I}}_m = \mathcal{G}(\mathbf{I}_v, \mathbf{e}_{[\mathcal{M}]}) \tag{2}$$

The autoencoder $\mathcal{G}(\cdot)$ consists of an encoder $\mathcal{F}_e$ and a decoder $\mathcal{F}_d$, both are stacked Transformer blocks. Here, a shared learnable mask token $\mathbf{e}_{[\mathcal{M}]}$ functions as the placeholder of masked patches, which are combined with the encoder's output and fed into the decoder. **Last**, an MSE loss $\mathcal{L}(\cdot)$ is computed at masked positions for self-supervision as $\mathcal{L} = \|\mathbf{I}_m - \hat{\mathbf{I}}_m\|_2^2$.

3.2 Motivation

To accurately score the quality and aesthetics of visual content, a broad range of VS-related factors necessitate examination, namely *high-level* attributes (*e.g.* , semantics, composition *etc.*) and *low-level* distortions (*e.g.* , blur, noise *etc.*). By analysing the insightful features of MIM, we believe the pretrained models have the potential to be both **quality-aware** and **aesthetics-aware**, described next.

First, it has been proved that MIM has the ability to comprehend the *high-level semantics* of the image [2, 24]. During pretraining, the large masking ratio forces the model to reconstruct the masked area provided with a few visible patches. Depicted by Fig. 2 (a), the pretrained model gives semantically plausible reconstruction even when 90% of the pixels are masked. **Second**, MIM is proved to be biased towards *low-level details* when reconstructing [51, 52] the *raw pixels*. Due to the perfect reconstruction of pixel values, the model focuses on intricate details (*e.g.* , texture with repeated patterns) besides understanding the content, allowing for a better perception of low-level distortion. To better illustrate the distortion-awareness of MIM, we separately apply blurring and noise to the same image. Reconstruction results in Fig. 2 (b) show that the pretrained model can perceive distortions in low and high frequency intervals, respectively.

Despite MIM has the potential to encompass VS-related factors comprehensively, unleashing its power on downstream VS tasks still presents a non-trivial endeavor. We dissect the MIM framework and identify three components that contribute to this gap:

Figure 4: Illustration of the gap in FC between SA-1B and SR datasets after introducing random cropping. Images in SA-1B are more likely to generate HFC crops compared to the images in DIV2K. Please zoom in for a better view.

- **Data**. ImageNet [11] has become the de-facto pretraining data in MIM studies. The images generally exhibit low resolution and lack intricate details. Over the years, supporting evidence from psychophysical studies has indicated the richness of details (*e.g.* , spatial complexity) in visual content directly impacts human eye's perception of quality and aesthetics [14, 23, 27, 82]. Thus, pretraining on data lacking details might not be sufficient for model to exploit fine-grained quality and aesthetics information. In all, curating **pretraining data** tailored for VS is of utmost importance.
- **Degradation**. MIM achieves excellent results in high-level tasks with simple degradations (*e.g.* , random cropping) [24, 91]. Thus, previous works pay less attention to the degradation design. Yet, simple degradations can only encompass VS-related factors presented in common scenario (*e.g.* , content editing), overlooking other factors introduced by various visual applications, such as compression, transmission, and unprofessional shooting. Therefore, *degradations that cover extensive VS-related factors need to be considered.*
- **Model**. HVS assesses quality and aesthetics in a *multi-scale* fashion [36]. Additionally, numerous previous works have proved the benefits of utilizing multi-scale features in other vision tasks [29, 49]. As a result, to mimic HVS and capture both fine-grained and coarse-grained VS-related factor effectively, the improvement of the **model structure** is exceedingly crucial.

3.3 Data

Demonstrated in Fig. 3, pretraining data of QPT V2 is curated from two criteria: **high resolution (HR)** and **high foreground coverage (HFC)**. As argued above, by reconstructing the rich textures and local structures within the HR images, models are prone to perceive a broad range of quality and aesthetics information during pretraining. In addition, FC is defined as the proportion of *foreground region* of the entire image. Since foreground region encodes way more semantics and texture than the background, pretraining on HFC images ensures the model's sensitivity to both high-level and low-level visual attributes.

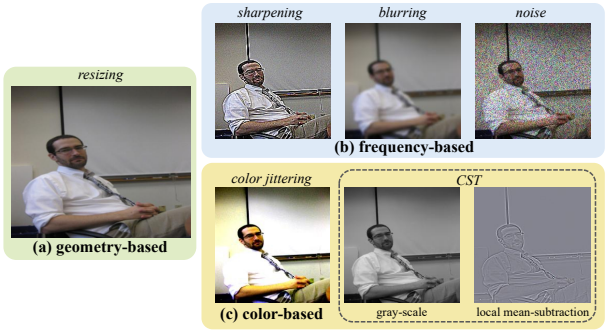


Figure 5: Illustration of the studied degradations, each transforms data stochastically.

Based on the two criteria, multiple datasets with various resolution and FC are investigated. We resort to SA-1B [38] as the *pretraining data source* for the following reasons. **First**, SA-1B has an average resolution of approximately 1600×2100, which is significantly higher than that of ImageNet. **Second**, although widely used datasets (*e.g.* , DIV2K [1], UnsplashFull [54] *etc.*) in super-resolution (SR) task possess higher resolution (>2K), SA-1B exhibits a significantly higher FC. To maintain the resolution of HR images while adapting to the small input size of the model (*e.g.* , 224×224), degradations $\mathcal{A}(\cdot)$ in Equ. 1 typically include random cropping, which further widens the gap between SA-1B and SR datasets in terms of FC. Fig. 4 highlights this difference. **Third**, SA-1B provides a straightforward criterion, namely *the number of objects per image*, which allows us to further filter the dataset to get images with higher FC. Eventually, the pretraining dataset for QPT V2 consists of 1.28 million HR images filtered from SA-1B, with each image containing 50 or more objects. The effectiveness of HR & HFC data on downstream VS tasks is validated, we refer the reader to Sec.4.3 for more details.

