Aesthetic-guided Outward Image Cropping

LEI ZHONG*, Nankai University, China
FENG-HENG LI*, Nankai University, China
HAO-ZHI HUANG, Xverse, China
YONG ZHANG, Tencent AI Lab, China
SHAO-PING LU†, Nankai University, China
JUE WANG, Tencent AI Lab, China

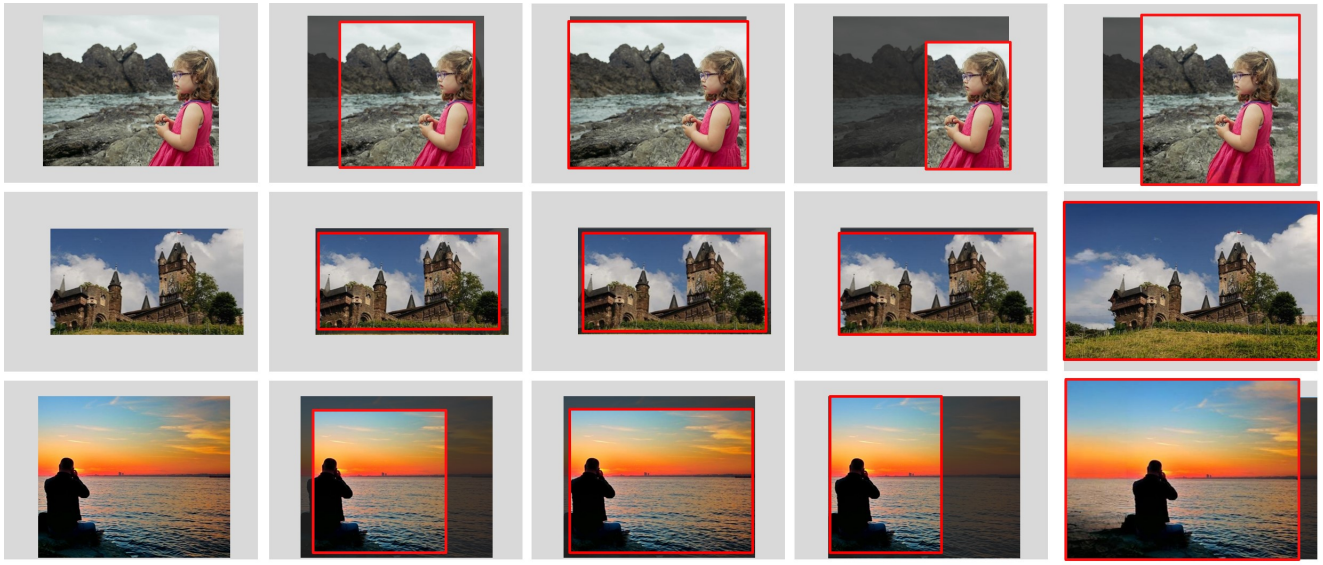

Fig. 1. Previous image cropping techniques are bound by the image border, thus may fail to produce a desirable result if a good composition does not exist within the image frame. We propose a new cropping operation called *Outward cropping*, which simultaneously expands the input image and creates a good composition from the modified FOV. We jointly consider the composition aesthetics and the quality of image extrapolation to achieve high-quality output.

Image cropping is a commonly used post-processing operation for adjusting the scene composition of an input photography, therefore improving its aesthetics. Existing automatic image cropping methods are all bounded by the image border, thus have very limited freedom for aesthetics improvement if the original scene composition is far from ideal, e.g. the main object is too close to the image border.

In this paper, we propose a novel, aesthetic-guided *outward image cropping* method. It can go beyond the image border to create a desirable composition that is unachievable using previous cropping methods. Our method first evaluates the input image to determine how much the content of the image should be extrapolated by a field of view (FOV) evaluation model. We then synthesize the image content in the extrapolated region, and seek an optimal aesthetic crop within the expanded FOV, by jointly considering the aesthetics of the cropped view, and the local image quality of the extrapolated image content. Experimental results show that our method can generate more visually pleasing image composition in cases that are difficult for previous image cropping tools due to the border constraint, and can also automatically degrade to an inward method when high quality image extrapolation is infeasible.

CCS Concepts: • **Computing methodologies** → **Computational photography**; **Image processing**; **Scene understanding**.

Additional Key Words and Phrases: image extrapolation, view composition, aesthetic evaluation, image cropping

ACM Reference Format:
Lei Zhong, Feng-Heng Li, Hao-Zhi Huang, Yong Zhang, Shao-Ping Lu, and Jue Wang. 2021. Aesthetic-guided Outward Image Cropping. *ACM Trans. Graph.* 40, 6, Article 211 (December 2021), 13 pages. https://doi.org/10.1145/3478513.3480566

*authors equally contribute to this work.
†Shao-Ping Lu is the corresponding author.

Authors' addresses: Lei Zhong, TKLNDST, CS, Nankai University, Tianjin, China, zhongleilz@icloud.com; Feng-Heng Li, TKLNDST, CS, Nankai University, Tianjin, China, lifengheng@foxmail.com; Hao-Zhi Huang, Xverse, Shenzhen, China, huanghz08@gmail.com; Yong Zhang, Tencent AI Lab, Shenzhen, China, zhangyong201303@gmail.com; Shao-Ping Lu, TKLNDST, CS, Nankai University, Tianjin, China, slu@nankai.edu.cn; Jue Wang, Tencent AI Lab, ShenZhen, China, arphid@gmail.com.

0730-0301/2021/12-ART211 $15.00
https://doi.org/10.1145/3478513.3480566

1 INTRODUCTION

View composition is one of the most important factors affecting image aesthetics. There are long-standing golden rules in photography for creating good compositions, such as the Rule of thirds, diagonal dominance, visual balance, avoiding distracting objects, etc. Given that capturing an image with perfect composition is hard even for professionals, image cropping becomes an essential step for improving image composition in the post-processing pipeline. It works by defining a rectangular region inside the input image as the final output and excludes the content outside the selected region. If done properly, this seemly simple interaction often yields dramatic improvement on image aesthetics, as it removes unwanted objects and re-positions the main subject according to composition rules.

Although manual image cropping tools are vastly available, extensive research has been conducted to develop automatic approaches for more intelligent image editing pipelines. Earlier methods rely on handcrafted features for evaluating composition [Barnes et al. 2009; Liu et al. 2010; Yan et al. 2013; Zhang et al. 2013], and more recently, deep learning becomes the de facto choice for developing more powerful learning-based cropping methods [Liang et al. 2017; Wei et al. 2018; Zeng et al. 2019]. In all these works, it is assumed that a good composition can be found within the original image frame, thus cropping is merely an inward process: the resulting image is always a sub-region in the original input. In practice, there are often cases where a good composition cannot be obtained by inward cropping, such as the examples shown in Fig. 1. In such examples, the main object either is too close to the image border or occupies a large portion of the image frame; thus the freedom of cropping becomes limited in order to keep the main object intact.

We argue that when adjusting an image composition, the cropping window should not be limited inside the field of view (FOV) of the given image, and a better composition can be created if we go beyond the border of the image when cropping. We call this operation *outward cropping*, and in this paper, we propose a novel approach to do it. In contrast to traditional cropping, which shrinks the image in all four directions, the outward cropping may shrink along with some directions with expanding along others to achieve a more favourable composition.

Several technical challenges need to be addressed in this outward cropping approach. First, the method should determine, based on the composition aesthetics of the input image, whether the FOV of the image needs to be expanded to find a good composition, or it can be found within the image. In the latter case, traditional inward cropping is sufficient to create a good composition. Second, if the FOV of the image needs to be expanded, the extrapolated part of the image should be visually realistic and semantically consistent with the origin image. Finally, when searching for a good composition outside the image, both the composition aesthetics and the quality of the extrapolated region need to be jointly considered. Image extrapolation is an ill-defined problem, and it cannot always produce high-quality results. When it is done well on an input image, the algorithm should pay more attention to composition aesthetics to take advantage of the expanded FOV. Otherwise, the algorithm should behave more conservatively to avoid introducing noticeable visual artifacts into the output. Therefore, how to balance composition aesthetics and extrapolation quality is a crucial issue.

To address the above challenges, we propose an aesthetic-guided outward cropping framework based on a holistic scene representation. Our framework consists of three main stages, as shown in Fig. 2. The first stage is to evaluate whether the FOV of the input image needs to be expanded, and furthermore, determine how much the image needs to be extrapolated using an FOV evaluation module. We then fill the extrapolated region with an image extrapolation neural network. Finally, considering both the composition aesthetics and quality of the extrapolated region, we employ a generative adversarial approach to look for an optimal crop that achieves a good trade-off between composition and extrapolation quality.

