Towards Smart Point-and-Shoot Photography

Jiawan Li[1,2,3,4] , Fei Zhou[1,2,3,4,*] , Zhipeng Zhong[5] , Jiongzhi Lin[1,2,3,4] , Guoping Qiu[6,7*]
[1] Shenzhen University,[2] Guangdong Provincial Key Laboratory of Intelligent Information Processing
[3]Guangdong-Hong Kong Joint Laboratory for Big Data Imaging and Communication
[4] Shenzhen Key Laboratory of Digital Creative Technology
[5] Loughborough University, [6] University of Nottingham,[7] Everimaging Ltd
lijiawan2022@email.szu.edu.cn,flying.zhou@163.com,2019281030@email.szu.edu.cn
Z.Zhong@lboro.ac.uk, guoping.qiu@nottingham.edu.cn

Abstract

Hundreds of millions of people routinely take photos using their smartphones as point and shoot (PAS) cameras, yet very few would have the photography skills to compose a good shot of a scene. While traditional PAS cameras have built-in functions to ensure a photo is well focused and has the right brightness, they cannot tell the users how to compose the best shot of a scene. In this paper, we present a first of its kind smart point and shoot (SPAS) system to help users to take good photos. Our SPAS proposes to help users to compose a good shot of a scene by automatically guiding the users to adjust the camera pose live on the scene. We first constructed a large dataset containing $320K$ images with camera pose information from 4000 scenes. We then developed an innovative CLIP-based Composition Quality Assessment (CCQA) model to assign pseudo labels to these images. The CCQA introduces a unique learnable text embedding technique to learn continuous word embeddings capable of discerning subtle visual quality differences in the range covered by five levels of quality description words $\{bad, poor, fair, good, perfect\}$. And finally we have developed a camera pose adjustment model (CPAM) which first determines if the current view can be further improved and if so it outputs the adjust suggestion in the form of two camera pose adjustment angles. The two tasks of CPAM make decisions in a sequential manner and each involves different sets of training samples, we have developed a mixture-of-experts model with a gated loss function to train the CPAM in an end-to-end manner. We will present extensive results to demonstrate the performances of our SPAS system using publicly available image composition datasets.

*Corresponding author

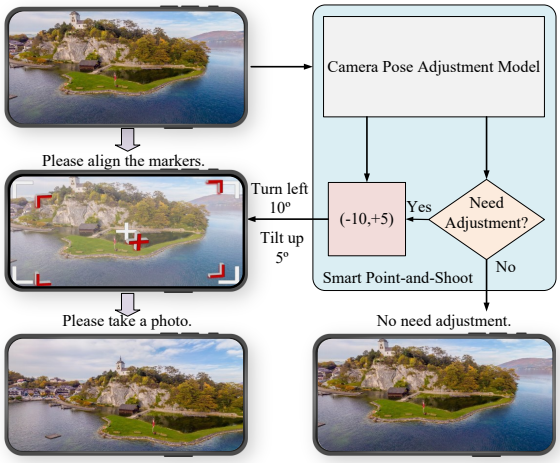
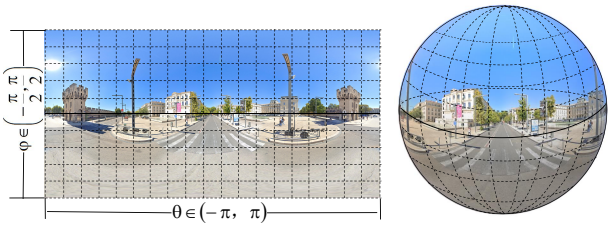


Figure 1. Given a view composed by the user, our Smart Point-and-Shoot (SPAS) system can predict camera pose adjustment suggestions so that the photo captured after applying the adjustment will have a better composition.

1. Introduction

Traditional Point-and-Shoot (PAS) cameras have built-in functions such as autofocus, autoexposure, and auto-flash to ensure a photograph is well focused and has the right brightness. However, these PAS cameras cannot tell the users how to compose the best shot of a scene. It is estimated that there are over 7 billion smartphones worldwide and every one is a PAS camera (in the context of this paper, smartphone and camera are used interchangeably). Although almost every smartphone user would routinely use their phones to take photos, very few would have the photography skill to compose a good shot of a scene. In this paper, we present a solution that automatically guides smartphone users to compose the best shot live on a scene.



(a) Equirectangular Projection (b) Sphere

Figure 2. Geometric explanation of the relationship between ERP (a) and the sphere (b).

Dataset	Year	Label	Scenes	Candidate Views	Camera Pose
ICDB[28]	2013	Best	950	1	N/A
HCDB[4]	2014	Best	500	1	N/A
GNMC[3]	2022	Best	10000	5	N/A
SACD[29]	2023	Best	2777	8	N/A
FCDB[2]	2017	Rank	1536	18	N/A
CPC[26]	2018	Rank	10800	24	N/A
GAICv1[31]	2019	Score	1236	86	N/A
GAICv2[32]	2020	Score	3336	86	N/A
UGCrop5K[23]	2024	Score	5000	90	N/A
PCARD(Ours)	2024	Score	4000	81	324000

Table 1. Image Composition datasets and PCARD.

For a given scene of interest and starting from an initial view a user points to, our Smart Point-and-Shoot (SPAS) system automatically recommends camera pose adjustment strategies and guide the user to rotate the camera upwards, downwards, rightwards or leftwards until the camera points to the best shot. In contrast to existing literature on automatic picture composition which is a post-processing procedure of cropping a photo that has already been taken from a fixed view, our SPAS is the first system that enables users to compose the best shot of a scene by guiding the users to adjust the camera pose live on the scene.

