sensors

MDPI

Article

Semantically Adaptive JND Modeling with Object-Wise Feature Characterization, Context Inhibition and Cross-Object Interaction

Xia Wang [1,2], Haibing Yin [1,2,*], Yu Lu [1], Shiling Zhao [1,2] and Yong Chen [3]

[1] School of Communication Engineering, Hangzhou Dianzi University, No. 2 Street, Xiasha, Hangzhou 310018, China
[2] Lishui Institute of Hangzhou Dianzi University, Nanmingshan Street, Liandu, Lishui 323000, China
[3] Hangzhou Arcvideo Technology Co., Ltd., No. 3 Xidoumen Road, Xihu, Hangzhou 310012, China
* Correspondence: yhb@hdu.edu.cn

Abstract: Performance bottlenecks in the optimization of JND modeling based on low-level manual visual feature metrics have emerged. High-level semantics bear a considerable impact on perceptual attention and subjective video quality, yet most existing JND models do not adequately account for this impact. This indicates that there is still much room and potential for performance optimization in semantic feature-based JND models. To address this status quo, this paper investigates the response of visual attention induced by heterogeneous semantic features with an eye on three aspects, i.e., object, context, and cross-object, to further improve the efficiency of JND models. On the object side, this paper first focuses on the main semantic features that affect visual attention, including semantic sensitivity, objective area and shape, and central bias. Following that, the coupling role of heterogeneous visual features with HVS perceptual properties are analyzed and quantified. Second, based on the reciprocity of objects and contexts, the contextual complexity is measured to gauge the inhibitory effect of contexts on visual attention. Third, cross-object interactions are dissected using the principle of bias competition, and a semantic attention model is constructed in conjunction with a model of attentional competition. Finally, to build an improved transform domain JND model, a weighting factor is used by fusing the semantic attention model with the basic spatial attention model. Extensive simulation results validate that the proposed JND profile is highly consistent with HVS and highly competitive among state-of-the-art models.

Keywords: just noticeable difference (JND); visual attention; semantic visual features; biased competition

Citation: Wang, X.; Yin, H.; Lu, Y.; Zhao, S.; Chen, Y. Semantically Adaptive JND Modeling with Object-Wise Feature Characterization, Context Inhibition and Cross-Object Interaction. *Sensors* **2023**, *23*, 3149. https://doi.org/10.3390/s23063149

Academic Editors: Guangtao Zhai, Xiongkuo Min, Menghan Hu and Wei Zhou

Received: 15 December 2022
Revised: 12 January 2023
Accepted: 15 January 2023
Published: 15 March 2023

1. Introduction

Regarding the human visual system (HVS) as a communication system with limited bandwidth and processing capacity, it is continuously receiving data input. As a result, the HVS can only sense variations in signal strength above a certain threshold, which is termed as JND [1]. Research on JND can be traced back to the experimental psychology of Ernst Weber [2] in the 19th century and was transferred to the field of digital multimedia at the end of the 20th century. Visual JND can build computational models by virtue of relevant physiology, psychology and neural research, combined with feature detection. Over the past decades, JND has proven to be a multi-factorial problem, including contrast sensitivity function (CSF), luminance adaptation (LA), masking effects and visual attention. An overview of these factors can be found in the survey by researchers in [3]. Using the computational domain as a classification criterion, there are two branches of existing JND models, i.e., the pixel domain (where JND thresholds are computed directly for each pixel) [4–13] and the transform domain (where the image is first transformed into a subband domain, and then JND thresholds are calculated for each subband) [14–24]. Nevertheless, both model types generally follow the same design philosophy of simulating the visual-masking effects of several elements before combining (multiplying or adding) them to obtain an

overall JND estimation. The pixel-domain JND model has undergone a protracted evolution. Based on the study of luminance and contrast, the pioneering work of Chou et al. [4] computed luminance adaptation and contrast masking (CM), using the winner as the JND thresholds. Yang et al. [5] considered the overlap effect between LA and CM, and proposed a nonlinear additivity model of masking (NAMM). Furthermore, the image can be decomposed into plain, texture and edge regions for more accurate modeling of CM [6]. Since the visual sensitivity decreases with increasing retinal eccentricity, the fovea masking was introduced [7]. Inspired by the internal generation mechanism, the masking effects of ordered and disordered components were quantified based on free energy theory [8], and the pattern masking was extended sequentially [9]. In addition, the RMS contrast was used as the spatial CSF of JND in the pixel domain [10]. Based on hierarchical predictive coding theory, self-information and information entropy were utilized to calculate the perceptual surprise and perceptual suppression effects at different levels to improve the accuracy of the JND model [12].

Transform coding is a widespread means of mainstream image/video coding, and transform domain JND modeling is also a significant research area. In 1992, Ahumada [14] proposed the first DCT domain JND model combining spatial CSF and LA. Based on [14], Watson [15] introduced the CM effect and proposed the DCTune model, which laid the research foundation for the DCT domain JND modeling. Subsequently, a more realistic LA function was proposed to improve JND estimation by integrating a block classification (plain, edge, and texture) strategy [16]. For video, corresponding video JND models for the DCT domain were proposed considering spatiotemporal CSF and eye movement compensation [17], LA effect based on gamma correction and CM effect based on more accurate block classification [18], texture complexity and frequency, visual sensitivity and visual attention [19], the various sizes of DCT blocks (from 4×4 to 32×32) [20], motion direction [21], fovea masking [22], temporal duration and residual fluctuations [23].

In recent years, machine-/deep-learning-based modeling of visual perception has become a new research trend due to the rapid development of deep neural networks [25–33]. To fill the vacancy of JND databases for image and video compression, several scholars have suggested MCL-JCI [34], JND-pano [35], MCL-JCV [36], VideoSet [37], etc. Based on these datasets, various modeling techniques for JND eventually emerged, most notably subjective data regression [25,27], binary classification [26], picture/video-wise JND (PWJND/VWJND) or satisfied user ratio (SUR) modeling [27–30], and finding appropriate weighting factors for the JND models [25,26]. In particular, motivated by the smoothing performance (non-smoothed regions have better capability for hiding noise than smoothed regions), Wu et al. [31] proposed an unsupervised learning method for generating JND images in the pixel domain. Considering that JND should be evaluated in the human brain perception domain, Jin et al. [33] presented an HVS-based signal degradation network and employed visual attention loss to further regulate JND estimation. In addition, JND modeling also extends to machine vision [38,39]; however, this specific topic is outside the scope of this paper. Visual attention is a key attribute of HVS [40], and models combined with visual attention can improve the accuracy of JND threshold. In a study by Chen et al. [7], attention-induced fixation was detected and a scheme was designed to reflect the increase in JND with increasing retinal eccentricity. Video JND estimation for smooth pursuit eye movements (SPEM) and non-SPEM situations was accomplished by investigating the relationship between JND thresholds and parameters (spatial frequency, eccentricity, retinal motion velocity) [22]. Considering the special focus of HVS on people, Wang et al. [41] combined Itti's [42] and Judd's [43] attention model (with face detector and person detector) to adjust the JND threshold.

The main feature of the HVS attention mechanism is the information-selection process, in which only a small number of visual signals are transmitted to the brain for processing. Indeed, the object-based attention theory suggests that humans are drawn to objects and advanced concepts [44]. Although Wang at al. [41] considered the advanced semantics of faces, there are often more advanced semantics in images than in plain faces. In any

case, traditional visual attention models mainly consider low-level features such as color, luminance, and orientation, while underestimating high-level semantic information, which leads to reduced JND model accuracy under the attention mechanism of HVS. This is mainly due to the difficulty of semantic feature extraction and fusion. Meanwhile, it is well known that deep neural networks are full of semantic information [45], but current JND models based on deep learning mainly try to construct PWJND/VWJND or SUR without fully considering semantic character behaviors.

In addition, neurons transmit and express visual information by consuming energy [46]. The structure and function of neural networks follow the essential principles of resource allocation and minimization of energy constraints. Therefore, not all stimuli trigger neuronal responses, i.e., biased competition for visual attention influences neuronal activity [47]. The theory of biased competition suggests that selectivity of one (or more) points in the visual process is caused by interaction between visual objects for neural representations. More interestingly, with the limited attentional resources, there will always be certain items and their visual features that win out in the competition. Therefore, it is inevitable to consider such interaction while constructing a semantics-based attention model.

From the above discussion, it is clear that semantics is crucial for accurate estimation of JND thresholds. Furthermore, it should be noted that this paper concentrates on the tuning scheme for DCT-based JND estimation because the majority of image/video processing, especially in image/video compression, is conducted in the DCT domain. Particularly, this paper argues that there are three semantic features to be considered in the attention allocation problem of videos, namely, object, context, and cross-object. Since it is troublesome to extract semantic attributes manually, this paper uses grad-CAM [48] and VNext [49] to extract semantic objects and features through deep learning. Based on this, attention effects induced by object-based semantic features are analyzed, including the sensitivity and size of semantic objects and central bias. Second, contextual information is another vital element that enhances instance differentiation and is considered as a form of center-surround contrast. Furthermore, considering limited attentional resources and infinite information, this paper describes the attentional competition effects of different semantic features (including focus intensity, spatial distance, and relative area) in terms of cross-objects to achieve an accurate model of semantic attention.

Specifically, inspired by the framework of [23], this paper proposes a statistical probability model corresponding to the feature parameters and quantifies the response of visual attention in a perceptual sense using information theory. Then, the attentional competition factor for calibrating the semantic attention model is proposed. Finally, the adaptive semantic attention weight is obtained by unifying and fusing the perceptual feature parameters, and the JND model is adjusted using this weight.

The rest of this paper is organized as follows. Section 2 section presents three features and describes the quantification strategies for these parameters. Section 3 section analyzes the interaction between the objects and context. Section 4 section depicts the fusion approaches of these parameters together with the competition properties of attention. Section 5 part details the proposed JND profile. Simulation results and conclusions are given in the Sections 6 and 7, respectively.

