Supplementary for Human-centric Image Cropping with Partition-aware and Content-preserving Features

Bo Zhang, Li Niu*, Xing Zhao, and Liqing Zhang

MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, China
{bo-zhang,ustcnewly,1033874657}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

In this document, we provide additional materials to supplement our main submission. We first present more implementation details in Section 1 and describe how to determine the main subject for human-centric image in Section 2. Next, we report the performance of our model using inaccurate human bounding box in Section 3 and evaluate two simplest baselines that use human bounding box to directly produce cropping box in Section 4. In Section 5, we study the impact of using different hyper-parameters in our method. Then, we present the results of evaluating our method on general image cropping and compare with the state-of-the-arts in Section 6, and compare the running speed and model complexity of different models in Section 7. Moreover, we show more qualitative results in Section 8.

1 Implementation Details

We conduct experiments on Ubuntu 18.04 equipped with an NVIDIA RTX 3090 GPU and 64GB RAM. We use the Adam optimizer [6] with a weight decay of $1e^{-4}$ to train the model for 80 epochs. We randomly select 64 crops for one image when training on GAICD dataset [15], while we use all the crops of an image when training on CPC dataset [14]. Following [8], we use the warmup [4] in the first five epochs to increase the learning rate from 0 to $1e^{-4}$ and decay the learning rate with a cosine annealing [9] in the following epochs.

2 Main Subject Determination

Our method is proposed to handle images with a single person and considers the other people as background when multiple subjects exist. In our implementation, given multiple human objects in an image, we take the person that appears in the ground-truth best crop of the image as the main subject. In real-world applications, we can adopt a simple approach to determine the main subject, by selecting the object with a larger area and closer to the image center through weighted ranking. Specifically, the rank score of a human object is defined as: $w*h+\alpha\big(1-\sqrt{(x-0.5)^2+(y-0.5)^2}\big)$ and we take the object with largest rank

* Corresponding author

score as the main subject. Here, given a human bounding box, w and h are its width and height normalized by the size of source image, respectively, and (x, y) indicates the normalized coordinates of the bounding box center. The weight α is set as 0.1 empirically. To verify the utility of this simple approach, we first select all images with more than one detected human objects from GAICD [15] and CPC [14] datasets, in which we obtain 60 and 216 samples, respectively. Then we use above approach to determine the main subject and compare it with the manually selected results. As a result, this approach achieves an accuracy of 98.0% on GAICD dataset and 97.6% on CPC dataset, demonstrating its utility in determining the main subject for human-centric images.

3 Performance using Inaccurate Human Bounding Box

Recall the proposed partition-aware feature depends on the human bounding box detected by Faster R-CNN [11] trained on Visual Genome [7], which can generally provide reliable human detection results. Nevertheless, the predicted human bounding box may still be inaccurate or even wrong. For quantitative study on the performance of human detection, we first define a detected bounding box, whose intersection over union (IoU) with the manually checked human bounding box is below 0.5, as a missed detection case. Note that we only keep at most one human bounding box per image and determine the main subject using the approach mentioned in Section 2 when detecting more than one human objects in an image. Meanwhile, a human-centric image without person detected is also regarded as a missed detection case and we treat it as a non-human-centric image to generate crops (see Section 3.4 of the main text).

In such case, we find that the detected human bounding boxes have a missed detection rate of 3.4% in the test sets of GAICD [15], FCDB [1], and FLMS [3] datasets, which have a total of 265 human-centric images. Then we evaluate the proposed model using the above inaccurate human bounding box, yielding the results: $\overline{SRCC}$=0.790, $\overline{Acc_5}$=59.3, $\overline{Acc_{10}}$=76.8 on GAICD dataset, IoU=0.745, $Disp$=0.0658 on FCDB and FLMS datasets, which are comparable with the performance of the proposed method using manually checked human bounding boxes (see Table 2 and Table 3 of the main text). This is probably because we perform partition on the feature map (with a large receptive field), rather than input image, which supports the model to be more robust to the human bounding boxes.

