3D View Optimization for Improving Image Aesthetics

Taichi Uchida
taisws5863@gmail.com

Yoshihiro Kanamori
kanamori@cs.tsukuba.ac.jp

Yuki Endo
endo@cs.tsukuba.ac.jp

University of Tsukuba

Abstract

Achieving aesthetically pleasing photography necessitates attention to multiple factors, including composition and capture conditions, which pose challenges to novices. Prior research has explored the enhancement of photo aesthetics post-capture through 2D manipulation techniques; however, these approaches offer limited search space for aesthetics. We introduce a pioneering method that employs 3D operations to simulate the conditions at the moment of capture retrospectively. Our approach extrapolates the input image and then reconstructs the 3D scene from the extrapolated image, followed by an optimization to identify camera parameters and image aspect ratios that yield the best 3D view with enhanced aesthetics. Comparative qualitative and quantitative assessments reveal that our method surpasses traditional 2D editing techniques with superior aesthetics.

1. Introduction

For visually appealing photography, it is imperative to deliberate on multiple elements, including composition and the conditions under which the photograph is captured. Post-capture, the aesthetic appeal of photographs can be augmented through the application of photo editing software, such as Adobe Photoshop. While professional photographers possess the acumen to factor in elements that enhance the aesthetic value of images, this remains a challenging endeavor for amateur photographers. In light of this, several methodologies have been introduced to automate the process of image composition editing [6, 10, 13, 14, 21], with a focus on improving aesthetic quality. Leveraging these advancements, individuals lacking specialized photography knowledge or skills can readily enhance the visual appeal of their images. However, current techniques mainly rely on 2D processing methods like cropping and stretching, which offer limited flexibility.

In this paper, we introduce an unprecedented approach that utilizes 3D manipulations to enhance image aesthetics further. We first outpaint the input image to widen the search space and reconstruct the 3D scene from the extrapolated image. We then search for parameters, such as camera translation, rotation, and field-of-view angle, for re-rendering the 3D scene to broaden the possibilities for aesthetic improvement. Image aesthetics is evaluated using a pre-trained network of aesthetics evaluation model [21]. During the parameter search, we find that a typical gradient-based optimization approach [12] easily falls into local minima. We thus introduce a black-box optimization method in evolutionary computation called *Covariance Matrix Adaptation Evolution Strategy* (CMA-ES) [8] to search for optimal parameters globally.

Our contributions are summarized as follows:
1. A pioneering insight to explore camera parameters and image aspect ratios to improve the aesthetics of the input image, and
2. A simple optimization procedure to search for optimal parameters globally.

Our approach provides a greater degree of freedom in enhancing aesthetics compared to traditional 2D methods. Through both qualitative and quantitative evaluations, we demonstrate the effectiveness of our method in elevating the aesthetic quality of images, showcasing its superiority in visually appealing photography.

2. Related Work

2.1. Aesthetics-guided Composition Editing

Contemporary studies on automatic composition editing for aesthetic enhancement predominantly focus on cropping techniques that trim input images. Such conventional cropping methods are constrained to adjusting the inner regions of an image for editing, particularly when the primary subject is close to the image's boundary, thus limiting the scope for aesthetic improvements. Zhong *et al.* [25] overcame this limitation by integrating a pre-cropping extrapolation of the area outside the original frame, thereby expanding the search space for aesthetic optimization.

Another method of changing the composition of an image is through *image retargeting* [3, 19], which resizes an

image without distortion while preserving the shape of the main subject. Liu *et al.* [15] proposed a method that automatically performs retargeting while moving the salient subjects in the image to more aesthetically pleasing positions.

While existing approaches to enhancing image aesthetics predominantly employ 2D image processing techniques, our method ventures into an unexplored domain to improve the aesthetics of input images through 3D scene exploration; we optimize camera position, gazing direction, field-of-view angle, and image aspect ratios, all the while preserving the 3D structure of the captured scene.

