Visual Attention Guided Pixel-Wise Just Noticeable Difference Model

ZHIPENG ZENG1,2, HUANQIANG ZENG1,2, (Senior Member, IEEE), JING CHEN1,2, (Member, IEEE), JIANQING ZHU3, (Member, IEEE), YUN ZHANG4, (Senior Member, IEEE), KAI-KUANG MA5 (Fellow, IEEE)

1School of Information Science and Engineering, Huaqiao University, Xiamen 361021, China
2Xiamen Key Laboratory of Mobile Multimedia Communications, Xiamen 361021, China
3College of Engineering, Huaqiao University, Quanzhou 362021, China
4Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China
5School of Electrical and Electronic Engineering, Nanyang Technological University, 639798 Singapore

Corresponding author: Huanqiang Zeng (e-mail: zeng0043@hqu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under the grants 61871434, 61802136, and 61602191, in part by the Natural Science Foundation for Outstanding Young Scholars of Fujian Province under the grant 2019J06017, in part by the Natural Science Foundation of Fujian Province under the grant 2017J05103, in part by the Fujian-100 Talented People Program, in part by the High-level Talent Innovation Program of Quanzhou City under the grant 2017G027, in part by the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University under the grants ZQN-YX403 and ZQN-PY418, and in part by the High-Level Talent Project Foundation of Huaqiao University under the grants 14BS201, 14BS204 and 16BS108.

ABSTRACT The just noticeable difference (JND) models in pixel domain are generally composed of luminance adaptation (LA) and contrast masking (CM), which takes edge masking (EM) and texture masking (TM) into consideration. However, in existing pixel-wise JND models, CM is not evaluated appropriately since they overestimate the masking effect of regular oriented texture regions and neglect the visual attention characteristic of human eyes for the real image. In this work, a novel JND model in pixel domain is proposed, where orderly texture masking (OTM) for regular texture areas (also called orderly texture regions) and disorderly texture masking (DTM) for complex texture areas (also called disorderly texture regions) are presented based on the orientation complexity. Meanwhile, the visual saliency is set as the weighting factor and is incorporated into CM evaluation to enhance JND thresholds. Experimental results indicate that compared with existing relevant JND profiles, the proposed JND model tolerates more distortion in the same perceptual quality, and brings better visual perception in the same level of the injected JND-noise energy.

INDEX TERMS Just noticeable difference, orientation complexity, visual attention.

I. INTRODUCTION Images/Videos are commonly explored in various multimedia services and become an indispensable part in people’s daily life. To provide a high quality of multimedia experience, there are many researches devoting to the development of image/video processing, image/video coding, and robust transmission technologies. Since human eyes are the ultimate receivers of images/videos in general, how to describe perceptual characteristics of human vision more precisely and efficiently has been drawing lots of attentions from both academic and industrial societies [1]–[4].

As is known, an important perceptual characteristic of human visual system (HVS) is that it presents limited visual sensitivity to the images/videos, only the pixel changes above a certain visibility threshold can be observed by human eyes [1]. To model this perceptual characteristic, the just noticeable difference (JND) model has been presented, in which the smallest perceptual visual threshold values of the human eyes for the input image are obtained [5], [6]. Therefore, the JND models are widely applied on variable kinds of perceptual-oriented image/video related tasks, such as perceptual compression [7]–[9], perceptual quality assessment [10], [11], watermarking [12], display [13], to name a few.

