Attentive Deep Stitching and Quality Assessment for 360° Omnidirectional Images

Jia Li, *Senior Member, IEEE*, Yifan Zhao, Weihua Ye, Kaiwen Yu, and Shiming Ge, *Senior Member, IEEE*

***Abstract*—360° omnidirectional images are very helpful in creating immersive multimedia contents, which enables a huge demand in their efficient generation and effective assessment. In this paper, we leverage an attentive idea to meet this demand by addressing two concerns: how to generate a good omnidirectional image in a fast and robust way and what is a good omnidirectional image for human. To this end, we propose an attentive deep stitching approach to facilitate the efficient generation of omnidirectional images, which is composed of two modules. The low-resolution deformation module aims to learn the deformation rules from dual-fisheye to omnidirectional images with joint implicit and explicit attention mechanisms, while the high-resolution recurrence module enhances the resolution of stitching results with the high-resolution guidance in a recurrent manner. In this way, the stitching approach can efficiently generate high-resolution omnidirectional images that are highly consistent with human immersive experiences. Beyond the efficient generation, we further present an attention-driven omnidirectional image quality assessment (IQA) method which uses joint evaluation with both global and local metrics. Especially, the local metric mainly focuses on the stitching region and attention region that mostly affect the Mean Opinion Score (MOS), leading to a consistent evaluation of human perception. To verify the effectiveness of our proposed assessment and stitching approaches, we construct a hybrid benchmark evaluation with 7 stitching models and 8 IQA metrics. Qualitative and quantitative experiments show our stitching approach generate preferable results with the state-of-the-art models at a 6× faster speed and the proposed quality assessment approach surpasses other methods by a large margin and is highly consistent with human subjective evaluations.**

Q1

***Index Terms*—360° omnidirectional image, image quality assessment (IQA), attentive deep stitching.**

The corresponding author is Shiming Ge.

Manuscript received April 21, 2019; revised September 3, 2019; accepted November 8, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61922006, Grant 61532003, and Grant 61772513, and in part by the Beijing Nova Program Z181100006218063. The work of S. Ge was supported by the Youth Innovation Promotion Association, Chinese Academy of Sciences. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Patrick LE CALLET. *(Corresponding author: Jia Li.)*

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: jiali@buaa.edu.cn).

Y. Zhao, W. Ye, and K. Yu are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: zhaoyf@buaa.edu.cn; yeweihua@buaa.edu.cn; kevinyu@buaa.edu.cn).

S. Ge is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China (e-mail: geshiming@iie.ac.cn).

I. Introduction

In the rapid development of virtual reality (VR) over the last decade, high-quality 360° omnidirectional images play an increasingly important role in producing multimedia contents, which requires natural immersions of real-world scenarios in head-mounted displays. Along with the boost of omnidirectional acquisition devices, there exists a huge demand in efficient omnidirectional image generation and accurate quality assessment, which can further be of great use in biology [1], [2], medical [3], modeling [4] and virtual reality [5]. To get the high-quality omnidirectional images, tens of models have been proposed to stitch the dual-fisheye images into 360° omnidirectional images. With the large amount of stitched images and stitching models, it further yields two important concerns: how to generate a good omnidirectional image in a fast and robust way and what is a good omnidirectional image for human?

In the view of the first concern, there exist two main categories of automatic stitching methods to generate omnidirectional images rather than manual calibration methods [4], [6]: direct stitching and feature-based stitching. Direct stitching approaches [1], [7], [8] have the advantage that they make full use of the available image data and hence can provide accurate registration but required a closed initialization. In contrast, the feature-based stitching methods [3], [9]–[11] do not require the complicated initialization procedure. They usually automatically detect invariant local features and construct a matching correspondence instead of manual registration. Some classical feature detectors [12], [13] usually perform well on conventional planar natural images but may lack invariant properties in handling dual-fisheye images, which may cause distortions or shape breakages in stitching regions.

Beyond the first concern, many image quality assessment (IQA) methods [14]–[18] have been proposed to address this problem. In the early researches of IQA, the evaluation methods mainly focus on the common daily images with many photometric quality indexes such as MSE [19], PSNR [20] and SSIM [21]. With the development of Convolutional Neural Networks (CNNs), some representative models [16], [22], [23] with deep features have been proposed. However, these models usually focus on the photometric quality indexes such as blurring, noise and color distortions, which may be not suitable for the omnidirectional images. Moreover, there are a few works [24], [25] on the quality assessment of panoramic images. For example, Yang *et al.* [17] proposed a light-weight model to evaluate the stitched panoramic images based on ghosting and

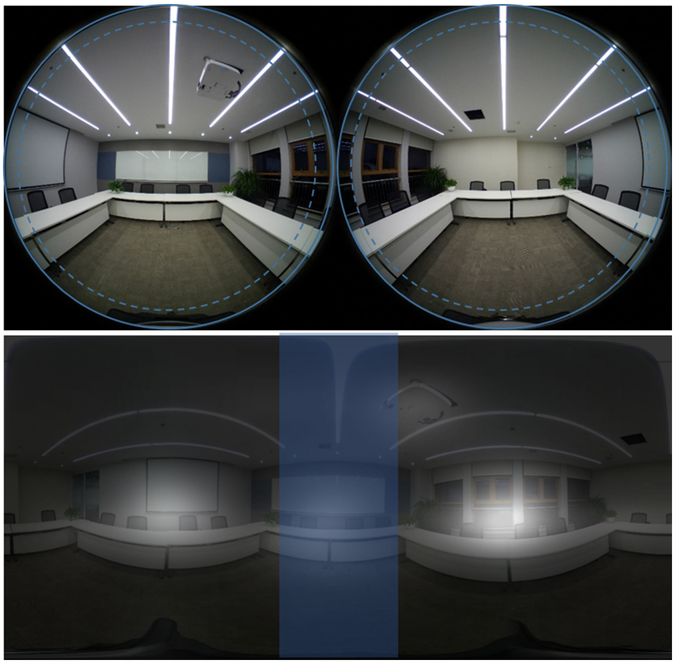


Fig. 1. Dual-fisheye image and its stitched 360° panoramic image with human attention (white) and stitching region focusing (blue). The image immersive experience to human is mainly affected by human attention mechanism and the distortions are most likely to happen in specific stitching regions.

structure inconsistency. These proposed metrics are designed for normal 2-D plane image stitching, such that cannot handle 360° omnidirectional images which are generated from the dual-fisheye images and have large distortion and information loss in the stitching areas.

In summary, these general stitching and assessment approaches usually treat every pixel of the stitching images equally. However, immersive experiences of omnidirectional images are affected by two attentive cues: attention region and stitching region. As shown in Fig. 1, attention region is mainly focused by human gaze while stitching region in the middle most likely happens distortions or shape breakage. To this end, we address the efficient generation and effective assessment of 360° omnidirectional images by two human perception-driven approaches, which is attentive deep stitching (ADS) and attentive quality assessment (AQA), respectively.

ADS adopts a progressive manner to perform efficient generation of 360° omnidirectional images, which is composed of two main modules along with an implicit-attention and explicit-attention mechanisms respectively. The first low-resolution deformation module learns the deformation features from the dual-fisheye image with multiple implicit-attention blocks. By combing the learned deformation features and the high-resolution dual-fisheye image, the second high-resolution recurrence module is conducted to assign the deformation relationship with the high-resolution pixel guidance. With the recurrent refinement scheme, a high-resolution omnidirectional image is obtained. At the end of these two modules, the explicit attention map of human gaze is introduced to regularize the consistency of stitching results and human subjective experience.

AQA is a novel full-reference quality assessment approach, which is designed to evaluate the quality of the stitched 360° omnidirectional images in accord with human perception. Based on the accurate cross-reference omnidirectional image dataset [26], we propose a joint approach to combining the local and global metrics, where the global metric mainly considers the environmental differences like color chromatism and the blind zone phenomenon. For the local metric, we develop an attentive sampling strategy to focus on attention region and stitching region, the two special regions that mostly affect the stitching quality for the attentive frequency and stitching distortions, respectively. To this end, We adopt the sparse reconstruction and appearance difference to represent the local metric and finally use the linear learning progress to match human subjective evaluations.

The contributions of this paper can be summarized as follows: 1) We propose a novel attentive deep stitching approach to facilitate the generation of high-resolution 360° omnidirectional images from dual-fisheye images in an end-to-end deep manner, which runs at a 6× faster speed than the state-of-the-art methods. 2) We propose an attentive quality assessment approach to automatically assess the stitching quality of 360° omnidirectional images, which provides more consistent evaluation to the human perception. 3) Qualitative and quantitative experiments are conducted to demonstrate the effectiveness of the proposed stitching approach while the proposed quality assessment approach is highly consistent with human subjective evaluation.

