Visual Redundancy Removal for Composite Images: A Benchmark Dataset and a Multi-Visual-Effects Driven Incremental Method

Miaohui Wang*, Rong Zhang, Lirong Huang, Yanshan Li

Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060
wang.miaohui@gmail.com, zhangrong2208@gmail.com, hlr.rongger@gmail.com, lys@szu.edu.cn

Abstract

Composite images (CIs) typically combine various elements from different scenes, views, and styles, which are a very important information carrier in the era of mixed media such as *virtual reality*, *mixed reality*, *metaverse*, *etc*. However, the complexity of CI content presents a significant challenge for subsequent visual perception modeling and compression. In addition, the lack of benchmark CI databases also hinders the use of recent advanced data-driven methods. To address these challenges, we first establish one of the earliest visual redundancy prediction (VRP) databases for CIs. Moreover, we propose a multi-visual effect (MVE)-driven incremental learning method that combines the strengths of hand-crafted and data-driven approaches to achieve more accurate VRP modeling. Specifically, we design special incremental rules to learn the visual knowledge flow of MVE. To effectively capture the associated features of MVE, we further develop a three-stage incremental learning approach for VRP based on an encoder-decoder network. Extensive experimental results validate the superiority of the proposed method in terms of subjective, objective, and compression experiments.

Introduction

Nowadays, multimedia applications such as *live streaming*, *video conferencing*, *virtual reality*, and *AI generation* (Kang et al. 2023) have greatly prompted an explosive growth in the creation of composite images (CIs) (Xie et al. 2023b). A typical CI is composed of various elements from different scenes, views, and styles, resulting in complex content structures. In addition, multimedia visual services are expected to provide high-quality visual experiences while minimizing data transmission delays (Huang et al. 2023), which presents a more challenging task for compressing CI data.

*This work was supported in part by the National Natural Science Foundation of China under Grants 61701310, 62372306 and 62076165, in part by the Natural Science Foundation of Guangdong Province under Grants 2022A1515011245, 2020KCXTD004 and 2022GXJK367, and in part by the Natural Science Foundation of Shenzhen City under Grant JCYJ20220809160139001. (*Corresponding author*: Miaohui Wang)

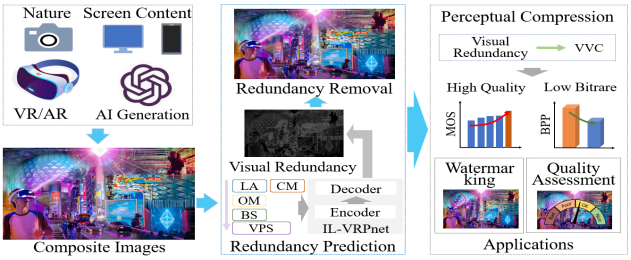


Figure 1: Applications of composite images (CIs) and illustration of the multi-visual effect (MVE)-driven visual redundancy prediction (*mveVRP*). A three-stage manner incremental learning approach is designed to model visual redundancy, which can be applied to *perceptual compression*, *watermarking enhancement*, and *quality assessment*.

Consequently, the transmission of large CIs has placed a heavy burden on communication and storage systems. Visual redundancy prediction (VRP) is an effective approach to improve the compression efficiency without compromising the quality of experience (QoE) (Qin et al. 2023; Yin et al. 2022). As depicted in Figure 1, by eliminating visual redundancy, compression efficiency can be greatly improved (Bai et al. 2022; Wang et al. 2022), thereby ensuring QoE and low latency for multimedia applications.

Existing VRP methods (Jiang et al. 2022; Wang et al. 2021) generally rely on the visual effects of the human visual system (HVS) to simulate the perception process. These methods are based on either hand-crafted or deep-learned paradigms, and they predict visual redundancy through linear and nonlinear superposition or deep feature fusion (Shen et al. 2021). However, there are several challenges in accurately modeling the complex HVS perception process in VRP. These challenges include representing various visual effects, understanding their correlation, and determining fusion mechanisms. Both knowledge-driven and data-driven methods have their limitations in VRP: 1) **Knowledge-driven**. This type of method is able to consider multiple visual effects related to perceptual redundancy simultaneously. However, the simple superposition and fusion of dif-

ferent visual effects can greatly jeopardize the accuracy of VRP. This issue becomes even more serious when dealing with complex and variable content, such as in the case of CI data. 2) **Data-driven**. This type of method directly learns visual redundancy, but it can suffer from label bias issues (Dong et al. 2021), especially when existing datasets contain very limited data. Additionally, the lack of human knowledge regarding visual effects can make it difficult to accurately simulate the perception of the HVS.

To address the aforementioned limitations, we begin by constructing a benchmark VRP database, specifically for CIs. Moreover, we establish a HVS-inspired fusion mechanism that incorporates multiple visual effects through incremental learning. This mechanism aims to imitate the continuous learning process of humans, allowing it to acquire various knowledge and maintain a stable system that combines both old and new knowledge (Qiang et al. 2023; Sun et al. 2020; Michieli and Zanuttigh 2021). The contributions of this work can be summarized as follows:

- To compensate for the lack of benchmark CI databases, we have established a visual redundancy database, namely composite image VRP database (*ciVRP-Set*). This database consists of 413 raw ultra-high-definition images, with a total of 10325 redundancy-removed samples. It is one of the earliest VRP databases for CIs.
- To effectively capture a strong feature association between different visual effects, we have designed a multi-visual effect (MVE)-based incremental VRP framework (*mveVRP*), where the feature association can be taken as a prompt, thereby filling the disadvantages in both handcrafted and data-driven schemes, resulting in more accurate performance of VRP for CIs.
- To represent the rules of MVE-based incremental learning, we have developed an HVS-driven knowledge flow and investigated a special regularization constraint. These components allow us to train our *mveVRP* in a three-stage manner. Further, we have constructed an incremental learning-based VRP network to predict the visual redundancy, namely *IL-VRPnet*, which includes a pioneer and a principal network. Experimental results validate its effectiveness on both traditional VRP dataset and our *ciVRP-Set*.

