Cross-Modal Scene Semantic Alignment for Image Complexity Assessment

Yuqing Luo[1]
27620195575393@stu.xmu.edu.cn
Yixiao Li[*2]
18335310648@163.com
Jiang Liu[1]
liuj137@cardiff.ac.uk
Jun Fu[1]
fujun@mail.ustc.edu.cn
Hadi Amirpour[3]
Hadi.Amirpour@aau.at
Guanghui Yue[4]
yueguanghui@szu.edu.cn
Baoquan Zhao[5]
zhaobaoquan@mail.sysu.edu.cn
Padraig Corcoran[1]
corcoranp@cardiff.ac.uk
Hantao Liu[1]
liuh35@cardiff.ac.uk
Wei Zhou[✉1]
zhouw26@cardiff.ac.uk

[1] School of Computer Science
Cardiff University
United Kingdom

[2] School of Mathematical Sciences
Beihang University
China

[3] Department of Information Technology
University of Klagenfurt
Austria

[4] School of Biomedical Engineering
Shenzhen University
China

[5] School of Artificial Intelligence
Sun Yat-sen University
China

Abstract

Image complexity assessment (ICA) is a challenging task in perceptual evaluation due to the subjective nature of human perception and the inherent semantic diversity in real-world images. Existing ICA methods predominantly rely on hand-crafted or shallow convolutional neural network-based features of a single visual modality, which are insufficient to fully capture the perceived representations closely related to image complexity. Recently, cross-modal scene semantic information has been shown to play a crucial role in various computer vision tasks, particularly those involving perceptual understanding. However, the exploration of cross-modal scene semantic information in the context of ICA remains unaddressed. Therefore, in this paper, we propose a novel ICA method called Cross-Modal Scene Semantic Alignment (CM-SSA), which leverages scene semantic alignment from a cross-modal perspective to enhance ICA performance, enabling complexity predictions to be more consistent with subjective human perception. Specifically, the proposed CM-SSA consists of a complexity regression branch and a scene semantic alignment branch. The complexity regression

[*] denotes the co-first author. [✉] denotes the corresponding author.

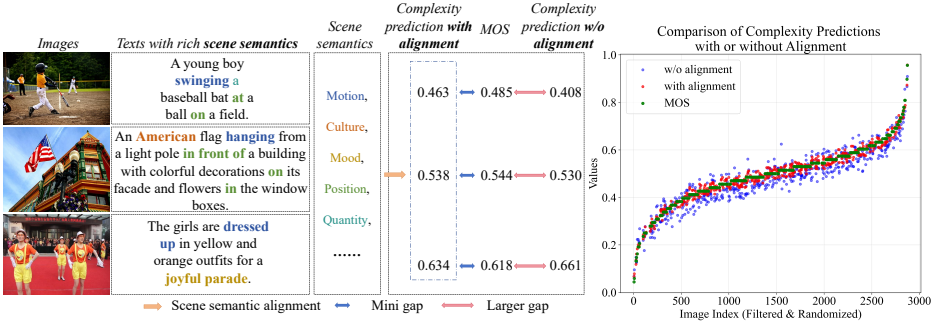

Figure 1: Examples of image complexity predictions with and without the guidance of our proposed cross-modal scene semantic alignment. Here, we adopt scene semantic information such as motion, culture, mood, spatial position, and quantity, which capture the relationships among the key visual entities (e.g., people, objects) within an image. The scatter plot visualizes the predictions of complexity scores on the IC9600 test set.

branch estimates image complexity levels under the guidance of the scene semantic alignment branch, while the scene semantic alignment branch is used to align images with corresponding text prompts that convey rich scene semantic information by pair-wise learning. Extensive experiments on several ICA datasets demonstrate that the proposed CM-SSA significantly outperforms state-of-the-art approaches. Codes are available at https://github.com/XQ2K/First-Cross-Model-ICA.

1 Introduction

The study of image complexity is particularly relevant to psychology and computer vision research [25, 34]. In psychology, image complexity plays a critical role in shaping visual aesthetics and influencing emotional responses [30, 45], while in computer vision, it serves as a fundamental attribute across various tasks. Automating image complexity assessment (ICA) has been proven beneficial for applications that include fire detection [24], image steganography [15], visual quality assessment [51], aesthetic image classification [35], and image enhancement [7]. However, the definition of image complexity remains a complex challenge. Subjectively, it represents the cognitive challenge that a human observer faces in interpreting or describing an image, factoring in global patterns and localized details, such as textures or fine structures [13, 32]. Objectively, image complexity encompasses the amount of detail, diversity, and structural variety [31].

In the literature, existing ICA methods can be broadly categorized into hand-crafted feature-based and deep learning-based methods. Among hand-crafted feature-based ICA approaches, early studies conceptualized complexity through statistical analysis of certain attributes of images. For example, multidimensional attributes such as object count, clutter, openness, symmetry, and color variety [31] are considered. Subsequent works employed machine learning models like gradient-boosted trees to regress complexity features [43], and aligned computational measures with subjective human perception [32]. Other studies explored spatial information and compression-based metrics [49], eye movement analysis [8], detailed feature designs capturing global, local, as well as salient characteristics [16], compression errors [29], and entropy [36, 41]. However, these methods have limited generaliza-

tion for the ICA task compared to deep learning-based approaches.

