

Received May 24, 2019, accepted June 22, 2019, date of publication June 27, 2019, date of current version July 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925430

Listwise View Ranking for Image Cropping

WEIRUI LU¹, XIAOFEN XING¹, (Member, IEEE), BOLUN CAI²,
AND XIANGMIN XU¹, (Member, IEEE)

¹School of Electronics and Information, South China University of Technology, Guangzhou 510000, China

²Tencent WeChat AI, Shenzhen, China

Corresponding author: Xiangmin Xu (xmxu@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U180120050, Grant U61702192, and Grant U1636218.

ABSTRACT Rank-based learning with the deep neural network has been widely used for image cropping. However, the performance of ranking-based methods is often poor and this is mainly due to two reasons: 1) image cropping is a listwise ranking task rather than a pairwise comparison and 2) the rescaling caused by pooling layer and the deformation in view generation damage the performance of composition learning. In this paper, we develop a novel model to overcome these problems. To address the first problem, we formulate the image cropping as a listwise ranking problem to find the best view composition. For the second problem, a refined view sampling (called RoIRefine) is proposed to extract refined feature maps for candidate view generation. Given a series of candidate views, the proposed model learns the Top-1 probability distribution of views and picks up the best one. By integrating refined sampling and listwise ranking, the proposed network called the listwise view ranking network (LVRN) achieves the state-of-the-art performance both in accuracy and speed.

INDEX TERMS Image cropping, listwise ranking, RoIRefine.

I. INTRODUCTION

Image cropping is a common photo manipulation process, which improves the overall composition by removing unwanted regions. Image cropping is widely used in photographic, film processing, graphic design, and printing businesses. Recent methods tend to learn photo composition and extract well-composed regions from ill-composed photo.

With the development of deep learning, most of researchers have devoted their efforts to proposing deep networks based on ranking approach. For ranking-based training, a number of candidate views in each image are labeled with the aesthetic ordering. Then, the image cropping task is formalized as classification of view pairs into two categories (correctly ranked and incorrectly ranked). Finally, sliding window [1], detector [8] or reinforcer [3] is adopted to finding the best view.

In [2], Chen *et al.* first investigated learning-to-rank methods for image cropping. View finding network (VFN) [1] based on a pairwise ranking layer is proposed to model the photo composition and crop image by sliding window. Wei *et al.* trained a view evaluation network (VEN) [4] with the pairwise siamese architecture. Inspired by knowledge

distillation [19] and anchor boxes [18], view proposal network (VPN) is proposed to transfer knowledge from VEN. In [3], reinforcement learning is adopted to crop image step by step, and each step is controlled by the aesthetic score generated by a pre-trained VFN.

However, these ranking-based cropping methods are often poor in performance, which is mainly due to two reasons:

First, pairwise training is unsuitable for image cropping process. In image cropping, the main goal is to pick up the best composed view from a list of candidate views. That is, image cropping is a listwise ranking task rather than pairwise comparison. In addition, pairwise training heavily depends on careful pair selection, because the samples with various distribution will result in training bias. Therefore, pairwise training significantly increases the computational complexity and make the training procedure unstable.

Second, coarse feature extracted from convolutional neural network (CNN) will affect the accuracy of model learning. Previous methods crop and warp the views in raw images or feature maps, and then calculate the rank score for each one. Pixel-accuracy is important in image cropping rather than object classification, and it will be reduced by warp operation. In addition, rescaling caused by the pooling layers will reduce the sampling resolution and damage the composition learning.

The associate editor coordinating the review of this manuscript and approving it for publication was Shuhan Shen.

To overcome these problems of image cropping, we propose a listwise ranking method with refined view sampling. In refined view sampling, a novel region of interest (RoI) operation called RoIRefine is proposed to extract refined feature maps of candidate views. Instead of carefully selecting the view pairs, we take advantage of all annotated views and train the model with a listwise ranking loss.

In summary, our main contributions are:

- We learn deep network for image cropping with listwise ranking.
- We propose a refined view sampling named RoIRefine to alleviate the problem of rescaling and distortion.
- The proposed model significantly outperforms the state-of-the-art methods in both accuracy and speed.