As shown in Figure 1, given an initial view, the camera pose adjustment model (CPAM) first evaluates whether the composition can be improved. If so, it predicts how the camera pose should be adjusted. Specifically, let θ, φ, and γ respectively denote the yaw, pitch, and roll angles of a camera pose $P = (\theta, \varphi, \gamma)$. Because it is unusual to roll the camera during shooting, it is reasonable to assume that the roll angle is fixed. The CPAM therefore suggests how to rotate along the vertical axis (change yaw angle θ) and how to rotate along the horizontal axis (change pitch angle φ). By providing camera pose adjustment suggestions during the shooting process, we can help the users to effectively improve the composition and take a good shot of the scene. The challenge now is how to construct the camera pose adjustment model (CPAM).

First of all, we require a suitably annotated dataset and then use the data to construct an intelligent model that can first determine if a given view's composition can be improved and if so how the camera pose should be changed in order to obtain an improved shot. The challenge of obtaining a large enough dataset is significant. Manually acquiring images of different camera poses from a variety of scenes and then annotate them with composition scores will be extremely time consuming and therefore is impracticable. In terms of the CPAM itself, it needs to perform sequential decision making on two tasks. It is therefore particularly important to model the relationship between the tasks to avoid conflicts arising from task discrepancies. In this paper, we have developed practical solutions to these problems.

To construct a dataset for the problem, we first take advantage of the availability of 360° images of Google Street View[1]. By exploiting the geometric explanation of the relationship between the equirectangular projection (ERP) of the 360° image and the sphere (see Figure 2), we discover that a panoramic image in the ERP format can be mapped onto the surface of a unit sphere. This mapping creates a complete 360° photographic environment where spherical coordinates (longitude θ and latitude φ) naturally correspond to the orientation of a virtual camera positioned at the sphere's center. Through this geometric correspondence, we can precisely control the camera's viewing direction using these spherical coordinates, enabling the generation of sample views with well-defined camera poses (θ, φ) via perspective projection. Based on this observation, we create the Panorama-based Composition Adjustment Recommendation dataset (PCARD). As shown in Table 1, the new PCARD contains $320K$ images with camera pose information from 4000 scenes. As far as we know, this is the first dataset created from the 360° images of Google Street View where each image contains the camera pose information. We will use the PCARD to develop a smart point and shoot (SPAS) solution.

For the $320K$ images in PCARD, it is necessary to assign each a quality label. Again manual approach is impractical. Instead we resort to images with composition quality ratings such as those in [32] to train a labeler to assign pseudo composition score labels to these images. One of the major challenges in developing the pseudo labeler is that neigbouring views have large overlapping regions and are very similar, therefore the labeler needs to have the ability to distinguish images with subtle differences. In this paper, we take full advantage of large language models (LLM) and have developed a CLIP-based Composition Quality Assessment (CCQA) model. As CLIP is sensitive to the choice of prompts and text descriptions of nuance visual differences are difficult, we abandon traditional subjective prompt set-

[1] https://www.google.com/streetview/

tings in favor of learnable text prompts. We have developed an effective method that learns continuous word embeddings capable of discerning subtle visual quality differences in the range covered by five levels of quality description words $\{bad, poor, fair, good, perfect\}$.

The Camera Pose Adjustment model (CPAM) performs two tasks in a sequential manner. Logically, it needs to first determine if the current view can be further improved and if so it then outputs the adjust suggestion in the form of two pose adjustment angles. This is a multitask learning problem but logically the decisions must be made in a sequential manner. Also, unlike normal multitask learning, the two learning tasks involve different training samples with one involves the full set and the other a subset of the training samples. To tackle this problem, we have developed a mixture-of-experts model with a gated loss function to train the CPAM in an end-to-end manner. In summary, this paper makes 4 major contributions:

- We present a first of its kind smart point and shoot (SPAS) system to help the billions of smartphone users to take good photographs. Our SPAS is the first in the literature that proposes to help users to compose a good shot of a scene by automatically guiding the users to adjust the camera pose live on the scene.
- We have constructed a large dataset containing $320K$ images with camera pose information from 4000 scenes by exploiting the availability of $360°$ images of Google Street View. This dataset which will be made publicly available and can be used for the task in this paper as well as other applications.
- We have developed an innovative CLIP-based Composition Quality Assessment (CCQA) model. The CCQA introduces a unique learnable text embedding technique to learn continuous word embeddings capable of discerning subtle visual quality differences in the range covered by five levels of quality description words $\{bad, poor, fair, good, perfect\}$.
- We have developed a camera pose adjustment model (CPAM) which first determines if the current view can be improved and if so it outputs the adjust suggestion in the form of two camera pose adjustment angles. The two tasks of CPAM make decisions in a sequential manner and each involves different sets of training samples, we have developed a mixture-of-experts model with a gated loss function to train the CPAM in an end-to-end manner.

2. Related Work

Image Composition dataset. For photo recommendation tasks, there exist some image cropping datasets [2–4, 23, 26, 28, 29, 31, 32] that can be categorized into two groups based on their annotation styles, as shown in Table 1. More details can be seen in the Supplementary Material.

Aesthetic-guided image composition. Image aesthetic quality assessment aims to quantify image aesthetic values, while image composition focuses on finding the most aesthetic view. While prior works [1, 14, 17, 18, 33] learn aesthetic-related features to evaluate composition quality, they lack recommendation capabilities. Instead, image cropping, which aims to find the most aesthetic sub-region through cropping boxes, has emerged as a promising direction. Existing methods [2, 7–9, 11–13, 15, 16, 19, 20, 23–26, 31, 32, 36] generally fall into two categories: score-based methods [2, 13, 19, 20, 23–26, 30–32] that evaluate candidate views using learned aesthetic knowledge, and coordinate regression-based [7–9, 11, 12, 15, 16, 36] methods that directly predict optimal cropping boxes through various learning strategies.

