IC9600: A Benchmark Dataset for Automatic Image Complexity Assessment

Tinglei Feng*, Yingjie Zhai*, Jufeng Yang, Jie Liang, Deng-Ping Fan, *Senior Member, IEEE*, Jing Zhang, Ling Shao, *Fellow, IEEE* and Dacheng Tao, *Fellow, IEEE*

Abstract—Image complexity (IC) is an essential visual perception for human beings to understand an image. However, explicitly evaluating the IC is challenging, and has long been overlooked since, on the one hand, the evaluation of IC is relatively subjective due to its dependence on human perception, and on the other hand, the IC is semantic-dependent while real-world images are diverse. To facilitate the research of IC assessment in this deep learning era, we built the first, to our best knowledge, large-scale IC dataset with 9,600 well-annotated images. The images are of diverse areas such as abstract, paintings and real-world scenes, each of which is elaborately annotated by 17 human contributors. Powered by this high-quality dataset, we further provide a base model to predict the IC scores and estimate the complexity density maps in a weakly supervised way. The model is verified to be effective, and correlates well with human perception (with the Pearson correlation coefficient being 0.949). Last but not the least, we have empirically validated that the exploration of IC can provide auxiliary information and boost the performance of a wide range of computer vision tasks. The dataset and source code can be found at https://github.com/tinglyfeng/IC9600.

Index Terms—Image complexity assessment, image attributes, large-scale well-annotated dataset.

1 Introduction

THE image complexity (IC) is defined as the intricacy contained within an image [1]. Objectively, IC can be considered as the amount of detail and variety in an image [2]. Subjectively, it is the degree of difficulty for a human audience to understand or describe an image [2], [3] regarding both global abstract and local details or textures. For example, as shown in Fig. 1, the plain sketch and open sky are in a lower IC, while the texture of architecture and the crowd of people are in a higher IC relatively. The overall IC of an image is impacted by the combination of such local areas with different IC levels. IC is an important perception in psychology, which can strongly influence the visual aesthetics [4], [5] and affective responses of viewers [6]. It is also a common and significant attribute in computer vision. An automatic prediction of IC has been proven helpful to multiple applications such as image segmentation [7], image steganography [8], webpage design [6], text detection [9], image enhancement [10], *etc*. Therefore, it is necessary and useful to explicitly evaluate the IC to mimic the human perception and facilitate the follow-up tasks.

To achieve this goal, several previous works present some heuristic metrics that can be leveraged for evaluating the IC, *e.g.*, image entropy [11], compression ratio between the original image size and its size of compressed format (*e.g.*, JPEG, GIF) [12], density of edge pixels in the whole image [13], count of the unique RGB colors in an image [14], *etc*. Afterward, machine learning methods such as SVM, random forest, and BP neural network [14], [15] are used to leverage a variety of basic image features to predict the IC. However, these algorithms are mainly based on small-scale datasets and focus on hand-crafted features, hindering their generalization capacity to evaluate the IC in practice.

The evaluation of IC is challenging due to the following reasons: (1) Real-world images are varied in an almost infinite number of patterns and scenarios, thus it is hard to robustly represent their IC based on the combination of hand-crafted features. (2) The IC is a kind of high-level

Fig. 1. Sample images of our *IC9600* dataset in different categories, such as abstract, scene, architecture, *etc*. 'S1'-'S5' denotes the distribution of complexity scores (1-5 point scale) annotated by 17 annotators. The images are ranked by the average scores (top on each image, normalized to $[0, 1]$).

- *Tinglei Feng and Yingjie Zhai contribute equally to this work.*
- *Tinglei Feng, Yingjie Zhai, Jufeng Yang and Deng-Ping Fan are with College of Computer Science, Nankai University. (Email: tinglyfeng@163.com, zhaiyingjier@163.com, yangjufeng@nankai.edu.cn, dengpingfan@mail.nankai.edu.cn)*
- *Jie Liang is with The Hong Kong Polytechnic University. (Email: liang27jie@163.com)*
- *Jing Zhang is with School of Computer Science, Faculty of Engineering, The University of Sydney. (Email: jing.zhang1@sydney.edu.au)*
- *Ling Shao is with Terminus Group, China. (Email: ling.shao@ieee.org)*
- *Dacheng Tao is with JD Explore Academy, China and The University of Sydney, Australia. (dacheng.tao@gmail.com)*
- *Corresponding author: Jufeng Yang.*

(subjective) concept relying on human perception, with a wide gap to those low-level features. Recently, deep convolutional neural networks (CNNs) have shown powerful representation and generalization ability to explicitly model the subjective human perception (*e.g.*, image aesthetic [16], image quality [17], *etc.*) in a data-driven manner. However, existing datasets of IC are all of small-scale with limited diversity since constructing an IC dataset is hard and time-consuming work (*e.g.*, each image needs to be labeled by enough annotators to reduce ambiguity). As shown in Tab. 1, those datasets are either small-scale, not publicly available, or limited to particular topics, which can hardly be suitable to power the effective data-driven and deep-based image analysis methods.

To facilitate the research of the IC in the deep learning era, we built a large-scale dataset termed *IC9600* with $9,600$ images. Each image is annotated by 17 well-trained contributors that are chosen through an elaborate complexity perception test. As shown in Fig. 1, the proposed dataset includes a variety of content categories, *i.e.*, abstract, advertisement, architecture, object, painting, person, scene, and transportation. The most relevant dataset is the SAVOIAS [18], while ours is built with a larger size, more diverse and application-oriented categories, and a more practical annotation solution under the scenario of a large amount of samples. With such diverse topics, we aim to support the training of deep and robust models, and provide comprehensive auxiliary representations to boost a wide range of related tasks.

Furthermore, based on the dataset, we propose a base model, namely *ICNet*, to extract powerful representations of IC and facilitate the other applications. *ICNet* is designed with two branches, *i.e.*, detail branch and context branch. The detail branch utilizes a shallow convolutional network to capture local representations from high-resolution images, while the context branch aims to excavate contextual information from an image with a smaller size via a deeper network. Then the informative features from two branches are combined and sent to the following two heads, of which one predicts an IC score that represents the overall complexity of an image while another outputs an IC map that depicts the local complexity intensity of the image. Experimental results demonstrate the effectiveness of our proposed method.

To further demonstrate the significance of IC in the computer vision, we make great efforts to explore the relations between the prior complexity knowledge and some specific tasks (*e.g.*, image aesthetic assessment [4], crowd counting [19], salient object detection [20], *etc.*) to improve their performance. Extensive experiments on multiple datasets indicate that IC can provide auxiliary information and we can effectively boost the performance of six sub-tasks by applying the IC in proper ways.

Our contributions can be summarized as follows:

- We built currently the largest well-annotated IC dataset, which addresses the urgent need of a large-scale database for IC assessment. Our proposed dataset contains $9,600$ images across 8 semantic categories. Each sample is annotated by multiple people to reduce the personal bias.
- We provide a baseline model to predict the image complexity scores and estimate the complexity density map of images in a weakly supervised manner. The model is designed to have two separate branches to extract detail and context features respectively. Moreover, we propose a spatial layout attention module to further improve the *ICNet*.
- We apply the proposed model to several computer vision tasks, which gives a preliminary exploration for the usage of IC in the deep learning era. Experiment results demonstrate that IC can be used as a primary image attribute to improve the performance of many tasks in computer vision.

2 Related Work

2.1 Image Complexity Analysis

Researchers have investigated the factors that influence the human perception of IC in psychology [2], [21], [22]. Oliva *et al.* [23] characterized the representation of IC as the number of objects, openness, clutter, symmetry, organization, and variety of colors. Forsythe [1] argued that familiarity is an important factor that influences the perception of IC, *e.g.*, observers tend to rate familiar shapes as less complex than they actually are. Purchase *et al.* [24] conducted an empirical study to investigate whether the IC could be quantified and if it could match participants' views of complexity. The study shows that it is challenging to define an explicit metric that adequately captures the human perception of IC.

Many algorithms have been developed to assess image complexity automatically [11], [25]. Some of them try to quantize IC with entropy. Stamps [21] investigated the relationship between the IC and the stimulus feature of entropy and finds the correlation between them is strong and linear. Based on the relation between image clutter and visual information, Rosenholtz *et al.* [26] proposed feature congestion and subband entropy to evaluate the IC. Machado *et al.* [12] claimed that simple images tend to have more redundant information while values of pixels in complex images are less predictable. In other words, simple images can normally be compacted to a smaller size, thus introducing compression ratio to describe the IC is practicable. Even though entropy is proven to correlate with IC, they can only roughly evaluate the amount of information an image carries. More specific descriptors (*i.e.*, hand-crafted features) will be needed to assess IC more precisely. In [27], the main factors affecting human visual perception are defined as the distribution of compositions, colors, and contents. Thus they designed 29 local, global, and salient region features to represent the above three factors. Apart from these, edge density [12], spatial information [28], visual attention [25], *etc.*, are also used as common metrics for calculating IC. Further, several machine learning methods have been leveraged to combine hand-crafted features to model the IC. For example, Sun *et al.* [4] adopted gradient boosted trees to regress the complexity features of composition, statistics, and distribution. Chen *et al.* [15] used the backpropagation neural network to establish the relation between IC and three features, *i.e.*, texture, edge, and region. Recently, Abdelwahab *et al.* [29] extracted features using a

pre-trained CNN and then use SVM to predict the complexity level. Similarly, Saraee *et al.* [18] proposed to predict image complexity with Ridge regression where the input image representations are extracted from pre-trained CNNs. Moreover, in the same work, an unsupervised activation energy (UAE) method is also developed to model image complexity based on the activations from the intermediate layer of deep neural networks.

Even though various methods have been proposed to investigate the IC, most of these methods are based on handcrafted features and traditional machine learning methods. A few recent methods employ the off-the-shelf pre-trained CNNs to extract image features, but do not train them in an end-to-end manner due to the lack of large-scale paired training data. Therefore, these methods can hardly correlate well with the high-level perception of IC. Moreover, due to the lack of benchmark datasets, most existing works conduct experiments and demonstrate performance on individual datasets that are constructed by themselves. As a result, the comparison might be biased. These datasets are of small-scale or have subjective bias (*i.e.*, each sample is annotated by limited observers), which may not get convincing conclusions. Therefore, a large-scale and high-quality IC benchmark is urgently needed for fair comparisons and to boost the development of IC assessment.

2.2 Subjective Visual Attributes Assessment

Image complexity, as well as many other image attributes, *e.g.*, image quality [34] and image aesthetic [35], is a subjective concept which varies from person to person. Here, we mainly investigate the successes of image quality assessment (IQA) and image aesthetic assessment (IAA) tasks since they have the most similarity to the IC assessment.