In recent years, deep learning-based ICA methods have leveraged neural networks to capture richer representations. Early works used the convolutional neural network (CNN) to extract features from texture and salient regions for complexity prediction [1, 6, 17, 38]. However, the limited dataset size constrained the evaluation of their effectiveness and generalization ability. To address this, the IC9600 dataset [12] was introduced, providing 9,600 labeled images along with a dual-branch CNN model, i.e., ICNet. To further reduce the annotation costs of constructing a large-scale dataset, Liu et al. [27] employed unsupervised contrastive learning based on MoCo v2 [5] for complexity representation learning. And attention-based models are further explored [3, 26, 28].

Hayes et al. [18] demonstrated through subjective perceptual experiments that scene semantic information involuntarily guides visual attention during search tasks. Thus, such information is closely linked to the subjective perception of images. Despite the above-mentioned advancements, existing ICA models, whether based on hand-crafted or CNN-derived features, have largely overlooked the exploration of scene semantic information. In particular, aspects such as object motion, cultural context, mood, spatial position, and quantity—factors that critically reflect the interrelationships among image key visual entities, as illustrated in Fig. 1—have not been effectively integrated into the current ICA modeling frameworks.

However, the significance of scene semantic information has been extensively demonstrated in various computer vision tasks. Specifically, Delaitre et al. [10] leveraged scene semantics by incorporating the distributional patterns of object-related super-pixels over time, capturing how people interact with objects, thereby improving pose estimation and semantic labeling tasks. Wu et al. [43] incorporated object and scene semantics, extracted via pre-trained detectors, into the video understanding pipeline, subsequently fusing these semantic cues to improve large-scale video understanding. Wei et al. [47] integrated textual scene semantic topic words into image captioning to enable the model to generate more accurate and scene-specific captions. Siahaan et al. [40] improved no-reference image quality assessment by using image semantic information (i.e., scene and object categories).

Inspired by the crucial role of scene semantic information in many other visual tasks, we aim to enhance the ICA task by explicitly exploring such information. Although such semantics could theoretically be extracted through pre-trained tasks like image classification or semantic segmentation, the ICA datasets [2, 12, 39] often lack the corresponding ground truth labels for these specific tasks. Thus, the absence of explicit semantic labels makes it challenging to directly incorporate scene semantic information into the complexity assessment process.

Since capturing scene semantic information from the vision-language framework is an efficient way [47]. To address the above gap and enhance the model's focus on scene semantic information, we propose a new ICA method called Cross-Modal Scene Semantic Alignment (CM-SSA). Our proposed CM-SSA adopts a dual-branch framework that refines visual features through image-text alignment guided by scene semantics. The complexity regression branch predicts complexity levels directly using image-prompt pair-wise learning based on complexity categories such as "high complexity" and "moderate complexity". Meanwhile, the scene semantic alignment branch refines image features by aligning them with scene descriptions automatically generated by instruction-tuned vision-language models such as InstructBLIP [9]. As shown in Fig. 1, incorporating the proposed cross-modal scene semantic alignment leads to complexity predictions that are more consistent with human perception.

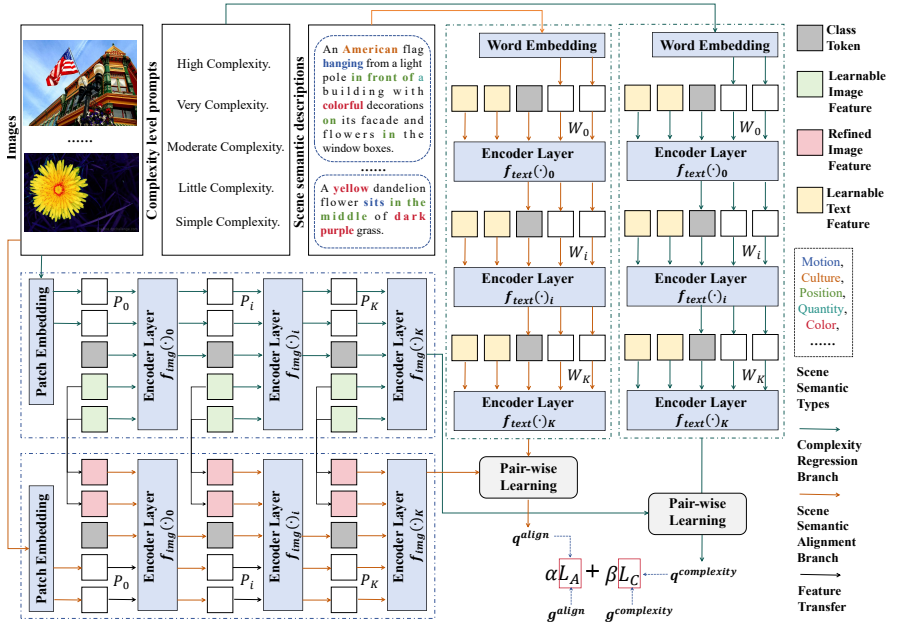

Figure 2: The framework of the proposed **Cross-Modal Scene Semantic Alignment (CM-SSA)**. The hyperparameters α and β control the relative weights of the **complexity regression branch** and **scene semantic alignment branch**. The alignment loss L_A encourages the image features to align with the scene semantic text features, using a ground truth of a vector of all ones (i.e., g^{align}) to ensure maximum alignment. The complexity loss L_C predicts image complexity levels, where the ground truth $g^{complexity}$ corresponds to the MOS (Mean Opinion Score) of image complexity.

