

AesCrop: Aesthetic-driven Cropping Guided by Composition

Yen-Hong Wong Lai-Kuan Wong

Faculty of Computing and Informatics, Multimedia University
 Persiaran Multimedia, Cyberjaya, Selangor, Malaysia.

1211101392@student.mmu.edu.my lkwong@mmu.edu.my

Abstract

Aesthetic-driven image cropping is crucial for applications like view recommendation and thumbnail generation, where visual appeal significantly impacts user engagement. A key factor in visual appeal is composition—the deliberate arrangement of elements within an image. Some methods have successfully incorporated compositional knowledge through evaluation-based and regression-based paradigms. However, evaluation-based methods lack globality while regression-based methods lack diversity. Recently, hybrid approaches that integrate both paradigms have emerged, bridging the gap between these two to achieve better diversity and globality. Notably, existing hybrid methods do not incorporate photographic composition guidance, a key attribute that defines photographic aesthetics. In this work, we introduce AesCrop, a composition-aware hybrid image-cropping model that integrates a VMamba image encoder, augmented with a novel Mamba Composition Attention Bias (MCAB) and a transformer decoder to perform end-to-end rank-based image cropping, generating multiple crops along with the corresponding quality scores. By explicitly encoding compositional cues into the attention mechanism, MCAB directs AesCrop to focus on the most compositionally salient regions. Extensive experiments demonstrate that AesCrop outperforms current state-of-the-art methods, delivering superior quantitative metrics and qualitatively more pleasing crops.

1. Introduction

Image cropping is the process of removing unwanted outer areas from an image. To crop an image in a visually appealing way, a solid understanding of image composition principles is essential. These principles guide the arrangement of elements within an image, ensuring balance, focus, and aesthetic harmony, which in turn directs the viewer’s attention to the intended subject and conveys the desired message

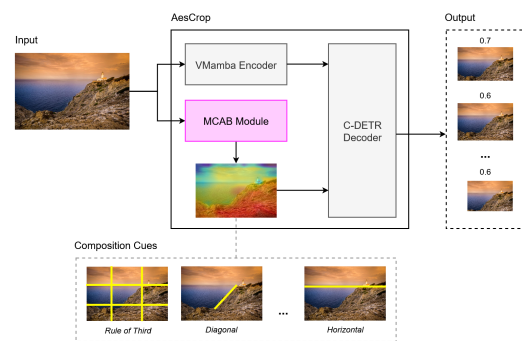


Figure 1. AesCrop is a composition-aware, hybrid image-cropping model that incorporates cues from multiple composition classes via the Mamba Composition Attention Bias (MCAB) and performs end-to-end rank-based image cropping.

or emotion. However, without sufficient knowledge of aesthetic and photographic principles, amateurs often produce undesirable and unappealing crops.

To address this gap, researchers introduced automatic image cropping to make aesthetic-driven cropping more accessible. Early image cropping approaches can be categorized primarily into two paradigms: evaluation-based [1, 10, 15–18] and regression-based [2–4, 11, 14] methods. Evaluation-based approaches generate a fixed number of candidate crops using heuristic rules, evaluate their aesthetic quality, and rank them using a separate aesthetics evaluation model. Regression-based methods directly regress the best candidate crop. In particular, incorporation of explicit compositional knowledge into image-cropping methods has been proven effective in enhancing the aesthetic appeal of cropped images in both evaluation-based [17] and regression-based [3, 14] approaches. However, both paradigms have limitations: the former lacks globality as heuristic-based crop candidates might miss the optimal choice, while the latter lacks diversity since generating a single crop may be insufficient. Recently, hybrid approaches [5, 19] that simultaneously regress and evalu-

ate multiple candidate crops have been proposed to achieve better diversity and globality. These methods can generate pleasing crops but none consider explicit compositional guidance, which could potentially further enhance the aesthetics quality of cropped images.

In this work, we introduce AesCrop, a novel, composition-aware hybrid cropping model that incorporates explicit compositional cues to guide cropping, as illustrated in Fig. 1. Capitalizing on the strength of Mamba’s encoder in its efficient long-range dependency modeling with linear computational complexity, we integrate a VMamba-based encoder [8] that generates rich embedding of the input image with a Conditional DETR-based transformer decoder [9] that regresses and scores multiple candidate crops using learnable anchor embeddings. To incorporate explicit compositional knowledge, we propose the Mamba Composition Attention Bias (MCAB), a novel mechanism that extracts composition cues from a VMamba-based [8] composition classifier and injects them into the transformer decoder by modulating its attention weights. MCAB guides the decoder to focus on compositionally salient regions, yielding crops that adhere to established compositional principles.