Although previous methods have achieved good results for cropping-based image composition tasks, image cropping is a post-processing exercise applied to already captured images where the viewpoints have already been fixed. It is not applicable in scenarios where the photographer needs to adjust the camera pose or position to capture the best view of a scene. In this work, we present a framework that automatically provides photographers with camera pose adjustment directions and guides the photographers to take the best shot of a given scene.

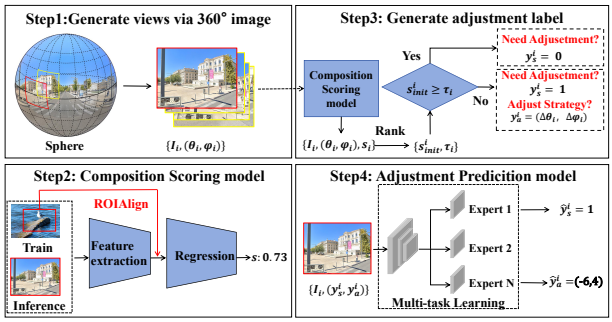


Figure 3. The overview of our method. Using perspective projection, we generate views from $360°$ images with camera poses (Step 1). We train a composition scoring model to evaluate image composition quality (Step 2) and design a composition quality score-guided method to generate camera pose adjustment labels (Step 3). Finally, a sequential multi-task MoE network predicts camera adjustments to improve image composition (Step 4).

3. Problem Definition and Overview

In general, a photographer assesses an initial view through the viewfinder and then adjusts the camera pose utilizing 3 degrees of freedom in the 3D world space (yaw θ, pitch φ, roll γ) to take the best shot.

Given an initial view $\boldsymbol{I}^i_{init}$ of the i^{th} scene, and a camera pose adjustment prediction model $f(\cdot)$, the problem can be

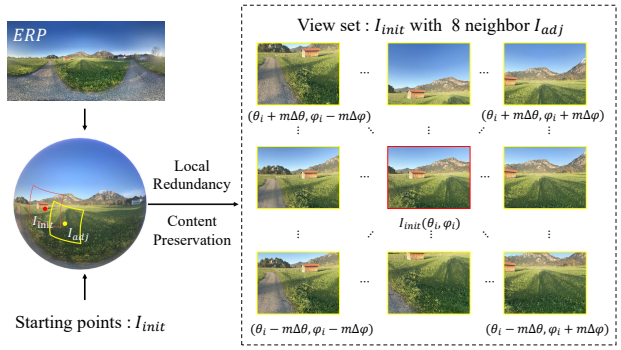


Figure 4. A multi-angle view generation method based on the 360° images.

formulated by

$$\left(\widehat{\boldsymbol{y}}_s^i, \widehat{\boldsymbol{y}}_a^i\right) = f\left(\boldsymbol{I}_{init}^i\right) \tag{1}$$

where $\widehat{\boldsymbol{y}}_s^i$ and $\widehat{\boldsymbol{y}}_a^i$ respectively represent the suggestion output and the adjustment output. $\widehat{\boldsymbol{y}}_s^i$ indicates whether the composition of an initial view $\boldsymbol{I}_{init}^i$ can be improved. If the composition can be improved, the adjustment predictor predicts the suitable camera pose adjustment strategy, which is $(\Delta\theta_i, \Delta\varphi_i, \Delta\gamma_i)$. In practice, it is unusual to roll the camera during the photography process, we therefore fix the roll angle γ. The camera pose and the camera pose adjustment strategy can be further simplified to (θ_i, φ_i) and $(\Delta\theta_i, \Delta\varphi_i)$. $\theta_i \in [-180°, 180°]$ and $\varphi_i \in [-90°, 90°]$. $\Delta\theta \in [-180°, 180°]$ represents the camera pose rotating left or right around the vertical axis, with rightward rotation being positive. $\Delta\varphi \in [-180°, 180°]$ represents the camera pose rotating up or down around the horizontal axis, with upward being positive. The pipeline of the whole approach is illustrated in Figure 3. First, to train the camera pose adjustment prediction model $f(\cdot)$, we create the Panorama-based Composition Adjustment Recommendation dataset $\mathcal{D}_{\mathcal{PCARD}} = \left\{\boldsymbol{I}_{init}^i, \boldsymbol{y}_s^i, \boldsymbol{y}_a^i\right\}_{i=1}^{N_{\text{scene}}}$ and present a pseudo-labeling method guided by composition quality scores to generate the camera pose adjustment labels $(\boldsymbol{y}_s^i, \boldsymbol{y}_a^i)$ (Sec. 4). Specially, we propose a CLIP-based Composition Quality Assessment (CCQA) model $h(\cdot)$ to evaluate the composition quality of views $\boldsymbol{I}$ (Sec. 5). Subsequently, the Camera Pose Adjustment model (CPAM) $f(\cdot)$ is illustrated in Sec. 6.

