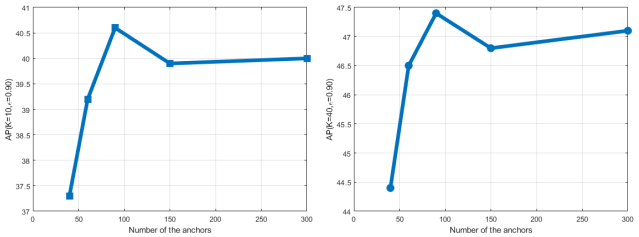
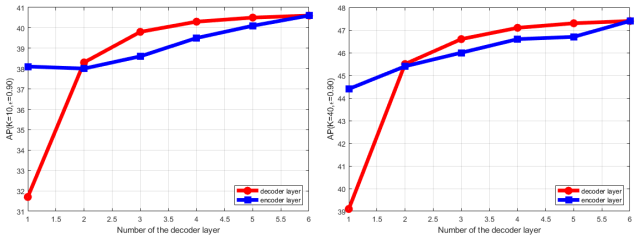
Supplementary Materials of "Rethinking Image Cropping: Exploring Diverse Compositions from Global Views"

Gengyun Jia, Huaibo Huang, Chaoyou Fu, Ran He*
School of Artificial Intelligence, University of Chinese Academy of Sciences
NLPR & CRIPAC, Institute of Automation, Chinese Academy of Sciences
{gengyun.jia, huaibo.huang}@cripac.ia.ac.cn, {chaoyou.fu, rhe}@nlpr.ia.ac.cn



(a) AP ($K = 10, \epsilon = 0.90$). (b) AP ($K = 40, \epsilon = 0.90$).

Figure 1. AP ($\epsilon = 0.90$) performances with different number of the learnable anchors.



(a) AP ($K = 10, \epsilon = 0.90$). (b) AP ($K = 40, \epsilon = 0.90$).

Figure 2. AP ($\epsilon = 0.90$) performances with different number of the encoder and decoder layers.

1. Analysis of Some Hyper-parameters

1.1. Number of Learnable Anchors

Our model uses a set of learnable anchors to regress the good crops. Theoretically, the only requirement for the anchor number N^q is that $N^q \geq N^g_{\max}$, where $N^g_{\max}$ is the maximum number of good crops in all images. Considering that the anchor number may influence the model training from multiple aspects, we test the different AP performances under different anchor numbers $\{40, 60, 90, 150, 300\}$ in the GAICv2 dataset [2] and show the results in Fig. 1. There are some notable phenomenons. Firstly, very small anchor numbers are harmful. For example, the AP ($K = 40$) at $N^q = 40$ is worse than that of $N^q = 90$ with a gap ΔAP=3.0. A possible reason is that it is difficult for small anchor numbers to learn enough information about good crops. Secondly, a very large anchor number has very slight negative influences. It can be seen that the AP ($k = 40$) performance drops with only 0.3 for anchor number 300. A possible explanation is that the focal loss automatically balances the gradients from hard and easy samples, which helps to ease the imbalance problem brought by the large anchor numbers.

1.2. Number of Transformer Encoder and Decoder Layers

The transformer decoder is the core module in the cDETR architecture. In each decoder layer, the learnable anchors absorb information from the input image features. Previous experiments use the default decoder layer number six as in [1]. In this section, we test the AP performances with different decoder layer numbers in the GAICv2 dataset and plot the results in Fig. 2. The figure shows that more decoder layers bring better results. However, only the first two layers provide significant improvements. The performances of the last two layers are nearly the same.

We further give the performances under different encoder layer numbers in Fig. 2. Unlike decoder, encoder only aims to learn global relations of input images. It is shown from the figure that although more encoder layers bring better results. The improvements are not as significant as decoder for the first two layers.

1.3. Forms of The Smoothing Mapping Function M

In the quality-guided label smoothing, a mapping function M is defined to turn the quality scores into soft labels. A truncated linear function is used in our model. In this section, we discuss another two types of the mapping function:

*corresponding author

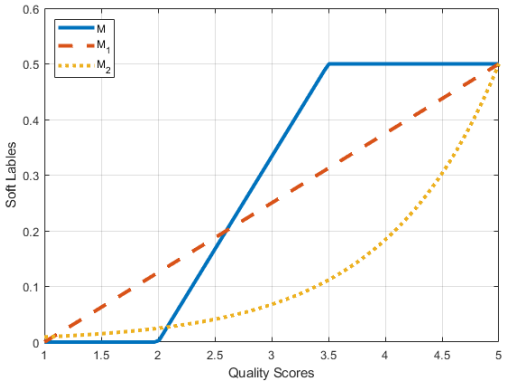


Figure 3. Different quality-label mapping functions.

Table 1. AP Performances of the three different mapping functions on the GAICv2 dataset.

