Beyond Image Borders: Learning Feature Extrapolation for Unbounded Image Composition

Xiaoyu Liu[1], Ming Liu[1](✉), Junyi Li[1], Shuai Liu, Xiaotao Wang, Lei Lei, Wangmeng Zuo[1,2]
[1]Harbin Institute of Technology, Harbin, China [2] Peng Cheng Laboratory, Shenzhen, China
liuxiaoyu1104@gmail.com, csmliu@outlook.com, nagejacob@gmail.com, wmzuo@hit.edu.cn

Abstract

For improving image composition and aesthetic quality, most existing methods modulate the captured images by striking out redundant content near the image borders. However, such image cropping methods are limited in the range of image views. Some methods have been suggested to extrapolate the images and predict cropping boxes from the extrapolated image. Nonetheless, the synthesized extrapolated regions may be included in the cropped image, making the image composition result not real and potentially with degraded image quality. In this paper, we circumvent this issue by presenting a joint framework for both unbounded recommendation of camera view and image composition (i.e., UNIC). In this way, the cropped image is a sub-image of the image acquired by the predicted camera view, and thus can be guaranteed to be real and consistent in image quality. Specifically, our framework takes the current camera preview frame as input and provides a recommendation for view adjustment, which contains operations unlimited by the image borders, such as zooming in or out and camera movement. To improve the prediction accuracy of view adjustment prediction, we further extend the field of view by feature extrapolation. After one or several times of view adjustments, our method converges and results in both a camera view and a bounding box showing the image composition recommendation. Extensive experiments are conducted on the datasets constructed upon existing image cropping datasets, showing the effectiveness of our UNIC in unbounded recommendation of camera view and image composition. The source code, dataset, and pretrained models is available at https://github.com/liuxiaoyu1104/UNIC.

1. Introduction

With the prevalence of electronic devices such as smartphones, taking photos has become a common activity in everyday life. Due to the lack of professional photography knowledge and skills, taking photos with harmonious image composition and high aesthetic quality is still difficult for non-professional users. As a remedy, image composition, which aims to find an aesthetic region of a scene, has attracted much attention in recent years.

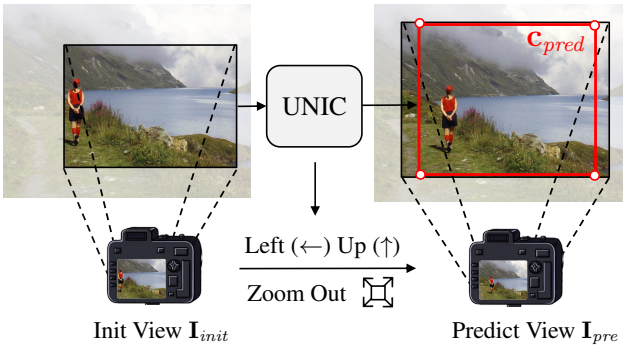

Figure 1. Illustration of our proposed UNIC for unbounded recommendation of camera view and image composition. On the left is the initial view provided by the user. Given the current view, our model can predict camera operations (*e.g.*, zoom out and the movement) and a image composition solution (*e.g.*, $\mathbf{c}_{pred}$). The prediction can be executed multiple times until convergence.

In order to facilitate the training of image composition models, several datasets [33, 38, 6] have been established, and each image comes along with one or multiple bounding boxes indicating the cropping schemes. Despite the notable progress, most existing image composition methods [33, 12, 18, 19, 38, 25, 43, 15] generally adopt a post-processing form on the already captured images, *i.e.*, they only adjust the composition in an image cropping manner. In other words, the captured images are modulated by striking out redundant content near the image borders. Nonetheless, a sub-optimal solution will inevitably be obtained when the best cropping is not entirely in the acquired image. To alleviate this issue, Zhong *et al*. [43] proposed to expand the image via out-painting, and then predict the cropped view on the expanded image. It is a practical solution in post-processing manner, but may suffers from out-painting artifacts.

To tackle the limitations of existing image composi-

tion methods, this paper proposes a novel framework for **un**bounded recommendation of camera view and **i**mage **c**omposition (*i.e*., UNIC). As shown in Fig. 1, the user initializes a view with the content of interest. Given the current view, our model finds a potential well-composed view and provides the corresponding camera movement operations, either inside or beyond the image borders. Note that solely performing camera view adjustment is not enough, since the aspect ratio is typically kept unchanged during the photography process. Therefore, our model also concurrently predicts a bounding box for cropping after camera movement operations. With our model, the user finally can get the most recommended camera view and the corresponding image composition bounding box as shown in Fig. 1.

For implementing the UNIC model, we further simplify the task of joint camera view adjustment and image composition into unbounded image composition by merging the outputs. In this way, the architecture of cropping based image composition methods can be deployed as the backbone, and we follow Jia *et al*. [16] to adopt the conditional-DETR [27] structure. In contrast to existing image cropping methods, we argue that our UNIC is more preferred and practical. First, existing image composition methods [33, 12, 18, 19, 38, 25, 15] are restricted to image cropping over the already captured images. The introduction of camera view adjustment can naturally circumvent the restriction of image borders by moving the camera or adjusting the optical zoom. Second, in comparison to outpainting [43], camera view adjustment can guarantee that the pixels outside the original borders are real and consistent with the pixels within the borders. Furthermore, new and real information can be introduced after each time of view adjustment. Thus, based on the result of last time, one can perform view adjustment and image composition for many times, which is also not supported by image cropping based methods.