3.4 Degradation

To comprehensively cover the VS-related factors, degradation **type** and **composition** are studied. Fig.5 visualizes all the degradations studied in this work. **First**, to account for VS-related factors introduced by *geometric transformation*, resizing is considered. **Second**, to cover the factors introduced by *frequency shift*, blurring, sharpening, and gaussian noise are studied. **Last**, to incorporate the factors introduced by *color changing*, color jittering and color space transformation (CST) are considered. Following [55], we employ four color spaces including RGB, LAB, HSV and grayscale. Fig.6 showcases the completeness of our degradation selection. In terms of degradation composition, two strategies are adpoted. **First**, we compose degradations *sequentially*. **Second**, inspired by recent progress in SR [85, 111], an *advanced* composition including shuffling, skipping and high-order operations is used to obtain complex degradations. Random cropping is applied after all the degradations by default.

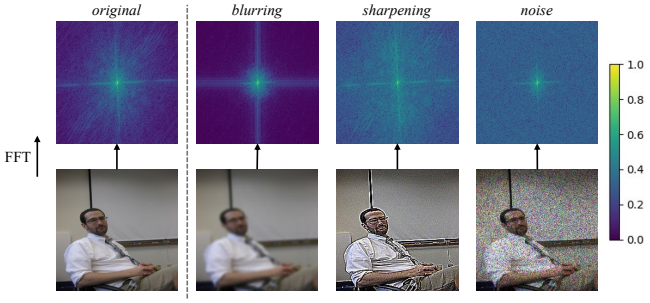


Figure 6: Illustration of the our comprehensive degradation selection. We only consider frequency-based degradations that result in *different* frequency distributions.

Benefiting from the comprehensiveness of our degradation selection, we discover that **CST** stands out as the most quality- and aesthetics-aware degradation. Previous NSS-based VS studies have demonstrated that the VS-related information exists in various color spaces [19, 79] and subsequent studies further proposed that the information of different color spaces are *complementary* to each other [55]. Therefore, we speculate that applying CST to the reconstruction target exposes a richer set of quality and aesthetics factors to the model, improving the data diversity during pretraining. Different from previous pretraining objectives based on contrastive learning, we further reveal the fact that QPT V2 does *not* benefit from the sequential or advanced composition of degradations. More details of both findings can be found in Sec.4.3.

3.5 Model

To perceive the quality and aesthetics information at different scales, *encoder architecture* and *multi-scale feature fusion* are considered. Regarding the selection of the encoder architecture, the common choices are ViT [13] and hierarchical backbones (*e.g.* , Swin [53] and HiViT [109]). Compared to ViT, hierarchical backbones are better at learning multi-scale features by leveraging image-related inductive biases. Thus, a representative hierarchical backbone HiViT is selected as the encoder.

There are three stages of different scales in HiViT. Upon that, we devise a *fusion module* to incorporate the multi-scale features output by different stages. The fusion process is described next. The hierarchical encoder $\mathcal{F}_e$ outputs features at multiple stages during pretraining, shown in Fig. 3. These features are denoted by $X = \{x_i\}_{1\le i\le N}$, where N represents the number of stage. **First**, x_i is processed by a *projection layer* $\mathcal{P}_i(\cdot)$, which aligns the feature space between outputs of different stages, as:

$$\overline{X} = \{\mathcal{P}_i(x_i)\}_{1\le i\le N} \tag{3}$$

Second, the projected features of all stages $\overline{X}$, are integrated by a fusion layer $\widetilde{F}(\cdot)$ as:

$$Y = \widetilde{F}(\overline{X}) \tag{4}$$

Y will be fed into the decoder $\mathcal{F}_d$ for pixel reconstruction. Note that the fusion process is only introduced during pretraining, without affecting the finetuning stage. More details of the architecture selection and feature fusion are in Sec.4.3.

4 EXPERIMENTS

In this section, experimental setups are first introduced in Sec.4.1. By comparing to existing SOTA methods in Sec.4.2, QPT V2 is evaluated on 11 benchmarks from all three VS tasks. Last, an in-depth ablation over QPT V2 is provided in Sec.4.3.

4.1 Evaluation Setups

Criteria. SRCC (Spearman rank correlation coef.) and PLCC (Pearson linear correlation coef.) are adopted as evaluation criteria for all three tasks, both ranging in [0, 1]. A larger SRCC indicates a better ranking between samples, and a larger PLCC shows a more accurate score prediction.

Benchmarks. 11 benchmarks are selected from IQA, VQA, and IAA to comprehensively evaluate the visual scoring ability of QPT V2. For *IQA*, three synthetically degraded datasets (TID2013 [62], LIVE [68], KADID [48]) and three datasets with real-world distortions (KonIQ10K [32], CLIVE [18], FLIVE [98]) are included. For *VQA*, we choose three public NR-VQA datasets, including LIVE-VQC [71], KoNViD-1k [31], and LSVQ [97]. For *IAA*, AVA [58] is selected for evaluation. The key designs of QPT V2 are ablated on FLIVE, LIVE-VQC, and AVA. For all the datasets without official splitting, we randomly split them into 80% for training and 20% for testing. The finetuning/evaluation procedure is conducted on 10 different splittings to avoid randomness, and the average SRCC and PLCC is reported.

Pretraining details. All the experiments are conducted on 4 NVIDIA V100 GPUs. The pretraining data, degradation and model are specified in Sec.3.3, Sec.3.4, and Sec.3.5, respectively. We randomly mask 75% of the pixels following [24] and the input image size is 224×224. The hyperparameter settings are inherited from [24].

Finetuning strategy. For *IQA*, we implement the regression head with a simple MLP (*e.g.* , two linear layers with a GeLU activation in between). Following [74], we resize the shorter edge of images to 340 while keeping the aspect ratio, then randomly crop sub-images with size 224×224. AdamW is adopted for optimization, with weight decay of 0.01. The initial learning rate is 2e-5 and decayed by cosine annealing without warmup. Pretrained models are finetuned for 200 epochs, and the checkpoint of the last epoch is selected for evaluation. When testing, we take the four corners and the center crops and average their predicted quality scores to obtain the final score.

