2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897834

DISCRIMINATE CLEARER TO RANK BETTER: IMAGE CROPPING BY AMPLIFYING VIEW-WISE DIFFERENCES

Zhiyu Pan[1] *Zhiguo Cao*[1] *Ke Xian*[1] *Hao Lu*[1] *Weicai Zhong*[2]

[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Huawei CBG Consumer Cloud Service Big Data Platform Dept

ABSTRACT

Image cropping aims to enhance the aesthetic quality of a given image by searching for the good cropping views. One common routine is to score and rank the candidate views by the neural network. The network is expected to discriminate the subtle view-wise differences. However, the image-wise differences and the ambiguity in the annotations render difficulties in discriminating the view-wise differences. To focus on the view-wise differences, we propose a feature spliter to build image-wise and view-wise feature and evaluate the candidate views only based on the view-wise feature. Then, we propose the ranking gain loss that alleviates the ambiguity in annotations to amplify the view-wise differences. The remarkable improvement compared with prior arts on public benchmarks illustrates that the view-wise differences matter in cropping view recommendation.

Index Terms— Image cropping, view-wise difference, ranking gain loss

1. INTRODUCTION

Image cropping is a low-cost technique of image recomposition [1], which is widely used in advertising, photography, film industry and so on. However, manual image recomposition is tedious and time-consuming. Therefore, automatic cropping view recommendation algorithms have been widely studied in the computer vision community [1, 2, 3, 4, 5, 6].

To implement the cropping view recommendation, prior works [1, 4, 5, 6] are in the paradigm of predicting the Mean Opinion Scores (MOS) to evaluate the aesthetic quality of all the candidate views, which are a certain sub-regions of an image. Then, the top-K views are recalled as the recommendation results. The MOS predictor is expected to discriminate between these content-similar candidate views based on the subtle differences between views. However, as illustrated in Fig. 1, the huge differences between images tend to overwhelm the view-wise differences. Besides, some different views in one image are of the same MOS annotation. This ambiguity makes it more challenging to discriminate between the candidate views.

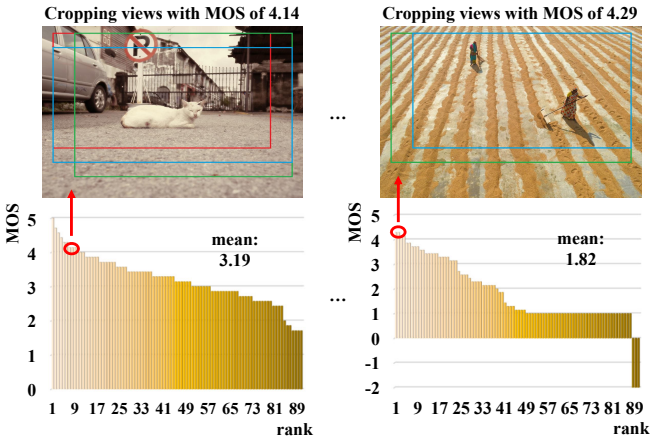

Fig. 1. **The ambiguity in MOS annotation and the MOS distribution variation.** In the first row, the ambiguity in MOS annotation is that the same MOS may correspond with different candidate views; In the second row, the huge image-wise differences about the content lead to the MOS distribution variation.

Facing these two problems, the question arises: *How to discriminate clearer between the candidate views?* We try to help the MOS predictor discriminate clearer in two aspects. To get the more discriminative feature, we propose to split the original feature into the image-wise and view-wise features, which depicts the image-wise and view-wise difference separately. The candidate cropping views are only evaluated by their corresponding view-wise features, which can avoid the influence of the image-wise differences. As for the ambiguity in the annotation, we design the ranking gain loss. By introducing the decided rank information into the loss function, different candidate views will not have the same label, which can alleviate the ambiguity in the annotation and amplify the view-wise differences.

To demonstrate the effectiveness of our two contributions, we built the View Ranking Network (VRN). On the GAIC benchmark [6], our model achieves promising improvement compared with state-of-the-art methods, which proves that

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).