2.2. Composition Editing with Depth Information

Traditional photo-editing tools struggle to incorporate 3D scene information, often failing to preserve spatial relationships between subjects during composition editing. CompZoom [4] addresses this by allowing manual adjustments to subject size and position through changes in focal length and camera position, utilizing a multi-view camera model and requiring multiple images from different focal lengths and positions of the same scene. Conversely, Liu *et al.* [16] introduced a composition-editing tool that leverages depth information from a single image, enabling adjustments in size, position, and perspective of objects while maintaining the scene's 3D structure, albeit necessitating a degree of expertise in photographic composition and familiarity with the tool for effective use. Our approach simplifies this process, offering an automatic composition editing method that manipulates camera parameters within a 3D scene reconstructed from a single image, making it accessible to users without in-depth photography knowledge.

2.3. Aesthetic View Exploration in 3D Space

Prior research on aesthetic view search within 3D spaces has predominantly utilized reinforcement learning to ascertain optimal camera positions and orientations in both real and virtual 3D environments. Xie *et al.* [22] introduced a technique for generating sequences of aesthetically appealing camera postures within a pre-constructed 3D scene, employing the same aesthetics evaluation model as AlZayer *et al.* [2]. In contrast, our approach focuses on identifying aesthetic views in the 3D scene reconstructed from a single input photograph, offering a novel perspective on enhancing image aesthetics.

2.4. Aesthetic Evaluation of Images

The task of image aesthetics evaluation involves categorizing images into distinct levels of aesthetics or quantifying an image's beauty with an aesthetics score. The advent of deep learning and the availability of extensive datasets for aesthetics assessment have led to deep learning-based methods outperforming traditional, manually crafted approaches in terms of accuracy and efficiency. These methods, noted for their robustness and minimal inference costs, have been effectively integrated into view-search algorithms within 3D spaces, as discussed in Section 2.3. Our approach leverages the aesthetics evaluation model called a *View Evaluation Net* (VEN) proposed by Wei *et al.* [21], which calculates aesthetics scores for images generated from initial viewpoints and various camera positions. This model, trained on a comprehensive dataset comprising over 1 million photo-to-picture pair comparisons, excels in determining the relative aesthetics among pairs of image crops, providing a solid foundation for our method's aesthetics-based view selection.

3. Our Method

Figure 1 illustrates the overview of our method. Our method commences with an image extrapolation of the input image utilizing Adobe Photoshop's generative fill function (Section 3.1). Subsequently, a depth map is generated from the extrapolated image via a monocular depth estimation technique [18]. Direct restoration of the 3D scene from the input image and depth map would reveal regions beyond the original field of view upon altering the viewpoint. We thus feed the extrapolated image and corresponding depth map into a 3D photo generation method [20] to reconstruct the 3D scene, resulting in a point cloud that supplements the color and depth information for the unseen areas of the input image. We then search for camera parameters, *i.e.*, translation, rotation, and field-of-view angle, to identify configurations that enhance aesthetic quality. An existing aesthetic evaluation model [21] assesses the aesthetics of images rendered from new viewpoints within this 3D scene. The optimization process seeks to maximize the aesthetic scores by adjusting the camera parameters and also explores variations in the image aspect ratio to broaden the search space. Further details on each step of our method are elaborated in the subsequent sections.

