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

Abstract

Existing image cropping works mainly use anchor evaluation methods or coordinate regression methods. However, it is difficult for pre-defined anchors to cover good crops globally, and the regression methods ignore the cropping diversity. In this paper, we regard image cropping as a set prediction problem. A set of crops regressed from multiple learnable anchors is matched with the labeled good crops, and a classifier is trained using the matching results to select a valid subset from all the predictions. This new perspective equips our model with globality and diversity, mitigating the shortcomings but inherit the strengthens of previous methods. Despite the advantages, the set prediction method causes inconsistency between the validity labels and the crops. To deal with this problem, we propose to smooth the validity labels with two different methods. The first method that uses crop qualities as direct guidance is designed for the datasets with nearly dense quality labels. The second method based on the self distillation can be used in sparsely labeled datasets. Experimental results on the public datasets show the merits of our approach over state-of-the-art counterparts.

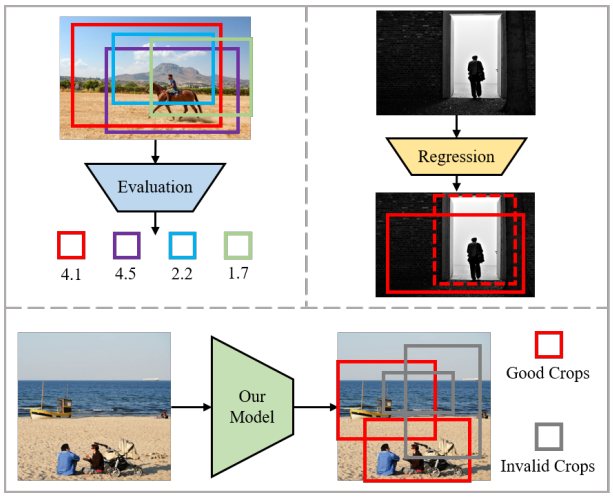


Figure 1. A diagram of different image cropping models. The anchor evaluation model (top left) can output diverse crops from pre-defined anchors, but these anchors cannot encompass good crops globally. The coordinate regression model (top right) crops from global image, but just outputs one good crop and ignores the others (e.g. the box with red dashed line). Our model (bottom) combines their strengths and overcomes their weaknesses, leading to diverse crops from global views.

1. Introduction

Image cropping has been widely used to improve image composition. Automatic image cropping is developed to make the technique friendly to amateurs and non-specialists. Previous works [3, 7, 26, 31, 47] usually merge the expert knowledge such as the "Rule of Thirds" into the models to guide the cropping. Such methods enable interpretability of the cropping process but are weak in learning sophisticated features. In recent years, many data-driven approaches built upon deep CNNs have been proposed. These methods are roughly categorized into anchor evaluation methods [4, 17, 20, 37–40, 43, 44] and coordinate regression methods [8, 11, 14–16, 24]. The former assigns quality scores to the pre-defined anchors. The latter directly regresses the coordinates of one crop on the input images. The anchor evaluation methods show the merits of generating diverse good crops. As shown in the top-left of Fig. 1, users can select an arbitrary number of crops according to the evaluated scores. But most anchor generation rules cannot search the cropping spaces globally, resulting in the possible omission of good crops. On the contrary, coordinate regression methods use a global view to cover all possibilities. But they only predict one crop from an image, which is insufficient for most images as shown in the top-right of Fig. 1. Besides, learning one best crop may cause ambiguity

*corresponding author

because it potentially assumes other crops are all bad.

In summary, cropping images is like picking fruits from a fruitful tree. Anchor evaluation methods give us many fruits to select (*Diversity*), but they only search part of the branches. Coordinate regression methods check every twigs (*Globality*) but only pick one fruit. Naturally, a question should be asked: *Can we combine the different advantages of the two methods to achieve both the Diversity and the Globality?* Inspired by the recently developed object detection models [2], we provide a new perspective by regarding image cropping as a set prediction problem. The core components of our set prediction model are the learnable anchors and the bipartite matching. Specifically, we randomly initialize a set of learnable anchors. After absorbing useful information from the input image features through a transformer model, these anchors are used to regress crops directly. To address the inequality between the pre-defined anchor number and the labeled good crop number, we employ the Hungarian algorithm to perform bipartite matching. According to the matching results, a classifier is trained to judge whether an anchor is valid. In short, the direct coordinate regression enables *Globality* and the multiple anchors with validity classification help to achieve *Diversity*. This framework successfully combines the different advantages of the anchor-based models and the regression models, as shown in the bottom of Fig. 1.

Despite the great advantages, a ghost of inconsistency hides between the regressed crops and the validity labels. According to the bipartite matching results, we assign hard labels $v = 0$ for the unmatched crops. Such a hard validity label forces the model to treat all the invalid crops equally, while these invalid crops fall into a wide quality range. The contradiction between the hard labels and the complex crop qualities causes inconsistency, which is harmful to the model training. To cope with this problem, we employ two different label smoothing methods to make the validity labels better reflect the crop qualities. The first method uses quality scores to guide the smoothing. The qualities of the invalid crops are estimated according to the local redundancy property [43] in the nearly dense labeled dataset. The second method employs self distillation [10] when the dense labels are unavailable considering the model itself has the potentials to learn knowledge about crop quality. The estimated validity probabilities are used to generate soft labels. We finally conduct sufficient experiments using various evaluation metrics on four datasets to validate the effectiveness of the proposed method. Our main contributions are as summarized as follows.