To obtain labels with low personal bias for subjective visual attributes, averaging the perception from multiple humans is verified to be currently a great practice in the area of IQA and IAA. Ponomarenko *et al.* [36] proposed a widely used IQA dataset named TID2008. The mean opinion scores (MOS) indicating image quality are collected from $256k$ individual human quality judgments. The recently proposed KADID-10k dataset [37] consists of more images and distortion types, and each image receives 30 degradation category ratings by crowdsourcing. For the IAA task, CUHK-PQ [38] is the earlier large-scale IAA dataset. Each image is annotated by ten viewers and is assigned a binary label that indicates high image aesthetic or not. Following this work, AVA [39] built by Murray *et al.* is currently the largest IAA dataset, in which each image is voted with scores from 1 to 10 by multiple people. Driven by these proposed datasets and supervisions of average human perception, many CNN-based models have been proposed to accurately assess the image quality or aesthetic. For example, Kang *et al.* [40] proposed a network consisting of one convolutional layer, two fully-connected layers, and one output node. This model is proved to be more effective for IQA than the traditional methods based on handcraft features. Bianco *et al.* [41] proposed the DeepBIQ model that predicts sub-regions scores and then averages them to estimate the image quality. Considering that the shape transformation such as resizing, cropping, or padding may damage the image aesthetic, Mai *et al.* [42] introduced a composition-preserving network that can directly process input images with original size and extract multi-scale features at the same time. Zhang *et al.* [43] introduced a gated peripheral-foveal CNN to mimic the human perception mechanism of aesthetic. This model can encode both the holistic information and fine-grained features.Besides, image cropping [44], group maximum differentiation competition [45], [46], and self-supervised feature learning [47] are also verified to be effective approaches for IAA or IQA. Following the existing successful experiences of IQA and IAA, we build the IC9600 benchmark dataset and propose a CNN-based model to facilitate the research of IC assessment.

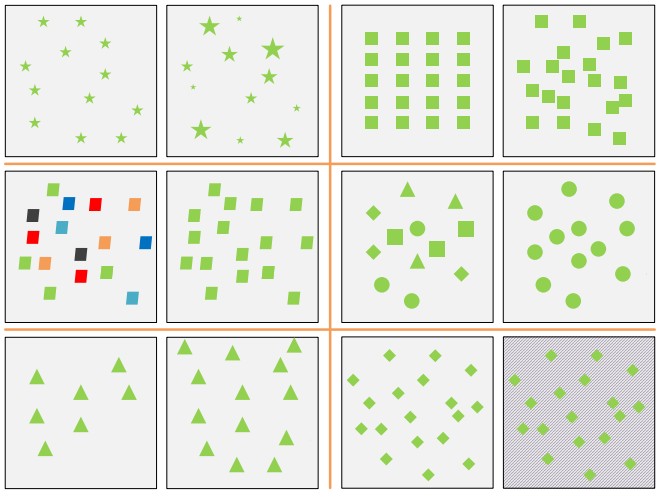


Fig. 2. Several sample pairs of our perception test. The two images in each pair are manually set to be in different IC levels. The candidates are asked to choose the more complex image in each pair.

2.3 Image Complexity Datasets

As shown in Tab. 1, several small datasets have been used for IC analysis. Specifically, Oliva *et al.* [23] collected 100 pictures of indoor scenes, and then annotate them by a three-times dichotomy (splitting each group into simple and complex groups). Iliyasu *et al.* [30] built the Corel 1000A dataset, in which images are annotated into three categories (*i.e.*, simple, normal, and complex). To study the automatic complexity metrics of graphical user interfaces (GUI), Miniukovich *et al.* [31] gathered screenshots of 140 webpages, and then ten graduate students are asked to rate the complexity on a 1-5 point scale. Afterward, Corchs *et al.* [32] collected 98 images of scenes and 122 real texture images to predict the complexity perception of real-world images. Recently, to identify factors that affect the IC perception of paintings, Fan *et al.* [33] constructed a complexity dataset containing 40 Chinese ink paintings. Guo *et al.* [14] collected 500 painting images with a 1-7 point scale. Further, Saraee *et al.* [18] created a dataset SAVOIAS with over $1,000$ images and unbiased ground truth labels for the IC analysis.

The SAVOIAS is the most similar to ours but our dataset is built with the following key significance. First, the size of our dataset is nearly 7 times larger than the SAVOIAS. Second, our dataset covers more diverse topics (8 versus 7) and these topics are more dedicated to real-world applications. Third, the annotations of our dataset cross the

TABLE 1
Overview of the current datasets for image complexity assessment. The acronym in line7 and line8 represent Abstract(Abs), Advertisement(Adv), Architecture(Arc), Art(Art), Id(Interior design), Obj(Object), Pai(Painting), Per(person), Sce(Scene), Sup(Suprematism), Tra(Transportation), Vi(Visualization and infographics).

#	Datasets	Year	Size	Image Type	Annotation Type	Public Available
1	Oliva *et al.* [23]	2004	100	Indoor scenes	1-8 point scale	No
2	Corel 1000A [30]	2013	1,000	Objects	Three categories	No
3	Miniukovich *et al.* [31]	2014	140	Webpages	1-5 point scale	No
4	Corchs *et al.* [32]	2016	220	Real-world scenes (98), real texture (122)	0-100 score	No
5	Fan *et al.* [33]	2017	40	Chinese ink painting	1-7 point scale	No
6	Guo *et al.* [14]	2018	500	Painting images	1-7 point scale	No
7	SAVOIAS [18]	2020	1,420	Adv, Art, Id, Obj, Sce, Sup, Vi	Pair-wise comparison	Yes
8	**IC9600 (Ours)**	2021	**9,600**	Abs, Adv, Arc, Obj, Pai, Per, Sce, Tra	1-5 point scale	Yes

whole dataset regardless of semantic categories while the ground truth scores in SAVOIAS are only comparable within the same category, which limits its further applications in general image complexity analysis. Note that the existing complexity-related datasets are all of small sizes and most of them are unavailable to the public. Therefore, it is necessary to build a large-scale dataset with diverse real-world scenarios to address the urgent need from the deep learning for the assessment of IC.

TABLE 2
The averaged PCC of each pair of groups with different settings. The equation $M \times N + K$ (1^{st} row) means we split the 17 annotations into $N+1$ groups, of which N groups contain M annotations for each and 1 group contains K annotations. The annotations in each group are averaged into one annotation. For the $N+1$ groups, we compute PCC for all the C_{N+1}^2 pairs. And the 2^{nd} row is the mean value of the C_{N+1}^2 PCCs.

Groups	1×17	$2\times7+3$	$3\times4+5$	$4\times3+5$	$5\times2+7$	$8\times1+9$
PCC	0.54	0.68	0.77	0.82	0.86	0.94

3 Proposed Dataset

3.1 Image Collection

• **Image Resources.** Our dataset contains eight categories including abstract, advertisement, architecture, object, painting, person, scene, and transportation. To build such a diverse dataset, we initially collect images for each category from several popular datasets. Specifically, we select abstract and architecture images from AVA [39], advertisement images from Image and Video Advertisements [48], object images from MS-COCO [49], painting images from JenAesthetics [35], person images from WiderPerson [50], scene images from Places365 [51], and transportation images from BDD100K [52].

• **Sampling Strategy.** To further improve the diversity of each semantic category, we choose images from each dataset to contain sub-categories as many as possible. For example, the Places365 [51] dataset contains 365 scenario categories. We randomly select 4 images from each scene category and get a total of $1,460$ images for the scene category. The sampling strategies of other categories are similar to this. After the sampling process, we get around $1,500$ images for each of the eight categories.

• **Removal of inappropriate images.** Note some images sampled from different datasets are identical copies or near-duplicate. Thus we remove the near-duplicates using the Image Deduplicator[1] tool. Then we filter out the images that have too many watermarks or are of low quality. After a multi-round checking and selection, we finally get a total of $9,600$ images.

3.2 Image Annotation

• **Annotation Guidance and Test Questions.** To ensure the quality of annotations, we elaborately select, train, and test the annotators as follows. First, we select all the candidates from vision-related university laboratories. Second, we train them with a detailed tutorial including the purpose of the study and the basic concept of IC. Third, to verify their abilities to distinguish different IC levels, we conduct a test consisting of a number of pairs of artificial images that are simulated with simple geometries (*e.g.*, triangle, square, and circle). Samples of these test pairs are shown in Fig. 2. The two images in each pair are manually set to be in different complexity levels based on the IC factors including texture, shape, object arrangement, *etc*. We finally get 20 qualified annotators with top performance in the test (all of them scoring at $90\%+$ accuracy). To guarantee the ability of annotators to distinguish multiple levels of IC, we carefully choose images in 5 complexity levels (*i.e.*, very simple, simple, medium, complex, and very complex) for each category according to the basic attributes of IC. We select 5 images for each complexity level. The annotators are asked to observe the difference between multiple complexity levels and take them as references when they annotate the images.

• **Annotation Environment.** Each annotator is asked to be in a quiet room without any interference from other people. They are required to pay attention to the annotation task and make their phones mute.

• **Annotating.** Following [4], [16], we ask each participant to annotate each image by using 1-5 point scales, with the complexity degrees ranging from very simple (*i.e.*, 1) to very complex (*i.e.*, 5). Each annotator should annotate 320 images in one day, which needs a total of 30 days. And they should observe an image for more than 10 seconds to have an accurate perception of the image. If they feel tired, they should stop the annotation and have a break. Besides, each time before their annotating, they are asked to review the guidance of the IC concept and multiple IC levels. Note that the pairwise comparison method applied in SAVOIAS dataset [18] is not suitable for our dataset since: (1) Unlike SAVOIAS (separately labels around 200 images for each category), our dataset is much larger ($9,600$ images). Therefore, the workload of pair-wise comparison for each annotator will exponentially increase, which is beyond the limits of our capability. (2) The multi-point Likert scale

1. https://github.com/idealo/imagededup

is verified to be reliable in many previous works related to the evaluation of subjective image attributes like image quality assessment [34] and image aesthetic assessment [16]. Additionally, the subjective bias from the multi-point Likert scale can be well reduced by the averaging strategy.

• **Outliers.** Following the previous work [53], annotators who have a very low agreement with other annotators are outliers. Specifically, for each annotator, we compute his average Pearson correlation coefficient (PCC) with other annotators. Finally, 3 annotators whose average Pearson correlation coefficient is lower than 0.4 are removed as outliers, and 17 annotations are used to calculate the final complexity score.

• **Subjectivity Removal.** We leverage the widely-used average strategy to reduce the subjectivity of annotators. Specifically, after annotating, the label set corresponding to the j-th sample can be denoted as $\{y_j^1, y_j^2, y_j^3, \cdots y_j^m\}$, where $y_j^i \in \{1, 2, 3, 4, 5\}$ and m is the total number of labels assigned to j-th image. The final complexity score l^j is acquired by averaging the m labels and is then normalized to $[0, 1]$, *i.e.*, $l^j = \sum_{i=1}^{m}(y_j^i - 1)/4m$. The final score acquired from the averaging strategy can effectively eliminate the bias caused by subjectivity. To verify it, we split the 17 annotations into several groups, and then compute the PCC (It is the ratio between the covariance of two variables and the product of their standard deviations, the formula can be found at 6) between each pair of groups. As shown in Tab. 2, with the number of annotations in each group increasing, the PCC is improved as well. And when we divide all the annotations into two groups, the PCC of them reaches 0.940. This phenomenon proves that the subjectivity can be reduced by averaging annotations from more observers. In addition to this calculated complexity score, we will also provide the label distribution of each sample for the community to research more characteristics of IC.