3.1. Preliminary Image Extrapolation

If we naïvely reconstruct a 3D scene from the input image, a slight change in camera position or gazing direction will reveal out-of-photographed regions beyond the edges of the input image. To avoid this, we extrapolate the input image in advance. Specifically, the input image is first uniformly resized such that the larger dimension, whether height h or width w of the input image, measures 512 pixels. Subsequently, we expand the image's borders by 256 pixels on all sides (*i.e.*, top, bottom, left, and right) utilizing Adobe Photoshop's generative fill function (without text prompts). This preparatory step not only enhances the search space for aesthetic improvements but also effectively minimizes the visibility of areas beyond the original photographic range.
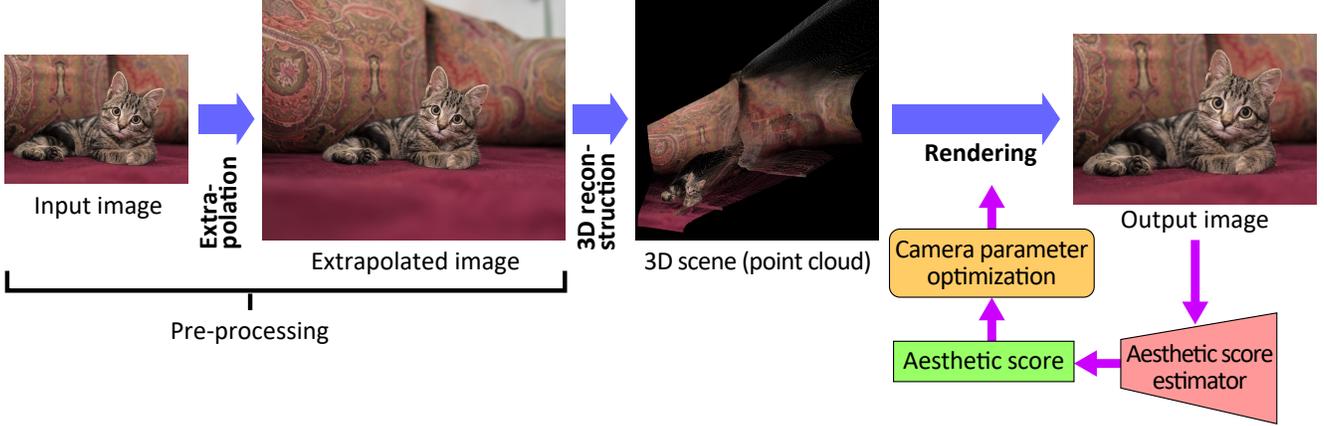
Figure 1. Overview of our method. The input image is first extrapolated in the pre-processing and then fed to a 3D reconstruction method [20] to obtain a point cloud that represents the 3D scene. Our method finds optimal camera parameters that maximize the aesthetic score [21] of the rendered image through an optimization loop.

3.2. Suppressing the Appearance of Out-of-photographed Regions

Unfortunately, the abovementioned image extrapolation is not perfect yet; a large change in camera translation or gazing direction may still reveal out-of-photographed regions.

To further suppress the appearance of such regions, our method integrates a regularization term defined with a binary mask rendered under new camera parameters. Let $\mathcal{P}$ be the colored point cloud obtained from the input image, Θ be the camera parameters, and $\mathcal{M}$ be the rendering function that outputs a binary mask. The regularization term is defined as follows:

$$\mathcal{L}_{mask} = \frac{1}{w \times h}\|\mathbf{1}_{w\times h} - \mathcal{M}(\mathcal{P}, \Theta)\|^2, \qquad (1)$$

where $\mathbf{1}_{w\times h}$ is an all-"1" image of $w \times h$ pixels.

3.3. Camera Parameter Optimization

Let $\mathcal{R}$ be a rendering function that outputs a color image from point cloud $\mathcal{P}$ under the camera parameter Θ, and $\mathcal{A}$ be an evaluation function that outputs an aesthetic score from a color image. We update the camera parameter Θ using the following equation.

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\ \mathcal{A}(\mathcal{R}(\mathcal{P}, \Theta)) - \lambda_{mask}\, \mathcal{L}_{mask}, \qquad (2)$$

where λ_{mask} is the weight of $\mathcal{L}_{mask}$ and we set $\lambda_{mask} = 10$.

Initially, we used the differentiable renderer of Py-Torch3D [11] as the rendering function $\mathcal{R}$ and the Adam optimizer [12] to optimize Equation (2). However, our preliminary experiment showed that this optimization easily falls into local minima. Therefore, we adopt a black-box optimization algorithm called CMA-ES [8] for a more global search. Note that this algorithm does not require gradient computation and thus differentiable rendering is not required.

The camera parameters to be optimized in this study are the camera translation vector $\mathbf{t} = (t_x, t_y, t_z)$, camera rotation angle $\theta = (\theta_{roll}, \theta_{pitch}, \theta_{yaw})$, and vertical field-of-view angle θ_{fovy}. Furthermore, to change the aspect ratio of the output image, we introduce coefficients s_w, s_h to control the output image size $(s_w w) \times (s_h h)$ and optimize these coefficients. A comparison with and without optimization of these coefficients is presented in the next section.