In summary, the main contributions of our work include:

- A content-aware image outward cropping method to expand the capabilities of traditional image cropping. To the best of our knowledge, it is the first image cropping method that allows the cropping window to extend outside the image border to find a visually aesthetic view.
- A generative adversarial approach to balance the composition aesthetics and the image quality of the extrapolated region.
- An extensive evaluation of the proposed method against existing image cropping methods and alternative baselines, both quantitatively and qualitatively, which demonstrates the effectiveness and characteristics of the proposed method.

2 RELATED WORK

2.1 Image extrapolation.

Extrapolating image content beyond image borders has gained considerable attention in image synthesis and computational photography. It predicts the content of an unknown region while maintaining semantic and structural coherency with the known region. Previous solutions can be divided into two subcategories: diffusion-based, patch-based methods [Zheng et al. 2019] and GAN-based methods. The former propagates pixel colors based on the isophote direction field [Ballester et al. 2001; Bertalmio et al. 2000] or global image statistics [Levin et al. 2003]. This approach often fails with highly-textured image regions. Patch-based methods synthesize missing regions by finding suitable patches from the input images [Barnes et al. 2009] or a pre-constructed dataset [Danon et al. 2019; Hays and Efros 2007]. They work well with repetitive textures but may fail when compatible patches are not available. By encoding the image region as a representative patch, Hu *et al.* [2013] introduce PatchNet to represent an image with a graph, where the geometric relationship between two regions as an edge. Wang *et al.* [2014; 2018] further propose data-driven image extrapolation, where graph nodes indicate region classification labels, and undirected graph edges represent spatial relationships. This problem was solved for the first time using the graph-based representation and corresponding subgraph matching algorithm, preserving the correct semantic structure of extrapolated results. Recently, GAN-based methods [Guo et al. 2020; Wang et al. 2019a; Yang et al. 2019] formulate the image extrapolation as an image-to-image translation task. Iizuka *et al.* [2017] present an adversarial training approach to ensure the generated images are both locally and globally consistent by two discriminators.

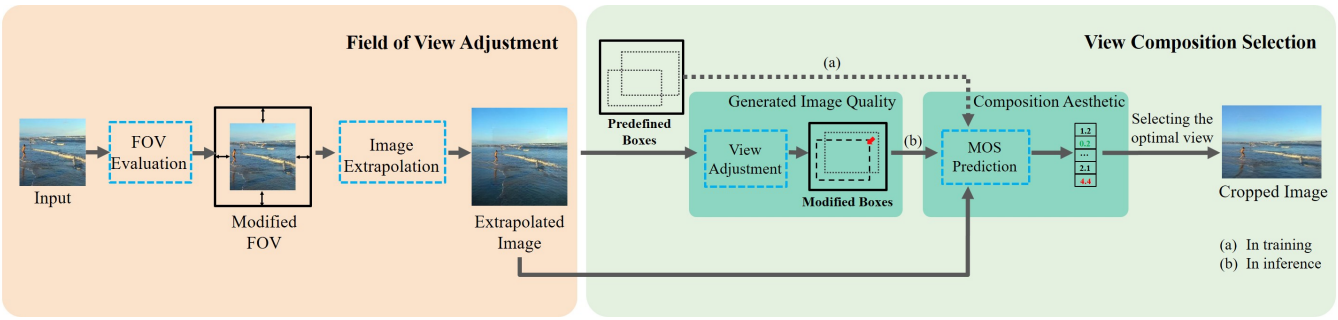

Fig. 2. Overview of the proposed outward cropping method. There are two main steps: Field of view Adjustment (left) and View Composition Selection (right). We first evaluate the FOV of the input image and determine how much the image needs to be extrapolated by a FOV evaluation module, then fill the unknown region using an image extrapolation module. To balance the generated image quality and composition aesthetics, we first adjust the positions of a pre-defined set of candidate boxes by maximizing the image quality inside each box and select the box with the highest aesthetic score as the final cropped region in the inference stage.

Teterwak *et al.* [2019] introduces semantic conditioning to module the behaviour of the discriminator. Zhao *et al.* [2021] propose co-modulated GANs, which bridges the gap between the modulated unconditional and image conditional generative models. It achieves excellent performance when filling large missing regions. In our approach, we use a similar network to fill the unknown region in the modified FOV.

2.2 Aesthetic image composition.

With the growing interest in improving the visual quality of digital photos, many image aesthetic enhancement techniques [Avidan and Shamir 2007; Gharbi et al. 2017; Hu et al. 2018] have been proposed. One of the essential factors in aesthetic image improvement is image composition. Previous image composition methods can be roughly categorized into cropping, warping, patch rearrangement operators, or a combination of the above ones. Warping [Liu et al. 2010] has been introduced to recompose images by relocating the salient objects. A triangular or quad mesh is constructed to represent the input image, and it is mapped to a target mesh with given aesthetic constraints like the rule of thirds (RT), visual balance (VB), and diagonal dominance (DD). Patch rearrangement [Barnes et al. 2009; Chang et al. 2014; Guo et al. 2012] methods divide the input image into non-overlapping or overlapping patches [Cho et al. 2009, 2008], and rearranges them to produce visually convincing results. The cut-and-paste methods [Bhattacharya et al. 2011; Zhang et al. 2013] explicitly extract the foreground objects and paste them onto the ideal placement according to aesthetic composition rules. However, foreground extraction is erroneous when the image contains complex background, leading to visual artifacts in the final composition. CompZoom [Badki et al. 2017] allows users to modify the image composition by manually changing the focal length and the camera position under a multi-perspective camera model. It requires a stack of images as input, while our method works with a single image.

2.3 Image cropping.

Automatic image cropping methods [Chen et al. 2016, 2017a; Esmaeili et al. 2017; Fang et al. 2014; Santella et al. 2006; Stentiford 2007; Suh et al. 2003; Zhang et al. 2012] seek for a rectangular cropping window to eliminate unwanted image objects, and at the same time properly positioning the main object to improve the aesthetics of the remaining region. Early image cropping methods [Santella et al. 2006; Stentiford 2007; Suh et al. 2003] are mostly based on attention mechanisms. They rely on saliency detection [Borji et al. 2019; Fan et al. 2020] to localize the main objects or the most informative region. However, tight cropping of the main objects may not guarantee visually pleasing results. Aesthetics-based cropping methods [Abeln et al. 2016; Chen et al. 2017b; Cornia et al. 2018; Tu et al. 2020; Wang and Shen 2017] utilize the image aesthetic characteristics or composition rules to improve the overall image quality. These methods use hand-crafted features to evaluate the quality of the candidate crops or adopt ranking models to rank them.

Thanks to the rapid development of deep learning techniques and newly developed datasets [Wei et al. 2018; Zeng et al. 2019], data-driven methods [Chen et al. 2017b; Cornia et al. 2018; Lu et al. 2019; Tu et al. 2020; Wang and Shen 2017; Zeng et al. 2019] have received increasing attention. Zeng *et al.* [2019] introduce a grid anchor-based formulation, making image cropping more efficient by reducing the searching space of candidate crops and defining more reliable evaluation metrics. Lu *et al.* [2019] formulate the image cropping as a listwise ranking problem and propose a refined view sampling to avoid the deformation in view generation. A meta-learning based cropping framework [Li et al. 2020a] is proposed to generate results with different aspect ratio requirements. The mutual relation between different candidate crops has been explored in [Li et al. 2020b] to find optimal compositions.

As mentioned earlier, although existing inward image cropping methods have achieved promising performance, they are limited by the image border and cannot produce satisfactory results when the main objects are too large or too close to the image border. In contrast, our method allows the candidate box to extend outside the image border, providing a higher degree of freedom for searching for better compositions.

3 METHOD

3.1 Overview

The pipeline of our aesthetic-guided outward cropping framework is shown in Fig. 2, which consists of two steps, *i.e.,* the field of view adjustment and the view composition selection. There are three main components: (1) field of view evaluation (Fig.3); (2) image extrapolation (Fig. 4); and (3) learning the cropping model. (1) and (2) are in the first step while (3) is in the second step.

Given an input image, our method first employs an FOV evaluation module to determine whether the FOV needs to be expanded from the perspective of image aesthetics. If it does, the evaluation module also determines how much the extrapolation should be. In this stage, we only focus on the perspective of image composition and ignores the impact of subsequent image extrapolation quality. Our method seeks for the minimum amount of content expansion of the input image for improving composition. This is under the consideration that as the amount of content extrapolation increases, it becomes more challenging to ensure high-quality image expansion.

