Multi-Scale Attention Network for Image Cropping

Tianpei Lian[1], Ke Xian[1], Zhiyu Pan[1], Chaoyi Hong[1], Zhiguo Cao[1], and Weicai Zhong[2]

[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
Wuhan 430074, PR China
[2]Huawei CBG Consumer Cloud Service Big Data Platform Dept
Xi'an 710075, PR China
[1]{tp_lian, kexian, zhiyupan, cyhong, zgcao}@hust.edu.cn
[2]zhongweicai@huawei.com

***Abstract*— Automatic image cropping is a completely practical but challenging task which aims to improve the aesthetic quality of an image by removing irrelevant areas. Most previous image cropping methods ignored compositional relationships among different regions of a given image. Global compositional relationships are extremely important for cropping models to decide whether to reserve a certain object of an input image. In this work, we propose a multi-scale attention network (MSANet) to address this issue. We employ three plug-and-play attention modules to catch the context on three different scales. The multi-scale attention (MSA) module ensures that our model perceives objects of different sizes and preserve needed areas. Moreover, we design a border-reserved grid anchor based formulation to better handle the situations where the subjects are at the edge of input images. The cosine similarity loss function is also utilized to acquire stable results. Extensive quantitative and qualitative experimental results show that our model is well aware of the compositional relationships of images. Compared to existing works, our multi-scale attention network achieves state-of-the-art performance with less time and lighter weights.**

***Keywords*—Image cropping; aesthetic quality; multi-scale attention; deep learning**

I. Introduction

Image cropping, aiming to enhance the aesthetic quality of a given image by finding the best cropping box, is a challenging but significant computer vision task with sufficient applications, such as image editing [1], photo-processing [2], view recommendation [3] and image thumbnailing [4]. Since cropping manually is tedious and expensive, automatic image cropping has been a great need and attracted the attention of many researchers [5]–[12].

With the rapid development of aesthetic assessment [13], [14], a number of deep image cropping methods have appeared in the past few years. Most of them divided the cropping process into two stages: candidates generation and aesthetic evaluation. Basically all methods focused on optimizing the means of generating the candidate cropping boxes. Zeng et al. [11], [12] propose a grid anchor based formulation and an efficient network, which greatly reduce the number of candidate boxes and make good results. Nevertheless, there exists a common defect for image cropping methods.

The corresponding author is Zhiguo Cao. E-mail:zgcao@hust.edu.cn. This work was funded by the DigiX Joint Innovation Center of Huawei-HUST and was supported in part by the National Natural Science Foundation of China (Grant No. 61876211).

(a) (b)

Fig. 1. Two successful cropping examples. The yellow boxes are the cropping boxes. As the white circles show, the left person in (a) is obviously unwanted and the man in (b) is closely related to the woman. Correctly understand the global compositional relationships among different regions is crucial for image cropping.

As shown in Fig. 1, the key to perfect cropping is correctly and fully understanding the global compositional relationships among different regions of input images. In Fig. 1(a), the left person (see the white circle) is obviously unwanted and should be removed. In Fig. 1(b), the woman is looking at her husband (see the white circle) who is fishing, and a good cropping should retain the man. It is worth noting that the size of the man in Fig. 1(b) is very small relative to the raw image. The scale factor should also be considered for image cropping task. It is regretful that no work has specially taken these factors into consideration so far.

In this paper, we propose a multi-scale attention network to perceive the global compositional relationships among various areas on different scales of an input image. The effectiveness of attention module has been proved in many computer vision tasks, like semantic segmentation [15], [16], object detection [17], [18], and depth prediction [19]. We employ three Spatial and Channel Squeeze & Excitation (scSE) Blocks [20] to form the multi-scale module. Each scSE block contains a channel attention and a spatial attention modules in parallel. In our multi-scale attention module, the three scSE blocks respectively act on three features maps of different scales.

Moreover, to handle the situation where the objects are at the edge of images, we propose a border-reserved grid anchor based formulation which improves the work in [11]. Since the ranking order of outputs is more important than the absolute scores, we utilize the cosine similarity loss function to acquire better results. In summary, our main contributions

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9326681

are as follows:

- We present a multi-scale attention image cropping network to focus on global compositional relationships among various regions of input images.
- A border-reserved grid anchor based formulation and the cosine similarity loss function are employed to get better cropping results.
- Extensive experimental results demonstrate that our method achieves state-of-the-art performances on GAICD [11] benchmark compared to existing works.

