Just Noticeable Distortion Profile Inference: A Patch-level Structural Visibility Learning Approach

Xuelin Shen, Zhangkai Ni, *Graduate Student Member, IEEE*, Wenhan Yang, *Member, IEEE*, Xinfeng Zhang, *Member, IEEE*, Shiqi Wang, *Member, IEEE*, and Sam Kwong, *Fellow, IEEE*

***Abstract*—In this paper, we propose an effective approach to infer the just noticeable distortion (JND) profile based on patch-level structural visibility learning. Instead of pixel-level JND profile estimation, the image patch, which is regarded as the basic processing unit to better correlate with the human perception, can be further decomposed into three conceptually independent components for visibility estimation. In particular, to incorporate the structural degradation into the patch-level JND model, a deep learning-based structural degradation estimation model is trained to approximate the masking of structural visibility. In order to facilitate the learning process, a JND dataset is further established, including 202 pristine images and 7878 distorted images generated by advanced compression algorithms based on the upcoming Versatile Video Coding (VVC) standard. Extensive experimental results further show the superiority of the proposed approach over the state-of-the-art. Our dataset is available at: https://shenxuelin-cityu.github.io/jnd.html.**

***Index Terms*—Just noticeable distortion, visual perception, deep neural network, perceptual video coding.**

I. Introduction

Thanks to the fast development of multimedia technology and network services, there has been a tremendous improvement in viewing experience in multimedia applications. As such, recent years have witnessed a strong demand for economizing the bandwidth and storage space of visual content. The next generation video coding standard Versatile Video Coding (VVC) [1] has substantially improved the coding efficiency compared to High Efficiency Video Coding (HEVC) [2]. However, there is still a strong desire to further investigate potential techniques for coding performance improvement.

Under these circumstances, the perceptual video coding (PVC), which aims at removing the perceptual redundancy in visual content, has recently emerged as an active research topic with great importance for both academia and industry realms. In particular, the just noticeable distortion (JND), which relies on the mechanism of Human Visual System (HVS) that humans cannot perceive small visual signal variations within certain visual contents, has been widely applied in perceptual image/video compression [3]–[6]. More specifically, JND indicates the maximum distortion that is allowed to be injected with the constraint that no visual quality variation is perceived. In the literature, the conventional JND models can be mainly categorized into pixel-wise models which calculate the JND threshold for each pixel straightforwardly in the pixel domain and sub-band JND models which calculate the JND threshold for each sub-band coefficient by transforming the image into frequency domain (*e.g.*, DCT, wavelet) representation.

Both the pixel-wise and sub-band JND models share the same pipeline: visibility maskings of contributing factors are firstly estimated, following which the JND threshold for each pixel/sub-band coefficient is finally derived by the non-linear combination or summation. Regarding the pixel-wise JND models, luminance adaptation (LA) and spatial contrast masking (CM) are dominant factors. An inchoate pixel domain method was proposed in [7], where LA and CM were combined to generate the JND masking. Afterward, the method proposed in [8] optimized the CM model with a Canny edge detector [9]. In [10], the pattern was considered a factor that affects the CM. The model was further improved in [11], where the local diversity orientation was calculated for modeling the pattern complexity. An edge profile reconstruction based JND model was especially established for screen content images by Wang *et al.* [12]. Regarding the frequency domain method, the commonly utilized factor on the sub-band domain is contrast sensitive function (CSF), associated with which the influences of luminance adaptation, contrast masking, and foveated masking can be jointly considered [13]–[18].

As mentioned above, the traditional JND models aim at predicting the JND threshold by gathering the visibility maskings of each corresponding factor. However, it has been widely hypothesized that visual perception is not performed at the pixel level [19], [20]. Moreover, calculating the JND threshold for each pixel/sub-band coefficient independently fails to capture the interaction and correlation among pixels within a visual perception area. This motivates us to model the JND from a new perspective based on patch decomposition. This philosophy also aligns the image coding process which coincidentally operates at each local region. The main contributions can be summarized as follows:

- A JND model is proposed to infer the JND profile with

This work was supported in part by the Hong Kong Research Grants Council (RGC) General Research Funds under Grant 9042816 (CityU 11209819), Grant 9042958 (CityU 11203820) and under Grant 9042957 (CityU 11203220), and in part by the Hong Kong Research Grants Council (RGC) Early Career Scheme under Grant 9048122 (CityU 21211018). *(Corresponding authors: Shiqi Wang and Sam Kwong).*

Xuelin Shen, Zhangkai Ni, Wenhan Yang, and Shiqi Wang are with the Department of Computer Science, City University of Hong Kong, Hong Kong 999077 (e-mail: xuelishen2-c@my.cityu.edu.hk; eezkni@gmail.com; wyang34@cityu.edu.hk; shiqwang@cityu.edu.hk).

Xinfeng Zhang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China (e-mail: xfzhang@ucas.ac.cn).

Sam Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong 999077, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China (e-mail: cssamk@cityu.edu.hk).

the patch decomposition strategy, such that the visibility masking factors for each component could be independently inferred and combined. In this paper, the JND profile refers to that the JND is a patch-level property of the reference patch. To be specific, in the proposed model, the patch would be regarded as basic element to calculate JND instead of pixel/sub-band coefficient, and the JND would be generated simultaneously for the whole patch. The dominant structural degradation that contributes to the JND profile is further obtained in a learning manner through a well trained deep neural network. The proposed scheme not only delivers an accurate estimation of the JND profile, but also sheds light on the development of perceptual image coding.

- To train the JND model, a database based on VVC is further established. In particular, the database is comprised of 202 image sets each of which contains a pristine image and the corresponding distortion version compressed by VTM 5.0 intra coding with QP ranging from 13 to 51. Furthermore, extensive subjective tests have been conducted to obtain the most appropriate QP point of each set under which the distortion caused by intra coding can be just noticed.
- Experimental results show that the proposed JND model can deliver an accurate estimation of the JND profile. In particular, the proposed scheme outperforms the state-of-the-art methods when approximating the JND profile in the domain of perceptual image coding.

The rest of this paper is organized as follows. In Section II, the related works and motivation are introduced. In Section III, the proposed patch-level structural visibility learning model is presented in detail. In Section IV, the dataset construction and methodology of subjective test are detailed. In Section V, extensive performance evaluation of the proposed JND model and other state-of-the-art models are performed and compared. Finally, Section VI draws the conclusion.

II. Related Works and Motivations

A. Related Works

1) Pixel-wise JND estimation: In the literature, the phenomenon of LA and CM is commonly adopted to calculate the JND threshold [21]. In particular, the LA is modeled according to the mechanism that the sensitivity of HVS differs with the variations of background luminance [22]. In [7], the LA model was established as a quasi-parabolic curve where the masking achieves its lowest point in the middle of the luminance range. In terms of CM, it originates from the straightforward observation that regions containing more nonuniform contents can tolerate more distortions comparing with those of homogeneous content. As for the visibility masking estimation, several factors have been considered relevant to CM. The luminance contrast is a straightforward factor and is usually modeled through the local gradient information. For example, in [7], the luminance contrast was derived through the strongest gradients among all directions, and a linear model was employed to infer the CM with the derived luminance contrast. Moreover, the linear model was further optimized with a logarithmic model based on the nonlinearity between the HVS response and contrast [23]. Furthermore, Wu *et al*. [10] modeled the structural uncertainty with pattern complexity based on the philosophy that HVS can usually predict the structure along with the understanding of visual content. Regarding the determination of the JND threshold, the inchoate methodology adopts the stronger masking effect between LA and CM [7]. Yang *et al*. [8] further optimized the JND model by a non-linear combination method based on LA and CM with an overlapping factor.

