Burst Image Quality Assessment: A New Benchmark and Unified Framework for Multiple Downstream Tasks

Xiaoye Liang[1], Lai Jiang[1], Minglang Qiao[1], Yichen Guo[1], Yue Zhang[1*], Xin Deng[1], Shengxi Li[1], Yufan Liu[2], Mai Xu[1]

[1]Beijing University of Aeronautics and Astronautics
[2]Institute of Automation, Chinese Academy of Sciences
{xiaoyeliang,yue_zhang}@buaa.edu.cn

Abstract

In recent years, the development of burst imaging technology has improved the capture and processing capabilities of visual data, enabling a wide range of applications. However, the redundancy in burst images leads to the increased storage and transmission demands, as well as reduced efficiency of downstream tasks. To address this, we propose a new task of Burst Image Quality Assessment (BuIQA), to evaluate the task-driven quality of each frame within a burst sequence, providing reasonable cues for burst image selection. Specifically, we establish the first benchmark dataset for BuIQA, consisting of 7,346 burst sequences with 45,827 images and 191,572 annotated quality scores for multiple downstream scenarios. Inspired by the data analysis, a unified BuIQA framework is proposed to achieve an efficient adaption for BuIQA under diverse downstream scenarios. Specifically, a task-driven prompt generation network is developed with heterogeneous knowledge distillation, to learn the priors of the downstream task. Then, the task-aware quality assessment network is introduced to assess the burst image quality based on the task prompt. Extensive experiments across 10 downstream scenarios demonstrate the impressive BuIQA performance of the proposed approach, outperforming the state-of-the-art. Furthermore, it can achieve 0.33 dB PSNR improvement in the downstream tasks of denoising and super-resolution, by applying our approach to select the high-quality burst frames.

Introduction

In recent years, burst imaging, a technique that rapidly captures multiple high-resolution frames in quick succession, has revolutionized visual data acquisition and processing, enabling unprecedented precision in both subjective and objective scene analysis. For subjective use, it helps capture the crucial shot of fleeting moments, such as sports events and wildlife photography. More importantly, for objective use, burst imaging enhances computational photography tasks, such as denoising and super-resolution, by merging multiple frames into a single high-quality output. However, unlike single images, burst images exhibit significant redundancy, leading to 1) increased storage and bandwidth costs, and 2) reduced efficiency in downstream processing tasks. Therefore, there exists a critical need for automatic key frame selection. For example, in album highlight generation, the most visually appealing and memorable frames are saved to reduce storage. Similarly, using a subset of the frames with minimal blur and noise can further improve the output quality of the downstream objective tasks like super-resolution. Therefore, in this paper, we propose a novel task of Burst Image Quality Assessment (BuIQA), which evaluates the subjective or objective quality of the individual frame within a burst sequence, enabling adaptive frame selection for specific downstream tasks.

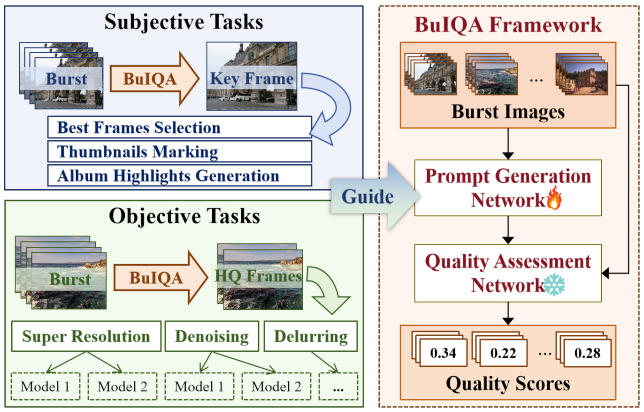

Figure 1: The illustration of the proposed BuIQA task and our unified framework based on task-aware prompt tuning.

Recently, Single Image Quality Assessment (SIQA) has received significant research interest. However, applying SIQA directly to burst images is challenging. First, compared with single image, the difference between burst frames is subtle and even visually imperceptible. Traditional SIQA approaches tailored for independent images are unable to assign the discriminative quality scores for consecutive frames. Second, as shown in Fig.1, burst images serve a wide range of downstream tasks, in which the quality standards vary a lot. For example, aesthetic factors dominate the task of highlight generation, while fidelity and detail preservation are important to denoising. As a result, existing SIQA approaches are constrained by standardized evaluation pipelines, limiting their flexibility and efficacy in addressing the diverse downstream tasks of burst images. In

*Corresponding author

this study, we propose a unified BuIQA framework with a task-aware prompt-tuning strategy, enabling fast adaptation to specific downstream tasks for more reliable quality assessment. Note that Video Quality Assessment (VQA) is also tailored for multiple frames as BuIQA, but it outputs overall video quality without discriminative score for each frame.

In this paper, we establish the first BuIQA dataset for multiple objective and subjective downstream tasks, comprising $7,346$ burst sequences. Inspired by the findings on the dataset, we propose a unified BuIQA framework for multiple downstream tasks, in a task-aware prompt-tuning manner. As shown in Fig.1, the proposed framework consists of a tunable prompt generation network and a quality assessment network. The prompt generation network includes a Task-Driven Prompt Generation (TPG) module, which learns task-specific priors via heterogeneous knowledge distillation. The quality assessment network then leverages these prompts to guide task-aware feature extraction and multi-scale attention, enabling accurate quality evaluation of burst frames. The main contributions are three-fold:

- We propose a novel task of BuIQA with a new benchmark of $7,346$ burst sequences, which contains a total of $45,827$ images and $191,572$ quality score annotations for multiple downstream tasks.
- We propose a new task-aware prompt-tuning approach for BuIQA, showcasing the excellent quality assessment performance and generalization ability over objective and subjective downstream tasks.
- We introduce a prompt generation network with heterogeneous knowledge distillation for learning the task priors, and a task-aware quality assessment network for enabling task-driven BuIQA.

