ProCrop: Learning Aesthetic Image Cropping from Professional Compositions

Ke Zhang[1], Tianyu Ding[2†], Jiachen Jiang[3], Tianyi Chen[2], Ilya Zharkov[2], Vishal M. Patel[1], Luming Liang[2†] *

[1]Johns Hopkins University
[2]Microsoft
[3]Ohio State University

Abstract

Image cropping is crucial for enhancing the visual appeal and narrative impact of photographs, yet existing rule-based and data-driven approaches often lack diversity or require annotated training data. We introduce ProCrop, a retrieval-based method that leverages professional photography to guide cropping decisions. By fusing features from professional photographs with those of the query image, ProCrop learns from professional compositions, significantly boosting performance. Additionally, we present a large-scale dataset of 242K weakly-annotated images, generated by out-painting professional images and iteratively refining diverse crop proposals. This composition-aware dataset generation offers diverse high-quality crop proposals guided by aesthetic principles and becomes the largest publicly available dataset for image cropping. Extensive experiments show that ProCrop significantly outperforms existing methods in both supervised and weakly-supervised settings. Notably, when trained on the new dataset, our ProCrop surpasses previous weakly-supervised methods and even matches fully supervised approaches.

Code and datasets —
https://bwgzk-keke.github.io/ProCrop/

Introduction

In visual arts, a well-composed photograph can captivate viewers and convey profound messages. Image cropping, the art of selectively removing peripheral areas from a photograph, is crucial for enhancing visual appeal and narrative potency. However, achieving aesthetically pleasing compositions through cropping is challenging due to the intricate interplay of various compositional elements (Liu et al. 2010; Obrador, Schmidt-Hackenberg, and Oliver 2010), especially for non-professionals and automated systems.

Existing automatic image cropping methods typically fall into two categories: those guided by composition rules in photography (Fang et al. 2014; Ni et al. 2013; Zhang et al. 2013) and data-driven approaches such as anchor-based (Li et al. 2020; Lian et al. 2021; Zeng et al. 2019, 2020; Wang et al. 2023) and coordinate regression-based (Guo et al. 2018; Hong et al. 2021; Li, Zhang, and Huang 2020; Liu et al. 2023) methods. Rule-based approaches often struggle to fully capture sophisticated features and complex compositions, being constrained by the very principles they're founded upon. Data-driven methods, while promising, face challenges due to their reliance on annotated datasets for training. Creating large-scale, diverse datasets of aesthetically pleasing compositions is labor-intensive and time-consuming. Currently, the largest available dataset for this task contains only about 10K images (see Table 1), which is insufficient to capture the diversity of compositions found in professional photography.

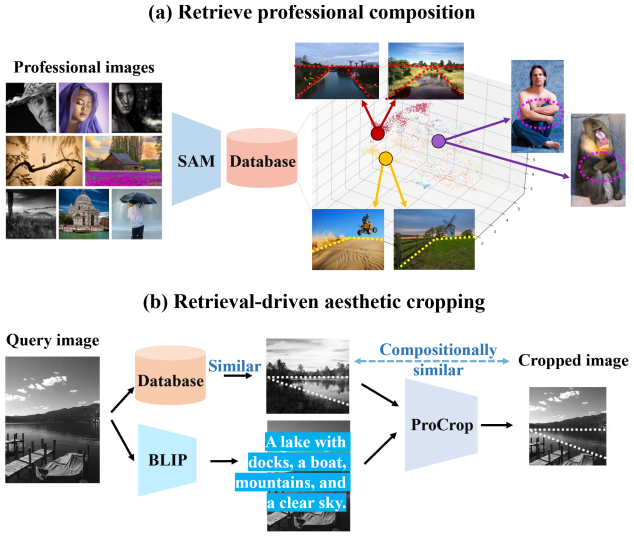


Figure 1: Overview of ProCrop: (a) Retrieves compositionally similar images from a curated database; (b) Uses them to guide aesthetic cropping.

In this paper, we introduce a novel retrieval-based image cropping approach that harnesses the wealth of existing professional photography. Inspired by retrieval augmentation in language models (Borgeaud et al. 2022; Guu et al. 2020) and the abundance of professional photography datasets, we learn from professional images with similar aesthetic compositions (see Figure 1). Our key insight is that professional photographers have already solved numerous compositional challenges through their experience and artistic vision. By

*† Corresponding author.

tapping into this knowledge base, we guide our model to align with professional standards. This approach addresses diversity limitations in rule-based methods while enhancing data-driven methods with external knowledge. Importantly, this requires no annotations for the reference database, ensuring its practicality. We demonstrate that integrating this retrieval-augmented concept into image cropping yields SOTA performance, underscoring its effectiveness.

Furthermore, we address the scarcity of high-quality aesthetic training data by developing a large-scale dataset through a weakly-supervised approach. Specifically, we leverage ControlNet (Zhang, Rao, and Agrawala 2023), a text-to-image diffusion model, to outpaint professional images, simulating cropped and uncropped pairs. Starting with AVA (Murray, Marchesotti, and Perronnin 2012) and unsplash-lite (Unsplash 2023), the large collection of professional images serving as expert labels (*i.e.*, good crops), we employ GPT-4 (Achiam et al. 2023) to infer textual layouts beyond original image boundaries and use SAM (Kirillov et al. 2023) to extract multi-scale compositional masks. These are then fed into ControlNet for image outpainting. Through an iterative refinement process, we generate diverse crop proposals, substantially expanding the available data. The resulting dataset comprises 242K annotated aesthetic images, significantly surpassing existing resources in scale and diversity (see Table 1). Our weakly supervised method trained on this dataset outperforms previous weakly supervised methods and achieves results comparable to fully supervised ones.

Our contributions are summarized as follows: (1) We propose ProCrop, a retrieval-based image cropping method that leverages professional photography knowledge to achieve aesthetically pleasing compositions. (2) We introduce a new dataset through a weakly-supervised, controlled approach. To the best of our knowledge, this is the largest dataset for aesthetic image cropping. The code and dataset will be made publicly available to facilitate future research. (3) Experiments show that our retrieval-based method significantly outperforms existing works. Notably, trained on our new dataset, it surpasses prior weakly-supervised methods and even matches fully supervised approaches.