2) The frequency domain JND estimation: The frequency domain JND models have also been widely investigated especially for visual compression as the visual signals are also represented in the frequency domain (*e.g.*, DCT domain). The derivations of JND thresholds for each subband are usually based on a fixed size block (*e.g.*, 8×8) through a linear combination with other modulation factors. The Contrast Sensitive Function (CSF) plays a dominant role in visibility threshold derivation, reflecting the bandpass characteristics of HVS in the spatial frequency domain and is modeled as an exponential function of the spatial contrast [24]. In terms of modulation factors, LA and CM are commonly utilized. For example, they can be modeled as continuous functions, such as power function [25] or 'U shape' curve [26]. In [15], Wei and Ngan utilized a simple piecewise function to calculate LA and model the CM as a categorizing function according to the richness of block texture information. Besides LA and CM, some works also considered foveated masking as contributing factors for JND estimation, indicating that the sensitivity of the HVS is directly related to retinal eccentricity from the attention point. The foveated masking was firstly modeled by Chen *et al.* in [27] where it was employed as an explicit form in the pixel domain JND model and incorporated into the frequency domain in [17].

3) Machine learning based JND estimation: Recent years have witnessed a dramatically increased interest in machine learning technologies. As a result, incorporating machine learning into human perceptual oriented domain has become a new research trend. In the image quality assessment domain, outstanding performances have been achieved by utilizing deep neural networks to model the HVS response [28]–[30]. As for the JND, a learning based DCT domain JND model was proposed in [31] where a linear regression-based model and a Convolutional Neural Network (CNN) based model was established to learn a compensatory factor to deal with the effect caused by quantization in video compression. Besides, efforts have also been devoted to establishing the database for deriving the image/video compression oriented picture/video wise JND threshold. More specifically, the picture/video-wise JND is the maximum quantization parameter under which the difference caused by codec can be barely noticed by HVS. In [19], the authors firstly established a video-wise JND dataset based on HEVC in which the JND threshold is generated through rigorous subjective testings. Then a machine learning based approach was proposed to predict the JND distribution based on extracted video features. Liu *et al.* in [20] further proposed a deep learning based JND predictor based on the MCL-JCI dataset [32]. The predictor was trained to

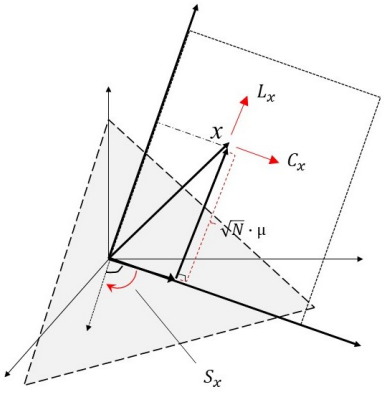

Fig. 1. Illustration of the derivation of visual factors for visibility masking estimation [33].

determine whether the picture coded by a certain quantization factor can be noticed by HVS.

B. Motivations

The underlying principle of most existing JND estimation approaches is using a linear or non-linear combination methodology to fuse the visual sensitivity estimation from multiple perspectives, *e.g.*, LA, CM and CSF. However, since the HVS is a nonlinear and highly complex system, how these mechanisms interact with each other has not been well understood. Moreover, the determination of the JND profile which highly depends on the local content relies on the local patch information. This motivates us to adopt the patch decomposition and fusion pipeline which represents an image patch with three independent components, including luminance, contrast and structure. The proposed patch level inference scheme benefits the derivation of the JND profile from multiple perspectives. First, contrary to most existing JND estimation methods, the proposed scheme adopts a decomposition and fusion pipeline, and the decomposed components are conceptually independent, such that the sensitivity for each component can be meaningfully derived and combined. Second, the proposed scheme handles the visibility derivation at the patch level, which is consistent with the image/video compression process, as well as the perception mechanism of visual information. Third, the derivation of barely visible distortion in terms of structural direction, which is the prominent component in determining the visual quality, could be feasibly achieved in a data-driven manner. As a result, the proposed approach naturally exploits the complementary roles of handcrafted and learning-driven near-threshold models in the JND profile inference.

III. The Proposed JND Model

In this section, we detail the proposed patch-wise structural visibility learning based JND model. We first describe the patch-level visual factor extraction framework, which benefits the estimation of visibility maskings as the adopted information lossless visual factor extraction approach relies on three independent and well established factors, including patch-wise luminance, contrast and structure. The first two with clear physical meanings can be represented by scalar values while the structure is the key factor that points to a specific direction and exhibits high sensitivity to noise. As such, visibility masking of each factor is feasible to be generated separately. Regarding the structure with large feasibility in visual masking estimation, a deep neural network (DNN) based model is established and trained, such that the structural sensitivity can be intelligently inferred. Finally, the JND profile can be generated by patch-wise combination and fusion of independent visibility maskings.

A. The Information Lossless Visual Factor Extraction

Inspired by the visual content representation in [33], we extract the visual factors through invertible patch decomposition. The process is illustrated in Fig. 1, where the $N \times N$ size patch x is represented as a directional vector in $N \times N$ dimension space. The patch x will be firstly orthogonally decomposed into two independent component vectors while one is always along the fixed diagonal direction and the other is highly content dependent:

$$x = \boldsymbol{\mu} + \|x - \boldsymbol{\mu}\| \cdot \frac{x - \boldsymbol{\mu}}{\|x - \boldsymbol{\mu}\|}, \tag{1}$$

where $\|\cdot\|$ denotes the ℓ_2 norm. Moreover, $\boldsymbol{\mu} = \mathbf{1} \times \mu$ denotes the first component vector along fixed direction $\mathbf{1}$ within which all coefficients equal to 1, and its length denoted as μ reveals the patch-level luminance intensity. In terms of the other component vector $x - \boldsymbol{\mu}$, it contains the texture information of x. In particular, $\|x - \boldsymbol{\mu}\|$ represents the patch-wise contrast information and the direction is the patch-wise structure which contains visually sensitive information of patch x. As such, three independent patch-wise visual factors: luminance L_x, contrast C_x and structure S_x are extracted by the information lossless decomposition $D(\cdot)$ simultaneously:

$$\{L_x, C_x, S_x\} = D(x), \tag{2}$$

where

$$L_x = \boldsymbol{\mu}; \quad C_x = \|x - \boldsymbol{\mu}\|; \quad S_x = \frac{x - \boldsymbol{\mu}}{\|x - \boldsymbol{\mu}\|}. \tag{3}$$

The main advantage of the information lossless visual factor extraction is that visibility maskings can be estimated separately and independently without any interactions and information loss. Moreover, it also enables the feasible JND profile generation by combining the visual maskings of these factors together.

B. The DNN Based Patch Structural Visibility Learning

The patch-wise structure is the dominant factor that influences the patch-wise visibility in practical applications (*e.g.,* visual content compression). However, estimating the structural visibility is an arduous task since the unit-length structure possesses an abundant degree of freedom. In the proposed patch-wise JND profile inference scheme, a DNN based model is established and trained to learn the structural visibility masking. In terms of the model training, a visual content compression oriented JND dataset is established, which will be

detailed in following Sec. IV. There are 202 pairs of pristine and corresponding picture-wise JND image pairs in the dataset, and the corresponding structural visibility maskings will be extracted and utilized for model training.

1) Model Architecture: The DNN-based structural learning model is illustrated in Fig. 2. In particular, the U-net [34] architecture is employed, which consists of the structural encoder and decoder. The whole architecture is fully convolutional and trained in an end-to-end manner. Regarding the encoder, it is stacked by cascaded convolutional layers and the residuals. Four strided convolutional layers are used to gradually reduce the image resolution and expand the channel size of features at the same time. Each convolutional layer is followed by a batch-normalization (BN) layer [35] and ReLU activation function. At the decoder side, the symmetric structure is used to up-sample the spatial resolution and reduce the size of channels. Additionally, we employ the long patch skip connection in order to remain the low-level visual features. Besides, to further improve performance, inspired by [36], five residual blocks with identical layout are connected and concatenated at the bottleneck. Each residual block consists of a convolutional layer with small 3×3 kernels and 256 channels followed by BN layer and ReLU activation. Short-path skip connection is employed to better refine features locally. In terms of the loss function, ℓ_1 Loss is employed in training since it is superior in terms of the perceptual information comparison [37].