For *VQA*, we follow the settings in [89] for finetuning. Also, the pretraining weight is inflated to adapt video input, as done in [75]. As for hyperparamters, AdamW is used with weight decay of 0.05 and mini-batch size of 16. The initial learning rate is set to 1e-3 and decay it with cosine annealing strategy. The pretrained models are finetuned for 30 epochs on $\text{LSVQ}_{\text{train}}$ followed by evaluating on $\text{LSVQ}_{\text{test}}$, $\text{LSVQ}_{1080\text{p}}$ and two other smaller datasets, LIVE-VQC and KoNViD-1k. We uniformly sample four 32-frame clips from an input video, and average the predicted quality scores as the final results.

For *IAA*, pretrained models are finetuned on AVA_{train} for 60 epochs and then evaluate on AVA_{test}, images are resized to 224×224

Table 1: Performance of existing SOTA methods and the proposed QPT V2 on three synthetic and three real-world IQA datasets. "-" means missing corresponding results in the original paper. The best and second-best results are bolded and underlined.

Method	Synthetic						Real-world					
	LIVE		TID2013		KADID		FLIVE		CLIVE		KonIQ10K	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE [57]	0.907	0.901	0.315	0.393	0.374	0.428	0.211	0.288	0.454	0.468	0.526	0.475
BRISQUE [56]	0.939	0.935	0.604	0.694	0.528	0.567	0.288	0.373	0.601	0.621	0.715	0.702
ILNIQE [104]	0.902	0.906	0.521	0.648	0.503	0.496	0.219	0.256	0.453	0.511	0.503	0.496
CORNIA [96]	0.947	0.950	0.678	0.768	0.516	0.558	-	-	-	-	-	-
HOSA [92]	0.946	0.950	0.735	0.815	0.618	0.653	-	-	-	-	-	-
DB-CNN [106]	0.968	0.971	0.816	0.865	0.851	0.856	0.554	0.652	0.844	0.862	0.878	0.887
HyperIQA [72]	0.962	0.966	0.840	0.858	0.852	0.845	0.535	0.623	0.855	0.871	0.908	0.921
CONRTIQUE [55]	0.960	0.961	0.843	0.857	**0.934**	**0.937**	0.580	0.641	0.854	0.890	0.896	0.901
Re-IQA [66]	0.970	0.971	0.804	0.861	0.872	0.885	0.645	**0.733**	0.840	0.854	0.914	0.923
MUSIQ [36]	-	-		-	-	-	0.566	0.661	-	-	0.916	0.928
TReS [22]	0.969	0.968	0.863	0.883	0.859	0.858	0.554	0.625	0.846	0.877	0.915	0.928
QPT [111]	-	-	-	-	-	-	0.610	0.677	0.895	**0.914**	**0.927**	**0.941**
QPT V2	**0.972**	**0.973**	**0.874**	**0.885**	0.897	0.896	**0.649**	0.684	**0.897**	0.902	0.913	0.930

Table 2: Performance of existing SOTA methods and the proposed QPT V2 on four in-the-wild VQA datasets. "-" means missing corresponding results in the original paper. The best and second-best results are bolded and underlined.

Method	Intra-dataset				Cross-dataset			
	LSVQ_{test}		LSVQ_{1080p}		LIVE-VQC		KoNViD-1k	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [56]	0.579	0.576	0.497	0.531	0.524	0.536	0.646	0.647
TLVQM [39]	0.772	0.774	0.589	0.616	0.670	0.691	0.732	0.724
VIDEVAL [78]	0.794	0.783	0.545	0.554	0.630	0.640	0.751	0.741
VSFA [45]	0.801	0.796	0.675	0.704	0.734	0.772	0.784	0.794
BVQA [44]	0.852	0.854	0.772	0.788	0.816	0.824	0.839	0.830
SimpleVQA [73]	0.867	0.861	0.764	0.803	-	-	0.860	-
$\text{PVQ}_{wo/patch}$ [97]	0.814	0.816	0.686	0.708	0.781	0.781	0.747	0.796
$\text{PVQ}_{w/patch}$ [97]	0.827	0.828	0.711	0.739	0.770	0.807	0.791	0.795
FastVQA [89]	0.876	0.877	0.779	0.814	0.823	0.844	0.859	0.855
Q-Align [90]	0.883	0.882	**0.797**	**0.830**	-	-	0.865	**0.877**
QPT V2	**0.886**	**0.889**	0.785	0.822	**0.827**	**0.853**	**0.866**	0.865

for evaluation. The regression head and hyperparameters are kept consistent with those in IQA.

4.2 Comparison with state-of-the-arts

IQA. We compare our QPT V2 with two groups of IQA methods, including 5 traditional methods and 7 deep learning-based methods. Results in Tab.1 show that QPT V2 achieves superior or comparable performances to current SOTA methods. Previous deep learning-based methods have achieved outstanding performances on three synthetically datasets. *Therefore, further improvements on these datasets can be challenging to attain.* Still, QPT V2 improves the results on LIVE and TID2013 (*e.g.* , **+1.1%** of SRCC on TID2013). Moreover, our method also reaches leading SRCC on FLIVE and CLIVE (+0.4% of SRCC on FLIVE), showcasing its ability to perceive real-world distortions effectively. Besides, Tab.1 includes methods that also harness the power of pretraining by desiging contrastive pretext tasks (*e.g.* , CONRTIQUE, Re-IQA, and QPT). For example, Re-IQA respectively learns a content-aware encoder on ImageNet-1K and a distortion-aware encoder on 758K distorted images. In comparison, QPT V2 consumes *less* pretraining data and achieves better performance.

VQA. We compare QPT V2 to three traditional methods and six deep learning-based methods. Results given in Tab.2 provide the following conclusions. First, the performances we obtain exceed all the traditional methods that rely on hand-crafted features by a large margin, and beat most data-driven methods on four VQA datasets. Second, under the *intra-dataset* setting, QPT V2 pushes the SRCC by 0.3% and PLCC by 0.7% on LSVQ_{test}, exhibiting accurate quality assessment. Third, under the *cross-dataset* setting, we surpass the current SOTAs as well (*e.g.* , with 0.4% and 0.9% gains in SRCC and PLCC on LIVE-VQC), presenting impressive generalization capability.

