Composing Photos Like a Photographer

Chaoyi Hong[†] Shuaiyuan Du[†] Ke Xian[†] Hao Lu[†] Zhiguo Cao[†,*] Weicai Zhong[‡]
[†]Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[‡]Huawei CBG Consumer Cloud Service Search Product & Big Data Platform Department
{cyhong, zgcao}@hust.edu.cn

Abstract

We show that explicit modeling of composition rules benefits image cropping. Image cropping is considered a promising way to automate aesthetic composition in professional photography. Existing efforts, however, only model such professional knowledge implicitly, e.g., by ranking from comparative candidates. Inspired by the observation that natural composition traits always follow a specific rule, we propose to learn such rules in a discriminative manner, and more importantly, to incorporate learned composition clues explicitly in the model. To this end, we introduce the concept of the key composition map (KCM) to encode the composition rules. The KCM can reveal the common laws hidden behind different composition rules and can inform the cropping model of what is important in composition. With the KCM, we present a novel cropping-by-composition paradigm and instantiate a network to implement composition-aware image cropping. Extensive experiments on two benchmarks justify that our approach enables effective, interpretable, and fast image cropping.

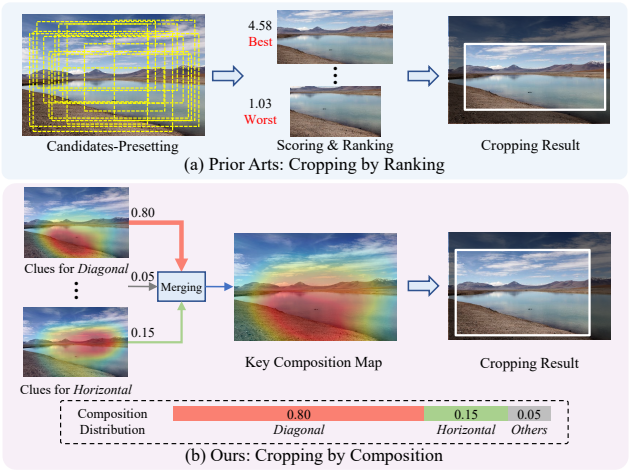

Figure 1. High-level ideas between prior arts and ours. (a) Prior image cropping methods follow a cropping-by-ranking paradigm, where no composition clue is modeled explicitly; (b) Our idea is to guide image cropping by composition. Our approach achieves this by discovering the key composition map and the composition distribution of the image, which can produce well-composed, interpretable cropping results.

1. Introduction

Professional photography is expensive, because a well-captured photo needs to take many factors into account. One of the most important factors is composition. Composition by definition refers to "*the nature of something's ingredients or constituents; the way in which a whole or mixture is made up*".[1] In photography, not only do those visual constituents of interest matter, but they compose following an aesthetic standard. Aesthetic composition, however, requires expert knowledge and extensive training. Can ordinary people compose photos like a photographer? The desire to popularize such knowledge and experience in everyday life has driven enthusiasm for an interesting research topic in computer vision—image cropping.

Given an arbitrarily captured image, image cropping aims to remove extraneous areas and to find a region in which visual constituents conform to aesthetic composition [38]. Aesthetic composition is an art. As the humanist photographer Henri Cartier-Bresson said, "*In a photograph, composition is the result of a simultaneous coalition, the organic coordination of elements seen by the eye.*" However, it is not immediately clear how such organic coordination can be achieved by image cropping.

Aesthetics-inspired approaches consider that good image cropping should be able to extract the most 'attractive' area, expecting that such an area obeys photographic composition rules and aesthetic attributes [6, 10]. However, these approaches typically follow a cropping-by-ranking paradigm, without explicit modeling of the composition rules. As

*Corresponding author.
[1]Definition by Oxford dictionary.

shown in Fig. 1(a), prior arts generate cropping results by presetting substantial candidate boxes, scoring the boxes per attention estimation or aesthetic assessment, and ranking them according to the scores. The box receiving the highest score is considered the final output. This paradigm assumes that aesthetic composition is automatically learned on a set of cropping views with relative aesthetic annotations [5, 21, 37, 42].

Does the assumption above hold in reality? A key insight of this work is that *good image cropping should model explicit, interpretable composition rules* [13]. Explicitness informs certain composition rules one image obeys; interpretability indicates major elements that determine the composition. They can provide image cropping with reasonable and reliable clues.

In this work, we introduce a new cropping-by-composition paradigm and propose a novel Composition-Aware Cropping Network (CACNet) to model the composition rules explicitly within the network (Fig. 1(b)). We assume the composition rules are learnable. This assumption is inspired by the observation that composition generally follows the basic, common arrangement of visual elements. For instance, the rule-of-thirds places important elements at the point where two horizontal and two vertical lines intersect; the horizontal rule arranges the major elements horizontally; the diagonal rule exploits leading diagonal lines to create dynamic senses. Since these leading elements play a critical role in composition, our idea is to learn a network to discover these elements to guide image cropping.

To this end, CACNet designs two branches: a composition branch and a cropping branch. The composition branch aims to learn the composition rules from data. This is achieved by categorizing images into 9 rules [17]: `rule-of-thirds` (RoT), `center` (Cen.), `horizontal` (Hor.), `symmetric` (Sym.), `diagonal` (Dia.), `curved` (Cur.), `vertical` (Ver.), `triangle` (Tri.), and `repeated pattern` (Pat.). These rules are then characterized by class activation maps (CAMs) [43]. Different CAMs are further squeezed to a 2D map we call the key composition map (KCM). As illustrated in Sec. 3.2, the KCM reveals the common laws hidden behind different composition rules. It is therefore used to inform the cropping branch of what is important in composition. The cropping branch finally exploits an anchor-point regressor to generate the cropping output conditioned on the KCM.

