Supplementary Materials for Image Cropping with Spatial-aware Feature and Rank Consistency

Chao Wang, Li Niu*, Bo Zhang, Liqing Zhang*
MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{wangchaojfj,ustcnewly,bo-zhang}@sjtu.edu.cn,zhang-lq@cs.sjtu.edu.cn

In this document, we provide more details as supplementary materials to our main submission. We first present the performance of the pair-wise ranking classifier in Section 1, proving that ranking knowledge is transferrable from labeled images to unlabeled images. Then, we compare our ranking consistency method with the previous works [2, 4] utilizing unlabeled data in Section 2. We further analyze the spatial-aware feature by changing the scale and position of the candidate crop to observe how placing the crop can get a high aesthetic score and show some cropping results in Section 3. We also analyze the influence of different hyper-parameter values in Section 4. In addition, more quantitative comparisons are shown in Section 5 . In the end, some failure cases are presented in Section 6.

1. Performance of Pair-wise Ranking Classifier

As described in Section 3.3 in the main paper, we train a pair-wise ranking classifier to transfer the ranking knowledge from labeled images to unlabeled images. Thus, the performance of the ranking classifier is important as the ranking knowledge can be transferred only if the classifier is well-trained. Otherwise, it may even harm the performance of the cropping model. We train the classifier using the labeled images in the training set of GAICD [5] and test its classification accuracy on both the training set and test set. We also report the ranking accuracy based on the predicted crop scores, which is referred to as basic model accuracy. Specifically, we train our model without rank consistency on the training images and apply the model to predict the crop scores in test images. Then, we obtain the rank of two crops according to their crop scores.

The results are shown in Table 1. We can observe that the ranking accuracy of the pair-wise classifier is better than the basic model(our model without rank consistency) in both the training set and the test set. Since the pair-wise ranking classifier is more concentrated on learning the relative ranks of candidate crops when trained on the training set, its ranking accuracy is higher than the basic model on the training set. When applying the trained classifier to the test set, the accuracy gap between the training set and the test set is smaller than that of basic model accuracy, and the ranking accuracy is higher than the basic model on the test set, which proves that the ranking knowledge is transferrable from the training set to test set. With transferred ranking knowledge, the rank consistency regularization on the unlabeled data is useful and can help improve the performance of the cropping model.

Model	Dataset	Ranking Accuracy
pair-wise classifier	training set	92.2%
basic model	training set	89.3%
pair-wise classifier	test set	88.7%
basic model	test set	82.8%

Table 1. The ranking accuracy on GAICD [5] training set and test set of our pair-wise ranking classifier and basic model. Basic model means we calculate the ranking accuracy based on the crop scores predicted by our model without using rank consistency.

2. Alternative Approaches to Use Unlabeled Data

In this section, we compare our ranking consistency method with other alternative approaches to use unlabeled data. We treat our model without rank consistency as the basic model. Then, we equip the basic model with different approaches to use unlabeled data.

Firstly, we use unlabeled test images in the same way as VFN [2]. Specifically, for each test image, we follow [2] to generate candidate crops including 8 boarder crops, 6 square crops, and one crop of the whole image. We send the images and the generated crops to the model to predict aesthetic scores. We adopt the same loss function as in [2], which assumes that the aesthetic score of the whole image is higher than the scores of other crops.

Secondly, we use unlabeled test images in the same way as [4], which is similar to commonly used self-training.

*Corresponding author

Unlabeled	$\overline{SRCC}\uparrow$	$\overline{PCC}\uparrow$	$\overline{Acc_5}\uparrow$	$\overline{Acc_{10}}\uparrow$
-	0.865	0.889	63.7	82.6
Ours	**0.872**	**0.893**	**64.8**	**83.3**
[2]	0.863	0.886	63.2	82.5
[4]	0.864	0.887	64.0	82.7

Table 2. The results obtained by using different approaches to use unlabeled test images.

Specifically, we first train the basic model to predict the pseudo labels of candidate crops in unlabeled test images. Then, we add these test crops with pseudo labels to the training set and retrain our basic model.

