No-Reference Video Quality Assessment with Heterogeneous Knowledge Ensemble

Jinjian Wu†
Yongxu Liu†
jinjian.wu@mail.xidian.edu.cn
yongxu.liu@stu.xidian.edu.cn
School of Artificial Intelligence
Xidian University

Weisheng Dong
wsdong@mail.xidian.edu.cn
School of Artificial Intelligence
Xidian University

Leida Li
ldli@xidian.edu.cn
School of Artificial Intelligence
Xidian University

Guangming Shi
gmshi@xidian.edu.cn
School of Artificial Intelligence
Xidian University

ABSTRACT
Blind assessment of video quality is still challenging even in this deep learning era. The limited number of samples in existing databases is insufficient to learn a good feature extractor for video quality assessment (VQA), while manually labeling a larger database with subjective perception is very labor-intensive and time-consuming. To relieve such difficulty, we first collect 3589 high-quality video clips as the reference and build a large VQA dataset. The dataset contains more than 300K samples degraded by various distortion types due to compression and transmission error, and provides weak labels for each distorted sample with several full-reference VQA algorithms. To learn effective representation from the weakly labeled data, we alleviate the bias of single weak label (i.e., single knowledge) via learning from multiple heterogeneous knowledge. To this end, we propose a novel no-reference VQA (NR-VQA) method with HETerogeneous Knowledge Ensemble (HEKE). Comparing to learning from single knowledge, HEKE can theoretically reach a lower infimum, and learn richer representation due to the heterogeneity. Extensive experimental results show that the proposed HEKE outperforms existing NR-VQA methods, and achieves the state-of-the-art performance. The source code will be available at https://github.com/Sissuire/BVQA-HEKE.

CCS CONCEPTS
• Computing methodologies → Image processing. Computer vision representations.

† Both authors contributed equally to this work, and should be considered co-first authors.
∗ Corresponding author.

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 ACM 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.

MM '21, October 20–24, 2021, Virtual Event, China
ACM ISBN 978-1-4503-8651-7/21/10...
https://doi.org/10.1145/3474085.3475550

1 INTRODUCTION
There are more than 1 billion hours of video viewed by users each day on YouTube, and online videos are expected to make up more than 82% of all consumer Internet traffic this year [4]. While videos are enriching our daily lives, the assessment of video quality is becoming more crucial. Reliable video quality assessment (VQA) methods are required in many video-related applications (e.g., compression, transcoding, delivering, etc.) to ensure the optimal trade-off between video quality and network condition. Due to the absence of reference information, no-reference VQA (NR-VQA) or blind VQA (BVQA) is more suitable and desirable in most cases, thus attracting much research attention.

Feature representation is of importance for NR-VQA task, and effective features would definitely facilitate reliable NR-VQA methods. Handcrafted features have been elaborately designed within various contexts in previous work. For example, they assumed that the video quality would affect some statistical distributions, such as the statistical characteristics of natural images/videos [19, 18], optical flow [13], discrete cosine transformation (DCT) coefficients [11], etc. However, these handcrafted features are insufficient for NR-VQA, and it is hard to elaborate the satisfactory features manually and carefully due to the complexity of subjective perception.

As deep learning has dominated many vision-related tasks, recent work in NR-VQA also tries to manipulate features with learning rather than designing [12, 35]. However, learning effective features from scratch requires a lot of data, which is not available in VQA. Current VQA databases [21, 26, 33, 16] contain only hundreds of samples, which is inadequate to learn effective and robust feature
representation. Even in the recent released databases of user generated content (UGC) [8, 22], the number of available samples is also small (1200 samples in [8], and 585 samples in [22]) comparing to that of parameters in convolutional neural networks (CNN). Even though features from other related tasks can be transferred to NR-VQA [1, 10], the differences between tasks restrict the further improvement of these methods. In general, the lack of data limits the advancement of current NR-VQA methods.

To alleviate the limitation of data and thus learn effective representation for NR-VQA, in this work, a WEakly Labeled Large-scale video quality dataSet (WELL-Set) is first built to meet the data requirement in deep learning models. The WELL-Set includes various compression artifacts and transmission error (packet loss), and results in over 320K samples, which would facilitate the representation learning of neural networks for NR-VQA task. Considering the difficulty of manually labeling, we resort to full-reference VQA (FR-VQA) methods, and the samples are automatically labeled with FR-VQA algorithms. Single FR-VQA algorithm is considered as single prior knowledge accounting for partial characteristic of human visual system (HVS). For example, multi-scale structural similarity index measure (MS-SSIM [27]) is supposed to measure the perceptual quality from the perspective of structural degradation, and spatiotemporal reduced-reference entropic differences (ST-RRED [23]) measures the video quality in the framework of information entropy. However, quality from subjective perception is supposed as an integration of many factors, thus partial knowledge would cause some certain bias and weaken the label. Therefore, in order to learn effective representation from weakly labeled data, we propose a novel NR-VQA with Heterogeneous Knowledge Ensemble (HEKE) to alleviate the bias from single knowledge (i.e., single weak label) as shown in Fig. 1. Comparing to learning from single knowledge, HEKE can reach a lower infimum, maintain the heterogeneity of knowledge, and learn richer feature representation for our task. With the WELL-Set and HEKE, a feature encoding network is trained for NR-VQA. After the training, the network can be directly adopted as representation extractor without fine-tuning when transferred to VQA databases. The fully connected (FC) layers are adopted to regress the features and form our NR-VQA model.