Extensive experiments on two benchmarks show that CACNet exhibits many appealing properties: i) *effective*: CACNet achieves state-of-the-art performance against recent competitors; ii) *interpretable*: the KCM can indicate why a cropping process makes sense; and iii) *fast*: since CACNet does not need to generate substantial candidate boxes, it can process images up to 155 FPS on a single RTX 2080 Ti GPU.

To our knowledge, we are among the first to explicitly model composition rules in the network and to bring the interpretability of composition into image cropping.

2. Related Work

What makes for good image cropping? Open literatures generally have two different opinions, i.e., image cropping guided by attention [2, 27, 29] and image cropping informed by aesthetics [5, 21, 33, 37, 42].

Attention-Guided Image Cropping. This line of approaches consider good image cropping to be able to preserve salient objects or informative regions. These approaches often have specific application scenarios such as thumbnails production [24, 30] and small-screen display [3, 7, 23]. 'Informativeness' is often estimated by saliency [3, 24, 30], energy functions [23], or human eye fixation [27]. Work of [24, 29, 30, 31, 36] aims to create a thumbnail cropping that summarizes the image. They mainly focus on estimating visual attention to render recognizable objects and on searching for a cropping window to encompass the image 'informativeness' [29].

Aesthetics-Informed Image Cropping. Aesthetics-based methods instead pursue the visual attractiveness of the cropped images. They assume aesthetic composition can be learned in comparative views and thus generally follow a two-stage pipeline: i) generating different candidates by scaling and shifting; ii) scoring all candidates by aesthetics and considering the top-1 to be the final output. A key question is *how to assess different candidates from almost the same visual content*. To this end, early works design hand-crafted features to manifest the aesthetic properties [10, 14, 25, 41]. Recently, the assessment of aesthetic composition has been approached by deep learning. VFN [5] proposes an aesthetics-aware ranking net, despite aesthetics is not explicitly modeled. DIC [34, 35] leverages both saliency prediction [12] and aesthetic assessment to generate content-preserved results. However, one open challenge is the existence of numerous candidates. Since each candidate needs to be scored by aesthetics, the methods can suffer from low efficiency. VPN [37] addresses this by accelerating the inference using knowledge distillation. GAIC [42] instead designs a grid anchor to reduce the searching space of candidates. Further, considering that the relations between candidates should be modeled, CGS [21] proposes a graph-based module with graph convolution. In addition to the two-stage pipeline, work of [18, 19, 22] formulates image cropping as a Markov decision-making process to imitate how humans make decisions. In contrast to the methods above, Mars [20] generates aspect-ratio-specified cropping and resorts to meta-learning for adaptive aspect ratio embedding.

Previous works assume that aesthetic composition can be automatically learned on comparative ranking views or

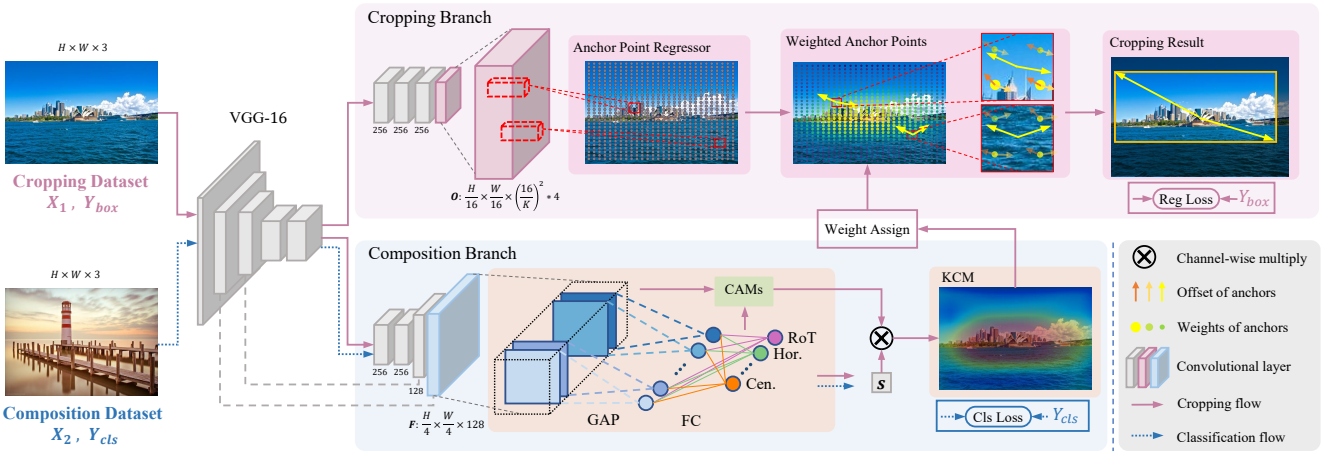

Figure 2. Technical pipeline of composition-aware image cropping. Our CACNet includes a backbone, an anchor-point-based cropping branch, and a composition branch. The cropping and composition branches are trained on two datasets, respectively. $\{X_2\}$ only flows to the composition branch, while $\{X_1\}$ flows to both. In the composition branch, the key composition map (KCM) is encoded by Class Activation Maps (CAMs). The KCM provides explicit and interpretable clues for the cropping branch by assigning weights to anchor points. These points then predict offsets to approximate the ground-truth boxes in an ensemble manner. K is the stride of anchor points.

learned by the aesthetic reward. By contrast, we believe that one should model the composition rules explicitly in the network, i.e., image cropping by composition.

