Image Semantic Quality Evaluation Model for Human-Machine Hybrid Intelligence: A Gradient-based Uncertainty Calculation Method

Ziyan Yue[1], Senyang Lu[2], Hong Lu[3*]
[1]Key Laboratory of Educational Informatization for Nationalities (YNNU), Ministry of Education, Yunnan Normal University, Kunming 650500, China
[2]Faculty of Art and Communication, Kunming University of Science and Technology, Kunming 650500, China
[3]Academy of Fine Arts, Nanjing Xiaozhuang University, Nanjing 211171, China
E-mail: luhongeys@126.com
[*]Corresponding author

Keywords: gradient, uncertainty calculation, human-machine hybrid, semantic distortion, image quality evaluation

Received: March 20, 2024

With the advancement of human-machine hybrid intelligence technology, the importance of images in interaction becomes increasingly high. The accurate evaluation of image semantic quality becomes crucial. However, traditional evaluation models may be limited in this environment. New methods are needed to improve evaluation accuracy. Then, an evaluation model for gradient-based uncertainty calculation method was proposed. The study conducted semantic distortion perception analysis at two levels. Firstly, overall, the recognition ability was analyzed by analyzing the average recognition accuracy of the dataset. Secondly, recognition ability analysis was conducted based on the confidence level of a single sample. Experiments showed that machines had a higher tolerance for distortion compared to humans. However, these machines were weaker in terms of generalization and stability. The proposed method performed well on the complex CIFAR100 dataset, achieving the lowest FPR of 95%, the highest TPR of 528%, and the lowest error detection rate of 3.65%. In addition, the accuracy of the proposed framework reached 68.03%, which was significantly better than 59.83% for humans and 40.16% for machines. The results indicated its ability to effectively combine the advantages of different decision-makers. This study is expected to provide new ideas for image quality evaluation, improving the application performance and user experience of images in multiple fields.

Povzetek: Predlagan je model za ocenjevanje semantične kakovosti slik, ki temelji na gradientni metodi izračuna negotovosti, z namenom izboljšati interakcijo med človekom in strojem.

1 Introduction

With the rapid development of social media, e-commerce, and digital media, people interact with a large amount of image content every day, including uploading, sharing, searching, and shopping. It is necessary to accurately evaluate the quality and semantic content of images to filter out junk images, improve the relevance of search results, and recommend related products to provide a better user experience [1]. Traditional image quality assessment (IQA) methods mainly focus on pixel level image quality, such as noise, blur, and distortion [2]. However, these methods often fail to capture the semantic content of the image and cannot determine whether the image meets the user's needs or contains important information. The gradient-based uncertainty calculation method, as a commonly used technique in deep learning and machine learning, can be used to evaluate the uncertainty of models [3-4]. Based on this, a subjective dataset is created to evaluate semantic distortion in monitoring distortion scenarios. Confidence measures are used to analyze the recognition ability of humans and machines on a single sample. Finally, the above content is applied to human-machine joint decision-making. A decision-making framework is designed. The research aims to develop more accurate and intelligent methods to improve the performance of applications such as understanding, searching, and retrieving image content. The innovation of the research lies in providing a new method suitable for human-machine collaborative decision-making and introducing gradient uncertainty calculation to more accurately estimate image quality.

The research consists of five parts. Part 1 introduces the research background, problems, and solutions of the image semantic quality evaluation model for human-machine hybrid intelligence. Part 2 reviews the current research status of image semantic quality evaluation models based on human-machine hybrid intelligence. The existing difficulties and methodological shortcomings are summarized. Part 3 establishes a human-machine hybrid intelligent image semantic quality evaluation model based on gradient uncertainty. Part 4

evaluated the performance of the model through comparative experiments and efficiency validation. Part 5 summarizes the research methods and proposes the shortcomings of the methods as well as future research directions.

The application of human-machine hybrid intelligent systems is becoming increasingly prominent in computer vision. The evaluation of image semantic quality involves multiple fields such as image processing, computer vision, and deep learning. Ensuring the evaluation accuracy of image semantic quality is important for system performance in numerous applications such as autonomous driving, facial recognition, and safety monitoring. Traditional image quality evaluation methods often fail to meet the requirements of human-machine hybrid intelligent systems due to their neglect of the uncertainty of deep learning models. This may lead to misleading results in practical applications. Therefore, an urgent issue that needs to be addressed is to improve the performance of image semantic quality evaluation models for human-machine hybrid intelligence. Some scholars have conducted a series of studies on this topic. Sara U et al. proposed structured similarity index method and feature similarity index method to measure the structural and feature similarity between the restored object and the original object based on perceptual comparison. Experiments showed that this method provided perceptual and saliency-based errors more easily understood [5]. Jang et al. proposed an automatic crack evaluation technique based on deep learning, aiming to achieve high-quality crack evaluation by utilizing semantic segmentation technology to process images. The experimental results showed that the method achieved a high accuracy rate of 90.92% and a high recall rate of 97.47% [6]. Researchers such as Fu proposed an evaluation method that combined rules and semantic logic based on deep learning semantic evaluation, aiming to provide evaluation regularity and semantic decoding. The experimental results showed that this method had the ability to automatically evaluate regularity and semantics and exhibited higher validation [7]. Liu et al. proposed a video reconstruction and semantic quality evaluation method based on the characteristics of upstream streaming media, aiming to further improve the accuracy of semantic evaluation. Experiments showed that block compression sensing required less sensing or storage resources in the front-end, achieving a lightweight observation matrix and supporting block by block or parallel transmission [8].

On the other hand, gradient-based uncertainty calculation methods are widely developed and applied in science, engineering, and machine learning. Giraud J et al. proposed a workflow for integrating geological modeling uncertainty information to solve the geological uncertainty information being used for local constraints. This experiment showed that this method significantly reduced the uncertainty of interpretation [9]. Ouziala et al. proposed a method for detecting small-scale faults involving parameter uncertainty, aiming to ensure optimal detection performance by optimizing thresholds. This experiment showed that this method improved the sensitivity of residuals to small faults and ensured optimal early detection [10]. Puzyrev et al. proposed a deterministic gradient-based method aimed at solving least squares optimization problems in high-dimensional parameter spaces. This experiment showed that the method exhibited excellent performance in multiple aspects such as accuracy, generalization ability, and training cost [11]. Pevey et al. proposed a gradient optimization design method for nuclear reactor core components. This method was based on continuous and discrete material neutronics objectives, aiming to fully utilize gradient information for design optimization. This experiment indicated that the accompanying gradient calculation method had potential application prospects in nuclear system design optimization [12].

In summary, there has been some development in image semantic quality evaluation for human-machine mixing. However, there are still problems with low model generalization ability, high computational complexity, and a lack of deep semantic understanding. On the other hand, gradient-based uncertainty calculation methods can be used to estimate the uncertainty of model output, thereby improving the reliability and predictive ability of the model. Then, this study proposes a gradient-based uncertainty calculation oriented human-machine hybrid intelligent image semantic evaluation model. This model is expected to improve the accuracy of evaluation, promote human-machine collaboration, and enhance the interpretability of intelligent systems. The literature review classification is shown in Table 1.

