Beyond Image Borders: Learning Feature Extrapolation for Unbounded Image Composition

(Supplementary Material)

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

The content of this supplementary material is organized as follows:

- Difference between our proposed UNIC model and Zhong *et al.* [10] in Sec. s1.
- Details of the model architecture in Sec. s2.
- Details of the learning objective in Sec. s3.
- Details of our proposed datasets for unbounded image composition in Sec. s4.
- Details of the evaluation metrics in Sec. s5.
- Additional qualitative results Sec. s6.

s1. Difference with Zhong *et al.* [10]

Since both our UNIC model and Zhong *et al.* [10] can give image composition results not fully lie in the camera view, we compare our UNIC with Zhong *et al.* [10] to show the difference. The working schemes of the two methods are given in Fig. s1.

Suitable Scenarios. As one can see, Zhong *et al.* [10] is designed to improve the composition of already taken images. On the contrary, our approach provides recommendations to adjust the camera for obtaining a new camera view, which is suitable during the photography process.

Authenticity. Since Zhong *et al.* [10] extrapolate the given image via out-painting methods and crop in the extrapolated image, the results may contain unrealistic out-painted areas, which affects the image quality of the final results. Our method instead guarantees that the final results are real images, and there are no concerns about authenticity.

Flexibility. Zhong *et al.* [10] highly rely on the already captured image, therefore the solution is greatly limited by the image out-painting methods. Executing their method multiple times will lead to a certain result, leaving no chance to adjust the image composition by users. Unlike their solution, our UNIC can be executed iteratively to obtain better results. Besides, the users can get involved in the photography process naturally.

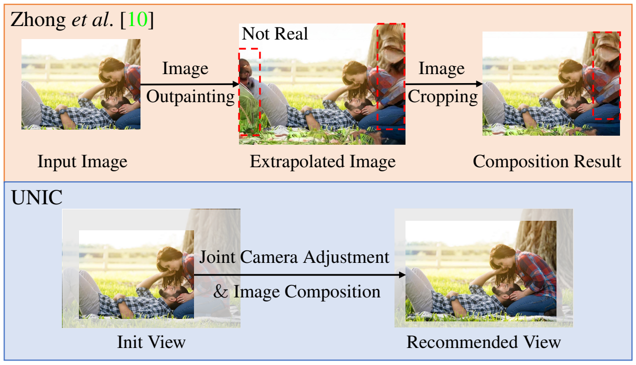

Figure s1. Pipeline of Zhong *et al.* [10] and Our UNIC. The extrapolated areas by Zhong *et al.* [10] are not real, where obvious artifacts can be observed. By contrast, our UNIC jointly performs camera adjustment and image composition.

s2. Model Architecture

Our UNIC contains a CNN backbone, a transformer encoder, a transformer decoder, and a feature extrapolation module. The details of the architecture are given as follows.

CNN Backbone. Following [1, 7], we use ResNet-50 [4] as the backbone network. For the initial image $\mathbf{I}_{init} \in \mathbb{R}^{3\times H_0\times W_0}$, we only use the last-layer feature $\mathbf{h}_{init}$ extracted by backbone, where $\mathbf{h}_{init} \in \mathbb{R}^{C\times H\times W}$, $H = \frac{H_0}{32}$, $W = \frac{W_0}{32}$, $C = 2048$.

Transformer Encoder. We first employ a 1×1 convolution to reduce the number of channels of the feature from 2048 to 256. The image features with encoded positional embeddings are rearranged into a sequence of feature tokens that can be fed into the encoder. The encoder is composed of

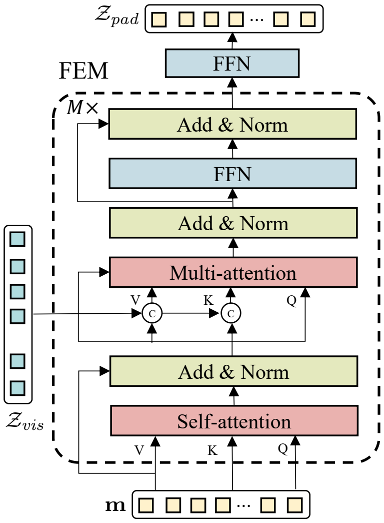

Figure s2. Illustrating the first layer of FEM. Later layers take the output of the previous layer as input.

six standard transform blocks including a multi-head self-attention module and a feed-forward network (FFN), where the query and key are provided by different feature tokens.