In summary, our main contributions are as follows:

- We present AesCrop, a novel hybrid image cropping model that incorporates explicit compositional knowledge to simultaneously achieve globality, diversity, and compositional awareness. Specifically, Mamba Composition Attention Bias (MCAB), a novel mechanism that extracts compositional cues from a Mamba-based classifier and injects them into attention weights is introduced to effectively guide the model to focus on compositionally important regions.
- Comprehensive experiments and ablation studies validate our architectural design choices and demonstrate that AesCrop consistently outperforms state-of-the-art methods in generating aesthetically pleasing crops that adhere more faithfully to professional photographic composition rules.

2. Related Work

Existing approaches can be broadly categorized into three paradigms: evaluation-based, regression-based, and hybrid-based methods.

Evaluation-Based. In this paradigm, the majority of works follow a two-step process: (1) generating a number of candidate crops using heuristic rules, and (2) ranking each candidate crop to determine the best crops. Within this framework, several representative approaches have pushed the state of the art by improving either candidate generation or crop scoring. Yan *et al.* [16] proposed a method that uses regional features taking changes into account to generate candidate crops, then scores these crops with an SVM. VFN [1] assumes that web images have the best aes-

thetics, allowing the model to learn from large-scale pairwise ranking in an unsupervised manner. VPN [15] learns to score image crops in real-time using a novel knowledge distillation framework. GAIC [18] reduces the candidate search space using special properties of image cropping, then ranks the crops using a CNN that models the RoI and RoD. TransView [10] utilizes a transformer-based architecture to model dependencies between visual elements inside and outside the bounding box for crop ranking. CLIPCropping [17] leverages CLIP to learn aesthetic and composition concepts from dedicated image-text pairs for crop ranking.

Regression-Based. In this paradigm, most approaches directly regress the coordinates of the optimal crop through a single-stage process. Within this paradigm, several methods have driven progress by refining regression strategies or by injecting domain knowledge into the prediction process. Guo *et al.* [2] introduced a cascaded regression method that progressively refines crop predictions through iterative adjustments. Mars [7] adapts a base model to diverse aspect ratio requirements by employing two meta-learners that dynamically generate model parameters based on target aspect ratios, enabling flexible and precise cropping across various formats. CACNet [3] utilizes Class Activation Maps (CAMs) to extract compositional features through a dedicated composition branch, which are then explicitly integrated into the cropping branch to enhance output quality. Pan *et al.* [11] operates on the principle that crops with similar boundaries share compositional characteristics, proposing C2C to regularize these features for boundary prediction that generalizes across both common and rare patterns. GenCrop [4] builds on the assumption that stock images exhibit optimal aesthetics, training its model on a large-scale image-crop dataset through a diffusion-based outpainting approach.

Hybrid-Based. The paradigm was first introduced by Jia *et al.* [5], who rethink image cropping through a hybrid method using Conditional DETR with a novel training strategy adapted from DETR. Their work introduced two label smoothing techniques to provide enhanced learning signals for crop ranking. Later, ClipCrop [19] extended this approach by incorporating a CLIP image encoder and adding conditioned embeddings to the decoder’s learnable anchors, enabling text- and image-conditional cropping. Notably, neither of these works in the hybrid paradigm has explicitly incorporated compositional knowledge.

3. Methodology

To address the lack of compositional guidance in the hybrid-based paradigm, we propose AesCrop, a novel hybrid-based, composition-aware image cropping framework. AesCrop features dual-stream VMamba encoders [8] that simultaneously generate the novel Mamba Composition Attention Bias (MCAB) and image embeddings, along with