2. Object-Wise Semantics Parameter Extraction and Quantification

As stated, the literature lacks a clear understanding of the mechanism of interaction between semantic perception and semantic features. This motivates us to formulate an effectual JND profile leveraging the HVS semantic traits, which is subject to challenges. The first issue is how to precisely extract and quantify the feature parameters that thoroughly depict the perceptual response of semantic HVS features in videos, such as semantic sensitivity, objective area, central bias.

2.1. Semantic Instances Extraction

Accurate extraction of high-level semantic features of videos helps to better visual attention modeling. On that basis, this paper uses grad-CAM for the extraction of high-level semantic features. In this network, the further the layer, the ampler the semantics is and the more spatial information can be preserved. In this respect, grad-CAM contains all the semantics of our target of interest. For more details, please see [48]. With grad-CAM, we can draw heat maps which are shown in Figure 1 below. Moreover, to effectively utilize the various properties of semantic objects, this paper uses VNext [49] to extract semantic objects, and the results are shown in Figure 2.

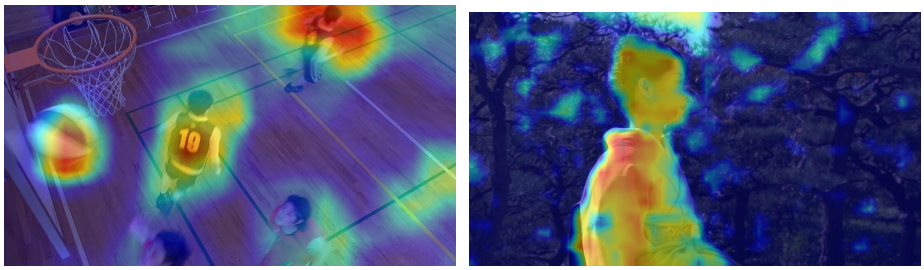


Figure 1. The grad-CAM heat-map results.

Figure 2. The VNext extraction results.

2.2. Semantic Sensitivity Quantification

The research [50] has shown that attention is focused on informative content. One definition of attention describes it as a limited resource for information processing, with the area of interest carrying the bulk of the information transmitted to the brain. Hence, Bruce et al. [51] pointed out that Shannon self-information can be used to outline attention. Whereupon, the attention value of a pixel located at (x, y) in a video frame is defined below.

$$A(x, y) = -\log_2 P(F = f(x, y)) \tag{1}$$

where F denotes a random variable of the feature of pixel (x, y), $f(x, y)$ represents the image features of pixel (x, y), $P(F = f(x, y))$ indicates the probability of feature F.

The concept of constructing statistically sparse representations of images appears to be fundamental to the primate visual system, as evidenced by a large body of research. That is, to some extent, computing the probability distribution of features is essentially a matter of statistical feature sparsity. If a pixel associated with a feature has a lower probability of appearing, it has a better chance of attracting attention because it conveys a greater amount of information. In view of this, the probability density function (PDF) of feature F is measured as follows by performing histogram statistics on the whole picture.

$$\text{PDF} = hist(F = f(x, y)) = P(F = f(x, y)) \tag{2}$$

Accordingly, the initial attention based on semantic sensitivity (m) can be calculated as follows.

$$I(m) = -\log_2 hist(f(m)) \tag{3}$$

It is worth noting that instead of using a uniform-fitting function, to better suit the image content, the histogram of each frame is adaptively fitted. Figure 3 shows the histogram of the 2nd frame in the HEVC standard test sequence "BasketballDrill" and its PDF fitting curve, separately.

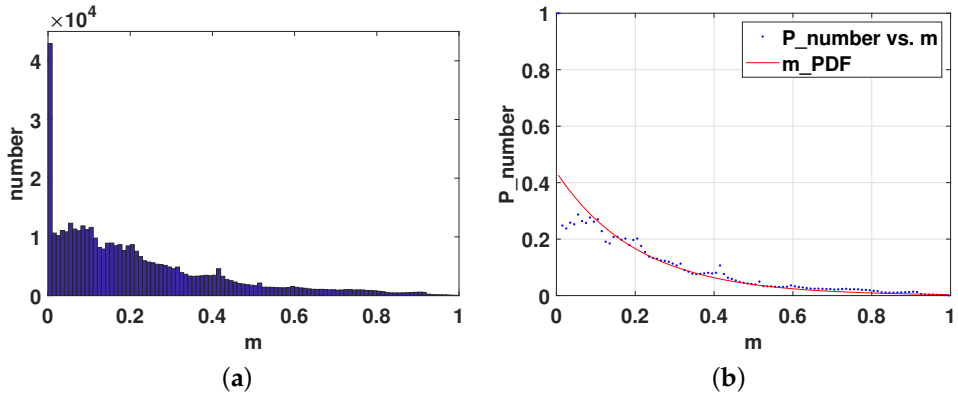


Figure 3. (**a**) A semantic sensitivity histogram; (**b**) A semantic sensitivity PDF curve.

2.3. Objective Area Quantification

The literature demonstrates that the larger the area, the higher the probability of attention, i.e., the probability of attention of semantic objects is proportional to size, however, this property becomes smooth after a certain saturation threshold [52]. This relationship can be summarized below using a piecewise function [53], with values within [0,1].

$$-\log_2(p(\eta_j)) = \begin{cases} -\frac{\ln(1-\eta_j)}{c_1} & 0 \le \eta_j \le t_1 \\ 1-\exp(-c_2 \cdot \eta_j) & t_1 < \eta_j \le t_2 \\ s_2 + c_3 \ln(1-\eta_j + t_2) & t_2 < \eta_j \le t_3 \\ s_3 \cdot \exp(-c_4(\eta_j - t_3)) & t_3 < \eta_j \le t_4 \\ 0 & t_4 < \eta_j \le 1.0 \end{cases} \tag{4}$$

where $0 < j \le N$, N is the number of semantic objects in an image frame, η_j represents the size of semantic object, which is calculated as: $\eta_j = B_j/(W \times H)$, B_j represents the number of objective pixels, and W, H refer to the width and height of the image frame. Figure 4 shows the relationship between area and attention.

Likewise, for the same size semantic objects, attention will be affected by the aspect ratio. As is known to all, the aspect ratio of current popular displays is 16:9. It follows that the closer the aspect ratio of a semantic object is to our vision field, the more it attracts attention.

Figure 5 is an illustration for the aspect ratio. For a more clear description, the blue rectangle on the left is labeled as A, the one on the right is B, and the container containing these two rectangles is C. Despite having the same area, A and B have various aspect ratios. The aspect ratios of A and C are 16:9, while B is 9:16. It is obvious from the example that rectangle A attracts more attention than rectangle B. Note that the rectangle is used merely for convenience of observation, and the results hold true other shapes as well.

Therefore, the joint effect of semantic object size, the aspect ratio (R_r) is used to measure the attention of semantic object size.

$$I(\eta_j) = -\log_2[p(\eta_j)] - \alpha_1 \log_2\left(\frac{R_{r_j}}{\beta_1}\right)^{\beta_2} \tag{5}$$

where α_1, β_1, β_2 are all adjustment control parameters, r_j denotes the aspect ratio of the j-th semantic object and is normalized to [0,1].

The model is consistent with HVS subjective perception that HVS perceptual attention intensity increases with the size of the object but there is a saturation point. The reason is

that when the object area is too large, the target and background are non-separable, i.e., the target is likely to be the background component at this point. In addition, targets with an aspect ratio closer to 16:9 attract greater HVS attention, which raises perceptual sensitivity and lowers JND thresholds.

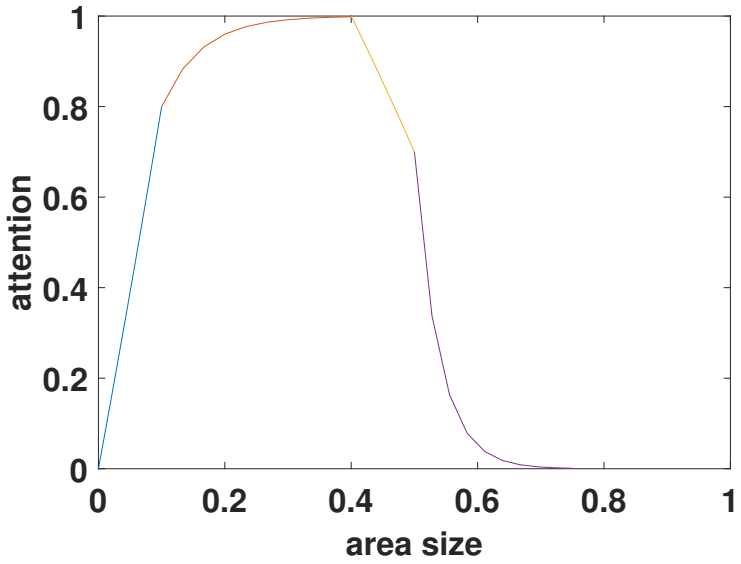


Figure 4. The function map of Equation (4).

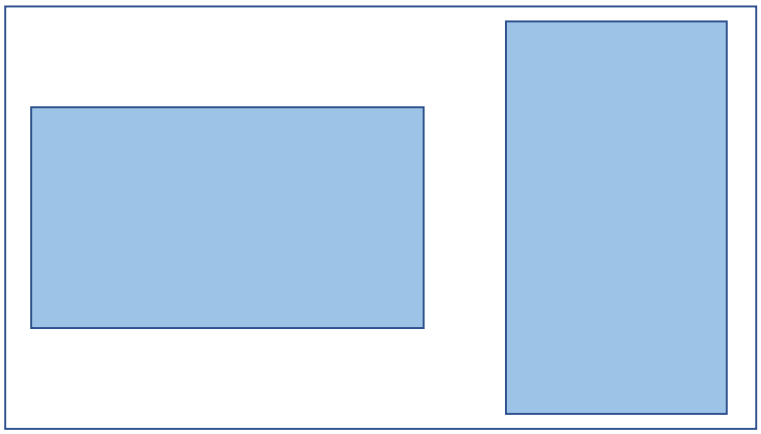


Figure 5. The illustration for the aspect ratio.