Moreover, our UNIC is free to go beyond image borders, yet directly predicting in unseen regions may lead to inferior results. To compensate for this, we further extend the camera field of view by extrapolation. Different from Zhong *et al*. [43], whose extrapolation was performed in the image domain, we choose feature extrapolation and use it for predicting camera movement and bounding box instead of synthesizing unseen content. Thus, we can get the content in novel views by moving the camera, and unseen content generation is not necessary. In comparison to the latent space specified for image composition, forcing the extrapolation into the image domain may bring redundant or even harmful information. Besides, the feature extrapolation module can be well integrated into our existing framework, avoiding the heavy computation burdens brought by extra modules such as the image decoder.

For training and evaluating the proposed model, we take the advantage of existing image cropping datasets [33, 38] and convert them into a more generalized form. Extensive experiments and ablation studies show the effectiveness of our UNIC, which can work well under diverse conditions.

To sum up, the contributions of this paper include,

- We propose a novel UNIC method for jointly performing unbounded recommendation of camera view and image composition. The user can adjust the current view following the recommendations to obtain images with higher aesthetic quality.
- We introduce a feature extrapolation module as well as an extrapolation loss term in the detection transformer framework, which improves the prediction accuracy, especially for out-of-image scenarios.
- Two unbounded image composition datasets are constructed upon existing image cropping ones. Experimental results show that our proposed method achieves superior performance against state-of-the-art methods.

2. Related Work

2.1. Image Composition

Image composition aims to find the most aesthetic photo of a scene, which is typically achieved by image cropping in the literature. Early works rely on saliency detection to preserve important content in the image [2, 9, 3, 31] or extract hand-crafted features from aesthetic characteristics or composition rules for predicting cropping schemes [2, 8, 40, 30, 35, 39]. Recently, a large number of methods address image cropping tasks in a data-driven manner. In general, existing methods can be broadly categorized as two groups, *i.e*., anchor evaluation [7, 33, 38, 43, 20] and cropping coordinate regression [12, 25, 18, 19, 15].

Anchor Evaluation. The general pipeline of anchor evaluation based methods is to generate candidate croppings and then rank different crops to obtain the final result. For example, Chen *et al*. [7] proposed paired ranking constraints to train an aesthetics-aware ranking net. Wei *et al*. [33] predicted scores efficiently by introducing a new knowledge transfer framework. Zeng *et al*. [38] constructed a novel grid anchor based formulation and a corresponding dataset for image cropping. CGS [21] explicitly utilized mutual relations between different candidate regions with a graph-based model. Besides, two tasks closely related to our method are worth mentioning. Zhong *et al*. [43] expanded the range of cropping windows outside the image border through image out-painting. However, the out-painting result may suffer from low visual quality and be inconsistent with the real-world scene. And some methods [28, 24] also tried to provide composition scores for the

current view when photographing with mobile devices, yet they lack the ability to recommend new camera views.

Coordinate Regression. Coordinate regression based methods directly obtain the coordinate of the cropping box. Some works [12, 25] directly designed an end-to-end network to predict the cropping boxes. Regarding image cropping as a consistent decision-making process, Li *et al*. [18, 19] introduced reinforcement learning to generate boxes. Composition rules were explicitly leveraged by Hong *et al*. [15], making the model work like a photographer. Based on object detection method [27], Jia *et al*. [16] predicted multiple crop schemes in a set prediction manner, which took model diversity and globalization into account.

In comparison to the aforementioned cropping based image composition methods, our solution performs unbounded recommendation of camera view and image composition, which can provide more flexibility in searching for better composition schemes.

2.2. Image Out-painting

In this work, we extrapolate the features for better prediction, which is closely related to image out-painting methods. Therefore, we briefly review the progress of out-painting tasks. Inspired by image in-painting methods, Sabini and Rusak [29] introduced the image out-painting task and trained a deep neural network framework adversarially. Wang *et al*. [32] designed an effective deep generative model termed SRN with practical context normalization module for image extrapolation. Some spatial-related loss terms are also proposed to improve the performance. For example, a recurrent content transfer model was proposed for spatial content prediction in NSIPO [36]. Based on StyleGAN2 [17], Zhao *et al*. [42] presented comodulated GANs, which utilized the difference between the unconditional and conditional generative models. Moreover, Ma *et al*. [26] decomposed the image out-painting task into two generation stages, *i.e*., semantic segmentation domain and image domain. More recently, transformer-based networks are incorporated to extend image borders. For example, Gao *et al*. [10] designed a transformer-based generative adversarial network with Swin transformer blocks [22]. QueryOTR [37] proposed a novel hybrid transformer that formulated out-painting problem as a sequence-to-sequence auto-regression problem. In this work, we extrapolate in the feature domain, which shows superior performance against image out-painting for our UNIC.

3. Method

3.1. Problem Definition and Overview

While existing cropping-based image composition methods predict a bounding box for image cropping, we extend the problem to joint unbounded recommendation of camera view and image composition. In specific, the user initializes a camera view $\mathbf{I}_{init}$ with field of view $\mathbf{v}_{init}$[1], which contains the subjects or scenes of interest. Given $\mathbf{I}_{init}$, we predict the actions (*e.g*., zoom in/out, move left/right, move up/down, *etc*.) for obtaining a new camera view $\mathbf{I}_{pred}$ located by $\mathbf{v}_{pred}$. In practice, the view with a high aesthetic score may not share the same aspect ratio as the camera, thus we concurrently predict a bounding box (denoted by $\mathbf{c}_{pred}$) for cropping in the adjusted camera view. With a model $f(\cdot)$, the problem can be formulated by

$$[\mathbf{v}_{pred}, \mathbf{c}_{pred}] = f(\mathbf{I}_{init}), \tag{1}$$

where the difference between $\mathbf{v}_{pred}$ and $\mathbf{v}_{init}$ indicates the camera movement actions.

