DDJND: Dual Domain Just Noticeable Difference in Multi-Source Content Images with Structural Discrepancy

Miaohui Wang[2], Zhenming Li[1], Wuyuan Xie[1]*

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060
[2]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060
wang.miaohui@gmail.com, 2310273052@email.szu.edu.cn, wuyuan.xie@gmail.com

Abstract

Most existing just noticeable difference (JND) methods primarily integrate specific masking effects in a single domain. However, these single-domain JND methods struggle with the structural discrepancies in multi-source content images, limiting their effectiveness in visual redundancy estimation. To address this issue, we propose a dual domain encoder that combines spatial and frequency features to comprehensively capture visual patterns. Our design includes spatial pattern balance and frequency detail correction modules to balance global and local patterns and correct low- and high-frequency distributions. Additionally, we develop a dual domain decoder to effectively extract multi-scale pattern redundancies and integrate them with detail redundancies in the frequency domain. Experiments demonstrate the effectiveness and robustness of our proposed method in handling structural discrepancies in multi-source content images.

1 Introduction

In the field of visual perception, just noticeable difference (JND) refers to the minimum difference in changes that the human visual system (HVS) can perceive (Menon et al. 2024). The performance and efficiency of JND estimation are closely related to the image content (Wu et al. 2017). Multi-source content (MSC) images may contain visual elements of multiple styles, scenes, and patterns (Wang et al. 2024a). These elements with multiple attributes may result in structural or property discrepancies, such as boundary artifacts, brightness contrast differences, color distortion, texture discontinuity, resolution and transparency differences, and perspective distortion, which affect the performance of existing JND methods for visual redundancy estimation.

To solve the problem of structural discrepancy, we first consider the sources of MSC images, which generally include three main content types: *1) AI-generated content (AIGC).* This is achieved through large-scale learning of the features of natural content images. AI algorithms capture and simulate various properties of natural content images, such as color distribution, lighting effects, and texture details, to reproduce the natural scene as realistically as possible. However, due to the limitations of the generation algorithms, AIGC images may still have structural problems such as texture discontinuity and inconsistent lighting. *2) Screen content (SC).* These images (*e.g.*, screen sharing, screen shots, charts, and text images) usually have high contrast and clarity, and contain a large number of straight lines, edges, and block areas (Wang et al. 2022a, 2021). When modeling these images, their unique structural characteristics and high-resolution requirements must be considered to ensure the fidelity of details. *3) User-generated content (UGC).* These images are usually taken or produced by users through personal devices such as smartphones and digital cameras. Due to the diversity of shooting equipment and environments, the quality and style of UGC images vary differently, as shown in Figure 1.

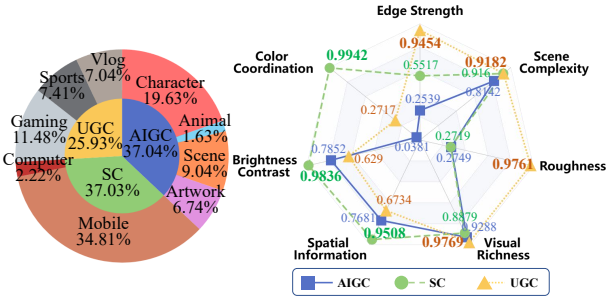

Figure 1: Statistics of anisotropy-related feature values of multi-source content (MSC) images including AIGC images, SC images, and UGC images in the five datasets.

Existing frequency domain-based methods (Wei and Ngan 2009; Bae and Kim 2013, 2014; Wang et al. 2020) usually implement JND estimation for natural content image by using discrete cosine transform (DCT), while some methods (Jiang et al. 2022) also use Karhunen-Loéve transform to effectively capture the global structural information that the HVS cannot perceive in the spatial domain. Frequency domain-based methods usually rely on transform processing of images, which assumes that the input images have certain statistical or structural properties. However, these assumptions may not be applicable to images from different content sources with complex structural discrepancies, as they may differ from the statistical or structural properties on

*Corresponding author: Wuyuan Xie

which these assumptions rely. In contrast, spatial domain-based methods process images directly at the pixel level, which can theoretically be more flexible in dealing with various visual features. However, these methods often rely on individual processing of image pixels, which may lead to a significant increase in computational complexity, especially when processing high-resolutions.

In summary, existing single-domain methods have many limitations: 1) *Single-domain methods are difficult to fully capture important information.* For example, frequency domain-based methods can effectively capture the periodic patterns and frequency distribution, but may overlook local details and texture information. Although spatial domain-based methods can accurately reflect the edges and details, they are difficult to capture the global spectral features. 2) *Single-domain methods are limited in their ability to comprehensively utilize feature information from both the frequency and spatial domains.* When processing data, single domain methods have high adaptability and optimization capabilities for specific domains, but often cannot fully utilize all available feature information in other domains. 3) *Single-domain methods lack the ability to fully simulate the multi-level processing by the HVS.* Our human eyes are empirically believed to be more influenced by various visual perceptions (*e.g.*, horizontal and vertical directions (Furmanski and Engel 2000), brightness contrast (Goodyear and Menon 1998), and stereopsis (Kane, Guan, and Banks 2014)), which requires comprehensive consideration of the image content and multi-level features when estimating JND. Therefore, single domain methods may perform poorly in terms of efficiency and accuracy when processing MSC images, and how to address these constraints remains an underexplored direction.