4. PCARD Database

4.1. Formulation

As shown in Figure 4, given an ERP image with spatial resolution $H \times W$, we transform it into a unit sphere S^3 with S^2 as its surface. Every point $(\theta, \phi) \in S^2$ is uniquely defined by its longitude $\theta \in [-\pi, \pi]$ and latitude $\phi \in [-\pi/2, \pi/2]$. In the spherical domain, this can be expressed as:

$$\begin{cases} \theta = \frac{2\pi u}{W} - \pi \\ \varphi = \frac{-\pi v}{H} + \frac{\pi}{2} \end{cases} \tag{2}$$

where $u \in [1, W]$ and $v \in [1, H]$. We assume that a virtual pinhole camera is positioned at the center of the sphere S^3. The visual content is then captured as a planar view $\boldsymbol{I}_{init}^i$ that is determined by the viewing angle $(\theta_{init}^i, \varphi_{init}^i)$, and field of view (fov_x, fov_y) of the camera through perspective projection[5]. By adjusting the camera pose we generate candidate views $\boldsymbol{I}_{adj}^i = \{I_i^j, (\theta_i^j, \varphi_i^j)\}_{j=1}^M$, where $M = \frac{360}{\Delta\theta} \times \frac{180}{\Delta\varphi}$ which is the number of candidate views, $\Delta\theta$ and $\Delta\varphi$ are the view adjustment step-sizes. Then the search space M is efficiently reduced by exploiting the *Content Preservation* and *Local Redundancy* properties. And to complete the dataset construction process, we propose a pseudo-labeling method guided by composition quality scores to generate the camera pose adjustment labels $(\boldsymbol{y}_s^i, \boldsymbol{y}_a^i)$ for the candidate views.

Content Preservation. Generally speaking, the adjusted view $\boldsymbol{I}_{adj}^i$ should preserve the main content of the initial view $\boldsymbol{I}_{init}^i$ to maintain the photographer's intended subject. Hence, we constrain the overlapping area between the adjusted next view $\boldsymbol{I}_{adj}^i$ and the initial view $\boldsymbol{I}_{init}^i$ to be no smaller than a certain proportion of $\boldsymbol{I}_{init}^i$. Note that it is directly defined on a sphere (the 360° images) rather than ERP or the tangent plane [35].

$$\text{SphOverlap}\ (S_{\text{adj}}, S_{\text{init}}) = \frac{A(S_{\text{adj}} \cap S_{\text{init}})}{A(S_{\text{init}})} > \lambda \tag{3}$$

where S_{adj} and S_{init} represent the spherical rectangles corresponding to $\boldsymbol{I}_{adj}^i$ and $\boldsymbol{I}_{init}^i$ in the 360° images respectively, $A(\cdot)$ is the area of the shape and $\lambda \in [0.5, 1)$.

Local Redundancy. Adjusting the camera pose to improve image composition is a problem with local redundancy because suboptimal solutions are also acceptable. Based on Moore neighborhood theory[22], we design a sampling matrix that captures 8 neighboring views around the current camera pose (θ_i, φ_i) at varying distances controlled by a multiplier m, as shown in Figure 4. To efficiently remove redundant candidate views, we set the sampling step sizes $\Delta\theta = \Delta\varphi = 5°$ following [37].

More detailed mathematical calculations can be found in the Supplementary Material.

4.2. Database Construction

We selected 20 countries from Street View Download 360 [2], with an average of 8 cities per country, resulting in a total download of over 150K 360° images in equirectangular projection (ERP) format. We designed a web player based on a Web3D library Three.js[3] to achieve 360° play-

[2] https://svd360.com/
[3] https://threejs.org/

back of the panoramic images. This allows us to inspect the panoramic images for obvious distortion or damage, and if none are present, select suitable initial views and record their camera poses. In the end, we retained 4,000 high-quality panoramic images. For each panoramic image, we first generate the initial view $\boldsymbol{I}^i_{init}$ according to the pre-recorded camera poses $(\theta^i_{init}, \varphi^i_{init})$ and then generate candidate views $\boldsymbol{I}^i_{adj} = \{I^j_i, (\theta^j_i, \varphi^j_i)\}^M_{j=1}$ following the Content Preservation and Local Redundancy in Sec. 4.1, where M is the size of candidate view set. In our final dataset, on average, $M = 81$ which we believe is a reasonable size for learning image composition.

4.3. Label Generation

We propose a labeling method guided by aesthetic scores to generate the camera pose adjustment labels $(\boldsymbol{y}^i_s, \boldsymbol{y}^i_a)$ of the view $\boldsymbol{I}^i_{init}$. To do this, we have designed a CLIP-based Composition Quality Assessment (CCQA) model which will be described in Sec. 5. Given an initial view $\boldsymbol{I}^i_{init}$ with camera pose $(\theta^i_{init}, \varphi^i_{init})$ and its corresponding candidate views $\boldsymbol{I}^i_{adj} = \{I^j_i, (\theta^j_i, \varphi^j_i)\}^M_{j=1}$, we use the CCQA model $h(\cdot)$ to assign numerical composition quality ratings to views: $s = h(I)$. We denoted that s^i_{init} represents the score for the initial view $\boldsymbol{I}^i_{init}$ while s^i_{adj} contains the scores for the candidate views $\boldsymbol{I}^i_{adj}$.

Then, we calculate the adaptive threshold τ_i for each scene i. This threshold is determined by ranking the scores of the candidate views in descending order and selecting the N^{th} score: $\tau_i = Top_N(s^i_{adj})$ where N is a fixed percentage of the total number of candidates M. In practice, we set $N = 25\%$, the detailed information will be discussed in Supplementary.

Finally, we generate the suggestion label $\boldsymbol{y}^i_s$ and adjustment label $\boldsymbol{y}^i_a$ leveraging the adaptive threshold τ_i:

$$\boldsymbol{y}^i_s = \begin{cases} 1, \text{ if } s^i_{init} < \tau_i \\ 0, \text{ if } s^i_{init} >= \tau_i \end{cases} \tag{4}$$

$$\boldsymbol{y}^i_a = \begin{cases} \left(\theta^i_{\text{best}}, \varphi^i_{\text{best}}\right) - \left(\theta^i_{\text{init}}, \varphi^i_{\text{init}}\right), & \text{if } \boldsymbol{y}^i_s = 1 \\ (0, 0), & \text{otherwise} \end{cases} \tag{5}$$

where $(\theta^i_{best}, \varphi^i_{best})$ represents the camera pose of the candidate view with the highest composition quality score.