- We rethink image cropping from the perspective of both globality and diversity: delving into all possibilities to find all good compositions.
- We regard image cropping as a set prediction problem, where multiple regressed crops with a validity classifier are used to match diverse good crops. This enables the globality and diversity.
- Two different label smoothing methods are developed into the set prediction method to deal with the inconsistency problem between the crops and validity labels.
- Extensive experiments are conducted to evaluate our model. The comparative experiments and the ablation study results prove the effectiveness of our model.

2. Related Works

2.1. Image Cropping

Aiming to improve image aesthetic quality, image cropping is different from other similar tasks that only preserve useful contents and structures such as image retargeting [33, 34] and graphcut [1, 41]. Therefore, aesthetic quality evaluation techniques [12, 30] are always involved. Early works [6,7,22,26,47] mainly use hand-craft aesthetic-related features with shallow classifiers [9]. Recently, deep neural networks have dominated many computer vision tasks from recognition to generation [13,27]. There are two major types of cropping methods. The anchor-based methods focus on the anchor generation and the anchor evaluation methods. Wang *et al.* [38] obtained candidates based on image saliency, and employed an AVA [30] pretrained network to evaluate crops. Wei *et al.* [39] proposed a new image cropping dataset and a method based on knowledge transfer. Zeng *et al.* [43] analyzed the redundancy property of crops and defined anchors based on a grid rule. Tu *et al.* [37] proposed to use composition and saliency aware score maps to evaluate crops, and a two-stage searching strategy was designed to find good views. Chen *et al.* [5] proposed to utilize good photos on the web to obtain bad crops by random cropping. Li *et al.* [17] pointed out that the mutual relations between different crops are key factors to improve the crop evaluation performances. The other category directly regresses cropping coordinates. Lu *et al.* [24] proposed an end-to-end network to achieve image cropping. Guo *et al.* [8] proposed to employ cascaded regression to regress the crop boundaries directly from the whole image. Different from other works, Li *et al.* [14, 15] employed the reinforcement learning method to obtain bounding boxes from the whole image, such that all the possible crops are covered. Hong *et al.* [11] proposed a model that uses different composition rules explicitly, making the model works like a photographer.

2.2. Label Smoothing

Label smoothing has shown effectiveness in many areas. Its functions could be categorized into three classes

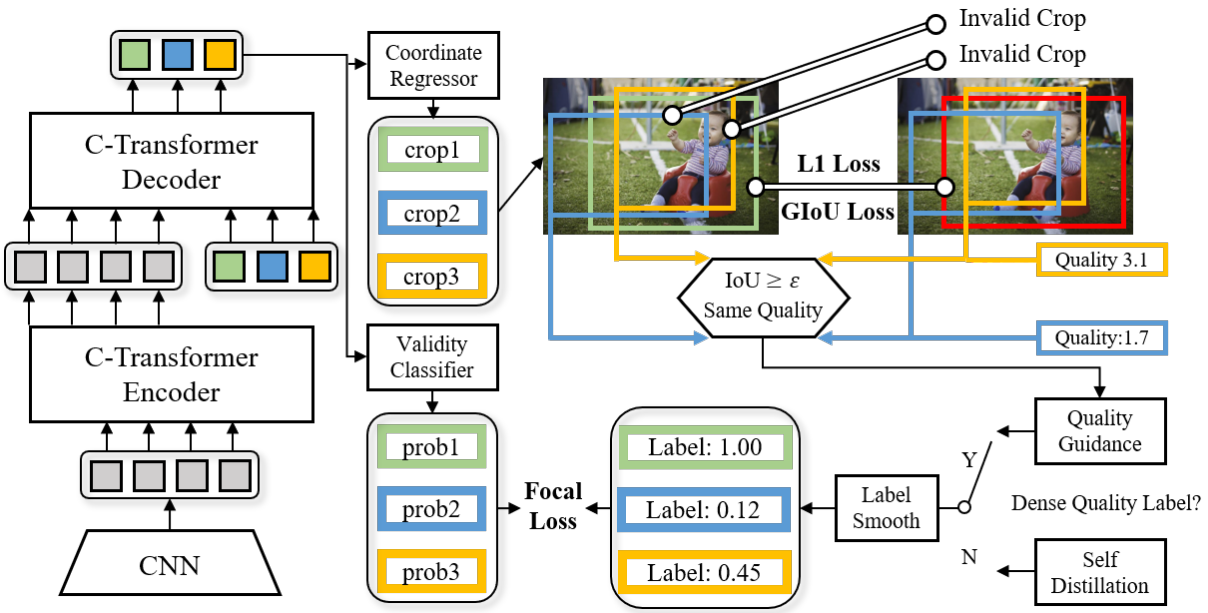


Figure 2. Framework of our model. The left part is the model architecture that contains a CNN backbone, a conditional transformer encoder, a conditional transformer decoder, and two prediction heads. The model predicts a crop and a validity probability from each anchor (denoted as colored squares). The right part depicts the bipartite matching and the label smoothing process. The predicted green crop is matched with the given good crop (red box). In the densely labeled dataset, we estimate the quality scores of the invalid crops by finding their high-IoU neighbors (the same-color crops in the top-right image) and use the scores to guide the label smoothing. In the sparsely labeled dataset, we use self distillation to smooth the labels.