3. Composition-Aware Image Cropping

3.1. Overview

Here we present an overview of our new image cropping paradigm—cropping by composition. This paradigm is inspired by the observation that composition in general is the arrangement of basic visual elements. These elements play a critical role in guiding image cropping. We therefore propose to explicitly explore and incorporate these composition clues in image cropping. In particular, a composition-aware cropping network (CACNet) is designed to implement the idea above. CACNet includes three parts: a backbone, a composition branch, and a cropping branch. Fig. 2 shows the technical pipeline of CACNet.

The backbone is shared by the composition branch and the cropping branch. The composition branch aims to learn composition rules by categorizing images to 9 composition rules [17]. Fig. 3 illustrates the different rules. To discover the key elements in images, class activation maps (CAMs) [43] are extracted and are further merged to form the KCM. We would like to highlight that, despite CAM is not a novel technique in this work, the idea of mining key composition rules using CAMs, to our knowledge, is novel to image cropping. For the cropping branch, an anchor-point regressor is exploited to predict offsets from anchor points to the ground-truth bounding box. Anchor points are weighted by the KCM, aiming to learn the composition mapping from the KCM to imitate expert annotations in an ensemble manner. CACNet is trained on both the cropping dataset [4] $\{X_1, Y_{box}\}$ and the composition dataset [17] $\{X_2, Y_{cls}\}$ simultaneously, where Y_{box} denotes the annotated cropping box and Y_{cls} the annotated composition classes.

3.2. Learning and Encoding Composition Rules

Learning Composition Rules. Composition rules are learned in the composition branch by categorizing images into different rules of composition. Exemplar images obeying different rules are illustrated in Fig. 3. One can observe that composition is the layout of different elements. For `RoT` and `center`, the dominant elements are at specific spatial position; rules of `horizontal`, `diagonal`, `curved`, and `vertical`, leverage leading lines or curves to convey messages of different moods (serenity, dynamism or strength); elements of `symmetric` and `pattern` follow symmetric or repeated pattern, respectively; `triangle` focuses on the arrangement of geometric shape or vanishing point. These elements in composition provide explicit and interpretable clues to guide image cropping. Moreover, this sub-task transfers composition knowledge into cropping and facilitates composition-aware cropping during training. It can be viewed as a source of auxiliary information to augment the feature representations. The capability to capture composition clues is thus naturally blended into the overall architecture.

Technically, the composition branch consists of a decoder, a global average pooling (GAP) layer, and a fully-connected (FC) layer, where the FC outputs the confidence score. In the composition dataset, training images may follow one or more (at most 3) composition rules. Images with more than one rule are trained multiple times for each ground-truth class. This training strategy is shown more ef-

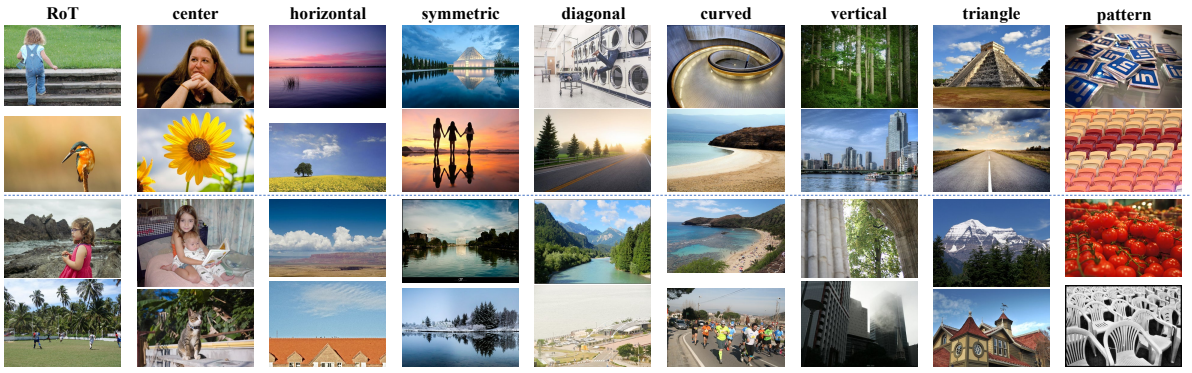

Figure 3. Exemplar images in different composition classes. The top two rows come from the training set of the composition dataset. The bottom two are those predicted by the composition branch in the cropping dataset.

fective than multi-label loss in experiments.

Encoding Rules in the Key Composition Map. Motivated by [43], CAMs are extracted to characterize corresponding composition rules. In our context, one important property of the CAM is that it can indicate the informative areas that determine each composition. Different CAMs are further encoded to a 2D map we call the Key Composition Map (KCM). The KCM informs the key regions of the overall composition. It reveals the common laws behind different composition rules and therefore can be used to guide the cropping branch of what is important in composition.