Related Work

To make this article more focused, we briefly review some milestone developments of VRP, including HVS-inspired methods and data-driven methods. Then, we provide the motivation of our *mveVRP*.

Visual Redundancy Prediction (VRP)

HVS-inspired VRPs are typically developed using physiological, psychological, and brain function studies that examine the visual characteristics of human eyes. These studies help determine the maximum visual redundancy threshold by mathematically combining various visual effects in either the pixel or transform domain. For example, luminance adaptation (LA) and contrast masking (CM) were two widely-used visual effects that have been deeply investigated since the pioneering work of Chou and Li (1995). Later, the edge masking effect was further considered by Yang et al. (2005) due to the higher sensitivity of HVS for distortions contained in edge areas. After that, various visual effects (Lin and Ghinea 2022) have been developed for the modeling of VRPs, such as disorderly masking (Wu et al. 2013), structural sensitivity (Wang et al. 2016), pattern masking (Wu et al. 2017), visual saliency (Hadizadeh, Rajati, and Bajić 2017), foveated masking (Chen and Wu 2019), oblique effect (Wang et al. 2022), and forward-backward modulation (Yin et al. 2023). On the other hand, these visual effects have also been studied in various transform domains, such as discrete cosine transform (Bae and Kim 2017), discrete wavelet transform (Wang et al. 2023), and karhunen-loeve transform (Jiang et al. 2022).

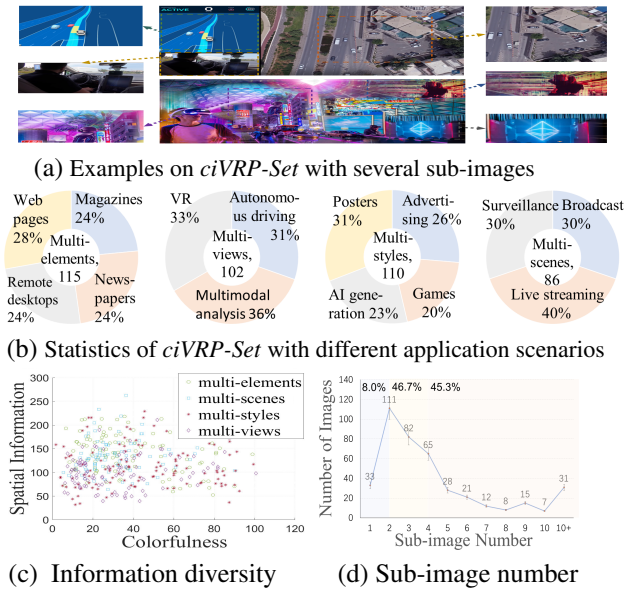
Data-driven VRPs usually utilize machine learning to automatically obtain feature representations of visual redundancy, thereby simulating various visual effects of HVS. For instance, perceptual redundancy was indirectly estimated by perceived visual quality level (Liu et al. 2019), perceptual quality prediction (Tian et al. 2021) based on the convolutional neural networks (CNN), and block level perceptual importance (Nami et al. 2023) through deep learning. Besides, recent approaches have also combined manual and deep features (Shen et al. 2021) or multimodal visual features (Xie et al. 2023a) to directly predict visual redundancy.

Motivation

Previous HVS-based VRPs usually combine multiple visual effects in either a linear or nonlinear manner. For example, NAMM (Yang et al. 2005) is frequently used to prevent overlapping effects of multiple visual characteristics in the pixel domain, while multiplication is commonly used in the transform domain. However, these simplistic combinations fail to consider the intricate relationship between multiple visual effects, making it challenging to accurately replicate the perceptual characteristics of the HVS. On the other hand, data-driven VRPs integrate visual representations at the feature level to estimate visual redundancy. Although deep models offer powerful nonlinear fitting capabilities, they lack direct access to the visual knowledge derived from physiological and psychological studies. Besides, the performance of these methods is limited by insufficient data.

Moreover, CIs are currently in full swing, and their complex contents pose a great challenge to existing HVS-based and data-driven VRPs. However, the lack of a visual redundancy dataset for CIs greatly hinders further research in this field. Therefore, our objective is to construct the first benchmark dataset *ciVRP-Set*.

Further, we propose to learn multiple visual effects knowledge through deep learning, and leverage its strong capabilities to deal with the potential relationships between various visual effects. However, traditional deep networks primarily focus on learning feature representations for specific tasks, which makes them less effective in capturing knowledge flow and more susceptible to the problem of "catastrophic forgetting". To address this issue, we will employ incremental learning (De Lange et al. 2022), which is designed to mitigate the effects of "catastrophic forgetting".



(a) Examples on *ciVRP-Set* with several sub-images

(b) Statistics of *ciVRP-Set* with different application scenarios

(c) Information diversity (d) Sub-image number

Figure 2: A statistical analysis of the proposed *ciVRP-Set*: (a) Examples of CI, (b) Image statistics in terms of *Multi-elements*, *Multi-views*, *Multi-styles*, and *Multi-scenes*, (c) Information diversity, and (d) sub-image number.

Specifically, with a series of the target database being arranged into a knowledge flow, incremental learning models will sequentially learn the knowledge flow, and the learning process of new tasks will be constrained using replay or regularization. We believe that this approach will enable us to effectively address the complex connections between multiple visual features in the context of *ciVRP-Set*. By doing so, we aim to overcome the limitations of both HVS-inspired and data-driven methods and develop a visual perception model that better fits the HVS modeling.

Proposed *ciVRP-Set*

In this section, we provide a detailed introduction to the proposed *ciVRP-Set*, including the collection and feature analysis of CIs and the labeling of visual redundancy.