Table 1: Literature review classification

Author	Method	Achieved goals	Disadvantage
Sara et.al [5]	Structured similarity index method and feature similarity index method	Measure structural and feature similarities between restored objects and original objects based on perceptual comparisons	From representation perspective, SSIM and FSIM are normalized, but MSE and PSNR are not.
Jang et.al [6]	Automatic crack assessment technology based on deep learning	Use semantic segmentation technology to process images to achieve high-quality crack assessment	Detection time is longer.

Fu et.al [7]	Evaluation method combining deep learning semantics and semantic logic	Provide evaluation regularity and semantic decodability	Pattern complexity increases decoding time.
Liu et.al [8]	Video reconstruction and semantic quality evaluation method based on the characteristics of upstream streaming media	Improve the accuracy of semantic evaluation	Detection perception accuracy is closely related to video quality.
Giraud et.al [9]	Workflow for integrating uncertainty information in geological modeling	Address issues where geological uncertainty information is used for local constraints	With all geological modeling, the model cannot account for geological units or faults that are not sampled by in-situ geological measurements, which can lead to biases in the final model.
Ouziala et.al [10]	A method for detecting micro-level faults involving parameter uncertainty	Ensure optimal detection performance by optimizing thresholds	The accuracy of residual error in detecting minor faults needs to be improved.
Puzyrev [11]	Deterministic gradient-based methods	Solve least squares optimization problems in high-dimensional parameter spaces	The inversion region has a significant impact on the results.
Pevey et.al [12]	A gradient optimization design method for nuclear reactor core components	Leverage gradient information for design optimization	Gradient-informed designs must scale as the dimensions of the design space increase.

2 Building an image semantic quality evaluation model on the ground of gradient uncertainty

A gradient-based uncertainty prediction method is proposed for out of distribution detection. In addition, a human-machine joint decision-making framework was designed for research. It combines the advantages of humans and machines in perceiving semantic distortion to improve decision accuracy.

2.1 Semantic distortion perception analysis on the ground of datasets

With the advancement of deep learning technology, machines are increasingly capable of semantic analysis of images. Machines include tasks such as object detection, positioning, and recognition. Nonetheless, distortion phenomena in images can negatively impact these analysis tasks. The specific effect is known as semantic distortion. Semantic distortion is different from traditional image quality distortion, which is not suitable for conventional image quality evaluation indicators. Therefore, research on this issue is particularly urgent. Semantic distortion is proposed for specific image semantic analysis tasks and needs to be explored in a specific application context [13]. The distortion can damage image quality. Therefore, IQA methods are needed to evaluate [14]. IQA methods are divided into subjective and objective categories. Subjective methods require human observers to evaluate, accurately but time-consuming and subject to interference. The performance of objective methods is usually measured by the similarity with subjective scores. The higher the similarity, the better the model performance. Objective image quality evaluation can be divided into three categories: full reference, semi-reference, and no reference [15]. The widely used objective quality evaluation methods currently include mean square error and structural similarity index. Mean square error is a measure of the average error at the pixel level of an image, and the calculation method for both is shown in equation (1).

$$\begin{cases} MSE = \dfrac{1}{WH}\sum_{j=1}^{H}\sum_{i=1}^{W}(I_{ref}(i,j) - I_{sts}(i,j)^2 \\ PSNR = 10\log\dfrac{D^2}{MSE} \end{cases} \tag{1}$$

In equation (1), I_{ref} represents the reference image, I_{sts} represents the distorted image, and D represents the range of pixel value dynamic transformation. MSE and $PSNR$ represent mean square error and structural similarity, respectively. Researchers have made efforts to improve the confidence score of deep learning models by conducting uncertainty predictions to reduce uncertainty. Therefore, the misleading high confidence predictions that may occur in models that do not conform to the distribution of training data can be addressed [16]. This is achieved by first classifying uncertainty and then dealing with it in a targeted manner. Figure 1 shows the classification of uncertainty.

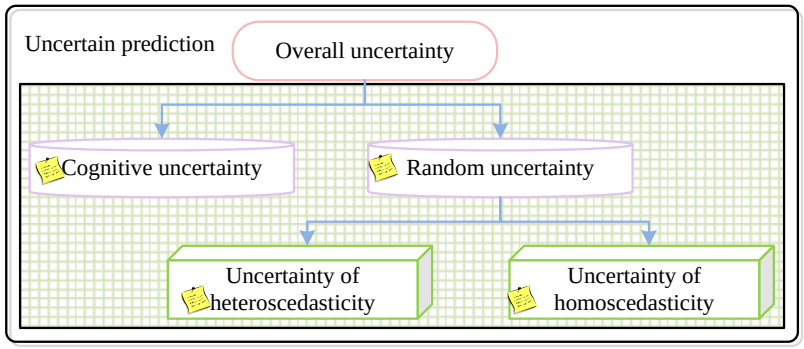


Figure 1: Classification of uncertainty

In Figure 1, cognitive uncertainty is caused by the uncertainty in the model parameter space. It can be reduced by increasing training data, usually caused by underfitting or dataset offset. After determining the uncertainty, this study needs to consider the perception differences among different populations, cultural backgrounds, and contexts. The semantic distortion perception in the dataset is further analyzed. When studying human semantic distortion perception, the first step is to select a reference image, then perform distortion processing on the image, followed by subjective experimental design, and finally eliminate outliers. In this study, Facenet is used for human face subset testing. Triplet is used to reduce intra-class spacing. The specific correlation loss function diagram is shown in Figure 2.

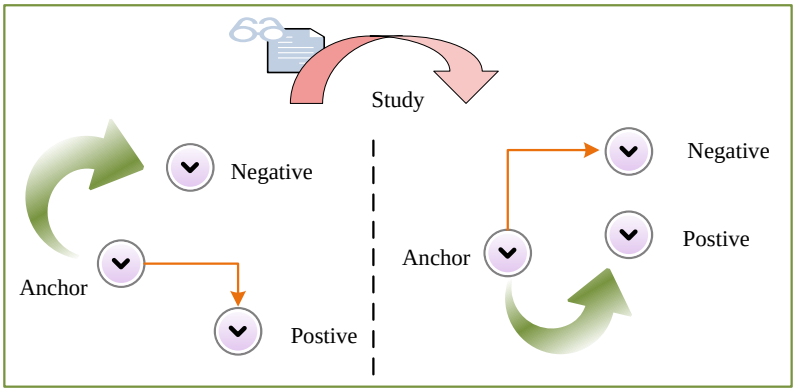


Figure 2: Schematic diagram of Triplet loss function in Facenet

In Figure 2, the distance between the anchor sample and all the same negative samples is greater than the distance between the anchor sample and all the same positive samples. The specific calculation expression is shown in equation (2).

$$\| f(a_i^a) - f(a_i^p) \|_2^2 + a < \| f(a_i^a) - f(a_i^n) \|_2^2 \\ \forall (f(a_i^a), f(a_i^p), f(a_i^n)) \in \Gamma \quad (2)$$

In equation (2), a represents the ideal distance maintained between positive and negative samples. a_i^a represents the anchor sample, a_i^p represents the positive sample, and a_i^n represents the negative sample. Γ represents the set of all possible triples in the training set. The loss function calculation method is shown in equation (3).

