2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

A Dual-Channel Framework for Blind Perceptual Quality Assessment in Bilateral Teleoperation

Zican Wang,[1] Xiao Xu,[1] Zhi Jin,[2] Dong Yang,[1] Eckehard Steinbach[1,3]

***Abstract*— This paper proposes a perceptual no-reference (blind) haptic quality assessment framework for predicting the Quality of Experience (QoE) in teleoperation systems with force feedback. The proposed approach employs a deep neural network that combines semantic and distortion-based channels. The semantic network generates a semantic vector that characterizes the interaction between the robot and its environment. Meanwhile, the distortion network decomposes complex noise introduced by control algorithms and communication artifacts into artificial noise of known types. To train the proposed network, we also construct an augmented dataset for perceptual quality assessment in teleoperation based on the subjective experiments. The dataset augmentation and the model are validated with real-world teleoperation tasks. Our experimental results demonstrate that the performance of our No-Reference (NR) haptic quality assessment model is comparable to or surpasses that of commonly used Full-Reference (FR) methods, achieving Spearman's Rank-Order Correlation scores above 0.85 for QoE prediction.**

I. Introduction

Teleoperation with haptic feedback has been extensively researched and applied in many fields, such as remote surgery and space exploration [1]. By providing force feedback from remote environments, teleoperation systems enhance an operator's perception of a robot's working environment and facilitate the natural transfer of professional skills [2]. However, signal distortions introduced during processing and transmission can significantly degrade the quality of force feedback. As a result, Quality of Experience (QoE), which quantifies the perceptual quality experienced by the operator in human-in-the-loop systems, becomes a critical performance metric for successful teleoperation tasks. Accurately assessing QoE requires extensive subjective experiments using teleoperation systems, such as the one shown in the top section of Fig. 1. However, these experiments are time-consuming and require costly hardware setups. To address this challenge, perceptual Haptic Quality Assessment (HQA) methods are designed to predict the subjective quality of feedback signals [3]. Their results are critical for dynamically adjusting teleoperation systems and optimizing new control methods.

Recent research on Quality Assessment (QA) methods in different domains, such as Image QA (IQA) and Audio QA (AQA), has focused on leveraging Deep Neural Networks (DNN) to predict the perceptual quality of signals without the need for subjective experiments [4]. DNNs, including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), have demonstrated exceptional accuracy and generalization capability in different QA tasks [5]. However, unlike IQA and AQA, HQA remains a relatively new topic driven by the fast development in robotics. As a result, perceptual HQA (pHQA) still lacks large-scale perceptual haptic datasets and dedicated HQA Neural Network (NN) models, which are the two primary focuses of this work.

Large DNN models have demonstrated superior generalization capabilities compared to traditional machine learning methods. However, they also require a larger volume of annotated data to achieve satisfactory performance. In the context of NR-AQA, NR-IQA, and NR-HQA, where subjective experiments are time-consuming to conduct, pseudo-Mean Opinion Score (MOS) serves as a common alternative to MOS with a trade-off between efficiency and accuracy. Pseudo-MOS is an intermediate metric derived from FR methods and can be utilized as ground truth for NR methods. For instance, Peak Signal-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are the commonly adopted pseudo-MOS metrics for NR-IQA [6]. [7] introduces HSSIM as a haptic application-specific SSIM, while [8] evaluates various objective FR metrics and proposes a new metric based on the Video Multi-method Assessment Fusion (VMAF) method [9]. DNN-based pseudo-MOS model can also be obtained by training on a smaller subjectively labeled dataset. The trained model is then utilized to generate pseudo-MOS labels for a larger set of unlabeled samples, which can be collected more efficiently through dataset augmentation or parallel experiments.

Although pseudo-MOS can be regarded as pHQA predictions, accurate and generalized pseudo-MOS methods rely on the existence of reference signals, which introduces an additional challenge. The non-reproducibility of human-in-the-loop interactions prevents us from collecting specific ref-

*This work has been supported in part by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG), and also in part funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/2 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of TUD Dresden University of Technology, and also in part by the joint Czech-Bavarian research project funded by the Bayerisch-Tschechische Hochschulagentur (BTHA/BAYHOST), Germany, under the project no. BTHA-JC-2024-31, and also in part by the Chinese Scholarship Council (CSC) with Funding Number #202106230099, and in part by the Sino-German Mobility Programme M-0421

[1] Authors are with Technical University of Munich, School of Computation, Information and Technology, Department of Computer Engineering, Chair of Media Technology, Munich Institute of Robotics and Machine Intelligence (MIRMI). {zican.wang, xiao.xu, dong.yang, eckehard.steinbach}@tum.de

[2] Zhi Jin is with Shenzhen Campus of Sun Yat-sen University and Sun Yat-Sen University, China. jinzh26@mail.sysu.edu.cn

[3] Eckehard Steinbach is also with the Centre for Tactile Internet with Human-in-the-Loop (CeTI) at TU Dresden.