Mapping Functions	AP ($\epsilon = 0.85$)		AP ($\epsilon = 0.90$)	
	$K = 10$	$K = 40$	$K = 10$	$K = 40$
M	**50.5**	56.8	**40.6**	**47.4**
M_1	50.3	**57.0**	39.2	46.5
M_2	49.7	55.3	38.8	45.6

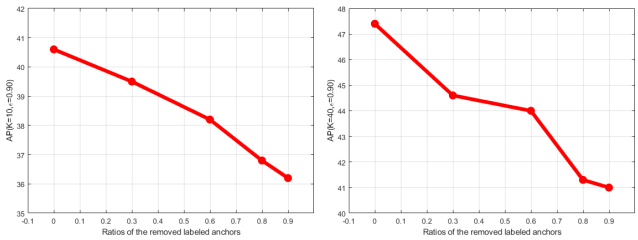
$$\widetilde{v}_i = M_1(s_i) = \mu \frac{s_i - s^{\min}}{s^{\max} - s^{\min}}, \tag{1}$$

$$\widetilde{v}_i = M_2(s_i) = \mu e^{(s_i - s^{\max})}, \tag{2}$$

M_1 is a linear function without truncation, M_2 is an exponential function. In the GAICv2 dataset, we have $s^{\min} = 1$ and $s^{\max} = 5$ and set $\mu = 0.5$. The two mapping functions are plotted in Fig. 3. The main difference between them and the original mapping function M is that they have no truncation in the quality score range. The comparison results listed in Table 1 give the following conclusions. Firstly, The exponential mapping function gives the worst performances. Secondly, the linear mapping function M_1 and the truncated linear mapping function M perform similar at $\epsilon = 0.85$. Thirdly, the truncated linear mapping function M obtains the best results at $\epsilon = 0.90$. These phenomenons show that the truncation is an important design, and the linear mapping function is better than the exponential mapping function.

2. Analysis of Labeled Anchors

Two label smoothing methods are designed for different situations in our model. The quality guidance method is used when the training images have nearly dense anchors labeled with quality scores. Insufficient labeled anchors will make it impossible to estimate the quality scores of some predicted invalid crops using local redundancy property. We conduct experiments to show the influences of different density levels of labeled anchors. Specifically, the labeled anchors are randomly removed with different ratios. The corresponding AP performances are given in Fig 4. It is shown that the performances drop significantly with more removed labeled anchors.



(a) AP ($K = 10, \epsilon = 0.90$). (b) AP ($K = 40, \epsilon = 0.90$).

Figure 4. AP ($\epsilon = 0.90$) performances with different ratio of removed labeled anchors.

Table 2. User Study.

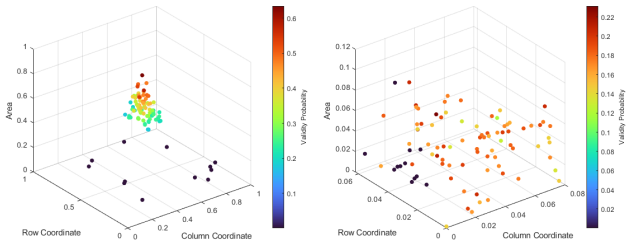
Models	GAICv2	Ours	VEN	VPN
Votes	158	**250**	232	243

3. Analysis of Learned Anchors

In our set prediction model, we use the learnable anchors to regress crops. It is worth studying the characters of these learned anchors. Therefore, we calculate the statistics of the crop center coordinates, the crop areas, and the validity probabilities for each of the 90 anchors on the GAICv2 testing images. The mean values and the standard deviations are shown in two scatter diagrams in Fig. 5a and Fig. 5b, respectively. The X, Y, and Z axes represent the center column coordinate, center row coordinate, and normalized area, respectively. Different colors of the scatter points indicate different validity probabilities. We can draw two conclusions from the figures. Firstly, Each anchor has its own focuses on the cropping location and the area. Fig. 5b shows that the regressed crops from one anchor only have slight differences on different images. Secondly, crops with higher areas and closer to the image center have higher validity probabilities. This phenomenon is consistent with the features of the GAICv2 dataset. Most images in this dataset are high-quality, making the qualities of large center crops always not bad. Besides, there is no small labeled crop in the training data. It is difficult for the model to learn to generate good small crops in this situation.

4. User Study

Considering that image cropping is a subjective task, we compare the proposed model trained on the CPC dataset with other models using user study. Specifically, 15 im-



(a) Mean values. (b) Standard deviations.

Figure 5. Scatter diagram of the anchor statistics. Each point represents a learned anchor. The X, Y, Z axes are the column coordinate, the row coordinate and the normalized area of the crops, respectively. Different colors represent different validity probabilities.

ages are randomly selected from the AVA dataset and 25 people are invited to select crops. Three models including GAICv2, VEN and VPN are employed as competitors. For each model, we select the top-2 crops based on the evaluated quality scores or the validity probabilities. Therefore, there are 8 crops for each image. Each user needs to select no more than 4 crops. Table 2 gives the voting results. It is shown that our model achieves the best results.