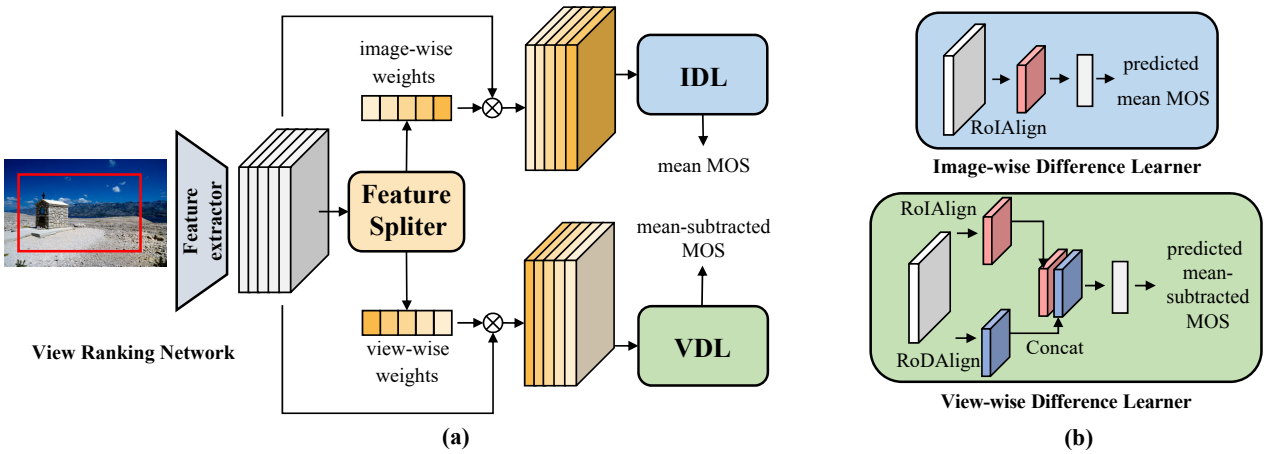

Fig. 2. **Overview of the view ranking network.** (a) The technical pipeline consists of the feature spliter, the view-wise difference learner (VDL) and the image-wise difference learner (IDL). The feature spliter recalibrates the backbone feature into two, which are further processed by the IDL and the VDL to predict the mean MOS and the mean-subtracted MOS respectively. At the inference stage, the IDL is dropped and only the predicted mean-subtracted MOS is used to evaluate the candidates. (b) The architecture of the VDL and IDL.

discriminating clearer between the candidate views helps to rank better and accounts for that view-wise differences matter in the cropping view recommendation task.

2. RELATED WORK

Existing methods can be generally classified into model-based methods and learning-based methods.

Model-based approaches conventionally design human aesthetic models by defining an energy function. Some works [2, 3, 7, 8, 9] model the aesthetics based on visual attention mechanism of human. They resort to saliency detection [10], face detection [11], and text detection [12] to mimic the human attention mechanism to establish the aesthetic model. Hence, the selected views always maximize the dominance of the salient region regardless of aesthetic rules. Some other works [2, 13, 14, 15, 16] model the aesthetic rules explicitly. They model the composition rules (e.g., the rule of thirds and visual balance) as handcrafted features to evaluate the aesthetic quality. However, the poor generality to the in-the-wild images shows that the human aesthetic model can be hardly covered by handcrafted features.

Learning-based approaches instead evaluate candidates by imitating how humans rank views. The common paradigm of these approaches is to evaluate the aesthetic quality of candidate views by a MOS predictor and recall the top-K views as the results. Due to the absence of MOS annotations, the VFN [4] is trained in a self-supervised manner based on the idea that random cropping violates the composition regulation of the well-composed image. After aesthetics evaluation datasets [5, 6, 17] are released, supervised models [1, 5, 6] become the mainstream. They rank candidate views according to the predicted MOSs and achieve acceptable performance on public benchmarks. But these approaches ignore the factors that violate the view-wise difference learning. Compared with these works, our approach ranks the views better by amplifying the view-wise differences to discriminate clearer.

3. PROPOSED APPROACH

To address the problems of overwhelmed view-wise differences and the ambiguity in the annotation, we propose the View Ranking Network (VRN), which supervised by a novel ranking gain loss. As shown in Fig. 2, after the feature extractor, there are two branches of the VRN. The original feature is reweighted by the feature spliter into image-wise and view-wise feature. Guided by the image-wise differences learner (IDL) and view-wise differences learner (VDL), the image-wise and view-wise feature can depict the image-wise and view-wise differences respectively. At the training stage, the IDL is supervised by the mean MOS of the views in the same image, and the VDL is supervised by the ranking gain loss. The ranking gain loss is the combination of the MOS and the rank information, which can alleviate the ambiguity. At the test stage, only the view-wise feature is used to predict the MOSs of every view, which isolates the influence of the image-wise differences.