The rest of this paper is organized as follows: Section II reviews related works and Section III proposes the attentive deep stitching method. In Section IV, the attentive quality assessment approach for omnidirectional stitching is proposed. We further conduct quantitative and qualitative experiments in Section V and finally conclude this paper in Section VI.

II. Literature Review

A. Omnidirectional Image Stitching

Classical stitching models: Image stitching techniques is now becoming a research hotspot with wide applications [7], [9], [27]. Classical stitching models, such as Stereoscopic Vision Projection (SVP) [28], Isometric Projection (IP) [29] and Equidistant Projection (EP) [30], have been widely used to generate the 360° omnidirectional images in an automatic way. Most modern digital cameras have added panoramic mode, including many mobile devices.

Classical image stitching methods can be roughly divided into two categories: camera calibration based image stitching and keypoint based image stitching. Recently several works [31]–[33], have made progress in improving traditional image stitching algorithm [11]. Charles *et al.* [34] solve the problem that the use of a single registration often leads to errors, especially in scenes with significant depth variation or object motion. With the portability and cheapness of the dual-fisheye camera, the research [35], [36] on fisheye image stitching becomes more and more applicable. Lo *et al.* [37] stitched the dual-fisheye image into a 360° panoramic image following the four basic steps of right angle transformation, feature extraction, mesh deformation and mixture.

Deep convolutional models: At present, most convolutional neural networks keep the image prototype and only extract special information. With the high demand on daily images, the traditional problems of pixel drift, such as image stitching and fisheye image distortion correction [35], [38], also need to be further studied. In recent years, a few researchers made attempts in solving the pixel drift issue with convolutional networks. Yin *et al.* [39] proposed an end-to-end multi-context collaborative deep network for removing distortions from single fisheye images that learns high-level semantics and low-level appearance features simultaneously to estimate the distortion parameters. Deng *et al.* [40] proposed restricted stitching convolution for semantic segmentation, which can effectively model geometric transformations by learning the shapes of convolutional filters.

B. Image Quality Assessment

Many IQA methods [41] have been proposed in the past decades, which can be roughly grouped into three categories. Some pioneer works for image IQA [15], [22], [42] focuses on both traditional IQA and common panoramic stitching IQA. In this paper, we mainly focus on the quality assessment omnidirectional images, which is a less-explored task with increasing demands.

Classical IQA metrics: Most of recent IQA researches focused on no-reference image quality assessment (NR-IQA) [43]–[48] and full-reference image quality assessment (FR-IQA) [18], [19], [49]–[51]. NR-IQA do not need specific reference image which is convenient to various image assessment task. In the process of FR-IQA, the assessment quality are compared with the results of reference image, For instance: MSE [19], PSNR [20], SSIM [52]. The assessment of immersive stitching IQA can also be adopted to full reference assessment method to some extent. However, the assessment result may be not so accurate and sometimes even opposite to human visual judgements.

Learnable IQA metrics: Due to the rapid development of deep learning in recent years, various existing problems can achieve better results on the basis of deep learning approaches. Therefore, many researchers use deep learning models to evaluate daily images in the field of image quality evaluation. For these deep learning models [15], [23], the biggest problem is that there is no suitable large dataset for training. Kang *et al.* [23] aimed at images which the dataset suitable for deep learning training, and proposed to use 32 $\times$ 32 patches for training. On the one hand, the method increased the amount of data through simplify image processing, on the other hand, the discontinuity of the main structure in the image may lead to inaccuracy. Liu *et al.* [16] trained a Siamese Network to sort and learn images, and learned the relationship between images by sharing the weight of the network. Some researchers adopted convolutional sparse coding to locate specific distortions [53]–[55] and designed the kernel to quantify the mixed effects of multiple distortion types in local regions.

Stitching IQA metrics: There are few researches in the image quality assessment of stitching images. For example, Yang *et al.* [17] solved the problem of ghosting and structural discontinuity in image stitching by using perceptual geometric error metric and local structure-guide metric, but for immersive image, the evaluation method is not comprehensive enough to detect the global color difference, and the conditions of blind zone. Huang *et al.* [56] proposed the quality evaluation of immersed images, mainly focusing on resolution and compression, neither the quality evaluation of stitching, nor on the image quality evaluation. In [53], the authors adopted convolutional sparse coding and compound feature selection which focuses on the stitching region for stitched image assessment. Moreover, some subjective omnidirectional video quality assessment methods [57], [58] have been proposed in this less-explored task.

III. Attentive Deep Stitching

A. Overview

The classical omnidirectional stitching problem is a transformation of optical refraction operation. The pixels of dual-fisheye images are transferred from the equirectangular coordinates to the two-dimensional plane stretching. Commonly, transforming dual fisheye image $\mathbf{x}$ to omnidirectional image $\mathbf{y}$ can be formally represented as:

$$\mathbf{y} = \mathcal{T}(\mathbf{x}; \theta), \tag{1}$$

where $\mathcal{T}: \mathbb{R}^{W\times H} \to \mathbb{R}^{W\times H}$ is the transformation function and θ denotes the parameters of stitching model. However, the transformation relationship varies significantly for different images and cameras.

Instead of the manual calibration or automatical registration, we advocate using deep convolutional neural networks to learn this transformation instead of hand-crafted designs. To this end, we propose an attentive deep stitching approach which efficiently solves the transformation $\mathcal{T}$ in two phases:

$$\tilde{\mathbf{f}} = \mathcal{F}(\tilde{\mathbf{x}}; \theta_F), \;\; \theta_F \subset \theta_L, \tag{2}$$

$$\mathbf{y} = \mathcal{H}(\tilde{\mathbf{f}}, \mathbf{x}; \theta_H), \tag{3}$$

where $\mathcal{F}$ is the low-resolution deformation module and $\mathcal{H}$ is the high-resolution recurrence module. $\tilde{\mathbf{x}}$ is the down-sampled input of dual-fisheye image $\mathbf{x}$. $\tilde{\mathbf{f}}$ are learnable deformation features of $\mathbf{x}$ and $\theta_{\{L,H\}}$ are learnable parameters of the low-resolution deformation and high-resolution recurrence phase, respectively. θ_F is used to extract the deformation features, which is a part of θ_L.

B. Low-Resolution Deformation

As shown in Fig. 2, the low-resolution deformation module aims to decode the transformation information $\mathcal{F}(\cdot)$ in a learning procedure. Verified many representative researches in panoramic attention, human gaze [59], [60] usually focuses on special regions which contain the most attractive information. Keeping these two cues in mind, we further develop an attention-based deformation learning process, which is jointly optimized with the implicit-attention and explicit-attention mechanisms.

Considering the computation cost and stitching efficiency, we resort to the low-resolution image $\tilde{\mathbf{x}}$ to learn the deformation information. Inspired by the successful U-Net [61] architecture, we develop the light-weighted deformation module which

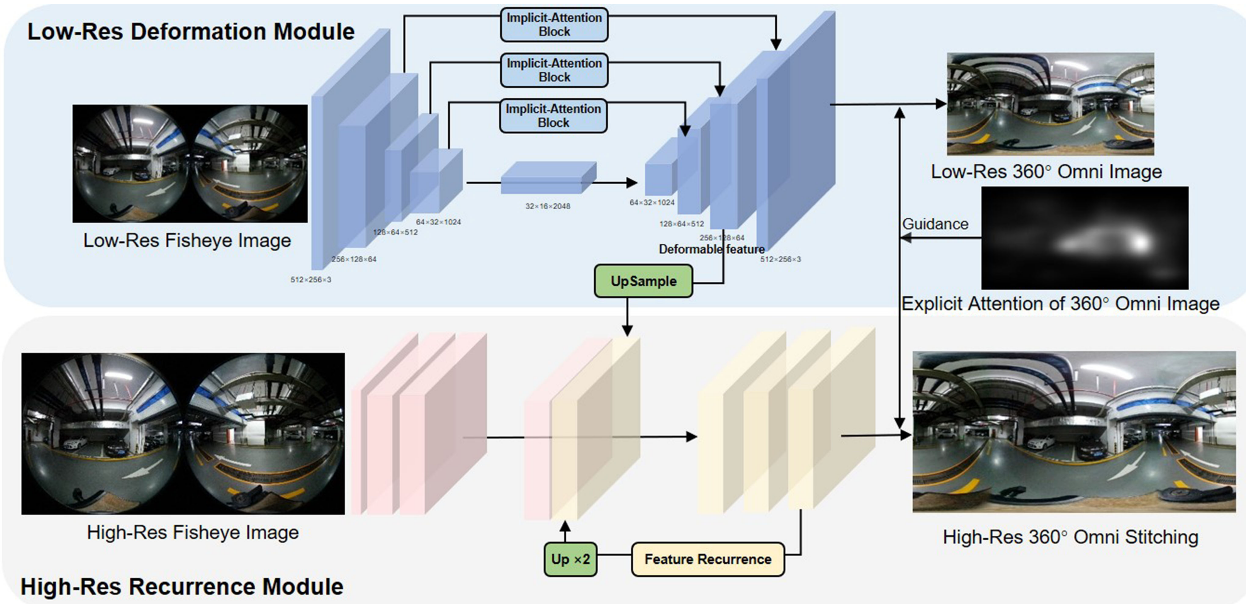


Fig. 2. Framework of proposed Attentive Deep Stitching (ADS). Our framework is mainly composed of two modules. The first deformation module is to learn the transformation rules from dual-fisheye to omnidirectional images with the joint human-supervised explicit attention and implicit attention mechanism. The second recurrence module utilizes the high-resolution fisheye image as a guidance to the stitching results in a recurrent manner.