5. Visualization of More Cropping Results

We give more cropping results in Fig. 6. The examples show that our model generates multiple good crops covering different aspect ratios, scales, and objects, and the redundant elements are removed. For example, the crops in the second row cover aspect ratios from smaller to larger than one. The second crop in the fifth row focuses on the left man in the two people. The three crops in the sixth row successfully remove the redundant objects in the bottom right of the original image.

References

[1] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. *arXiv preprint arXiv:2108.06152*, 2021. 1

[2] 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

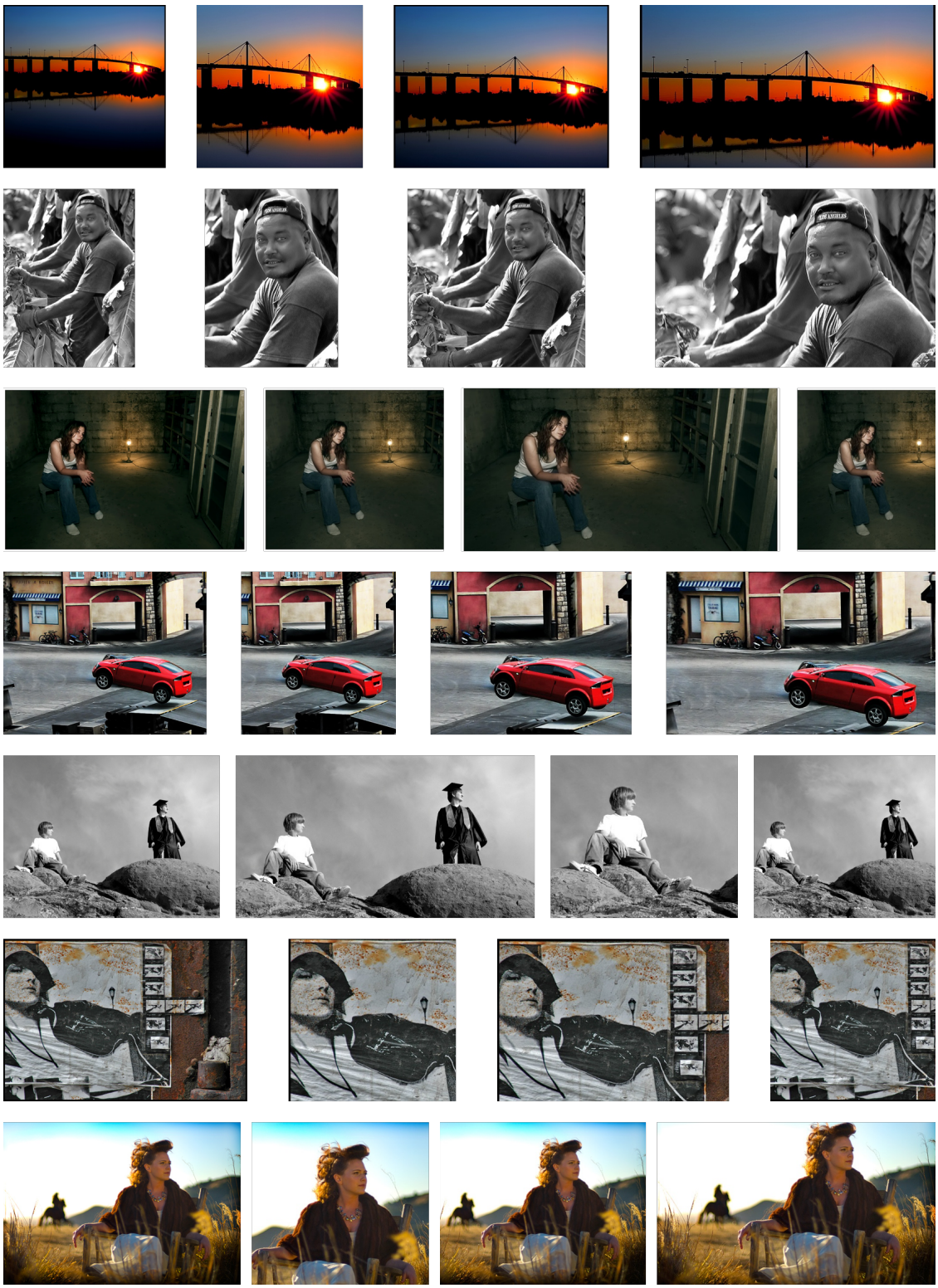


Figure 6. Some cropping results. The first image in each row is the original input image, and the rest three are the cropped images selected from the top-10 validity score predictions. We use the model trained on the CPC dataset.