Next, our method fills the expanded region by an image extrapolation module, and pass it to the final cropping module to produce a good cropping. However, image extrapolation is an ill-posed problem, and visual artifacts could be introduced in this process in difficult cases. If strong artifacts appear, the overall image quality would significantly degrade even if the composition itself is satisfactory. We thus design our cropping module in such a way that it balances the visual quality of extrapolated region and the overall composition aesthetics.

3.2 Field of View Evaluation

Although many successful image aesthetics assessment methods have been proposed, they cannot be directly applied to solve the unique problems that we encounter in this new task: (1) whether a good composition can be found within the given image, or extrapolation is needed; and (2) in the latter case, how much the image should be expanded for finding a good composition within.

Though general composition rules exist, *e.g.,* Rule of thirds [Grill and Scanlon 1990], we argue that the composition assessment is highly correlated with the image content, *i.e.,* the main factors affecting the composition quality differ in scenes. For example, for portrait images, most of the attention should be paid to the foreground person. For landscape images, on the other hand, the composition focus should mainly be on the interactions among different elements in the image. Directly applying the same existing rules to each image may lead to sub-optimal results. We thus propose to use neural networks to learn composition rules that are adaptive to the image content and use the learned model for FOV evaluation.

Specifically, we regard the FOV evaluation as a multi-classification task where the model predicts one of the pre-defined expansion ratios as the output label for a given image. Though converting FOV evaluation into a regression task is a straightforward solution, it brings great difficulties to the training of the model. Because it is subjective to accurately determine how much the image's content needs to be extrapolated to find a good composition, and the annotations are generally not available. Besides, to simplify the problem, in this step we assume the four borders of the image are extrapolated with the same ratio.

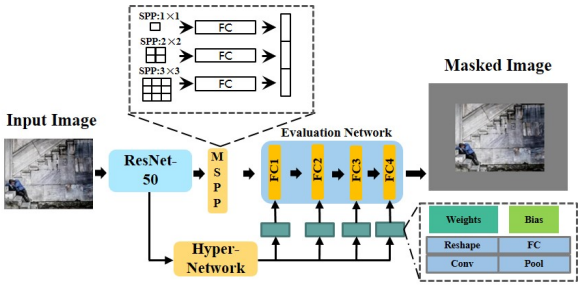

Fig. 3. Overview of our FOV evaluation pipeline. Given an image, we first extract semantic features from ResNet-50, and then import them to the hyper-network to generate weights for the evaluation network, which finally estimates how much the image content needs to be extrapolated.

As mentioned above that the composition rules should depend on the image content; we learn a hyper network to explicitly capture the rules by generating weights for an FOV evaluation network with the image content. As shown in Fig. 3, we first extract image semantics using a pre-trained backbone network, then utilize the hyper network to dynamically generate weights for the evaluation network that maps image semantics to one of the pre-defined ratios. Intuitively, the generated weights can be interpreted as the learned composition rules relative to the image content. Finally, the evaluation network predicts whether the current FOV needs to be expanded and determines how much the expansion should be.

3.2.1 Network Architecture. We employ a hyper network [Klocek et al. 2019] architecture following [Su et al. 2020], which consists of three 1×1 convolution layers and four weight generating branches. The branches generate weights and biases for the fully connected (FC) layers of the evaluation network composed of four FC layers. As shown in Fig. 3, given extracted features, FC weights are generated by a convolution and reshape operation, and FC biases are generated through a pooling and FC operation. We choose a pre-trained ResNet-50 [He et al. 2016] as the backbone network to extract semantic features. Motivated by the MNA-CNN-Scene [Mai et al. 2016] and spatial pyramid pooling (SPP) [He et al. 2015], we utilize multiple SPP modules (MSPP) to learn the multi-scale localization information for image compositions.

3.2.2 Loss Function. We pre-define five expansion ratios for the multi-classification task, *i.e.,* 0%, 12.5%, 25%, 37.5%, and 50%. Each one is treated as a categorical label. 0% indicates that no adjustment of FOV is needed. The FOV evaluation module is trained with the categorical cross-entropy loss.

3.2.3 Training. To create training data for the FOV evaluation module, we employ the existing image cropping dataset GAICD [Zeng et al. 2019] to generate sample images and their ground truth labels. The candidate box with the highest mean opinion score (MOS) of an image is selected as the target view. We first resize the image to 256×256 and randomly center-crop the result to a smaller size of $\alpha \times \alpha$, where $\alpha \in [128, 256)$. We then calculate the maximal distance of the four boundaries between the center-cropped image and the

target view and divide the maximum value by 16 to get the FOV evaluation label. The FOV evaluation module is trained using Adam with a learning rate of 1×10^4 and a batch size of 48 for 500 epochs.

3.3 Extrapolated Region Synthesis

Image extrapolation is a notably difficult task. With the recent progress of Generative Adversarial Networks (GANs) [Goodfellow et al. 2014], recent approaches treat image extrapolation as an image-to-image translation (I2I) task and have achieved good results. In our approach, we design a GAN-based image extrapolation method.

Inspired by the success of the StyleGAN2 [Karras et al. 2020] in high-resolution image generation, we utilize StyleGAN2 as our base generator. We formulate image extrapolation as embedding the image and mask to the latent space of StyleGAN2 and retrieving the latent code to synthesize a whole image. As shown in Fig. 4, we learn a conditional encoder to map the input image and its mask to a series of style vectors which are fed into StyleGAN2 for extrapolation.

3.3.1 Network Architecture. We mainly employ the same architecture with co-modulated GANs [Zhao et al. 2021], including the StyleGAN2 generator and the conditional encoder. Specifically, a local context discriminator ensures semantic consistency between the extrapolated region and the known region. Please refer to our supplemental material for the detailed architecture.

3.3.2 Loss Function. Like [Iizuka et al. 2017], we introduce a local adversarial loss to ensure the local consistency of the extrapolated regions and a perceptual loss to improve the perceptual quality of the results. Given a ground-truth high-resolution image x and a binary mask M with 0s for known pixels and 1s for unknown ones, we first generate a masked image x_z:

$$x_z = x \odot (1 - M), \tag{1}$$

where $\odot$ is an element-wise multiplication operator. To ensure that the extrapolation regions are realistic, we utilize a global discriminator D_g and a local discriminator D_l. The adversarial losses are:

$$L^g_{adv}(G, E, D_g) = E_{x\sim\mathbb{P}_x}[\log(D_g(x))] + E_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[\log(1 - D_g(\hat{x}))], \tag{2}$$

$$L^l_{adv}(G, E, D_l) = E_{e\sim\mathbb{P}_e}[\log(D_l(e))] + E_{\hat{e}\sim\mathbb{P}_{\hat{e}}}[\log(1 - D_l(\hat{e}))], \tag{3}$$

where $\hat{x} = G(E(x_z))$ is the output returned by the encoder $E(\cdot)$ and generator $G(\cdot)$, $e = x \odot M$ and $\hat{e} = \hat{x} \odot M$. Then, the entire loss for global and local discriminators is:

$$L_{adv} = (L^g_{adv} + L^l_{adv})/2. \tag{4}$$

For the pixel-wise reconstruction loss, we employ L_1 to minimize reconstructed difference between x and $\hat{x}$ by:

$$L_{rec} = \|\hat{x} - x\|_1 . \tag{5}$$

The perceptual loss is defined as:

$$L_{per} = \|\phi_l(\hat{x}) - \phi_l(x)\|_2 , \tag{6}$$

where $\phi_l(\cdot)$ denotes the feature activation function at the l-th layer of the VGG-19 networks $\phi(\cdot)$. In our work, we fix $l = 4$ for calculating the perceptual loss.

The total loss of our image extrapolation module is:

$$L_{total} = \lambda_{adv} L_{adv} + \lambda_{rec} L_{rec} + \lambda_{per} L_{per}, \tag{7}$$

where λ_{adv}, λ_{rec} , and λ_{per} are trade-off hyperparameters. We set λ_{adv}=0.02, λ_{rec}=10, and λ_{per}=1 in our experiments.

3.3.3 Training. The image extrapolation module is trained with high-resolution images (512×512) of the Place365-Challenge dataset [Zhou et al. 2017]. Following [Teterwak et al. 2019], we select the top 50 classes of the above dataset as our dataset, which covers most typical photography scenes. In our dataset, the last ten images of each class are extracted as the test set, and the remaining images are training set. The image extrapolation is trained with Adam with a learning rate of 0.002 and a batch size of 4 for 50k iterations.