To address the structural discrepancies in MSC images, we propose a dual-domain approach that combines frequency and spatial domain features, thereby improving the accuracy of JND estimation. The main contributions of our work can be summarized as follows:

- We design a dual-domain network to extract image features from both frequency and spatial domains. The frequency and spatial feature encoders guide each other through knowledge distillation, enabling them to share and integrate feature information effectively.
- We introduce a spatial pattern balance module that processes global and local pattern features in the spatial domain, allowing our model to understand the broader context while also focusing on local areas. Additionally, we develop a frequency detail correction module that handles low- and high-frequency detail features in the frequency domain, effectively balancing and adjusting attention to various frequency features.
- We design a multi-scale and multi-domain visual information decoding module, utilizing the more interpretable Kolmogorov-Arnold network (KAN). The nonlinear learnable methods in KAN provide a better explanation for the multi-level processing of visual information across spatial and frequency domains, leading to more reasonable and reliable adaptive estimation.

2 Related Work

Transform Domain-based Methods

Compared with spatial domain-based methods, relatively few methods have been proposed to process original images in the transform domain. (Wei and Ngan 2009) proposed a new spatio-temporal JND model based on discrete cosine transform (DCT). (Bae and Kim 2013) made the previous model more accurate in processing high-frequency details and low-frequency smooth areas by weighting different frequency components in the DCT domain. To adapt to a variety of block sizes, (Bae and Kim 2014) combined the consideration of different blocks and frequency characteristics with the DCT. In addition, (Wang et al. 2020) took into account the combined effect of multiple masking effects to more accurately reflect the characteristics of the HVS. (Jiang et al. 2022) utilized the convergence characteristics of the energy of the Karhunen-Loéve transform coefficients to adaptively and effectively capture the microstructure information that the HVS cannot perceive in the spatial domain, avoiding direct modeling of the complex interactions of masking effects. These frequency domain-based methods usually rely on certain statistical or structural characteristics, without taking into account the structural discrepancies.

Spatial Domain-based Methods

Many existing methods are implemented in the spatial domain, including many handcrafted methods (Wu et al. 2017) and learning-based methods (Xie et al. 2023b). In order to calculate the directional diversity within a local area in the spatial domain, (Wu et al. 2017) incorporated pattern complexity into the model calculation. (Shen et al. 2020) approached human visibility perception by decomposing image patches in the spatial domain into luminance, contrast and structure, and then incorporating structure degradation. To better adapt to the perceptual characteristics of the HVS, (Huang, Zhang, and Wang 2023) distinguished the certainty feature from the uncertainty information and further processed the uncertainty. (Xie et al. 2023b) comprehensively utilized the luminance, color, and segmentation modal information to explore the patterns of structural discrepancies in composite images. Additionally, multiple image modality information have been introduced to explore potential complementary effects (Xie et al. 2023a; Wang et al. 2024b). To implicitly predict the JND map to characterize complex masking effects, (Jiang et al. 2024) conceptualizes the JND as a difference map between the original spatial image and its corresponding critical perceptually lossless image. Spatial domain-based methods mainly focus on the relationship between local pixels and tend to ignore the global statistical characteristics, which may have disadvantages in global structure and noise resistance.

Summary

Inspired by previous studies, we explore the JND modeling from a dual domain perspective. In the spatial domain, local correlations and repetitive patterns are crucial, and local image structures in terms of orientation, brightness contrast, and sensitivity can be learned. In the frequency do-

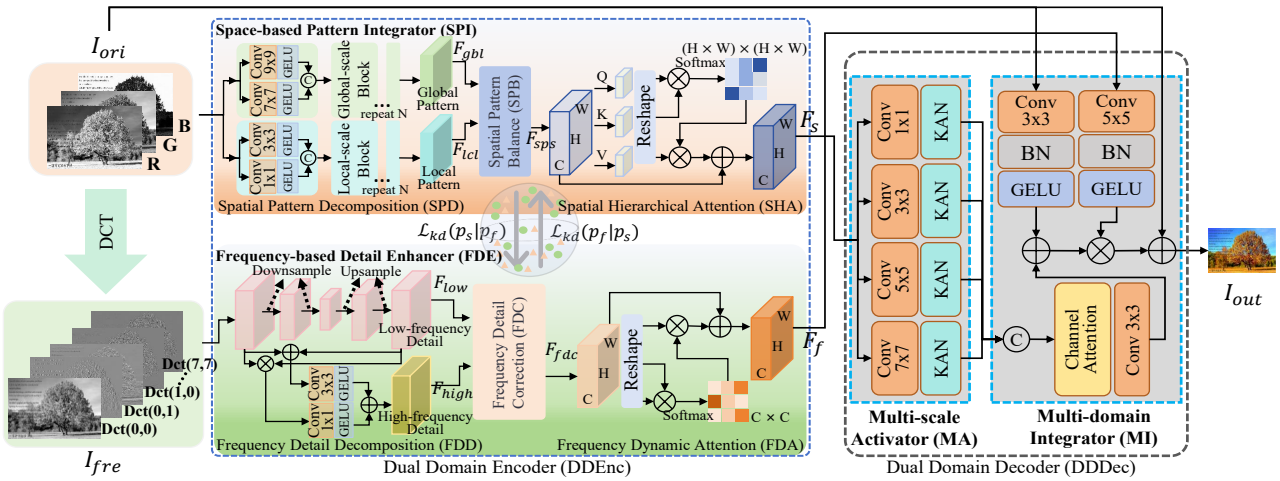

Figure 2: Pipeline of the proposed dual domain just noticeable difference (DDJND) model. This model consists of the dual domain encoder (DDEnc) that includes the space-based pattern integrator (SPI) and the frequency-based detail enhancer (FDE), as well as the dual domain decoder (DDDec) that includes the multi-scale activator (MA) and the multi-domain integrator (MI).

main, high-frequency edge textures and low-frequency flat area features can be directly processed, but the capture of subtle structures is not accurate enough. Therefore, we comprehensively utilize the information in the frequency and spatial domains for complementary learning of feature representations, so as to improve the accuracy and robustness of visual redundancy estimation.