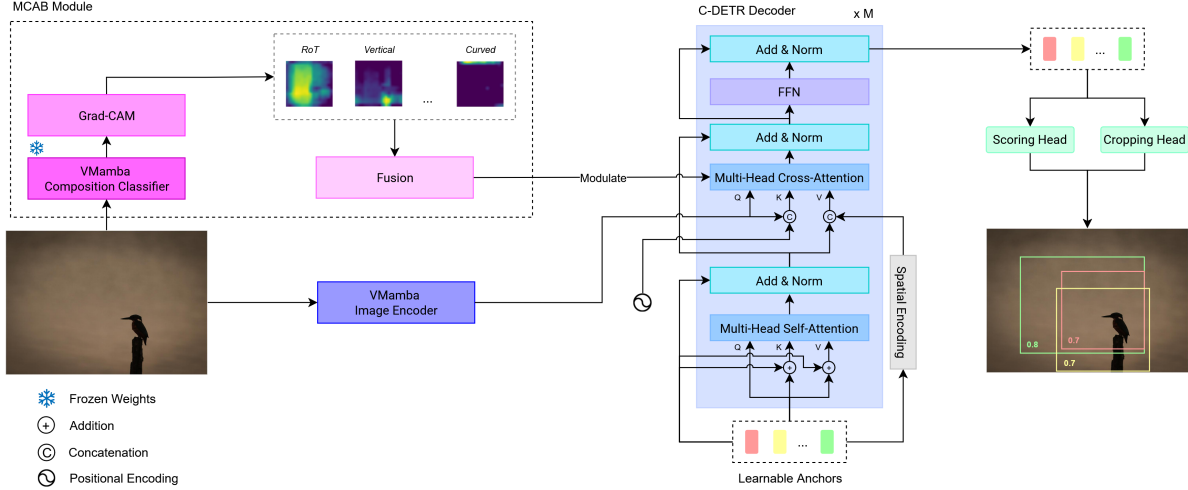


Figure 2. **The architecture of AesCrop.** The input image is processed by dual-stream encoders: (1) a module that generates the Mamba Composition Attention Bias (MCAB) and (2) an encoder that produces image embeddings. A decoder then refines learnable queries via self-attention, aggregates those queries with the image embeddings through MCAB-modulated cross-attention, and processes the result through feed-forward networks. The final embeddings are passed to prediction heads that output crop boxes and quality scores.

a Conditional DETR-based decoder inspired by the design of Jia et al. [5]. The complete architecture is illustrated in Fig. 2.

Our framework begins with a dual-stream encoding architecture designed to jointly capture both visual semantics and compositional structure. The visual processing stream utilizes a VMamba image encoder $\mathcal{F}_{\text{enc}}$ that transforms the input image $\mathbf{I}$ into comprehensive image embeddings $\mathbf{E}$, which capture essential visual features for crop selection:

$$\mathbf{E} = \mathcal{F}_{\text{enc}}(\mathbf{I}) \quad (1)$$

Simultaneously, the compositional stream generates the composition attention bias, denoted by $\mathbf{B}$ through the $\mathcal{F}_{\text{MCAB}}$ module, encoding explicit compositional guidance:

$$\mathbf{B} = \mathcal{F}_{\text{MCAB}}(\mathbf{I}) \quad (2)$$

These components, along with the learnable anchor queries $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^N$, are processed by a Conditional DETR (C-DETR) [9] decoder $\mathcal{F}_{\text{dec}}$ with M layers. Each decoder layer first applies self-attention to model global dependencies among the learnable anchors queries, establishing their contextual relationships. The layer then performs MCAB-modulated cross-attention, where the contextualized queries interact with the image embeddings $\mathbf{E}$ under the guidance of the compositional bias $\mathbf{B}$ generated by MCAB. This attention mechanism dynamically prioritizes compositionally salient regions by adjusting attention weights based on the composition priors from $\mathbf{B}$. Following these cross attention operations, the features undergo non-linear transformation through feed-forward networks (FFNs). Through M decoder layers, the decoder progressively integrates visual features, spatial positions, and

compositional knowledge into the final output embeddings $\mathbf{O} = \{\mathbf{o}_i\}_{i=1}^N$, where $\mathbf{o}_i$ represents the refined embedding for a candidate crop:

$$\mathbf{O} = \mathcal{F}_{\text{dec}}(\mathbf{E}, \mathbf{B}, \mathbf{Q}) \quad (3)$$