II. Related Work

In this section, we roughly divide the mainstream image cropping methods into two categories: attention-based and aesthetic-based methods.

A. Attention-Based Methods

Early image cropping methods [21]–[24] devote to preserving the areas that attract the most attention of observers. It is obvious that these methods follow a two-stage pipeline. Firstly, a trained saliency detection model (e.g., [25], [26]) is used to get a saliency map. Secondly, the cropping box containing the highest saliency value is chosen to be the final cropping result. However, high saliency values do not equal to high aesthetic qualities, these methods can not perform well compared to aesthetic-based methods.

B. Aesthetic-Based Methods

As aesthetic assessment methods [13], [14] develop rapidly, aesthetic-based cropping methods gradually move into the mainstream. Similarly, these methods mainly contain two stages: candidates generation and aesthetic evaluation. At first, plentiful candidate boxes are generated using certain rules. Then all the candidates are evaluated by a trained aesthetic assessment network, and the cropped image with the highest aesthetic quality is picked out to be the final result. The main difference between these methods is the way to generate candidate boxes. Wang et al. [8] utilize a saliency detection to get an inital box, and 1296 boxes are generated around it. A sliding window search strategy is used to produce abundant anchor boxes in VFN [7]. Too many candidate boxes lead to huge time consumption due to the reuse of aesthetic assessment networks. Several works try to reduce the number of candidates and avoid the reuse of aesthetic assessment networks. A2-RL [9] formulates image cropping as a sequential decision-making process with a deep reinforcement learning network. VPN [3] employs a knowledge distilled network and directly outputs aesthetic scores for all candidate boxes through only one inference. Zeng et al. [11], [12] propose a grid anchor based formulation to reduce the number of candidates to less than 100, and an efficient aesthetic assessment network is proposed to score all candidate boxes at once. Besides, GAICD dataset and a series of metrics are established in their works to evaluate the effectiveness of image cropping.

However, global compositional relationships which matter a lot to image cropping have not been valued in above methods. To make up this shortcoming, we propose a multi-scale attention network based on the work in [12]. Besides, a border-reserved grid anchor based formulation and cosine loss function are employed in this work.

III. Methodology

In this section, we will clarify the multi-scale attention network, the border-reserved grid anchor based formulation, and the training of our cropping model.

A. Multi-Scale Attention Network

The architecture of our multi-scale attention network is shown in Fig. 2, which is designed on the basis of the network in [12]. The model contains 3 parts: feature extraction, multi-scale attention, and aesthetic score prediction modules.

- We use a MobileNetv2 [27] as the backbone network to extract multi-scale features of a given image.
- Different from the work in [12], we parallelly place three scSE [20] blocks to respecitvely act on three scales. The pipeline of scSE block is presented in Fig. 3, each scSE block contains channel attention and spatial attention branches, which can fully grasp the global compositional relationships between different regions of an input image. The features output by the MSA module are adjusted to the same spatial resolution through bilinear upsampling and downsampling operations. Then these features are concatenated together and sent to the aesthetic score prediction module after a convolutional layer with 1×1 kernels.
- The aesthetic score prediction module completely follows the work in [12]. The RoIAlign and RoDAlign operations respectively transform the region of interest and the discarded region to feature maps with a resolution of 9×9. Then aesthetic scores are produced for all candidate boxes.

B. Border-Reserved Grid Anchor Based Formulation

The grid anchor based formulation of image cropping is proposed in [11], which generates typical anchor boxes for each image and greatly reduce the number of boxes. However, there exists a visible problem in this manner. As shown in Fig. 4(a), the grid centers are chosen to be corners of candidate cropping boxes. It means that the edges of each image are surely cut off. Suppose the spatial resolution of an image is 800×800, and $M = N = 12$, each bin represent 67 pixels on both w and h axes. Then the areas containing 34 pixels on both axes at the edges of images are unavoidable to be cut off. It obviously performs poorly when the main objects are at the edge of images.

