MetaJND: A Meta-Learning Approach for Just Noticeable Difference Estimation

Miaohui Wang[2] , **Yukuan Zhu**[2] , **Rong Zhang**[2] and **Wuyuan Xie**[1]*

[1]College of Computer Science and Software Engineering, Shenzhen University

[2]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University

{wang.miaohui, kra169169, zhangrong2208, wuyuan.xie}@gmail.com

Abstract

The modeling of just noticeable difference (JND) in supervised learning for visual signals has made significant progress. However, existing JND models often suffer from limited generalization due to the need for large-scale training data and their constraints to certain image types. Moreover, these models primarily focus on a single RGB modality, ignoring the potential complementary impacts of multiple modalities. To address these challenges, we propose a new meta-learning approach for the JND modeling, called MetaJND. We introduce two key visual-sensitive modalities like saliency and depth, and leverage a self-attention mechanism for effective interdependence of multi-modal features. Additionally, we incorporate meta-learning for the modality alignment, facilitating dynamic weight generation. Furthermore, we perform hierarchical fusion through multi-layer channel and spatial feature rectification. Experimental results on four benchmark datasets demonstrate the effectiveness of our MetaJND. Moreover, we have also evaluated its performance in compression and watermarking applications, observing higher bit-rate savings and better watermark hiding capabilities.

1 Introduction

In visual signal analysis, just noticeable difference (JND) measures the smallest change or difference that the human visual system (HVS) can perceive. Such measurements effectively quantify the visual redundancy existing in visual signals [Bai *et al.*, 2022], which can be used to achieve *perceptual coding* [Cheng *et al.*, 2021], thereby saving more bit-rate without damaging visual quality [Nami *et al.*, 2022; Wang *et al.*, 2021]. This is of great importance due to the increasing demand for high visual quality in digital visual services like *online video streaming, social media sharing, virtual reality, etc.* Furthermore, JND can also be widely used in various perceptual image and video processing tasks, such as *enhancement* [Nikzad *et al.*, 2021], *compression* [Qi *et al.*, 2023], *digital watermarks* [Qu *et al.*, 2023], and *quality assessment* [He *et al.*, 2022].

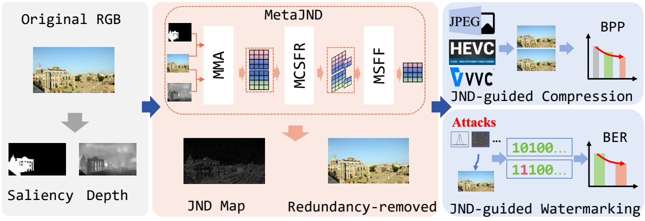


Figure 1: *Illustration of just noticeable difference (JND) and its applications.* The proposed meta-learning JND framework jointly explores the RGB, saliency, and depth modalities to predict the visual redundancy based on a multi-modal meta alignment (MMA) module, a multi-layer channel-space feature rectification (MCSFR) module, and a multi-scale feature fusion (MSFF) module.

Traditional JND models are typically driven by the HVS knowledge. Knowledge-driven approaches primarily compute JND by combining various HVS visual effects, such as brightness adaptation [Yang *et al.*, 2005], edge and texture sensitivity [Liu *et al.*, 2010], free energy principle [Wu *et al.*, 2013], pattern complexity diversity [Wu *et al.*, 2017], central foveal effect [Chen and Wu, 2019], feedforward and feedback modulation effects [Yin *et al.*, 2023], and blur sensitivity effect [Wang *et al.*, 2022]. However, how to accurately simulate multiple visual effects and their interactions remain challenging for these methods due to the limited knowledge of the HVS.

Recently, data-driven JND models [Shen *et al.*, 2020] have emerged as a new trend due to their ability to automatically learn descriptive features from visual data. The performance of these approaches heavily relies on the amount and distribution of datasets. However, there is a limitation in popular JND datasets such as PWJND [Shen *et al.*, 2020] and KonJND1k [Lin *et al.*, 2022], as they have very limited data available due to the high cost of subjective JND labeling. This limitation can greatly impact the accuracy and generalization of the training models. To address this issue, [Xie *et al.*, 2023] proposed a multi-modal network called hmJND-Net for the JND estimation, which utilized RGB, depth, saliency, and segmentation modalities to solve the problem of data scarcity.