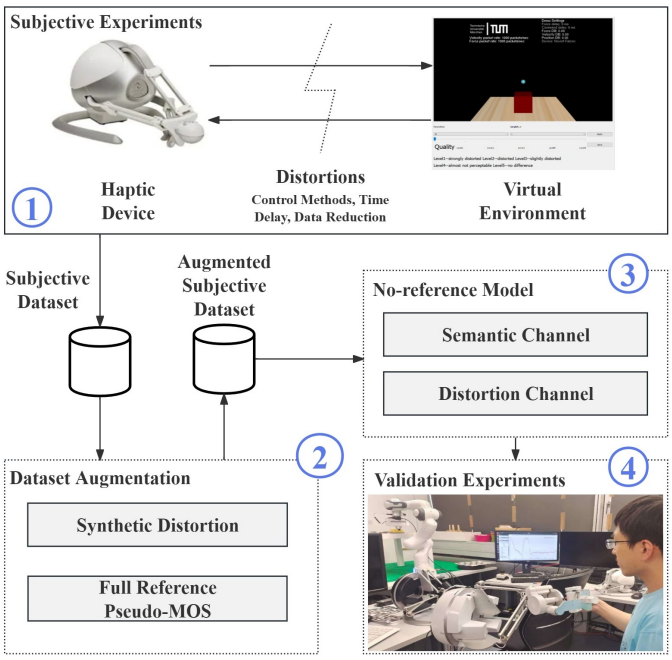

Fig. 1. Overview of the proposed method. The original subjective dataset is collected through a subjective experiment with a virtual environment for reproducibility. Then, the dataset is augmented in a FR manner to obtain pseudo-MOS values. The augmented dataset is used to train a dual-channel network to obtain the NR perceptual quality prediction. The trained model is validated with a robot arm and real-world scenarios for validity.

erence signals during the subjective experiments. Moreover, in certain scenarios like rescue operations and deep-sea exploration, obtaining reference signals from the teleoperation system is challenging due to network limitations or working conditions. Therefore, practical HQA should be designed as NR approaches.

As a step toward overcoming these challenges, this paper proposes an augmented dataset with FR pseudo-MOS and an NR dual-channel HQA model that predicts the perceptual quality of force feedback signals while integrating both distortion and semantic information. Inspired by prior works in IQA and Video QA (VQA), the design of the dual channel is motivated by the observation that distortion information is the primary factor contributing to quality degradation. Whereas in teleoperation, the movements and interactions of the follower robot, represented by semantic information, also influence the perceptual quality of the human operator. The distortion channel identifies the types and levels of signal distortions, whereas the semantic channel interprets the signals into interaction and movement tokens, accounting for complex human factors, such as the free-energy principle [10] and perceptual significance [11].

The proposed dataset is augmented from our existing subjective dataset. Artificial distortions commonly encountered in teleoperation, such as artifacts from haptic data compression, communication delay, and packet loss, are introduced during augmentation for the distortion analysis channel.

Our results reveal that jointly leveraging semantic and distortion features improves the performance of the NR-HQA model, achieving results comparable to or surpassing common FR methods. To further assess the effectiveness of our approach, we designed validation experiments in realistic scenarios. Different types of haptic devices and robotic arms were employed to demonstrate the robustness and practical applicability of the proposed method. The main contributions of this work can be summarized as follows:

- A novel large-scale augmented subjective teleoperation dataset, specifically designed for distortion analysis, is proposed for NR pHQA based on our existing subjective dataset.
- We introduce a novel NR dual-channel network that utilizes distortion and semantic information in teleoperation feedback signals. Its effectiveness is assessed through both quantitative analysis and validation experiments in generalized real-world teleoperation scenarios.

II. Related Work

Many practical and effective QA methods have been proposed within the scope of AQA, IQA, and VQA. AQA resembles HQA as they both focus on predicting the QoE of time series [12]. A typical NR approach based on FR pretraining is distortion analysis, which decomposes authentic noise into synthetic components. Synthetic distortions are easy to reproduce, and their levels are controllable. By analyzing the relationship between authentic and synthetic noise, researchers can augment subjective datasets and generate pseudo-MOS for training [13]. [14] proposes a network with two sequential sub-networks, where one network analyzes the distortion and the other generates the QoE results. Although distortion analysis is widely applied in the context of IQA, it remains underexplored in HQA, which motivates us to incorporate it into our work.

Compared with FR methods, Blind Quality Assessment (BQA) methods (also known as NR methods) lack the crucial reference information needed for quality assessment and distortion analysis. To address this problem, additional sources of information are introduced to support the QA without reference. For example, [15] proposed a Blind VQA (BVQA) system based on Human Visual System (HVS) for user generated video content. Furthermore, psychological effects are also employed to enhance the performance of BQA [16]. These methods evaluate the overall quality of the signals rather than the difference between the reference and distorted signals. However, the complexity of accurately modeling human physiological and psychological systems imposes limitations on the performance of BQA methods.

Another source of the additional information for BQA is semantic information within the distorted signals. With the development of Deep Learning (DL) methods, NR semantic analysis for different signal types has become increasingly effective and practical. For instance, [17] proposes a blind IQA method that integrates both semantic information and HVS. These studies inspired us to develop a semantic subnetwork for NR-HQA.

Perceptual quality assessment for teleoperation is a relatively new topic in robotics. Research in this area primarily focuses on various FR methods for haptic feedback across different working scenarios. In [18], the authors analyze the cognitive performance for teleoperation in industrial manufacturing scenarios based on different interactive interfaces and requirements. Our previous work [8] introduced an Multi-method Assessment Fusion (MAF)-based method that combines multiple FR objective metrics, such as PSNR and SSIM, to predict QoE for teleoperation force feedback. Although this approach enhances HQA model performance compared to single objective metrics, it still relies on reference signals, which may not be available in certain teleoperation tasks. Building upon this foundation, we propose a novel NR approach utilizing dual-channel networks.