The main contributions of this work are summarized as below:

- We build a large-scale video dataset containing over 320K video clips. Videos are distorted with various compression and transmission artifacts, and are weakly labeled. 
- A novel method (HEKE) is proposed to learn effective representation from weakly labeled data for NR-VQA. Comparing to methods learning from single knowledge, HEKE enables a lower infimum and enriches the representation.
- With the well trained representation extractor, a NR-VQA method based on HEKE is established. Extensive experiments demonstrate that the feature learned from HEKE is representative for NR-VQA, and the proposed method achieves a high consistency with subjective perception.

2 RELATED WORK

NR-VQA methods, generally measuring the degradation from compression and transmission error, have attracted much attention in the last decade. Before the prevalence of deep learning, handcrafted features have been exhaustively explored for NR-VQA, and many of them are based on the statistical characteristics. VideoBLINDS [19] builds a spatiotemporal natural scene statistics (NSS) model for videos. The work in [11, 36] utilizes the statistical distributions in DCT domain to characterize the perceptual degradation. Besides, based on the hypothesis that distortions affect the statistics of optical flow, an optical flow-based NR-VQA is also proposed in [13]. More recently, the spatiotemporal statistics of mean subtracted and contrast normalized (MSCN) coefficients, as well as 3D Gabor coefficients, are also studied [18]. Besides the statistic-oriented features, VIDEOD [14] constructs a completely BVQA based on the same idea of NIQE [15], TL-VQM [9] and VIDEVAL [25] integrate many related features to deal with UGC-VQA. Even though these handcrafted features are elaborated and manipulated carefully, NR-VQA is such a challenging task that it is almost impossible to exhaust all the possible features, and the difficulty of feature designing blocks the further improvement.

With the limitation of feature designing and prevalence of deep learning, recently, more attention is attracted in feature learning for NR-VQA. Due to the constraint of data in VQA, most of the methods extract features with network pretrained from other tasks (e.g., image quality assessment (IQA) or image classification). The work in [30, 1] directly extracts frame-wise features with IQA models, and the work in [10, 3] extracts frame-wise features with classification-based ResNet model. However, these feature extractors are not trained for NR-VQA, thus the capability of representation is limited. To deal with the shortage of data and train a VQA-specific model, the work in [32] adopts cube-based training to augment the data, and the work in [35] trains the model with single weak label. However, both of the two methods are trained within the database, and the trained models are database-specific. Since the number of samples...
in the database is small, the diversity and generalization of the learned features are limited. Further, single weak label from single knowledge is strongly biased, which also weakens the capability of learned representation. VMEON [12] adopts the similar weakly supervised learning with single weak label but enhances the model with other auxiliary tasks. Even though auxiliary tasks might relieve the bias from single knowledge, the method only considers limited compression artifacts, which limits the utility. Also, the data and the pretrained model are not publicly available, which is not friendly to promote the advancement of NR-VQA.

In general, the shortage of data in VQA makes it hard to train a database-agnostic and completely VQA-specific model for NR-VQA. Still, an effective and publicly available representation extractor is required. To tackle this problem, in this work, we first build a weakly labeled large-scale video dataset to alleviate the limitation of data. To learn effective representation extractor from the weakly labeled data, we propose HEKE to weaken the bias from single knowledge and learn richer representation. The trained model can be directly utilized as feature extractor for NR-VQA.

3 A LARGE-SCALE VIDEO DATASET

The number of videos in current VQA databases can not guarantee the effectiveness and generalization of the learned representation. To alleviate the limitation of data in NR-VQA, WELL-Set, a weakly labeled large-scale video dataset, is built in this work. In this section, we first give a view of data generation in the dataset, and then provide the detailed data labeling.

3.1 Data Generation

We manually collect thousands of videos with at least 720p from both UGC and professional generated content (PGC). Most of the source videos are from websites (e.g., YouTube and Bilibili), and the others are from movies. To form the dataset, 861 videos are manually selected from the source videos, and each video is guaranteed to be without obvious artifacts. Based on the 861 source videos, we generate 3589 video clips with 10s duration as the reference. The clips from the same video can be totally different since scene change exists. To relieve the burden of data storage, video clips are downsized to 768 × 432, 432 × 768, 640 × 480 or 480 × 640 according to the original aspect ratio. Fig. 2 gives some sample frames of the selected videos.