Data Collection. We collect 413 representative CIs from various application scenarios, such as *video games, posters, website pages, live streaming, online conferences, surveillance, newspapers, advertising magazines, autonomous driving, virtual reality, etc*. The image resolution is 1920×1080. Figure 2 (a) exhibits two toy examples on *ciVRP-Set* with several sub-images in different positions, sizes, views, and scenes. We also provide the number of images in each application scenario as shown in Figure 2 (b). Besides, we provide the spatial information and colorfulness of the collected images to evaluate the information diversity of *ciVRP-Set* in Figure 2 (c). As seen, our *ciVRP-Set* covers a wide range of the image space. Moreover, we investigate the number of the sub-image, and the statistical results are shown in Figure 2 (d). As seen, more than 90% of CIs have more than 2 sub-images, which is quite different from natural images.

Feature Analysis. We organize and analyze the proposed *ciVRP-Set*, summarizing four distinct four distinct features:

- **Multi-scenes**. CIs in *video surveillance, live streaming, and online conferences* often contain sub-images from multiple different scenes.
- **Multi-views**. CIs in *medical images and autonomous driving* usually contain multiple modalities, angle, or dimensionality to provide more abundant information.
- **Multi-elements**. CIs in *newspapers, web pages, and remote desktops* commonly consist of various elements, such as text, images, and icons.
- **Multi-styles**. CIs in *advertising, magazines, games, and posters* often need to be in various styles to better attach the attention of users. The statistical percentage of each feature can also be found in Figure 2 (b).

Visual Redundancy Annotation. To annotate the maximum visual redundancy contained in CIs, we first encode the original images on the latest compression standard, Versatile Video Coding (VVC). The quantization parameter (QP) is ranging from 25 to 49 on the reference *VTM-6.0* software, generating 10325 redundancy-removed images. According to the standard ITU-R BT.500-13 (*e.g.*, subject, display, environment), we have conducted subjective experiments to determine the maximum redundancy-removed image.

Specifically, two tested images are randomly juxtaposed on a 65-inch monitor, where one of them is the original image and the other one is the redundancy-removed (*i.e.*, decoded) image. We invite 20 subjects to participate in the viewing test, including 12 males and 8 females, aged from 18 to 40. Each subject is asked to indicate whether they perceive a quality difference between the two juxtaposed images by selecting "Yes" or "No". Each redundancy-removed image has a total of 20 subjective scores, where the one (quantized by $\mathtt{qp}$) is regarded as the ground-truth $\mathbf{I}_{gt}$ (also known as *just noticeable difference*) should satisfy this conditions: its average score is closest to 0 and that of its neighboring image (quantized by $\mathtt{qp}\pm1$) is not equal to 0.

Proposed *mveVRP*

In this section, we present the proposed multi-visual effect (MVE)-driven VRP method. We first formulate the problem of incremental learning for *mveVRP*. Next, we generate five visual effect subsets based on their mathematical modeling to facilitate the incremental training of MVE. Then, we design incremental learning rules for MVE features extraction and fusion. Moreover, we construct an encoder-decoder network as the backbone of *IL-VRPnet*, to accomplish the incremental learning of VRP.

Problem Formulation

When dealing with the strong connection between MVE, *IL-VRPnet* aims to maintain the knowledge obtained from previous visual effects while learning new visual effect data. Specifically, when training on the current visual effect data $\mathcal{V}_t$, it is assumed that the previously learned visual effect

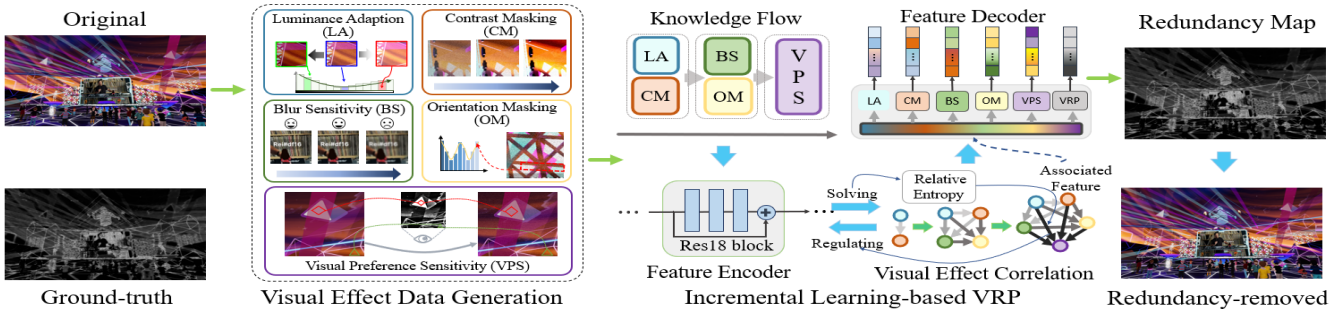


Figure 3: Overall framework of the proposed visual redundancy prediction (VRP) approach. The proposed *mveVRP* method mainly consists of a multiple visual effect data generation process and a incremental learning-based VRP module. Please zoom in for better details.

data $\{\mathcal{V}_k\}_{k=1}^{t-1}$ cannot be directly accessed. The learning task is to optimize the trained model $\mathcal{T}(\mathcal{V}; \mathbf{w})$ while maintaining the performance of all learned visual effects. In order words, it needs to minimize the distance between the predicted and the ground-truth maps, and optimizes the following function:

$$\sum_{k=1}^{t} \mathcal{T}(\mathcal{V}_k; \mathbf{w}) = \sum_{k=1}^{t} \left(\frac{1}{|\mathcal{V}_k|} \sum_{(x,q)\in\mathcal{V}_k} \mathcal{D}(g_{\mathbf{w}}(x), q) \right), \quad (1)$$

where x refers to a training sample and $|\mathcal{V}_k|$ denotes the total number of samples in the k-th visual effect. q denotes the corresponding label, $g_{\mathbf{w}}$ represents a deep network with the learned parameters $\mathbf{w}$, and $\mathcal{D}(\cdot)$ represents a distance function.