3 Proposed Dual Domain Framework

In this section, we provide a detailed description of the proposed a dual domain just noticeable difference (DDJND) model, which is divided into a dual domain encoder (DDEnc) and a dual domain decoder (DDDec).

Overview

Problem formulation. Our overall task requires the use of mutual knowledge distillation to comprehensively optimize the interaction of spatial- and frequency-domain information to more comprehensively guide the DDJND model for joint training. The optimization process of alternating distillation loss combines the advantages of spatial and frequency domains. In addition, optimizing the feature loss that focuses on pixel level similarity helps to preserve detailed structures. By combining the alternating distillation loss and the feature loss, we can minimize a specially designed loss function $\mathcal{L}_{all}$, which is expressed as

$$\mathcal{L}_{all} = \lambda_1 \cdot \mathcal{L}_{kd} + \lambda_2 \cdot \mathcal{L}_{fea}, \tag{1}$$

where λ_1 and λ_2 represent the weights of two loss functions, which are set to 1. $\mathcal{L}_{kd}$ represents the alternating distillation loss, which is specifically defined as (21). $\mathcal{L}_{fea}$ represents the feature loss, which is defined as mean square error between the network output I_{out} and the ground-truth I_{gt}. By optimizing this objective function, our DDJND model achieves better results in balancing global structure and local perception, thereby improving the estimation performance in MSC images.

Network Architecture. The network architecture of our DDJND model is shown in Figure 2, which can be formulated as follows:

$$I_{out} = DDJND\left(I_{ori}, I_{fre}\right), \tag{2}$$

where I_{ori} denotes the original image, I_{fre} denotes the frequency domain image obtained by applying DCT to I_{ori}, and I_{out} represents the estimated JND map (*i.e.*, visual redundancy).

The entire DDJND framework can be divided into the DDEnc and DDDec modules. First, the spatial and frequency domain features are fed into the DDEnc module for the mutual knowledge distillation. By sharing knowledge between the spatial and the frequency domains, joint feature representations can be learned, which makes up for the shortcomings of a single-domain model. This process can be expressed as follows:

$$F_s, F_f = DDEnc\left(I_{ori}, I_{fre}\right). \tag{3}$$

After the DDEnc module, two important features are obtained: One is the spatial features compensated by frequency domain information, and the other is the frequency domain features compensated by spatial information. We aim to integrate and decode these compensation features, and finally perform a reasonable range regularization process to obtain I_{out}, which is defined as follows:

$$I_{out} = \theta\left(DDDec\left(F_s, F_f\right)\right), \tag{4}$$

where θ represents a regularization process.

Dual Domain Encoder (DDEnc)

In this section, we describe the important components of our DDEnc, including the space-based pattern integrator (SPI) and the frequency-based detail enhancer (FDE). The FDE module is composed of frequency detail decomposition (FDD), frequency detail correction (FDC) and frequency dynamic attention (FDA) modules, while the SPI module

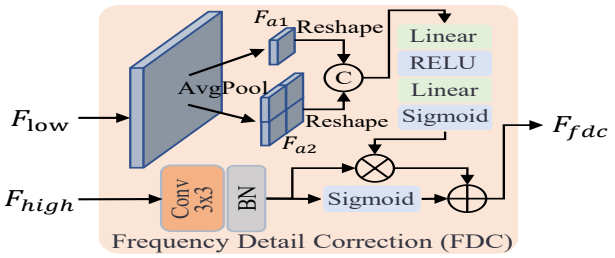

Figure 3: Frequency detail correction (FDC) module. The input is the low-frequency feature F_{low} and high-frequency feature F_{high} obtained through the frequency detail decomposition (FDD) module, and the output is the corrected fusion feature F_{fdc}.

is composed of spatial pattern decomposition (SPD), spatial pattern balance (SPB) and spatial hierarchical attention (SHA) modules.

Frequency Detail Decomposition (FDD). To extract frequency domain features, we first perform DCT on each image. Specifically, we first divide each input image into non-overlapping 8×8 image blocks. Subsequently, a two-dimensional DCT is performed on each image block, so that the pixel values are decomposed into different frequency components. When the DCT coefficient $F_{(u,v)}$ is set to zero, the certain frequency component is selectively ignored, and then the inverse DCT is performed to obtain $\tilde{f}_{(i,j)}$:

$$\tilde{f}_{(i,j)} = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1} C_{(u)}C_{(v)}F_{(u,v)} cos\left[\frac{(2i+1)\pi}{2N}u\right] cos\left[\frac{(2j+1)\pi}{2N}v\right], \quad (5)$$

where $C_{(u)}$ and $C_{(v)}$ are normalization coefficients. When setting $F_{(u,v)}$=0, the residual before and after the performing (5) is calculated as the frequency domain input I_{fre}:

$$I_{fre} = \left| f_{(i,j)} - \tilde{f}_{(i,j)} \right|. \quad (6)$$

To further extract the low-frequency details, we use a 5×5 convolution and a maximum pooling to reduce its size, and then restore the original image size by upsampling. The low-frequency feature F_{low} can be extracted by

$$F_{low} = Upsample^2\left(ConvMaxPoolReLu_5^2\left(I_{fre}\right)\right), \quad (7)$$

where $Upsample^2$ represents performing two upsamplings, $ConvMaxPoolReLu_5^2$ represents performing two 5×5 convolutions, max pooling, and `ReLU` activation operations.