Feature Extrapolation Module (FEM). The FEM is a stack of decoder layers to predict the padded features conditioned on visible ones. The query is initialized by a learnable variable $\mathbf{m}$ and is transformed into padded features $\mathcal{Z}_{pad}$ by multiple decoder layers and a feed-forward network. The detailed structure of the decoder layer of FEM is illustrated in Fig. s2. In each decoder layer, apart from self-attention between different padded feature embeddings, we also calculate cross-attention between visible features and padded features to utilize the visible information to learn feature extrapolation. The key and value for cross attention consist of visible features $\mathcal{Z}_{vis}$ and the output of the self-attention layer.

Transformer Decoder. A group of learnable anchors are transformed into output embeddings by the decoder and are futher converted into bounding boxes and their corresponding confidence labels by two heads. The decoder is standard architecture of the transformer, which perform through multiple multihead self-attention between different decoder embeddings to remove redundant boxes and and cross attention aggregating image information to predict the boxes.

s3. Learning Objective

Unbounded Regression on Set Prediction. Our network predicts a set of N bounding boxes, but the number of predicted boxes is not consistent with the number of ground truth views N^{gt} (typically $N^{gt} \ll N$). Following [1, 5, 7], we use set prediction to compute the loss in a reasonable way. First, we pad the ground truth set of views with $\varnothing$ (invalid view) to a set of size N as follows:

$$\begin{cases} \{\mathbf{c}^i = [x, y, w, h], \mathbf{p}^i = 1\}, & 1 \leq i \leq N^{gt} \\ \{\mathbf{c}^i = \varnothing, \mathbf{p}^i = 0\}, & N^{gt}+1 \leq i \leq N \end{cases}, \tag{s1}$$

then we use a bipartite matching to find an one-to-one index mapping $\sigma \in \mathfrak{S}_N$ for these two set to minimize the matching cost $\mathcal{L}_{\text{comp}}$ which is the same as the loss function,[1]

$$\sigma^* = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}_{\text{comp}} \left(\{\mathbf{c}_{pred}^{\sigma(i)}, \mathbf{p}_{pred}^{\sigma(i)}\}, \{\mathbf{c}^i, \mathbf{p}^i\} \right). \tag{s2}$$

Those predicted views having a matching with ground-truth valid views (*i.e.*, $\mathbf{c}_i \neq \varnothing$) contribute to the regression loss, IoU loss, and focal loss (with $\mathbf{p}_i = 1$). For other views without a valid matching (*i.e.*, $\mathbf{c}_i = \varnothing$), they only contribute to the focal loss (with $\mathbf{p}_i = 0$). With the above definitions, the loss function of our UNIC can be rewritten as

$$\begin{aligned} \mathcal{L}_{\text{comp}} = \sum_{i=1}^{N} & f_{\text{bool}}(\mathbf{c}_i \neq \varnothing) \mathcal{L}_{\text{reg}}(\mathbf{c}_{pred}^{\sigma^*(i)}, \mathbf{c}^i) + \\ & f_{\text{bool}}(\mathbf{c}_i \neq \varnothing) \lambda_{\text{IoU}} \mathcal{L}_{\text{IoU}}(\mathbf{c}_{pred}^{\sigma^*(i)}, \mathbf{c}^i) + \\ & \lambda_{\text{focal}} \mathcal{L}_{\text{focal}}(\mathbf{p}_{pred}^{\sigma^*(i)}, \mathbf{p}^i), \end{aligned} \tag{s3}$$

where $f_{\text{bool}}(*)$ equals to 1 when the condition $*$ is satisfied otherwise 0.