Related Work

Aesthetic image cropping Aesthetics image cropping aims to enhance the visual appeal of images by learning aesthetic composition via comparative views. Image cropping methods can be broadly categorized into rule-based and data-driven approaches. Rule-based methods (Hong et al. 2021; Fang et al. 2014; Ni et al. 2013; Zhang et al. 2013) rely on handcrafted features and techniques like saliency detection (Vig, Dorr, and Cox 2014) or specific aesthetic rules (Liu et al. 2010; Nishiyama et al. 2009; Zhang et al. 2005). While effective at content preservation, they often struggle with nuanced compositions. Data-driven approaches, which now dominate the field, include anchor-based methods (Li et al. 2020; Chen et al. 2017c; Tu et al. 2020; Zeng et al. 2019; Wei et al. 2018; Wang and Shen 2017; Lian et al. 2021; Zeng et al. 2020) that evaluate candidate regions, and coordinate regression methods (Chen et al. 2017c; Li et al. 2018; Guo et al. 2018; Hong et al. 2021; Li, Zhang, and Huang 2020; Liu et al. 2023) that directly predict crop boundaries. A crucial aspect of data-driven methods is their dependence on large-scale supervised training. Widely used datasets such as GAICv1 (Zeng et al. 2019), GAICv2 (Zeng et al. 2020), CPC (Wei et al. 2018), FCDB (Chen et al. 2017a), and SACD (Yang et al. 2023) are labor-intensive and expensive to create. Recently, (Hong et al. 2024) attempted to address these issues by outpainting professional images. However, their approach is constrained to single crop suggestions, faces reliability issues with outpainted content, and is not publicly accessible. In this work, we present a large-scale dataset of weakly-annotated images, generated by out-painting professional images and iteratively refining diverse crop proposals. Our composition-aware approach yields high-quality and diverse crop proposals.

Datasets	Year	Venue	# of Images	# of Annotations	
				Avg	Total
ICDB (Yan et al. 2013)	2013	CVPR	1,000	1	1000
FLMS (Fang et al. 2014)	2014	ACM	500	10	5000
FCDB (Chen et al. 2017b)	2017	WACV	1,743	1	1743
CPC (Wei et al. 2018)	2018	CVPR	10,797	24	259,128
GAICv1 (Zeng et al. 2019)	2019	CVPR	1,036	90	93,240
GAICv2 (Zeng et al. 2020)	2020	TPAMI	2,626	90	236,340
SACD (Yang et al. 2023)	2023	CVM	2,777	8	22,216
UGCrop5K (Ko, Jin, and Kim 2024)	2024	AAAI	5,000	90	450,000
Ours	2026	AAAI	242,000	8	1,936,000

Table 1: Summary of datasets for image cropping.

Retrieval augmentation. Retrieval augmentation (Asai et al. 2023; Blattmann et al. 2022; Borgeaud et al. 2022; Guu et al. 2020; Qin et al. 2022) provides an effective approach to improve model performance without expanding model parameters or requiring additional training data. Instead of storing all knowledge within model parameters, these techniques utilize external database to fetch relevant information on demand. A typical method is to fetch k-nearest neighbors from a pre-computed embedding space to provide supplementary input. This strategy has demonstrated success across various domains, including language models (Guu et al. 2020), diffusion models (Qin et al. 2022), and layout generation (Horita et al. 2024).

Method

Overview Image cropping aims to enhance the composition of photographs that may not have been captured professionally. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, our goal is to predict a series of crop rectangles with high aesthetic scores, denoted as $\{(\boldsymbol{b}_n, s_n)\}_{n=1}^{N}$, where $\boldsymbol{b}_n \in [0,1]^4$ is the bounding box in normalized coordinates, s_n is the aesthetic quality, and N is the number of predicted crops. We use the image as input and its ground truth crops or pseudo labels for supervision. This task presents significant challenges due to the intricate interplay of various compositional elements, such as subject positioning, adherence to the rule of thirds, and the use of leading lines. Inspired by how retrieval augmentation has improved the quality of language models and image synthesis, we propose a novel module for retrieval-based aesthetic image cropping. Our approach learns from professional compositions without requiring additional annotations on the retrieval database, significantly improving

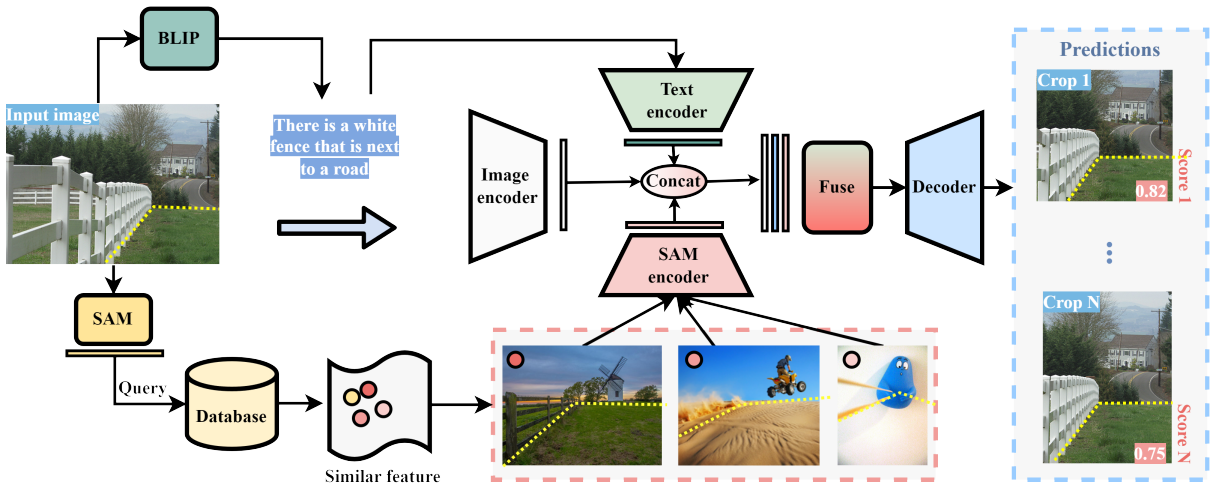


Figure 2: The pipeline of ProCrop. Given an input image, ProCrop retrieves compositionally similar professional images and generates a textual description, which guide the model to produce aesthetically enhanced crops along with aesthetic scores.

the quality of generated compositions. Furthermore, we introduce a composition-aware approach to generate a large dataset. This method offers multiple high-quality crop proposals guided by aesthetic principles, enhancing the learning process and ultimate performance.