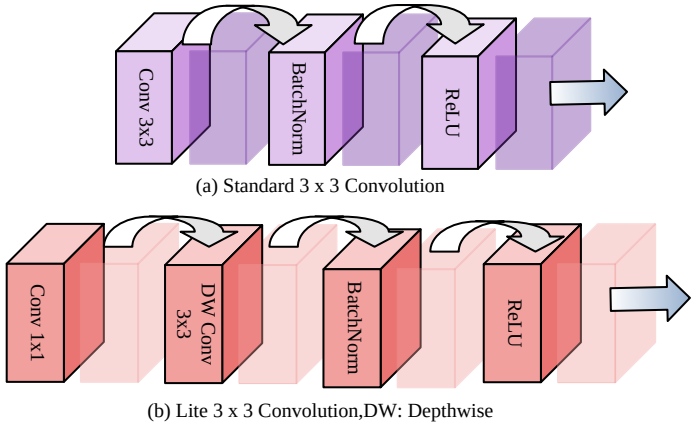
$$\tau = \sum_i^N [\| f(a_i^a) - f(a_i^p) \|_2^2 - \| f(a_i^a) - f(a_i^n) \|_2^2 + a] \quad (3)$$

In equation (3), τ represents the value of the loss function. The Omni-scale network (OSNet) is used to conduct the pedestrian subset test after the face subset test. The network references the image source Market-1501 data set. The test accuracy reaches 93%. OSNet is a full-scale learning structure for person re-identification, which contains a residual module composed of multiple convolution streams. Each convolution stream is responsible for feature detection at a certain scale. In addition, a new unified aggregation gate is introduced in the network to dynamically fuse multi-scale features with input-related channel weights. Its specific calculation formula is shown in equation (4).

$$y = x + \tilde{x}, s.t. \tilde{x} = F(x) \quad (4)$$

In equation (4), x represents the given input value, F represents the mapping function, and $\tilde{x}$ represents the learning residual. Figure 3 shows the convolutional

comparison of Lite3*3 in OSNet, which is a standard 3*3 convolutional kernel, after introducing a row aggregation gate.



(a) Standard 3 x 3 Convolution

(b) Lite 3 x 3 Convolution,DW: Depthwise

Figure 3: Comparison intention between standard 3*3 convolution and Lite3*3 convolution in ONet

In Figure 3, unlike the standard 3*3 convolution layer, the Lite3*3 convolution layer uses depthwise separable convolution to reduce the amount of network parameters. The scale of the feature is represented by an index to achieve multi-scale feature learning. All convolution streams are dynamically fused through a unified aggregation gate. The weights of different scales are dynamically adjusted according to the input samples. The residual expression that the network needs to learn is shown in equation (5).

$$\tilde{x} = \sum_{t=1}^{T} G(F^t(x)) \odot F^t(x) \tag{5}$$

In equation (5), $\odot$ represents the Hadamard product, and $G(F^t(x))$ represents the channel weight coefficient. After pedestrian recognition, the study further conducted license plate detection and recognition.

2.2 Analysis of single sample semantic distortion perception on the ground of gradient uncertainty calculation method

In Section 2.1, semantic distortion is evaluated by computing the recognition accuracy on a specific dataset. Although this method is simple to operate, it has some obvious limitations. For example, accuracy is a statistical result based on a large number of samples, which cannot reveal subtle differences between individual samples. The results of accuracy are easily affected by the characteristics of the selected data set. Then, the study proposes to introduce confidence as a new metric to measure semantic distortion. Confidence represents the probability of prediction correctness. Confidence not only reflects the strength of the recognition ability, but also can be directly calculated based on the model's prediction of the current input sample. Confidence provides the possibility of in-depth analysis of a single sample [17]. At the human level, confidence is defined by calculating the proportion of individuals in the population that correctly identify the sample. At the machine level, confidence is related to the uncertainty predictions of the deep learning model. Therefore, the experiment proposes a new gradient-based uncertainty prediction method specifically for outlier detection. In addition, the research proposes a joint decision-making framework that integrates human and machine perception. This method aims to utilize the complementary advantages of the two on different samples to improve the overall decision-making accuracy. The formula for calculating human confidence is shown in equation (6).

$$U_H = \frac{1}{|\Omega_H|} \sum_{i \in \Omega_H} 1(h_i = y) \tag{6}$$

In equation (6), Ω_H represents the human subject population, y represents the label corresponding to the sample, and h_i represents the recognition result of the i-th human subject. This study uses a deep learning model to predict human confidence in different data types. The model is adjusted to output a scalar value representing human confidence. The training process uses random gradient descent and sets some hyperparameters. During data processing, data augmentation operations that may affect human confidence are avoided. The study uses human confidence in subjective datasets as the true score and the output of the model as the predicted value. The main evaluation indicator for model performance is the correlation between the true and predicted values, usually using Spearman rank correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC). SROCC is a nonlinear correlation coefficient used to measure the correlation between two variables, whose values are obtained by ranking the data. Specifically, the SROCC calculation method is shown in equation (7).

$$SROCC = 1 - \frac{6\sum_{i=1}^{N}(rank(x_i) - rank(y_i))^2}{N(M^2 - 1)} \tag{7}$$

In equation (7), $rank(x_i)$ represents the arrangement order of x in all sequences. PLCC is a widely used linear correlation coefficient used to measure the linear relationship between two sets of data. The calculation expression is shown in equation (8).

$$PLCC = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(x_i - \overline{x})^2}\sqrt{(y_i - \overline{y})^2}} \quad (8)$$

In equation (8), x_i represents the true score of the i th image, and y_i represents the test value of the i th image. $\overline{x}, \overline{y}$ represent the average values corresponding to both. In practical applications, this study needs to first perform nonlinear fitting between the test scores and the true values. In this experiment, a logistic regression model is used for fitting, as shown in equation (9).

$$y = \beta_1(0.5 - \frac{1}{1+e^{\beta_2(x-\beta_3)}}) \_ \beta_4 x + \beta_5 \quad (9)$$

In equation (9), x represents the predicted score before fitting. y represents the predicted score after fitting. The fitting parameters are represented by $\{\beta_i \mid i = 1,2,3,4,5\}$. A gradient-based uncertainty prediction method is proposed to address the high complexity and impracticability of deterministic prediction methods in machine semantic distortion perception. The study first observes the feature sparsity of out of distribution samples, and then further utilizes the Jacobian matrix to analyze the relationship between feature sparsity and gradient norm. Then, the network output is obtained based on the network input and the network linear layer, as shown in equation (10).