We investigate the distortion types among current VQA databases, and implement various compression artifacts as well as transmission error (packet loss). A brief description of distortion types is given in the following.

- **Compression Artifacts**: We implement various compression artifacts with FFmpeg [6], including HEVC, H.264, MPEG-4, MPEG-2, MJPEG, MJ2K and Snow. According to the CSIQ database [26], white noise is also introduced with H.264 compression in the dataset. As H.264 and HEVC are very popular and widely used today, we adopt different compression modes for these two compression methods (i.e., average bitrate (ABR)-based, quantization parameter (QP)-based, and constant rate factor (CRF)-based). As for MPEG-4, we adopt ABR and QP separately to control the quality variation. Therefore, there are total 13 different distortion types from compression.

- **Transmission Error**: Similar to various VQA databases [21, 26, 33, 5], packet loss with H.264 compression is simulated. We compress the video clips in H.264 with two different CRF settings and implement packet loss with a two state Gilbert’s model [5]. Different CRF settings simulate the variable conditions (fair or worse). When the network situation is fair, the videos are compressed with a normal CRF, and vice versa.

All the reference clips are suffered from the compression and transmission artifacts with six degradation levels. The encoding parameters are carefully tuned to ensure a wide range of quality variation. Thus, the WELL-Set contains 3589 × 15 × 6 = 323010 samples. Tab. 1 lists a brief overview of several commonly used VQA databases and the WELL-Set. For simplicity, the compression artifacts are denoted as CMP, and transmission error is denoted as TR in the table. The established WELL-Set makes a large improvement considering both of the various content and the numerous samples.

3.2 Data Labeling

Manually labeling the massive data with mean opinion scores (MOS) or difference MOS (DMOS) is almost impossible. Hence, rather than subjectively labeling, the objective FR-VQA methods are adopted for the automatic judgment. The following gives several fast and robust FR-VQA methods for data labeling:

- **MS-SSIM** [27]: MS-SSIM is a popular IQA method, and measures the degradation from the perspective static structure.

---

**Table 1: A brief overview of existing VQA databases. Denote compression artifacts as CMP and transmission error as TR.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ref. Videos</th>
<th>Dist. Types</th>
<th>Dist. Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIVE [21]</td>
<td>10</td>
<td>CMP &amp; TR</td>
<td>150</td>
</tr>
<tr>
<td>CSIQ [26]</td>
<td>12</td>
<td>CMP &amp; TR</td>
<td>216</td>
</tr>
<tr>
<td>IVPL [33]</td>
<td>10</td>
<td>CMP &amp; TR</td>
<td>128</td>
</tr>
<tr>
<td>IVC-IC [17]</td>
<td>60</td>
<td>CMP</td>
<td>240</td>
</tr>
<tr>
<td>EPFL-PoliMI [5]</td>
<td>12</td>
<td>TR</td>
<td>144</td>
</tr>
<tr>
<td>KoNViD [8]</td>
<td>1200</td>
<td>authentic</td>
<td>1200</td>
</tr>
<tr>
<td>LIVE-VQC [22]</td>
<td>585</td>
<td>authentic</td>
<td>585</td>
</tr>
<tr>
<td>WELL-Set</td>
<td>3589</td>
<td>CMP &amp; TR</td>
<td>323010</td>
</tr>
</tbody>
</table>
Training with HEKE

Heterogeneous Knowledge

Feature Encoder

Divide & Conquer

Hierarchical Features Regression

VQA database

Weak labels

Figure 3: The framework of our method. A feature encoder is first trained with HEKE in the WELL-Set $D$. The proposed method employs a "divide and conquer" method to learn heterogeneous representation from the weak labels. $y^*$ is the real label but cannot be available in $D$, and $y^k_w$ is the weak label from corresponding heterogeneous knowledge where $k = 1 \sim 4$. Comparing to learning from only single knowledge (e.g., $y^k_w$), HEKE can acquire richer and more diverse representation simultaneously due to the ensemble of heterogeneous knowledge. The well trained encoder is directly employed in NR-VQA to extract representative features without fine-tuning. A hierarchical feature regression is adopted for the prediction of video quality.

- **GMSD** [31]: GMSD is also an effective but efficient IQA method, which measures the degradation from the perspective of similarity deviation.
- **ST-GMSD**: This method is originally from the work in [28]. It follows the same idea of GMSD but via spatiotemporal gradient magnitude, thus characterizing the degradation of videos from the perspective of space-time gradients.
- **ST-RRED** [23]: ST-RRED also measures video quality in a general spatiotemporal dimension, but with the framework of information entropy.