3.4 Learning to Crop

Some existing image cropping methods [Liu et al. 2010; Lu et al. 2020] directly predict the position of the target cropping box, while others [Li et al. 2020b; Wei et al. 2018; Zeng et al. 2019] choose the box from a pre-defined candidate list. We choose the latter strategy given its simplicity and robust performance. Note that our task is significantly different from traditional crop evaluation. Previous methods score the candidate boxes based on real image content. Differently, in this work, since the image has been extrapolated before this step when evaluating the candidate boxes, we need to consider the composition aesthetics and the local image quality of the extrapolated content inside each box.

We propose an image cropping method that balances composition aesthetics and the image quality when selecting a good view to crop. Our framework consists of two main branches: the view adjustment module and the mean opinion score (MOS) prediction module. The former refines the position of each box based on local image synthesis quality, while the latter further predicts the MOS of each candidate box based on the composition aesthetics.

3.4.1 View Adjustment. The input of the view adjustment module is an extrapolated image with pre-defined candidate boxes and the output of that is the offset of each box. Compared with the MOS prediction with data annotation, it is more challenging to evaluate the image quality of the extrapolated content, given the lack of annotations. We therefore employ a weakly supervised method to address this problem. Formally, given an extrapolated image I_e with a set of candidate views $V = \{v_1, ..., v_N\}$, the modified candidate views $V^\psi = \left\{v^\psi_1, ..., v^\psi_N\right\}$ can be obtained through the view adjustment module Ψ. The view adjustment module predicts a set of offsets $T_{(X,Y)} = \left\{t_{(x_1,y_1)}, ..., t_{(x_N,y_N)}\right\}$ for candidate views to move to the neighboring regions with the best image quality. v^ψ can be seen as drawn independently from distribution of extrapolated images $\mathbb{P}$. Equally, corresponding ground-truth images $G = \{g_1, ..., g_N\}$ can be seen as drawn independently from a distribution of real images $\mathbb{Q}$. The view adjustment module can be optimized by:

$$\overline{\Psi} = \arg\min_{\Psi} D\left(\mathbb{P}_\psi, \mathbb{Q}\right), \tag{8}$$

where $\mathbb{P}_\Psi$ represents $\mathbb{P}$ is parameterized by Ψ and D represents the dissimilarity between distributions $\mathbb{P}_\Psi$ and $\mathbb{Q}$. We minimize the above formula to make the candidate box move to the area with the best image quality in the neighbouring region. According to to [Lu et al. 2020], the divergence based approach can be utilized to

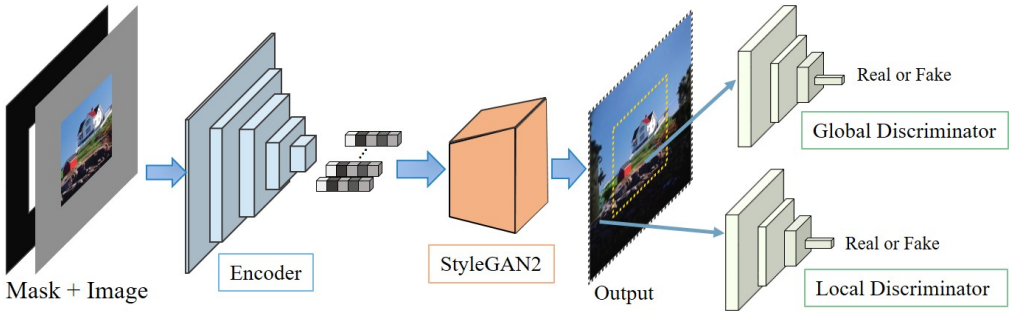

Fig. 4. Overview of image extrapolation architecture. The conditional encoder first embeds a masked image to a latent code in StyleGAN2's latent domain. StyleGAN2 then recovers the partial images and generates diverse samples in the extrapolated region. The global discriminator takes the entire image as input, while the local discriminator takes the extrapolated region as input. Both discriminators are trained to distinguish between the real and generated images.

calculate the distribution dissimilarity. Moreover, minimizing distribution dissimilarity can be converted to minimize the divergence based on the extrapolated images in the candidate views v^{ψ} and the corresponding ground-truth images G.

Therefore, we utilize the generative adversarial approach to solve the divergence minimization problem according to [Nowozin et al. 2016]. The view adjustment module acts as a generator to move the candidate view to a region with good image quality, and the discriminator determines whether the input images are from the real data distribution or the generated data distribution.

3.4.2 MOS Prediction. The input of the MOS prediction module is an extrapolated image with boxes refined by the view adjustment module and the output of that is the score of each candidate box. This procedure is similar to [Zeng et al. 2019], where the candidate boxes are scored based on the composition aesthetics. However, we further incorporate the image quality evaluation into the cropping pipeline. That is, instead of directly selecting the optimal result from the pre-defined candidate boxes, we first modify the position of each candidate box with good image quality of each candidate's view, then seek an optimal composition aesthetic from such modified candidate boxes. Please refer to our supplemental material for the ablation study of the order of the view adjustment module and the MOS prediction module.

3.4.3 Network Architecture. Similar to [Zeng et al. 2019], we utilize the same network architecture to extract features of different boxes. The MOS prediction module and view adjustment module share the same features extraction module. The MOS prediction module consists of two fully connected layers, and the view adjustment module consists of three fully connected layers.

3.4.4 Loss Function. We adopt the Huber loss to train the MOS prediction network for extrapolated images and corresponding real images. $g_{i,j}$ and $p_{i,j}$ are the ground-truth MOS and predicted score of the j-th crop of extrapolated image i, respectively. $p_{r,j}$ is that of the corresponding real image r. The Huber loss is defined as:

$$L_{ij} = \begin{cases} \frac{1}{2}(g_{i,j} - p_{i,j})^2, & \text{if } |g_{i,j} - p_{i,j}| \le 1, \\ |g_{i,j} - p_{i,j}| - \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{9}$$

Table 1. Ablation study of FOV evaluation on the GAICD dataset.

Method	*ACC*	$Acc_{1/5}$	$Acc_{1/10}$	$\overline{SRCC}$
2-class	**98.45**	41.5	60.0	0.706
5-class	90.21	**58.0**	73.0	**0.803**
9-class	85.23	42.5	60.5	0.700
17-class	55.59	42.0	61.0	0.698
33-class	30.48	43.0	62.5	0.705
65-class	13.98	48.0	65.5	0.708
regression	-	29.0	48.5	0.601
resnet-50	80.23	54.5	**77.5**	0.773
w/o SPP	85.05	46.5	63.0	0.693

Our Huber loss is defined as the combination of L_{ij} and L_{rj}:

$$L_{hub} = \frac{1}{2}L_{ij} + \frac{1}{2}L_{rj}. \tag{10}$$

Given an modified candidate view v_e^{ψ} from an extrapolated image I_e and a corresponding view v_r for real image I_r, we calculate the adversarial loss as:

$$L_{adv} = E_{I_r}[\log(D(v_r))] + E_{I_e}[\log(1 - D(v_e^{\psi}))], \tag{11}$$

where D is the discriminator and G is the generator. The generator consists of the region feature extraction module and the view adjustment module. Totally, the overall loss L_{total} of cropping model learning is:

$$L_{total} = \lambda_{hub}L_{hub} + \lambda_{adv}L_{adv}, \tag{12}$$

where $\lambda_{hub}$and λ_{adv} are set to 1 and 0.1, respectively.

3.4.5 Training. The view adjustment module and the MOS prediction module are trained simultaneously on the GAICD dataset. We take 1,036 images for training and 200 images for testing. We employ the same pre-defined anchors in the GAICD dataset to search for the optimal views. The Adam optimizer with α=0.0001, β_1=0.5, and β_2=0.9 is employed to train the view composition selection module for 80 epochs. The MOS prediction module computes the MOS scores on the pre-defined boxes in the training stage, and it predicts the MOS scores on the modified boxes during inference. In our experiment, the maximum displacement of the candidate box the view adjustment module can adjust is 1/10 of the length and width of the candidate box.

Table 2. Comparison with image extrapolation methods on the Place365 dataset."–" means that result is not available.