To obtain high-frequency information, we perform addition compensation and multiplication scaling operations on F_{low} and I_{fre}. After passing them through 1×1 and 3×3 convolutions and `GELU` activation layers, they are added together to obtain the high-frequency feature F_{high}:

$$F_{high} = CG_1(F_{low} \otimes I_{fre}) \oplus CG_3(F_{low} \oplus I_{fre}), \quad (8)$$

where CG_k indicates the use of k×k convolution and a `GELU` activation function. $\oplus$ and $\otimes$ represent element-wise sum and element-wise multiplication, respectively.

Frequency Detail Correction (FDC). To adjust and optimize the decomposed high- and low-frequency features, we design this FDC module to dynamically adjust the weights of different frequency components according to image contents, as shown in Figure 3. F_{low} and F_{high} are used as the inputs to the FDC module to adjust the importance of different frequency components, helping our DDJND model to perceive different scales and textures. Specifically, without considering the batch size, F_{low} is subjected to two average poolings to obtain $F_{a1} \in \mathbb{R}^{C\times1\times1}$ and $F_{a2} \in \mathbb{R}^{C\times2\times2}$, which are reshaped into $\mathbb{R}^{C}$ and $\mathbb{R}^{4C}$ and then concatenated to form $\mathbb{R}^{5C}$. After passing through the last `Sigmoid` layer in the multilayer perceptron (MLP), a weight matrix $\in \mathbb{R}^{C\times1\times1}$ of low-frequency features is obtained, which is then multiplied with the high-frequency matrix obtained after passing F_{high} through a convolution and regularization layer to obtain a frequency fusion feature map $\in \mathbb{R}^{C\times H\times W}$. F_{high} is obtained by a `Sigmoid` layer, and the weight matrix $\in \mathbb{R}^{C\times H\times W}$ of the high-frequency detail features is added to the frequency fusion feature map to obtain the detail correction feature matrix $F_{fdc} \in \mathbb{R}^{C\times H\times W}$, which is defined as:

$$\begin{aligned} F_{fdc} = MLP\left(Cat(AvgPool_1(F_{low}), AvgPool_2(F_{low}))\right) \otimes \\ BN\left(Conv_3(F_{high})\right) \oplus Sigmoid\left(BN(Conv_3(F_{high}))\right), \end{aligned} \quad (9)$$

where Cat represents the concatenation operation. $AvgPool_k$ represents the pooling layer with the image size k×k after average pooling. $Conv_3$ represents the layer using a 3×3 convolution kernel. BN represents the batch normalization.

Frequency Dynamic Attention (FDA). To improve the quality of feature representations, we further design a FDA module that automatically learns and adjusts the weights between different channels by strengthening interdependent feature mappings. Specifically, we reshape F_{fdc} into size $C \times N$ and then perform the matrix multiplication between F_{fdc} and its transpose, aiming to capture the interaction information between different channels. Finally, we apply the `Softmax` for normalization to obtain the channel attention map $CA \in \mathbb{R}^{C\times C}$. The FDA process is defined as follows:

$$CA_{ij} = \frac{exp\left(F_{fdc}^i \cdot (F_{fdc}^j)^T\right)}{\sum_{i=1}^{C}\sum_{j=1}^{C} exp\left(F_{fdc}^i \cdot (F_{fdc}^j)^T\right)}, \quad (10)$$

where CA_{ij} is used to quantify the influence of the i-th row channel on the j-th column channel in F_{fdc}. We perform a matrix multiplication on CA_{ij} and F_{fdc}, and reshape the result into $\mathbb{R}^{C\times H\times W}$. Then we introduce an adjustable parameter β like position attention, and add it to F_{fdc} to get the final output $F_f \in \mathbb{R}^{C\times H\times W}$, which is defined as:

$$F_f = \beta \times CA_{ij} \times F_{fdc} + F_{fdc}. \quad (11)$$

Spatial Pattern Decomposition (SPD). The analysis of global and local patterns in the spatial domain is critical for the subsequent balancing of spatial patterns and adjustment of attentions. To decompose the global and local patterns, we concatenate the outputs of the parallel 9×9 and 7×7 convolutional layers as a global-scale block (GB) to extract global

pattern features, and concatenate the outputs of the parallel 3×3 and 1×1 convolutional layers as a local-scale block (LB) to extract local pattern features. Extracting global pattern features can be defined as (12). Specifically, GB and LB are repeated N times and $F_{gbl} = F_{gbl}^{N}$.

$$F_{gbl}^{k} = \begin{cases} I_{ori}, & if\ k = 0 \\ Cat\left(GB_9(F_{gbl}^{k-1}), GB_7(F_{gbl}^{k-1})\right), & else\ if\ 1 \le k \le N \end{cases}, \quad (12)$$

where Cat represents the concatenation operation according to the channel dimension. GB_k represents a global scale block containing a convolutional layer of size k×k and a GELU activation function. Similarly, extracting local pattern features can be defined as (13) and $F_{lcl} = F_{lcl}^{N}$.

$$F_{lcl}^{k} = \begin{cases} I_{ori}, & if\ k = 0 \\ Cat\left(LB_3(F_{lcl}^{k-1}), LB_1(F_{lcl}^{k-1})\right), & else\ if\ 1 \le k \le N \end{cases}, \quad (13)$$

where LB_k represents a local scale block containing a convolutional layer with the kernel size of k×k and a GELU activation function.

