Towards Smart Point-and-Shoot Photography

Supplementary Material

In this document, we provide more details as supplementary materials to our main submission. We first present the related work. We present the mathematical details of view generation from equirectangular panoramas to perspective views in our proposed Panorama-based Composition Adjustment Recommendation dataset (PCARD) in Sec. 2, followed by the definition and calculation of spherical overlap and spherical IoU metrics in Sec. 3. In Sec. 4, we provide comprehensive statistical information of our proposed PCARD, including its taxonomic structure and detailed label analysis. Sec. 5 presents extensive ablation studies on our CLIP-based Composition Quality Assessment (CCQA) model. Finally, in Sec. 6, we describe our subjective evaluation setup and provide additional qualitative results demonstrating the effectiveness of our approach across diverse scenarios.

1. Related Work

The image cropping (ICDB) dataset [9] and the human crop (HCDB) dataset [3] contain a small number of images that were manually annotated with the best cropping boxes by multiple professional photographers. Since the best cropping generated in this manner relies entirely on the annotators' experience without explicit constraints, Christensen and Vartakavi [2] constructed an aspect ratio-aware image cropping (GNMC) dataset, where each image includes optimal cropping annotations in different aspect ratios (16:9, 3:4, 4:3, 2:2, 1:1).The subject-aware composition (SACD) dataset [10] contains 2777 images and more than 24,000 candidate views, where each image is annotated with 8 optimal cropping boxes. However, the limited number of annotated crops is not conducive to the training of a robust image composition model. Therefore, some other datasets [1, 5, 7, 10–12] were created with dense annotations. The primary paradigm of creating these datasets is first generating a large number of candidate views, followed by experts annotating them using a pair-wise ranking strategy [1, 7] or a direct scoring approach [5, 11, 12]. The flicker-cropping (FCDB) dataset [1] contains 1743 images and 31430 annotated pairs of candidate views while the comparative photo composition (CPC) dataset [7] contains 10800 images, with 24 candidate views for each image and generates more than 1 million view pairs. The labeled cropping windows all have high aesthetic value with a certain focused subject. Different from the pair-wise strategy, the grid-anchor-based image cropping (GAICv1, GAICv2) dataset [11, 12] provides an average of 86 fixed candidate views for each image, where each candidate view is assigned an aesthetic quality score. The user-generated content crop (UGCrop5K) dataset [5] consists of 45000 exhaustively annotated candidate views on 5K images.

Dataset	Year	Label	Scenes	Candidate Views	Camera Pose
ICDB[9]	2013	Best	950	1	N/A
HCDB[3]	2014	Best	500	1	N/A
GNMC[2]	2022	Best	10000	5	N/A
SACD[10]	2023	Best	2777	8	N/A
FCDB[1]	2017	Rank	1536	18	N/A
CPC[7]	2018	Rank	10800	24	N/A
GAICv1[11]	2019	Score	1236	86	N/A
GAICv2[12]	2020	Score	3336	86	N/A
UGCrop5K[5]	2024	Score	5000	90	N/A

Table 1. Image Composition datasets.

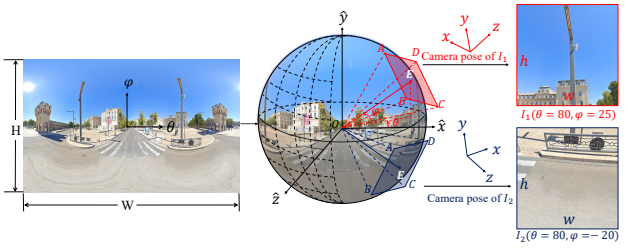


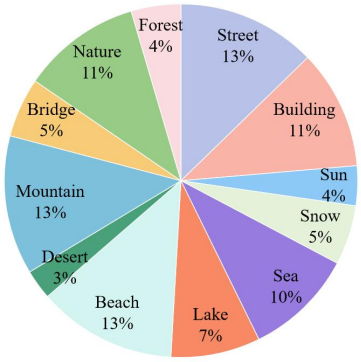
Figure 1. View generation. An equirectangular (ERP) image can be mapped to a perspective view using camera field of view (fov_x, fov_y) and viewing direction $(\theta_{0,0}, \varphi_{0,0})$ parameters.