Finally, when predicting the visual redundancy, $\mathcal{T}(\cdot)$ is required to adopt the fused knowledge from the MVE flow and detect the visual redundancy more reasonably. The predicted redundancy map, $\mathbf{I}_{vrp}$ acquired from original images $\mathbf{I}_{ori}$, can be formulated as:

$$\mathbf{I}_{vrp} = \mathcal{T}(\mathbf{I}_{ori}; \mathbf{w}). \quad (2)$$

Visual Effect Data Generation

Based on the feature analysis of *ciVRP-Set* as discussed in Section , we consider five visual effects to represent visual redundancy in our incremental learning paradigm, including *luminance adaption* (LA), *contrast masking* (CM), *blur sensitivity* (BS), *orientation masking* (OM), and *visual preference sensitivity* (VPS). Among them, LA and CM have been widely proven to be efficient in detecting the overall visual redundancy. BS and OM (Wang et al. 2022) have shown a powerful capacity to reveal the visual redundancy in sharp edges, especially in text and thin lines. VPS is employed to represent the limited visual attention that occurs when observing complex content, especially in multiple sub-images.

Specifically, we generate five visual effect subsets based on *ciVRP-Set*, where the mathematical modeling and computation method are provided as follows: 1) **LA** is measured with respect to the average luminance and the luminance variation (Wang et al. 2016). 2) **CM** is determined with the edge height and the average background luminance (Yin et al. 2023). 3) **BS** is measured with the difference between the original and blurred images with the maximum standard deviation (Ferzli and Karam 2009). 4) **OM** is simulated as suggested in (Wang et al. 2022). 5) **VPS** is characterized with local binary patterns (LBP) (Chen et al. 2021).

MVE-based Incremental Learning Rules

To continuously manage various types of visual effects, we design incremental learning rules for the modeling of VRP. Specifically, we first organize the MVE knowledge flow based on the hierarchical cognitive process of HVS. Then, we determine the correlation between five visual features using the relative entropy. Based on the computed correlations, we incorporate regularization constraints when learning the knowledge of new visual effects, which allows us to obtain the corresponding MVE features in *mveVRP*.

Knowledge Flow. HVS follows a hierarchical cognitive procedure: Initially, we perceive some basic image features such as brightness and contrast; Then, we combine the underlying features including visual blur and direction information to form the secondary global perception; Finally, we generate an understanding of the image content and visual preference. Drawing inspiration from this process, we divide the visual effects into three groups and conduct the training in a sequential manner: LA and CM subsets KF_1, BS and OM subsets KF_2, and the VPS subset KF_3.

Visual Effect Correlation. During the training of the k-th visual effect, we need to calculate the correlation between the extracted features of the k-th and $k-1$ visual effects. Specifically, we first train a network to learn the visual knowledge of the k-th visual effect. Then, we obtain visual effect features $\mathbf{F}_{ve}^{(k)}$ and $\mathbf{F}_{ve}^{(k-1)}$ with knowledge of the $k-1$ visual effects, respectively. Subsequently, we compute the relative entropy $d_{sim}^{(k)}$ between these two kinds of features:

$$d_{sim}^{(k)} = \sum \mathbf{F}_{ve}^{(k)} log \left(\frac{\mathbf{F}_{ve}^{(k)}}{\mathbf{F}_{ve}^{(k-1)}} \right). \quad (3)$$

Finally, the correlation ratio $r^{(k)}$ of the k-th visual effect and the previous $k-1$ visual effects is determined by:

$$r^{(k)} = \frac{1}{1 + \alpha \cdot e^{-\beta \cdot d_{sim}^{(k)}}}, \quad (4)$$

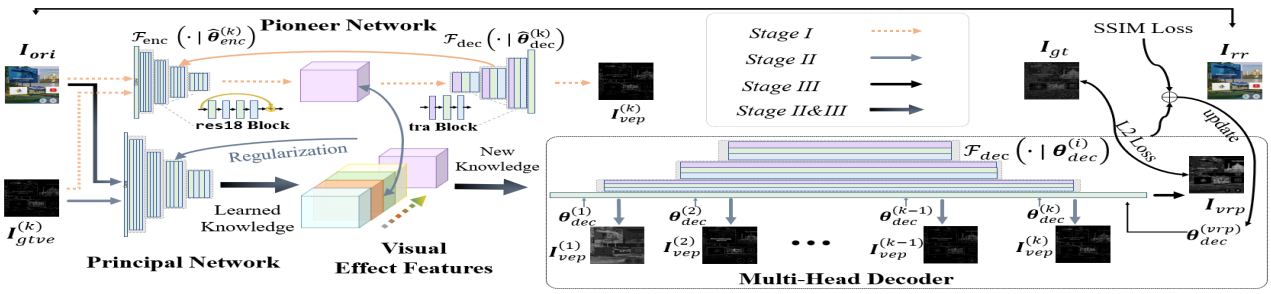


Figure 4: Incremental learning network *IL-VRPnet* for visual redundancy threshold inference. *IL-VRPnet* consists of a pioneer network and a principal network, and they share a similar structure for different learning tasks.

where α and β are two constant scalars. In the experiments, they are empirically set to 0.2 and 5, respectively.