4. Experiments

4.1. Experimental settings

We implemented our method using Python, PyTorch, Py-Torch3D, and Optuna [1]. Experiments were conducted using the test images of the following three datasets: FCDB [5], FLMS [7], and GAICD [23]. These three datasets contain 309, 500, and 500 test images, respectively. Image extrapolation was performed on all of the test images in advance. The optimization took approximately 5 minutes per image on an Intel Core i7-5960X CPU.

We adopted CMA-ES as the optimization algorithm because it gave the best results among the algorithms implemented in Optuna. The number of optimization steps was set to 2,000, and optimization was terminated if the evaluated value did not decrease by 0.001 or more in the last 500 steps.

The search ranges of parameters were set as follows. The camera translation vector was set as $t_x, t_y \in [-0.1, 0.1]$ and $t_z \in [-0.5, 0.5]$, the camera rotation angles were $\theta_{roll}, \theta_{pitch}, \theta_{yaw} \in [-10°, 10°]$, the vertical field-of-view angle was $\theta_{fovy} \in [-10°, 10°]$. The coefficients s_w, s_h were initialized as 1, and the search range was set to $[0.1, 2]$.

Table 1. Ablation study of our method. We compare three variants, *i.e.*, optimized using (i) Adam [12], (ii) CMA-ES [8], and (iii) CMA-ES with image scaling (*i.e.*, image aspect ratio optimization) with three image datasets. Boldface indicates best values.

Method	FCDB	FLMS	GAICD	Total
Input (before opt.)	0.84	1.18	0.77	0.94
Adam	1.30	1.64	1.22	1.40
CMA-ES w/o scaling	3.22	3.21	2.81	3.06
CMA-ES w/ scaling	**3.55**	**3.50**	**3.11**	**3.36**

Table 2. Quantitative comparison with existing 2D cropping methods. We calculated aesthetic scores using VEN [21], SAMPNet [24], and TANet [9] as evaluation metrics.

Method	VEN scores ↑				SAMPNet scores ↑				TANet scores ↑			
	FCDB	FLMS	GAICD	Total	FCDB	FLMS	GAICD	Total	FCDB	FLMS	GAICD	Total
Input	0.84	1.18	0.77	0.94	2.87	3.14	3.31	3.14	0.49	0.50	0.54	0.51
VPN [21]	0.60	0.89	0.55	0.69	2.84	3.11	3.24	3.10	0.48	0.49	0.52	0.50
VEN [21]	0.86	1.12	0.82	0.94	2.86	3.11	3.27	3.11	0.48	0.49	0.52	0.50
CGS [14]	0.23	0.46	0.11	0.27	2.80	3.10	3.24	3.08	0.48	0.48	0.52	0.49
CAC [10]	0.30	0.60	0.22	0.38	2.81	3.10	3.24	3.08	0.48	0.49	0.54	0.50
UNIC [17]	0.64	0.94	0.60	0.74	2.85	3.15	3.31	3.14	0.49	0.50	0.54	0.51
Ours	**3.55**	**3.50**	**3.11**	**3.36**	**2.96**	**3.24**	**3.35**	**3.21**	**0.51**	**0.52**	**0.55**	**0.53**

4.2. Ablation Study

We conducted an ablation study of our method. Specifically, we compared three variants: optimized using (i) Adam as a gradient-based optimizer, (ii) CMA-ES as a black-box optimizer, and (iii) CMA-ES with scale adjustment (*i.e.* image aspect ratio optimization). As an evaluation metric, we adopted aesthetic scores estimated using VEN [21]. While VEN's scores are unbounded, the larger score indicates more aesthetic. Table 1 shows a quantitative comparison of the three variants and the input images. We can see Adam only slightly improves the scores from the input images, but the margins are quite small. Contrarily, CMA-ES outperforms Adam by a large margin, and the scores are further improved by accounting for image size adjustment.

This trend can also be confirmed from the qualitative comparison shown in Figure 3. Adam improves the scores only slightly but the resultant images look almost the same as the input images, which indicates that Adam fell into local minima. Using CMA-ES, the scores are drastically improved and the 3D views are largely changed from the input images. Note that the camera's viewing directions are clearly changed in the first and third rows of Figure 3 while preserving the 3D scene structure, which cannot be accomplished using traditional 2D cropping methods. The scores are further improved by accounting for the image scaling to adjust image aspect ratios. Also note that the aspect ratio might change significantly, for example, a vertical image became horizontal; see the bottom example in Figure 3. Hereafter we denote CMA-ES with image scaling as our method.