II. RELATED WORK

A. IMAGE CROPPING

Image cropping is a common operation in image editing, which aims to find views with good photo composition. A lot of methods have been proposed towards automating this task. Previous cropping methods, in general, can be divided into attention-based or aesthetic-based approaches. The attention-based methods focus on finding the most visually important area in the original image. For example, [24] trained a simple classifier on an annotated image database for generating attention maps. In [23], visual saliency information, face and skin color detection results are combined for placing bounding box in image cropping. For those aesthetic-based methods, they emphasize the general attractiveness of cropped image. Reference [14] proposed a aesthetic photo cropping system which combines three models: visual composition, boundary simplicity and content preservation. A set of aesthetic quality classifiers were trained to discriminate the quality of candidate windows [8]. With the development of datasets labeled by comparative aesthetic score, ranking-based methods are adopted to grade the composition of candidate windows [2], [17]. Recently, ranking-based methods together with other novel framework (e.g. knowledge transfer [4] and reinforcement learning [3]) have achieved the state-of-the-art performance.

B. LEARNING TO RANK

Ranking is widely used in information retrieval [21], recommender systems [20] and software engineering [22]. In learning-to-rank task, training data consists of lists of items with some partial order which is specified between items in each list. Most ranking algorithms are categorized into three groups by their input representation and loss function: the pointwise, pairwise, and listwise approach [9].

Pointwise approaches assume that each item in the training data has a numerical or ordinal score. Then the learning-to-rank problem can be approximated by a regression problem. Ordinal regression and classification algorithms can be used to predict the score of a single item. For example, the perceptron ranking (PRank) algorithm was proposed to

find a rank-prediction rule that assigns each instance a rank order [10].

Pairwise approach formalizes the learning task as comparison of object pairs into two categories (correctly and incorrectly). RankNet [11] learned a rank rule by using gradient descent methods and a natural probabilistic cost function on pairs of examples. RankBoost [15] used boosting to train ranking model by minimizing classification errors on instance pairs.

Listwise approaches try to directly optimize the value over all items on training data. ListNet [12] tried to define a listwise loss function for learning to rank and introduces two probability models, respectively referred to as permutation probability and Top-1 probability. Suppose that π is a permutation on the n objects. We write the permutation as $\pi = \langle \pi(1), \pi(2), \dots, \pi(n) \rangle$ and $\pi(1) = j$ means the j object is ranked on top one in π permutation. The set of all possible permutations of n objects is denoted as Ω_n . Besides, we defines $\Phi(\cdot)$ as an increasing and strictly positive function. Then, given the list of scores s , ListNet defines the probability of permutation π as

$$P_s(\pi) = \prod_{j=1}^n \frac{\Phi(s_\pi(j))}{\sum_{k=1}^n \Phi(s_\pi(k))} \quad (1)$$

and the Top-1 probability of object j is defined as

$$P_s(j) = \sum_{\pi(1)=j, \pi \in \Omega_n} P_s(\pi), \quad (2)$$

Thus, from Eq. (1) and (2), we can obtain

$$P_s(j) = \frac{\Phi(s_j)}{\sum_{k=1}^n \Phi(s_k)}, \quad (3)$$

where s_j is the score of object $j = 1, 2, \dots, n$.

In general, with the use of Top-1 probability, cross entropy is used to represent the distance between the two given score lists.

III. THE PROPOSED APPROACH

In this paper, we propose a listwise view ranking network (LVRN) for image cropping. As illustrated in Figure 1, a refined view sampling (called RoIRefine) extracts high-resolution features to rank candidate views with listwise loss.

A. LISTWISE VIEW RANKING

To address the shortcut of pairwise approaches, we formulate composition learning as a listwise ranking problem. In this paper, the proposed model listwisely ranks the candidate views and picks up the best one.

Given a set of annotation images $D = \{d_1, d_2, \dots, d_m\}$, each image d_i consists of a list of candidate views $V_i = \{v_i^1, v_i^2, \dots, v_i^n\}$, where m is the number of images and n is the number of views. For each view v_i^n in the i -th image, a rank score y_i^n is labeled to represent the relative degree of view composition. For instance, the number of views n is 24 in CPC dataset labeled with listwise protocol. In the view