5. CLIP-based Composition Quality Assessment

We introduce our CLIP-based Composition Quality Assessment (CCQA) model illustrated in Figure 5. The model is trained on the GAICv2 dataset [32] that pairs each image x with multiple cropping view v and their corresponding composition quality scores s.

Image encoder. Given an image x and a set of view v, the image encoder creates global feature maps from x by

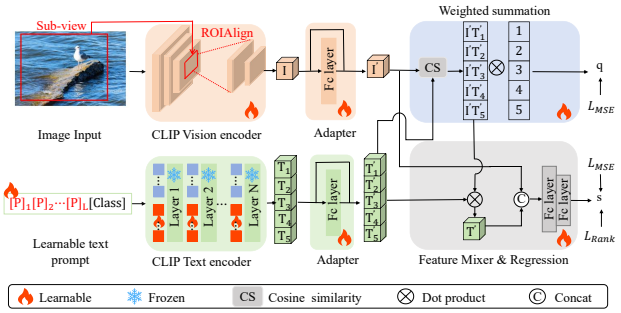


Figure 5. CLIP-based Composition Quality Assessment model.

a trainable CLIP image encoder's first 3 blocks, then uses RoIAlign to extract sub-view feature which are further encoded by CLIP's final block to obtain sub-view embeddings that denoted as visual embedding I.

Learnable prompt. The design of prompts can greatly impact performances. CLIP is sensitive to the choice of prompts, therefore, we abandon traditional subjective prompt settings in favor of a learnable prompt strategy. These learnable text prompts T are defined as follows:

$$T = [P]_1[P]_2 \dots [P]_L[Class] \tag{6}$$

Each $[P]_l (l \in \{1, \dots, L\})$ is a learnable word embedding in the text prompt templates with the same 512 dimensionality as the CLIP word embedding, L represents the number of context tokens. $Class$ is one of five-level quality description words $\{bad, poor, fair, good, perfect\}$.

Feature adapters and Weighted summation. We introduce learnable feature adapters to better leverage CLIP's prior knowledge and enhance visual-text feature synergy. The adapted features I' and T' are obtained by applying residual adaptation and normalization to the visual embedding I and text embeddings T respectively.

The quality weights W_i are computed by applying softmax to the cosine similarities between adapted image feature I' and five class prompts $\{T_i'\}^5_{i=1}$[27, 34].

$$W_i = \frac{\exp\left(I'^{\top} T_i'/\sigma\right)}{\sum_{j=1}^{5} \exp\left(I'^{\top} T_j'/\sigma\right)}, \tag{7}$$

where σ is the temperature parameter. The assessment score q of the given image is calculated as:

$$q = \sum_{i=1}^{5} W_i \times C_i, \tag{8}$$

where $\{C_i\}^5_{i=1}$ are the numerical scores of the five-level quality description words which are set to $1, 2, 3, 4$ and 5 with a lower numerical value corresponds to a lower quality class word.

Feature mixers and regression. To better enable the CLIP features to discern subtle differences in aesthetic quality across a series of similar photos, we obtain the weighted text features F_t through calculating the dot product between the quality weights W_i and the adapted text features $\{T_i{}'\}_{i=1}^{5}$ of the five prompts:

$$F_t = \sum_{i=1}^{5} W_i \times T_i{}', \tag{9}$$

The final score $\hat{s}$ is predicted by passing the concatenated weighted text features F_t and adapted image features I' through an MLP.

Optimization. The CCQA uses a multi-task loss function. The first task enforces the predicted scores to be close to their ground truth scores:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}(\hat{s}_i - s_i)^2 \tag{10}$$

where $\hat{s}_i$ and s_i represent the score predicted by CCQA and the ground truth.

The second task enforces the predicted scores of different views to have the same ranking order as that of the ground truth scores. We therefore also incorporate a ranking loss $\mathcal{L}_2$ to explicitly model the ranking relationship.

$$\mathcal{L}_2 = \frac{\sum_{i,j} \max\left(0, -sign\left(s_i - s_j\right)\left(\left(\hat{s}_i - \hat{s}_j\right) - \left(s_i - s_j\right)\right)\right).}{N(N-1)/2} \tag{11}$$

where $sign(\cdot)$ is the standard sign function.

The third task $\mathcal{L}_3$ enforces consistency between q (cosine similarity-based weighting computed according to Eq. (8)), and the ground truth scores s_i.

$$\mathcal{L}_3 = \frac{1}{N}\sum_{i=1}^{N}(q_i - s_i)^2 \tag{12}$$

The total loss function can be summarized as

$$\mathcal{L}_{CCQA} = \mathcal{L}_1 + \mathcal{L}_2 + \alpha * \mathcal{L}_3 \tag{13}$$

where the hyperparameters α is used to balance different losses (set to 0.1 in this paper).

6. Camera Pose Adjustment Model

Given an image, the Camera Pose Adjustment model (CPAM) $f(\cdot)$ produces two outputs. Firstly, the output of the suggestion predictor $\widehat{\boldsymbol{y}}_s^i$ predicts whether a view adjustment should be performed. Suggestion predictor is a binary classification head. Secondly, the output of the adjustment predictor $\widehat{\boldsymbol{y}}_a^i$ predicts how to adjust the camera pose when the suggestion predictor indicates an adjustment is needed.