[45] including label regularization, label relation mining and noisy label learning. The three functions always do not work alone [42]. Many efforts have been made from simple uniform smoothing [35] to some more complex forms [10, 19, 21, 45] by mining the relations between different data, classes or learning stages. There are also some works trying to explain the mechanisms and the relations to other techniques [25, 29]. Label smoothing in our work is a kind of label relation mining method by making crops of closer qualities have closer labels.

3. Method

3.1. Image Cropping Based on The Set Prediction

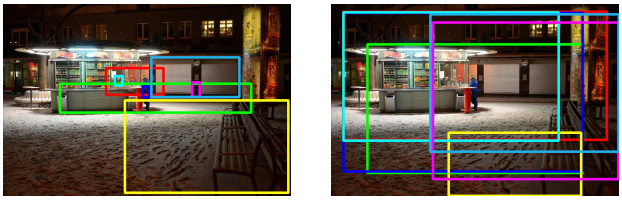
We introduce our model by showing its evolution path from the traditional coordinate regression models. From this perspective, only the *Diversity* needs to be added since the *Globality* has been equipped. Therefore, we decompose our task into two sub-tasks. The first sub-task enables the model to predict a fixed number of multiple crops and the second sub-task relaxes the fixed number to the arbitrary numbers.

To achieve the first goal, we employ multiple input features, *i.e.*, anchors, to regress multiple crops. Specifically, we randomly initialize a set of learnable anchors $q_i \in \mathbb{R}^C$, where $i \in \{1, 2, ..., N^q\}$. A model that takes both the anchors and the images as inputs has two different functions. The first function exchanges information between different anchors. The second function transmits information from the input images to the anchors. Finally, a regression head is employed to predict a crop $\hat{b}_i \in \mathbb{R}^4$ from the anchor q_i.

In the first sub-task, we define a sufficiently large anchor number N^q. But two new problems arise. First, the number of the good crops N^{B_j} from the j-th image may not reach N^q, i.e., $N^{B_j} < N^q$. Second, different images may have different number of good crops, i.e., $N^{B_j} \neq N^{B_i}$. Therefore, in the second sub-task, we create an auxiliary binary classifier to find different valid subsets from all the N^q regressed crops. A classifier output $\hat{v}_i^j$ represents the validity probability of the regressed crop y_i in the input image I_j.

To train the validity classification and the coordinate regression, bipartite matching is performed between the labeled good crops and the anchor predictions using the Hungarian algorithm the same as [2]. Specifically, the N^{B_j} good crops are padded to N^q. After padding, we have the ground-truth label set $Y^j = \{y_i | i = 1, 2, ..., N^q\}$, where y_i contains the coordinates of the good crops b_i and the validity labels v_i,

$$y_i = \begin{cases} \{b_i = [c_x, c_y, w, h],\ v_i = 1\} & 1 \leq i \leq N^{B_j} \\ \{b_i = \varnothing,\ v_i = 0\} & N^{B_j} + 1 \leq i \leq N^q \end{cases} \quad (1)$$



(a) Initial crops. (b) Learned crops.

Figure 3. Comparison between the initial crops and the learned crops. We show the crops from the randomly selected 7 invalid anchors.

where $[c_x, c_y, w, h]$ represents the center coordinate, the width and the height of the crop, respectively. The bipartite matching finds an index mapping $\sigma \in \mathfrak{S}_{N_q}$ such that the matching cost $\mathcal{L}_{\text{match}}\left(y_i, \hat{y}_{\sigma(i)}\right)$ is minimized:

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_{N_q}} \sum_{i}^{N} \mathcal{L}_{\text{match}}\left(y_i, \hat{y}_{\sigma(i)}\right) \tag{2}$$

In our model, we use the summation of three losses, including the coordinate regression loss, the generalized IoU loss, and the focal loss following [28]:

$$\begin{aligned}\mathcal{L}_{\text{match}}\left(y_i, \hat{y}_{\sigma(i)}\right) =& \mathcal{L}_{\text{reg}}(b_i, \hat{b}_{\sigma(i)}) + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}}(b_i, \hat{b}_{\sigma(i)}) \\ &+ \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}(v_i, \hat{v}_{\sigma(i)}),\end{aligned} \tag{3}$$

where λ_{iou} and λ_{bce} are the trade-off parameters between different losses. Note that $\mathcal{L}_{\text{iou}} = \mathcal{L}_{\text{reg}} = 0$ when $b_i = \varnothing$. We find that the focal loss is critical to prevent the model from degenerating to a naive solution. Finally, the model parameters are updated by minimizing the loss under the optimal match $\mathcal{L}_{\text{match}}\left(y_i, \hat{y}_{\hat{\sigma}(i)}\right)$.

3.2. Label Smoothing

Our set prediction model uses a validity classifier to select crops from predictions. Simple binary labels are assigned to the regressed crops according to the bipartite matching results. This label setting may not be optimal for the inconsistency between the labels and the crops. In this subsection, we propose to use the label smoothing method to tackle the problem.