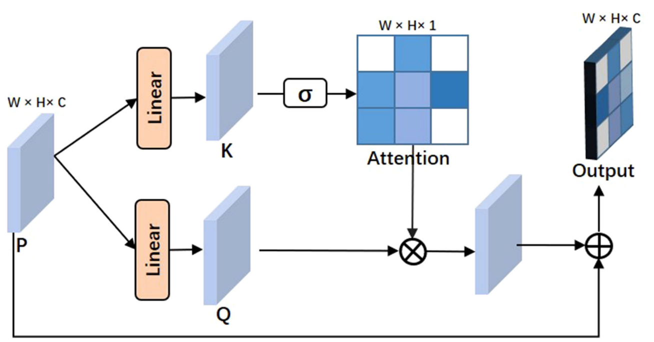


Fig. 3. Implicit-attention block. σ: element summation in channel dimension and a softmax operation. $\otimes$: scalar production. $\oplus$: element-wise sum.

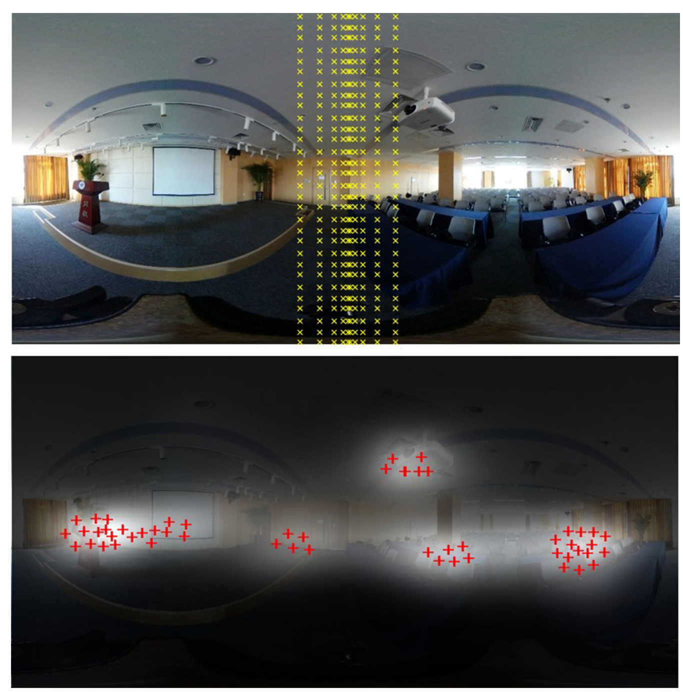


Fig. 4. Illustration of local attentive sampling. First row: gaussian-based stitching region sampling. Second row: attention-based region sampling.

consists of a contracting path and an expansive path. Based on this design, we add multiple implicit-attention blocks as the transition layers to pass through the low-level attention to the high-level decoders, which attaches more importance in attention region and can further eliminate the gradient loss. The detailed architecture of implicit-attention block is shown in Fig. 3. σ denotes the channel-wise summation and softmax operation. With the input feature $\mathbf{P} \in \mathbb{R}^{W \times H \times C}$, this block can be formally represented as:

$$
\begin{aligned}
\mathbf{V} &= tanh(\mathbf{w}_v \mathbf{P} + \mathbf{b}_v), \\
\mathbf{K} &= tanh(\mathbf{w}_k \mathbf{P} + \mathbf{b}_k),
\end{aligned}
\tag{4}
$$

where $\mathbf{w}_k, \mathbf{b}_k, \mathbf{w}_v, \mathbf{b}_v$ are the parameters and $\odot$ is the scalar-product operation. After getting this transformed features, the final output $\mathbb{S}(\mathbf{P})$ with the implicit attention is formulated as:

$$
\begin{aligned}
\mathbb{S}(\mathbf{P}_{i,j,k}) &= \frac{e^{\mathbf{M}_{i,j}}}{\sum_{i,j} e^{\mathbf{M}_{i,j}}} \odot \mathbf{Q}_{i,j,k} + \mathbf{P}_{i,j,k}, \\
\mathbf{M}_{i,j} &= \sum_{k} \mathbf{K}_{i,j,k}.
\end{aligned}
\tag{5}
$$

After that, the outputs $\mathbb{S}(\mathbf{P})$ pass from these implicit-attention blocks to the corresponding features in decoder with a concatenation operation.

In addition to the implicit-attention mechanism, the explicit attention is usually generated by human gaze estimation, which provides reliable region attention information. To this end, we resort to the state-of-the-art SALICON [62] approach to estimate human gaze. We train the gaze model on large daily image datasets and finetune the network with omnidirectional image annotations.

With the reliable generation of explicit attention model, we develop a region sensitive MSE loss $\mathcal{L}_L$ to attach more importance on attention regions:

$$\mathcal{L}_L(\theta_L) = \frac{1}{2\,N}\sum_{i=1}^{N}\mathbf{W}_i\|\mathcal{F}(\tilde{\mathbf{x}}_i;\theta_L) - \tilde{\mathbf{y}}_i\|^2, \quad \mathbf{W}_i = \frac{1}{1+e^{-\mathbf{A}(\tilde{\mathbf{y}}_i)}}, \tag{6}$$

where N is the total number of elements and $\mathcal{F}(\tilde{\mathbf{x}}_i;\theta_L)$ denotes the low-resolution prediction of the deformation module. More exactly, θ_L is the deformation module with the final prediction, and θ_F is used to obtain the transformation features. $\mathbf{W}_i$, $\tilde{\mathbf{x}}_i$, and $\tilde{\mathbf{y}}_i$ are the ith element of the explicit-attention weight $\mathbf{W}$, the low-resolution dual-fisheye image $\tilde{\mathbf{x}}$ and low-resolution omnidirectional image $\tilde{\mathbf{y}}$, respectively. $\mathbf{A}(\tilde{\mathbf{y}}_i)$ is the ith attention value of omnidirectional image $\tilde{\mathbf{y}}$, which is normalized in [0, 1].

Regularized with these two attention mechanisms, our deformation module is developed with $d = 32$ times down-sample operations with chained max-pooling. In this manner, each pixel in the highest level takes the responsibility to learn the transformation from $k^2 \times d^2$ pixels of the original input, where k is the kernel size of the current feature map. Moreover, we explore chained max-pooling operation in each down-sample operation and the $tanh(\cdot)$ activation at the end of the network, which are suitable to emphasize the local extremum and greatly maintain the transformation information. The $tanh(\cdot)$ function also accelerates the convergence speed of our network.

C. High-Resolution Recurrence

After the initialized deformation, we develop a progressive high-resolution generation module with a recurrent manner. As shown in Fig. 2, the high-resolution fisheye image passes through a feature encoder (view in red) to obtain the accurate pixel guidance. In another way, the feature from the second last feature map (features before output) is obtained with a up-sampling operation as the deformation guidance. The high-resolution pixel guidance and deformation guidance are then concatenated with 1×1 convolutions. Finally, these fused features pass through a hourglass network without down-sampling operations to get the higher-resolution output.