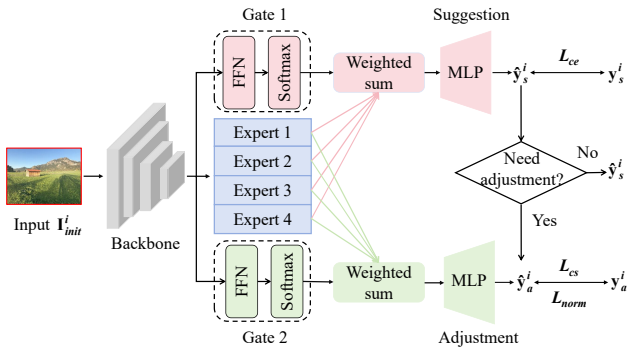


Figure 6. Camera Pose Adjustment model.

The adjustment predictor has a regression head respectively predicting the variables $\Delta\theta$ and $\Delta\varphi$.

A key challenge is the sequential dependency between these tasks - the adjustment prediction is only meaningful when the suggestion predictor indicates adjustment is needed. This creates an imbalanced training scenario where only a subset of samples contribute to the adjustment task, potentially causing conflicts between tasks due to different sample spaces and gradient frequencies.

To resolve this problem, we adopt a multi-gate mixture of experts architecture, which allows each task to adaptively control parameter sharing through task-specific gates, enabling the model to learn task-specific features while maintaining shared knowledge where beneficial. Each task can dynamically assign different weights to experts, mitigating the conflicts caused by imbalanced training.

Specifically, as shown in Figure 6. Given an image, let $x \in \mathbb{R}^D$ denote the shared features extracted by the ResNet backbone. Our Camera Pose Adjustment model (CPAM) consists of M experts E_m: $\mathbb{R}^D \rightarrow \mathbb{R}^D$ and task-specific gates G_t: $\mathbb{R}^D \rightarrow \mathbb{R}^M$, where $t \in [1, 2]$ indicates different tasks. Each gate follows the softmax design:

$$G_t(x) = \text{Softmax}\left(\text{FFN}_t(x)\right) \tag{14}$$

where FFN_t represents a task-specific feed-forward network. The output feature f_t of each task branch is computed as:

$$f_t = \sum_{i=1}^{m} G_t(x)_i \cdot E_i(x) \tag{15}$$

Finally, these task-specific features f_t are processed through separate MLP layers to generate the suggestion prediction $\widehat{\boldsymbol{y}}_s^i$ and the adjustment prediction $\widehat{\boldsymbol{y}}_a^i$

Optimization. The CPAM uses a multi-task loss function. For the suggestion prediction task, we adopt the cross-entropy loss function $\mathcal{L}_{ce}$:

$$\mathcal{L}_{suggest} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{ce}(\widehat{\boldsymbol{y}}_s^i, \boldsymbol{y}_s^i) \tag{16}$$

For the adjustment prediction task, the loss function is:

$$\mathcal{L}_{cs} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\widehat{\boldsymbol{y}}_a^i \cdot \boldsymbol{y}_a^i}{\|\widehat{\boldsymbol{y}}_a^i\|\|\boldsymbol{y}_a^i\|} \tag{17}$$

$$\mathcal{L}_{norm} = \frac{1}{N}\sum_{i=1}^{N}(\|\widehat{\boldsymbol{y}}_a^i\| - \|\boldsymbol{y}_a^i\|)^2 \tag{18}$$

$$\mathcal{L}_{adjust} = \mathcal{L}_{cs} + \mathcal{L}_{norm} \tag{19}$$

The total loss function can be summarized as

$$\mathcal{L}_{CPAM} = \mathcal{L}_{suggest} + \mathbf{1}_{(\boldsymbol{y}_s=\mathbf{1})}\mathcal{L}_{adjust} \tag{20}$$

where $\mathbf{1}_{(\boldsymbol{y}_s=\mathbf{1})}$ is an indicator function that determines during training, the gradients of the adjustment predictor are backpropagated only for samples where a suggestion should be provided.

7. Experiments

7.1. Implementation Details

Training. We use CLIP (RN50) [21] as the backbone of CCQA, with RoIAlign size of 14×14. The CCQA model trained for 120 epochs using Adam optimizer [10] with learning rate 5×10^{-6}. For CPAM, we adopt ImageNet pretrained ResNet50 [6] and train it for 50 epochs using Adam with learning rate 1×10^{-4} and weight decay 1×10^{-4}.

Datasets. We train CCQA on GAICv2 [32] (3,636 images, 86 views per image) and evaluate its generalization on CPC [26] (10,800 images, 24 views per image). Our PCARD dataset is divided into training and test sets with an 8:2 ratio for CPAM training and evaluation.

Evaluation Metrics. We use the AUC (Area under receiver operating characteristics curve) to evaluate the performance of the suggestion predictor. This metric measures how accurately a model triggers suggestions. Then, we evaluate the accuracy of the adjustment predictor using cosine similarity (CS) and MAE, where cosine similarity (CS) measures how close the predicted adjustment direction is to the actual adjustment direction, and MAE measures the precision of the adjustment predictor. We adopt Intersection over Union (IoU) to quantify the accuracy of view adjustment predictions. Notably, the IoU is computed on the spherical panorama surface. More details can be seen in Supplementary Material.