Motivated by the discussions above, this paper presents a

Corresponding author: Wuyuan Xie

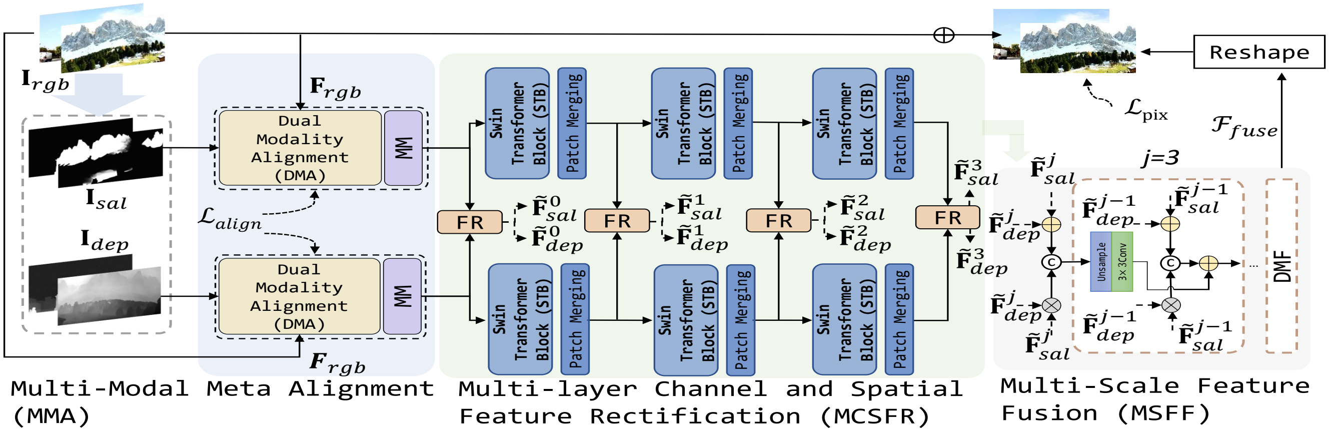

$\mathbf{I}_{rgb}$
$\mathbf{I}_{sal}$
$\mathbf{I}_{dep}$
$\mathbf{F}_{rgb}$
Dual Modality Alignment (DMA)
MM
$\mathcal{L}_{align}$
Swin Transformer Block (STB)
Patch Merging
FR
$\tilde{\mathbf{F}}^0_{sal}$
$\tilde{\mathbf{F}}^0_{dep}$
$\tilde{\mathbf{F}}^1_{sal}$
$\tilde{\mathbf{F}}^1_{dep}$
$\tilde{\mathbf{F}}^2_{sal}$
$\tilde{\mathbf{F}}^2_{dep}$
$\tilde{\mathbf{F}}^3_{sal}$
$\tilde{\mathbf{F}}^3_{dep}$
$\mathcal{L}_{\text{pix}}$
Reshape
$\mathcal{F}_{fuse}$
j=3
$\tilde{\mathbf{F}}^j_{sal}$
$\tilde{\mathbf{F}}^j_{dep}$
$\tilde{\mathbf{F}}^{j-1}_{dep}$
$\tilde{\mathbf{F}}^{j-1}_{sal}$
Unsample
3×3Conv
C
DMF
Multi-Modal Meta Alignment (MMA)
Multi-layer Channel and Spatial Feature Rectification (MCSFR)
Multi-Scale Feature Fusion (MSFF)

Figure 2: *Overall framework of our MetaJND.* RGB, saliency, and depth modalities are three inputs. The multi-modal meta alignment (MMA) module is then used to enhance the multi-modal alignment. Multi-layer channel and spatial feature rectification (MCSFR) and multi-Scale feature fusion (MSFF) modules are used to integrate the latent JND features. *Please zoom-in the electronic version for better details.*

multi-modal-driven meta-learning JND framework, namely MetaJND, to improve the accuracy and generalization. There are two key differences between MetaJND and [Xie *et al.*, 2023]. First, [Xie *et al.*, 2023] includes the segmentation modality to provide object category information. However, this can lead to issues when attempting to match category information with the JND threshold, potentially misleading model training. For example, as depicted in Figure 1, different buildings are with the same object category, but they have different JND thresholds. In contrast, MetaJND avoids this problem by excluding the segmentation modality. Second, [Xie *et al.*, 2023] combines auxiliary modalities and aligns them with the RGB modality. This approach faces two potential issues: 1) Only a few top-layer features are fused and aligned, while a significant number of features from different layers are ignored. 2) It is difficult to guarantee all auxiliary modalities are efficiently aligned and utilized. In contrast, MetaJND first employs the attention mechanism and meta-learning to separately align the features of the RGB and each auxiliary modality. Then, we introduce a multi-layer channel and spatial feature rectification to fuse the aligned features, which can comprehensively utilize the information of hierarchical features.

The main contributions of this paper are threefold: 1) To improve the limited generalization of deep JND models, we have proposed a meta-learning JND model at the pixel-level, namely MetaJND. By using meta-learning, we can make online adjustments to the RGB, saliency, and depth modalities, ensuring a higher generalization. 2) To effectively align features from different modalities, we have devised a novel multi-modal meta alignment (MMA) module that combines attention mechanism and meta-learning. This module combines attention mechanisms and meta-learning techniques to efficiently align the RGB features with each auxiliary modality. 3) To fully exploit complementary information from hierarchical features, we have designed a multi-layer channel and spatial feature rectification (MCSFR) module and a multi-scale feature fusion (MSFF) module to merge the aligned multi-modal features, predicting more accurate JND.

2 Related Work

In this section, we first briefly review some representative data-driven JND models, including single-image and multi-image modality JND methods. We then provide a detailed explanation of the motivation behind our MetaJND.

2.1 Single-Image Modality Methods

The single-image modality JND models are typically trained on the original RGB data and their distorted versions. This type of method predicts JND in either picture or block level with deep networks, like convolutional neural networks (CNN). Picture-level JND (PJND) prediction distinguishes the critical lossless image from several distorted images with different levels of noises. For instance, perceptually lossless/lossy binary classifier [Liu *et al.*, 2019], satisfied user ratio (SUR) prediction [Lin and Ghinea, 2022], and quality factors prediction [Tian *et al.*, 2020] have been developed for the PJND estimation. In contrast, block-level JND (BJND) prediction is conducted on the image blocks instead of the whole picture. For instance, block-level structural degradation [Shen *et al.*, 2020] and four-stage block-level framework [Nami *et al.*, 2022] have been introduced to predict BJND.

2.2 Multiple-Image Modality Methods

The multiple-image modality JND models aim at utilizing not only the RGB image information but also auxiliary modalities to predict the JND. For example, [Wu *et al.*, 2020] explored the JND modeling by introducing visual attention and pattern complexity. Based on the gradient modality and the class activation mapping (CAM) modality, [Jin *et al.*, 2021] designed signal degradation networks to simulate human perception. Additionally, [Xie *et al.*, 2023] fused aligned depth, saliency, and segmentation modalities to estimate the JND threshold.

2.3 Motivation

Due to the lack of sufficient data in existing JND datasets, both single-image modality and multiple-image modality JND models have been developed to improve the prediction performance. The former simplifies the JND estimation by predicting at the picture or block level, while the latter increases the amount of data by incorporating auxiliary modalities. Among them, multiple-image modality methods show more promise in accurately estimating pixel-level JND.

However, the modeling of the JND framework using multiple-image modality is still in its infancy. The main challenges lie in the alignment and fusion of multiple modalities that may differ significantly from each other. Traditional learning strategies struggle to simultaneously align different pairs of modalities, such as RGB-depth and RGB-saliency, due to their distinct characteristics. To address these challenges, we employ meta-learning [Hu *et al.*, 2019; Zhang *et al.*, 2021; Ma *et al.*, 2022; Ye *et al.*, 2021] to model the JND framework, which enables adaptability to the features of various modalities. We believe that employing meta-learning cannot only improve the alignment of different modalities, but also enhance generalization capabilities.

3 Proposed Method

In this section, we present the details of the proposed MetaJND. As shown in Figure 2, our MetaJND mainly consists of a multi-modal alignment (MMA) module, a multi-layer channel and spatial feature rectification (MCSFR) module, and a multi-scale feature fusion (MSFF) module.