In this manner, the deformation branch provides the transformation regulation $\mathcal{F}$ and the high-resolution input provides the pixel-level guidance $\mathcal{G}$. In sth stage, we concatenate the up-sampling transformation regulation $\mathcal{F}$ and high-resolution $\mathcal{G}$ to decode them with a high-resolution hourglass decoder $\varphi_s(\cdot)$, which is composed of 8 convolutional layers with the 3×3 kernels. This recurrent manner in sth stage can be formulated as:

$$\mathcal{H}_s(\tilde{\mathbf{x}}_s, \tilde{\mathbf{x}}_{s+1}) = \begin{cases} \varphi_s(\mathcal{G}(\tilde{\mathbf{x}}_{s+1}) \circledast \mathcal{F}(\tilde{\mathbf{x}}_s;\theta_F)), & \text{if } s = 1 \\ \varphi_s(\mathcal{G}(\tilde{\mathbf{x}}_{s+1}) \circledast \mathcal{H}_{s-1}(\cdot)), & \text{if } 1 < s \leq S, \end{cases} \tag{7}$$

where the $\tilde{\mathbf{x}}_s$ is the down-sampled input of the sth scale and $\circledast$ is the feature concatenate operation with 1×1 convolutions. $\mathcal{H}(\cdot)$ is the output feature in the sth iteration. S is the maximum stage with largest input resolutions, which is set as 3 in our experiments. The first iteration adopts deformation feature from our first module and the following iterations adopts the high-resolution features of the last stage as guidance. At the end of the third iteration, the width and height of the recurrence $\mathcal{F}$ are fixed for the computation limitation. We recommend using this result as the final prediction considering the time efficiency. Our recurrence network follows the common stage-wise training process, the final loss in the sth stage can be represented as:

$$\mathcal{L}_H(\theta_H^s) = \frac{1}{2\,N}\sum_{i=1}^{N}\mathbf{W}_i\|\mathcal{H}(\tilde{\mathbf{x}}_{i,s}, \tilde{\mathbf{x}}_{i,s+1};\theta_H^s) - \tilde{\mathbf{y}}_{i,s+1}\|^2, \tag{8}$$

where $s = 1 \dots S$ denotes the iteration stage. The weight $\mathbf{W}_i$ is generated by the explicit attention map in Eqn. (6). With the recurrent iterations from low-resolution to high-resolution, a finer stitching result is obtained with an end-to-end inference.

IV. Attentive Quality Assessment

After the promise of the stitching models, the main concern is how to evaluate these models on the stitched images. In this section, we propose a novel approach to evaluate the less-explored omnidirectional stitching task with joint local and global quality assessment metrics. Both global and local indexes are full-reference metrics which are evaluated with the cross-reference ground-truth. The local indexes mainly focus on attentive and the stitching seam region, while the global ones mainly focus on the environmental immersion experience.

A. Local Attentive Assessment

The main distortions in the omnidirectional images are most likely to happen in the regions near the stitching seam. In contrast, the regions far from the stitching seam are usually with fewer distortions. On the other hand, human gaze [63]–[65] usually focuses on regions with special patterns, which drives us not to treat every pixel equally. Specially, the stitching images and ground-truth reference may have some slight degree changes, which is not suitable to align pixels in stitching image and ground-truth.

To this end, we sample the patches instead of per-pixel matching in both stitched images and ground-truth image. The stitching regions Ω_{sti} are sampled with gaussian sigma-criterion in Eqn. (9):

$$\mathbf{R}_s(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}\right), \mathbf{x} \in \Omega_{sti}, \tag{9}$$

where the region indicators μ is set as the 0.5 times width of the stitching region. To sampling more patches in stitched images and eliminate the pixel shifting, we set σ for stitching regions and reference regions as 220 and 350 respectively. The patch size is set as 8×8 and sampled by using a sliding-window strategy. Moreover, the human attention $\mathbf{R}_a(\mathbf{x})$ is sampled with the most brightness scores in attention map, which is calculated by the same SALICON [62] model fine-tuned on the omni-directional image. Similarly, The ground-truth patches $\mathbf{D}_a(\mathbf{x})$ and $\mathbf{D}_s(\mathbf{x})$ are obtained. With the summarization of attentive sampled patches, two meaningful metrics of local regions are further proposed.

1) Sparse reconstruction: To robustly measure the region similarity at various levels of details, we propose to adopt the sparse reconstruction errors as the local metric. The foundation of this metric is that the similar patches can represent each other with a minimal length of the sparse code. To this end, an over-complicated sparse dictionary $\mathbf{D} = \{\mathbf{D}_a, \mathbf{D}_s\}$ is constructed with the ground-truth patches and the stitched patches are stacked as $\mathbf{R} = \{\mathbf{R}_a, \mathbf{R}_s\}$. Our solving procedure of the minimal reconstruction code $\mathbf{X}^*$ can be formulated as:

$$\mathbf{X}^* = \text{argmin}_{\mathbf{X}} \frac{1}{2}||\mathbf{R} - \mathbf{D}\mathbf{X}||_F^2 + \lambda||\mathbf{X}||_1. \tag{10}$$

This can be easily solved with the Online Dictionary Learning [66]. With the optimized sparse representation $\mathbf{X}^*$, we further adopt the SVD decomposition to the principal component with $\mathcal{F}_{PCA}$. The final score is evaluated with the L1-norm:

$$\mathbf{X}^* = \sum_{i=1}^{r} \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T,$$

$$\mathbf{E}_{\text{sparse}} = -\sum_{i=1}^{r} ||\mathcal{F}_{PCA}(\mathbf{\Sigma}_i)||_1, \tag{11}$$

where $\mathbf{\Sigma}$ is decomposed with singular values to represent the sparse reconstructions.

2) Appearance similarity: To evaluate the appearance similarity, we resort to the commonly used Gray-Level Co-occurrence Matrix (GLCM) [67] in degrees of [45, 90, 135, 180] to extract the texture features. With these features, we adopt the histogram calculation to measure the texture similarity between the sampled patches. In this manner, we calculate the cosine similarity between the divided bins in Eqn. (12).

$$\mathbf{E}_{\text{app}} = \sum_{d=1}^{4} \sum_{i}^{n} \sum_{j}^{n} \cos(\mathbf{h}_i, \mathbf{h}_j) ||\mathbf{h}_i||_F^2 ||\mathbf{h}_j||_F^2, \tag{12}$$

where $\mathbf{h}_i$ is the ith histogram bin of the GLCM matrix with $n = 10$ devisions and d indicates the four degrees in GLCM matrix.

B. Global Environmental Assessment

To evaluate the environmental immersion experience, we further propose two global metrics to qualify global regions of the stitched images, which can be summarized as:

1) Color chromatism: Most of the stitching methods adjust some optical parameters to match two images, which could further bring in some chromatic aberrations. To evaluate the point-wise color difference, we adopt the SIFT [68] matching to find the pixel correspondences between the stitched image and ground-truth image. For each matched pair of points, we compute the K-nearest neighbor to eliminated mismatches, and we denote $\mathbf{S}$ and $\mathbf{G}$ as the sampled patches of stitching regions and referenced ground-truth regions, respectively. The sift matching procedure can be formulated as:

$$\mathbf{G}^* = \text{argmin}_{\mathbf{G}} ||\mathbf{S} - \mathcal{H}_{\text{sift}}(\mathbf{G}_i)||_F^2, \quad i = 1 \ldots K. \tag{13}$$

After matching every $\mathbf{S}$ with the nearest $\mathbf{G}$, the color chromatism score can be calculated as:

$$\mathbf{E}_{\text{color}} = -\sum_{i=1}^{M} \sum_{k=1}^{C} \lambda \frac{||\mathbf{S}_{ik} - \mathbf{G}_{ik}^*||_F^2}{M \times C}, \tag{14}$$

where M is the number of corresponding pairs and C is the number of channels. λ is set as 100 to balance the final score.

2) Blind zone: The blind zones are the blank areas with the information loss during transformation processes, which affect the visual comfortableness in immersive experiences. To accurately measure the impact of blind zones, we propose an attention-weighted blind zone evaluation metric. We adopt the same SALICON model [62] to calculate the attentive regions with groundtruth. The generated attention map $\mathbf{w}_i^b$ for blind zone is generated with the ground-truth omni-directional image. The value of $\mathbf{w}_i^b$ are normalized in [0, 1]. The score of $\mathbf{E}_{\text{blind}}$ can be calculated as:

$$\mathbf{E}_{blind} = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{B_i}}{1 + e^{-\mathbf{w}_i^b}}, \tag{15}$$

where $\mathbf{B_i}$ denotes the ith pixel of blind zones masks, which is set as 1 when it is in blind zone. N denotes the number of pixels in the stitched image. The region $\mathbf{B_i}$ can be simply calculated with the bottom and top stitching region with continuous blank areas (zero-value pixels). If the attentive region with important message are missing, the $\mathbf{E}_{blind}$ will generate lower scores.