IAA. We select 11 deep learning-based methods for comparison. Tab.3 indicates that our method significantly surpasses previous SOTA results on AVA dataset, reaching 0.865 (**+4.3%**) of SRCC and

Table 3: Performance of existing SOTA methods and the proposed QPT V2 on AVA dataset. The best and second-best results are bolded and underlined.

Method	AVA_test	
	SRCC	PLCC
NIMA [15]	0.612	0.636
MLSP [30]	0.756	0.757
AFDC [7]	0.649	0.671
MUSIQ [36]	0.726	0.738
MaxViT [76]	0.708	0.745
CLIP-IQA+ [83]	0.619	0.586
Aesthetic Predictor [42]	0.721	0.723
TANet [28]	0.758	0.765
GAT$_{\times 3}$-GATP [21]	0.762	0.764
LIQE [108]	0.776	0.763
VILA [37]	0.774	0.774
Q-Align [90]	0.822	0.817
QPT V2 (60% finetuning data)	0.766	0.780
QPT V2	**0.865**	**0.875**

Table 4: Comparisons of end-to-end finetuning evaluation using different pretext tasks on CLIVE and LIVE-VQC, and AVA.

Pretext task	CLIVE		LIVE-VQC		AVA	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
QPT V2	0.645	0.684	0.827	0.853	0.865	0.875
QPT [111]	0.610	0.677	-	-	-	-
MoCo [25]	0.578	0.629	0.819	0.828	0.707	0.712
Supervised	0.556	0.604	0.810	0.825	0.704	0.690
w/o	0.451	0.475	0.696	0.731	0.545	0.552

0.875 (**+5.8%**) of PLCC. It is worth noting that Q-Align [90] leverages the power of large multi-modality models (LMMs). In comparison, our work introduces a new pretraining paradigm, and exhibits lower computation and smaller model size. The advantages become more evident when comparing to methods *without utilizing LMMs* (*e.g.* , **+8.9%** of SRCC and **+10.1%** of PLCC). The finetuning data amount is further reduced to investigate the power of QPT V2. The results show that QPT V2 achieves parity with some previous SOTA methods (*e.g.* , LIQE , VILA) using only **60%** finetuning data, realizing a more data-efficient transfer. Both LIQE and VILA solve IAA by using auxiliary knowledge in text description. In comparison, QPT V2 achieves SOTA results without the assistance of text modality.

QPT V2 vs. other pretext tasks. We compare QPT V2 with four pretext tasks, including QPT, MoCo, ImageNet-1K supervised and train-from-scratch, on three representative VS benchmarks. Note that both supervised training and train-from-scratch use the same encoder backbone as QPT V2, which is HiViT. MoCo and QPT are based on semantic-aware and quality-aware contrastive learning, respectively. The results in Tab.4 verify the superiority of QPT V2. In addition, QPT V2 also achieves better performances than the supervised training and the one without pretrained weights in all three VS tasks.

4.3 Ablation Studies

Table 5: Ablation on resolution and foreground coverage of pretraining data. IN1K and UF denote ImageNet-1K and UnsplashFull for simplicity.

Source	HR	HFC	FLIVE		LIVE-VQC		AVA	
			SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
IN1K	✗	✓	0.617	0.653	0.812	0.825	0.759	0.780
UF	✓	✗	0.602	0.631	0.799	0.828	0.778	0.801
SA-1B	✓	✓	0.645	0.684	0.827	0.853	0.865	0.875

Table 6: Ablation on single degradation type, each transforms data stochastically.

Deg.	FLIVE		LIVE-VQC		AVA	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
None	0.616	0.664	0.813	0.836	0.832	0.820
Resizing	0.593	0.621	0.797	0.815	0.774	0.752
Blurring	0.628	0.664	0.813	0.833	0.827	0.831
Sharpening	0.617	0.650	0.803	0.820	0.801	0.786
Noise	0.602	0.614	0.793	0.810	0.773	0.747
CST	0.645	0.684	0.827	0.853	0.865	0.875
Color jittering	0.623	0.649	0.809	0.826	0.788	0.792

Table 7: Ablation on different forms of degradation composition. CST and B denote color space transform and blurring for simplicity.

Comp.	Deg.	FLIVE		LIVE-VQC		AVA	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
None	CST	0.645	0.684	0.827	0.853	0.865	0.875
	B	0.628	0.664	0.813	0.833	0.827	0.831
Sequential	B→CST	0.645	0.671	0.806	0.824	0.821	0.840
	CST→B	0.637	0.674	0.815	0.838	0.855	0.874
Advanced	All	0.603	0.652	0.797	0.811	0.820	0.839

Effectiveness of HR & HFC data. We demonstrate the effectiveness of HR & HFC data in QPT V2 by comparing to models pretrained on data with differenet resolution and FC. Following conclusions can be drawn from Tab.5. **First**, pretraining on data with both HR and HFC leads to best downstream performances. **Second**, high resolution matters. When the foreground coverage is generally high, pretraining on HR images yields noticeably superior performance on *all three* representative VS datasets. *e.g.* , +2.9% on FLIVE, +2.2% on LIVE-VQC and +10.1% on AVA. **Last**, HFC is essential. When the resolution is relatively high, HFC data always prevail. *e.g.* , +4.8% on FLIVE, +2.7% on LIVE-VQC , and +8.1% on AVA.