3.3 Properties of Dataset

• **Human Rating Consistency.** The perception of visual attributes, *e.g.*, IC, might be different across people due to the subjective nature of human perception, yet will be stable under a distribution of multiple trials. Here, we make an attempt to demonstrate the reliability of the subjective annotation of multiple annotators. Following the common standards [54], [55], we leverage the Pearson correlation coefficient, Spearman's correlation coefficients, and Kendall's tau correlation between each pair of annotations, and evaluate their statistical significance of the correlation with respect to a null hypothesis of uncorrelated responses. Results show that the average PCC, Spearman's correlation coefficient (SRCC), and Kendall's tau correlation are 0.54, 0.53, and 0.48, and at a significance of 0.01, the p-value is less than 0.01 for all pairs, which demonstrates the consistency between annotators. Besides, similar to the crowdsourcing assessment studies for image quality assessment and image aesthetic assessment [53], [56], we calculate the intra-class correlation coefficient (ICC) of our annotation. ICC is the most widely used indicator to measure the inter-rater reliability. A high ICC shows that most variance originates from differences in the images, but not by individual differences in the evaluations by the annotators, thus indicating a high

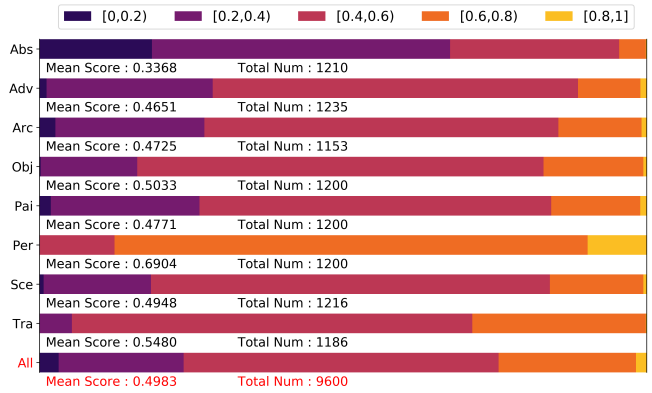


Fig. 3. The distribution of annotations of the proposed *IC9600* dataset. We equally divide the scores from 0 to 1 into five intervals and compute the ratio of the scores lying within these intervals for each category (three-letter acronym). The distribution is shown in the above-stacked bars with the corresponding mean score and total number below. The same information of the overall dataset is also shown in the last row.

degree of consistency among annotators. We use the same one way random model of ICC as the [53], [56]. Experiment shows the ICC is 0.518, which is better than the results of [53] (0.46) and [56] (0.403). It also demonstrates the reliability and consistency of our annotation.

• **Distribution of Annotations.** All images are randomly divided into a training and a testing set in a ratio of $0.7 : 0.3$. For each of the eight semantic categories, the score distribution is shown in Fig. 3. The category distribution of our dataset is relatively balanced since each category contains $\sim 1,200$ images. However, the distribution of complexity in different semantic categories varies from each other. For the *abstract* category, the number of scores below 0.2 is larger than any other category, which results in a minimum mean score. It is reasonable because the contents of the abstract image are mostly simple geometries. Note that images in the *person* category are derived from WiderPerson [50], which is a dense pedestrians dataset. The people in the pictures are presented in a crowded and disordered manner, making the pictures very complicated. Thus, the mean score of the *person* category is much higher than others and none of the scores are lower than 0.4. This explanation can also be adapted to the high scores of the *transportation* category. As shown in the last row of Fig. 3, we can see that nearly half of the images are assigned to medium scores (*i.e.*, $[0.4, 0.6)$), and the whole dataset presents a symmetric Gaussian distribution centered at around 0.5, which reflects the IC distribution in the real world. Additionally, we calculate the mean and variance of annotations for each sample and split them into five intervals according to their mean values. The variance distributions of each split area are shown in the boxplot of Fig. 4. We can observe that the samples with the complexity score between 0.8 and 1.0 have relatively higher variances, which indicates that IC in this area may be slightly harder for human beings to explicitly distinguish. Nevertheless, most of the variances are lower than 0.04, which proves the high annotation consistency of our dataset. We hope such a diverse dataset can help researchers explore more general IC attributes of images and more broad applications of IC in various computer vision tasks.

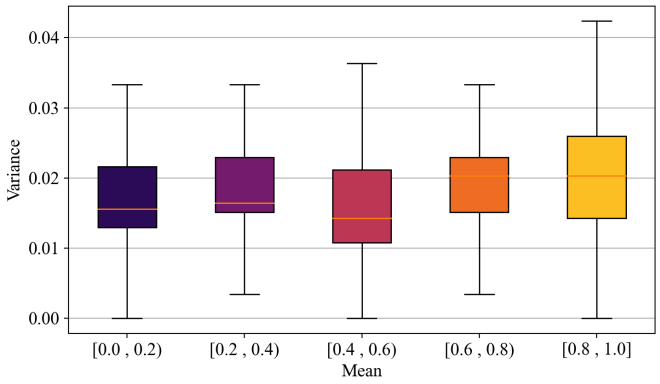


Fig. 4. Distribution of variances. All the samples are split into five intervals according to their mean values. To make the boxplot clearer, the few outliers are omitted.

4 Predicting Image Complexity

In order to set a baseline for the following research and verify the reasonability of the proposed dataset, we design a base model, namely *ICNet*, in this paper. As shown in Fig. 5, *ICNet* consists of two branches for feature extraction. The detail branch captures spatial and detailed feature representations while the global branch encodes the context and high-level information. The two kinds of features are then concatenated and sent to the following two heads for IC map and IC score predictions respectively. Moreover, each intermediate feature is refined by a spatial layout attention module (SLAM) that is specifically designed for IC feature scaling, and thus generates more effective representations. We will detail the *ICNet* next.

4.1 Two-Branch Extractor

The IC is a basic image attribute that depends on both the low-level representations and the high-level semantic information in the whole image. It indicates that the designed model is expected to have the ability to excavate the two kinds of essential features from an image at the same time. Given a deep convolutional neural network (DCNN), Zeiler *et al.* [58] project the activations of features back to the pixels of the input image. They visually demonstrate that the features from shallow layers of DCNN are normally activated by simple patterns like edges, corners, angles, *etc*. In contrast, the responses of deeper layers are basically determined by more abstract semantic information like a dog's face. Based on the empirical phenomenon of DCNN, we propose a two-branch network for feature extraction.

Both the two branches of our model are modified from a ResNet18 [57]. More specifically, we separate the ResNet18 into four stages. The first stage downsamples the image to $1/4$ resolution feature maps and the downsampling factors of the following three stages are $1/8$, $1/16$ and $1/32$ respectively. Both the adaptive average pooling and fully-connected layer at the end of ResNet18 are dropped. Note that all the stages are pre-trained on ImageNet [59] dataset, the initial model has a strong ability to capture general image features. As for the context branch, it comprises all four stages. The image fed into it is of low resolution (*i.e.*, 256×256). In this way, this head with deep layers yields a relatively large reception field compared to the small input size. Thus it can encode the context and high-level abstract feature representations. In contrast, the detail branch aims to capture detailed spatial information, thus it borrows only the first two stages of ResNet18 and accepts an image with a large size (*i.e.*,512×512). The downsampling factor of this branch is only $1/8$, which can produce high-resolution feature maps with the size of 64×64.

To combine the two feature maps generated from each branch, we adopt a concatenation operation on the channel dimension. Note that the spatial dimensions of the two concatenated maps need to be identical, which implies that we should either upsample the maps from the context branch or downsample the maps from the detail branch. In this case, the spatial information encoded in the detail features is shown to be an important factor of IC in the following sections. Thus we upsample the context maps and concatenate them to fully utilized both the low-level and high-level features while preserving the spatial details.

4.2 Spatial Layout Attention Module

In general, images with more textures, edges, items, *etc*., are deemed more complicated. Thus a common practice is averaging the extracted features to a vector and sending it to the multi-layer perceptron (MLP) followed by a sigmoid function at the end to predict the complexity score. In this way, the vector only encodes the average intensity of each feature in the image. It loses the essential factor, *i.e.*, the spatial layout of features, that largely determines IC. As shown in Fig. 6, the two images with ground truth scores shown below are picked from the *abstract* category in our dataset. It is obvious that the right image has more colors and lines, but the left image is annotated with a higher IC score. The main difference between the two images is the spatial layout of image elements, which makes the perception difference of IC. More specifically, the image on the left is formed with winding and disordered lines while which on the right image are uniform and consistent. Therefore, the annotators tend to give a higher score to the left image.

Inspired by the above observation, we propose an attention module that can adaptively scale the activation intensity by excavating the spatial layout information, termed spatial layout attention module (SLAM). As shown in Fig. 5, given a feature maps $\boldsymbol{F} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$, we first flatten it along the spatial dimension into a 2-D map $\boldsymbol{F_{fl}} \in \mathbb{R}^{\mathrm{C}\times(\mathrm{HW})}$. For any single channel of index i, the vector $\boldsymbol{F_{fl}^{i}} \in \mathbb{R}^{\mathrm{HW}}$ encodes the activations of the feature embedded in i_{th} channel at each spatial location. To learn the way of how the layout of features affects the IC, we send $\boldsymbol{F_{fl}}$ to a MLP, resulting in a vector $\boldsymbol{s} \in \mathbb{R}^{\mathrm{C}}$. This operation is independent of channel dimension, which means for an index of channel dimension i, the input is $\boldsymbol{F_{fl}^{i}}$, and the output is a scalar s^i. The MLP layer consists of two linear layers, each of which is followed by an activation function, which can be denoted as:

$$s^i = \sigma_1(\boldsymbol{W_1} \times \sigma_0(\boldsymbol{W_0} \times \boldsymbol{F_{fl}^{i}} + \boldsymbol{b_0}) + \boldsymbol{b_1}), \tag{1}$$

where the σ_0 and σ_1 are ReLU and sigmoid respectively, $\boldsymbol{W_0} \in \mathbb{R}^{(\mathrm{HW})\times 512}$ and $\boldsymbol{W_1} \in \mathbb{R}^{512\times 1}$ are the weights of the two linear layers, of which $\boldsymbol{b_0} \in \mathbb{R}^{512}$ and $\boldsymbol{b_1} \in \mathbb{R}^{1}$ are the biases. Since the spatial layout is independent of different features, thus the weights and biases of the MLP are shared through the channel dimension.