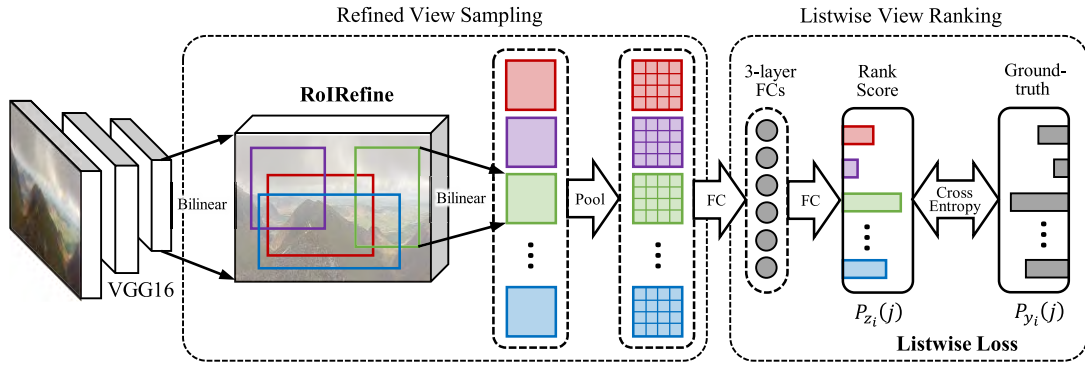


FIGURE 1. The framework of listwise view ranking for image cropping. The model first applies VGG16 [6] (truncated before the last max pooling layer) as the backbone to extract the feature of input images. Given a series of candidate boxes, the refined view sampling integrates RoI features in each bounding boxes. And then we use 3-layer full connection (FC) to generate the ordering distribution, and apply cross entropy to measure the listwise distance.

ranking network, we denote the rank function as $f(\cdot)$, which takes a view v_i^j (sampled from image d_i) as input and then outputs a rank score $f(v_i^j)$. For the i -th image, we can obtain a list of scores $Z_i = \{f(v_i^1), f(v_i^2), \dots, f(v_i^n)\}$ from the list of views V_i . Therefore, the ranking function $f(\cdot)$ can be optimized by minimizing the loss between Z_i and ground-truth scores Y_i .

Instead of pairwise approaches carefully selecting the training pairs, listwise learning removes the training bias as all candidate views are seen in each iteration. Even so, there are still a few view biases in the list – the best composed view is important than the worse ones. To address this problem, a nonlinear transformation $\Phi(\cdot)$ is adopted to amplify the effect of the best one. We define $\Phi(\cdot)$ as a common increasing function $\exp(\cdot)$:

$$\Phi(f(v_i^j)) = \exp(f(v_i^j)) \quad (4)$$

According to Eq. (3), we rewrite the output scores to Top-1 probability as

$$P_{Z_i}(j) = \frac{\exp(f(v_i^j))}{\sum_{k=1}^n \exp(f(v_i^k))}. \quad (5)$$

Similarly, the ground-truth scores is rewrote as P_Y . Following [12], we employ cross entropy as metric to minimize the distance between output probability P_Z and ground-truth probability P_Y . The loss function is defined as

$$L(Y, Z) = - \sum_{i=1}^m \sum_{j=1}^n P_{Y_i}(j) \log(P_{Z_i}(j)). \quad (6)$$

The ranking function $f(\cdot)$ can be simply found by minimizing the loss function $L(Y, Z)$. Once the ranking function $f(\cdot)$ is learned, we simply use it to calculate the rank scores and crop the images from candidate views. The Image Cropping method is shown in Algorithm 1.

B. REFINED VIEW SAMPLING

Coarse features extracted from CNN backbone limit the performance of image cropping. Previous methods generate

Algorithm 1 Listwise View Ranking Algorithm

Input:

Image Data d
Candidate boxes $B = \{b_1, b_2, \dots, b_n\}$;
Ranking model $f(\cdot)$;

Output:

Top-1 Candidate view v_{top1}

- 1: according to the candidate boxes, crop views $V = \{v_1, v_2, \dots, v_n\}$ from Image Data d
- 2: calculate the ranking scores $S = \{s_1, s_2, \dots, s_n\}$ of candidate views V , $s_i = \frac{\exp(f(v_i))}{\sum_{k=1}^n \exp(f(v_k))}$
- 3: calculate the index of ranking scores S : $Index = \text{argmax}(S)$
- 4: select the top-1 candidate view: $v_{top1} = V[Index]$
- 5: **return** v_{top1} ;

candidate views from images and then warp them to a fixed size (e.g., 227×227 in VFN). However, warping is not suitable for composition learning and make the view deformed. The deformation of feature seriously damages the common composition rule, such as golden ratio, golden spiral and rule of thirds. In additional, the rescaling and multiple down-sampling in the CNN backbone make the model insensitive to view contents.