7.2. Objective Evaluation

Exploration of different expert numbers in CPAM. To investigate the optimal number of experts in our Camera Pose Adjustment model, we conducted ablation studies by varying the number of experts M from 1 to 5. As shown in Table 2, we can observe that: **(a)** the model achieves the best overall performance when $M = 2$. The suggestion predictor demonstrates the highest AUC of 79.3%, and the adjustment predictor shows superior performance across all metrics; **(b)** increasing the number of experts beyond two leads to a gradual decline in performance across all metrics. This degradation might be attributed to the increased model complexity in expert predictions; **(c)** despite having similar suggestion prediction performance (AUC scores of 78.4% and 78.7% respectively), the $M = 3$ configuration demonstrates superior adjustment prediction capability compared to the $M = 1$. This is because when $M = 1$, the gating network becomes ineffective, so the CPAM lacks the dynamic expert weighting mechanism that is crucial for the mixture of experts; **(d)** we also report the IoU metrics for both true positives (TP) and all predicted adjustment cases (TP+FP) from the suggestion predictor. Notably, when $M \geq 2$, our adjustment predictor can still generate reasonable adjustments even when the suggestion predictor makes mistakes.

M	AUC↑	TP			TP+FP
		CS↑	MAE↓	IoU ↑	IoU ↑
1	<u>78.7</u>	0.401	0.524	0.604	0.601
2	**79.3**	**0.415**	**0.507**	**0.613**	**0.617**
3	78.4	<u>0.408</u>	<u>0.515</u>	<u>0.606</u>	<u>0.612</u>
4	77.3	0.398	0.52	0.597	0.601
5	76.7	0.368	0.541	0.591	0.597

Table 2. Ablation study of Camera Pose Adjustment model. (TP: True Positive, FP: False Positive).

Loss	AUC↑	TP			TP+FP
		CS↑	MAE↓	IoU ↑	IoU ↑
A	76.1	0.391	0.507	0.602	0.605
B	78.5	0.408	0.507	0.606	0.611
C	**79.3**	**0.415**	**0.507**	**0.613**	**0.617**

Table 3. Ablation Study of Loss Functions in Camera Pose Adjustment Model. (TP: True Positive, FP: False Positive)

Exploration of different loss functions in CPAM. To further validate the rationality of the loss functions (Eq. (19)) for the adjustment predictor, we compared the results of two other loss functions on the PCARD dataset, as shown in Table 3. A represents replacing Eq. (19) with MSE loss, treating the camera pose adjustment prediction as a standard regression problem. B denoted replacing Eq. (18) with MSE loss, which treats the camera pose adjustment prediction as a regression of direction and coordinates, using cosine similarity to supervise the alignment of predicted and labeled directions, and MSE for the coordinate regression. C is the loss functions adopted in our paper, treating camera pose adjustment prediction as a strict spatial vector prediction, which involves supervision of both the vector direction and the vector magnitude. The results demonstrate the effectiveness of our designed loss functions.

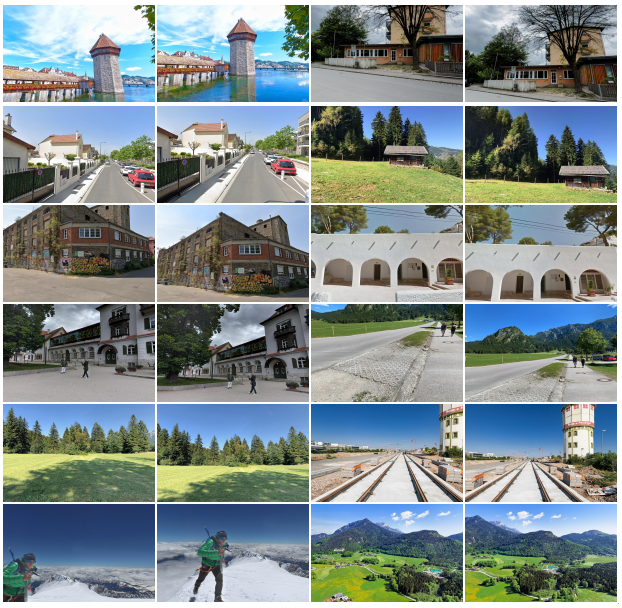


Figure 7. Qualitative examples. Each pair shows the original image (left) and the result of the adjustment (right).

Generalization Capability Validation of CCQA.To evaluate the generalization capability of our proposed CCQA model, and demonstrate the reliability of the scoring order in our PCARD dataset, we conducted rigorous experiments on additional unseen datasets. Specifically, we trained the model on the GAICv2 dataset[32] and then tested it directly on the unseen CPC dataset[26]. The averaged top-k accuracy ($\overline{Acc_k}$) and weighted average top-k accuracy($\overline{Acc_k^w}$) for both k=5 and k=10 as evaluation metrics are reported in Table 4. There are no discarded regions in the images of our PCARD dataset. Therefore, the networks were appropriately modified to adapt to our dataset, $*$ indicates that we removed RODAlign from these networks designed for image cropping tasks. The best generalization capability results are marked in bold and the second generalization capability results are marked with underlines. We can see that our proposed CCQA achieves the best performance on all metrics, the results demonstrate that the CCQA model exhibits good generalization capability, suggesting that utilizing this model to provide aesthetic scoring for our PCARD dataset is a reliable approach.

The ablation studies of CCQA and analysis experiments on PACRD dataset labels will be discussed in the Supplementary Material.

7.3. Subjective Evaluation

To further demonstrate the effectiveness of our proposed framework, we design an annotation toolbox and conduct two sets of user studies. First, we select 100 images from our dataset and show the raters the image both before and after applying the suggested camera adjustment strategy to evaluate whether the suggested camera adjustment strategy effectively improves the composition of the original image. Second, we select another 50 image pairs from our dataset and show the raters the original image and candidate image to evaluate the accuracy of whether a camera pose adjustment should be suggested. To make the comparison fair, we invited 25 students to participate in the user study. The subjects are asked which image has the better composition or if they cannot tell. The order of the two images is chosen randomly to avoid bias. The results are in Table 5 and the qualitative examples are shown in Figure 7. When a camera pose adjustment suggestion is provided, our framework effectively improves the composition of images in most cases (64.0%), with erroneous adjustment suggestions accounting for about 27.0%. When no suggestion is needed, our model has a high success rate (82.0%), and it only wrongly judges the need for a suggestion about 14.0% of the time. More qualitative results can be provided in Supplementary Material.