The results are presented in Table 2. By comparing row 2 with row 3 and row 4, we can see that two alternative ways [2, 4] to use unlabeled test images cannot exceed our proposed rank consistency. The performance of [2] even drops slightly compared with row 1, which could be explained as follows. [2] assumes that the aesthetic score of the whole image is higher than the scores of other crops, which does not always hold. The approach [4] only achieves comparable results with row 1, probably because self-training as in [4] may not introduce new knowledge when retraining the basic model. Therefore, even in the transductive learning setting, the unreasonable or naive ways to use unlabeled test images could not bring much performance gain. In contrast, our proposed method can mine the inherent ranking knowledge, helping to transfer the knowledge more effectively.

3. Analyses of the Spatial-aware Feature

In this section, we provide some intuitive illustrations of how the spatial-aware feature judges the aesthetic quality by considering the spatial relation between candidate crops and aesthetic elements. Specifically, we vary the scale and position of the candidate crop and report the predicted scores only using spatial-aware features, which are obtained by multiplying the spatial-aware features with the corresponding weights of the last fully connected layer and adding half bias.

We also select the example image shown in Section 4.3 in our main text and the results are shown in Figure 1. Row 1 is the best crop predicted with aesthetic scores contributed by spatial-aware features. We can find that the crop encloses the semantic edges and salient objects as much as possible. Row 2 to row 5 are crops in different positions with the same size. They cut through some edges or salient objects, leading to the dropping scores. The last two rows contain the two persons but have different sizes. Their scores are higher than the above four rows. We can observe that the last row has relatively higher composition quality considering the position of the people in the image, so its score is relatively higher. The above results demonstrate how the model places the crops and verify the effectiveness of our proposed spatial-aware feature.

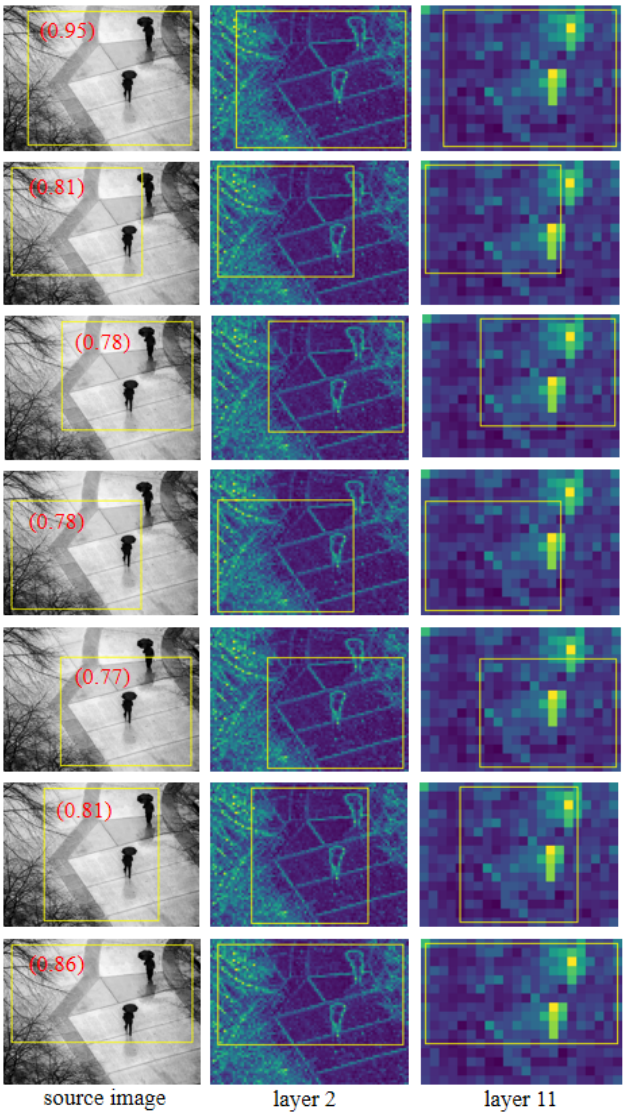

Figure 1. The visual analyses of spatial-aware features. The first column shows the crops in the source images with their aesthetic scores (in red) contributed by spatial-aware features. The second and the third columns show the feature maps of layer 2 and layer 3 respectively. The first row is the predicted top-1 crop, and we vary the crop position from row 2 to row 5. In row 6 and row 7, we vary the scale and aspect ratio of the crop.