Related Work

Burst Image Processing Burst image processing refers to the technique of merging multiple frames from a burst sequence to generate a single high-quality output, which has been widely adopted in various computer vision tasks. For instance, several studies (Ehret et al. 2019; Liu et al. 2023; Guo et al. 2025) utilize burst images with varying exposures to achieve high dynamic range imaging. Similar approaches (Kokkinos and Lefkimmiatis 2019; Monod, Delon, and Veit 2021) leverage inter-frame context to suppress the noise in the current frame. Recently, an attention-based fusion (Bhat et al. 2021) and Swin Transformer (Luo et al. 2022) architectures are respectively introduced for super-resolving burst images. However, the existing approaches process the whole burst sequence without considering the varying quality of individual frames, which limits efficiency and substantially degrades performance.

Visual Quality Assessment Visual quality assessment is typically categorized into three main approaches: Full Reference (FR) (Kim and Lee 2017; Wu et al. 2023a), Reduced Reference (RR) (Liu et al. 2017; Min et al. 2018), and No Reference (NR) (Lin and Wang 2018; Liu, Van De Weijer, and Bagdanov 2017; Wen et al. 2024b,a). For instance, FRIQA (Wu et al. 2023a) is proposed for FR IQA by integrating low-level and high-level feature fusion. Similarly, a RR IQA model is presented based on the free-energy principle to improve image quality evaluation (Liu et al. 2017). Besides, Hallucinated-IQA (Lin and Wang 2018), a NR IQA method, leverages adversarial learning for quality evaluation. However, the above quality assessment approaches cannot be directly applied to BuIQA.

Prompt-tuning Prompt-tuning is a transfer learning technique that adapts pre-trained models to new tasks by optimizing a small set of parameters, mainly into two categories: explicit (Radford et al. 2019, 2021; Wang et al. 2023) and implicit prompts (Jiang et al. 2024; Zhou et al. 2022b,a; Jia et al. 2022). Explicit prompts rely on observable task-specific inputs, such as text or images. For example, GPT-2 (Radford et al. 2019) performs various natural language processing tasks via textual prompts without task-specific training. Wang *et al.* apply CLIP to assess image aesthetics using natural language supervision (Wang, Chan, and Loy 2023). In contrast to explicit prompts, VPT (Jia et al. 2022) introduces learnable prompt tokens to adapt vision transformers to specific tasks with minimal overhead. For BuIQA, implicit prompts are more reasonable due to the difficulty of defining the explicit evaluation criteria for diverse downstream tasks.

Dataset and Analysis

In this paper, a BuIQA dataset is established for multiple downstream tasks, consisting of two sub-datasets: Burst Image Objective Quality Assessment (BI-OQA) and Burst Image Subjective Quality Assessment (BI-SQA). In total, our BuIQA dataset comprises $7,346$ burst sequences with $45,827$ images and $191,572$ annotated quality scores.

Dataset Establishment

BI-OQA: In BI-OQA, we first collect real-world burst image datasets, BurstSR (Bhat et al. 2021) and HDR+ (Hasinoff et al. 2016) for two denoising and super-resolution, respectively. After that, we apply downstream task models to infer the relative importance of different frames within a burst sequence, which are used as ground-truth quality scores. Specifically, 4 benchmark models for denoising, *i.e.*, HDR21 (Monod, Delon, and Veit 2021), INN (Kokkinos and Lefkimmiatis 2019), DBD (Godard, Matzen, and Uyttendaele 2018) and BPN (Xia et al. 2020), and 4 for super-resolution, *i.e.*, EBSR (Luo et al. 2021), BSRT (Luo et al. 2022), DBSR (Bhat et al. 2021) and BIP (Dudhane et al. 2022a), are adopted as different downstream scenarios, since each benchmark model has its specific evaluation criterion. Moreover, to ensure consistency across tasks, we follow DBSR (Bhat et al. 2021) and further synthesize 1,204 RAW burst sequences (14 frames each) for all the downstream scenarios. Subsequently, frame-level scores are obtained via comparative experiments. Finally, our BI-OQA contains a total of 2,237 burst sequences with 30,543 images and 176,288 annotations.

BI-SQA: The BI-SQA dataset is constructed by collecting and refining burst sequences and subjective quality scores

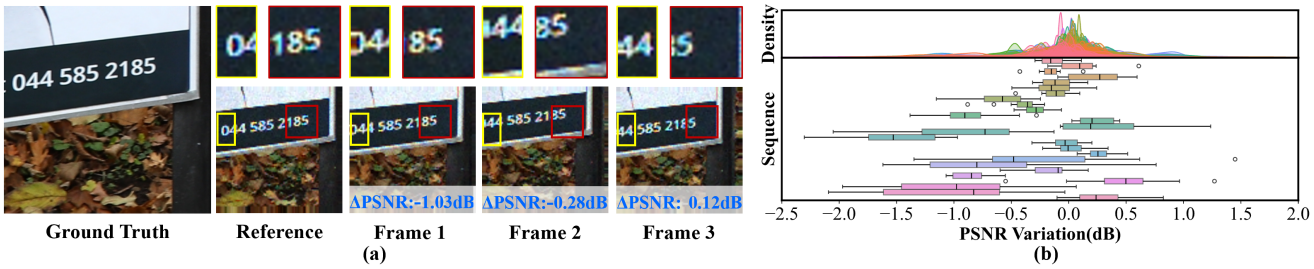

Figure 2: (a) A burst sequence from our BI-OQA dataset. Blue values indicate PSNR variations when the corresponding frame is excluded from the input sequence. (b) The boxplots and density curves of the PSNR variations when removing a frame from the sequence. The boxes and curves in different colors present the results of different sequences, and the black dots are outliers.

from two datasets, *i.e.*, Photo Triage (Chang et al. 2016) and SPAQ (Fang et al. 2020). Specifically, 4,175 burst sequences with 11,314 burst images, as well as corresponding quality scores are extracted from the training and validation sets of the Photo Triage dataset. SPAQ, which contains individually annotated smartphone images, is grouped visually similar samples to form 934 burst sequences with 3,970 images. The original quality scores are normalized to the range [0, 1] for consistency. Finally, the BI-SQA dataset contains a total of 5,109 burst sequences with 15,284 images.