Spatial Pattern Balance (SPB). To enable the DDJND to learn structural patterns, we need to balance F_{gbl} and F_{lcl} obtained in the SPD module. We expect our DDJND to dynamically adjust the global and local features according to the actual image content, so that the DDJND can adaptively balance these two features with different contents. The designed details of the spatial pattern balance module are shown in Figure 4. Specifically, N spatial separable convolutional blocks (SSCBs) are used to extract pattern features more finely from different dimensions, which can be defined as:

$$\left\{F_g^i, F_l^i\right\} = \begin{cases} \left\{F_{gbl}, F_{lcl}\right\}, & if\ i = 0 \\ SSCB^i\left(F_g^{i-1}, F_l^{i-1}\right), & else\ if\ 1 \le i \le N \end{cases}. \quad (14)$$

To enhance the expression of local detail features and make it easier for our DDJND model to identify regions of different hierarchical structures, two convolutional layers with different sizes and one concatenation are used to adjust the pattern feature $F_{gl}^{'}$, which is specifically defined as:

$$\begin{aligned} F_{gl}^{'} &= Cat\left(CG_1(F_{gl}), CG_5(F_{gl})\right), \\ \text{s.t.} \quad F_{gl} &= Cat(F_g^N, F_l^N) \end{aligned}, \quad (15)$$

where CG_k represents a network layer containing a convolution layer with the kernel size of k×k and a GELU activation function. $F_{gl}^{'}$ is element-wise multiplied with its own enhanced specific features, which can be expressed by

$$F_{sps} = F_{gl}^{'} \otimes CG_3\left(F_{gl}^{'}\right), \quad (16)$$

where $\otimes$ represents the element-wise multiplication.

Spatial Hierarchical Attention (SHA). Without considering the batch size, we first input F_{sps} into a convolutional layer to generate three new feature maps Q, K and V, respectively. $\{Q, K, V\} \in \mathbb{R}^{C\times H\times W}$, which is defined as follows:

$$\{Q, K, V\} = Conv\left(F_{sps}\right). \quad (17)$$

We reshape the three tensors Q, K, and V into the size of $N \times C$, where $N = H \times W$. Then we perform matrix multiplication on Q and K^T, and apply the Softmax function to normalize the matrix multiplication results to obtain the spatial hierarchical attention graph $PA_{ij} \in \mathbb{R}^{N\times N}$. The correlation degree between different spatial positions is defined as follows:

$$PA_{ij} = \frac{exp(Q_i \cdot K_j^T)}{\sum_{i=1}^{N}\sum_{j=1}^{N} exp(Q_i \cdot K_j^T)}, \quad (18)$$

where PA_{ij} is used to measure the influence of the i-th row position of the tensor Q on the j-th column position of the tensor K. The closer the feature representations of two positions are, the higher the correlation between them. Then we perform the matrix multiplication on PA_{ij} and V, and reshape the product result to $\mathbb{R}^{C\times H\times W}$. To adjust the weight of the output features, we further introduce an adjustable parameter γ and perform a sum operation with F_{sps} to obtain the final output $F_s \in \mathbb{R}^{C\times H\times W}$, which is defined as follows:

$$F_s = \gamma \times PA_{ij} \times V + F_{sps}, \quad (19)$$

where the initial value γ is set to 0.5, and is continuously adjusted to a reasonable weight during training.

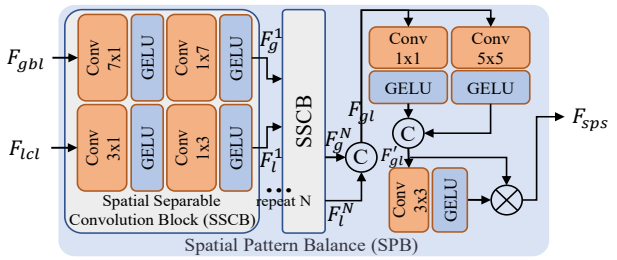

Figure 4: Spatial pattern balance (SPB) module. The input is the global pattern feature F_{gbl} and local pattern feature F_{lcl} obtained after the spatial pattern decomposition (SPD) module, and the output is the balanced fusion feature F_{sps}.

Dual Domain Knowledge Distillation. We introduce a new knowledge distillation framework between frequency- and spatial-domain features. This framework consists of two core modules: one is dedicated to processing frequency domain features and the other is dedicated to analyzing spatial features.

The traditional knowledge distillation method usually transfers the knowledge from the teacher to the student network, while the student network passively receives the knowledge. However, this unidirectional knowledge transfer method may limit the learning ability of students online, making it difficult for them to surpass the teacher network. The mutual knowledge distillation strategy (Kim et al. 2021; Liu et al. 2023; Yang et al. 2023) breaks this unidirectionality, allowing teachers and students to learn from each other and transfer knowledge to each other.