4. Hyper-parameter Analyses

Recall that we have the following hyper-parameters: η (score margin applied when training the pair-wise classifier), P (the number of crop pairs sampled when training

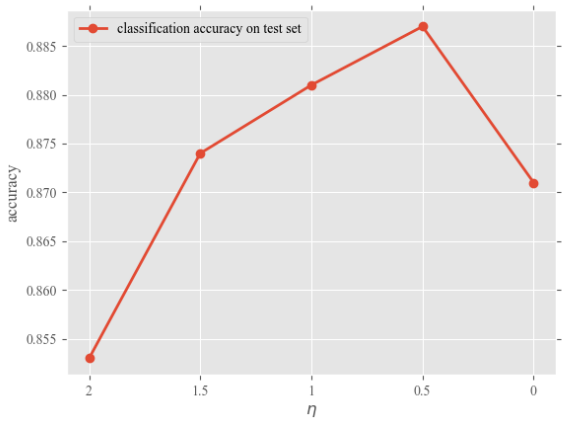

Figure 2. Classification accuracy variation on GAICD [5] test set of our method with different hyper-parameter η (score margin applied when training the pair-wise classifier).

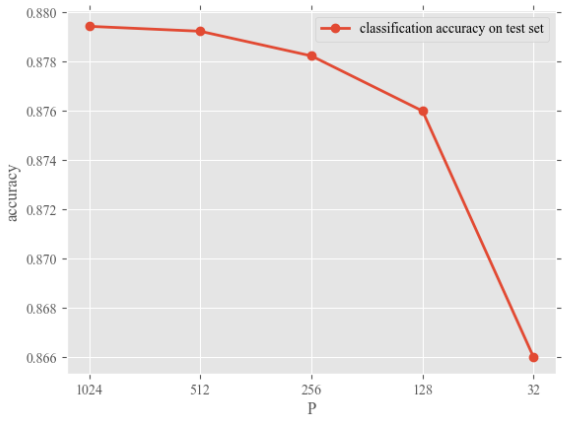

Figure 3. Classification accuracy variation on GAICD [5] test set of our method with different hyper-parameter P (the number of crop pairs sampled when training the pair-wise classifier).

the pair-wise classifier), λ_{cls} (weight of classification loss), and δ (margin in the consistency loss). We vary each hyper-parameter and fix other hyper-parameters and get the results via cross validation.

We first test different values of η and P report the classification accuracy of the pair-wise classifier as they influence the performance of the pair-wise classifier significantly. We set η in the range of $[0, 2]$ with an interval of 0.5, and $P = 32, 128, 256, 512, 1024$. We test the classification accuracy on GAICD [5] test set and the results are shown in Figure 2 and Figure 3. We can see that when $\eta = 0.5$, the classifier has the highest predicting accuracy, and it is relative stable when η is in the range [0.5, 1.5]. It can be interpreted that if the score margins are too large

λ_{cls}	$\overline{Acc_5}\uparrow$	$\overline{Acc_{10}}\uparrow$	$\overline{SRCC}\uparrow$
0.01	63.4	82.5	0.864
0.1	64.5	83.0	0.867
1	**64.8**	**83.3**	**0.872**
10	64.0	82.8	0.866
100	62.9	82.1	0.863

Table 3. Performance of our model with different λ_{cls} (weight of classification loss).

δ	$\overline{Acc_5}\uparrow$	$\overline{Acc_{10}}\uparrow$	$\overline{SRCC}\uparrow$
0	64.6	83.0	0.869
0.1	**64.8**	**83.3**	**0.872**
1	64.2	82.9	0.865
1.5	63.4	82.5	0.864
2	63.1	82.1	0.861

Table 4. Performance of our model with different δ (margin in the consistency loss).

or too small, the crop pairs selected to train the classifier may introduce some noise and disturb the classifier to rank the crops accurately. When increasing the number of crop pairs P, the performance of the classifier gets higher, but the growth rate declines dramatically as $P > 256$. We finally choose $P = 256$ for training considering the computational cost and the performance comprehensively.