4.3. Comparison with Existing 2D Cropping Methods

Next, we compared our method with existing 2D cropping methods. The compared methods are VPN [21], VEN [21], CGS [14], CAC [10], and UNIC [17]. As evaluation metrics, in addition to VEN [21], we also evaluated aesthetic scores using SAMPNet [24] and TANet [9].

Table 2 shows the quantitative evaluation result. Surprisingly, even though our method uses only VEN as an aesthetic score estimator during optimization, our results are consistently better than other methods when evaluated using not only VEN but also SAMPNet and TANet.

Figures 5, 4, and 6 show the results of qualitative comparison. The relative image sizes in each row are maintained from the orignal sizes using uniform scaling. The aesthetic scores under images are calculated using VEN. It can be observed that the existing methods attempted to extract the most salient objects within the rectangle regions, while our method strived to find views that capture the whole 3D scenes, presumably because VEN prefers such views. This strategic difference is evident in the decisive differences in the under-image aesthetic scores estimated using VEN.

4.4. User Study

To further validate our method, we conducted a user study with 15 test image sets (*i.e.*, seven sets from Figure 5, six from Figure 4, and two from the first to the second rows of Figure 6). We compared the top-four methods, including the original images, in reference to Table 2, *i.e.*, Input, VEN, UNIC, and Ours. 64 subjects were requested to pick the most aesthetically pleasing result for each test set, mentioning the importance of image composition. Consequently, the selection percentages are: Input 12.5%, VEN 24.3%, UNIC 17.3%, and Ours 45.9%, which means that ours outperforms other methods.

This user study also revealed an interesting tendency of subjects' preferences. Overall, the subjects preferred images with a wider field of view to grasp the 3D scene structures, while they preferred images focusing on human subjects, if there were any in the scene. VEN also seemingly prefers a wider field of view, and thus, our optimized images have such views, resulting in the best selection percentage in this user study. This result stems from the characteristics of VEN, and a comprehensive investigation including other aesthetic score estimators is beyond the scope of our paper, which should be an exciting future work.

4.5. Limitations

Currently, our method has the following limitations.

Dependency on image extrapolation quality. Our method heavily relies on the quality of prior image extrapo-

Input	Ours	Input	Ours

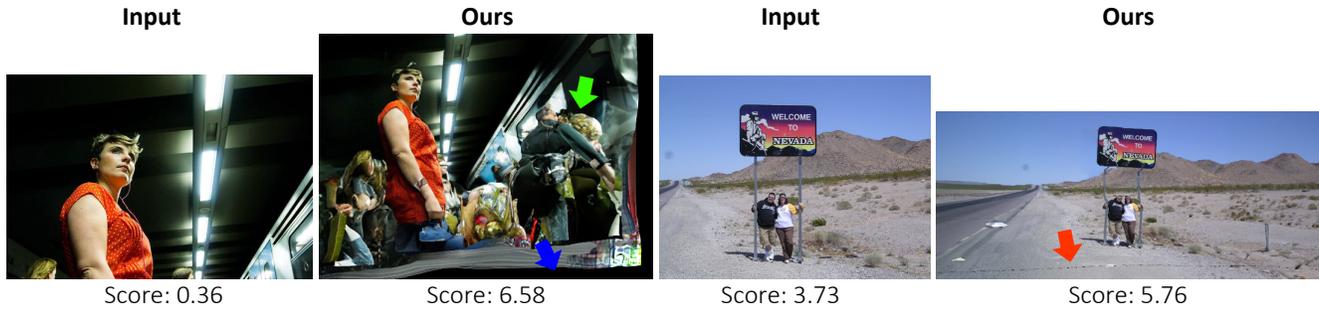| Score: 0.36	Score: 6.58	Score: 3.73	Score: 5.76

Figure 2. Failure cases. Blue arrow: unphotographed regions. Green arrow: unsatisfactory objects created by image extrapolation. Red arrow: gap in the point cloud.

lation. We currently use Adobe Photoshop's generative fill, but it often yields unsatisfactory extrapolated objects (see the green arrow in Figure 2), particularly when sufficient information for extrapolation is unavailable in the original image. More advanced extrapolation techniques would alleviate this defect in the future.