ProCrop: Retrieval-driven aesthetic cropping

To effectively leverage professional images, we introduce a retrieval module that addresses two key challenges: (1) retrieving reference images from a database based on their compositional features, and (2) fusing the retrieved features into a final augmented representation. Our approach is inspired by the assumption that compositional features can be characterized by line combinations in images (Lee et al. 2018; Ko, Jin, and Kim 2024). To capture these line compositions in retrieved images, we employ SAM (Kirillov et al. 2023), which offers richer, more precise boundaries without relying on direct semantic mask extraction, compared to CLIP (Radford et al. 2021) or saliency map (Hoh, Zhang, and Dodgson 2023; Horita et al. 2024). More details in Appendix D.

Feature retrieval. Let $\mathcal{V}$ represent the database of professional images. For an input image I, we aim to identify professional images with the most similar compositional characteristics. We use the SAM encoder to extract features from both the query image I and each image in the professional database $\mathcal{V}$, yielding f_I and $\boldsymbol{F} = \{f_{\widetilde{I}} \mid \widetilde{I} \in \mathcal{V}\}$, respectively. Based on feature similarity, we retrieve the top-K most similar compositional features in $\boldsymbol{F}$, represented by $\boldsymbol{R} \in \mathbb{R}^{K \times m \times d}$, where m is the flattened spatial dimension and d is the feature dimension. To streamline the training process, we precompute and cache the SAM feature embeddings for each training image along with their K most similar counterparts from $\mathcal{V}$. These image-embedding pairs are indexed in a database, with ElasticSearch (Huggingface 2024) serving as the retrieval engine for efficient similarity-based matching. In contrast to existing retrieval methods (Fu et al. 2024; Horita et al. 2024) that primarily emphasize category similarity, our method is specifically designed to capture compositional characteristics. By using SAM embeddings and our streamlined retrieval pipeline, we effectively learn compositional knowledge from professional photographs.

Feature fusion. Given the retrieved top-K image features $\boldsymbol{R}$ from $\mathcal{V}$, we fuse them with the query image's embedded features to guide the cropping process. While directly utilizing the SAM embedding f_I is feasible, SAM's computational overhead leads to slow training and inference. Instead, we adopt an encoder architecture similar to Conditional DETR (cDETR) (Meng et al. 2021), which offers superior training convergence and inference efficiency while maintaining comparable performance. We denote this query image feature as $\bar{f}_I \in \mathbb{R}^{p \times d}$, where p represents the flattened spatial dimension specific to this encoder. To effectively fuse $\bar{f}_I$ with $\boldsymbol{R}$, we employ a learnable projection head $\Pi(\cdot)$ that transforms $\boldsymbol{R}$ to match the spatial-channel dimensions of $\bar{f}_I$. The final feature fusion is achieved through: $f_R = \text{Concat}(\bar{f}_I, \Pi(\boldsymbol{R}), f_c)$, where f_c denotes the cross-attended feature obtained by using $\bar{f}_I$ as the query and $\Pi(\boldsymbol{R})$ as both key and value. The resulting fused feature f_R is subsequently fed into the rest of the pipeline, incorporating compositional knowledge retrieved from professional photography.

Motivated by the natural ability of language to highlight salient image regions, we enhance the model by integrating multi-modal features with the fused image features. For an input image I, we first employ BLIP (Li et al. 2022) to generate compositional text descriptions that explicitly capture the desired objects and their spatial arrangements. We then leverage BLIP to extract multi-modal embeddings $\boldsymbol{M} \in \mathbb{R}^{m' \times d}$ from these image-text pairs, where m' is the flattened spatial dimension specific to the BLIP encoder. We precompute this process for all training images. The multi-modal feature fusion is then computed as $f_M = \text{Concat}(\bar{f}_I, \Pi'(\boldsymbol{M}), f'_c)$, where $\Pi'(\cdot)$ denotes a learnable projection head for harmonizing the feature dimensions, and f'_c represents the cross-attended feature derived by utilizing $\bar{f}_I$ as the query and

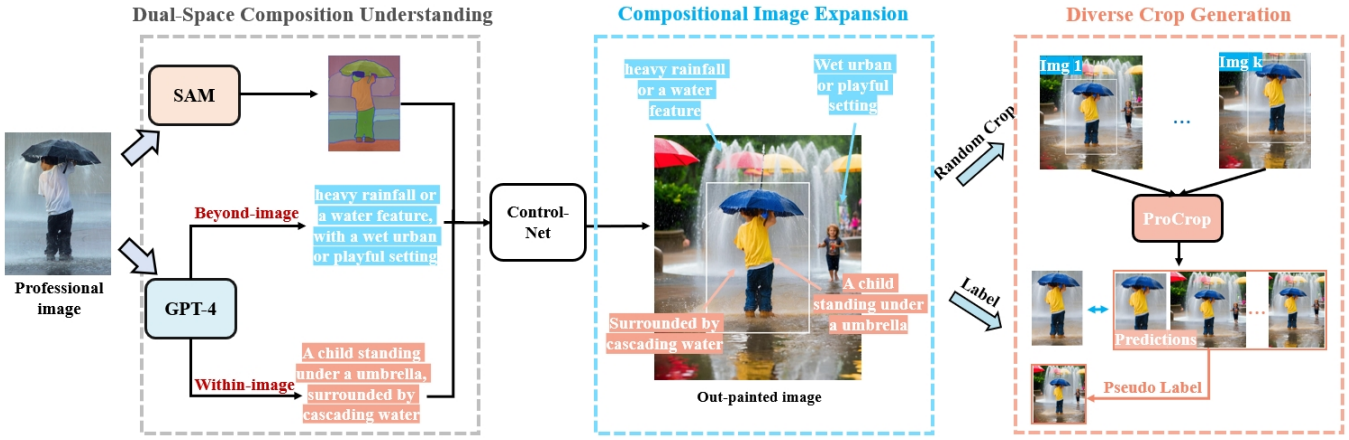


Figure 3: Composition-aware dataset generation. Professional images undergo three stages to create diverse image-crop pairs.