$$F(a) = \Phi_L(W_L\Phi_{m-1}(W_{L-1}...\Phi_1(W_{1a}))) \quad (10)$$

In equation (10), a represents the network input, $F(a)$ represents the output, $\{W_i \mid W_i = [W_{i,1...,W_i,d_i}]\}$ represents the network linear layer, and $\{\Phi_i \mid i = 1,...,L\}$ represents the nonlinear layer. The linear correction unit layer is shown in equation (11) by combining the relationship between gradient norm and network nonlinear layer analysis feature sparsity.

$$\Phi_i(z) = \Phi_{\mathrm{ReU}L}(z) = \max(0, z), i = 1,...,L-1 \quad (11)$$

In equation (11), ReUL represents the activation layer. The nonlinear layer derivative expression in the Jacobian matrix is shown in equation (12).

$$\frac{\partial h_i}{\partial W_i h_{i-1}} = diag[1(w_{i,j}h_i(x) > 0)], j = 1,...,d_i \quad (12)$$

In equation (12), in backpropagation, the zero value of the Jacobian matrix is positively correlated with the sparse matrix output by the ReLU layer. The gradient norm is negatively correlated with the sparse matrix output [18]. Usually, backpropagation requires label data to calculate the gradient of the loss function, but no labels are available during the testing phase. To address this issue, the study considers introducing a loss function for gradient retrieval in the unlabeled case. Firstly, this study perturbs the output to a small amplitude of ε. Then, a new loss function is introduced, as shown in equation (13).

$$\begin{cases} \tilde{F} = (1+\varepsilon)F \\ \tilde{\tilde{\Gamma}}_W = W \cdot (\tilde{F} - F) = \varepsilon W \cdot F \end{cases} \quad (13)$$

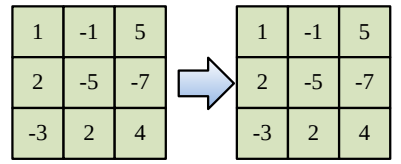
In equation (13), W represents the channel total coefficient corresponding to the output shape. This study designs a loss function for adaptive adjustment of input samples, especially when dealing with "out of distribution" samples. The disturbance of the loss function directly affects the gradient size of backpropagation. Previous studies have shown that predicting probability distributions to some extent reflects uncertainty [19]. Therefore, the disturbance amplitude ε is not a uniform value for all samples, which is related to the network output F(x). The expression for calculating disturbance amplitude is shown in equation (14).

$$\varepsilon = A(F(x)) \quad (14)$$

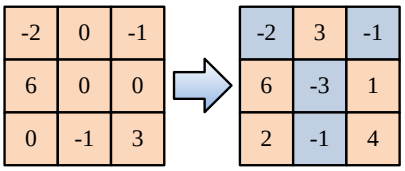
In equation (14), $A(F(x))$ represents the scalar statistic of the predicted probability distribution. This method is based on the correlation between input gradient and "out of distribution". That is, the larger the input gradient, the greater the sample difference, which is used to measure the "out of distribution" probability [20]. This method uses backpropagation to calculate the input gradient, as shown in equation (15).

$$U(x) = E_W[\varphi(\frac{\partial \tilde{\Gamma} W}{\partial x})] \quad (15)$$

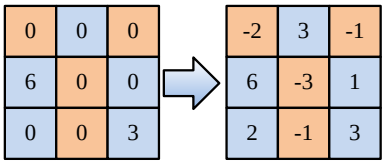
In equation (15), εW represents the calculated mean of the weight coefficients. This study focuses on the size of the gradient norm. However, the mutual cancellation of positive and negative gradients in backpropagation may reduce the gradient norm. This reduces the significance of analyzing the differences between in distribution and out of distribution samples [21]. To address this issue, this study uses a backpropagation optimization strategy. It is based on the directed backpropagation method, which separates positive and negative gradients by truncating the gradient flow. The specific schematic diagram is shown in Figure 4.



(a) Forward pass



(b) Backward pass:backpropagation



(c) Backward pass:guidedbackpropagation

Figure 4: Schematic diagram of guided backpropagation method

In Figure 4, the core idea of the directed backpropagation method is to truncate the negatively activated gradient in the ReLU layer, making it zero in backpropagation and no longer affecting gradient propagation. Deep learning technology develops rapidly in different fields. Due to the uncertainty of models, decisions in practical applications are not always reliable, especially in situations where high accuracy is required. Therefore, this study proposes a new human-machine joint decision-making framework, as shown in Figure 5.

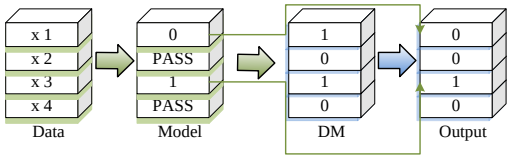
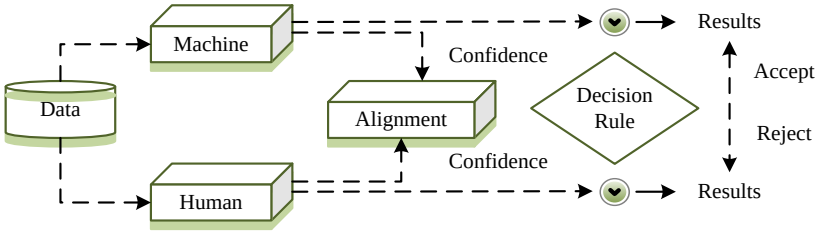
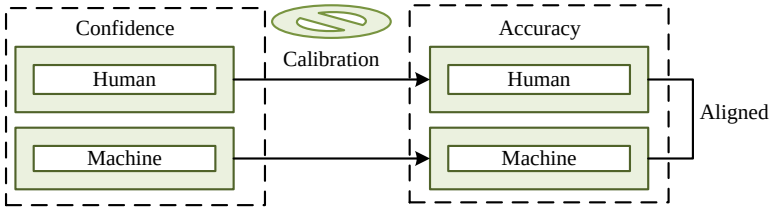


Figure 5: Schematic diagram of human-computer interactive decision-making process

In Figure 5, the proposed human-computer hybrid decision-making method includes a deep neural network and an external decision maker (DM), which is usually a human. The decision-making flow is in the form of a cascade and is divided into two steps. First, the DNN model can give a prediction result or choose to reject it. If choosing to reject, the second step is carried out. Otherwise, the system outputs the model prediction result. Second, if choosing to reject, the system redirects the input to the external decision-maker for a second judgment and outputs the prediction result of the external decision-maker. Although this modeling method is simple, it can still describe a large part of the decision-making system including multiple decision-makers. To better control the decision-making process, the study proposes a "human-computer hybrid decision-making framework", which requires the confidence of human and deep neural network models to be calibrated in order to compare their confidence. The relevant framework and confidence are shown in Figure 6.