3.1. View Ranking Network

As shown in Fig. 2, the VRN includes three modules: the VDL, the IDL and the feature spliter. Similar to the GAIC [6], the VDL is composed of the RoIAlign and the RoDAlgin [6]. In the IDL, to focus on the global information, we only adopt the RoIAlign architecture. The feature spliter, which is shown in Fig. 3, is inspired by the SE block [18]. In the feature spliter, a global average pooling is first used to encode global information, and then two fully-connected layers recalibrate

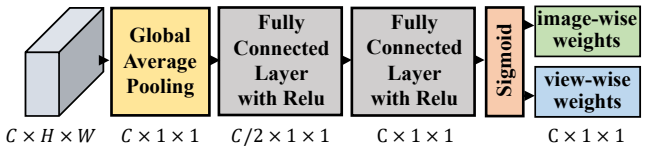

Fig. 3. **View-wise difference amplifier.** The input of VDA is the feature map from the backbone. The feature is compressed and resumed in channel to encode the channel-wise information. After the sigmoid layer, the channel-wise mutually-exclusive weights are generated.

the global feature into the view-wise weights. We assume that the view-wise and image-wise differences are mutually-exclusive. Hence, we make the sum of view-wise weights and image-wise weights equaling to 1 at each channel to get the image-wise weights. Weighted by the image-wise and view-wise weights, the global feature is transformed to the image-wise and view-wise feature, which are then fed into the IDL and the VDL respectively. For one certain candidate view, the IDL will generate the image-wise MOS, the score to judge the image-wise quality, and the VDL will generate the view-wise MOS, which is the score to evaluate the view-wise quality.

At the training stage, the IDL and the VDL are respectively supervised by image-wise and view-wise clues. As shown in Fig. 1, the image-wise differences can change the MOS distribution. Hence, the statistics of the MOSs of the views in one image can be treated as the image-wise clues. We choose the MOS mean as the image-wise clue. Which means that the image-wise MOSs of all the candidate views in one image are all expected to be the average MOS. We use the $\ell 1$ loss to supervise the image-wise MOSs. The view-wise MOSs, which are generated by the VDL, are supervised by the combination of the $\ell 1$ loss about the mean-subtracted MOSs and the proposed ranking gain loss. The details about the overall loss function will be introduced in Sec. 3.2. At the inference stage, the branch of IDL is abandoned. Only the view-wise MOSs are used to evaluate the aesthetic quality.

3.2. Ranking Gain Loss

The ranking gain loss is designed to introduce the relative ordering information into the supervision signal to amplify the view-wise differences. We transform the NDCG metric [19] to a generalized ranking gain, which is denoted by

$$\beta = \sum_{i=1}^{N} \frac{e^{\sigma_i} - 1}{\ln(i+1)}, \tag{1}$$

where i represents the ground truth rank of each candidate, which means that β is a rank-weighted metric. σ_i denotes the similarity between the ith candidate and the retrieval target, defined by

$$\sigma_i = \begin{cases} s_i, & s_i < \hat{s}_i \\ \hat{s}_i - (s_i - \hat{s}_i), & s_i \geq \hat{s}_i \end{cases}, \tag{2}$$

where s_i represents the predicted score of the ith candidate view, and $\hat{s}_i$ denotes the ith ground truth MOS. Eq. (2) can be simplified as

$$\sigma_i = \hat{s}_i - |s_i - \hat{s}_i| \ . \tag{3}$$

Hence, a large β means that the candidate view is in the relatively correct order with accurately predicted score. With the purpose of maximizing β, the ranking gain loss can be formulated by

$$L_r = e^{-\beta} \ . \tag{4}$$

The output of VDL is supervised by the ranking gain loss and the ℓ_1 loss. The overall loss is

$$L = L_i + L_v + \lambda L_r \ , \tag{5}$$

where L_i and L_v are the ℓ_1 loss for IDL and VDL respectively, and λ is a presetting coefficient. The ablation study in Sec. 4 shows the effectiveness of the ranking gain loss.