To maximize the space occupation of $\mathbf{c}_{pred}$ in $\mathbf{v}_{pred}$, we can define the relationship between $\mathbf{v}_{pred}$ and $\mathbf{c}_{pred}$, *i.e*., they share the same center position,

$$(x^{\mathbf{v}}_{pred}, y^{\mathbf{v}}_{pred}) = (x^{\mathbf{c}}_{pred}, y^{\mathbf{c}}_{pred}), \tag{2}$$

and have the same width and/or height,

$$\begin{cases} w^{\mathbf{v}}_{pred} = w^{\mathbf{c}}_{pred}, & w^{\mathbf{c}}_{pred}/h^{\mathbf{c}}_{pred} \geq w^{\mathbf{v}}_{pred}/h^{\mathbf{v}}_{pred} \\ h^{\mathbf{v}}_{pred} = h^{\mathbf{c}}_{pred}, & w^{\mathbf{c}}_{pred}/h^{\mathbf{c}}_{pred} \leq w^{\mathbf{v}}_{pred}/v^{\mathbf{c}}_{pred} \end{cases}, \tag{3}$$

and $\mathbf{v}_{pred}$ will coincide with $\mathbf{c}_{pred}$ when they have the same aspect ratio. Note that the camera view ratio is typically kept unchanged during the photography process, without loss of generality, in this paper we assume the camera view ratio to be $w^{\mathbf{v}} : h^{\mathbf{v}} = 4:3$ or $w^{\mathbf{v}} : h^{\mathbf{v}} = 3:4$. Then given Eqns. (2) and (3), $\mathbf{v}_{pred}$ can be naturally derived from $\mathbf{c}_{pred}$. Thus, we simplify the problem defined in Eqn. (1) as,

$$\mathbf{c}_{pred} = f(\mathbf{I}_{init}), \tag{4}$$

which can also be easily generalized to other camera view ratios or even adjustable ratios.

3.2. Unbounded Regression Model

With the simplified task in Eqn. (4), $f(\cdot)$ can be regarded as a generalized image cropping model which allows the predicted bounding box $\mathbf{c}_{pred}$ to exceed the image borders. As such, we can implement the UNIC model based on existing image cropping models [12, 25, 18, 19, 15]. In particular, Jia *et al*. [16] have successfully applied DETR-like architectures [1, 27] in image cropping tasks, which enables global interactions via the attention mechanism, and the set prediction settings also benefit the diversity of the results. Therefore, we follow Jia *et al*. [16] and adopt conditional-DETR [27] as a base model for implementing $f(\cdot)$.

[1]We represent the position and size of bounding boxes by four values $[x, y, w, h]$, where (x, y) is the center coordinate, while w and h are width and height, respectively. The axes are normalized to $[0, 1]$ w.r.t. $\mathbf{I}_{init}$.

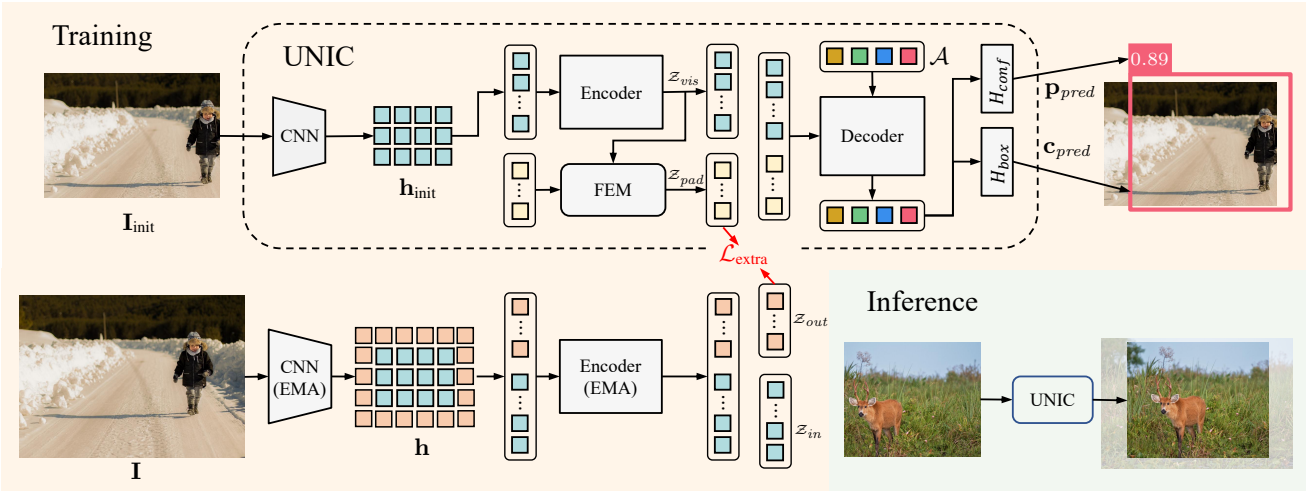

Figure 2. Architecture of the proposed UNIC framework. It adopts a cDETR-like encoder-decoder architecture [27] to predict aesthetic plausible view $\mathbf{c}_{pred}$ from initial view $\mathbf{I}_{init}$. To mitigate the difficulty in predicting $\mathbf{c}_{pred}$ beyond image borders, a feature extrapolation module is deployed to predict the invisible tokens $\mathcal{Z}_{pad}$ from visible ones $\mathcal{Z}_{vis}$. The FEM is supervised by tokens extracted from larger view $\mathbf{I}$ with the exponential moving averaged CNN and encoder during training.