Effectiveness of quality- and aesthetics-aware degradation. Tab.6 displays the downstream performances after applying six different degradations to the reconstructed target in QPT V2. Results obtained without employing any form of degradation serve as the baseline. **First**, the CST degradation incorporated in QPT V2 performs the best, demonstrating its quality- and aesthectics-awareness. **Second**, blurring brings a slight improvement in IAA (+0.5% of SRCC and +1.1% of PLCC). Recent MIM studies find that

removing the high-frequency components of pixels helps the model to focus on semantics, benefiting downstream high-level tasks [51]. Thus, we attribute the gains to the fact that IAA places greater emphasis on high-level visual attributes compared to IQA and VQA [90]. **Last**, the geometry-based resizing, frequency-based sharpening and noise, and color jittering impair the downstream performances on all three benchmarks. This suggests that, altering the spatial layout of the reconstruction target, enriching its high-frequency details, corrupting its frequency spectrum, or perturbing its color are all detrimental to model's quality- and aesthetics-awareness.

To study the effect of degradation composition, two top-performing degradations in Tab.6, namely, CST and blurring are arranged sequentially. Also, all six degradations are included in an advanced composition protocol discussed in Sec.3.4. The results in Tab.7 indicate two findings. **First**, CST and blurring cannot synergize when arranged sequentially, which leads to slight inferior results. **Second**, QPT V2 does not benefit from a complicated degradation space. Above findings are inconsistent with those in contrastive learning, we attribute them to the distinction between two pretraining paradigms.

Effectiveness of multi-scale model. Tab.8 validates the effectiveness of the *encoder architecture selection*. With similar model capacity, hierarchical backbones outperform plain ViT in *all three* VS tasks. For example, Swin-T reaches 0.868 of SRCC on AVA dataset, 4.2% higher than ViT-S. Since two hierarchical models exhibit similar downstream performances, we opt for HiViT-T as the encoder in QPT V2 for better training efficiency. Though increasing the model capacity might potentially yield better results, we did not use larger models out of tradeoffs, which can be done in future work.

To validate the effectiveness of the multi-scale feature fusion strategy proposed in Sec.3.5, we fuse features at different stages, and the downstream results are given in Tab.9. By default, the output of the last stage (stage 3) is always fed to the decoder. Tab.9 indicates that multi-scale feature fusion always provides benefits for VS tasks. Particularly, fusing features from shallow stage (stage 1) yields the most significant gains on three downstream datasets. Due to the inclusion of more low-level details in shallow layer features, we believe that fusing these features assist the model in better perceiving low-level VS-related factors. Additionally, the implementation choices of the projection layer $\mathcal{P}_i(\cdot)$ and the fusion layer $\widetilde{F}(\cdot)$ specified in Sec.3.5 are discussed in Tab.10. Following conclusions can be drawn: **First**, a simple linear layer is sufficient to project representation into the same feature space. A more complex MLP (*e.g.* , Linear-GeLU-Linear structure) cannot bring improvement while introducing non-negligible computational overhead. We think the non-linearity may increase the optimization difficulty as for pretraining. **Second**, weighted average pooling is better suited for integrating projected features compared to the simple summation.

Impact of pretraining data amount . We study the impact of data amount on QPT V2 by using 20%, 50% and 100% percentages of the pretraining data. Given by Tab.11, the performances on three downstream datasets continue to improve as the data amount increases. Surprisingly, we find that even when using only **50%** of the data, QPT V2 still achieves comparable performances to SOTA methods in IAA (*e.g.* , 0.770 of SRCC, 0.774 of PLCC on AVA dataset), demonstrating its strong aesthetics perception capabilities. Furthermore, using different data sources to organize more HR & HFC data remains an open question and could be explored in future work.

Table 8: Ablation on the selection of the encoder architecture. MS denotes to multi-scale for simplicity.

Model	MS	Param	FLIVE		LIVE-VQC		AVA	
			SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
ViT-S	✗	22M	0.614	0.651	0.809	0.835	0.822	0.832
Swin-T	✓	28M	0.647	0.671	0.828	0.850	0.868	0.863
HiViT-T	✓	19M	0.645	0.684	0.827	0.853	0.865	0.875

Table 9: Ablation on the location for feature fusion.

Stage 1	Stage 2	FLIVE SRCC	FLIVE PLCC	LIVE-VQC SRCC	LIVE-VQC PLCC	AVA SRCC	AVA PLCC
✗	✗	0.643	0.650	0.814	0.837	0.848	0.858
✓	✗	0.645	0.684	0.827	0.853	0.865	0.875
✗	✓	0.643	0.671	0.819	0.838	0.842	0.863
✓	✓	0.654	0.672	0.818	0.838	0.854	0.869

Table 10: Ablation on different implementatons of the multi-scale feature fusion. Linear and MLP represents linear and MLP projection layer, while Pool and Sum denote the fusion strategies of weighted-average pooling and summation.

Linear	MLP	Pool	Sum	FLIVE SRCC	FLIVE PLCC	LIVE-VQC SRCC	LIVE-VQC PLCC	AVA SRCC	AVA PLCC
✓		✓		0.645	0.684	0.827	0.853	0.865	0.875
✓			✓	0.638	0.667	0.804	0.821	0.820	0.830
	✓	✓		0.656	0.682	0.820	0.848	0.858	0.864

Table 11: Impact of data amount for the pretext task of QPT V2, using different percentages of the pretraining dataset.

Percentage	FLIVE SRCC	FLIVE PLCC	LIVE-VQC SRCC	LIVE-VQC PLCC	AVA SRCC	AVA PLCC
20%	0.529	0.533	0.772	0.796	0.586	0.602
50%	0.610	0.647	0.805	0.829	0.770	0.774
100%	0.645	0.684	0.827	0.853	0.865	0.875

5 CONCLUSION

We propose QPT V2, a novel MIM-based pretraining paradigm crafted for visual scoring tasks, aiming at alleviating the obstacle of insufficient annotated data. To enhance the quality- and aesthetics-awareness of the pretraining objective, we provide a meticulous analysis over the vanilla MIM framework and make targeted improvements on three key components: pretraining data, degradation, and model structure. After pretraining, QPT V2 achieves SOTA results on 11 downstream benchmarks in IQA, VQA, and IAA, demonstrating impressive capability and generalization ability. In all, we hope this work will inspire the community to reflect and explore the possibility of different pretraining paradigms in the context of visual scoring.