Some FR-VQA algorithms may provide more accurate labels but the computational burden is huge, thus they are not appropriate for the labeling with massive data. Let $x_i$ denote the $i$-th video clip in the dataset $D$, and $y^k_w$ denote the $k$-th weak label corresponding to $x_i$. Then, $D = \{(x_i, y^k_w) | k = 1, ..., K; i = 1, ..., N\}$, where $N$ is the number of samples in $D$, and $K$ is the number of adopted FR-VQA algorithms ($K = 4$ in this work).

## 4 HETEROGENEOUS KNOWLEDGE ENSEMBLE

Since the data is weakly labeled in WELL-Set $D$, how to learn representative features from the weak labels is still an open challenge. The single weak label from single knowledge is strongly biased, since it only considers part of the valuable information. To alleviate the bias and learn a good feature encoder, we propose to learn representative features with HEKE. In this section, the mathematical formulation of the main idea is first given. Then, we provide the details of HEKE. Finally, we give the implementation of the proposed NR-VQA. The framework of our method is given in Fig. 3.

### 4.1 Formulation

Let $h(\theta; D)$ denote a learning model for NR-VQA within the dataset $D$ defined in 3.2, where $\theta$ is the model parameter. Given an input $x$, $h(x)$ is the predicted quality of the input video, and we omit $\theta$ and $D$ for simplicity. We define $\hat{h}(x) = E_D[h(x)]$ as the expectation of $h(x)$ in dataset $D$. Then, given $y^*$ as the ground truth of $x$, the prediction error of model $h$ in dataset $D$ can be formulated as

$$
\mathcal{E}(h; D) = E_D[(h(x) - y^*)^2]
= E_D[(h(x) - \hat{h}(x) + \hat{h}(x) - y^*)^2]
= E_D[(h(x) - \hat{h}(x))^2] + E_D[(\hat{h}(x) - y^*)^2]
$$

$$
+bias^2(h) + variance(h)
= variance(h) + bias^2(h).
$$

Considering $y^*$ is the true label without noise for simplicity, $\mathcal{E}(h; D)$ can be decomposed to $variance(h)$ and $bias^2(h)$ according to the bias-variance decomposition [7, 2].
As $y^*$ is unavailable in $\mathcal{D}$, only the weak label $y_w$ can be acquired. Then, we have
\[
\text{bias}^2(h) = \mathbb{E}_D[(\hat{h}(x) - y^*)^2] \\
= \mathbb{E}_D[(\hat{h}(x) - y_w + y_w - y^*)^2] \\
= \mathbb{E}_D[(\hat{h}(x) - y_w)^2] + \mathbb{E}_D[2(\hat{h}(x) - y_w)(y_w - y^*)] \\
+ \mathbb{E}_D[(y_w - y^*)^2] + \mathbb{E}_D[2(\hat{h}(x) - y_w)(y_w - y^*)].
\]

The first term $\text{bias}^2(h)$ denotes the offset between the learning model and the weak label. $\mathcal{H}(h, y_w)$ is thought to be the infimum of $\text{bias}^2(h)$, as well as that of $\mathcal{E}(h; \mathcal{D})$. It is assumed that $\mathcal{H}(h, y_w) \geq 0$, because when $y_w \rightarrow y^*$, $\mathcal{H}(h, y_w) \rightarrow 0$, and the infimum reaches the optimal: $\inf_{\mathcal{E}(h; \mathcal{D})} \rightarrow 0$; otherwise, $\inf_{\mathcal{E}(h; \mathcal{D})} > 0$.

Single label from single knowledge to strong bias, which corresponds to the first term of $\mathcal{H}(h, y_w)$. Let $\mathcal{E}(h; \mathcal{D}; k)$ denote the error of the model learned from the $k$-th knowledge, where $k = 1 \sim K$ as given in 3.2, then the infimum of $\mathcal{E}(h; \mathcal{D}; k)$ is
\[
\inf_{\mathcal{E}(h; \mathcal{D}; k)} = \mathcal{H}(h, y^*_w) \\
= \mathbb{E}_D[(y^*_w - y^*)^2] + \mathbb{E}_D[2(\hat{h}(x) - y^*_w)(y^*_w - y^*)] \\
= \mathbb{E}_D[(y^*_w - y^*)^2 + 2(\hat{h}(x) - y^*_w)(y^*_w - y^*)] \\
= \frac{1}{N} \sum_{i=1}^{N} (y^*_w - y^*_i)^2 + 2(\hat{h}(x_i) - y^*_w)(y^*_w - y^*_i) \\
= \frac{1}{N} \sum_{i=1}^{N} I(i, k),
\]
where $N$ is the number of samples in dataset $\mathcal{D}$.