Smooth Label. It is worth noting that N^{gt} is typically much smaller than N, therefore, the invalid views whose $\mathbf{p}_i = 1$ have major contributions to $\mathcal{L}_{\text{focal}}$ in Eqn. (s3). To remedy this problem, Jia *et al.* [5] have proposed two strategies to smooth the labels for invalid views, *i.e.*, quality guidance and self-distillation. The quality guidance strategy is more feasible for densely annotated datasets (*e.g.*, GAICD [9]), which uses the annotated quality score to get the smooth labels of invalid views. For a predicted invalid view, we calculate the IoU between the view and all labeled views, and regard the quality score of maximum-IoU neighbor view as the quality score of the predicted box. Then we can map the quality score to the soft labels by a linear function. As for the self-distillation strategy, the smooth labels are generated by the model whose parameters are exponential moving averages (EMA) of the model parameters. Jia *et al.* [5] use these strategies for GAICD and CPC, respectively.

In this paper, we argue that, at the beginning of the training process, the model is not well trained, thus the self-distillation strategy is unable to generate stable smooth labels. On the contrary, at the later phase of the training process, most predicted views have decent quality but are unable to match with limited ground-truths, forcing $\mathbf{p}_i$ to 0 or

[1] More intuitively, we can regard σ as a shuffle operation, and $\mathfrak{S}_N$ is all shuffle solutions. We aim to find a shuffle solution σ^* such that the loss $\mathcal{L}_{\text{comp}}$ is minimized.

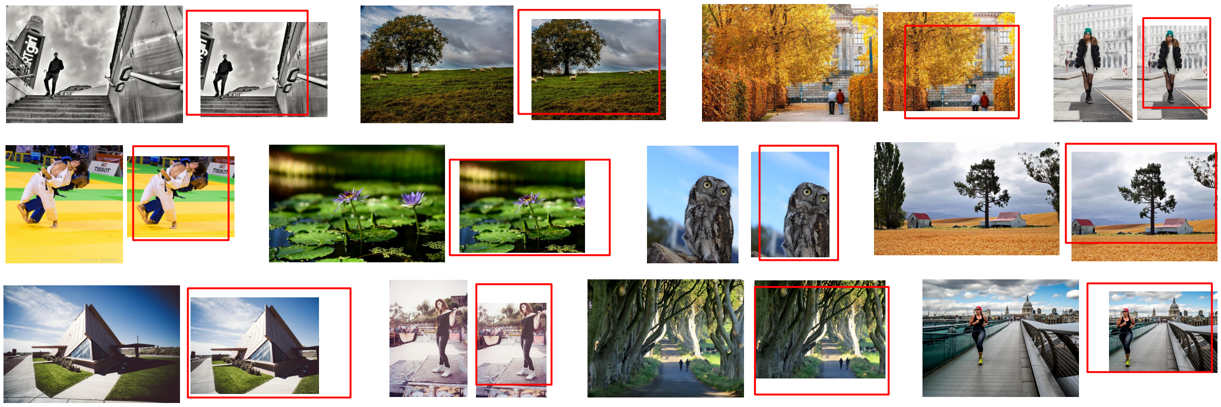

Figure s3. Some sample images from our datasets. The left is the full-view image from the GAICD [9], and the right is the image from our dataset. The ground-truth views are circled by the red box

a manually set label is unreasonable. Therefore, we propose to combine these two strategies based on our observations, and adopt the quality guidance strategy at first, then switch to the self-distillation strategy afterward.