3.1 Problem Formulation

The multi-modal data typically originate from different sensors or data sources, providing distinct visual information and features. However, these modalities also have distinct representations, distributions, and feature spaces. For instance, the RGB modality covers complete but low-level information, while the depth and saliency modalities contain incomplete but high-level semantic information. Therefore, it is necessary to align the extracted shallow features from different modalities using an alignment network $\mathcal{F}_{\text{align}}$. This alignment is achieved by applying the same loss function, $\mathcal{L}_{\text{align}}$, to both the RGB modality and each auxiliary modality. The aligned features are then fed into a fusion network, $\mathcal{F}_{\text{fuse}}$. The JND threshold is predicted by upsampling the fused features. The optimization problem is to minimize the mixed loss of $\mathcal{L}_{\text{align}}$ and the final pixel distance $\mathcal{L}_{\text{pix}}$ between the prediction and ground truth:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{align}} + \alpha \times \mathcal{L}_{\text{pix}}. \tag{1}$$

3.2 Multi-modal Meta Alignment

To align multi-modal features for effective information fusion, we propose a multi-modal meta alignment (MMA) module, which consists of a dual-modality alignment (DMA) module and a meta-learning module (MM), as illustrated in Figure 3. DMA aligns the shallow feature of the RGB and the auxiliary saliency or depth modality, while MM further makes adaptive adjustments to the aligned features, which improve the generalization of MetaJND.

Dual-Modality Alignment

In the dual-modality alignment (DMA) module, we employ the Swin Transformer (ST) to align each pair of modalities. Considering both its complexity and efficiency, we have adopted the Swin-B version [Liu *et al.*, 2021]. Before the DMA, we apply a 3×3 convolution network to individually extract shallow features (*i.e.*, $\mathbf{F}_{rgb}$, $\mathbf{F}_{dep}$, and $\mathbf{F}_{sal}$) from the RGB, depth, and saliency modalities (*i.e.*, $\mathbf{I}_{rgb}$, $\mathbf{I}_{dep}$, and $\mathbf{I}_{sal}$). The shallow features are then fed into the DMA module, which adopts a dual-stream structure to align the RGB and auxiliary modality features. Additionally, the multi-head cross-attention (MCA) is incorporated to guide the network to consider the correlation of different modalities and learn their interrelations. In practice, the RGB and auxiliary modality features are separately passed through two consecutive ST blocks with a windowed multi-head attention (W-MCA) and a sliding window multi-head attention (SW-MCA). Q, K, and V are computed by mapping through the weight matrix W, and the calculation process is as follows:

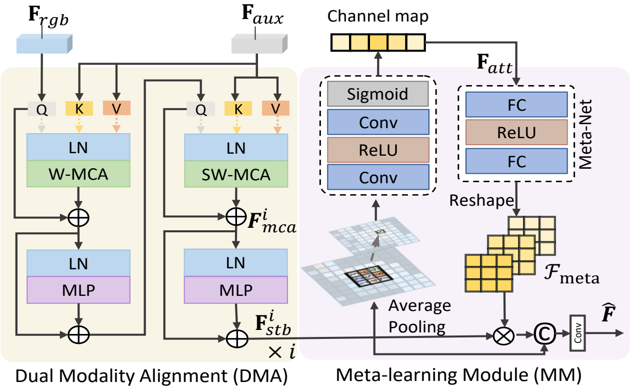


Figure 3: *Illustration of the proposed multi-modal meta alignment (MMA).* The inputs are the RGB modality feature $\mathbf{F}_{rgb}$ and the auxiliary features $\mathbf{F}_{aux}$ that consists of $\mathbf{F}_{dep}$ and $\mathbf{F}_{sal}$, while the outputs are the meta-aligned features $\hat{\mathbf{F}}^i_{dep}$ and $\hat{\mathbf{F}}^i_{sal}$, respectively.

$$\{Q, K, V\} = \left\{\mathbf{F}_{rgb}W^Q, \text{Conv}(\mathbf{X})W^K, \text{Conv}(\mathbf{X})W^V\right\}, \tag{2}$$

where $\mathbf{X}$ represents $\mathbf{F}_{dep}$ or $\mathbf{F}_{sal}$, and Conv$(\cdot)$ denotes the convolution operation.

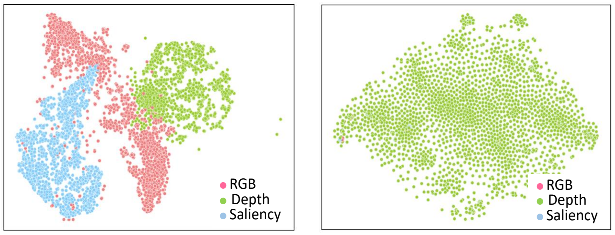
Subsequently, we add a multi-layer perceptron (MLP) and a GELU layer to enhance the feature tokens generated by the MCA. A layer normalization (LN) with the residual connections is added before the MCA and MLP blocks. Consequently, the entire process can be expressed by

$$\begin{cases} \mathbf{F}^i_{mca} = \text{MCA}(LN(K,Q,V)) + Q, \\ \mathbf{F}^i_{stb} = \text{MLP}\left(LN\left(\mathbf{F}^i_{mca}\right)\right) + \mathbf{F}^i_{mca}. \end{cases}, \tag{3}$$

where $\mathbf{F}^i_{mca}$ and $\mathbf{F}^i_{stb}$ represent the output features of the MCA blocks and the i-th ST block, respectively.

Meta-learning Module

When performing the DMA, shallow features from three modalities are mapped to a shared feature space. However, the limited JND-specific dataset hinders training a network to create a complete feature space that is suitable for unseen inputs, resulting in low generalization ability. To help the network learn a more general representation, we introduce a meta-learning module, which can generate adjusters for different inputs.



(a) Unaligned features

(b) Aligned features

Figure 4: *Visualization of unaligned and aligned multi-modal features*: (a) shows the distribution of unaligned features of three modalities, and (b) shows the distribution of features aligned by the MMA module.

MM is mainly comprised of a channel attention block and a meta-net. The channel attention block explores dependencies between the two channels to generate attention features, denoted as $\mathbf{F}_{att} = \mathcal{F}_{att}(\mathbf{F}^i_{stb})$. For convenience, we take SEnet [Hu *et al.*, 2018] as the backbone of $\mathcal{F}_{att}$. The meta-net is constructed with two fully-connected (FC) layers, with a ReLU layer in between. The adjusters generated by the meta-net are based on $\mathbf{F}_{att}$. During the feed forward process, the first FC layer extracts features from the inputs, while the second FC layer reshapes these extracted features into an adjuster, $\mathcal{F}_{\text{meta}}$. The process of generating adjusters can be defined by

$$\mathcal{F}_{\text{meta}} = \Phi(\text{FC}(\text{ReLU}(\text{FC}(\mathbf{F}_{att})))), \tag{4}$$

where FC$(\cdot)$ and ReLU$(\cdot)$ represent the FC and ReLU layer, respectively. $\Phi(\cdot)$ represents the reshape operation.