Network Design. In specific, as shown in Fig. 2, following cDETR [27], the base model contains a CNN backbone, a transformer encoder, a transformer decoder, as well as two heads H_{pred} and H_{conf} for predicting candidate bounding boxes and their corresponding confidence, respectively. The initial view $\mathbf{I}_{init}$ is extracted into deep feature $\mathbf{h}_{init}$ with the CNN backbone, which is reorganized into patches. Then the patches with positional embeddings attached according to their spatial positions are processed by the transformer encoder, resulting in a group of latent features denoted by $\mathcal{Z}_{vis}$. Finally, the transformer decoder and two head branches predict candidate image composition results from a group of learnable anchors denoted by $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$. Specifically, the bounding box head H_{box} generates the coordinate of n possible bounding boxes from the anchors (*i.e.*, $\mathbf{c}_{pred}$), and the confidence head H_{conf} predicts the confidence (or possibility) for each candidate bounding box (denoted by $\mathbf{p}_{pred}$).

However, as one can see, $\mathcal{Z}_{vis}$ only contains the feature of visible parts in the range of $\mathbf{v}_{init}$. For improving the prediction accuracy beyond the initial camera view $\mathbf{v}_{init}$, a feature extrapolation module (FEM) is inserted into the base model. The FEM is intended to predict the latent features outside $\mathbf{v}_{init}$, and the padded features are denoted by $\mathcal{Z}_{pad}$. For predicting patches in $\mathcal{Z}_{pad}$, a learnable token $\mathbf{m}$ is fed into the FEM together with the positional embeddings. We will give more details about the FEM in Sec. 3.3.

Model Training. For training the UNIC model, we design the learning objective for composition mainly following Jia *et al.* [16], *i.e.*,

$$\begin{aligned}\mathcal{L}_{\text{comp}} = \mathcal{L}_{\text{reg}}(\mathbf{c}_{pred}, \mathbf{c}) + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}}(\mathbf{c}_{pred}, \mathbf{c}) \\ + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}(\mathbf{p}_{pred}, \mathbf{p}),\end{aligned} \tag{5}$$

where λ_{IoU} and λ_{focal} are hyper-parameters for balancing different loss terms. Note that there may exist multiple ground-truths for each $\mathbf{I}_{init}$, and the number of ground-truths might be different from the number of predicted bounding boxes. Following [16, 27, 1], we find the corresponding ground-truth for the predicted bounding boxes $\mathbf{c}_{pred}$ via bipartite matching. In this way, only the results with a corresponding ground-truth contribute to the regression loss $\mathcal{L}_{\text{reg}}$ and IoU loss $\mathcal{L}_{\text{IoU}}$.

Another key factor is the construction of $\mathbf{p}$. An intuitive way is to assign 1 or 0 according to the existence of ground-truth bounding box for the i-th prediction result. Jia *et al.* [16] further proposed two strategies to generate soft labels for GAICD [38] and CPC [33], respectively. In specific, the one for GAICD [38] is a soft label according to the aesthetic score of the ground-truth (denoted by quality guidance), while for CPC [33] whose labels are more sparse, they use the prediction of the exponential moving averaged model as ground-truth (denoted by self-distillation). In this work, we find that adopting the quality guidance strategy at first can stabilize the training process, and switching to the self-distillation strategy afterward further promotes the performance. More details about the learning objectives are given in the supplementary material.

3.3. Feature Extrapolation Module

To obtain an image composition result that may exceed the range of the initial view, Zhong *et al.* [43] expand the image by out-paining and predict the cropping scheme on the expanded image. However, the out-painting manner may lead to unreal and inconsistent regions in the final image composition results, and there may be redundant or even harmful information in the generated pixels. On the con-

trary, the space of the latent features $\mathcal{Z}_{vis}$ is dedicated to image composition tasks, which motivates us to extrapolate in the feature space of $\mathcal{Z}_{vis}$.

Recent advances in masked image modeling [4, 41, 13, 34, 5] have achieved significant performance in predicting the representations of masked patches from visible parts of an image. Inspired by their architectural design and learning schemes, we build our FEM module by stacking 6 transformer blocks. The visible features $\mathcal{Z}_{vis}$ are involved in the feature extrapolation process through the cross-attention mechanism in the transformer block, and the detailed structure of the FEM is given in the supplementary material.

For training the FEM, it is insufficient if solely relies on the image composition loss $\mathcal{L}_{\text{comp}}$. In order to provide extra supervision for FEM, we leverage the full view image $\mathbf{I}$ (the initial view $\mathbf{I}_{init}$ is extracted from $\mathbf{I}$), and obtain the full view latent features $\mathcal{Z}$ via the CNN backbone and transformer encoder. Then $\mathcal{Z}$ is split into two categories, *i.e.*, $\mathcal{Z}_{in}$ in the range of $\mathbf{I}_{init}$ and $\mathcal{Z}_{out}$ outside $\mathbf{I}_{init}$. As such, we can construct another supervision with $\mathcal{Z}_{out}$ for the extrapolation via FEM, where a robust smooth-ℓ_1 loss [11] is adopted, *i.e.*,

$$\mathcal{L}_{\text{extra}} = \text{smooth-}\ell_1(\mathcal{Z}_{pad}, sg(\mathcal{Z}_{out})), \tag{6}$$

where $sg(\cdot)$ is the stop gradient operator. Note that to improve training stability, the parameters of the CNN backbone and transformer encoder for extracting $\mathcal{Z}$ are from the exponential moving averages (EMA) of the corresponding UNIC parameters. The overall learnng objective for training the UNIC model is defined by,

$$\mathcal{L} = \mathcal{L}_{\text{comp}} + \mathcal{L}_{\text{extra}}. \tag{7}$$

3.4. Unbounded Image Composition Dataset

Even though there are several datasets [33, 38, 6] for image composition tasks, all of them are intended for cropping based image composition tasks, and there is no publicly available dataset for unbounded image composition. To make full use of the aesthetic annotations in existing image cropping datasets, we recreate an unbounded image composition dataset based on GAICD [38] and CPC [33].