Data Analysis

Findings 1: The importance of different frames in a burst sequence varies significantly for the downstream tasks.

Analysis: We evaluate frame impact by computing PSNR variation when removing each input. Fig. 2(a) presents an example by the EBSR model (Luo et al. 2021). We can observe that frame 3 fails to capture the digit "0" due to positional displacement, and shows a certain level of degradation for the digit "5". Consequently, excluding frame 3 improves performance, while removing frames 1–2 degrades it. Then, we conduct an analysis on the entire BI-OQA dataset. The results are shown in Fig. 2(b), where the upper and lower panels depict the density distribution and the boxplot of PSNR variation. Quantitatively, 40% of the frames are redundant ($\leq$ 0.1 dB change), 25% cause severe drops (up to 4 dB) and should be retained, while discarding 35% low-quality frames yields slight gains.

Findings 2: The ground-truth quality scores of burst frames remain consistent across sequences of different lengths.

Analysis: We investigate whether frame quality scores remain stable under varying sequence lengths. The quality rank of each frame is compared for visualization and analysis. As shown in Fig. 3(a), most frame rankings are unchanged despite changes in input length, with only a few exhibiting minor shifts. Furthermore, we calculate the PLCC values between ranking pairs under different sequence lengths. The PLCC results are over 0.7 for all pairs of our BI-OQA dataset, indicating a strong consistency. The results indicate that the ground-truth quality scores of the burst frames are robust to changes in sequence length, reinforcing the reliability of quality score annotations in our dataset.

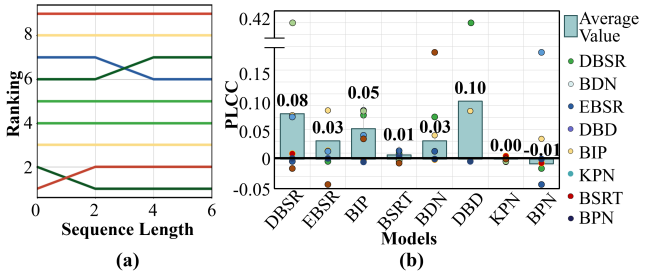

Figure 3: (a) Swimlane diagrams of sequence length and quality ranking, in which different colors represent the specific burst frames. (b) Histogram of PLCC among quality scores generated by different models and downstream tasks.

Findings 3: The objective quality score of burst frame varies significantly across different downstream tasks and models.

Analysis: Based on the BI-OQA dataset, we explore the consistency of objective quality scores across different downstream tasks and model architectures. We first obtain the quality scores for each burst sequence through different models: KPN (Mildenhall et al. 2018), BDN (Dudhane et al. 2022b), DBD (Godard, Matzen, and Uyttendaele 2018) and BPN (Xia et al. 2020) for denoising task; and BIP (Dudhane et al. 2022a), EBSR (Luo et al. 2021), BSRT (Luo et al. 2022), DBSR (Bhat et al. 2021) for super-resolution task. Then, we compute the PLCC values between the quality scores from each pair of models. The results are shown in Fig. 3(b), from which we can observe that the average PLCC values of all pairs are less than 0.1. The above results demonstrate that the objective quality score of a burst frame is highly correlated with downstream tasks and models.

The Proposed Approach

In this section, we propose a unified framework for BuIQA. As illustrated in Fig. 4, our framework consists of a prompt generation network and a quality assessment network. For prompt generation, a TPG module is constructed to generate task-driven prompt **P** based on the input burst sequence **B**. Specifically, for objective downstream tasks, the TPG module learns priors of the tasks via knowledge distillation. Note