Model	SRN				Boundless				SpiralNet				*Ours*			
	50%	37.5%	25%	12.5%	50%	37.5%	25%	12.5%	50%	37.5%	25%	12.5%	50%	37.5%	25%	12.5%
PSNR↑	15.35	-	-	-	15.81	17.46	19.47	22.66	15.83	15.63	17.11	20.63	**15.98**	**18.17**	**21.16**	**26.34**
SSIM↑	0.501	-	-	-	0.495	0.640	0.755	0.894	0.496	0.546	0.700	0.863	**0.535**	**0.656**	**0.788**	**0.927**
FID↓	107.4	-	-	-	104.73	83.76	69.79	64.98	90.59	84.17	70.82	52.24	**61.50**	**47.41**	**32.12**	**18.05**

Table 3. Ablation study of Image Cropping Module on the GAICD dataset.

Method	w/o FOV					w/o VA					w/o Cropping					Ours				
	50%	62.5%	75%	87.5%	100%	50%	62.5%	75%	87.5%	100%	50%	62.5%	75%	87.5%	100%	50%	62.5%	75%	87.5%	100%
IOU ↑	0.724	0.617	0.460	0.342	0.253	**0.825**	**0.802**	**0.802**	**0.815**	0.810	0.633	0.630	0.635	0.643	0.636	0.720	0.720	0.705	0.724	**0.811**
BDE ↓	0.066	0.110	0.193	0.287	0.413	**0.041**	**0.047**	**0.047**	**0.044**	0.047	0.102	0.103	0.102	0.100	**0.017**	0.067	0.067	0.071	0.066	0.046
FID ↓	**66.168**	68.403	78.054	91.043	102.260	73.472	48.467	28.822	11.574	6.964	86.071	65.293	43.064	25.061	6.530	66.390	**41.448**	**20.216**	**9.881**	**6.602**
SSIM ↑	**0.690**	0.372	0.347	0.330	0.311	0.638	0.730	0.804	**0.902**	0.878	0.564	0.647	0.732	0.857	0.870	0.679	**0.772**	**0.833**	0.896	**0.879**
PSNR ↑	**18.274**	15.432	14.336	13.461	12.856	17.395	19.914	23.176	30.672	84.534	16.426	18.323	20.578	25.135	84.321	18.146	**20.808**	**25.238**	**33.296**	**84.542**

Table 4. Comparison with the state-of-the art methods on the sub-GAICD dataset. "–" means that result is not available. As the input image size decreases, the degradation of our approach is slower than the existing methods.

Method	*BDE* ↓	*IoU* ↑	Method	*BDE* ↓	*IoU* ↑	Method	*BDE* ↓	*IoU* ↑
A2RL (87.5%)	0.077	0.693	VEN (87.5%)	0.090	0.643	VPN (87.5%)	0.114	0.607
A2RL (75.0%)	0.109	0.606	VEN (75.0%)	0.110	0.558	VPN (75.0%)	0.098	0.449
A2RL (62.5%)	0.150	0.469	VEN (62.5%)	0.151	0.414	VPN (62.5%)	0.086	0.309
A2RL (50.0%)	0.198	0.325	VEN (50.0%)	0.198	0.279	VPN (50.0%)	0.079	0.190
WSIC (87.5%)	0.071	**0.733**	GAIC (87.5%)	0.076	0.681	Ours (87.5%)	**0.066**	0.724
WSIC (75.0%)	0.103	0.647	GAIC (75.0%)	0.139	0.533	Ours (75.0%)	**0.071**	**0.705**
WSIC (62.5%)	0.148	0.497	GAIC (62.5%)	0.210	0.547	Ours (62.5%)	**0.067**	**0.720**
WSIC (50.0%)	0.199	0.336	GAIC (50.0%)	0.281	0.245	Ours (50.0%)	**0.067**	**0.720**

Table 5. Comparison with the state-of-the-art methods on the GAICD dataset. "–" means that result is not available.

Method	$Acc_{1/5}$ ↑	$Acc_{2/5}$ ↑	$Acc_{3/5}$ ↑	$Acc_{4/5}$ ↑	$Acc_{1/10}$ ↑	$Acc_{2/10}$ ↑	$Acc_{3/10}$ ↑	$Acc_{4/10}$ ↑	$\overline{SRCC}$ ↑	*BDE* ↓	*IoU* ↑
A2RL	23.0	-	-	-	38.5	-	-	-	-	0.077	0.693
VPN	40.0	-	-	-	49.5	-	-	-	-	0.117	0.597
WSIC	27.0	-	-	-	46.5	-	-	-	-	0.071	0.693
VEN	40.5	36.5	36.7	36.8	54.0	51.0	50.4	48.4	0.621	0.101	0.617
GAIC	53.5	51.5	49.3	46.5	71.5	70.0	67.0	65.5	0.735	**0.041**	**0.823**
Ours	**60.5**	**57.0**	**52.0**	**49.4**	**74.5**	**73.3**	**72.0**	**70.9**	**0.763**	0.046	0.811

4 EVALUATION AND RESULTS

We conduct extensive experiments to demonstrate the effectiveness of the proposed aesthetic-guided outward image cropping framework. First, We analyze and evaluate each module of our framework against existing solutions and its ablations. We then compare our whole system with previous image cropping methods.

4.1 Field of View Evaluation Module

To verify the effectiveness of our FOV module, we conduct an ablation study to investigate the contribution of multiple SPP modules and the architecture of the hyper network. We utilize the classification accuracy (*i.e., Acc*) to evaluate whether the predicted FOV can cover the labeled best composition of the original image. $Acc_{1/5}$, $Acc_{1/10}$ and $\overline{SRCC}$ [Zeng et al. 2019] are reliable metrics to evaluate the performance of composition view selection. We use these metrics to evaluate the influence of the different settings of the FOV evaluation module on the final composition selection.

4.1.1 Ablation study. First, we compare the performance when regarding FOV evaluation as a multi-classification or a regression task. As shown in Table 1, all variants of multi-classification achieve better performance than that of regression. Therefore, our system is finally not formulated in a regression style. Second, we study the effect of formulating FOV evaluation as a 2,5,7,9,17,33, or 65-class classification problem. From Table 1, as the number of classes increases, the accuracy of evaluation decreases rapidly. 5-class classification achieves the best performance in $Acc_{1/5}$ and $\overline{SRCC}$ metric. We thus adopt this setting in subsequent experiments.

Third, we conduct an experiment by removing the multiple SPP modules from the system. From Table 1, the results suggest that the SPP module improves the performance on all metrics, proving the effectiveness of incorporating the localization information. Finally, we then investigate the effectiveness of the proposed hyper-network architecture by replacing it with an alternative approach. Because we treat FOV evaluation as a classification task, the straightforward

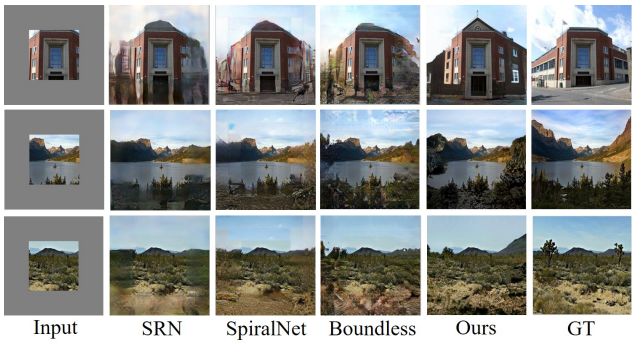

Fig. 5. Example image extrapolation results, comparing three state-of-the-art image extrapolation methods with ours. The results of existing methods have obvious artifacts, while our method ensures that the extrapolated content in the unknown region is realistic and semantically consistent with the origin image.

Fig. 6. Comparison with w/o FOVE. The results of w/o FOVE contain noticeable artifacts, which will seriously affect the subjective aesthetic.

approach is to fine-tune an existing classification network on our dataset. We choose the ResNet-50 with SPP modules to conduct 5-class classification experiments for comparison. The results show that our network performs better than ResNet-50 on all metrics except for $Acc_{1/10}$, proving the superiority of our hyper network architecture.