4 Comparison with the Simplest Baselines

As described in Section 3.2 of the main text, we use the human bounding box to derive partition-aware feature from the basic feature and apply partition-aware feature to help improve human-centric image cropping. Apart from the proposed method, naturally, there are some other approaches to leverage human bounding box for human-centric image cropping. We take two simplest

Table 1. Comparison with the simplest baselines on human-centric images in FCDB [1] and FLMS [3] datasets. Baseline_A takes the human bounding box as output. Baseline_B selects a crop that contains a larger human area and places the person closer to the center from pre-defined candidates

Method	Backbone	Training Data	IoU↑	Disp↓
Baseline_A	-	-	0.3634	0.1600
Baseline_B	-	-	0.3903	0.1541
Ours(basic)	VGG16	CPC	0.7263	0.0695
Ours	VGG16	CPC	**0.7469**	**0.0648**

ways as baselines: directly take the human bounding box as cropping box (denoted by "Baseline_A") and select a crop that contains a larger human area and places the person closer to the crop center from the pre-defined candidate crops through weighted ranking (denoted by "Baseline_B"). Specifically, given the human bounding box B_h and a candidate crop B_c, "Baseline_B" calculates the rank scores of this candidate as: $IoU(B_h, B_c) + \sqrt{2} - Dist(B_h, B_c)$, in which $IoU(B_h, B_c)$ is the intersection over union between the candaite crop and human bounding box: $area(B_h \cap B_c)/area(B_h \cup B_c)$, and $Dist(B_h, B_c)$ denotes the euclidean distance between their centers normalized by the size of source image. After calculating rank scores of all candidates, "Baseline_B" takes the candidate crop with largest rank score as the best crop.

Due to only producing one cropping box, these two baselines cannot be evaluated using the evaluation metrics of GAICD dataset [15], *i.e.*, $\overline{SRCC}$, $\overline{Acc_5}$, and $\overline{Acc_{10}}$. So we evaluate them on human-centric images in FCDB [1] and FLMS [3] datasets using IoU and Disp as evaluation metrics, whose results are reported in Table 1. For comparison, in Table 1, we also display the performance of our method and its basic version ("Ours(basic)"). We can see that those simplest baselines perform significantly worse than our basic model and are also inferior to existing state-of-the-arts (see Table 3 of the main text), which may attribute to the plain equal treatment of all regions outside the human for human-centric image cropping. In contrast, by using partition-aware, our model can treat different regions in a crop differently conditioned on the human information.

5 Hyper-parameter Analysis

Recall that we have a trade-off parameter λ in Eqn.(7) of the main text, which is set as 1 via cross-validation by splitting 20% training samples as validation set. We report the results on test set in Figure 1 when λ varies from 0 to 100, using the average Spearman's rank-order correlation coefficient ($\overline{SRCC}$) and averaged top-5 accuracy ($\overline{Acc_5}$). Comparing the result without content loss ($\lambda = 0$) and the result with $\lambda = 1$, we can see a clear gap between their performance. When $\lambda = 0$, the quality of predicted heatmap cannot be guaranteed without the

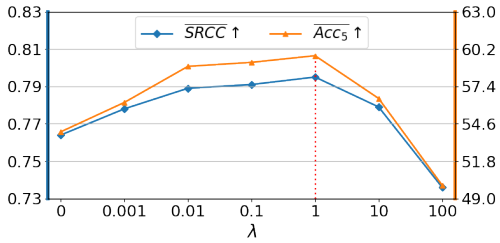

Fig. 1. Performance variation of our method with different trade-off parameter λ in Eqn.(7) of the main text on GAICD dataset. The dashed vertical line denotes the default value used in other experiments

supervision of pseudo ground-truth and the low-quality heatmap may harm the performance. When λ becomes larger than 1, the performance begins to drop. The experimental results in Figure 1 demonstrate that our model is robust when setting λ in the range of [0.01,1].