III. Methodology

As shown in Fig.1, our proposed framework can be broadly divided into four components, namely subjective experiments, dataset augmentation, Blind HQA (BHQA) model training, and validation experiments. The subjective experiments generate the original QoE dataset for bilateral teleoperation. Dataset augmentation expands the original dataset with artificial distortions in an FR manner. The distortion channel and the semantic channel are then trained separately on the augmented dataset, and the Multi-Layer Perception (MLP) component of the BHQA model is trained while keeping the parameters of the semantic and distortion channels frozen. Finally, validation experiments are conducted to assess and analyze the results. The remainder of this section is also organized according to this framework.

Our HQA problem can be formulated as a function $\hat{F}: \{\mathbf{s}_L\}_{K_t} \mapsto \{\hat{m}\}_{K_t}$, where $\mathbf{s}_L$ denotes signals with a length of L. K_t denotes the size of the test dataset, and $\{\hat{m}\}_{K_t}$ represents the predicted MOS values. Our objective is to maximize evaluation metrics such as the Spearman's Rank-order Correlation $SROCC(\{m\}_{K_n}, \{\hat{m}\}_{K_n})$.

A. Subjective Dataset Establishment

To create a subjectively evaluated dataset for haptic feedback signals, we generate distorted force feedback signals using the open-source haptic codec reference software developed by the IEEE 1918.1.1 working group [19]. This dataset serves as the ground truth for establishing an objective quality assessment metric. Specifically, fifteen haptic signals are recorded as reference signals, and three control methods, namely Time Domain Passivity Control (TDPA) [20], energy-based TDPA [21], and power-based TDPA [22], are chosen as the control methods for teleoperation. Additionally, time delay, deadband compression [23], and packet loss are selected as the primary network artifacts.

A total of 2715 feedback signals are collected with different combinations of distortions and control methods. Conducting subjective experiments for all the signals is overwhelmingly time-consuming, so 240 signals are evenly sampled according to the control method and distortion level. We invited 20 subjects and conducted 4800 subjective evaluations across 240 distorted force feedback signals in total. After the experiments, outlier detection and SROCC tests are performed, and all subjects' responses passed these tests. The resulting MOS values are treated as the ground truth for our subjective dataset.

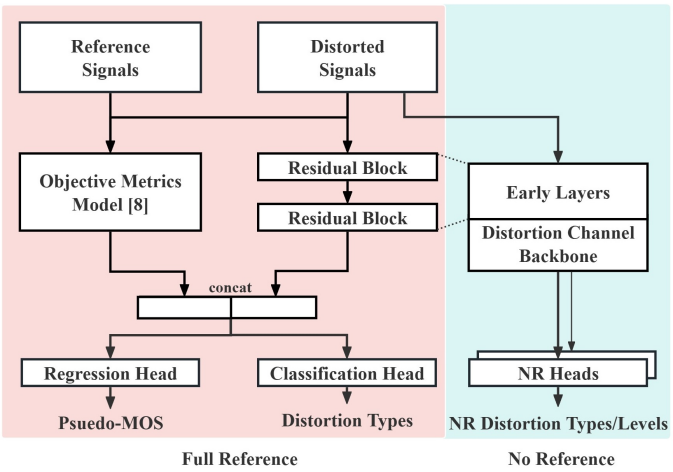

Fig. 2. The ResNet–MLP [24] architecture is employed in both the pseudo-MOS generation and the NR distortion prediction phases. The red part represents the FR networks and models used to generate pseudo-MOS for unannotated augmented samples. The residual blocks from this module are transferred to the early layers of the ResNet backbone in the NR distortion channel. The ResNet backbone together with the MLP heads are trained in a NR manner and serve as the distortion channel in Fig. 3.

The subjective experiments are conducted with a haptic device and a virtual environment. Participants were instructed to follow the recorded motion displayed in the interaction animation on the screen while perceiving the distorted feedback signals. This experimental design offers high reproducibility, efficiency, and controllability. In contrast to real robot teleoperation scenarios, which are employed in the validation experiments, this setup enables the collection of multiple subjective scores from different participants for a fixed distorted signal. It also reduces the time required to explain to non-professional participants and reset the experimental setup, which is critical given that thousands of tests are needed.

B. Pseudo-MOS Generation and Dataset Augmentation

To augment the original smaller subjectively annotated dataset, synthetic distortions of different types and levels are introduced into the subjective dataset D_{subj} to generate a synthetic dataset, denoted as $D_{syn} = \{\mathbf{s}_L^{syn}; \overrightarrow{D}_{ist}\}_{K^{dist}}$, where $\overrightarrow{D}_{ist}$ represents the type and level of the distortion, and K^{dist} denotes the size of the dataset. $\mathbf{s}_L^{syn}$ stands for the signals with synthetic distortions. Distortion types that are common in teleoperation applications are selected, and the parameters of different distortion types, such as the loss rate in packet loss distortion, determine the distortion levels.