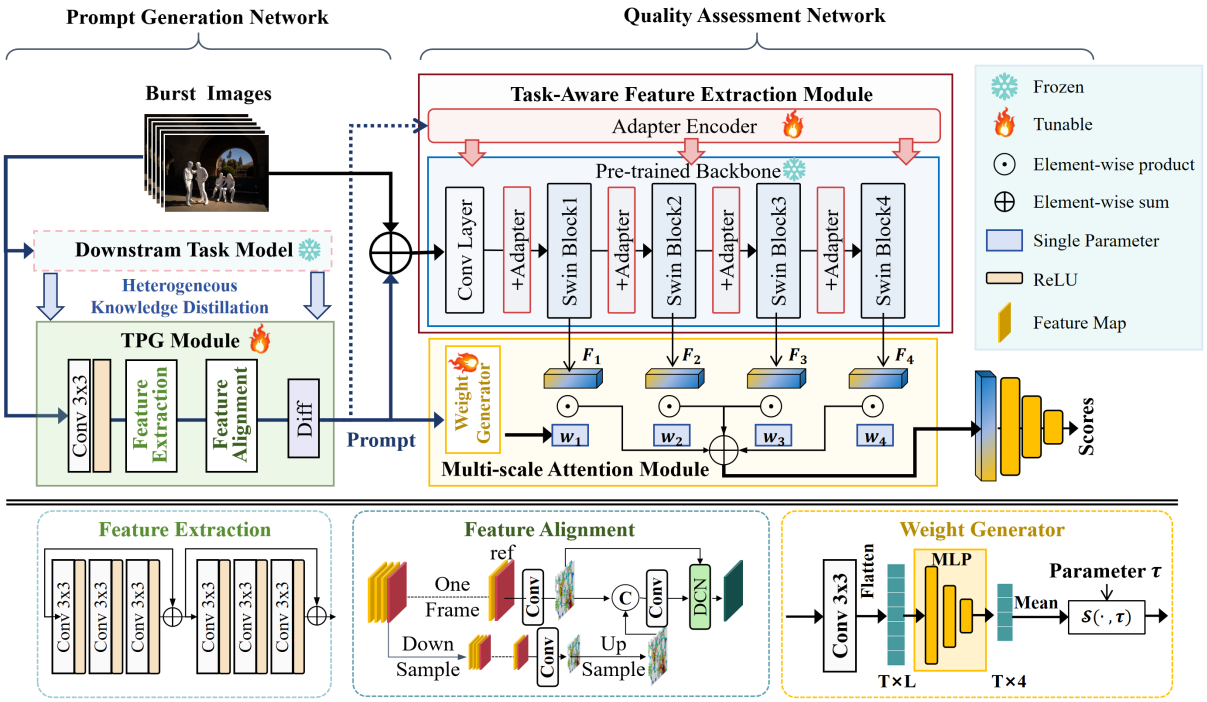

Figure 4: Illustration of our proposed framework.

that for subjective tasks, it learns the subtle difference between frames of burst sequence, without the need of distillation. Given the generated prompt $\mathbf{P}$ and the burst sequence $\mathbf{B}$, the quality assessment network first extracts task-specific burst features through the proposed task-aware feature extraction module, which is composed of a frozen pretrained backbone and a learnable adapter encoder. In this way, we can effectively adapt the knowledge-rich backbone to extract features well aligned with the downstream tasks. Subsequently, an multi-scale attention module is devised to transform the prior information in the prompt into multi-scale attention weights, which are applied to the multi-scale features to produce refined feature representation $\mathbf{F}_{\mathbf{fused}}$. Finally, a multi-layer perceptron (MLP) is employed on $\mathbf{F}_{\mathbf{fused}}$ to predict the task-aware quality scores $\hat{\mathbf{S}}$ for all frames of the input burst sequence.

Prompt Generation Network

We propose the TPG Module to learn prior information of downstream tasks. Given an input burst sequence $\mathbf{B} \in \mathbb{R}^{T\times H\times W\times 3}$, the TPG Module generates a task-driven prompt $\mathbf{P} \in \mathbb{R}^{T\times H\times W\times 3}$, enabling the following quality assessment network to quickly adapt to specific downstream tasks. Specifically, the TPG module first extracts shallow features $\mathbf{Fea}_1$ from the input burst sequence using a convolutional layer followed by a ReLU activation. Then, these features are passed through a deep feature extraction module, which is composed of two residual blocks, to capture high-level features $\mathbf{Fea}_2$:

$$\mathbf{Fea}_2 = E_{FE}(\mathbf{Fea}_1). \tag{1}$$

To further emphasize the subtle differences among burst images, we devise a feature alignment module to align each frame with the feature of the reference frame:

$$\mathbf{Fea}_3 = E_{FA}(Concat(\mathbf{Fea}_2, \mathbf{Fea}_2^{ref}))))). \tag{2}$$

After that, we highlight the differences between each frame and the reference frame through a differential operation, where the aligned feature $\mathbf{Fea}_3$ are subtracted from the reference feature $\mathbf{Fea}_3^{ref}$:

$$\mathbf{P} = \mathbf{Fea}_3 - \mathbf{Fea}_3^{ref}. \tag{3}$$

Here, $\mathbf{P}$ is taken as the output prompt of the TPG module, which embeds the subtle difference among frames of burst sequence. The prompt $\mathbf{P}$ is then used to guide our quality assessment network for predicting quality scores of various downstream tasks.

Heterogeneous Knowledge Distillation. *Finding 3* indicates that the quality scores of different downstream models show weak correlation, even within the same task. Therefore, we propose to leverage knowledge distillation to learn priors from downstream task models. In traditional homogeneous knowledge distillation, the student model usually has similar structure with the teacher model. However, our framework requires the unified student model to learn from diverse downstream task teacher models with different architectures. To solve this, we propose a relational heterogeneous knowledge distillation approach, as illustrated in Fig. 5. Specifically, we employ a similar map $\mathbf{Map}_{\text{sim}}$ to represent the shared components, and a difference map $\mathbf{Map}_{\text{diff}}$ to represent complementary differences between the features

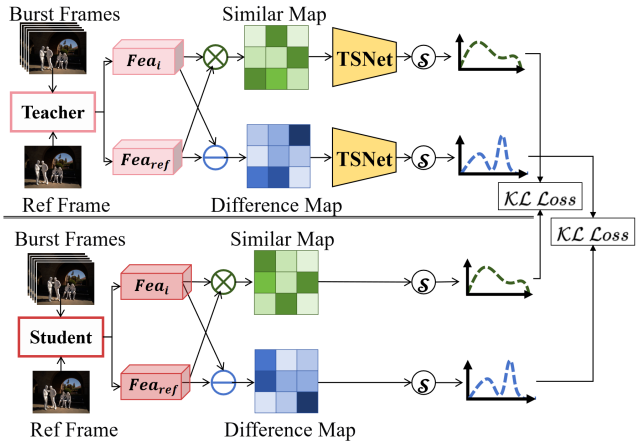

Figure 5: The illustration of our knowledge distillation.

from teacher and student models:

$$\mathbf{Map}_{\text{sim}} = \mathbf{Fea}_{\text{ref}} \odot \mathbf{Fea}_{\text{i}}, \tag{4}$$

$$\mathbf{Map}_{\text{diff}} = \mathbf{Fea}_{\text{ref}} - \mathbf{Fea}_{\text{i}}. \tag{5}$$