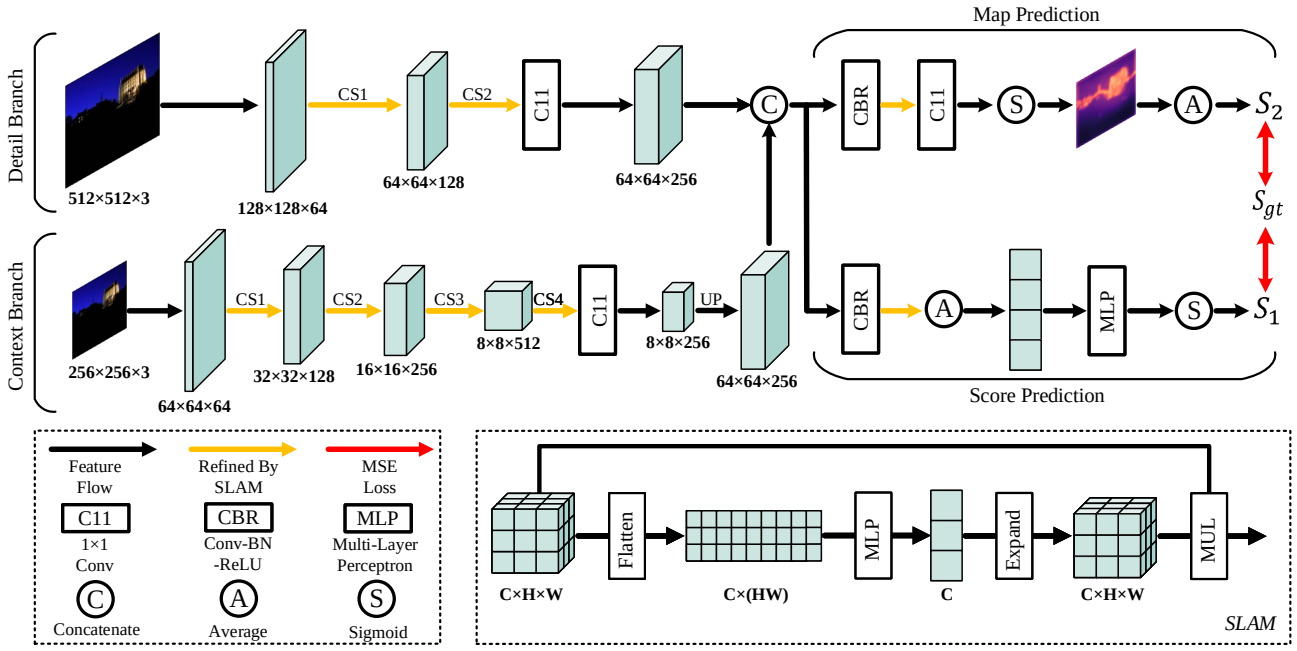


Fig. 5. Pipeline of our proposed *ICNet*. The model consists of a shallow and a deep extractor modified from a ResNet18 [57]. The "CS(N)" on the arrow stands for N_{th} convolutional stage of ResNet18. The detail branch captures spatial and low-level features from high-resolution image, while the context branch fed a smaller image extracts context and high-level representations. The two kinds of features are then concatenated and sent to the following two heads for map prediction and score prediction. Besides, the feature maps with an orange arrow behind them are refined by our proposed spatial layout attention module (SLAM), which can help to scale features according to their spatial layout and produce more effective representations for IC assessment.

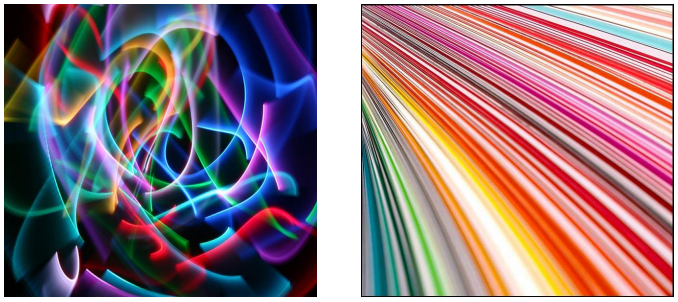


Fig. 6. Two images picked from the *abstract* category of our dataset. The number under each image is the ground truth complexity score. Even though the image on the right contains more colors and lines, the left image is annotated with a higher score, due to the irregular and disordered layout of lines in it.

Through (1), the MLP has the flexibility of evaluating how much the spatial layout of i-th feature in $\boldsymbol{F}$ affects the IC, *i.e.*, s^i. If the layout of this feature is regular and uniform, then we should suppress this feature. In contrast, the feature should be amplified when of which the layout is disordered and messy. To satisfy this requirement, we expand $\boldsymbol{s}$ on the spatial dimension to the same size of $\boldsymbol{F}$, resulting in $\boldsymbol{S}$. Then a simple element-wise multiplication is operated on the two 3-D tensors to produce the final output $\boldsymbol{O}$. The overall operations of our proposed SLAM can be denoted as:

$$\boldsymbol{O} = \boldsymbol{F} \cdot o_3(o_2(o_1(\boldsymbol{F}))), \tag{2}$$

where o_1, o_2, and o_3 denote flattening, MLP, and expanding operators respectively.

We add the SLAM behind the feature maps followed by an orange arrow shown in Fig. 5. Note that we downsample the feature maps whose size is of above 32×32 to 32×32 to reduce the computation cost.

4.3 Predicting Complexity Score and Map

In this paper, we explore two types of IC modalities. One is the IC score that describes the overall complexity of the whole image while another is a complexity map depicting the complexity intensity of the local area.

To produce the two kinds of IC modalities, we send the concatenated feature maps to the following two heads, *i.e.*, the map prediction head and the score prediction head. In order to fuse and balance features in different levels, we set a Conv-BN-ReLU block in the entrance of both the two heads. Afterward, a SLAM is employed to scale feature in each channel according to its spatial layout. To predict the global IC score S_1, the feature maps refined by SLAM are then sent to a global average pooling layer, yielding a feature vector as the input of the following MLP layer and sigmoid function.

For the map prediction head, feature maps from SLAM are followed by a 1×1 convolution and a sigmoid function behind, which projects them to a single-channel map, or in another word, the complexity map. However, a problem arises here, *i.e.*, similar to the segmentation tasks, a ground truth IC map is needed to evaluate the pixel-level regression loss for backpropagation. But it is hard to annotate such an IC intensity map. To overcome this dilemma, we propose a simple weakly-supervised method by learning the complexity map from only the ground truth IC score S_{gt}. We average the generated complexity map to a scalar S_2 and then calculate the distance between it and S_{gt}. This way, this head can implicitly learn to predict the local IC intensity so

TABLE 3
Comparison with 10 traditional methods and 4 deep-based methods (we train and test them **on our proposed dataset**). ↑ (↓) represents the larger or smaller is better, respectively. The methods with superscript U denote that they are unsupervised methods. The 'N' in the table means that this metric is not applicable to the UAE method.

	Methods	Metrics			
		PCC ↑	SRCC ↑	RMSE ↓	RMAE ↓
Traditional	IC^U [60]	-0.006	0.053	0.343	0.552
	CR^U [12]	0.228	0.314	0.196	0.405
	FC^U [32]	0.459	0.439	0.342	0.558
	EN^U [32]	0.479	0.458	0.385	0.600
	SE^U [32]	0.534	0.498	0.136	0.327
	NR^U [32]	0.556	0.541	0.188	0.394
	ED^U [14]	0.569	0.491	0.226	0.427
	AR^U [61]	0.571	0.481	0.234	0.445
	HOG+SVR [62]	0.689	0.689	0.118	0.299
	SIFT+SVR [63]	0.885	0.861	0.069	0.242
Deep-based	UAE^U [18]	0.651	0.635	N	N
	SAE [18]	0.865	0.860	0.074	0.240
	AlexNet [59]	0.924	0.920	0.064	0.222
	ResNet18 [57]	0.935	0.928	0.061	0.222
	HyperIQA [17]	0.935	0.935	0.067	0.229
	P2P-FM [55]	0.940	0.936	0.056	0.208
	ICNet (Ours)	**0.949**	**0.945**	**0.053**	**0.205**

as to minimize the distance between the average prediction and the ground truth.

During training, we optimize the distance between the predicted score and ground truth, which is calculated by the mean square error (MSE) loss:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{j=1}^{N}(S_1 - S_{gt})^2, \tag{3}$$

$$\mathcal{L}_2 = \frac{1}{N}\sum_{j=1}^{N}(S_2 - S_{gt})^2, \tag{4}$$

where N is the total number of samples in a batch, and the overall loss is computed from $\mathcal{L}_1$ and $\mathcal{L}_2$:

$$\mathcal{L} = \lambda \times \mathcal{L}_1 + (1-\lambda) \times \mathcal{L}_2, \tag{5}$$

where λ controls the weight of the two-part losses.

5 Experiments and Results

5.1 Experimental Settings

• **Implementation Details.** We implement the proposed model based on the PyTorch [64] framework using two NVIDIA GTX 1080TI GPUS. We use the pre-trained model on the ImageNet [59] dataset to initialize the parameters of each stage of the two-branch extractor. The mini-batch (batch size is 64) stochastic gradient descent (SGD) is used to optimize the model. The momentum is set to 0.9 and the weight decay is 0.001. We set the initial learning rate to 0.05 and divide it by 5 every 10 epochs. The overall training time is about 1 hour for 30 epochs. The model is trained using the default training-test split as proposed above. All the training images are augmented by random horizontal flipping. Besides, we set λ to 0.9 to get balanced performance between the two IC modalities. As for the evaluation, we use the S_1 as the final score prediction.

• **Evaluation Metrics.** We use Pearson correlation coefficient (PCC) [32], Spearman's correlation coefficient (SRCC) [16], root mean square error (RMSE), and root mean absolute error (RMAE) to evaluate the methods.

PCC is defined as:

$$\rho(X,Y) = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{\sum_{i=1}^{N}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{N}(Y_i-\mu_Y)^2}}, \tag{6}$$

where X and Y represent the predicted scores and the corresponding ground truth subjective scores. μ_X and μ_Y are the mean of X and Y. N is the total image number. SRCC is computed from:

$$\rho' = 1 - 6\frac{\sum_{i=1}^{N}(r_i - r_i')^2}{N^3 - N}, \tag{7}$$

where r_i and r_i' represent the rank of the i-th item when predicted scores and ground truth scores are sorted in descending order. Besides, RMSE is calculated by:

$$r(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2}, \tag{8}$$

and RMAE is represented by:

$$m(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i|}. \tag{9}$$

• **Contenders.** We compare the proposed method with several traditional methods and deep-based methods. Traditional methods include image colorfulness (IC) [60], compression ratio (CR, JPEG format) [12], feature congestion (FC) [32], entropy(EN) [32], subband entropy (SE) [32], number of regions (NR) [32], edge density (ED) [14], autoregressive model (AR) [61], support vector regression with histogram of orientated gradients (HOG+SVR) [62], [65], and support vector regression with scale-invariant feature transform (SIFT+SVR) [63]. Deep methods contain the classic models (*i.e.*, AlexNet [59] and ResNet18 [57]). Besides, we compare with two other methods (*i.e.*, HyperIQA [17] and P2P-FM [55]) used for image quality assessment. Further, the methods proposed in [18] that are specifically designed for IC assessment based on CNNs are also added to our comparisons, including the Unsupervised Activation Energy (UAE) method and the Supervised Activation Energy (SAE) method. The UAE method averages the feature maps from the intermediate activation layer (*e.g.*, ReLU) of pre-trained models to a single activation score representing the complexity of an image. Since the averaged activation can only represent relative scores within images in the same category, only the PCC and SRCC are employed to evaluate this method. Similarly, the SAE method also first extracts features from intermediate layers of pre-trained CNNs, where the difference is that these features are then sent to a Ridge regression model and trained with the supervision of ground truth complexity scores. Besides, according to [18], the VggNet yields the best performance compared with other architectures such as ResNet, DenseNet, EfficientNet, *etc*. Therefore, our reproduced experiments are conducted on VggNet. Since [18] does not provide the details of which layer they use for features extractions, we report our results by choosing features that produce the best performances. We train and test these methods on our proposed dataset

TABLE 4
Comparison with 10 traditional methods and 4 deep-based methods (we train them **on our proposed dataset** and test **on the small-scale SAVOIAS dataset [18]**).