2. View Generation

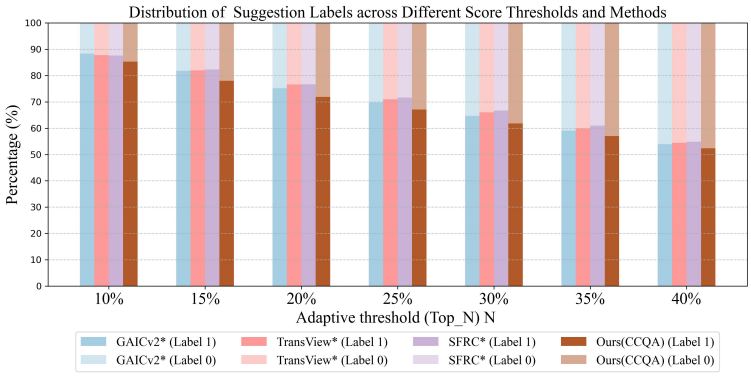
As shown in Figure 1, we can generate a perspective view I from an Equirectangular(ERP) image given two key parameters: (1) the camera field of view (fov_x, fov_y), and (2) the viewing direction $(\theta_{0,0}, \varphi_{0,0})$ that defines the viewpoint E in the spherical domain S^2. The generated view I has a spatial resolution of $h \times w$ in the 2D plane. Given a pixel located at (i, j) $(i \in [1, w], j \in [1, h])$ in the view I, we transform it into a 3D point $(x_{i,j}, y_{i,j}, z_{i,j})$ in the camera coordinate system through inverse perspective projection:

$$\begin{bmatrix} x_{i,j} \\ y_{i,j} \\ z_{i,j} \end{bmatrix} = \begin{bmatrix} f_x & 0 & i_0 \\ 0 & f_y & j_0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} \tag{1}$$

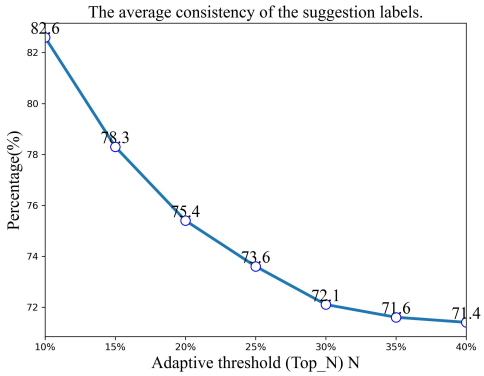
$$\begin{cases} f_x = \frac{w}{2\tan\left(\frac{fov_x}{2}\right)} \\ f_y = \frac{h}{2\tan\left(\frac{fov_y}{2}\right)} \end{cases} \tag{2}$$



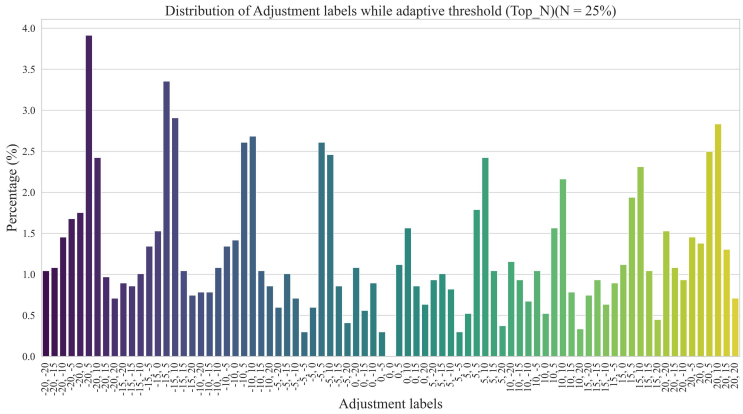
(a) Taxonomic structure.