(a) Research the proposed resolution intention



(b) Confidence alignment diagram

Figure 6: The proposed resolution intention and confidence alignment diagram

In Figure 6, the study needs to calculate the confidence scores of the human and deep neural network models for each input sample separately, which are adjusted to the same measurement scale. Then, decision rules are designed to generate the final decision result based on these two confidence scores. It ensures that the

confidence and accuracy distributions of human and deep neural network models are as consistent as possible on specific datasets.

3 Performance verification of a human-machine hybrid intelligent image semantic quality evaluation model on the ground of gradient uncertainty calculation method

The aim of this study is to create a comprehensive monitoring scene dataset that includes faces, pedestrians, and license plates. Three different distortion methods: JPEG compression, BPG compression, and motion blur are considered. The main objective of the study is to require participants to identify target objects in distorted images while excluding abnormal data to ensure the reliability of the obtained data.

3.1 Verification of semantic distortion perception performance on the ground of gradient uncertainty and datasets

This study conducted facial and pedestrian recognition tasks. Participants needed to select images from a template library that match the faces or pedestrians in distorted images. For facial recognition, this study divided the images into 10 groups, including different hairstyles and genders, to exclude other interfering factors. For pedestrian recognition, this study divided the images into 8 groups based on the color of the clothing. As for the license plate recognition task, the study used a single choice experiment, requiring participants to input complete license plate numbers, including province abbreviations, letters, and numbers. It was only considered correct recognition when the input matched the actual license plate perfectly. Figure 7 shows the trend of the average recognition accuracy of human subjects and deep neural network models as the distortion increases under different tasks and distortions.

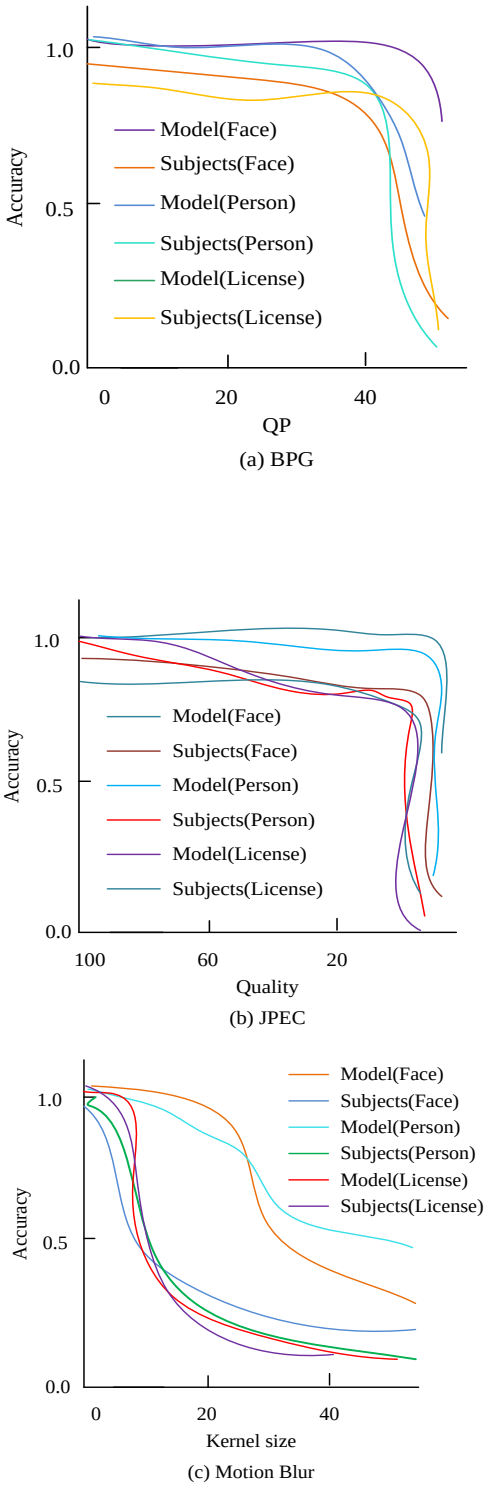
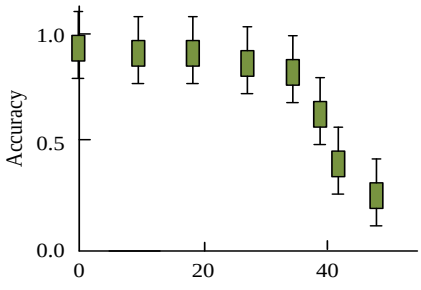
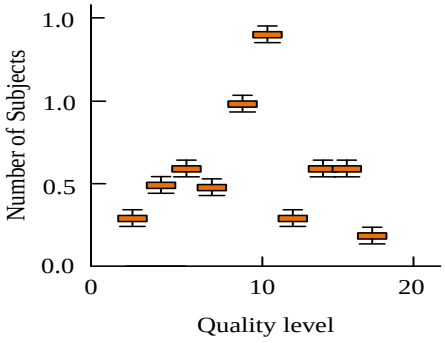


Figure 7: The average recognition accuracy of human and DNN models varies with distortion

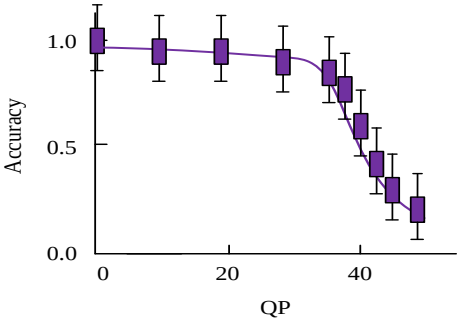
Figure 7(a) shows the results of image processing using the BPG compression method. As the image distortion value increased, the obtained image accuracy gradually changed from high accuracy to low accuracy. At the beginning of the experiment, the accuracy of the image reached 1.0. When the distortion value reached more than 40, the accuracy of the image was less than 0.5 and approached 0. Figure 7(b) shows the results of JPEG compression on image processing. When the image quality decreased, the accuracy of the image also shrank from the initial 1.0 and approached 0.0, not equal to 0.0. Figure 7(c) shows the results of image processing by three methods of motion blur. As the model kernel size became larger, the accuracy of the model on the image showed a trend of fluctuation. When the kernel size was 40, the accuracy of the image began to level off. It is worth noting that the DNN model performed better than humans in facial recognition and pedestrian recognition tasks under the three processing methods, especially in cases of severe image distortion. However, in the task of license plate recognition, the advantage of DNN model was relatively small. This indicates that the DNN model is more robust in dealing with image distortion. Figure 8 shows the differences in recognition accuracy among some different human participants.