To learn representative features from weakly labeled data, we propose to learn the model with heterogeneous knowledge ensemble. Then, the model can not only learn from single weak label $y^*_w$ but the heterogeneous labels $\{y^*_w|k = 1, \ldots, K\}$. Similarly, let $\mathcal{E}(h; \mathcal{D}; e)$ denote the error of the model learned with HEKE, then
\[
\inf_{\mathcal{E}(h; \mathcal{D}; e)} = \frac{1}{N} \sum_{i=1}^{N} I(i, \omega_i) \\
\leq \inf_{\mathcal{E}(h; \mathcal{D}; k)}
\]
where
\[
\omega_i = \arg\min_k I(i, k), \quad k \in \{1, \ldots, K\}.
\]

The intuitive explanation is that, since HEKE is guided by multiple heterogeneous knowledge, the advantage of every single knowledge would be maintained. As the knowledge is partial and labels are weak, single weak label can not always achieve the optimal among the heterogeneous weak labels. Therefore, HEKE would achieve a lower infimum comparing to learning from single knowledge.

### 4.2 Method

Our goal is to train an effective feature encoder network serving for NR-VQA. To this end, we propose HEKE to learn from weakly labeled data as shown in Fig. 3, we implement HEKE with a “divide and conquer” manner, and the details are given in this subsection.

Let $f$ denote our feature encoder network, which is expected to be a CNN-based architecture without FC layers for simplicity. Then, the representation from the encoder is acquired as $\mathcal{F} = f(x)$ with $C$ channels. That is, $\mathcal{F} = \{f_c|x = 1, \ldots, C\}$. To enable learning from heterogeneous knowledge, a “divide and conquer” manner [34] is introduced for an efficient implementation. Specifically, considering $K$ types of heterogeneous knowledge, there are $K$ weak labels (i.e., $y^*_w, \ldots, y^*_w$) corresponding to the input. Firstly, $\mathcal{F}$ is divided into $K$ subsets, and each subset keeps $\frac{C}{K}$ channels from $\mathcal{F}$. Then, The weak labels from heterogeneous knowledge are conquered respectively by the subsets with $K$ regressors $\mathcal{G} = \{G_k|k = 1, \ldots, K\}$.

Since the input $x$ is a sample of video frames, the dimension of $\mathcal{F}$ can be either $C \times T \times H \times W$ or $C \times H \times W$, where $T$ is the temporal depth, $H$ is the height of the features and $W$ is the width. After the dividing, the features in each subset are pooled by average-pooling and max-pooling. Then the pooled features within subsets are regressed by $\mathcal{G}$ respectively. For simplicity, we consider $\mathcal{G}_3$ as a three-layer FC network. Let $y^*_k$ denote the output of $\mathcal{G}_k$, then the loss function of $\mathcal{G}_k$ is given as
\[
\mathcal{L}_k = \begin{cases} 
0.5(y^*_k - y^*_w)^2 & \text{if } |y^*_k - y^*_w| < 1 \\
|y^*_k - y^*_w| - 0.5 & \text{otherwise}.
\end{cases}
\]

$\mathcal{L}_k$ is known as Huber loss or smooth $L_1$ loss. Then the whole model is optimized with the overall loss function
\[
\mathcal{L} = \sum_{k=1}^{K} \mathcal{L}_k.
\]

### 4.3 NR-VQA

Once the encoder $f$ is trained, it is directly adopted for representation extraction without fine-tuning. An intuitive method is to extract the features of the last layer in $f$, and regress the features into video quality. To this end, a global pooling is adopted to pool the feature maps. Formally, the pooled feature is given as
\[
z = GP(\mathcal{F}) = GP(f(x)).
\]

Then, a regressor predicts the quality as
\[
q = g(z),
\]
where $g$ is the regressor, and generally modeled with FC layers.

The intuitive method would generally achieve a moderate result. In this work, motivated by the work in [29], we also explore the hierarchical features for quality regression. Considering a commonly used ResNet-like architecture (others would be similar), the hierarchical features from conv2$_x \sim$ conv5$_x$ are extracted and globally pooled. Let $\text{res}_1 \sim \text{res}_4$ denote the features extracted from the 4 layers respectively, then the pooled hierarchical features are concatenated as
\[
z_k = GP(\text{res}_1) \oplus GP(\text{res}_2) \oplus GP(\text{res}_3) \oplus GP(\text{res}_4),
\]
where $\oplus$ presents the concatenation operation. Similar as the single-layer regression given in Eq. (9), the predicted quality is also given as $q = g(z_k)$, where $g$ is a three-layer FC network in our work.

To better capture the distribution of features thus benefiting NR-VQA, the global pooling $GP$ is given as $GP = [GP_{\text{ave}}, GP_{\text{max}}, GP_{\text{std}}]$. That is, the three pooling methods (i.e., average-pooling, max-pooling, and standard deviation-pooling) are
adopted and the pooled features are concatenated together. We believe that the richer representation would be acquired with the comprehensive pooling.

