CIT. Journal of Computing and Information Technology, Vol. 31, No. 1, March 2023, 21–37
doi: 10.20532/cit.2023.1005695

A Visual Cortex-Attentive Deep Convolutional Neural Network for Digital Image Design

Lei Zheng

Shanghai Lida University, Shanghai, China

With the proliferation of advanced visualization techniques in visual communication, enhancing digital image quality remains a persistent challenge. This study presents a sophisticated Convolutional Neural Network (CNN) model to optimize image processing. The model incorporates a multi-stage architecture attentive to biological visual pathways. Inter-subnetwork connections enable integrated feature learning, guided by adaptive weighting of luminance, color, orientation, and edge maps. Spatial and channel attention modules further enrich feature interplay. When evaluated on the LIVE 3D Phase dataset, the approach demonstrates marked improvements, with saliency maps closely mirroring human visual perception. Pearson Correlation Coefficient and Histogram Intersection metrics exceed conventional models, at 0.6486 and 0.7074, respectively. Testing across distortion types reveals strong agreement with subjective rankings, confirming the model's effectiveness. By combining automated feature extraction with insights from visual cortex mechanisms, this bio-inspired CNN framework significantly enhances image optimization and quality. The scalable approach provides a foundation for next-generation computer vision and machine learning applications.

ACM CCS (2012) Classification: Computing methodologies → Machine learning → Machine learning approaches → Neural networks

Keywords: HSV, CNN, visual communication, bilinear combination

1. Introduction

Digital image technology is a transformative field that has reshaped how we capture, store, and manipulate visual information. It emerged in the late 20th century as a departure from analog imaging methods, utilizing pixels to create images with resolutions determined by pixel count. Digital cameras, powered by image sensors like CCD and CMOS, capture and convert light into digital data, while software tools enable image processing, enhancing quality and correcting imperfections. Images are stored in formats like JPEG, PNG, and TIFF, using compression algorithms to balance quality and file size. These images are displayed on various devices with evolving display technologies. This technology finds applications in photography, medicine, computer vision, and entertainment, and its future promises further innovations through AI and computational photography.

In step with unceasing scientific and technological innovation, three-dimensional images have begun to prominently feature in the public's purview, finding wide-ranging utility in realms such as film, design, remote sensing, and beyond [1–3]. In the context of specific applications, Convolutional Neural Networks (CNNs) not only afford a more precise reflection of image processing prowess but also exhibit exemplary adaptability [4]. The training of CNNs is inherently straightforward; amidst the treatment of distorted images, network inputs predominantly consist of multi-dimensional image inputs. Consequently, images can be directly introduced into the network, eliminating the need, as in conventional recognition algorithms, for subsequent data extraction from the image.

Moreover, within the ambit of CNNs, the commendable phenomenon of weight sharing prevails, allowing judicious regulation of training parameters through computer vision algorithms.

This enables a prudent reduction of parameter standards, endowing us with the twin virtues of finely tuned capacity management and potent aberration processing, thereby guaranteeing the ultimate level of image generalization [5–6]. Under these circumstances, image information can harmoniously align with CNN computations, necessitating the sole parameterization of CNN's final resolution and the digitized image to ensure parameter congruence, thus satisfying utilization prerequisites. Furthermore, the amalgamation of feature vectors lays bare the extent of monocular view distortion, thereby efficaciously reflecting image content quality [7]. While stereo images diverge from their planar counterparts, the perceptual quality model takes on heightened complexity. The quality score of the encompassing stereo image pair derives not merely from the content quality of the left and right views. Beyond the distinct role of binocular signals, the impact of depth information on stereo image quality evaluation demands due consideration. As a result of disparate horizontal eye positions, the retinal portrayal of a shared object in both eyes exhibits minute disparities. This binocular parallax emerges as a pivotal determinant, furnishing depth information to human visual perception.

Owing to inherent inter-individual disparities, the perception of a shared image begets divergent subjective experiences, attributed to the mutable physiological architecture of the Human Visual System (HVS). The sway of psychological constituents within the visual system bears inexorable relevance. In juxtaposition with 2D image quality assessment, the scrutiny of stereo images entails a broader panorama, encompassing the ramifications of depth information and the perceptual attributes of the HVS. As it stands, prevailing algorithms for appraising stereo image quality suffer from a paucity of robust predictive efficacy. This is predominantly attributable to the embryonic state of biological theories underpinning human visual system research. The existing shallow models remain ill-equipped to simulate the intricate information processing mechanisms coursing through the visual cortex [8–9]. Leveraging image enhancement techniques serves to augment the gamut of image greyscale, culminating in the assured attainment of contour lucidity and heightened contrast. Concomitantly, these techniques bolster the divergence in greyscale between the target subject and the backdrop of interference, achieving a pronounced partition and judicious manipulation of the target within this spectrum. The employment of existing image enhancement methodologies engenders a perceptible elevation in the visual quality of images, constituting an overarching avenue in image processing [10]. Simultaneously, the crux of this image enhancement lies in the elevation of the comprehensive visual impact of image presentation. Thus, the transmutation of image content into a format conducive to facile human or machine comprehension and analysis emerges as a requisite.

Regrettably, existing algorithms solely direct their focus towards the binocular visual attributes ingrained within the confines of the primary visual cortex, offering no investigatory probe into the province of alternate visual regions. Contemporary inquiries within the realm of visual analytics predominantly gravitate towards the Convolutional Neural Network (CNN) domain, entailing the orchestration of network architecture, channel allocation, convolutional kernel dimensions, and the like. Yet, these pursuits remain bereft of the systematic compass and reinforcement provided by the theory of visual neural mechanisms. Furthermore, although three-dimensional (3D) images, in comparison to their two-dimensional (2D) counterparts, promise an immersive user experience, the convoluted realm of image acquisition, coding, transmission, and storage introduces a medley of noise—attributable to environmental constraints and hardware limitations—that invariably mars the integrity of the original images.

In light of these circumstances, the principal contribution of this study emerges as follows:

1. The design of an interactive CNN model, underpinned by a multi-stage framework embellished with an attention mechanism. This augmentation entails the integration of a convolutional block attention module and a self-channel interaction module into the conventional CNN architecture. This innovative augmentation serves to forge spatial information correlations and foster feature channel interactions, thereby amplifying the model's efficacy in image optimization.

2. Amidst the terrain of visual feature extraction, a novel attention model for feature segmentation takes center stage. This avant-garde model incorporates supplementary edge features and orchestrates feature integration, leveraging a meticulously optimized weight distribution relationship. This strategic maneuver culminates in an ameliorated precision of image extraction.

2. Related Works

Computer vision orchestrates the acquisition, analysis, comprehension, and decision-making of intricate scenes through the conduit of hardware devices such as computers and the prism of pattern recognition technology. This interdisciplinary pursuit traverses diverse domains encompassing mathematics, neuroinformatics, computer science, and beyond. Its quintessential objective resides in attaining parity with, or potentially exceeding, the echelons of human visual prowess. Scholars have sequentially put forth an array of CNN-based models tailored for contour detection. These endeavors predominantly draw inspiration from prevailing classification network models, such as VGG [11] and ResNet [12], thereby underscoring the endeavor to refine decoding network architectures. Consequently, the contour attributes harvested via CNNs evolve along a spectrum ranging from fine-grained to coarse-grained, from local to global. This metamorphosis bequeaths the network with an escalating capacity to convey semantically profound, object-level contour characteristics [13].

Concurrently, CNNs boast formidable nonlinear classification capabilities attributed to the infusion of nonlinear activation functions. This augmentation endows the network with the capacity to apprehend intricate contextual nuances within high-dimensional domains [14]. In 2014, Zhang *et al.* [15] introduced the classic supervised learning algorithm Part-based R-CNN, which is based on the object detection algorithm R-CNN [16]. Initially, a bottom-up region algorithm is employed to generate a set of candidate bounding boxes. Subsequently, the R-CNN algorithm is used to detect these candidate regions and provide scoring values. Weakly supervised image classification does not rely on additional manual annotations but only uses image category labels. Through techniques such as high-order feature encoding and attention mechanisms, it enables the localization and prediction of local regions. Fu *et al.* [17] proposed a representative Recurrent Attention CNN, which operates on a principle similar to continuously zooming in the camera lens when observing a scene, thereby enhancing attention in regions requiring focus.