2) Training Data: The 202 pristine and corresponding JND image pairs $p_1 - p_{202}$ in the dataset described in Sec. IV are utilized for producing patches in model training, validation and testing. To be specific, 162 pairs are cropped into 64×64 size patches without overlapping and used for model training, while the remaining 10 and 30 pairs are for validation and testing, respectively. Besides, the preprocessing of the training set is necessary since the model is established to learn the structural variations. First, we discard the plain patches by comparing C_x with a given threshold. In particular, the reference patch x will be considered a plain patch when satisfying,

$$C_x < 0.3 \times 10^3, \tag{4}$$

and such a pair would be discarded. Subsequently, during the training of the proposed patch-wise structural learning model, we avoid the effects caused by different luminance and contrast values. To be specific, assuming $\overline{x}$ is a patch from the image holding the just noticeable profile, the corresponding ground truth patch $\widehat{x}$ utilized for training is generated by modifying its patch luminance and contrast according to the reference patch x:

$$\widehat{x} = D^{-1}(L_x, C_x, S_{\overline{x}}). \tag{5}$$

Here, $D^{-1}(\cdot)$ denotes the inverse operation of the information lossless decomposition $D(\cdot)$ defined in Eqn. (2). It is worth mentioning that although in the normal training process, the luminance and contrast of patches would also be excluded by normalization and zero-mean in the data argumentation process, the preprocessing step in Eqn. (5) would be helpful in terms of model visualization. After preprocessing, around 75000 patch pairs are generated for model training.

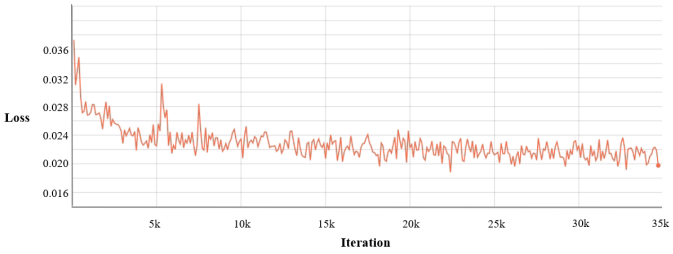

Fig. 3. Illustration of the training performance in terms of the training loss.

3) Training and Validation: Truncated normal initializer is employed in this work to initialize the network weights and Adam [38] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ is employed to optimize the objective function. The learning rate is initialized to 1×10^{-4}, which will linearly decay to 0 from the 50-*th* epoch. The mini-batch size is set to 200. We implement the networks in Pytorch 1.3.0 and Python 3.5.2. Then training is performed on the machine with NVIDIA GTX2080Ti GPU and memory 64G. The loss curve during training is demonstrated in Fig. 3, where x-axis and y-axis are the number of iterations and the training loss, respectively. As such, the predictor converges in the training process.

C. Visibility Masking Estimation of Luminance and Contrast

Regarding the visibility masking of patch luminance, it can be straightforwardly estimated based on the psychophysical HVS characteristics. Herein, the pixel-wise LA model which was firstly proposed in [7] is extended to the patch-level,

$$M_{L_x} = \begin{cases} 17 \times \left(1 - \sqrt{\frac{\mu}{127}}\right) + 3, & \text{if} \quad \mu < 127, \\ \frac{3 \times (\mu - 127)}{128} + 3, & \text{if} \quad \mu \geq 127. \end{cases} \tag{6}$$

Therefore, the HVS tolerable changed patch luminance $\widetilde{L_x}$ is computed by:

$$\widetilde{L_x} = L_x + r \times M_{L_x}, \tag{7}$$

where r is randomly set as 1 or -1 and adapted at the image level.

As for the visibility masking of patch contrast, inspired by [12], it is estimated by a divisive normalization strategy based on the assumption that corresponding visibility masking is determined by the normalized contrast rather than the absolute values. As such, the contrast masking can be modeled with a factor:

$$f_C = \frac{C_x - \Delta C_x}{C_x + \Delta C_x}, \tag{8}$$

where C_x and ΔC_x are the patch contrast and corresponding variation, respectively. An example is illustrated in Fig. 4, where the patch contrast is changed according to a given f_C value. It can be clearly observed that the changing of signal contrast leads to the change of the pixels' dispersion degree and the luminance distribution interval. Furthermore, the edges as well as the whole patch will be dimmed or sharpened as a result.

As such, the visibility masking of patch contrast is gauged by the determination of f_C. In this work, the factor f_C is

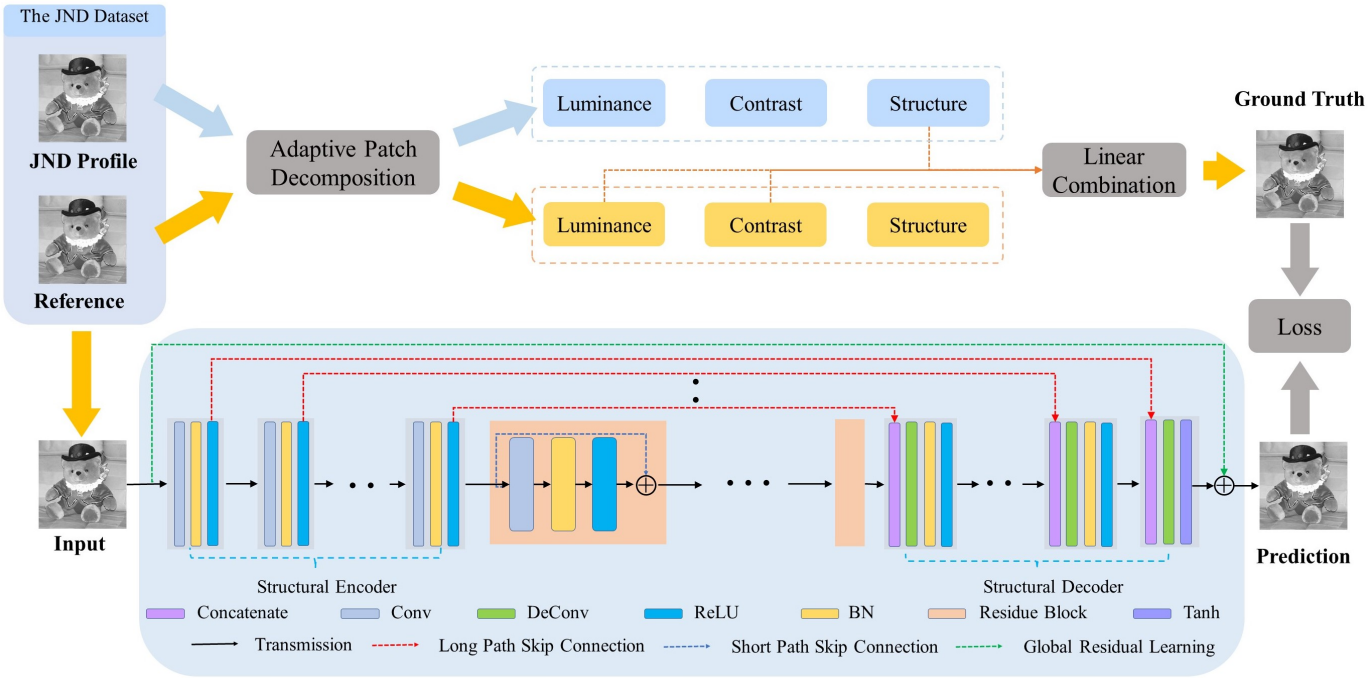

Fig. 2. Architecture of the DNN based structural visibility learning model.

(a) $f_C = 0.43$ (b) $f_C = 0.69$ (c) $f_C = 1$ (d) $f_C = 1.32$ (e) $f_C = 1.66$

Fig. 4. Illustration of changing patch contrast along the factor f_C.