In summary, the main contributions of this paper are as follows:

1. We propose the first cross-modal metric, namely Cross-Modal Scene Semantic Alignment (CM-SSA), for image complexity assessment.
2. We introduce a dual-branch framework, including the complexity regression branch and the scene semantic alignment branch, for exploiting cross-modal scene semantic information to assist the complexity prediction.
3. Experiments on several ICA datasets, including IC9600, VISC-C, and SAVOIAS, demonstrate that our proposed CM-SSA outperforms state-of-the-art methods. Ablation studies further validate the effectiveness of each proposed key component.

2 Related Work

Existing ICA methods can be broadly categorized into hand-crafted feature-based and deep learning-based approaches. Among the hand-crafted approaches, Oliva et al. [31] conceptualized visual complexity as a multidimensional representation involving factors such as the number of objects, clutter, openness, symmetry, organization, and variety of colors. Sun et

al. [43] adopted gradient-boosted trees to regress image complexity features like composition, statistical properties, and distribution. Purchase et al. [32] quantified visual complexity by aligning it with participants' subjective perception, concluding that while the subjective concept of "complexity" is consistent across individuals and groups, it is not easily correlated with straightforward computational metrics. Yu et al. [49] found a strong association between spatial information metrics and compression-based complexity measures. Silva et al. [8] investigated the relationship between human eye movements and various computational attention models in the context of ICA. Guo et al. [16] considered how global and local characteristics in paintings influence overall impressions and detailed perceptions. Penousal et al. [29] used edge detection techniques, as well as metrics based on image compression errors, to estimate visual complexity. Additionally, entropy-based ICA methods have also been actively explored. Stamps et al. [41] investigated the relationship between visual diversity and statistical entropy. Ruth et al. [36] proposed three visual clutter metrics (feature congestion, subband entropy, and edge density), which improved the generalization of visual search models and highlighted the crucial role of color diversity in visual clutter estimation. However, these methods rely on predefined features and prior assumptions, limiting their adaptability and semantic understanding.

Recently, deep learning-based ICA methods have exploited neural networks for richer representations. For example, Chen et al. [6] used networks to analyze texture and salient regions, while Abdelwahab et al. [1] and Saraee et al. [38] employed the CNN with SVM or ridge regression to predict the visual complexity. Ivanovici et al. [20] explored the CNN for modeling fractal image complexity using modified ResNet-18. However, the effectiveness and generalization of these models are limited due to the lack of large-scale ICA datasets. Therefore, the IC9600 dataset [17] was introduced as the first large-scale ICA dataset, containing 9,600 images. To avoid the high annotation cost of constructing the large-scale IC9600 dataset, Liu et al. [27] then utilized unsupervised learning based on the MoCo v2 [5] for the representation learning of image complexity. However, these methods often use shallow networks and single-image modality as input, failing to capture high-level semantics and multi-modal cues.

3 Proposed CM-SSA Method

In this section, we introduce our proposed ICA metric, namely the Cross-modal Scene Semantic Alignment (CM-SSA), designed to guide the image encoder in extracting more complexity-related scene semantic information through cross-modal alignment between images and corresponding scene semantic descriptions. Fig. 2 shows the overall framework of the proposed CM-SSA, which consists of two branches: the complexity regression branch and the scene semantic alignment branch. Both branches utilize a cross-modal pair-wise learning between images and texts to fine-tune the contrastive language-image pretraining (CLIP) model. We begin by briefly introducing the image and text encoders of the CLIP model, followed by a detailed explanation of the proposed CM-SSA framework.

3.1 CLIP Model

The CLIP model [33] maps images and text into a shared high-dimensional embedding space for cross-modal alignment. The image encoder, typically based on a Vision Transformer [11] or ResNet [19], processes an input image X^{Img} by tokenizing it into M patches. Each patch

is embedded into a d_v-dimensional latent vector, resulting in a patch matrix $P_0 \in \mathbb{R}^{M \times d_v}$. A learnable class token $c_0 \in \mathbb{R}^{d_v}$ is then added to P_0, and the sequence $[c_0, P_0]$ is fed into the initial transformer block $f_{\text{img}(\cdot)_0}$. The process is recursively repeated $K-1$ times across subsequent blocks:

$$[c_i, P_i] = f_{\text{img}}([c_{i-1}, P_{i-1}])_i, \quad i = 1, \ldots, K. \tag{1}$$

The final image representation $Z^{\text{Img}} \in \mathbb{R}^D$ is obtained by projecting the last class token c_K into a D-dimensional latent space. $f_{\text{img}}(\cdot)$ denotes the overall image encoder. Similarly, the text encoder $f_{\text{text}}(\cdot)$, based on a transformer architecture, tokenizes and embeds the input textual prompt X^{Text} into a sequence of L words, each mapped into a d_l-dimensional embedding. The initial word embedding matrix $W_0 = [w_0^1, \ldots, w_0^L] \in \mathbb{R}^{L \times d_l}$ is then passed through K transformer layers as follows:

$$[W_i] = f_{\text{text}}(W_{i-1})_i, \quad i = 1, \ldots, K. \tag{2}$$

The final text representation $Z^{\text{Text}} \in \mathbb{R}^D$ is obtained by projecting the last token w_K^L into the same latent space as the image representation.