For view generation, there are three common RoI-aware operation shown in Figure 2:

- **RoIPool** [13] (Figure 2(a)) is a standard operation for extracting a small feature map from each RoI. The quantized RoI is subdivided into spatial bins, and finally feature values covered by each bin are aggregated. The quantizations introduce misalignments between the RoI and the extracted features.
- **RoIAlign** [5] (Figure 2(b)) removes the harsh quantization of RoIPool, properly aligning the extracted features. In each RoI, RoIAlign uses bilinear interpolation to compute the exact features at four regularly sampled locations, and aggregates the RoI features using max/average pooling.

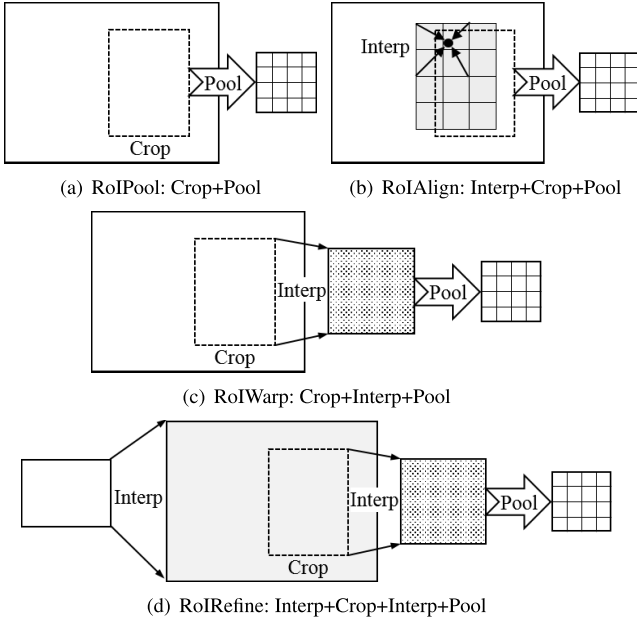


FIGURE 2. Illustrations of RoI-aware operations and our RoIRefine.

- **RoIWarp** (Figure 2(c)) operation is proposed in [16]. Unlike RoIAlign, RoIWarp crops a feature map region and warps it into a target size by interpolation. Even though RoIWarp also adopts bilinear resampling, it overlooks the alignment of floating-number RoI.

These RoI-aware operations are widely used in object detection and instance segment, but unsuitable for image cropping. Inspired by RoIAlign and RoIWarp, we propose an **RoIRefine** layer shown in Figure 2(d) to extract high-quality features for reducing deformation. Our proposed change is simple: we sample the full-map features and resample the RoI-aware features to reduce deformation. The first bilinear interpolation improves the sampling resolution. Although the first bilinear interpolation does not increase additional information, the improvement of resolution makes the features sensitive to floating-number RoI. Without the first interpolation, we cannot achieve the float coordinate in the feature map, which means candidate boxes shift or rescale. In the other word, interpolation implements finer sampling with float-quantization. The second resampling avoids inconsistent between the feature maps and candidate views. In this paper, we simply upsample the full-map features to $2\times$ size and resample the RoI-aware features to the size of 14×14 . Considering the trade-off between performance and efficiency, $2\times$ upsampling is the best choice as larger scale upsampling ($4\times$ or $8\times$) hardly improves performance. Compare to previous RoI-aware operations, RoIRefine leads to large improvements as shown in Section IV-C2.

C. IMPLEMENTATION

In this paper, we initialize the backbone CNN with VGG16 pre-trained on ImageNet [7]. All weights of the three FC layers are initialized with normal distribution (zero mean and 0.01 standard deviation), bias are set to zero and the

channels are set to 1024, 512 and 1, respectively. The proposed model is trained on CPC dataset [4] including 10,797 images, each with 24 candidate views. We directly rank 24 views and assign the order of views as ground-truth rank score.

During training, the images are resized to 224×224 regardless of its original size. Resizing the original image to fixed size is to fit the VGG-16 pretrained on 224×224 images (ImageNet), and is beneficial to model fine-tuning. Although resizing the original image does result global deformation, but its effect is weak. Global deformation does not affect listwise ranking because the ranking objects are views instead of original images. Every candidate views in one list only have the same global deformation, and listwise ranking loss is not sensitive to global deformation.