6 Evaluation on General Image Cropping

In Section 4.4 of the main text, we have shown that the proposed method outperforms existing state-of-the-art methods on the human-centric images of several benchmarks. Recall our method can perform image cropping on both human-centric and non-human-centric images (see Section 3.4 of the main text), and the proposed content-preserving feature is also useful for general image cropping, because preserving important content is also crucial for non-human-centric images. So we evaluate our model on the whole test set of GAICD dataset [15], which contains 50 human-centric images and 150 non-human-centric images. In Table 2, we compare our method with different state-of-the-arts on GAICD and use three metrics for performance evaluation: the average Spearman's rank-order correlation coefficient ($\overline{SRCC}$) and averaged top-N accuracy ($\overline{Acc_N}$) for both $N = 5$ and $N = 10$. For our method, we additionally report the results of a basic version ("Ours(basic)") without using partition-aware feature or content-preserving feature. We can see that Ours(basic) yields similar results with GAIC(ext) [16], which attributes to that they adopt the same region feature extractor (RoI+RoD). Ours outperforms Ours(basic) significantly, which demonstrates the effectiveness of content-preserving feature for the general image cropping task. Among these methods, CGS [8] is the most competitive one, probably because it exploits mutual relations between different crops.

Despite that the proposed approach outperforms the state-of-the-art methods including CGS [8] when evaluating on human-centric images, it does not achieve the best performance on the whole test set of GAICD. One possible reason is that non-human-centric image cropping cannot benefit from the proposed partition-aware feature, while there are significantly more non-human-centric images than human-centric ones in the test set of GAICD with a ratio of 3:1. The other

Table 2. Quantitative comparison on the whole test set of GAICD dataset [15]. GAIC(ext) [16] is the extension of GAIC[15]. The results of other methods are from original papers, except for the results of VFN and VEN are from [15]. Two best results for each metric are highlighted in boldface

Method	Backbone	Training Data	$\overline{SRCC}\uparrow$	$\overline{Acc_5}\uparrow$	$\overline{Acc_{10}}\uparrow$
VFN [2]	AlextNet	Flickr	0.450	26.7	38.7
VEN [14]	VGG16	CPC	0.621	37.6	50.9
ASM-Net [13]	VGG16	GAICD	0.766	54.3	71.5
GAIC(ext) [16]	MobileNetV2	GAICD	0.783	57.2	**75.5**
CGS [8]	VGG16	GAICD	**0.795**	**59.7**	**77.8**
Ours(basic)	VGG16	GAICD	0.777	54.3	71.0
Ours	VGG16	GAICD	**0.793**	**58.6**	74.5

reason may lie in the generation mechanism of candidate crops in GAICD [15], which constrains the area of candidate crops to preserve the major content of the source image. Recall the content-preserving feature is proposed to augment our method by learning how well each candidate crop preserves the important content. For the images that place the important content in the center, the performance gain of the proposed content-preserving feature would be limited by the aforementioned mechanism. In practice, the area of good crops is varying and the candidate crops should be sampled across different sizes and positions.

Apart from GAICD dataset, we also evaluate on the whole FLMS dataset [15], which has total 500 images including 39 human-centric images, and present results in Table 3, in which we train the model on CPC dataset [14], and use intersection of union (IoU) and boundary displacement (Disp) as evaluation metrics following [14]. Ours outperforms Ours(basic) significantly, which again proves that our modification is also beneficial for general image cropping. Among these baselines, CACNet [5] achieves the best performance, probably because it is trained to directly regress the best crop without using predefined candidate crops. Then, our model also performs comparably with the strongest CACNet in terms of IoU and Disp. Moreover, the proposed method can perform favorably against CGS [8] on FLMS, which verifies the utility of content-preserving feature.