In view of this problem, we make some changes to the formulation. As shown in Fig. 4(b), we use intersections of grids instead of grid centers as corners of boxes. However, if the vertex of a source image becomes the corner of one box, the RoDAlign will be meaningless. As a result, we leave some blank on the four sides of each image. The lengths of blank area on w and h axes are respectively dw and dh times the grid size for each image. In our experiments, we set $M = N = 12$, $m = n = 4$, and $dw = dh = 0.25$. For one 800×800 image,

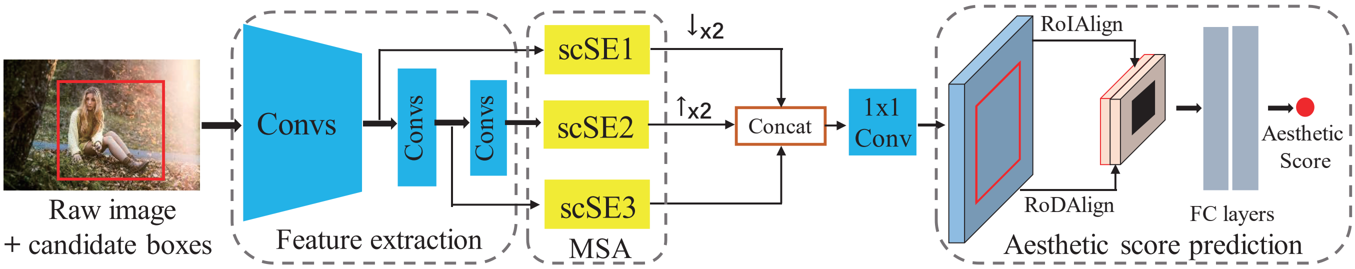

Fig. 2. The architecture of our cropping model. The whole framework is designed on the basis of the network proposed in [12]. The feature extraction module is MobileNetv2 [27]. "scSE1", "scSE2", and "scSE3" denote three scSE [20] blocks. "↑ ×2" and "↓ ×2" denote bilinear upsampling and downsampling, respectively.

TABLE I. Ablation study on multi-scale attention (MSA) module and cosine loss function (cos_loss) based on GAICD [11] dataset. "no" and "✓" respectively represent the components are used or not used. The boldface numbers mean the best results.

MSA	cos_loss	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$SRCC$	PCC
no	no	62.0	58.3	54.5	52.0	78.5	75.2	72.7	70.8	0.782	0.806
no	✓	64.0	58.5	54.3	51.4	80.5	76.2	73.5	71.4	0.786	0.806
✓	no	63.5	59.8	55.7	53.0	79.5	77.2	74.7	72.1	**0.788**	**0.813**
✓	✓	**65.0**	**61.0**	**56.5**	**54.1**	**82.0**	**78.5**	**75.7**	**72.6**	0.781	0.805

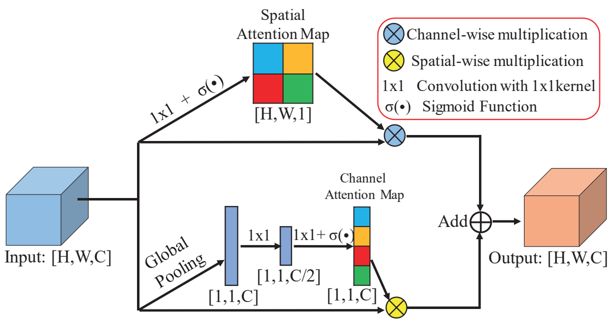

Fig. 3. The pipeline of scSE [20] block. A channel attention and a spatial attention modules are placed parallelly in it.

the number of pixels on each side that are surely cut off is $\frac{800}{12+0.25+0.25} \times 0.25 = 16$, which is significantly smaller than 34. Similarly, we place some constraints on the area and aspect ratio for each potential cropping box in the test stage, which are shown as follows:

$$S_c \geq \lambda S_I, \qquad \alpha_1 \leq \frac{H_c}{W_c} \leq \alpha_2 \tag{1}$$

where S_c, S_I denote the areas of a cropping box and the corresponding original image. W_c, H_c are the width and height of a cropping box. We set λ, α_1, α_2 to 0.5, 0.5, and 2 respectively, which is the same as the work in [11], [12].