In specific, for a sample in image cropping datasets, a full-view image is provided with one or multiple ground-truth bounding boxes. All these ground-truths are located in the range of the full-view image, making it infeasible for unbounded image composition tasks. As a remedy, we randomly sample a bounding box (*i.e.*, $\mathbf{v}_{init}$) in the full-view image, then the ground-truths may not fully lie in the range of $\mathbf{v}_{init}$. In other words, the ground-truths for cropping based image composition are adapted to unbounded image composition tasks.

However, randomly sampling $\mathbf{v}_{init}$ with no constraints may be improper in particular situations. For example, if the interested object is outside of $\mathbf{v}_{init}$, it is unreasonable to require that the predicted bounding box can cover that object. Therefore, we set up some rules as follows when recreating the unbounded image composition dataset.

Size of $\mathbf{I}_{init}$. To ensure the initial camera view contains sufficient image content, we set the lower bound of the height and width of $\mathbf{I}_{init}$ as

$$h^{\mathbf{v}}_{init} \geq \alpha \cdot h \text{ and } w^{\mathbf{v}}_{init} \geq \alpha \cdot w, \tag{8}$$

where h and w denote the height and width of full-view image $\mathbf{I}$, and α is the scale threshold empirically set to 0.7.

Position of $\mathbf{I}_{init}$. Apart from high aesthetic qualities, an important property of the ground-truth bounding boxes is that they well describe the range of desired objects or scenes. To ensure that the initial view contains the desired objects or scenes, we constrain the intersection of union (IoU) of the initial view $\mathbf{v}_{init}$ and the ground-truth $\mathbf{v}$. Specifically, the constraint is defined as,

$$\text{IoU}(\mathbf{v}_{init}, \mathbf{v}) \geq \beta, \tag{9}$$

where the threshold β is set to 0.7 in this paper.

Aspect ratio of $\mathbf{I}_{init}$. Considering that the camera view ratio is typically kept unchanged during the photography process, without loss of generality, we sample $\mathbf{I}_{init}$ with an aspect ratio of $4:3$, which is the most common setting for DLSRs and smartphones.

Since the cameras could take photos vertically or horizontally, we have

$$w^{\mathbf{v}}_{init}:h^{\mathbf{v}}_{init} = 4:3 \text{ or } w^{\mathbf{v}}_{init}:h^{\mathbf{v}}_{init} = 3:4. \tag{10}$$

4. Experiments

4.1. Implementation Details

Datasets. We adopt two widely-used datasets for training, *i.e.*, GAICD [38] and CPC [33]. GAICD [38] is a grid anchor based image cropping dataset, where each image has exhaustive annotations for the cropping candidates. It contains 2,636 images for training, 200 images for validation, and 500 images for testing. We train our model on the training split and evaluate it on the testing split. CPC [33] dataset is sparsely annotated for training purposes only, which contains 10,800 images with 24 annotated views per image. We evaluate our model trained with CPC [33] on FLMS dataset [9] following [16]. Both datasets are pre-processed for unbounded image composition as shown in Sec. 3.4.

Evaluation metrics. The camera view recommendation accuracy could be measured with the intersection of union (IoU) and boundary displacement (Disp) between the predicted view and the ground-truth view with the highest aesthetic score following [33]. However, there may exist multiple human-annotated ground-truth bounding boxes in each

Table 1. Quantitative comparison for unbounded image composition on GAICD [38] and FLMS [9] datasets. The best results are highlighted with **bold**. The results marked by † and ‡ are retrained with our data or reproduced in our framework, respectively.

Method	GAICD						FLMS	
	$Acc_{1/5}$		$Acc_{1/10}$		IoU ↑	Disp ↓	IoU ↑	Disp ↓
	ϵ =0.90	ϵ =0.85	ϵ =0.90	ϵ =0.85				
VFN [7]	0.6	5.2	1.7	9.5	0.577	0.124	0.622	0.122
VEN [33]	2.6	8.9	3.4	11.5	0.600	0.095	0.688	0.065
GAIC [38]	7.2	21.8	10.6	31.5	0.683	0.074	0.723	0.060
CGS [21]	7.2	25.8	10.9	33.5	0.682	0.074	0.703	0.064
A2-RL [18]	6.9	22.9	11.2	34.7	0.686	0.076	0.731	0.059
CACNet† [15]	16.9	49.1	25.8	60.7	0.779	0.052	0.813	0.044
Zhong *et al.* ‡ [43]	22.3	53.5	28.7	67.2	0.795	0.050	0.818	0.042
Jia *et al.* † [16]	21.4	48.0	26.8	57.2	0.786	0.052	0.817	0.042
Ours	**23.2**	**59.0**	**32.7**	**72.8**	**0.801**	**0.048**	**0.828**	**0.040**

image, these metrics ignore such situations, which limits their flexibility. As a remedy, $Acc_{K/N}$ calculates how many of K predicted views falls into the N ground-truth views with highest score. Therefore, we adopt $Acc_{K/N}$ as another evaluation metric for grid annotated GAICD [38]. As the predicted views may not align perfectly with the predefined grid views, we follow [16] and regard two crops the same when their IoU is sufficiently large. Two thresholds $\epsilon = \{0.9, 0.85\}$ are used in this paper. For FLMS dataset [9] without grid annotation, we use IoU and Disp metrics.

