Just Noticeable Visual Redundancy Forecasting: A Deep Multimodal-Driven Approach

Wuyuan Xie[1], Shukang Wang[1], Sukun Tian[2], Lirong Huang[1], Ye Liu[3], Miaohui Wang[1]*

[1] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University
[2] Peking University
[3] Nanjing University of Posts and Telecommunications
wuyuan.xie@gmail.com, sukhum169@hotmail.com, yeliu@njupt.edu.cn, wang.miaohui@gmail.com

Abstract

Just noticeable difference (JND) refers to the maximum visual change that human eyes cannot perceive, and it has a wide range of applications in multimedia systems. However, most existing JND approaches only focus on a single modality, and rarely consider the complementary effects of multimodal information. In this article, we investigate the JND modeling from an end-to-end homologous multimodal perspective, namely hmJND-Net. Specifically, we explore three important visually sensitive modalities, including saliency, depth, and segmentation. To better utilize homologous multimodal information, we establish an effective fusion method via summation enhancement and subtractive offset, and align homologous multimodal features based on a self-attention driven encoder-decoder paradigm. Extensive experimental results on eight different benchmark datasets validate the superiority of our hmJND-Net over eight representative methods.

1 Introduction

Just noticeable difference (JND) reflects the maximum visual redundancy (also known as *visibility threshold*) that cannot be perceived by the human visual system (HVS). In multimedia community, visibility threshold effectively quantifies the perceptual redundancy in visual signals, which has a wide range of applications in image compression and processing systems (Wang et al. 2022). For instance, JND can guide the removal of redundant information in multimedia compression (Zhang et al. 2016; Tian et al. 2021), improving the coding efficiency without affecting its perceived quality. It can also improve the accuracy of quality assessment (Seo, Ki, and Kim 2020), as well as increase the strength of watermark embedding (Liu et al. 2021a). Based on the modeling approach, the existing JND methods can be roughly divided into 1) HVS-guided models (Bae, Kim, and Kim 2016; Wu et al. 2017; Chen and Wu 2019) and 2) learning-based models (Huang et al. 2017; Ki et al. 2018; Liu et al. 2019; Shen et al. 2020; Tian et al. 2021).

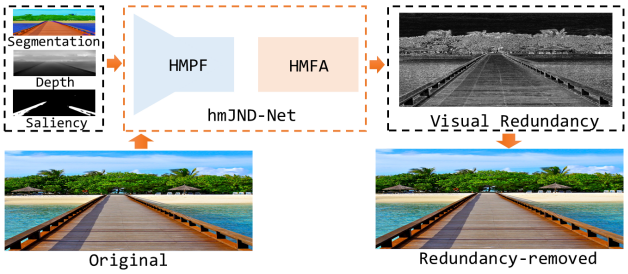


Figure 1: Illustration of the proposed multimodal-driven visual redundancy forecasting network. The saliency, depth, segmentation, and RGB modalities are jointly investigated to predict the visual redundancy based on a homologous-modality prior fusion (HMPF) module and a homologous-modality feature alignment (HMFA) module.

The HVS-guided models mainly explore the visibility threshold in the pixel or transform domain, both of which share a basic process (Liu, Zhan, and Wang 2020): first modeling some prominent visual masking effects, and then combining them with a nonlinear fusion method. Based on the properties of HVS, the influence factors considered in existing methods are luminance adaptation (LA), spatial contrast masking (CM), visual saliency (Wang et al. 2014), contrast sensitivity (Bae, Kim, and Kim 2016), pattern complexity (Wu et al. 2017), and foveation effect (Chen and Wu 2019), *etc*. However, the HVS-guided approaches have at least two shortcomings: 1) Due to the fact that our knowledge for HVS is limited, we cannot obtain all possible influence factors to predict the just noticeable visual redundancy; 2) the potential interrelationship between different HVS factors cannot

*This work was supported in part by the National Natural Science Foundation of China under Grants 61701310 and 61902251, in part by the Natural Science Foundation of Shenzhen City under Grants 20200805200145001 and JCYJ20220809160139001, and in part by the Natural Science Foundation of Guangdong Province under 2021A1515011877 and 2022A1515011245. (*Corresponding author*: *Miaohui Wang*)

be accurately characterized by existing analytic models.

With the development of deep learning, some preliminary studies (Liu et al. 2019; Shen et al. 2020; Tian et al. 2021) aim to address the problems of the HVS-guided model described above. Several encoder-based JND datasets have been constructed, which are essential for data-driven approaches. In labeling the visibility threshold map, each subject evaluates a series of distorted images or videos produced by a typical codec with different quality factors. For instance, (Jin et al. 2016) constructed a JPEG-compression-based dataset MCL-JCI, where each source image was encoded 100 times by the JPEG codec from 1 to 100, and the decoded images were evaluated by volunteers through a dichotomous search strategy to determine the picture-wise label data. (Shen et al. 2020) further constructed a large dataset based on the versatile video coding (VVC) codec. In addition, (Wang et al. 2016) constructed a video-wise dataset. However, the learning-based JND models rely heavily on manually labeled datasets, and face the problem of insufficient data. In view of this, we try to explore the training data by introducing homologous multimodalities to further improve the prediction accuracy of visual redundancy.

Inspired by the above discussion, we investigate an end-to-end homologous multimodal-driven visual redundancy forecasting network, which incorporates saliency, depth, segmentation, and RGB image modalities as shown in Figure 1. More specifically, we firstly acquire depth, saliency, and segmentation modalities from the input data. Then, we fuse those prior multimodal features via a channel attention network, and subsequently use a transformer network to align the RGB modality with the fused prior multimodal features to obtain the final visibility threshold.

The main contributions compared with the previous methods are summarized as follows. We for the first time investigate homologous multimodal information in a supervised visibility threshold modeling, and devise a novel end-to-end homologous multimodal JND forecasting network (namely hmJND-Net). The proposed hmJND-Net alleviates the problem of insufficient training data as well as improves the prediction accuracy. To better utilize the three prior modalities including depth, saliency, and segmentation, we further develop a new homologous-modality prior fusion (HMPF) module to early fuse these prior features, and explore a homologous-modality feature alignment (HMFA) module with a transformer structure to align RGB image modality and three prior modalities. Experimental results on eight benchmark datasets verify that our hmJND-Net outperforms representative schemes by providing higher accuracy and more bit-rate saving.