C. End-to-End Training of the Cropping Model

We train our model on GAICD [11] dataset, each image of which has about 85 cropping boxes marked with aesthetic scores. In the training stage, each picture and randomly selected 64 candidate boxes of it are sent to the cropping network, then 64 aesthetic scores are predicted for all boxes. In [11], [12], only smooth L_1 loss $\mathcal{L}_{sml1}$ is used to supervise the training of our model, which can be formulated as:

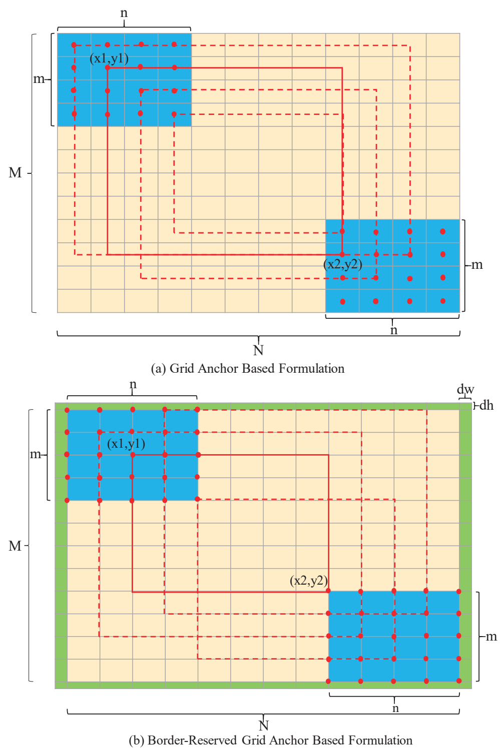

Fig. 4. Illustration of the border-reserved grid anchor based formulation of image cropping. (a) shows the diagram of grid anchor based formulation which is proposed in [11]. (b) shows the diagram of border-reserved grid anchor based formulation which improves the manner in (a). M and N are the number of bins for grid partition. And m, n represent the adopted range of anchors. dw and dh define the range of border to be surely cut off. (x_1, y_1) and (x_2, y_2) are the corners of one cropping box.

$$\mathcal{L}_{sml1} = \frac{1}{64}\sum_{i=1}^{64} l_i \tag{2}$$

Fig. 5. Qualitative comparison of whether to use the multi-scale attention module (MSA). The baseline model means the model without MSA module which is trained using only smooth L_1 loss function, i.e., the network in [12]. MSA module can help the model perceive global compositional relationships and correctly retain the needed areas.

in which l_i is the smooth L_1 loss between the predicted score of the $i-th$ box p_i and its groundtruth score g_i. It can be calculated as:

$$l_i = \begin{cases} 0.5e_i^2, & \text{if } |e_i| < 1, \\ |e_i| - 0.5, & \text{otherwise}, \end{cases} \tag{3}$$

where e_i is the error, which is defined as $e_i = g_i - p_i$.

For the cropping model, it is more important to get the right ranking order than predict the same scores as groundtruth. Considering this fact, the cosine similarity loss function is employed in this work, which is defined as:

$$\mathcal{L}_{cos} = 1 - \frac{\mathbf{g} \cdot \mathbf{p}}{|\mathbf{g}| \cdot |\mathbf{p}|} \tag{4}$$

where $\mathbf{g}$, $\mathbf{p}$ denote the vectors of groundtruth and predicted scores respectively. Then the overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{sml1} + \lambda_c \mathcal{L}_{cos} \tag{5}$$

where λ_c is the coefficient of cosine loss function. In this work, we set λ_c to 0.4 through experiments.

IV. Experiments

A. Experimental Setting

*1) **Implementation Details**:* We conduct our model on Pytorch, and Adam optimizer with default parameters is used to train our model. The short side of each input is resized to 256 with a constant aspect ratio. Some data augmentation strategies (e.g., horizontally flip and color jittering) are utilized in the training stage, which is the same as the work in [12]. In this paper, we carry out our method and other approaches on one RTX 2080Ti GPU and an Intel Xeon Silver 4210 CPU at 2.2GHz.