Training details. The amount of data in the image cropping datasets is not sufficiently enough for training DETR-like models from scratch. Thus, we initialize the CNN backbone with ImageNet pre-trained weights [14]. The layer numbers of the transformer encoder and decoder are both set to 6. During training, we take views with an aesthetic score larger than 4 in GAICD [38] and that larger than 2 in CPC dataset [33] as ground-truth views. The trade-off parameters λ_{IoU} and λ_{focal} are set to 0.4 and 0.1, respectively. The model is trained with an AdamW [23] optimizer with weight decay of 1×10^{-4} for 50 epochs. The initial learning rates for the CNN backbone and the transformer encoder/decoder are set to 1×10^{-5} and 1×10^{-4}, which are decreased to 1/10 at the 30-th epoch. We apply data augmentation via color jittering and resizing following [27].

4.2. Results of Unbounded Image Composition

Due to the lack of competing methods for unbounded image composition, we adopt several state-of-the-art image cropping methods with source code publicly available, including anchor evaluation based methods, *i.e.*, VFN [7], VEN [33], GAIC [38], and CGS [21], as well as coordinate regression based methods, *i.e.*, A2-RL [18], CACNet [15], and Jia *et al.* [16]. Among them, the anchor evaluation based methods require the cropped image for scoring. Directly applying them for unbounded image composition tasks will lead to poor results due to the incomplete image for views exceeding the initial view borders. Therefore, we show the results of cropping based image composition for these anchor evaluation based methods. As for coordinate regression based methods, we show cropping based results of A2-RL [18] since it is based on VFN [7], and retrain CACNet [15] and the method of Jia *et al.* [16] with our training data. Zhong *et al.* [43] can predict cropping schemes via image extrapolation, which is the most similar method to our UNIC. Since the source code is unavailable, we reimplement their method in our framework, where the extrapolation module is replaced by a StyleGAN2 [17] based image out-painting model [42].

Quantitative comparison. We conduct comprehensive experiments to assess the effectiveness of the proposed method, and the quantitative results are shown in Tab. 1. From the $Acc_{1/5}$ and $Acc_{1/10}$ metrics with two IoU thresholds $\epsilon \in \{0.90, 0.85\}$ on GAICD [38], we can see that anchor evaluation based methods [7, 33, 38, 21] are limited by the border of the current view. Regression based methods [18, 15, 16] show inferior results as they are not properly designed for unbounded image composition tasks. Our method instead shows significant improvement for unbounded image composition tasks compared to the previous methods. The IoU and Disp metrics in GAICD [38] and FLMS [9] datasets also demonstrate the effectiveness of the proposed UNIC model.

Qualitative comparison. The qualitative results of competing methods are shown in Fig. 3. Anchor evaluation methods [7, 38, 21] are restricted by the boundary of the initial view, which cannot adjust the camera toward a larger view and show inferior results. After training with our dataset, the regression based methods [16] could predict outward views, but due to the implicit regression from inbound contents, their accuracy in the out-of-border regions is also limited. For [43], the extrapolated regions may be included in the cropped image, which harms the aesthetic quality such as the abnormal arm in the first row and the artifacts near

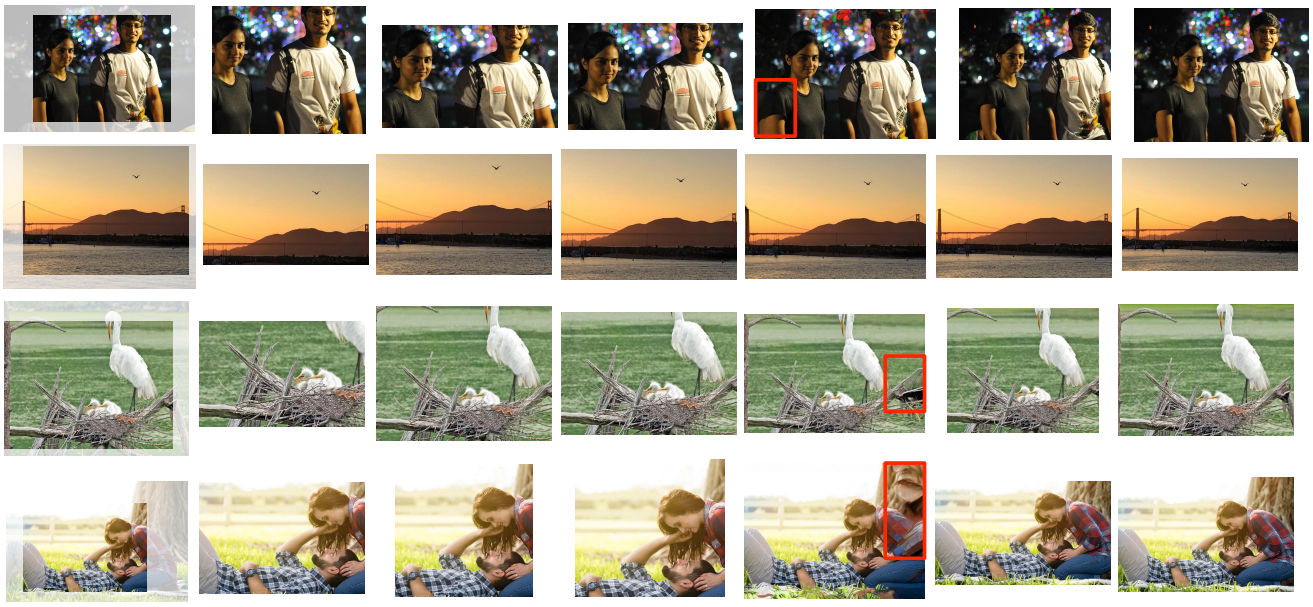

Figure 3. Qualitative comparison with other methods. Our method goes beyond the border of the image to predict a well-composed region with the main objects in reasonable places.

Table 2. Quantitative comparison for image cropping on GAICD [38] and FLMS [9] datasets. $Acc_{1/5}$ and $Acc_{1/10}$ are calculated according to [16].