The vanguard of edge extraction was established by Roberts, who initially harnessed a 2x2 2D cross-gradient operator [18] to delineate edges along diagonal trajectories. However, due to the operator's inherent simplicity, it encapsulates scant orientation information and lacks symmetry with regard to the central pixel point. Subsequently, a 3x3 convolutional kernel was enlisted to compute gradients within localized regions, facilitating the identification of horizontal, vertical, and diagonal edges [19].

The renowned Canny edge detector [20–21], acclaimed as one of the paramount edge detection algorithms, orchestrates the extraction of edges from greyscale images. This methodological orchestration harmonizes Gaussian fuzzy operators with Sobel gradient operators, subsequently refining edge widths through subsequent post-processing enhancement algorithms, exemplified by non-maximal suppression. The zenith of this progression culminates in the formulation of a hysteresis thresholding algorithm. This stratagem seamlessly amalgamates edges that are feeble with those that are robust, yielding an edge map that strives for utmost comprehensiveness. Nevertheless, these nascent methodologies grappled with distinguishing responses originating from textures as opposed to those emanating from contours, inherently manifesting imprecision and unsuitability for contemporary applications.

An image's depiction of a scene, compounded by the infusion of noise, engenders alterations in the external information it encapsulates. The human eye capitalizes on the observation of these feature signal variations to discern image quality. This perceptual acumen, wielded by the visual system, affords a nuanced grasp of image quality [22]. As image distortion escalates, the likelihood of compromising high-quality features from the original image concurrently increases. For instance, severe Gaussian white noise distortion detrimentally affects color attributes, while Gaussian blurring distortion

erodes image sharpness. To comprehensively apprehend image distortion cues in an objective manner, a deeper investigation into the pivotal attributes within the image becomes imperative.

Extant Non-Reference Subjective Image Quality Assessment (NR-SIQA) algorithms exhibit commendable performance, often incorporating multiple feature descriptors. This inclusive approach effectively captures the state of image distortion, thus facilitating precise quality evaluations. Cao *et al.* [23] harnessed features like local amplitude, local phase, and visual saliency from distorted stereo image pairs as binocular attributes. These binocular energy response features were subsequently mapped to principal features through a statistical regression model, which in turn correlated with subjective quality scores. Si *et al.* [24], on the other hand, extracted monocular features for statistical quality perception from the left and right views separately. They fused multiple features from summation signals and entropy attributes of difference signals, which were harnessed to train a Support Vector Regression (SVR) model.

Moreover, the sphere of deep learning presently showcases evident advantages in image feature extraction. Through the automatic learning of image signals via multi-scale convolutional kernels, deep learning can amass more dimensional and intrinsic information. Concurrently, traditional manual extraction methods retain certain low-dimensional visual features overlooked by CNN models. The amalgamation of these dual feature extraction techniques yields a performance boost for NR-SIQA algorithms, accentuating their efficacy [25].

Within the realm of image analysis, feature coding algorithms, spearheaded by CNNs, have yielded commendable outcomes due to their robust feature representation capabilities. Nevertheless, these models primarily confine their utilization to the terminal layer of features, a practice that results in excessively high-dimensional feature outputs. Moreover, they neglect the treatment of their inherent attention graph, a shortcoming that impairs the localization of differentiated target components, thus rendering the models susceptible to overfitting [26–27]. In pursuit of elevating the efficacy of image feature representation, researchers have embarked upon the exploration of feature fusion from distinct networks. This endeavor has ushered in the concept of bilinear combination, whereby features emanating from two networks undergo element-wise product operations, engendering a novel feature representation. This refined representation is believed to more adeptly encapsulate disparate patterns and structures within images, thereby augmenting the model's classification performance. For instance, Cao and Zhang [28] achieved a noteworthy elevation in accuracy for a specific image classification task through the adept application of bilinear combination to the output features of two CNN-based networks. Similarly, the work of Sadeghzade [29] delved into the ramifications of deploying bilinear combination across diverse image datasets, proffering a comprehensive analysis of its merits. Subsequent inquiries have explored avenues to optimize the utility of bilinear combination. Zhang *et al.* [30] proffered an adaptive weight allocation strategy to more equitably balance the import of disparate network outputs during the fusion process. In parallel, Wang [31] interwove the bilinear combination concept with an attention mechanism, enabling the model to autonomously assimilate and accentuate pivotal feature insights.

3. Methods

3.1. Visual Cortex

Broadly speaking, the intricate realm of information processing within the precincts of the visual cortex encompasses two principal visual pathways: the dorsal pathway and the ventral pathway. The dorsal pathway comprises a succession of cerebral regions (V1-V2-V3-V5-PC) stretching from the occipital lobe to the parietal cortex (PC). This pathway takes up the mantle of processing visual information linked to motion and depth. In contrast, the ventral pathway encompasses an array of cerebral regions (V1-V2-V4-IT) extending from the occipital lobe to the inferior temporal cortex (IT). Its chief purview encompasses the processing of contours, hues, and letter forms. The literature furnishes a model explicating the segregated information processing dynamics in cortical domains [32]. Figure 1 offers a succinct graphical representation of this model.

Certainly, the intricate tapestry of cortical areas is far more interconnected than delineated earli-

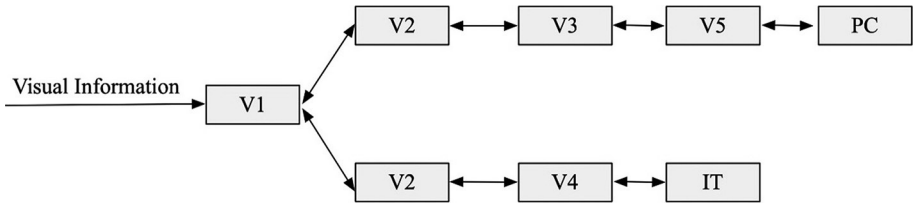

Figure 1. Schematic diagram of the visual pathway.

er. As an illustrative instance, area V1 simultaneously projects to regions like V3 and V5, revealing the intricate web of connections between these visual domains. This paper, in essence, underscores the pervasive interlinkages among these visual regions. The coexistence of bidirectional projections within each visual pathway underscores an exceptionally high level of internal regional correlation. It is imperative to note that extensive cross-projections traverse between the two pathways, as evidenced by the interactions between regions V3 and V4 in the diagram.

Guided by the discernible visual attributes associated with the division of labor among cortical regions in image processing, it becomes evident that the construction of perceptual quality in the realm of stereoscopic imagery is inseparably tied to the concerted interplay of these cortical realms. The quintessential query that looms large revolves around the judicious integration of this intricate process of visual perception into the fabric of the Subjective Image Quality Assessment (SIQA) algorithm—an inquiry meriting profound contemplation.

3.2. Algorithmic Framework

Within this paper, a novel interactive CNN model boasting a multi-stage framework enriched by an attention mechanism is meticulously fashioned. The skeletal architecture of this network is provided in Figure 2. The steps of its construction are as follows: Initially, a shallow convolutional structure undertakes the preliminary extraction of content features from the summation image, monocular image, and difference image. These extracted features undergo fusion, culminating in a low-level visual output signal entrenched within the V1 region. Subsequently, a dual-channel deep convolutional structure is meticulously crafted to usher in the high-level feature extraction of the amalgamated images. This mirrors the process wherein the intrinsic, sensitized information gleaned from the primary visual domain finds renewed utility and interaction as it navigates the terrain of the high-level visual arena (spanning V2 to V5).