In summary, although the focus of this work is human-centric image cropping, our method can still achieve competitive performance on the general image cropping task.

7 Running Speed and Model Complexity

A practical image cropping model should have fast speed and acceptable computational complexity for real-time implementation. We compare the running speed in terms of frame-per-second (FPS) and model complexity of our method with different state-of-the-arts, and report results in the last three columns of Table 3. All models are tested on the same PC with i9-10920X CPU, 64G RAM and

Table 3. Quantitative comparison on the whole FLMS dataset [15]. All methods use VGG16 [12] as backbone. The results of other methods are from original papers. To measure the efficiency, we also report FPS and model complexity of different models. Two best results for each metric are highlighted in boldface

Method	Training	IoU↑	Disp↓	FPS↑	FLOPs	Parameters
VEN[14]	CPC	0.837	0.041	10	15.39G	40.93M
ASM-Net[13]	CPC	0.849	0.039	102	64.36G	14.95M
LVRN[10]	CPC	0.843	-	270	15.39G	40.93M
GAIC[15]	GAICD	0.834	0.041	**299**	20.07G	16.31M
CGS[8]	GAICD	0.836	0.039	174	20.08G	21.25M
CAC-Net[5]	FCDB	**0.854**	**0.033**	**323**	16.26G	18.93M
Ours(basic)	CPC	0.832	0.042	211	20.17G	17.90M
Ours	CPC	**0.850**	**0.034**	128	20.25G	19.47M

one NVIDIA RTX 3090 GPU, and our method as well as other methods that do not provide codes is implemented under the Pytorch toolbox. Following previous works [15,8], the running speed of all methods is evaluated on GAICD dataset [15], in which each image has about 86 candidate crops, and our method is tested on the human-centric images of GAICD dataset using both partition-aware and content-preserving features. Moreover, the reported FLOPs of different models are calculated on image with the resolution used in their original papers.

There are three points worth noting: 1) VEN [14] and LVRN [10] adopt the same network architecture, leading to the same FLOPs and number of parameters. 2) The high FPS of CAC-Net [5] can be attributed to its one-stage regression manner, which is significantly different from the general pipeline used in our method. 3) For the methods that do not provide codes, including ASM-Net [13], CGS [8], and CAC-Net [5], their FLOPs and number of parameters are calculated based on our implementations, which may be different from them reported in their original papers due to different implementation details. In Table 3, we can see that the proposed method runs slower than some state-of-the-art methods yet still at 128 FPS, which enables our model to be applied to practical applications with real-time implementation requirement. Regarding the model complexity, the FLOPs (the number of multiply-adds) and number of parameters of our model are more than the most efficient GAIC [15] and CAC-Net [5], but comparable with most other competitors. Furthermore, the comparison between our model and its basic version ("Ours(basic)") indicates the low computational resources of additional modules, *i.e.*, 80M FLOPs and 1.57M parameters, which is almost ignorable compared to the computational cost of our basic model.

8 More Qualitative Results

To better demonstrate the effectiveness of our method, we present more qualitative comparisons with the state-of-the-art methods on the human-centric images