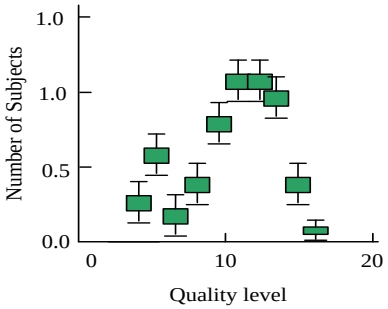
(a) Face BPG



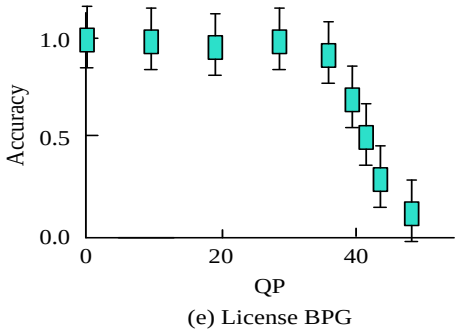
(b) Face BPG



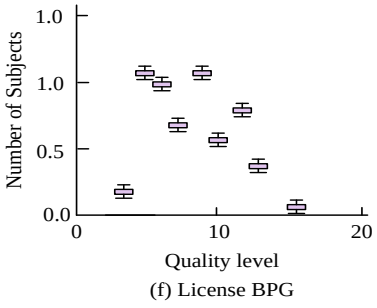
(c) Person BPG



(d) Person BPG



(e) License BPG



(f) License BPG

Figure 8: Human individual recognition accuracy curve and subjective recognition threshold distribution graph

Figures 8 (a), 8 (c), and 8 (e) represent the average, maximum, and minimum recognition accuracy of human individuals under QP. Figures 8 (b), 8 (d), and 8 (f) represent histograms of the distribution of subjective recognition thresholds for human individuals. Under the same level of distortion, the recognition accuracy of different individuals varied greatly, indicating that different individuals had different impacts when facing image distortion. In summary, machines are more robust

in dealing with image distortion compared to humans. However, they may be relatively weak in terms of generalization and stability. In addition, there are significant differences between human individuals.

3.2 Verification of semantic distortion perception performance on the ground of gradient uncertainty and single sample

The relationship between out of distribution samples and gradients was analyzed to validate the design based on gradient uncertainty. Firstly, starting from the sparsity of ReLU output features, the study investigated the correlation between feature sparsity and out of distribution samples. Subsequently, the connection between feature sparsity and gradient was established through the network Jacobian matrix. The experiment used CIFAR-10 and CIFAR-100 as in distribution datasets. TinyImageNet, LSUN, and iSUN were used as out of distribution datasets. The evaluation adopted indicators such as FPRat95% TPR, Detection Error, and AUROC. Table 1 shows the performance comparison between the proposed uncertainty prediction method and the complex method.

Table 2: Performance comparison of the proposed uncertainty prediction method with that of current high-complexity methods

Method	Knowledge	Complexity	Detection error	AUROCT	FPR@ 95%TPR
ODIN	OOD Val	Black-box	24	91	44
Malahanobis	IND Val	White-Box	7	97	13
Ours	None	White-Box	3	96	5
Margin-based ensemble	IND Train+00D Val	Retraining	8	97	16
Outlier exposure	OOD Val	Retraining	/	83	57

In Table 1, the study further compared the proposed method with the more complex DenseNet model, using CIFAR-100 as the in-distribution dataset and LSUN (r) as the out of distribution dataset. On the more challenging CIFAR-100 dataset, the studied method achieved the lowest FPRat95% TPR and detection error rate, which performed best among all methods. The experiment used confidence interval analysis to quantify the uncertainty of the evaluation results to enhance the credibility of the results. Specifically, the experiment calculated a 95% confidence interval for each performance indicator, ensuring the consistency and reliability of the evaluation experimental results. Then, the performance of the model was robust even under different experimental conditions. The method was applied to different data sets for training and testing, which effectively enhanced the generalization ability of the model and the repeatability of experiments. Figure 9 shows the complexity comparison of the proposed method with methods such as ONID, Mahalanobis, Softmax, etc.

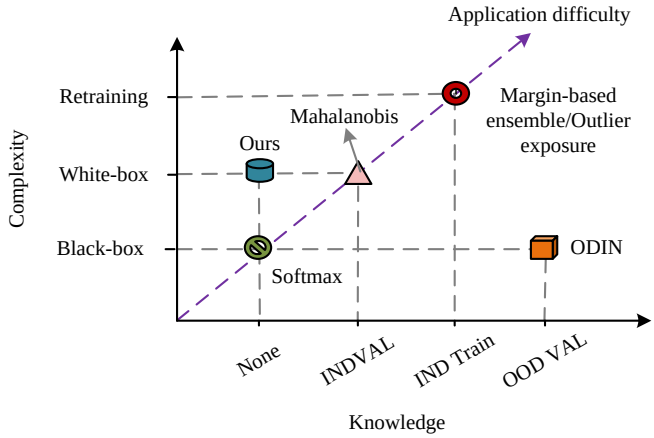


Figure 9: Complexity comparison between the proposed method and other methods

According to Figure 9, the proposed method achieved similar or even better performance while significantly reducing complexity compared with the state-of-the-art method. In addition, the study further used the constructed semantic dataset to simulate human predictions. The study randomly selected the predicted results of 20 human subjects as human predictions in actual scenarios. To eliminate the impact of randomness, each image was tested multiple times and the average value was taken as the final result. Table 2 shows the best

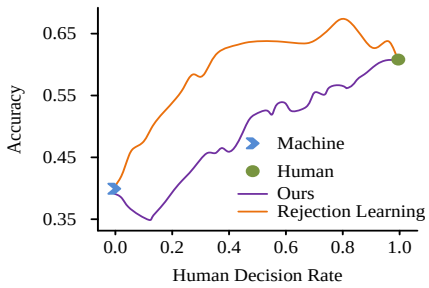
performance comparison between the proposed method and the refusal learning framework.

Table 3: The optimal performance comparison of the proposed framework and the rejection learning framework

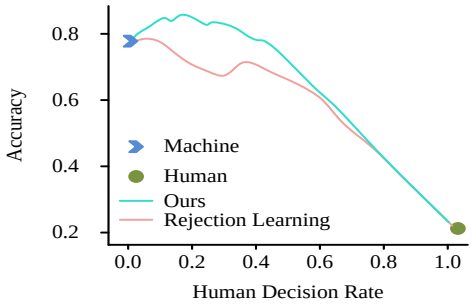
Method	Framework	Effectiveness of human decision	Accuracy	Human decision rate
License	Human	/	24.2	/
	Ours	54	86.6	19
	Rejection learning	49	78.3	5
	Machine	/	75.8	/
Person	Human	/	59.8	/
	Ours	34	68.0	81
	Rejection learning	19	59.8	100
	Machine	/	40.1	/

In Table 2, the optimal accuracy of the framework reached 68.03% when humans were superior to machines. The result was significantly higher than the accuracy of 59.83% for humans and 40.16% for machines. This indicates that the framework can effectively combine the advantages of different decision-makers. In contrast, the optimal accuracy of rejecting learning frameworks was 59.84%, which was only slightly higher than the accuracy of 59.83% in humans. This indicates that its performance limit does not exceed the performance of a single human or machine decision-maker, which cannot reflect the advantages of human-machine collaboration.