We firstly analyze the existence of the inconsistencies. During training, only the valid anchors are assigned with the ground-truth good crops. The rest anchors are left without any supervision signal for the cropping coordinate regression. But this does not mean that these invalid anchors only output meaningless noises. This is because an invalid anchor in an input image may become valid for some other images. When the model is sufficiently trained, many anchors can regress meaningful crops. Fig. 3 shows an example of the changes on the invalid anchor predictions. In the initial stage, most invalid anchors only generate crops of poor quality. But in the late training stage, the qualities of the crops from these anchors are significantly improved. In such a situation, the validity labels become quality labels, indicating that a given crop is either good ($v_i = 1$) or bad ($v_i = 0$). But only one level of "bad" is not reasonable to depict the invalid crops, especially when we use a strict criterion to define good crops. For example, some datasets [39,43,44] employ continuous mean opinion scores ranging from 1 to 5 from multiple users to describe the qualities of crops. If we set the criterion $s \geq 4$ to define the good crops, a crop with a score $s_i = 1.2$ and a crop with a score $s_j = 3.8$ are both bad crops, although their qualities are highly different. This is the source of the inconsistency between the regressed crops and the validity labels.

To deal with this problem, we propose to smooth the validity labels to reflect the crop qualities better. Two different methods are proposed to adapt to different situations. Detailed introductions of them are given below.

3.2.1 Quality Guidance

Recalling that our goal is to improve the label-quality consistency, the direct solution is to use the qualities of the invalid crops and make better crops have smoother labels. However, such qualities are unavailable. A substitution is to find a reliable method to estimate the qualities. Fortunately, when the training data has nearly dense crops labeled with quality scores, a property called local redundancy [43] can help us to realize this target. This property shows that human perceptions are not sensitive to small changes in cropping scales and the cropping locations. In other words, the qualities of two crops with an extremely high Intersection over Union (IoU) are very likely to be the same. In practice, we first calculate the IoU between a regressed invalid crop and all the crops labeled with quality scores in the training image. Then we check if the maximum IoU is larger than a given threshold ϵ. Once the condition is satisfied, we can directly transfer the quality score from the maximum-IoU neighbor to the invalid crop. Finally, we define a truncated linear function M to map the quality scores s_i to the soft labels $\widetilde{v}_i$:

$$\widetilde{v}_i = M(s_i) = \begin{cases} 0, & \text{if } s_i \leq s^l \\ \mu\frac{s_i - s^l}{s^u - s^l}, & \text{if } s^l < s_i \leq s^u \\ \mu, & \text{if } x \geq s^u \end{cases} \tag{4}$$

where μ is an upper bound of the smooth labels to ensure sufficiently large label gaps between the valid and the invalid anchors. s^l and s^u are two quality score thresholds. Crops whose quality scores are lower than s^l will directly use label 0, and the labels for the crops with quality scores higher than s^u are μ.

3.2.2 Self Distillation

In the first method, the local redundancy property helps us quickly estimate the invalid crops' qualities. However, once the training data does not provide nearly dense crops with quality scores, most invalid crops cannot find their high-IoU neighbors. In this situation, the model itself is the only thing we can count on. Some previous works [10] find a model itself can learn the relations between classes even it is trained with the one-hot hard labels in classification tasks. According to this property, we employ the self distillation method [18, 36, 46] to smooth the labels. Specifically, we start to train our model using the default hard labels. When the model converges at a good point, a new validity classification loss is added, using the predicted validity probabilities from the momentum averaged model as the soft labels. Given a well-trained model F^t at the training iteration t, we define the momentum averaged model $\overline{F}^t$ as:

$$\overline{F}^t = \theta F^t + (1-\theta)F^{t-1} \tag{5}$$

where θ is the moving average decay. The validity classification loss becomes:

$$\mathcal{L}^t_{\text{validity}} = \mathcal{L}_{\text{focal}}(v_i, \hat{v}_{\sigma(i)}) + \mathcal{L}_{\text{focal}}(\overline{F}^t(X_j, q_{\sigma(i)}), \hat{v}_{\sigma(i)}) \tag{6}$$

3.3. Model Architecture

Our model architecture is similar to the Conditional-DETR (cDETR) [28]. The model consists of three parts, a CNN backbone, a transformer encoder, and a transformer decoder. The CNN backbone and the transformer encoder extract features from the input images. The transformer decoder takes both the encoder outputs and the learnable anchors as inputs to perform both self-attention and cross-attention, The self-attention modules exchange knowledge between different anchors, and the cross-attention modules transmit image features to the anchors. Finally, a regression head and a classification head are employed to estimate the cropping coordinates and the validity probabilities.

4. Experiments

4.1. Datasets

FLMS dataset [7] contains 500 images, and each image is annotated with no more than 10 high-quality crops. This dataset is only used for testing.

CPC dataset [39] is a large dataset containing 10,797 images. There are four groups of crops on each image corresponding to four kinds of aspect ratios, and each group has six crops. Six AMT workers annotate each crop with scores ranging from 0 to 4 through a two-stage annotation pipeline. This dataset is only used for training. We randomly select 1,000 images as the validation data.