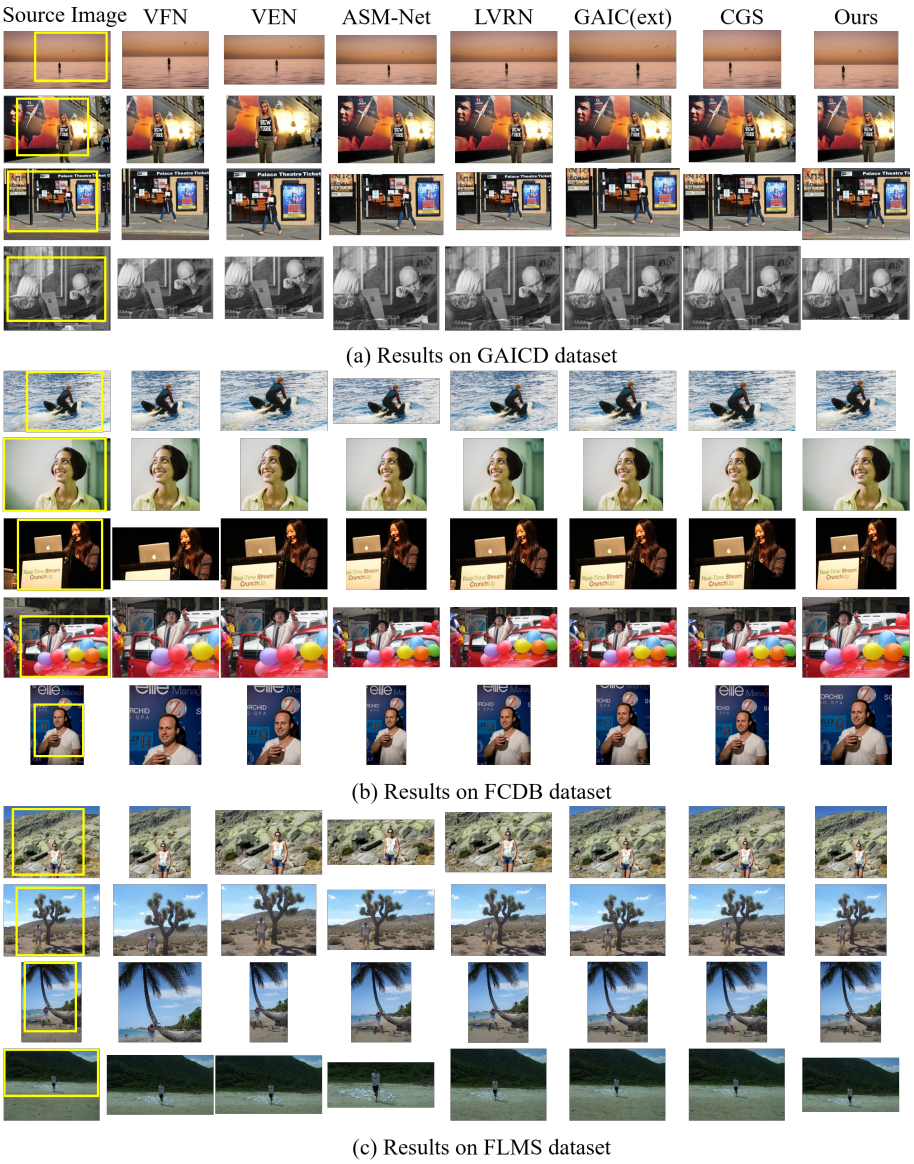

Fig. 2. Qualitative comparison of different methods on the human-centric images of GAICD [15], FCDB [1], and FLMS [3] datasets. We show the best crops predicted by different methods, demonstrating that our method can generate better results close to the ground-truth best crops (yellow)

of GAICD [15], FCDB [1], and FLMS [3] datasets. In Figure 2, we show the source image, the ground-truth best crop, and the predicted best crops by different methods. For the images in FLMS dataset that are assigned with more than one ground-truth crop, we randomly select one to draw on the image.

It can be seen that the proposed method generally produces more appealing cropping results close to the ground truth, which attributes to more reasonable content preservation and removal. For example, in the second row on GAICD dataset, most compared approaches cannot entirely remove the extraneous content on the right side, while the proposed method works well. In the second row on FCDB dataset, considering that the person looks at the upper right, it is supposed to preserve more corresponding area in the returned best crop. However, most methods fail on it, but our method accomplishes it correctly. Those examples further validate the utility of our method for human-centric image cropping.