2.4. Central Bias Quantification

It is stated that most images have the foreground object in the center of the image frame [43]. It follows that when a person looks at an image, he or she habitually looks at the center of the image first. Eye-movement experiments have shown that HVS tends to focus on the center of the image, and the deviation from the target center to the image center, i.e., the central bias, has been considered a noteworthy prior in attention modeling [44]. Based on this prior knowledge, many attention modeling algorithms based on central bias use various methods to increase the attention value at the image center location as a way to highlight the salient targets in the image.

Nevertheless, the feedback brought by the central bias is not always favorable. For some unconventional cases where the target is off-center in the image, attention enhancement at the image center can lead to over/under estimation of the target far from the image center. In general, the problem with the classical center-priority based attention modeling algorithm is that it tends to lack flexibility, which leads to significant inaccuracies when the foreground targets of the image are not positioned as expected.

The formation of a top-notch attention map is the ultimate goal of attention modeling. The complexity of attention modeling is greatly reduced and the quality of the final attention map is improved if the locations of the salient targets are approximated before proceeding to formal attention modeling.

In this letter, the object (M) closest to the image center is calculated, and the central bias is constructed with M as the experienced center of vision. The closer an object is to the visual center, the higher probability it is to be noticed, i.e., the attention probability of semantic objects is inversely proportional to the central bias distance, which is denoted as follows.

$$P(A|d_j) = 1 - \log 2\left[1 + \left(\frac{d_j}{\omega_1}\right)^{\omega_2}\right] \tag{6}$$

where d_j is the distance from the semantic object to M, and normalized to [0,1].

$$d_j = mean(\sum_{f \in O_j} \left\|(x_f, y_f) - (x_M, y_M)\right\|^2) \tag{7}$$

here, O_j is a semantic object, f denotes the pixel of O_j, (x_f, y_f) represents the coordinates corresponding to f, (x_M, y_M) means the center coordinates of M, $\|\cdot\|$ is the euclidean distance, ω_1, ω_2 serve as the adjustment parameters.

As illustrated in Figure 6, the modified central bias result is better in line with HVS perception. Since the target M in the figure is closest to the image center, it obviously draws the highest attention, which is shown by the brightest brightness. In this instance, the target M stands in for the image center as the experienced center of vision and the closer the other objects are to the target M, the more attention they attract, and the brighter they are presented in the figure.

(**a**) (**b**)

Figure 6. (**a**) The traditional central bias map; (**b**) The modified central bias map.

3. Context-Wise Attentional Inhibition Quantification

Given that HVS still beats the state-of-the-art computer vision systems, modeling visual attention in natural scenes is a hot topic of research today. How is such great effectiveness achieved? The ability of HVS to utilize the context of a scene to direct attention before most items are recognized appears to be a key factor compared to artificial systems [54].

How does the context of a scene serves to focus attention on objects in the scene? Davenport and Potter [55] suggest that there is an interaction between objects and context during scene processing, which is supported by research. On this basis, this paper conceptualizes the local spatial arrangement of the image background as an alternative to the context and focuses primarily on the complexity of the patterns.

In most situations, the region between the target and its background is heterogeneous and has a complex organization. The density of the visual information presented is normally used to describe the complexity of the visual background. In general, a scene is complex when its background displays a large amount of information and the unpredictability of this information is considerable [56]. Typically, the complexity of the visual context usually has

a detrimental impact on the visual task. The higher the visual complexity, the longer the response latency of the nerve cells and the stronger the degree of inhibition of the target [57].

As the contextual complexity of the target suppresses the attention strength, the more complex the pattern of the background, the stronger the suppression, while this is going on, the findings of numerous human visual field experiments indicate that the human binocular visual field is approximately circular (oval), with the proximate form depicted in Figure 7 [58]. Hence, to represent the inhibition of each semantic target background, we utilize the average of the contextual complexity in the circumcircle of semantic targets to simplify the calculation. Since entropy reflects the chaotic degree of a system, we estimate contextual complexity by local entropy. As a result, attention and contextual complexity are related in the following way.

$$P(\bar{A}|u_j) = \log 2\left[1 + \left(\frac{u_j}{\xi_1}\right)^{\xi_2}\right] \tag{8}$$

where u_j is the contextual complexity, and normalized to [0,1].

$$u_j = mean(entfilt(circle(O_j) - O_j)) \tag{9}$$

here, ξ_1 and ξ_2 serve as the scaling factors, $circle(\cdot)$ denotes circumcircle, $entfilt(\cdot)$ is the function to calculate the local entropy.

Figure 8 shows the inhibitory effect of contextual complexity on semantics.

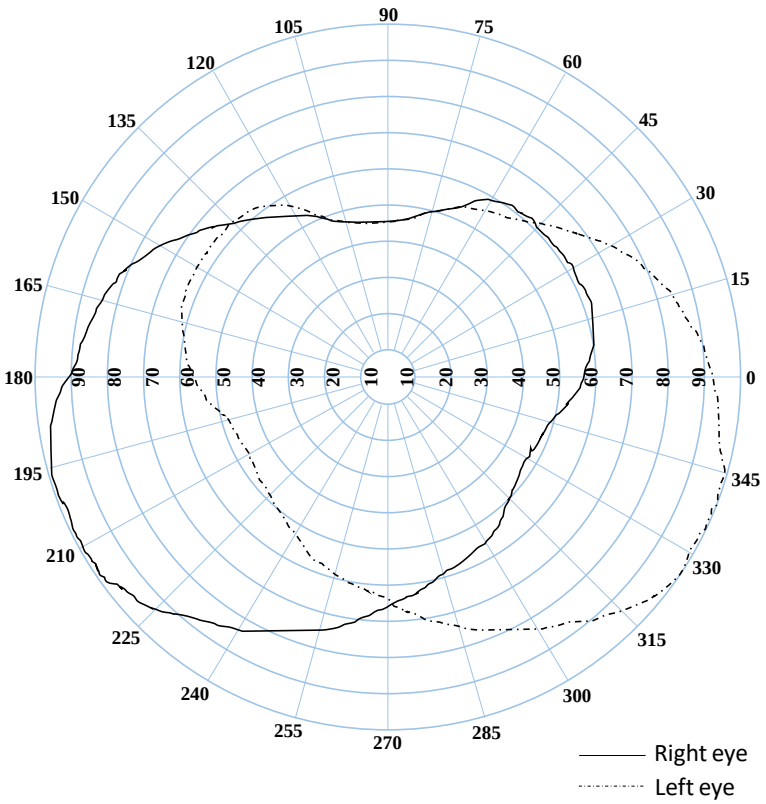
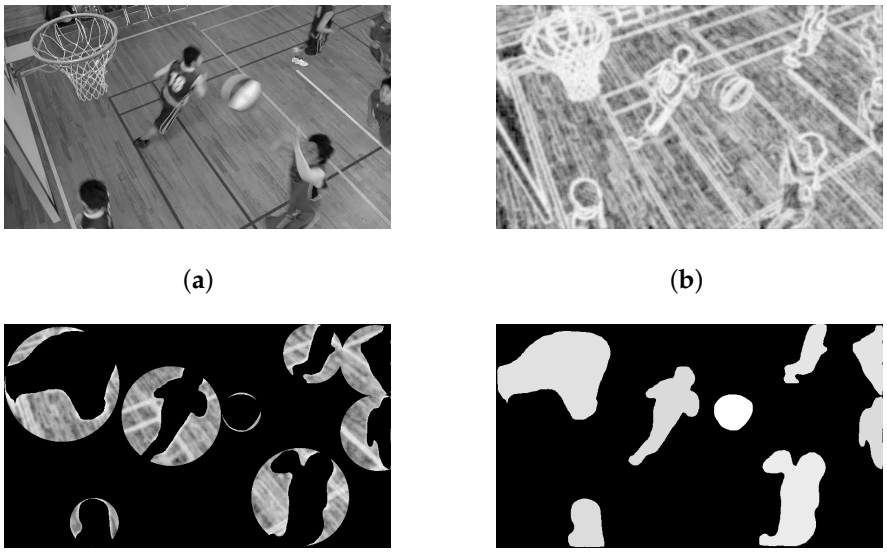


Figure 7. Areas of vision field.



(**a**) (**b**)

(**c**) (**d**)

Figure 8. (**a**) The 212th original frame of the "Basketball Drill"; (**b**) The complexity map of the whole frame; (**c**) The contextual complexity map of the semantics; (**d**) The inhibitory effect of contextual complexity on semantics.

4. Fusion Strategies and Cross-Object-Wise Attentional Competition

Furthermore, given elaborately chosen feature parameters, the second sore point is how to quantify the interaction among these feature parameters, i.e., how to fuse these heterogeneous feature parameters.

Consolidating the semantics-based attention model is the aim of this part. In general, different feature maps do not have the same attributes, and different features contain complementary global contextual information and local detailed information between them. Consequently, to obtain reliable results for the attention distribution, it is essential to select the appropriate weights for each feature map.

In this paper, to estimate the weights of the feature maps above, the gradient change value of each feature map is first calculated, where the greater the variation of the data, the more information can be gleaned. The gradient variation is then used to estimate the relative weights of these feature maps. Finally, utilizing their individual weights, the feature maps are normalized and combined into a single attention map.

$$att_{pre} = \frac{\lambda_1 \cdot I(m) + \lambda_2 \cdot I(\eta) + \lambda_3 \cdot P(A|d)}{P(\bar{A}|u)} \tag{10}$$

here, $\lambda_i (i = 1, 2, 3)$ serves as the weighting factor, and is calculated as follows.

$$\lambda_i = \frac{vg(z_i)}{\sum_{i=1}^{3} vg(z_i)} \tag{11}$$

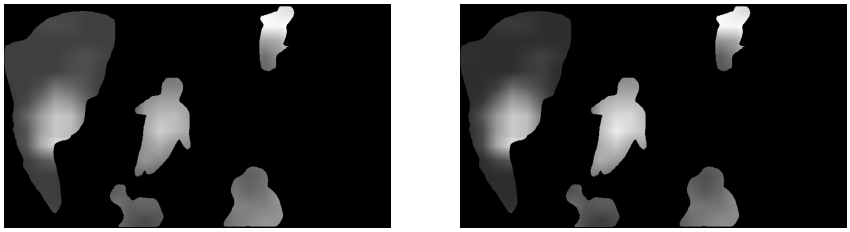
$$vg(z_i) = \frac{\max(g(z_i)) - \min(g(z_i))}{mean(g(z_i))} \tag{12}$$

$$g(z_i(O_m)) = \sum_{m=1, m \neq n}^{N} |z_i(O_m) - z_i(O_n)| \tag{13}$$

where $z_i \in \{I(m), I(\eta), P(A|d)\}$ is the attention map of different semantic features.