We start with a brief revisit of the CAM. In the composition branch, the decoder extracts C-dimensional feature maps $\mathbf{F}$. $\mathbf{F}$ is then transformed to a vector $\mathbf{f}$ after a GAP layer and is further transformed to a confidence score $\mathbf{s}$ by a FC and a softmax layer. CAMs, denoted by $\{\mathbf{M}_n, (n = 1, 2, ..., N)\}$, are extracted with weights $\mathbf{W} \in \mathbb{R}^{C \times N}$ of the FC layer projected back on $\mathbf{F}$, where N denotes the number of composition rules. Hence, the CAM $\mathbf{M}_n$ for the rule n takes the form

$$\mathbf{M}_n = \sum_{c=1}^{C} w_{c,n} \cdot \mathbf{F}_c \,, \tag{1}$$

where $w_{c,n}$ is the value of $\mathbf{W}$ indexed by (c, n). $\mathbf{M}_n$ reveals the discriminative composition evidence of identifying the input image to the composition rule n.

The KCM is merged by CAMs. Considering that images may obey one or more composition rules (e.g., an image of a person standing by the sea may follow both the rule of Hor. and RoT.), the KCM should take more than one rule into account. To fuse global composition elements, the KCM is extracted as a weighted sum of different normalized CAMs, i.e., each $\mathbf{M}_n$ is weighted by the confidence score s_n

$$\mathrm{KCM} = \sum_{n=1}^{N} s_n \phi(\mathbf{M}_n) \,, \tag{2}$$

where $\phi(\cdot)$ is the normalization function that normalizes the value of CAMs to $[0, 1]$. The KCM is then upsampled to the input image size. Fig. 4 illustrates the generation of KCM. It highlights the most informative composition regions, which provide interpretable cues as guidelines for cropping.

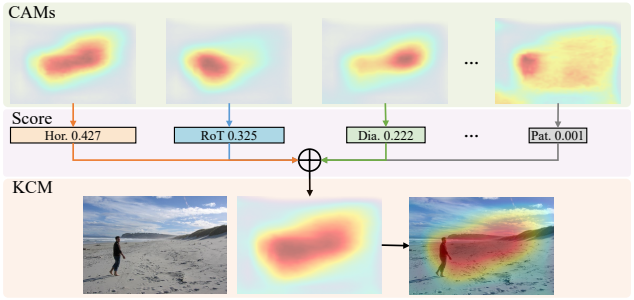

Figure 4. Generation of KCM. Different CAMs are weighted by the predicted confidence score and are further merged to KCM.

3.3. Cropping Informed by KCM

The cropping branch aims to model the relation between the KCM and a well-composed cropping. Since image cropping aims to change the relative position of dominant composition regions, it can be better accomplished when informed by explicit composition clues. The question is how to inject the KCM into the cropping model. Inspired by anchor-point methods [9, 32, 39] in object detection and pose estimation, local anchor-point regressors can be combined for better generalization. The KCM can assign weights to anchor points to strengthen/suppress the impacts of different anchors.

Local anchor points are uniformly set on the input image with a stride of K. In the cropping branch, the regression head predicts offsets $\mathbf{O}$ from each anchor point to four coordinates of a ground truth box. Suppose that, the offsets $\mathbf{O}$ is predicted at $\frac{1}{M}$ resolution of the input size, each spatial unit of $\mathbf{O}$ corresponds to M^2 pixels on the input image and $(\frac{M}{K})^2$ anchor points. An anchor point $a \in A$ with spatial position $p(a) = (i, j)$ on the input image estimates offset $o_k(a)$ for coordinate k. We assume that a good composition layout can be modeled by constructing relations between

the KCM and the expert cropping result. Anchors within the dominant composition regions should have more impacts on the final prediction. Each anchor a is therefore weighted by the KCM, defined by

$$w(a) = \frac{\mathrm{KCM}(p(a))}{\sum_{a \in A} \mathrm{KCM}(p(a))}\,, \tag{3}$$

and all anchors assigned with different weights jointly predict each boundary b_k of the cropping box

$$b_k = \sum_{a \in A} w(a)(p(a) + o_k(a))\,. \tag{4}$$

3.4. Multi-Task Network Optimization

Training. The cropping and composition branches are optimized simultaneously on the corresponding datasets. In each training iteration, X_1 and X_2 are sequentially fed into the backbone. Then X_1 passes through both branches while X_2 only goes into the composition branch. Similar to the training strategy in [40], the network parameters are updated using the accumulated gradients. The cropping loss $\mathcal{L}_{crop}$ uses the smooth ℓ_1 loss [26] between the predicted box and the ground-truth box. The composition loss $\mathcal{L}_{com}$ is defined by the cross entropy between the predicted confidence score s and the composition label. The whole network is trained in an end-to-end manner by minimizing the following loss function

$$\mathcal{L} = \lambda \mathcal{L}_{crop} + (1 - \lambda)\mathcal{L}_{com}\,, \tag{5}$$

where λ is a balancing factor.

Inference. Given a source image, CACNet outputs i) a cropping box, ii) a KCM that highlights the discriminative composition, and iii) one or more composition rules the image follows. The co-presentation of KCM, composition distribution, and the cropping box provides viewers with composition guidance more than a simple cropping result.