*2) **Evaluation Data and Metrics**:* A set of new metrics have been proposed in [11], [12], and they can only be employed in GAICD [11] dataset. We test our method on GAICD dataset, which has 200 test images. Each test image is annotated about 85 candidate boxes with marked scores. In the testing stage, each image and all boxes of it are inputted into the network, and the predicted aesthetic scores are output for the input boxes. Then we can calculate the metrics between the groundtruth scores and predictions.

We adopt the metrics proposed in [11], [12], which contains 3 categories: ranking correlation metrics (i.e., $SRCC$ and PCC), best return metrics (i.e., $Acc_{K/N}$), and rank weighted best metrics (i.e., $Acc^w_{K/N}$).

TABLE II. The results of different coefficients of cosine loss function λ_c on GAICD [11].

λ_c	$Acc_{1/5}$	$Acc_{4/5}$	$Acc_{1/10}$	$Acc_{4/10}$	$SRCC$	PCC
0.2	62.5	**51.7**	78.5	71.0	0.773	0.797
0.4	**64.0**	51.4	**80.5**	**71.4**	**0.786**	**0.806**
0.6	62.5	50.9	79.0	70.6	0.781	0.802

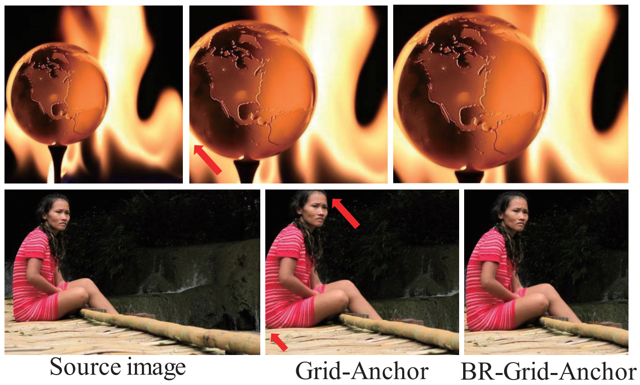

Fig. 6. The effectiveness of border-reserved grid anchor based formulation. Using border-reserved grid anchor based formulation can better reserve the edges of objects.

B. Ablation Study

*1) **Multi-Scale Attention Module**:* To show the effectiveness of our multi-scale attention module, we carry out quantitative and qualitative experiments. Table. I shows the results of cropping models with different modules. In our experiments, we treat the model without the MSA module and cosine loss as the baseline. It can be seen that MSA can obviously improve the cropping accuracy. When both MSA module and cosine loss function are employed in our model, we can get the best results. Fig. 5 shows the effectiveness of MSA module using two examples. The edge of the glider in the left image and the fisherman in the right image (see red arrows) are truncated by the baseline, and they are preserved well by the model with MSA module. Particularly, though the fisherman in the right image is extremely tiny, our multi-scale attention network can get a good result.

*2) **Cosine Similarity Loss Function**:* As Table. I shows, the use of cosine similarity loss function can improve the accuracy of our model. Different coefficients of cosine loss function λ_c (see Equ. 5) has a significant influence on evaluation results. Table. II presents the results of different λ_c, and the preferable λ_c is 0.4.

*3) **Border-Reserved Grid Anchor Based Formulation**:* As the candidate boxes are fixed when we evaluate our model

TABLE III. The comprehensive comparisons between the mainstream image cropping methods on cropping effectiveness, model size, and running speeds.