Existing JND models can be roughly classified into two categories according to the JND threshold calculating domain: the pixel-wise JND models (e.g., [1], [14], [15]) and the subband-based (e.g., DCT or wavelet transform)
JND models (e.g., [16]–[18]). Compared to the subband-domain JND models, the pixel-wise ones can be calculated directly and avoid the subband transformation, which would be more convenient and cost-effective to estimate the JND thresholds. Base on that, the objective of this work is to design an effective pixel-wise JND model to accurately describe characteristics of HVS on images. Pixel-wise JND models commonly take the luminance adaptation (LA) and contrast masking (CM) into account. Note that LA reflects the masking effect of the HVS in respect of the luminance of the background, while CM reflects the visibility attenuation of one contrast at the presence of another contrast. Some early JND models, like the one developed by Chou et al. [19] overlooked the interaction between these two masking effects, resulted in a rough JND estimation. Based on Chou et al. [19], Yang et al. [20] exploited a nonlinear additivity model to reduce the overlapping effects between LA and CM. Since these two methods overestimated the masking effects in the edge regions, Liu et al. [15] decomposed one input image into two images, one is named structural image and the other is the textural image, followed by performing edge masking (EM) estimation and texture masking (TM) estimation, respectively. Considering that the CM effect is not comprehensively evaluated, Wu et al. [21] proposed the disorderly concealment effect based on free-energy principle for JND estimation. Motivated by the observation that the HVS is highly sensitive to the repeated pattern in visual signal, Wu et al. [1] introduced the concept of pattern complexity to decide the total masking effects. With image saliency information, Hadizadeh et al. [22] developed a saliency-guided JND model by the normalized Laplacian pyramid.

According to the research of cognitive psychology and neuroscience, HVS is usually motivated to fetch the visual regularities for perception and understanding [23], [24]. And in the local receptive field of the image, the visual cortex displays distinct orientation selectivity mechanism for visual content representation and extraction [25], [26], which also indicates that orientation regularity (also called low orientation complexity) plays a significant role in the process of visual perception. Inspired by these, in our JND model, the textural image proposed by Liu et al. [15] is further decomposed into two portions according to the regularity of the texture. For regular texture regions, an orderly texture masking (OTM) is exploited; for disorderly textural portions, the disorderly texture masking (DTM) is used. Furthermore, based on the visual attention mechanism, the higher the visual saliency the higher priority of being processed by HVS, and people’s eyes will focus on the saliency areas for a relatively long time. Therefore, the visual saliency regarded as the adjustment factor is incorporated into the proposed CM estimation. Combined with luminance adaptation (LA), the proposed JND model is established. Experimental results show that the proposed JND model is well correlated with the HVS perception and outperforms the relevant pixel-wise JND methods.

The remaining sections of this paper are arranged as follows. In Section II, the proposed JND model is presented in detail. Section III provides the experimental results and analyses. The conclusion is summarized in Section IV.

II. PROPOSED JND MODEL

As illustrated in Fig. 1, the main body of the proposed pixel-wise JND model consists of three modules, namely LA, $CM_s$ and $NAMM$, where LA and $NAMM$ modules are referred to [14], while $CM_s$ module encircled with the red dashed line is the contribution of this work. In order to predict contrast masking (CM) precisely, we estimate edge masking (EM), orderly texture masking (OTM) and disorderly texture masking (DTM) from structural image $u$, orderly textural image $v_o$ and disorderly textural image $v_d$, respectively, instead of obtaining CM estimation from the whole image $f$ at first hand. Meanwhile, for the sake of visual attention mechanism of HVS [27], [28], the bottom-up saliency model [29] for non-local spatial redundancy is treated as the weight coefficient to adjust the CM values perceptually.

A. THE CONTRAST MASKING MODEL BASED ON VISUAL ATTENTION

As [30] denotes, contrast masking represents the visibility reduction of one visual component at the presence of another. Based on the visual attention of image/video content, the sensitivity of HVS is diverse in different image areas for CM evaluation. The CM presented by Liu et al. [15] is composed of edge masking (EM) for edge regions and texture masking (TM) for textural areas, respectively. However, TM overestimates the masking effect of homogeneous textural regions. More precise estimation is needed adapting to regular oriented texture region and homogeneous one. Therefore, there are...
two textural masking estimation in the proposed CM module. One is for regular oriented texture region, named orderly texture masking (OTM); the other is for complex texture region, named disorderly texture masking (DTM). With OTM and DTM, the TM can be estimated more accurately. Moreover, a saliency adjustment factor $u_s$ is introduced concerning about the visual attention of HVS to adjust the CM module.