$\Pi'(\boldsymbol{M})$ as both key and value. Details on text embeddings are provided in Appendix C of the Supplementary Material.

We concatenate the multi-modal feature f_M with the retrieved feature f_R and pass the fused feature into a decoder, where $\bar{f}_I$ is injected once during the fusion step. Following (Jia et al. 2022), our decoder processes both the input features and learnable anchors through parallel regression and classification heads. This architecture generates N crop proposals, each accompanied by its aesthetic score, which can be expressed as: $\text{Decoder}(f_R, f_M) \mapsto \{(\boldsymbol{b}_n, s_n)\}_{n=1}^{N}$.

Composition-aware dataset generation

Data-driven image cropping rely on annotated datasets for training. However, high-quality datasets containing images and their aesthetic crops are scarce due to the labor-intensive nature. To address this, we develop an automated pipeline for generating large-scale cropping datasets in a weakly-supervised manner, as shown in Figure 3. Our dataset encompasses diverse image categories, professional crop proposals, and compositional descriptions. The pipeline leverages quality-validated professional photographs from public sources. We employ language and segmentation foundation models to encode compositions both within and beyond image boundaries. These are then fed into a text-to-image model to generate outpainted images, simulating uncropped and cropped image pairs. More details in Appendix B.1 .

Dual-space composition understanding. For *within-image* compositions, we prompt GPT-4 to analyze compositional elements and identify salient subjects that attract human attention. We incorporate SAM-generated segmentation masks to ensure semantic consistency between input and generated content. For *beyond-image* compositions, GPT-4 predicts potential content outside image boundaries and describes the broader context. Our experiments show that these beyond-image compositional descriptions are essential for effective outpainting, as shown in Figure 4. While (Hong et al. 2024) proposes an outpainting approach for weakly annotated data, their reliance on image captions leads to artifacts like extraneous objects or unnatural grid patterns. In contrast, our composition-aware prompting strategy generates more coherent and visually plausible results.

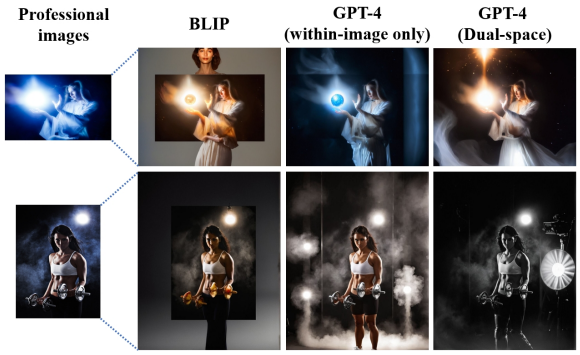


Figure 4: Comparison of out-painting shows our dual-space GPT-4 approach yields the most realistic results.

Compositional image expansion. We randomly downscale the professional image and enlarge it to create a canvas with dimensions between 700 and 1024 pixels. The outpainting process feeds the canvas I_c, GPT-4 generated text descriptions T, and multi-scale SAM masks S into a pretrained ControlNet (Zhang, Rao, and Agrawala 2023) to produce the output $I' = \text{ControlNet}(I_c, S, T)$. **Diverse crop generation.** Instead of the single crop proposal naturally arising from the original and outpainted images, we develop an iterative refinement process that creates high-quality, varied crop proposals (see Figure 5) through a model-in-the-loop approach. We generate random crops from expanded images, ensuring the preservation of original content while varying in size and aspect ratio. These random crops serve as initial training inputs, with their corresponding original image regions acting as labels. We train a ProCrop model using these image-crop pairs. The model then enters an iterative cycle where it automatically generates crop proposals for each query image. These proposals undergo a curation process that selects a diverse set adhering to established aesthetic principles. During this iterative refinement process, we dynamically rank the aesthetic scores of the crop set. The top-k crops are then utilized as pseudo labels, significantly enhancing the diversity

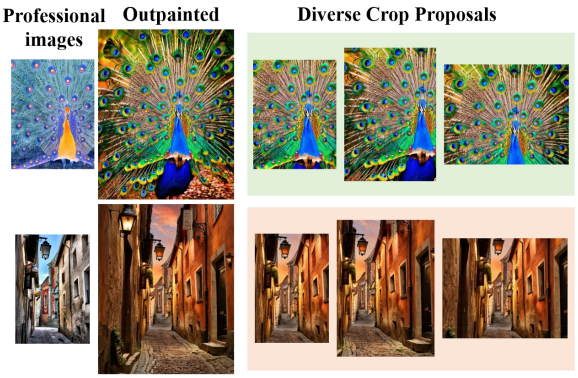


Figure 5: Outpainting examples and multiple crop proposals, which are pseudo-labels generated via model-in-the-loop.

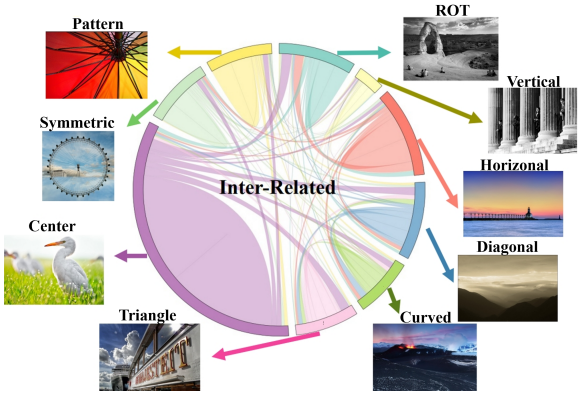


Figure 6: Distribution of compositional layouts across CAD.

Composition	Original		Out-painted		Total
	AVA	UnSplash	AVA	UnSplash	
RoT	5376	2203	15050	5652	20702
Vertical	1897	707	6023	1990	8013
Horizontal	4322	5755	19706	8428	28134
Diagonal	5339	2455	15023	487	15510
Curved	2998	1330	10661	4447	15108
Triangle	5147	2646	11816	3172	14988
Center	19665	8066	80150	13162	93312
Symmetric	1669	1360	16105	5562	21667
Pattern	3182	449	17588	2658	20246
Total	49595	24971	192122	49942	**242064**

Table 2: Distribution of composition categories in CAD.