4. EXPERIMENTS

4.1. Dataset and Metrics

Experiments are conducted on the GAIC dataset [6] that contains a training set of 1036 images with $89,519$ annotated views and a test set with 200 images. The $Acc_{K/N}$ metrics proposed along with the GAIC dataset are employed to evaluate the performance. The $Acc_{K/N}$ metrics indicate how many the predicted top-K views fall in the set of views with top-N annotated MOSs. We set $K = 1$ to show the cropping performance and compute the $\overline{Acc_{K/N}} = \frac{1}{N}\sum_{K=1}^{N} Acc_{K/N}$ to show the recommendation performance.

4.2. Implementation Details

The feature extractor is based on MoblieNetV2 [20], and we reduce the dimension of the output feature map to 32. Our candidate views are defined similar to the ones proposed in GAIC. During training, we set the $\lambda = 3$. The input is one image with its 64 corresponding candidate views. With the Adam optimizer, our model is trained with the learning rate of $1e-4$ for 60 epochs.

4.3. Comparison with State-of-the-Art Methods

Quantitative results on the GAIC dataset are illustrated in Table 1. The significant improvement of our view ranking model on the $Acc_{1/5}$ and $Acc_{1/10}$ metrics shows that the cropping results of our model are obviously more in line with human preferences. As for the $\overline{Acc_5}$ and $\overline{Acc_{10}}$ metrics, our VRN

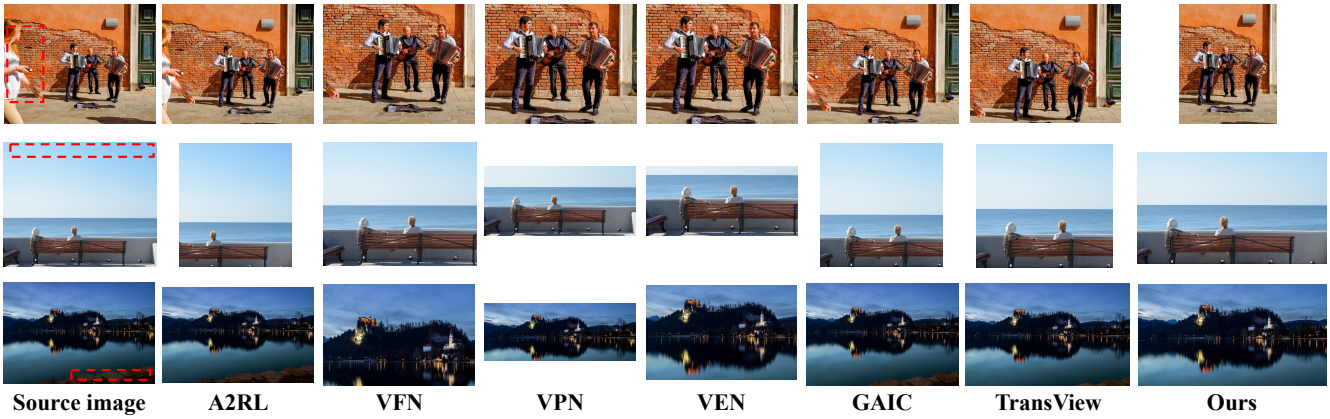

Fig. 4. **Qualitative comparison with other state-of-the-art methods.** The red boxes indicate the redundant areas.

Table 1. Quantitative comparison with other state-of-the-art methods on the GAIC dataset.

Model	$Acc_{1/5}$	$\overline{Acc_5}$	$Acc_{1/10}$	$\overline{Acc_{10}}$
VPN [5]	40.0	-	49.5	-
VFN [4]	27.0	26.7	39.0	38.7
VEN [5]	40.5	37.6	54.0	50.9
GAIC [6]	53.5	50.2	71.5	68.5
GraphConv [1]	63.0	59.7	81.5	77.8
Transview [21]	68.5	**63.9**	83.0	78.0
Ours	**69.5**	62.1	**83.5**	**78.4**

Table 2. Ablation study on the GAIC dataset for the feature spliter (FS) and the ranking gain loss (RGL).

FS	RGL	$Acc_{1/5}$	$\overline{Acc_5}$	$Acc_{1/10}$	$\overline{Acc_{10}}$
		59.0	55.3	79.5	74.5
✓		61.5	55.5	81.5	75.3
	✓	64.0	58.5	79.6	74.7
✓	✓	**69.5**	**62.1**	**83.5**	**78.4**

still achieves comparable performance. For a more intuitive comparison, the predicted top-1 views are visualized in Fig. 4. It is clear that our model can not only remove redundant areas but also align better with the aesthetic rules (e.g., subject dominance, golden ratio, and visual balance).