4. Experiments

4.1. Datasets and Implementation Details

Composition Dataset. To train the composition branch, we use the KU-PCP dataset [17], which consists of 4244 outdoor photographs. We split the data into a training set of 3169 images and a validation set of 1075 images. Each image is categorized into 9 photographic composition classes: `rule-of-thirds` (RoT), `center` (Cen.), `horizontal` (Hor.), `symmetric` (Sym.), `diagonal` (Dia.), `curved` (Cur.), `vertical` (Ver.), `triangle` (Tri.), and `repeated pattern` (Pat.). Since an image may follow multiple composition rules, each sample is given with at least one label (≤ 3 labels).

Table 1. Quantitative comparisons against state-of-the-art methods on the FCDB [4] and FLMS [11]. Best performance is in **bold**, and the second best is underlined. To measure the efficiency, we report FPS on our hardware. For methods that do not provide codes, we cite their reported FPS and indicate using †.

Method	FCDB		FLMS		FPS↑
	BDE↓	IoU↑	BDE↓	IoU↑	
non-real time (FPS<30)					
VFN[5] (MM'17)	0.084	0.685	-	-	0.4
VEN[37] (CVPR'18)	0.072	0.735	0.041	0.837	0.3
DIC[34] (ICCV'17)	0.090	0.630	0.057	0.810	4.4†
DIC*[35] (TPAMI'18)	0.080	0.650	0.052	0.830	4.4†
A2-RL[18] (CVPR'18)	0.089	0.664	0.045	0.821	2.6
A3-RL[19] (TIP'19)	0.077	0.696	-	0.839	24.3†
ASMNet[33] (AAAI'20)	**0.068**	**0.749**	0.039	0.849	1.1
real time (FPS≥30)					
VPN[37] (CVPR'18)	0.073	0.711	0.044	0.836	96.2
GAIC[42] (CVPR'19)	0.081	0.674	0.041	0.834	129.8
CGS[21] (CVPR'20)	-	-	0.039	0.836	100.0†
CACNet (Ours)	0.069	0.718	**0.033**	**0.854**	**155.0**

Cropping Datasets. Two image cropping benchmarks, *i.e.*, FCDB [4] and FLMS [11], are used for evaluation. FCDB contains 1743 images, among which, 1395 images are used for training and 348 for testing. For each image, a single ground truth box is annotated. For FLMS, we use all 500 images for testing. Each image is annotated with 10 cropping boxes. We adopt the metric of the top-1 maximum overlap following previous works [11, 14] on this dataset.

Implementation Details. CACNet is trained using the Adam optimizer [16] with weight decay of $1\mathrm{e}{-4}$. The initial learning rate is set to $3.5\mathrm{e}{-4}$ and decays by $\times 0.1$ every 5 epochs. The backbone exploits all convolution blocks excluding the last max pool layer of VGG-16 [28] pre-trained on ImageNet [8]. The composition branch is of $\frac{1}{4}$ input resolution, while the cropping branch is of $\frac{1}{16}$ input resolution. We augment data with random horizontal flipping. The balancing factor λ is set to 0.7. Sensitivity experiments of λ and details on CACNet technical architecture are placed in the supplementary materials. All experiments are conducted on a single NVIDIA RTX 2080Ti GPU.

4.2. Comparison With State of the Art

Here we compare CACNet against existing state-of-the-art approaches. Following [18, 19, 37], we use the average intersection-over-union (IoU) and the average boundary displacement error (BDE) as metrics. Also, we report the frames per second (FPS) to evaluate the efficiency.

Quantitative Results. Table 1 reports the quantitative results. Our method exhibits not only good generalization performance but also fast inference speed. On both the FLMS and FCDB datasets, CACNet outperforms or at least is on par with state-of-the-art methods. Note that CACNet yields a substantial performance improvement on

Figure 5. Qualitative comparison of different methods. Our method generates better composed results close to ground truths.

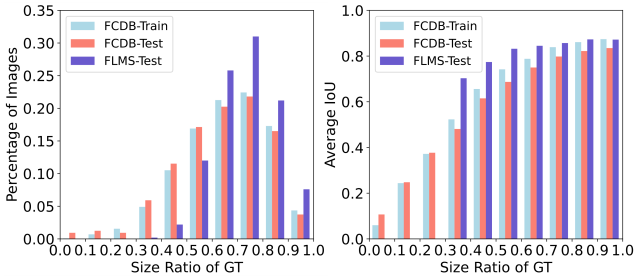

Figure 6. *Left:* Percentage of images vs. size ratio of GT (size of GT box / size of source image). One can see data imbalance in GT size ratio. *Right:* Average IoU vs. size ratio of GT. CACNet performs well on images with relatively larger annotations, but not the case on small-size annotated ones.

the FLMS dataset that is not seen during training. The good generalization may be the introduction of the composition branch and the explicit KCM embedding of professional photographic knowledge, which leads to better composition-aware feature learning. The good generalization also implies the network indeed captures underlying composition rules, rather than simply exploiting data biases. CACNet also runs at rates of 155.0 FPS. The high efficiency can boil down to the one-stage regression manner of CACNet, which avoids many time-consuming procedures used in other methods such as per-candidate scoring. Particularly, CACNet significantly outperforms the previous top-3 fast methods on both datasets.