Loss Terms. We employ the learning objective for composition including a regression loss $\mathcal{L}_{\text{reg}}$ and a generalized IoU loss $\mathcal{L}_{\text{IoU}}$ for bounding box regression, and a focal loss $\mathcal{L}_{\text{focal}}$ for predicting the confidence.

$$\begin{aligned}\mathcal{L}_{\text{comp}} &= \mathcal{L}_{\text{reg}}(\mathbf{c}_{pred}, \mathbf{c}) + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}}(\mathbf{c}_{pred}, \mathbf{c}) \\ &\quad + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}(\mathbf{p}_{pred}, \mathbf{p}).\end{aligned} \tag{s4}$$

The regression loss is defined as an ℓ_1 loss to supervise $\mathbf{c}_{pred}$, *i.e.*, a four-dimensional vector consisting of the box center coordinates and its height and width, *i.e.*,

$$\mathcal{L}_{\text{reg}}(\mathbf{c}_{pred}, \mathbf{c}) = \|\mathbf{c}_{pred} - \mathbf{c}\|_1, \tag{s5}$$

while the IoU loss is defined by,

$$\mathcal{L}_{\text{IoU}}(\mathbf{c}_{pred}, \mathbf{c}) = 1 - \left(\frac{|\mathbf{c}_{pred} \cap \mathbf{c}|}{|\mathbf{c}_{pred} \cup \mathbf{c}|} - \frac{|\mathbf{c}_c - \mathbf{c}_{pred} \cup \mathbf{c}|}{|\mathbf{c}_c|}\right), \tag{s6}$$

where $|\cdot|$ means area, and $\mathbf{c}_c$ is the smallest enclosing convex object for $\mathbf{c}_{pred}$ and $\mathbf{c}$. And the focal loss is,

$$\begin{aligned}&\mathcal{L}_{\text{focal}}(\mathbf{p}_{pred}, \mathbf{p}) = -|\mathbf{p}_{pred} - \mathbf{p}|^{\beta} \\ &((1 - \mathbf{p}_{pred})\log(1 - \mathbf{p}) + \mathbf{p}_{pred}\log(\mathbf{p})),\end{aligned} \tag{s7}$$

where $\beta = 2$.

s4. Unbounded Image Composition Dataset

We construct the unbounded image composition (UIC) dataset based on the existing datasets (*i.e.*, GAICD [9], CPC [8] & FLMS [3]). In our UIC dataset, the ground-truth composition may not fully lie in the range of the image, and the rules for generating the dataset are given in the main manuscript. Some examples of the UIC dataset are shown in Fig. s3.

s5. Evaluation Metrics

In this section, we provide the formulations of the evaluation metrics used in our paper, *i.e.*, $Acc_{K/N}$, IoU, and Disp.

$Acc_{K/N}$ is the accuracy (or ratio) of predicted K views fall into the N ground-truth views. In specific, for image i, we define the set of annotated views with the top N of quality score as $C_N^i = \{\mathbf{c}^{i1}, \ldots, \mathbf{c}^{ij}, \ldots, \mathbf{c}^{iN}\}$. And a model returns K views with the highest confidence scores denoted by $\mathbf{c}^{ik}_{pred}$. Then $Acc_{K/N}$ can be defined by

$$Acc_{K/N} = \frac{1}{TK}\sum_{i=1}^{T}\sum_{k=1}^{K} f_{\text{bool}}\Big(\max_{\mathbf{c}^{ij}\in B_N^i} \left\{\text{IoU}\left(\mathbf{c}^{ik}_{pred}, \mathbf{c}^{ij}\right)\right\} \geq \epsilon\Big), \tag{s8}$$

where $\epsilon \in \{0.90, 0.85\}$, $f_{\text{bool}}(*)$ equals to 1 when the condition $*$ is satisfied otherwise 0.