2 Related Work

In this section, we first briefly review some representative just noticeable visual redundancy models, including HVS-guided and learning-based approaches. Then, we introduce the motivation of the proposed hmJND-Net.

2.1 HVS-guided Methods

As mentioned earlier, the basis for HVS-guided models is derived from the properties of the HVS. In the pixel domain, the LA and CM effects are widely-used: The LA effect responds to the different sensitivities of the HVS to visual signals with different background intensities, while the CM effect responds to small changes of the HVS to uniform or non-uniform content (*i.e.*, visibility suppression). For instance, (Chou and Li 1995) proposed a classical JND framework that modeled LA as a quasi-parabolic model and derived CM from the local gradient intensity. Based on it, (Yang et al. 2005) developed a nonlinear additivity model for masking (NAMM) module to reduce the superposition effect between two masking effects. (Liu et al. 2010) further decomposed the original image into structural and textural components to estimate edge masking and texture masking, which aimed to avoid the over-estimation of the edge and texture regions. (Wu et al. 2013) established an auto-regressive model based on the free-energy principle to predict the order and disorder regions, and employed the disorder concealment effect in the visibility threshold estimation. Later, they utilized the diversity of orientations to measure the pattern complexity and derive the pattern masking (Wu et al. 2017). Recently, (Chen and Wu 2019) further introduced the foveation effect and the asymmetric visual sensitivity into the maximum visual redundancy modeling. (Jiang et al. 2022) explored the characteristics of Karhunen–Loève transform (KLT) coefficient to derive a perceptually lossless prediction.

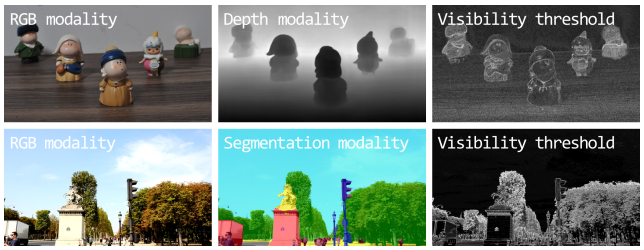


Figure 2: Examples of the depth modality and the segmentation modality.

Besides the pixel domain models, the transform domain methods have emerged to facilitate multimedia applications. The contrast sensitivity function (CSF) plays a dominant role in the transform domain JND modeling, which responds to the frequency characteristics of HVS. For instance, (Peterson and Ahumada Jr 1992) modeled the CSF component as an exponential function of spatial contrast. (Zhang, Lin, and Xue 2005) revised the LA and CM models based on a block classification in the discrete cosine transform (DCT) domain. (Wei and Ngan 2009) used a gamma correction to compensate for the LA and CM effects in the DCT domain. (Bae, Kim, and Kim 2016) established the CM effect based on the structural contrast of DCT blocks with multiple sizes, designed a DCT-based local redundancy probability to estimate the distribution of the transform coefficients, and introduced it into perceptual video coding.

2.2 Learning-based Methods

With the development of deep learning, learning-based JND estimation has become a new trend in this field. For in-

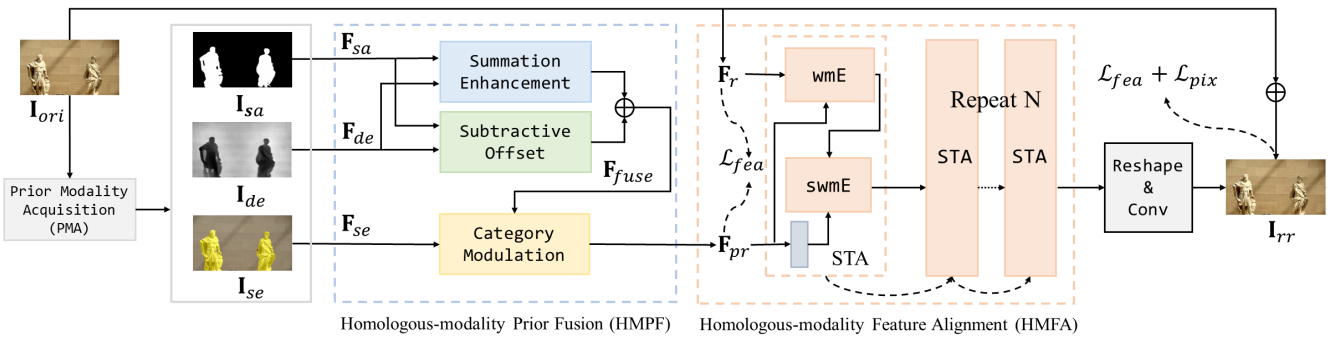


Figure 3: Pipeline of the proposed multimodal-based visual redundancy framework. It mainly consists of a prior modality acquisition (PMA) module, a homologous multimodal prior fusion (HMPF) module, and a homologous multimodal feature alignment (HMFA) module. Please zoom in the electronic version for better details.

stance,(Huang et al. 2017) proposed to learn the mean value of group-based visibility threshold distributions based on extracted video features. (Ki et al. 2018) established an energy reduction-based model, which used a linear regression and convolutional neural network (CNN)-based models to reduce the impact of the compressed artifacts. (Liu et al. 2019) developed a deep binary classifier to predict whether an encoded image was perceptually distorted with respect to its original one. (Shen et al. 2020) decomposed the input image into three components (*i.e.*, luminance, contrast, and structure), and then trained a deep CNN degradation model to predict the visibility threshold. (Tian et al. 2021) designed a block-level visual redundancy algorithm, which employed AlexNet to predict the encoding parameters for each image block. (Wang et al. 2022) predicted the visibility threshold by considering the oblique correction effect, where a learned end-to-end CNN mapping was finally established between the visibility threshold and quality factor.

2.3 Motivation

Recently, multimodal data is usually used to portray an object from different sources or different perspectives. Data obtained from different sources refers to heterogeneous multimodal information (Yu et al. 2021; Min et al. 2020). Data obtained from a single source refers to homologous multimodal information (Wang et al. 2018; Shao and Wang 2021; Xie, Huang, and Wang 2022). How to fuse homologous multimodal data is a popular research topic in multimedia image processing. Many studies have employed the homologous multimodal information to improve the performance of image processing systems. For example, (Wang et al. 2018) introduced a segmentation modality to address the problem of unrealistic super-resolution textures. (Cheng, Fu, and Yang 2020) adopted a depth modality to extract the high and low resolution patches to train a super-resolution network. (Shao and Wang 2021) utilized a saliency modality as the supervision for detail reconstruction in image inpainting tasks.