3.2 Cross-Modal Scene Semantic Alignment for ICA

We aim to align images and their corresponding scene semantic information in a shared embedding space, enabling the rich scene semantics in the text to guide the learning of complexity-related features in the image embedding. A straightforward implementation involves fine-tuning the CLIP model by inputting the image and its corresponding scene semantic textual description to directly regress to complexity levels.

Since the IC9600 dataset [12] for the ICA task does not contain scene semantic textual descriptions, we first generate such descriptions X^{Scene} for each input image X^{Img} using the InstructBLIP [9]. Examples of these descriptions include phrases such as "An American flag hanging from a light pole in front of a building with colorful decorations on its facade and flowers in the window boxes", as shown in Fig. 2.

Subsequently, we input the image-text pair $[X^{\text{Img}}, X^{\text{Scene}}]$ into the CLIP model's image encoder and text encoder, respectively, obtaining the corresponding embeddings $[Z^{\text{Img}}, Z^{\text{Scene}}]$. These embeddings are aligned in a shared high-dimensional embedding space through image-text pair-wise learning. Specifically, let the input image set be $\{X_i^{\text{Img}}, i = 1, 2, \ldots, N\}$, where N denotes the number of input images in a batch, and let the corresponding textual description set be $\{X_i^{\text{Scene}}, i = 1, 2, \ldots, N\}$. The embeddings obtained through the image and text encoders can be expressed as:

$$[Z_i^{\text{Img}}, Z_i^{\text{Scene}}] = [f_{\text{img}}(X_i^{\text{Img}}), f_{\text{text}}(X_i^{\text{Scene}})], \quad i = 1, 2, \ldots, N. \tag{3}$$

Inspired by AGIQA [14], we utilized additional learnable prompts for better learning capability, as shown in Fig. 2. The cosine similarity between the image and text embeddings is then computed to facilitate alignment:

$$s_i^{complexity} = \frac{Z_i^{\text{Img}} \odot Z_j^{\text{Scene}}}{\|Z_i^{\text{Img}}\| \cdot \left\|Z_j^{\text{Scene}}\right\|}, i, j \in \{1, 2, \ldots, N\}, \tag{4}$$

and the result is regressed to predict the complexity score:

$$q_i^{complexity} = \frac{e^{s_i^{complexity}}}{\sum_{j=1}^{N} e^{s_j^{complexity}}}, i \in 1, 2, \ldots, N. \tag{5}$$

However, **the experimental results of this direct approach reveal a significant gap from the dual-branch framework illustrated in the ablation study in Table 3.** We believe this is because scene semantic textual descriptions, while reflecting the scene content of images, do not explicitly align with specific complexity levels. Consequently, aligning scene semantic text with complexity predictions is less effective. To address this, we propose the CM-SSA framework consisting of both a complexity regression branch (i.e., Branch_C) and a scene semantic alignment branch (i.e., Branch_A).

In the complexity regression branch, textual prompts of complexity levels $X_k^{\text{Complexity}}, k = 1,2,3,4,5$ (*i.e.,* Simple Complexity, Little Complexity, Moderate Complexity, Very Complexity, High Complexity) are utilized. Following the calculation processes of Eq. (3, 4, 5), the complexity predictions are calculated as follows:

$$\begin{aligned}
&[Z_i^{\text{Img}}, Z_k^{\text{Complexity}}] = [f_{\text{img}}(X_i^{\text{Img}}), f_{\text{text}}(X_k^{\text{Complexity}})], \\
&s_{ik}^{complexity} = \frac{Z_i^{\text{Img}} \odot Z_k^{\text{Complexity}}}{\|Z_i^{\text{Img}}\| \cdot \left\|Z_k^{\text{Complexity}}\right\|}, \\
&q_i^{complexity} = \frac{e^{s_{i1}^{complexity}}}{\sum_{k=1}^{5} e^{s_{ik}^{complexity}}}, i \in \{1,2,\ldots,N\}, k \in \{1,2,3,4,5\}.
\end{aligned} \tag{6}$$

The complexity loss is calculated by the MSE loss:

$$L_C = \frac{1}{N}\sum_{i=1}^{N}\left\|q_i^{complexity} - g_i^{complexity}\right\|_2^2, \tag{7}$$

where $g^{complexity}$ is the MOS of image complexity.

Furthermore, to involve the scene semantic information in refining the image feature learning, we conceive the scene semantic alignment branch that fully aligns images with corresponding scene semantic textual descriptions.