4.4. Ablation Study

Ablation study is conducted on the GAIC dataset to reveal the contribution of the feature spliter and the ranking gain loss. Results are shown in Table 2. Our baseline is to directly predict the MOS based on concatenated RoIAlign features and RoDAlign features. Results in Table 2 illustrate that the feature spliter and the ranking gain loss have complementary contribution for discriminating the candidate views, which shows that amplifying view-wise difference in multiple aspects can collaboratively help to rank better and boost the recommendation performance.

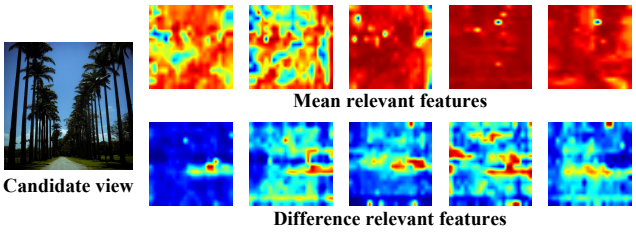

Fig. 5. Examples of the image-wise and view-wise difference relevant features.

4.5. Further Analysis

We further analyze why our pipeline can boost the performance of recommendation. The examples of image-wise and view-wise difference relevant features (the input of IDL and VDL) are visualized in Fig. 5. It can be seen that the areas with high responses in the image-wise difference relevant features cover the whole region of the candidate views. It shows that the global information contributes greatly to the mean. As for the view-wise difference relevant features, they have more structural information, which implies that the view-wise difference relevant feature is more discriminative to represent the composition of candidate views.

5. CONCLUSION

In this paper, we argue that view-wise differences matter in cropping view recommendation. We propose a feature spliter to disentangle the view-wise feature from the original feature. As for the ambiguity in MOS annotation, which makes it challenging to capture view-wise differences, we propose a ranking gain loss to amplify view-wise difference by introducing the rank information. Experiments show that our model achieves remarkable improvement compared with other state-of-the-art methods, which also demonstrates the importance of amplifying the view-wise differences in cropping view recommendation task.

6. REFERENCES

[1] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang, "Composing good shots by exploiting mutual relations," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4213–4222.

[2] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *Proc. ACM international conference on Multimedia*, 2014, pp. 1105–1108.

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

[4] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma, "Learning to compose with professional photographs on the web," in *Proc. ACM international conference on Multimedia*, 2017, pp. 37–45.

[5] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5437–5446.

[6] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang, "Reliable and efficient image cropping: A grid anchor based approach," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5949–5957.

[7] Li-Qun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang, and He-Qin Zhou, "A visual attention model for adapting images on small displays," *Multimedia Systems*, vol. 9, no. 4, pp. 353–364, 2003.

[8] Bongwon Suh, Haibin Ling, Benjamin B Bederson, and David W Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proc. ACM symposium on User interface software and technology*. 2003, pp. 95–104, Association for Computing Machinery.

[9] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. IEEE Conference on International Conference on Computer Vision*. IEEE, 2009, pp. 2232–2239.

[10] Eleonora Vig, Michael Dorr, and David Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.

[11] Henry Schneiderman and Takeo Kanade, "A statistical method for 3d object detection applied to faces and cars," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2000, pp. 746–751.

[12] Xiangrong Chen and HongJiang Zhang, "Text area detection from video frames," in *Pacific-Rim Conference on Multimedia*. Springer, 2001, pp. 222–228.

[13] Mingju Zhang, Lei Zhang, Yanfeng Sun, Lin Feng, and Weiying Ma, "Auto cropping for digital photographs," in *Proc. IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 4–pp.

[14] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato, "Sensation-based photo cropping," in *Proc. ACM international conference on Multimedia*, 2009, pp. 669–672.

[15] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian, "Learning to photograph," in *Proc. ACM international conference on Multimedia*, 2010, pp. 291–300.

[16] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang, "Learning the change for automatic image cropping," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 971–978.

[17] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.

[18] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[19] Kalervo Järvelin and Jaana Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," in *Proc. ACM SIGIR Forum*. ACM New York, NY, USA, 2017, pp. 243–250.

[20] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[21] Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong, "Transview: Inside, outside, and across the cropping view boundaries," in *Proc. IEEE Conference on International Conference on Computer Vision*, 2021, pp. 4218–4227.