Inspired by the above efforts, we for the first time investigate the homologous multimodal-based JND modeling (*i.e.*, saliency, depth, segmentation) in this paper: 1) **Saliency modality.** Due to the fact that visual acuity decreases when the distance increases from the fovea of the human retina, visibility threshold increases with eccentricity. Thus, it is a natural choice to reduce less visual redundancy for salient regions and reduce more visual redundancy for non-salient regions. 2) **Depth modality.** The image patches close to the camera will be clearer, and as the depth is deeper, the image content will shrink and become blurred. As shown in Figure 2, there is a relationship between the visual redundancy and their depth, where a deeper depth tolerates more visual redundancy than the shallower one. Based on this prior knowledge, we can reduce more visual redundancy in the deeper depth regions based on the depth modality. 3) **Segmentation modality.** We find that different objects have various visibility thresholds. As shown in Figure 2, the visibility thresholds of trees and houses with complex textures are significantly higher than those of the flat sky and wall. Therefore, we introduce the category information into our framework via semantic segmentation modality, which reduces different redundancies for different objects.

3 Proposed Multimodal Framework

In this section, we describe the proposed end-to-end homologous multimodal visibility threshold forecasting network in detail, which consists of a prior modality acquisition (PMA) module, homologous-modality prior fusion (HMPF) module, and homologous-modality feature alignment (HMFA) module as shown in Figure 3.

3.1 Overview

Problem Formulation. Unlike existing visibility threshold methods, we perform a pixel-wise homologous multimodal network in an end-to-end manner, and hence our overall task can be expressed by minimizing a specially designed optimization problem as

$$\min_{I_{rr}} \left(\mathcal{L}_{overall} \left(\mathcal{L}_{fea}(F_r, F_{pr}), \mathcal{L}_{pix}(I_{rr}, I_{gt}) \right) \right), \quad (1)$$

where $\mathcal{L}_{overall}(\cdot)$ denotes the overall loss function. $\mathcal{L}_{overall}(\cdot)$ is obtained by weighting the similarity measure $\mathcal{L}_{fea}$ between the RGB features F_r and the prior fused features F_{pr}, and the pixel distance $\mathcal{L}_{pix}$ between the output redundancy-removed image I_{rr} and its ground-truth I_{gt}.

More specifically,

$$\begin{aligned}&\mathcal{L}_{overall}\left(\mathcal{L}_{fea}(F_r, F_{pr}), \mathcal{L}_{pix}(I_{rr}, I_{gt})\right)\\&= \lambda_{fea} \times \mathcal{L}_{fea}(F_r, F_{pr}) + \lambda_{pix} \times \mathcal{L}_{pix}(I_{rr}, I_{gt})\\&= \frac{\lambda_{fea}}{h_f \times w_f \times c_f}\|F_r - F_{pr}\|_1 + \frac{\lambda_{pix}}{h_p \times w_p \times c_p}\|I_{rr} - I_{gt}\|_2^2\end{aligned}, \quad (2)$$

where (h_f, w_f, c_f) denotes the height, width, and the number of channels of the prior fused features. (h_p, w_p, c_p) denotes the height, width, and the number of channels of the predicted ones. $\mathcal{L}_{fea}$ denotes the pixel-wise L_1 loss between the RGB features and the prior fused features to measure their similarity, and minimizes the distance between them for the multimodal feature alignment. $\mathcal{L}_{pix}$ denotes the pixel-wise L_2 loss between the output redundancy-removed image and the ground-truth. λ_{fea} and λ_{pix} are the corresponding weights.

Architecture. We investigate a total of four modalities in the proposed hmJND-Net, including RGB image modality, saliency modality, depth modality, and segmentation modality. Figure 3 shows the overall network architecture, which is formulated as

$$I_{rr} = \mathcal{M}_{JND}\left(I_{ori}, \mathcal{M}_{EX}(I_{hm})\right), \quad (3)$$

where I_{hm} denotes the homologous multimodal information obtained via the original RGB modality I_{ori}, including saliency modality I_{sa}, depth modality I_{de}, and segmentation modality I_{se}. Generally, Eq. (3) can be decomposed into two sub-tasks: a multimodal feature extraction module $\mathcal{M}_{EX}$ and a multimodal forecasting module $\mathcal{M}_{JND}$.

The multimodal feature extraction, $\mathcal{M}_{EX}$, consists of the PMA and HMPF modules. First, the original RGB image modality is fed into the PMA module to generate the corresponding three prior modalities, and the shallow features are extracted by a three-layer convolution. Then the shallow features are fed into the HMPF module, $\mathcal{M}_{HMPF}$, for the fusion of the prior modality features. $\mathcal{M}_{EX}(\cdot)$ can be expressed as

$$F_{pr} = \mathcal{M}_{HMPF}\left(\mathcal{M}_{PMA}(I_{ori})\right), \quad (4)$$

where F_{pr} represents the prior fused features after a homologous multimodal fusion module, $\mathcal{M}_{PMA}(\cdot)$ denotes a PMA module, and $\mathcal{M}_{HMPF}(\cdot)$ denotes a HMPF module.

After the HMPF process, F_{pr} and F_r are fed together into the HMFA module for the feature alignment. First, the pixel-wise distance between F_{pr} and F_r is minimized, and then several Swin Transformer Alignment (STA) blocks are used to further align the F_{pr} and F_r features. Subsequently, the aligned features are fed into a `Reshape` module. Finally, the initial input is connected for the residual learning. $\mathcal{M}_{JND}(\cdot)$ can be expressed as

$$I_{rr} = \theta\left(\mathcal{M}_{RE}(\mathcal{M}_{HMFA}(F_{pr}, F_r))\right), \quad (5)$$

where F_r denotes the intermediate features of the RGB image modality extracted by a three-layer convolution, $\mathcal{M}_{HMFA}(\cdot)$ denotes a HMFA module, $\mathcal{M}_{RE}(\cdot)$ denotes a `Reshape` module, and $\theta(\cdot)$ denotes the normalization operation to restrict the output to a reasonable scale.