Figure 9 displays the initial attention map with different weighting factors. As illustrated, compared to map (a), the map (b) with the gradient variation based weights is more

consistent with the HVS subjective perception. This is partly due to the fact that we fully take the content characteristics of different feature maps into account.



(**a**) (**b**)

Figure 9. (**a**) The initial attention map with the identical weights; (**b**) The initial attention map with the gradient variation based weights.

However, there is a competitive mechanism that manifests itself as a relative enhancement of responses to task-relevant objects or a relative inhibition of neglected objects within the brain [59]. That is, when modeling the semantic attention model, it is not sufficient to consider only the semantic object itself, but also to take full account of the biased competition between different objects.

- Focused Intensity: when multiple stimuli are present in the visual field at the same time, inhibitory competition in the visual cortex affects the allocation of attention, which probably stem from the inability to bias the interaction toward a particular object [60]. In reality, the brain's representation of information is essential for human vision, and regardless of the distance between neurons or neuronal populations, they do not operate independently, but constantly interact with each other. Because their activities are interconnected and competing, they cannot provide "independent attentional resources" [61]. That is, when there are multiple semantic objects in an image, attention to each semantic object is impacted by the other objects. The higher the intensity of the focus, the more pronounced the suppression of the other objects;
- Spatial Distance: event-related potentials, functional MRI investigations as well as single-cell recordings in monkeys [62] and humans [63] have pointed out that neural enhancement in attentional focus may be followed by neural inhibition in peripheral regions. Detection is slower and discrimination performance is poorer at interference locations close to the target compared to interference locations far from the target [64]. In the receptive fields of cells, shifting attention from one stimulus to another can have a strong effect when two stimuli are in close proximity to each other. For stimuli that are far apart, the effect is much less [59]. In other words, the more spatially distant the objects are from each other, the smaller the inhibitory impact;
- Relative Area: since an object generally has several neighbors, the relative area of an object shows with its neighbors affects its attention competition. The rationale for selecting relative area is that, even if the object has an attractive intrinsic area, it may not stand out unless it exhibits the greatest contrast if all of its neighbors also have attractive areas [53]. In particular, the higher the area contrast, the more strongly other objects will be suppressed.

Based on the points discussed above, the attentional competition weight is defined as follows:

$$C_A(k) = \left(\sum_{k=1, k \neq l}^{N} D_\tau(k,l) times D_\nu(k,l) \right) \cdot S_\delta(k) \tag{14}$$

in which, $D_\tau(k,l)$, $D_v(k,l)$ are the average focused intensity distance and the spatial proximity distance between the k-th, l-th semantic objects, respectively.

$$D_\tau(k,l) = \exp\left(\frac{F(O_k) - F(O_l)}{F(O_k) + F(O_l)}\right) \tag{15}$$

$$F(O_k) = mean(att_{pre}(O_k)) \tag{16}$$

where $F(O_k)$ denotes the average focused intensity fo the k-th object.

$$D_v(k,l) = \exp\left(\frac{\mu}{\|(x_k, y_k) - (x_l, y_l)\|^2}\right) \tag{17}$$

where μ is a scaling factor, (x_k, y_k) and (x_l, y_l) are the coordinates of the center point of the k-th, l-th semantic objects.

$$S_\delta(k) = 1 - \exp(-12 \cdot G(k)) \tag{18}$$

where $S_\delta(k)$ represents the relative area competition weight of the k-th semantic object, $G(k)$ is the relative area of the k-th object with respect to its neighbors, see [53] for more details.

As a result, the semantics-based attention model is constructed by melting the attentional competition weight.

$$Sem_{att} = att_{pre} \cdot C_A \tag{19}$$

Figure 10a–c takes the 2nd frame of the "Basketball Drill" sequence as an illustration, showing the attention map of semantic sensitivity, objective area and central bias. Figure 10d displays the inhibition effect of the contextual complexity. Figure 10e shows the attentional competition result. Figure 10f exhibits the final attention map. Intuitively, brighter areas indicate higher attention, inhibition and competition.

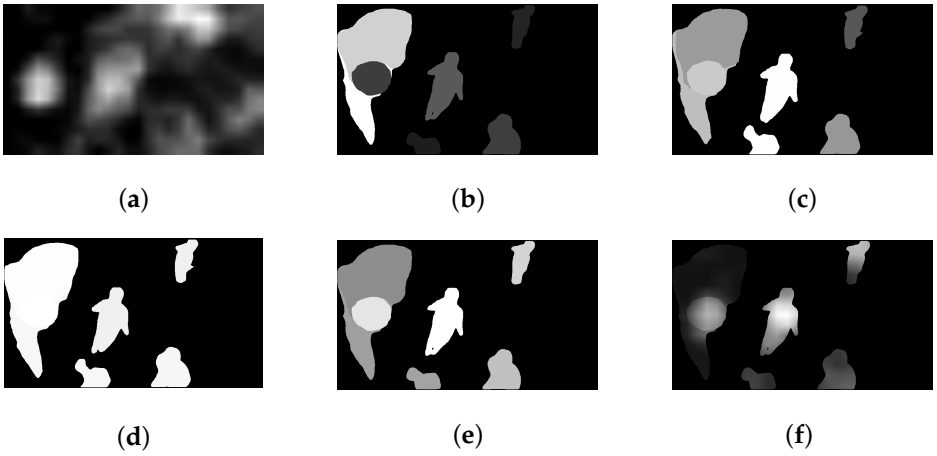


Figure 10. (**a**–**c**) The attention map of the semantic sensitivity, objective area, central bias; (**d**) The inhibition effect of the contextual complexity; (**e**) The attentional competition map; (**f**) The semantic attention map.

5. Semantic-Based Spatio-Temporal Transform Domain JND Profile

As mentioned above, this paper combines semantic sensitivity, objective area, central bias, contextual complexity and attentional competition to measure semantic perceptual attention. Then, based on the model in [65], considering the impact of semantics, this study modifies the spatial attention factor, aiming for a more accurate JND model. The framework of the proposed JND profile is shown in Figure 11.

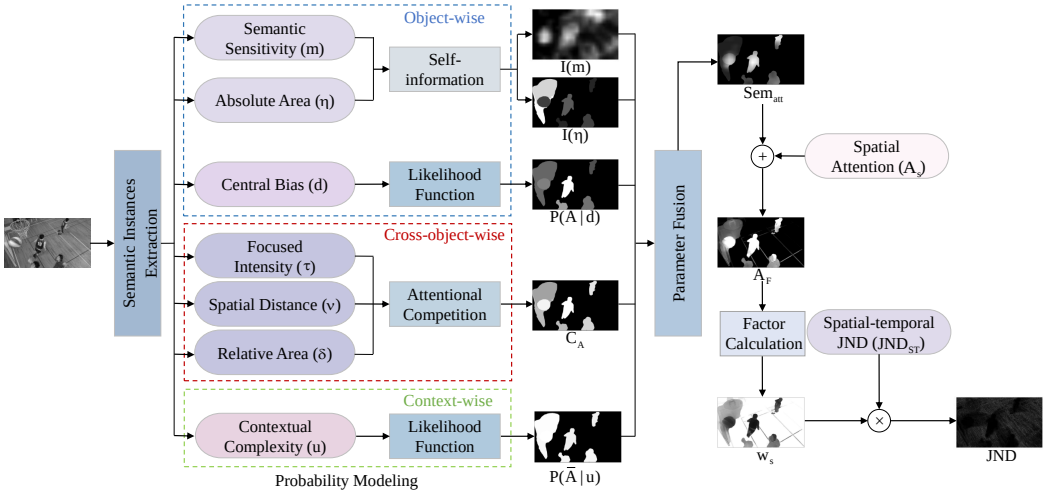


Figure 11. A briefoverview of the proposed spatio-temporal JND model.

By introducing the weighting factor of the semantics-based attention, we propose the following JND model.

$$JND(t,n,i,j) = JND_{st}(t,n,i,j) \cdot w_s(t,n) \tag{20}$$

here, $JND_{st}(t,n,i,j)$ is the spatio-temporal JND threshold of coefficient (i,j) of the n-th block in the t-th frame, considering the LA, CSF and masking effects [23].

Taking the spatio-temporal JND threshold JND_{st} as the basis, this work proposes a patch-wise weighting factor, the attentional weight $w_s(t,n)$, accounting for the visual perceptual attention of semantics, aiming at developing a more accurate JND model. Patch level $w_s(t,n)$ is determined as the mean of pixel-wise adjustment factors in the spatial domain, $w_s(t,i,j)$, of the n-th image block in the t-th frame.

To be specific, by following [65], we estimate spatial attention $A_s(x)$, and combine semantic attention $Sem_{att}(x)$ to get semantics-based spatial attention $A_F(x)$ using NAMM [5].

$$A_F(x) = A_s(x) + Sem_{att}(x) - 0.3 \cdot \min(A_s(x), Sem_{att}(x)) \tag{21}$$

In general, the larger the visual attention $A_F(x)$, the smaller the JND threshold. Intuitively, it seems sensible to apply the sigmoid-like function to normalize $A_F(x)$ to [0,1] and measure the spatial attention weight $w_s(t,n)$ below.

$$w_s(t,n) = \vartheta_1 \cdot \left(1 - \frac{1}{1+\exp(-\vartheta_2 \cdot A_F(t,n))}\right) \tag{22}$$

where ϑ_1, ϑ_2 are normal constants, $A_F(t,n)$ is the mean of the n-th block of $A_F(x)$. According to subjective experiments, ϑ_1 and ϑ_2 are set to 2.5 and 2, respectively.