The input to this branch is the image-scene semantic description pair $[X_i^{\text{Img}}, X_i^{\text{Scene}}], i = 1,2,\ldots,N$. The cosine similarity and alignment level between the image and scene semantic text are computed as:

$$\begin{aligned}
&s_i^{scene} = \frac{Z_i^{\text{Img}} \odot Z_j^{\text{Scene}}}{\|Z_i^{\text{Img}}\| \cdot \left\|Z_j^{\text{Scene}}\right\|}, \\
&q_i^{align} = \frac{e^{s_i^{scene}}}{\sum_{j=1}^{N} e^{s_j^{scene}}}, i, j \in \{1,2,\ldots,N\}.
\end{aligned} \tag{8}$$

To ensure full alignment of the image with the high-level scene semantic information, we set the ground truth alignment levels g^{align} to be 1. This follows the image-text alignment dataset [23], where the alignment score is scaled from 0 to 1. Thus, we set the maximum alignment level as 1.

$$L_A = \frac{1}{N}\sum_{i=1}^{N}\left\|q_i^{align} - g_i^{align}\right\|_2^2, \tag{9}$$

The above loss calculation assumes that the textual descriptions generated by InstructBLIP perfectly align with the corresponding images. The final loss is calculated as follows:

$$L = \alpha L_A + \beta L_C, \tag{10}$$

where α and β are the weight hyperparameters that determine the proportion of two branches, which are empirically set to 0.1 and 0.9, respectively.

Table 1: Performance comparisons on the IC9600, VISC-C, and Savoias datasets. The best-performing results are highlighted in bold. The calculation of model Flops is based on the input image shape $3 \times 512 \times 512$, and the inference speed is taken from an average of 100 inferences. * The CLICv2 is pretrained on IC1M dataset and obtained by linear probing on the IC9600. * The Savoias results of DINO-v2 are pretrained on IC9600.

Dataset	IC9600			VISC-C			Savoias			Complexity		
Model	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	Params/M	Flops/G	Speed/ms
CR [29]	0.314	0.228	0.196	-	-	-	0.305	0.271	0.257	-	-	-
ED [16]	0.491	0.569	0.226	-	-	-	0.449	0.467	0.273	-	-	-
DBCNN [38]	0.871	0.879	0.071	0.779	0.783	0.086	0.768	0.770	0.147	15.31	86.22	23.69
NIMA [30]	0.838	0.555	0.194	0.810	0.803	0.125	0.781	0.771	0.210	54.32	13.18	25.00
HyperIQA [32]	0.926	0.933	0.204	0.734	0.739	0.181	0.801	0.798	0.293	27.38	107.83	29.64
CLIPIQA [34]	0.897	0.898	0.078	0.781	0.796	0.122	0.779	0.794	0.101	-	61.07	21.76
TOPIQ [3]	0.938	0.944	0.049	0.803	0.811	0.079	0.838	0.832	0.123	45.20	37.26	14.43
CNet [21]	0.870	0.873	-	-	0.716	-	-	-	-	-	-	-
*CLICv2 [28]	0.927	0.933	-	-		-	-	-	-	-	-	-
*DINO-v2 [1]	0.847	0.851	0.086	-		-	0.675	0.689	-	-	-	-
ICCORN [23]	0.951	0.954	0.048	0.766	0.784	0.332	-	-	-	-	-	-
ICNet [12]	0.937	0.946	0.049	0.818	**0.814**	0.079	0.865	0.849	0.121	20.33	28.40	4.56
CM-SSA	**0.958**	**0.961**	**0.009**	**0.823**	0.805	**0.018**	**0.883**	**0.875**	**0.026**	205.45	52.66	35.43

4 Experiments

4.1 Dataset and Evaluation Metric

We conduct experiments on three publicly available image complexity assessment datasets: IC9600 [12], VISC-C [22], and Savoias [37]. IC9600 is the first large-scale ICA dataset, containing 9,600 images across eight semantic categories, each annotated by 17 trained annotators. VISC-C includes 800 images with corresponding human-rated complexity scores. Savoias comprises over 1,400 images from seven categories, covering a broad range of low- and high-level visual features.

Following ICNet [12], we evaluate model performance using four metrics: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SRCC), Root Mean Squared Error (RMSE), and Root Mean Absolute Error (RMAE). PLCC and SRCC assess correlation with subjective scores, while RMSE and RMAE quantify prediction errors. Together, these metrics offer a comprehensive evaluation of prediction accuracy and consistency.

4.2 Implement Details

We adopt a CLIP model based on ViT-B/32, setting the length of the learnable prompts to 8. The datasets are randomly divided into training and testing sets with an 8:2 ratio. During training, the batch size per iteration is 64, and the model is optimized using the Adam algorithm with a learning rate of 10^{-4} over 50 epochs. In the testing phase, complexity scores are predicted using a patch-based evaluation method [14]. All experiments are implemented in PyTorch on a V100 GPU.

4.3 Performance Comparisons

Following the protocols of the ICNet [12], we conduct a performance comparison of the proposed CM-SSA with two hand-crafted feature-based ICA methods, two deep learning-based ICA methods, and five advanced image quality assessment (IQA) models: The ICA baselines include compression ratio (CR) [29], edge density (ED) [16], CNet [21], CLICv2 [28],

Table 2: Cross-dataset validation. The best performances are highlighted in bold.