Table 1. $\text{ACC}_{1/N}$ Performances on both the GAICv1 and the GAICv2 datasets.

Models	GAICv1		GAICv2	
	ACC_5	ACC_{10}	ACC_5	ACC_{10}
A2-RL [14]	23.0	38.5	23.2	39.5
VPN [39]	40.0	49.5	36.0	48.5
VFN [5]	27.0	39.0	26.6	40.6
VEN [39]	40.5	54.0	37.5	50.5
GAICv1 [43]	53.5	71.5	65.8	82.4
GAICv2 [44]	-	-	68.2	85.8
ASM-Net [37]	54.3	71.5	-	-
Li *et al.* [17]	63.0	81.5	-	-
MFDM [40]	66.5	83.0	-	-
TransView [32]	-	-	69.0	85.4
Ours ($\epsilon = 0.85$)	**81.5**	**91.0**	**85.0**	**92.6**
Ours ($\epsilon = 0.90$)	65.5	74.5	**72.0**	**86.0**

GAICv1 dataset [43] has 1,036 images for training and 200 images for testing. Each image contains at most 90 crops generated by a pre-defined grid-anchor rule. Each crop is annotated with a score ranging from 1 to 5. This dataset is extended to **GAICv2** [44], in which the number of training, validation and testing images are 2,636, 200 and 500 respectively. There is no official split of validation data in GAICv1, so we randomly select 36 images from its training set for validation. As for the GAICv2 dataset, we use its official protocol.

4.2. Evaluation Metrics

IoU is the most commonly adopted metric in the previous works. However, some works [43,44] point out that it is not reliable. Therefore, besides the IoU metric, we further employ the $\text{ACC}_{K/N}$ metric [43] to evaluate our model. Only $K = 1$ is used since it cannot be ensured that the number of good crops in an image is larger than 1. Given the ground-truth good crop set $B_j = \{b_1, ..., b_{N^{B_j}}\}$ and the regressed crop set $\hat{B}_j = \{\hat{b}_1, ..., \hat{b}_{N^{\hat{B}_j}}\}$ with the top $N^{\hat{B}_j}$ validity scores, the $\text{ACC}_{1/N}$ is defined as follows in our model:

$$\text{ACC}_{1/N} = \frac{1}{T}\sum_{j=1}^{T} \mathbb{1}(\max_{\hat{b}^j \in \hat{B}^j_N} \{F_{IoU}(b^j_{\text{best}}, \hat{b}^j)\} \geq \epsilon), \tag{7}$$

where b^j_{best} is the crop with the highest quality score in B^j, and $\hat{B}^N_j$ indicates that $N^{\hat{B}_j} = N$ for all j. $\mathbb{1}(*)$ equals to 1 when the condition $*$ is satisfied otherwise 0. ϵ is a pre-defined IoU threshold. When the IoU between two crops is sufficiently large, the qualities of the two crops can be regarded the same according to the local redundancy property [43]. Two thresholds $\epsilon \in \{0.85, 0.90\}$ are used in our experiments.

Table 2. AP Performances of different top-K predictions on the GAICv2 dataset.

Models	AP		
	$K = 5$	$K = 10$	$K = 40$
VEN [39]	20.2	25.5	34.7
GAICv2 [44]	24.3	33.8	42.2
Ours($\epsilon = 0.85$)	**38.2**	**50.5**	**56.8**
Ours($\epsilon = 0.90$)	**30.3**	**40.6**	**47.4**

Table 3. IoU performances on the FLMS dataset.

Models	IoU
Fang *et al.* [7]	0.740
ABP+AA [38]	0.810
VPN [39]	0.835
VEN [39]	0.837
GAICv2 [44]	0.836
Ours	**0.838**

$\mathrm{ACC}_{1/N}$ is sometimes limited because it only reflects the recall performance of the best crop, while we define more than one good crop in many images. Therefore, we further use the average precision (AP) metric calculated by averaging the different precisions under the different recalls. This metric has been widely used in object detection models, and can better reflect the overall performances. Our implementation is based on the COCO API. Please visit https://github.com/cocodataset/cocoapi for more details about this metric.

4.3. Implementation Details

Training and evaluation details: We follow the training details in the cDETR [28]. The optimizer is ADAMW [23] with 10^{-4} weight decay. The learning rate is 10^{-4}, and the CNN backbone uses a lower learning rate 10^{-5}. The model is trained for 50 epochs and the learning rate is divided by ten at the 40-th epoch. In the self distilled label smoothing, we start distillation at the 40-th epoch, reduce learning rate at the 50-th epoch, and stop training at the 60-th epoch.

Dataset settings: We use the data augmentations similar as [43]. Multi-scale augmentation the same as [2] is also used. In the GAICv1 and GAICv2 datasets, we define the crops whose quality scores are higher than four as the ground-truth good crops. Unless otherwise specified, quality guidance is used to smooth labels in the two datasets. As for the CPC dataset, the quality score threshold is two and we smooth labels using the self distillation.

Model setting: The loss trade-off parameters are set as $\lambda_{\text{iou}} = \lambda_{\text{focal}} = 0.4$. We use 90 anchors for all the datasets. In the quality guided label smoothing, we set $s^l = 2$, $s^u = 3.5$ and $\mu = 0.5$. In the self distilled label smoothing, the moving average rate θ is set as 0.5.