Method	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$SRCC$	PCC
A2-RL [9]	23.0	-	-	-	38.5	-	-	-	-	-
VPN [3]	40.0	-	-	-	49.5	-	-	-	-	-
VFN [7]	27.0	28.0	27.2	24.6	39.0	39.3	39.0	37.3	0.450	0.470
VEN [3]	40.5	36.5	36.7	36.8	54.0	51.0	50.4	48.4	0.621	0.653
GAIC [11]	53.5	51.5	49.3	46.6	71.5	70.0	67.0	65.5	0.735	0.762
Baseline [12]	62.0	58.3	54.5	52.0	78.5	75.2	72.7	70.8	**0.782**	**0.806**
Ours	**65.0**	**61.0**	**56.5**	**54.1**	**82.0**	**78.5**	**75.7**	**72.6**	0.781	0.805
Method	$Acc^w_{1/5}$	$Acc^w_{2/5}$	$Acc^w_{3/5}$	$Acc^w_{4/5}$	$Acc^w_{1/10}$	$Acc^w_{2/10}$	$Acc^w_{3/10}$	$Acc^w_{4/10}$	$Model\ Size$	FPS
A2-RL [9]	15.3	-	-	-	25.6	-	-	-	142.4M	2.55
VPN [3]	19.5	-	-	-	29.0	-	-	-	285.1M	54.2
VFN [7]	16.8	13.6	12.5	11.1	25.9	22.1	20.7	19.1	116.4M	0.36
VEN [3]	20.0	16.1	14.2	12.8	30.0	25.9	24.2	23.8	319.8M	0.06
GAIC [11]	37.6	33.9	31.5	30.0	53.7	49.4	48.4	46.9	52.9M	**275**
Baseline [12]	38.9	37.3	36.5	36.7	56.5	54.6	53.5	53.3	**10.7M**	152
Ours	**41.9**	**41.9**	**38.7**	**38.4**	**58.7**	**58.5**	**56.3**	**55.2**	12.0M	138

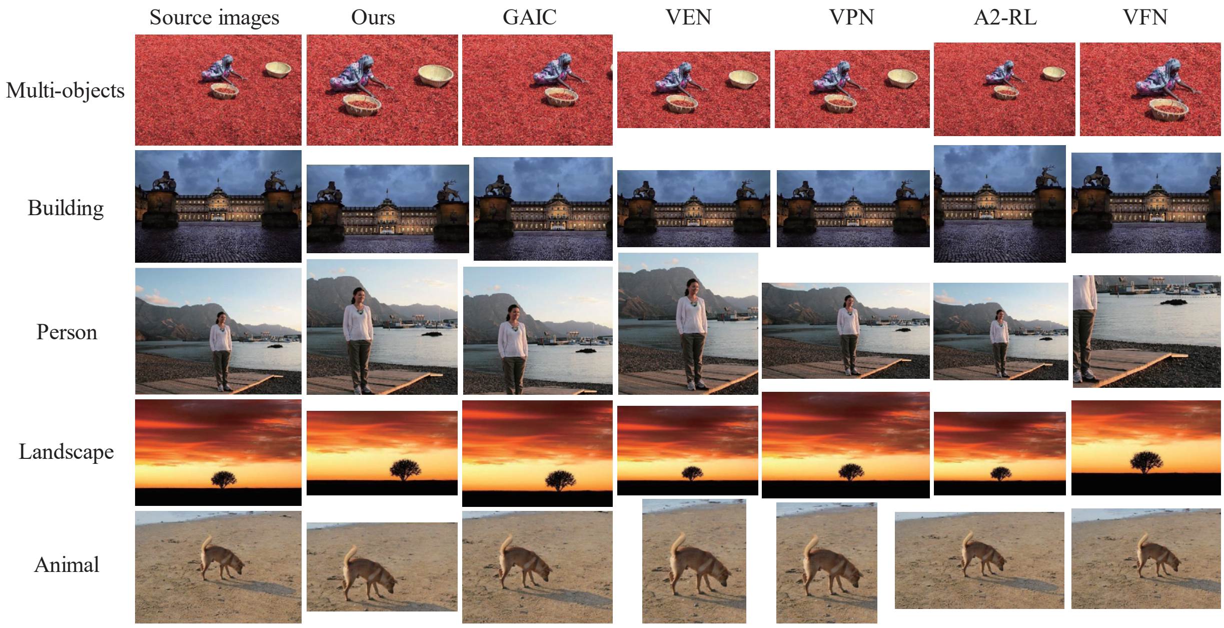

Fig. 7. Qualitative comparison between different methods. For the methods which can output multiple boxes, the shown images are the returned top-1 results.

using the metrics mentioned in Section IV-A2, we test our border-reserved grid anchor based formulation through visualization results. Fig. 6 shows the comparison between grid anchor based formulation (Grid-Anchor) and border-reserved grid anchor based formulation (BR-Grid-Anchor). The lefts are source images, objects of which are near the edges. The cropping results show that our model with border-reserved grid anchor based formulation tries not to damage the edges of objects. Specifically, the full ball is retained in the upper image. In the bottom image, the head and back of the women are well kept by our model with border-reserved grid anchor based formulation.

