Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

QuatHashing: Perceptually Hashing with Quaternion Discrete Wavelet Transform for Reduced-Reference Visual Security Assessment

Yannan Ren
Shandong Jiaotong University
Jiantao Wang
Shandong Normal University
Hua Wu
wuhua@sdjtu.edu.cn

Shandong Jiaotong University

Research Article

Keywords: Perceptually Hashing, Quaternion Discrete Wavelet Transform, Perceptually Encryption, Visual Security Assessment, Human Visual System

Posted Date: July 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10287744/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal on Information Security on November 6th, 2026. See the published version at https://doi.org/10.1186/s13635-026-00248-4.

QuatHashing: Perceptually Hashing with Quaternion Discrete Wavelet Transform for Reduced-Reference Visual Security Assessment

Yannan Ren[1], Jiantao Wang[2], Hua Wu[1*]

[1]School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan, 250357, Shandong, China.
[2]School of Communication and Electronic Engineering, Shandong Normal University, Jinan, 250358, Shandong, China.

*Corresponding author(s). E-mail(s): wuhua@sdjtu.edu.cn;

Abstract

Visual security assessment (VSA) is an important task of perceptual image encryption and plays a key role in evaluating the effectiveness of encryption schemes. This paper proposes a perceptual hashing algorithm with quaternion discrete wavelet transform (QDWT) for reduced-reference (RR) VSA. The directional complementary sub-band representation is exploited to reorganize the image into horizontal and vertical complementary sub-bands. Since the pixel rearrangement does not discard any information, the distortion introduced by encryption is preserved in the complementary sub-bands while the directional image information is explicitly represented. The complementary sub-bands are modeled as two pure quaternion images and decomposed by QDWT to extract perceptually relevant features from the horizontal and vertical channels. The low-frequency structural features, texture features and multi-scale structural features are further obtained to describe the residual visual information from both directional and multi-scale perspectives. As the human eye is more sensitive to visually salient regions, global and local attention mechanisms are introduced to emphasize these regions and generate compact perceptual feature sequences. The perceptual hash is finally determined by quantizing and combining the horizontal and vertical feature sequences, and the Euclidean distance between the hashes of the reference and encrypted images is used as the final visual security score. Experimental results demonstrate that the proposed QuatHashing achieves more consistent performance than existing IQA and VSA methods, particularly for encrypted images with low and moderate visual quality.

Keywords: Perceptually Hashing, Quaternion Discrete Wavelet Transform, Perceptually Encryption, Visual Security Assessment, Human Visual System

1 Introduction

With the development of the Internet, image is widely used in the Internet due to its advantages of being the most concise, fastest, easiest to recognize, most expressive and transcending language barriers between countries. With that comes a variety of visual information security concerns,researchers have also started to pay attention to the protection of image information[1–4]. Therefore, perceptually image encryption technology, as a technology to protect the security of visual information of images, has received a lot of attention[5–7]. Initially, researchers developed image encryption technologies to protect image information generally in a fully encrypted manner[8–11]. Since the image contains a lot of redundant information, in many cases, if the redundant information is not encrypted, the visual information security of the image will not be affected. This shows that the traditional full encryption method is not completely suitable for the encryption of image data[12, 13]. The perceptually image encryption technology does not encrypt the image globally like traditional encryption methods, but encrypts part of the image data according to the needs, which reduces the amount of encrypted information and improves the encryption efficiency. Also, perceptually image encryption technologies can simultaneously maintain and prevent misuse of the essential visual properties of the original data. These advantages make perceptually image encryption technologies have high application value in image security protection.

However, how to measure the security of perceptually image encryption technology is a problem that needs to be improved at present. Evaluating the visual security of encrypted images can measure the effectiveness of perceptually image encryption technologies[14, 15]. Perceptually encryption technology can be continuously optimized according to the evaluation results of visual security. Traditional image encryption security analysis methods, such as differential analysis[16], statistical analysis[17], adjacent pixel correlation analysis[18], analyzes the entire image and focus on evaluating whether the encrypted image achieves semantic security. But perceptually image encryption is to partially encrypt the image, and the concept of semantic security does not apply here. Therefore, traditional methods for image encryption security analysis are not applicable to visual security assessment (VSA) of perceptually image encryption[5]. Existing solutions usually use image quality assessment (IQA) indexes to assess the visual security of encrypted images. For example, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM)[19], and Visual Fidelity (VIF)[20].

The purpose of IQA is to assess the quality of an image. For a perceptually encrypted image, the low quality of the perceptually encrypted image does not exactly correspond to its higher security. Since humans are the ultimate recipients of images, people's subjective assessment are the most authoritative in assessing the security of encrypted image[5]. However, subjective assessment is a time-consuming and laborious job. Therefore, people urgently need a more effective objective assessment indicators to approximate subjective assessment and achieve the effect of subjective assessment. In response to the above problems, researchers have designed a series of visual security assessment methods specifically for perceptually image encryption. For example, LSS[21], ESS[21], LFBVS[22], VSI-Canny[23] and VSASW[24]. In general, the existing VSAs can be divided into the following three types. (1) Full-reference(FR)VSA: the input is a reference image and an encrypted image. (2) Reduced-reference(RR)VSA:

the input is representative information of the reference image and its encrypted image. (3)No-reference(NR)VSA: all that is entered is the encrypted image. RR-VSA evaluates the security of the image by extracting specific information from the image, which is less informative and more practical. This paper investigates a novel perceptually hashing algorithm based on quaternion discrete wavelet transform for RR VSA.

Image hashing is a useful image processing technique[25–27]. It effectively extracts a compact hash sequence based on visual content from the input image, which can be used to represent the input image itself. Due to the advantages of low storage capacity and fast similarity calculation, image hashing is widely used in image copy detection, image retrieval, image tampering detection, and image content authentication[28, 29]. Robustness and discrimination are the two main characteristics of image hashing. Specifically, robustness requires that the hash sequence of the original image be similar to that of its deformed hash sequence[30]. Differentiation requires completely different hash sequences for different images. These two attributes are interrelated and mutually conditioning. An effective image hashing algorithm should strike an ideal balance between the two. In addition, for the application of RR VSA, the perceptual image hash must satisfy the distortion sensitivity, that is, the hashing algorithm needs to perceive the degree of distortion on similar images. In addition, the human visual system is the ultimate receiver of image information. The human visual system is sensitive to the orientation information of images, and has the characteristics of multi-scale observation in observation. Based on the above considerations, a novel perceptual hashing algorithm is developed to evaluate the visual security of perceptually encrypted images. Our main contributions are as follows:

- The directional complementary sub-band representation is exploited to reorganize the input image into horizontal and vertical complementary sub-bands for extracting perceptual features. Since the pixel rearrangement does not discard any information, the distortion introduced by perceptual encryption is preserved in the complementary sub-bands. As the horizontal and vertical sub-bands make the directional information of the image explicit, the features extracted from these sub-bands can better reflect the visual differences between the reference image and the encrypted image.
- A double-channel QDWT-based feature extraction method is developed to capture perceptually relevant image information. Since the horizontal and vertical complementary sub-bands are modeled as two pure quaternion images and decomposed in the QDWT domain, the low-frequency structure, texture and multi-scale structural features can be extracted from the decomposed coefficients to describe the remaining visual content of encrypted images. As this design combines directional analysis with multi-scale representation, it is consistent with the orientation sensitivity and multi-scale perception characteristics of the human visual system.
- An attention-guided reduced-reference perceptual hashing assessment scheme is constructed for visual security assessment. Global attention and local attention are introduced to emphasize visually sensitive regions, such as salient structures, edge contours, and texture variations. The attention-guided horizontal and vertical features are then converted into compact hash sequences, and the Euclidean distance between the hash sequences of the reference and encrypted images is used as the

final visual security score. Experimental results on public encrypted image databases show that the proposed QuatHashing method provides stable and competitive performance compared with existing IQA and VSA methods.

The rest of this article is organized as follows. Section 2 reviews related works. Section 3 describes the details of QuatHashing. Section 4 presents hash construction and similarity assessment. Section 5 shows the experimental evaluation of QuatHashing Section 6 concludes the paper.

2 Related Work

Many methods have been proposed for the VSA of perceptually image encryption, and these methods are roughly divided into two categories: the first category is the image quality assessment (IQA)[31], and the second category is the visual security assessment (VSA)[32] specially developed for the security assessment of perceptually encrypted images.Over the past few decades, a large number of objective assessments have been proposed. Such as peak signal-to-noise ratio (PSNR), structural similarity (SSIM)[19] and visual fidelity (VIF)[20]. Some researchers directly use the well-known IQAs to evaluate the degree of visual security of perceptually encrypted images[16].These IQAs' goal to assess the quality of images, but a low-quality image doesn't necessarily mean it has strong security, so when IQAs are used for the visual security of low-quality perceptually encrypted images, the performance is not satisfactory.

Therefore, the VSA specially developed for the security assessment of perceptually encrypted images came into being. Mao and Wu[21] proposed ESS and LSS to calculate edge similarity and brightness similarity between original and encrypted images. ESS first divides the image into blocks and calculates the gradient direction of the image block, then calculates the direction difference between the original image and the encrypted image corresponding to the image block to represent the edge similarity, and finally integrates the edge similarity of all image blocks to obtain the edge similarity of the image. LSS first divides the image into blocks and calculates the average brightness of the image blocks, calculates the brightness similarity of the corresponding image blocks of the original image and the encrypted image, and finally averages the brightness similarity of all image blocks to obtain the brightness similarity.