The boundary displacement error (Disp.) is defined by

$$\text{Disp} = \frac{1}{4}\sum_{j} \|\mathbf{b}^{j}_{pred} - \mathbf{b}^{j}\|_1 \tag{s9}$$

where $\mathbf{b}$ denotes all boundaries of the bounding box, and $\mathbf{b}_j$ is the normalized coordinate of that boundary (*e.g.*, the x-axis of the left and right boundary). Finally, the IoU is defined by,

$$\text{IoU} = \frac{|\mathbf{c}_{pred} \cap \mathbf{c}|}{|\mathbf{c}_{pred} \cup \mathbf{c}|}. \tag{s10}$$

s6. Additional qualitative results.

In Fig. s4, we present the visualization comparison with other existing cropping methods on FLMS [3] dataset. Meanwhile, we give additional visualization results in Fig. s5 on GAICD [9] dataset. And it can be seen that our method usually produces more appealing results by unbounded image composition.

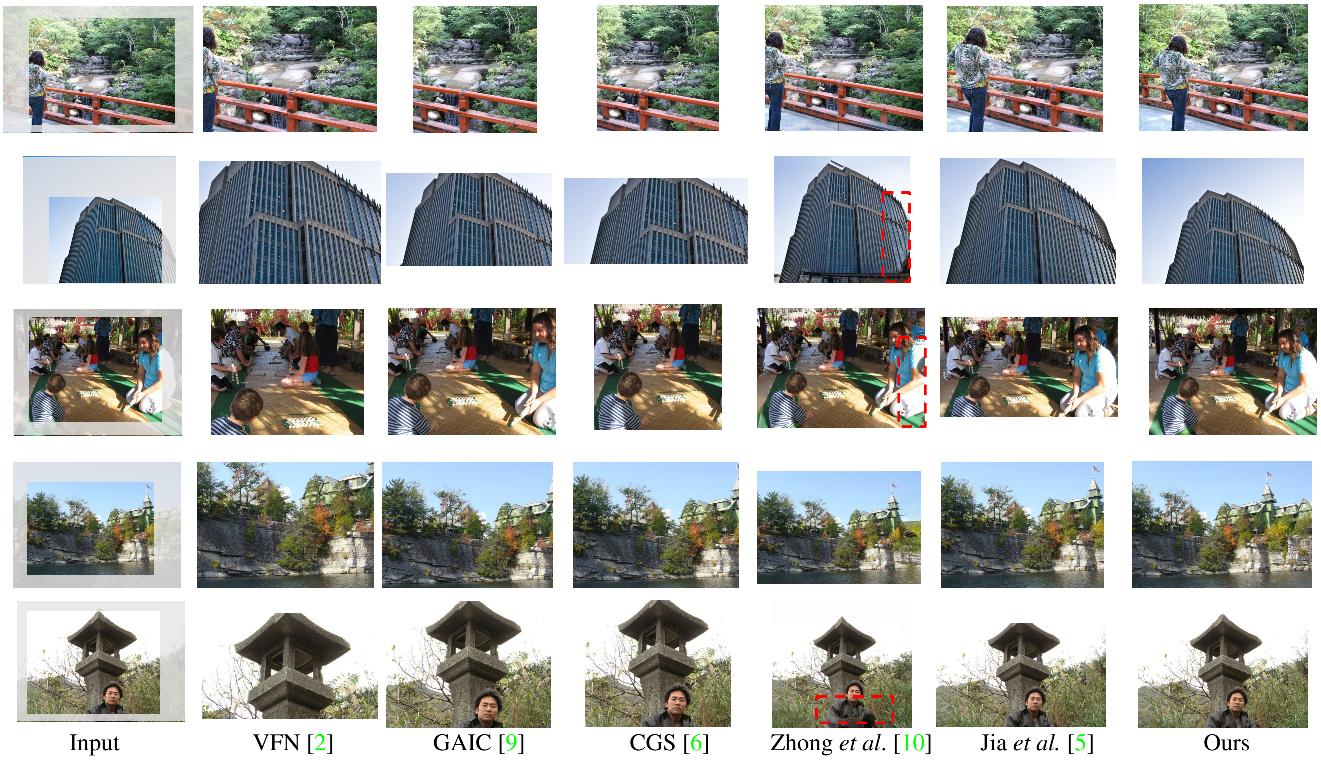

Figure s4. Qualitative comparison with other methods on FLMS [3] dataset.