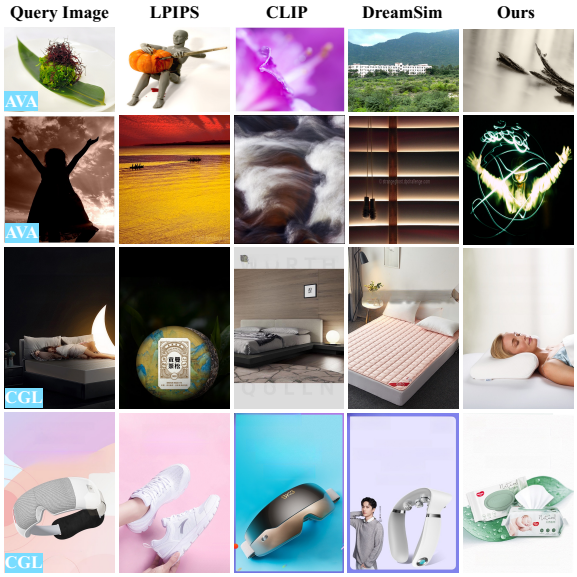


Figure 7: Our retrieval on CGL and AVA by leveraging line composition for more compositionally relevant matches..

of our crop annotations and ultimately improving the model's ability to generalize across various cropping scenarios.

Experiments

Setup

(1) Retrieval datasets. We employ two datasets for image retrieval: CGL (Zhou et al. 2022) and AVA (Murray, Marchesotti, and Perronnin 2012). CGL consists of 60,548 e-commerce posters, primarily featuring cosmetics and clothing advertisements with relatively simple compositional layouts. On the other hand, AVA is a significantly larger dataset containing 255,000 images with more complex scenarios and diverse compositional arrangements. From AVA, we select the top 55,000 images based on aesthetic scores to form professional retrieval set. More details in Appendix B.
(2) Cropping datasets. We utilize five datasets for image cropping: GAICv1 (Zeng et al. 2019), GAICv2 (Zeng et al. 2020), CPC (Wei et al. 2018), FLMS (Fang et al. 2014), and SACD (Yang et al. 2023). GAIC and CPC serve as small and mid-sized training datasets, respectively, while SACD and FLMS are used for evaluation in zero-shot transfer experiments. The GAICv1 dataset contains 1,036 training and 200 testing images, with each image offering up to 90 crop proposals generated using a predefined grid-anchor system. GAICv2 is an extended version, consisting of 2,636 training images, 200 validation images, and 500 testing images. These proposals are rated on a 1-5 scale by six annotators through a two-stage process and organized into four aspect ratio groups, each containing six crops. The CPC dataset is a larger collection of 10,797 images, serving as a mid-sized benchmark for training supervised image cropping models. The FLMS dataset consists of 500 images, each accompanied by up to 10 high-quality crop proposals, and is exclusively used for testing purposes. Following (Hong et al. 2024), we utilize the test set of SACD for evaluation, which provides six to eight annotated cropping windows per image, focusing on aesthetic quality to ensure well-composed subjects. **Curation of CAD dataset:** We source professional images from AVA (Murray, Marchesotti, and Perronnin 2012) and Unsplash Lite (Unsplash 2023). From AVA, we select the top 55,000 images based on their aesthetic scores. The Unsplash Lite dataset contributes 25,000 high-quality, nature-themed photographs, which are available for both commercial and non-commercial use. Using these 80,000 curated professional images as a foundation, we generate 242,000 synthetic images that meet our quality standards through automatic filtering (Hong et al.

2024). The distribution of compositional layouts in CAD is shown in Figure 6 and table 2. **(3)Implementation details:** Following cDETR (Meng et al. 2021) and recent works (Jia et al. 2022; Hong et al. 2024), we optimize our model using AdamW optimizer with a weight decay of 10^{-4}. The learning rate is set to 10^{-4}, with a reduced rate of 10^{-5} for the CNN backbone. The model trains for 500 epochs. In the weakly-supervised setting with our curated CAD, we divide training into two stages: Stage 1 (first 100 epochs) initializes the model weights, while Stage 2 (remaining 400 epochs) involves crop prediction and dynamic ranking to generate diverse pseudo-labels. **(4) Evaluation Metrics.** We adopt three evaluation metrics, including Intersection-over-Union (IoU), boundary displacement (Disp), and top-N accuracy ($\text{ACC}_{K/N}$), following (Hong et al. 2024; Yang et al. 2023; Zhang et al. 2022). IoU and Disp provide objective and consistent comparisons, while $\text{ACC}_{K/N}$ reflects human perception. Specifically, for $\text{ACC}_{K/N}$, we define the best crops of an image as those ranked within the top-N by mean opinion scores (MOS) from human ratings. $\text{ACC}_{K/N}$ then measures how many of the top-K predicted crops fall within this top-N MOS set. This makes $\text{ACC}_{K/N}$ highly correlated with user study results. Following (Su et al. 2024; Wang et al. 2023), we report the average top-k accuracy ($\overline{\text{ACC}_k}$) for $k = 5$ and $k = 10$. When predicted views do not align exactly with predefined grid views, we consider two crops equivalent if IoU exceeds a threshold of $\epsilon = 0.85$, as in (Liu et al. 2023).

Comparative assessment

We first conduct comparative analysis on retrieval approaches and then evaluate our ProCrop model performance in both supervised and weakly-supervised settings.

Retrieval approaches analysis. We compare our SAM-based retrieval against SOTA embeddings (DreamSim (Fu et al. 2024), OpenCLIP (Cherti et al. 2023)) and established learned metrics like LPIPS (Zhang et al. 2018). Our evaluation uses examples from CGL (Zhou et al. 2022) and AVA (Murray, Marchesotti, and Perronnin 2012) datasets, where for each query image, we compute similarities across the dataset and retrieve the nearest neighbors based on each metric. As shown in Figure 7, existing methods either focus on fine-grained visual features (LPIPS emphasizing background color) or broader semantic attributes (DreamSim and OpenCLIP focusing on object categories). In contrast, our SAM-based retrieval uniquely excels at identifying compositional similarities across diverse visual styles, demonstrating effective generalization without relying on category.