$\mathcal{F}_{\text{meta}}$ enables the network to adjust to different attention features dynamically. Specifically, it uses the adjuster to assign weights to the aligned features in the DMA, followed by a convolution operation. After the MM module, the output features of the MMA module are represented by

$$\hat{\mathbf{F}}_{dep/sal} = \text{Conv}_{1\times1}\left(\text{Concat}\left(\mathbf{F}^i_{stb} \times \mathcal{F}_{\text{meta}}, \mathbf{F}^i_{stb}\right)\right), \tag{5}$$

where $\hat{\mathbf{F}}_{dep/sal}$ represents the final aligned RGB-depth or RGB-saliency feature, and Concat$(\cdot)$ represents a concatenation operation.

To verify the effectiveness of the proposed MMA module, we have conducted a t-SNE visualization of aligned and unaligned features as depicted in Figure 4. In the scatter plot before alignment, the features of the three modalities form separate clusters, with each cluster representing a different modality. This indicates that the distributions of the different modalities are relatively independent when not aligned. However, after performing the MMA, the features of the three modalities become clustered together and overlap. This finding suggests that the proposed MMA has the ability to map the initially dispersed modality features into a shared feature space, resulting in a more concentrated feature distribution.

3.3 Multi-layer Channel and Spatial Feature Rectification

The fusion process, $\mathcal{F}_{\text{fuse}}$, plays a crucial role in fully utilizing the multi-modal information to improve the estimation performance. However, existing methods tend to only fuse features from the top layer but discard hierarchical features from the intermediate layers. As the high-layer has larger receptive fields, it generates coarse-grained features that capture the overall shape and structure. In contrast, the low-layer benefits from smaller receptive fields and can extract fine-grained features with more details. Additionally, the aligned features $\hat{\mathbf{F}}_{dep}$ and $\hat{\mathbf{F}}_{sal}$ may still contain noises due to the uncorrelated features extracted from different modalities.

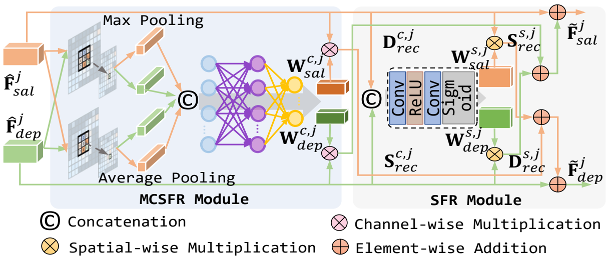


Figure 5: *Illustration of the proposed multi-layer channel and spatial feature rectification (MCSFR)*. The MCSFR module consists of a multi-layer channel rectification and a spatial feature rectification.

To address these issues, we propose a multi-layer channel and spatial feature rectification (MCSFR) module. This module extracts multi-scale features and rectifies them before performing $\mathcal{F}_{\text{fuse}}$. It contains a feature transformation (FT) module and a feature rectification (FR) module. The FT module utilizes ST blocks and patch merging layers to transform features into different scales. The FR module corrects latent features with a multi-layer channel rectification module and a spatial feature rectification module. Our MCSFR is shown in Figure 5.

Multi-layer Channel Rectification

Multi-layer channel rectification (MCR) aims to filter the noise in different channels. Specifically, we both apply the global max-pooling and the average-pooling to both the aligned saliency feature $\hat{\mathbf{F}}^j_{sal}$ and depth feature $\hat{\mathbf{F}}^j_{dep}$, where j means the j-th transformation. This module retains more information and better represents different aspects of the input modalities. Subsequently, the latent features are concatenated into a single tensor $\hat{\mathbf{F}}^j_{mcr}$. Then, $\hat{\mathbf{F}}^j_{mcr}$ is fed into a multi-layer perceptron (MLP) and a Sigmoid layer, and two sets of trainable weights (*i.e.*, $\mathbf{W}^{c,j}_{dep}$ and $\mathbf{W}^{c,j}_{sal}$) are obtained for rectifying the original features. This process can be expressed as follows:

$$\mathbf{W}^{c,j}_{dep}, \mathbf{W}^{c,j}_{sal} = \mathcal{F}_{\text{split}}(\text{Sigmoid}(\text{MLP}(\hat{\mathbf{F}}^j_{mcr}))), \tag{6}$$

where $\mathcal{F}_{\text{split}}(\cdot)$ refers to the process of splitting the output of MLP into two parts. Then, the multi-layer channel rectification process is formulated as:

$$\mathbf{D}^{c,j}_{rec} = \mathbf{W}^{c,j}_{dep} \otimes \hat{\mathbf{F}}^j_{dep}, \quad \mathbf{S}^{c,j}_{rec} = \mathbf{W}^{c,j}_{sal} \otimes \hat{\mathbf{F}}^j_{sal}, \tag{7}$$

where $\otimes$ denotes an element-wise multiplication, $\mathbf{D}^{c,j}_{rec}$ represents the rectified RGB-depth features, and $\mathbf{S}^{c,j}_{rec}$ represents the rectified RGB-saliency features after the multi-layer channel rectification.

Spatial Feature Rectification

In the process of the spatial feature rectification, the input features are first concatenated together. Then, the concatenated

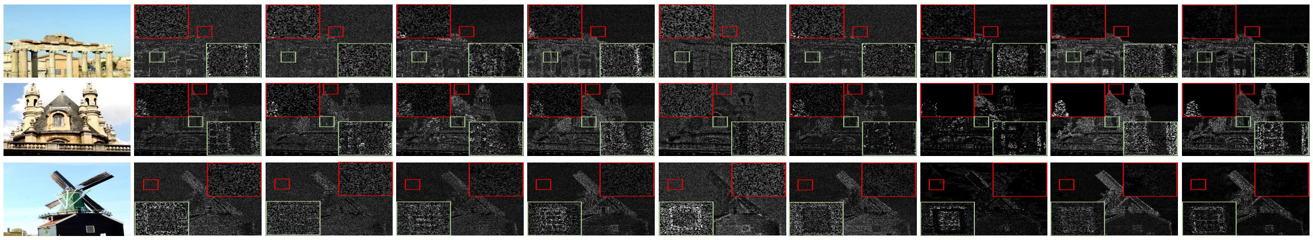


Figure 6: *Visual comparison of JND maps obtained by nine different methods*. Brighter magnitude indicates higher predicted visual redundancy, where smooth image regions usually contain less visual redundancy. From left to right: *Original*, *Liu2010TCSVT*, *Wu2013TIP*, *Wu2017TIP*, *Chen2019TCSVT*, *Shen2020TIP*, *Wang2022TII*, *Jiang2022TIP*, *Xie2023AAAI*, and *Proposed*

spatial feature, $\mathbf{F}^j_{sfr}$, is obtained along the channel dimension through an MLP, with the calculation formula as follows:

$$\mathbf{F}^j_{sfr} = \text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{Conv}_{1\times1}\left(\text{Concat}\left(\hat{\mathbf{F}}^j_{dep}, \hat{\mathbf{F}}^j_{sal}\right)\right)\right)\right). \tag{8}$$