Next, we train the network as in Fig. 2 to classify distortion types and generate pseudo-MOS in an FR manner. The trained model is denoted as

$$F_{dist}^{T}: \{\mathbf{s}_L^{dist}; \mathbf{f}_N\}_{K^{dist}} \mapsto \{T_{\hat{D}ist}\}_{K^{dist}},$$

$$F_{dist}^{MOS} : \{\mathbf{s}_L^{dist}; \mathbf{f}_N\}_{K^{dist}} \mapsto \{MOS_{\hat{D}ist}\}_{K^{dist}},$$

where $\{T_{\hat{D}ist}\}_{K^{dist}}$ and $\{MOS_{\hat{D}ist}\}_{K^{dist}}$ represent the predicted distortion types and pseudo-MOS from the classification and regression heads. A Cross Entropy (CE) loss function is applied between the predicted and the ground truth distortion types, and Mean Square Error (MSE) is applied for the regression head. The reference signals are used here to generate FR features and signal errors with the distorted signals, as depicted in the red part of Fig. 2. The residual layers extract latent vectors as hidden features. Hand-crafted features, such as MSE and SSIM, are concatenated with the latent vectors, along with compound features obtained from the MAF method [8]. The MLP regression head processes the extended feature vector to learn the pseudo-MOS and the classification head learns the distortion types.

Finally, the model is trained in an NR setting, as illustrated by the blue section in Fig. 2. The early residual blocks retain parameters from the previously trained FR model, while the subsequent layers together with the MLP classification and regression heads are optimized to produce NR type and level predictions. The trained ResNet–MLP model is then incorporated into the distortion channel of the BHQA model, with its parameters frozen to preserve the learned representations.

Additionally, the original dataset is also augmented with a conditional Generative Adversarial Network (GAN) [25], which generates signals similar to those in the dataset given a specified pseudo-MOS. Only distortions with authentic distortions from the original subjective dataset are used to train the conditional GAN.

C. Dual-channel Blind Haptic Quality Assessment Model

As shown in Fig.3, the proposed dual-channel BHQA model comprises a Seq2Seq structure [26] semantic channel, a Resnet-MLP distortion channel, and an integration MLP that aggregates the learned information from both channels to generate the predicted perceptual quality. The dual-channel BHQA model operates in an NR manner, requiring only the distorted signal as input. Since the distortion channel is based on the ResNet structure, which utilizes CNN, the distortion signal must be either trimmed or resampled to maintain a consistent length.

The distortion channel is composed of the pretrained ResNet–MLP module and a fusion head. The ResNet backbone, together with the classification and regression heads, is trained in an NR setting as described in Fig. 2. The predicted distortion type is transformed into a one-hot encoded vector and fused with the predicted distortion level through the fusion head, which is jointly trained with the final MLP QoE predictor.

During the subjective experiments, the operator was instructed to follow the cursor's movement and make contact with a spring on the table. Accordingly, the interaction status tokens were designed to align with this task. Seven status tokens, namely Free Moving (FM), Slow Pushing (SP), Keeping Position (KP), Slow Leaving (SL), Fast Shaking (FS), Bouncing Back (BB), and Hard Contact (HC), are defined to describe different statuses of the contact interaction of the robot.

All 15 reference signals are annotated semi-automatically. The token sequences are initially generated based on movement speed and interaction force and then manually refined. As a result, 15 sequences of status tokens are created as the ground truth for training the Seq2Seq model. To expand the training dataset, force signals with low levels of synthetic distortion are also included, maintaining the same token sequence order as their corresponding reference signals.

The input signals are separated into short segments, similar to word embedding vectors, and subsequently fed into the Long Short-Term Memory (LSTM). The output of the Seq2Seq model is not fixed in length. A maximum output length is defined for the status token sequence. If the generated output exceeds this maximum length, it is truncated; if it is shorter, the empty slots are filled with End-Of-Sequence (EOS) tokens. The output token sequence is then encoded by a one-hot encoder and flattened. The semantic information, which stands for the interaction and movement status of the robots, is represented by this fixed-length semantic vector.

Finally, the fixed-length semantic vector, representing the movement sequence of the feedback signal, is concatenated with the distortion vector and fed into the integration MLP. During the training phase of the MLP, the parameters of both the semantic and distortion channels remain frozen, except for those of the fusion head in the distortion channel. The integration MLP utilizes the pseudo-MOS as the labels. Four datasets, namely the augmented dataset, the GAN-augmented dataset, and the dataset incorporating both augmentation methods, are trained separately to assess the effect of augmentation.

IV. Experiments and Results

A. Experimental Settings

For our subjective experiments, the Force Dimension Falcon device is used as the haptic device, while the haptic codec reference software [19] is employed to simulate the communication network and interactions with virtual environments. All networks are trained using CUDA version 12.2 on an NVIDIA GeForce 4090 GPU.

The original dataset consists of 2715 samples, including 240 distorted signals annotated with the subjective experiments described in Section III. We selected 6 distortions as artificial distortions: white noise, Poisson noise, communication delay, low-pass filtering, packet loss, and deadband. These distortions are commonly encountered in teleoperation scenarios and can be reliably reproduced in a controlled manner. For each of these 6 distortions, 100 distortion levels are generated and applied to 15 undistorted signals, expanding the dataset to 11715 samples. The distortion levels are selected based on FR features, such as MSE and PSNR, of the distorted signals. Distorted signals with the highest distortion levels are limited to have FR features comparable to those of the most distorted authentic signals. Additionally,