(b) Distribution of Suggestion Labels.



(c) The average consistency of the suggestion labels.



(d) Distribution of Adjustment Labels based on CCQA.

Figure 2. Statistics of the proposed PCARD dataset.

where f_x and f_y are the horizontal and the vertical focal lengths. Following [13], we standardize all views with a vertical FOV of $fov_y = 60°$ and a fixed spatial resolution of $h \times w = 768 \times 1024$. The principal point (i_0, j_0) represents the pixel coordinates of the center point on the view I, where $i_0 = (w-1)/2, j_0 = (h-1)/2$.

To align the camera coordinate system (x, y, z) with the world coordinate system $(\hat{x}, \hat{y}, \hat{z})$, we apply two successive rotations:

$$\begin{bmatrix} \hat{x}_{i,j} \\ \hat{y}_{i,j} \\ \hat{z}_{i,j} \end{bmatrix} = \mathcal{R}_y\left(\theta_{0,0}\right) \mathcal{R}_x\left(\varphi_{0,0}\right) \begin{bmatrix} x_{i,j} \\ y_{i,j} \\ z_{i,j} \end{bmatrix} \tag{3}$$

where $\mathcal{R}_y(\theta_{0,0})$ represents the rotation matrix of angle $\theta_{0,0}$ along y-axis, $\mathcal{R}_x(\varphi_{0,0})$ represents the rotation matrix of angle $\varphi_{0,0}$ along x-axis. The rotated coordinates $(\hat{x}_{i,j}, \hat{y}_{i,j}, \hat{z}_{i,j})$ are then converted to spherical coordinates (longitude $\theta_{i,j}$ and latitude $\varphi_{i,j}$):

$$\begin{cases} \theta_{i,j} = \arctan\left(\frac{\hat{x}_{i,j}}{\hat{z}_{i,j}}\right) \\ \varphi_{i.j} = \arcsin\left(\hat{y}_{i,j}\right) \end{cases} \tag{4}$$

Finally, we map these spherical coordinates to pixel coordinates $(u_{i,j}, v_{i,j})$ in the ERP image domain with width W and height H:

$$\begin{cases} u_{i,j} = \left(\frac{\theta_{i,j}}{2\pi} + \frac{1}{2}\right) W \\ v_{i,j} = \left(-\frac{\varphi_{i,j}}{\pi} + \frac{1}{2}\right) H \end{cases} \tag{5}$$

Through the series of coordinate transformations, we establish a complete mapping from the source ERP image to the target perspective view, enabling accurate view generation from any given viewpoint.

3. Spherical overlap and Spherical IoU

Given a spherical rectangle $S_i(\theta_i, \varphi_i, \alpha_i, \beta_i)$, the area of the shape is $A(\cdot)$:

$$A\left(S_i\right) = 4\arccos\left(-\sin\frac{\alpha_i}{2}\sin\frac{\beta_i}{2}\right) - 2\pi, \text{ for } i \in \{1, 2\} \tag{6}$$

where θ_i and φ_i denote the polar angle, α_i and β_i represent the horizontal and vertical field of view. The overlapping region between two spherical rectangles is most likely not a standard spherical rectangle but rather an irregular spherical polygon, making the calculation of Area $A(S_i \cap S_j)$ quite complex. However, we can utilize the fact that the boundaries of the two spherical rectangles are great circle arcs and can be used to calculate the area of the overlapping region [8]:

$$A\left(S_i \cap S_j\right) = \sum_{i=1}^{n} \omega_i - (n-2)\pi \tag{7}$$

where n is the number of sides of the spherical polygon defined by the intersection region, ω_i is the angle of the spherical polygon, which is the angle between the planes on the adjacent boundaries.