Next, we test different values of λ_{cls} and δ and report the cropping metrics of our models. We separately set $\lambda_{cls} = 0.01, 0.1, 1, 10, 100$ and $\delta = 0, 0.1, 0.5, 1, 1.5, 2$. The results are shown in Table 3 and Table 4. We can see that when $\lambda_{cls} = 1$ and $\delta = 0.1$, our model has the best performance and our method is relatively robust to λ_{cls} and δ when setting them in a reasonable range.

5. More Qualitative Comparisons

Firstly, in order to gain an intuition on how each component of our model improves cropping results, we show some examples of GAICD [5] test set using the basic model(row 1 in Table 4 of the main paper) and our proposed method with only spatial-aware feature component and rank consistency component respectively in Figure 4 and Figure 5. From left to right in Figure 4, the first image shows that the basic model cuts through the tree branches which are represented as semantic edges in the feature maps. However, our proposed model shifts the crop box to the left to include the whole branch part, improving the crop score from 3.4 to 4.3 with the help of the spatial-aware feature. The same phenomenon occurs in the second image. In the third and fourth images, our model expands the cropping region to include the salient objects in the images, preventing important

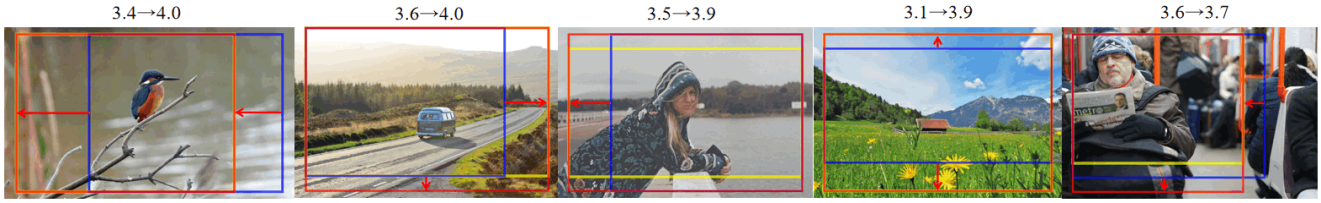

Figure 4. Qualitative comparison on GAICD [5] test set between the basic model(row 1 in Table 4 of the main paper) and our proposed method with only spatial-aware feature component. The annotated best crops are in yellow, the predicted best crops by the basic model and our proposed method are in blue and red respectively. The numbers above the images are their predicted scores.

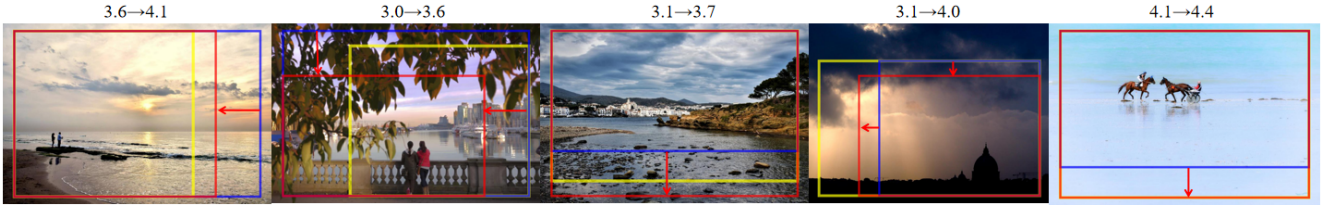

Figure 5. Qualitative comparison on GAICD [5] test set between the basic model(row 1 in Table 4 of the main paper) and our proposed method with only rank consistency component. The annotated best crops are in yellow, the predicted best crops by the basic model and our proposed method are in blue and red respectively. The numbers above the images are their predicted scores.

objects from being cut through by the cropping box and the aesthetic scores are improved through the spatial-aware feature. In the last image, the predicted crop of our proposed model shrinks to the left relative to the basic model and cuts out the non-salient person on the right as a distraction. As mentioned in the main paper, our proposed model can place important semantic edges and salient objects in a more appropriate position in the cropping box with the help of the spatial-aware feature, and try to avoid cutting through or cutting off them. In Figure 5, the images show the improved crop results by our proposed model with only rank consistency. With the transferred ranking knowledge, our model can rank crops more accurately and the aesthetic scores of predicted best crops are lifted. However, it is hard to restrictively summarize the cases in which rank consistency could improve the cropping results because the ranking is determined by many complex factors.