Incremental Regularization Constraint. Based on the correlation ratio $r^{(k)}$, we further design a regularization constraint to reduce significant changes on important parameters of previously learned knowledge to protect them from being corrupted. Considering that the larger amplitude of the parameter reflects a stronger response, we use the absolute value to measure the importance. Specifically, given the network parameters $\boldsymbol{\theta}_{enc}^{(k-1)}$ after the training of the first $k-1$ visual effects, we perform the gradient flow based on the regularization constraint $\mathcal{F}_{grad}\left(\boldsymbol{\theta}_{enc}^{(k-1)}\right)$ that can be formulated as:

$$\mathcal{F}_{grad}\left(\boldsymbol{\theta}_{enc}^{(k-1)}\right)=\begin{cases} False, & \left|\boldsymbol{\theta}_{enc}^{(k-1)}\right| \geq r^{(k)}\cdot\boldsymbol{\theta}_{enc,max}^{(k-1)} \\ True, & \left|\boldsymbol{\theta}_{enc}^{(k-1)}\right| < r^{(k)}\cdot\boldsymbol{\theta}_{enc,max}^{(k-1)} \end{cases}, \quad (5)$$

where $|\cdot|$ returns the amplitude of parameters, and $\boldsymbol{\theta}_{enc,max}^{(k-1)}$ refers to the maximum amplitude of learned parameters. $False$ indicates that the backpropagation gradient is frozen, while $True$ means that the backpropagation gradient can update the relevant parameters.

Incremental Learning-based VRP Network

To implement the incremental learning rules, we construct an incremental learning-based VRP (*IL-VRPnet*) model. Specifically, we employ an encoder-decoder structure as the backbone of *IL-VRPnet*, and design a three-stage MVE knowledge-increment scheme. A detailed description is provided below.

Network Architecture. The proposed incremental network consists of a pioneer network and a principal network, both of which have a similar network structure as shown in Figure 4. The pioneer network is designed to pre-learn the knowledge of visual effects, then assists the incremental learning of the principal network. The principal network aims to continuously acquire the knowledge flow of MVE, and predict the visual redundancy based on the associated visual effect features. To achieve this, the encoder employs the `resnet-18` $\mathcal{F}_{res18}\left(\cdot|\boldsymbol{\theta}_{res18}\right)$ (He et al. 2016) to extract MVE features $\mathbf{F}_{ve}$ from the original image $\mathbf{I}_{ori}$. It can be formulated as

$$\mathbf{F}_{ve}=\mathcal{F}_{enc}\left(\mathbf{I}_{ori}\left|\boldsymbol{\theta}_{enc}\right.\right)=\mathcal{F}_{res18}\left(\mathbf{I}_{ori}\left|\boldsymbol{\theta}_{res18}\right.\right). \quad (6)$$

The decoder stacks multiple transposed convolutional blocks and builds a transposed network $\mathcal{F}_{tra}\left(\cdot|\boldsymbol{\theta}_{tra}\right)$ to decode the corresponding features. It can be expressed as

$$\mathcal{F}_{dec}\left(\mathbf{F}_{ve}\left|\boldsymbol{\theta}_{dec}\right.\right)=\mathcal{F}_{tra}\left(\mathbf{F}_{ve}\left|\boldsymbol{\theta}_{tra}\right.\right). \quad (7)$$

Incremental Knowledge. The three-stage incremental learning scheme contains a pioneer network training stage, a principal network incremental training stage, and a VRP stage.

Stage I: In the pioneer network training stage, we input the training sample $\mathbf{I}_{ori}$ and its label $\mathbf{I}_{gtve}^{(k)}$ of the k-th visual effect, and update the parameters $\hat{\boldsymbol{\theta}}_{enc}^{(k)}$ and $\hat{\boldsymbol{\theta}}_{dec}^{(k)}$ by minimizing the ℓ_1 distance between the prediction and the associated label:

$$\min_{\hat{\boldsymbol{\theta}}_{enc}^{(k)},\ \hat{\boldsymbol{\theta}}_{dec}^{(k)}}\left\|\mathcal{F}_{dec}\left(\mathcal{F}_{enc}\left(\mathbf{I}_{ori}\left|\hat{\boldsymbol{\theta}}_{enc}^{(k)}\right.\right)\left|\hat{\boldsymbol{\theta}}_{dec}^{(k)}\right.\right)-\mathbf{I}_{gtve}^{(k)}\right\|_1. \quad (8)$$

Stage II: In the principal network incremental training stage, we train the model on the MVE subsets with the guidance of five feature encoders obtained from the *Stage I*. Specifically, we calculate the relative entropy $d_{sim}^{(k)}$ and the normalized correlation ratio $r^{(k)}$ as shown in Eq. (3) and Eq. (4). The regularization constraint is performed when optimizing learned parameters with Eq (8). After this stage, we acquire an encoder $\mathcal{F}_{enc}\left(\cdot\left|\boldsymbol{\theta}_{enc}^{(k)}\right.\right)$ that is able to extract highly fused features, and obtain k decoders $\mathcal{F}_{dec}\left(\cdot\left|\boldsymbol{\theta}_{dec}^{(i)}\right.\right),\ i=1,2\ldots k$ to decode different visual effects.

Stage III: In the VRP stage, we input the origin images $\mathbf{I}_{ori}$ and the ground-truth of visual redundancy $\mathbf{I}_{gt}$ into the principal network with the MVE knowledge, and only update $\boldsymbol{\theta}_{dec}^{(vrp)}$ to minimize the mixed loss as follows:

$$\min_{\boldsymbol{\theta}_{dec}^{(vrp)}}\ \left\|\mathbf{I}_{vrp}-\mathbf{I}_{gt}\right\|_2+\lambda\left(1-SSIM\left(\mathbf{I}_{ori},\mathbf{I}_{rr}\right)\right), \quad (9)$$

where $\mathbf{I}_{vrp}$ denotes a predicted visual redundancy map, $\mathbf{I}_{rr}$ denotes a predicted redundancy-removed image, and $SSIM\left(\cdot\right)$ represents the Structural Similarity (SSIM) measurement. $\|\cdot\|_2$ represents the ℓ_2 distance, and λ is empirically set to 0.042.