Therefore,

$$\text{SphOverlap}\ \left(S_{\text{adj}}, S_{\text{init}}\right) = \frac{A(S_{\text{adj}} \cap S_{\text{init}})}{A(S_{\text{init}})} \tag{8}$$

$$\text{SphIoU}\ \left(S_{\text{adj}}, S_{\text{init}}\right) = \frac{A(S_{\text{adj}} \cap S_{\text{init}})}{A(S_{\text{adj}}) + A(S_{\text{adj}}) - A(S_{\text{adj}} \cap S_{\text{init}})} \tag{9}$$

where S_{adj} and S_{init} represent the spherical rectangles corresponding to $\boldsymbol{I}^i_{adj}$ and $\boldsymbol{I}^i_{init}$ in the 360° images respectively.

4. Statistics of the PACRD

Taxonomic structure. To better explore the aesthetic diversity of the PCARD, we manually divided it into 12 categories, namely Street, Building, Sun, Snow, Sea, Lake, Beach, Desert, Mountain, Bridge, Nature, and Forest, as shown in Figure 2 (a). It is worth noting that, unlike common image composition datasets with single semantic information, these categories are not mutually exclusive, as individual images may contain multiple semantic elements. This is because the images in PCARD have richer semantic information, with overlaps between categories.

Label analysis. First, to evaluate the reliability and consistency of our pseudo-labeling method, we conducted the comparative analysis using four composition scoring models: our proposed CCQA model and three representative models (GAICv2* [11], TransView* [4], and SFRC* [6])[1] following the main paper. These models were selected for their diverse technical approaches, ranging from RoIAlign and RoDAlign feature fusion, visual elements dependencies modeling to spatial-aware feature and transductive learning. We then analyzed the distribution of the suggested labels $\boldsymbol{y}^i_s$ generated by these four composition scoring models under the pseudo-labeling method based on different score thresholds. The visualization results are shown in Figure 2 (b). It can be observed that if the score thresholds are set consistently, the distributions of the suggested labels generated based on different scoring models exhibit similar patterns, indicating the stability of our pseudo-labeling approach. Furthermore, we calculated the consistency of suggestion labels generated by CCQA and three other models, as shown in Figure 2 (c). The average consistency of suggestion labels reaches over 70%. This high level of agreement among different models demonstrates both the robustness of our proposed pseudo-label generation method and the reliability of labels generated based on the CCQA model. In particular, considering both the balanced distribution and average consistency of suggestion labels, we chose $N = 25\%$ in practice. Figure 2 (d) show the distribution of adjustment labels $\boldsymbol{y}^i_a$ based on CCQA while $N = 25\%$. If the composition can be improved, the distribution of adjustment labels aligns with the eight-neighborhood candidate adjustment space defined by the dataset with a step size of $\Delta\theta = \Delta\varphi = 5°$, indicating that our proposed pseudo-label generation method and CCQA model can be reasonably applied to generate adjustment labels without missing potential adjustment spaces.

[1] *Exclude the RoDAlign branch for a fair comparison.

No.	WS	FMR	LP	$\overline{Acc_5}$	$\overline{Acc_{10}}$	$\overline{Acc^w_5}$	$\overline{Acc^w_{10}}$
1	✓			49.4	65.5	34.7	49.2
2		✓		48.3	64.7	34	43
3	✓	✓		49.4	65.8	35	49.4
4	✓		✓	51.5	68	36.4	51.5
5		✓	✓	50.4	66.7	35.7	50.3
6	✓	✓	✓	**56.1**	**72.6**	**39.8**	**55.5**

Table 2. Ablation study of different components in CCQA. "LP", "FMR", and "WS" are short for Learnable prompt, Feature mixers and regression, and Weighted summation respectively.

No.	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\overline{Acc_5}$	$\overline{Acc_{10}}$	$\overline{Acc^w_5}$	$\overline{Acc^w_{10}}$
1	✓			34.8	56.6	23.1	39.1
2		✓		45.5	63.8	30.1	46.1
3	✓		✓	41.1	60.2	27.5	43.1
4		✓	✓	50.1	68.5	33.7	50.7
5	✓	✓		48.8	65.4	34.3	50.7
6	✓	✓	✓	**56.1**	**72.6**	**39.8**	**55.5**

Table 3. Ablation study of different loss functions in CCQA.