4.4 Implementation

The architecture of f is actually model-agnostic. However, to capture spatiotemporal characteristics on VQA but reduce the memory consumption, we adopt the 18-layer R(2+1)D [24] as the backbone network. Thus, C is 512, and given K = 4, the number of channels of each subset is \( \frac{C}{K} = 128 \). The dimension of hidden layers in regressors is set to \([128, 32]\) for both of \( g_k \) and \( g \). We sparsely and uniformly sample 12 short snippets for each 10s video as the input of the network, and each snippet contains 8 frames in RGB color.

In our implementation, the hierarchical features regression is adopted by default. In the experiment of ablation study, we will give the comparison about the regression with features from different layers. We use PyTorch to implement our method, and Adam to optimize the model. The learning rate for training from scratch with HEKE is set to 1e-3, and when the model is fine-tuned on VQA databases, the learning rate is set to 1e-4. The encoder is trained on the WELL-Set only once and fixed during experimental validation. The random seed is fixed during validation to keep the reproducibility.

5 EXPERIMENTAL RESULTS

5.1 Experimental Protocols

5.1.1 Database. In the experiments, 6 subjective VQA databases covering various contents and distortion types are adopted to verify the methods. The following gives a brief review for the databases.

**LIVE Video database** [21] contains 150 distorted videos from 10 reference videos with a resolution of 768 × 432. There are four distortion types including compression and transmission.

**CSIQ video database** [26] contains 216 distorted videos from 12 reference videos with a resolution of 832 × 480. There are six distortion types including compression, transmission, and white noise.

**IVPL video database** [33] contains 128 distorted videos from 10 reference videos with a resolution of 1920 × 1088. There are four distortion types including compression and transmission.

**IVC-IC database** [17] contains 240 distorted videos from 60 reference videos with a resolution of 640 × 480. The database is to explore the influence of various video content and contains compression artifact only.

**EPFL-PoliMI database** [5] contains 144 distorted videos from 12 reference videos with the resolution of both 704 × 576 and 352 × 288. The database is to study the influence of transmission error, thus only packet loss is included. We denote EPFL-PoliMI as EP-Po for short.

**LIVE-Mobile database** [16] contains 200 distorted videos from 10 reference videos with a resolution of 1280 × 720. The database contains both compression and transmission artifacts. In the experiments, we exclude the distorted video with frame-freezing and adopt 160 distorted videos. We denote LIVE-Mobile as LIVE-M for short.

5.1.2 Criteria. We adopt three widely used criteria the evaluate the performance of methods, which are: Pearson linear correlation coefficient (PLCC), Spearman rank-order correlation coefficient (SRCC), and root mean square error (RMSE). PLCC ans SRCC measure the correlation between the predicted quality by methods and the ground truth, and RMSE indicates the relative error. A better method would result in higher PLCC and SRCC, but lower RMSE.

In the experimental tables, we use ↑ (↓) to denote the tendency that the higher (lower) the better, thus resulting in PLCC↑, SRCC↑, and RMSE↓.

5.1.3 Setting. Since most of existing NR-VQA methods are based on learning, if there’s no additional elucidation, the performance is obtained with cross validation in the experiments. Videos in the dataset are split according to the reference, of which 20% are used for validation. For a fair and general performance comparison, in each database (except for IVC-IC), we exhaust all the possible combinations for training/validation splitting, and run the model to get more robust performance. That is, as for LIVE, IVPL and LIVE-Mobile, the training-validation procedure would be repeated.
Table 3: Performance comparison of existing NR-VQA methods among databases

<table>
<thead>
<tr>
<th>Crit.</th>
<th>PR-VQA</th>
<th>NR-VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MS-SSIM</td>
<td>ST-REDD</td>
</tr>
<tr>
<td>LIVE</td>
<td>0.8287</td>
<td>0.8623</td>
</tr>
<tr>
<td></td>
<td>0.8608</td>
<td>0.8418</td>
</tr>
<tr>
<td></td>
<td>5.6059</td>
<td>5.6563</td>
</tr>
<tr>
<td>CSIQ</td>
<td>0.7660</td>
<td>0.8214</td>
</tr>
<tr>
<td></td>
<td>0.7771</td>
<td>0.8091</td>
</tr>
<tr>
<td>IVPL</td>
<td>0.7822</td>
<td>0.8000</td>
</tr>
<tr>
<td></td>
<td>0.8148</td>
<td>0.8091</td>
</tr>
<tr>
<td></td>
<td>0.6004</td>
<td>0.6167</td>
</tr>
<tr>
<td>IVC-IC</td>
<td>0.9100</td>
<td>0.8971</td>
</tr>
<tr>
<td></td>
<td>0.9249</td>
<td>0.8769</td>
</tr>
<tr>
<td></td>
<td>0.4055</td>
<td>0.5118</td>
</tr>
<tr>
<td>EP-Po</td>
<td>0.9557</td>
<td>0.9578</td>
</tr>
<tr>
<td></td>
<td>0.9655</td>
<td>0.9326</td>
</tr>
<tr>
<td></td>
<td>0.3156</td>
<td>0.4481</td>
</tr>
<tr>
<td>LIVE-M</td>
<td>0.8214</td>
<td>0.8908</td>
</tr>
<tr>
<td></td>
<td>0.8890</td>
<td>0.9399</td>
</tr>
<tr>
<td></td>
<td>0.4141</td>
<td>0.3095</td>
</tr>
</tbody>
</table>