Here, $\mathbf{Fea}_i$ denotes the feature of i-th frame. Since the architecture of the teacher model is variable, inspired by Hint. *et al.* (Romero et al. 2014), we map the teacher features to the student feature space using TSNet:

$$\mathbf{Map}_{\mathbf{x}}^{\mathbf{t}} = TSNet(\hat{\mathbf{Map}}_{\mathbf{x}}^{\mathbf{t}}), \mathbf{x} \in \{\text{sim}, \text{diff}\}, \tag{6}$$

where $\mathbf{t}$ indicates the teacher model. Then, all maps are normalized via a softmax function to obtain probability distributions:

$$\mathbf{D}_{\mathbf{x}}^{\mathbf{y}} = Softmax(\mathbf{Map}_{\mathbf{x}}^{\mathbf{y}}), \mathbf{y} \in \{\mathbf{t}, \mathbf{s}\}. \tag{7}$$

Here, $\mathbf{s}$ indicates the student model. Finally, we adopt the Kullback-Leibler (KL) divergence loss function to supervise the knowledge distillation process:

$$\mathcal{L}_{\text{Dist}} = \sum_{\mathbf{x} \in \{\mathbf{sim}, \mathbf{diff}\}} \mathbf{D}_{\mathbf{x}}^{\mathbf{t}} \cdot \log\left(\frac{\mathbf{D}_{\mathbf{x}}^{\mathbf{t}}}{\mathbf{D}_{\mathbf{x}}^{\mathbf{s}}}\right) + (1 - \mathbf{D}_{\mathbf{x}}^{\mathbf{t}}) \cdot \log\left(\frac{1 - \mathbf{D}_{\mathbf{x}}^{\mathbf{t}}}{1 - \mathbf{D}_{\mathbf{x}}^{\mathbf{s}}}\right). \tag{8}$$

Note that the computation in Eq. (8) is conducted in an element-wise manner.

Quality Assessment Network

The Quality Assessment Network consists of two key components: a task-aware feature extraction module and a multi-scale attention module. The former consists of a pre-trained backbone and an adapter encoder for task-aware feature extraction, and the latter one is devised to generate multi-scale attention weights for feature refinement. More details of the two modules are discussed as follows.

Task-aware feature extraction module. We adopt Swin Transformer as the pre-trained backbone. To enhance its task-aware adaptability, we introduce an adapter encoder to transfer the backbone for specific tasks. Specifically, the adapter encoder learns task-specific representations from the prior information in the prompt $\mathbf{P}$, and its outputs are combined with the multi-scale features of the backbone and fed into the subsequent layer:

$$\mathbf{F}_i = Block_i(\mathbf{F}_{i-1} + Adapter_i(\mathbf{P})), \tag{9}$$

where i represents $i-th$ block in the pre-trained backbone. This process yields four-level of features $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ at different scales, capturing both low-level details (e.g., noise and blur) and high-level semantics (e.g., composition and tone).

Multi-scale attention module. The required features for BuIQA vary at different tasks: assessments for subjective tasks focus on both low-level (e.g., noise, blur) and high-level attributes (e.g., composition, color), while assessment for objective tasks rely more on pixel-level information due to minimal high-level variations across frames. In this paper, we infer these differences based on the task-driven prompt. Specifically, we design a multi-scale attention module, which takes the prompt as input and generates attention weights to dynamically fuses the multi-scale features. Specifically, the module generates 4 attention weight values, which are normalized via softmax to produce attention weights:

$$\mathbf{w} = Softmax(E_{wg}(\mathbf{P})). \tag{10}$$

The weighted features are then fused to generate refined representation:

$$\mathbf{F}_{\mathbf{fused}} = \sum_{i=0}^{4} \mathbf{w}_i \cdot \mathbf{F}_i. \tag{11}$$

Finally, an MLP is applied on the refined representation to predict the final quality score $\hat{\mathbf{S}}$ of the sequence:

$$\hat{\mathbf{S}} = MLP(\mathbf{F}_{\mathbf{fused}}), \tag{12}$$

where $\hat{\mathbf{S}} = \{\hat{S}_1, \hat{S}_2, ..., \hat{S}_n\}$, n is the length of the sequence.

Margin Loss function. To further emphasize differences between burst frames, we employ a margin loss (Wang et al. 2019) function to train our quality assessment network. Specifically, the margin loss function is defined as:

$$\mathcal{L}_{\text{pair}}(i, j) = ReLU((S_i - S_j) - (\hat{S}_i - \hat{S}_j)), \tag{13}$$

where i and j denote the indices of images from different groups, respectively. S_i represents the ground truth score, $\hat{S}_i$ are the predicted score. For example, if frame i is of higher quality than frame j (e.g., $S_i - S_j > 0$), the loss penalizes incorrect rankings (e.g., $\hat{S}_i - \hat{S}_j < 0$) and underestimated differences (e.g., $0 < \hat{S}_i - \hat{S}_j < S_i - S_j$). *Finding 2* reveals that only a small number of frames undergo ranking changes when using different sequence lengths, and these changes are limited to a small range.