REFERENCES

[1] Eirikur Agustsson and Radu Timofte. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=p-BhZSz59o4

[3] Thomas Barnett, Shruti Jain, Usha Andra, and Taru Khurana. 2018. Cisco visual networking index (vni) complete forecast update, 2017–2022. *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation* (2018), 1–30.

[4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2018. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Trans. Image Process.* 27, 1 (2018), 206–219. https://doi.org/10.1109/TIP.2017.2760518

[5] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *CoRR* abs/2303.04226 (2023). https://doi.org/10.48550/ARXIV.2303.04226 arXiv:2303.04226

[6] Pengfei Chen, Leida Li, Yipo Huang, Fengfeng Tan, and Wenjun Chen. 2019. QoE Evaluation for Live Broadcasting Video. In *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*. IEEE, 454–458. https://doi.org/10.1109/ICIP.2019.8802978

[7] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. 2020. Adaptive Fractional Dilated Convolution Network for Image Aesthetics Assessment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 14102–14111. https://doi.org/10.1109/CVPR42600.2020.01412

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. *CoRR* abs/2003.04297 (2020). arXiv:2003.04297 https://arxiv.org/abs/2003.04297

[9] Alexandre G. Ciancio, André Luiz N. Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. 2011. No-Reference Blur Assessment of Digital Pictures Based on Multifeature Classifiers. *IEEE Trans. Image Process.* 20, 1 (2011), 64–75. https://doi.org/10.1109/TIP.2010.2053549

[10] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 3953)*, Ales Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer, 288–301. https://doi.org/10.1007/11744078_23

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[14] Karen Egiazarian, Jaakko Astola, Nikolay Ponomarenko, Vladimir Lukin, Federica Battisti, and Marco Carli. 2006. New full-reference quality metrics based on HVS. In *Proceedings of the second international workshop on video processing and quality metrics*, Vol. 4. 4.

[15] Hossein Talebi Esfandarani and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Trans. Image Process.* 27, 8 (2018), 3998–4011. https://doi.org/10.1109/TIP.2018.2831899

[16] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 19358–19369. https://doi.org/10.1109/CVPR52729.2023.01855

[17] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual Quality Assessment of Smartphone Photography. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 3674–3683. https://doi.org/10.1109/CVPR42600.2020.00373

[18] Deepti Ghadiyaram and Alan C. Bovik. 2016. Massive Online Crowdsourced Study of Subjective and Objective Picture Quality. *IEEE Trans. Image Process.* 25, 1 (2016), 372–387. https://doi.org/10.1109/TIP.2015.2500021

[19] Deepti Ghadiyaram and Alan C. Bovik. 2016. Perceptual Quality Prediction on Authentically Distorted Images Using a Bag of Features Approach. *CoRR* abs/1609.04757 (2016). arXiv:1609.04757 http://arxiv.org/abs/1609.04757

[20] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. 2017. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (2017), 2061–2077.

[21] Koustav Ghosal and Aljosa Smolic. 2022. Image Aesthetics Assessment Using Graph Attention Network. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*. IEEE, 3160–3167. https://doi.org/10.1109/ICPR56361.2022.9956162

[22] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. 2022. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*. IEEE, 3989–3999. https://doi.org/10.1109/WACV51458.2022.00404

[23] Prateek Gupta, Priyanka Srivastava, Satyam Bhardwaj, and Vikrant Bhateja. 2011. A modified PSNR metric based on HVS for quality assessment of color images. In *2011 International Conference on Communication and Industrial Application*. IEEE, 1–4.

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[27] Lihuo He, Fei Gao, Weilong Hou, and Lei Hao. 2014. Objective image quality assessment: a survey. *Int. J. Comput. Math.* 91, 11 (2014), 2374–2388. https://doi.org/10.1080/00207160.2013.816415

[28] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. 2022. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 942–948. https://doi.org/10.24963/IJCAI.2022/132

[29] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. 2019. Effective aesthetics prediction with multi-level spatially pooled features. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9375–9383.

[30] Vlad Hosu, Bastian Goldlücke, and Dietmar Saupe. 2019. Effective Aesthetics Prediction With Multi-Level Spatially Pooled Features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 9375–9383. https://doi.org/10.1109/CVPR.2019.00960

[31] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *Ninth International Conference on Quality of Multimedia Experience, QoMEX 2017, Erfurt, Germany, May 31 - June 2, 2017*. IEEE, 1–6. https://doi.org/10.1109/QOMEX.2017.7965673

[32] Vlad Hosu, Hanhe Lin, Tamás Szirányi, and Dietmar Saupe. 2020. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Trans. Image Process.* 29 (2020), 4041–4056. https://doi.org/10.1109/TIP.2020.2967829

[33] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. 2022. MILAN: Masked Image Pretraining on Language Assisted Representation. *CoRR* abs/2208.06049 (2022). https://doi.org/10.48550/ARXIV.2208.06049 arXiv:2208.06049

[34] Mariko Isogawa, Dan Mikami, Kosuke Takahashi, Daisuke Iwai, Kosuke Sato, and Hideaki Kimata. 2019. Which is the Better Inpainted Image?Training Data Generation Without Any Manual Operations. *Int. J. Comput. Vis.* 127, 11-12 (2019), 1751–1766. https://doi.org/10.1007/S11263-018-1132-0

[35] Le Kang, Peng Ye, Yi Li, and David S. Doermann. 2014. Convolutional Neural Networks for No-Reference Image Quality Assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 1733–1740. https://doi.org/10.1109/CVPR.2014.224

[36] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17,*

2021. IEEE, 5128–5137. https://doi.org/10.1109/ICCV48922.2021.00510

[37] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. 2023. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 10041–10051. https://doi.org/10.1109/CVPR52729.2023.00968

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment Anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

[39] Jari Korhonen. 2019. Two-Level Approach for No-Reference Consumer Video Quality Assessment. *IEEE Trans. Image Process.* 28, 12 (2019), 5923–5938. https://doi.org/10.1109/TIP.2019.2923051