Figure 12a,b display the obtained attention map and the corresponding weight map. Figure 12c shows the final spatio-temporal JND threshold map. In Figure 12a,c, brighter areas indicate higher visual attention or masking, while the opposite is true for Figure 12b.

(**a**) (**b**) (**c**)

Figure 12. (**a**) The semantic-based spatial attention $A_F(x)$ map; (**b**) The spatial attention weight $w_s(x)$ map; (**c**) The proposed JND threshold map.

6. Experimental Results

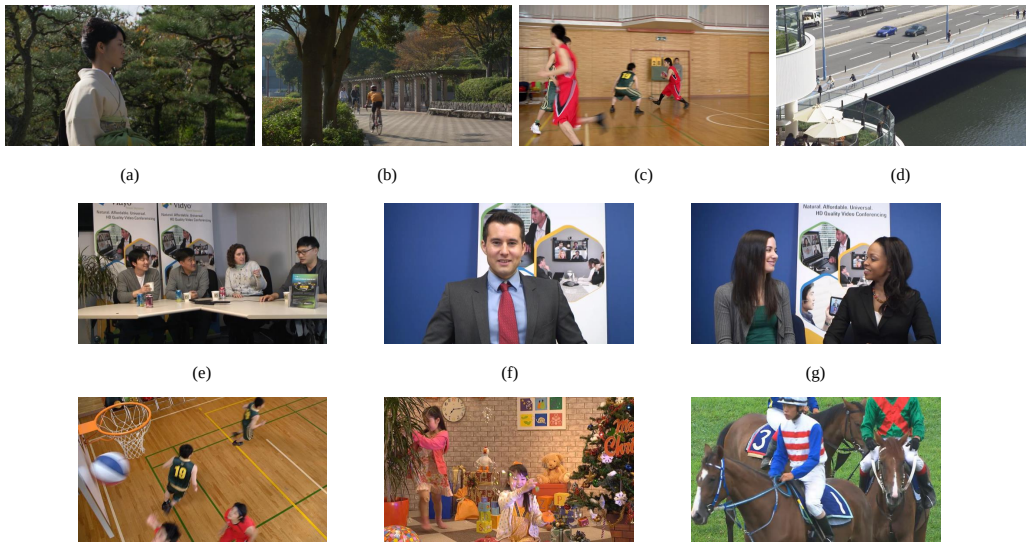
6.1. Comparison of Model Performance

An ideal JND model should, in a sense, distribute noise more fairly, concealing more of it, with acceptable perceptual quality. Thus, we add coefficient-wise JND-guided noise to video sequences, as described in [22], to assess the effectiveness.

$$\hat{R}(t,n,i,j) = R(t,n,i,j) + \rho \cdot rand(t,n,i,j) \cdot JND(t,n,i,j) \tag{23}$$

where $R(t,n,i,j)$ is the transform coefficients of original sequence, $\hat{R}(t,n,i,j)$ is the JND noise contaminated coefficients, ρ regulates the energy of JND noise, and $rand(t,n,i,j) \in \{+1,-1\}$ is a bipolar random noise.

We evaluate the effectiveness of the proposed work in comparison to four benchmark models, namely Bae 2017 [22], Zeng 2019 [66], Wang 2020 [12], Xing 2021 [23], and Li 2022 [67]. Ten videos with diverse semantics were chosen from the HEVC standard sequences in order to meet the requirements of different resolutions. Four of them are videos with 1920 × 1080 full HD resolution, namely, "Kimono1", "ParkScene", "Basketball Drive" and "BQTerrace", as shown in Figure 13a–d. Three of them, "FourPeople", "Johnny", "KristenAndSara", are 1280 × 720 resolution videos, as shown in Figure 13e–g. The remaining three, "Basketball Drill", "PartyScene", and "RaceHorses", are videos with 832 × 480 resolution, as shown in Figure 13h–j.



(a) (b) (c) (d)

(e) (f) (g)

(h) (i) (j)

Figure 13. The HEVC standard test sequences. (**a**) Kimono1; (**b**) ParkScene; (**c**) Basketball Drive; (**d**) BQTerrace; (**e**) FourPeople; (**f**) Johnny; (**g**) KristenAndSara; (**h**) Basketball Drill; (**i**) PartyScene; (**j**) RaceHorses.

Using Peak Signal-to-Noise Ratio (PSNR) as an objective evaluation criterion, a lower PSNR indicates a better ability to mask noise. However, despite that PSNR is the most popular

objective quality evaluation metric, the results of a large number of empirical-psychological studies have demonstrated that PSNR scores fall short of properly capturing HVS perception due to the involvement of visual physiological and psychological mechanisms. Furthermore, the human eye is the receiver of the final visual signal, and therefore, subjective quality metrics must be taken into account when evaluating the proposed JND model.

In this paper, the visual quality of JND-contaminated videos was assessed by recruiting 17 subjects with normal or corrected normal vision, using the subjective viewing test in [68]. Specifically, the monitor used for the display was a 27-in LED monitor, while the view distance was six times the height of the video frame. The difference between the original and processed sequences was observed using the Double Stimulus Continuous Quality Scale (DSCQS) method [23].

Figure 14 illustrates the testing procedure of the DSCQS. In the experiment, for each presentation, the reference and test sequences are arranged in a pseudo-randomized form. During the voting period, participants are expected to rate the quality of each of the two videos. The visual quality of the JND-contaminated videos is then measured by calculating the MOS difference (DMOS) between the original and matched processing sequences. The calculation is described below.

$$\text{DMOS} = MOS_{JND} - MOS_{ORI} \tag{24}$$

where MOS_{ORI} and MOS_{JND} are the measured average opinion score values for the original and test videos, respectively. Five quality scales are used, i.e., excellent (80–100), good (60–80), fair (40–60), poor (20–40), and bad (0–20). The smaller the value of DMOS, the better the visual quality of the JND-polluted video.

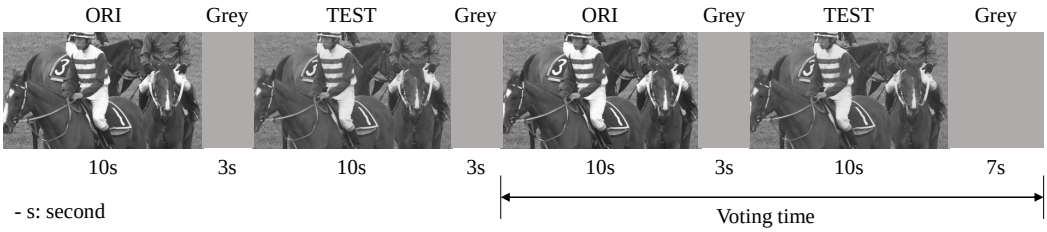


Figure 14. DSCQS method, where the original sequence (ORI) and test sequence (TEST) are pseudo-randomly ordered for each presentation.

The detailed performance results of different JND models are shown in Table 1. From the panoramic viewpoint, it can be found that the proposed JND model has the best performance on PSNR and DMOS for all videos except the "RaceHorses" sequence. This is mainly due to the large proportion of foreground semantic objects in the "RaceHorses" sequence, while our model guides only a small amount of noise into the foreground semantic object region. Therefore, the PSNR value of the proposed JND model is slightly higher than this of Xing, 2021. In addition, as shown in Table 2, the Video Multimethod Assessment Fusion (VMAF) scores [69] of the noise-contaminated videos of different JND models are measured, and the proposed model obtains optimal or suboptimal scores. The larger VMAF scores indicate that the subjective quality of the noise-contaminated sequences is closer to that of the original sequences. The results in Table 2 further verify the superiority of the proposed model in the VMAF metric scenario. Meanwhile, the average PSNR values of Bae 2017, Zeng 2019, Wang 2020, Xing 2021, Li 2022 and the proposed JND profile are 28.36 dB, 30.08 dB, 28.90 dB, 27.83 dB, 30.95 dB, and 27.03 dB, respectively. The average DMOS values of these are 12.49, 23.47, 24.08, 12.51, 19.59 and 9.42, respectively. The average VMAF values of these are 90.09, 87.71, 90.98, 97.30, 88.96 and 97.39, respectively. To visualize the data, Figure 15 displays the bar graphs of the average PSNR, VMAF and accompanying DMOS results for test sequences of various resolutions. Apparently, the noise-contaminated video generated by our model achieves the best perceptual quality (smallest DMOS score) and the largest distortion (lowest PSNR value) compared to the

other four models. These experimental results validate the superiority of our JND profile in guiding noise injection.

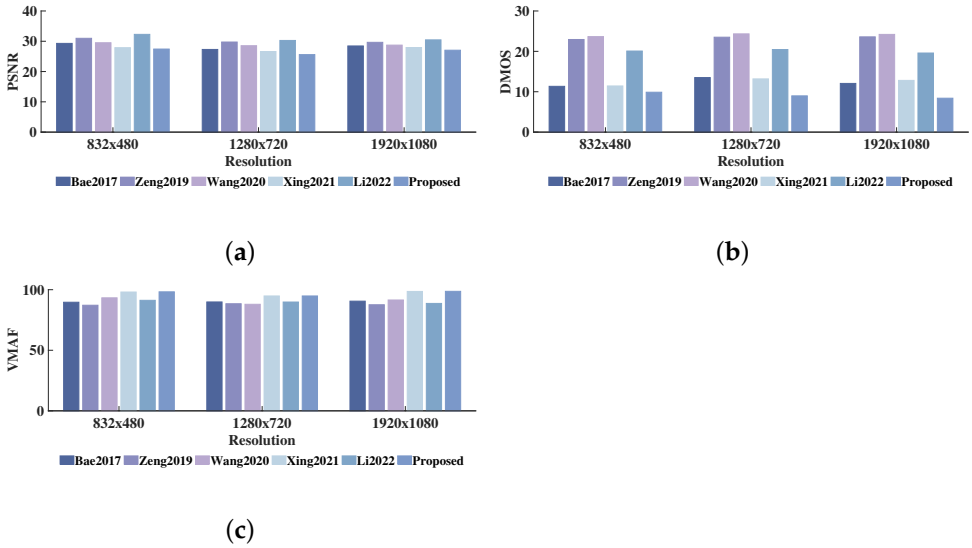


Figure 15. (**a**) Average PSNR values in Table 1 for 832 × 480, 1280 × 720, and 1920 × 1080; (**b**) Average DMOS values in Table 1 for 832 × 480, 1280 × 720, and 1920 × 1080; (**c**) Average VMAF scores in Table 2 for 832 × 480, 1280 × 720, and 1920 × 1080 [12,22,23,66,67].