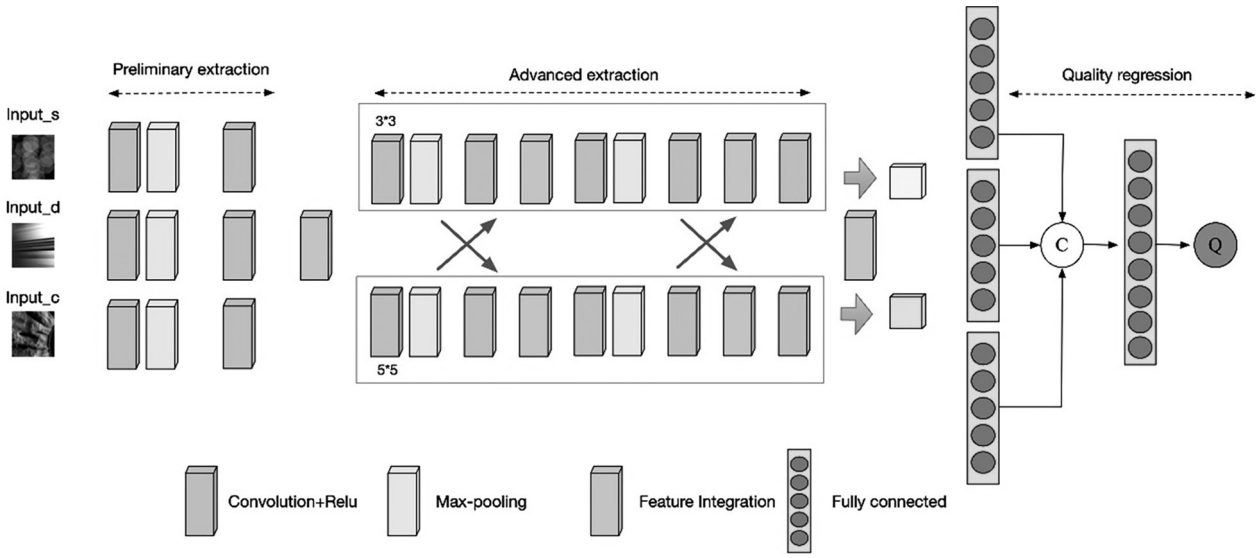

Figure 2. Algorithmic framework.

In the sphere of feature integration, the allocation of weights is orchestrated through an enhanced particle swarm optimization algorithm. The brightness, color, direction, and edge saliency maps are harmoniously superimposed via linear combination, thereby yielding a composite that mirrors the weight allocation outcomes. Ultimately, all these feature vectors coalesce within the embrace of a fully connected layer, culminating in the return of a perceptual quality score.

The bedrock of the model consists of a dual-tiered serial configuration of sub-networks tailored to emulate the image information processing paradigms witnessed within the advanced visual arenas—specifically, the dorsal and ventral pathways. Each of these sub-networks is endowed with five convolutional layers, with the convolutional layer pathways incrementally expanding: 32, 32, 48, 48, and 64. This systematic progression guarantees that as the network structure delves deeper, the capacity to encapsulate deeper facets of feature imagery is concomitantly fortified.

To enable each channel to assimilate multi-scale features from distinct receptive domains, the convolution kernel dimensions within the two sub-networks are set at 33 and 55, correspondingly. The convolution operation proceeds with a sliding step of 1, while the output image size of the convolutional layer remains constant via the employment of the "same" strategy. Following each convolutional layer, an activation function in the form of Rectified Linear Unit (ReLU) is enlisted to accentuate the nonlinear representation prowess of the network. Concurrently, a strategic inclusion of three maximum pooling layers contributes to parameter reduction, wherein the pooling window dimensions stand at 22 and the stride is set at 2. This orchestration imparts the final output feature image a scaled-down size, amounting to one-eighth of the initial input image.

Moreover, with consideration for the projective interconnections between visual pathways, our design encompasses multiple facets of information sharing within the network, epitomized by interactive connections between the sub-networks. This interaction mechanism draws parallels with the aforementioned fusion methodology for three-input images, yet their roles diverge. The former ensures the simultaneous ingestion of all features extracted from the three binocular images by both sub-networks. Conversely, the latter exploits the internal outputs of one sub-network as shared information across the other sub-network.

Within the intricate landscape of the interaction process, a pivotal operation known as Concatenation plays a vital role in enhancing the fusion of features within our neural network architecture. This operation seamlessly combines the feature output image generated by a convolutional layer within one sub-network with the maximally pooled output image derived from the corresponding position within the preceding layer of the adjacent sub-network. This harmonious amalgamation of information serves as a critical juncture in the network's ability to process and comprehend complex data.

To further consolidate this amalgamated information, a 1x1 convolutional kernel is employed. This convolutional operation is meticulously configured to preserve the original channel count, ensuring that the quantity of features within each sub-network layer remains consistent. This strategic decision helps maintain the network's capacity to capture and represent intricate patterns and nuances within the data.

The true innovation lies in this supplementation of feature content with additional information, a concept that is at the core of the network's effectiveness. By fusing feature maps from different layers and sub-networks, the model gains access to a broader spectrum of information, enabling it to discern intricate relationships and patterns that might otherwise go unnoticed. This augmentation of data is a testament to the network's adaptability and its ability to harness the wealth of information at its disposal.

It is crucial to emphasize that the network possesses an inherent ability to learn and adapt. Through the training process, it discerns the significance of various pieces of information and learns to prioritize what is truly efficacious for the task at hand. This innate ability allows the network to sift through the influx of information and distill it down to the most pertinent and influential features, ultimately optimizing its performance in complex tasks.

This interactive approach is carried out twice within the model, wielding the Fusion and In-

teraction (FI) module four times to seamlessly merge convolutional output images across disparate levels. To illustrate, the output of the first pooling layer in one sub-network is seamlessly linked with the output of the second convolutional layer in the other sub-network. The interactive processing model bestows upon a solitary visual channel the capacity to not only assimilate features within its own scope but also to integrate image information from an alternate channel. This symphony of cross-channel information exchange echoes a semblance of a cortical model of visual perception, mirroring the interplay observed within the human visual system.

3.3. Integration of Visual Features

In feature delineation, Canny edge features are additionally delineated, where weights are assigned by the improved particle swarm optimization algorithm, and brightness, colour, direction and edge salient maps are linearly superimposed according to the results of weight assignment.

To enhance the feature delineation stage through visual attention, a further augmentation is introduced, encompassing the delineation of edge features. This augmentation culminates in the computation of four distinct types of feature maps: luminance, color, orientation, and edge. While substantial research has delved into the computation of luminance, color, and orientation feature maps, this particular section delves exclusively into the intricate realm of edge saliency map computation, as shown in Figure 3.

Edge features are computed using the Canny operator on the intensity map $I(\sigma)$, $\sigma \in \{0, 1, ..., 8\}$ is convoluted to obtain, and $E(\sigma)$ is used to denote the edge pyramid. The edge feature map is calculated as follows:

$$E(c, s) = |E(c) \Theta E(s)| \tag{1}$$

where, the $c \in \{2, 3, 4\}$ denotes the high resolution scale factor. $s = c + \delta$ denotes the low resolution scale factor. $\delta \in \{3, 4\}$ represents the centre-periphery difference operator. Θ represents the centre-surround difference operator.

The weights of different features are reasonably assigned. The following formula was used for calculating Edge saliency map $\overline{E}$.

$$\overline{E} = \bigoplus_{c=2} \overset{c+4}{\oplus} N\left(E\left(c, s\right)\right) \tag{2}$$

In Formula (2), $\oplus$ denotes cross-scale fusion, where feature maps at different scales are interpolated and adjusted to the same scale and then summed. N denotes the normalisation operation.

When integrating features, this paper presents the brightness feature salient maps $\overline{I}$, colour feature salient maps $\overline{C}$, direction feature salient map $\overline{O}$ and edge features $\overline{E}$, and then linearly superimpose them to calculate the final visual saliency map S. The final visual saliency map is calculated as follows:

$$S = \alpha_1 \cdot N\left(\overline{I}\right) + \alpha_2 \cdot N\left(\overline{C}\right) + \alpha_3 \cdot N\left(\overline{O}\right) + \alpha_4 \cdot N\left(\overline{E}\right) \tag{3}$$

where the α_1, α_2, α_3 and α_4 are the weights of luminance, colour, orientation and edge, respectively, which are all between 0 and 1, and $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$ are the weights of brightness, colour, direction and edge, respectively, all between 0 and 1, and $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$.