4.4. Comparisons with Previous Methods

Quantitative comparison: We firstly compare different models on the GAICv1 [43] and the GAICv2 [44] datasets using the $\mathrm{ACC}_{1/5}$ and the $\mathrm{ACC}_{1/10}$ metrics. Table 1 shows the results [1] on the two datasets. We directly show the results of the previous works reported in their papers. We can observe significant performance improvements on the metrics when $\epsilon = 0.85$, indicating the great superiority of our model. Even when $\epsilon = 0.90$, our model still obtains the best performance on the GAICv2 dataset. We attribute the success on the $\mathrm{ACC}_{1/N}$ metric to two characteristics of our model. Firstly, compared with the anchor-based models, our model concentrates on finding the good crops instead of evaluating all the crops of the entire quality range. The latter method may distract the model, especially when the high-quality crops only make up a small fraction of all the anchors. Secondly, unlike traditional regression models with a single output crop, our model generates multiple good crops adaptively, covering different compositions, scales and preferences more comprehensively. Therefore, the crop with the highest quality score is more likely to be detected by our model.

We further give the results of the AP metric. The open-sourced anchor-based models [39, 44] are employed as the competitors since this metric is inappropriate for the traditional regression models. To calculate the AP metric, we need to select the predictions with the top-K validity probabilities. We set three different K values of $\{5, 10, 40\}$ and show the results in Table 2. It is shown that our model outperforms the competitors in all the settings.

Finally our model is compared with the previous models on the FLMS dataset using the IoU metric. The model is trained on the CPC dataset. The experimental results in Table 3 show that our model achieves similar results compared with the previous works. Since this metric is not always reliable as analyzed in [43]. The results only reflect the rough performances.

Qualitative comparison: We use some qualitative comparisons to show the advantages of our model against the traditional methods. The model used here is trained on CPC dataset with the label smoothing based on the self distillation. The crops are selected from the top 10 outputs according to the validity probabilities. The used images are chosen from the AVA [30] dataset and are ensured not to exist in the training data. As we have analyzed before, the coordinate regression models that only crop from a single view lack *Diversity*. Fig. 5 shows two examples that our model finds two good crops for each input image, while the A2_RL [14] model only generates one crop. The traditional

[1] Note that $\mathrm{ACC}_{1/N}$ is abbreviated as ACC_N to adapt to the table scale

Figure 4. **Globality**: In each pair of images, the left one is the input image and the right one is a crop from our model predictions. When using the anchors in [43], the four anchor corners need to be inside the corresponding four white transparent regions shown in each image. However, the bounding boxes (red dashed line) of the crops do not satisfy the requirement. This means that our model can generate crops that do not exist in the pre-defined anchors in [43].

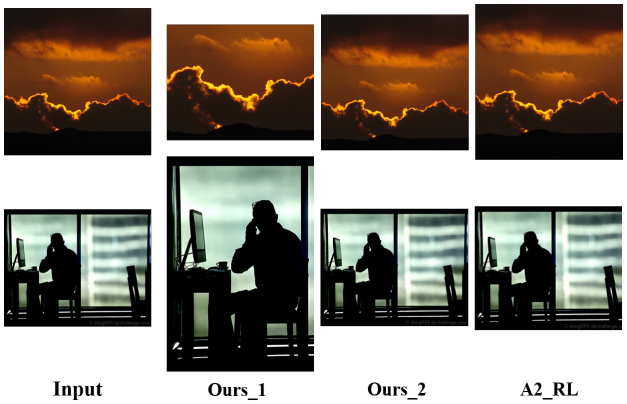


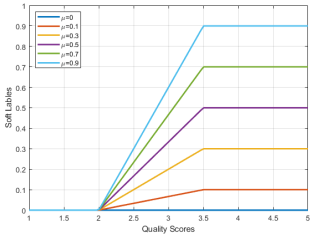
Figure 5. **Diversity**: Compared with the previous coordinate regression models that only crops from one single view, our model can generate multiple good crops.

anchor evaluation models need carefully designed anchors. However, sometimes there are still neglected good crops, resulting in the lack of the *Globality*. We use six examples in Fig. 4 to show that our model overcomes such a weakness. When using the grid anchors defined in [43], the four corners of the anchors need to be inside the corresponding four white transparent regions shown in the left of the image pairs in Fig. 4. But the crops from our model, shown in the right of the image pairs and denoted as the red dashed boxes on the left input images, do not follow the anchor generation rules. This means that our model generates some good crops that do not exists in the grid anchors. One reason that the grid anchors lack *Globality* is that they uses a strong assumption called content preservation. Sometimes this assumption does not hold in real-world applications since the amateur users do not always put the key objects in the central regions.