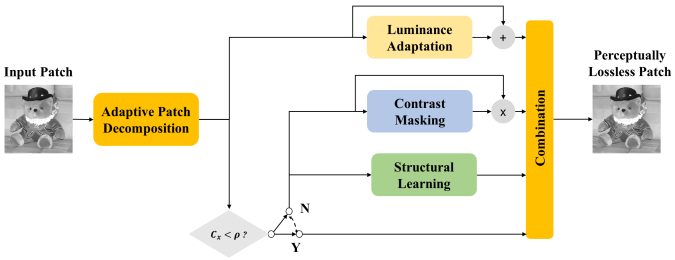

Fig. 5. Flowchart of the patch-wise JND profile generation.

empirically decided as 0.9. Subsequently, the perceptually lossless distorted patch contrast will be generated:

$$\widetilde{C_x} = C_x \times (\frac{2 \cdot f_C}{1 + f_C}). \tag{9}$$

It is worth mentioning that the $\widetilde{C_x}$ indicates the decreased contrast. Herein, we mainly consider the degradation of the contrast which is common in most signal processing and compression applications. Moreover, we differentiate the smooth and non-smooth patches, and a threshold ρ of patch contrast is set as a criterion to determine whether the current patch is plane or not. As such, the perceptually lossless distorted patch

Fig. 6. Examples of selected source images in our dataset.

contrast is modeled as a piece-wise function,

$$\widetilde{C_x} = \begin{cases} C_x, & \text{if} \quad C_x < \rho, \\ C_x \times (\frac{2 \cdot f_C}{1 + f_C}), & \text{if} \quad C_x \geq \rho, \end{cases} \tag{10}$$

where ρ is set to 0.3×10^3.

D. The Patch-wise JND Profile

The whole pipeline of patch-wise JND profile generation is illustrated in Fig. 5, which includes three stages. The first two stages adaptively decompose the patch and perform the estimation of visibility masking. Finally, the inferred JND patch can be generated by:

$$x_{JND} = D^{-1}(\widetilde{L_x}, \widetilde{C_x}, \widetilde{S_x}), \tag{11}$$

where $\widetilde{S_x}$ is the patch-wise structure from the DNN based structural learning model.

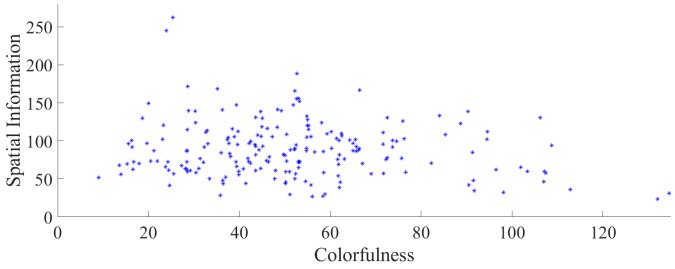

Fig. 7. The distribution of spatial information and colorfulness of source images in the dataset.

TABLE I
THE NUMBER OF SOURCE IMAGES IN EACH CATEGORY OF THE DATASET.

Content	Number
Outdoor	42
Indoor	20
Landscape	28
Nature	32
People	20
Objects	22
Building	18

IV. THE PROPOSED JND DATASET

A. Dataset Description

In our proposed JND database, both reference images and corresponding distorted versions are included. In particular, 202 lossless images are selected as reference images from the open-source RAISE database [39], which cover a wide variety of image contents, including building, outdoor scenes, objects, *et al.*, as shown in Table I. All the selected reference images are cropped to a uniform aspect ratio of 16:9 and then down-sampled to a fixed size in 1920×1080 using Lanczos interpolation method. Fig. 6 provides some representative samples of the reference images in the database. Moreover, the spatial information (SI) [40] and colorfulness (CF) [41] are calculated to quantitatively characterize the image content. For better visualization, Fig. 7 provides the SI and CF distributions of all 202 reference images in our JND database. One can observe that the distributions of SI and CF span the entire range with reasonable separations. Regarding distorted versions, they are generated by applying the video coding reference platform of VVC (VTM 5.0 [1]). For each reference image, the QP range [13,51] is empirically selected, which contains 39 levels of quantization parameters and covers the entire range of visual quality from imperceptible to severely annoying. As a result, 7878 distorted images are generated in our database.

B. Subjective Testing

As specified in ITU-R BT.500-13 [42], the single stimulus and double stimulus are the widely used subjective test approaches. The intention of our subjective test is to seek the just noticeable distortion level with the pristine image as the reference, such that the double stimulus approach is employed.

1) Subjective Testing Environment: The subjective testing was conducted in a controlled laboratory environment. The Samsung Q7F 55-inch smart TV (with a resolution of 3840×2160) was employed as the display device and the test software ran on a computer equipped with 16 GB RAM and 64-bit Windows operating system. The reference image and corresponding distorted image were randomly presented in a "side-by-side" fashion, and the invited subject was required to determine if any differences can be noticed between the two images, as shown in Fig. 8. Only one subject participated in the test at a time. In addition to setting the viewing distance according to [42], the subject was required to sit right in front of the monitor such that eyes and the middle of the monitor are maintained at the same height.

Fig. 8. A screenshot of the interface for conducting subjective evaluations.

2) Test Procedure: Each subject will be presented with an instruction prior to the formal test regarding how to perform these evaluations. In order to efficiently determine the appropriate QP point that exactly matches the JND profile, the binary search procedure described in [43] is utilized. More specifically, for one set which consists of one reference image and corresponding 39 distorted images, subjects will be asked to determine if any differences between the two images shown on the screen. It will take several turns to find the JND profile x_n for the n-*th* set by gradually shrinking the dynamic searching range $[x_l, x_r]$, and the searching details are summarized in Algorithm 1. The 202 reference images and corresponding 7878 distorted images were randomly presented and divided into 4 sessions without any overlapping for subjective testing. Each subject was assigned all 4 sessions for subjective evaluation and a break was required after each session. Twenty subjects with the age between 20 and 30 years old and normal vision participated in subjective tests, including 13 males and 7 females. Finally, we could obtain the final evaluation results from these 20 subjects.

C. Outlier Removal

As mentioned above, each image set in our JND database was evaluated by 20 individuals to obtain the corresponding raw opinion results. These results need to be further processed to generate the final JND profile. The obtained corresponding JND profile in each image set can thus be served as ground truth for training the proposed JND prediction model. To pursue the goal that the obtained results are as reliable as possible, the outlier detection methods outlined in [44] and [45] have been employed in our work.

TABLE II
EVALUATION OF THE PREDICTION ACCURACY.

Model	Training Set	Test Set	$PSNR^S_{pre}$ (dB)	$\Delta PSNR^S$ (dB)
G_1	$p_{41} - p_{202}$	$p_{21} - p_{40}$	38.62	2.67
G_2	$p_1 - p_{39}$ & $p_{81} - p_{202}$	$p_{61} - p_{80}$	38.24	2.85
G_3	$p_1 - p_{79}$ & $p_{121} - p_{202}$	$p_{101} - p_{120}$	39.08	2.03
G_4	$p_1 - p_{119}$ & $p_{161} - p_{202}$	$p_{141} - p_{160}$	37.83	3.62
G_5	$p_1 - p_{161}$	$p_{173} - p_{202}$	37.28	3.95

Algorithm 1 Testing procedure to determine the best QP that approaches the JND profile

Input:
A set of test images compressed with QP with in the range [13, 51] and the pristine image x_o.
Output:
The best QP that approaches the JND profile x_n.
Initialize: $x_l = 13, x_r = 51$
while flag **do**
  $x_c = (x_l + x_r)/2$;
  if x_o and x_c have quality difference **then**
    $x_r \Leftarrow x_c$;
    if x_r - $x_l < 2$ **then**
      flag $\Leftarrow$ false;
      $x_n \Leftarrow (x_l + x_r)/2$;
    end if
  else
    $x_l \Leftarrow x_c$;
    if x_r - $x_l < 2$ **then**
      flag $\Leftarrow$ false;
      $x_n \Leftarrow (x_l + x_r)/2$;
    end if
  end if
end while