[40] Lukas Krasula, Yoann Baveye, and Patrick Le Callet. 2020. Training Objective Image and Video Quality Estimators Using Multiple Databases. *IEEE Trans. Multim.* 22, 4 (2020), 961–969. https://doi.org/10.1109/TMM.2019.2935687

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[42] LAION. 2023. *aesthetic-predictor*. https://github.com/LAION-AI/aesthetic-predictor

[43] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. 2022. Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*. IEEE, 1139–1148. https://doi.org/10.1109/CVPRW56347.2022.00123

[44] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. 2022. Blindly Assess Quality of In-the-Wild Videos via Quality-Aware Pre-Training and Motion Perception. *IEEE Trans. Circuits Syst. Video Technol.* 32, 9 (2022), 5944–5958. https://doi.org/10.1109/TCSVT.2022.3164467

[45] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality Assessment of In-the-Wild Videos. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2351–2359. https://doi.org/10.1145/3343031.3351028

[46] Dingquan Li, Tingting Jiang, and Ming Jiang. 2021. Unified Quality Assessment of in-the-Wild Videos with Mixed Datasets Training. *Int. J. Comput. Vis.* 129, 4 (2021), 1238–1257. https://doi.org/10.1007/S11263-020-01408-W

[47] Dingquan Li, Tingting Jiang, Ming Jiang, Vajira Lasantha Thambawita, and Haoliang Wang. 2021. Reproducibility Companion Paper: Norm-in-Norm Loss with Faster Convergence and Better Performance for Image Quality Assessment. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 3615–3618. https://doi.org/10.1145/3474085.3477937

[48] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. KADID-10k: A Large-scale Artificially Distorted IQA Database. In *11th International Conference on Quality of Multimedia Experience QoMEX 2019, Berlin, Germany, June 5-7, 2019*. IEEE, 1–3. https://doi.org/10.1109/QOMEX.2019.8743252

[49] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[50] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. 2017. RankIQA: Learning from Rankings for No-Reference Image Quality Assessment. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 1040–1049. https://doi.org/10.1109/ICCV.2017.118

[51] Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. 2023. PixMIM: Rethinking Pixel Reconstruction in Masked Image Modeling. *CoRR* abs/2303.02416 (2023). https://doi.org/10.48550/ARXIV.2303.02416 arXiv:2303.02416

[52] Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. 2023. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5361–5372.

[53] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

[54] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. 2021. NTIRE 2021 Learning the Super-Resolution Space Challenge. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 596–612. https://doi.org/10.1109/CVPRW53098.2021.00072

[55] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. 2022. Image Quality Assessment Using Contrastive Learning. *IEEE Trans. Image Process.* 31 (2022), 4149–4161. https://doi.org/10.1109/TIP.2022.3181496

[56] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* 21, 12 (2012), 4695–4708. https://doi.org/10.1109/TIP.2012.2214050

[57] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* 20, 3 (2013), 209–212. https://doi.org/10.1109/LSP.2012.2227726

[58] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. IEEE Computer Society, 2408–2415. https://doi.org/10.1109/CVPR.2012.6247954

[59] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. 2016. CVD2014 - A Database for Evaluating No-Reference Video Quality Assessment Algorithms. *IEEE Trans. Image Process.* 25, 7 (2016), 3073–3086. https://doi.org/10.1109/TIP.2016.2562513

[60] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. 2023. What Do Self-Supervised Vision Transformers Learn?. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=azCKuYyS74

[61] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *CoRR* abs/2208.06366 (2022). https://doi.org/10.48550/ARXIV.2208.06366 arXiv:2208.06366

[62] Nikolay N. Ponomarenko, Oleg Ieremeiev, Vladimir V. Lukin, Karen O. Egiazarian, Lina Jin, Jaakko Astola, Benoît Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. 2013. Color image database TID2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing, EUVIP 2013, Paris, France, June 10-12, 2013*. IEEE, 106–111. https://ieeexplore.ieee.org/document/6623960/

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[64] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. http://proceedings.mlr.press/v139/ramesh21a.html

[65] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir D. Bourdev. 2019. Learned Video Compression. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 3453–3462. https://doi.org/10.1109/ICCV.2019.00355

[66] Avinab Saha, Sandeep Mishra, and Alan C. Bovik. 2023. Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 5846–5855. https://doi.org/10.1109/CVPR52729.2023.00566

[67] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

[68] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Trans. Image Process.* 15, 11 (2006), 3440–3451. https://doi.org/10.1109/TIP.2006.881959

[69] Eero P Simoncelli and Bruno A Olshausen. 2001. Natural image statistics and neural representation. *Annual review of neuroscience* 24, 1 (2001), 1193–1216.

[70] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

[71] Zeina Sinno and Alan Conrad Bovik. 2019. Large-Scale Study of Perceptual Video Quality. *IEEE Trans. Image Process.* 28, 2 (2019), 612–627. https://doi.org/10.1109/TIP.2018.2869673

[72] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 3664–3673. https://doi.org/10.1109/CVPR42600.2020.00372

[73] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. 2022. A Deep Learning based No-reference Quality Assessment Model for UGC Videos. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 856–865. https://doi.org/10.1145/3503161.3548329

[74] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. 2023. Blind Quality Assessment for in-the-Wild Images via Hierarchical Feature Fusion and Iterative Mixed Database Training. *IEEE J. Sel. Top. Signal Process.* 17, 6 (2023), 1178–1192. https://doi.org/10.1109/JSTSP.2023.3270621

[75] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html

[76] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan C. Bovik, and Yinxiao Li. 2022. MaxViT: Multi-axis Vision Transformer. In *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV (Lecture Notes in Computer Science, Vol. 13684)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 459–479. https://doi.org/10.1007/978-3-031-20053-3_27

[77] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content. *IEEE Trans. Image Process.* 30 (2021), 4449–4464. https://doi.org/10.1109/TIP.2021.3072221

[78] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content. *IEEE Trans. Image Process.* 30 (2021), 4449–4464. https://doi.org/10.1109/TIP.2021.3072221

[79] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. RAPIQUE: Rapid and Accurate Video Quality Prediction of User Generated Content. *CoRR* abs/2101.10955 (2021). arXiv:2101.10955 https://arxiv.org/abs/2101.10955

[80] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018). arXiv:1807.03748 http://arxiv.org/abs/1807.03748

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[82] Bo Wang, Zhibing Wang, Yupeng Liao, and Xinggang Lin. 2008. HVS-based structural similarity for image quality assessment. In *2008 9th International Conference on Signal Processing*. IEEE, 1194–1197.