Tong et al.[22] proposed a local feature-based LFBVS to assess the visual security of images or videos. For a given original image and encrypted image, first divides it into 16×16 blocks, extracts the color features and edge features of each block, calculates the color similarity and edge similarity of each block, and uses a linear weighting function The local similarity of the image is obtained by integrating it; then the similarity of each block is weighted to obtain the final image similarity.

T. Xiang et al.[23] proposed the VSI-Canny by computing the edge and texture similarity between the original image and the perceptually encrypted image. First, the edge features of the original image and the encrypted image are extracted by the Canny operator with different thresholds, and the edge similarity is calculated, and then the edge similarity of the different thresholds is linearly weighted to obtain the edge similarity of the original image and the encrypted image. The co-occurrence matrix extracts the texture features of the image and calculates the texture similarity.

Then, the adaptive linear method is used to weight the edge similarity and texture similarity to obtain the final visual security index.

The performance of these existing VSAs for evaluating the visual security of encrypted images needs to be further improved. Among them, Figure 1(d–f) illustrates the performance of different VSAs on images from the IVC-SelectEncrypt[33] database. Figure 1(d) is the original image, and Figure 1(e-f) are the two encrypted images of Figure 1(d). Clearly, Figure 1(f) is more security than Figure 1(e). However, the LSS values and VSI-Canny values of Figure 1(f) are higher than those of Figure 1(e). As mentioned above, existing VSAs still have some shortcomings: they don't fully consider the visual characteristics of encrypted images, and they don't take into account the role of hashing can play in VSAs. Addressing the above will improve the accuracy of visual security assessment.



Fig. 1 Performance analysis of IQA method on IVC-SelectEncrypt database and PEID database: (**a**)original image from PEID database. (**b**)encrypted image,PSNR = 17.73,SSIM = 0.53. (**c**)encrypted image,PSNR = 23.34,SSIM = 0.65.(**d**)original image from IVC-SelectEncrypt database.(**e**)encrypted image with relatively weak visual protection.(**f**)encrypted image with stronger visual protection.

3 Proposed QuatHashing Method

The framework of our proposed hashing method is shown in Figure 2. Our algorithm consists of the following four steps: (1) A preprocessing operation on the input image, in which bicubic interpolation is used to convert the size of the input image to $T \times T$. (2) Pixelize the preprocessed image, and three horizontal complementary sub-bands and three vertical complementary sub-bands are generated for feature extraction. (3) QDWT is used to extract image orientation features from directional complementary sub-bands. (4) The quantized horizontal and vertical features are concatenated

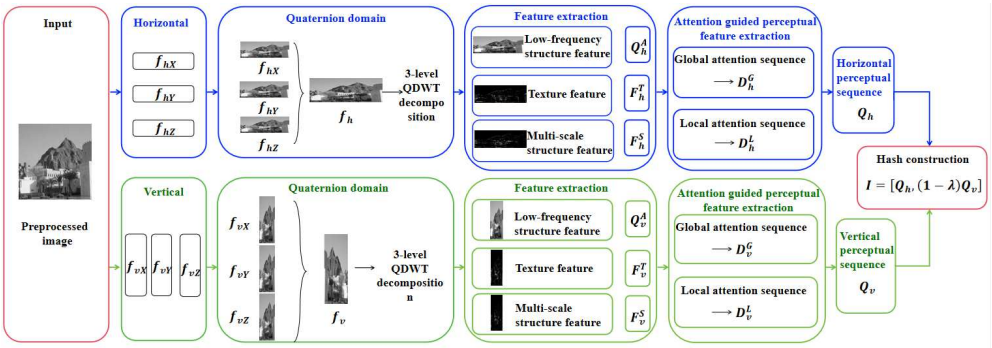


Fig. 2 The framework of the proposed QuatHashing method.

Table 1 Main symbols used in this paper.

Symbols	Meaning	Symbols	Meaning
q	a quaternion	k_1, k_2, k_3	three imaginary units of a quaternion
f_{hX}, f_{hY}, f_{hZ}	horizontal complementary sub-bands	f_{vX}, f_{vY}, f_{vZ}	vertical complementary sub-bands
f_h, f_v	pure quaternion images of the horizontal and vertical channels	t	directional channel index, $t \in \{h, v\}$
σ	QDWT decomposition level, $\sigma \in \{1, 2, 3\}$	Q_t^A	low-frequency structure feature of channel t
$Q_{t,\sigma}^H, Q_{t,\sigma}^V, Q_{t,\sigma}^D$	directional high-frequency sub-bands at level σ	Q_{XY}, Q_{XZ}, Q_{YZ}	complementary components
$E_{H,t}, E_{V,t}, E_{D,t}$	cross-scale texture energies	F_t^T	texture feature of channel t
$F_{\sigma,t}^{(Q_{XY}, Q_{XZ}, Q_{YZ})}$	structural feature at scale σ	F_t^S	multi-scale structural feature of channel t
$\vec{F}$	joint feature vector	G	global saliency response
$\mathcal{B}_i$	the i-th full-height strip	b, F	strip width and number of strips, $F = W/b$
$D^{*,G}, D^{*,L}$	global and local feature sequences	$\hat{D}^{*,G}, \hat{D}^{*,L}$	relativized feature sequences
Q_h, Q_v	horizontal and vertical feature sequences	λ	weight balancing the two directional channels
l	initial hash sequence	$\mathbf{L}$	final integer hash sequence

together to construct a hash sequence. The details of our hashing algorithm will be described in the following subsections. For reference, the relevant symbols used in this paper are summarized in Table 1.

3.1 Directional Complementary Sub-bands

The breakdown of horizontal and vertical orientations captures all the major directional features in the image. Horizontal sub-bands can effectively extract horizontal features such as horizon lines and building edges, while vertical sub-bands are more

suitable for capturing vertical elements such as building facades and trees. This omnidirectional feature extraction provides a comprehensive perspective for subsequent image analysis, which can obtain richer and more diverse data about the image structure. In practice, three horizontal complementary sub-bands and three vertical complementary sub-bands are obtained in pixels, and the process is shown in Figure 2.

The three adjacent pixels are divided into a group, and the 1st, 2nd, and 3rd pixels of each group are recombined to obtain the horizontal complementary sub-bands f_{hX}, f_{hY} and f_{hZ}, respectively. Similarly, the three adjacent pixels in the vertical direction are grouped and recombined to obtain the vertical complementary sub-bands f_{vX}, f_{vY} and f_{vZ}. Here, X, Y and Z are used to index the three complementary sub-bands, so that they can be clearly distinguished from the three QDWT decomposition levels.

3.2 Double Channel QDWT

A quaternion[34] is a linear combination of a real scalar and three imaginary units, and its significance is as follows:

$$q = a + bk_1 + ck_2 + dk_3 \tag{1}$$

where $a, b, c, d \in R$, k_1, k_2 and k_3 are three orthogonal imaginary parts that conform to the following calculation rules:

$$\begin{cases} k_1k_2 = k_3, k_2k_3 = k_1, k_3k_1 = k_2 \\ k_2k_1 = -k_3, k_3k_2 = -k_1, k_1k_3 = -k_2 \\ k_1^2 = k_2^2 = k_3^2 = k_3k_1k_2 = -1 \end{cases} \tag{2}$$

The conjugate and modulus values of any quaternion can be expressed as:

$$\bar{q} = a - bk_1 - ck_2 - dk_3 \tag{3}$$

$$|q| = \sqrt{q\bar{q}} = \sqrt{a^2 + b^2 + c^2 + d^2} \tag{4}$$

In the human visual system, the orientation information of the image plays a vital role. Studies have shown that the human brain is able to efficiently process orientation information in images, allowing for rapid recognition of the shape, structure, and spatial positioning of objects. In order to better extract the image orientation information, we propose a quaternion model, which takes the directional complementary sub-band as the three imaginary parts of the quaternion. Each pixel can be represented as a pure quaternion, i.e., a quaternion with a real part of 0. The image orientation information of the two channels can be represented by pure quaternions:

$$f_h = f_{hX}k_1 + f_{hY}k_2 + f_{hZ}k_3 \tag{5}$$

$$f_v = f_{vX}k_1 + f_{vY}k_2 + f_{vZ}k_3 \tag{6}$$

f_{hX}, f_{hY} and f_{hZ} denote the three horizontal complementary sub-bands, while f_{vX}, f_{vY} and f_{vZ} denote the three vertical complementary sub-bands. For the directional channel $t \in \{h, v\}$, f_{tX}, f_{tY} and f_{tZ} denote the three complementary sub-bands of

channel t, respectively. We then define the image orientation information as a QDWT of size $M \times N$, which is represented as follows:

$$QW_{\rho}^{L}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x,y=0} zf\rho_{j_0,m,n} \tag{7}$$

$$QW_{\psi}^{\{H,V,D\}}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x,y=0} zf\psi_{j_0,m,n}^{\{H,V,D\}} \tag{8}$$

where ρ and ψ are scale functions and wavelet functions, j_0 is an arbitrary initial scale, and z is a pure quaternion in unit. QDWT divides the image orientation information into a low-frequency part and a high-frequency part, the low-frequency part QW_{ρ}^{L} usually contains the main information of the signal, and the high-frequency part is related to noise, interference and small details in the image. We represent the direction features extracted by QDWT as follows:

$$Q_{final} = Q_t + Q_{XY}k_1 + Q_{XZ}k_2 + Q_{YZ}k_3 \tag{9}$$

where Q_t is the real part and Q_{XY}, Q_{XZ} and Q_{YZ} are the three imaginary components, which are defined as follows:

$$\begin{aligned} Q_t &= -\alpha DWT(f_{tZ}) - \beta DWT(f_{tY}) - \gamma DWT(f_{tX}) \\ Q_{XY} &= \beta DWT(f_{tX}) - \gamma DWT(f_{tY}) \\ Q_{XZ} &= -\alpha DWT(f_{tX}) + \gamma DWT(f_{tZ}) \\ Q_{YZ} &= \alpha DWT(f_{tY}) - \beta DWT(f_{tZ}) \end{aligned} \tag{10}$$

Where DWT$(\cdot)$ stands for the traditional discrete wavelet transform. α, β, γ represent the weighting coefficients.