We trained the network for 10 epochs using stochastic gradient descent (SGD) with momentum of 0.9 and learning rate of 0.001 that decays by 0.1 after 4 epochs. The batch-size is set to 50 that means each mini-batch including 50×24 candidate views cropped from 50 images. Early stopping was adopted based on validation results on FCDB dataset [2].

IV. EXPERIMENTS

We validate the effectiveness of the proposed model on two public image cropping databases (FCDB [2] and FLMS [14]). We also compare the time efficiency on a GPU to existing image cropping models in Table 4.

A. EXPERIMENTAL SETTINGS

To evaluate our model, we utilize the sliding window strategy of [1] to generate candidate views and choose the views with best rank score $f(v_i^j)$. Here we set the size of search windows among $[0.6, 0.65, 0.7, \dots, 0.9]$ and the aspect ratio among $[1:1, 3:4, 4:3, 9:16, 16:9]$. To refine candidate views, we adopt non-maximum suppression (NMS) based on overlap ratio between candidate views and original image to generate 1,745 candidate boxes.

1) FCDB DATASET

FCDB contains 348 test images and each image is labeled by a photography hobbyist. To evaluate the generalization ability of our model, we adopt the same metrics as previous works [2], [4], including intersection-over-union (IoU) and boundary displacement (Disp). The IoU can be computed as

$$IoU = \frac{Area_i^{gt} \cap Area_i^{pred}}{Area_i^{gt} \cup Area_i^{pred}}, \quad (7)$$

where $Area_i^{gt}$ and $Area_i^{pred}$ denote the area of the ground-truth and best-ranking crop view, respectively. Boundary displacement is given by

$$Disp = \sum_{j=1}^4 \frac{\|\hat{B}_i^j - B_i^j\|}{4}, \quad (8)$$

TABLE 1. Quantitative evaluation on FCDB dataset.

Methods	Avg. IoU	Avg. Disp
AesRankNet [17]	0.4843	0.1400
RankSVM [2]	0.6020	0.1060
VFN+SW [1]	0.6328	0.0982
A2-RL [3]	0.6633	0.0892
VPN [4]	0.6641	0.0858
LVRN (Ours)	0.7100	0.0735

TABLE 2. Quantitative evaluation on FLMS dataset.

Methods	Top-1 Max IoU
Fang et al. [14]	0.6998
VFN+SW [1]	0.7265
ABP+AA [8]	0.8100
A2-RL [3]	0.8204
VPN [4]	0.8233
LVRN (Ours)	0.8434

where $\hat{B}_i^j$ and B_i^j denote the four corresponding edges between the ground-truth and best crop view, respectively.

2) FLMS DATASET

FLMS contains 500 test images and each image has 10 annotations from 10 different persons. The evaluation metric is a little different as it has more annotations for each image than FCDB. Following previous methods, Top-1 maximum IoU is chosen as the evaluation metric. Top-1 means to pick up the best cropping views to compute the result. We compute the IoU between the ground-truth and Top-1 views, and then choose the maximum IoU as final results.

B. QUANTITATIVE EVALUATION

In this section, we study the cropping accuracy of our model with the state-of-the-art methods. We evaluate the performance on FCDB and FLMS dataset. VFN uses the ground truth window as the candidate views which leads to remarkable improvement, and VPN performs a post-processing by discarding small views to improve performance. For comparison fairness, the results (shown in Table 1 and Table 2) are evaluated without ground-truth windows and post-processing as [3].

1) FCDB DATASET

As shown in Table 1, we evaluate the cropping performance on FCDB dataset. Besides of the methods discussed above (VFN, A2-RL and VPN), we choose two other pairwise learning-to-rank methods as baselines. AesRankNet [17] is proposed to rank photo aesthetics modeled by a pairwise loss function. RankSVM [2] uses AlexNet to extract aesthetic features and find the best cropping window among candidate views. According to Table 1, the proposed model achieves the best IoU and Disp scores compared to the others.

2) FLMS DATASET

We also evaluate on FLMS dataset and the results are shown in Table 2. Following [3], we choose Top-1 maximum IoU (Max IoU) as metric to represent cropping accuracy.

TABLE 3. Candidate number analysis on FCDB dataset.

Methods	Avg. Candidate	Avg. IoU	Avg. FPS
VFN+SW	137	0.6328	0.77
VFN+SW+	500	0.6395	0.22
VFN+SW++	1125	0.6442	0.10
LVRN (Ours)	344	0.6773	197
	919	0.6841	153
	1745	0.7100	125

TABLE 4. Efficiency evaluation on FCDB dataset.