	Methods	Metrics			
		PCC ↑	SRCC ↑	RMSE ↓	RMAE ↓
Traditional	IC^U [32]	0.230	0.243	0.290	0.485
	CR^U [12]	0.271	0.305	0.257	0.452
	FC^U [32]	0.430	0.456	0.259	0.454
	EN^U [32]	0.448	0.466	0.375	0.567
	SE^U [32]	0.352	0.345	0.261	0.454
	NR^U [32]	0.580	0.595	0.244	0.438
	ED^U [14]	0.467	0.449	0.273	0.460
	AR^U [61]	0.497	0.485	0.261	0.454
	HOG+SVR [62]	0.380	0.350	0.253	0.447
	SIFT+SVR [63]	0.704	0.695	0.185	0.382
Deep-based	UAE^U [18]	0.763	0.763	N	N
	SAE [18]	0.750	0.750	0.189	0.394
	AlexNet [59]	0.819	0.818	0.183	0.387
	ResNet18 [57]	0.843	0.845	0.177	0.380
	HyperIQA [17]	0.826	0.831	0.184	0.390
	P2P-FM [55]	0.836	0.842	0.179	0.383
	ICNet (Ours)	**0.866**	**0.868**	**0.176**	**0.379**

using the default training-test split. For algorithms without released codes, we reproduce the codes according to their papers.

5.2 Quantitative Results

We show the evaluation results of different methods in Tab. 3. From this table, we can draw the following conclusions. First, previous methods based on hand-crafted features can hardly outperform those deep-based methods. Most of them have a low PCC under 0.6 while deep methods are all over 0.9. We can observe the PCC of IC in the first row of Tab. 3 is very close to 0, revealing that the change of colors in an image may not be a crucial factor of image complexity. Other methods like CR, FC, NR, *etc*., are shown to have better correlations with complexity. Note that each of them is only from a single component view of IC, they can be used to assess IC in a specific perspective, but are not comprehensive metrics to thoroughly measure this abstract concept. Among these traditional methods, handcrafted feature of SIFT with a support vector regression yields the best performance. It indicates that SIFT with the ability to extract a wide range of spatial features at different scales correlates well with IC, but being only a local feature descriptor limits its further performance. Second, with the support of our large-scale and high-quality dataset, CNNs can extract high-level representations that model the human perception of IC better than low-level features. This fact has been proved by the high performance of deep-based methods shown in Tab. 3. Even the vanilla DCNNs simply modified from a classification network (*i.e.*, AlexNet and ResNet) outperform the traditional methods by a big margin.

Third, our model exceeds all compared methods in terms of the four metrics, which proves the superiority of *ICNet*. The HyperIQA and P2P-FM perform worse than our proposed method, because these methods specifically designed for image quality assessment overlook some factors, *e.g.*, detail and context information, that influence the human perception of IC. The proposed *ICNet* can avoid this problem and model IC from the views of both detail and context, thus achieving favorable results. Note that while both the UAE and SAE methods show significant performance gaps from our method, the results are quite reasonable. First, the UAE method does not employ any human assessment but purely relies on the assumption that the activations of complexity areas are normally higher. Even though this hypothesis may be partially correct, the produced IC scores are averaged from whole activation maps that do not take the elements layout of an image into consideration, which is critical for IC assessment proved by the experiments of our paper, thus can not excavate high-level IC factors. For the SAE method, due to the lack of training data in SAVOIAS dataset [18], the authors only employ pre-trained DCNN for feature extractions and linear Ridge regression for score predictions, where the DCNN is fixed in the whole process, which limits its potential learning abilities.

Note that for comparisons within unsupervised methods or between them and supervised methods, the PCC and SRCC evaluating relative IC activations are more representative than RMAE and RMSE due to dissimilar IC intensity calibrations for different methods. Nevertheless, for PCC and SRCC we can observe from Tab. 3 that unsupervised methods are normally inferior to supervised methods, which makes sense since the methods supervised by our large-scale and well-annotated dataset are given the huge advantages of progressively learning to capture more broad IC factors, which are more effective and complete than those relying on partial and pre-defined features from limited perspectives, thus yield favorable performances.

Besides, to verify the powerful generalization ability of our model, we also provide the results of different methods on the recent SAVOIAS [18] dataset in Tab. 4. Note this dataset is small-scale and the ground truth scores are separately annotated for each category (with a total of seven categories). Thus, we train each method on all the samples of our dataset, test them on the SAVOIAS, and report the mean results of seven categories. Results show that the proposed *ICNet* can also exceed the compared methods, which proves the significant generalization ability of our model.

We also provide the results of our *ICNet* evaluated on each semantic category in our proposed dataset, as shown in Fig. 7. We can observe that the performance of abstract category is relatively lower on each metric. We conjecture that it may come from the wide variety of content in *abstract* images, which makes it hard for the model to predict consistent results. Besides, SRCC of *Person* is obviously the lowest among the eight categories, while in terms of RMSE and RMAE, its performance is relatively better. It is reasonable since the complexity distribution of *Person* presented in Fig. 3 are mostly on a high complexity interval, *i.e.*, $[0.6, 0.8)$. Hence it will be easier to predict a concrete score, but explicitly predicting the complexity rank of each image is more difficult than other categories, resulting in the lowest SRCC. On the contrary, the distribution of *Painting* images in Fig. 3 is symmetrical across five complexity intervals, thus it acquires the best performance of SRCC.

To make it easier to understand the IC, we show some visualization results predicted by our model in Fig. 8. We can find that the predicted scores are very close to the ground truth scores. The right image of each pair is the complexity map predicted from the detail branch. For visualization, we upsample the predicted map to the size of

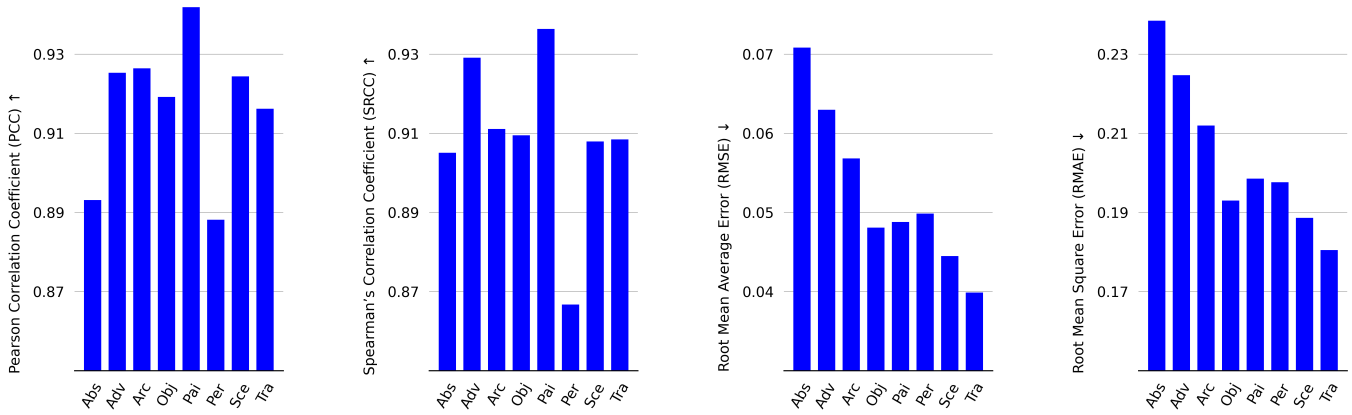


Fig. 7. Performance of *ICNet* on each semantic category in our proposed dataset using the metrics PCC, SRCC, RMSE, and RMAE.

TABLE 5
Ablation analysis of two branches and SLAM.

Settings	Metrics			
	PCC ↑	SRCC ↑	RMSE ↓	RMAE ↓
Only detail branch	0.929	0.927	0.063	0.222
Only context branch	0.939	0.933	0.061	0.219
Both branches	0.944	0.939	0.058	0.215
Both branch + SE [66]	0.944	0.939	0.059	0.218
Both branch + CBAM [67]	0.943	0.942	0.059	0.217
Both branch + BAM [68]	0.944	0.938	0.057	0.214
Both branch + SLAM	**0.949**	**0.945**	**0.053**	**0.205**

the image by the bilinear interpolation and then α-blend it with the input image. From these maps, we observe that our model can explicitly find the visually complex regions in an image. We can also find that the complex areas mostly focus on the positions that have a large amount of objects, textures, edges, variations, *etc*., and are hard for a person to explicitly describe. These maps have great potential to give the models (machines) guidance for understanding the complexity distribution in an image and may be applied to a variety of tasks, *e.g.*, image cropping, automatic drive, image generation, advertisement designing, *etc*., in the future. The last row of the figure shows two kinds of failure cases. The first and second samples have many local textures so that the model tends to predict a higher score. Most of the area of the third example is blank, thus the model predicts a lower complexity score. Collecting more images of such kind samples may help to address these failure cases in the future. Besides, we plot the training dynamics of the predicted map in Fig. 9. We can observe that in the earlier period, the complexity heat map is uniformly distributed on the entire image regardless of its content since the initial model has limited direct knowledge of IC. With the progress of training, the model is supervised to predict the global complexity from averaging local areas, which indirectly requires fine-grained predictions for each pixel. Therefore, there emerge high complexity areas revolving around foreground objects, and the boundaries between low and high complexity pixels are refined from coarser to smoother. In the end, from learning the entire dataset with thousands of images, the model is instructed to learn the patterns of constructing a reasonable complexity map to minimize the loss between averaged activations and ground truth scores, thus the predicted complexity heat map can precisely reflect the pixel-level complexity of the whole image.

5.3 Ablation Study

We investigate the effectiveness of each component in our proposed model. The results are presented in Tab. 5. When only using the detail branch, we get a PCC of 0.929, which is lower than only using the context branch (PCC is 0.939). Even though the detail branch is fed with a high-resolution image, it mostly captures low-level and spatial information while the context branch can extract abstract and high-level representations from small-scale image. And the results indicate that the context information is more crucial in IC evaluation. Nevertheless, they are both crucial cues for people to perceive IC. Therefore, when we combine the two branches, the performance can be improved to 0.944 (PCC). Moreover, when we insert the proposed SLAM into the intermediate layers, the spatial layout information of features is taken into consideration to refine features, thus the PCC is further improved to 0.949. We have also investigated other attention mechanisms by replacing SLAM with squeeze and excitation (SE) [66], convolutional block attention module (CBAM) [67], and bottleneck attention module (BAM) [68] respectively. The results shown in Tab. 5 are similar to just using only the context and detail branches. We speculate that these attention mechanisms are designed for general feature extraction but fail to take account of the intrinsic traits of IC, thus suffering from poor generalization in the IC assessment task. While the proposed SLAM is specifically designed for the IC assessment, it utilizes the spatial layout information and can further improve the performance.

6 Applications and Discussion

In this section, we introduce several possible applications of IC in multiple computer vision tasks. Applications of IC in three ways and six sub-tasks are demonstrated. We have also discussed more broad potential applications of image complexity applying to many other areas outside of computer vision or image processing.