The final processing stage employs two specialized prediction heads that operate on each output embedding $\mathbf{o}_i$. The crop prediction head f_{crop} generates normalized bounding box coordinates $\hat{\mathbf{b}}_i \in [0, 1]^4$ for each candidate crop, consisting of the predicted center coordinates $(\hat{c}_x^i, \hat{c}_y^i)$ and dimensions $(\hat{w}^i, \hat{h}^i)$. Simultaneously, the scoring head f_{score} produces a normalized quality score $\hat{v}_i \in [0, 1]$ for the corresponding candidate crop:

$$f_{\text{crop}}(\mathbf{o}_i) = \hat{\mathbf{b}}_i = [\hat{c}_x^i, \hat{c}_y^i, \hat{w}^i, \hat{h}^i] \quad (4)$$

$$f_{\text{score}}(\mathbf{o}_i) = \hat{v}_i \quad (5)$$

3.1. Mamba Composition Attention Bias

The Mamba Composition Attention Bias (MCAB) module enables AesCrop to incorporate compositional knowledge by adaptively weighting compositionally salient regions. Specifically, the composition attention bias $\mathbf{B}$ is generated through a multi-step process via $\mathcal{F}_{\text{MCAB}}$. First, a frozen VMamba composition classifier pretrained on the KU-PCP dataset [6], which consists of nine image composition classes, is used to output a probability p_i for each class i . For each predicted class i , Grad-CAM [13] produces a corresponding Class Activation Map (CAM) $\mathbf{C}_i$ by computing gradients of the composition class score with respect to the feature maps. Each $\mathbf{C}_i$ is a heatmap that matches the

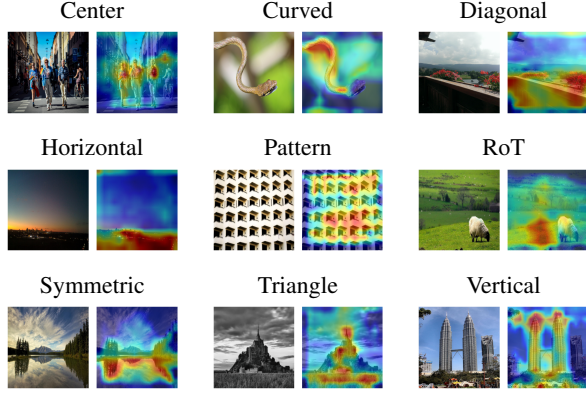


Figure 3. Class activation maps of nine images representing distinct composition classes.

dimension of the input image, with values normalized between $[0, 1]$ from least important to most important, highlighting the regions most relevant to the composition class i . Examples of these CAMs are illustrated in Fig. 3.

The CAMs are then aggregated through probability-weighted averaging in the fusion block to generate composition attention bias, $\mathbf{B}$:

$$\mathbf{B} = \sum_{i=1}^N p_i \cdot \mathbf{C}_i \quad (6)$$

This operation produces $\mathbf{B}$, a unified global heatmap that captures the attentive regions corresponding to each composition class. Then, $\mathbf{B}$ is normalized to the range $[0, 1]$. The resulting $\mathbf{B}$ is downsampled and flattened to spatially align with the image encoder’s output dimensions, yielding the final MCAB.

Traditionally, cross-attention between queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ is computed as:

$$\text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (7)$$

where $\mathbf{Q}\mathbf{K}^T$ represents the attention weights, a matrix of shape $|\mathbf{Q}| \times |\mathbf{K}|$ whose element at the i -th row and j -th column indicates how much query i attends to key j . In Ae-sCrop, this determines how much learnable anchor i focuses on image patch j .

To inject composition knowledge, we modify the cross-attention mechanism as follows:

$$\text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \log \mathbf{B} \right) \mathbf{V} \quad (8)$$

Here, $\mathbf{B}$ undergoes log-scaling to amplify the weight of compositionally salient regions relative to non-salient ones, yielding $\log \mathbf{B}$. This log-scaled term is then added to the original attention weights, biasing the model toward compositionally important regions by adjusting the attention of

learnable anchor i toward image patch j . This modification directs the decoder’s focus to areas rich in compositional elements, significantly improving the model’s ability to generate crop proposals that adhere to compositional principles.