Despite being trained on the FCDB dataset, CACNet does not achieve the best performance on the corresponding test set. One possible reason is the data imbalance in the size of ground truth (GT) cropping, as shown in Fig. 6. One can also observe performance degrades in images with small-size annotations. Unlike other candidates-based methods, such as VFN, VEN, VPN, ASMNet, and CGS, that can densely preset many small-size candidates, CACNet can suffer from data imbalance and has little chance to fit small-size cropping. The other reason may lie in the learning mechanism of CACNet. It learns to crop from the source composition then models potential spatial composition rules from relations between the KCM and expert annotations. Therefore, CACNet prefers to elaborate composition from a global perspective, rather than cropping a locally salient or aesthetic area. In practice, people tend to preserve the major contents of source images after cropping. That is, the cropping box should not be too small, which is also argued in [42].

Qualitative Analysis. Qualitative comparisons are shown in Fig. 5. One can observe: i) VFN and VEN are prone to crop part of the main subjects, breaking the overall balance and composition; ii) VPN tends to produce compacted cropping results; iii) A2RL cannot cut out unnecessary distractions in some cases; iv) GAIC can generate relatively appealing results and maintain the major contents, but the results still cannot be called a sufficiently good composition; and v) CACNet, by contrast, learns to construct spatial composition and can stably produce well-composed cropping results close to ground truths.

We further demonstrate the interpretability of CACNet by showing the KCM, weighted anchor points, and the composition distributions in Fig. 7. The KCM reveals the spatial composition clues of source images. For instance, the KCM highlights the leading lines of the `diagonal` rule or the dominant subjects of the `center` and `RoT` rule. It can be used to guide ordinary people (especially newbies) to learn what fundamental elements should be focused on and how to improve composition with them. In addition, the histogram distributions also provide a comparison of composition before and after cropping. Both the KCM and the composition distributions provide our method with interpretable evidences for composition-aware cropping. Fig. 8 shows the KCM and cropping results of 9 composition rules. More qualitative results can refer to the supplementary materials.

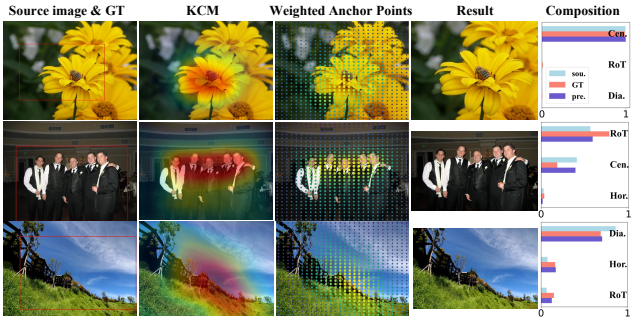

Figure 7. Interpretable image cropping. Discriminative composition clues are highlighted by the KCM. The right column shows the predicted top-3 composition distributions of the source images, GT, and the cropping results, respectively.

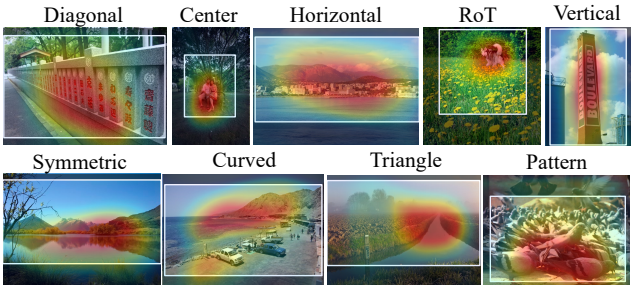

Figure 8. Results of interpretable cropping of 9 composition rules.

4.3. Ablation Study

To justify each design choice of CACNet, we conduct ablation studies on the FCDB. In particular, we use 1195 images for training and the held-out images for validation.

Comparison With the Baseline. Compared with the baseline that directly regresses the GT cropping box, our full model incorporates an anchor-point regressor, a composition branch, and the anchor-point weight assignment with KCM. The results of different combinations of these components are shown in Table 2. One can observe:

- *The KCM benefits image cropping* (B4 vs. B5): Compared with training without the KCM where each anchor point shares the same weight, the anchor-point weight assignment with KCM improves the BDE by 4.0% and IoU by 1.7%. This corroborates the importance of injecting basic composition concepts into image cropping and of the modeling of composition rules with the KCM.
- *Learning composition rules is helpful* (B1 vs. B3, B2 vs. B4): With the composition branch introduced to the baseline, there is an improvement of 16.0% BDE and 5.1% IoU. Compared with the model addressing only the cropping task, CACNet is trained in a multi-task manner and can be augmented by inductive knowledge transfer [1, 15], thus reducing the risk of overfitting. Particularly, learning to distinguish composition rules allows spatial relations to be explored for image cropping.

Table 2. Ablation study on each module of CACNet. **base.**: baseline that directly regresses the ground-truth cropping box; **anc.**: anchor-point regressor; **CB**: composition branch; **KCM**: anchor-point weight assignment with KCM; Δ represents the relative improvement compared with the baseline.

No.	base.	anc.	CB	KCM	BDE↓	IoU↑	ΔBDE↑	ΔIoU↑
B1	✓				0.075	0.703	-	-
B2	✓	✓			0.069	0.720	8.0%	2.4%
B3	✓		✓		0.063	0.739	16.0%	5.1%
B4	✓	✓	✓		0.061	0.746	18.7%	6.1%
B5	✓	✓	✓	✓	**0.058**	**0.758**	**22.7%**	**7.8%**

Table 3. Ablation study on the impact of composition.