Train	IC9600				VISC-C				Savoias			
Test	VISC-C		Savoias		IC9600		Savoias		IC9600		VISC-C	
Model	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
DBCNN	0.685	0.674	0.606	0.626	0.742	0.750	0.525	0.538	0.773	0.787	0.644	0.652
NIMA	0.566	0.297	0.603	0.240	0.714	0.390	0.575	0.540	0.818	0.725	0.630	0.588
HyperIQA	0.711	0.688	0.669	0.669	0.668	0.668	0.550	0.554	0.761	0.748	0.643	0.629
CLIPIQA	0.702	0.686	0.600	0.608	0.406	0.399	0.487	0.481	0.789	0.775	0.550	0.568
TOPIQ	0.759	0.730	0.659	0.648	0.695	0.700	0.594	0.597	0.806	0.804	0.674	0.685
ICNet	0.753	0.712	**0.728**	**0.716**	**0.718**	**0.739**	**0.664**	**0.660**	0.805	0.807	**0.712**	**0.709**
CM-SSA	**0.760**	**0.740**	0.710	0.712	0.577	0.556	0.541	0.486	**0.838**	**0.828**	0.702	0.681

Table 3: Performance comparison of different branch configurations. The text prompts are coarse-grained complexity prompts when the single complexity regression branch Branch_C is utilized, and the text prompts are fine-grained textual prompts when the single text-driven scene semantic branch Branch_A is utilized. The best results are highlighted in bold.

Branch_C	Branch_A	SRCC ↑	PLCC ↑	RMSE↓	RMAE↓
✓	×	0.947	0.951	0.010	0.008
×	✓	0.915	0.922	0.013	0.010
✓	✓	**0.958**	**0.961**	**0.009**	**0.007**

Table 4: Ablation study on the prompt levels in Branch_C. 3 levels refer to ["Simple Complexity, Moderate Complexity, High Complexity"], while 7 levels refer to ["Very Low Complexity, Fairly Low Complexity, Slightly Low Complexity, Moderate Level Complexity, Slightly High Complexity, Fairly High Complexity, Very High Complexity"].

Class Name	SRCC ↑	PLCC↑	RMSE↓	RMAE ↓
3 levels	0.951	0.954	0.010	0.008
5 levels	**0.958**	**0.961**	**0.009**	**0.007**
7 levels	0.955	0.960	**0.009**	**0.007**

DINO-v2 [3], ICCORN [17], and ICNet [12]. The IQA baselines include DBCNN [50], NIMA [44], HyperIQA [42], CLIPIQA [46], and TOPIQ [4].

As shown in Table 1, hand-crafted feature-based ICA methods (CR, ED) underperform significantly across all datasets due to their limited capacity to model high-level semantics. While several IQA methods (e.g., DBCNN, NIMA, and HyperIQA) show moderate performance, they fall short on ICA tasks as they are not specifically designed for complexity perception. In contrast, deep ICA models (ICCORN, ICNet, and CM-SSA) achieve consistently superior results, highlighting the importance of task-specific learning. Notably, our proposed CM-SSA outperforms all other methods across datasets, demonstrating state-of-the-art performance with competitive computational efficiency.

4.4 Cross-Dataset Evaluation

Table 2 presents cross-dataset validation results, evaluating the generalization capability of various models when trained and tested on different datasets. The proposed model consistently demonstrates stronger robustness across most cross-dataset scenarios. For example, when trained on IC9600 and tested on VISC-C, our model achieves the highest SRCC and PLCC values of 0.760 and 0.740, respectively. However, CM-SSA exhibits reduced generalization performance when trained on VISC-C and tested on other datasets. This may be due to the limited scale and diversity of the VISC-C dataset, especially given the relatively large number of parameters in CM-SSA that make it more susceptible to generalizing from a small dataset to larger ones. These results confirm that the proposed method generalizes competitively across diverse datasets with varying characteristics.

Table 5: Ablation study on the prompt models and lengths on IC9600 dataset.

Prompt	SRCC ↑	PLCC↑	RMSE↓	RMAE ↓
BLIP	0.947	0.955	**0.009**	**0.007**
BLIP2	0.952	0.957	0.010	**0.007**
InstructBLIP-Short	0.952	0.956	0.010	0.008
InstructBLIP-Medium	**0.958**	**0.961**	**0.009**	**0.007**
InstructBLIP-Long	0.953	0.957	0.010	0.008

Table 6: Ablation study on the hyper-parameters α and β on IC9600 dataset.

α	β	SRCC ↑	PLCC ↑	RMSE ↓	RMAE ↓
0.1	0.9	**0.958**	**0.961**	**0.009**	**0.007**
0.3	0.7	0.952	0.956	0.010	0.008
0.5	0.5	0.952	0.955	0.010	0.008
0.7	0.3	0.951	0.954	0.010	0.008
0.9	0.1	0.950	0.954	0.010	0.008

4.5 Ablation Tests

- **Validity on branches**. Table 3 presents a performance comparison between the two branch configurations for image complexity assessment. Using only the complexity regression branch yields relatively high performance, indicating its significant role. In contrast, relying solely on the scene semantic alignment branch results in lower performance, highlighting limited standalone effectiveness. Meanwhile, combining both branches yields the best results, emphasizing the complementary nature of the two branches. This confirms that although scene semantic information is less effective for direct complexity regression, it plays a crucial role in refining image features to be more perceptually relevant to scene semantic information.