3.3 Low-frequency Structure Feature Extraction

The QDWT low-frequency sub-band retains the main structural information of the image, such as object contours, regional distribution and overall brightness variation. In visual security assessment, if the encrypted image still preserves the low-frequency structure of the original image, the human eye may still recognize the main image content even when fine texture details are partially destroyed. Therefore, the low-frequency structural difference is an important factor for measuring visual security. The extraction process of low-frequency features is shown in Figure 3.

For each directional channel $t \in \{h, v\}$, the pure quaternion image f_t is first decomposed by a three-level QDWT. At each decomposition level, the QDWT separates the input signal into one low-frequency approximation sub-band and several high-frequency directional sub-bands. The high-frequency directional sub-bands $Q_{t,\sigma}^{H}$, $Q_{t,\sigma}^{V}$ and $Q_{t,\sigma}^{D}$, where $\sigma \in \{1, 2, 3\}$, are mainly used for texture feature extraction in Section 3.4.1. The complementary sub-band correlation components Q_{XY}, Q_{XZ} and Q_{YZ} are used for multi-scale structural feature extraction in Section 3.4.2. In contrast, the low-frequency structure feature is directly obtained from the final low-frequency approximation sub-band after the three-level QDWT decomposition.

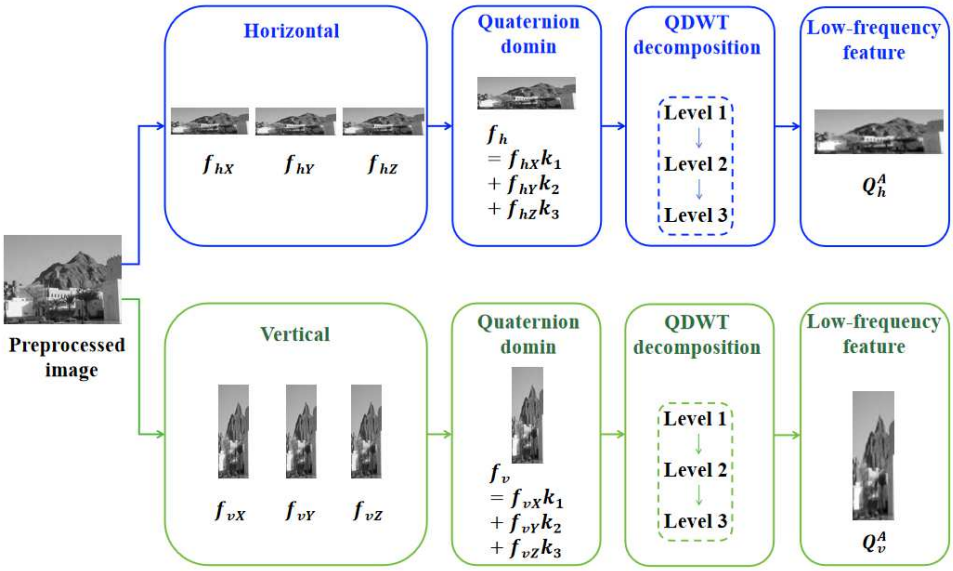


Fig. 3 The process of low-frequency feature extraction

Let Q_t^A denote the low-frequency approximation sub-band of directional channel t after the third-level QDWT decomposition. Since it is obtained from the deepest approximation branch, Q_t^A contains the coarsest-scale structural information of the directional channel, including the global luminance distribution, main contours and regional layout. Different from the texture and multi-scale structural features extracted from high-frequency or correlation components, Q_t^A is directly taken as the low-frequency structure feature map without additional inverse QDWT reconstruction. This is because the subsequent attention-guided feature extraction in Section 3.5 is performed directly on the feature map through strip-wise pooling and statistical modeling.

In this way, Q_t^A provides a compact description of the low-frequency structure preserved in the image. For a perceptually encrypted image, if the main low-frequency contours and regional brightness distribution are still similar to those of the reference image, the image content may remain visually recognizable. Therefore, the difference between the low-frequency structure feature maps of the reference image and the encrypted image can effectively reflect whether the main visual structure has been destroyed. Together with the texture feature map F_t^T and the multi-scale structural feature map F_t^S obtained in the following subsections, Q_t^A is fed into the attention-guided perceptual feature extraction module, where the global and local feature sequences are obtained in a unified manner.

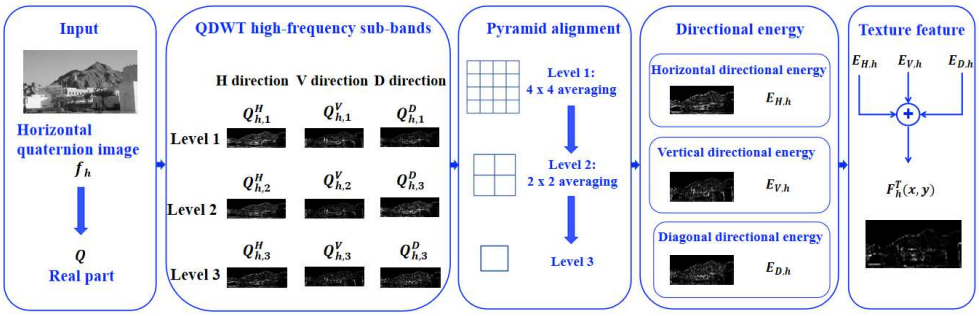


Fig. 4 The process of texture feature extraction in the horizontal channel.

3.4 QDWT-Based Texture and Multi-Scale Structural Feature Extraction

The low-frequency structure feature mainly reflects the overall content of the image, but visual security assessment also needs to pay attention to whether edges, textures and local directional structures remain in the encrypted image. As a description of the repeated patterns within image regions, texture carries content information that the human eye can readily perceive. Even if the overall structure is destroyed, if obvious edge contours or texture distributions still exist in the encrypted image, the human eye may still infer part of the content of the original image. Therefore, we further extract texture features from the QDWT high-frequency coefficients, and construct multi-scale structural features using the complementary sub-band correlation components.

3.4.1 Texture Feature Extraction

For the directional channel $t \in \{h, v\}$, the high-frequency directional sub-bands obtained by QDWT decomposition contain horizontal, vertical and diagonal directional information at different scales, in which the local texture details of the image are mainly concentrated. The 1st, 2nd and 3rd level directional high-frequency sub-bands, denoted by $Q^H_{t,\sigma}$, $Q^V_{t,\sigma}$ and $Q^D_{t,\sigma}$ with $\sigma \in \{1, 2, 3\}$, are selected for texture feature extraction. Taking the horizontal channel as an example, the texture feature extraction process is shown in Figure 4.

Since different decomposition levels have different spatial resolutions, in order to fuse multi-level texture responses at a unified scale, local averaging is first performed on the 1st and 2nd level directional sub-bands. Specifically, the 1st level directional sub-bands are divided into 4×4 blocks for averaging, and the 2nd level directional sub-bands are divided into 2×2 blocks for averaging, so that they have the same spatial resolution as the 3rd level directional sub-bands. The 3rd level directional sub-bands are already at the target resolution and are therefore used directly in the subsequent energy fusion. In the following formulation, $H_{3,t}$, $V_{3,t}$ and $D_{3,t}$ denote the 3rd level horizontal, vertical and diagonal directional high-frequency responses of channel t,

respectively. The corresponding averaged directional coefficients are calculated as:

$$\begin{aligned}\left(H_{1,t}^{ave}, V_{1,t}^{ave}, D_{1,t}^{ave}\right) &= \frac{1}{16}\sum_{4\times4}\left(Q_{t,1}^{H}, Q_{t,1}^{V}, Q_{t,1}^{D}\right), \\ \left(H_{2,t}^{ave}, V_{2,t}^{ave}, D_{2,t}^{ave}\right) &= \frac{1}{4}\sum_{2\times2}\left(Q_{t,2}^{H}, Q_{t,2}^{V}, Q_{t,2}^{D}\right), \\ \left(H_{3,t}, V_{3,t}, D_{3,t}\right) &= \left(Q_{t,3}^{H}, Q_{t,3}^{V}, Q_{t,3}^{D}\right).\end{aligned} \tag{11}$$