Evaluation under supervised setting. We evaluate our model against various baselines trained on GAICv1 (Zeng et al. 2019), GAICv2 (Zeng et al. 2019), and CPC (Wei et al. 2018). For models trained on GAICv1 and GAICv2, we evaluate using $\text{ACC}_{1/5}$ and $\text{ACC}_{1/10}$ on their respective test sets. For models trained on CPC, we measure IoU on the FLMS dataset. To ensure fair comparison, we exclude text embeddings from feature fusion. A key feature of our approach is the integration of the retrieval module, which fetches 10 similar images from the top-rated 55,000 images in AVA during both training and inference. Table 3 shows that our method significantly outperforms previous approaches across all datasets and metrics, demonstrating the effectiveness of guidance from retrieved professional image compositions.

Methods	GAICv2			
	$\text{ACC}_{1/5}(\uparrow)$	$\overline{\text{ACC}_5}(\uparrow)$	$\text{ACC}_{1/10}(\uparrow)$	$\overline{\text{ACC}_{10}}(\uparrow)$
A2-RL (Li et al. 2018)	23.2	26.4	39.5	40.1
VFN (Chen et al. 2017c)	26.6	26.4	40.6	40.1
VEN (Wei et al. 2018)	37.5	50.5	35.5	48.6
CGS (Li et al. 2020)	63.0	59.7	81.5	77.8
GAICv2 (Zeng et al. 2020)	68.2	63.1	84.4	81.6
TransView (Pan et al. 2021)	69.0	63.9	85.4	82.4
HCIC (Zhang et al. 2022)	-	63.8	-	81.3
Jia *et al* (Jia et al. 2022)	85.0	-	92.6	-
Chao *et al* (Wang et al. 2023)	70.0	64.8	86.8	83.3
S^2CNet (Su et al. 2024)	-	64.0	-	82.7
Ours ($\epsilon = 0.85$)	**85.4**	**81.8**	**94.2**	**91.2**

Methods	GAICv1		FLMS	
	$\text{ACC}_{1/5}(\uparrow)$	$\text{ACC}_{1/10}(\uparrow)$	IOU (↑)	Disp(↓)
A2-RL (Li et al. 2018)	23.0	38.5	0.821	0.045
VFN (Chen et al. 2017c)	27.0	39.0	0.577	0.124
VPN (Wei et al. 2018)	40.0	49.5	0.835	-
VEN (Wei et al. 2018)	40.5	54.0	0.837	0.041
CGS (Li et al. 2020)	63.0	81.5	0.836	0.039
GAICv1 (Zeng et al. 2019)	53.5	71.5	-	-
ASM-Net (Tu et al. 2020)	54.3	71.5	-	-
Jia *et al* (Jia et al. 2022)	81.5	91.0	0.838	0.037
UNIC (Liu et al. 2023)	-	-	0.840	0.037
Ours($\epsilon = 0.85$)	**86.0**	**94.5**	**0.843**	**0.036**

Table 3: Comparison under supervised setting. We compute our metrics and report comparative results based on (Liu et al. 2023; Jia et al. 2022; Wang et al. 2023; Su et al. 2024).

.

Methods	Trained on	WS	IOU	Disp
LVRN (Lu et al. 2019)	CPC	×	0.6962	0.0765
GAIC (Zeng et al. 2020)	GAICD	×	0.7124	0.0696
CACNet (Li et al. 2020)	FCDB,KUPCP	×	0.7109	0.0716
HCIC (Zhang et al. 2022)	GAICD	×	0.7120	0.0683
HCIC (Zhang et al. 2022)	CPC	×	0.7109	0.0712
VPN (Wei et al. 2018)	CPC+AADB	×	0.7164	0.0663
VPN (Wei et al. 2018)	Flickr	✓	0.6690	0.0887
VPN (Wei et al. 2018)	Unsplash	✓	0.6555	0.0775
Gencrop (Hong et al. 2024)	Unspash	✓	0.7301	0.0632
Ours (w/o rtr.)	CAD	✓	0.7035	0.0722
Ours (N=1)	CAD	✓	**0.7303**	**0.0610**
Ours (N=2)	CAD	✓	0.7546	0.0541
Ours (N=3)	CAD	✓	0.7678	0.0506

Table 4: Comparison with supervised and WS baselines on SACD (from (Hong et al. 2024)); N = number of crops.

Evaluation under weakly-supervised (WS) setting. We evaluate ProCrop, trained on our large-scale CAD dataset, on the unseen subject-aware SACD dataset (zero-shot transfer). Table 4 compares our approach with previous subject-aware methods on supervised and WS benchmarks. Unlike prior methods using sliding-window ensembles that are later combined into a single output, ProCrop generates diverse, aesthetic crops in a single pass. With 90 predicted crops, our highest-scoring crop outperforms ensemble outputs of existing methods (e.g., GAIC, CACNet, Gencrop) in both IOU and Disp metrics. Our method further excels in generating multiple effective crop candidates. Notably, our full model with the retrieval module significantly outperforms the variant without retrieval, highlighting the effectiveness of retrieval guidance in this WS scenario. We visually compare crops produced by our method with those generated by existing approaches. Notably, our predicted crops effectively capture the salient subject while enhancing the aesthetic quality.