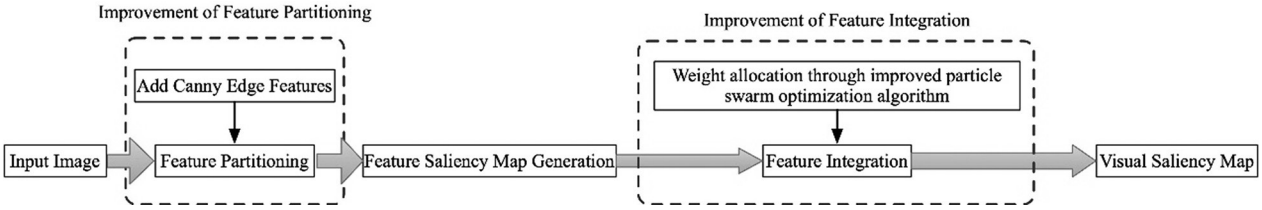

Figure 3. Visual feature extraction process.

Considering that there are many different combinations of weights when assigning weights. If every weight combination is traversed, although the most reasonable weight assignment result can be found, it cannot be done in a limited time. In order to efficiently determine the reasonable weight combinations, the particle swarm optimization algorithm is used to assign the weights of brightness, colour, direction and edge feature salient maps.

3.4. Channel Attention Module

The novel Convolutional Block Attention Module (CBAM), as elucidated in this study, embodies an attention mechanism module that ingeniously fuses spatial attention and channel attention [33]. This astute amalgamation translates to superior outcomes when compared to attention mechanisms that singularly emphasize either channel or spatial attributes. The architecture of CBAM revolves around two distinct sub-modules: the Channel Attention Module and the Spatial Attention Module. The former calculates attention maps from the channel perspective, while the latter orchestrates attention map computation grounded in spatial characteristics. The ensuing step involves the multiplication of these calculated attention maps with the corresponding feature elements, ultimately yielding salient feature maps. This composite mechanism endeavors to synergistically optimize both channel-specific and spatially-informed attention, thereby elevating the model's efficacy in discerning salient information.

Given the input features $F \in \mathbb{R}^{C\times H\times W}$,we sequentially infer a 1D channel attention map $M_C \in \mathbb{R}^{C\times 1\times 1\times 1}$ and 2D spatial attention $M_C \in \mathbb{R}^{1\times H\times W}$, the overall attention process can be summarised as follows:

$$F' = M_C(F) \otimes F \tag{4}$$

$$F'' = M_S(F) \otimes F \tag{5}$$

where $\otimes$ denotes the element-wise multiplication and F'' is the final output.

Within convolutional networks, feature outputs are often channeled into a multifarious array of channels, each endowed with a varying degree of influence over the ultimate target. Consequently, the pursuit of refined attention distribution mandates a concentration of attention onto these pivotal channels. Conventionally, global pooling is leveraged to achieve this, but CBAM introduces an advanced approach. CBAM capitalizes on the potency of global maximum and global average pooling, employing both in tandem. These pooled outcomes subsequently traverse through a nonlinear transformation conducted via a fully connected network. The ensuing phase involves feature summation and subsequent reactivation, which collectively bestow attention weights. This intricate orchestration manifests as an endeavor to foster heightened attention focus on channels that wield significant sway over the overarching objective, culminating in a more precise channel-specific attention allocation.

The features are first compressed in the spatial dimension using MaxPool and AvgPool to obtain two different spatial contextual descriptions: the F^C_{avg} and F^C_{max}, then these two different spatial context descriptions are computed using a shared network consisting of MLPs to obtain a channel attention map $M_C \in \mathbb{R}^{C\times 1\times 1\times 1}$.

$$\begin{aligned} M_c(F) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F))) \\ &\quad + \mathrm{MLP}(\mathrm{MaxPool}(F)) \\ &= \sigma(W_1(W_0(F^C_{avg})) \\ &\quad + W_1(W_0(F^C_{max}))) \end{aligned} \tag{6}$$

In Formula 6, σ denotes the sigmoid function, W_0 and W_1 are the MLP weights. The MLP network parameters are shared.

The max pooling layer and the average pooling layer are used to generate two 2D feature description maps F^C_{avg}, $F^C_{max} \in \mathbb{R}^{1\times H\times W}$. The spatial attention weights can be obtained by connecting the two outputs in the channel dimension, convoluting them with a convolution kernel of size 7×7, and then activating them with a sigmoid function. The formula of spatial attention is as follows:

$$\begin{aligned} M_S(F) &= \sigma(f^{7\times 7}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])) \\ &= \sigma(f^{7\times 7}(F^C_{avg}, F^C_{max})) \end{aligned} \tag{7}$$

where the $f^{7\times 7}$ denotes the convolution kernel size of 7×7, and σ denotes the sigmoid activation function.

4. Experiments and Analysis

4.1. Experimental Environment

The experimental setting encompasses Python 3.6 as the programming language, an i7-6800k processor as the computational engine, accompanied by a 32 GB RAM. The system uses a a GTX2080Ti GPU, and Linux operating system. The deep learning framework used was PyTorch.

For the purpose of testing the proposed methods we use the publicly accessible LIVE 3D Phase dataset. Our approach to harnessing the potential of this dataset is characterized by an elaborate and meticulous stratagem. To construct a robust and versatile foundation for our research, we employ a careful partitioning technique. Eighty percent of this voluminous corpus is judiciously selected to form the crucible of our training set. Within this extensive training set lies a wealth of diverse examples, each contributing to the development of our predictive models. Here, the algorithms decipher intricate patterns, recognize trends, and internalize the nuances of the data through a process of continuous learning.

The remaining 20% of the dataset takes on the vital role of the testing crucible. This subset serves as the ultimate litmus test for our models, as it comprises unseen data instances that our algorithms must confront. In this realm, our models are put to the challenge of applying the knowledge distilled from the training data to make accurate predictions and classifications. It is in this testing crucible that the true mettle of our models is evaluated, and their generalization abilities are scrutinized.

Every data point, from its acquisition and preprocessing to its incorporation into either the training or testing set, is meticulously handled. Metadata, timestamps, sensor calibrations, and contextual information are all carefully preserved to maintain the dataset's authenticity and integrity. This ensures an unbridgeable chasm between the realms of training and testing, eliminating any potential for data leakage or overlap that could compromise the validity of our predictive outcomes.

4.2. Evaluation Indicators

PLCC (Pearson Linear Correlation Coefficient) and HI (Histogram Intersection) were used to validate the accuracy and performance of the visual saliency map calculation results. These metrics can help to assess the degree of match between the computed results and the real data, thus measuring the effectiveness of the algorithm.

PLCC is used to measure the degree of linear relationship between two variables. Here, it is used to compare the correlation between the generated saliency map and the true saliency map, which is calculated by the formula:

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{n}(X_i - \mu X)(Y_i - \mu Y)}{n\sigma X \sigma Y} \tag{8}$$

where, the n is the sample size. μX and μY are the means of the generated saliency maps and the true saliency maps converted to one-dimensional vectors, σX and σY are their standard deviations.

HI is used to measure the similarity between two histograms. Here, it is used to compare the degree of distributional similarity between the generated saliency map and the true saliency map, which is computed as follows:

$$HI = \sum_{i=1}^{n} \min\left(H(X_i), H(Y_i)\right) \tag{9}$$

where $H(X_i)$ is the number of pixels in the first interval of the generated saliency map i number of pixels in the interval, $H(Y_i)$ is the number of pixels in the first interval of the true significance map, i is the number of pixels in the first interval of the true significance map, and n is the number of intervals.

4.3. Results

4.3.1. Model Training Process

The optimal weights obtained after experimentation to get the luminance, colour, orientation and edge features are $\alpha_1 = 0.0334$, $\alpha_2 = 0.1306$, $\alpha_3 = 0.3901$ and $\alpha_4 = 0.4459$, respectively, and

the performance of the model is recorded for the training and testing iterations as shown in Figure 4.