Table 4: Cross-database performance (SRCC↑) comparison when trained on LIVE

<table>
<thead>
<tr>
<th>Crit.</th>
<th>CSIQ</th>
<th>IVPL</th>
<th>IVC-IC</th>
<th>EP-Po</th>
<th>LIVE-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBLIINDS</td>
<td>0.0346</td>
<td>0.2439</td>
<td>0.1329</td>
<td>0.3776</td>
<td>0.7362</td>
</tr>
<tr>
<td>VSFA</td>
<td>0.3177</td>
<td>0.3872</td>
<td>0.5484</td>
<td>0.4614</td>
<td>0.7016</td>
</tr>
<tr>
<td>TL-VQM</td>
<td>0.1133</td>
<td>0.2853</td>
<td>0.1258</td>
<td>0.0975</td>
<td>0.2767</td>
</tr>
<tr>
<td>NSTSS</td>
<td>0.1170</td>
<td>0.1984</td>
<td>0.1257</td>
<td>0.0760</td>
<td>0.3792</td>
</tr>
<tr>
<td>VIDEVAL</td>
<td>0.0355</td>
<td>0.3230</td>
<td>0.3499</td>
<td>0.0585</td>
<td>0.3997</td>
</tr>
<tr>
<td>HEKE</td>
<td>0.7778</td>
<td>0.8223</td>
<td>0.8574</td>
<td>0.8592</td>
<td>0.8347</td>
</tr>
</tbody>
</table>

Table 5: Cross-database performance (SRCC↑) comparison when trained on CSIQ

<table>
<thead>
<tr>
<th>Crit.</th>
<th>VBLIINDS</th>
<th>IVPL</th>
<th>IVC-IC</th>
<th>EP-Po</th>
<th>LIVE-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBLIINDS</td>
<td>0.2889</td>
<td>0.1411</td>
<td>0.6262</td>
<td>0.1963</td>
<td>0.4974</td>
</tr>
<tr>
<td>VSFA</td>
<td>0.1758</td>
<td>0.3559</td>
<td>0.7485</td>
<td>0.4917</td>
<td>0.6981</td>
</tr>
<tr>
<td>TL-VQM</td>
<td>0.1398</td>
<td>0.3011</td>
<td>0.4082</td>
<td>0.2955</td>
<td>0.3622</td>
</tr>
<tr>
<td>NSTSS</td>
<td>0.0245</td>
<td>0.2031</td>
<td>0.6304</td>
<td>0.1588</td>
<td>0.1784</td>
</tr>
<tr>
<td>VIDEVAL</td>
<td>0.1830</td>
<td>0.4108</td>
<td>0.5504</td>
<td>0.0735</td>
<td>0.1574</td>
</tr>
<tr>
<td>HEKE</td>
<td>0.7457</td>
<td>0.8661</td>
<td>0.8802</td>
<td>0.8402</td>
<td>0.8413</td>
</tr>
</tbody>
</table>

for 45 times since there are 10 reference videos in each database and the number of combinations for validation set is 45. Similarly, as for CSIQ and EPFL-PoliMi, we repeat the procedure 66 times for an exhaustive validation. Due to the number of combinations in IVC-IC is large enough, we adopt 50 rounds which is randomly split but kept the same for all the methods. The performance in each database is adopted as the median criteria among these combinations.

As for SVR-based methods, we adopt the widely used package LIBSVM and radius basis function (RBF) as the kernel. A grid-based search is used to set the optimal parameters within each training-validation procedure.

5.2 Performance Comparison

In this subsection, the performance is verified in two aspects. In the first phase, since our feature encoder is trained in the WELL-Set, a completely blind performance comparison is conducted to verify the effectiveness of the encoder. In the second phase, the cross-validation experiment within each database is given.

In the first phase, the well trained model is directly examined on the databases with a completely blind manner. In this comparison, NIQE [15] and VIIDEO [14] are compared since the two methods are "completely" BVQA while most of the others are regression-based where the ground truth is required. We introduce two models: one is the model learning from single knowledge (denoted as HEKE1) and the other is learned from 4 kinds of heterogeneous knowledge (denoted as HEKE4). We also introduce several FR methods (e.g., PSNR, and MOVIE [20]) for the comparison and the result is shown in Tab. 2 (The best two performances of completely BVQA are highlighted in bold). In the experiment, HEKE4 is learned from the weak label generated by ST-GMSD, and HEKE4 is guided by the four methods followed, each of which corresponds to a regressor \( G_k \) as given in 4.2. From the table, compared with NIQE and VIIDEO, it is suggested that HEKE4 shows a large improvement, and performs more consistently with subjective perception across the databases. Also, HEKE4 outperforms PSNR and is competitive to SSIM among the databases in a completely blind manner even though the two methods are FR-based. This result is encouraging since HEKE4 is a completely blind method but achieves the performance as good as FR-VQA methods do.