Secondly, we show some specific examples where we pick some images with multiple salient objects from the GAICD [5] test set and observe how our model improves the cropping results for such images. From left to right in Figure 6, the first one shows that the crop of our method preserves the two birds and the outline of the mountain as much as possible. In the second image, our method does not cut through the two doors. As in the third one, our method put the person and the houselet in better place. The last two images show that our method excludes the distractions(the tree in the fourth image and the car in the last image). Compared with the basic model, we can see that with the help of the spatial-aware feature and rank consistency, our model improves the ability to deal with multi-salient objects cropping situations.

In the end, we show some more qualitative results between our proposed method and other state-of-the-art methods on GAICD [5] dataset and FCDB [1] dataset. Similar to Section 4.3 in the main paper, we show top-1 crops obtained by VFN [2], VEN [4], VPN [4], CGS [3], GAIC [5], and our method. On the GAICD test set, we use the predefined anchor boxes [5]. As for the FCDB dataset, we use the pre-defined sliding windows as candidate crops. We select the top-1 predicted crop as the best crop. The results on the two datasets are shown in Figure 8 and Figure 9 respectively. We can find that on both datasets, our method is able to generate more appealing crops that are closer to the ground-truth than other approaches.

6. Failure Cases

As discussed in Section 5 in our main submission, although our method can produce visually appealing crops, there still exist some failure cases. In Figure 7, we show some images in GAICD [5] test set whose predicted top-1 crops (red box) are far from their ground-truth top-1 crops (yellow box).

We observe that when cropping the images of landscape, our model tends to generate broad views that preserve most of the scene. For relatively complex scenes, our model tends to include more semantic edges and salient objects compared with other methods. As shown in Figure 7, in the first two images, our method chooses the largest anchor box as the best one that encloses lines and objects as many as possible and cuts out distractions little. In the third image, the

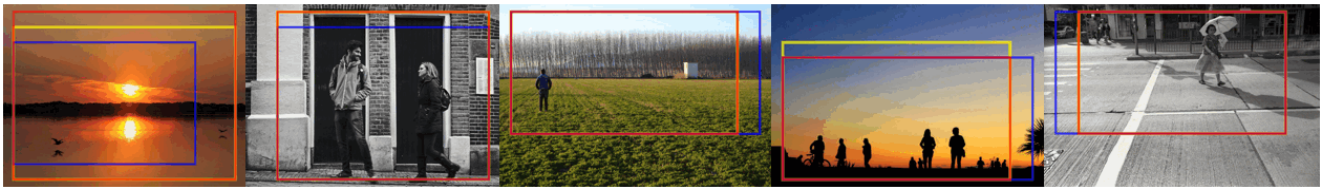

Figure 6. Qualitative comparison on images with multiple salient objects from GAICD [5] test set between the basic model(row 1 in Table 4 of the main paper) and our whole model. We present the images with their annotated best crops in yellow, predicted best crops by the basic model in blue, and predicted best crops by our proposed method in red. The predicted scores of our whole model are higher than those of the basic model.

Figure 7. Some failure cases in GAICD [5] test set. We present the images with their annotated best crops (yellow bounding box) and the predicted best crops (red bounding box) by our method.

predicted crop preserves almost all the shoreline while the annotated crop cuts through the shoreline at the right bottom corner. The last image contains a relatively complex scene and the predicted crop also encloses some distractions including the sunshade and the miniature tree. One possible explanation is that the spatial-aware feature in our cropping model improperly penalizes the crops which cut through thus semantic edges and salient objects.

References

[1] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *WACV*, 2017. 4, 6

[2] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *ACMMM*, 2017. 1, 2, 4

[3] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *CVPR*, 2020. 4

[4] 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 *CVPR*, 2018. 1, 2, 4

[5] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 3, 4, 5, 6

Figure 8. More qualitative comparison on GAICD [5] test set. We show the annotated best crop (yellow bounding box) in the source image in the left column and top-1 crops obtained by different methods in the rest of the columns.