Table 1. Performance comparison of different JND models.

Sequences	**Bae 2017 [22]**		**Zeng 2019 [66]**		**Wang 2020 [12]**		**Xing 2021 [23]**		**Li 2022 [67]**		**Proposed**	
	PSNR	**DMOS**	**PSNR**	**DMOS**	**PSNR**	**DMOS**	**PSNR**	**DMOS**	**PSNR**	**DMOS**	**PSNR**	**DMOS**
BasketballDrill	27.24	14.24	31.93	22.18	30.05	22.94	26.19	11.59	33.32	19.82	25.22	10.88
PartyScene	29.40	10.18	30.43	23.88	28.91	24.53	27.94	12.24	32.04	19.35	27.10	10.06
RaceHorses	31.27	9.59	30.58	22.76	29.57	23.47	29.50	10.47	31.48	21.06	29.96	8.71
FourPeople	26.86	14.41	29.06	24.41	28.38	24.94	25.90	14.88	29.86	19.29	25.30	11.94
Johnny	27.76	13.65	30.89	21.76	29.20	22.47	28.98	11.71	31.26	17.94	26.91	7.29
KristenAndSara	27.28	12.47	29.30	24.12	28.01	25.59	24.94	12.94	29.69	20.12	24.56	7.71
Kimono1	27.54	13.94	28.17	25.06	27.86	25.64	26.55	14.76	27.93	23.29	25.94	7.41
ParkScene	26.61	13.00	28.25	23.18	27.60	24.00	24.94	13.12	28.62	20.53	24.56	8.18
BasketballDrive	31.10	10.76	32.29	22.47	30.56	23.29	30.52	10.65	33.42	16.24	29.15	9.88
BQTerrace	28.54	10.53	29.89	23.65	28.90	23.88	29.73	12.71	31.88	18.29	28.54	8.12
Average	28.36	12.49	30.08	23.47	28.90	24.08	27.83	12.51	30.95	19.59	27.03	9.42

Table 2. Video multimethod assessment fusion scores.

Sequences	**Bae 2017 [22]**		**Zeng 2019 [66]**		**Wang 2020 [12]**		**Xing 2021 [23]**		**Li 2022 [67]**		**Proposed**	
	PSNR	**VMAF**	**PSNR**	**VMAF**	**PSNR**	**VMAF**	**PSNR**	**VMAF**	**PSNR**	**VMAF**	**PSNR**	**VMAF**
BasketballDrill	27.24	87.98	31.93	87.07	30.05	92.37	26.19	97.99	33.32	88.40	25.22	98.78
PartyScene	29.40	87.64	30.43	84.06	28.91	88.86	27.94	96.36	32.04	85.40	27.10	96.03
RaceHorses	31.27	93.36	30.58	90.40	29.57	98.65	29.50	99.91	31.48	91.61	29.96	99.92
FourPeople	26.86	89.66	29.06	87.48	28.38	87.63	25.90	93.07	29.86	88.54	25.30	93.10
Johnny	27.76	89.92	30.89	89.18	29.20	89.21	28.98	95.80	31.26	90.83	26.91	95.56
KristenAndSara	27.28	90.01	29.30	88.56	28.01	87.07	24.94	95.62	29.69	90.11	24.56	95.84
Kimono1	27.54	90.01	28.17	85.44	27.86	85.71	26.55	99.82	27.93	85.53	25.94	99.71
ParkScene	26.61	86.76	28.25	83.15	27.60	89.11	24.94	95.15	28.62	83.29	24.56	95.66
BasketballDrive	31.10	93.25	32.29	91.15	30.56	95.80	30.52	99.92	33.42	92.52	29.15	99.93
BQTerrace	28.54	92.08	29.89	90.56	28.90	95.40	29.73	99.40	31.88	93.40	28.54	99.35
Average	28.36	90.09	30.08	87.71	28.90	90.98	27.83	97.30	30.95	88.96	27.03	97.39

In order to compare these five JND models more clearly, Figure 16 takes the 4th frame of the "Kimono1" sequence as an example. In general, HVS tends to pay more attention to the semantic targets [65]. In Figure 16a, the woman is prominent relative to the background. Therefore, a lower JND value should be applied in this region to inject less noise. We mainly focus on the head and body to allow for facilitate visual comparison. Apparently, there is considerable visible noise in Figure 16d–g, while the eyes and ears in Figure 16c are blurred. In contrast, Figure 16h is significantly clearer with the lowest PSNR value. As far as the body is concerned, the noise in Figure 16j–m is plainly apparent without exception, while in Figure 16o it is almost undetectable. As a consequence, the above objective and subjective evaluations validate that the proposed semantic attention-based weighting model is effective and excellent.

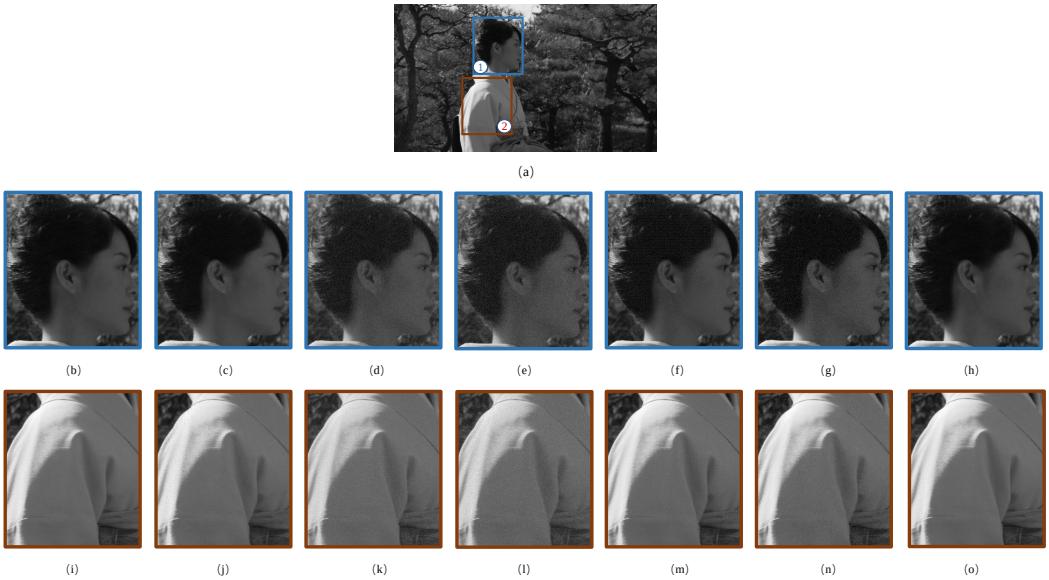


Figure 16. The 4th frame of the "Kimono1" sequence. (**a**) The original image; (**b**–**h**) are the enlarged images of the sub-region ①; (**i**–**o**) are the enlarged images of the sub-region ②. In turn, they are the original patch and the distorted versions processed by Bae 2017, Zeng 2019, Wang 2020, Xing 2021, Li 2022 and the proposed JND model, respectively. The PSNR values of the distorted versions are 27.53 dB, 28.17 dB, 27.76 dB, 27.10 dB, 27.98 dB and 26.69 dB, respectively.

As mentioned in the introduction, high-level semantic objects other than humans often appear in image and video frames. Frame 70 of the "Basketball Drill" sequence is depicted in Figure 17 as an illustration. Clearly, this is a scene depicting basketball training. Naturally, in this case, the basketball and the ball frame are obviously of interest in addition to the basketball players. With respect to the basketball frame, Figure 17c–g show significant distortion, while Figure 17h exhibits strong performance. For the basketball, Figure 17j–l,n depict considerable distortion, Figure 17m shows significant distortion at the lower right of the basketball, while the noise in Figure 17o is barely visible.

6.2. Ablation Experiment and Analysis

To verify the effectiveness of each module in this algorithm, this paper performed ablation experiments from three perception modules: object, context, and cross-object, and evaluated their impacts on the perception results.

The experimental results in Table 3 demonstrate that a combination of the three modules achieves optimal results, while the other combinations do not achieve the desired effect. In other words, when considering the influence of semantics on JND, one of the three aspects of object, context and cross-object features is indispensable. It is worth noting that since the context and cross-object modules are calculated based on object, combinations that do not contain object module are discarded. In practical applications, different kinds of

videos focus on different scales of semantic features, and the combination of three modules achieves better generality.