Methods	Avg. Candidate	Avg. IoU	Avg. FPS
VFN+SW	137	0.6328	0.77
A2-RL	13.56	0.6633	4.08
VPN	895	0.6641	75
LVRN (Ours)	1745	0.7100	125

In addition to ranking-based methods, two classification-based methods are also compared on FLMS dataset. Fang et al. [14] learns an aesthetic-based cropping model by discriminative classifier training. In [8], attention box prediction (ABP) network and aesthetics assessment (AA) network are proposed to model the photo assessment problem as aesthetic quality classification. From experiments in Table 2, we can see that our model outperforms other methods in cropping accuracy.

3) TIME EFFICIENCY

To validate the time efficiency, we compare the time cost between our model and the state-of-the-art methods (VFN, A2-RL and VPN) on FCDB dataset. All the results in Table 4 are evaluate on the same perform with one NVIDIA GeForce 1080 GPU.

The selection of candidate views plays an important role for image cropping. In Table 4, *Candidate* means the number of bounding boxes used to find the best view (in VFN and VPN) or extract the evaluation feature (in A2-RL). In general, the model using most candidate views in evaluation can most likely find the best results shown in Table 3. From Table 4, the proposed method uses the **most candidate views** in the **least time** (120+ frames per second) and achieves the **best accuracy**.

C. PERFORMANCE ANALYSIS

1) PERFORMANCE OF LISTWISE LEARNING

To illustrate the effectiveness of listwise learning, we design a contrast experiment shown in this section. As described in Section III, we build the VGG16-based networks without RoIRefine using different ranking losses (pairwise and listwise). We train the networks on CPC dataset and compute the rank scores for all candidate views once time. Following [1], [4], the pairwise loss is defined as

$$L(v_i^1, v_i^2) = \max\{0, 1 + f(v_i^1) - f(v_i^2)\}, \quad (9)$$

where v_i^1 and v_i^2 is two views selected in the same image d_i , and v_i^1 is preferred more than v_i^2 . For listwise training,