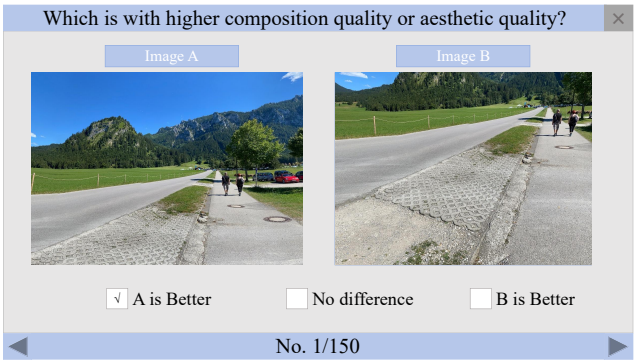


Figure 3. Illustration of the annotation toolbox.

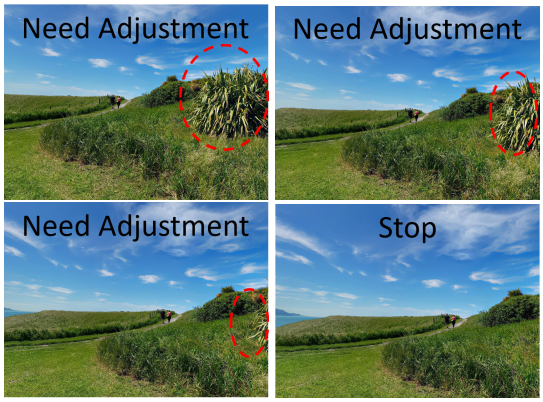


Figure 4. Examples of composition improvement trajectory: 3 steps.

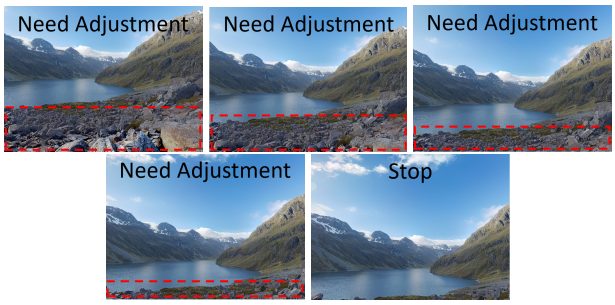


Figure 5. Examples of composition improvement trajectory: 4 steps.

5. Ablation study of the CCQA

To comprehensively evaluate our proposed CLIP-based Composition Quality Assessment model (CCQA) and demonstrate the reliability of the scoring order in our PCARD dataset, we conduct extensive ablation studies to analyze the contribution of each component and loss function. The experiments are performed by training on the GAICv2 dataset [12] and testing generalization on the unseen CPC dataset [7].

Model architecture.We first investigate the impact of different architectural components in CCQA. As shown in Table 2, we systematically evaluate three key components: Learnable prompt (LP), Feature mixers and regression (FMR), and Weighted summation (WS). The baseline model with only FMR achieves $\overline{Acc_5}$ of 48.3%, $\overline{Acc_{10}}$ of 64.7%, $\overline{Acc_5^w}$ of 34% and $\overline{Acc_{10}^w}$ of 43%. Adding WS (No.1) or combining it with FMR (No.3) shows incremental improvements. The introduction of Learnable prompt (LP) significantly enhances performance, particularly when combined with other components. The full model incorporating all three components (No.6) achieves the best performance across all metrics, with $\overline{Acc_5}$ of 56.1%, $\overline{Acc_{10}}$ of 72.6%, $\overline{Acc_5^w}$ of 39.8% and $\overline{Acc_{10}^w}$ of 55.5% on unseen data.