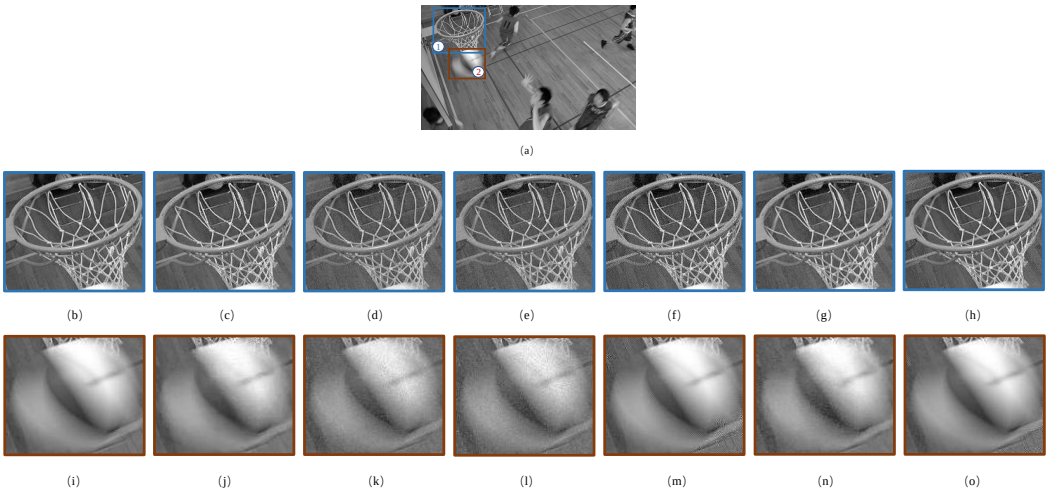


Figure 17. The 70th frame of the "Basketball Drill" sequence. (**a**) The original image; (**b**–**h**) are the enlarged images of the sub-region ①; (**i**–**o**) are the enlarged images of the sub-region ②. In turn, they are the original patch and the distorted versions processed by Bae 2017, Zeng 2019, Wang 2020, Xing 2021, Li 2022 and the proposed JND model, respectively. The PSNR values of the distorted versions are 27.24 dB, 32.12 dB, 30.63 dB, 26.16 dB, 33.42 dB and 25.42 dB, respectively.

Table 3. Performance Comparison of Different Combinations of Modules.

Modules			PSNR		
Object	**Context**	**Cross-Object**	**832 × 480**	**1280 × 720**	**1920 × 1080**
✓			26.67 29.95 31.45	27.85 32.07 29.51	28.41 27.91 31.09 29.80
✓	✓		25.66 28.30 30.32	26.07 29.26 27.48	26.53 25.02 30.19 29.12
✓		✓	25.56 28.22 30.28	26.10 28.78 28.24	26.32 25.67 30.70 28.97
✓	✓	✓	25.22 27.10 29.96	25.30 26.91 24.56	25.94 24.56 29.15 28.54

7. Conclusions

A novel semantic-based JND model was proposed in this paper by thoroughly mining and characterizing the semantic feature parameters that affect attention. The interaction between semantic visual features and HVS was investigated from object, contextual and cross-object perspectives. This analysis included HVS responses induced by semantic sensitivity, objective area and central bias, as well as perceptual suppression brought about by contextual complexity and cross-object competition for attention. In conjunction with underlying attention in the spatial domain, perceptual attention to stimuli was measured with information-theoretic support and incorporated into a patch-level weighting factor. Using the semantic-based attentional weight, the JND model for the spatiotemporal trans-

form domain was modified. The experimental results demonstrate the effectiveness of the proposed JND model with superior performance and stronger distortion-hiding ability compared to the state-of-the-art JND model.

Existing JND models only consider the effects of unimodal signals, while the study of cross-modal JND remains an open problem. However, there exists great incentive to transfer this issue from the laboratory to real-world adoption. In the future, we will further investigate multimodal asynchronous perception, seek a more unified approach to fusing multi-modal feature parameters, and obtain more accurate JND thresholds.

Author Contributions: Conceptualization, X.W., H.Y.; data curation, X.W.; funding acquisition, H.Y.; methodology, X.W., Y.L., Y.C.; project administration, H.Y.; software, X.W., S.Z.; Supervision, H.Y.; writing—original draft, X.W.; writing—review and editing, H.Y., Y.L., S.Z., Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province(2022C01068), NSFC 61972123, and NSFC 62031009.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Macknik, S.L.; Livingstone, M.S. Neuronal Correlates of Visibility and Invisibility in the Primate Visual System. *Nat. Neurosci.* **1998**, *1*, 144–149. [CrossRef] [PubMed]
2. Carlson, N.R. *Psychology: The Science of Behavior*; McGraw-Hill: New York, NY, USA , 1987.
3. Wu, J.; Shi, G.; Lin, W. Survey of Visual Just Noticeable Difference Estimation. *Front. Comput. Sci.* **2019**, *13*, 4–15. [CrossRef]
4. Chou, C.H.; Li, Y.C. A Perceptually Tuned Subband Image Coder Based on the Measure of Just-noticeable-distortion Profile. *IEEE Trans. Circuits Syst. Video Technol.* **1995**, *5*, 467–476. [CrossRef]
5. Yang, X.; Lin, W.; Lu, Z.; Ong, E.P.; Yao, S. Just-noticeable-distortion Profile with Nonlinear Additivity Model for Perceptual Masking in Color Images. *IEEE Int. Conf. Acoust. Speech Signal Process.* **2003**, *3*, III-609.
6. Liu, A.; Lin, W.; Paul, M.; Deng, C.; Zhang, F. Just Noticeable Difference for Images with Decomposition Model for Separating Edge and Textured Regions. *IEEE Trans. Circuits Syst. Video Technol.* **2010**, *20*, 1648–1652. [CrossRef]
7. Chen, Z.; Guillemot, C. Perceptually-Friendly H.264/AVC Video Coding Based on Foveated Just-Noticeable-Distortion Model. *IEEE Trans. Circuits Syst. Video Technol.* **2010**, *20*, 806–819. [CrossRef]
8. Wu, J.; Shi, G.; Lin, W.; Liu, A.; Qi, F. Just Noticeable Difference Estimation for Images with Free-energy Principle. *IEEE Trans. Multimed.* **2013**, *15*, 1705–1710. [CrossRef]
9. Wu, J.; Li, L.; Dong, W.; Shi, G.; Lin, W.; Kuo, C.C.J. Enhanced Just Noticeable Difference Model for Images with Pattern Complexity. *IEEE Trans. Image Process.* **2017**, *26*, 2682–2693. [CrossRef]
10. Jakhetiya, V.; Lin, W.; Jaiswal, S.; Gu, K.; Guntuku, S.C. Just Noticeable Difference for Natural Images Using RMS Contrast and Feed-back Mechanism. *Neurocomputing* **2018**, *275*, 366–376. [CrossRef]
11. Chen, Z.; Wu, W. Asymmetric foveated just-noticeable-difference model for images with visual field inhomogeneities. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4064–4074. [CrossRef]
12. Wang, H.; Yu, L.; Liang, J.; Yin, H.; Li, T.; Wang, S. Hierarchical Predictive Coding-based JND Estimation for Image Compression. *IEEE Trans. Image Process.* **2020**, *30*, 487–500. [CrossRef] [PubMed]
13. Cui, X.; Peng, Z.; Chen, F.; Jiang, G.; Yu, M. Perceptual ultra-high definition video coding based on adaptive just noticeable distortion model. *Displays* **2022**, *75*, 102301. [CrossRef]
14. Ahumada, A.J., Jr.; Peterson, H.A. Luminance-model-based DCT Quantization for Color Image Compression. *Hum. Vis. Vis. Process. Digit. Disp. III* **1992**, *1666*, 365–374.
15. Watson, A.B. DCTune: A Technique for Visual Optimization of DCT Quantization Matrices for Individual Images. *Soc. Inf. Disp. Dig. Tech. Pap.* **1993**, *24*, 946–946.
16. Zhang, X.; Lin, W.; Xue, P. Improved Estimation for Just-noticeable Visual Distortion. *Signal Process.* **2005**, *85*, 795–808. [CrossRef]
17. Jia, Y.; Lin, W.; Kassim, A.A. Estimating Just-noticeable Distortion for Video. *IEEE Trans. Circuits Syst. Video Technol.* **2006**, *16*, 820–829. [CrossRef]
18. Wei, Z.; Ngan, K.N. Spatio-Temporal Just Noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain. *IEEE Trans. Circuits Syst. Video Technol.* **2009**, *19*, 337–346.
19. Wang, H.; Wang, L.; Hu, X.; Tu, Q.; Men, A. Perceptual video coding based on saliency and just noticeable distortion for H. 265/HEVC. In Proceedings of the International Symposium on Wireless Personal Multimedia Communications, Sydney, NSW, Australia, 7–10 September 2014 ; pp. 106–111.
20. Bae, S.H.; Kim, J.; Kim, M. HEVC-based perceptually adaptive video coding using a DCT-based local distortion detection probability model. *IEEE Trans. Image Process.* **2016**, *25*, 3343–3357. [CrossRef]