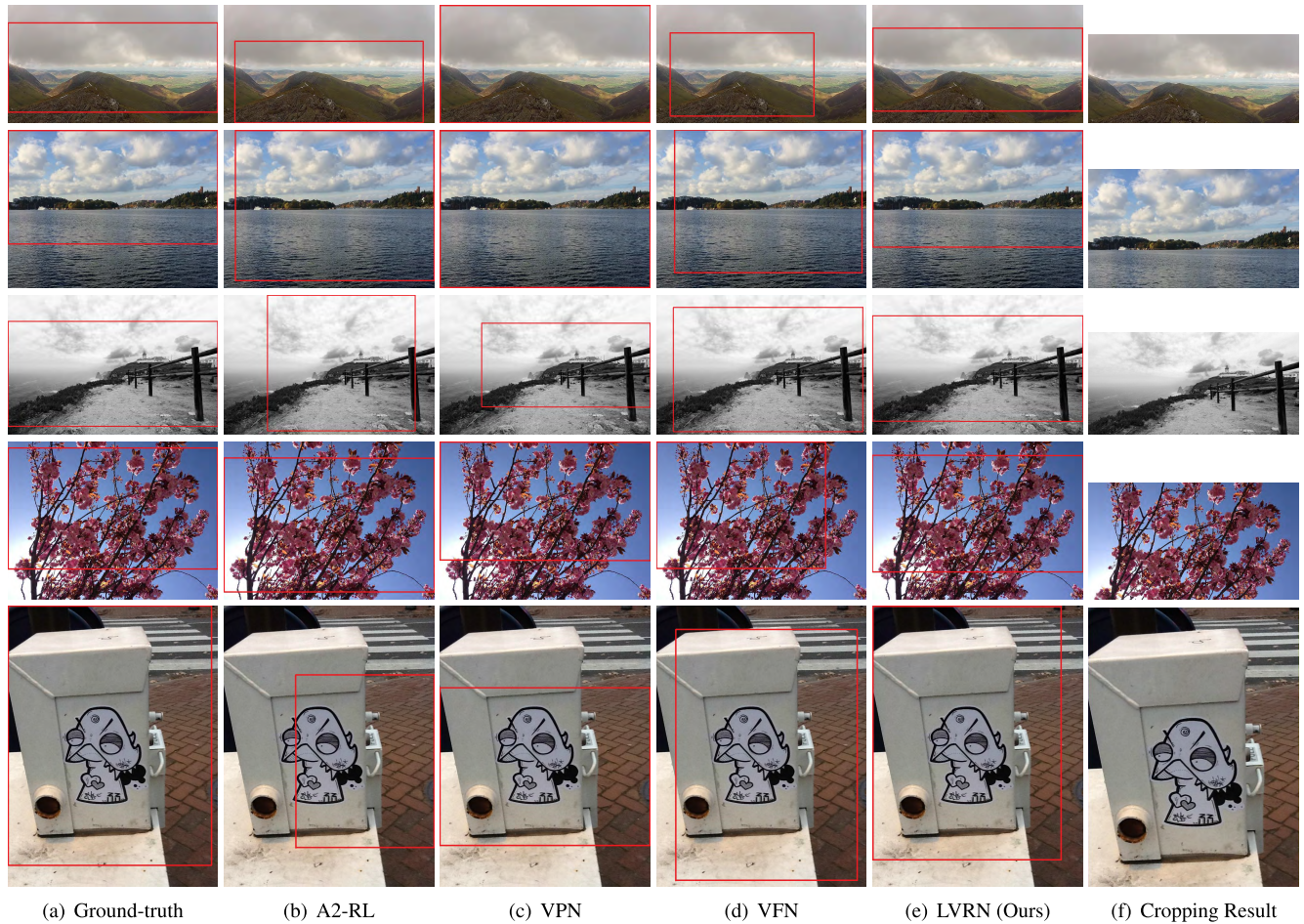


FIGURE 3. Qualitative visualization on FCDB dataset. In columns (a) to (e), the best crops determined by different methods are drawn as red rectangles. The column (f) shows the cropping results of our model.

TABLE 5. Performance analysis of listwise learning on FCDB dataset.

Ranking Loss	Avg. IoU
Pairwise + w/o selection (260W+)	0.5355
Pairwise + simple selection (130W+)	0.5568
Pairwise + careful selection	0.6080
Listwise	0.6204

the training setting is the same as the Section IV-C.2 except that RoIRefine is not used.

The result of pairwise training heavily depends on pair selection because the samples with various distribution will result training bias. In order to compare as much as detailed, we train three models with different selection methods shown in Table 5. Without pair selection, there are more than 2.6 million pairs in CPC dataset. With simple pair selection, we set a threshold 0.5 to drop the pairs with a minor gap of rank score, and generate 1.3 million pairs. With careful pair selection, we train the model following [4]. The markedly improvement in Table 5 shows that listwise training overcomes the problem of the pairwise training.

2) PERFORMANCE OF ROIREFINE

In this section, we study the improvement of the proposed RoIRefine. In order to show the contribution of using the

TABLE 6. Performance analysis of RoIRefine on FCDB dataset.

RoI Operation	Avg. IoU(Pairwise)	Avg. IoU(Listwise)
w/o	0.6080	0.6204
RoIPool	0.6526	0.6706
RoIAlign	0.6709	0.6956
RoIWarp	0.6732	0.6997
RoIRefine	0.6882	0.7100

listwise loss function when using the better view sampling(RoIRefine), we train and evaluate eight models using different ranking loss and different RoI-aware operations. The experiment results of RoIRefine and three RoI-aware operations on FCDB are shown in Table 6. The differences between these RoI-aware operations is the number and place of interpolation shown in Figure 2. RoIPool aggregates the view feature after RoI-crop without any interpolation; RoIAlign aligns the feature using interpolation before RoI-crop; RoIWarp resamples the feature using interpolation after RoI-crop. Inspired by RoIAlign and RoIWarp, RoIRefine adopts bilinear interpolation before and after RoI-crop.

RoI-aware operations reduce the deformation caused by traditional view generation and achieve the markedly improvement about 5.0% IoU. Without interpolation,

RoIPool results feature deformation and achieves the worst performance of four RoI-aware operations. RoIAlign and RoIWarp removes the harsh quantization of RoI boundaries, and improve IoU by about 2.0% to 2.5% over RoIPool. RoIRefine combines pre-interpolation (RoIAlign) and post-interpolation (RoIWarp) to refine the view feature, and achieves a gain about 1.5% IoU than RoIAlign and 1.0% IoU than RoIWarp. The experiment results demonstrate that high-quality features extracted by RoIRefine can overcome the problems of rescaling and deformation.