C. Joint Assessment With Human Guided Classifier

With the proposed two local metrics and two global ones, we further introduce human subjective evaluations to supervise our linear classifier. We use the concluded pair-wise evaluation scores as the final results, which is the Mean Opinion Score (MOS) collected in the CROSS dataset [26].

The aim of our classifier is to provide learnable weight to make the metric consistent with the human subjective assessment. To this end, we adopt the multiple linear regression (MLR) [69], [70] to fit the human subjective ground-truth. Stack vector $\mathbf{x} = \{\mathbf{E}_{\text{app}}, \mathbf{E}_{\text{sparse}}, \mathbf{E}_{\text{color}}, \mathbf{E}_{\text{blind}}\}$ with the proposed metrics above, we adopt MOS scores as the ground-truth $\mathcal{M}$, The weight-balance parameters $\boldsymbol{\beta}$ can be learned by generalized least squares estimation, which are shown in Eq. (16):

$$\begin{aligned}\mathcal{M} &= \boldsymbol{\beta} \cdot \mathbf{x}, \\ \boldsymbol{\beta}^* &= \operatorname{argmin}_{\boldsymbol{\beta}} (\mathbf{x}^T \boldsymbol{\Omega}^{-1} \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\Omega}^{-1} \mathcal{M},\end{aligned} \tag{16}$$

where $\boldsymbol{\Omega}$ is the covariance matrix of residual error. Finally, the final assessment scores can be calculated as:

$$\hat{\mathcal{R}} = \boldsymbol{\beta}^* \cdot \mathbf{x}, \tag{17}$$

which can be further used to rank different stitching results.

V. Experiments

A. Experiments Settings

1) Cross-Reference Dataset: We conduct our experimental on the CROSS dataset [26], which contains 292 fisheye images as quaternions. Each quaternion is composed of images captured from standard quarters of 0, 90, 180 and 270 degrees. Taking two images in opposite directions for stitching, the other two images can provide high-quality ground-truth references. In this manner, a high-quality ground truth stitching image is obtained.

To make a fair comparison of the state-of-the-art models, we randomly select 192 fisheye images as the training set from 12 different scenarios and the rest 100 images as the test set. Moreover, all the existing dual-fisheye images and corresponding $360°$ panoramic images are mirrored horizontally and vertically to obtain four times the number of original. In order to verify the robustness of our network, we only use the original images and the horizontal images in training process but add the vertical images in the test.

2) Evaluation Criterion: Our evaluation criterion is composed of two systems. The first criterion system is composed of the five most commonly used quality assessment evaluation metrics to match the mean opinion score (MOS) provided by [26], which conducted pair-wise ranking scores with 14,847 comparisons. The evaluation metrics include Cosine Similarity (CS), the Pearson Rank Correlation Coefficient (PRCC), the Spearman's Rank Order Correlation Coefficient (SROCC), Kendall Rank Correlation Coefficient (KRCC) and Root Mean Square Error (RMSE). We also adopt the quality metric evaluation framework [74] to assess our AQA approach. The detailed formulation of these evaluation metrics can be found in Table II. The MOS scores are stacked as vectors to calculate correlation similarities with IQA methods.

The second criterion system is adopted to evaluate the quality of stitched images. Despite the proposed omnidirectional quality assessment, 9 widely-used IQA methods are adopted for our benchmark, including classical methods MSE [19], PSNR [20], SSIM [49], no-reference quality assessment metrics, BRISQUE [43], NIQE [73], PIQE [14], CCF [47], CEIQ [48], and current method based machine learning, CNNIQA [15].

3) Implementation Details: Our deep deformation model is built with 10 convolutional blocks with (3×3, ReLU, batch norm), followed by a 2×2 max-pooling after each block in encoder and upsampling in decoder. The high-resoluton recurrence module is built with the same convolutional blocks and finally designed with a tanh function at the end. The whole model is trained on a single NVIDIA GeForce GTX 1080 GPU and a single Intel i7-6700 CPU. The learning rate of deformation module and recurrence module are both starting with 0.001 and reduces to half of that when the validation loss reaches a plateau. The deformation module is trained with the resolution of 512×256. Owing to the limitation of GPU memory, we adopt the 2048×1024 as final high-resolution output and the recurrence stage as 3 iterations to get the final output in our experiments. The deformation stage is trained for 30 k iteration and each recurrent stage are trained for 15 k iterations. The adopted attention model SALICON [62] is pre-trained on the SALICON fixation dataset [75], which contains 20,000 annotations of daily images [76]. We then fine-tuned on the fixation annotations of our training set with 10 k iterations with a finetuned $lr = 1e-6$. The groundtruth annotation follows the original dataset [75].

B. The Omnidirectional Stitching Benchmark

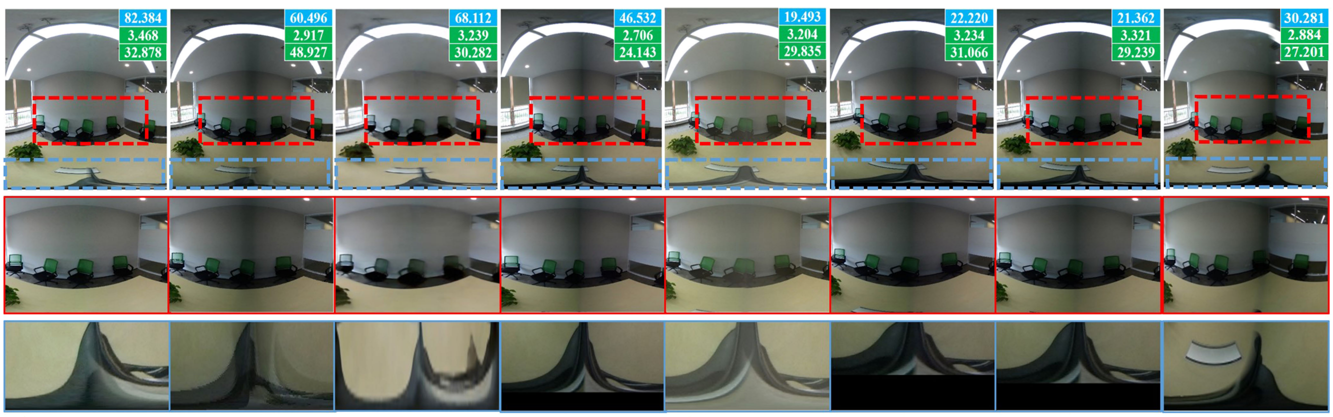
We adopt 7 widely-used state-of-the-art stitching methods to construct our benchmark, including Samsung Gear 360 [71], OpenSource [72], Stereoscopic Vision Projection (SVP) [28], Isometric Projection (IP) [29] and Equidistant Projection (EP) [30], ManMethod (Manual Method) and our proposed Attentive Deep Stitching (ADS), which finally yields 1344 stitched images in total for comparisons.

We firstly use the 7 compared IQA methods to evaluate results of selected 7 stitching models, which finally yields 49 scores, as shown in Table I. To compare with these IQA indexes, we use the ranking order (view in blue) to evaluate these methods. From which we can see that in most of the proposed indexes, our proposed deep stitching method generate the preferable results, comparing to the time-costing or labour-consume methods. Our proposed method ranks the first place in referenced IQA metrics, which demonstrate the stitching deformation results of our method. However, our proposed losses a lot details in No-reference IQA comparisons such as BRISQE [43], NIQE [73] and PIQE [14] because of the resolution limitation. Most of the stitching methods are conducted in the resolution of 5792×2896, which our stitching result are generated by upsampling from 2048×1024, which may loss many details in the non-referenced IQA methods. we will further expound this in Section V-D.