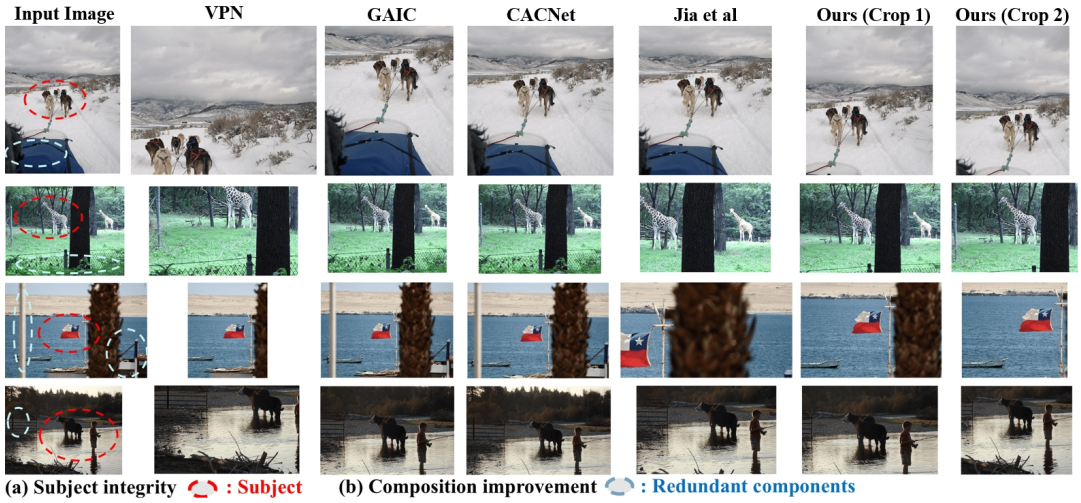
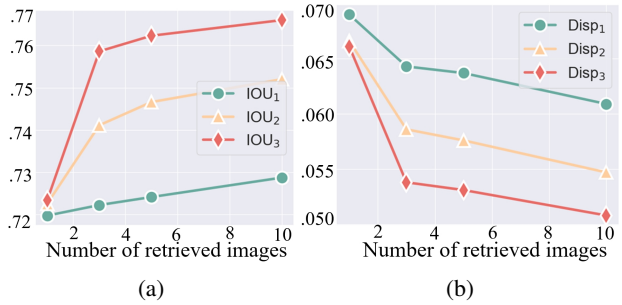


Figure 8: Qualitative comparison: Our method keeps main subjects (red) and removes distractions (blue) for better aesthetics.



(a) (b)

Figure 9: Impact of retrieval count: IOU and Disp improve with more retrieved images, evaluated on top-i crops.

Retrieval				GAICv1	
Retrieve set	Set size	Image	Annotation	ACC_5	ACC_{10}
-	-	-	-	0.815	0.910
GAICv1	1000	✓	×	0.805	0.920
GAICv1	1000	×	✓	0.820	0.920
GAICv1	1000	✓	✓	0.834	0.915
CPC	10000	✓	✓	0.840	0.940
AVA	55000	✓	×	**0.860**	**0.945**

Table 5: Results across different retrieval sources.

Ablation study

We present ablations on retrieval sources, number of retrieved images, and our model components. Further ablations on retrieval encoder, feature alignment, crop number, efficiency, transferability are provided in Appendix A. **(1) Retrieval from different datasets:** Table 5 shows five ablations on GAICv1. Retrieving only GAICv1 images matched the baseline, but using retrieved GAICv1 image-label pairs improved results. Adding CPC images for retrieval boosted performance, and including professional AVA data gave the best results, highlighting the value of diverse, high-quality retrievals for ProCrop. **(2) Impact of retrieval image count.** Figure 9 illustrates how the number of retrieved images affects model performance. Models trained on our CAD dataset and evaluated on the SACD dataset show similar values for IOU_1, IOU_2, and IOU_3, when only one image is retrieved. As the retrieval count increases, greater diversity in crop compositions leads to significant improvements in both IOU and Disp metrics. Performance stabilizes around ten retrieved images, benefiting from diverse layout information. **(3) Components of ProCrop.** Table 6 evaluates the effectiveness of ProCrop components in the weakly-supervised setting, focusing on image retrieval and text embeddings. Results show that incorporating image retrieval leads to notable improvements in average IoU (0.7489 vs. 0.7103) and Disp (0.0555 vs. 0.0667) metrics. The addition of text embeddings further enhances performance, demonstrating the effectiveness of our proposed strategies.

Retrieve	Text	Metric	N=1	N=2	N=3	Avg
×	×	IOU (↑)	0.7035	0.7114	0.7160	0.7103
✓	×		0.7287	0.7520	0.7660	0.7489
✓	✓		**0.7303**	**0.7546**	**0.7678**	**0.7509**
×	×	Disp (↓)	0.0722	0.0647	0.0632	0.0667
✓	×		**0.0609**	0.0547	0.0508	0.0555
✓	✓		0.0610	**0.0541**	**0.0506**	**0.0552**

Table 6: Ablations of ProCrop components (N = crop proposals) under weak supervision on SACD dataset.

Conclusion

This work presents a novel composition-aware cropping framework that leverages professional images with similar aesthetic compositions. Our key contributions include a retrieval-based approach integrating features from professional images with query image embeddings, along with a large-scale compositional-aware cropping dataset. Through comprehensive evaluation across image retrieval, supervised, and weakly-supervised image cropping tasks, our results demonstrate SOTA, showcasing robust applicability across various benchmarks.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Asai, A.; Min, S.; Zhong, Z.; and Chen, D. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, 41–46.

Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; and Ommer, B. 2022. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35: 15309–15324.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Chen, Y.-L.; Huang, T.-W.; Chang, K.-H.; Tsai, Y.-C.; Chen, H.-T.; and Chen, B.-Y. 2017a. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *2017 IEEE winter conference on applications of computer vision (WACV)*, 226–234. IEEE.

Chen, Y.-L.; Huang, T.-W.; Chang, K.-H.; Tsai, Y.-C.; Chen, H.-T.; and Chen, B.-Y. 2017b. Quantitative Analysis of Automatic Image Cropping Algorithms:A Dataset and Comparative Study. In *IEEE WACV 2017*.

Chen, Y.-L.; Klopp, J.; Sun, M.; Chien, S.-Y.; and Ma, K.-L. 2017c. Learning to compose with professional photographs on the web. In *Proceedings of the 25th ACM international conference on Multimedia*, 37–45.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Fang, C.; Lin, Z.; Mech, R.; and Shen, X. 2014. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, 1105–1108.

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2024. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. *Advances in Neural Information Processing Systems*, 36.

Guo, G.; Wang, H.; Shen, C.; Yan, Y.; and Liao, H.-Y. M. 2018. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Transactions on Multimedia*, 20(8): 2073–2085.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Hoh, W. K.; Zhang, F.-L.; and Dodgson, N. A. 2023. Salient-centeredness and saliency size in computational aesthetics. *ACM Transactions on Applied Perception*, 20(2): 1–23.