Overall, the interactive convolutional network model with the introduction of the multi-stage and attention mechanism model converges faster, does not show convergence jittery state, and presents a stable convergence state of the accuracy and loss values after the 6th iteration. The faster convergence speed means fast and robust fusion of the model, which can avoid overfitting or falling into the local optimum.

4.3.2. Comparison of Feature Extraction Results

After obtaining 1200 PLCC and HI metrics corresponding to luminance, colour, orientation and edges based on the adopted dataset, the corresponding PLCC and HI for each feature were averaged, and the results of the calculations are shown in Figure 5.

A greater magnitude in the mean of the PLCC serves as an indicator of heightened correlation between the respective feature saliency map and the established visual saliency map. Similarly, an elevated mean value in the HI signifies enhanced resemblance between the corresponding feature saliency map and the standard visual saliency map. The outcomes unveiled that edge features exhibited the most substantial PLCC and HI indices, measuring at 0.6206 and 0.0179, sequentially. These findings follow a descending order of correlation and similarity, namely: edge, orientation, color, and luminance.

Furthermore, for comparative analysis, the models of GBVS, AIM, CAS, and DVA were thoughtfully selected, with the outcomes delineated in Figure 6.

GBVS (Graph-Based Visual Saliency) [34], a graphical model for visual saliency, constructs a graph to represent inter-pixel relationships in an image, employing graph analysis to derive saliency. It embraces multiple attributes such as chromaticity, luminosity, orientation, and local contrast, thereby conducting a thorough exploration of local and global image features to craft the saliency map.

AIM (Attention by Information Maximization) [35], grounded in the principles of information theory, operates as a model for visual attention by maximizing information gain. It orchestrates a transformation of the image into a multiscale representation, followed by the derivation of a saliency map through the maximization of mutual information at every image scale.

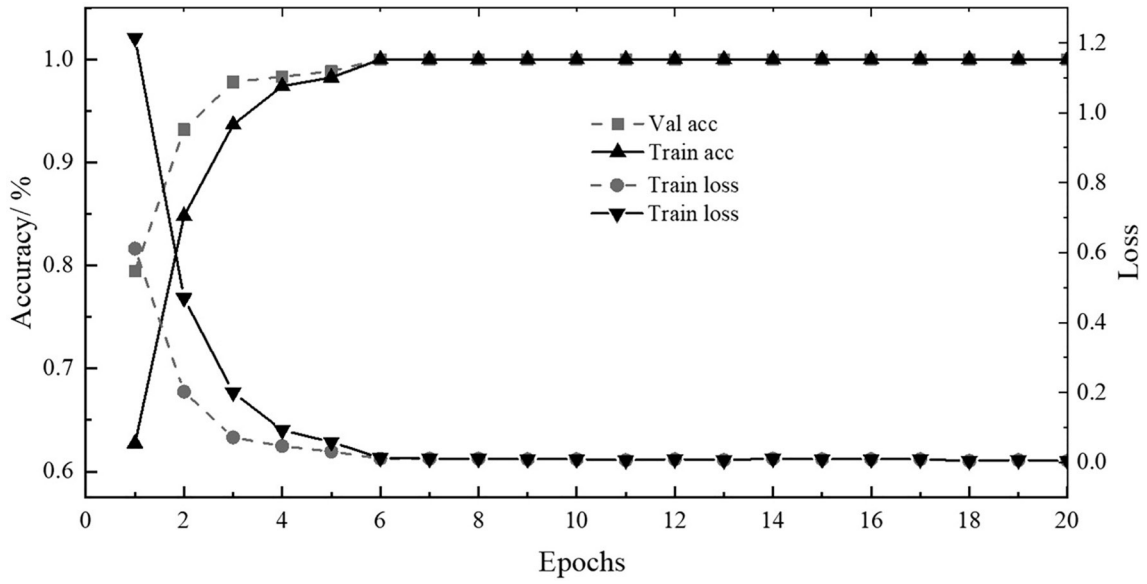

Figure 4. Model training process.

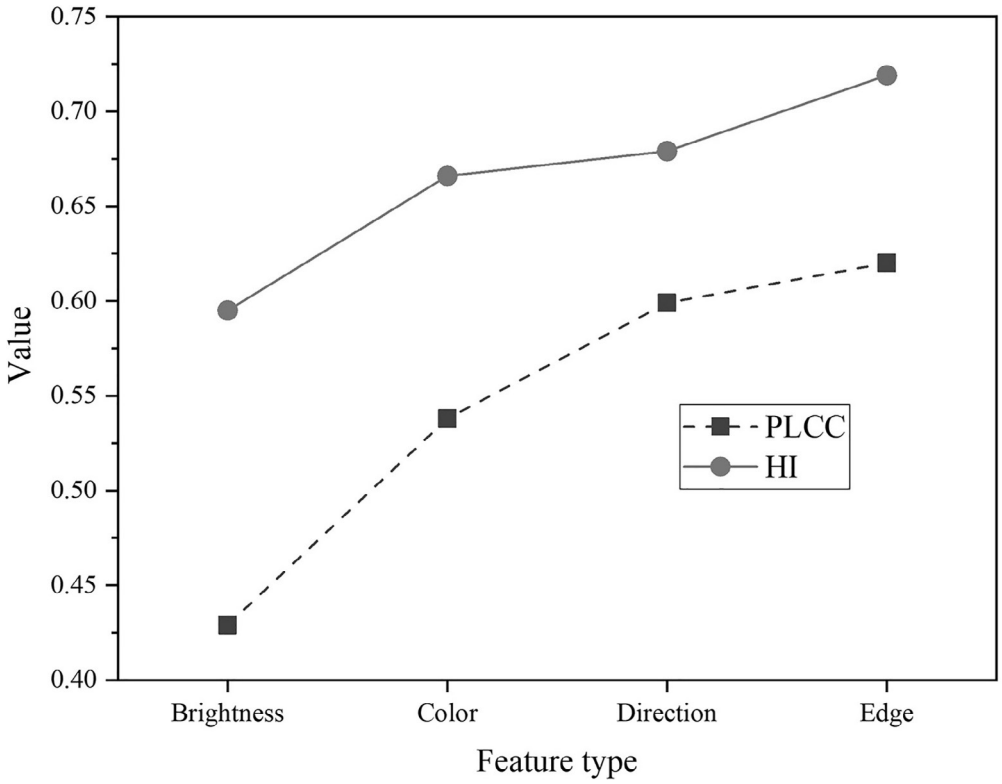

Figure 5. Comparison of the effect of different feature extraction.

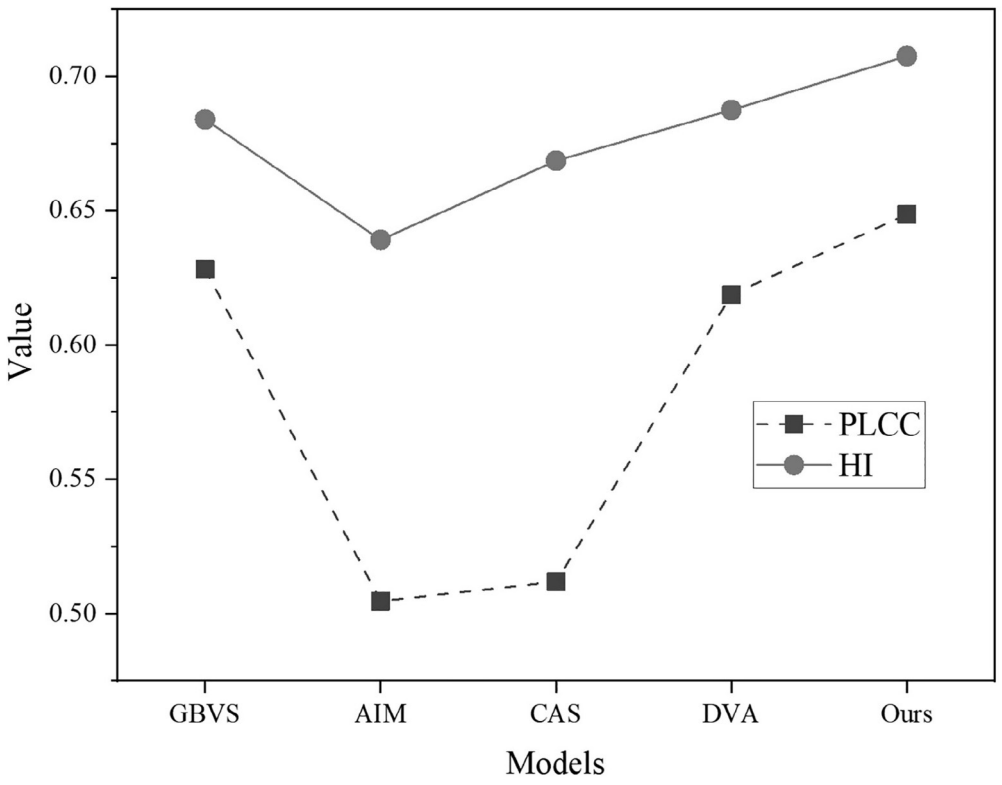

Figure 6. Model comparison results.