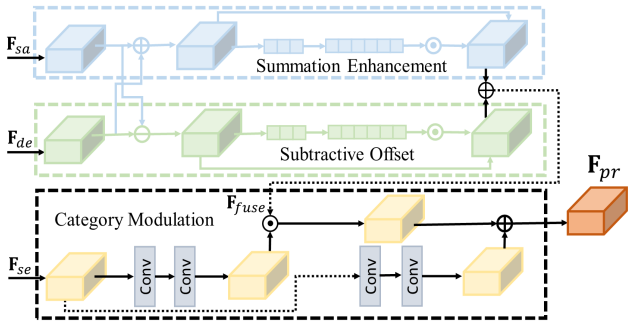


Figure 4: Homologous-modality Prior Fusion. The inputs are saliency F_{sa}, depth F_{de} and segmentation F_{se} modalities, and the output is the prior fused feature F_{pr}.

3.2 Homologous-modality Prior Fusion

In order to make full use of the prior multimodal information, we develop a special homologous-modality prior fusion (HMPF) module as shown in Figure 4. Before the HMPF process, the acquired prior modalities will be extracted by a three-layer 3×3 convolution to obtain the corresponding shallow feature $F_x(i.e., x = \{sa, de, se\})$. In other words, I_{sa}, I_{de}, and I_{se} will be embedded into F_{sa}, F_{de}, and F_{se}, respectively. Given that the depth and saliency modalities on the visibility threshold estimation are not always positively correlated, image regions with a deeper depth may belong to salient, while those with shallower depth may belong to non-salient. Therefore, to capture the homologous multimodal complementary information, hmJND-Net needs to learn not only the consistency between the depth and saliency modalities but also the difference between them.

Based on the above analysis, we propose a two-submodules strategy, summation enhancement $\mathcal{M}_{SE}$ and subtractive offset $\mathcal{M}_{SO}$, to learn the complementary relationship between these two modalities, and then modulate the fused feature map F_{fuse} by introducing the category information. Specifically, we obtain the subtractive feature F_{sub} with the difference information by subtracting F_{sa} and F_{de}, and the summation feature F_{sum} with the consistency information by adding F_{sa} and F_{de}, respectively.

Inspired by the channel attention network SENet (Hu, Shen, and Sun 2018), we further perform the channel attention weighting on these two intermediate features to obtain the enhanced feature F_e and the offsetting feature F_o, and then fuse them to obtain F_{fuse}, which can be formulated as

$$F_{fuse} = F_e \oplus F_o, \quad (6)$$

$$F_e = F_{sum} \otimes \mathcal{F}^s_{fc}(\mathcal{F}^r_{fc}(\mathcal{F}_{gp}(F_{sum}))) \oplus F_{sum}, \quad (7)$$

$$F_o = F_{sub} \otimes \mathcal{F}^s_{fc}(\mathcal{F}^r_{fc}(\mathcal{F}_{gp}(F_{sub}))) \oplus F_{sub}, \quad (8)$$

where $\oplus$ and $\otimes$ denote an element-wise summation and multiplication, respectively. $\mathcal{F}_{gp}(\cdot)$ denotes a global pooling, and $\mathcal{F}^x_{fc}(\cdot)$ denotes a fully-connected layer followed by a `Sigmoid` normalization or a `ReLU` activation layer.

After obtaining the fusion features of saliency and depth modalities, we need to incorporate the category information. Inspired by the spatial feature transform (Wang et al. 2018),

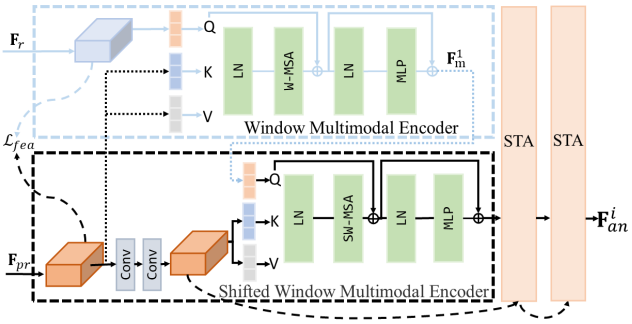
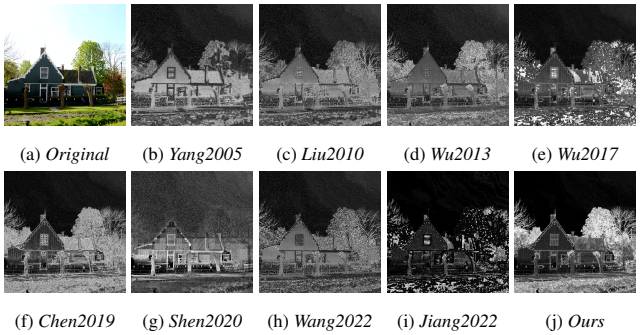


Figure 5: Homologous-modality Feature Alignment. The inputs are the prior fused feature F_{pr} and RGB modality feature F_r, and the output is the aligned feature F_{an}^i.



(a) *Original* (b) *Yang2005* (c) *Liu2010* (d) *Wu2013* (e) *Wu2017*

(f) *Chen2019* (g) *Shen2020* (h) *Wang2022* (i) *Jiang2022* (j) *Ours*

Figure 6: Comparisons of the predicted visual redundancy for 9 representative approaches. In Figures (b)-(j), brighter pixels represent higher redundancies.

we propose to use the semantic segmentation modality, and modulate the fused features based on it. This fusion process can be formulated as

$$F_{pr} = F_{fuse} \otimes \mathcal{F}_{conv}(F_{se}) \oplus \mathcal{F}_{conv}(F_{se}), \tag{9}$$

where $\mathcal{F}_{conv}(\cdot)$ denotes a two-layer convolution. The fused features of all three modalities will be fed into the next HMFA module.

3.3 Homologous-modality Feature Alignment

After obtaining F_{pr}, we need to align the RGB modality feature F_r with the prior multimodal feature F_{pr} due to the large difference between them. For this purpose, we further design a homologous-modality feature alignment (HMFA) module, as shown in Figure 5. We measure the similarity of these two features as an explicit alignment, and perform an implicit alignment via a Swin Transformer Alignment (STA) block which is inspired by (Liu et al. 2021b). The main purpose behind it is to use the self-attention mechanism to learn the alignment relationship between F_r and F_{pr}.