Method	GAICD		FLMS	
	$Acc_{1/5}$	$Acc_{1/10}$	IoU ↑	Disp ↓
VFN [7]	26.6	40.6	0.577	0.124
VEN [33]	37.5	48.5	0.837	0.041
GAIC [38]	68.2	85.8	0.834	0.041
CGS [21]	63.0	81.5	0.836	0.039
A2-RL [18] (ϵ=0.90)	7.6	12.6	0.821	0.045
A2-RL [18] (ϵ=0.85)	28.6	43.2		
CACNet [15] (ϵ=0.90)	50.7	66.0	**0.854**	**0.033**
CACNet [15] (ϵ=0.85)	78.0	89.3		
Jia *et al.* [16] (ϵ=0.90)	72.0	86.0	0.838	0.037
Jia *et al.* [16] (ϵ=0.85)	85.0	92.6		
Ours (ϵ=0.90)	**74.7**	**89.6**	0.840	0.037
Ours (ϵ=0.85)	**87.2**	**95.5**		

the woman in the fourth row. In contrast, our method not only learns to predict beyond image borders, but also predicts a more accurate and aesthetically pleasurable view by extrapolation in the feature space. More qualitative results are given in the supplementary material.

4.3. Results for Image Cropping

Although our UNIC is delicately designed for unbounded image composition tasks, it can degrade to an image cropping model with the absence of the FEM. Tab. 2 shows the results for image cropping task on the original GAICD [38] and FLMS [9] datasets. One can see that our method outperforms all existing methods that are specifically designed for image cropping tasks on the GAICD [38] dataset and achieves comparable performance to the state-of-the-art methods on the FLMS [9] dataset, showing the effectiveness of the proposed UNIC framework.

Table 3. Ablation study on the extrapolation (Extra.) strategies. Extrapolation in the feature space achieves the best results.

Method	$Acc_{1/5}$ ($\epsilon = 0.90$)	$Acc_{1/5}$ ($\epsilon = 0.85$)
Ours w/o Extra.	22.6	48.1
Ours w/ SRN [32]	19.9	51.0
Zhong *et al.* [43]	22.3	53.5
Ours w/ QueryOTR [37]	23.1	53.7
Ours w/ Feature Extra.	**23.2**	**59.0**

4.4. Ablation Study

Effects of extrapolation strategy. As illustrated in Sec. 3.3, we apply extrapolation in the feature space to boost the performance of our unbounded regression model. In this subsection, we make detailed experiments to assess the effects of different extrapolation strategies, *e.g.*, no extrapolation, image extrapolation, and feature extrapolation. We take the UNIC without FEM as the model with no extrapolation, and several state-of-the-art out-painting methods [32, 42, 37] are applied to the input image for evaluating the image-level extrapolation. As shown in Tab. 3, image-level extrapolation may suffer from generative arti-

Table 4. Ablation study on the $\mathcal{L}_{\text{extra}}$ for feature extrapolation.

Type	$Acc_{1/5}$ ($\epsilon = 0.90$)	$Acc_{1/5}$ ($\epsilon = 0.85$)
MSE	22.1	56.1
Cosine Distance	23.1	57.9
KL-Divergence	**23.8**	56.4
Smooth-$\mathcal{L}_1$	23.2	**59.0**

Table 5. Ablation study on multi-step adjustment.

	step=1	step=2	step=3
$Acc_{1/5}$ ($\epsilon = 0.90$)	16.9	19.3	**19.3**
$Acc_{1/5}$ ($\epsilon = 0.85$)	48.2	51.8	**54.2**

facts due to the extrapolation model, as our model with SRN [32] exhibits a performance drop in $Acc_{1/5}$ ($\epsilon = 0.90$) and limited improvement on $Acc_{1/5}$ ($\epsilon = 0.85$). With more powerful generative models [17, 37], image-level extrapolation shows consistent improvement. Nonetheless, our model with feature extrapolation benefits from recent advances in mask image modeling [4] and end-to-end training, which shows the best results. It achieves a 2.5% improvement on $Acc_{1/5}$ ($\epsilon = 0.90$) and a 22.6% improvement on $Acc_{1/5}$ ($\epsilon = 0.85$) over the base model, which demonstrates the effectiveness of extrapolation in the feature space. More analysis and visual results are provided in the supplementary material.

Effects of FEM loss. In order to assess the effects on the loss function of the FEM for feature extrapolation, we experiment on several commonly used loss functions for regression, *e.g.*, mean square error (MSE), cosine distance, KL-divergence, and smooth-ℓ_1. As shown in Tab. 4, smooth-ℓ_1 yields the best overall performance. Thus we choose smooth-ℓ_1 loss for our FEM in this paper.

Effects of multi-step adjustment. The camera view predicted from our regression model may not be the most aesthetic view with the unseen regions, but is expected to move closer toward it. Based on the above idea, the camera view could be further improved with multi-step adjustment. Concretely, we first apply our model to the initial view captured by the camera, predict camera operations, and perform adjustments. Then the same process is applied on the new view after adjustment, which could be operated multiple times. Tab. 5 shows the results of multi-step adjustment. We note that we use images in GAICD [38] as the whole scene and a crop within the images as the initial view in our experiment. Multi-step adjustment may exceed the border of the scene, so we select 83 large images from GAICD [38] to avoid this problem, thus the results in Tab. 5 are not consistent with other tables in the paper. From the table, the performance is promoted with increased adjustment steps, which demonstrates the effectiveness of the multi-step adjustment. As shown in Fig. 4, our model predicts better views with increased adjustment steps to approach the ground-truth view circled by the red box.

Figure 4. Visual comparison for multi-step adjustment. Our model predicts better view with increased adjust steps to approach the ground-truth view within the red box.