The philosophy is that it is allowed to have a different understanding of the JND profile for each subject, however, the majority of the subjects should draw similar conclusions on the appropriate QP values. Therefore, the outlier detection is essential in identifying and rejecting outliers in the raw results. The statistical-based outlier rejection method suggested in [42] examines the consistency of raw results between a particular subject and all other subjects in a testing session. For ease of description, let Q_n^m be the raw results of m-th subject on n-th image set, where m=1,2,...,M and n=1,2,...,N. The Z-score of subject m can be obtained by,

$$Z_n^m = \frac{Q_n^m - \mu_n}{\sigma_n}, \tag{12}$$

where μ_n and σ_n denote the mean and standard deviation vector against all subjects for the n-th image set:

$$\mu_n = \frac{1}{M}\sum_{m=1}^{M} Q_n^m; \sigma_n = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(Q_n^m - \mu_n)^2}. \tag{13}$$

As a result, the Z-score of subject m can be defined as:

$$z^m = [z_1^m, z_2^m, z_3^m, ..., z_N^m]. \tag{14}$$

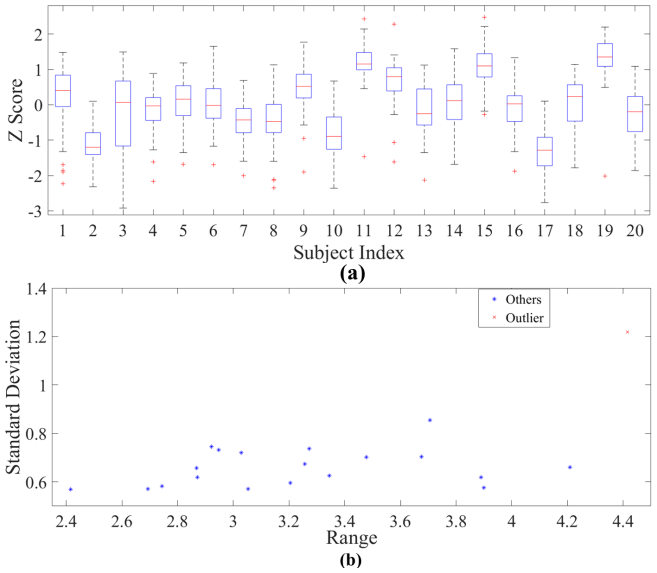

Fig. 9. Example of Z-vector based outlier detection, (a) the boxplot of Z-score in one session; (b) the dispersion plot of Z-vectors of all subjects in the same session; the sample marked as red cross is from subject #3.

The Z-score z^m indicates the consistency of m-th subject with all other subjects in the same test session. Subsequently, the outlier detection method proposed in [44] are strictly followed where the outlier can be identified according to the range R_m and standard derivation D_m:

$$R_m = max(z^m) - min(z^m), \quad D_m = std(z^m). \tag{15}$$

As shown in Fig. 9, since both R_3 and D_3 are larger, the subject #3 is determined to be an outlier and discarded.

After removing the samples from outliers, it is necessary to further check whether distribution of the post-processed samples follow the Gaussian distribution. The kurtosis based validation recommend in [45] is adopted. Regarding the n-th image set, the kurtosis is measured by,

$$\beta = \frac{\rho_{n,4}}{(\rho_{n,2})^2}, \tag{16}$$

where

$$\rho_{n,t} = \frac{\sum_{m=1}^{M'}(Q_n^m - \gamma_n)^t}{M'}, t = 2, 4, \tag{17}$$

$$\gamma_n = \frac{1}{M'}\sum_{m=1}^{M'} Q_n^m. \tag{18}$$

Here M' denotes the number of subjects after removing the outliers in a certain session. The distribution of the scores can

be regarded as normally distributed if the computed kurtosis value β falls between 2 and 4 [42]. All sets pass the normality verification after removing the outliers in our dataset.

Finally, the QP point that leads to the JND profile for the n-th image is determined as follows,

$$Q_n = \sum_{m=1}^{M'} Q_n^m. \tag{19}$$

V. Experimental Results

A. Evaluation of Structural Visibility Learning

In the subsection, we first evaluate the proposed structural visibility learning model. For cross validation, we split the dataset into training and testing image pairs in five ways and the corresponding trained models are denoted as G_1-G_5, as shown in Table II. In order to avoid blocking artifacts, we allow the processed patches to have overlaps. To be specific, when generating the testing patches, there would be 10-pixel-width overlap on the top and left of the current patch. Subsequently, when generating the final predicted images, linear interpolation is employed in overlapped regions. Besides, to better demonstrate the performance of the structural visibility learning model, the image-level structural learning accuracy is evaluated. To be specific, the luminance and contrast factors of the predicted images and JND images will be firstly unified according to the original images. We denote the PSNR between the original image and the predicted image as $PSNR_{pre}^S$, and the PSNR between the original image and the ground truth JND image as $PSNR_{JND}^S$. The structural learning accuracy $\Delta PSNR^S$ is given by,

$$\Delta PSNR^S = |PSNR_{pre}^S - PSNR_{JND}^S|, \tag{20}$$

where $|\cdot|$ denotes the absolute value and the results are shown in Table II.

Promising results are achieved where the predicted structure preserves the original structure with the $PSNR_{pre}^S$ to be over 37dB on average, and the average structural learning accuracy is around 3dB. An example is shown Fig. 10, where the $PSNR_{pre}^S$ and $PSNR_{JND}^S$ are 35.39dB and 37.10dB, respectively. And almost no perceptual difference can be observed in the comparison.

B. Evaluation of the Proposed JND Model

1) Experimental Settings: Extensive experiments are carried out to evaluate the performance of the proposed JND model. In particular, we adopt 19 pristine images which have no overlap in the JND dataset described in Sec. IV. The captured scenes of these testing images include nature, people, building and indoor, as shown in Fig. 11. Four existing pixel-wise JND models are employed to validate the proposed JND model's performance, including Yang *et al.*'s model [46], Liu *et al.*'s model [4] and Wu *et al.*'s models [10], [11].

2) Objective Evaluation of Distortion Masking: First, the error tolerance capacities of JND models are evaluated. In terms of the four pixel-wise compared JND models, the corresponding distorted images are produced by directly injecting the JND noise into pristine images. From this perspective, PSNR between the pristine images and JND images are regarded as the evaluation criterion. The results are shown in Table III.

TABLE III
Comparison of JND models in terms of PSNR.

Index	Yang *et al.* [46]	Liu *et al.* [4]	Wu *et al.* [10]	Wu *et al.* [11]	Proposed
1	32.59	26.95	43.92	30.85	28.84
2	30.42	29.31	34.68	30.53	31.76
3	33.33	31.94	33.35	29.74	32.97
4	31.57	28.93	46.49	32.05	29.69
5	29.41	30.75	29.14	26.43	32.20
6	30.44	30.18	30.74	28.65	33.11
7	29.83	29.95	45.87	34.23	31.50
8	33.50	32.49	34.43	30.11	32.87
9	29.96	31.14	39.24	32.30	32.67
10	28.90	29.05	40.57	32.58	30.90
11	34.31	28.41	28.21	26.55	31.61
12	27.81	29.40	41.23	32.90	33.21
13	31.41	32.40	33.21	29.86	33.65
14	30.13	29.76	41.72	33.41	30.40
15	28.32	28.32	33.40	29.18	29.75
16	32.85	32.85	35.22	30.92	33.88
17	27.57	27.57	40.11	31.60	31.16
18	31.24	31.24	43.00	32.37	32.36
19	29.63	29.63	39.80	32.33	31.07
Average	30.02	30.78	38.12	30.87	31.77

TABLE IV
Scoring criteria for the subjective test.

Description	Score
The right one is much worse that the left one	-3
The right one is worse that the left one	-2
The right one is slightly worse that the left one	-1
The right one has the same quality as the left one	0
The right one is slightly better that the left one	1
The right one is better that the left one	2
The right one is much better that the left one	3

3) Subjective testing: Subjective testing is also conducted to compare the perceptual quality. Before the subjective testing, the distorted images generated from the anchor model $\widehat{I}$ is first adjusted in order to maintain the same objective distortion

Fig. 11. Images for JND models comparison.