6.1 As an Auxiliary Task

Multi-task learning is a general and intuitive idea, its reliable improvement for deep learning models has been proved in many influential works [69], [70]. Its main idea is to add more supervision information related to the main task and optimize them simultaneously to help the model

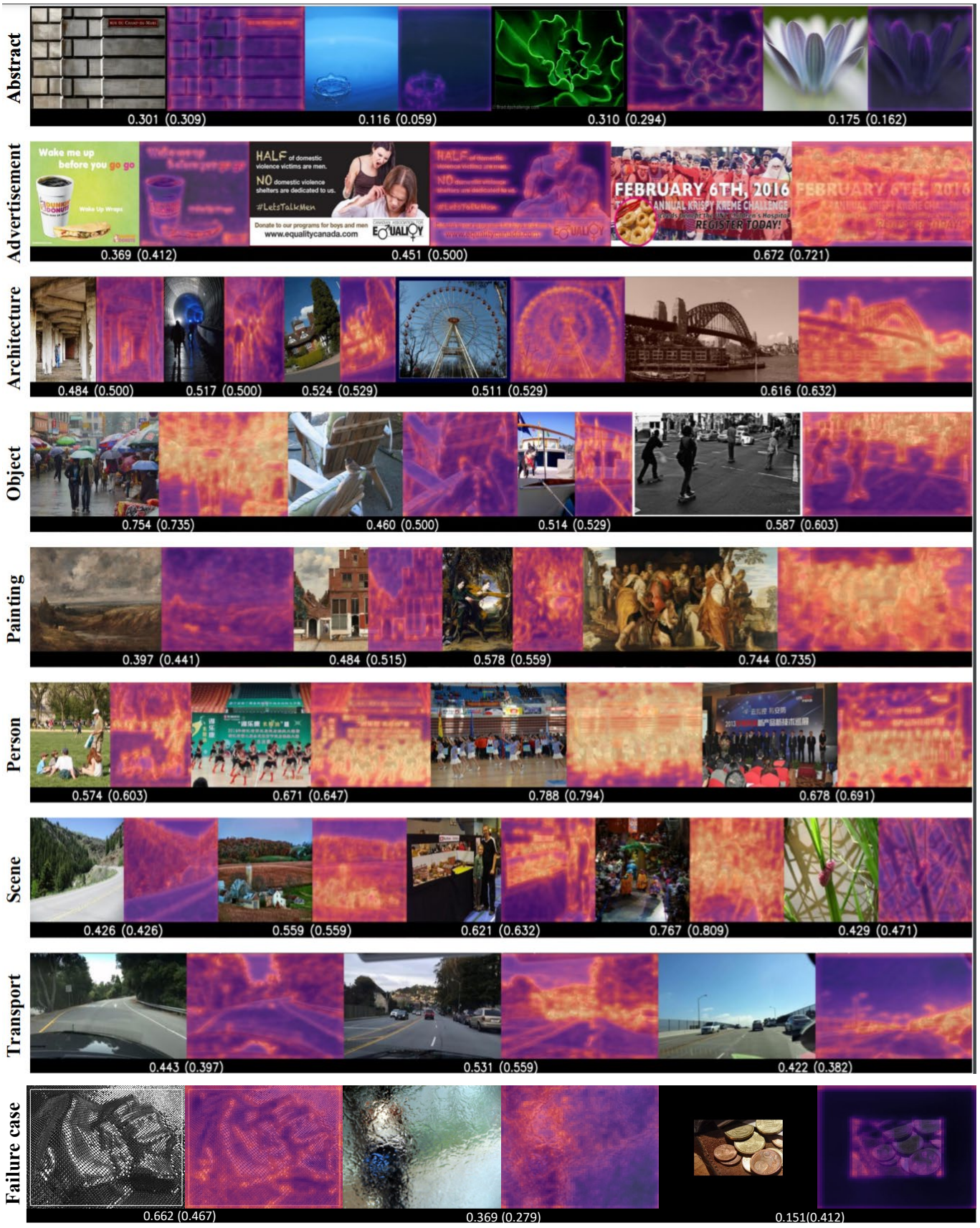


Fig. 8. Visualization results of some images from our test set. The left (right) image of each pair is the input image (predicted complexity map). The number in the bracket represents the predicted complexity score from our model while the number outside the bracket is its ground truth score (normalized to $0-1$) labeled by the annotators. The last row shows several failure cases, of which the predicted score is relatively far from the ground truth score.

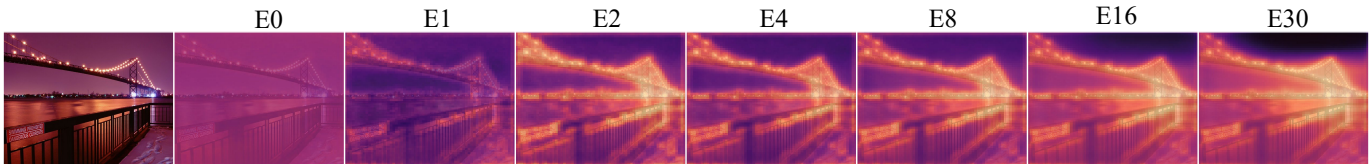


Fig. 9. Training dynamics of the predicted maps. 'E' represents the 'Epoch'. Initially, The model in the first epoch (E0) can hardly predict reasonable IC for each pixel. After several epochs of training, local areas with intricate textures and irregular layout have shown higher activation while the clean and flat background (*e.g.*, sky) is assigned lower IC score. In the following training, the whole IC heat map has been progressively refined to be smoother and finer and presents clear distinctions between simple and complex areas.

learn more robust features, thus improving the generalization ability. Here, we treat the complexity assessment as an auxiliary task, and generate the supervision signals automatically using our trained *ICNet*. We validate the effectiveness of adding the auxiliary complexity assessment sub-branch on four vision tasks.

• **Image Aesthetic Assessment.** IC has been verified to be an critical indicator for aesthetic assessment by previous works [71]–[73]. They find that complexity has negative impacts on the appraisal of aesthetics. Motivated by this, we try to improve the performance of image aesthetic assessment by collaboratively optimizing this task with the IC predicting. Specifically, in our experiment, the base model is set to a common ResNet18 network by changing the last fully-connected layer to output the aesthetic score or aesthetic classes. The multi-task model is modified from the base model by adding another branch behind the last average pooling layer to predict the IC score. For training the multi-task model, we generate the ground truth complexity score of each image by using the proposed *ICNet* trained on our dataset. We train and test the base model and multi-task model using the same experimental settings on the AADB [16] and CUHKPQ [74] dataset. Both the pipelines of the base model and multi-task model are shown in Fig. 10. Experimental results are shown in the first row of Tab. 6. We can observe that both the PCC and SRCC of AADB and CUHKPQ datasets are improved by 1-2 percent, which demonstrates the effectiveness of IC in improving the performance of image aesthetic assessment by leveraging such an auxiliary branch.

• **Image Quality Assessment.** Image quality assessment is also a subjective task that is closely related to human perception. The degree of IC correlates a lot with the perceptual image quality. For example, the complexity of an image gets higher when high-frequency noise is injected, while it drops as low-frequency blur is introduced [61]. Thus, we also apply the IC to the image quality assessment on the LIVEC [34] and KADID [37] dataset in a multi-task training way. The experiments (ResNet18 backbone) are similar to that of the above image aesthetic assessment. The pipeline of the base model and multi-task model is shown in Fig. 10. Experiment results are shown in the second row of Tab. 6. Similar to image aesthetic assessment, when we collaboratively optimize the two tasks, the backbone is forced to extract more general and robust features so that they can satisfy the demands of both the two tasks. Thus when we introduce another IC predicting branch, the performance of IQA can be also improved.

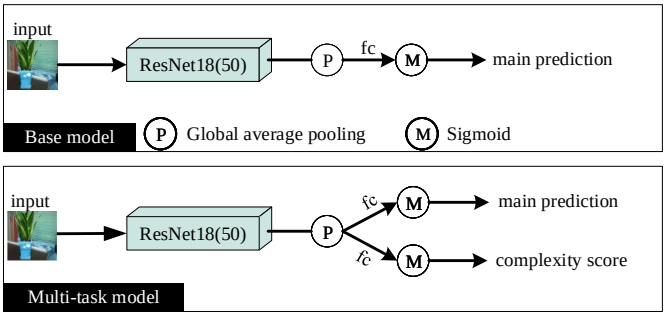


Fig. 10. Illustration of the base and multi-task model for the tasks of image quality assessment, image aesthetic assessment, and image classification.

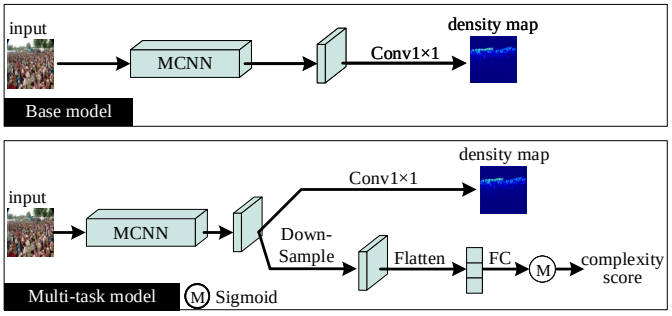


Fig. 11. Illustration of the base model (MCNN) and multi-task framework for the task of crowd counting.

• **Image Classification** The common image classification tasks may also benefit from being supervised by IC scores. Since IC is sometimes a discriminative feature of images belonging to different categories, learning to excavate effective features for IC assessment also contributes to the classification task. We conduct experiments on the Tsinghua Dogs Dataset [75] using the ResNet50. As shown in Fig. 10, we add a subbranch at the end of ResNet50 to regress an IC score and backpropagate the MSE error that measures the distance between the predicted score and the ground truth score. The improvement of employing the additional IC information is shown in the fifth row of Tab. 6.

• **Crowd Counting.** Crowd counting aims at estimating the crowd count from an individual image. It is obvious that the more people the picture contains, the more complex the picture is. Thus the features that determine the complexity of an image can be also utilized for crowd counting. For this task, we conduct experiments using the Multi-column CNN (MCNN) [19]. MCNN consists of three parallel columns and predicts the crowd density map (task1) using the merged

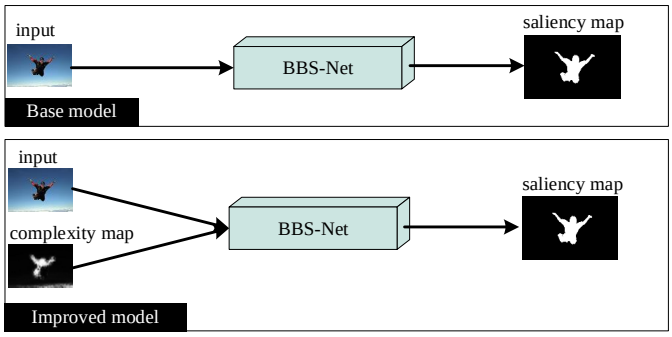


Fig. 12. Illustration of the base model (BBS-Net) and the improved model for the task of salient object detection.

feature maps from these columns. We improve the MCNN by inserting another head (task2) behind the merged feature maps. The head contains a downsampling layer to reshape the feature map to a small size, a flatten layer to flatten the feature map, and a fully-connected layer to predict the complexity score. In this way, the model can be collaboratively optimized by the two tasks. Here the ground truth complexity score is generated by our *ICNet*. The multi-task model uses the same training settings as the base model. The pipeline of training the original MCNN and multi-task MCNN is shown in Fig. 11. The training framework is modified from [76] and the evaluated WorldExpo10 [77] and UCF QNRF [78] dataset are also acquired from it. Results on the third row of Tab. 6 show that the parallel complexity predicting task can reduce the error of crowd counting (with lower MAE and RMSE).