Specifically, we adopt knowledge distillation loss during the dual domain characteristics interaction. Firstly, we calculate the probability distributions p_s and p_f for the spatial and frequency domain models:

$$p_{s,f} = Softmax\left(\frac{F_{s,f}}{T}\right), \quad (20)$$

where T represents the temperature coefficient that controls the `Softmax` function. We alternate the probability distributions as the loss calculations for teacher and student:

$$\mathcal{L}_{kd} = KL\left(p_s, p_f\right) \cdot \frac{T^2}{B} + KL\left(p_f, p_s\right) \cdot \frac{T^2}{B}, \quad (21)$$

where B represents the batch size and $KL(\cdot)$ represents the Kullback-Leibler divergence. The goal of this loss function is to make the output of the student and teacher model in terms of distribution, thereby achieving dual domain feature interaction.

Dual Domain Decoder (DDDec)

We integrate F_s and F_f into the dual domain feature decoding process. Specifically, the dual domain decoder consists of two parts: a multi-scale activator (MA) and a multi-domain integrator (MI).

To enhance the spatial domain features, we adopt the more interpretable KAN as the key part of the MA module. The non-linear learnable approach in KAN can better explain the multi-level processing of visual information in spatial pattern distribution. MA activates features of different scales, which is used to coordinate the decoding of frequency domain features for the subsequent MI integration. Specifically, each convolution layer $Conv_i (i \in \{1, 3, 5, 7\})$ followed by a KAN processes the spatial pattern integration features, and finally outputs the result set CK of all scale activations:

$$CK = \left\{KAN\left(Conv_i\left(F_s\right)\right)\right\}, i \in 1, 3, 5, 7. \quad (22)$$

MI decodes the output of MA by integrating the frequency domain detail enhancement features. Specifically, we concatenate the activation results of all scales of spatial visual features according to the channel dimension, and dynamically adjust the importance of different channels through a channel attention layer and a 3×3 convolutional layer to capture key information. We pass the frequency domain detail enhancement features and original image into convolutional layers. Then the original image features are summed with the spatial multi-scale activation features, multiplied with the frequency domain detail enhancement features, and finally summed with the original image features:

$$\begin{aligned} I_{out} = &(Conv_3(CA(Concat(CK))) \\ &\oplus GELU(BN(Conv_3(I_{ori})))) \\ &\otimes GELU(BN(Conv_5(F_f))) \oplus I_{ori} \end{aligned}, \quad (23)$$

where CA represents the channel attention module. *Concat* represents the concatenation operation. $Conv_k$ means that the convolution layer uses a convolution kernel size of k×k. *GELU* is an activation function. $\oplus$ means element-wise sum and $\otimes$ means element-wise multiplication.

4 Experimental Validations

Experimental Protocols

Comparison methods. We compare the proposed DDJND with several existing transform-domain and spatial-domain based methods, including *Wei2009* (Wei and Ngan 2009), *Wu2013* (Wu et al. 2013), *Ki2018* (Ki et al. 2018), *Shen2020* (Shen et al. 2020), *Jiang2022* (Jiang et al. 2022) and *Jiang2024* (Jiang et al. 2024).

Datasets. To verify the effectiveness and robustness, we conduct qualitative and quantitative experiments on three JND datasets, including *MCL-JCI* (Jin et al. 2016), *SHEN2020* (Shen et al. 2020) and *KonJND1k* (Lin et al. 2022). Also, we employ different content datasets, including the AIGC dataset (*DiffusionDB* (Wang et al. 2022b)), the SC dataset (*SCID* (Yang, Fang, and Lin 2015) and *SIQAD* (Ni et al. 2017)), and the UGC dataset (*Youtube UGC* (Wang, Inguva, and Adsumilli 2019)).

Evaluation metrics. Since each JND dataset has a certain bias and tendency, we calculate the average noise level of the first JND point (Lin and Ghinea 2022). Specifically, the injected noise levels of their first JND points are as follows: the average mean square error (MSE) of the *MCL-JCI* dataset is 26.836, that of the *SHEN2020* dataset is 16.121, and that of the *KonJND1k* dataset is 14.091. For AIGC, SC, and UGC datasets without ground-truth images, we control the JND map predicted by each JND model to the same noise level (MSE=100). Under the above conditions of injected noise levels, we use structural similarity index measure (SSIM) (Wang et al. 2004) and learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018) to effectively evaluate the overall performance of predicting visual redundancy.

Implementation details. In the SPD module, we set the number of layers N to 5. The key parameters λ_1 and λ_2 in the loss optimization are both set to 1. In all SSCBs, the convolution kernel sizes used for F_{gbl} are 7×1 and 1×7, while the convolution kernel sizes used for F_{lcl} are 3×1 and 1×3. Our model is implemented in *Pytorch* on the SHEN2020 dataset. During the data processing stage, each image is cropped into 224×224 and the ratio of the training set, validation set, and test set is 8:1:1. During the training and validation phases, our DDJND is trained for 40 epochs on 1 *NVIDIA GeForce RTX 3090* with the batch size of 4. The Adam optimizer is used with an initial learning rate of 1*e*-5.