3.2. Hybrid Cropping Training Strategy

Training a hybrid cropping network without fixed anchors or post-processing steps such as non-maximum suppression requires an effective matching strategy between predictions and ground truth crops. Following Jia *et al.* [5], we first isolate a set of high-quality ground truths, denoted y_{good} , by selecting those with mean opinion scores $s_i \geq 4$. Additionally, we pad y_{good} with empty objects so as to match the cardinality of $\hat{y}$. Next, we employ the Hungarian Algorithm to establish a one-to-one correspondence $\hat{\sigma}$ between our predicted crops $\hat{y}$ and these high-quality annotations. This correspondence is chosen to minimize the aggregate matching loss:

$$\hat{\sigma} = \arg \min_{\sigma} \sum_{i=1}^N \mathcal{L}_{\text{match}}(\hat{y}_i, y_{\text{good}, \sigma(i)}) \quad (9)$$

Here, the matching loss $\mathcal{L}_{\text{match}}$ integrates three components:

$$\begin{aligned} \mathcal{L}_{\text{match}}(y_{\text{good}, i}, \hat{y}_{\sigma(i)}) = & \mathcal{L}_{\text{reg}}(\hat{b}_i, b_{\text{good}, \sigma(i)}) \\ & + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}(\hat{b}_i, b_{\text{good}, \sigma(i)}) \\ & + \lambda_{\text{focal}} \mathcal{L}_{\text{focal}}(\hat{v}_i, v_{\text{good}, \sigma(i)}) \end{aligned} \quad (10)$$

where $\mathcal{L}_{\text{reg}}$, $\mathcal{L}_{\text{GIoU}}$, and $\mathcal{L}_{\text{focal}}$ are the L1 regression loss, the Generalized IoU loss, and the focal loss, respectively, with λ_{GIoU} and λ_{focal} as weighting hyperparameters.

For predicted crops matched to non-empty, high-quality ground truths, we supervise both their bounding boxes and quality scores using those matched annotations. However, the majority of predictions will be matched to an empty object. To provide additional learning signals for these predictions, any unmatched prediction that nevertheless overlaps substantially ($\text{IoU} \geq 0.85$) with a ground truth crop is supervised by assigning it a soft label estimated from the quality score of its high-IoU neighbor, thus exploiting local redundancy. Finally, for all remaining predictions that neither match a high-quality crop nor exhibit substantial overlap, we set their quality scores to zero, providing explicit negative supervision for low-quality crops.

4. Experiments

4.1. Datasets

We use GAIC [18] dataset as our cropping dataset, as it is the only sufficiently densely annotated dataset suitable for training hybrid-based cropping models [5]. Specifically, we adopt GAICv2, which offers a larger number of images compared to its predecessor, GAICv1. Additionally,

we use KU-PCP [6] as the composition dataset to train our VMamba composition classifier.

GAICv2 [18] contains 3,336 images, split into 2,636 training samples, 200 validation samples, and 500 test samples. Each image includes up to 90 candidate crops annotated with mean opinion scores (MOS) ranging from 1 to 5. The dataset demonstrates strong annotator agreement, with 94.25% of crops having a score standard deviation below 1.

KU-PCP [6] consists of 4,244 images, divided into 3,169 training samples and 1,075 testing samples. Each image is annotated with one or more of the nine composition classes: *rule of thirds*, *center*, *horizontal*, *vertical*, *symmetric*, *diagonal*, *curved*, *triangle*, and *pattern*.

4.2. Evaluation Metrics

We evaluate our method using the Return K of Top N accuracy metric, $Acc_{K/N}$ [18], adapted for hybrid-based approaches [5]. The metric is formally defined as:

$$Acc_{K/N} = \frac{1}{TK} \sum_{i=1}^T \sum_{j=1}^K \mathbb{I} \left(\max_{g \in S_i(N)} \text{IoU}(c_{i,j}, g) \geq \epsilon \right) \quad (11)$$

where $\mathbb{I}(\cdot)$ is an indicator function that returns 1 when the condition is true and 0 otherwise, ϵ is the IoU threshold for crop matching, T is the total number of test examples, $c_{i,j}$ represents the top j -th predicted crop for the i -th example, and $S_i(N)$ denotes the set of top N ground truth crops for the i -th example.

We additionally employ the averaged metric $\overline{Acc}_N$, computed as:

$$\overline{Acc}_N = \frac{1}{|S_K|} \sum_{K \in S_K} Acc_{K/N} \quad (12)$$

where S_K represents the set of K values under consideration, and $|S_K|$ denotes its cardinality.

4.3. Implementation Details

Dataset Configuration: Both cropping and composition datasets undergo similar preprocessing pipelines, with cropping datasets requiring additional steps. All images are first resized to a fixed 512×512 resolution to preserve high-quality details for accurate predictions. During training, we apply multiple data augmentation techniques to enhance model robustness, including random variations in saturation, hue, and contrast. Following augmentation, we normalize all images using ImageNet’s mean and standard deviation values to ensure training stability. For cropping datasets, we perform additional processing by normalizing ground truth coordinates to the $[0,1]$ range and converting them to the standardized c_x, c_y, w, h format.