6.2 As a Complexity Modality

The proposed model *ICNet* can generate a pixel-level complexity map from the map prediction head. Such a fine complexity map can be considered as a kind of modality that provides guidance for models to understand the local complexity degree of an image.

• **Salient Object Detection.** IC has a vital influence on the salient object detection [20]. We can observe from Fig. 8 that the areas with high complexity are usually largely overlapped with salient objects. Also, it is usually easy to separate foreground objects from background regions for visually simple images, while it seems hard to segment the salient objects for complex images. To verify whether the IC information can help to find salient objects, we make two experiments on the DUTS [79] and PASCAL-S [80] dataset. The BBS-Net [84] is chosen as our base model. It is designed to find salient objects from the RGB modality and depth modality by feeding them into two branches respectively. Since BBS-Net provides an efficient and general multi-modality extracting and fusing strategy. The original depth channel can be simply replaced by other modalities. For the base model, we make the input of the depth branch be zeros. To leverage the complexity modality, we make the input of the depth branch to be the complexity map generated by our *ICNet*. We then train and test the two models using the same settings. The pipeline of this task is shown in Fig. 12. Tab. 6 shows that the model with complexity modality input (*i.e.*, 'Ours' in the table) has a higher value of max F-measure and lower MAE value, compared with the base model. It proves that the IC modality helps to segment salient objects by providing prior guidance for the model in the areas that are simple or complex in an image.

6.3 As a Prior Weight

It is intuitive that a visually complex image may be hard to recognize or segment. To address this, we can set high weights for the complex images or the complex local areas to make the model focus more on them.

• **Image Classification.** Here, we attempt to consider the IC score of an image as the prior knowledge to represent the degree of difficulty for the model to correctly classify it. Since image classification is an image-level task, we utilize the image-level complexity score to weight each sample. We conduct an experiment using the ResNet50 [57] on the food-101 dataset [81]. Specifically, for the base model, we use the cross-entropy loss to optimize the model, the loss is defined as $\ell_{ce} = -(\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j \ln p_i^j)/N$, where N is the total image number, C is the total categories. If j is the ground truth label, $y_i^j = 1$, otherwise $y_i^j = 0$. p_i^j is the output of the final softmax layer. While for the improved model, we modify the loss to $\ell'_{ce} = -(\sum_{i=1}^{N} w_i \sum_{j=1}^{C} y_i^j \ln p_i^j)/N$, where w_i represents the weight of sample i, and is the same as the complexity score predicted by the proposed *ICNet*. In this way, the prior complex images will be given higher loss weights to optimize them. Results in the fifth row of Tab. 6 show that we can effectively improve the performance of classification task by leveraging the prior IC weight.

• **Image Segmentation.** For image segmentation, we give pixel-level weights for each image according to the IC map generated by our model (*i.e.*, high complexity areas are given higher weights) when calculating the losses. We conduct experiments on the PASCAL VOC2012 [82] and CityScapes [83] dataset. The representative Deeplabv3 segmentation model [85] is employed for comparison. For the base model, the pixel-wise cross entropy loss mask $\boldsymbol{L}$ for an image is defined as $\boldsymbol{L} = -\sum_{j=1}^{C} \boldsymbol{G_j} \odot \ln \boldsymbol{Y_j}$, where C is the total classes, $\boldsymbol{G_j}$ and $\boldsymbol{Y_j}$ are the ground truth map and predicted map for the class j, $\odot$ means element-wise product. We back propagate the average loss of the mask $\boldsymbol{L}$, denoting as $\ell = \frac{1}{H}\frac{1}{W}\sum_{i=1}^{H}\sum_{j=1}^{W} \boldsymbol{L}_{ij}$, where H and W are the height and width of the mask. For the improved model, the loss mask $\boldsymbol{L'}$ is calculated by: $\boldsymbol{L'} = -\boldsymbol{W} \odot \sum_{j=1}^{C} \boldsymbol{G_j} \odot \ln \boldsymbol{Y_j}$, where $\boldsymbol{W}$ represents the complexity map produced by *ICNet*. The last row of Tab. 6 shows that using the IC map can help to improve the performance of image segmentation.

6.4 Applications Outside of Computer Vision

We believe that except for what we mentioned above, image complexity can be applied to more broad areas. And we have thoroughly reviewed potential applications from a wide range of related works.

• **Psychology.** Understanding visual complexity (VC) serves as a medium of studying the underlying human perception. Gestalt psychology (gestaltism) [86] originated from discovering the connections between sensory input and perceptual complexity intensity roots VC as the foundation of revealing how the human brain perceives experiences. Researches from the single visual form, visual array, to visual display [87] in this area are mostly boiled down to

TABLE 6
IC boosts the performance of a variety of vision tasks. 'maxF' and 'Acc' represent the max F-measure and accuracy. 'Base'= baseline methods. 'Ours' = Base + IC information.

Tasks	Datasets	Metrics	Base	Ours	Dataset	Metrics	Base	Ours
Image Aesthetic Assessment	AADB [16]	PCC ↑ SRCC ↑	0.702 0.693	**0.713** **0.705**	CUHKPQ [74]	ACC↑	0.856	**0.878**
Image Quality Assessment	LIVEC [34]	PCC ↑ SRCC ↑	0.842 0.806	**0.851** **0.818**	KADID [37]	PCC ↑ SRCC ↑	0.706 0.724	**0.730** **0.745**
Crowd Counting	WorldExpo10 [77]	MAE ↓ RMSE ↓	19.33 28.64	**16.83** **25.04**	UCF-QNRF [78]	MAE↓ RMSE↓	276 425	**250** **386**
Salient Object Detection	DUTS [79]	maxF ↑ MAE ↓	0.855 0.043	**0.865** **0.039**	PASCAL-S [80]	maxF ↑ MAE ↓	0.893 0.071	**0.899** **0.068**
Image Classification	Food-101 [81]	ACC↑	0.814	**0.834**	Tsinghua Dogs [75]	ACC↑	0.804	**0.818**
Image Segmentation	VOC2012 [82]	mIoU ↑ pixACC ↑	0.594 0.858	**0.610** **0.872**	CityScapes [83]	mIoU ↑ pixACC ↑	0.612 0.923	**0.623** **0.925**

excavating and understanding the internal VC mechanisms. Besides, VC has been proven to affect a wide variety of psychologic areas, including attention and emotional systems [88], mechanisms of visual pattern encoding [89], and visual memorability [18], *etc*.

• **Arts.** Visual complexity largely dictates the appraisal of arts. For instance, a strong positive linear correlation between complexity and building appearance has been found in [90]. Gartus *et al.* [91] suggest that complexity largely influences abstract patterns through dimensions of quantity and structure. Besides, relationships between aesthetics of drawing, painting, photography *etc*., and VC are still under broad studies [4], [5], [92], [93]. Therefore, an accurate IC evaluation tool will benefit a wide area of arts in helping assess intrinsic arts value.

• **Webpage and Advertisement Design.** Numerous studies prove that complexity plays a critical role in the webpage and advertisement design. Pieters*et al.* [94] find that dense visual feature complexity in advertisements hurts customers' attention and attitude towards the brand. Similarly, excessive complexity of background in live streaming is also shown to have negative impacts on individuals' purchase intention [95]. Besides, complexity has shown more significance in users' satisfaction when shopping with mobile devices [96]. In addition, simplicity and clearness have become modern webpage design trends since less complexity is demonstrated to have more attractions [6], [97]. All of the above studies imply that VC should be carefully controlled in commercial activities.

• **Discussions.** The applications we listed above are mostly outside computer vision and cross a wide range of areas. Among them, an automatic and reliable IC assessment method is urgently needed since : First, most researchers assess IC by employing traditional methods such as edge detection or image compression, which can not precisely reflect comprehensive image complexity, and thus may result in biased conclusions. Second, without reliable IC assessment tools, the complexity of an image in some research is mostly collected from multiple human ratings, which is inflexible, expensive, and time-consuming, thus obstructing large-scale IC applications. To address the above issues, we believe our high-quality dataset and reliable IC assessment model will bring a huge favor to boost further IC research and potential applications.

7 Conclusion

In this paper, we introduce the challenging and long overlooked problem of image complexity assessment. We first address the most critical data deficiency problem by building a large-scale benchmark dataset consisting of $9,600$ carefully labeled images from diverse categories. Based on this dataset, we then provide a baseline model, called *ICNet*, to evaluate the complexity score of images, which can achieve a high PCC with the human perception. Additionally, we make an attempt to apply the complexity evaluation model to six tasks and experimental results demonstrate that IC can help to improve their performance. We hope the proposed dataset, model, and exploration of applications can encourage and promote the further research of IC.

Acknowledgements: This work is supported by the National Key Research and Development Program of China Grant (NO.2018AAA0100400), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), NSFC (NO.61876094, U1933114, 61929104), and Fundamental Research Funds for the Central Universities. Dr. Jing Zhang is supported by the Australian Research Council project FL-170100117.

References

[1] A. Forsythe, "Visual complexity: is that all there is?" in *International Conference on Engineering Psychology and Cognitive Ergonomics*, 2009.
[2] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity." *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6, no. 2, pp. 174–215, 1980.
[3] C. Heaps and S. Handel, "Similarity and features of natural textures." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, no. 2, pp. 299–320, 1999.
[4] L. Sun, T. Yamasaki, and K. Aizawa, "Relationship between visual complexity and aesthetics: application to beauty prediction of photos," in *European Conference on Computer Vision*, 2014.
[5] J. McCormack and A. Lomas, "Deep learning of individual aesthetics," *Neural Computing and Applications*, vol. 33, no. 1, pp. 3–17, 2021.
[6] A. N. Tuch, J. A. Bargas-Avila, K. Opwis, and F. H. Wilhelm, "Visual complexity of websites: Effects on users' experience, physiology, performance, and memory," *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 703–715, 2009.

[7] F. Meng, H. Li, K. N. Ngan, L. Zeng, and Q. Wu, "Feature adaptive co-segmentation by complexity awareness," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4809–4824, 2013.

[8] R. Grover, D. K. Yadav, D. Chauhan, and S. 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*, 2018.

[9] M. Li and M. Bai, "A mixed edge based text detection method by applying image complexity analysis," in *World Congress on Intelligent Control and Automation*, 2012.

[10] C. Yanyan, W. Huijuan, and M. Xinjiang, "Digital image enhancement method based on image complexity," *International Journal of Hybrid Information Technology*, vol. 9, no. 6, pp. 395–402, 2016.

[11] P. Li, Y. Yang, W. Zhao, and M. Zhang, "Evaluation of image fire detection algorithms based on image complexity," *Fire Safety Journal*, vol. 121, pp. 103 306–103 317, 2021.

[12] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, "Computerized measures of visual complexity," *Acta Psychologica*, vol. 160, pp. 43–57, 2015.

[13] L. Dai, K. Zhang, X. S. Zheng, R. R. Martin, Y. Li, and J. Yu, "Visual complexity of shapes: a hierarchical perceptual learning model," *The Visual Computer*, 2021.

[14] X. Guo, Y. Qian, L. Li, and A. Asano, "Assessment model for perceived visual complexity of painting images," *Knowledge-Based Systems*, vol. 159, pp. 110–119, 2018.