- **Validity on prompt level in** Branch_C. Table 4 displays the ablation on prompt levels. The results reveal that a finer level of text prompt in Branch_C can lead to better complexity prediction.

- **Validity on prompt length in** Branch_A. Table 5 compares different prompt lengths and models. InstructBLIP outperforms BLIP and BLIP2, likely because BLIP-based models generate shorter prompts. Regarding prompt length, both short and long prompts underperform the medium ones, possibly due to insufficient information in short prompts and hallucinations in long prompts.

- **Validity on hyperparameters**. Table 6 presents the ablation results on the hyperparameters (i.e., α, β in eq. 10). The results illustrate that the Branch_C contributed most to the final complexity regression, while increasing the proportion of α may result in a slight performance drop.

5 Conclusion

In this paper, we present the Cross-Modal Scene Semantic Alignment (CM-SSA), the first cross-modal ICA framework. The proposed CM-SSA addresses the shortcomings of existing ICA metrics, which largely depend on hand-crafted and deep learning-based features limited to single-modality visual information. Specifically, our proposed CM-SSA is a dual-branch architecture. The complexity regression branch leverages pair-wise learning between images and their corresponding complexity level prompts to predict complexity scores. To further enhance the model's understanding of scene semantic information, the scene semantic alignment branch is designed to align image features with scene semantic descriptions. This alignment refines visual prompts with high-level text-driven scene semantics, enriching the complexity assessment process. Experimental results on several ICA datasets demonstrate that the proposed CM-SSA outperforms state-of-the-art methods. Ablation studies further validate the contributions of each branch and the generalization performance across datasets.

References

[1] Mohamed A. Abdelwahab, Abdullah M. Iliyasu, and Ahmed S. Salama. Leveraging the Potency of CNN for Efficient Assessment of Visual Complexity of Images. In *ICIP Theory, Tools and Applications*, pages 1–8, 2019.

[2] Erdem Akagunduz, Adrian G Bors, and Karla K Evans. Defining image memorability using the visual memory schema. *IEEE TPAMI*, 42(9):2165–2178, 2019.

[3] Luigi Celona, Gianluigi Ciocca, and Raimondo Schettini. On the Use of Visual Transformer for Image Complexity Assessment. *Proceedings Copyright*, 640:647, 2024.

[4] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE TIP*, 2024.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[6] Yan-Qin Chen, Jin Duan, Yong Zhu, Xiao-Fei Qian, and Bo Xiao. Research on the image complexity based on neural network. In *International Conference on Machine Learning and Cybernetics*, volume 1, pages 295–300, 2015.

[7] Yanyan Cheng, Huijuan Wang, and Xinjiang Mi. Digital Image Enhancement Method based on Image Complexity. *International Journal of Hybrid Information Technology*, 9:395–402, 2016.

[8] Matthieur Perreira Da Silva, Vincent Courboulay, and Pascal Estraillier. Image complexity measure based on visual attention. In *IEEE ICIP*, pages 3281–3284, 2011.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *NeurIPS*, 2023.

[10] Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. Scene semantics from long-term observation of people. In *ECCV*, pages 284–298, 2012.

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. IC9600: A benchmark dataset for automatic image complexity assessment. *IEEE TPAMI*, 45(7):8577–8593, 2023.

[13] Alexandra Forsythe. Visual Complexity: Is That All There Is? In *Lecture Notes in Computer Science, Engineering Psychology and Cognitive Ergonomics*, page 158–166, Dec 2008. doi: 10.1007/978-3-642-02728-4_17.

[14] Jun Fu, Wei Zhou, Qiuping Jiang, Hantao Liu, and Guangtao Zhai. Vision-language consistency guided multi-modal prompt learning for blind AI generated image quality assessment. *IEEE Signal Processing Letters*, 2024.

[15] Ridhima Grover, Dinesh Kumar Yadav, D.K. Chauhan, and Suraj Kamya. Adaptive Steganography via Image Complexity Analysis using 3D Color Texture Feature. In *International Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity*, page 1–5, Oct 2018.

[16] Xiaoying Guo, Yuhua Qian, Liang Li, and Akira Asano. Assessment model for perceived visual complexity of painting images. *Knowledge-Based Systems*, 159:110–119, 2018.

[17] Xiaoying Guo, Lu Wang, Tao Yan, and Yanfeng Wei. Image Visual Complexity Evaluation Based on Deep Ordinal Regression. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 199–210, 2023.

[18] Taylor R Hayes and John M Henderson. Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin & Review*, 26:1683–1689, 2019.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[20] Mihai Ivanovici, Cosmin Hatfaludi, and Radu-Mihai Coliban. Color fractal texture image complexity estimation using a convolutional neural network. In *International Symposium on Electronics and Telecommunications*, pages 1–4, 2020.