Methods \ Datasets	*MCI-JCL*		*Shen2020*		*ciVRP-Set*		*Average*	
	PSNR ↓	SSIM ↑	PSNR ↓	SSIM ↑	PSNR ↓	SSIM ↑	PSNR ↓	SSIM ↑
Wang16TIP (Wang et al. 2016)	32.4477	0.8897	30.9624	0.8925	33.3540	0.8744	32.2547	0.8855
Wu17TIP (Wu et al. 2017)	31.9915	0.9011	30.6406	0.9037	32.5582	0.8742	31.7301	0.8931
Chen19TCSVT (Chen and Wu 2019)	31.4986	0.9070	30.3490	0.9108	31.4062	0.8811	31.0846	0.8996
Shen20TIP (Shen et al. 2021)	31.8017	0.8715	31.1341	0.9074	32.0921	0.8833	31.6760	0.8874
Wang22TII (Wang et al. 2022)	32.0945	0.8852	30.3418	0.8911	32.8045	0.8722	31.7469	0.8828
Jiang22TIP (Jiang et al. 2022)	30.7473	0.9162	30.1409	0.9284	30.8630	0.9296	30.5837	0.9247
Xie23AAAI (Xie et al. 2023a)	30.7109	0.9050	29.9875	0.9377	30.8625	0.9243	30.5203	0.9223
Proposed	**30.4911**	**0.9196**	**29.9312**	**0.9403**	**30.6433**	**0.9359**	**30.3552**	**0.9319**

Table 1: Objective quality comparison results on three datasets in terms of two widely-used quality indicators. The best result in each column is highlighted in bold. "*Average*" provides the mean result for each row.

Validation and Application

In this section, we have conducted objective, subjective, and compression experiments to evaluate the performance of our *mveVRP*. In addition, we have performed the ablation experiment to demonstrate the effectiveness of our three-stage incremental learning scheme as well as the encoder-decoder structure.

Experimental Setup

Implementation Details. The generation of visual effect subsets is based on the software platform *MATLAB R2021b*. We have implemented our *mveVRP* on *PyTorch* with the GPU device *NVIDIA Tesla A100*. Specifically, during the training stage, the optimizer is set to *Adam* with the default parameters, and the input sample size is cropped into 256×256. The batch size is set to 64, and the learning rate is set to 0.0001. During the testing stage, the whole image is taken as the input.

Noise ejection is a widely-used way (Wang et al. 2016; Shen et al. 2021; Xie et al. 2023a) to evaluate the impairment tolerance capability of a VRP model. $\mathbf{I}_{vrp}$ is injected into the original image $\mathbf{I}_{ori}$:

$$\mathbf{I}_{con} = \mathbf{I}_{ori} + \gamma \cdot \mathbf{s} \times \mathbf{I}_{vrp}, \tag{10}$$

where $\mathbf{I}_{con}$ represents the contaminated $\mathbf{I}_{ori}$, γ is an adjuster controlling the amplitude of injected noise, and $\mathbf{s}$ is a matrix with the same size of $\mathbf{I}_{ori}$ and randomly takes ± 1.

Comparison Methods and Datasets. In addition, we select 7 representative advances, including *Wang16TIP* (Wang et al. 2016), *Wu17TIP* (Wu et al. 2017), *Chen19TCSVT* (Chen and Wu 2019), *Shen20TIP* (Shen et al. 2021), *Wang22TII*(Wang et al. 2022), *Jiang22TIP* (Jiang et al. 2022), and *Xie23AAAI* (Xie et al. 2023a) for comparison.

Besides *ciVRP-Set*, we also select another two mainstream visual redundancy datasets with 1920×1080 resolution, including *MCL-JCI* (Jin et al. 2016) (50 reference images in natural scenes) and *SHEN2020* (Shen et al. 2021) (202 reference images in natural scenes) to evaluate the generalization capacity of our methods. We have divided *ciVRP-Set* into the training, validation, and testing sets at the ratio of 8:1:1. Meanwhile, *MCL-JCI* and *SHEN2020* are only used for the cross-dataset testing.

Objective Performance

The well-known PSNR and the image quality metric SSIM are adopted as two objective indicators to measure the impairment tolerance capability as defined in Eq. (10). In the experiments, we compute the PSNR and SSIM values between $\mathbf{I}_{ori}$ and $\mathbf{I}_{con}$. A promising VRP method should achieve a lower PSNR and higher SSIM, indicating that it predicts a larger amount of visual redundancy while maintaining the similar quality.

Table 1 provides the average PSNR and SSIM results. As seen, our *mveVRP* achieves the lowest PSNR and the highest SSIM in all three datasets, which demonstrates the superior performance of the proposed VRP method. Besides, it is worth noting that when most previous methods (except *Xie23AAAI* and *Jiang22TIP*) predict visual redundancy on *ciVRP-Set*, there is an obvious performance degradation (*e.g.*, higher PSNR and lower SSIM) than on two natural image datasets. This finding also suggests that complex content structures in CIs are associated with lower performance using traditional methods.