1) The RTV model for EM measurement
It is known that an original image $f$ can be represented by a structural image $u$ (containing large-scale subjects like piecewise smooth and sharp edge) and a textural one $v$ (containing fine-scale details which usually have periodicity and oscillation). That is $f = u + v$. The relative total variation (RTV) model is exploited to effectively obtain the structural and textural information of the image [31]. The RTV model is defined as:

$$
\arg \min_u \sum_p (u_p - f_p)^2 + \lambda \left( \frac{\mathcal{M}_x(p)}{\mathcal{N}_x(p) + \varepsilon} + \frac{\mathcal{M}_y(p)}{\mathcal{N}_y(p) + \varepsilon} \right)
$$

where $f$ and $u$ represent the input image and the output structural image, respectively. $p$ denotes the index for 2-D image pixel. $\lambda$ is a weighting factor and $\varepsilon$ is a small positive value to avoid zero denominator. $\mathcal{M}_x(p)$ and $\mathcal{M}_y(p)$ mean windowed total variations in the $x$ and $y$ directions, which are expressed as:

$$
\mathcal{M}_x(p) = \sum_{q \in R(p)} w_{p,q} \cdot |\partial_x u_q|
$$

$$
\mathcal{M}_y(p) = \sum_{q \in R(p)} w_{p,q} \cdot |\partial_y u_q|
$$

$\mathcal{N}_x(p)$ and $\mathcal{N}_y(p)$ denote the overall spatial variation in the $x$ and $y$ directions, which are defined as:

$$
\mathcal{N}_x(p) = \sum_{q \in R(p)} w_{p,q} \cdot (\partial_x u_q)
$$

$$
\mathcal{N}_y(p) = \sum_{q \in R(p)} w_{p,q} \cdot (\partial_y u_q)
$$

where $q$ belongs to $R(p)$, the rectangular region centered at pixel $p$. $w_{p,q}$ is a weighting function, which is written as:

$$
w_{p,q} \propto \exp \left( -\frac{(x_p - x_q)^2 + (y_p - y_q)^2}{2\sigma^2} \right)
$$

where $\sigma$ adjusts the spatial scale of the window, which affects $\mathcal{M}(p)$ and $\mathcal{N}(p)$ directly.

The parameters $\lambda$ and $\sigma$ are adjusted to extract the structure image from the original image [31]. The value ranges of $\lambda$ and $\sigma$ are set as $[0.01, 0.03]$ and $(0, 8]$, respectively. When $\lambda$ is larger, the structural image will be fuzzier and the texture details can be retained completely. And the parameter $\sigma$ plays an opposite role compared to $\lambda$. When $\sigma$ is greater, it can make the structural image keep more fine-scale details and suppress the texture. In this paper, $\lambda$ and $\sigma$ are set as 0.01 and 3, respectively, referring to [31]. From Fig. 2a and Fig. 2b, the structural image $u$ and textural image $v$ can be achieved by RTV model.

Therefore, EM estimation for structural image $u$ is calculated as follows,

$$EM^u(x, y) = C_s^u
$$

where $C_s^u$ indicates the spatial contrast of $u$, and $C_s$ denotes the maximum luminance difference within the $5 \times 5$ neighborhood of $u$ [19].

2) The orientation complexity for OTM and DTM estimation
Based on the analyses above, the orientation complexity is used to split textural image $v$ obtained by RTV into orderly textural image $v_o$ for OTM estimation and disorderly textural image $v_d$ for DTM estimation, respectively.

As analysed by [32], the orientation selectivity based pattern can be described as the organization of neighbor pixels. The local perceptive region $\psi(3 \times 3)$ is related to the interactions among the orientation $\theta(x)$ of pixels in $\psi$. The similarities of pixels preferred orientation is calculated. More specific, if the orientation similarity of region $\psi$ is high, it may be a region with regular orientation. On the contrary, if
the similarity of region $\psi$ is low, it may be an irregular orientation region. It has been revealed that dissimilar orientations cause strong masking effect, the higher the dissimilarity, the stronger the masking effect. When orientation difference is larger than a certain threshold, the masking effect is obviously improved.