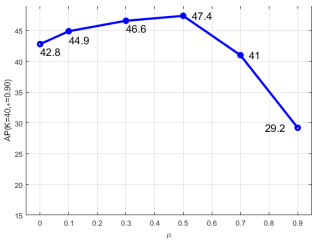
4.5. Label Smoothing Analysis

Quality guidance: We show the influences of the upper bound μ in Eq. (4) in the label smoothing based on the quality guidance (QG). Theoretically, higher μ brings better consistency. The model does not use label smoothing when setting $\mu = 0$. The AP performances with $\epsilon = 0.90$ and $K = 40$ under different μ are plot in Fig. 6b. The mapping functions under different μ are given in Fig. 6a. We can observe two phenomenons. Firstly, compared with the binary hard labels, the quality-guided smoothed labels effectively improve the performances. The AP metric is improved from 42.8 to 47.4. Secondly, overly smoothed labels are very harmful. When the upper bound is very close to 1, the performances drop drastically. For example, the AP is only 29.2 when $\mu = 0.9$. The phenomenons show that the label smoothing has both positive and negative influences. The negative influence may originate from the damage to the model discriminability. If μ is close to 1, the small label margin will make it difficult for the model to distinguish the invalid and the valid crops.

Self distillation: In the self distillation (SD) method, we focus on two questions: (1) Is this label smoothing method effective? (2) What are the characters of the self-learned soft labels? To answer the first question, we perform experiments on both the GAICv2 and the CPC datasets since this method can work on both of them. In the CPC dataset, we randomly select 1000 images to test the model performances. The results in Table 4 shows that the AP ($\epsilon = 0.85$) metric obtains consistent improvements on both datasets. The effectiveness of this label smoothing method is proved. For the second question, our main concern is the relations between the learned soft labels and the crop qualities. Therefore, we use the method described in Section 3.2.1 to estimate the quality scores of the regressed crops if possible and plot the crops in scatter diagrams whose x-axis represents the estimated quality scores and the y-axis rep-



(a) Quality-label mapping functions M under different μ



(b) AP performances under different μ.

Figure 6. Influences of the label smoothing upper bound μ. We test 6 different μ values $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ (a) shows the mapping function M under the different μ. (b) shows the AP ($\epsilon = 0.90, K = 40$) performances under the different μ.

Table 4. AP ($\epsilon = 0.85$) Performances of the models with and without the self distillation on two datasets.

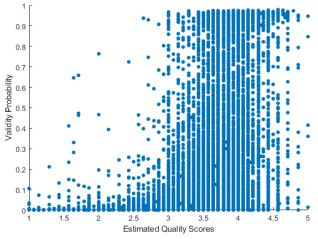
Models	GAICv2		CPC	
	$K = 10$	$K = 40$	$K = 10$	$K = 40$
w/o SD	47.9	54.8	26.3	26.6
w/ SD	**49.0**	**55.5**	**27.0**	**27.8**

resents the learned soft labels. Only the invalid crops are plotted in Fig. 7b for the training images in the GAICv2. As for the testing images, all the crops with the estimated validity probabilities are shown in Fig. 7a. The two figures show significant positive correlations between the quality scores and the soft labels. Most high validity probabilities belong to the high-quality crops, while nearly all the low-quality crops only have low validity probabilities. This phenomenon further proves our analysis that the label smoothing improves the model performance by improving the consistency between the validity labels and the regressed crops.

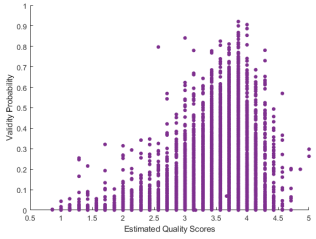
Comparisons between the two methods: We finally compare the two different label smoothing methods on the GAICv2 dataset using the AP metrics. Theoretically, the quality-guided method should perform better than the self distillation method. This is because the former can directly use the precisely estimated quality scores. In contrast, the latter only uses the learned knowledge from the model itself. The experimental results that the quality guidance method performs better in all conditions in Table 5 prove our deduction. We also observe that the performance gaps are larger at $\epsilon = 0.90$. This phenomenon further validates the merits of the direct quality guidance.

5. Limitations and Broader Impacts

A major limitation of our model is the cropping scale bias. We notice that the model trained on the GAIC dataset tends to generate large scale crops and ignore small good crops. We will work on this problem in our future works. A possible negative impact is the influences to personal aesthetics. The abuse of image cropping models may make people concentrate obsessively on the common cropping patterns and neglect the highly different personal preferences.



(a) GAICv2 testing data.



(b) GAICv2 training data.

Figure 7. Quality-validity scatter diagram. The abscissa is the estimated quality score ranging from 1 to 5 using the method described in section 3.2.1, and the ordinate is the probability from the validity classifier. In both training and testing data, we can observe positive correlations between the validity probabilities and the quality scores.

Table 5. AP Performances of the two different label smoothing methods on the GAICv2 dataset.

Smooth Methods	AP ($\epsilon = 0.85$)		AP ($\epsilon = 0.90$)	
	$K = 10$	$K = 40$	$K = 10$	$K = 40$
QG	**50.5**	**56.8**	**40.6**	**47.4**
SD	49.0	55.5	37.7	43.5

6. Conclusion

In this paper, the weaknesses of the traditional image cropping models are analyzed. We propose a new perspective that regards image cropping as a set prediction problem to mitigate their shortcomings. The set prediction model directly regresses multiple crops and automatically estimates their validity. However, we find the original set prediction lacks consistency between the regressed crops and the validity labels. We propose two kinds of methods to ease the inconsistency problem in different situations. The quality guidance method directly uses the estimated quality scores, and the self distillation method extracts knowledge from the model itself. Sufficient experimental results prove the effectiveness of different modules and show the merits over the previous methods.