Additionally, to take a close look at the superiority of the proposed method, we compare the cropping results of our method and its basic version ("Ours(basic)") without using partition-aware feature or content-preserving feature in Figure 3. To study the sensitivity to human of our model, we show the cropping results of our/basic models on the images with different face orientations and postures in Figure 3 (a) and (b), respectively, from which we see that the proposed model can adaptively generate well-composed cropping according to the human information and generally produce more reliable crops than the basic model. For example, in the first two rows of Figure 3 (a), when the person looks to the right (*resp.*, left), our model preserves more content on the right (*resp.*, left) of human, yielding visually pleasing crops. Meanwhile, in the third row, when the person looking straight ahead, our model places the human in the crop center and removes the distracting contents, leading to good composition with visual balance. However, for above images, the basic model performs inflexible and typically places the human closer to the crop center regardless of different face orientations, resulting in inferior cropping results. Similarly, in Figure 3 (b), the proposed model is capable of incorporating human posture and performing flexible content preservation/removal for the area around person, which can generate more appealing crops than the basic model. Their performance differences can be attributed to the partition-aware feature, which enables our model to treat different regions in a crop differently conditioned on the human information.

Besides human information, we also evaluate our/basic models on the images containing the important content that are often ignored by saliency detection methods, such as interesting objects (*e.g.*, landmark) and the objects that person interacts with. In Figure 3 (c), we show the results on the images with landmark behind the person. It can be seen that the our model can roughly capture the entire landmark in the predicted crops, while the basic model may fail to include the landmark very well. We conjecture that this is primarily driven by the important content estimation adopted in the proposed model, which is supervised by the pseudo heatmap that highlights the attractive objects (*e.g.*, landmark) in the background and the objects that person interacts with (see