Thus, the complexity $P_C(x)$ of orientation selectivity based pattern of a local region $\psi$ ($3 \times 3$) is calculated with the histogram $H_m(x)$ of orientations $\theta(x)$ by quantifying $\theta(x)$ with the interval $T=12^\circ$ [1], [33],

$$P_C(x) = \sum_{m=1}^{M} ||H_m(x)||_0$$  \hspace{1cm} (6)

where $|| \cdot ||_0$ represents the $L_0$ norm and $M$ indicates the limit number of $\theta(x)$, and the histogram $H_m(x)$ is defined as follows:

$$H_m(x) = \sum_{x \in \psi(x)} \delta(\theta(x), m)$$  \hspace{1cm} (7)

where $\delta(\cdot)$ represents the pulse function, and for which

$$\delta(\theta(x), m) = \begin{cases} 1, & \text{if } \theta(x) = m \\ 0, & \text{if } \theta(x) \neq m \end{cases}$$  \hspace{1cm} (8)

The results shown in [1] illustrated that the orientation complexity $P_C(x)$ of regular region is low, and for the irregular region, the corresponding complexity $P_C(x)$ is high.

To split the textural image $v$ properly, value “1” is regarded as the threshold of $P_C(x)$ to obtain orderly textural image $v_o$ and disorderly textural image $v_d$.

$$\begin{cases} v_o, & \text{if } P_C(x) = 1 \\ v_d, & \text{if } P_C(x) \neq 1 \end{cases}$$  \hspace{1cm} (9)

When $P_C(x)$ equals to “1”, each pixel in the local region $\psi$ ($3 \times 3$) has similar $\theta(x)$, which illustrates that the orderly textural image has the homogeneous pattern complexity. The orderly textural image $v_o$ and the disorderly textural image $v_d$ are shown in Fig. 3a and Fig. 3b.

Hence, OTM and DTM evaluation can be computed as follows:

$$\begin{align*} OTM^{v_o}(x, y) &= C^{v_o}_s \\ DTM^{v_d}(x, y) &= C^{v_d}_s \end{align*}$$  \hspace{1cm} (10)

where $C^{v_o}_s$ indicates the spatial contrast for orderly textural image $v_o$, and $C^{v_d}_s$ denotes the spatial contrast for disorderly textural image $v_d$.

3) The saliency adjustment factor estimation

In order to estimate CM values more accurately, visual saliency, which is the perceptual characteristic of HVS, is added to adjust the proposed CM measurement.

The saliency model [29] is adopted to determine the saliency by removing redundant contents instead of measuring the significance.

The visual saliency model is evaluated by:

$$S(x) = \sum_{j=1}^{J} \sum_{k=1}^{K} w_{jk} \hat{H}_{jk}(x)$$  \hspace{1cm} (11)

FIGURE 4: (a) Original Lena image. (b) The saliency map $\hat{S}(x)$ of Lena image.

where $J$ and $K$ denote the number of pyramid levels and the number of image channels, respectively. $w_{jk}$ is the normalizing coefficients for each channel and scale, which is set as $w_{jk} = 1/\max_x \hat{H}_{jk}(x)$. As for $\hat{H}_{jk}$, it refers to saliency estimation provided by the redundancy reduction as follows,

$$\hat{H}_{jk} = (1 - \rho(x)) \hat{H}(x)$$  \hspace{1cm} (12)

where $\rho(x)$ represents the redundancy coefficient of pixel $x$, and $\hat{H}(x)$ refers to the entropy of pixel $x$.

In this paper, $S(x)$ is normalized as $\hat{S}(x) \in [0, 1]$ to get the final saliency map [34]. As shown in Fig. 4b, the brighter the region of $\hat{S}(x)$ is, the closer the pixel value of $\hat{S}(x)$ to value "1", and the higher degree of saliency is. Then, a threshold is set as 0.5 to binarize the final saliency map $\hat{S}(x)$ into “saliency” area and “non-saliency” area. Since HVS is more sensitive to changes in the “saliency” area, we use the saliency factor $u_S$ to adjust the CM value adaptively in “saliency” area and “non-saliency” area. The saliency adjustment factor $u_S$ is defined as follows.