Loss function. We further examine the effectiveness of our loss function design, particularly focusing on the intermediate feature constraint $\mathcal{L}_3$. As shown in Table 3, we compare our complete CCQA model with a variant without the $\mathcal{L}_3$ loss term. The baseline CCQA without $\mathcal{L}_3$ achieves $\overline{Acc_5}$ of 48.8%, $\overline{Acc_{10}}$ of 65.4%, $\overline{Acc_5^w}$ of 34.3% and $\overline{Acc_{10}^w}$ of 49% on unseen data. By incorporating $\mathcal{L}_3$, our complete model shows substantial improvements across all metrics. These results demonstrate that constraining intermediate features through $\mathcal{L}_3$ is crucial for improving the generalization capability of CCQA.

6. Subjective Evaluation

6.1. Annotation box

The interface of our annotation toolbox for user studies is shown in Figure 3. The annotation toolbox was specifically designed to facilitate efficient and unbiased comparison of image compositions. The tool presents two images side-by-side for direct comparison, ensuring consistent evaluation conditions across all participants.

6.2. More qualitative results

This section shows more qualitative results.

The sequence of operation of SPAS is as follows: (i) Decide if the current composition can be improved by performing the suggestion prediction. (ii) If the suggestion predictor outputs 0, then the user will take a photo and the process is complete. (iii) If the suggestion predictor outputs 1, the system predicts the adjustment angles and the user follows the instruction to adjust the camera pose. (iv) Go to (i) and the process continues.Depending on the scene and initial view, it takes different number of steps, typically 3 to 6 in our data. Figure 4 and Figure 5 show examples of typical trajectories of improvement.

To comprehensively evaluate our CPAM model's per-

Figure 6. Examples of well-composed images requiring no adjustment.

formance, we present two sets of qualitative results that demonstrate its intelligent decision-making capabilities across different scenarios. As shown in Figure 6, we show cases where the source compositions are already well-crafted. These images, spanning various scenes including coastal landscapes, mountain views, and rural paths. Figure 7 presents paired examples of images requiring compositional improvements, along with CPAM's adjustment results. Each pair consists of the source image and our model's optimized view. The adjustments demonstrate CPAM's effectiveness in various challenging scenarios: reframing landscapes to better emphasize focal points, optimizing horizon placement in outdoor scenes, and improving the balance of architectural elements. Notably, the adjustments are subtle yet meaningful, showing CPAM's ability to make refined modifications while preserving the essential character of each scene.These results collectively highlight CPAM's dual capability: maintaining already-optimal compositions while making appropriate adjustments when needed. This discriminative behavior is crucial for practical applications, where Smart Point-and-Shoot (SPAS) systems must be both effective and judicious in their interventions. The diverse range of scenarios in both figures also demonstrates the model's robust generalization across different photographic contexts and composition challenges.

References

[1] 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. 1

[2] Casper L Christensen and Aneesh Vartakavi. An experience-based direct generation approach to automatic image cropping. *IEEE Access*, 9:107600–107610, 2021. 1

[3] 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. 1

[4] 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

[5] 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. 1

[6] 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. 3

[7] 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. 1, 4

[8] Hang Xu, Qiang Zhao, Yike Ma, Xiaodong Li, Peng Yuan, Bailan Feng, Chenggang Yan, and Feng Dai. Pandora: A panoramic detection dataset for object with orientation. In *European conference on computer vision*, pages 237–252. Springer, 2022. 3

[9] 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. 1

[10] 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. 1

[11] 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. 1, 3

[12] 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. 1, 4

[13] Yucheng Zhu, Guangtao Zhai, Yiwei Yang, Huiyu Duan, Xiongkuo Min, and Xiaokang Yang. Viewing behavior supported visual saliency predictor for 360 degree videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4188–4201, 2021. 2