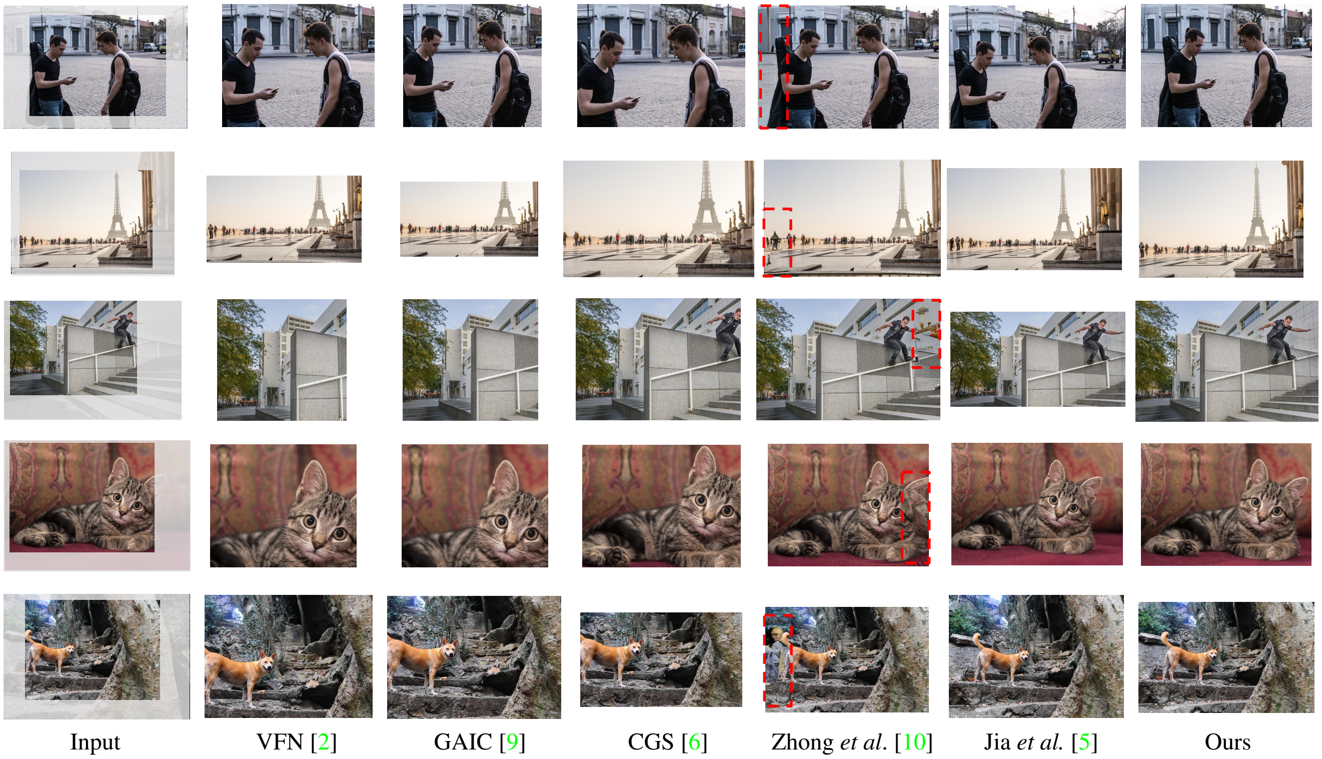

Figure s5. Qualitative comparison with other methods on GAICD [9] dataset.

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. 1, 2

[2] 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. 4

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

[4] 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. 1

[5] 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, 4

[6] 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. 4

[7] 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. 1, 2

[8] 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. 3

[9] 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. 2, 3, 4

[10] 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, 4