7. Acknowledgement

This work is partially funded by National Natural Science Foundation of China (Grant No. U21B2045, U20A20223), Youth Innovation Promotion Association CAS (Grant No. Y201929) and CCF-Baidu Open Fund (Grant NO. 2021PP15002000).

References

[1] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning*, pages 874–883. PMLR, 2020. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 3, 6

[3] 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*, 9(4):353–364, 2003. 1

[4] 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 *IEEE Winter Conference on Applications of Computer Vision*, pages 226–234, 2017. 1

[5] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *ACM International Conference on Multimedia*, pages 37–45, 2017. 2, 5

[6] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer, 2006. 2

[7] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACM International Conference on Multimedia*, pages 1105–1108. ACM, 2014. 1, 2, 5, 6

[8] Guanjun Guo, Hanzi Wang, Chunhua Shen, Yan Yan, and Hong-Yuan Mark Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Transactions on Multimedia*, 20(8):2073–2085, 2018. 1, 2

[9] Ran He, Bao-Gang Hu, and Xiao-Tong Yuan. Robust discriminant analysis based on nonparametric maximum entropy. In *Asian Conference on Machine Learning*, pages 120–134. Springer, 2009. 2

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 5

[11] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7057–7066, 2021. 1, 2

[12] Gengyun Jia, Peipei Li, and Ran He. Theme-aware aesthetic distribution prediction with full-resolution photographs. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[14] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8193–8201, 2018. 1, 2, 5, 6

[15] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *IEEE Transactions on Image Processing*, 2019. 1, 2

[16] Debang Li, Junge Zhang, and Kaiqi Huang. Learning to learn cropping models for different aspect ratio requirements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12685–12694, 2020. 1

[17] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4213–4222, 2020. 1, 2, 5

[18] Peipei Li, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Deep label refinement for age estimation. *Pattern Recognition*, 100:107178, 2020. 5

[19] Weizhi Li, Gautam Dasarathy, and Visar Berisha. Regularization via structural label smoothing. In *International Conference on Artificial Intelligence and Statistics*, pages 1453–1463. PMLR, 2020. 3

[20] Tianpei Lian, Zhiguo Cao, Ke Xian, Zhiyu Pan, and Weicai Zhong. Context-aware candidates for image cropping. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1479–1483. IEEE, 2021. 1

[21] Julian Lienen and Eyke Hüllermeier. From label smoothing to label relaxation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI, Online, February 2-9, 2021*. AAAI Press, 2021. 3

[22] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. Optimizing photo composition. *Computer Graphics Forum*, 29(2):469–478, 2010. 2

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[24] Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin. An end-to-end neural network for image cropping by learning composition from aesthetic photos. *arXiv preprint arXiv:1907.01432*, 2019. 1, 2

[25] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020. 3

[26] Matthew Ma and Jinhong K Guo. Automatic image cropping for mobile device with built-in camera. In *First IEEE Consumer Communications and Networking Conference, 2004. CCNC 2004.*, pages 710–711. IEEE, 2004. 1, 2

[27] Xin Ma, Xiaoqiang Zhou, Huaibo Huang, Gengyun Jia, Zhenhua Chai, and Xiaolin Wei. Contrastive attention network with dense field estimation for face completion. *Pattern Recognition*, 124:108465, 2022. 2

[28] 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. 4, 5, 6

[29] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019. 3

[30] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012. 2, 6

[31] Bingbing Ni, Mengdi Xu, Bin Cheng, Meng Wang, Shuicheng Yan, and Qi Tian. Learning to photograph: A compositional perspective. *IEEE Transactions on Multimedia*, 15(5):1138–1151, 2013. 1

[32] Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong. Transview: Inside, outside, and across the cropping view boundaries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4218–4227, 2021. 5

[33] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. A comparative study of image retargeting. In *ACM SIGGRAPH Asia*, pages 1–10, 2010. 2

[34] Vidya Setlur, Saeko Takagi, Ramesh Raskar, Michael Gleicher, and Bruce Gooch. Automatic image retargeting. In *Proceedings of the 4th international conference on Mobile and ubiquitous multimedia*, pages 59–68, 2005. 2

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 5

[37] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12104–12111, 2020. 1, 2, 5

[38] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *IEEE International Conference on Computer Vision*, pages 2186–2194, 2017. 1, 2, 6

[39] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2018. 1, 2, 4, 5, 6

[40] Yifei Xu, Wujiang Xu, Mian Wang, Li Li, Genan Sang, Pingping Wei, and Li Zhu. Saliency aware image cropping with latent region pair. *Expert Systems with Applications*, 171:114596, 2021. 1, 5

[41] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*, 2021. 2

[42] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. 3

[43] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5949–5957, 2019. 1, 2, 4, 5, 6, 7

[44] 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, 4, 5, 6

[45] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021. 3

[46] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019. 5

[47] Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and Nicu Sebe. Weakly supervised photo cropping. *IEEE Transactions on Multimedia*, 16(1):94–107, 2013. 1, 2