After the scale alignment, the directional responses of the 1st, 2nd and 3rd levels are fused to construct the cross-scale texture energy. Since the texture feature is subsequently aggregated only by a variance-based statistic and the directional energy is already well characterized in the QDWT coefficient domain, reconstructing the coefficients back into the spatial domain through IQDWT is not required. The aligned directional coefficients are squared and accumulated over the three scales to construct the cross-scale texture energies of the horizontal, vertical and diagonal directions, which are respectively defined as:

$$\begin{aligned}E_{H,t} &= \left(H_{1,t}^{ave}\right)^2 + \left(H_{2,t}^{ave}\right)^2 + \left(H_{3,t}\right)^2, \\ E_{V,t} &= \left(V_{1,t}^{ave}\right)^2 + \left(V_{2,t}^{ave}\right)^2 + \left(V_{3,t}\right)^2, \\ E_{D,t} &= \left(D_{1,t}^{ave}\right)^2 + \left(D_{2,t}^{ave}\right)^2 + \left(D_{3,t}\right)^2.\end{aligned} \tag{12}$$

where $E_{H,t}$, $E_{V,t}$ and $E_{D,t}$ represent the cross-scale texture energies in the horizontal, vertical and diagonal directions of the directional channel t, respectively. Each energy accumulates the squared directional responses over the three scales, so that stable texture structures that are consistent across scales are enhanced while unstable weak responses are suppressed. The three directional energies are then summed to obtain the QDWT texture feature map of the channel:

$$F_t^T = E_{H,t} + E_{V,t} + E_{D,t}, \tag{13}$$

where F_t^T represents the QDWT texture feature of the directional channel t, which is used to describe the directional texture, edge contours and local detail variation in the image. In this paper, the three directional energy components $E_{H,t}$, $E_{V,t}$ and $E_{D,t}$ are also regarded as the horizontal, vertical and diagonal components of the texture feature, respectively, and are further used in the joint feature modeling of the attention-guided extraction stage. Compared with reconstructing the coefficients into the spatial domain before the statistic, the coefficient-domain energy representation is more direct and retains the same texture-discriminative information.

3.4.2 Multi-Scale Structural Feature Extraction

In addition to the directional high-frequency sub-bands, the complementary subband correlation components in QDWT also contain important structural information.

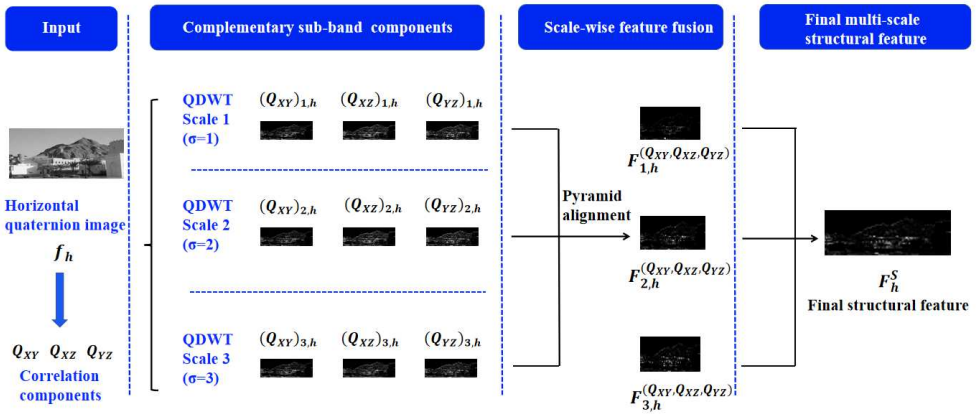


Fig. 5 The extraction process of multi-scale structural features in the horizontal channel.

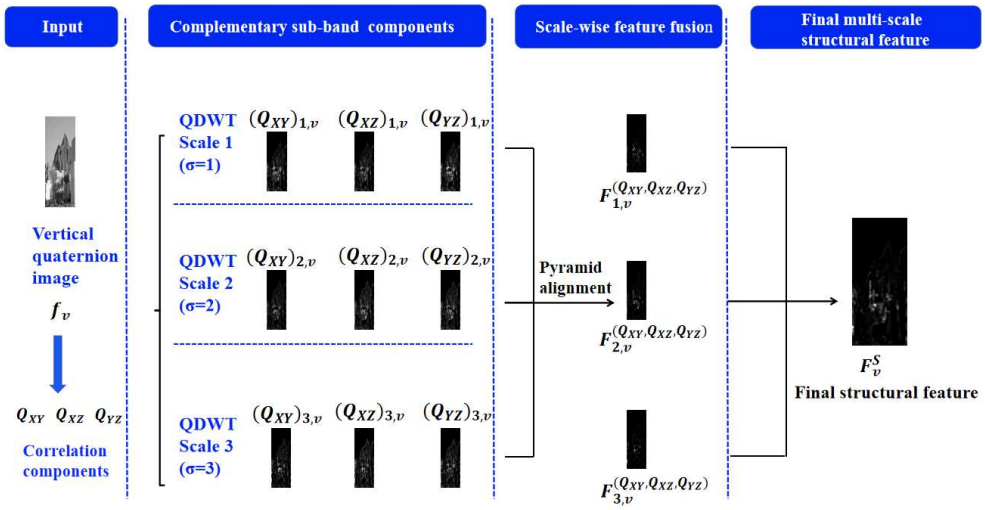


Fig. 6 The extraction process of multi-scale structural features in the vertical channel.

Specifically, Q_{XY}, Q_{XZ} and Q_{YZ} describe the differences and correlations between the three complementary sub-bands. Since these complementary sub-bands come from different sampling positions of the same grayscale image, their correlation components can reflect the variation of local structures at different sampling positions. The process of multi-scale feature extraction is shown in Figs. 5 and 6.

For the directional channel t, each decomposition level $\sigma \in \{1, 2, 3\}$ produces its own set of correlation components, denoted by $(Q_{XY})_{\sigma,t}$, $(Q_{XZ})_{\sigma,t}$ and $(Q_{YZ})_{\sigma,t}$, which are selected for structure feature extraction. Consistent with the texture feature, the structure feature is subsequently aggregated only by a variance-based statistic.

Since the correlation components from different decomposition levels have different spatial resolutions, they are aligned to the spatial resolution of the 3rd level before cross-scale fusion. Specifically, the structural responses from the 1st and 2nd levels are aligned by the same local averaging strategy used in texture feature extraction, while the 3rd level response is directly used as the target scale. In this way, the structural responses from the three levels can be represented on a unified coefficient grid.

The correlation components are squared and summed in the coefficient domain to obtain the structure feature at scale σ:

$$F_{\sigma,t}^{(Q_{XY},Q_{XZ},Q_{YZ})} = (Q_{XY})_{\sigma,t}^2 + (Q_{XZ})_{\sigma,t}^2 + (Q_{YZ})_{\sigma,t}^2, \quad \sigma \in \{1,2,3\}. \tag{14}$$

Here, $F_{\sigma,t}^{(Q_{XY},Q_{XZ},Q_{YZ})}$ represents the structural response of channel t at scale σ after the scale-consistent representation. The three correlation components describe the structural relationships among the complementary sub-bands from different perspectives, and their energy fusion can highlight the local structural variations caused by different sampling positions. The structure features of the directional channel t at multiple scales are further summed to obtain the multi-scale structural feature map of the channel:

$$F_t^S = \sum_{\sigma=1}^{3} F_{\sigma,t}^{(Q_{XY},Q_{XZ},Q_{YZ})}. \tag{15}$$

where F_t^S represents the multi-scale structural feature of the directional channel t. Here, the per-component energies $\sum_{\sigma=1}^{3}(Q_{XY})_{\sigma,t}^2$, $\sum_{\sigma=1}^{3}(Q_{XZ})_{\sigma,t}^2$ and $\sum_{\sigma=1}^{3}(Q_{YZ})_{\sigma,t}^2$ are also regarded as the XY, XZ and YZ components of the structural feature, denoted $F_t^{S,XY}$, $F_t^{S,XZ}$ and $F_t^{S,YZ}$ respectively, and are further used in the joint feature modeling of the attention-guided extraction stage.Compared with the texture feature F_t^T, F_t^S places more emphasis on the structural relationship between the complementary sub-bands, and the two can describe the perceptible information that may remain in the encrypted image from different perspectives.

3.5 Attention-Guided Perceptual Feature Extraction

After obtaining the low-frequency structure feature map Q_t^A, the QDWT texture feature map F_t^T and the multi-scale structural feature map F_t^S of each directional channel $t \in \{h, v\}$, it is necessary to further consider the importance of different regions to visual perception. Since the human eye is more sensitive to salient edges, local textures and structurally abrupt regions, we construct, for each feature map, a global attention sequence and a local feature sequence, both of which are obtained through strip-wise pooling. In this way, the global statistical distribution and the local responses are characterized from two complementary perspectives, and the visually salient regions contribute more to the final hash features. The following construction is performed on the horizontal channel and the vertical channel separately; for brevity, the channel index t is omitted in this subsection. For a unified description, let $M \in \{Q^A, F^T, F^S\}$ denote any one of the three feature maps, and let the feature maps be partitioned along the width direction into $F = W/b$ non-overlapping strips $\mathcal{B}_i$ of size $H \times b$.