$$u_S = \begin{cases} 1 - \hat{S}(x), & \hat{S}(x) \geq 0.5 \\ 1, & \hat{S}(x) < 0.5 \end{cases}$$  \hspace{1cm} (13)

4) The proposed CM model

As aforementioned analyses, the preliminary CM evaluation is calculated as:

$$CM(x, y) = EM^{v_o}(x, y) + OTM^{v_o}(x, y) + DTM^{v_d}(x, y)$$  \hspace{1cm} (14)

where

$$\begin{align*} EM^{v_o}(x, y) &= C^{v_o}_s \cdot W_c \\ OTM^{v_o}(x, y) &= C^{v_o}_s \cdot W_{v_o} \\ DTM^{v_d}(x, y) &= C^{v_d}_s \cdot W_{v_d} \end{align*}$$  \hspace{1cm} (15)

Note that to distinguish the effect of EM, OTM and DTM to the contrast masking, $W_c, W_{v_o}$ and $W_{v_d}$ are regarded as the weight coefficients of the estimation, which are set to 1, 2, and 3, respectively [35].
FIGURE 5: (a) JND map of Liu’s model [15], (b) JND map of Wu’s model [21], (c) JND map of Our proposed model, (d) Original Baboon image, (e) The saliency map of Baboon image.

Combined with the saliency adjustment factor considering the visual attention, the final CM estimation is established as follows:

\[ CM_s(x, y) = CM(x, y) \cdot u_S \]  \hspace{1cm} (16)

**B. LUMINANCE ADAPTATION**

It is well known that the human eyes is less sensitive to the distortion of darkness. With the increase luminance, the sensitivity of HVS to image changes may be improved. Therefore, a luminance adaptation (LA) model [20] is designed to adapt to the HVS.

\[ LA(x, y) = \begin{cases} 
17 \times (1 - \sqrt{f(x, y)/127}) + 3, & \text{if } f(x, y) \leq 127 \\
3 \times (f(x, y) - 127)/128 + 3, & \text{otherwise} 
\end{cases} \]  \hspace{1cm} (17)

where \((x, y)\) is the coordinate in the image.

**C. THE PROPOSED JND MODEL**

Since LA and CM are usually integrated into the pixel-wise JND model for overall JND estimation via the nonlinear additivity model for masking (NAMM) [20], the proposed JND model is established by:

\[ JND = LA + CM_s - C_{lc} \times \min\{LA, CM_s\} \]  \hspace{1cm} (18)

where \(C_{lc}\) is used to settle the overlapping impact between \(LA\) and \(CM_s\). As for \(C_{lc}\), it is set as 0.3, same as that in [20].

For illustration purpose, the JND maps for Liu et al. [15], Wu et al. [21] and the proposed model are displayed in Fig. 5a, Fig. 5b, and Fig. 5c, respectively. From these JND maps, it’s obvious that Liu’s model overvalued the visual masking in some regions around the baboon’s nose with low orientation complexity, and although Wu’s model considers the concealment effect in disorderly textural regions, its model still underestimates the visual redundancy in some areas with high orientation complexity, such as the baboon’s fur. By contrast, our model shown in Fig. 5c estimates visual redundancy more accurately based on orientation complexity and visual attention of HVS.