To address these minor variations, we introduce a "grouping rank" approach. To be specific, frames with similar scores are grouped together, and their rankings are treated as equivalent, while frames without ranking changes retain their original ranks. This grouping approach ensures that slight score fluctuations do not disturb the overall ranking structure, thereby enhancing the robustness and reliability

Metrics	Approach	Denoising				Super-resolution			
		HDR21	INN	DBD	BPN	EBSR	BSRT	DBSR	BIP
R0	Baseline	0.266	0.139	0.352	0.104	0.087	0.077	0.051	0.074
	PAU	0.397	0.139	0.601	0.220	0.143	0.028	0.041	0.010
	SPAQ	0.276	0.177	0.175	0.108	0.087	0.035	0.035	0.020
	PAUQA	0.369	0.196	0.317	0.038	0.111	0.023	0.034	0.061
	ELTA	0.197	0.041	0.079	0.149	0.017	0.015	0.021	0.030
	ESFD	0.206	0.406	0.098	0.137	0.002	0.004	0.003	0.005
	FasterVQA	0.247	0.141	0.060	0.218	0.061	0.043	0.039	0.033
	KVQ	0.421	0.098	0.352	0.026	0.053	0.007	0.034	0.018
	Ours	**0.810**	**0.447**	**0.768**	**0.367**	**0.486**	**0.269**	**0.282**	**0.292**
R0.02	Baseline	0.230	0.046	0.404	0.107	0.240	0.124	0.150	0.236
	PAU	0.231	0.074	0.602	0.257	0.313	0.157	0.154	0.197
	SPAQ	0.508	0.150	0.233	0.186	0.274	0.098	0.132	0.228
	PAUQA	0.391	0.244	0.333	0.202	0.259	0.096	0.137	0.163
	ELTA	0.260	0.136	0.091	0.277	0.201	0.110	0.107	0.245
	ESFD	0.284	0.438	0.100	0.234	0.211	0.083	0.108	0.034
	FasterVQA	0.261	0.152	0.063	0.338	0.239	0.124	0.147	0.170
	KVQ	0.565	0.100	0.562	0.099	0.227	0.105	0.170	0.202
	Ours	**0.862**	**0.487**	**0.756**	**0.427**	**0.553**	**0.315**	**0.363**	**0.350**
R0.05	Baseline	0.283	-0.003	0.483	0.137	0.389	0.154	0.262	0.298
	PAU	0.532	0.090	0.611	0.243	0.445	0.201	0.270	0.288
	SPAQ	0.691	0.167	0.288	0.188	0.423	0.151	0.242	0.271
	PAUQA	0.372	0.264	0.356	0.235	0.426	0.148	0.271	0.213
	ELTA	0.321	0.117	0.097	0.331	0.408	0.138	0.211	0.253
	ESFD	0.310	0.441	0.082	0.290	0.366	0.128	0.225	0.230
	FasterVQA	0.252	0.141	0.090	0.353	0.407	0.169	0.277	0.270
	KVQ	0.645	0.161	0.606	0.250	0.375	0.159	0.271	0.123
	Ours	**0.934**	**0.497**	**0.743**	**0.435**	**0.658**	**0.337**	**0.390**	**0.489**
R0.1	Baseline	0.318	0.036	0.515	0.165	0.510	0.219	0.278	0.275
	PAU	0.656	0.130	0.542	0.321	0.528	0.253	0.327	0.244
	SPAQ	0.795	0.178	0.230	0.161	0.561	0.210	0.302	0.241
	PAUQA	0.436	0.256	0.366	0.337	0.529	0.186	0.334	0.203
	ELTA	0.337	0.083	0.148	0.334	0.525	0.214	0.268	0.226
	ESFD	0.417	0.473	0.090	0.295	0.525	0.154	0.285	0.005
	FasterVQA	0.303	0.141	0.111	0.436	0.529	0.232	0.324	0.222
	KVQ	0.714	0.197	0.578	0.376	0.514	0.199	0.329	0.177
	Ours	**0.969**	**0.523**	**0.661**	**0.483**	**0.718**	**0.373**	**0.488**	**0.565**

Table 1: The performance of our and compared approaches. The best and second best results are in **bold** and underlined.

of quality assessment. Let $\mathbf{G}$ represents groups, and the final loss is computed as the mean of $\mathcal{L}_{\text{pair}}$ over all image pairs:

$$\mathcal{L}_{\text{Mrg}} = \sum_{(i,j)\in\mathbf{Pair}} \mathcal{L}_{\text{pair}}(i,j)/N_P, \tag{14}$$

where $\mathbf{Pair} = \{(i,j)|i \in \mathbf{G}_k, j \notin \mathbf{G}_k, k \in \mathbb{Z}^+, k \leq N_G\}$. N_G and N_P are the lengths of $\mathbf{G}$ and $\mathbf{Pair}$, respectively. Finally, the overall loss can be formulated as:

$$\mathcal{L}_{\text{fnl}} = \alpha\mathcal{L}_{\text{Dist}} + \beta\mathcal{L}_{\text{Mrg}}, \tag{15}$$

where α and β are hyper-parameters for balancing each individual loss.

Experimental Results

In this section, we conduct experiments to evaluate our method for BuIQA over both subjective and objective downstream tasks. Specifically, both the proposed BI-OQA and BI-SQA datasets are randomly split into training and test sets at a ratio of 4:1. To balance different loss terms, the hyper-parameters α, β are set to $1, 10$, respectively. For training our model, the Adam optimizer is adopted for parameter optimization, and the initial learning rate is set to 1×10^{-3}. Since BuIQA is a newly proposed task without comparative approaches, we first apply VGGNet-16 (Simonyan and Zisserman 2014) as the baseline model, supervised by our margin loss. Moreover, we compare our approach with 7 state-of-the-art quality assessment approaches, including 5 IQA approaches, *i.e.*, PAU (Huang et al. 2022), SPAQ (Fang et al. 2020), PAUQA (Lin, Dong, and Dong 2025), ELTA (Liu et al. 2024) and ESFD (Dong et al. 2025), and 2 VQA approaches, *i.e.*, FasterVQA (Wu et al. 2023b) and KVQ (Qu et al. 2025). Note that all the baseline and compared approaches are trained and evaluated under the same settings as our approach.