CAS (Context-Aware Saliency) [36], a saliency computation algorithm, bestows significance upon contextual insight. This notion revolves around the impact exerted by encompassing regions upon the saliency of the present region. The algorithm efficiently captures salient facets within the image by meticulously accounting for the attributes and connections of neighboring regions.

DVA (Dynamic Visual Attention) [37], a model of dynamic visual attention, encompasses temporal fluctuations and motion intricacies. Saliency computation within DVA hinges upon the scrutiny of video sequences, reliant upon motion characteristics and target fluctuations. This characteristic equips DVA with an upper hand in processing scenes with dynamic attributes.

From the aforementioned findings, it becomes apparent that the optimization methodology articulated within this paper yields the most substantial averages for both PLCC and HI metrics, attaining 0.6486 and 0.7074 correspondingly. In light of the model contrast outcomes, it becomes evident that the processing of digital imagery through the proposed optimization scheme resonates more harmoniously with HVS compared to the conventional GBVS, AIM, and CAS approaches.

Furthermore, it is noteworthy that the DVA model's responsiveness could be contingent upon the selection of model architectures and hyperparameters. The intricacy inherent to the DVA model might result in escalated computational expenses. In contrast, juxtaposed with the DVA model, our enhanced approach demonstrates judicious segmentation of distinct feature weights, obviating the necessity for extensive datasets. This trait bestows commendable accuracy and stability to the computed visual saliency maps, particularly for natural images. Consequently, the efficacy of the advanced visual attention model stands validated.

4.3.3. Image Quality Evaluation

The assessment encompassed a test set of partially distorted stereo images, meticulously scrutinized to delineate disparities between the empirical subjective evaluation scores and the prognosticated scores from the model. Within this study, the global efficacy of the proposed approach was gauged utilizing DMOS, with outcomes meticulously depicted in Figures 7 and 8. Along the horizontal axis lie the project-

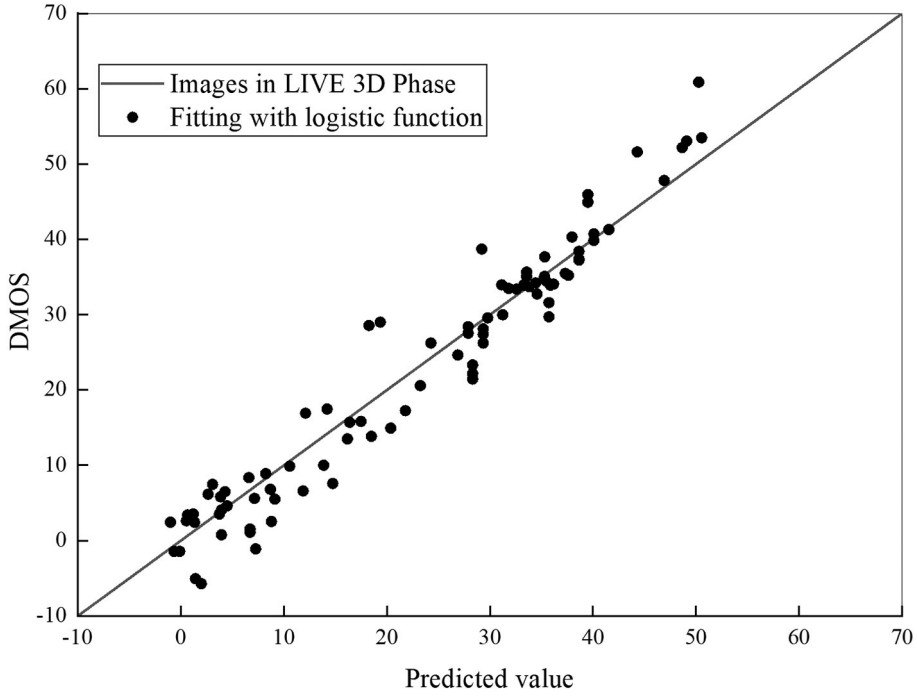

Figure 7. Comparison of image quality evaluation results (with improvement).

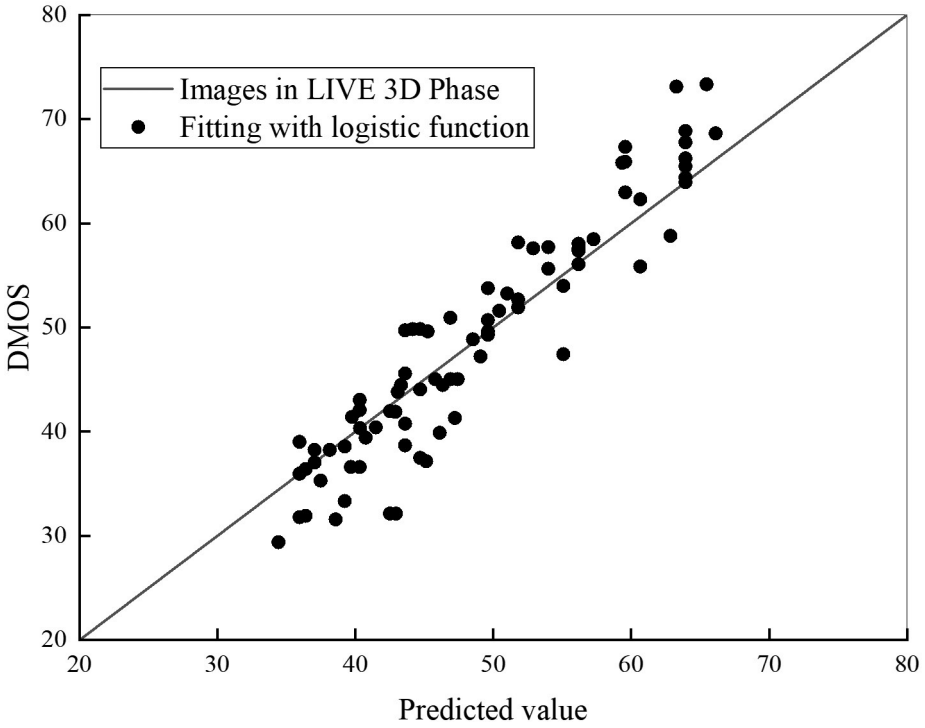

Figure 8. Comparison of image quality evaluation results (without improvement).

ed quality scores for the distorted images, while the corresponding DMOS values are plotted along the vertical axis. Each distinct symbol featured within the graph denotes a collection of stereo images perturbed by a particular form of distortion.

The scatter plot configuration divulges a compelling trend wherein the subjective and objective scores exhibit a notable convergence. This convergence underscores the high degree of correspondence between the devised scheme and subjective evaluations, persisting across diverse distortion types. The investigation encompassed a comprehensive spectrum of image content attributes, encompassing luminosity, chromaticity, directionality, edge characteristics, in addition to sophisticated visual features autonomously acquired through CNNs.

A block sampling strategy was employed for every ensemble of distorted stereo images, meticulously tailored to align with the requisites of CNN model refinement. Concomitantly, quality scores for all images were acquired by aggregating the projected values of their image blocks. The ramifications of this localized averaging upon ultimate perceptual quality might potentially hinge upon the dimensions of the image blocks. Thus, it becomes imperative to examine the impact of this sampling strategy upon algorithmic efficacy.