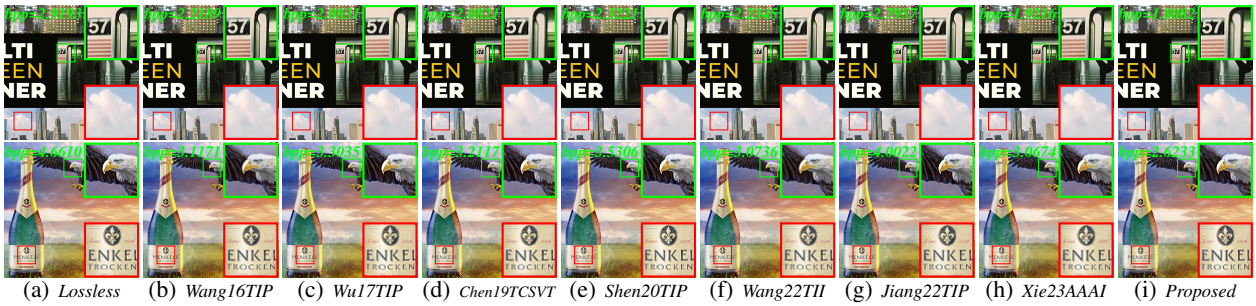
Subjective Performance

To demonstrate the visual quality, we conducted subjective experiments to compare the perceptual quality of noise injection guided by the eight VRPs. Specifically, we adjust γ in Eq. (10) to maintain the same level of injected noise (PSNR=30) and use the mean opinion score (MOS) as a subjective indicator. The subjective experiment follows the similar setting in Section , where each subject is asked to evaluate the image quality with a score range of [-3, +3]. The average MOS values for the eight methods (see Table 1) are -0.9913, -0.8613, -0.7550, -1.3413, -0.9188, -0.9775, -0.7038, and -0.5825, respectively. However, we have also noticed over-estimation in some facial regions, which is probably caused by that facial knowledge is not incorporated into the incremental training.

Ablation Study

In this section, we conduct additional experiments to evaluate the effectiveness of incremental learning and the backbone network structure.

Effect of Incremental Learning Stages. We have investigated four different training schedules, including direct



(a) *Lossless* (b) *Wang16TIP* (c) *Wu17TIP* (d) *Chen19TCSVT* (e) *Shen20TIP* (f) *Wang22TII* (g) *Jiang22TIP* (h) *Xie23AAAI* (i) *Proposed*

Figure 5: Visual comparisons of VVC compression guided by 8 VRPs. The red and green boxes show the zoom-in results.

Method	KF_1		$KF_1 \rightarrow KF_2$		$KF_1 \rightarrow KF_2 \rightarrow KF_3$	
	PSNR↓	SSIM↑	PSNR↓	SSIM↑	PSNR↓	SSIM↑
DT	30.7777	0.9243	30.7777	0.9243	30.7777	0.9243
ST	30.8999	0.9308	30.8884	0.9217	30.7840	0.9240
CT	30.9093	0.9278	30.8443	0.9259	30.7773	0.9266
VIT	**30.8345**	**0.9351**	**30.7531**	**0.9356**	**30.6433**	**0.9359**

Table 2: Ablation study on the effect of training stages.

Backbone structures	PSNR↓	SSIM↑
Unet	30.8844	0.9226
VGG	30.8013	0.9340
IL-VRPnet	**30.6433**	**0.9359**

Table 3: Ablation study of the backbone network structures.

training (DT), consecutive training(CT), separate training (ST), and proposed visual-effect incremental training (VIT) on our *ciVRP-Set*. DT means a direct training on the visual redundancy ground-truth without *Stage I* and *Stage II*. ST means a separate training with visual effect groups without *Stage II*. CT denotes the inherited learning of the MVE subsets in turns without any regulation in *Stage II*. We compare the objective performance in Table 2. As seen, VIT achieves the highest SSIM, which demonstrates the effectiveness of the proposed learning scheme.

Effect of the Network Structure. We first replace our *IL-VRPnet* with *Unet*-like (Ronneberger, Fischer, and Brox 2015), whose encoder is a *VGG*-like (Simonyan and Zisserman 2015) structure. Then, we replace the `resnet18` encoder with the *VGG*-like encoder. We train and test the above backbone networks on *ciVRP-Set*. Table 3 provides the results of SSIM and PSNR, indicating the performance decline when using *Unet* and *VGG* as replacements. This further shows the effectiveness of the proposed *IL-VRPnet*.

Compression Application

We integrate the proposed *mveVRP* into the VVC platform and conduct compression experiments on *ciVRP-Set*. In VVC, a smaller prediction residual value typically leads to a lower compressed bitrate. However, blindly reducing the prediction residuals may degrade the perceptual quality. We utilize the VRP to reduce the residuals. Specifically, we modify the prediction residuals as follows.

$$\widehat{\mathbf{R}}(i,j) = \begin{cases} 0, & if\ |\mathbf{R}(i,j)| \leq \mathbf{I}_{vrp}(i,j)\ and\ \sigma_{ij}^2 > \sigma_{block}^2, \\ min(\mathbf{R}(i,j) \cdot \frac{|\mathbf{I}_{vrp}(i,j)|}{\sigma_{block}^2},\ \mathbf{R}(i,j) + \mathbf{I}_{vrp}(i,j)),\ else\ if\ \mathbf{R}(i,j) < 0, \\ max(\mathbf{R}(i,j) \cdot \frac{|\mathbf{I}_{vrp}(i,j)|}{\sigma_{block}^2},\ \mathbf{R}(i,j) - \mathbf{I}_{vrp}(i,j)),\ else. \end{cases}, \quad (11)$$

where $\widehat{\mathbf{R}}(i,j)$ denotes the processed prediction residual and $\mathbf{R}(i,j)$ denotes the prediction residual at pixel (i,j). σ_{ij}^2 denotes the variance of the same pixel position, which is calculated from the surrounding 3×3 pixels. σ_{block}^2 denotes the variance of the current encoding block in VVC.

We have also implemented the VRP-guided compression on the VVC software *VTM 6.0*. To show the improvement in compression efficiency, we calculate the average bit-rate saving between the VVC lossless compression and the VRP-guided one. The VRP-guided results in Table 1 are -26.80%, -30.19%, -31.12%, -22.36%, -26.47%, -13.74%, -32.71%, and -35.08%, respectively. Our *mveVRP* achieves the highest bit rate savings. Moreover, we show the visual performance and bits per pixel (`bpp`) after compression in Figure 5. As seen, our method obtains the lowest `bpp` under the same perceived quality compared to the VVC lossless coding.