Ablation Study

To verify the effectiveness of each module in our proposed DDJND, we have conducted ablation experiments on two sub-modules in the DDEnc (including SPI and FDE modules) and two sub-modules in the DDDec (including MA and MI modules), as shown in Table 1. Specifically, we use a 3×3 convolution as an alternative to the SPI, FDE, and MA modules, respectively. The alternative to the MI module is to concatenate the output results of the MA module according to the channel dimension to form a spatial activation feature, and then multiply it by the frequency domain dynamic feature F_f and add the associated result to the original image feature to produce the final output.

Qualitative Results

After injecting noise, flatter areas should tolerate less noises and their brightness in the JND map should be darker due to its greater susceptibility to visual perception. More complex areas that are often less sensitive to cause visual perception should tolerate more noises. Based on this knowledge, we can reasonably evaluate the estimation accuracy of a JND model of visual redundancy prediction. Figure 5 shows the

SPI	FDE	MA	MI	SSIM↑	LPIPS↓
-	-	-	-	0.8583	0.3296
✓	-	-	-	0.8641	0.3265
-	✓	-	-	0.8607	0.3254
-	-	✓	-	0.8740	0.3183
-	-	-	✓	0.9191	0.2594
✓	✓	-	-	0.8910	0.2615
✓	-	✓	-	0.9056	0.2332
-	✓	-	✓	0.8932	0.2417
-	-	✓	✓	0.9132	0.2242
✓	✓	✓	✓	0.9283	0.2212

Table 1: Ablation experiments of the SPI, FDE, MA, and MI modules on the Shen2020 dataset. '-' denotes the current module is disabled, while '✓' the current module is enabled.

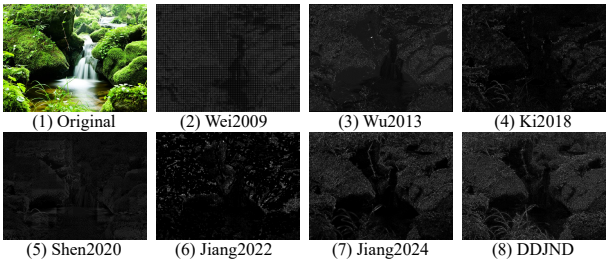

Figure 5: Visual comparison of JND predictions over seven approaches.

qualitative results on seven JND methods. As seen, our proposed DDJND model has better prediction performance.

Quantitative Results

To systematically evaluate the overall performance of the proposed DDJND in estimating visual redundancy after integrating spatial patterns and frequency domain details, we have conducted quantitative experiments. The average SSIM and LPIPS results are calculated on seven JND methods and three JND datasets, as shown in Table 2. As seen, our DDJND is able to predict visual redundancy more correctly than the other methods. The main reasons can be that our method simultaneously analyzes spatial and frequency features after integrating spatial patterns and frequency details.

Cross-dataset Comparison

There are significant structural discrepancies in MSC images. To verify the robustness of our JND model on MSC images, we have additionally conducted a cross-dataset experiment including AIGC, SC, and UGC datasets. Specifically, we control the ejected noise level to MSE=100, and use SSIM and LPIPS to evaluate the overall performance of our proposed DDJND for visual redundancy estimation, as shown in Table 3. As seen, our method still has better prediction results on different image contents, exhibiting good adaptability and robustness.

Methods	JND Datasets					
	MCL-JCI		SHEN2020		KonJND1k	
	SSIM↑	**LPIPS↓**	**SSIM↑**	**LPIPS↓**	**SSIM↑**	**LPIPS↓**
Wei2009	0.8618	0.3314	0.9009	0.3267	0.9076	0.2499
Wu2013	0.8584	0.3275	0.8970	0.3362	0.8997	0.2503
Ki2018	0.8625	0.3312	0.9193	0.2327	0.9178	0.2819
Shen2020	0.8544	0.3399	0.9084	0.3504	0.9036	0.2592
Jiang2022	0.8956	0.2300	0.9200	0.2703	0.9358	0.1501
Jiang2024	0.9020	0.3032	0.9251	0.1043	0.9409	0.2028
DDJND	0.9124	0.2045	0.9283	0.2212	0.9492	0.1452

Table 2: Comparison of the JND prediction in terms of the average SSIM and LPIPS on three JND datasets.

Methods	MSC Datasets					
	AIGC		SC		UGC	
	SSIM↑	**LPIPS↓**	**SSIM↑**	**LPIPS↓**	**SSIM↑**	**LPIPS↓**
Wei2009	0.8167	0.2573	0.7612	0.3212	0.7461	0.4860
Wu2013	0.8094	0.2604	0.7829	0.2982	0.7317	0.4999
Ki2018	0.8484	0.2583	0.8118	0.2871	0.7646	0.4888
Shen2020	0.8195	0.2737	0.7346	0.3431	0.7462	0.5000
Jiang2022	0.8472	0.2188	0.7463	0.2852	0.7867	0.4262
Jiang2024	0.8648	0.1941	0.8305	0.2504	0.8076	0.3964
DDJND	0.8715	0.1883	0.8432	0.2352	0.8206	0.3753

Table 3: Cross-dataset comparison of the JND prediction in terms of the average SSIM and LPIPS on three MSC image datasets (*i.e.*, AIGC, SC, and UGC).