Hong, C.; Du, S.; Xian, K.; Lu, H.; Cao, Z.; and Zhong, W. 2021. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7057–7066.

Hong, J.; Yuan, L.; Gharbi, M.; Fisher, M.; and Fatahalian, K. 2024. Learning Subject-Aware Cropping by Outpainting Professional Photos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2175–2183.

Horita, D.; Inoue, N.; Kikuchi, K.; Yamaguchi, K.; and Aizawa, K. 2024. Retrieval-Augmented Layout Transformer for Content-Aware Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 67–76.

Huggingface. 2024. ElasticSearch. https://www.elastic.co/cn/elasticsearch. Accessed: 2024-11-14.

Jia, G.; Huang, H.; Fu, C.; and He, R. 2022. Rethinking image cropping: Exploring diverse compositions from global views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2446–2455.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Ko, J.; Jin, D.; and Kim, C.-S. 2024. Semantic Line Combination Detector. *arXiv preprint arXiv:2404.18399*.

Lee, J.-T.; Kim, H.-U.; Lee, C.; and Kim, C.-S. 2018. Photographic composition classification and dominant geometric element detection for outdoor scenes. *Journal of Visual Communication and Image Representation*, 55: 91–105.

Li, D.; Wu, H.; Zhang, J.; and Huang, K. 2018. A2-RL: Aesthetics aware reinforcement learning for image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8193–8201.

Li, D.; Zhang, J.; and Huang, K. 2020. Learning to learn cropping models for different aspect ratio requirements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12685–12694.

Li, D.; Zhang, J.; Huang, K.; and Yang, M.-H. 2020. Composing good shots by exploiting mutual relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4213–4222.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lian, T.; Cao, Z.; Xian, K.; Pan, Z.; and Zhong, W. 2021. Context-aware candidates for image cropping. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1479–1483. IEEE.

Liu, L.; Chen, R.; Wolf, L.; and Cohen-Or, D. 2010. Optimizing photo composition. In *Computer graphics forum*, volume 29, 469–478. Wiley Online Library.

Liu, X.; Liu, M.; Li, J.; Liu, S.; Wang, X.; Lei, L.; and Zuo, W. 2023. Beyond Image Borders: Learning Feature Extrapolation for Unbounded Image Composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13023–13032.

Lu, W.; Xing, X.; Cai, B.; and Xu, X. 2019. Listwise view ranking for image cropping. *IEEE Access*, 7: 91904–91911.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3651–3660.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415.

Ni, B.; Xu, M.; Cheng, B.; Wang, M.; Yan, S.; and Tian, Q. 2013. Learning to photograph: A compositional perspective. *IEEE Transactions on Multimedia*, 15(5): 1138–1151.

Nishiyama, M.; Okabe, T.; Sato, Y.; and Sato, I. 2009. Sensation-based photo cropping. In *Proceedings of the 17th ACM international conference on Multimedia*, 669–672.

Obrador, P.; Schmidt-Hackenberg, L.; and Oliver, N. 2010. The role of image composition in image aesthetics. In *2010 IEEE International Conference on Image Processing*, 3185–3188. IEEE.

Pan, Z.; Cao, Z.; Wang, K.; Lu, H.; and Zhong, W. 2021. Transview: Inside, outside, and across the cropping view boundaries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4218–4227.

Qin, X.; Dai, H.; Hu, X.; Fan, D.-P.; Shao, L.; and Van Gool, L. 2022. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, 38–56. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Su, Y.; Cao, Y.; Deng, J.; Rao, F.; and Wu, Q. 2024. Spatial-Semantic Collaborative Cropping for User Generated Content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4988–4997.

Tu, Y.; Niu, L.; Zhao, W.; Cheng, D.; and Zhang, L. 2020. Image cropping with composition and saliency aware aesthetic score map. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12104–12111.

Unsplash. 2023. Unsplash-lite Dataset. https://unsplash.com/data. Accessed: 2023-12-15.

Vig, E.; Dorr, M.; and Cox, D. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2798–2805.

Wang, C.; Niu, L.; Zhang, B.; and Zhang, L. 2023. Image Cropping With Spatial-Aware Feature and Rank Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10052–10061.

Wang, W.; and Shen, J. 2017. Deep cropping via attention box prediction and aesthetics assessment. In *Proceedings of the IEEE international conference on computer vision*, 2186–2194.

Wei, Z.; Zhang, J.; Shen, X.; Lin, Z.; Mech, R.; Hoai, M.; and Samaras, D. 2018. Good view hunting: Learning photo composition from dense view pairs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5437–5446.

Yan, J.; Lin, S.; Bing Kang, S.; and Tang, X. 2013. Learning the change for automatic image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 971–978.

Yang, G.-Y.; Zhou, W.-Y.; Cai, Y.; Zhang, S.-H.; and Zhang, F.-L. 2023. Focusing on your subject: Deep subject-aware image composition recommendation networks. *Computational Visual Media*, 9(1): 87–107.

Zeng, H.; Li, L.; Cao, Z.; and Zhang, L. 2019. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5949–5957.

Zeng, H.; Li, L.; Cao, Z.; and Zhang, L. 2020. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1304–1319.

Zhang, B.; Niu, L.; Zhao, X.; and Zhang, L. 2022. Human-centric image cropping with partition-aware and content-preserving features. In *European Conference on Computer Vision*, 181–197. Springer.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, L.; Song, M.; Yang, Y.; Zhao, Q.; Zhao, C.; and Sebe, N. 2013. Weakly supervised photo cropping. *IEEE Transactions on Multimedia*, 16(1): 94–107.

Zhang, M.; Zhang, L.; Sun, Y.; Feng, L.; and Ma, W. 2005. Auto cropping for digital photographs. In *2005 IEEE international conference on multimedia and expo*, 4–pp. IEEE.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhou, M.; Xu, C.; Ma, Y.; Ge, T.; Jiang, Y.; and Xu, W. 2022. Composition-aware graphic layout GAN for visual-textual presentation designs. *arXiv preprint arXiv:2205.00303*.