Similarly, $\mathbf{F}^j_{sfr}$ is fed into a Sigmoid layer and divided into two parts, $\mathbf{W}^{s,j}_{dep}$ and $\mathbf{W}^{s,j}_{sal}$. The detailed process is expressed as:

$$\mathbf{W}^{s,j}_{dep}, \mathbf{W}^{s,j}_{sal} = \mathcal{F}_{split}(\text{Sigmoid}(\mathbf{F}^j_{sfr})). \tag{9}$$

The spatial rectification is then expressed by

$$\mathbf{D}^{s,j}_{rec} = \mathbf{W}^{s,j}_{dep} \otimes \hat{\mathbf{F}}^j_{dep}, \mathbf{S}^{s,j}_{rec} = \mathbf{W}^{s,j}_{sal} \otimes \hat{\mathbf{F}}^j_{sal}, \tag{10}$$

where $\mathbf{D}^{s,j}_{rec}$ and $\mathbf{S}^{s,j}_{rec}$ represent the rectified RGB-depth and RGB-saliency features after the spatial rectification, respectively.

Finally, the overall rectified features $\tilde{\mathbf{F}}^j_{sal}$ and $\tilde{\mathbf{F}}^j_{dep}$ after the *j*-th transformation are obtained by

$$\begin{cases} \tilde{\mathbf{F}}^j_{dep} = \hat{\mathbf{F}}^j_{dep} + \lambda_c \times \mathbf{D}^{c,j}_{rec} + \lambda_s \times \mathbf{D}^{s,j}_{rec}, \\ \tilde{\mathbf{F}}^j_{sal} = \hat{\mathbf{F}}^j_{sal} + \lambda_c \times \mathbf{S}^{c,j}_{rec} + \lambda_s \times \mathbf{S}^{s,j}_{rec}, \end{cases}, \tag{11}$$

where λ_c and λ_s are two hyper-parameters, and they are set to 0.5. j belongs to $\{0, 1, 2, 3\}$.

3.4 Multi-Scale Feature Fusion

After the MCSFR module, the rectified features are further fused for the final JND estimation. To fully take advantage of the multi-scale information provided by different layers, we have designed a multi-scale feature fusion (MSFF) module, which consists of four dual modality fusion (DMF) modules. DMF first fuses the multi-scale rectified features $\tilde{\mathbf{F}}^j_{dep}$ and $\tilde{\mathbf{F}}^j_{sal}$ using addition, multiplication, and concatenation operations:

$$\mathbf{F}^j_{cat} = \text{Concat}\left(\left(\tilde{\mathbf{F}}^j_{dep} + \tilde{\mathbf{F}}^j_{sal}\right), \left(\tilde{\mathbf{F}}^j_{dep} \times \tilde{\mathbf{F}}^j_{sal}\right)\right), \tag{12}$$

where $\mathbf{F}^j_{cat}$ represents the early fused features after the concatenation.

Next, we progressively aggregate the early fused features and upsample them into shallow features for the final JND estimation, which can be expressed as:

$$\mathcal{F}^j_{\text{fuse}} = \begin{cases} \text{Conv}_{3\times3}\left(\text{Upsampling}\left(\mathbf{F}^j_{cat}\right)\right), & j = 0, \\ \text{Conv}_{3\times3}\left(\text{Upsampling}\left(\mathbf{F}^j_{cat} + \mathcal{F}^{j-1}_{\text{fuse}}\right)\right), & j = 1, 2, 3. \end{cases} \tag{13}$$

where Upsampling$(\cdot)$ denotes a 2× upsampling, which employs a bilinear interpolation. The predicted JND map is then obtained by feeding the fused feature $\mathbf{F}_{fuse}$ into an unsampling network.

4 Experimental Validations

In this section, we present the implementation details of our MetaJND and report the experimental comparison with eight representative methods. Moreover, we conduct an ablation study on the impact of using different modalities and modules.

4.1 Experimental Protocols

Dataset Description

In the experiments, we have trained the MetaJND using the benchmark dataset PWJND [Shen *et al.*, 2020]. This dataset contains 202 source images with a resolution of 1920×1080, where each original image is labeled with a compressed version with Versatile Video Coding (VVC). We randomly split the whole dataset into training, validation, and test sets with a ratio of 8:1:1.

To verify the generalization ability of different models, we further select three additional JND benchmark datasets, including MCL-JCI [Wang *et al.*, 2016], KonJND-1k [Lin *et al.*, 2022], and MDTJND [Liu *et al.*, 2023] for testing. In addition, we employ BASNet [Qin *et al.*, 2019] and DPT-Hybrid [Ranftl *et al.*, 2021] to generate the saliency and the depth modalities, respectively.

Training Details

Our MetaJND-Net is implemented on the *PyTorch* with all weights initialized using a truncated normal initializer. We use the default parameters of the *Adam* optimizer, such as β_1 = 0.9 and β_2 = 0.999. Before the training, we crop the input images into 224×224 with a random cropping and a random rotation. We set the batch size to 8 and the initial learning rate to 1e-5. The model training is then conducted on *NVIDIA GeForce RTX 3090* GPU, which takes approximately 16 hours for 200 epochs.