Composition Accuracy (%)	w/o KCM		w/ KCM	
	BDE↓	IoU↑	BDE↓	IoU↑
50.0%	0.070	0.720	0.067	0.730
60.0%	0.070	0.719	0.063	0.739
70.0%	0.069	0.722	0.063	0.743
88.2% (top)	0.070	0.721	**0.062**	**0.746**

Table 4. Ablation study on the composition loss function. **BCE**: binary cross entropy; **CE**: cross entropy.

Composition Loss	BDE↓	IoU↑
multi-label (BCE)	0.061	0.749
single-label (CE)	**0.058**	**0.758**

- *Anchor-point regressor works* (B1 vs. B2, B3 vs. B4): The anchor-point regressor improves results by 8.0% BDE and 2.4% IoU, which verifies the effectiveness of one-stage regression. Compared with the baseline, local spatial context information can be better explored [39].

Impact of Composition. A key opinion of this work is that encoding composition rules can facilitate composition-aware image cropping. Here we further verify the effectiveness of the KCM and investigate the impact of the composition branch (CB). We design an experiment as follows: i) training the backbone and the CB until the composition accuracy reaches a fixed value; ii) freezing the CB; and iii) fine-tuning the backbone and the cropping branch. Table 3 reports the results of training with and without the KCM with different frozen accuracy rates of CB. Surprisingly, when training without the KCM, the learning of the CB has little impact on cropping. In this case, the encoding of composition rules only affects the pre-trained backbone. When training with the KCM, the cropping accuracy increases with increased composition accuracy. This finding further highlights the importance of the KCM. Higher composition accuracy enables more accurate KCM generation, and therefore more effective image cropping.

Additionally, compared with the training strategy above, it is more effective to jointly train the two branches (B4 and B5 in Table 2). We speculate that the features of the two tasks can be better aggregated in this way. We also find

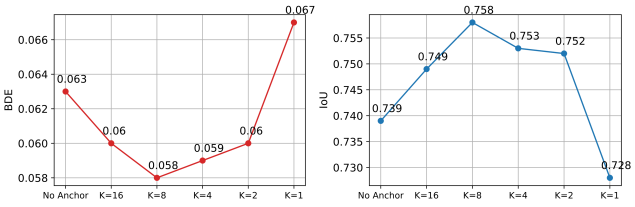

Figure 9. Impact of anchor-point stride K. Both BDE and IoU are reported. Smaller K corresponds to denser anchor points. The performance reaches a peak when $K = 8$.

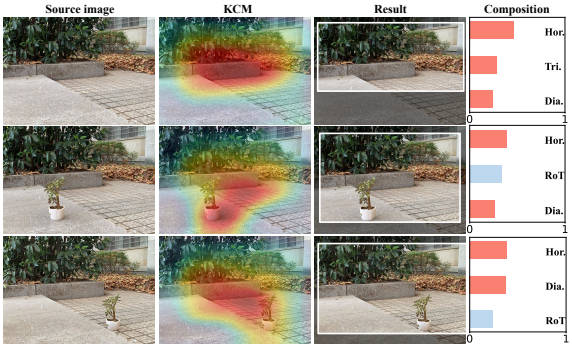

Figure 10. Changed cropping with changed composition. 3 source images are taken under the same scene but with different positions of a potted plant. CACNet can adaptively generate well-composed cropping according to the key elements of composition.

that, when training the CB, duplicating multi-label images for each ground-truth class can be more helpful than using a multi-label loss (Table 4). The reason is likely that most images are single-labeled. Since the classification performance is crucial to the cropping branch, we show the corresponding results in the supplementary materials.

Impact of Anchor-Point Stride K**.** Fig. 9 compares results with different strides of K. Performance is improved when no-anchor training is altered to anchor-point training. CACNet achieves the best BDE and IoU when $K = 8$, while its performance degrades when K further decreases. This indicates that dense anchor points can be harmful to the performance. In our view, the KCM trained in a weakly-supervised manner cannot provide accurate pixel-level weight assignment, but instead a global coarse activation. A coarse KCM cannot support dense anchor points. $K = 8$ is sufficient.

4.4. Composition-Aware Image Cropping

A good cropping method can adaptively re-compose images when the elements of composition change. We demonstrate an interesting application of our method. As shown in Fig. 10, we fix the camera and take 3 photos under the same scene. The only difference is where we place the potted plant. One can find that CACNet can be aware of the key elements of composition and generate composition-aware cropping results.

Table 5. User study results. A good cropping method favors more 'Good' and less 'Bad'.

Method	**Ours**	GAIC	VPN	VEN	A2RL	VFN
Good (%)	**48.1**	34.1	39.3	29.6	21.5	13.2
Normal (%)	**46.9**	52.7	46.8	34.1	54.8	22.8
Bad (%)	**5.0**	13.2	13.9	36.3	23.7	64.0

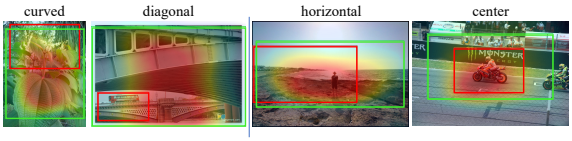

Figure 11. Failure cases. The top-1 predicted composition rule is shown above each image. GT and the predicted results are drawn in red and green boxes, respectively.

4.5. User Study

Since composition is subjective, we further conduct an user study to validate different methods. We randomly choose 155 images, 95 from the FLMS [11] and 60 from the test set of FCDB [4]. 15 people with photographing experience are invited. Results in Table 5 show that most users favor our method. More details can refer to the supplementary materials.