3.5.1 Global Attention Feature Sequence

The feature distribution of all positions is first analyzed from a global statistical perspective. The three types of feature responses, together with their directional and correlation components, are stacked into a joint feature vector at each position. The low-frequency structure feature Q^A describes the coarse structural information of the image. The texture components $F^{T,H}$, $F^{T,V}$ and $F^{T,D}$ correspond to the horizontal, vertical and diagonal texture energy components obtained from the QDWT directional high-frequency sub-bands. The structural components $F^{S,XY}$, $F^{S,XZ}$ and $F^{S,YZ}$ correspond to the structural responses derived from the three complementary sub-band correlation components Q_{XY}, Q_{XZ} and Q_{YZ}. The joint feature vector at each position is defined as:

$$\vec{F} = \left[\, Q^A,\ F^{T,H},\ F^{T,V},\ F^{T,D},\ F^{S,XY},\ F^{S,XZ},\ F^{S,YZ} \,\right]^\top, \tag{16}$$

whose dimension is $n = 7$. Compared with using only the texture and structure responses, the joint feature vector provides a richer description of the local content, so that the subsequent multivariate statistical modeling becomes effective. Let $\vec{F}_{ave}$ be the mean of all joint feature vectors in the whole image, and Σ be the covariance matrix. According to the multivariate normal distribution, the probability density of the joint feature vector at each position is defined as:

$$p\left(\vec{F}\right) = \frac{\exp\left[-\frac{1}{2}\left(\vec{F} - \vec{F}_{ave}\right)^\top \Sigma^{-1}\left(\vec{F} - \vec{F}_{ave}\right)\right]}{(2\pi)^{n/2}|\Sigma|^{1/2}}. \tag{17}$$

where Σ^{-1} represents the inverse of the covariance matrix, $|\Sigma|$ represents its determinant, and n represents the dimension of the joint feature vector. The probability density $p(\vec{F})$ measures how common the feature of the current position is in the global distribution. If the probability density of a position is low, it means that the feature of this position is relatively rare compared with the whole image, which usually corresponds to regions with abrupt edges, enhanced textures or abnormal local structures. Based on this, the global saliency response is defined as:

$$G = \left(\log\left(p\left(\vec{F}\right)^{-1}\right)\right)^{1/2}, \tag{18}$$

where the inverse probability ensures that statistically rare and thus salient positions obtain higher responses.

The global saliency response G is shared by the three feature maps. To convert it into a compact feature sequence, an average pooling operation with a full-height strip kernel is applied along the width direction:

$$s_G^{(i)} = \mathrm{AP}_{H\times b}[G] = \frac{1}{Hb}\sum_{\mathcal{B}_i} G, \quad i = 1, 2, \ldots, F, \tag{19}$$

which yields the shared saliency sequence $D^G = \left[s_G^{(1)}, s_G^{(2)}, \ldots, s_G^{(F)}\right]$. The average pooling preserves the overall saliency distribution of each strip and reflects the global background-related information. To obtain the global attention sequence of a specific feature map, each of the low-frequency, texture and structural feature maps is modulated by the saliency response G and then pooled strip by strip in the same way as (19), giving the global attention feature sequences $D^{A,G}$, $D^{T,G}$ and $D^{S,G}$, respectively. In this manner, the global saliency is consistently injected into all three features.

3.5.2 Local Attention Feature Sequence

Relying solely on global statistics may ignore the strong response regions within local windows. To capture the local variation of each feature map, a local feature sequence is further constructed by strip-wise pooling, where the pooling operator is selected according to the property of the feature map. For the smooth low-frequency feature map, the variance is more sensitive to local structural fluctuation than the peak response. Therefore, the variance pooling is adopted for Q^A:

$$s_{L,A}^{(i)} = \frac{1}{Hb-1} \sum_{\mathcal{B}_i} \left(Q^A - \bar{\mathcal{B}}_i^A\right)^2, \quad i = 1, 2, \ldots, F, \tag{20}$$

where $\bar{\mathcal{B}}_i^A$ is the mean of Q^A over the strip $\mathcal{B}_i$. For the high-frequency texture and structural feature maps, the local texture or structure mutation is better captured by the peak response. Therefore, a max pooling operation with the full-height strip kernel is applied to F^T and F^S, respectively:

$$s_L^{(i)} = \mathrm{MP}_{H\times b}[\,\cdot\,] = \max_{\mathcal{B}_i}(\,\cdot\,), \quad i = 1, 2, \ldots, F. \tag{21}$$

Accordingly, the local feature sequences $D^{A,L}$, $D^{T,L}$ and $D^{S,L}$ of the low-frequency, texture and structural feature maps are obtained, where $D^{A,L}$ is computed by variance pooling and $D^{T,L}$, $D^{S,L}$ by max pooling.

To suppress the global offset introduced by a uniform encryption operation, each feature sequence is relativized by subtracting its own mean and taking the absolute value. Taking the texture sequences as an example:

$$\hat{D}^{T,G}(i) = \left| s_G^{(i)} - \overline{D^{T,G}} \right|, \qquad \hat{D}^{T,L}(i) = \left| s_L^{(i)} - \overline{D^{T,L}} \right|, \tag{22}$$

where $\overline{D^{T,G}}$ and $\overline{D^{T,L}}$ are the means of the corresponding sequences. The low-frequency sequences $D^{A,G}$, $D^{A,L}$ and the structural sequences $D^{S,G}$, $D^{S,L}$ are relativized in the same manner to obtain $\hat{D}^{A,G}$, $\hat{D}^{A,L}$, $\hat{D}^{S,G}$ and $\hat{D}^{S,L}$. For each directional channel t, six feature sequences are thus obtained from the low-frequency, texture and structural feature maps, which are used for the subsequent hash construction.

4 Hash Similarity Assessment

After the above steps, three types of features can be obtained for both the horizontal channel and the vertical channel: the low-frequency structure feature, the texture feature, and the multi-scale structural feature, each represented by a global attention sequence and a local feature sequence. For the horizontal direction, the features are directly concatenated to form the horizontal feature sequence:

$$Q_h = \left[\hat{D}_h^{A,G},\ \hat{D}_h^{A,L},\ \hat{D}_h^{T,G},\ \hat{D}_h^{T,L},\ \hat{D}_h^{S,G},\ \hat{D}_h^{S,L}\right], \tag{23}$$

where $\hat{D}_h^{A,G}$ and $\hat{D}_h^{A,L}$ represent the horizontal global and local low-frequency features, $\hat{D}_h^{T,G}$ and $\hat{D}_h^{T,L}$ represent the horizontal global and local texture features, and $\hat{D}_h^{S,G}$ and $\hat{D}_h^{S,L}$ represent the horizontal global and local structural features. Similarly, for the vertical direction, the features are directly concatenated to form the vertical feature sequence:

$$Q_v = \left[\hat{D}_v^{A,G},\ \hat{D}_v^{A,L},\ \hat{D}_v^{T,G},\ \hat{D}_v^{T,L},\ \hat{D}_v^{S,G},\ \hat{D}_v^{S,L}\right], \tag{24}$$

where the symbols are defined analogously to those of the horizontal direction. The initial hash sequence is generated from the direction features computed by QDWT:

$$l = \left[\lambda Q_h,\ (1-\lambda)Q_v\right], \tag{25}$$

where l is the initial hash sequence of length d, d is the sum of the sequence lengths of Q_h and Q_v, and λ is used to adjust the relative importance of the horizontal and vertical directional features, and is closely related to the performance of the VSA method.

Suppose $l(i)$ represents the i-th element in the initial hash sequence, $1 \le i \le d$. Since $l(i)$ is a decimal number that requires many bits to store, for example, a floating-point number requires 32 bits and a double number requires 64 bits, the element $l(i)$ is quantized as an integer $L(i)$ to reduce the storage space:

$$L(i) = \text{round}\left[l(i) \times 100 + 0.5\right], \tag{26}$$

where round$[\cdot]$ is the rounding operation. Therefore, the final hash sequence $\mathbf{L}$ is a sequence of d integers:

$$\mathbf{L} = [L(1), L(2), \ldots, L(d)]. \tag{27}$$

We use the L_2 norm to measure the similarity of two image hashes. Suppose $\mathbf{L}_1 = [L_1(1), L_1(2), \ldots, L_1(d)]$ and $\mathbf{L}_2 = [L_2(1), L_2(2), \ldots, L_2(d)]$ are the hash sequences of the original image and the encrypted image, respectively. The L_2 norm of the two hash sequences is defined as:

$$S(\mathbf{L}_1, \mathbf{L}_2) = \sqrt{\sum_{i=1}^{d} \left[L_1(i) - L_2(i)\right]^2}, \tag{28}$$

where $L_1(i)$ and $L_2(i)$ are the i-th elements of $\mathbf{L}_1$ and $\mathbf{L}_2$, respectively. In general, a smaller S means that the two images fed into the hash sequences are more similar, and thus the visual security of the encrypted image is lower.

5 Test Databases

To assess the performance of QuatHashing, we test it on two widely used encrypted image databases, IVC-SelectEncrypt[33] and PEID[35]. The IVC-SelectEncrypt database includes 200 encrypted images generated from eight ordinary images through five different encryption algorithms based on five different security levels. The PEID database contains 1080 encrypted images, of which 20 ordinary images use ten perceptually encryption schemes. Both databases used a subjective survey approach, allowing each encrypted image to receive a mean opinion score (MOS), which indicates its visual security score. Therefore, MOS can be regarded as the ground truth for the visual security score predicted by our method.

Table 2 Overall performance on IVC-SelectEncrypt database.