4.2 Image Extrapolation Module

We compare our image extrapolation module with three state-of-the art image extrapolation methods, including Boundless [Teterwak et al. 2019], SpiralNet [Guo et al. 2020], and SRN [Wang et al. 2019b]. We use the source code provided by the authors to re-train their models on the Place356-Challenge dataset [Zhou et al. 2017], and use the best performing model for comparison. We extrapolate the four sides of the image in equal proportions to evaluate the above methods on the test set of Place356-Challenge dataset. Experiments

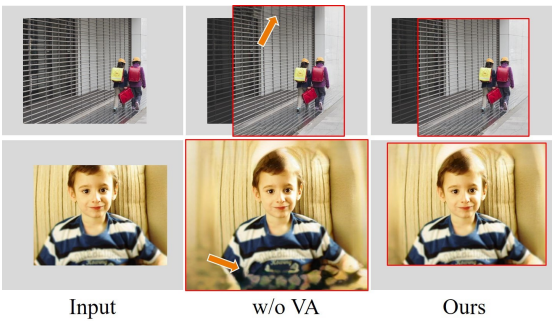

Fig. 7. Comparison with w/o VA. The results of w/o VA achieve a well-composed composition, but it includes artifacts or blurred image parts. Our view adjustment module can effectively improve the image quality of the extrapolated content by moving the pre-defined candidate boxes towards high image quality regions.

Fig. 8. Comparison with w/o Cropping. Our cropping module can successfully eliminate the unsatisfactory and redundant parts.

include four settings, *i.e.*, the outermost 50%, 37.5%, 25%, and 12.5% of the pixels in the image are masked respectively to be filled. We use three commonly used metrics to evaluate the visual realism and semantic consistency of the filled region, including the Structural similarity measure (SSIM), Peak signal-to-noise ratio (PSNR), and Frechet Inception Distance (FID).

Some visual results are shown in Fig. 5. The results of SRN maintain semantic consistency well, but the extrapolated regions are blurry and can be easily identified as synthesized. Since the SpiralNet progressive extrapolates the image spirally, there are several noticeable boundary artifacts in the extrapolated region. The Boundless approach tends to generate visible raindrop-shape artifacts. In contrast, our approach generates more convincing results with realistic details while maintaining semantic consistency.

The quantitative results are shown in Table 2. Due to the limitation of the feature expansion module, SRN can only achieve 50% extrapolation. Therefore, we only calculate the metric of the SRN method at the scale of 50%. The results show that our method achieves better scores than other methods under all experimental settings. Specifically, our extrapolated images form closer distributions to the real testing set, with FID scores achieve 44% improvement on average at all mask scales.

Fig. 9. Comparisons with the state-of-art methods in the validation set of GAICD. VEN [Wei et al. 2018] and VPN [Wei et al. 2018] tend to crop out the salient objects. GAIC [Zeng et al. 2019] and WSIC [Lu et al. 2020] sometimes directly return the input image. In contrast, our method produces a visually appealing composition by outward cropping.

Fig. 10. Comparisons with GAIC under different aspect ratios.

Fig. 11. Comparisons with the state-of-art methods in the original images of GAICD.

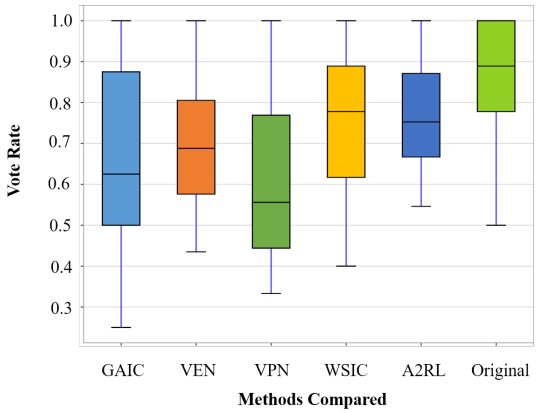

Fig. 12. A user study is illustrated in the box plot, using five summaries – minimum, first quartile, median, third quartile, and maximum. The original images and results of each method are compared with that of ours separately. If the voting rate is significantly higher than 0.5, it means our method surpasses the compared one.

4.3 Image Cropping Module

In the image cropping module, we utilize the same region feature extraction module (ROI+ROD) as the GAIC [Zeng et al. 2019]. We compare our cropping module against the following variants:

- *w/o FOV*: bypass the FOV evaluation module, *i.e.*, directly extrapolates image borders to the maximum value before cropping.
- *w/o VA*: bypass the view adjustment module, *i.e.*, directly evaluate and select from the pre-defined candidate boxes.
- *w/o Cropping*: bypass the entire cropping module and directly output the extrapolated image.

We evaluate the quality of tlhe extrapolated region by calculating the *PSNR*, *SSIM*, and *FID* between the selected view and the corresponding real image. The *IoU* and *BDE* are used as composition aesthetic metrics to evaluate the performance of aesthetic view selection. Note that here $Acc_{K/N}$ and $\overline{SRCC}$ are not selected, because their values are based on the annotated scores of the pre-defined candidate boxes instead of the modified ones generated by our view adjustment module. In Table 3, w/o FOV performs poorly in both composition aesthetic metrics and image quality metrics. It proves that directly expanding image borders to the maximum value increases the difficulty of finding a well-composed view, and at the same time, increases the possibility that extrapolation artifacts are included in the result. The results of our proposed method are slightly worse than the w/o FOV evaluation on the 50% scale, because our method is affected by the accuracy of FOV evaluation (*i.e.*, 96.5% accuracy on the 50% scale). In contrast, w/o FOV evaluation directly extrapolates image's content at the maximum pre-defined expansion ratio before cropping (the maximum expansion ratio is 50% in our settings). w/o VA performs slightly better than our method in composition aesthetic metrics but worse than our method in image quality metrics. This is expected as our method tries to strike a balance between improving composition aesthetics and minimizing visual artifacts. In fact, we treat the latter at a higher priority because extrapolation artifacts could significantly reduce the realism of the output image. Such examples are shown in Fig. 6 and Fig. 7, where our method effectively filters out artifacts and blurred image regions, resulting in an overall well-composed view with no visual artifacts. w/o Cropping also performs worse than our method in both composition aesthetic metrics and image quality metrics, indicating the necessity of cropping. Fig. 8 shows that the cropping module can successfully eliminate the unsatisfactory and redundant region introduced by extrapolation.

4.4 Comparisons with the state-of-the-art

4.4.1 Experimental Setup. We compare our method with several state-of-the-art image cropping methods that have source code publicly available. For a fair comparison, we compare different methods in two scenarios. The first one is the case that a good composition cannot be found within the original image frame, for which our method is mainly designed for. The second scenario is

that a satisfactory composition view exists in the input image, which ideally could be well handled by previous methods. For the first scenario, we first center-crop the original images with a smaller size (87.5%, 75%, 62.5%, and 50% resizing ratio, respectively) to build four sub-datasets. The labeled cropping box with the highest MOS score of the original image is used as the target view of the corresponding sub-image. We utilize the Intersection over Union (*IoU*) and the Boundary Displacement Error (*BDE*) as metrics to assess the similarity between the predicted and the target views. We also compute $Acc_{K/N}$ and $\overline{SRCC}$ to evaluate the predicted results. $\overline{SRCC}$ measures the rank similarity between the annotated and predicted scores of candidate boxes. $Acc_{K/N}$ computes in average how many of the top-K boxes predicted by the model fall into the top-N annotated boxes as a way to evaluate whether the returned top-K boxes are acceptable. We set N to either 5 or 10 and set K = 1,2,3,4 for both N = 5 and N = 10. More details of these metrics can be seen in [Zeng et al. 2019].

We select the methods of A2RL [Li et al. 2018], VPN [Wei et al. 2018], VEN [Wei et al. 2018], GAIC [Zeng et al. 2019] and WSIC [Lu et al. 2020] for comparison. Because A2RL and WSIC only produce one predicted box and the pre-defined candidate boxes of VPN are different from us, we only compare $Acc_{1/5}$, $Acc_{1/10}$ for them. Because many candidate boxes of the original image cross the boundary of the sub-image, finding the candidate box in the original image that is nearest to the predicted box in the sub-image will bring errors. $Acc_{k/N}$ and $\overline{SRCC}$ are not utilized in the scenario 1.

4.4.2 Scenarios 1. As shown in Table 4, our method outperforms existing methods with higher *IoU* and lower *BDE*. When the input image size decreases to 75%, it is difficult for A2RL, WSIC, and VPN to obtain good values of the above metrics. The results of VEN are with relatively low values in *BDE* and high values in *IoU* without much fluctuation due to the fact that VEN tends to focus on main objects. As the input image size further decreases, it becomes more challenging to find an optimal composition; thus the calculated performance of the previous methods degrades rapidly. However, The performance degradation of our method is much slower, thanks to its unique ability of finding a good composition outside the input image frame. Even when the input image size becomes 50%, our method still achieves 78% and 206% improvement than that of GAIC in terms of the *IoU* and *BDE* metrics, respectively.