[21] Cameron Kyle-Davidson, Adrian G Bors, and Karla K Evans. Predicting human perception of scene complexity. In *IEEE ICIP*, pages 1281–1285, 2022.

[22] Cameron Kyle-Davidson, Elizabeth Yue Zhou, Dirk B Walther, Adrian G Bors, and Karla K Evans. Characterising and dissecting human perception of scene complexity. *Cognition*, 231: 105319, 2023.

[23] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. *IEEE TCSVT*, 34(8):6833–6846, 2024.

[24] Pu Li, Yi Yang, Wangda Zhao, and Miao Zhang. Evaluation of image fire detection algorithms based on image complexity. *Fire Safety Journal*, 121:103306, 2021.

[25] Wentian Li. On the Relationship Between Complexity and Entropy for Markov Chains and Regular Languages. *Complex Syst*, 5, Jun 2000.

[26] Yixiao Li, Xiaoyuan Yang, Yuqing Luo, Hadi Amirpour, Hantao Liu, and Wei Zhou. Unlocking implicit motion for evaluating image complexity. *Displays*, 90:103131, 2025.

[27] Shipeng Liu, Liang Zhao, Dengfeng Chen, and Zhanping Song. Contrastive learning for image complexity representation. *arXiv preprint arXiv:2408.03230*, 2024.

[28] Shipeng Liu, Liang Zhao, and Dengfeng Chen. CLICv2: Image Complexity Representation via Content Invariance Contrastive Learning. *arXiv preprint arXiv:2503.06641*, 2025.

[29] Penousal Machado, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal. Computerized measures of visual complexity. *Acta Psychologica*, 160:43–57, Aug 2015.

[30] Jon McCormack and Andy Lomas. Deep Learning of Individual Aesthetics. *Neural Computing and Applications*, page 3–17, Dec 2020.

[31] Aude Oliva, Michael L. Mack, Mochan Shrestha, and Angela Peeper. Identifying the Perceptual Dimensions of Visual Complexity of Scenes. In *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society*, pages 1041–1046, 2004.

[32] Helen Purchase, Euan Freeman, and John Hamer. An Exploration of Visual Complexity. In *Lecture Notes in Computer Science*, pages 200–213, July 2012. doi: 10.1007/978-3-642-31223-6_22.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[34] Ilhan Raman, Evren Raman, and Biran Mertan. A standardized set of 260 pictures for turkish: Norms of name and image agreement, age of acquisition, visual complexity, and conceptual familiarity. *Behavior Research Methods*, 46(2):588–595, Jun 2014.

[35] Juan Romero, Penousal Machado, Adrian Carballal, and Antonino Santos. Using complexity estimates in aesthetic image classification. *Journal of Mathematics and the Arts*, 6(2-3):125–136, 2012.

[36] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *Journal of Vision*, page 17, Aug 2007.

[37] Elham Saraee, Mona Jalal, and Margrit Betke. Savoias: A diverse, multi-category visual complexity dataset. *arXiv preprint arXiv:1810.01771*, 2018.

[38] Elham Saraee, Mona Jalal, and Margrit Betke. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195:102949, 2020.

[39] Elham Saraee, Mona Jalal, and Margrit Betke. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, page 102949, 2020.

[40] Ernestasia Siahaan, Alan Hanjalic, and Judith A Redi. Semantic-aware blind image quality assessment. *Signal Processing: Image Communication*, 60:237–252, 2018.

[41] Arthur E. Stamps. Entropy, Visual Diversity, and Preference. *The Journal of General Psychology*, 129(3):300–320, Jun 2002.

[42] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *CVPR*, June 2020.

[43] Litian Sun, Toshihiko Yamasaki, and Kiyoharu Aizawa. Relationship Between Visual Complexity and Aesthetics: Application to Beauty Prediction of Photos. In *ECCV Workshops*, page 20–34, Dec 2014.

[44] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE TIP*, 27(8):3998–4011, 2018.

[45] Alexandre N. Tuch, Javier A. Bargas-Avila, Klaus Opwis, and Frank H. Wilhelm. Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *International Journal of Human-Computer Studies*, 67(9):703–715, 2009.

[46] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, pages 2555–2563, 2023.

[47] Haiyang Wei, Zhixin Li, Feicheng Huang, Canlong Zhang, Huifang Ma, and Zhongzhi Shi. Integrating scene semantic knowledge into image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(2):1–22, 2021.

[48] Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, and Leonid Sigal. Harnessing object and scene semantics for large-scale video understanding. In *CVPR*, pages 3112–3121, 2016.

[49] Honghai Yu and Stefan Winkler. Image complexity and spatial information. In *International Workshop on Quality of Multimedia Experience*, pages 12–17, 2013.

[50] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE TCSVT*, 30(1):36–47, 2020.

[51] Wei Zhou, Hadi Amirpour, Christian Timmerer, Guangtao Zhai, Patrick Le Callet, and Alan C Bovik. Perceptual visual quality assessment: Principles, methods, and future directions. *arXiv preprint arXiv:2503.00625*, 2025.