The qualitative results are shown in Fig. 5, our model generates favorable results with a fast inference procedure. The proposed evaluation scores are shown in the top left corner (view in blue). The NIQE [73] and PIQE [14] are shown in green color in the second and third row, respectively. On the one hand, the proposed AQA method is sensitive to the local distortions and global color chromatism, while NIQE [73] and PIQE [14] do not generate favorable results to match the visual experience. On the other hand, our proposed attentive deep stitching approach shows less breakage and color chromatism, while other methods show apparent stitching error in stitching regions. Moreover,

TABLE I
JOINT BENCHMARKING OF 8 STITCHING MODELS WITH 10 IQA METRICS. ADS: ATTENTIVE DEEP STITCHING METHOD. AQA: PROPOSED ATTENTIVE QUALITY ASSESSMENT. THE ASSESSMENT RANKS OF THE 1ST AND 2ND PLACE IN EACH ROW ARE VIEW IN **BOLD** AND $\underline{\text{UNDERLINED}}$. ↑: THE HIGHER THE BETTER. ↓: THE LOWER THE BETTER

	Method	SamsungG. [71]	OpenSource [72]	SVP [28]	ManMethod	IP [29]	EP [30]	MLS [33]	**ADS**
No-Reference	CNNIQA [15] ↑	$\underline{21.12}$	19.02	19.52	18.56	20.76	19.33	17.93	**35.81**
	CEIQ [48] ↑	**3.438**	3.291	3.220	3.262	3.383	3.344	3.384	$\underline{3.405}$
	PIQE [14] ↑	$\underline{32.25}$	**45.60**	23.83	29.38	30.19	28.34	27.99	28.33
	CCF [47] ↑	16.56	**19.27**	12.41	15.10	16.36	13.90	$\underline{17.76}$	14.39
	BRISQUE [43] ↓	30.02	31.18	**15.79**	$\underline{21.74}$	31.67	24.73	27.06	45.02
	NIQE [73] ↓	3.443	2.969	$\underline{2.772}$	3.226	3.230	3.306	**2.283**	3.550
Full-Reference	MSE [19] ↓	0.077	$\underline{0.056}$	0.088	0.058	0.113	0.106	0.052	**0.033**
	PSNR [20] ↑	12.25	12.57	10.94	$\underline{13.04}$	9.77	10.05	12.90	**15.90**
	SSIM [49] ↑	0.636	0.603	0.604	$\underline{0.698}$	0.533	0.557	0.624	**0.708**
	AQA ↑	**86.76**	64.80	49.57	25.19	26.16	25.76	27.12	$\underline{75.23}$



(a) SamsungGear (b) OpenSource (c) **ADS** (d) SVP (e) ManMethod (f) IP (g) EP (h) MLS

Fig. 5. Benchmark evaluations. The proposed evaluation scores are shown in the top left corner (view in blue). The NIQE [73] and PIQE [14] are shown in green color in the second and third row, respectively. Our proposed deep stitching method with the second highest scores show fewer distortions and color chromatism, especially in the attention regions in the second row.

benefiting from our network architecture, the blind zones in our results more also smaller than the state-of-the-art models.

C. Analysis of Image Quality Assessment

To further evaluate the effectiveness of the proposed quality assessment metric, we further use the five commonly used metrics (e.g., CS, PRCC, SROCC, KRCC, RMSE) to evaluate our IQA scores. We adopt the pair-wise ranking of 14,847 images comparisons as the MOS results. As shown in Table III, Our evaluation show a large superior margin comparing to the second best result of PIQE [14] in CS, PRCC, SROCC and KRCC. However, our proposed IQA method generates comparable results to the PIQE with a slightly lower of 3%, while most of the classical metrics failed to handle this kind of images. Despite the existing evaluation metrics in Table II, we adopt the MOS evaluation framework of [74] to transform the subjective MOS scores into pair-wise significance. From Table IV, our proposed AQA method is also highly matched with the human subjective evaluations with the AUC value of 0.965 and surpasses the state-of-the-art methods.

To verify the robustness of the proposed AQA algorithm, we split our dataset into different parts and test our algorithm under various scenarios. The qualitative results are shown in Table V, where our model show robustness in the various conditions. Joint analyzing with Table III, the lowest results of our proposed algorithms still show superiority to the state-of-the-arts. From Table V, we can easily conclude that indoor-scenes are more challenging than the outdoors mainly because of the various object and light changes. To further lightness parameters in our assessment metrics, we divided the test images into two groups, the natural-light and no-natural (e.g., indoor light,

TABLE II
MATHEMATICAL FORMULATION OF 5 CORRELATION METRICS

Metrics	Mathematical Formulation
CS	$\mathbf{R}_{xy} = \frac{\mathbf{X}\cdot\mathbf{Y}}{\lVert\mathbf{X}\rVert\lVert\mathbf{Y}\rVert} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$
PRCC	$\mathbf{R}_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$
KRCC	$\mathbf{R}_{xy} = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$
SROCC	$\mathbf{R}_{xy} = \frac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i (x_i-\overline{x})^2 \sum_i (y_i-\overline{y})^2}}$
RMSE	$\mathbf{R}_{xy} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2}$

TABLE III
MOS SIMILARITIES OF 10 STATE-OF-THE-ART IQA METHODS WITH 5 CORRELATIONS CRITERIONS. ↑: THE HIGHER THE BETTER. ↓: THE LOWER THE BETTER

Metrics	MSE [19]	PSNR [20]	SSIM [49]	BRISQUE [43]	NIQE [73]	PIQE [14]	CNN [15]	CCF [47]	CEIQ [48]	**AQA**
CS ↑	0.798	0.832	0.782	0.867	0.871	0.895	0.832	0.887	0.848	**0.948**
PRCC ↑	0.012	0.158	0.089	0.336	0.354	0.476	0.158	0.460	0.279	**0.742**
SROCC ↑	0.012	0.158	0.089	0.336	0.354	0.476	0.158	0.460	0.279	**0.742**
KRCC ↑	0.024	0.143	0.071	0.238	0.262	0.365	0.103	0.389	0.216	**0.596**
RMSE ↓	2.807	2.558	2.938	2.263	2.220	**2.003**	2.567	4.319	5.765	2.067

TABLE IV
MOS SIMILARITIES OF 8 STATE-OF-THE-ART IQA METHODS USING PAIR-WISE MOS FRAMEWORK [74]

Method	MSE [19]	PSNR [20]	SSIM [49]	BRISQUE [43]	NIQE [73]	PIQE [14]	CNNIQA [15]	**AQA**
AUC	0.823	0.861	0.800	0.939	0.922	0.940	0.865	**0.965**

TABLE V
COMPARISONS WITH HUMAN SUBJECT SUBJECTIVE EVALUATIONS. EVALUATION SCORES: TOTAL WINING PROPORTIONS VIA PAIR-WISE COMPARISONS

Metrics	Indoor	Outdoor	Natural-light	No Natural-light
CS ↑	0.947	0.951	**0.961**	0.933
PRCC ↑	0.735	0.757	**0.804**	0.667
SROCC ↑	0.735	0.757	**0.804**	0.667
KRCC ↑	0.571	**0.657**	0.635	0.508
RMSE ↓	2.119	1.943	**1.571**	2.667

TABLE VI
MOS SIMILARITY EVALUATIONS WITH LOCAL INDICATORS. AQA (ALL): AQA METHOD WITH ALL LOCAL AND GLOBAL METRICS

Method	Similarity with MOS
AQA (All)	**0.948**
Sparse Reconstruction	0.794
Appearance Similarity	0.811
Global Color Chromatism	0.803
SSIM	0.782
MSE	0.798
CNNIQA	0.832

TABLE VII
ABLATION STUDY OF ADS WITH STATE-OF-THE-ART IQA METRICS

Model	SSIM	PSNR	NIQE	AQA
Progressive-1	0.77	15.52	**4.34**	38.36
W/o Attention	0.72	15.35	4.36	41.23
Progressive-All	**0.83**	**16.29**	4.36	**51.56**

TABLE VIII
TIME COST OF ADS AND STATE-OF-THE-ART MODELS WITH THE SAME RESOLUTION OF 2048 × 1024 STITCHED OUTPUT

Method	Time per hundred images
EP [30]	322.3s
IP [29]	375.9s
SVP [28]	362.3s
OpenSource [72]	133.5±4.5 s
ADS	**21.95s**

streetlight). Our assessment still faces a challenge in handling these conditions, with a sharp 17% drop in sensitive PRCC and SROCC metrics, while getting acceptable results in CS metric (3% lower).

We use single indicator to evaluate the scores and conduct CS similarity with MOS scores. The results can be found in Table VI. With our single appearance similarity indicators, the CS similarity with MOS can be 0.811, which is higher than the classical MSE and SSIM indicators. The single global color chromatism indicator also shows 0.803 similarity with MOS. With our full-reference AQA algorithm, the full algorithm reach the performance of 0.948. This also verifies that our local and global module are complementary and can boost the performance together.

We further evaluate the time efficiency, and compare our method with state-of-the-art IQA approaches. The execution time of our method is evaluated on a single Intel I7-6700 CPU. For a single 2048 × 1024 image, our proposed AQA method costs 5.38 seconds and CCF [47] costs 6.98 seconds per image. While the classical SSIM and PSNR indexes with lower performance cost 0.604 s and 0.275 s per image respectively.