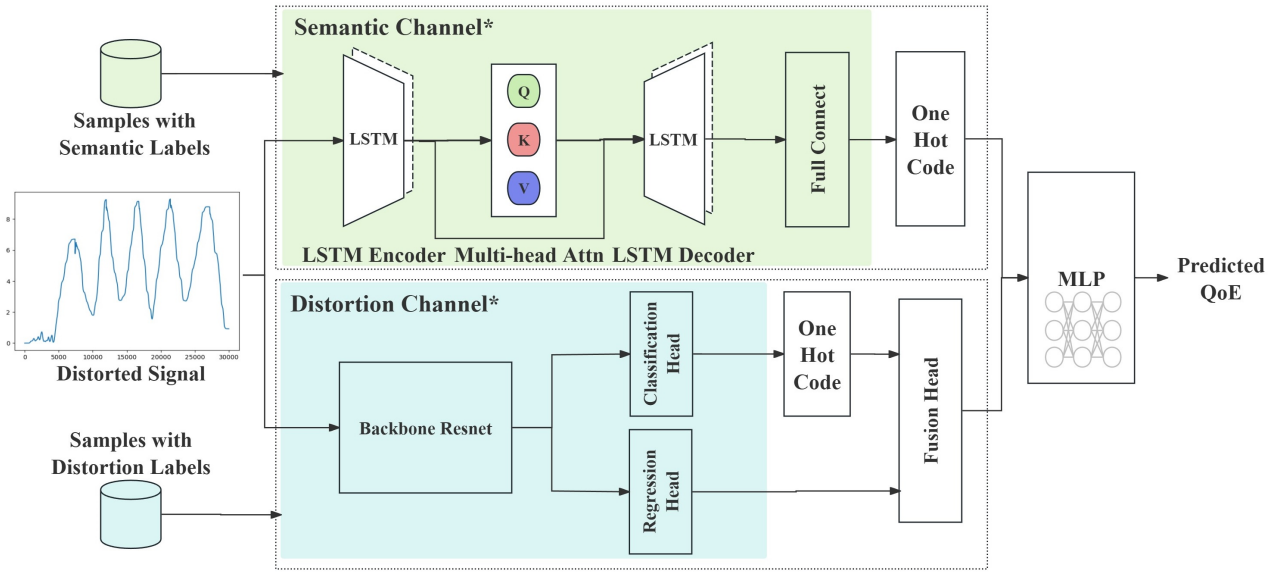

Fig. 3. The proposed dual-channel BHQA model is composed of a semantic channel and a distortion channel. The semantic channel adopts a Seq2Seq [26] architecture with multi-head attention as its backbone, while the distortion channel is built on a one-dimensional ResNet backbone. In separate trainings, the green section of the semantic channel is optimized with samples annotated with semantic information, and the blue section of the distortion channel is optimized with samples labeled with distortion types and levels. Their parameters are frozen (marked by $*$) during the final training of the QoE prediction MLP.

400 signals are synthesized for each of the 15 undistorted signals using the conditional GAN, with 10 levels of pseudo-MOS. Thus, the final size of the augmented pseudo-MOS dataset is 17715.

The original experiments involved 20 subjects, consisting of 14 males and 6 females. Among them, 7 were familiar with teleoperation, 11 had only heard of the concept, and 2 had no prior knowledge. The validation experiments were conducted with 8 new additional subjects (7 males and 1 female), all of whom were acquainted with teleoperation. All experimental procedures and associated questionnaires have been approved by the Ethics Committee of the Technical University of Munich, under certificate number 2023-401-S-NP.

The original data were partitioned into training and test subsets, with 10% reserved for testing. The test set was fixed across all experiments to ensure consistency. To mitigate overfitting, a dropout rate of 0.2 was applied to the MLP layers, and early stopping was applied based on test loss, with training terminated if no improvement for five consecutive epochs.

B. Distortion Analysis and Pseudo-MOS Generation

The rank correlations, namely SROCC and Pearson's Linearity Correlation Coefficient (PLCC), between the ground-truth MOS and pseudo-MOS on the subjective dataset without augmentation are calculated to assess the performance of the ResNet-MLP model in pseudo-MOS generation. The results are presented in Tab. I and compared with FR objective metrics and MAF-based integrated methods. The rows "Adaboost" and "RandomForest" in Tab. I represent the MAF-based method using different machine learning models for metrics integration.

TABLE I
RANK CORRELATIONS BETWEEN GROUND-TRUTH AND PSEUDO-MOS COMPARED WITH FR METHODS(THE HIGHER THE BETTER)

	SROCC	PLCC
Resnet-MLP	**0.94**	**0.96**
Adaboost [8]	0.87	0.89
RandomForest [8]	0.72	0.83
HSSIM	0.52	0.55
segSNR	0.33	0.13

The results indicate that MAF-based methods demonstrate significantly better performance than single objective metrics because they integrate multiple objective metrics. The proposed ResNet-MLP model outperforms MAF-based methods as it leverages the ResNet architecture to automatically extract latent features, thereby capturing more information from the signal. The high SROCC and PLCC values indicate that the pseudo-MOS exhibits strong consistency with the ground-truth MOS and can be utilized as the ground truth in subsequent network training.

C. QoE Prediction Results

Tab. II presents the SROCC of the prediction results of the proposed BHQA model in comparison with FR methods. All methods are evaluated across four datasets: the original subjectively annotated dataset, the Distortion Augmented (DA) dataset, the dataset augmented with GAN, and the dataset augmented with both DA and GAN. The SROCC score is calculated between the predicted values and the ground-truth MOS and pseudo-MOS. The *Original* column represents the subjectively annotated dataset where MOS

is available. In contrast, the DA dataset, GAN-augmented dataset, and DA&GAN dataset utilize pseudo-MOS as the ground truth. The second and third rows correspond to the results of MAF-based feature integration methods using the Adaboost and random forest models, respectively.