Evaluation on Objective Tasks

We evaluate the quality assessment performance for objective tasks, *i.e.*, denoising and super-resolution on the proposed BI-OQA dataset. We leverage the SRCC to mea-

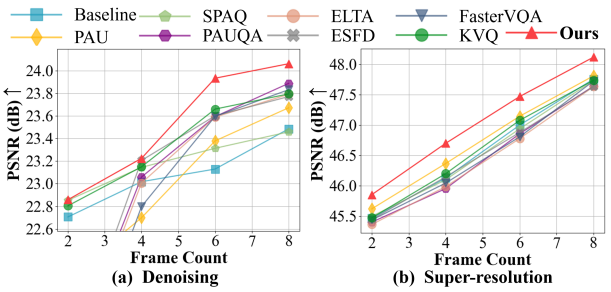

Figure 6: Performance comparison on downstream tasks.

sure the quality assessment performance. Besides, recall that *Finding 2* demonstrates that frames with comparable quality scores contribute almost equally to downstream tasks. In other words, for the evaluation of BuIQA, it is reasonable to treat frames with comparable scores as the same importance level for downstream tasks. Inspired by this, we employ a more flexible evaluation metric to introduce relaxed thresholds: R0, R0.02, R0.05, and R0.1. Taking R0.02 as an example, if the gap between the predicted and ground-truth quality scores is less than 0.02, it would be considered equivalent. The larger threshold indicates a more flexible evaluation metric. Tab. 1 shows the results in BI-OQA, where our approach outperforms all compared approaches in terms of SRCC, under all 2 downstream tasks and 8 annotations by models. Specifically, for EBSR model of SR, our approach promotes SRCC by at least 0.343, 0.240, 0.213 and 0.157 under the relaxed thresholds of R0, R0.02, R0.05, and R0.1, respectively. Similar results can be found for denoising task. The above results validate the effectiveness of our approach on BuIQA.

Besides, we also evaluate the practical utility of our BuIQA approach over the two downstream tasks, *i.e.*, INN (Kokkinos and Lefkimmiatis 2019) for denoising and EBSR (Luo et al. 2021) for super-resolution, via the performance gain in terms of PSNR. Here, we select the burst frames according to the quality scores from our and compared approaches. Specifically, as can be seen in Fig. 6, for frame count M, only the frames with top-M predicted scores are input to the downstream models. It shows that under different frame counts, our approach outperforms almost all compared approaches in both the denoising and super-resolution tasks. The above results validate the practical utility of our approach on downstream tasks.

Evaluation on Subjective Tasks

Here, we evaluate the quality assessment for subjective downstream tasks, *i.e.*, subjective frame selection on BI-SQA dataset. For evaluation, we leverage the widely used pairwise accuracy metric (Chang et al. 2016), which compares the scores of each frame pair for each burst image. As can be seen in Tab. 2, our approach outperforms all compared approaches on both the Photo Traige (Chang et al. 2016) and SPAQ (Fang et al. 2020) datasets. For instance, our approach achieves at least 0.026 and 0.016 accuracy promotion on the two datasets, respectively. The above results validate the effectiveness of our approach on subjective downstream tasks in terms of IQA accuracy and frame selection.

Approach		Photo Traige	SPAQ
IQA	Baseline	0.631±0.127	0.501±0.087
	PAU	0.632±0.074	0.524±0.079
	SPAQ	0.584±0.135	0.566±0.072
	PAUQA	0.639±0.124	0.501±0.103
	ELTA	0.680±0.127	0.510±0.094
	ESFD	0.665±0.128	0.538±0.091
VQA	FasterVQA	0.587±0.142	0.503±0.093
	KVQ	0.533±0.144	0.472±0.112
BuIQA	**Ours**	**0.706±0.121**	**0.582±0.067**

Table 2: Results of pairwise accuracy on BI-SQA dataset

Settings				Metrics
Prompt		Quality assessment		SRCC
TPG	$\mathcal{L}_{\text{Dist}}$	MS features	MS attention	
		✓	✓	0.500
✓		✓	✓	0.643
✓	✓	✓		0.470
✓	✓		✓	0.773
✓	✓	✓	✓	0.818

Table 3: Ablation results of our approach.

Ablation Studies

We perform ablation experiments to assess the effectiveness of our prompt generation network and quality assessment network. For the prompt generation network, we test our model without two main components: (1) TPG module, (2) distillation loss $\mathcal{L}_{\text{Dist}}$. For the quality assessment network, we also test without two components: (1) multi-scale (MS) features in task-aware feature extraction module, (2) MS attention module. In each case, the other network is kept fixed with its complete configuration. As can be seen in Tab. 3, all settings result in performance drops. Notably, the disregard of distillation loss and MS attention degrades SRCC by at least 0.175 and 0.348, respectively. These results confirm the efficacy of each component in our design.

Conclusion

In this paper, we introduced the novel task of BuIQA, which evaluates the task-driven quality of individual frames within burst sequences, enabling effective frame selection for both subjective and objective downstream tasks. We established the first benchmark dataset for BuIQA. Based on comprehensive analysis on our dataset, we uncovered key insights into how frame quality affects downstream performance. Guided by the findings, we proposed a unified BuIQA framework with a task-aware prompt-tuning strategy, integrating a prompt generation network and a quality assessment network for adaptive evaluation. Extensive experiments across downstream tasks demonstrated that BuIQA outperforms existing approaches and significantly boosts denoising and super-resolution performance.

Acknowledgements

This work was supported by NSFC under Grants 62231002,62401027, 62372024, 62206011, 62450131, the Fundamental ResearchFunds for the Central Universities, and the Academic Excellence Foundation of BUAA for PhD Students.

References

Bhat, G.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2021. Deep burst super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9209–9218.

Chang, H.; Yu, F.; Wang, J.; Ashley, D.; and Finkelstein, A. 2016. Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, 35(4): 1–10.

Dong, G.; Liao, X.; Li, M.; Guo, G.; and Ren, C. 2025. Exploring semantic feature discrimination for perceptual image super-resolution and opinion-unaware no-reference image quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28176–28187.