First, a HMFA module is composed of several STA blocks, and each STA block aligns the RGB modality with the prior homologous multimodal feature based on a multi-headed attention mechanism.

$$F_{an} = \mathcal{B}_{STA}(F_{pr}, F_r), \tag{10}$$

where F_{an} denotes the aligned output feature by the STA module. $\mathcal{B}_{STA}(\cdot)$ denotes a STA block, which contains two feature encoders including a window multimodal encoder (wmE) and a shifted window multimodal encoder (swmE). The STA blocks are organized by Eq. (11).

$$F_{an}^{i+1} = \begin{cases} \mathcal{B}_{sw}^i(Q = F_m^i, K, V = \mathcal{F}_{conv}^i(F_{pr})) & i = 1 \\ \mathcal{B}_{sw}^i(Q = F_m^i, K, V = \mathcal{F}_{conv}^{2i-1}(F_{pr})) & n \geq i > 1 \end{cases}, \tag{11}$$

where $\mathcal{B}_w^i(Q, K, V)$ and $\mathcal{B}_{sw}^i(Q, K, V)$ denote the *i*-th wmE and swmE, respectively. (Q, K, V) is the query, key, and value in the self-attention mechanism. F_m^i denotes the output of the *i*-th wmE, which is defined by

$$F_m^i = \begin{cases} \mathcal{B}_w^i(Q = F_r, K, V = F_{pr})) & i = 1 \\ \mathcal{B}_w^i(Q = F_{an}^{i-1}, K, V = \mathcal{F}_{conv}^{2i-2}(F_{pr}))) & n \geq i > 1 \end{cases}, \tag{12}$$

where $\mathcal{F}_{conv}^i(\cdot)$ denotes the *i*-th convolution block, and each convolution block consists of a two-layer convolution.

The difference between wmE and swmE is mainly in the shifted window mechanism for the interaction between different windows, which are used interchangeably in the STA module. The benefit of such a structure strengthens the memory of the deep network by repeatedly connecting the prior multimodal feature via the self-attention mechanism, gradually capturing the global information of multimodal features, learning the homologous multimodal interrelationships, and aligning the RGB and prior modalities.

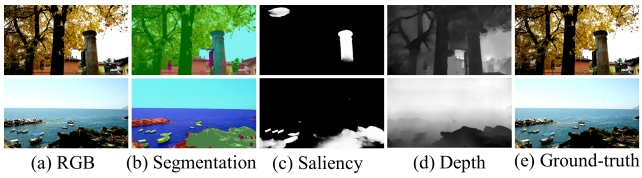
4 Experimental Validations

4.1 Experimental Protocols

Dataset Description. We trained hmJND-Net based on the latest benchmark dataset (Shen et al. 2020), which covers various image contents, including outdoor, indoor, landscape, nature, people, objects and buildings. The dataset consists of 202 high-definition original images with the size of 1920×1080 and 7878 redundancy-removed images by VVC. Each original image has 39 encoded versions with different redundancy-removed levels, which is selected by the subjective experiments.

The three modalities used in the experiments were obtained from the original images, with the depth modality generated by DPT-Hybrid (Ranftl, Bochkovskiy, and Koltun 2021), the saliency modality generated by BASNet (Qin et al. 2019), and the segmentation modality generated by BEIT (Bao et al. 2021). We provide some typical examples as shown in Figure 7.

We randomly select image pairs from the dataset as the training set, validating set, and testing set based on the ratio of **8:1:1**. To demonstrate the generalization capability, we also independently conduct the testing experiments of seven additional benchmark datasets, including *CSIQ* (Larson and Chandler 2010) (30 original natural-content images with the size of 512×512), *KADID-10K* (Lin, Hosu, and Saupe 2019) (81 original natural-content images with the size of 512×384), *LIVE* (Sheikh 2005) (29 original natural-content images with various resolutions), *TID2013* (Ponomarenko et al. 2015) (25 original natural-content images with the size



(a) RGB (b) Segmentation (c) Saliency (d) Depth (e) Ground-truth

Figure 7: Examples of four modalities used in the experiments. The right-most column is the ground-truth of I_{rr}.

Saliency	Depth	Segmentation	$PSNR_{rr}^{gt}[+]$	$SSIM_{rr}^{gt}[+]$
✗	✗	✗	33.8645	0.9571
✓	✗	✗	34.3585	0.9634
✗	✓	✗	34.3719	0.9637
✗	✗	✓	34.1246	0.9631
✓	✓	✓	**34.5038**	**0.9652**

Table 1: Ablation study of three additional modalities on the proposed hmJND-Net.

of 512×384), *SCID* (Yang, Fang, and Lin 2015) (40 original screen-content images with the size of 1280×720), *SIQAD* (Ni et al. 2017) (20 original screen-content images with various resolutions), and *MCL-JCI* (Jin et al. 2016) (50 original natural-content images with the size of 1920×1080).

Evaluation Metrics. A better JND approach is able to hide more noise. To evaluate the noise tolerance performance, we randomly eject noise into each pixel guided by the predicted visibility threshold.

$$I_{con} = I_{ori} + \alpha \times \gamma \times I_{vt}, \tag{13}$$

where I_{con} denotes the JND-contaminated image and I_{ori} denotes the original image. α denotes an adjustment factor used to adjust the noise level. γ represents a random matrix with the same size as I_{ori}, and its value is 1 or -1. I_{vt} denotes the predicted visibility threshold by the corresponding method. In the experiments. I_{vt} represent the visibility threshold, which is the difference between I_{ori} and I_{rr}.

In the experiments, we adjust α in Eq. (13) to ensure all JND-contaminated images at the same noise level (*e.g.*, mean square error (MSE)=100). To better demonstrate the comparison performance, we evaluate the prediction accuracy of visibility threshold at the image level. Specifically, we use the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity (SSIM) as the baseline indicators. We denote the PSNR result between the redundancy-removed image and the ground-truth image as $PSNR_{rr}^{gt}$, and the SSIM result as $SSIM_{rr}^{gt}$.