We additionally constructed a validation dataset, including 454 images, by manually cropping the images of the GAICD dataset so that salient objects are close to the border of the images. Visual comparisons on the validation dataset are shown in Fig. 9. VEN and VPN tend to crop out the salient objects, but their global compositions are not satisfactory in many cases. GAIC and WSIC sometimes directly return the input image. A2RL slightly cuts some essential objects, which is undesirable. Our method overall produces visually appealing composition by outward cropping.

We conduct experiments on the validation dataset to verify which cases are more important in practice. The result of the FOV evaluation shows that the scales of 50%, 62.5%, 75%, 87.5%, and 100% are with 0.66%, 20.93%, 38.55%, 32.92%, and 7.05% of the total images, respectively. There are few cases when half of the image content needs to be extrapolated . As for the cases where scales are lower than 50%, most images are severely destroyed and even humans could not easily recognize the original contents. Therefore, those cases lower than 50% are ignored in this paper.

A well-designed image cropping system should be able to obtain visually pleasing results under different aspect ratios. We compare our method against GAIC, which employs the same pre-defined candidate boxes as ours, under three widely used aspect ratios, i.e., 16:9, 4:3, and 1:1. As shown in Fig. 10, our method generates more visually pleasing compositions than GAIC by outward cropping.

4.4.3 Scenarios 2. For the second scenario, we conduct experiments on the original images of the GAICD dataset. To avoid changing the position of the predefined candidate box, we do not use the view adjustment module when calculating $Acc_{K/N}$ and $\overline{SRCC}$. The quantitative results are shown in Table. 5. It can be observed that A2RL and WSIC do not perform well in terms of the $Acc_{K/N}$ score. This may be because WSIC is trained in a weakly supervised way where the ground-truth labels are not used. A2RL is supervised by an aesthetic classifier that cannot accurately capture the difference between different views within one image. As shown in Fig. 11, WSIC fails to remove unimportant elements in some cases. The performance of the VPN is slightly worse than the VEN because it is supervised by the VEN during training. Our method achieves comparable results with GAIC, performing better in $Acc_{K/N}$ and $\overline{SRCC}$, and slightly worse in *IoU* and *BDE*. We also conduct experiments on the other public datasets. More comparison results are shown in the supplementary material.

4.4.4 Running Speed. Our model is tested on a machine with Intel(R) Core(TM) i7-7800K CPU@3.50GHz, 64Gb Memory, and one Nvidia 2080Ti GPU. We employ the frame-per-second (FPS) metric to compute the running speed of our method. In general, our overall model runs about 15 FPS, wherein the FOV evaluation, image extrapolation, and image cropping modules are with 66, 25, and 100 FPS, respectively. Replacing the image extrapolation module with a lightweight network can further increase the running speed of our method.

4.4.5 Discussions. It is worth noting that no adjustment is applied in our method for the two scenarios. In practice, given an input image, the algorithm decides on itself whether or not to cross the original image border to find a good composition, depending on the image content. If the algorithm decides that no expansion is needed, it gracefully falls back to a standard inward cropping method that achieves comparable results to the state-of-the-art.

As mentioned earlier, some previous methods use image warping and seam caving instead of cropping to improve view composition. In Fig 13, we visually compare our method against a warping-based method [Liu et al. 2010] and seam carving based method [Guo et al. 2012] . The results of [Liu et al. 2010] and [Guo et al. 2012] contain noticeable deformation, and the relative positions of the objects are not well maintained. Our method produces more realistic results than existing methods.

5 USER STUDY

Due to the subjective nature of image aesthetics, we design a user study to further evaluate our method. Considering that allowing the

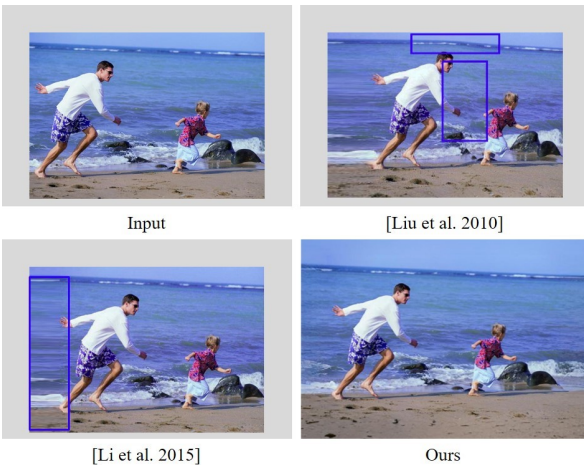

Fig. 13. Comparison with other retargeting methods. The results of [Li et al. 2015; Liu et al. 2010] contain noticeable distortions.

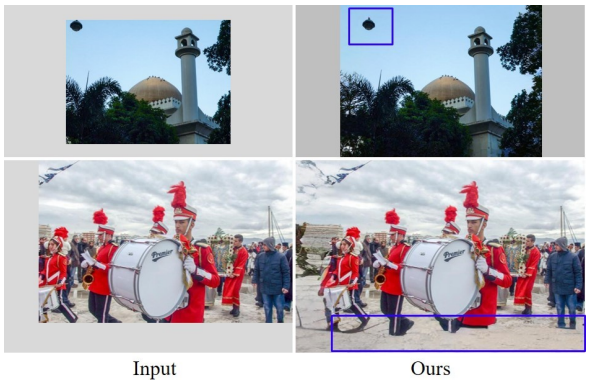

Fig. 14. Example failure cases. Our method may fail to generate semantically meaningful regions when the objects miss essential parts.

participant to rank or score various methods will increase the difficulty of the participant's selection and thus affect the validity of the results, we design an online questionnaire with pairwise A/B tests. We randomly select 38 input images, including the need for inward cropping and outward cropping, and generate 190 cropped images by five different methods, including ours. The original images and results of each method are compared with that of ours separately, yielding a total of 190 image pairs. The original image is included to evaluate whether our method improves the image composition of the original image. The human subjects are asked to pick one that is more visually appealing from each pair.

We recruited 121 human subjects in our university with diverse academic backgrounds to participate in the study, including 60 males and 61 females, and their ages range from 18 to 40. Instead of computing the average across 38 questions, we use the minimum, first quartile, median, third quartile, and maximum values to display the statistical distribution of each method.

The voting rate is defined as the percentage of choices that respond to the preferred images produced by our method. For instance, if 12 out of 20 people mark our output images as the best one, the voting rate is 60%. We collect the voting rate of 38 questions and construct a box plot in Fig. 12. The result shows that by effectively improving the original image composition, our method outperforms existing methods by a large margin.

We conducted z-tests on the user study results. Compared against the results from GAIC, VEN, VPN, WSIC, and A2RL, ours are preferred by 68.0%, 69.5%, 60.9%, 74.7%, and 75.5% of the participants, respectively. Paired z-tests further prove ours significantly outperforms GAIC ($z = 5.122, p < 0.001$), VEN ($z = 7.951, p < 0.001$), VPN ($z = 3.810, p < 0.001$), WSIC ($z = 9.239, p < 0.001$), and A2RL ($z = 10.924, p < 0.001$).

6 LIMITATION

Our method may fail to produce a satisfactory result in several special scenarios. When the main object in the image is missing essential parts or global context, the image extrapolation module may generate locally consistent but semantically wrong extrapolation results. Consequently, the cropping result is not very pleasing. Two such examples are shown in Fig. 14, where our method fails to synthesize a realistic pole (up) and people's legs (down). Having a more powerful, semantic image extrapolation method would help eliminate such artifacts, which is beyond the scope of this paper. A possible improvement would be that the FOV evaluation could better predict these situations and reduce the need for image extrapolation. Our method is also limited by the cropping operation itself. When the input image contains multiple objects that are occluding each other, it is challenging to obtain a good composition from the given viewpoint. Further relaxation on changing the viewpoint could give the method more freedom to generate a good composition.

7 CONCLUSION

In this paper, we have presented a novel aesthetic-guided outward cropping approach, which can go beyond the image border to obtain a well-composed composition. Our method has the following major characteristics: (1) given an image, the algorithm can determine by itself whether to expand the current view and, if yes, by how much; and (2) it achieves a good trade-off between composition aesthetics and image extrapolation quality so that the final output is both visually pleasing and artifact-free. Extensive experiment results show that our method can generate a more visually pleasing composition than existing image cropping methods, especially when the original FOV lacks an aesthetic composition.

ACKNOWLEDGMENTS

We would like to thank all reviewers for their valuable comments. This work was funded partially by NSFC (No. 61972216 and No. 62111530097), Tianjin NSF (18JCYBJC41300 and 18ZXZNGX00110), and BNRist (BNR2020KF01001).