This study further conducted accuracy analysis. Figure 10 shows the accuracy of the proposed framework and refusal learning framework under different thresholds.



(a) Human superiority over machines



(b) Machines are superior to humans

Figure 10: Accuracy rates of the proposed framework and rejection learning framework under different thresholds.

Figure 10(a) shows the changes in human superiority over machines. The optimal accuracy of this framework was significantly higher than human accuracy and machine accuracy. The accuracy of this framework was always better than that of the rejection learning framework under the same ratio of human judgments. Figure 10(b) shows the curve change of machines outperforming humans. Since the human accuracy was lower than the proposed model, the performance of the framework might inevitably show an overall downward trend as the proportion of human decisions increased.

However, the accuracy of the proposed framework was always better than the rejection learning framework. Nonetheless, the framework still showed significant performance improvements compared to machines, a level that rejection learning frameworks could not achieve. Finally, the model was applied to the quality evaluation of a certain graphic semantics to analyze the impact of different factors on the accuracy of the model. The results are shown in Figure 11.

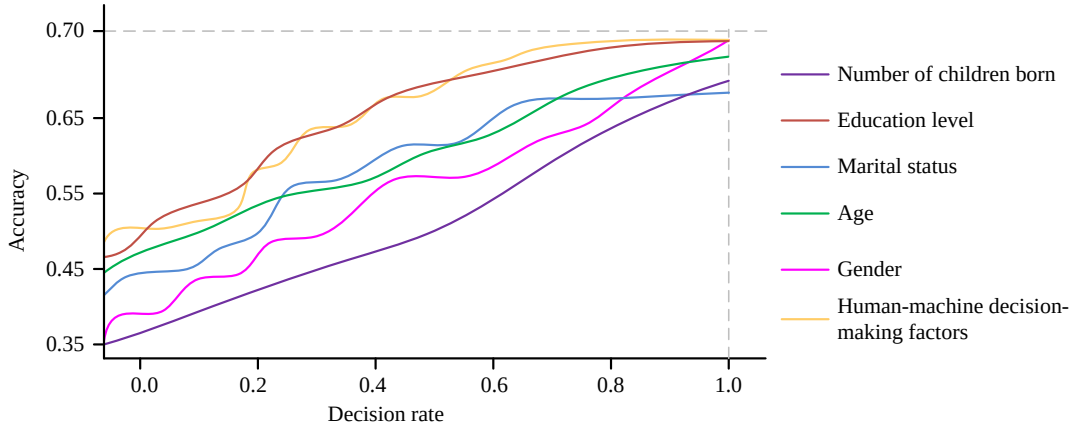


Figure 11: Impact of human factors

In Figure 11, as the decision-making rate of different factors increased, the accuracy of the model under the influence of the corresponding factors also showed different growth rates. When the decision rate reached 1.0, the model accuracy rate under the influence of all factors had the highest value. The education level between different statistical populations had the greatest impact on the image semantic quality evaluation of accuracy. The children born to different statistical populations had the greatest impact. The impact of quantity on accuracy was smaller than other factors. This may be because human beings' education level directly affects human beings' understanding and discrimination of different things and changes their views on things. There was overlap between the impact of human-machine decision-making factors on accuracy and human factors. The final model accuracy was less than 0.70. However, the connection between human-machine decision-making factors and human factors was very close. The accuracy obtained was higher than that under the influence of human factors. This can also be directly shown from the accuracy obtained. Therefore, human-machine decision-making factors can be used to evaluate the semantic quality of images.

4 Discussion and conclusion

4.1 Discussion

The image semantic quality evaluation model based on the proposed gradient-based uncertainty calculation method was tested in different scenarios. The performance was compared with different methods. The proposed method showed superior performance than traditional methods in key indicators such as accuracy and distortion recognition perception. The proposed model showed stronger robustness, especially in the face of complex background and noise interference. When humans were better than machines, the best accuracy of the proposed model framework reached 68.03%, which was significantly higher than the accuracy of humans of 59.83% and the accuracy of machines of 40.16%. This framework could effectively combine different decisions with the advantage of the person. In contrast, the best accuracy of the rejection learning framework was 59.84%, which was only slightly higher than the human accuracy of 59.83%. Its performance upper limit did not exceed the performance of a single human or machine decision-maker and could not reflect the advantages of the human-machine collaboration. For example, the automatic IQA model based on hybrid deep neural network was proposed by Chan K et al. Although the average correlation of the model reached 0.57, the image accuracy was not as good as the proposed model [22]. In addition, although the IQA algorithm based on the HCL framework and NR-IQA proposed by Wang et al. had strong generalization capabilities, there were still certain challenges in extracting distorted image information [23].

The performance difference may be mainly due to the fact that the proposed model is better able to handle blurred and discontinuous regions in the image by introducing gradient-based uncertainty calculation, thereby improving the accuracy of the evaluation. In addition, the computational efficiency of the proposed model was optimized. Compared with traditional methods, the proposed model reduced computational time while maintaining high accuracy, which had important practical significance in real-time application scenarios. The proposed model provides new contributions to image semantic quality assessment, especially in terms of the rationality of using confidence as a measure of semantic distortion. The resulting model is applied in a human-machine joint decision-making framework and achieves superior performance.

4.2 Conclusion

With the explosive growth of digital media content, automated image quality evaluation is important for content management such as automatic sorting, filtering, and recommendation systems. It is worth noting that the current semantic evaluation models have problems such as weak noise resistance, insufficient diversity and universality. This study selected monitoring scenarios and

selected three common objects: faces, pedestrians, and license plates to further improve the accuracy of image semantic quality evaluation. Three common distortion types: JPEG compression, BPG compression, and motion blur to test the accuracy of human recognition of these objects were selected. Then, a subjective perception database of semantic distortion was built. In addition to recognition accuracy, this study also introduced confidence to measure the recognition ability of human or deep neural network models for individual samples. This is to provide a deeper analysis of the perceptual effects of semantic distortion. The experimental results showed that machines were more robust in distortion compared to humans, but performed poorly in generalization and stability. The study adopted fine-grained semantic object classification, which meant that local detail features were more crucial, explaining why machines were more robust than humans. The research method achieved an optimal accuracy of 68.03%, significantly higher than human accuracy of 59.83% and machine accuracy of 40.16%. This indicates that the proposed method can effectively combine the advantages of different decision-makers. This study may lack in-depth research and evaluation of user subjective experience. Understanding users' expectations and preferences for image quality can provide important references for model improvement. Future research can enhance the design of user surveys and subjective evaluations.