 Besides, as shown in the table, single knowledge has a strong bias. For example, ST-GMSD performs better on LIVE, but worse on CSIQ comparing to GMSD. With HEKE, the biases of methods are alleviated and more representative features are learned in the HEKE4. As is given in the table, the regressor of ST-GMSD in HEKE4 performs
better on LIVE over that of GMSD, while the situation is reversed on CSIQ. Note that we adopt a “divide and conquer” manner to address the ensemble, thus both of the knowledge, regardless of GMSD or ST-GMSD (or the others), are maintained in the model. The heterogeneity of knowledge guarantees the diversity and generalization ability of the learned representation.

In the second phase, the cross-validation experiment within databases is conducted, which follows the settings given in 5.1.3. The proposed method is compared with three FR-VQA methods (MS-SSIM [27], ST-RRED [21] and MOVIE [20]), three NR-VQA methods (VBLiINDS [19], VIIDEO [14] and NSTSS [18]), and several UGC-VQA methods (VSFA [10], TL-VQM [9], and VIDEVAL [25]). The result of the performance comparison is given in Tab. 3. Due to the diversity in databases, some methods achieve good performances in some databases (e.g., EPFL-PoliMI and LIVE-Mobile) but fail in the others (e.g., CSIQ and LIVE). Comparing to existing methods, our proposed HEKE-VQA keeps stable and robust performances among the databases, which demonstrate that the features learned from heterogeneous knowledge is representative and effective. It is also suggested that the learned feature encoder is more suited for NR-VQA comparing to ResNet, which is adopted in VSFA for feature extraction. The main goal of this work is to learn a representative feature encoder specified for NR-VQA, thus filling the gap that there’s no effective and task-specific feature encoder available. Both of the two experimental results demonstrate the effectiveness of HEKE and the learned representation.

### 5.3 Cross-Database Validation

Cross-validation within databases may suffer from overfitting since the data within single database shares the similar distribution. To further verify the generalization of NR-VQA methods, the cross-database experiment is conducted. For example, when a model is trained on LIVE, the model should be validated on CSIQ, IVPL, IVC-IC, PEFL-PoliMI, and LIVE-Mobile simultaneously. In Tab. 4 and Tab. 5, we give the cross-database performance (SRCC) comparison when models are trained on LIVE and CSIQ respectively. Videos in each database is believed to be in the same distribution. From the tables, it is shown that current NR-VQA methods trained on one distribution can hardly deal with the videos in the other distribution. Even though some methods perform well in the cross-validation experiment as given in Tab. 3, the capability of generalization is still not good. From the cross-database experiments, it is clear that the proposed method obtains a good capability of generalization due to the representative features we have learned.

### 5.4 Ablation Study

In this subsection, we explore the effectiveness of the hierarchical features in the learned model. Tab. 6 gives the performance (SRCC) of various types of hierarchical combination. We fix features from res4 since they are the last layer in the encoder. From the table, it is shown that with the hierarchical features, the performance on LIVE is largely improved. However, the performances on IVC-IC and IVPL also show that the prediction error might be increased due to the feature redundancy. The nonuniformity among the databases makes NR-VQA challenging enough, but the proposed method can always keep a stable and robust performance irrespective of features from hierarchy or simply the last layer. This ablation study also demonstrates the effectiveness of representation from the learned feature encoder.

### 6 CONCLUSION

Due to limitation of available data, a good feature encoder is still a gap for NR-VQA even in this deep learning era. To fill the gap, we have first built a large-scale dataset to break through the restriction of data. Over 320K video samples have been generated and weakly labeled. With the dataset, we have proposed a novel method HEKE to learn representative features from weakly labeled data. Comparing to learning from single knowledge, the proposed HEKE is thought to achieve lower infimum and learn more diverse and representative features. We have provided a completely blind VQA model and a feature encoder accounting for video compression and transmission. Extensive experiments have demonstrated the effectiveness of both the proposed method and the representation from the learned feature encoder.

Since the model is trained from compression and transmission artifacts, the feature encoder suffers from a bias when transferred to UGC-VQA. Therefore, strictly speaking, the well established model in this work concentrates only on video compression and packet loss. We hope this work would contribute to the advancement of conventional VQA and the related video society.

### 7 ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (No. 62022063, No. 61772388 and No. 61632019) and the National Key R&D Program of China (2018AAA0101400).