Figure 5. Visualization of failure cases.

5. Limitation and Future Work

Although UNIC predicts well-composed views in most scenarios, it may encounter failure in certain circumstances. As shown in the left of Fig. 5, an unexpected people near the border appears in the predicted view, which is unseen in the initial camera view and may the affect the aesthetics quality of the predicted view. This could be addressed with multi-step adjustment as the predicted view becomes stable. The right example shows the camera view adjustment operations are limited to shifting and zooming in or out in this paper, it's hard to adjust the camera view without camera pose adjustment. Besides, more scene information (*e.g.*, depth) could be leveraged for better camera view recommendation. We leave these problems as future work.

6. Conclusion

In this paper, we propose a novel framework for **UN**bounded **I**mage **C**omposition, *i.e.*, UNIC. Different from previous image cropping methods that improve the composition in a post-process manner, UNIC provides recommendations for camera view adjustment during photographing. To improve the model accuracy beyond borders, we introduce a feature extrapolation module based on recent advances in mask image modeling. To assist the model training and evaluation, we construct unbounded image composition datasets based on existing image cropping ones. Extensive experiments demonstrate that our UNIC achieves better performance against the state-of-the-art methods in both image cropping and unbounded image composition tasks.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2022YFA1004103 and the National Natural Science Foundation of China (NSFC) under Grant No. U19A2073.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020. 3, 4

[2] Yuan-Yang Chang and Hwann-Tzong Chen. Finding good composition in panoramic scenes. In *IEEE International Conference on Computer Vision*, pages 2225–2231, 2009. 2

[3] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 507–515, 2016. 2

[4] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 5, 8

[5] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124, 2022. 5

[6] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 226–234, 2017. 1, 5

[7] 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*, pages 37–45, 2017. 2, 6, 7

[8] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. Learning to photograph. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 291–300, 2010. 2

[9] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1105–1108, 2014. 2, 5, 6, 7

[10] Penglei Gao, Xi Yang, Rui Zhang, Kaizhu Huang, and Yujie Geng. Generalised image outpainting with u-transformer. *arXiv preprint arXiv:2201.11403*, 2022. 3

[11] Ross Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 5

[12] Guanjun Guo, Hanzi Wang, Chunhua Shen, Yan Yan, and Hong-Yuan Mark Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Transactions on Multimedia*, pages 2073–2085, 2018. 1, 2, 3

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[15] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7057–7066, 2021. 1, 2, 3, 6, 7

[16] Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2446–2455, 2022. 2, 3, 4, 5, 6, 7

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 3, 6, 8

[18] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8193–8201, 2018. 1, 2, 3, 6, 7

[19] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *IEEE Transactions on Image Processing*, pages 5105–5120, 2019. 1, 2, 3

[20] Debang Li, Junge Zhang, and Kaiqi Huang. Learning to learn cropping models for different aspect ratio requirements. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12685–12694, 2020. 2

[21] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4213–4222, 2020. 2, 6, 7

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Hao Lou, Heng Huang, Chaoen Xiao, and Xin Jin. Aesthetic evaluation and guidance for mobile photography. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2780–2782, 2021. 2

[25] Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin. An end-to-end neural network for image cropping by learning composition from aesthetic photos. *arXiv preprint arXiv:1907.01432*, 2019. 1, 2, 3

[26] Ye Ma, Jin Ma, Min Zhou, Quan Chen, Tiezheng Ge, Yuning Jiang, and Tong Lin. Boosting image outpainting with se-

mantic layout prediction. *arXiv preprint arXiv:2110.09267*, 2021. 3

[27] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *IEEE International Conference on Computer Vision*, pages 3651–3660, 2021. 2, 3, 4, 6

[28] Yogesh Singh Rawat and Mohan S Kankanhalli. Context-aware photography learning for smart mobile devices. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, pages 1–24, 2015. 2

[29] Mark Sabini and Gili Rusak. Painting outside the box: Image outpainting with gans. *arXiv preprint arXiv:1808.08483*, 2018. 3

[30] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H Hsu, and Shao-Yi Chien. Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia*, 14(3):833–843, 2012. 2

[31] Jin Sun and Haibin Ling. Scale and object aware image thumbnailing. *International journal of computer vision*, pages 135–153, 2013. 2

[32] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019. 3, 7, 8

[33] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2018. 1, 2, 4, 5, 6, 7

[34] Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at what you see: Masked image modeling without reconstruction. *arXiv preprint arXiv:2211.08887*, 2022. 5

[35] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–978, 2013. 2

[36] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *IEEE International Conference on Computer Vision*, pages 10561–10570, 2019. 3

[37] Kai Yao, Penglei Gao, Xi Yang, Jie Sun, Rui Zhang, and Kaizhu Huang. Outpainting by queries. In *European Conference on Computer Vision*, 2022. 3, 7, 8

[38] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5949–5957, 2019. 1, 2, 4, 5, 6, 7, 8

[39] Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and Nicu Sebe. Weakly supervised photo cropping. *IEEE Transactions on Multimedia*, pages 94–107, 2013. 2

[40] Luming Zhang, Mingli Song, Qi Zhao, Xiao Liu, Jiajun Bu, and Chun Chen. Probabilistic graphlet transfer for photo cropping. *IEEE Transactions on Image Processing*, pages 802–815, 2012. 2

[41] Xinyu Zhang, Jiahui Chen, Junkun Yuan, Qiang Chen, Jian Wang, Xiaodi Wang, Shumin Han, Xiaokang Chen, Jimin Pi, Kun Yao, et al. Cae v2: Context autoencoder with clip target. *arXiv preprint arXiv:2211.09799*, 2022. 5

[42] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 3, 6, 7

[43] 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*, pages 1–13, 2021. 1, 2, 4, 6, 7