TABLE II
SROCC RESULTS OF DIFFERENT METHODS OVER DATASETS WITH DIFFERENT AUGMENTATION METHODS(THE HIGHER THE BETTER)

	Original	DA	GAN	DA&GAN
Proposed BHQA(NR)	**0.85**	**0.88**	0.80	**0.84**
Adaboost(FR) [8]	**0.87**	**0.91**	0.82	**0.83**
RandomForest(FR) [8]	**0.85**	0.84	**0.86**	0.74
HSSIM(FR)	0.57	0.41	0.33	0.26
segSNR(FR)	0.39	0.45	0.26	0.24

From Tab. II, we observe that among the FR methods, MAF-based approaches outperform simple objective metrics, with the Adaboost-based model achieving slightly better performance than the random forest model. The proposed NR methods demonstrate performance comparable to that of MAF-based FR methods. A notable observation is that the proposed method performs worse on the GAN-augmented dataset. This can be explained as the proposed method is an NR method. If the reference data represents force feedback with inherently lower QoE, such as hard contact with a solid surface or rapid consecutive pushing and releasing with large movements, the model may misclassify these movements as authentic distortions with the absence of the reference signal.

D. Validation Experiments

To further evaluate the effectiveness of the proposed model, validation experiments were conducted in real-world teleoperation scenarios. A Panda robotic arm was employed as the follower to interact with the environment, while two different haptic devices, the Force Dimension Omega 6 and Sigma 7, were used as the leader devices. The leader haptic device and the follower robot were operated through separate controller PCs connected via Ethernet. The follower robot was controlled using an impedance controller with fixed parameters, including damping and stiffness.

Two key differences exist between the subjective experiments used to construct the original dataset and the validation experiments. First, the subjective experiments in virtual environments relied on recorded animation and force signals, with participants instructed to passively follow the movement and perceive the feedback. In contrast, reproducing identical motion and force feedback in real-world teleoperation is difficult due to hardware errors and environmental uncertainties. Therefore, in the validation experiments, participants were instructed to actively perform tasks without requiring identical movements and force feedback signals across different subjects.

Second, in the validation experiments, distortions were generated naturally by the control system and hardware. Time delay, deadband, and packet loss were selected as the main network distortions. When the time delay was below 100 ms, mechanical damping was sufficient to stabilize the teleoperation system, and therefore, TDPA was not applied. Once the time delay exceeded 100 ms, TDPA was employed to ensure system stability. The intrinsic delay between the controllers was less than 1 ms, and the natural packet loss rate was below 1%. To simulate the degraded network conditions commonly encountered in teleoperation scenarios such as space exploration and remote rescue, time delay was introduced through a delay buffer, while packet loss was simulated using a random packet discarder on the sending side.

The validation task was designed as vertically pressing a sponge while perceiving the force feedback. Such a pressing task is commonly used in teleoperation research to evaluate the quality of force feedback, and it reflects industrial applications of teleoperation, including remote scanning and haptic digital twins [27]. In addition, this task was chosen to align with the subjective dataset, which consists of one-dimensional force feedback signals used for training our model.

TABLE III
SROCC, PLCC, AND RMSE FOR VALIDATION EXPERIMENTS (MOS FROM OMEGA 6 HAPTIC DEVICE ARE USED AS THE GROUND-TRUTH)

	SROCC↑	PLCC↑	RMSE↓
MOS(Sigma 7)	1.0	0.99	0.26
Proposed BHQA(NR)	0.96	0.95	0.57
MLP Pseudo-MOS(FR)	0.96	0.98	0.48
HSSIM	0.64	0.43	1.31

All 8 participants were instructed to perform 7 subjective validation experiments with varying levels of distortions for each haptic device, and one additional experiment without distortions was included as a reference, with its QoE score set to the maximum. The reference force feedback signals were measured with a force–torque sensor mounted at the end of the Panda arm. For validation, prediction results were obtained from three approaches: the FR perceptual metric HSSIM, the FR ResNet–MLP pseudo-MOS model, and the proposed NR pHQA model.

Tab. III summarizes the results of the validation experiments. The MOS collected with the Omega 6 haptic device was used as the ground truth. From the first row of Tab. III, it can be observed that the MOS collected with the Sigma 7 device is highly consistent with the ground truth, indicating that different haptic devices with similar mechanical configurations have only a marginal effect on perceptual quality, particularly in terms of rank correlation. The remaining results further show that the pseudo-MOS predictions closely align with the ground-truth MOS, and that the proposed NR method achieves comparable performance.

V. DISCUSSION

Tab. IV presents the results of the ablation experiments. In these experiments, the semantic channel, the attention module within the semantic channel, and the distortion

TABLE IV

SROCC RESULTS OF ABLATION EXPERIMENTS FOR STRUCTURE COMPONENTS (THE HIGHER THE BETTER)

	Original	DA	GAN	DA&GAN
Proposed BHQA	**0.83**	**0.86**	**0.78**	**0.80**
Semantic w/o. Attention	0.77	0.82	0.71	0.76
w/o. Semantic	0.70	0.71	0.53	0.74
w/o. Distortion	0.50	0.42	0.36	0.44

channel are sequentially deactivated to assess their individual contributions to the BHQA model.

The first observation is that all models with specific components removed has worse performance than the complete BHQA model. This indicates that both the semantic and distortion channels contribute significantly to the performance of the dual-channel network. The second observation is that the removal of the attention module has a relatively minor impact. This is because the training dataset remains relatively small, preventing the attention mechanism from achieving its full potential. As the dataset size increases, the impact of removing the attention module becomes more obvious.