5 Conclusion

This paper presents a dual-domain encoder based on knowledge distillation, which includes a pattern integrator and a detail enhancer. These components are designed to decompose, correct, and reallocate attention to spatial- and frequency-domain information, enabling our DDJND model to more comprehensively capture descriptive texture features related to structural discrepancies in MSC images. Our focus is on the spatial pattern balance and frequency detail correction modules within the encoder, thereby improving adaptability to structural discrepancies. Additionally, we introduce a dual-domain decoder that effectively extracts redundant regions at different scales while collaboratively perceiving both coarse- and fine-grained detail redundancy in the frequency domain, thus effectively addressing structural discrepancies. Qualitative and quantitative experiments on several JND datasets and MSC cross-datasets demonstrate the effectiveness and robustness of our proposed DDJND.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62472290 and 62372306, and in part by the Natural Science Foundation of Guangdong Province under Grants 2024A1515011972, 2023A1515011197 and 2022A1515011245.

References

Bae, S.-H.; and Kim, M. 2013. A novel DCT-based JND model for luminance adaptation effect in DCT frequency. *IEEE Signal Processing Letters*, 20(9): 893–896.

Bae, S.-H.; and Kim, M. 2014. 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*, 23(8): 3227–3240.

Furmanski, C. S.; and Engel, S. A. 2000. An oblique effect in human primary visual cortex. *Nature Neuroscience*, 3(6): 535–536.

Goodyear, B. G.; and Menon, R. S. 1998. Effect of luminance contrast on BOLD fMRI response in human primary visual areas. *Journal of Neurophysiology*, 79(4): 2204–2207.

Huang, L.; Zhang, R.; and Wang, M. 2023. Just Noticeable Difference Estimation for Screen Content Images: A Content Uncertainty-guided Approach. In *IEEE International Conference on Multimedia and Expo (ICME)*, 372–377.

Jiang, Q.; Liu, F.; Wang, Z.; Wang, S.; and Lin, W. 2024. Rethinking and Conceptualizing Just Noticeable Difference Estimation by Residual Learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Jiang, Q.; Liu, Z.; Wang, S.; Shao, F.; and Lin, W. 2022. Toward top-down just noticeable difference estimation of natural images. *IEEE Transactions on Image Processing*, 31: 3697–3712.

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.

Kane, D.; Guan, P.; and Banks, M. S. 2014. The limits of human stereopsis in space and time. *Journal of Neuroscience*, 34(4): 1397–1408.

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.

Kim, J.; Hyun, M.; Chung, I.; and Kwak, N. 2021. Feature fusion for online mutual knowledge distillation. In *IEEE International Conference on Pattern Recognition (ICPR)*, 4619–4625.

Lin, H.; Chen, G.; Jenadeleh, M.; Hosu, V.; Reips, U.-D.; Hamzaoui, R.; and Saupe, D. 2022. Large-scale crowdsourced subjective assessment of picturewise just noticeable difference. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9): 5859–5873.

Lin, W.; and Ghinea, G. 2022. Progress and opportunities in modelling just-noticeable difference (JND) for multimedia. *IEEE Transactions on Multimedia*, 24: 3706–3721.

Liu, S.; Yin, S.; Qu, L.; Wang, M.; and Song, Z. 2023. A Structure-aware Framework of Unsupervised Cross-Modality Domain Adaptation via Frequency and Spatial Knowledge Distillation. *IEEE Transactions on Medical Imaging*.

Menon, V. V.; Rajendran, P. T.; Feldmann, C.; Schoeffmann, K.; Ghanbari, M.; and Timmerer, C. 2024. JND-aware Two-pass Per-title Encoding Scheme for Adaptive Live Streaming. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2): 1281–1294.

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.

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.

Wang, H.; Yu, L.; Yin, H.; Li, T.; and Wang, S. 2020. An improved DCT-based JND estimation model considering multiple masking effects. *Journal of Visual Communication and Image Representation*, 71: 102850.

Wang, M.; Liu, X.; Xie, W.; and Xu, L. 2021. Perceptual Redundancy Estimation of Screen Images via Multi-domain Sensitivities. *IEEE Signal Processing Letters*, 28: 1440–1444.

Wang, M.; Xu, Z.; Liu, X.; Xiong, J.; and Xie, W. 2022a. Perceptually Quasi-lossless Compression of Screen Content Data via Visibility Modeling and Deep Forecasting. *IEEE Transactions on Industrial Informatics*, 18(10): 6865–6875.

Wang, M.; Zhang, R.; Huang, L.; and Li, Y. 2024a. Visual Redundancy Removal for Composite Images: A Benchmark Dataset and a Multi-Visual-Effects Driven Incremental Method. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 10189–10197.

Wang, M.; Zhu, Y.; Zhang, R.; and Xie, W. 2024b. Meta-JND: A Meta-Learning Approach for Just Noticeable Difference Estimation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 3151–3159.

Wang, Y.; Inguva, S.; and Adsumilli, B. 2019. YouTube UGC dataset for video compression research. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 1–5.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2022b. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.; Wang, S.; Tian, S.; Huang, L.; Liu, Y.; and Wang, M. 2023a. Just Noticeable Visual Redundancy Forecasting: A Deep Multimodal-driven Approach. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 2965–2973.

Xie, W.; Wang, S.; Zhang, R.; and Wang, M. 2023b. Visual Redundancy Removal of Composite Images via Multimodal Learning. In *ACM International Conference on Multimedia (ACM MM)*, 6765–6773.

Yang, C.; An, Z.; Zhou, H.; Zhuang, F.; Xu, Y.; and Zhang, Q. 2023. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10212–10227.

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

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.