4.6. Analysis of Failure Cases

While CACNet in most cases can generate well-composed, interpretable cropping, it can fail in some circumstances. Some interesting failure cases are shown in Fig. 11 where the composition rule and the KCM are not accurately predicted and the ground truth box is relatively small. This reveals a limitation of our method: it tends to preserve major visual contents.

5. Conclusion

Inspired by composition rules in photography, we present a novel cropping-by-composition paradigm to implement composition-aware image cropping. We show that explicit modeling of composition rules benefits image cropping. To model such rules, we introduce the concept of the KCM and implement it in a composition-aware cropping network CACNet. Extensive results justify that CACNet enables effective, interpretable, and fast image cropping, providing a strong and efficient baseline to the image cropping community. We also showcase additional analyses to highlight the role of composition. For future work, we plan to explore flexible cropping models applicable to different size ratios. Our work also opens a door for composition-specific cropping.

Acknowledgements. This work was funded by the DigiX Joint Innovation Center of Huawei-HUST.

References

[1] Jonathan Baxter. A model of inductive bias learning. *J. Artif. Intell. Res.*, 12:149–198, 2000. 7

[2] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 507–515, 2016. 2

[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 Syst.*, 9(4):353–364, 2003. 2

[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 *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pages 226–234, 2017. 3, 5, 8

[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 *Proc. ACM Int. Conf. Multimedia*, pages 37–45, 2017. 2, 5

[6] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. Learning to photograph. In *Proc. ACM Int. Conf. Multimedia*, pages 291–300, 2010. 1

[7] Gianluigi Ciocca, Claudio Cusano, Francesca Gasparini, and Raimondo Schettini. Self-adaptive image cropping for small displays. *IEEE Trans. Consum. Electron.*, 53(4):1622–1627, 2007. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 5

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proc. Int. Conf. Comput. Vis.*, pages 6569–6578, 2019. 4

[10] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proc. ACM Int. Conf. Multimedia*, pages 1105–1108, 2014. 1, 2

[11] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proc. ACM Int. Conf. Multimedia*, pages 1105–1108, 2014. 5, 8

[12] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1072–1080, 2015. 2

[13] Paul Jonas. *Photographic composition simplified*, volume 149. Amphoto, 1976. 2

[14] Yueying Kao, Ran He, and Kaiqi Huang. Automatic image cropping with aesthetic map and gradient energy map. In *Proc. ICASSP*, pages 1982–1986, 2017. 2, 5

[15] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7482–7491, 2018. 7

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *J. Vis. Commun. Image Represent.*, 55:91–105, 2018. 2, 3, 5

[18] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8193–8201, 2018. 2, 5

[19] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *IEEE Trans. Image Process.*, 28(10):5105–5120, 2019. 2, 5

[20] Debang Li, Junge Zhang, and Kaiqi Huang. Learning to learn cropping models for different aspect ratio requirements. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12685–12694, 2020. 2

[21] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4213–4222, 2020. 2, 5

[22] Zhuopeng Li and Xiaoyan Zhang. Collaborative deep reinforcement learning for image cropping. In *Proc. Int. Conf. Multimedia and Expo*, pages 254–259, 2019. 2

[23] Matthew Ma and Jinhong K Guo. Automatic image cropping for mobile device with built-in camera. In *Proc. IEEE Consum. Commun. Networking Conf.*, pages 710–711, 2004. 2

[24] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Proc. Int. Conf. Comput. Vis.*, pages 2232–2239, 2009. 2

[25] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato. Sensation-based photo cropping. In *Proc. ACM Int. Conf. Multimedia*, pages 669–672, 2009. 2

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Adv. Neural Inform. Process. Syst.*, pages 91–99, 2015. 5

[27] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *Proc. SIGCHI Conf. Hum. Fact. Comput. Syst.*, pages 771–780, 2006. 2

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Represent.*, 2015. 5

[29] Fred Stentiford. Attention based auto image cropping. In *Proc. Int. Conf. Comput. Vis. Syst.*, 2007. 2

[30] Bongwon Suh, Haibin Ling, Benjamin B Bederson, and David W Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proc. Annu. ACM Symp. User Interface Softw. Technol.*, pages 95–104, 2003. 2

[31] Jin Sun and Haibin Ling. Scale and object aware image thumbnailing. *Int. J. Comput. Vis.*, 104(2):135–153, 2013. 2

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proc. Int. Conf. Comput. Vis.*, pages 9627–9636, 2019. 4

[33] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *Proc. AAAI*, pages 12104–12111, 2020. 2, 5

[34] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *Proc. Int. Conf. Comput. Vis.*, pages 2186–2194, 2017. 2, 5

[35] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1531–1544, 2018. 2, 5

[36] Wenguan Wang, Jianbing Shen, Yizhou Yu, and Kwan-Liu Ma. Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Trans. Vis. Comput. Graph.*, 23(8):2014–2027, 2016. 2

[37] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5437–5446, 2018. 2, 5

[38] Wikipedia contributors. Cropping (image) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Cropping_(image)&oldid=847382681, 2018. [Online; accessed 17-October-2020]. 1

[39] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proc. Int. Conf. Comput. Vis.*, pages 793–802, 2019. 4, 7

[40] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 675–684, 2018. 5

[41] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 971–978, 2013. 2

[42] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5949–5957, 2019. 2, 5, 6

[43] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2921–2929, 2016. 2, 3, 4