Method	$\overline{Acc_5}$	$\overline{Acc_{10}}$	$\overline{Acc_5^w}$	$\overline{Acc_{10}^w}$
TransView*[20]	<u>51.1</u>	66.4	36.7	50.7
GAICv2*[32]	50.9	<u>66.5</u>	36.6	<u>50.8</u>
SFRC* [25]	51	65.9	<u>36.8</u>	50.6
CCQA(Ours)	**56.1**	**72.6**	**39.8**	**55.5**

Table 4. Comparison of the generalization ability of different composition scoring models on the CPC dataset.

Which is better	Suggestion	Adjustment
After/Original	82.0%	64.0%
Before/Candidate	14.0%	27.0%
No difference	4.0%	9.0%

Table 5. Subjective evaluation results on our dataset PCARD.

8. Concluding remarks

We have presented a new smart point and shoot (SPAS) solution to help smartphone users to take better photographs. We have made several contributions in this paper including a large dataset with $320K$ images from 4000 scenes where each image containing camera pose information. We have also developed an image quality labeler that can discern subtle image quality difference as well as a camera pose adjustment model that using a mixture of experts solution to accomplish two sequential tasks of guiding a user to compose a good shot of a scene.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62271323 and U22B2035, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012956 and 2023B1212060076, and in part by the Shenzhen Research and Development Program under Grant JCYJ20220531102408020 and KJZD20230923114209019.

References

[1] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. Adaptive fractional dilated convolution network for image aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14114–14123, 2020. 3

[2] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 226–234. IEEE, 2017. 2, 3

[3] Casper L Christensen and Aneesh Vartakavi. An experience-based direct generation approach to automatic image cropping. *IEEE Access*, 9:107600–107610, 2021. 2

[4] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1105–1108, 2014. 2, 3

[5] Jaouad Hajjami, Jordan Caracotte, Guillaume Caron, and Thibault Napoleon. Arucomni: detection of highly reliable fiducial markers in panoramic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 634–635, 2020. 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 7

[7] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7057–7066, 2021. 3

[8] James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2175–2183, 2024.

[9] Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2455, 2022. 3

[10] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[11] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8193–8201, 2018. 3

[12] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *IEEE Transactions on Image Processing*, 28(10):5105–5120, 2019. 3

[13] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4213–4222, 2020. 3

[14] Dong Liu, Rohit Puri, Nagendra Kamath, and Subhabrata Bhattacharya. Composition-aware image aesthetics assessment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3569–3578, 2020. 3

[15] Xiaoyu Liu, Ming Liu, Junyi Li, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Beyond image borders: Learning feature extrapolation for unbounded image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13023–13032, 2023. 3

[16] Peng Lu, Jiahui Liu, Xujun Peng, and Xiaojie Wang. Weakly supervised real-time image cropping based on aesthetic distributions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 120–128, 2020. 3

[17] Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4535–4544, 2017. 3

[18] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 497–506, 2016. 3

[19] Shijia Ni, Feng Shao, Xiongli Chai, Hangwei Chen, and Yo-Sung Ho. Composition-guided neural network for image cropping aesthetic assessment. *IEEE Transactions on Multimedia*, 25:6836–6851, 2022. 3

[20] Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong. Transview: Inside, outside, and across the cropping view boundaries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4218–4227, 2021. 3, 8

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[22] Pratibha Sharma, Manoj Diwakar, and Niranjan Lal. Edge detection using moore neighborhood. *International Journal of Computer Applications*, 61(3), 2013. 4

[23] Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-semantic collaborative cropping for user generated content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4988–4997, 2024. 2, 3

[24] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12104–12111, 2020.

[25] Chao Wang, Li Niu, Bo Zhang, and Liqing Zhang. Image cropping with spatial-aware feature and rank consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10052–10061, 2023. 8

[26] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5437–5446, 2018. 2, 3, 7, 8

[27] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 5

[28] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–978, 2013. 2, 3

[29] Guo-Ye Yang, Wen-Yang Zhou, Yun Cai, Song-Hai Zhang, and Fang-Lue Zhang. Focusing on your subject: Deep subject-aware image composition recommendation networks. *Computational Visual Media*, 9(1):87–107, 2023. 2, 3

[30] Quan Yuan, Leida Li, and Pengfei Chen. Aesthetic image cropping meets vlp: Enhancing good while reducing bad. *Journal of Visual Communication and Image Representation*, page 104316, 2024. 3

[31] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5949–5957, 2019. 2, 3

[32] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1304–1319, 2020. 2, 3, 5, 7, 8

[33] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. *arXiv preprint arXiv:2104.03133*, 2021. 3

[34] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 5

[35] Pengyu Zhao, Ansheng You, Yuanxing Zhang, Jiaying Liu, Kaigui Bian, and Yunhai Tong. Spherical criteria for fast and accurate 360 object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12959–12966, 2020. 4

[36] Lei Zhong, Feng-Heng Li, Hao-Zhi Huang, Yong Zhang, Shao-Ping Lu, and Jue Wang. Aesthetic-guided outward image cropping. *ACM Transactions on Graphics (TOG)*, 40(6): 1–13, 2021. 3

[37] Zizhuang Zou, Mao Ye, Xue Li, Luping Ji, and Ce Zhu. Stable viewport-based unsupervised compressed 360° video quality enhancement. *IEEE Transactions on Broadcasting*, 2024. 4