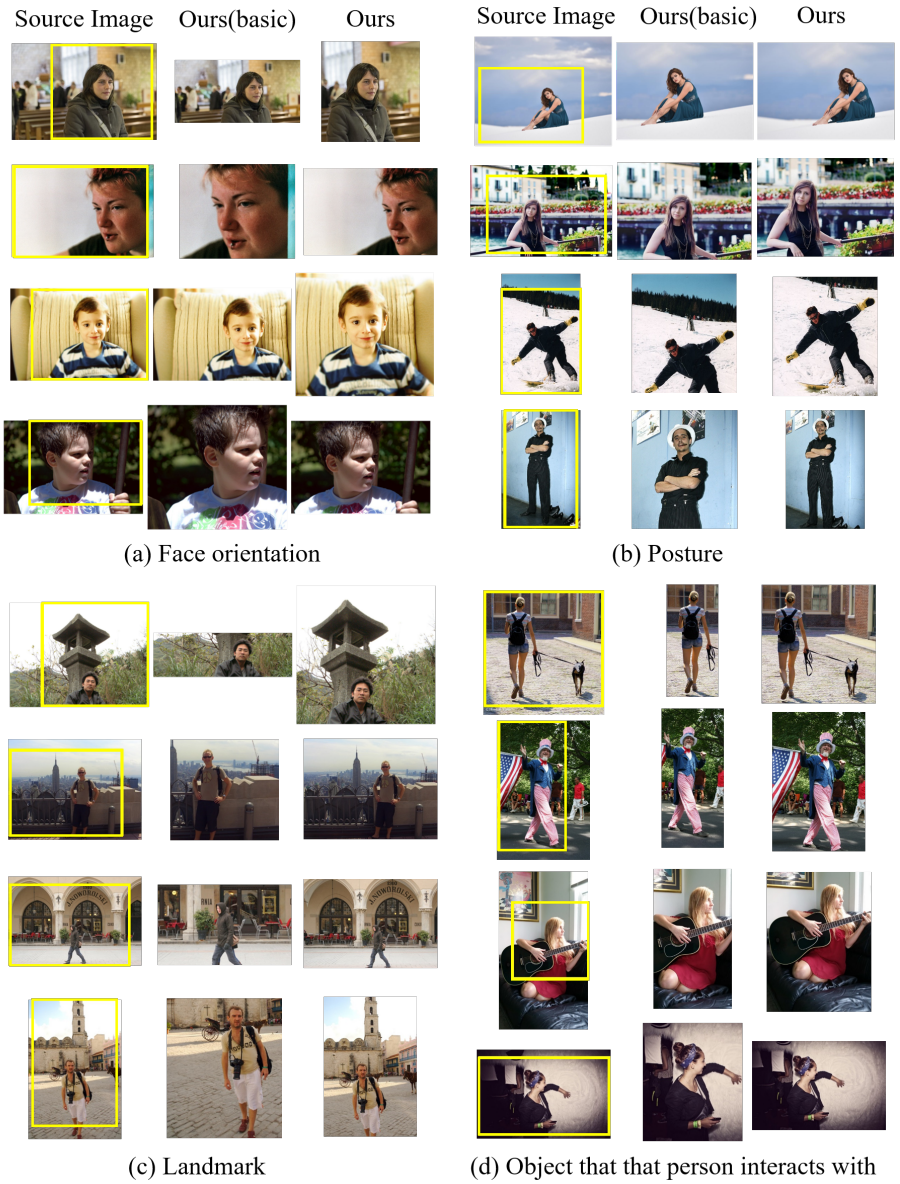

Fig. 3. Qualitative comparison of our model and its basic version without using partition-aware feature or content-preserving feature. The examples are divided into four groups according to the concerned content (under each subfigure), *e.g.*, group (a) evaluates model on the images with various face orientations. The yellow box indicates the ground-truth best crop

Section 3.3 of the main text). In Figure 3 (d), compared with the basic model, our model can preserve these objects that person interacts with (*e.g.*, dog, flag, and guitar) better, which benefits from the aforementioned pseudo heatmap and graph-based region relation mining that exploits the mutual relation between different regions for important content estimation (see Section 4.6 of the main text). Those qualitative comparisons further confirm the effectiveness of the proposed partition-aware and content-preserving features on human-centric image cropping task.

References

1. Chen, Y.L., Huang, T.W., Chang, K.H., Tsai, Y.C., Chen, H.T., Chen, B.Y.: Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In: WACV (2017)
2. Chen, Y.L., Klopp, J., Sun, M., Chien, S.Y., Ma, K.L.: Learning to compose with professional photographs on the web. In: ACMMM (2017)
3. Fang, C., Lin, Z., Mech, R., Shen, X.: Automatic image cropping using visual composition, boundary simplicity and content preservation models. In: ACMMM (2014)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
5. Hong, C., Du, S., Xian, K., Lu, H., Cao, Z., Zhong, W.: Composing photos like a photographer. In: CVPR (2021)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
7. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D., et al.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
8. Li, D., Zhang, J., Huang, K., Yang, M.H.: Composing good shots by exploiting mutual relations. In: CVPR (2020)
9. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts (2017)
10. Lu, W., Xing, X., Cai, B., Xu, X.: Listwise view ranking for image cropping. IEEE Access **7**, 91904–91911 (2019)
11. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. PAMI **39**, 1137–1149 (2015)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
13. Tu, Y., Niu, L., Zhao, W., Cheng, D., Zhang, L.: Image cropping with composition and saliency aware aesthetic score map. In: AAAI (2020)
14. Wei, Z., Zhang, J., Shen, X., Lin, Z., Mech, R., Hoai, M., Samaras, D.: Good view hunting: Learning photo composition from dense view pairs. In: CVPR (2018)
15. Zeng, H., Li, L., Cao, Z., Zhang, L.: Reliable and efficient image cropping: A grid anchor based approach. In: CVPR (2019)
16. Zeng, H., Li, L., Cao, Z., Zhang, L.: Grid anchor based image cropping: A new benchmark and an efficient model. PAMI **PP**(01), 1–1 (2020)