To guarantee the alignment efficiency, we have developed a meta-alignment loss $\mathcal{L}_{\text{align}}$, which is calculated by a pixel-wise L1 distance between the input RGB feature $\mathbf{F}_{rgb}$ and the aligned feature $\hat{\mathbf{F}}_y$. With α setting to 1, the total loss function

Dataset Method	MCL_JCI			KonJND_1k			MDTJND			PWJND		
	SSIM ↑	LPIPS ↓	MS-SSIM ↑	SSIM ↑	LPIPS ↓	MS-SSIM ↑	SSIM ↑	LPIPS ↓	MS-SSIM ↑	SSIM ↑	LPIPS ↓	MS-SSIM ↑
Liu2010TCSVT [Liu *et al.*, 2010]	0.8958	0.0927	0.9848	0.8833	0.1099	0.9841	0.9205	0.0637	0.9893	0.9143	0.0575	0.9885
Wu2013TIP [Wu *et al.*, 2013]	0.8988	0.0887	0.9853	0.8839	0.1082	0.9840	0.9221	0.0591	0.9893	0.9197	0.0543	0.9888
Wu2017TIP [Wu *et al.*, 2017]	0.9252	0.0666	0.9890	0.9128	0.0821	0.9880	0.940	0.0452	0.9916	0.9406	0.0331	0.9915
Chen2019TCSVT [Chen and Wu, 2019]	0.9370	0.0620	0.9911	0.9232	0.0834	0.9898	0.9447	0.0485	0.9926	0.9448	0.0349	0.9922
Shen2020TIP [Shen *et al.*, 2020]	0.8878	0.0961	0.9833	0.8786	0.1155	0.9826	0.9072	0.0734	0.9872	0.9085	0.0709	0.9864
Jiang2022TIP [Jiang *et al.*, 2022]	0.9376	0.0466	0.9903	0.9341	0.0527	0.9903	0.9466	0.0295	0.9922	0.9416	0.0348	0.9910
Wang2022TII [Wang *et al.*, 2022]	0.9039	0.0842	0.9859	0.8981	0.0963	0.9861	0.9317	0.0556	0.9907	0.9311	0.0420	0.9908
Xie2023AAAI [Xie *et al.*, 2023]	0.9402	0.0535	0.9915	0.9348	0.0691	0.9911	0.9523	0.0366	0.9934	0.9579	0.0276	0.9938
MetaJND (ours)	**0.9526**	**0.0435**	**0.9929**	**0.9526**	**0.0490**	**0.9932**	**0.9642**	**0.0256**	**0.9946**	**0.9632**	**0.0271**	**0.9944**

Table 1: Performance comparisons of our MetaJND with eight representative methods on four benchmark datasets in terms of SSIM, LPIPS, and MS-SSIM. "↑" means the higher the better, while "↓" means the lower the better. The best results in each column are highlighted in **bold**.

in (1) can be expressed as:

$$\mathcal{L}_{\text{overall}} = \sum_{y \in dep,sal} \parallel \mathbf{F}_{rgb} - \hat{\mathbf{F}}_y \parallel_1 + \alpha \times \parallel \mathbf{I}_{gt} - \mathbf{I}_{jnd} \parallel_2^2, \quad (14)$$

where $\mathbf{I}_{gt}$ refers to the ground-truth, while $\mathbf{I}_{jnd}$ denotes the predicted JND map.

Evaluation Settings

We have conducted the commonly-used noise injection experiments [Xie *et al.*, 2023; Jiang *et al.*, 2022] to evaluate the accuracy of the JND estimation. Specifically, we randomly injected noise into each pixel with the guidance of $\mathbf{I}_{jnd}$, which can be formulated as:

$$\mathbf{I}_{noised} = \mathbf{I}_{ori} + \mu \times \beta \times \mathbf{I}_{jnd}, \quad (15)$$

where $\mathbf{I}_{noised}$ represents the JND-guided noised image, β randomly takes ± 1, and μ is an adjuster scalar. A better JND model can tolerate more noise, thereby achieving better visual quality under the same noise level. In our experiments, we adjust μ to control the amplitude of the injected noise, aiming for a specific peak signal-to-noise ratio (PSNR). Specifically, the noise levels of the first JND level for four datasets are PSNR=35.05dB, 33.18dB, 38.35dB, and 35.12dB, respectively. To achieve consistency, we control the injected noise level to the mean value (*e.g.*, PSNR=35.42 dB) of these four datasets.

To facilitate quantitative comparisons, we have used three additional widely-used perceptual metrics to assess the image quality: the structural similarity index (SSIM), the multi-scale structural similarity (MS-SSIM), and the learned perceptual image patch similarity (LPIPS) [Zhang *et al.*, 2018].

4.2 Performance Comparisons

To demonstrate the overall performance, we provide the qualitative and quantitative results of our MetaJND and eight representative methods, including *Liu2010TCSVT* [Liu *et al.*, 2010], *Wu2013TIP* [Wu *et al.*, 2013], *Wu2017TIP* [Wu *et al.*, 2017], *Chen2019TCSVT* [Chen and Wu, 2019], *Shen2020TIP* [Shen *et al.*, 2020], *Jiang2022TIP* [Jiang *et al.*, 2022], *Wang2022TII* [Wang *et al.*, 2022], and Xie2023AAAI [Xie *et al.*, 2023].

Subjective Qualitative Results

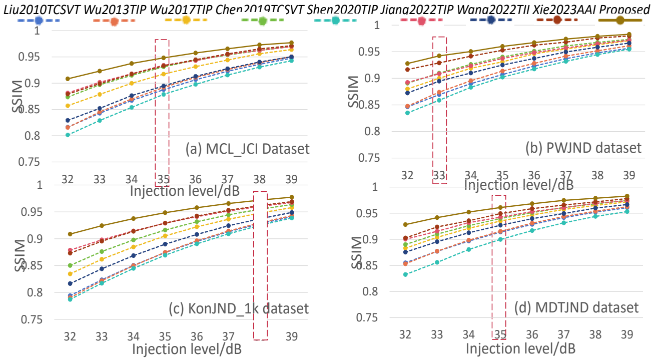


Figure 7: *Noise tolerance comparisons with various noise injection levels*. The red box represents the first JND point for each dataset.

Figures 6 (b)-(j) show the subjective visual comparison results for JND estimation of nine methods, where brighter pixels indicate higher tolerance redundancy. In (b)-(i), higher visual redundancy is observed in smooth background regions (highlighted by the red box) and intricate structural regions (highlighted by the green box). As human vision exhibits higher perceptual sensitivity to uniformly smooth and structural regions, lower visual redundancy should be estimated in these areas. Consequently, our MetaJND, benefiting from meta-learning and multi-layer fusion, achieves more precise JND estimation in these sensitive regions.

Objective Quantitative Results

Table 1 presents the SSIM, LPIPS, and MS-SSIM results for nine JND models. Our MetaJND demonstrates stable performance across all four datasets, demonstrating its superior capability in noise tolerance and generalization to different types of data.

In addition, to address concerns of fairness in the use of the noise injection level at the PSNR=35.42dB, we have expanded our experiments to include a wider range of noise levels for performance comparison. For instance, we have considered the noise injection level at various PSNRs, including {32, 33, 34, 35, 36, 37, 38, and 39}. In Figure 7, we have plotted the performance curve. As seen, our MetaJND consistently outperforms the other methods.