The third observation is that turning off either the semantic or distortion channel leads to a substantial decline in performance, resulting in noticeably worse results than those of FR methods in Tab. II. Moreover, the distortion channel contributes more significantly to the model's performance. This can be caused by the fact that authentic distortions in force feedback signals are the primary factor affecting QoE. Furthermore, when distortion levels are excessively high, the semantic channel cannot extract effective information from the feedback signals.

Fig. 4 and Fig. 5 present two groups of distorted signals along with their corresponding MOS (if subjectively annotated), pseudo-MOS, and predicted results obtained from the FR method Adaboost-MAF and the proposed BHQA method.

In the left figure of Fig. 4, the authentic distorted data is shown alongside its reference signal. The distorted data receives a high MOS score because, from the operator's perspective, its pattern closely resembles that of the reference signal, and the consistent error between the two is hardly perceptible. Both the pseudo-MOS and the BHQA model demonstrate high scores, whereas the prediction result from Adaboost-MAF is relatively low. This discrepancy arises because main objective metrics, such as MSE and PSNR, give lower QoE scores and contribute negatively to evaluations.

The distorted signal in the right figure is affected by synthetic time delay and does not have a subjectively annotated MOS. The pseudo-MOS is high, which aligns with expectations, as human operators typically cannot perceive small time delays. While the FR Adaboost-MAF assigns a relatively low score for the same reason observed in the left figure, the BHQA model demonstrates stable performance.

The second case study in Fig. 5 illustrates two examples characterized by a high frequency of pushing and releasing movements. The distorted signal in the left figure is generated by the TDPA control algorithm, which is commonly employed in teleoperation systems to maintain stability under network delay. A distinctive feature of the distortion introduced by TDPA is the occurrence of sudden force jumps during the releasing movement. This force feedback signal receives a low MOS score, as human operators frequently experience disturbances due to these sudden force jumps. However, the FR Adaboost-MAF model assigns a relatively higher score because the force jumps are brief and have minimal impact on many FR objective metrics. In contrast, both the pseudo-MOS and the prediction result from the BHQA model align closely with the MOS.

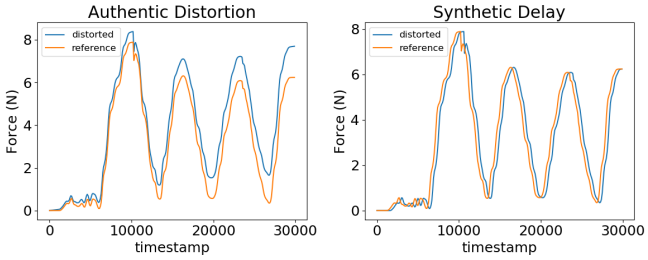

Fig. 4. Case study for the proposed BHQA model. The left distorted signal has MOS = 91.5, pseudo-MOS = 88.3, Adaboost-MAF = 69.2, and BHQA = 92.9. The right synthetic time-delayed distorted signal has pseudo-MOS = 93.2, Adaboost-MAF = 84.6, and BHQA = 92.7.

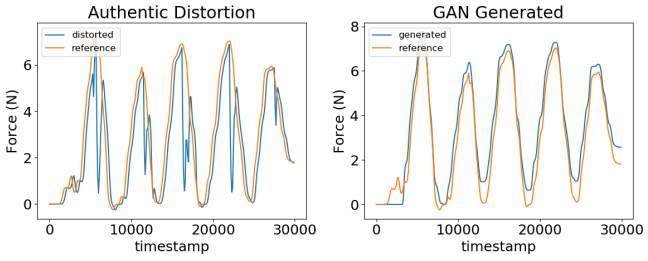

Fig. 5. Another case study for the BHQA model. The left distorted signal has MOS = 67.2, pseudo-MOS = 68.1, Adaboost-MAF = 87.0, and BHQA = 62.2. The right GAN-generated signal is generated with a target pseudo-MOS of 89.8, resulting in Adaboost-MAF = 93.0 and BHQA = 74.6.

In the right figure, a force feedback signal is generated by a conditional GAN with a target pseudo-MOS of 90.0. Since the generated signal closely resembles the reference signal, Adaboost-MAF assigns a high score. However, the BHQA model performs worse in this case, providing a lower score that is inconsistent with the pseudo-MOS. This discrepancy arises due to the semantic channel in the BHQA model. As the signal represents multiple pushing and releasing movements, the BHQA model generates several semantic tokens corresponding to bouncing back, leading to a lower score. This case demonstrates a scenario where BHQA performs worse compared to Adaboost-MAF due to its NR nature, and the situation happens more often in the GAN augmented dataset. This phenomenon can also be observed in Tab. II. To address this issue, our future work will incorporate additional information to enhance the model's awareness of the control scenario and working conditions.

VI. Conclusion

In this paper, we propose a dual-channel BHQA model based on semantic and distortion analysis. Our proposed NR method achieved comparable or better performance on our augmented subjective QoE dataset for bilateral teleoperation compared to FR methods. Our experimental results demonstrate that both the semantic and distortion channels contribute to QoE prediction, and the model's performance is comparable to or exceeds certain FR methods. In order to utilize DNN models in HQA, we augment our dataset with both synthetic noise and a GAN-based model. We analyzed the prediction results using rank correlation metrics and detailed case studies, and further discussed the performance of the proposed model under different experimental conditions. Through the validation experiments, we evaluated not only the effectiveness of the pseudo-MOS annotated dataset but also the robustness and general applicability of the proposed model. These analyses demonstrate its potential for practical deployment in real-world applications.