Metrics	PSNR	SSIM	VIF	LSS	ESS	LFBVS	VSI-Canny	VSASW	Proposed
μ_D	0.156	0.122	0.170	0.222	0.343	0.254	0.154	0.098	**0.049**
σ_D	0.101	0.184	0.101	0.188	0.188	0.174	0.089	0.099	**0.081**
$\max_D$	0.394	0.689	0.404	0.520	0.399	0.479	0.353	**0.343**	0.587
PLCC	0.915	0.891	0.939	0.919	0.901	0.891	0.942	0.956	**0.967**
SRCC	0.911	0.869	0.939	0.932	0.909	0.891	0.935	0.954	**0.964**
KRCC	0.743	0.702	0.791	0.786	0.747	0.712	0.776	0.815	**0.846**
RMSE	0.543	0.645	0.455	0.658	0.506	0.601	0.446	0.388	**0.304**

Table 3 Overall performance on PEID database.

Metrics	PSNR	SSIM	VIF	LSS	ESS	LFBVS	VSI-Canny	VSASW	Proposed
μ_D	0.189	0.431	0.169	0.098	0.158	0.116	0.333	**0.096**	0.153
σ_D	0.144	0.239	0.077	**0.044**	0.137	0.105	0.210	0.059	0.136
$\max_D$	0.538	0.853	0.534	0.771	0.479	0.530	0.882	0.774	**0.462**
PLCC	0.825	0.845	0.854	0.823	0.808	0.724	0.850	0.886	**0.928**
SRCC	0.796	0.831	0.910	0.796	0.772	0.618	0.726	0.870	**0.920**
KRCC	0.616	0.657	0.768	0.616	0.592	0.458	0.557	0.691	**0.864**
RMSE	1.056	0.999	**0.645**	1.061	1.065	1.248	0.984	0.866	0.668

5.1 Overall Performance Analysis

The metrics for the overall performance of all IQA and VSAs on the IVC-SelectEncrypt and PEID databases are listed in Table 2 and Table 3, respectively, and the best

performing metrics are marked in bold. In IVC-SelectEncrypt database, our proposed method has the best μ_D, σ_D, PLCC[36], SRCC[37] and KRCC[38] and RMSE[39] values. In PEID database, it has the best $\max_D$,PLCC, SRCC, KRCC values. It is clear that QuatHashing performs better with IVC-SelectEncrypt and PEID than the other methods. More than half of the indicators show optimization.

5.2 Parameter Setting and Scatter Plot

λ is used to adjust the relative importance of the horizontal and vertical directional features, and is closely related to the performance of the VSA method. Different λ will lead to different performance of the VSA method. As shown in Table 4, when the value of λ is 0.8, the VSA method performs the best.

Table 4 The performance of QuatHashing under different values of PEID database.

λ	**PLCC**	**SRCC**	**KRCC**	**RMSE**
1	0.919	0.901	0.723	0.724
0.9	0.923	0.903	0.725	0.719
0.8	**0.928**	**0.920**	**0.864**	**0.668**
0.7	0.925	0.914	0.763	0.686
0.2	0.922	0.913	0.758	0.726
0.1	0.919	0.909	0.755	0.731
0	0.917	0.885	0.723	0.756

In order to show the prediction of our method more vividly, we show the scatter plot of MOS versus predicted scores in Figure 7. Compared with other methods, the predictions of our method are closer to the real MOS.

5.3 Performance Analysis of Different Mass Ranges

Perceptually encrypted images are usually concentrated in the lower quality range, so this section assesses the performance of these VSAs in different image quality ranges (i.e. low, moderate, and high range).

The comparison results of the two test databases in different image quality ranges are shown in Table 5 and Table 6. On the IVC-SelectEncrypt database, as shown in Table 5, in the low quality range, VIF has the lowest $\max_D$ value, LSS shows the best performance on RMSE, QuatHashing has the best performance on μ_D , σ_D, PLCC, SRCC and KRCC. In the moderate quality range, PSNR has the lowest $\max_D$ value, and QuatHashing is in μ_D ,σ_D , PLCC, SRCC, KRCC and RMSE. In the high quality range, QuatHashing shows the best performance on PLCC. On the PEID database, as shown in Table 6, in the low quality range, VIF has the lowest σ_D and $\max_D$ values, VSASW shows the best performance on RMSE, QuatHashing has the best performance on μ_D , PLCC, SRCC and KRCC. In the moderate quality range, VIF has the lowest μ_D value, FBVS has the lowest σ_D and $\max_D$ values, VSASW has the

Table 5 Performance of different quality ranges on the IVC-SelectEncrypt database.

	Metrics	PSNR	SSIM	VIF	LSS	ESS	LFBVS	VSI-Canny	VSASW	Proposed
low quality	μ_D	0.216	0.367	0.283	0.311	0.218	0.179	0.183	0.155	**0.144**
	σ_D	0.112	0.199	0.137	0.211	0.307	0.256	0.099	0.092	**0.083**
	$\max_D$	0.394	0.689	**0.283**	0.876	0.312	0.479	0.353	0.306	0.362
	PLCC	0.684	0.680	0.900	0.935	0.760	0.586	0.900	0.912	**0.942**
	SRCC	0.703	0.528	0.889	0.902	0.673	0.662	0.826	0.899	**0.905**
	KRCC	0.524	0.389	0.711	0.746	0.582	0.479	0.635	0.733	**0.794**
	RMSE	0.381	0.384	0.228	**0.185**	0.574	0.620	0.228	0.246	0.218
mid quality	μ_D	0.129	0.168	0.222	0.290	0.136	0.141	0.183	0.154	**0.034**
	σ_D	0.174	0.125	0.138	0.116	0.258	0.273	0.090	0.094	**0.029**
	$\max_D$	**0.255**	0.393	0.404	0.423	0.376	0.312	0.277	0.306	0.372
	PLCC	0.514	0.474	0.583	0.503	0.602	0.388	0.503	0.713	**0.746**
	SRCC	0.531	0.348	0.521	0.495	0.577	0.321	0.473	0.629	**0.723**
	KRCC	0.367	0.244	0.367	0.345	0.324	0.253	0.324	0.422	**0.489**
	RMSE	0.286	0.294	0.271	0.289	0.602	0.639	0.288	0.234	**0.217**
high quality	μ_D	0.130	**0.038**	0.186	0.114	0.220	0.198	0.119	0.164	0.241
	σ_D	0.076	**0.044**	0.093	0.117	0.331	0.277	0.068	0.080	0.073
	$\max_D$	0.301	**0.178**	0.368	0.424	0.399	0.309	0.259	0.334	0.360
	PLCC	0.694	0.740	0.596	0.253	0.588	0.646	0.605	0.650	**0.788**
	SRCC	0.694	**0.732**	0.606	0.494	0.599	0.646	0.606	0.638	0.728
	KRCC	0.520	**0.558**	0.445	0.360	0.442	0.466	0.437	0.469	0.492
	RMSE	0.316	**0.295**	0.353	0.425	0.564	0.549	0.352	0.333	0.367

Table 6 Performance of different mass ranges on PEID database.

	Metrics	PSNR	SSIM	VIF	LSS	ESS	FBVS	VSI-Canny	VSASW	Proposed
low quality	μ_D	0.152	0.300	0.030	0.389	0.170	0.049	0.197	0.097	**0.028**
	σ_D	0.170	0.159	**0.019**	0.152	0.095	0.039	0.105	0.041	0.064
	$\max_D$	0.527	0.716	**0.142**	0.753	0.376	0.283	0.643	0.225	0.363
	PLCC	0.593	0.577	0.793	0.641	0.573	0.315	0.432	0.804	**0.811**
	SRCC	0.574	0.617	0.698	0.531	0.484	0.287	0.196	0.711	**0.785**
	KRCC	0.405	0.452	0.532	0.380	0.355	0.204	0.135	0.533	**0.593**
	RMSE	0.534	0.542	0.449	0.510	0.573	0.619	0.599	**0.403**	0.463
mid quality	μ_D	0.187	0.527	**0.091**	0.432	0.398	0.116	0.441	0.290	0.284
	σ_D	0.145	0.161	0.054	0.192	0.195	**0.038**	0.161	0.127	0.116
	$\max_D$	0.499	0.850	0.387	0.771	0.394	**0.276**	0.725	0.565	0.379
	PLCC	0.355	0.407	0.617	0.295	0.387	0.376	0.474	**0.623**	0.598
	SRCC	0.320	0.368	0.604	0.418	0.289	0.257	0.475	0.601	**0.637**
	KRCC	0.221	0.256	0.418	0.288	0.198	0.170	0.333	0.430	**0.457**
	RMSE	0.531	0.519	0.503	0.543	0.534	0.539	0.500	0.519	**0.328**
high quality	μ_D	0.210	0.511	**0.159**	0.295	0.278	0.286	0.393	0.317	0.366
	σ_D	0.172	0.321	**0.133**	0.180	0.163	0.172	0.264	0.198	0.136
	$\max_D$	0.538	0.853	0.534	0.628	0.479	0.530	0.882	0.774	**0.432**
	PLCC	0.850	0.849	0.838	0.518	0.840	0.693	**0.860**	0.480	0.685
	SRCC	0.795	0.781	0.785	0.632	0.766	0.629	**0.820**	0.459	0.673
	KRCC	0.608	0.593	0.584	0.466	0.568	0.433	**0.639**	0.321	0.540
	RMSE	0.473	0.329	0.312	0.532	**0.310**	0.412	0.317	0.498	0.468

Table 7 Performance of different encryption methods (SRCC and KRCC).