Appearance of out-of-photographed regions. Although we introduced a mask regularization term to suppress the appearance of out-of-photographed regions (see Section 3.2), unphotographed regions (*i.e.* the black background) do appear in the rendered images when the camera moves significantly (see the blue arrow in Figure 2). Image extrapolation with much larger regions might sometimes help, but it also increases the risk of the appearance of unsatisfactory objects. A hard constraint defined by the viewing frustum of the initial camera might fix the problem.

Dependency on 3D reconstruction and rendering. Our method also relies on the quality of 3D reconstruction and its rendering. Distorted objects appear if the quality of 3D reconstruction is low. Also, we observe gaps on object surfaces (see the red arrow in Figure 2) because the point cloud is locally sparse or looks sparse when viewed from specific viewpoints. A view-dependent splatting-based technique might fill such gaps.

5. Conclusions

In this paper, we have explored an unprecedented approach in improving image aesthetics by optimizing (i) camera parameters in the 3D scene reconstructed from a single input image and (ii) the output image size. We showed that a gradient-based optimizer easily falls into local minima, while a black-box optimizer successfully finds good solutions. Our resultant images often have drastically different views from those of the input images, which cannot be accomplished using traditional 2D cropping methods. Qualitative and quantitative evaluations as well as the user study

demonstrated that our method is more effective in improving image aesthetics than existing methods.

We are confident that our study is a siginificant first step toward the concept of "*aesthetic retrospective rephotography*," which means that we can re-take a photograph by going back to the time of photoshooting to change the camera settings for more aesthetically pleasing photography. In future work, we would like to explore the vast search space of various camera settings even further, *e.g.*, bokeh with depth-of-field control, tone and contrast control with different exposure levels and camera response curves, motion blur, and vignette effects.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2623–2631. ACM, 2019. 3

[2] Hadi AlZayer, Hubert Lin, and Kavita Bala. Autophoto: Aesthetic photo capture using reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 944–951, 2021. 2

[3] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10, 2007. 1

[4] Abhishek Badki, Orazio Gallo, Jan Kautz, and Pradeep Sen. Computational zoom: a framework for post-capture image composition. *ACM Trans. Graph.*, 36(4):46:1–46:14, 2017. 2

[5] 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 *WACV 2017*, pages 226–234. IEEE Computer Society, 2017. 3

[6] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 37–45, New York, NY, USA, 2017. Association for Computing Machinery. 1

[7] 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*, page 1105–1108, New York, NY, USA, 2014. Association for Computing Machinery. 3

[8] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001. 1, 3, 4

[9] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 942–948, 2022. 4

[10] 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 (CVPR)*, pages 7057–7066, 2021. 1, 4

[11] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3D deep learning with PyTorch3D. In *SIGGRAPH Asia 2020 Courses*. Association for Computing Machinery, 2020. 3

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1, 3, 4

[13] 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. 1

[14] 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. 1, 4

[15] Ligang Liu, Yong Jin, and Qingbiao Wu. Realtime aesthetic image retargeting. In *6th International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging (CAe 2010)*, pages 1–8. Eurographics Association, 2010. 2

[16] Sean J Liu, Maneesh Agrawala, Stephen DiVerdi, and Aaron Hertzmann. ZoomShop: Depth-aware editing of photographic composition. In *Computer Graphics Forum*, pages 57–70. Wiley Online Library, 2022. 2

[17] 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 (ICCV)*, pages 13023–13032, 2023. 4

[18] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44 (3):1623–1637, 2022. 2

[19] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. A comparative study of image retargeting. *ACM Trans. Graph.*, 29(6), 2010. 1

[20] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3D photography using context-aware layered depth inpainting. In *CVPR 2020*, pages 8025–8035. Computer Vision Foundation / IEEE, 2020. 2, 3

[21] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2018. 1, 2, 3, 4

[22] Desai Xie, Ping Hu, Xin Sun, Soren Pirk, Jianming Zhang, Radomír Mech, and Arie E Kaufman. Gait: Generating aesthetic indoor tours with deep reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7409–7419, 2023. 2