The experimentation encompassed the segmentation of images into three distinct dimensions, sequentially subjected to the unaltered model for iterative training and assessment. The outcomes of these comparative trials are graphically presented in Figure 9, outlining the effects of the sampling strategy upon algorithmic performance.

As we delve into the intricacies of Figure 9, it becomes evident that the various dimensions chosen for testing purposes exerted a negligible impact on the model's overall performance. This observation is supported by the marginal discrepancies observed in the metric values across the different dimension settings. The resilience displayed by our proposed model in the face of alterations in image block size parameters is a noteworthy characteristic. It underscores the adaptability and robustness of the model, highlighting its capacity to maintain consistent performance regardless of the specific dimensions chosen for analysis. This remarkable quality is

a testament to the versatility of our algorithm, which can seamlessly accommodate changes in the size and scale of the input data without compromising its effectiveness.

Figure 9 serves as a visual testament to the stability and reliability of our model across a range of image block size variations. This adaptability is a valuable asset, particularly in scenarios where data may exhibit inherent variability or where flexibility in parameter settings is crucial. The insights gained from these trials not only contribute to our understanding of the model's behavior but also reinforce its suitability for a wide spectrum of practical applications, where consistent and robust performance is paramount.

4.4. Discussion

The presented experimental outcomes elucidate our proposed model's intricate enhancements, which include fostering interactive interconnections between channels to encapsulate the intricate realm of depth perception within the higher visual cortex. A strategic augmentation surfaces by incorporating edge features that align more fervently with the discerning tendencies of HVS during the feature delineation phase. This augmentation is further refined through the integration of Canny edge characteristics. Additionally, in the feature integration juncture, the model accommodates the diverse sensitivities of HVS to distinct attributes, thus imparting augmented precision to the computed visual saliency maps.

The architecture of the dual-channel model, equipped with multi-tiered interaction processing, mirrors the intricate trajectory of actual human visual signal processing. This assimilation, coupled with a synthesis of HVS visual perceptual traits, manifests a substantial upsurge in the evaluation algorithm's efficacy. Nonetheless, the intricate task of gauging the degree of distortion within stereoscopic images solely from the amalgamated content and depth information remains a formidable challenge. Such an approach might not comprehensively encapsulate the intricate nuances of the Human Visual System's perceptual traits. This nuance distinguishes it from certain prevailing deep interaction network models, signifying an ongoing gap in alignment with these models.

Promising avenues for future research encompass addressing the intricacies tied to images harboring densely distributed small targets that pose challenges for user annotation. This endeavor might lead to the development of clas-

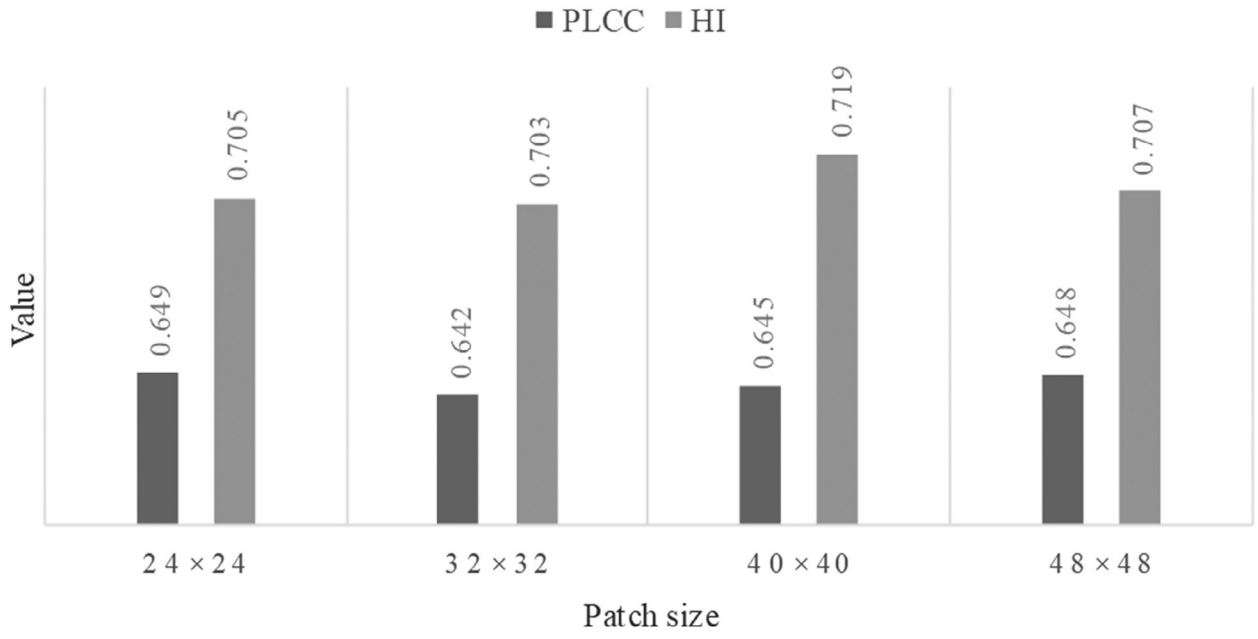

Figure 9. Comparison of model training effect under different image sizes.

sification models featuring numerous convolutional layers, engendering expansive model capacity and parameters. Subsequent research endeavors could gravitate towards refining existing classification network models, culminating in lightweight encoding networks that epitomize accuracy and efficiency, all the while drawing inspiration from visual neural mechanisms.

5. Conclusion

This study presented a novel CNN model for optimizing digital image design by integrating insights from deep learning and visual cortex analysis. The dual-channel architecture with multi-scale convolutions mimics the hierarchical processing in biological vision. Inter-network connections allow complementary feature sharing, further enriched by attention modules targeting spatial and channel correlations. Adaptive integration of saliency maps for key attributes including luminance, color, orientation, and edges improves alignment with human perceptual tendencies.

Comprehensive quantitative testing on benchmark datasets demonstrates significant performance gains over existing methods. The model achieves a Pearson Correlation Coefficient of 0.6486 and Histogram Intersection of 0.7074, along with strong agreement in subjective image assessments across distortion types. These results highlight the efficacy of combining automated feature learning with knowledge from neuroscience to enhance computer vision.

By emulating the intricacies of visual information flow, this bio-inspired model marks a pivotal advancement in replicating human-level sensitivity to complex attributes. The work provides an adaptable framework to continue pushing the frontiers of image optimization and quality. Extending the synergistic fusion of deep learning and biological principles holds immense promise for tackling long-standing challenges in computer vision and pattern recognition. Overall, this study helps blaze the trail toward more sophisticated, human-centered AI technologies.

References

[1] F. Ye and Y. Li, "Research on 3D Modelling Software Maya Digital Media Animation Assisted Brain Surgery Technology", *Journal of Imaging Science & Technology*, vol. 65, no. 2, pp. 2–7, 2021. http://dx.doi.org/10.2352/J.ImagingSci.Technol.2021.65.2.020401

[2] Y. Yuan, "Modeling of Stereoscopic Images in 3D Environmental Art Design. International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2022). and Data Processing (ICCAID 2022)," SPIE, 12604: 1260402, 2023. http://dx.doi.org/10.1117/12.2674870

[3] B. Zhang *et al.*, "Three-dimensional Convolutional Neural Network Model for Tree Species Classification Using Airborne Hyperspectral Images", *Remote Sensing of Environment*, vol. 247, 2020. http://dx.doi.org/10.1016/j.rse.2020.111938

[4] B. Xi *et al.*, "Few-shot Learning with Class-covariance Metric for Hyperspectral Image Classification", *IEEE Transactions on Image Processing*, vol. 31, pp. 5079–5092, 2022. http://dx.doi.org/10.1109/TIP.2022.3192712

[5] Z. Zhao *et al.*, "A Battle of Network Structures: An Empirical Study of CNN, Transformer, and MLP", arXiv preprint arXiv:2108.13002, 2021. http://dx.doi.org/10.48550/arXiv.2108.13002

[6] D. Pan *et al.*, "Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning", *Frontiers in Neuroscience*, vol. 14, p. 259, 2020. http://dx.doi.org/10.3389/fnins.2020.00259