4.3 Ablation Study

In this section, we have conducted the ablation experiments to verify the effects of two additional modalities and our designed modules on the dataset PWJND [Shen *et al.*, 2020]. The selected (unselected) modality or module is denoted by "✓" ("×").

Modalities			Indicators		
RGB	Depth	Saliency	SSIM	LPIPS	MS-SSIM
✓	×	×	0.9472	0.0371	0.9921
✓	×	✓	0.9599	0.0281	0.9939
✓	✓	×	0.9592	0.0280	0.9938
✓	✓	✓	0.9632	0.0271	0.9944

Table 2: Ablation study of two additional modalities.

MetaJND Modules			Indicators		
MMA	MCSFR	MSFF	SSIM	LPIPS	MS-SSIM
×	×	×	0.9384	0.0497	0.9905
✓	×	×	0.9476	0.0375	0.9922
✓	✓	×	0.9597	0.0290	0.9938
✓	✓	✓	0.9632	0.0271	0.9944

Table 3: Ablation study of our MMA, MCSFR, and MSFF modules.

Effects of Auxiliary Modalities

We have conducted four experiments to validate the necessity of the depth and saliency modalities as shown in Table 2. For example, to evaluate the impact of the depth modality, we replace it with the RGB modality. As seen, selecting all three modalities simultaneously leads to the greatest performance improvement, highlighting the necessity of these two auxiliary modalities.

Effects of MMA, MCSFR, and MSFF

To demonstrate the importance of the proposed MMA, MCSFR, and MSFF modules, we have also conducted a series of ablation experiments. The results are provided in Table 3. In the experiments, we disable the unselected module and replace it with a simple network. For example, the MMA module is replaced by a network that incorporates a concatenation operation, three 3×3 convolution layers, and one 1×1 convolution layer. The MCSFR module is replaced with a three-layer ST block, while the MSFF module is simply replaced with concatenation. As seen, the best performance is achieved when all three modules are enabled simultaneously.

5 JND Applications

5.1 Compression Application

JND has been widely-applied in image coding to enhance the perceptual quality. In this section, we have integrated MetaJND into three encoding standards, including JPEG, HEVC, and VVC for compression applications. We have employed the same compression methods as described in [Xie *et al.*, 2023]. Figure 8 shows the visual comparisons of JND-guided compression. As seen, all JND-guided compression methods provide nearly identical visual quality. However, our MetaJND approach achieves the highest bit-per-pixel (BPP) savings, with an average of 24.11% coding gain on four datasets.

5.2 Watermarking Application

Image watermarking is commonly used to achieve copyright protection, content verification, and identity authentication.

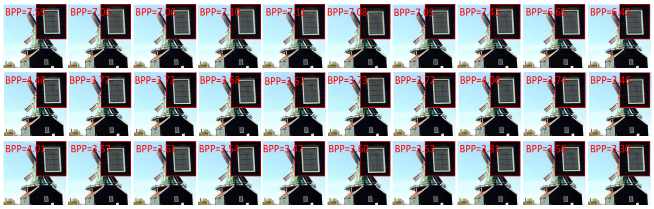


Figure 8: *JND-guided compression performance of nine methods.* From the second column to the ninth column: *Liu2010TCSVT*, *Wu2013TIP*, *Wu2017TIP*, *Chen2019TCSVT*, *Shen2020TIP*, *Wang2022TII*, *Jiang2022TIP*, *Xie2023AAAI*, and *Proposed*. From the first row to the third row: JPEG, HEVC, and VVC.

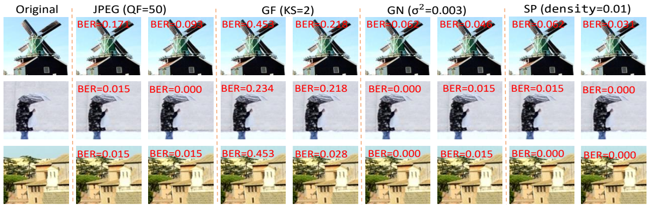


Figure 9: *JND-guided watermarking performance of four attacks.* In each attack type, the left is *Jia2021MM*, while the right is *Jia2021MM+MetaJND*.

As suggested in [Fang *et al.*, 2023], the JND map is introduced as a weight matrix in the loss function in the luminance channel to improve the deep watermarking model. The robustness of watermarking is measured by the bit error rate (BER) performance. We evaluate the watermarking performance under four types of attacks, including JPEG (quality factor (QF) =50), Gaussian filter (kernel size (KS)=2), Gaussian noise (σ^2=0.003), and Salt & Pepper (`density`=0.01). As seen in Figure 9, our MetaJND-guided method archives lower BER than [Fang *et al.*, 2023] under the same PSNR level as in Table 1. The average BER value of our MetaJND-guided method is 0.0913 on four datasets, while that of [Fang *et al.*, 2023] is 0.0951.

6 Conclusion

In this paper, we propose a new meta-learning framework, called MetaJND, for estimating just noticeable differences at the pixel level. MetaJND leverages multiple modalities, including RGB, saliency, and depth, to enhance both its accuracy and generalization. To achieve this, we have developed a multi-modal meta alignment module that combines attention mechanisms and meta-learning to align features from different modalities. Additionally, we have explored a multi-layer channel feature rectification module to correct intermediate features before conducting the multi-modal fusion. Furthermore, we have designed a multi-scale feature fusion module to merge the aligned features obtained from different layers. Experimental results on four benchmark datasets, along with comparisons to eight representative JND methods, demonstrate the superior performance and enhanced generalization capabilities of our MetaJND.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61701310 and 62372306, in part by the Natural Science Foundation of Shenzhen City under Grant 20220809160139001, and in part by the Natural Science Foundation of Guangdong Province under Grants 2024A1515011972, 2021A1515011877, and 2022A1515011245.

References

[Bai *et al.*, 2022] Yuanchao Bai, Xu Yang, Xianming Liu, Junjun Jiang, Yaowei Wang, Xiangyang Ji, and Wen Gao. Towards end-to-end image compression and analysis with transformers. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 104–112, 2022.

[Chen and Wu, 2019] Zhenzhong Chen and Wei Wu. 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, 2019.

[Cheng *et al.*, 2021] Zhengxue Cheng, Ting Fu, Jiapeng Hu, Li Guo, Shihao Wang, Xiongxin Zhao, Dajiang Zhou, and Yang Song. Perceptual Image Compression using Relativistic Average Least Squares GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1895–1900, 2021.

[Fang *et al.*, 2023] Han Fang, Zhaoyang Jia, Hang Zhou, Zehua Ma, and Weiming Zhang. Encoded feature enhancement in watermarking network for distortion in real scenes. *IEEE Transactions on Multimedia*, 25:2648–2660, 2023.