D. QUALITATIVE VISUALIZATION

As shown in Figure 3, there are five groups of qualitative results generated by different methods on FCDB dataset. Obviously, it is very intuitive comparison that our model can extract better view than the others.

For A2-RL (Figure 3(b)), reinforcement learning is sensitive to initial status and iteration step, resulting unstable performance shown in the second and fifth images. VPN (Figure 3(c)) uses 895 anchor boxes including the origin image, and tends to select the full image shown in the first two images. Because of high computational complexity, VFN cannot apply a mount of candidate views to achieve high-accuracy results shown in (Figure 3(d)). Comparing Figure 3(a) and Figure 3(e), we can see that our predicted boxes are close to ground-truth. In the last column, the results cropped by our method have better visual quality than the origin images.

V. CONCLUSIONS

Image cropping is a common photo manipulation process, which improves the overall composition by removing unwanted regions. In this paper, we formulate the learning of photo composition as a list-wise ranking problem to overcome the problem of pairwise-based approaches. Furthermore, a novel RoIRefine operation is proposed to extract high-quality features for view generation. The experiment results on two common datasets show that our method creates new state-of-the-art results with faster speed of 120 + FPS.

In the future work, we will study multi-task learning to combine composition evaluation and boxes regression. Unfortunately, how to design the multi-task loss is a problematic issue. Inspired of the success of detection framework, RCNN-like [13] or SSD-like [18] method will be our first choice.

REFERENCES

- [1] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the Web," in *Proc. 25th Int. Conf. Multimedia*, 2017, pp. 37–45.
- [2] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen, "Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 226–234.
- [3] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-RL: Aesthetics aware reinforcement learning for image cropping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8193–8201.

- [4] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5437–5446.
- [5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [8] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2205–2213.
- [9] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, 2009.
- [10] K. Crammer and Y. Singer, "Pranking with ranking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 641–647.
- [11] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 89–96.
- [12] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 129–136.
- [13] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [14] C. Fang, Z. Lin, R. Mech, and X. Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *Proc. 22nd Int. Conf. Multimedia*, 2014, pp. 1105–1108.
- [15] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *J. Mach. Learn. Res.*, vol. 4, pp. 933–969, Nov. 2003.
- [16] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3150–3158.
- [17] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 662–679.
- [18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
- [19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: <https://arxiv.org/abs/1503.02531>
- [20] H. Li, R. Hong, D. Lian, Z. Wu, M. Wang, and Y. Ge, "A relaxed ranking-based factor model for recommender system from implicit feedback," in *Proc. IJCAI*, 2016, pp. 1683–1689.
- [21] T. Yao, F. Long, T. Mei, and Y. Rui, "Deep semantic-preserving and ranking-based hashing for image retrieval," in *Proc. IJCAI*, 2016, pp. 3931–3937.
- [22] J. Xuan and M. Monperrus, "Learning to combine multiple ranking metrics for fault localization," in *Proc. IEEE Int. Conf. Softw. Maintenance Evol.*, Sep./Oct. 2014, pp. 191–200.
- [23] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini, "Self-adaptive image cropping for small displays," *IEEE Trans. Consum. Electron.*, vol. 53, no. 4, pp. 1622–1627, Nov. 2007.
- [24] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2232–2239.



WEIRUI LU is currently pursuing the B.S. degree with the School of Electronic and Information Engineering, South China University of Technology, China. His research interests include computer vision, machine learning, and image processing.



XIAOFEN XING received the B.S., M.S., and Ph.D. degrees from the South China University of Technology, China, in 2001, 2004, and 2013, respectively, where she has been an Associate Professor with the School of Electronic and Information Engineering, since 2007. Her main research interests include image/video processing, human computer interaction, and video surveillance.



XIANGMIN XU (M'13) received the Ph.D. degree from the South China University of Technology, Guangzhou, China, where he is currently a Full Professor with the School of Electronic and Information Engineering. His current research interests include image/video processing, human computer interaction, computer vision, and machine learning.

...



BOLUN CAI received the M.Eng. and Ph.D. degrees from the South China University of Technology, China, in 2016 and 2019, respectively. He is currently a Senior Researcher with Tencent WeChat AI. His research interests include computer vision, machine learning, and image processing.