[7] C. Zhuang *et al.*, "ACDNet: Adaptively Combined Dilated Convolution for Monocular Panorama Depth Estimation", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 3653–3661, 2022. http://dx.doi.org/10.1609/aaai.v36i3.20278

[8] H. Peng H *et al.*, "Automatic Aesthetics Evaluation of Robotic Dance Poses Based on Hierarchical Processing Network", *Computational Intelligence and Neuroscience*, vol. 2022, 2022. http://dx.doi.org/10.1155/2022/5827097

[9] P. Sciortino and C. Kayser, "Steady State Visual Evoked Potentials Reveal a Signature of the Pitch-size Crossmodal Association in Visual Cortex", *NeuroImage*, vol. 273, 2023. http://dx.doi.org/10.1016/j.neuroimage.2023.120093

[10] Y. Sun *et al.*, "Low-illumination Image Enhancement Algorithm Based on Improved Multi-scale Retinex and ABC Algorithm Optimization", *Frontiers in Bioengineering and Biotechnology*, vol. 10, 2022. http://dx.doi.org/10.3389/fbioe.2022.865820

[11] N. Thakur *et al.*, "Deep Learning-based Parking Occupancy Detection Framework Using ResNet and VGG-16", *Multimedia Tools and Applications*, pp. 1–24, 2023. http://dx.doi.org/10.1007/s11042-023-15654-w

[12] D. Yang *et al.*, "An Overview of Edge and Object Contour Detection", *Neurocomputing*, vol. 488, pp. 470–493, 2022. http://dx.doi.org/10.1016/j.neucom.2022.02.079

[13] T. Lei *et al.*, "Local and Global Feature Learning with Kernel Scale-adaptive Attention Network for VHR Remote Sensing Change Detection", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7308–7322, 2022. http://dx.doi.org/10.1109/JSTARS.2022.3200997

[14] Z. Wang *et al.*, "Hybrid cGAN: Coupling Global and Local Features for SAR-to-Optical Image Translation", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022. http://dx.doi.org/10.1109/TGRS.2022.3212208

[15] N. Zhang *et al.*, "Part-based R-CNNs for Fine-grained Category Detection", *European conference on computer vision*, pp. 834–849, 2014. http://dx.doi.org/10.1007/978-3-319-10590-1_54

[16] R. Girshick *et al.*, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. http://dx.doi.org/10.1109/cvpr.2014.81

[17] J. Fu *et al.*, "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446. http://dx.doi.org/10.1109/cvpr.2017.476

[18] L. G. Roberts, "Machine Perception of Three-dimensional Solids", *Massachusetts Institute of Technology*, 1963. http://dx.doi.org/10.1016/0045-7949(85)90050-1

[19] R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", *New York: Wiley*, vol. 137, no. 3, pp. 442–443, 1973. http://dx.doi.org/10.2307/2344977

[20] Y. Lu *et al.*, "Application and Improvement of Canny Edge-detection Algorithm for Exterior Wall Hollowing Detection Using Infrared Thermal Images", *Energy and Buildings*, vol. 274, 2022. http://dx.doi.org/10.1016/j.enbuild.2022.112421

[21] S. Zhao *et al.*, "Non-Contact Crack Visual Measurement System Combining Improved U-Net Algorithm and Canny Edge Detection Method with Laser Rangefinder and Camera", *Applied Sciences*, vol. 12, no. 20, 2022. http://dx.doi.org/10.3390/app122010651

[22] S. Yang *et al.*, "Maniqa: Multi-dimension Attention Network for No-reference Image Quality Assessment", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.1191–1200, 2022. http://dx.doi.org/10.48550/arXiv.2204.08958

[23] L. Cao *et al.*, "Client-Oriented Blind Quality Metric for High Dynamic Range Stereoscopic Omnidirectional Vision Systems", *Sensors*, 2022, vol. 22, no. 21, p. 8513. http://dx.doi.org/10.3390/s22218513

[24] J. Si *et al.*, "A No-reference Stereoscopic Image Quality Assessment Network Based on Binocular Interaction and Fusion Mechanisms", *IEEE Transactions on Image Processing*, vol. 31, pp. 3066–3080, 2022. http://dx.doi.org/10.1109/TIP.2022.3164537

[25] C. Liu *et al.*, "Multi-scale ResNet and BiGRU Automatic Sleep Staging Based on Attention Mechanism", *Plos one*, vol. 17, no. 6, 2022. http://dx.doi.org/10.1371/journal.pone.0269500

[26] Z. Wang *et al.*, "High-quality Image Compressed Sensing and Reconstruction with Multi-scale Dilated Convolutional Neural Network", *Circuits, Systems, and Signal Processing*, vol. 42, no. 3, pp. 1593–1616, 2023. http://dx.doi.org/10.1007/s00034-022-02181-6

[27] J. Wang *et al.*, "Multi-classification of UWB Signal Propagation Channels Based on One-dimensional Wavelet Packet Analysis and CNN", *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8534–8547, 2022. http://dx.doi.org/10.1109/TVT.2022.3172863

[28] P. Cao *et al.*, "Improving Deep Learning Based Second-Order Side-Channel Analysis With Bilinear CNN", *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3863–3876, 2022. http://dx.doi.org/10.1109/TIFS.2022.3216959

[29] A. Sadeghzadeh and M. B. Islam, "BiSign-Net: Fine-grained Static Sign Language Recognition based on Bilinear CNN", in *Proceedings of the 2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). IEEE*, 2022, pp.1–4. http://dx.doi.org/10.1109/ISPACS57703.2022.10082808

[30] X. Zhang *et al.*, "Hierarchical Bilinear Convolutional Neural Network for Image Classification", *IET Computer Vision*, vol. 15, no. 3, pp. 197–207, 2021. http://dx.doi.org/10.1049/cvi2.12023

[31] Y. Wang *et al.*, "Aircraft Image Recognition Network Based on Hybrid Attention Mechanism", *Computational Intelligence and Neuroscience*, vol. 2022, 2022. http://dx.doi.org/10.1155/2022/4189500

[32] M. Kern *et al.*, "Blink- and Saccade-related Suppression Effects in Early Visual Areas of the

Human Brain: Intracranial EEG Investigations During Natural Viewing Conditions", *NeuroImage*, vol. 230, 2021. http://dx.doi.org/10.1016/j.neuroimage.2021.117788

[33] Y. Zhang *et al.*, "ANC: Attention Network for COVID-19 Explainable Diagnosis Based on Convolutional Block Attention Module", *CMES-Computer Modeling in Engineering & Sciences*, vol. 127, no. 3, pp. 1037–1058, 2021. http://dx.doi.org/10.32604/cmes.2021.015807

[34] R. K. Kumar *et al.*, "Guiding Attention of Faces Through Graph Based Visual Saliency (GBVS)", *Cognitive Neurodynamics*, vol. 13, pp. 125–149, 2019. http://dx.doi.org/10.1007/s11571-018-9515-z

[35] C. Gu *et al.*, "Image Search with Text Feedback by Deep Hierarchical Attention Mutual Information Maximization", in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4600–4609. http://dx.doi.org/10.1145/3474085.3475619

[36] M. Ahmadi *et al.*, "Context-aware Saliency Detection for Image Retargeting Using Convolutional Neural Networks", *Multimedia Tools and Applications*, vol. 80, pp. 11917–11941, 2021. http://dx.doi.org/10.1007/s11042-020-10185-0

[37] X. Hou and L. Zhang, "Dynamic Visual Attention: Searching for Coding Length Increments", *Advances in Neural Information Processing Systems 21*, 2008.

Received: August 2023
Revised: October 2023
Accepted: October 2023

Contact addresses:
Lei Zheng
Shanghai Lida University
Shanghai
China
e-mail: zhenglei2000@hotmail.com

Zheng Lei is a lecturer at Shanghai Lida College. He obtained his Master's degree in Software Engineering from Shanghai Jiao Tong University in 2013. Since 2011, he has published more than 10 journal papers and authored a monograph. His current research interests include optical visual image art, interactive art, and digital media.