21. Wan, W.; Wu, J.; Xie, X.; Shi, G. A Novel Just Noticeable Difference Model via Orientation Regularity in DCT Domain. *IEEE Access* **2017**, *5*, 22953–22964. [CrossRef]
22. Bae, S.H.; Kim, M. A DCT-Based Total JND Profile for Spatiotemporal and Foveated Masking Effects. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1196–1207. [CrossRef]
23. Xing, Y.; Yin, H.; Zhou, Y.; Chen, Y.; Yan, C. Spatiotemporal Just Noticeable Difference Modeling with Heterogeneous Temporal Visual Features. *Displays* **2021**, *70*, 102096. [CrossRef]
24. Jiang, Q.; Liu, Z.; Wang, S.; Shao, F.; Lin, W. Towards Top-Down Just Noticeable Difference Estimation of Natural Images. *IEEE Trans. Image Process.* **2022**, *31*, 3697–3712. [CrossRef] [PubMed]
25. Ki, S.; Bae, S.H.; Kim, M.; Ko, H. Learning-based Just-noticeable-quantization-distortion Modeling for Perceptual Video Coding. *IEEE Trans. Image Process.* **2018**, *27*, 3178–3193. [CrossRef] [PubMed]
26. Hadizadeh, H.; Heravi, A.R.; Bajić, I.V.; Karami, P. A Perceptual Distinguishability Predictor for JND-noise-contaminated Images. *IEEE Trans. Image Process.* **2018**, *28*, 2242–2256. [CrossRef] [PubMed]
27. Zhang, X.; Yang, C.; Wang, H.; Xu, W.; Kuo, C.C.J. Satisfied-user-ratio Modeling for Compressed Video. *IEEE Trans. Image Process.* **2020**, *29*, 3777–3789. [CrossRef] [PubMed]
28. Liu, H.; Zhang, Y.; Zhang, H.; Fan, C.; Kwong, S.; Kuo, C.C.J.; Fan, X. Deep Learning-based Picture-wise Just Noticeable Distortion Prediction Model for Image Compression. *IEEE Trans. Image Process.* **2019**, *29*, 641–656. [CrossRef]
29. Lin, H.; Hosu, V.; Fan, C.; Zhang, Y.; Mu, Y.; Hamzaoui, R.; Saupe, D. SUR-FeatNet: Predicting the Satisfied User Ratio Curve for Image Compression with Deep Feature Learning. *Qual. User Exp.* **2020**, *5*, 1–23. [CrossRef]
30. Wu, Y.; Wang, Z.; Chen, W.; Lin, L.; Wei, H.; Zhao, T. Perceptual VVC quantization refinement with ensemble learning. *Displays* **2021**, *70*, 102103. [CrossRef]
31. Wu, Y.; Ji, W.; Wu, J. Unsupervised Deep Learning for Just Noticeable Difference Estimation. In Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–6.
32. Jin, J.; Yu, D.; Lin, W.; Meng, L.; Wang, H.; Zhang, H. Full RGB Just Noticeable Difference (JND) Modelling. *arXiv* **2022**, arXiv:2203.00629.
33. Jin, J.; Xue, Y.; Zhang, X.; Meng, L.; Zhao, Y.; Lin, W. HVS-Inspired Signal Degradation Network for Just Noticeable Difference Estimation. *arXiv* **2022**, arXiv:2208.07583.
34. Jin, L.; Lin, J.Y.; Hu, S.; Wang, H.; Wang, P.; Katsavounidis, I.; Aaron, A.; Kuo, C.C.J. Statistical Study on Perceived JPEG Image Quality via MCL-JCI Dataset Construction and Analysis. *Electron. Imaging* **2016**, *2016*, 1–9. [CrossRef]
35. Liu, X.; Chen, Z.; Wang, X.; Jiang, J.; Kowng, S. JND-Pano: Database for Just Noticeable Difference of JPEG Compressed Panoramic Images. In *Pacific Rim Conference on Multimedia , Proceedings of the 19th Pacific-Rim Conference on Multimedia, Hefei, China, 21–22 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 458–468.
36. Wang, H.; Gan, W.; Hu, S.; Lin, J.Y.; Jin, L.; Song, L.; Wang, P.; Katsavounidis, I.; Aaron, A.; Kuo, C.C.J. MCL-JCV: A JND-based H.264/AVC Video Quality Assessment Dataset. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 1509–1513.
37. Wang, H.; Katsavounidis, I.; Zhou, J.; Park, J.; Lei, S.; Zhou, X.; Pun, M.O.; Jin, X.; Wang, R.; Wang, X.; et al. VideoSet: A Large-scale Compressed Video Quality Dataset Based on JND Measurement. *J. Visual Commun. Image Represent.* **2017**, *46*, 292–302. [CrossRef]
38. Jin, J.; Zhang, X.; Fu, X.; Zhang, H.; Lin, W.; Lou, J.; Zhao, Y. Just Noticeable Difference for Deep Machine Vision. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3452–3461. [CrossRef]
39. Lin, W.; Ghinea, G. Progress and Opportunities in Modelling Just-Noticeable Difference (JND) for Multimedia. *IEEE Trans. Multimed.* **2021**, *24*, 3706–3721. [CrossRef]
40. Liu, H.; Shi, S.; Bai, R.; Liu, Y.; Lian, X.; Shi, T. A brain-inspired computational model for extremely few reference image quality assessment. *Displays* **2023**, *76*, 102331. [CrossRef]
41. Wang, H.; Yu, L.; Wang, S.; Xia, G.; Yin, H. A Novel Foveated-JND Profile Based on an Adaptive Foveated Weighting Model. In Proceedings of the 2018 IEEE Visual Communications and Image Processing (VCIP), Taichung, Taiwan, 9–12 December 2018; pp. 1–4.
42. Itti, L.; Koch, C. A Saliency-based Search Mechanism for Overt and Covert Shifts of Visual Attention. *Vision Res.* **2000**, *40*, 1489–1506. [CrossRef] [PubMed]
43. Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to Predict Where Humans Look. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 2106–2113.
44. Borji, A.; Sihite, D.N.; Itti, L. What Stands out in a Scene? A Study of Human Explicit Saliency Judgment. *Vision Res.* **2013**, *91*, 62–77. [CrossRef] [PubMed]
45. Wang, C.; Wang, C.; Li, W.; Wang, H. A brief survey on RGB-D semantic segmentation using deep learning. *Displays* **2021**, *70*, 102080. [CrossRef]
46. Wang, R.; Zhang, Z. Energy Coding in Biological Neural Networks. *Cognit. Neurodyn.* **2007**, *1*, 203–212. [CrossRef]
47. Feldman, H.; Friston, K.J. Attention, Uncertainty, and Free-energy. *Front. Hum. Neurosci.* **2010**, *4*, 1–23. [CrossRef]
48. Selvaraju, R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 618–626.

49. Wu, J.; Jiang, Y.; Zhang, W.; Bai, X.; Bai, S. SeqFormer: A Frustratingly Simple Model for Video Instance Segmentation. *arXiv* **2021**, arXiv:2112.08275.
50. Long, Y.; Jin, D.; Wu, Z.; Zuo, Z.; Wang, Y.; Kang, Z. Accurate Identification of Infrared Ship in Island-Shore Background Based on Visual Attention. In Proceedings of the 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2022; pp. 800–806.
51. Bruce, N.D.; Tsotsos, J.K. Saliency, Attention, and Visual Search: An Information Theoretic Approach. *J. Vis.* **2009**, *9*, 1–24. [CrossRef] [PubMed]
52. Shen, Z.; Zhang, L.; Li, R.; Hou, J.; Liu, C.; Hu, W. The effects of color combinations, luminance contrast, and area ratio on icon visual search performance. *Displays* **2021**, *67*, 101999. [CrossRef]
53. Syeda-Mahmood, T.F . *Attentional Selection in Object Recognition* ; Technical Report 1420; MIT Computer Science & Artificial Intelligence Laboratory: Cambridge, MA, USA, 1993.
54. Wu, C.C.; Wick, F.A.; Pomplun, M. Guidance of Visual Attention by Semantic Information in Real-world Scenes. *Front. Psychol.* **2014**, *5*, 54. [CrossRef]
55. Davenport, J.L.; Potter, M.C. Scene Consistency in Object and Background Perception. *Psychological science* **2004**, *15*, 559–564. [CrossRef]
56. Beck, M.R.; Lohrenz, M.C.; Trafton, J.G. Measuring Search Efficiency in Complex Visual Search Tasks: Global and Local Clutter. *J. Exp. Psychol. Appl.* **2010**, *16*, 238. [CrossRef] [PubMed]
57. Caroux, L.; Mouginé, A. Influence of Visual Background Complexity and Task Difficulty on Action Video Game Players' Performance. *Entertain. Comput.* **2022**, *41*, 100471. [CrossRef]
58. Dimond, S.J.; Farrington, L.; Johnson, P. Differing Emotional Response from Right and Left Hemispheres. *Nature* **1976**, *261*, 690–692. [CrossRef]
59. Duncan, J. EPS Mid-Career Award 2004: Brain Mechanisms of Attention. *Quarterly Journal of Experimental Psychology* **2006**, *59*, 2–27. [CrossRef]
60. Scalf, P.E.; Basak, C.; Beck, D. Attention Does More than Modulate Suppressive Interactions: Attending to Multiple Items. *Exp. Brain Res.* **2011**, *212*, 293–304. [CrossRef]
61. Scalf, P.E.; Torralbo, A.; Tapia, E.; Beck, D.M. Competition Explains Limited Attention and Perceptual Resources: Implications for Perceptual Load and Dilution Theories. *Front. Psychol.* **2013**, *4*, 243. [CrossRef] [PubMed]
62. Schall, J.D.; Sato, T.R.; Thompson, K.G.; Vaughn, A.A.; Juan, C.H. Effects of Search Efficiency on Surround Suppression During Visual Selection in Frontal Eye Field. *J. Neurophysiol.* **2004**, *91*, 2765–2769. [CrossRef] [PubMed]
63. Müller, N.G.; Kleinschmidt, A. The Attentional 'Spotlight's' Penumbra: Center-surround Modulation in Striate Cortex. *Neuroreport* **2004**, *15*, 977–980. [CrossRef]
64. Mounts, J.R. Evidence for Suppressive Mechanisms in Attentional Selection: Feature Singletons Produce Inhibitory Surrounds. *Percept. Psychophys.* **2000**, *62*, 969–983. [CrossRef]
65. Yan, Y.; Ren, J.; Sun, G.; Zhao, H.; Han, J.; Li, X.; Marshall, S.; Zhan, J. Unsupervised Image Saliency Detection with Gestalt-laws Guided Optimization and Visual Attention Based Refinement. *Pattern Recognit.* **2018**, *79*, 65–78. [CrossRef]
66. Zeng, Z.; Zeng, H.; Chen, J.; Zhu, J.; Zhang, Y.; Ma, K.K. Visual Attention Guided Pixel-Wise Just Noticeable Difference Model. *IEEE Access* **2019**, *7*, 132111–132119. [CrossRef]
67. Li, J.; Yu, L.; Wang, H. Perceptual redundancy model for compression of screen content videos. *IET Image Process.* **2022**, *16*, 1724–1741. [CrossRef]
68. Yang, X.; Lin, W.; Lu, Z.; Ong, E.P.; Yao, S.S. Just Noticeable Distortion Model and Its Applications in Video Coding. *Signal Process. Image Commun.* **2005**, *20*, 662–680. [CrossRef]
69. Li, Z.; Aaron, A.; Katsavounidis, I.; Moorthy, A.; Manohara, M. Toward a practical perceptual video quality metric. *Netflix Tech Blog* **2016**, *6*, 2.