We employ the MS-SSIM and the mean opinion score (MOS) to measure the visual quality of contaminated images generated by different models (Zhai and Min 2020). MS-SSIM is an objective full-reference image quality indicator with a score in the range of [0,1]. MS_{con}^{ori} denotes the MS-SSIM between the JND-contaminated image and the original one. Correspondingly, MOS evaluates the visual quality difference, denoted as MOS_{con}^{ori}, which is obtained by a subjective experiment.

In summary, $PSNR_{rr}^{gt}$ and $SSIM_{rr}^{gt}$ measure the forecasting accuracy of redundancy-removed images. MS_{con}^{ori} and MOS_{con}^{ori} measure the perceptual quality of JND-contaminated images. A higher value means a better result for all four metrics.

Subjective Test. The subjective experiment is conducted mainly based on ITU-R BT.500-13 standard. We invite 20 subjects to participate in the subjective test, 12 males and 8 females, aging from 18 to 40. Two images are juxtaposed on a 65-inch monitor, where one of them is the original image and the other one is the JND-contaminated image. The subjective progress is the same as in (Shen et al. 2020). After obtaining the subjective results, we employ a statistical-based outlier rejection method suggested in (Shen et al. 2020) to examine the consistency of raw results in the experiments. After removing the outlier data, the average result is considered as the final MOS score in the following section.

Implementation Details. hmJND-Net has been implemented on *PyTorch* with all weights initialized by the *truncated normal initializer*, and the *Adam* optimizer with the default parameters (*e.g.*, β_1 = 0.9 and β_2 = 0.999). For the HMFA module, we have executed 3 sets of STA blocks. We have trained the hmJND-Net with a mini-batch size of 16 for 200 epochs on an *Nvidia Tesla A100* GPU, which takes about 10 hours. The initial learning rate is 1e-4, which will linearly decay to 0 from the 200-*th* epoch. Due to the large image size (1920×1080) in the dataset, an adaptive partition is performed on all input images, with the partition size of 224×224. Apart from the above operations, we do not use any additional special data augmentation methods.

4.2 Ablation Study

To verify the necessity of the introduced prior modalities and the validity of the proposed multimodal fusion and alignment modules, we perform the ablation study on the benchmark dataset (Shen et al. 2020). In Table 1 and Table 2, the symbols "✓" and "✗" mean selected and unselected, respectively.

Effects of Three Modalities. To verify the necessity of the introduced three modalities, we have conducted five related experiments as tabulated in Table 1. In the experiments, the substitution of a certain modality is done by repeating the enabled modality. The experimental results show that the best results are achieved when three modalities are used simultaneously, which verify the necessity of the introduced modalities.

Effects of HMPF and HMFA. To validate the effectiveness of our HMPF and HMFA modules, we have also conducted four additional experiments as provided in Table 2. In the experiments, the alternative to HMPF is to simply concatenate all prior modalities together, and then use a three-layer 3×3 convolution and a one-layer 1×1 convolution to reduce the number of features.

The alternative to HMFA is to first concatenate the prior multimodal features and the RGB modality together, and then use the channel attention mechanism based on SENet

HMPF	HMFA	$PSNR_{rr}^{gt}[+]$	$SSIM_{rr}^{gt}[+]$
✗	✗	33.8348	0.9594
✗	✓	34.2866	0.9633
✓	✗	34.0914	0.9608
✓	✓	**34.5038**	**0.9652**

Table 2: Ablation study of the proposed HMPF and HMFA modules.

and a one-layer 1×1 convolution to align them. The experimental results show that the best performance can be achieved when both two modules are enabled simultaneously, which verify the effectiveness of the proposed HMPF and HMFA modules.

4.3 Overall Performance Comparisons

Qualitative Analysis. Figure 6 demonstrates a visual comparison of "I05" on SHEN2020 dataset, where Figure 6 (a) is the original image, and Figure 6 (b)-(j) provide the results of nine predicted visual redundancy, where brighter pixels mean higher redundancies. The detailed information for each method (*e.g.*, *Yang2005*) is provided in Table 3.

As seen, existing representative methods may predict high visibility thresholds on the salient objects or flat regions, which leads to worse visual quality. The main reason is that smooth and flat regions are more sensitive, and a small change may attract the attention of HVS. Due to the joint guidance of the multimodal information, hmJND-Net tolerates more noise in texture regions (*e.g.*, trees and grasses) which are not easily noticed by HVS, and tolerates less noise in smooth and flat sensitive regions. Thus, our hmJND-Net has a better redundancy prediction as shown in Figure 6 (j).

Quantitative Analysis. The quantitative experiments are carried out on the eight widely-used image datasets as described in Sec. 4.1. In the experiments, the noise ejection level is MSE=100 for all testing images by Eq. (13).

Table 3 provides the MS_{con}^{ori} results of nine representative models on eight benchmark datasets. It can be seen that the proposed method achieves the highest average MS-SSIM results under the same noise level on all eight datasets.

In addition, we also conduct a subjective experiments under the same noise level (MSE=100) in terms of MOS_{con}^{ori} on the SHEN2020 dataset. The average MOS_{con}^{ori} results of each model (from top to bottom in Table 3) is -1.5071, -1.7214, -1.5929, -1.1429, -1.0572, -1.8964, -1.1842, -1.0833 and -0.5536, respectively. As seen, our hmJND-Net achieves the highest average MOS result.

4.4 JND-guided Compression Performance

In this section, we have incorporated our hmJND-Net into the compression application, including the widely-used JPEG, and two state-of-the-art compression standards (*e.g.*, HEVC and VVC).

JND-guided JPEG Compression. In JND-guided JPEG compression, one widely-used manner is to remove visual redundancy of an input image before encoding, which can save bit-rate while maintaining the same visual quality. Specifically, each 8×8 block is processed by the predicted visibility threshold as follows:

$$\widehat{I}_{ori}(\mathbf{p}) = \begin{cases} \overline{I}_b, & if\ |I_{ori}(\mathbf{p}) - \overline{I}_b| \le I_{vt}(\mathbf{p}), \\ I_{ori}(\mathbf{p}) + I_{vt}(\mathbf{p}), & if\ I_{ori}(\mathbf{p}) - \overline{I}_b < -I_{vt}(\mathbf{p}), \\ I_{ori}(\mathbf{p}) - I_{vt}(\mathbf{p}), & otherwise. \end{cases}, \quad (14)$$

where $I_{ori}(\mathbf{p})$ denotes a pixel magnitude at the position $\mathbf{p}$, $I_{vt}(\mathbf{p})$ denotes the related visual threshold value, $\overline{I}_b$ denotes the mean luminance value of the current 8×8 block, and $\widehat{I}_{ori}(p)$ denotes the pre-processed output.