Model Pretraining: AesCrop employs a two-stage pre-training approach before final training. First, we pretrain the

composition classifier on KU-PCP [6] and keep its weights frozen during main training to ensure stable generation of MCAB. Second, we pretrain the remaining components on COCO-minitrain [12], a carefully selected 25,000 image subset of the COCO object detection dataset. This addresses the limited sample size in existing cropping datasets while avoiding the computational burden of the full 100,000+ image COCO training set. The COCO-minitrain subset maintains the statistical properties of the complete dataset while offering efficient training. After both pretraining stages complete, we initialize AesCrop’s corresponding components with the learned weights.

Model Configuration: AesCrop predicts $N = 90$ crops per image, aligning with the maximum number of annotations available in the GAICv2 [18] dataset. The architecture employs VMamba-T variant for its compactness and efficiency and $M = 6$ decoder layers, with loss weights set to $\lambda_{\text{focal}} = \lambda_{\text{GIoU}} = 0.4$. The model is optimized using AdamW, which adapts learning rates dynamically for stable training. All experiments run on an NVIDIA RTX 4090 GPU with a batch size of 16. The initial learning rate of 10^{-4} is maintained for 40 epochs, then divided by ten for the final 10 epochs.

4.4. Comparison with Existing Works

Quantitative Comparison: For quantitative evaluation, we employ the $Acc_{K/N}$ and $\overline{Acc}_N$ metrics for $K \in \{1, 2, 3, 4\}$ and $N \in \{5, 10\}$ on the GAICv2 [18] testing set with $\epsilon = 0.90$, similar to Jia *et al.*’s [5] experimental settings. The quantitative results for the existing works were obtained directly from their papers, and the comparisons are summarized in Tab. 1. The results show that the proposed model, AesCrop outperforms state-of-the-arts methods across all evaluated metrics, with a substantial performance gain of 7.4% in $Acc_{1/5}$ and 4.7% in $Acc_{1/10}$. This impressive performance demonstrates the effectiveness of the MCAB module is guiding AesCrop to consider composition salient regions in generating aesthetically appealing crops.

Qualitative Comparison: Next, we performed a qualitative comparison of AesCrop with two publicly available models, GAIC [18] and Jia *et al.* [5], on six test images from the GAIC [18] dataset. For GAIC [18], we generate the crops using the pretrained weights. As Jia *et al.* [5] did not provide pretrained weights, we retrained their model using their original configuration, adjusting the input size to 512×512 to align with AesCrop for efficient training. For each image, we compare the generated crops with the highest score for each model, alongside the top three ground truth crops with the highest Mean Opinion Scores (MOS), all of which have a score of at least 4. This strategy ensures that we only compare with crops of high quality. Figure 4 presents a qualitative comparison of these results.

Table 1. Quantitative Comparison. The best results are marked in **bold** and the second best results are marked with underline.

Model	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$\overline{Acc}_5$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$\overline{Acc}_{10}$
VFN [1]	26.6	26.5	26.7	25.7	26.4	40.6	40.2	40.3	39.3	40.1
VPN [15]	36.0	-	-	-	-	48.5	-	-	-	-
VEN [15]	37.5	35.0	35.3	34.2	35.5	50.5	49.2	48.4	46.4	48.6
GAIC [18]	68.2	64.3	61.3	58.5	63.1	84.4	82.7	80.7	<u>78.7</u>	81.6
TransView [10]	69.0	<u>66.9</u>	61.9	57.8	63.9	85.4	<u>84.1</u>	<u>81.3</u>	78.6	82.4
CLIPCropping [17]	70.0	66.7	<u>63.0</u>	<u>60.0</u>	<u>64.9</u>	<u>87.5</u>	83.7	80.5	78.5	<u>82.5</u>
Jia <i>et al.</i> [5]	<u>72.0</u>	-	-	-	-	86.0	-	-	-	-
AesCrop (Ours)	79.4	75.8	71.5	68.3	73.7	92.2	90.4	87.1	84.8	88.6