REFERENCES

Jonas Abeln, Leonie Fresz, Seyed Ali Amirshahi, I Chris McManus, Michael Koch, Helene Kreysa, and Christoph Redies. 2016. Preference for well-balanced saliency in details cropped from photographs. *Frontiers in human neuroscience* 9 (2016), 704.

Shai Avidan and Ariel Shamir. 2007. Seam Carving for Content-Aware Image Resizing. *ACM Trans. Graph.* 26, 3 (2007).

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

Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. 2001. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.* 10, 8 (2001), 1200–1211.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* 28, 3 (2009), 24.

Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image inpainting. In *SIGGRAPH*. 417–424.

Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2011. A holistic approach to aesthetic enhancement of photographs. *ACM Trans. Multim. Comput.* 7, 1 (2011), 1–21.

Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. 2019. Salient object detection: A survey. *Computational visual media* 5, 2 (2019), 117–150.

Hui-Tang Chang, Yu-Chiang Frank Wang, and Ming-Syan Chen. 2014. Transfer in photography composition. In *Proc. ACM MM*. 957–960.

Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. 2016. Automatic image cropping: A computational complexity study. In *Proc. CVPR*. 507–515.

Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. 2017a. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *Proc. WACV*. 226–234.

Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. 2017b. Learning to compose with professional photographs on the web. In *Proc. ACM MM*. 37–45.

Taeg Sang Cho, Shai Avidan, and William T Freeman. 2009. The patch transform. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 8 (2009), 1489–1501.

Taeg Sang Cho, Moshe Butman, Shai Avidan, and William T Freeman. 2008. The patch transform and its applications to image editing. In *Proc. CVPR*. IEEE, 1–8.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Trans. Image Process.* 27, 10 (2018), 5142–5154.

Dov Danon, Hadar Averbuch-Elor, Ohad Fried, and Daniel Cohen-Or. 2019. Unsupervised natural image patch learning. *Computational Visual Media* 5, 3 (2019), 229–237.

Seyed A Esmaeili, Bharat Singh, and Larry S Davis. 2017. Fast-at: Fast automatic thumbnail generation using deep neural networks. In *Proc. CVPR*. 4622–4630.

Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. 2020. S4Net: Single stage salient-instance segmentation. *Computational Visual Media* 6, 2 (2020), 191–204.

Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. 2014. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proc. ACM MM*. 1105–1108.

Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. 2017. Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph.* 36, 4 (2017), 1–12.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. NeurIPS*. 2672–2680.

Tom Grill and Mark Scanlon. 1990. *Photographic composition*. Amphoto Books.

Dongsheng Guo, Hongzhi Liu, Haoru Zhao, Yunhao Cheng, Qingwei Song, Zhaorui Gu, Haiyong Zheng, and Bing Zheng. 2020. Spiral Generative Network for Image Extrapolation. In *Proc. ECCV*. Springer, 701–717.

YW Guo, Mingming Liu, TT Gu, and WP Wang. 2012. Improving photo composition elegantly: Considering image similarity during composition optimization. In *Comput. Graph. Forum.*, Vol. 31. Wiley Online Library, 2193–2202.

James Hays and Alexei A Efros. 2007. Scene completion using millions of photographs. *ACM Trans. Graph.* 26, 3 (2007), 4–es.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 9 (2015), 1904–1916.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*. 770–778.

Shi-Min Hu, Fang-Lue Zhang, Miao Wang, Ralph R. Martin, and Jue Wang. 2013. PatchNet: A Patch-Based Image Representation for Interactive Library-Driven Image Editing. *ACM Trans. Graph.* 32, 6 (2013), 196:1–12.

Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2018. Exposure: A white-box photo post-processing framework. *ACM Trans. Graph.* 37, 2 (2018), 1–17.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Trans. Graph.* 36, 4 (2017), 1–14.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proc. CVPR*. 8110–8119.

Sylwester Klocek, Łukasz Maziarka, Maciej Wołczyk, Jacek Tabor, Jakub Nowak, and Marek Śmieja. 2019. Hypernetwork functional image representation. In *International Conference on Artificial Neural Networks*. 496–510.

Anat Levin, Assaf Zomet, and Yair Weiss. 2003. Learning How to Inpaint from Global Image Statistics.. In *ACM Trans. Graph.*, Vol. 1. 305–312.

Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. 2018. A2-RL: Aesthetics aware reinforcement learning for image cropping. In *Proc. CVPR*. 8193–8201.

Debang Li, Junge Zhang, and Kaiqi Huang. 2020a. Learning to Learn Cropping Models for Different Aspect Ratio Requirements. In *Proc. CVPR*. 12685–12694.

Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. 2020b. Composing Good Shots by Exploiting Mutual Relations. In *Proc. CVPR*. 4213–4222.

Ke Li, Bo Yan, Jun Li, and Aditi Majumder. 2015. Seam carving based aesthetics enhancement for photos. *Signal Process Image Commun.* 39 (2015), 509–516.

Yuan Liang, Xiting Wang, Song-Hai Zhang, Shi-Min Hu, and Shixia Liu. 2017. PhotoRecomposer: Interactive photo recomposition by cropping. *IEEE Trans. Vis. Comput. Graph.* 24, 10 (2017), 2728–2742.

Ligang Liu, Yong Jin, and Qingbiao Wu. 2010. Realtime Aesthetic Image Retargeting. *Computational aesthetics* 10 (2010), 1–8.

Peng Lu, Jiahui Liu, Xujun Peng, and Xiaojie Wang. 2020. Weakly Supervised Real-time Image Cropping based on Aesthetic Distributions. In *Proc. ACM MM*. 120–128.

Weirui Lu, Xiaofen Xing, Bolun Cai, and Xiangmin Xu. 2019. Listwise view ranking for image cropping. *IEEE Access* 7 (2019), 91904–91911.

Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-preserving deep photo aesthetics assessment. In *Proc. CVPR*. 497–506.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-GAN: training generative neural samplers using variational divergence minimization. In *Proc. NeurIPS*. 271–279.

Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. In *SIGCHI*. 771–780.

Fred Stentiford. 2007. Attention based auto image cropping. In *International Conference on Computer Vision Systems: Proceedings (2007)*.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proc. CVPR*. 3667–3676.

Bongwon Suh, Haibin Ling, Benjamin B Bederson, and David W Jacobs. 2003. Automatic thumbnail cropping and its effectiveness. In *Proc. UIST*. 95–104.

Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. 2019. Boundless: Generative adversarial networks for image extension. In *Proc. ICCV*. 10521–10530.

Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. 2020. Image Cropping with Composition and Saliency Aware Aesthetic Score Map. 12104–12111.

Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R Martin, and Shi-Min Hu. 2014. Biggerpicture: data-driven image extrapolation using graph matching. *ACM Trans. Graph.* 33, 6 (2014), 1–13.

Miao Wang, Ariel Shamir, Guo-Ye Yang, Jin-Kun Lin, Guo-Wei Yang, Shao-Ping Lu, and Shi-Min Hu. 2018. BiggerSelfie: Selfie video expansion with hand-held camera. *IEEE Trans. Image Process.* 27, 12 (2018), 5854–5865.

Wenguan Wang and Jianbing Shen. 2017. Deep cropping via attention box prediction and aesthetics assessment. In *Proc. ICCV*. 2186–2194.

Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. 2019a. Wide-context semantic image extrapolation. In *Proc. CVPR*. 1399–1408.

Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. 2019b. Wide-Context Semantic Image Extrapolation. In *Proc. CVPR*. 1399–1408.

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. 2018. Good view hunting: Learning photo composition from dense view pairs. In *Proc. CVPR*. 5437–5446.

Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. 2013. Learning the change for automatic image cropping. In *Proc. CVPR*. 971–978.

Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. 2019. Very long natural scenery image prediction by outpainting. In *Proc. ICCV*. 10561–10570.

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. 2019. Reliable and efficient image cropping: A grid anchor based approach. In *Proc. CVPR*. 5949–5957.

Fang-Lue Zhang, Miao Wang, and Shi-Min Hu. 2013. Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Trans. Multimedia* 15, 7 (2013), 1480–1490.

Luming Zhang, Mingli Song, Qi Zhao, Xiao Liu, Jiajun Bu, and Chun Chen. 2012. Probabilistic graphlet transfer for photo cropping. *IEEE Trans. Image Process.* 22, 2 (2012), 802–815.

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. 2021. Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In *Proc. ICLR*.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic Image Completion. In *Proc. CVPR*. 1438–1447.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2017), 1452–1464.