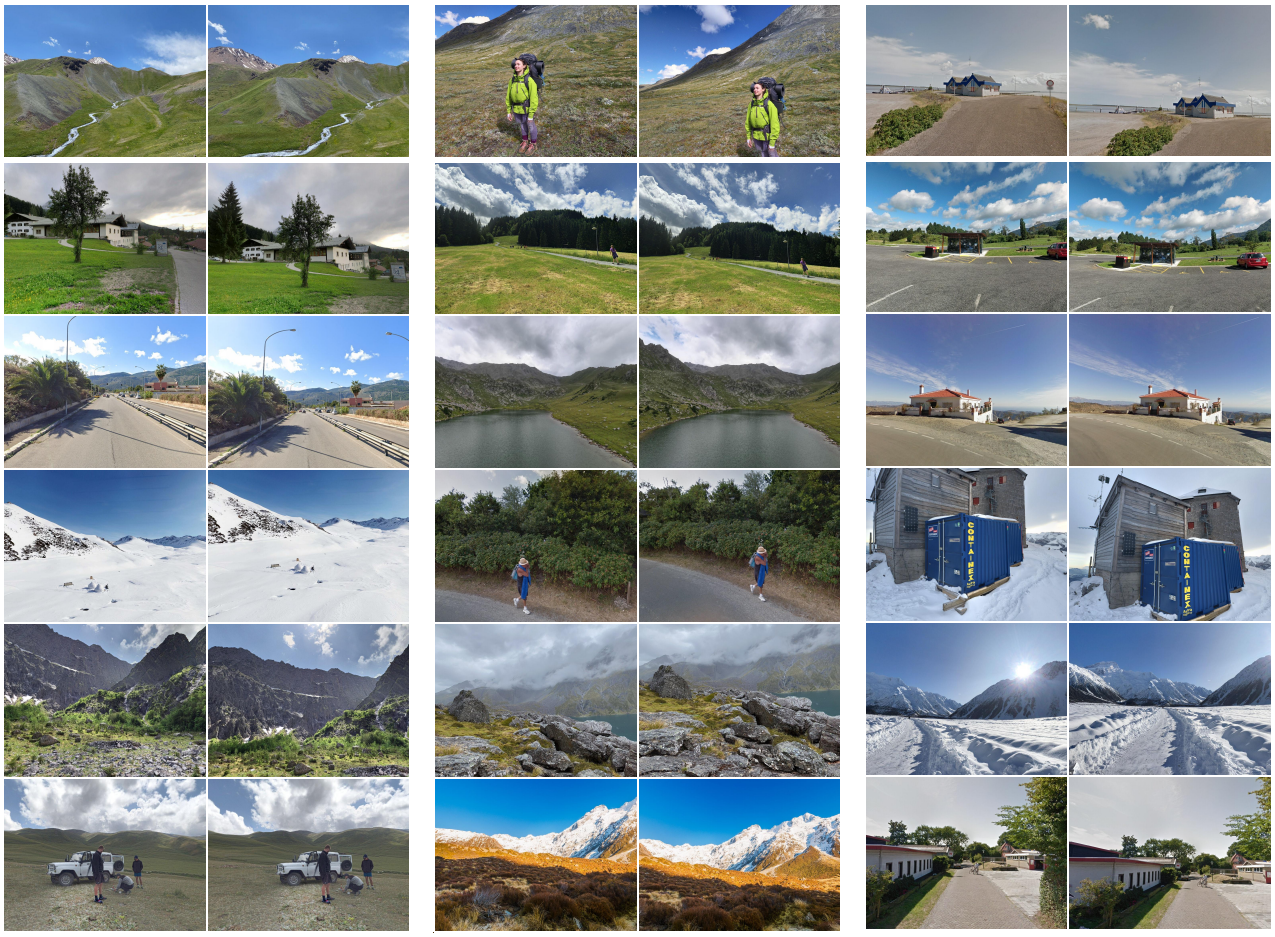


Figure 7. Demonstration of CPAM's adjustment capabilities: Before-and-after comparison on images requiring composition optimization.