The visual comparison reveals that AesCrop can generate more aesthetically pleasing crops compared to the other models in most cases. In the first image, only AesCrop preserved the negative space in front of the children, illustrating a higher correlation with the ground truth crops. Negative space, an often neglected composition rule in automatic cropping approaches, defines the focal point and maintain a sense of balance within the composition. In the second image, AesCrop was also the sole method to remove distracting elements along the left and right borders, effectively directing focus to the Ferris wheel. For the third image, both AesCrop and Jia *et al.*'s [5] method successfully eliminated the black border at the bottom, unlike GAIC [18]. Comparatively, AesCrop yielded a more aesthetically pleasing result with its placement of the snail adhering to the rule of thirds. In the fourth image, AesCrop and GAIC[18] retained the full blue arch which plays a crucial role in framing the scene, while Jia *et al.* [5] cropped it into half, breaking the natural framing effect. The fifth image showed comparable performance across all models. However, in the sixth image, none of the methods managed to include the rightmost section of the building in their crops.

We attribute AesCrop's superior performance to the MCAB module. To further analyze the role of MCAB, we selected two examples from Fig. 4 and visualize the composition attention bias in Fig. 5. From the first example, we can observe that MCAB places heavy emphasis on the negative space beside the children, an important compositional region that other models failed to capture. In the second example, MCAB highlights the blue arch, enabling the model to preserve the crucial image framing element. These attention bias heatmaps show that MCAB consistently guides AesCrop to focus on compositionally salient regions, resulting in higher-quality crops.

Overall, AesCrop demonstrates competitive performance, matching or exceeding existing models in most evaluations. It reliably avoids composition failures present in other models, notably reducing over-cropping of key semantic regions and demonstrating a stronger adherence to

Table 2. Impact of MCAB.

MCAB	$Acc_{1/5}$	$Acc_{1/10}$
✗	76.2%	90.8%
✓	79.4%	92.2%

Table 3. Impact of Object Detection Pretraining.

Pretrained	$Acc_{1/5}$	$Acc_{1/10}$
✗	72.8%	89.6%
✓	79.4%	92.2%

professional photographic composition rules. We attribute this success to two key innovations: (1) VMamba's ability to generate rich, comprehensive image encodings, and (2) the Mamba Composition Attention Bias (MCAB) mechanism, which effectively directs the model's focus toward compositionally significant areas.

4.5. Ablation Study

Importance of Mamba Composition Attention Bias:

The MCAB enhances AesCrop by dynamically modulating attention weights using learned compositional principles. Quantitative analysis in Tab. 2 shows that removing MCAB reduces performance by 3.2% on $Acc_{1/5}$ and 1.4% on $Acc_{1/10}$, verifying its importance. This performance gap demonstrates that explicit composition modeling through MCAB is crucial for generating aesthetically optimal crops.

Importance of Object Detection Pretraining:

AesCrop leverages object detection pretraining to boost feature extraction capabilities. As shown in Tab. 3, this strategy yields significant gains, with the pretrained model outperforms its non-pretrained counterpart by 6.6% on $Acc_{1/5}$ and 2.6% on $Acc_{1/10}$. This improvement reflects the pretrained model's superior object localization skills, which enable it to focus on learning aesthetic importance rather than basic visual recognition during cropping-specific training.

MCAB aggregation technique: The MCAB attention prior is generated as a probability-weighted average of



Figure 4. Visual comparison of top three ground truth crops with predicted crops.

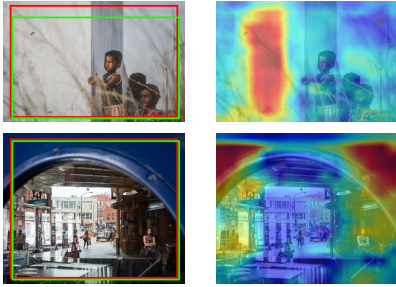


Figure 5. (left) Source image with ground truth (green box), and AesCrop's best predicted crops (red box) and (right) the corresponding composition attention priors generated by MCAB.

class activation maps (CAMs). Quantitative comparisons in Tab. 4 show that this approach outperforms aggregation approach that considers only the highest-probability CAM, demonstrating that probability-weighted method can better preserve compositional diversity. On the other hand, the max-probability alternative discards valuable information and introduces noise. This result reinforces the importance of diversity in photographic composition, whereby multiple composition rules can be applied to enhance the aesthetics appeal of a photograph.

Number of Decoder Layers: We empirically determined the optimal number of decoder layers by evaluating configurations from 1 to 6 layers. As Fig. 6 demonstrates,

Table 4. Impact of MCAB Aggregation Technique.