(a) **(b)**

Fig. 10. Comparisons between the JND image (a) and predicted image (b) from structural visibility learning.

TABLE V
QUALITY COMPARISON THROUGH SUBJECTIVE TESTING.

	vs. Yang *et al.* [46]		*vs.* Liu *et al.* [4]		*vs.* Wu *et al.* [10]		*vs.* Wu *et al.* [11]		*vs.* Pristine Image	
Index	Ave	Std	Ave	Std	Ave	Std	Ave	Std	Ave	Std
1	1.00	0.489	0.90	0.444	1.70	0.489	0.85	0.410	-0.35	0.489
2	0.40	0.410	0.50	0.470	0.86	0.410	0.30	0.552	-0.80	0.410
3	0.85	0.307	0.80	0.223	1.15	0.444	0.50	0.444	-0.25	0.443
4	0.60	0.510	0.75	0.410	1.26	0.510	0.45	0.447	-0.45	0.510
5	0.55	0.447	0.50	0.366	1.00	0.489	0.45	0.410	-0.35	0.410
6	0.30	0.410	0.15	0.489	0.88	0.513	0.20	0.470	-0.50	0.513
7	0.55	0.550	0.50	0.470	1.49	0.410	0.45	0.489	-0.20	0.410
8	0.80	0.410	0.15	0.410	1.05	0.502	0.10	0.513	-0.40	0.502
9	0.40	0.324	0.50	0.307	0.85	0.502	-0.20	0.502	-0.60	0.502
10	0.55	0.366	0.80	0.410	1.32	0.470	0.25	0.510	-0.30	0.470
11	0.35	0.410	0.50	0.513	0.70	0.510	0.20	0.489	-0.45	0.510
12	-0.15	0.444	-0.05	0.587	0.70	0.307	-0.45	0.510	-0.90	0.307
13	0.30	0.552	0.45	0.550	0.62	0.410	-0.30	0.513	-0.80	0.410
14	0.40	0.458	0.60	0.410	0.90	0.502	0.60	0.410	-0.60	0.502
15	0.60	0.587	0.85	0.502	0.99	0.510	0.60	0.366	-0.55	0.510
16	0.15	0.513	0.30	0.489	1.05	0.489	0.40	0.443	-0.35	0.489
17	0.30	0.502	0.30	0.502	0.30	0.470	0.40	0.470	-0.30	0.472
18	0.55	0.447	0.45	0.458	1.20	0.510	0.65	0.410	-0.55	0.510
19	0.25	0.510	0.15	0.510	1.15	0.470	0.50	0.410	-0.30	0.470
Average	0.460	0.455	0.494	0.448	1.064	0.469	0.310	0.461	-0.473	0.469

level with ours in terms of PSNR,

$$\widehat{I} = I + \varepsilon \times r \times \tau(I), \tag{21}$$

where ε is the corresponding scaling factor, r is the pixel-wise random controller which takes -1 or 1 randomly, and $\tau(I)$ is the JND map generated by the corresponding method. Twenty subjects who are different from the subjects involved in the creation of the dataset in Sec. IV participate in the testing. The testing environment is same as that descried in Sec. IV. The scoring criteria in the subjective experiments are shown in Table IV. After specific instructions and training sessions, subjects are asked to offer their opinions on the subjective quality of the images following the rating criteria. Specifically, image pairs are randomly showed in the parallel way on the same screen. The subjective scores are finally converted with the unified order that the right is the original image as the reference and the left is the JND-injected version. Subsequently, the difference between the compared method and the proposed method is calculated. Each pair needs to be repeatedly scored twice by each subject. The average values across different subjects are calculated to demonstrate the visual quality comparison, which is illustrated in Table V. It is worth mentioning that the positive numbers denote that our proposed model is superior, and vice versa.

4) Performance analysis: Regarding distortion masking demonstrated in Table III, we can see that the proposed JND generation profile is at a similar level compared to other state-of-the-art pixel-wise JND models. Moreover, the subjective results exhibited in Table V have demonstrated that the proposed JND profile is superior to the conventional JND models. In general, there are several reasons behind this. First, the patch decomposition method provides an analytical representation of visual information by independent visual factors, which means that all the corresponding visibility maskings

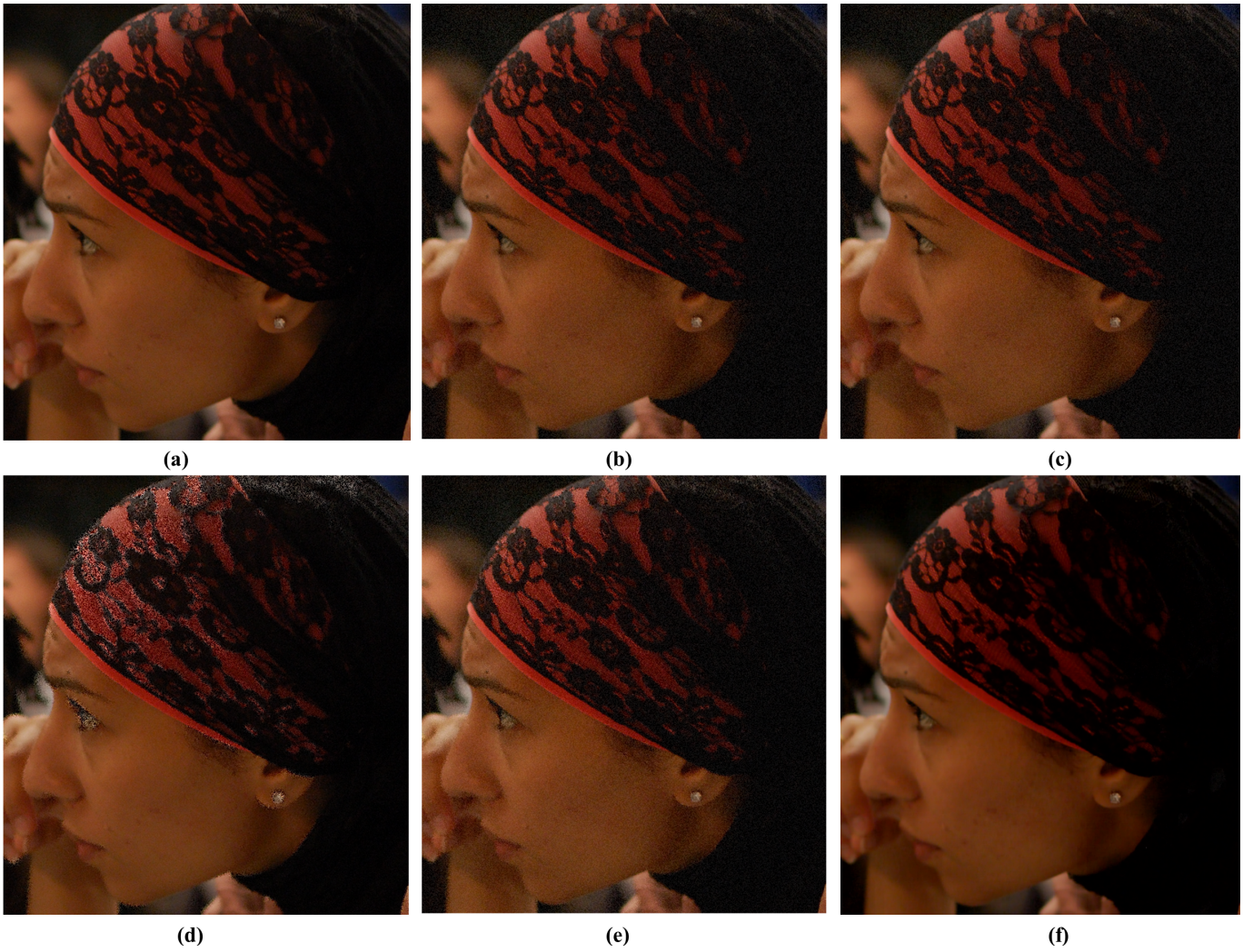

Fig. 12. Visual comparison of different JND models. (a) pristine image, (b) Yang *et al.* [46], (c) Liu *et al.* [4], (d) Wu *et al.* [10], (e) Wu *et al.* [11], (f) the proposed method. The five JND distorted images have the same noise level (PSNR= 28.84dB).

are involved in JND profile generation. This leads to the proposed JND model being able to tolerate more distortions and achieve better subjective performance while maintaining the same PSNR comparing with the conventional models. Second, estimating the JND profile at the patch-level instead of pixel/sub-band level aligns with the HVS mechanism. Third, the learning based approach improves the model's applicability which means the proposed model is able to achieve favorable performances in patches of different contents. These lead to accurate JND estimation.