Dudhane, A.; Zamir, S. W.; Khan, S.; Khan, F. S.; and Yang, M.-H. 2022a. Burst Image Restoration and Enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5749–5758.

Dudhane, A.; Zamir, S. W.; Khan, S.; Khan, F. S.; and Yang, M.-H. 2022b. Burst image restoration and enhancement. In *Proceedings of the ieee/cvf Conference on Computer Vision and Pattern Recognition*, 5759–5768.

Ehret, T.; Davy, A.; Arias, P.; and Facciolo, G. 2019. Joint demosaicking and denoising by fine-tuning of bursts of raw images. In *Proceedings of the ieee/cvf international conference on computer vision*, 8868–8877.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3677–3686.

Godard, C.; Matzen, K.; and Uyttendaele, M. 2018. Deep burst denoising. In *Proceedings of the European conference on computer vision (ECCV)*, 538–554.

Guo, Y.; Xu, M.; Jiang, L.; Deng, X.; Zhang, Y.; and Liu, Y. 2025. Compressed Image Super-resolution based on Invertible Degradation and Restoration. *Pattern Recognition*, 112532.

Hasinoff, S. W.; Sharlet, D.; Geiss, R.; Adams, A.; Barron, J. T.; Kainz, F.; Chen, J.; and Levoy, M. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6): 1–12.

Huang, J.; Zhang, L.; Gong, Y.; Zhang, J.; Nie, X.; and Yin, Y. 2022. Series photo selection via multi-view graph learning. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06. IEEE.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European conference on computer vision*, 709–727. Springer.

Jiang, Y.; Yan, X.; Ji, G.-P.; Fu, K.; Sun, M.; Xiong, H.; Fan, D.-P.; and Khan, F. S. 2024. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1): 17.

Kim, J.; and Lee, S. 2017. Deep blind image quality assessment by employing FR-IQA. In *2017 IEEE International Conference on Image Processing (ICIP)*, 3180–3184. IEEE.

Kokkinos, F.; and Lefkimmiatis, S. 2019. Iterative residual cnns for burst photography applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5929–5938.

Lin, B.; Dong, J.; and Dong, X. 2025. Perception-Aware Underwater Image Quality Assessment: Dataset, Perceptual Quality Scores and Assessment Network. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lin, K.-Y.; and Wang, G. 2018. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 732–741.

Liu, L.; He, S.; Ming, A.; Xie, R.; and Ma, H. 2024. ELTA: an enhancer against long-tail for aesthetics-oriented models. In *Forty-first International Conference on Machine Learning*.

Liu, S.; Zhang, X.; Sun, L.; Liang, Z.; Zeng, H.; and Zhang, L. 2023. Joint hdr denoising and fusion: A real-world mobile hdr image dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13966–13975.

Liu, X.; Van De Weijer, J.; and Bagdanov, A. D. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, 1040–1049.

Liu, Y.; Zhai, G.; Gu, K.; Liu, X.; Zhao, D.; and Gao, W. 2017. Reduced-reference image quality assessment in free-energy principle and sparse representation. *IEEE Transactions on Multimedia*, 20(2): 379–391.

Luo, Z.; Li, Y.; Cheng, S.; Yu, L.; Wu, Q.; Wen, Z.; Fan, H.; Sun, J.; and Liu, S. 2022. Bsrt: Improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 998–1008.

Luo, Z.; Yu, L.; Mo, X.; Li, Y.; Jia, L.; Fan, H.; Sun, J.; and Liu, S. 2021. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 471–478.

Mildenhall, B.; Barron, J. T.; Chen, J.; Sharlet, D.; Ng, R.; and Carroll, R. 2018. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2502–2510.

Min, X.; Gu, K.; Zhai, G.; Hu, M.; and Yang, X. 2018. Saliency-induced reduced-reference quality index for natural scene and screen content images. *Signal Processing*, 145: 127–136.

Monod, A.; Delon, J.; and Veit, T. 2021. An analysis and implementation of the hdr+ burst denoising method. *Image Processing On Line*, 11: 142–169.

Qu, Y.; Yuan, K.; Xie, Q.; Sun, M.; Zhou, C.; and Wang, J. 2025. KVQ: Boosting Video Quality Assessment via Saliency-guided Local Perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2150–2160.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wang, B.; Vesdapunt, N.; Sinha, U.; and Zhang, L. 2019. Real-time burst photo selection using a light-head adversarial network. *IEEE Transactions on Image Processing*, 29: 3065–3077.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2555–2563.

Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6830–6839.

Wen, S.; Jiang, L.; Qiao, M.; Xu, M.; Deng, X.; Li, S.; and Duan, Y. 2024a. A Novel Patch Selection Approach for Quality Assessment on Short-Form Videos. In *2024 16th International Conference on Wireless Communications and Signal Processing (WCSP)*, 1361–1367.

Wen, S.; Qiao, M.; Jiang, L.; Xu, M.; Deng, X.; and Li, S. 2024b. MT-VQA: A Multi-task Approach for Quality Assessment of Short-form Videos. In *Proceedings of the 3rd Workshop on Quality of Experience in Visual Multimedia Applications*, 30–38.

Wu, C.; Liao, X.; Yue, H.; Xu, X.; Wei, X.; Wu, D.; and Zhou, M. 2023a. Full-reference image quality assessment via low-level and high-level feature fusion. *International Journal of Pattern Recognition and Artificial Intelligence*, 37(11): 2354016.

Wu, H.; Chen, C.; Liao, L.; Hou, J.; Sun, W.; Yan, Q.; Gu, J.; and Lin, W. 2023b. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15185–15202.

Xia, Z.; Perazzi, F.; Gharbi, M.; Sunkavalli, K.; and Chakrabarti, A. 2020. Basis prediction networks for effective burst denoising with large kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11844–11853.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.