[23] H. Zeng, L. Li, Z. Cao, and L. Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5942–5950, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 3

[24] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. In *BMVC 2021*, page 144. BMVA Press, 2021. 4

[25] 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. 1

Input	Adam	CMA-ES w/o scaling	CMA-ES w/ scaling (Ours)

Score: 1.99	Score: 1.99	Score: 6.53	Score: 7.13
Score: 0.089	Score: 0.036	Score: 6.14	Score: 6.87
Score: 0.94	Score: 1.65	Score: 3.48	Score: 4.53
Score: 0.68	Score: 0.87	Score: 5.22	Score: 5.52
Score: 2.49	Score: 3.30	Score: 4.27	Score: 4.52

Figure 3. Ablation study. We compare three variants; optimized using (i) Adam (2nd column), (ii) CMA-ES (3rd column), and (iii) CMA-ES with image scaling (4th column). The aesthetic scores (the larger, the better) under images are calculated using VEN.

Input	VPN	VEN	CGS	CAC	UNIC	Ours
Score: 1.95	Score: 1.36	Score: 1.82	Score: 1.17	Score: 1.46	Score: 2.13	Score: 4.66
Score: 1.05	Score: 1.00	Score: 0.95	Score: -0.21	Score: 0.23	Score: 1.80	Score: 2.93
Score: 1.74	Score: 1.62	Score: 1.63	Score: 0.90	Score: 1.39	Score: 0.35	Score: 3.71
Score: -0.22	Score: -0.76	Score: -0.93	Score: -1.31	Score: -0.60	Score: -1.50	Score: 4.23
Score: 0.99	Score: 0.43	Score: 1.05	Score: 0.73	Score: 0.74	Score: 0.78	Score: 4.38
Score: 0.53	Score: 0.49	Score: -0.45	Score: 0.36	Score: -0.64	Score: 0.63	Score: 2.71

Figure 4. Qualitative comparison of our method with existing 2D cropping methods.

Input	VPN	VEN	CGS	CAC	UNIC	Ours
Score: 0.045	Score: 0.19	Score: 0.078	Score: -0.056	Score: 0.94	Score: 0.30	Score: 3.15
Score: -0.27	Score: -0.24	Score: -1.01	Score: -0.78	Score: 0.20	Score: -0.37	Score: 2.77
Score: 1.01	Score: 1.15	Score: 0.90	Score: 0.80	Score: 0.10	Score: 0.85	Score: 3.01
Score: -1.10	Score: -0.72	Score: -1.87	Score: -1.68	Score: 1.62	Score: -0.43	Score: 1.53
Score: 1.77	Score: 2.64	Score: 1.87	Score: 0.12	Score: 0.29	Score: 0.39	Score: 4.40
Score: 1.30	Score: 2.54	Score: 1.14	Score: 0.12	Score: 0.29	Score: 2.48	Score: 5.14
Score: 0.52	Score: -0.89	Score: 0.061	Score: -1.13	Score: -0.98	Score: -0.18	Score: 4.67

Figure 5. Qualitative comparison of our method with existing 2D cropping methods.

Input	VPN	VEN	CGS	CAC	UNIC	Ours
Score: 1.98	Score: 1.83	Score: 1.83	Score: 1.75	Score: 1.95	Score: 2.47	Score: 4.12
Score: 0.42	Score: -0.45	Score: 0.21	Score: -1.43	Score: -0.39	Score: -0.51	Score: 2.62
Score: 0.96	Score: 1.22	Score: 0.81	Score: -0.042	Score: 1.95	Score: 0.83	Score: 4.25
Score: 3.20	Score: 3.32	Score: 3.35	Score: 2.31	Score: 2.83	Score: 0.88	Score: 5.13
Score: 2.33	Score: 2.33	Score: 2.35	Score: 1.26	Score: 1.62	Score: -0.78	Score: 4.07
Score: 1.71	Score: 1.29	Score: 1.24	Score: 0.97	Score: 0.78	Score: 0.80	Score: 4.90

Figure 6. Qualitative comparison of our method with existing 2D cropping methods.