In particular, an intuitive comparison is illustrated in Fig. 12, where Fig. 12(a) is the pristine image, and Fig. 12(b)-(f) are JND distorted images of different models at the same objective quality in terms of PSNR. Advantages of proposed patch-wise JND model against conventional pixel-wise models can be observed. By estimating the structural variation instead of estimating visibility masking of each pixel separately, the white noise like distortion especially in plain and background regions (*e.g.*, face and background regions in Fig. 12(b) and (c)) are avoided in proposed model. Moreover, the conventional JND models share the same methodology that assigns more distortions into HVS non-sensitive regions. However, visual features extracted at pixel-level can't always provide reasonable and accurate modeling of HVS mechanism. Visual features extracted from pixel-level may have conflict with the global features within the actual HVS perceive region. As a result, JND thresholds will be overestimated in these regions. An example can be observed in the eye region in Fig. 12. From the pixel-level perspective, the eye region is considered as high-contrast and irregular region because of the sharp edges around the eye and the disordered eyelashes. As a result, more distortions will be assigned to this region. However, from the perspective of the whole region, the classification is not accurate since this region has few structural uncertainties. Besides, this region is also a saliency region in terms of human perception. As such, corresponding JND threshold are overestimated as shown in Fig. 12(d) and (e). By extracting features from patch-level, the proposed model provides accurate and reasonable understanding of visual information. The proposed learning-based model ensures all predicted patches are perceptually lossless by strictly mining the training dataset. Moreover, in terms of the distortion structure, as shown in

TABLE VI
EVALUATION OF THE PROPOSED PERCEPTUALLY LOSSLESS CODING.

Index	QP25 (bpp)	QP27 (bpp)	QP29 (bpp)	Proposed (bpp)
1	0.21	0.12	0.09	0.06
2	0.55	0.45	0.36	0.24
3	0.61	0.52	0.44	0.28
4	0.19	0.10	0.08	0.06
5	1.14	0.98	0.82	0.54
6	1.28	1.11	0.94	0.79
7	0.09	0.06	0.05	0.03
8	0.51	0.41	0.32	0.20
9	0.29	0.23	0.18	0.13
10	0.24	0.17	0.12	0.08
11	1.72	1.54	1.36	0.94
12	0.23	0.15	0.10	0.05
13	0.61	0.51	0.41	0.26
14	0.16	0.07	0.05	0.03
15	0.58	0.47	0.37	0.24
16	0.69	0.51	0.37	0.23
17	0.20	0.16	0.13	0.09
18	0.22	0.18	0.14	0.10
19	0.35	0.24	0.17	0.10
Average	0.52	0.42	0.34	0.23
bpp saving	**55.7%**	**45.2%**	**32.3%**	-
Subjective Score	**44.2**	**48.9**	**52.1**	-

Fig. 12(f), we can see the distortions from the proposed model are quite similar with quantization caused noise. This ensures that the proposed JND model can provide reliable and accurate guidance for PVC scheme that incorporates the JND model into intra coding.

C. Perceptually Lossless Coding with the Proposed JND Profile

In the literature, JND models are usually employed in image/video coding frameworks to improve coding efficiency by removing perceptual redundancy of the visual content. In this subsection, we further examine the perceptually lossless coding by incorporating the proposed JND profile into intra coding. More specifically, in this framework, the patch-wise JND profile would provide guidance for CTU-level coding parameter assignment. As mentioned in Sec. IV, in image/video coding processing, the content variations noticed by HVS are mostly caused by structural change meanwhile changes along luminance and contrast are far below corresponding visibility maskings. Here, we denote x as the original CTU with size 64×64, S_x as the corresponding structure and $\widetilde{S_x}$ is the HVS tolerated structural produced by the proposed structural visibility learning model. As such, we aim to find the optimal QP denoted by q,

$$\text{argmax } q, \quad \text{subject to}: |S_x - S_{x_q}| \leq |S_x - \widetilde{S_x}|, \qquad (22)$$

where x_q is the compressed CTU with QP equals to q, S_{x_q} is the corresponding structure and $|\cdot|$ denotes ℓ_1 distance. We adopt the multi-pass coding strategy proposed by [47] to approach the best q with a binary searching method at the CTU level. As for plane CTUs, corresponding QPs can be directly inherited from nearby CTUs. The framework is implemented on VTM 5.0 intra. As for the anchors, the VTM 5.0 intra coding with constant frame-level QPs of 25, 27 and 29 are employed. In order to provide more comprehensive demonstrations, we also conduct the two-alternative-forced-choice (2AFC) based subjective test to compare the perceptual qualities of different methods. In this experiment, image pairs from our coding framework and the employed anchors are shown side-by-side with a random placement. Twenty subjects are asked to choose one with better quality. Specifically, we aim at examining whether the proposed perceptually lossless coding framework can increase the coding efficiency in comparison with VTM while keeping almost the same subjective quality level.

The results are shown in Table VI in terms of bits per pixel (bpp) and subjective score which demonstrates the percentage regarding the favor of the images coded by the proposed method. The results show that the patch-wise JND profile based perceptually lossless coding framework can achieve about 32.3%- 45.2% bit-rate savings while keeping almost the same subjective quality level comparing with VTM 5.0 intra. This further provides evidences that the proposed scheme has great potentials in the perceptual image coding.

VI. CONCLUSIONS

We have proposed a JND model that is particularly designed to estimate visibility masking at the patch level. The novelty of the model lies in representing visual contents with a meaningful combination of three visual factors including luminance, contrast and structure. Besides, another key contribution in our work is that an intra coding oriented JND database is established and served as the ground truth for learning the DNN based structural visibility learning model. Extensive experiments conducted over conventional pixel-wise JND models and our proposed JND profile have clearly demonstrated that the proposed JND profile delivers the highest performance in terms of JND estimation accuracy in accordance with what the HVS perceives. The JND model is further incorporated into VVC intra to provide a perceptually lossless coding scheme, and experimental results have shown that the proposed model can significantly improve the coding efficiency without sacrificing the perceptual quality.

REFERENCES

[1] B. Bross, J. Chen, and S. Liu, "Versatile Video Coding (Draft 5)," *JVET N1001*, 2019.

[2] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec 2012.

[3] J. Nikil, J. James, and S. Robert, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.

[4] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, "Just noticeable difference for images with decomposition model for separating edge and textured regions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1648–1652, 2010.

[5] X. Zhang, S. Wang, G. Ke, W. Lin, S. Ma, and W. Gao, "Just-noticeable difference-based perceptual optimization for JPEG compression," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 96–100, 2016.

[6] C. Mak and K. Ngan, "Enhancing compression rate by just-noticeable distortion model for H. 264/AVC," in *2009 IEEE International Symposium on Circuits and Systems*. IEEE, 2009, pp. 609–612.

[7] C. Chouand and Y. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Transactions on circuits and systems for video technology*, vol. 5, no. 6, pp. 467–476, 1995.

[8] X. Yang, W. Ling, Z. Lu, E. Ong, and S. Yao, "Just noticeable distortion model and its applications in video coding," *Signal Processing Image Communication*, vol. 20, no. 7, pp. 662–680, 2005.