Fundings

The research is supported by Key Project of Science and Technology Research of Ministry of Education: Research on Intelligent Network Teaching Model Based on Ontology and Agent (No.: 210210), National Natural Science Foundation of China “Research on ontology correction”(No.: 60903131), 2023 Innovation and Entrepreneurship Research Fund Project of Yunnan Normal University "Research on the Construction of Normal College Students' Intelligent Educational Literacy Evaluation Index System for Improving Employment Competence".

References

[1] H. T. Wang, L. Wang, F. Q. Lai, and J. Y. Zhang, “Investigation of image segmentation effect on the accuracy of reconstructed digital core models of coquina carbonate,” Applied Geophysics, vol. 17, no. 4, pp. 501-512, 2020. https://doi.org/10.1007/s11770-020-0846-2

[2] S. Zhao, P. Wang, Q. Cao, H. Song, and W. Li, “Weakly supervised salient object detection based on image semantics,” Journal of Computer-Aided Design & Computer Graphics, vol. 33, no. 2, pp. 270-277, 2021. https://doi.org/10.3724/SP.J.1089.2021.18318

[3] A. Williams, “Human-centric functional computing as an approach to human-like computation,” Artificial Intelligence and Applications, vol. 1, no. 2, pp. 118-137, 2023. https://doi.org/10.47852/bonviewAIA2202331

[4] F. Ecer, and E. Aycin, “Novel comprehensive MEREC weighting-based score aggregation model for measuring innovation performance: The case of G7 countries,” Informatica, vol. 34, no. 1, pp. 53-83, 2023. https://doi.org/10.15388/22-INFOR494

[5] U. Sara, M. Akter, and M. S. Uddin, “Image quality assessment through FSIM, SSIM, MSE and PSNR-A comparative study,” Computer and Communication (English), vol. 7, no. 3, pp. 8-18, 2019. https://doi.org/10.4236/JCC.2019.73002

[6] K. Jang, Y. K. An, B. Kim, and S. Cho, “Automated crack evaluation of a high-rise bridge pier using a ring-type climbing robot,” Computer-Aided Civil and Infrastructure Engineering, vol. 36, no. 1, pp. 14-29, 2021. https://doi.org/10.1111/mice.12550

[7] K. Fu, Y. Zhang, and X. Lin, “The automatic evaluation of regularity and semantic decodability in wallpaper decorative patterns,” Perception, vol. 48, no. 8, pp. 731-751, 2019. https://doi.org/10.1177/0301006619862142

[8] H. Liu, R. Huang, and H. Yuan, “Survey on compressive sensing video stream for uplink streaming media,” Journal of Image and Graphics, vol. 26, no. 7, pp. 1545-1557, 2021. https://doi.org/10.11834/jig.200487

[9] J. Giraud, M. Lindsay, V. Ogarko, M. Jessell, and E. Pakyuz-Charrier, “Integration of geoscientific uncertainty into geophysical inversion by means of local gradient regularization,” Solid Earth, vol. 10, no. 1, pp. 193-210, 2019. https://doi.org/10.5194/se-10-193-2019

[10] M. Ouziala, Y. Touati, S. Berrezouane D. Benazzouz, and B. Ouldbouamama, “Optimized fault detection using bond graph in linear fractional transformation form,” Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, vol. 235, no. 8, pp. 1460-1471, 2021. https://doi.org/10.1177/0959651820985617

[11] V. Puzyrev, “Deep learning electromagnetic inversion with convolutional neural networks,” Geophysical Journal International, vol. 218, no. 2, pp. 817-832, 2019. https://doi.org/10.1093/gji/ggz204

[12] J. Pevey, B. Hiscox, A. Williams, O. Chvala, V. Sobes, and J. W. Hines, “Gradient-informed design optimization of select nuclear systems,” Nuclear Science and Engineering: The Journal of the American Nuclear Society, vol. 196, no. 12, pp. 1559-1571, 2022. https://doi.org/10.1080/00295639.2021.1987133

[13] X. Li, S. Li, S. Liu, and D. He, “A malicious webpage detection algorithm based on image semantics,” Traitement du Signal, vol. 37, no. 1, pp. 113-118, 2020. https://doi.org/10.18280/ts.370115

[14] J. Wu, J. Zeng, W. Dong, G. Shi, and W. Lin, “Blind

image quality assessment with hierarchy: Degradation from local structure to deep semantics," Journal of Visual Communication and Image Representation, vol. 58, pp. 353-362, 2019. https://doi.org/10.1016/j.jvcir.2018.12.005

[15] Z. Jin, D. Yu, Z. Yuan, and L. Yu, "MCIBI++: soft mining contextual information beyond image for semantic segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 5, pp. 5988-6005, 2023. https://doi.org/10.48550/arXiv.2209.04471

[16] F. Liu, M. Huang, W. Pedrycz, and H. Zhao, "Group decision making based on flexibility degree of fuzzy numbers under a confidence level," IEEE Transactions on Fuzzy Systems: A Publication of the IEEE Neural Networks Council, vol. 29, no. 6, pp. 1640-1653, 2021. https://doi.org/10.1109/TFUZZ.2020.2983663

[17] Y. Fang, B. Luo, and T. Zhao, "ST-SIGMA: Spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting," CAAI Transactions on Intelligence Technology, vol. 7, no. 4, pp. 744-757, 2022. https://doi.org/10.1049/cit2.12145

[18] X. Xia, X. He, and L. Feng, "Semantic translation of face image with limited pixels for simulated prosthetic vision," Information Sciences: An International Journal, vol. 609, no. 2, pp. 507-532, 2022. https://doi.org/10.1016/j.ins.2022.07.094

[19] K. Liu, Z. Ye, H. Guo, D. Cao, L. Chen, and F. Y. Wang, "FISS GAN: A generative adversarial network for foggy image semantic segmentation," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 8, pp. 1428-1439, 2021. https://doi.org/10.1109/JAS.2021.1004057

[20] K. Shimoyama, and S. Kawai, "A kriging-based dynamic adaptive sampling method for uncertainty quantification," Transactions of the Japan Society for Aeronautical and Space Sciences, vol. 62, no. 3, pp. 137-150, 2019. https://doi.org/10.2322/tjsass.62.137

[21] V. Puzyrev, "Deep learning electromagnetic inversion with convolutional neural networks," Geophysical Journal International, vol. 218, no. 2, pp. 817–832, 2019. https://doi.org/10.1093/gjiz204

[22] K. Y. Chan, H. K. Lam, and H. Jiang, "A genetic programming-based convolutional neural network for image quality evaluations," Neural Computing and Applications, vol. 34, no. 18, pp. 15409-15427, 2022. https://doi.org/10.1007/s00521-022-07218-0

[23] J. Wang, Z. Chen, C. Yuan, B. Li, W. Ma, and W. Hu, "Hierarchical curriculum learning for no-reference image quality assessment," International Journal of Computer Vision, vol. 131, no. 11, pp. 3074-3093, 2023. https://doi.org/10.1007/s11263-023-01851-5