D. Analysis of Attentive Deep Stitching

To evaluate the effectiveness of our proposed Attentive Deep Stitching (ADS), we compare the stitching time with four state-of-the-art automatic stitching methods. It can be conclude that our stitching method runs 15 times faster than the classical Stereoscopic Vision Projection (SVP) [28], Isometric Projection (IP) [29] and Equidistant Projection (EP) [30]. Thanks to the two-stage lightweight design, our method runs over 6 times faster than the state-of-the-art OpenSource [72] toolbox, which is shown in Table VIII.

To verify the design of our proposed network architecture, we conduct ablation study in the proposed ADS. We randomly select a half from the test set for this validation with state-of-the-art IQA metrics. As shown in Table VII, the first row of Progressive-1 denotes the first iterated output of the recurrence network. With 3 times recurrence, the stitching results boost PSNR from 15.52 to 16.29. While our model without the attention mechanism generates much coarser results than the final model in the third line. Our proposed method is consistent with human subjective

Fig. 6. Results of different iterative resolutions. The first row is the stitching results in the iteration of 512×256. The second row shows the final results with resolutions of 2048×1024. The proposed progressive module provide increasing details, especially in the attention regions, e.g., cars in the first row and widows in the second and third row.

Fig. 7. Visualized results of our ADS algorithm. ADS w/o Attention: ADS algorithm without attention module. ADS-Final: the final results with proposed attention mechanism.

evaluations while the NIQE [73] is not sensitive to this kind of image quality. The proposed AQA is also sensitive to the image quality improvement for varying from 38.36 to 51.56 in the progressive iterations. The visualized ablation results of our attention mechnism can be found in Fig. 7. The corresponding attention map can be found in third column. Compared to the second column with our full model, the results without attention mechanism lose many details in local regions, especially the attentive regions.

The qualitative results are shown in Fig. 6. Comparing the first row with fewer iterations and the second row, the proposed progressive module provide increasing details, especially in the attention regions (e.g., cars in the first row and widows in the second and third row). However, with the limitation of GPU memory, obtaining results with higher resolutions is still a challenging task, which is the largest limitation of our module.

The last thing we want to emphasize is the choice of attention algorithms. As shown in Table IX, we adopt three different ways in generating attention regions for stitching algorithms. We adopt the super-pixel based attention generation method [77] designed for 360° omni-directional images. The second row and the third row indicate the SALICON model trained with or without the omnidirectional benchmark [78] images. It can be easily concluded that the results with deep SALICON model [62] show better performance than the classical method [77]. With more accurate saliency annotations in 360° scenarios, most of the IQA metrics show a slight performance boost.

TABLE IX
ADS With Different Attention Methods. The Best Performances Are in Bold. ↑: The Higher the Better. ↓: The Lower The Better. †: Trained With Salient 360 Dataset [78]

Model	MSE↓	PSNR↑	SSIM↑	BRIS.↓	NIQE↓
ADS (suppix [77])	0.041	14.67	**0.710**	55.56	3.76
ADS (SALICON)	**0.033**	15.90	0.708	57.64	3.55
ADS (SALICON)†	**0.033**	**15.91**	0.699	**55.02**	**3.12**

VI. Conclusions

In this paper, we mainly address two concerns with increasing demands in 360° omnidirectional images: how to generate a good omnidirectional image in a fast and robust way and what is a good omnidirectional image for human? To address these two concerns, we develop two human perception-driven approaches, which are attentive deep stitching (ADS) and attentive quality assessment (AQA) for omnidirectional images. Firstly, our progressive attentive deep stitching model consists of two modules, the first to learn the deformation information, the second to progressively enhance the perceptive ability in resolution. To achieve this, we propose a joint implicit and explicit attention mechanism to make our results consistent with human subjective evaluations. Secondly, to accurately evaluate the stitching results, we develop a novel attentive quality assessment approach for 360° omnidirectional images, which consists of two local sensitive metrics to focus on the human attention and stitching region and two global ones on environmental immersions. Qualitative and Quantitative experiments show that our stitching approach generates preferable results with the state-of-the-arts at a 6× faster speed. Moreover, the proposed attentive quality assessment approach for omnidirectional images surpasses the state-of-the-art methods by a large margin and is highly consistent with human subjective evaluations.

References

[1] J. Chalfoun *et al.*, "Mist: Accurate and scalable microscopy image stitching tool with stage modeling and error minimization," *Sci. Rep.*, vol. 7, 2017, Art. no. 4988.

[2] E. A. Semenishchev, V. V. Voronin, V. I. Marchuk, and I. V. Tolstova, "Method for stitching microbial images using a neural network," *Proc. SPIE*, vol. 10221, 2017, Art. no. 102210.

[3] D. Li, Q. He, C. Liu, and H. Yu, "Medical image stitching using parallel sift detection and transformation fitting by particle swarm optimization," *J. Med. Imag. Health Informat.*, vol. 7, no. 6, pp. 1139–1148, Oct. 2017.

[4] L. Barazzetti, M. Previtali, and F. Roncoroni, "3D modelling with the samsung gear 360," *ISPRS-Int. Archives Photogramm., Remote Sens. Spatial Inf. Sci.*, pp. 85–90, 2017. Ref [4]: (Vol. 42, No. 2W3, pp. 85-90)

[5] D. Chapman and A. Deacon, "Panoramic imaging and virtual realityłfilling the gaps between the lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 53, no. 6, pp. 311–319, 1998.

[6] G. Payen de La Garanderie, A. Atapour Abarghouei, and T. P. Breckon, "Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360-degree panoramic imagery," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 789–807.

[7] T. Z. Xiang, G. S. Xia, X. Bai, and L. Zhang, "Image stitching by line-guided local warping with global similarity constraint," *Pattern Recognit.*, pp. 481–497, 2018. Ref [7]: Vol 83 (2018): 481-497.

[8] W. Ye, K. Yu, Y. Yu, and J. Li, "Logical stitching: A panoramic image stitching method based on color calibration box," in *Proc. IEEE Int. Conf. Signal Process.*, 2018, pp. 1139–1143.

[9] I. C. Lo, K. T. Shih, and H. H. Chen, "Image stitching for dual fisheye cameras," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 3164–3168.

[10] T. Ho and M. Budagavi, "Dual-fisheye lens stitching for 360-degree imaging," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 2172–2176.

[11] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, 2007.

[12] P. C. Ng and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, 2003.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, May 2006, pp. 404–417.

[14] V. N. , P. D. , M. C. Bh., S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 21st Nat. Conf. Commun.*, Feb./Mar. 2015, pp. 1–6.

[15] L. Kang, P. Ye, Y. Li, and D. S. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1733–1740.

[16] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 1040–1049.

[17] G. Cheung, L. Yang, Z. Tan, and Z. Huang, "A content-aware metric for stitched panoramic image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Venice, Italy, Oct. 2017, pp. 2487–2494.

[18] Y. Niu, H. Zhang, W. Guo, and R. Ji, "Image quality assessment for color correction based on color contrast similarity and color value difference," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 28, no. 4, pp. 849–862, Apr. 2018.

[19] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[20] A. Tanchenko, "Visual-PSNR measure of image quality," *J. Visual Commun. Image Representation*, pp. 874–878, 2014. Ref[20]: Volume 25 Issue(5).

[21] Z. Wang *et al.*, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[22] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[23] L. Kang, P. Ye, Y. Li, and D. S. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process.*, Quebec City, QC, Canada, Sep. 2015, pp. 2791–2795.

[24] G. Cheung, L. Yang, Z. Tan, and Z. Huang, "A content-aware metric for stitched panoramic image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Venice, Italy, Oct. 2017, pp. 2487–2494.

[25] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2018, pp. 1–5.

[26] J. Li, K. Yu, Y. Zhao, Y. Zhang, and L. Xu, "Cross-reference stitching quality assessment for 360 omnidirectional images," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2360–2368.

[27] O. S. Vaidya and S. T. Gandhe, "The study of preprocessing and postprocessing techniques of image stitching," in *Proc. Int. Conf. Adv. Commun. Comput. Technol.*, Feb. 2018, pp. 431–435.

[28] B. Maneshgar, L. Sujir, S. P. Mudur, and C. Poullis, "A long-range vision system for projection mapping of stereoscopic content in outdoor areas," in *Proc. 12th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Porto, Portugal, Feb./Mar. 2017, pp. 290–297.

[29] D. Cai, X. He, and J. Han, "Isometric projection," in *Proc. 22nd AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, Jul. 2007, pp. 528–533.