[9] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, , no. 6, pp. 679–698, 1986.

[10] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi, "Just noticeable difference estimation for images with free-energy principle," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1705–1710, 2013.

[11] J. Wu, L. Li, W. Dong, G. Shi, W. Lin, and Kuo. C-C Jay, "Enhanced just noticeable difference model for images with pattern complexity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2682–2693, 2017.

[12] S. Wang, L. Ma, Y. Fang, Lin W, S. Ma, and W. Gao, "Just noticeable difference estimation for screen content images," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3838–3851, 2016.

[13] Y. Jia, W. Linand, and A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 7, pp. 820–829, 2006.

[14] X. Zhang, W. Lin, and P. Xue, "Just-noticeable difference estimation with pixels in images," *Journal of Visual Communication and Image Representation*, vol. 19, no. 1, pp. 30–41, 2008.

[15] Z. Wei and K. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 337–346, 2009.

[16] S. Bae and M. Kim, "A novel generalized DCT-based jnd profile based on an elaborate CM-JND model for variable block-sized transforms in monochrome images," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3227–3240, 2014.

[17] S. Bae and M. Kim, "A DCT-based total JND profile for spatiotemporal and foveated masking effects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1196–1207, 2016.

[18] S. Ki, S. Bae, M. Kim, and H. Ko, "Learning based just-noticeable-quantization-distortion modeling for perceptual video coding," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3178–3193, 2018.

[19] Q. Huang, H. Wang, S. Lim, H. Kim, S. Jeong, and C-C Jay Kuo, "Measure and prediction of HEVC perceptually lossy/lossless boundary QP values," in *2017 Data Compression Conference (DCC)*. IEEE, 2017, pp. 42–51.

[20] H. Liu, Y. Zhang, H. Zhang, C. Fan, S. Kwong, C-C Jay Kuo, and X. Fan, "Deep learning-based picture-wise just noticeable distortion prediction model for image compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 641–656, 2019.

[21] J. Wu, G. Shi, and W. Lin, "Survey of visual just noticeable difference estimation," *Frontiers of Computer Science*, vol. 13, no. 1, pp. 4–15, 2019.

[22] J. Tim, C. Mark, R. Hermann D. Jonathan B L. Stephen M, K. William L, and S. Joshua H, "A synaptic mechanism for retinal adaptation to luminance and contrast," *Journal of Neuroscience*, vol. 31, no. 30, pp. 11003–11015, 2011.

[23] M. Jourlin, Ma. Carré, J. Breugnot, and M. Bouabdellah, "Logarithmic image processing: additive contrast, multiplicative contrast, and associated metrics," in *Advances in Imaging and Electron Physics*, vol. 171, pp. 357–406. Elsevier, 2012.

[24] J. Ahumada, J. Albert J, and H. Peterson, "Luminance-model-based DCT quantization for color image compression," *Human vision, visual processing, and digital display III*, vol. 1666, pp. 365–374, 1992.

[25] B. Andrew Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," *Sid International Symposium Digest of Technical Papers*, vol. 24, pp. 946–946, 1993.

[26] X. Zhang, W. Lin, and P. Xue, "Improved estimation for just-noticeable visual distortion," *Signal Processing*, vol. 85, no. 4, pp. 795–808, 2005.

[27] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 6, pp. 806–819, 2010.

[28] K. Woojae, A. Nguyen, S. Lee, and A. Bovik, "Dynamic receptive field generation for full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4219–4231, 2020.

[29] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[30] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1676–1684.

[31] S. Ki, S. Bae, M. Kim, and H. Ko, "Learning-based just-noticeable-quantization-distortion modeling for perceptual video coding," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3178–3193, 2018.

[32] JY. Lin L. Jin, P. Wang I. Katsavounidis A. Aaron S. Hu, H. Wang, and C-C Jay Kuo, "Statistical study on perceived jpeg image quality via mcl-jci dataset construction and analysis," *Electronic Imaging*, vol. 2016, no. 13, pp. 1–9, 2016.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[36] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[37] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2599–2613, 2018.

[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] D. Nguyen, D. Tien, P. Cecilia, C. Valentina, and B. Giulia, "RAISE: A raw images dataset for digital image forensics," in *Proceedings of the 6th ACM Multimedia Systems Conference*. ACM, 2015, pp. 219–224.

[40] "Subjective video quality assessment methods for multimedia applications," *ITU Telecom. Standardization Sector of ITU*, 1999.

[41] W. Stefan, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[42] "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 500–13, 2012.

[43] H. Wang, C-C Jay, et al., "MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1509–1513.

[44] H. Wang, I. Katsavounidis, J. Zhou, P. Jeonghoon, S. Lei, X. Zhou, M. Pun, X. Jin, R. Wang, X. Wang, et al., "Videoset: A large-scale compressed video quality dataset based on jnd measurement," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 292–302, 2017.

[45] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K. Ma, "Esim: Edge similarity for screen content image quality assessment," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4818–4831, 2017.

[46] X. Yang, W. Lin, Z. Lu, E. On, and S. Yao, "Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 6, pp. 742–752, 2005.

[47] Y. Li, H. Liu, and Z. Chen, "Perceptually-lossless image coding based on foveated-JND and H. 265/HEVC," *Journal of Visual Communication and Image Representation*, vol. 40, pp. 600–610, 2016.

Xuelin Shen received the B.E. and M.S. degrees in the School of Data and Computer Science from Sun Yat-Sen University, China in 2014 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, China (Hong Kong SAR). His research interests include video coding, computer vision and deep learning.

Zhangkai Ni received the M.E. degree in communication engineering from the School of Information Science and Engineering, Huaqiao University, Xiamen, China, in 2017. He was a Research Engineer with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2017 to 2018. He is currently a Ph. D. candidate with the Department of Computer Science, City University of Hong Kong, Hong Kong. His current research interests include computer vision, image processing, unsupervised learning, and quality assessment.

Wenhan Yang (M'18) received the B.S degree and Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, in 2012 and 2018. He is currently a postdoctoral research fellow with the Department of Computer Science, City University of Hong Kong. Dr. Yang was a Visiting Scholar with the National University of Singapore, from Sep. 2015 to Sep. 2016 and from Sep. 2018 to Nov. 2018. His current research interests include deep-learning based image processing, bad weather restoration, related applications and theories.

Xinfeng Zhang (M'16) received the B.S. degree in computer science from the Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. From 2014 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore. From Oct. 2017 to Oct. 2018, he was a Post-Doctoral Fellow with the School of Electrical Engineering System, University of Southern California, Los Angeles, CA, USA. From Dec. 2018 to Aug. 2019, he was a Research Fellow with the department of Computer Science, City University of Hong Kong.

He currently is an Assistant Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. He authored more than 100 refereed journal/conference papers and received the Best Paper Award of IEEE Multimedia 2018, the Best Paper Award at the 2017 Pacific-Rim Conference on Multimedia (PCM) and the Best Student Paper Award in IEEE International Conference on Image Processing 2018. His research interests include video compression, image/video quality assessment, and image/video analysis.

Shiqi Wang (M'15) received the B.S. degree in computer science from the Harbin Institute of Technology in 2008 and the Ph.D. degree in computer application technology from Peking University in 2014. From 2014 to 2016, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From 2016 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong, China (Hong Kong SAR). He has authored over 40 technical proposals in ISO/MPEG, ITU-T, and AVS standards. His research interests include image/video compression, analysis, and quality assessment.

Sam Kwong (F13) received the B.S. and M.S. degrees in electrical engineering from State University of New York at Buffalo in 1983, University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from University of Hagen, Germany, in 1996. From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada. He joined Bell Northern Research Canada as a Member of Scientific Staff. In 1990, he became a Lecturer at the Department of Electronic Engineering, City University of Hong Kong, where he is currently a Chair Professor of the Department of Computer Science, City University of Hong Kong, Kowloon, China (Hong Kong SAR). His research interests are video and image coding and evolutionary algorithms.