JND-Guided HEVC/VVC Compression. In HEVC or VVC standards, we first obtain the JND-guided residual value, and the residual block is then encoded by the HEVC or VVC encoder. Specifically, we compute the JND-guided residual value as follows:

$$\widehat{R}(\mathbf{p}) = \begin{cases} 0, & if\ |R(\mathbf{p})| \le I_{vt}(\mathbf{p})\ and\ \sigma^2(\mathbf{p}) > \sigma^2, \\ \min(R(\mathbf{p}) \times \frac{|I_{vt}(\mathbf{p})|}{\sigma^2}, R(\mathbf{p}) + I_{vt}(\mathbf{p})), & else\ if\ R(\mathbf{p}) < 0, \\ \max(R(\mathbf{p}) \times \frac{|I_{vt}(\mathbf{p})|}{\sigma^2}, R(\mathbf{p}) - I_{vt}(\mathbf{p})), & otherwise. \end{cases}, \quad (15)$$

where $\sigma^2(\mathbf{p})$ and σ^2 denote the related variances of a local range and the encoding block at the position $\mathbf{p}$, $R(\mathbf{p})$ denotes the residual value obtained by HEVC or VVC, and $\widehat{R}(\mathbf{p})$ denotes the processed prediction residual value.

Compression Performance Comparisons. We have conducted the compression experiments on *SHEN2020*, and four representative methods are compared in terms of bits-per-pixel (bpp). For example, the average bpp results of JPEG lossless, (Chen and Wu 2019) (*Chen19JPEG*), (Wang et al. 2022) (*Wang22JPEG*), (Jiang et al. 2022) (*Jiang22JPEG*), and our hmJND-Net (*OursJPEG*) are 4.3031, 3.4440, 3.5604, 4.0643, and 3.3771 on the JPEG encoder. The average bpp results of *HEVC* lossless, *Chen19HEVC*, *Wang22HEVC*, *Jiang22HEVC*, and *OursHEVC* are 5.0411, 3.6557, 3.6366, 4.2253, and 3.4144 on the HEVC encoder. The average bpp results of *VVC* lossless, *Chen19VVC*, *Wang22VVC*, *Jiang22VVC*, and *OursVVC* are 4.5304, 2.9958, 2.7764, 3.0934, and 2.6328 on the VVC encoder. With the similar visual quality, our hmJND-guided coding method saves the average bit-rate by 32.3809%.

Figure 8: Visual comparisons of four state-of-the-art JND-guided methods on three different encoding platforms.

Method \ Dataset	*CSIQ*	*KADID-10K*	*LIVE*	*TID2013*	*SCID*	*SIQAD*	*MCL-JCI*	*SHEN2020*	**Average**
Yang2005 (Yang et al. 2005)	0.9591	0.9513	0.9543	0.9519	0.9525	0.9686	0.9352	0.9510	0.9530
Liu2010 (Liu et al. 2010)	0.9629	0.9547	0.9576	0.9548	0.9558	0.9712	0.9370	0.9495	0.9554
Wu2013 (Wu et al. 2013)	0.9598	0.9541	0.9551	0.9520	0.9676	0.9831	0.9365	0.9504	0.9573
Wu2017 (Wu et al. 2017)	0.9647	0.9617	0.9629	0.9599	0.9650	0.9761	0.9490	0.9593	0.9623
Chen2019 (Chen and Wu 2019)	0.9690	0.9659	0.9658	0.9630	0.9691	0.9765	0.9567	0.9634	0.9662
Shen2020 (Shen et al. 2020)	0.9558	0.9454	0.9477	0.9452	0.9463	0.9609	0.9301	0.9418	0.9466
Wang2022 (Wang et al. 2022)	0.9677	0.9613	0.9616	0.9586	0.9677	0.9820	0.9412	0.9581	0.9623
Jiang2022 (Jiang et al. 2022)	0.9599	0.9689	0.9620	0.9591	0.9724	0.9772	0.9576	0.9627	0.9650
Ours	**0.9709**	**0.9737**	**0.9695**	**0.9666**	**0.9795**	**0.9880**	**0.9584**	**0.9710**	**0.9722**

Table 3: The average MS_{con}^{ori} results between 9 representative methods on 8 benchmark datasets. The best result in each row is highlighted in bold.

In addition, Figure 8 shows the visual compression performance comparisons of four methods on "I05". As seen, our hmJND-Net-guided compression results provide nearly the same visual quality as the lossless coding schemes, but our hmJND-guided approach saves the bit-rate by 28.4016% on average.

5 Conclusion

In this paper, we present an end-to-end visual redundancy forecasting network based on the homologous multimodal learning. Specifically, we consider three prior modalities to enrich the training data from different perspectives. To effectively utilize the multimodal information, a homologous-modality prior fusion (HMPF) module is developed to fuse the prior modalities feature based on summation enhancement and subtractive offset. In addition, we explore a homologous-modality feature alignment (HMFA) module to align the intermediate features from different domains based on a self-attention driven encoder-decoder module. Experimental results on eight different benchmark datasets demonstrate the superiority of the method qualitatively and quantitatively. We believe that the proposed multimodal-driven approach will help to improve the quality of service for the image compression and processing systems.

References

Bae, S.-H.; Kim, J.; and Kim, M. 2016. HEVC-based perceptually adaptive video coding using a DCT-based local distortion detection probability model. *IEEE Transactions on Image Processing*, 25(7): 3343–3357.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations (ICLR)*, 1469–1479.

Chen, Z.; and Wu, W. 2019. Asymmetric foveated just-noticeable-difference model for images with visual field inhomogeneities. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11): 4064–4074.

Cheng, X.; Fu, Z.; and Yang, J. 2020. Zero-shot image super-resolution with depth guided internal degradation learning. In *Springer European Conference on Computer Vision (ECCV)*, 265–280.