Aggregation Technique	$Acc_{1/5}$	$Acc_{1/10}$
Max	76.4%	89.0%
Average	79.4%	92.2%

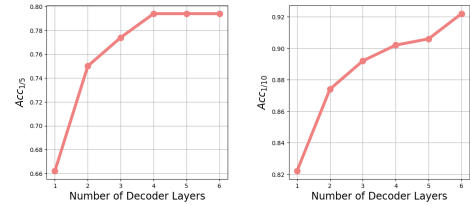


Figure 6. Performance across different decoder depths for $Acc_{1/5}$ (left) and $Acc_{1/10}$ (right).

both accuracy metrics peak at 6 layers, achieving 79.4% for $Acc_{1/5}$ and 92.2% for $Acc_{1/10}$.

5. Limitations

Although AesCrop demonstrates remarkable performance, it is not without limitations. Since AesCrop was trained on the GAIC [18] dataset, where crops are generated by heuristic rules, it inherits the pattern, and thus struggles to achieve true globality.

Moreover, while the MCAB modulates decoder attention according to the composition structure of the whole image,

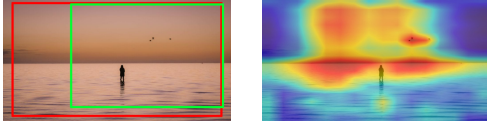


Figure 7. (left) Source image with ground truth (green box), and AesCrop’s best predicted (red box) crops, and (right) AesCrop composition priors generated by MCAB.

the optimal crop may not always align with it. For example, Figure 7 shows a case where the MCAB highlights the central horizontal region, leading AesCrop to prioritize that area, even though the optimal crop does not rely on this particular compositional feature.

6. Conclusion

We introduce AesCrop, a hybrid cropping framework that incorporates compositional knowledge via Mamba Composition Attention Bias. By directing decoder attention toward compositionally salient regions, AesCrop achieves superior alignment with human annotations across all metrics, achieving state-of-the-arts performance. Extensive experiments validate the impact of each architectural component while revealing two key limitations: heuristic-induced biases from training data and occasional over-prioritization of compositional cues. These insights suggest promising directions for future work to enhance global adaptability and further refine aesthetic and compositional integration.

References

- [1] 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 Multimedia 2017*, 2017. 1, 2, 6
- [2] 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
- [3] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7053–7062, 2021. 1, 2
- [4] James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by out-painting professional photos, 2024. 1, 2
- [5] Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2436–2445, 2022. 1, 2, 3, 4, 5, 6, 7
- [6] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *Journal of Visual Communication and Image Representation*, 55:91–105, 2018. 3, 5
- [7] Debang Li, Junge Zhang, and Kaiqi Huang. Learning to learn cropping models for different aspect ratio requirements. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12682–12691, 2020. 2
- [8] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model, 2024. 2
- [9] Depu Meng, Xiaokang Chen, ZeJia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. *CoRR*, abs/2108.06152, 2021. 2, 3
- [10] Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong. Transview: Inside, outside, and across the cropping view boundaries. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4198–4207, 2021. 1, 2, 6
- [11] Zhiyu Pan, Yinpeng Chen, Jiale Zhang, Hao Lu, Zhiguo Cao, and Weicai Zhong. Find beauty in the rare: Contrastive composition feature clustering for nontrivial cropping box regression, 2023. 1, 2
- [12] Nermin Samet, Samet Hicsonmez, and Emre Akbas. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *European Conference on Computer Vision (ECCV)*, 2020. 5
- [13] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 3
- [14] Tengfei Shi, Chenglizhao Chen, Yuanbo He, Wenfeng Song, and Aimin Hao. Joint probability distribution regression for image cropping. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 990–994, 2023. 1
- [15] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 6
- [16] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–978, 2013. 2
- [17] Quan Yuan, Leida Li, and Pengfei Chen. Aesthetic image cropping meets vlp: Enhancing good while reducing bad. *Journal of Visual Communication and Image Representation*, 105:104316, 2024. 1, 2, 6
- [18] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model, 2019. 1, 2, 4, 5, 6, 7
- [19] Zhihang Zhong, Mingxi Cheng, Zhirong Wu, Yuhui Yuan, Yinqiang Zheng, Ji Li, Han Hu, Stephen Lin, Yoichi Sato, and Imari Sato. Clipcrop: Conditioned cropping driven by vision-language model. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 294–304, 2023. 1, 2