[15] Y.-Q. Chen, J. Duan, Y. Zhu, X.-F. Qian, and B. Xiao, "Research on the image complexity based on neural network," in *International Conference on Machine Learning and Cybernetics*, 2015.

[16] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *IEEE International Conference on Computer Vision*, 2017.

[17] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[18] E. Saraee, M. Jalal, and M. Betke, "Visual complexity analysis using deep intermediate-layer features," *Computer Vision and Image Understanding*, vol. 195, pp. 102 949–102 968, 2020.

[19] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[20] M. Liu, K. Gu, G. Zhai, and P. Le Callet, "Visual saliency detection via image complexity feature," in *IEEE International Conference on Image Processing*, 2016.

[21] S. Arthur, "Entropy, visual diversity, and preference," *The Journal of General Psychology*, vol. 129, no. 3, pp. 300–320, 2002.

[22] N. Gauvrit, F. Soler-Toscano, and H. Zenil, "Natural scene statistics mediate the perception of image complexity," *Visual Cognition*, vol. 22, no. 8, pp. 1084–1091, 2014.

[23] A. Olivia, M. L. Mack, M. Shrestha, and A. Peeper, "Identifying the perceptual dimensions of visual complexity of scenes," in *The Annual Meeting of the Cognitive Science Society*, vol. 26, no. 26, 2004, pp. 1041–1044.

[24] H. C. Purchase, E. Freeman, and J. Hamer, "Predicting visual complexity," in *International Conference on Appearance*, 2012.

[25] M. P. Da Silva, V. Courboulay, and P. Estraillier, "Image complexity measure based on visual attention," in *IEEE International Conference on Image Processing*, 2011.

[26] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *Journal of Vision*, vol. 7, no. 2, pp. 17–17, 2007.

[27] X. Guo, T. Kurita, C. M. Asano, and A. Asano, "Visual complexity assessment of painting images," in *IEEE International Conference on Image Processing*, 2013.

[28] H. Yu and S. Winkler, "Image complexity and spatial information," in *International Workshop on Quality of Multimedia Experience*, 2013.

[29] M. A. Abdelwahab, A. M. Iliyasu, and A. S. Salama, "Leveraging the potency of cnn for efficient assessment of visual complexity of images," in *International Conference on Image Processing Theory, Tools and Applications*, 2019.

[30] A. M. Iliyasu, A. K. Al-Asmari, M. A. AbdelWahab, A. S. Salama, M. A. Al-Qodah, A. R. Khan, P. Q. Le, and F. Yan, "Mining visual complexity of images based on an enhanced feature space representation," in *IEEE International Symposium on Intelligent Signal Processing*, 2013.

[31] A. Miniukovich and A. De Angeli, "Quantification of interface visual complexity," in *International Working Conference on Advanced Visual Interfaces*, 2014.

[32] S. E. Corchs, G. Ciocca, E. Bricolo, and F. Gasparini, "Predicting complexity perception of real world images," *PloS one*, vol. 11, no. 6, pp. 1–22, 2016.

[33] Z. B. Fan, Y.-N. Li, J. Yu, and K. Zhang, "Visual complexity of chinese ink paintings," in *ACM Symposium on Applied Perception*, 2017.

[34] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[35] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, and C. Redies, "Jenaesthetics subjective dataset: analyzing paintings by subjective scores," in *European Conference on Computer Vision*, 2014.

[36] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[37] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *International Conference on Quality of Multimedia Experience*, 2019.

[38] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *IEEE International Conference on Computer Vision*, 2011.

[39] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[40] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[41] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.

[42] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[43] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2815–2826, 2019.

[44] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2018.

[45] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 851–864, 2018.

[46] Z. Wang and K. Ma, "Active fine-tuning from gmad examples improves blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[47] K. Sheng, W. Dong, M. Chai, G. Wang, P. Zhou, F. Huang, B.-G. Hu, R. Ji, and C. Ma, "Revisiting image aesthetic assessment via self-supervised feature learning," in *AAAI Conference on Artificial Intelligence*, 2020.

[48] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka, "Automatic understanding of image and video advertisements," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014.

[50] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "Widerperson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 380–393, 2019.

[51] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[52] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.

[53] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality

assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[54] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*, 2016.

[55] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[56] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1338–1350, 2016.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[58] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012.

[60] E. Alghamdi, E. Velloso, and P. Gruba, "Auvana: An automated video analysis tool for visual complexity," *OSF Preprints*, 2021.

[61] K. Gu, J. Zhou, J.-F. Qiao, G. Zhai, W. Lin, and A. C. Bovik, "No-reference quality assessment of screen content pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4005–4018, 2017.

[62] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[63] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999.

[64] B. Steiner, Z. DeVito, S. Chintala, S. Gross, A. Paszke, F. Massa, A. Lerer, G. Chanan, Z. Lin, E. Yang *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019.

[65] Z. Xie, J. Liu, C. Liu, Y. Zuo, and X. Chen, "Optical and sar image registration using complexity analysis and binary descriptor in suburban areas," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[66] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[67] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision*, 2018.

[68] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[69] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *International Joint Conference on Artificial Intelligence*, 2016.

[70] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[71] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *British Journal of Psychology*, vol. 95, no. 4, pp. 489–508, 2004.

[72] R. Reber, N. Schwarz, and P. Winkielman, "Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?" *Personality and Social Psychology Review*, vol. 8, no. 4, pp. 364–382, 2004.

[73] K. N. Ochsner, "Are affective events richly recollected or simply familiar? the experience and process of recognizing feelings past." *Journal of Experimental Psychology: General*, vol. 129, no. 2, p. 242, 2000.

[74] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[75] D.-N. Zou, S.-H. Zhang, T.-J. Mu, and M. Zhang, "A new dataset of dog breed images and a benchmark for finegrained classification," *Computational Visual Media*, vol. 6, no. 4, pp. 477–487, 2020.

[76] J. Gao, W. Lin, B. Zhao, D. Wang, C. Gao, and J. Wen, "C^3 framework: An open-source pytorch code for crowd counting," *arXiv preprint arXiv:1907.02724*, 2019.

[77] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[78] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *European Conference on Computer Vision*, 2018.

[79] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[80] Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, and G. Zhao, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *IEEE International Conference on Computer Vision*, 2019.

[81] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.

[82] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[83] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[84] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *European Conference on Computer Vision*, 2020.

[85] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[86] W. Köhler, "Gestalt psychology," *Psychologische Forschung*, vol. 31, no. 1, pp. XVIII–XXX, 1967.

[87] D. C. Donderi, "Visual complexity: a review." *Psychological Bulletin*, vol. 132, no. 1, p. 73, 2006.

[88] N. Sadeh and E. Verona, "Visual complexity attenuates emotional processing in psychopathy: Implications for fear-potentiated startle deficits," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 12, no. 2, pp. 346–360, 2012.

[89] S. F. Chipman and M. J. Mendelson, "Influence of six types of visual structure on complexity judgments in children and adults." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 5, no. 2, p. 365, 1979.

[90] D. Hussein, "A user preference modelling method for the assessment of visual complexity in building façade," *Smart and Sustainable Built Environment*, vol. 9, no. 4, pp. 483–501, 2020.

[91] A. Gartus and H. Leder, "Predicting perceived visual complexity of abstract patterns using computational measures: The influence of mirror symmetry on complexity perception," *PloS one*, vol. 12, no. 11, pp. 1–22, 2017.

[92] A. Forsythe, M. Nadal, N. Sheehy, C. J. Cela-Conde, and M. Sawey, "Predicting beauty: Fractal dimension and visual complexity in art," *British Journal of Psychology*, vol. 102, no. 1, pp. 49–70, 2011.

[93] A. Tuch, S. Kreibig, S. Roth, J. Bargas-Avila, K. Opwis, and F. Wilhelm, "The role of visual complexity in affective reactions to webpages: Subjective, eye movement, and cardiovascular responses," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 230–236, 2011.

[94] R. Pieters, M. Wedel, and R. Batra, "The stopping power of advertising: Measures and effects of visual complexity," *Journal of Marketing*, vol. 74, no. 5, pp. 48–60, 2010.

[95] X. Tong, Y. Chen, S. Zhou, and S. Yang, "How background visual complexity influences purchase intention in live streaming: The mediating role of emotion and the moderating role of gender," *Journal of Retailing and Consumer Services*, vol. 67, p. 103031, 2022.

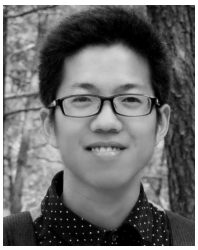
[96] S. Sohn, B. Seegebarth, and M. Moritz, "The impact of perceived visual complexity of mobile online shops on user's satisfaction," *Psychology & Marketing*, vol. 34, no. 2, pp. 195–214, 2017.

[97] A. Krishen, "Perceived versus actual complexity for websites: their relationship to consumer satisfaction," *The Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, vol. 21, 2008.

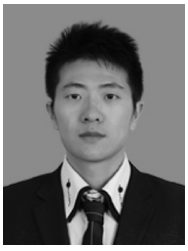
Tinglei Feng is currently pursuing a Master's degree at the College of Computer Science, Nankai University, Tianjin, China. His research interests center on computer vision and pattern recognition.

Yingjie Zhai received his Master's degree from Nankai University. His current research interests include deep learning and computer vision, especially salient object detection and image restoration.

Jufeng Yang received the Ph.D. degree from Nankai University, Tianjin, China, in 2009. He is currently a full professor in the Department of Computer Science, Nankai University and was a visiting scholar with the Vision and Learning Lab, University of California, Merced, USA, from 2015 to 2016. His recent interests include affective computing, image retrieval, fine-grained classification, and medical image recognition.



Jie Liang is currently pursuing his Ph.D. degree from The Hong Kong Polytechnic University. His current research interests include computer vision, machine learning, pattern recognition, and optimization



Deng-Ping Fan received his PhD degree from Nankai University in 2019. He joined Inception Institute of AI in 2019. He has published about 25 top journal and conference papers such as TPAMI, CVPR, ICCV, ECCV, etc. His research interests include computer vision and visual attention, especially on RGB salient object detection (SOD), RGB-D SOD, Video SOD, CoSOD. He won the Best Paper Finalist Award at IEEE CVPR 2019, the Best Paper Award Nominee at IEEE CVPR 2020.

Jing Zhang (Member, IEEE) is currently a Research Fellow at the School of Computer Science, The University of Sydney. He has published more than 40 papers on prestigious conferences and journals, such as CVPR, ICCV, NeurIPS, International Journal of Computer Vision (IJCV), and IEEE Transactions on Image Processing (TIP). His research interests include computer vision and deep learning. He is a Senior Program Committee Member of the AAAI Conference on Artificial Intelligence and the International Joint Conference on Artificial Intelligence. He serves as a reviewer for many journals and conferences.

Ling Shao (Fellow, IEEE) is the Chief Scientist of Terminus Group and the President of Terminus International. He was the founding CEO and Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, deep learning, medical imaging and vision and language. He is a fellow of the IEEE, the IAPR, the BCS and the IET.

Dacheng Tao (Fellow, IEEE) is currently the Inaugural Director of the JD Explore Academy and a Senior Vice President of JD.com, Inc. He mainly applies statistics and mathematics to artificial intelligence and data science. His research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences. He is a fellow of the Australian Academy of Science, AAAS, and ACM. He received the 2015 Australian Scopus-Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award.