Chou, C.-H.; and Li, Y.-C. 1995. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(6): 467–476.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141.

Huang, Q.; Wang, H.; Lim, S. C.; Kim, H. Y.; Jeong, S. Y.; and Kuo, C.-C. J. 2017. Measure and prediction of HEVC perceptually lossy/lossless boundary QP values. In *IEEE Data Compression Conference (DCC)*, 42–51.

Jiang, Q.; Liu, Z.; Wang, S.; Shao, F.; and Lin, W. 2022. Towards Top-Down Just Noticeable Difference Estimation of Natural Images. *IEEE Transactions on Image Processing*, 1608–1612.

Jin, L.; Lin, J. Y.; Hu, S.; Wang, H.; Wang, P.; Katsavounidis, I.; Aaron, A.; and Kuo, C.-C. J. 2016. Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis. *Electronic Imaging*, 2016(13): 1–9.

Ki, S.; Bae, S.-H.; Kim, M.; and Ko, H. 2018. Learning-based just-noticeable-quantization-distortion modeling for perceptual video coding. *IEEE Transactions on Image Processing*, 27(7): 3178–3193.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *SPIE Journal of electronic imaging*, 19(1): 1116–1137.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3.

Liu, A.; Lin, W.; Paul, M.; Deng, C.; and Zhang, F. 2010. Just noticeable difference for images with decomposition model for separating edge and textured regions. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(11): 1648–1652.

Liu, H.; Zhang, Y.; Zhang, H.; Fan, C.; Kwong, S.; Kuo, C.-C. J.; and Fan, X. 2019. Deep learning-based picture-wise just noticeable distortion prediction model for image compression. *IEEE Transactions on Image Processing*, 29: 641–656.

Liu, X.; Zhan, X.; and Wang, M. 2020. A novel edge-pattern-based just noticeable difference model for screen

content images. In *IEEE International Conference on Signal and Image Processing (ICSIP)*, 386–390.

Liu, Y.; Liu, J.; Argyriou, A.; Ma, S.; Wang, L.; and Xu, Z. 2021a. 360-Degree VR Video Watermarking Based on Spherical Wavelet Transform. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(1): 1–23.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision (ICCV)*, 10012–10022.

Min, X.; Zhai, G.; Zhou, J.; Farias, M. C.; and Bovik, A. C. 2020. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing*, 29: 6054–6068.

Ni, Z.; Ma, L.; Zeng, H.; Chen, J.; Cai, C.; and Ma, K.-K. 2017. ESIM: Edge similarity for screen content image quality assessment. *IEEE Transactions on Image Processing*, 26(10): 4818–4831.

Peterson, H. A.; and Ahumada Jr, A. J. 1992. Luminance-model-based DCT quantization for color image compression. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, 365–374.

Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Elsevier Signal Processing: Image Communication*, 30: 57–77.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7479–7489.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 12179–12188.

Seo, S.; Ki, S.; and Kim, M. 2020. A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7): 2602–2616.

Shao, H.; and Wang, Y. 2021. Generative image inpainting with salient prior and relative total variation. *Elsevier Journal of Visual Communication and Image Representation*, 79: 101–110.

Sheikh, H. 2005. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality. Accessed: 2022-02-03.

Shen, X.; Ni, Z.; Yang, W.; Zhang, X.; Wang, S.; and Kwong, S. 2020. Just noticeable distortion profile inference: a patch-level structural visibility learning approach. *IEEE Transactions on Image Processing*, 30: 26–38.

Tian, T.; Wang, H.; Kwong, S.; and Kuo, C.-C. J. 2021. Perceptual image compression with block-level just noticeable difference prediction. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(4): 1–15.

Wang, H.; Gan, W.; Hu, S.; Lin, J. Y.; Jin, L.; Song, L.; Wang, P.; Katsavounidis, I.; Aaron, A.; and Kuo, C.-C. J. 2016. MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset. In *IEEE International Conference on Image Processing (ICIP)*, 1509–1513.

Wang, H.; Wang, L.; Hu, X.; Tu, Q.; and Men, A. 2014. Perceptual video coding based on saliency and just noticeable distortion for H. 265/HEVC. In *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 106–111.

Wang, M.; Xu, Z.; Liu, X.; Xiong, J.; and Xie, W. 2022. Perceptually Quasi-Lossless Compression of Screen Content Data via Visibility Modeling and Deep Forecasting. *IEEE Transactions on Industrial Informatics*, 6865–6875.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 606–615.

Wei, Z.; and Ngan, K. N. 2009. Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(3): 337–346.

Wu, J.; Li, L.; Dong, W.; Shi, G.; Lin, W.; and Kuo, C.-C. J. 2017. Enhanced just noticeable difference model for images with pattern complexity. *IEEE Transactions on Image Processing*, 26(6): 2682–2693.

Wu, J.; Shi, G.; Lin, W.; Liu, A.; and Qi, F. 2013. Just noticeable difference estimation for images with free-energy principle. *IEEE Transactions on Multimedia*, 15(7): 1705–1710.

Xie, W.; Huang, T.; and Wang, M. 2022. MNSRNet: Multimodal Transformer Network for 3D Surface Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12703–12712.

Yang, H.; Fang, Y.; and Lin, W. 2015. Perceptual quality assessment of screen content images. *IEEE Transactions on Image Processing*, 24(11): 4408–4421.

Yang, X. K.; Lin, W. S.; Lu, Z. K.; Ong, E. P.; and Yao, S. S. 2005. Just noticeable distortion model and its applications in video coding. *Elsevier Signal Processing: Image Communication*, 20(7): 662–680.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI Conference on Artificial Intelligence*, volume 35, 3208–3216.

Zhai, G.; and Min, X. 2020. Perceptual image quality assessment: a survey. *Springer Science China Information Sciences*, 63(11): 1–52.

Zhang, X.; Wang, S.; Gu, K.; Lin, W.; Ma, S.; and Gao, W. 2016. Just-noticeable difference-based perceptual optimization for JPEG compression. *IEEE Signal Processing Letters*, 24(1): 96–100.

Zhang, X. H.; Lin, W. S.; and Xue, P. 2005. Improved estimation for just-noticeable visual distortion. *Elsevier Signal Processing*, 85(4): 795–808.