C. Comparisons to State-of-the-art Approaches

*1) **Quantitative Comparisons**:* A series of experiments are conducted to confirm the superiority of our multi-scale attention network over other approaches. Table. III shows the results of different image cropping methods on GAICD [11] benchmark. As A2-RL [9] and VPN [3] can not support the candidate boxes in GAICD dataset, only $Acc_{1/N}$ and $Acc^w_{1/N}$ metrics can be calculated. When we calculate the indexes, the output box is approximated to the closest candidate box in GAICD. The results mainly contain two aspects: cropping accuracy and cropping efficiency.

Cropping Accuracy As Table. III shows, our multi-scale attention network outperforms other methods by a noticeable margin in almost all the indexes. As for $SRCC$ and PCC metrics, our model gets comparable results to the baseline [12] model. The results demonstrate the effectiveness of our improvements in this paper.

Cropping Efficiency In addition to accuracy, efficiency is also an important measurement of a cropping method. The model sizes and running speeds of different methods are also shown in Table. III. To achieve fair comparisons, all the methods are evaluated on the same hardware conditions, and the spatial resolution of each image is fixed at 800×800. It is shown that our model runs faster with smaller size than most methods. It seems strange that GAIC [11] runs faster than the baseline [12], this is due to that depth-wise convolutions in MobileNetv2 [27] can not show the advantage in GPU with high parallelism. Compared to the baseline, the efficiency of our model hardly decreases, which further demonstrates the efficiency of our MSA module.

*2) **Qualitative Comparisons:*** Fig. 7 shows the visual results of different methods on image cropping. There are five typical scenes: multi-objects, building, person, landscape, and animal. It can be seen that our model can understand the global context information and handle the details well. For example, our model can fully retain the symmetrical structure in the second image. And in the third image, some blank are suitably left in the right and the feet of the women are well reserved.

V. Conclusion

In this work, we have proposed a multi-scale attention network for image cropping. Compared to previous works, we focus on the global compositional relationships among different regions on different scales of an input image. A multi-scale attention module is employed in our model to acquire the key information. Besides, a border-reserved grid anchor based formulation is designed to prevent the edges of objects being truncated. To acquire more stable results, cosine similarity loss function is used to supervise the training of our model. Extensive experiments demonstrate the superiority of our model in both accuracy and efficiency over existing works.

References

[1] Yuan-Yang Chang and Hwann-Tzong Chen, "Finding good composition in panoramic scenes," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 2225–2231.

[2] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 37–45. [Online]. Available: https://doi.org/10.1145/3123266.3123274

[3] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, "Good view hunting: Learning photo composition from dense view pairs," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] S. A. Esmaeili, B. Singh, and L. S. Davis, "Fast-at: Fast automatic thumbnail generation using deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] J. Yan, S. Lin, S. Bing Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 971–978.

[6] C. Fang, Z. Lin, R. Mech, and X. Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1105–1108.

[7] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 37–45.

[8] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2186–2194.

[9] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-rl: Aesthetics aware reinforcement learning for image cropping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8193–8201.

[10] D. Li, H. Wu, J. Zhang, and K. Huang, "Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5105–5120, 2019.

[11] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Reliable and efficient image cropping: A grid anchor based approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[12] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Grid anchor based image cropping: A new benchmark and an efficient model," *arXiv preprint arXiv:1909.08989*, 2019.

[13] J.-T. Lee and C.-S. Kim, "Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1191–1200.

[14] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[16] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[17] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, "Attention couplenet: Fully convolutional attention coupling network for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 113–126, 2019.

[19] R. Li, K. Xian, C. Shen, Z. Cao, H. Lu, and L. Hang, "Deep attention-based classification network for robust depth prediction," in *Computer Vision – ACCV 2018*, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham: Springer International Publishing, 2019, pp. 663–678.

[20] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 421–429.

[21] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Systems*, vol. 9, no. 4, pp. 353–364, 2003.

[22] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proceedings of the 16th annual ACM symposium on User interface software and technology*, 2003, pp. 95–104.

[23] F. Stentiford, "Attention based auto image cropping," in *International Conference on Computer Vision Systems : Proceedings (2007)*, 2007.

[24] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2232–2239.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[26] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," pp. 4510–4520, 2018.