[He *et al.*, 2022] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 942–948, 2022.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[Hu *et al.*, 2019] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-SR: A magnification-arbitrary network for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1575–1584, 2019.

[Jiang *et al.*, 2022] Qiuping Jiang, Zhentao Liu, Shiqi Wang, Feng Shao, and Weisi Lin. Toward top-down just noticeable difference estimation of natural images. *IEEE Transactions on Image Processing*, 31:3697–3712, 2022.

[Jin *et al.*, 2021] Jian Jin, Xingxing Zhang, Xin Fu, Huan Zhang, Weisi Lin, Jian Lou, and Yao Zhao. Just noticeable difference for deep machine vision. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3452–3461, 2021.

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

[Lin *et al.*, 2022] Hanhe Lin, Guangan Chen, Mohsen Jenadeleh, Vlad Hosu, Ulf-Dietrich Reips, Raouf Hamzaoui, and Dietmar Saupe. Large-scale crowdsourced subjective assessment of picturewise just noticeable difference. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5859–5873, 2022.

[Liu *et al.*, 2010] Anmin Liu, Weisi Lin, Manoranjan Paul, Chenwei Deng, and Fan Zhang. 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, 2010.

[Liu *et al.*, 2019] Huanhua Liu, Yun Zhang, Huan Zhang, Chunling Fan, Sam Kwong, C-C Jay Kuo, and Xiaoping Fan. Deep learning-based picture-wise just noticeable distortion prediction model for image compression. *IEEE Transactions on Image Processing*, 29:641–656, 2019.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[Liu *et al.*, 2023] Yaxuan Liu, Jian Jin, Yuan Xue, and Weisi Lin. The First Comprehensive Dataset with Multiple Distortion Types for Visual Just-Noticeable Differences. *arXiv preprint arXiv:2303.02562*, pages 1–6, 2023.

[Ma *et al.*, 2022] Ruijun Ma, Shuyi Li, Bob Zhang, Leyuan Fang, and Zhengming Li. Flexible and generalized real photograph denoising exploiting dual meta attention. *IEEE Transactions on Cybernetics*, pages 1–13, 2022.

[Nami *et al.*, 2022] Sanaz Nami, Farhad Pakdaman, Mahmoud Reza Hashemi, and Shervin Shirmohammadi. BL-JUNIPER: A CNN-assisted framework for perceptual video coding leveraging block-level JND. *IEEE Transactions on Multimedia*, pages 1–16, 2022.

[Nikzad *et al.*, 2021] Mohammad Nikzad, Yongsheng Gao, and Jun Zhou. Attention-based Pyramid Dilated Lattice Network for Blind Image Denoising. In *International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 931–937, 2021.

[Qi *et al.*, 2023] Linfeng Qi, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Motion Information Propagation for Neural Video Compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6111–6120, 2023.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7479–7489, 2019.

[Qu *et al.*, 2023] Xinghua Qu, Xiang Yin, Pengfei Wei, Lu Lu, and Zejun Ma. AudioQR: deep neural audio watermarks for QR code. In *International Conference on*

International Joint Conferences on Artificial Intelligence (IJCAI), pages 6192–6200, 2023.

[Ranftl *et al.*, 2021] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021.

[Shen *et al.*, 2020] Xuelin Shen, Zhangkai Ni, Wenhan Yang, Xinfeng Zhang, Shiqi Wang, and Sam Kwong. Just noticeable distortion profile inference: A patch-level structural visibility learning approach. *IEEE Transactions on Image Processing*, 30:26–38, 2020.

[Tian *et al.*, 2020] Tao Tian, Hanli Wang, Lingxuan Zuo, C-C Jay Kuo, and Sam Kwong. Just noticeable difference level prediction for perceptual image compression. *IEEE Transactions on Broadcasting*, 66(3):690–700, 2020.

[Wang *et al.*, 2016] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset. In *IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513, 2016.

[Wang *et al.*, 2021] Menglu Wang, Xueyang Fu, Zepei Sun, and Zheng-Jun Zha. JPEG artifacts removal via compression quality ranker-guided networks. In *International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 566–572, 2021.

[Wang *et al.*, 2022] Miaohui Wang, Zhuowei Xu, Xueqin Liu, Jian Xiong, and Wuyuan Xie. Perceptually quasi-lossless compression of screen content data via visibility modeling and deep forecasting. *IEEE Transactions on Industrial Informatics*, 18(10):6865–6875, 2022.

[Wu *et al.*, 2013] Jinjian Wu, Guangming Shi, Weisi Lin, Anmin Liu, and Fei Qi. Just noticeable difference estimation for images with free-energy principle. *IEEE Transactions on Multimedia*, 15(7):1705–1710, 2013.

[Wu *et al.*, 2017] Jinjian Wu, Leida Li, Weisheng Dong, Guangming Shi, Weisi Lin, and C-C Jay Kuo. Enhanced just noticeable difference model for images with pattern complexity. *IEEE Transactions on Image Processing*, 26(6):2682–2693, 2017.

[Wu *et al.*, 2020] Yuhao Wu, Weiping Ji, and Jinjian Wu. Unsupervised deep learning for just noticeable difference estimation. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, 2020.

[Xie *et al.*, 2023] Wuyuan Xie, Shukang Wang, Sukun Tian, Lirong Huang, Ye Liu, and Miaohui Wang. Just Noticeable Visual Redundancy Forecasting: A Deep Multimodal-driven Approach. *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2965–2973, 2023.

[Yang *et al.*, 2005] Xiaokang Yang, Weisi Lin, Zhongkhang Lu, Ee Ping Ong, and Susu Yao. Just noticeable distortion model and its applications in video coding. *Elsevier Signal processing: Image Communication*, 20(7):662–680, 2005.

[Ye *et al.*, 2021] Shuquan Ye, Dongdong Chen, Songfang Han, Ziyu Wan, and Jing Liao. Meta-PU: An arbitrary-scale upsampling network for point cloud. *IEEE Transactions on Visualization and Computer Graphics*, 28(9):3206–3218, 2021.

[Yin *et al.*, 2023] Haibing Yin, Hongkui Wang, Li Yu, Junhui Liang, and Guangtao Zhai. Feedforward and Feedback Modulations Based Foveated JND Estimation for Images. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(5):1–23, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[Zhang *et al.*, 2021] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-DETR: Image-level few-shot object detection with inter-class correlation exploitation. *arXiv preprint arXiv:2103.11731*, pages 1–14, 2021.