[83] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 2555–2563. https://doi.org/10.1609/AAAI.V37I2.25353

[84] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. *CoRR* abs/2208.10442 (2022). https://doi.org/10.48550/ARXIV.2208.10442 arXiv:2208.10442

[85] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*. IEEE, 1905–1914. https://doi.org/10.1109/ICCVW54120.2021.00217

[86] Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. YouTube UGC Dataset for Video Compression Research. In *21st IEEE International Workshop on Multimedia Signal Processing, MMSP 2019, Kuala Lumpur, Malaysia, September 27-29, 2019*. IEEE, 1–5. https://doi.org/10.1109/MMSP.2019.8901772

[87] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. 2022. Masked Feature Prediction for Self-Supervised Visual Pre-Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 14648–14658. https://doi.org/10.1109/CVPR52688.2022.01426

[88] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. 2022. MVP: Multimodality-Guided Visual Pre-training. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX (Lecture Notes in Computer Science, Vol. 13690)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 337–353. https://doi.org/10.1007/978-3-031-20056-4_20

[89] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. FAST-VQA: Efficient End-to-End Video Quality Assessment with Fragment Sampling. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 13666)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 538–554. https://doi.org/10.1007/978-3-031-20068-7_31

[90] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. 2023. Q-ALIGN: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *arXiv preprint arXiv:2312.17090* (2023).

[91] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. SimMIM: a Simple Framework for Masked Image Modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 9643–9653. https://doi.org/10.1109/CVPR52688.2022.00943

[92] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David S. Doermann. 2016. Blind Image Quality Assessment Based on High Order Statistics Aggregation. *IEEE Trans. Image Process.* 25, 9 (2016), 4444–4457. https://doi.org/10.1109/TIP.2016.2585880

[93] Minrui Xu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shiwen Mao, Zhu Han, Abbas Jamalipour, Dong In Kim, Xuemin Shen, Victor C. M. Leung, and H. Vincent Poor. 2023. Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services. *CoRR* abs/2303.16129 (2023). https://doi.org/10.48550/ARXIV.2303.16129 arXiv:2303.16129

[94] Dan Yang, Veli-Tapani Peltoketo, and Joni-Kristian Kämäräinen. 2019. CNN-Based Cross-Dataset No-Reference Image Quality Assessment. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE, 3913–3921. https://doi.org/10.1109/ICCVW.2019.00485

[95] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*. IEEE, 1190–1199. https://doi.org/10.1109/CVPRW56347.2022.00126

[96] Peng Ye, Jayant Kumar, Le Kang, and David S. Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. IEEE Computer Society, 1098–1105. https://doi.org/10.1109/CVPR.2012.6247789

[97] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan C. Bovik. 2021. Patch-VQ: 'Patching Up' the Video Quality Problem. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 14019–14029. https://doi.org/10.1109/CVPR46437.2021.01380

[98] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. 2020. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 3572–3582. https://doi.org/10.1109/CVPR42600.2020.00363

[99] Junyong You. 2021. Long short-term convolutional transformer for no-reference video quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2112–2120.

[100] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=Ee277P3AYC

[101] Xiangxu Yu, Zhengzhong Tu, Zhenqiang Ying, Alan C. Bovik, Neil Birkbeck, Yilin Wang, and Balu Adsumilli. 2022. Subjective Quality Assessment of User-Generated Content Gaming Videos. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*. IEEE, 74–83. https://doi.org/10.1109/WACVW54805.2022.00013

[102] Jiquan Yuan, Xinyan Cao, Changjin Li, Fanyi Yang, Jinlong Lin, and Xixin Cao. 2023. PKU-I2IQA: An Image-to-Image Quality Assessment Database for

AI Generated Images. *CoRR* abs/2311.15556 (2023). https://doi.org/10.48550/ARXIV.2311.15556 arXiv:2311.15556

[103] Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Dong Uk Kim, Sung-Ho Bae, Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon, and Choong Seon Hong. 2023. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? *CoRR* abs/2303.11717 (2023). https://doi.org/10.48550/ARXIV.2303.11717 arXiv:2303.11717

[104] Lin Zhang, Lei Zhang, and Alan C. Bovik. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Trans. Image Process.* 24, 8 (2015), 2579–2591. https://doi.org/10.1109/TIP.2015.2426416

[105] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. 2019. RankSRGAN: Generative Adversarial Networks With Ranker for Image Super-Resolution. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 3096–3105. https://doi.org/10.1109/ICCV.2019.00319

[106] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. 2020. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE Trans. Circuits Syst. Video Technol.* 30, 1 (2020), 36–47. https://doi.org/10.1109/TCSVT.2018.2886771

[107] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2020. Learning To Blindly Assess Image Quality In The Laboratory And Wild. In *IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020*. IEEE, 111–115. https://doi.org/10.1109/ICIP40778.2020.9191278

[108] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 14071–14081. https://doi.org/10.1109/CVPR52729.2023.01352

[109] Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. 2023. HiViT: A Simpler and More Efficient Design of Hierarchical Vision Transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=3F6I-0-57SC

[110] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. 2023. MD-VQA: Multi-Dimensional Quality Assessment for UGC Live Videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 1746–1755. https://doi.org/10.1109/CVPR52729.2023.00174

[111] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. 2023. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22302–22313.

[112] Kai Zhao, Kun Yuan, Ming Sun, and Xing Wen. 2023. Zoom-VQA: Patches, Frames and Clips Integration for Video Quality Assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 1302–1310. https://doi.org/10.1109/CVPRW59228.2023.00137