SRCC									
Parameter	PSNR	SSIM	VIF	LSS	ESS	LFBVS	VSI-Canny	VSASW	Proposed
trad	0.965	0.968	0.964	0.948	0.933	0.960	0.951	0.968	**0.980**
trunc	0.888	0.879	**0.953**	0.948	0.913	0.907	**0.953**	0.921	0.946
iwind_ec	0.909	0.924	0.941	0.887	0.859	0.886	0.927	**0.965**	0.940
iwind_nec	0.910	0.953	0.951	0.907	0.946	0.946	0.934	0.953	**0.985**
res	0.915	0.886	0.955	0.955	0.926	0.916	0.944	0.960	**0.960**
enc01	0.906	0.927	0.897	**0.951**	0.540	0.452	0.864	0.937	0.833
enc02	0.951	0.911	**0.963**	0.835	0.846	0.627	0.887	0.938	**0.949**
enc03	0.923	0.898	0.977	0.944	0.966	0.862	0.975	**0.978**	0.946
enc04	0.911	0.916	0.916	0.908	0.828	0.907	0.738	**0.962**	**0.962**
enc05	0.929	0.936	0.940	0.924	0.789	0.821	0.927	**0.946**	0.897
enc06	0.970	0.965	0.972	0.968	0.961	0.914	0.972	0.976	**0.979**
enc07	0.956	0.977	0.983	0.958	0.975	0.959	0.982	0.971	**0.984**
enc08	0.083	0.666	0.191	0.200	0.325	0.360	**0.710**	0.246	0.623
enc09	0.235	**0.760**	0.582	0.187	0.364	0.339	0.155	0.521	0.736
enc10	0.760	0.943	0.980	0.631	0.846	0.678	0.632	0.930	**0.965**
KRCC									
Parameter	PSNR	SSIM	VIF	LSS	ESS	LFBVS	VSI-Canny	VSASW	Proposed
trad	0.864	0.864	0.854	0.820	0.799	0.852	0.818	0.854	**0.890**
trunc	0.731	0.710	0.803	**0.829**	0.801	0.798	0.826	0.764	0.718
iwind_ec	0.738	0.766	**0.800**	0.751	0.724	0.753	0.779	0.811	0.785
iwind_nec	0.770	0.837	0.829	0.820	0.827	0.822	0.798	0.837	**0.917**
res	0.778	0.717	0.839	0.842	0.787	0.765	0.825	**0.846**	**0.846**
enc01	0.735	0.774	0.759	**0.807**	0.473	0.386	0.709	0.759	0.650
enc02	**0.809**	0.760	0.785	0.650	0.668	0.453	0.711	0.764	0.808
enc03	0.749	0.741	0.825	0.795	0.813	0.724	0.872	**0.873**	0.813
enc04	0.742	0.750	0.741	0.736	0.694	0.713	0.538	0.769	**0.819**
enc05	0.774	0.790	0.783	0.767	0.613	0.689	0.783	**0.830**	0.733
enc06	0.847	0.833	0.852	0.845	0.844	0.825	**0.853**	0.853	0.841
enc07	0.805	0.869	0.885	0.818	0.876	0.859	0.884	0.862	**0.897**
enc08	0.066	0.528	0.144	0.145	0.256	0.259	**0.558**	0.184	0.481
enc09	0.162	**0.582**	0.493	0.126	0.289	0.276	0.082	0.380	0.561
enc10	0.589	0.807	0.823	0.473	0.723	0.649	0.468	0.788	**0.858**

Table 8 Performance of different encryption methods (PLCC and RMSE).

PLCC									
Parameter	PSNR	SSIM	VIF	LSS	ESS	LFBVS	VSI-Canny	VSASW	Proposed
trad	0.955	0.969	0.969	0.955	0.922	0.957	0.955	0.968	**0.986**
trunc	0.916	0.907	**0.980**	0.970	0.946	0.949	0.974	0.958	0.962
iwind_ec	0.957	0.981	0.983	0.941	0.916	0.967	0.982	0.984	**0.988**
iwind_nec	0.931	0.965	0.960	0.907	0.924	0.948	0.950	0.971	**0.976**
res	0.905	0.931	0.978	0.955	0.949	0.939	0.961	0.961	**0.983**
enc01	0.904	0.957	0.948	0.952	0.796	0.583	0.944	0.946	**0.966**
enc02	0.967	0.950	0.970	0.888	0.861	0.636	0.888	0.939	**0.968**
enc03	0.964	0.956	**0.985**	0.970	0.979	0.917	0.984	**0.985**	0.982
enc04	0.940	0.928	0.950	0.939	0.816	0.911	0.883	0.967	**0.971**
enc05	0.917	0.947	0.955	0.921	0.812	0.853	**0.956**	**0.956**	0.941
enc06	0.982	0.979	0.988	0.982	0.978	0.926	0.984	**0.989**	0.979
enc07	0.980	0.982	**0.987**	0.968	0.982	0.964	0.979	**0.987**	**0.987**
enc08	0.108	0.610	0.272	0.273	0.387	0.364	**0.717**	0.229	0.636
enc09	0.426	**0.760**	0.715	0.296	0.619	0.475	0.656	0.698	0.729
enc10	0.824	0.966	**0.987**	0.750	0.933	0.809	0.804	0.969	0.971

RMSE									
Parameter	PSNR	SSIM	VIF	LSS	ESS	LFBVS	VSI-Canny	VSASW	Proposed
trad	0.250	0.208	0.207	0.250	0.264	0.248	0.249	0.211	**0.203**
trunc	0.510	0.536	0.246	0.308	0.316	0.293	0.287	0.365	**0.236**
iwind_ec	0.434	0.280	0.264	0.483	0.512	0.491	0.274	0.258	**0.212**
iwind_nec	0.435	**0.312**	0.334	0.502	0.493	0.375	0.370	**0.312**	0.336
res	0.480	0.412	0.234	**0.211**	0.501	0.487	0.198	0.311	0.258
enc01	0.509	0.346	0.328	0.363	0.716	0.903	0.393	0.418	**0.319**
enc02	0.424	0.516	**0.410**	0.760	0.749	0.856	0.762	0.517	0.422
enc03	0.481	0.533	0.326	0.442	0.383	0.412	0.324	**0.312**	0.423
enc04	0.526	0.573	0.442	0.531	0.473	0.459	0.722	0.419	**0.417**
enc05	0.520	0.420	0.402	0.508	0.529	0.513	**0.385**	0.402	0.393
enc06	0.363	0.391	0.362	0.358	0.371	0.402	0.345	**0.339**	0.352
enc07	0.386	0.366	0.367	0.483	0.375	0.396	0.391	0.415	**0.336**
enc08	0.428	0.341	0.796	0.414	0.369	0.483	**0.230**	0.418	0.325
enc09	1.408	**1.011**	1.129	1.487	1.245	1.378	1.174	1.115	1.041
enc10	1.067	0.489	0.463	1.246	0.574	0.796	1.121	0.463	**0.353**

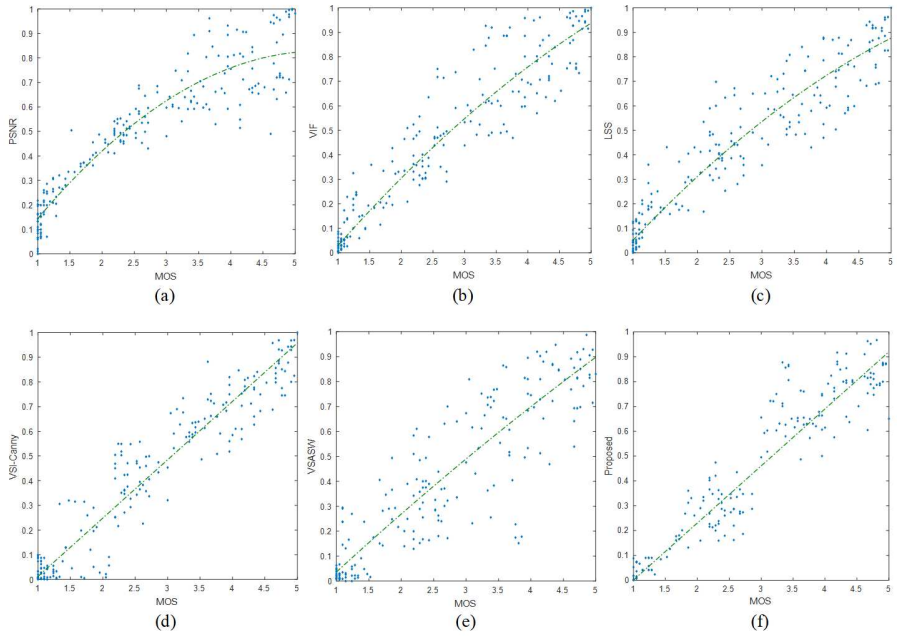


Fig. 7 Scatter plots between QuatHashing and state-of-the-art assessment methodologies on the IVC-Select Encrypt database. (a)PSNR, (b)VIF, (c)LSS, (d)VSI-Canny, (e)VSASW, (f)Proposed.

best performance on PLCC, and QuatHashing is in SRCC, KRCC and RMSE. In the high quality range, QuatHashing shows the best performance on $\max_{\mathrm{D}}$.

5.3.1 Performance Analysis of Different Encryption Algorithms

We also conducts performance tests on VSAs for different types of encrypted images on two test databases to more comprehensively assess the performance of VSAs. There are 15 different encryption types in both databases. Table 7 and Table 8 report the performance results for all encryption types present in the test database, respectively.