Conclusion

This paper presents a novel multi-visual effect (MVE)-driven visual redundancy prediction (*mveVRP*) for composite images (CIs). Specifically, we first establish a benchmark dataset called *ciVRP-Set* to facilitate future investigation on VRP for CIs. To better model the complex association and fusion mechanism of MVE, we propose an incremental learning method that captures the knowledge flow of visual effects and generates strong descriptive features to guide deep VRP. Further, we design incremental learning rules to mitigate the issue of "catastrophic forgetting" when learning multiple visual knowledge. Finally, we implement the incremental learning scheme in a three-stage manner using an encoder-decoder network structure. The effectiveness of our proposed method is demonstrated through extensive objective, subjective, and compression experiments. We believe that this exploration will contribute to the advancement of compression efficiency and enhance the visual experience in popular multimedia applications such as virtual reality (VR) and artificial intelligence (AI) generation.

References

Bae, S.-H.; and Kim, M. 2017. A DCT-Based Total JND Profile for Spatiotemporal and Foveated Masking Effects. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6): 1196–1207.

Bai, Y.; Yang, X.; Liu, X.; Jiang, J.; Wang, Y.; Ji, X.; and Gao, W. 2022. Towards end-to-end image compression and analysis with transformers. In *AAAI Conference on Artificial Intelligence (AAAI)*, 104–112.

Chen, H.; Miao, F.; Chen, Y.; Xiong, Y.; and Chen, T. 2021. A Hyperspectral Image Classification Method Using Multifeature Vectors and Optimized KELM. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14(1): 2781–2795.

Chen, Z.; and Wu, W. 2019. Asymmetric Foveated Just-Noticeable-Difference Model for Images With Visual Field Inhomogeneities. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11): 4064–4074.

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

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2022. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385.

Dong, S.; Hong, X.; Tao, X.; Chang, X.; Wei, X.; and Gong, Y. 2021. Few-shot class-incremental learning via relation knowledge distillation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 1255–1263.

Ferzli, R.; and Karam, L. J. 2009. A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB). *IEEE Transactions on Image Processing*, 18(4): 717–728.

Hadizadeh, H.; Rajati, A.; and Bajić, I. V. 2017. Saliency-Guided Just Noticeable Distortion Estimation Using the Normalized Laplacian Pyramid. *IEEE Signal Processing Letters*, 24(8): 1218–1222.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, Y.; Chen, B.; Qin, S.; Li, J.; Wang, Y.; Dai, T.; and Xia, S.-T. 2023. Learned Distributed Image Compression with Multi-Scale Patch Matching in Feature Domain. In *AAAI Conference on Artificial Intelligence (AAAI)*, 4322–4329.

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(1): 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.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up GANs for Text-to-Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–36.

Lin, W.; and Ghinea, G. 2022. Progress and Opportunities in Modelling Just-Noticeable Difference (JND) for Multimedia. *IEEE Transactions on Multimedia*, 24(1): 3706–3721.

Liu, H.; Zhang, Y.; Zhang, H.; Fan, C.; Kwong, S.; Kuo, C.-C. J.; and Fan, X. 2019. Deep learning-based picture-wise just noticeable distortion prediction model for image compression. *IEEE Transactions on Image Processing*, 29(1): 641–656.

Michieli, U.; and Zanuttigh, P. 2021. Continual Semantic Segmentation via Repulsion-Attraction of Sparse and Disentangled Latent Representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1114–1124.

Nami, S.; Pakdaman, F.; Hashemi, M. R.; and Shirmohammadi, S. 2023. BL-JUNIPER: A CNN-Assisted Framework for Perceptual Video Coding Leveraging Block-Level JND. *IEEE Transactions on Multimedia*, 25(1): 5077–5092.

Qiang, S.; Hou, J.; Wan, J.; Liang, Y.; Lei, Z.; and Zhang, D. 2023. Mixture Uniform Distribution Modeling and Asymmetric Mix Distillation for Class Incremental Learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 9498–9506.

Qin, G.; Hu, R.; Liu, Y.; Zheng, X.; Liu, H.; Li, X.; and Zhang, Y. 2023. Data-Efficient Image Quality Assessment with Attention-Panel Decoder. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2091–2100.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Shen, X.; Ni, Z.; Yang, W.; Zhang, X.; Wang, S.; and Kwong, S. 2021. Just Noticeable Distortion Profile Inference: A Patch-Level Structural Visibility Learning Approach. *IEEE Transactions on Image Processing*, 30(1): 26–38.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 1–14.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; and Wang, H. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*, 8968–8975.

Tian, T.; Wang, H.; Kwong, S.; and Kuo, C.-C. J. 2021. Perceptual Image Compression with Block-Level Just Noticeable Difference Prediction. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(4): 1–15.

Wang, C.; Li, S.; Liu, Y.; Meng, L.; Zhang, K.; and Wan, W. 2023. Cross-scale feature fusion-based JND estimation for robust image in DWT domain. *OPTIK*, 272(1): 1–21.

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(1): 1440–1444.

Wang, M.; Xu, Z.; Liu, X.; Xiong, J.; and Xie, W. 2022. Perceptually quasi-lossless compression of screen content data via visibility modeling and deep forecasting. *IEEE Transactions on Industrial Informatics*, 18(10): 6865–6875.

Wang, S.; Ma, L.; Fang, Y.; Lin, W.; Ma, S.; and Gao, W. 2016. Just noticeable difference estimation for screen content images. *IEEE Transactions on Image Processing*, 25(8): 3838–3851.

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)*, 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 (MM)*, 13749–13765.

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

Yin, G.; Wang, W.; Yuan, Z.; Han, C.; Ji, W.; Sun, S.; and Wang, C. 2022. Content-variant reference image quality assessment via knowledge distillation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 3134–3142.

Yin, H.; Wang, H.; Yu, L.; Liang, J.; and Zhai, G. 2023. Feedforward and Feedback Modulations Based Foveated JND Estimation for Images. *ACM Transactions on Multimedia Computing, Communications and Applications*, 1(1): 1–22.