**III. EXPERIMENTAL RESULTS AND ANALYSIS**

**A. EXPERIMENTAL SETTING**

1) Test Image

In our experiments, twelve commonly-used test images are adopted to comprehensively evaluate the performance of various JND models [1], [15], [36]. These images are of the resolution 512×512 and contain a variety of visual content and spatial complexity, as shown in Fig. 6.
TABLE 1: Objective quality comparison of the proposed model and two pixel-wise JND models in terms of PSNR (dB)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>33.29</td>
<td>34.35</td>
<td>34.64</td>
</tr>
<tr>
<td>I2</td>
<td>32.32</td>
<td>33.00</td>
<td>33.95</td>
</tr>
<tr>
<td>I3</td>
<td>34.92</td>
<td>35.77</td>
<td>36.18</td>
</tr>
<tr>
<td>I4</td>
<td>29.70</td>
<td>30.33</td>
<td>30.73</td>
</tr>
<tr>
<td>I5</td>
<td>33.96</td>
<td>34.55</td>
<td>35.26</td>
</tr>
<tr>
<td>I6</td>
<td>33.68</td>
<td>34.88</td>
<td>35.16</td>
</tr>
<tr>
<td>I7</td>
<td>33.97</td>
<td>35.04</td>
<td>34.25</td>
</tr>
<tr>
<td>I8</td>
<td>34.25</td>
<td>34.92</td>
<td>34.93</td>
</tr>
<tr>
<td>I9</td>
<td>36.15</td>
<td>36.37</td>
<td>37.04</td>
</tr>
<tr>
<td>I10</td>
<td>32.73</td>
<td>34.69</td>
<td>34.93</td>
</tr>
<tr>
<td>I11</td>
<td>33.83</td>
<td>34.56</td>
<td>34.64</td>
</tr>
<tr>
<td>I12</td>
<td>34.65</td>
<td>35.32</td>
<td>36.06</td>
</tr>
<tr>
<td>Average</td>
<td>33.62</td>
<td>34.48</td>
<td>34.81</td>
</tr>
</tbody>
</table>

2) Evaluation Procedure

Intuitively, an ideal JND model should be able to tell how to conceal the noise in the image as much as possible with the acceptable image quality. In other words, to inject the noise with the same energy, a better JND model will put more noise into the regions with higher visual redundancy while less noise into the regions with lower perceptual redundancy for achieving the better perceptual quality. According to various pixel-wise JND models, we add the JND noise to each pixel of the test images $I$ for measuring its performance as suggested in [21]:

$$\hat{I}(x, y) = I(x, y) + \eta \cdot \xi \cdot JND(x, y)$$ (19)

where $(x, y)$ means the spatial coordinate of the pixel in image, $\hat{I}(x, y)$ denotes the contaminated image by injecting the JND guided noise, the parameter $\eta \in \{-1, +1\}$ is randomly decided to avoid the occurrence of noise change in fixed pattern, and $\xi$ is used to adjust the JND noise injection energy to ensure the noised images contaminated by different JND models at the same level of noise energy.

The contaminated test images resulted from various JND models are compared with the original test images in terms of PSNR and through the subjective quality assessment to evaluate the performances of various JND models. Note that with the same perceived quality (measured by SSIM), the higher the injected-JND-noise energy (measured by PSNR) is, the more reliable the JND model is.

B. PERFORMANCE COMPARISON

1) Objective Quality Comparison

Table 1 shows the objective quality comparison of the proposed JND model and two existing pixel-wise JND models [15], [21] in terms of PSNR. It can be easily seen that the proposed JND model is able to, on average, achieve the lowest PSNR and also the lowest PSNR for all the test images. Compared with Liu [15] and Wu [21], the additional redundancy yielded by the proposed JND model is 0.86 dB and 1.19 dB, respectively. This study shows the superiority of the proposed JND model, which can tolerate more distortions and exploit the visual redundancy more accurately.

2) Subjective Quality Comparison

In addition, subjective quality comparison is also performed to demonstrate the effectiveness of our proposed JND model. The subjective quality assessment tests are conducted by exploring the adjectival categorical judgment method and strictly following the ITU-R BT.500-11 standard [37]. The evaluation platform is the desktop PC, which is equipped with a 23-inch LED monitor (with a resolution of 1920 × 1080), 8 GB RAM, and 64-bit Windows operating system. The evaluation process is conducted indoors, under a normal lighting condition. In each test, two contaminated images by two JND models under comparison presented to the assessor will be judged as one of seven opinion levels, as shown in Fig. 7. These two contaminated images will include the one contaminated by the proposed JND model and the one contaminated by other JND models under comparison, and they will be randomly posed as the left or the right images at the same time. Seven discrete scales from -3 to +3 will be used to reflect the degree of difference of the subjective quality between the left and right images according to their corresponding definitions in Fig. 7. Twenty assessors were invited to evaluate the subjective quality of all the image pairs, and the assessor is allowed to response after observing the images at least 4 seconds [36].