Our model establishes a solid foundation for the BHQA problem while maintaining a flexible network structure. In future work, we plan to expand our dataset with additional subjective experiments, a more powerful testbed, and data collected under a wider range of teleoperation conditions. The assessed task will be extended from one-dimensional to multi-dimensional interactions, and the model will be redesigned to capture compound effects in multi-dimensional force feedback. With the enlarged dataset, we also aim to explore larger pretrained models and transfer learning techniques to further improve prediction performance. Moreover, we also aim to develop a real-time, online NR HQA system capable of making QoE predictions simultaneously while the operator is controlling the robot. The predicted QoE can be leveraged to modify the control parameters to provide a better human-in-the-loop experience to the operator.

References

[1] R. V. Patel, S. F. Atashzar, and M. Tavakoli, "Haptic feedback and force-based teleoperation in surgical robotics," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 1012–1027, 2022.

[2] D. Wei, B. Huang, and Q. Li, "Multi-view merging for robot teleoperation with virtual reality," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8537–8544, 2021.

[3] A. Pastor and P. Le Callet, "Towards guidelines for subjective haptic quality assessment: A case study on quality assessment of compressed haptic signals," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1667–1672, 2023.

[4] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2864–2878, 2023.

[5] S. Wang, Y. Lin, M. Hao, H. Xu, and Q. Tian, "Interference quality assessment of speech communication based on deep learning," *IEEE Transactions on Reliability*, vol. 71, no. 2, pp. 1011–1021, 2022.

[6] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[7] R. Hassen and E. Steinbach, "Hssim: An objective haptic quality assessment measure for force-feedback signals," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2018.

[8] Z. Wang, F. Mei, X. Xu, and E. Steinbach, "Towards subjective experience prediction for time-delayed teleoperation with haptic data reduction," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 129–134, 2022.

[9] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A fusion-based video quality assessment (fvqa) index," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pp. 1–5, 2014.

[10] M. J. Ramstead, C. Hesp, A. Tschantz, R. Smith, A. Constant, and K. Friston, "Neural and phenotypic representation under the free-energy principle," *Neuroscience Biobehavioral Reviews*, vol. 120, pp. 109–122, 2021.

[11] P. Hinterseer, S. Hirche, S. Chaudhuri, E. Steinbach, and M. Buss, "Perception-based data reduction and transmission of haptic data in telepresence and teleaction systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 588–597, 2008.

[12] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Objective assessment of perceptual audio quality using visqolaudio," *IEEE Transactions on Broadcasting*, vol. 63, no. 4, pp. 693–705, 2017.

[13] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "Graphiqa: Learning distortion graph representations for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 25, pp. 2912–2925, 2023.

[14] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[15] P. Kancharla and S. S. Channappayya, "Completely blind quality assessment of user generated video content," *IEEE Transactions on Image Processing*, vol. 31, pp. 263–274, 2022.

[16] K. Sim, J. Yang, W. Lu, and X. Gao, "Blind stereoscopic image quality evaluator based on binocular semantic and quality channels," *IEEE Transactions on Multimedia*, vol. 24, pp. 1389–1398, 2022.

[17] J. Ma, J. Wu, L. Li, W. Dong, X. Xie, G. Shi, and W. Lin, "Blind image quality assessment with active inference," *IEEE Transactions on Image Processing*, vol. 30, pp. 3650–3663, 2021.

[18] C. Zheng, K. Wang, S. Gao, Y. Yu, Z. Wang, and Y. Tang, "Design of multi-modal feedback channel of human–robot cognitive interface for teleoperation in manufacturing," *Journal of Intelligent Manufacturing*, vol. 36, no. 6, p. 4283, 2025.

[19] "Ieee standard for haptic codecs for the tactile internet," *IEEE Std 1918.1.1-2024*, pp. 1–127, 2024.

[20] M. Panzirsch, H. Singh, X. Xu, A. Dietrich, T. Hulin, E. Steinbach, and A. Albu-Schaeffer, "Enhancing the force transparency of the energy-reflection-based time-domain passivity approach," *IEEE Transactions on Control Systems Technology*, vol. 33, no. 1, pp. 181–188, 2025.

[21] J.-H. Ryu, C. Preusche, B. Hannaford, and G. Hirzinger, "Time domain passivity control with reference energy following," *IEEE Transactions on Control Systems Technology*, vol. 13, no. 5, pp. 737–742, 2005.

[22] Y. Ye, Y.-J. Pan, and Y. Gupta, "A power based time domain passivity control for haptic interfaces," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 7521–7526, 2009.

[23] X. Xu, C. Schuwerk, B. Cizmeci, and E. Steinbach, "Energy prediction for teleoperation systems that combine the time domain passivity approach with perceptual deadband-based haptic data reduction," *IEEE Transactions on Haptics*, vol. 9, no. 4, pp. 560–573, 2016.

[24] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, pp. 119–133, 2019.

[25] Z. Wang, X. Xu, D. Yang, Z. Wang, S. Shtaierman, and E. Steinbach, "Haptic dataset augmentation with subjective qoe labels using conditional generative adversarial network," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5072–5078, 2023.

[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[27] M. Panzirsch, J.-H. Ryu, and M. Ferre, "Reducing the conservatism of the time domain passivity approach through consideration of energy reflection in delayed coupled network systems," *Mechatronics*, vol. 58, pp. 58–69, 2019.