From Table 7, it can be seen that QuatHashing has better monotonic performance (SRCC and KRCC have higher values). More specifically, QuatHashing has the most SRCC optimal values (8 times) and KRCC optimal values (6 times), and this value is higher than other metrics. From Table 8, it can be found that QuatHashing still has the most PLCC optimal values (8 times) and RMSE optimal values (7 times).

6 Conclusion

In this paper, a reduced-reference visual security assessment method based on perceptual hashing with quaternion discrete wavelet transform, termed QuatHashing, is proposed for perceptually encrypted images. The reference image and the perceptually encrypted image are first reorganized into horizontal and vertical complementary sub-bands, which retain the distortion introduced by the encryption operation while

separating the directional information of the image. The complementary sub-bands of each direction are then modeled as a pure quaternion image and decomposed by a double-channel QDWT, from which the low-frequency structure feature, the texture feature and the multi-scale structural feature are extracted to describe the visual content remaining in the encrypted image from both directional and multi-scale perspectives. On this basis, a global attention sequence and a local feature sequence are constructed for each feature through strip-wise pooling, so that the visually sensitive regions contribute more to the final hash. Finally, the horizontal and vertical feature sequences are quantized and concatenated into the perceptual hash, and the Euclidean distance between the hashes of the reference and encrypted images is taken as the visual security score. Experiments on the IVC-SelectEncrypt and PEID databases show that QuatHashing achieves better performance than existing IQA and VSA methods, and the advantage is more obvious on encrypted images in the low and moderate quality ranges.

Declarations

Ethics Approval

Not applicable.

Funding

Not applicable.

References

[1] Liu, Y., Zhang, J., Han, D., Wu, P., Sun, Y., Moon, Y.S.: A multidimensional chaotic image encryption algorithm based on the region of interest. Multimedia Tools and Applications **79**, 17669–17705 (2020)

[2] Wan, W., Wang, J., Li, J., Meng, L., Sun, J., Zhang, H., Liu, J.: Pattern complexity-based jnd estimation for quantization watermarking. Pattern Recognition Letters **130**, 157–164 (2020). Image/Video Understanding and Analysis (IUVA)

[3] Wan, W., Wang, J., Zhang, Y., Li, J., Yu, H., Sun, J.: A comprehensive survey on robust image watermarking. Neurocomputing **488**, 226–247 (2022)

[4] Wang, R., GU, L., Li, Y., Wang, J., Wan, W., Sun, J., Cheung, S.-c.S.: Scia-gan: Robust image watermarking via spatial-channel interaction attention and feature preservation. Expert Systems with Applications **328**, 132890 (2026)

[5] Xiang, T., Yang, Y., Liu, H., Guo, S.: Visual security evaluation of perceptually encrypted images based on image importance. IEEE Transactions on Circuits and Systems for Video Technology **30**(11), 4129–4142 (2019)

[6] Lv, J.: Research on privacy and secure storage protection of personalized medical data based on hybrid encryption. EURASIP Journal on Information Security **2024**, 1 (2024)

[7] Hasan, M.K., Hasan, M.M., Al-Qirim, N., Abdullah, S.N.H.S., Islam, S., Razzaque, M.A.: Review on data privacy and security for iot-based multifunctional layers of cyber-physical systems in smart grids. Eurasip Journal on Information Security **2026**, 9 (2026)

[8] Xiong, L., Han, X., Yang, C.-N., Shi, Y.-Q.: Robust reversible watermarking in encrypted image with secure multi-party based on lightweight cryptography. IEEE Transactions on Circuits and Systems for Video Technology **32**(1), 75–91 (2021)

[9] Hafsa, A., Sghaier, A., Malek, J., Machhout, M.: Image encryption method based on improved ecc and modified aes algorithm. Multimedia Tools and Applications **80**, 19769–19801 (2021)

[10] Wang, R., Lingchen, G., Li, J., Wang, J., Sun, J., Wan, W.: Genran: Genfusion-guided reversible anonymization network for face privacy preserving. Information Fusion **121**, 103120 (2025)

[11] Rovida, L., Leporati, A.: Encrypted image classification with low memory footprint using fully homomorphic encryption. Cryptology ePrint Archive (2024)

[12] Zhao, Y., Yang, C., Tang, B.: Construction of perceptual security dataset for video selective encryption based on double-blind subjective experiment. In: 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE), pp. 30–35 (2021). IEEE

[13] Wen, W., Huang, M., Zhang, Y., Fang, Y., Zuo, Y.: Visual security index combining cnn and filter for perceptually encrypted light field images. ACM Transactions on Multimedia Computing, Communications and Applications **20**(1), 1–15 (2023)

[14] Xiong, J., Zhu, X., Yuan, J., Shi, R., Gao, H.: Perceptual visual security assessment by fusing local and global feature similarity. Computers & Electrical Engineering **91**, 107071 (2021)

[15] Yang, Y., Xiang, T., Lv, X., Guo, S., Zeng, T.: The illusion of visual security: Reconstructing perceptually encrypted images. IEEE Transactions on Circuits and Systems for Video Technology (2023)

[16] Qing-li, C., Guo, H., Hong-yin, Q.: A hybrid differential algorithm for image enhancement. In: 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), pp. 96–99 (2020). IEEE

[17] Kumar, V., Das, A.K., Sinha, D.: Statistical analysis of the unsw-nb15 dataset

for intrusion detection. In: Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019, pp. 279–294 (2020). Springer

[18] Shahna, K., Mohamed, A.: A novel image encryption scheme using both pixel level and bit level permutation with chaotic map. Applied Soft Computing **90**, 106162 (2020)

[19] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

[20] Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on image processing **15**(2), 430–444 (2006)

[21] Mao, Y., Wu, M.: Security evaluation for communication-friendly encryption of multimedia. In: 2004 International Conference on Image Processing, 2004. ICIP'04., vol. 1, pp. 569–572 (2004). IEEE

[22] Tong, L., Dai, F., Zhang, Y., Li, J.: Visual security evaluation for video encryption. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 835–838 (2010)

[23] Xiang, T., Guo, S., Li, X.: Perceptual visual security index based on edge and texture similarities. IEEE Transactions on Information Forensics and Security **11**(5), 951–963 (2016)

[24] Wu, Z., Zhang, K., Ren, Y., Li, J., Sun, J., Wan, W.: Visual security assessment via saliency-weighted structure and orientation similarity for selective encrypted images. Security and Communication Networks **2021**, 1–16 (2021)

[25] Du, L., Ho, A.T., Cong, R.: Perceptual hashing for image authentication: A survey. Signal Processing: Image Communication **81**, 115713 (2020)

[26] Yan, C., Gong, B., Wei, Y., Gao, Y.: Deep multi-view enhancement hashing for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(4), 1445–1451 (2020)

[27] Li, Y., Gemert, J.: Deep unsupervised image hashing by maximizing bit entropy. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2002–2010 (2021)

[28] Luo, X., Wang, H., Wu, D., Chen, C., Deng, M., Huang, J., Hua, X.-S.: A survey on deep hashing methods. ACM Transactions on Knowledge Discovery from Data **17**(1), 1–50 (2023)

[29] Gu, L., Shen, X., Sun, J., Liu, Y., Li, J., Li, Z., Cheung, S.-C.S., Wan, W.: Dual prototypes-based personalized federated adversarial cross-modal hashing. IEEE

Transactions on Circuits and Systems for Video Technology **35**(12), 12846–12860 (2025)

[30] Liang, X., Tang, Z., Xie, X., Wu, J., Zhang, X.: Robust and fast image hashing with two-dimensional pca. Multimedia Systems **27**(3), 389–401 (2021)

[31] Tang, Z., Huang, Z., Yao, H., Zhang, X., Chen, L., Yu, C.: Perceptual image hashing with weighted dwt features for reduced-reference image quality assessment. The Computer Journal **61**(11), 1695–1709 (2018)

[32] Hofbauer, H., Uhl, A.: Identifying deficits of visual security metrics for images. Signal Processing: Image Communication **46**, 60–75 (2016)

[33] Stütz, T., Pankajakshan, V., Autrusseau, F., Uhl, A., Hofbauer, H.: Subjective and objective quality assessment of transparently encrypted jpeg2000 images. In: Proceedings of the 12th ACM Workshop on Multimedia and Security, pp. 247–252 (2010)

[34] Pei, S.-C., Ding, J.-J., Chang, J.-H.: Efficient implementation of quaternion fourier transform, convolution, and correlation by 2-d complex fft. IEEE transactions on signal processing **49**(11), 2783–2797 (2001)

[35] Guo, S., Xiang, T., Li, X., Yang, Y.: Peid: A perceptually encrypted image database for visual security evaluation. IEEE Transactions on Information Forensics and Security **15**, 1151–1163 (2019)

[36] Liu, Q., Yuan, H., Hamzaoui, R., Su, H., Hou, J., Yang, H.: Reduced reference perceptual quality model with application to rate control for video-based point cloud compression. IEEE Transactions on Image Processing **30**, 6623–6636 (2021)

[37] Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3677–3686 (2020)

[38] Mahajan, P., Jakhetiya, V., Abrol, P., Lehana, P.K., Subudhi, B.N., Guntuku, S.C.: Perceptual quality evaluation of hazy natural images. IEEE Transactions on Industrial Informatics **17**(12), 8046–8056 (2021)

[39] Hodson, T.O.: Root mean square error (rmse) or mean absolute error (mae): When to use them or not. Geoscientific Model Development Discussions **2022**, 1–10 (2022)