[30] D. Schneider, E. Schwalbe, and H. G. Maas, "Validation of geometric models for fisheye lenses," *ISPRS J. Photogrammetry Remote Sens.*, vol. 64, no. 3, pp. 259–266, 2009.

[31] W. Li, C.-B. Jin, M. Liu, H. Kim, and X. Cui, "Local similarity refinement of shape-preserved warping for parallax-tolerant image stitching," *Institution Eng. Technol.*, pp. 661–668, 2017. Ref[31]: Volume 12 issue 5

[32] T. Xiang, G.-S. Xia, and L. Zhang, "Image stitching with perspective-preserving warping," 2016, *arXiv:1605.05019*.

[33] T. Ho, I. D. Schizas, K. Rao, and M. Budagavi, "360-degree video stitching for dual-fisheye lens cameras based on rigid moving least squares," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 51–55.

[34] C. Herrmann *et al.*, "Robust image stitching with multiple registrations," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[35] L. Yao, Y. Lin, C. Zhu, and Z. Wang, "An effective dual-fisheye lens stitching method based on feature points," in *Proc. Int. Conf. Multimedia Model.*, Jan. 2019, pp. 665–677.

[36] J. Tan, G. Cheung, and R. Ma, "360-degree virtual-reality cameras for the masses," *IEEE Multimedia*, vol. 25, no. 1, pp. 87–94, Jan./Mar. 2018.

[37] I.-C. Lo, K.-T. Shih, and H. H. Chen, "Image stitching for dual fisheye cameras," in *Proc. Int. Conf. Image Process.*, Oct. 2018, pp. 3164–3168.

[38] X. Li, Y. Pi, Y. Jia, Y. Yang, Z. Chen, and W. Hou, "Fisheye image rectification using spherical and digital distortion models," *Multispectral Image Acquisition Process. Anal.*, 2018, Art. no. 106070G.

[39] X. Yin, X. Wang, J. Yu, P. Fua, M. Zhang, and D. Tao, "Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 469–484.

[40] L. Deng, M. Yang, H. Li, T. Li, B. Hu, and C. Wang, “Restricted deformable convolution based road scene semantic segmentation using surround view cameras,” 2018, *arXiv:1801.00708*.

[41] C. Li, M. Xu, S. Zhang, and P. L. Callet, “State-of-the-art in 360° video/image processing: Perception, assessment and compression,” 2019, *arXiv:1905.00161*.

[42] Y. Qian, D. Liao, and J. Zhou, “Manifold alignment based color transfer for multiview image stitching,” in *Proc. IEEE Int. Conf. Image Process.*, Melbourne, Australia, Sep. 2013, pp. 1341–1345.

[43] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[44] C. Richardt, Y. Pritch, H. Zimmer, and A. Sorkine-Hornung, “Megastereo: Constructing high-resolution stereo panoramas,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 1256–1263.

[45] J. Zaragoza, T. Chin, Q. Tran, M. S. Brown, and D. Suter, “As-projective-as-possible image stitching with moving DLT,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1285–1298, Jul. 2014.

[46] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1098–1105.

[47] Y. Wang *et al.*, “An imaging-inspired no-reference underwater color image quality assessment metric,” *Comput. Elect. Eng.*, vol. 70, pp. 904–913, 2018.

[48] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, “No-reference quality assessment of contrast-distorted images based on natural scene statistics,” *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 838–842, Jul. 2015.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[50] Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[51] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[52] L. Liu, H. Dong, H. Huang, and A. C. Bovik, “No-reference image quality assessment in curvelet domain,” *Signal. Process Image Commun.*, vol. 29, pp. 494–505, 2014.

[53] S. Ling, G. Cheung, and P. L. Callet, “No-reference quality assessment for stitched panoramic images using convolutional sparse coding and compound feature selection,” in *Proc. IEEE Int. Conf. Multimedia Expo*, San Diego, CA, USA, Jul. 2018, pp. 1–6.

[54] Y. Yuan, Q. Guo, and X. Lu, “Image quality assessment: A sparse learning way,” *Neurocomputing*, vol. 159, pp. 227–241, 2015.

[55] C. Zhang, J. Pan, S. Chen, T. Wang, and D. Sun, “No reference image quality assessment using sparse feature representation in two dimensions spatial correlation,” *Neurocomputing*, vol. 173, pp. 462–470, 2016.

[56] M. Huang *et al.*, “Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression,” *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 6039–6050, Dec. 2018.

[57] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, “Assessing visual quality of omnidirectional videos,” *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[58] M. Xu, C. Li, Y. Liu, X. Deng, and J. Lu, “A subjective visual quality assessment method of panoramic videos,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 517–522.

[59] Y. Rai, P. L. Callet, and P. Guillotel, “Which saliency weighting for omni directional image quality assessment?” in *Proc. 9th Int. Conf. Quality Multimedia Experience*, Erfurt, Germany, May/Jun. 2017, pp. 1–6.

[60] E. Upenik, M. Rerábek, and T. Ebrahimi, “Testbed for subjective evaluation of omnidirectional visual content,” in *Proc. Picture Coding Symp.*, Nuremberg, Germany, Dec. 2016, pp. 1–5.

[61] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, Oct. 2015, pp. 234–241.

[62] X. Huang, C. Shen, X. Boix, and Q. Zhao, “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 262–270.

[63] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O’Connor, “Saltinet: Scan-path prediction on 360 degree images using saliency volumes,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2331 2338.

[64] Y. Rai, J. Gutiérrez, and P. Le Callet, “A dataset of head and eye movements for 360 degree images,” in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 205–210.

[65] M. Startsev and M. Dorr, “360-aware saliency estimation with conventional image saliency predictors,” *Signal Process., Image Commun.*, vol. 69, pp. 43–52, 2018.

[66] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *J. Mach. Learn. Res.*, vol. 11, no. Jan, pp. 19–60, 2010.

[67] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[68] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. 7th Int. Conf. IEEE Comput. Vis.*, 1999, pp. 1150–1157.

[69] C. Li, A. C. Bovik, and X. Wu, “Blind image quality assessment using a general regression neural network,” *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 793–799, May 2011.

[70] K. J. Preacher, P. J. Curran, and D. J. Bauer, “Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis,” *J. Educational Behav. Statist.*, pp. 437–448, 2006. Ref[70]:Volume 31 Issue 4

[71] S. Group, “Samsung gear 360 stitching toolbox,” 2017. [Online]. Available: https://www.samsung.com/global/galaxy/gear-360

[72] Github.com, “Dualfisheye,” 2016. [Online]. Available: https://github.com/ooterness/DualFisheye.

[73] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ”completely blind” image quality analyzer,” *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[74] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, “On the accuracy of objective image and video quality models: New methodology for performance evaluation,” in *Proc. IEEE 8th Int. Conf. Quality Multimedia Experience*, 2016, pp. 1–6.

[75] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “Salicon: Saliency in context,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1072–1080.

[76] T.-Y. Lin *et al.*, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 740–755.

[77] Y. Fang, X. Zhang, and N. Imamoglu, “A novel superpixel-based saliency detection model for 360-degree images,” *Signal Process., Image Commun.*, vol. 69, pp. 1–7, 2018.

[78] J. Gutiérrez, E. J. David, A. Coutrot, M. P. Da Silva, and P. Le Callet, “Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360 contents,” in *Proc. IEEE 10th Int. Conf. Quality Multimedia Experience*, 2018, pp. 1–3.

Jia Li (M’12–SM’15) received the B.E. degree from Tsinghua University, Beijing, China, in 2005, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in June 2014, he used to conduct research in Nanyang Technological University, Peking University and Shanda Innovations. He is the author or coauthor of more than 70 technical articles in refereed journals and conferences such as TPAMI, IJCV, TIP, CVPR, and ICCV. His research interests include computer vision and multimedia big data, especially the learning-based visual content understanding. He is a senior member of CIE and CCF. More information can be found at http://cvteam.net

Yifan Zhao received the B.E. degree from the Harbin Institute of Technology, Harbin, China, in Jul. 2016. He is currently working toward the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include computer vision and image understanding.

Weihua Ye is currently working toward the M.S. Degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include computer vision and virtual reality.

Kaiwen Yu is currently working toward the M.S. degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include image quality assessment and deep learning.

Shiming Ge (M'13–SM'15) received the B.S. and Ph.D degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He is also the member of Youth Innovation Promotion Association, Chinese Academy of Sciences. Prior to that, he was a Senior Researcher and Project Manager in Shanda Innovations, a Researcher in Samsung Electronics and Nokia Research Center. His research mainly focuses on computer vision, data analysis, machine learning and AI security, especially efficient learning models and solutions toward scalable applications.