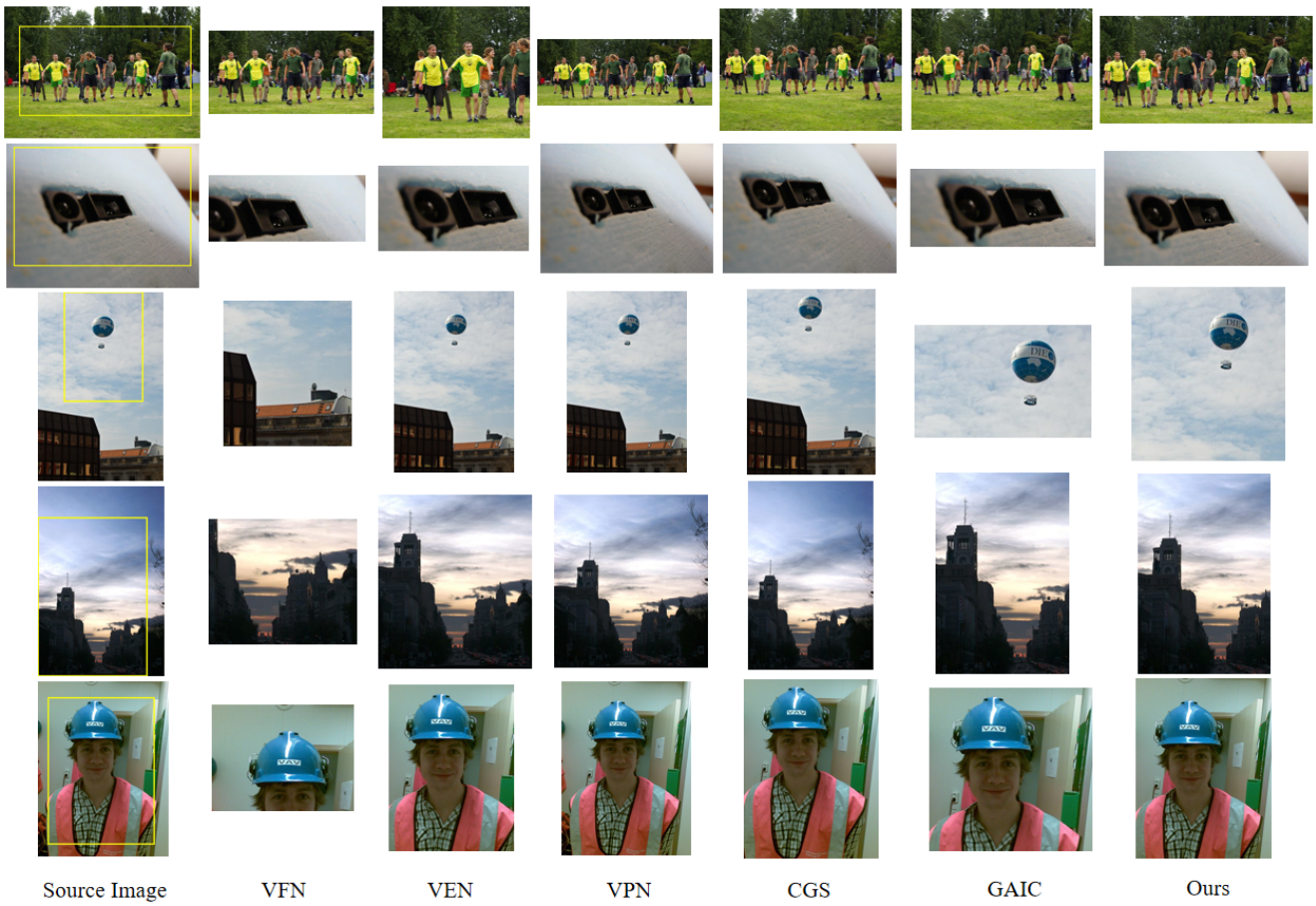

Figure 9. More qualitative comparison on FCDB [1] test set. We show the annotated best crop (yellow bounding box) in the source image in the left column and top-1 crops obtained by different methods in the rest of the columns.