Table 2 shows the subjective quality comparison of the proposed JND model and two existing pixel-wise JND models [15], [21], where $m$ denotes the mean value of the subjective scores and $SD$ means their standard deviation. Moreover, the bar graph is also shown in Fig. 8 to have a clearer illustration. Note that a larger negative (or positive) subjective score demonstrates that the image processed by our proposed JND model has much better (or worse) perceived quality than that processed by other JND models under comparison. Firstly, we can see from Table 2 that the standard deviations of the subjective scores are quite small.
By means of RTV model and orientation complexity, a real
based on elaborate image decomposition and the saliency.
In this work, a novel pixel-wise JND model was proposed
our JND model is superior to Liu’s and Wu’s models.
eyes, there seems similar among them. Thus it can be seen
numbers, which are relatively sensitive to HVS and given
areas circled by green ellipse, the visual redundancy of the
regions according to free energy principle. However, for the
al.’s model, it emphasizes the masking effect of disorderly
noise. While for the image in Fig. 9b processed by Wu et
actually only the unpredicted texture regions can hide much
the texture regions to tolerate much distortion is highlighted,
it maintains fairly good edge information, the function of
the image in Fig. 9a dealt with Liu et al.’s model, although
better than Fig. 9a and Fig. 9b. Tracing it to its cause, for
eyes are sensitive, the visual effect in Fig. 9 (c) is obviously
achieves better subjective quality.
(i.e., nearly 1), showing that the subjective evaluation results
from twenty assessors are stable and reliable. Then, as shown
in Table 2 and Fig. 8, the average mean values of two groups
of comparison tests are all negative, i.e., -0.245 and -0.094,
respectively, meaning that the contaminated images resulted
from the proposed method have overall better subjective
quality than that of other JND models [15], [21]. In other
words, the proposed JND model consistently outperforms
other JND models [15], [21].

Moreover, we further take I9 as an example to show the
corresponding contaminated images resulted from different
JND models, as displayed in Fig. 9. Note that the same noise
energy is injected into the original I9 with different kinds of
JND noise. It can be observed that the proposed model
achieves better subjective quality.

For the areas encircled with green ellipse to which human
eyes are sensitive, the visual effect in Fig. 9 (c) is obviously
better than Fig. 9a and Fig. 9b. Tracing it to its cause, for
the image in Fig. 9a dealt with Liu et al.’s model, although
it maintains fairly good edge information, the function of the
texture regions to tolerate much distortion is highlighted,
actually only the unpredicted texture regions can hide much
noise. While for the image in Fig. 9b processed by Wu et
al.’s model, it emphasizes the masking effect of disorderly
regions according to free energy principle. However, for the
areas circled by green ellipse, the visual redundancy of the
numbers, which are relatively sensitive to HVS and given
more attention, are overvalued. And for the regions encircled
with red ellipses which are dark and insensitive to human
eyes, there seems similar among them. Thus it can be seen
that our JND model is superior to Liu’s and Wu’s models.

IV. CONCLUSION
In this work, a novel pixel-wise JND model was proposed
based on elaborate image decomposition and the saliency.
By means of RTV model and orientation complexity, a real
image is split into three portions, namely, structural image,
orderly textural image and disorderly textural image for EM,
OTM and DTM estimation, respectively. Considering visual
attention of HVS, we proposed CMs for contrast masking
estimation combining based on the saliency. From the results
of PSNR comparison test and subjective quality comparison,
our proposed model is better than the related existing JND
models. Furthermore, with the advantage of our model, it will
have effective improvement in video coding, image quality
evaluation, image watermarking and so on.

REFERENCES
noticeable difference model for images with pattern complexity,” IEEE
Transactions on Image Processing, vol. 26, no. 6, pp. 2682–2693, June
2017.
difference estimation for screen content images,” IEEE Access, vol. 6, pp.
guided rate distortion optimization for high efficiency video coding,”
models of human perception,” Proceedings of the IEEE, vol. 81, no. 10,
FIGURE 9: The subjective comparison of JND noise-injected images resulted from different JND models (Taking I9 as an example): (a) Liu [15]; (b) Wu [21]; and (c) Proposed.